การลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร

PHRASE-TREE-BASED FEATURE REDUCTION FOR TEXT CATEGORIZATION

ปริสุทธิ์ จิตต์ภักดี
PARISUT JITPAKDEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2550

KMITL-2007-IT-M-001-128

การลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร

PHRASE-TREE-BASED FEATURE REDUCTION FOR TEXT CATEGORIZATION

ปริสุทธิ์ จิตต์ภักดี

PARISUT JITPAKDEE

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2550

KMITL-2007-IT-001-128

PHRASE-TREE-BASED FEATURE REDUCTION FOR TEXT CATEGORIZATION

PARISUT JITPAKDEE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2007

KMITL-2007-IT-001-128

หัวข้อวิทยานิพนธ์	การลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร
นักศึกษา	นางสาวปริสุทธิ์ จิตต์ภักดี
รหัสประจำตัว	47066405
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2550
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้ นำเสนอวิธีการและอัลกอริทึมใหม่ในการลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร ในการจำแนกประเภทเอกสาร เอกสารที่เรารู้ประเภทของพวกมันแล้ว จะถูกแทนให้อยู่ในรูปของเวกเตอร์ของฟีเจอร์ ฟีเจอร์เหล่านี้เป็นกุญแจสำคัญของประสิทธิภาพในการจำแนกประเภทเอกสาร เซตของฟีเจอร์ที่ดีจะต้องให้ความถูกต้องในการจำแนกประเภทสูง และมีขนาดเล็ก โดยการใช้วลีเป็นฟีเจอร์นั้นดีกว่าการใช้คำเดี่ยว เพราะวลีสามารถให้ความหมายได้มากกว่าคำเดี่ยว ซึ่งต้นไม้วลีที่นำเสนอนี้จะสร้างฟีเจอร์วลี และแสดงถึงความสัมพันธ์ระหว่างวลีและวลีย่อยของพวกมัน ซึ่งช่วยลดความซ้ำซ้อนของฟีเจอร์ทำให้ได้ฟีเจอร์ที่มีประสิทธิภาพ จากการทดลองเบื้องต้น วิธีการที่นำเสนอนี้ให้เซตของฟีเจอร์ที่มีขนาดเล็กกว่าวิธีการแบบเดิม และยังทำให้ความถูกต้องในการจำแนกประเภทเอกสารเพิ่มขึ้นอีกด้วย

Thesis Title	Phrase-Tree-based Feature Reduction for Text Categorization
Student	Ms. Parisut Jitpakdee
Student ID.	47066405
Degree	Master of Science
Programme	Information Technology
Year	2007
Thesis Advisor	Assoc.Prof.Dr.Worapoj Kreesuradej

ABSTRACT

This thesis proposes a new method and an algorithm for feature reduction based on phrase tree for text categorization. In text categorization, the documents that we know their categories will be represented as vectors of features. These features are the key of the performance of text categorization. The good features set must give the high accuracy of categorization and have small size. Using phrases as features is better than using words, because phrase gives the meaning more than single word. This proposed phrases tree generates phrase features and represents the relationship of these phrases and their sub phrases that help to remove redundant of feature to make the efficient features set. From preliminary experiments, the proposed method gives smaller set of features than conventional methods. Furthermore, the classification model obtained from the propose method provides higher accuracy.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยการให้คำปรึกษาพร้อมทั้งคำชี้แนะแนวทางการแก้ปัญหาของงานวิจัยตั้งแต่เริ่มต้นจนเสร็จสมบูรณ์จาก รศ.ดร.วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์ของท่านและขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณคุณย่าและญาติพี่น้องทุกท่าน ที่คอยสนับสนุนและคอยให้กำลังใจมาโดยตลอด

ขอขอบคุณคณาจารย์คณะเทคโนโลยีสารสนเทศทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ต่างๆ เพื่อนำไปใช้ประโยชน์ในการทำวิจัย

ขอขอบคุณเจ้าหน้าที่คณะเทคโนโลยีสารสนเทศทุกท่าน ที่ให้คอยให้ความสะดวกในการทำงาน

ขอขอบคุณพี่ๆ และเพื่อนๆ ห้องปฏิบัติการ Data Mining & Data Exploration Laboratory ที่คอยให้ความช่วยเหลือ คำแนะนำ และให้กำลังใจซึ่งกันและกันตลอดมา

อนึ่ง งานวิจัยที่นำเสนอในงานวิทยานิพนธ์ฉบับนี้นั้น ส่วนหนึ่งกระทำภายใต้ห้องปฏิบัติการ Data Mining & Data Exploration Laboratory คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

คุณค่าและประโยชน์จากวิทยานิพนธ์นี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ปริสุทธิ์ จิตต์ภักดี

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญรูป

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันมีการใช้คอมพิวเตอร์ และอินเตอร์เน็ตกันอย่างแพร่หลาย ทำให้ข้อมูลที่อยู่ในรูปแบบอิเล็กทรอนิกส์มีปริมาณเพิ่มขึ้นอย่างรวดเร็ว โดยเฉพาะข้อมูลที่อยู่ในรูปแบบเอกสารหรือข้อความ เช่น ข้อมูลงานวิจัย ข้อมูลข่าว และหนังสืออิเล็กทรอนิกส์ ดังนั้นจึงมีความต้องการในการจัดการกับข้อมูลเหล่านี้ เพื่อให้สามารถนำมาใช้ประโยชน์ได้อย่างมีประสิทธิภาพ

การจำแนกประเภทเอกสาร (Text Categorization หรือ Text Classification) เป็นการจัดเอกสารตามลักษณะเนื้อให้สอดคล้องกับประเภทเอกสารต่างๆ ที่มีอยู่ ทำให้เอกสารสามารถถูกจัดเก็บอย่างมีระบบ สืบค้นได้ง่าย ซึ่งการจำแนกประเภทเอกสารนั้นยังได้ถูกนำมาประยุกต์ใช้ในงานต่างๆ เช่น การกรองเอกสาร กลไกการค้นหาหัวข้อ การจำแนกประเภทเว็บ การสร้างดัชนีแบบอัตโนมัติในระบบการเทคนิคค้นคืนสารสนเทศ

ส่วนที่สำคัญอย่างหนึ่งในการจำแนกประเภทเอกสาร คือ การแทนเอกสาร (Document Representation) ซึ่งทำให้เอกสารอยู่ในรูปแบบที่จะถูกจำแนกประเภทต่อไป การแทนเอกสารที่ใช้กันอย่างแพร่หลายคือ การแทนเอกสารโดยใช้คำเดี่ยว (Single words) เป็นฟีเจอร์ (Feature) เรียกว่า Bags of word [1] ซึ่งแต่ละเอกสารจะถูกแทนเป็นเวกเตอร์ โดยแต่ละไอเทมในเวกเตอร์คือค่าที่เป็นตัวเลขสอดคล้องกับคำนั้นๆ ซึ่งอาจจะเป็นความถี่ของคำในเอกสารนั้น หรือเป็นค่าไบนารี หรือเป็นค่าน้ำหนักต่างๆ โดยเวกเตอร์ของเอกสารทั้งหมด เรียกว่า Vector Space Model [2] ซึ่งมีขนาดหรือมิติคือ จำนวนเอกสาร x จำนวนฟีเจอร์

การใช้คำเดี่ยวนี้ มีข้อด้อยหลายอย่างด้วยกัน ยกตัวอย่างเช่น ทำให้ขาดความสัมพันธ์ระหว่างคำ เช่น ลำดับของเทอม และคำๆ หนึ่งยังสามารถมีได้หลายความหมายถ้าปรากฏในเนื้อหาที่ต่างกัน วิธีหนึ่งในการแก้ปัญหาดังกล่าว คือ การใช้วลี (Phrases) เป็นฟีเจอร์แทนการใช้คำเดี่ยว วลีสามารถลดปัญหาความไม่แน่นอนของความหมายของคำเดี่ยวได้ ยกตัวอย่างเช่น คำว่า “Java” มี 2 ความหมายด้วยกัน คือ ภาษาโปรแกรมชนิดหนึ่ง และเป็นชื่อเกาะของประเทศอินโดนีเซีย แต่ถ้าเราใช้วลีเป็น “Java Script” หรือ “Java Island” ก็จะได้ความหมายที่ชัดเจนมากกว่า

ดังนั้น จึงมีการนำวลีมาใช้เป็นฟีเจอร์ในการแทนเอกสาร แต่ในการศึกษาการจำแนกประเภทเอกสารโดยใช้วลีเป็นฟีเจอร์เพียงอย่างเดียวนั้น ผลลัพธ์ที่ได้กลับมีความถูกต้องลดลงน้อยกว่าการใช้คำเดียว ซึ่งเหตุผลหลักก็คือ มีจำนวนวลีต่างๆ มาก แต่ว่าแต่ละวลีกลับมีความถี่น้อยมาก และวลีดังกล่าวยังมีความซ้ำซ้อนกันอยู่มาก คือมีวลีจำนวนมากที่มีความหมายเหมือนกัน จึง

มีผู้นำคำเดี่ยวและวลี มาใช้เป็นฟีเจอร์ร่วมกัน ซึ่งช่วยเพิ่มความถูกต้องในการจำแนกประเภทเอกสารมากขึ้น

วลีที่ใช้ในเป็นฟีเจอร์นั้นแบ่งได้ 2 แบบ คือ วลีทางภาษา (Syntactic Phrase) วลีในลักษณะนี้ จะถูกต้องตามหลักไวยากรณ์ของภาษา และใช้วิธีการทางภาษาธรรมชาติ (NLP) ในการแยกวลีออกมา ส่วนอีกแบบหนึ่งคือ วลีทางสถิติ (Statistical Phrase) หมายถึง เซต หรือ ลำดับของคำใดๆ ที่ปรากฏต่อเนื่องกันในเอกสาร โดยมีค่านัยสำคัญทางสถิติ ตัวอย่างวลีประเภทนี้เช่น การสร้าง n-gram ซึ่งคือ ลำดับของคำใด ๆ ที่ปรากฏต่อเนื่องกันในเอกสาร

ปัญหาของวลีทางภาษา คือ ในการสกัดวลี (Phrase Extraction) ทางภาษานั้นทำได้ยากกว่าวลีทางสถิติ และวลีทางภาษาที่ถูกตามหลักไวยากรณ์นั้นจะมีความหมายแคบเกินไป เช่น วลี "*k-nearest neighbor categorization algorithm*" จะปรากฏในแต่ละเอกสารน้อยมาก แต่ถ้าเราใช้วลีย่อยเป็น "*k-nearest neighbor*" หรือ "*categorization algorithm*" จะมีการกระจายของค่าในการแทนเอกสารดีกว่า เพราะมีโอกาสในการปรากฏมากกว่าในแต่ละเอกสารทำให้ค่าที่ถูกแทนในการแทนเอกสารมีค่าน้อย แต่ถ้าวลีนั้นสั้นไป ก็จะเกิดปัญหาคล้ายกับการใช้คำเดี่ยว คือ มีความสามารถในการแบ่งแยกประเภทต่ำ โดยเฉพาะอย่างยิ่งเมื่อประเภทของข้อมูลมีความใกล้เคียงกันมาก เช่น วลี "*drama player voting*" อยู่ในข่าวบันเทิง ส่วนวลี "*football player voting*" อยู่ในข่าวกีฬา ถ้าเราใช้ฟีเจอร์เป็น "*player voting*" ก็ทำให้ประสิทธิภาพในการจำแนกว่าอยู่ในข่าวประเภทไหนลดลง

การใช้วลีทางสถิติเป็นการแก้ปัญหาอย่างหนึ่งของการใช้วลีทางภาษา เพราะไม่ต้องคำนึงถึงหลักไวยากรณ์ทางภาษา เช่น การสร้าง n-gram ซึ่งคือ ลำดับของคำใด ๆ ที่ปรากฏต่อเนื่องกันในเอกสาร โดยส่วนใหญ่จะใช้ n=2 เรียกว่า Bi-gram แต่ปัญหาก็คือ วลีที่สร้างได้มีจำนวนมาก ที่เป็นวลีขยะ คือ วลีที่มีอำนาจในการจำแนกประเภทต่ำ และในการสร้าง n-gram นั้นจำเป็นที่จะต้องกำหนดขนาดสูงสุดของวลีไว้ คือ n ทำให้บางครั้งมีวลีที่มีความสามารถในการจำแนกประเภทที่ดี แต่มีขนาดมากกว่า n ก็จะไม่ถูกใช้

วิธีการแก้ปัญหาที่กล่าวมาก็คือ จะต้องมีวิธีการที่ดีในการสร้างวลี และต้องทำการเลือกเอาเฉพาะวลีที่สามารถแบ่งแยกประเภทของเอกสารได้ดีเท่านั้น โดยทำการคัดเลือกฟีเจอร์ (Feature Selection หรือ Feature Reduction) [3] เพื่อช่วยลดขนาดของฟีเจอร์ที่ใช้ และช่วยเพิ่มประสิทธิภาพการจำแนกประเภทเอกสาร

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

วิทยานิพนธ์ฉบับนี้ มีวัตถุประสงค์เพื่อพัฒนาวิธีการ และอัลกอริทึมในการสร้างฟีเจอร์วลีและเลือกวลี เพื่อลดขนาดของฟีเจอร์วลีที่จะถูกนำมาใช้แทนเอกสาร สำหรับการจำแนกประเภทเอกสาร โดยประยุกต์ใช้ทฤษฎีการสร้างฟีเจอร์ และการคัดเลือกฟีเจอร์ ร่วมกับโครงสร้างข้อมูล

แบบต้นไม้ ที่ช่วยทำให้การจำแนกประเภทเอกสารมีประสิทธิภาพดียิ่งขึ้น ทั้งในแง่ของขนาดโดยการลดมิติเพื่อการทำงานที่เร็วขึ้น และในแง่ของความถูกต้องในการจำแนกประเภทเอกสาร

1.3 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

งานวิจัยนี้ได้นำทฤษฎีและเทคนิคต่างๆมาประยุกต์ใช้ประกอบด้วย

1. เทคนิคการจำแนกประเภทเอกสาร ตั้งแต่การเตรียมเอกสาร การสร้างแบบจำลองในการจำแนกประเภทไปจนถึงการวัดประสิทธิภาพ
2. เทคนิคการแทนเอกสาร
3. เทคนิคการเลือกฟีเจอร์
4. เทคนิคการดำเนินการโครงสร้างข้อมูลแบบต้นไม้

1.4 ขอบเขตการวิจัย

ในการวิจัยนี้จะออกแบบและพัฒนาวิธีการและอัลกอริทึมที่ใช้โครงสร้างต้นไม้ ร่วมกับเทคนิคการเลือกฟีเจอร์ เพื่อคัดเลือกฟีเจอร์ที่เป็นวลี ที่มีประสิทธิภาพ และลดจำนวนของฟีเจอร์ที่เป็นวลีที่ใช้ในการแทนเอกสาร

1.5 ขั้นตอนของการศึกษา

ในขั้นตอนของการศึกษานี้ ได้แสดงลำดับการทำงานตั้งแต่เริ่มต้นจนถึงสิ้นสุดการทำงานดังรายละเอียดต่อไปนี้

1.5.1 ศึกษาทฤษฎีและงานวิจัยจากเอกสารบทความต่างๆ ที่เกี่ยวข้องกับการทำงานวิจัย

1.5.2 กำหนด หัวข้อ วัตถุประสงค์ และขอบเขตการทำงานวิจัย

1.5.3 ออกแบบและวิเคราะห์อัลกอริทึมใหม่

1.5.4 พัฒนาโปรแกรม โดยใช้ซอฟท์แวร์ MATLAB รุ่น 6.5 พร้อมทั้งแก้ไขข้อผิดพลาด และทดสอบการทำงานของอัลกอริทึม

1.5.5 เตรียมข้อมูลที่ใช้งานจริง เพื่อนำมาทดสอบการทำงานของอัลกอริทึม

1.5.6 ทดลองกับข้อมูลที่ใช้งานจริง พร้อมทั้งวัดประสิทธิภาพการทำงานของอัลกอริทึม

1.5.7 รวบรวมผลการทดลองจากการทำงานของอัลกอริทึม

1.5.8 วิเคราะห์และสรุปผลการทดลอง

1.5.9 เรียบเรียงเอกสารประกอบวิทยานิพนธ์

งานวิจัยนี้มุ่งเน้นศึกษาในส่วนของฟีเจอร์ที่เป็นวลีที่ใช้ในการแทนเอกสาร โดยพยายามสร้าง และเลือกใช้ฟีเจอร์ที่เป็นวลี เฉพาะที่มีประสิทธิภาพเพื่อลดขนาดของฟีเจอร์วลีที่ใช้ และนำโครงสร้างต้นไม้มาประยุกต์ใช้ในการทำงาน

ในบทที่ 2 จะกล่าวถึง ทฤษฎีต่างๆ ที่เกี่ยวข้องกับการทำงานวิจัยที่นำเสนอนี้

บทที่ 2
ทฤษฎีพื้นฐาน และงานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้จะกล่าวถึงทฤษฎีพื้นฐานต่างๆ ที่เกี่ยวข้องในการวิจัย พื้นฐานของ การจำแนกประเภทเอกสาร การลดฟีเจอร์ และแสดงถึงการจำแนกประเภทเอกสารโดยใช้ฟีเจอร์ในลักษณะต่างๆ พร้อมทั้งแสดงถึงงานวิจัยที่เกี่ยวข้อง และปัญหาที่เกิดขึ้น

2.1 การจำแนกประเภทเอกสาร (Text Categorization)

การจำแนกประเภทเอกสาร คือ การจัดเอกสารไปยังประเภทต่างๆ ตามลักษณะของเนื้อหานั้น ๆ ว่าสอดคล้องกับเอกสารในประเภทไหนที่มีอยู่ โดยการจำแนกประเภทเป็นงานหนึ่งของดาต้าไมนิง (Data Mining) ที่ทำการค้นหาสารสนเทศที่มีอยู่จากข้อมูลที่มีปริมาณมาก ซึ่งการจำแนกประเภทอยู่ในส่วนของการทำนาย (Prediction) โดยหมายถึง การนำตัวอย่างจากประสบการณ์ที่ผ่านมาที่เรารู้คำตอบ แน่ชัดแล้ว มาใช้ในกรณีในอนาคต [24] ในการจำแนกประเภทตัวอย่างที่เรารู้แน่ชัดแล้วก็คือ เอกสารที่ถูกให้ป้ายชื่อของประเภทไว้ (Label) การจำแนกประเภทแตกต่างจากการจัดกลุ่ม (Clustering) ตรงที่การจัดกลุ่ม เราไม่รู้เลยว่าข้อมูลอยู่ในปรเภทไหน และแบ่งได้กี่ประเภท

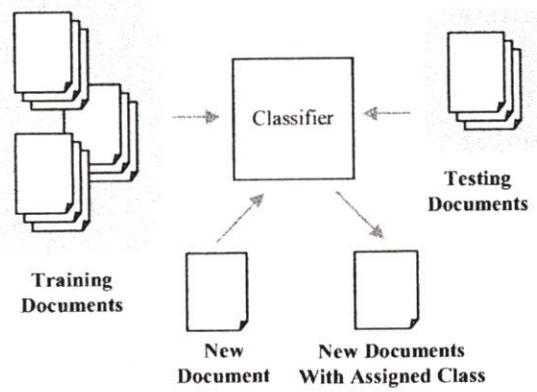
Sebastiani [4] ได้นิยามการจำแนกประเภทเอกสารไว้ว่า การจำแนกประเภทเอกสาร คือ งานของการให้ค่าบูลีนกับแต่ละคู่ระหว่างเอกสารกับคลาส $(d_j, c_i) \in D \times C$ โดย D คือ เซตของเอกสาร และ C คือ เซตของคลาส

การจำแนกประเภทถูกนำมาประยุกต์ใช้ในหลายๆ งานด้วยกัน [4] เช่น

1. Document Organization: เช่น ในงานหนังสือพิมพ์ มีความจำเป็นที่จะต้องการจำแนกประเภทของข่าวที่ผ่านมา เพื่อความง่ายในการค้นหาในอนาคต ในกรณีที่เหตุการณ์ที่เกิดขึ้นใหม่มีเกี่ยวข้องกับข่าวที่ผ่านมา ประเภทที่เป็นไปได้ เช่น "*Sport*", "*Entertainment*"
2. การกรองเอกสาร (Text Filtering): เช่น ในการกรอง e-mail โดยแบ่ง e-mail เป็น 2 ประเภท คือ e-mail ที่เราสนใจ กับ e-mail ที่เราไม่สนใจ หรือ e-mail ที่เป็นขยะ กับ e-mail ที่ไม่เป็นขยะ
3. การแบ่งประเภทแบบมีลำดับชั้นของเว็บเพจ

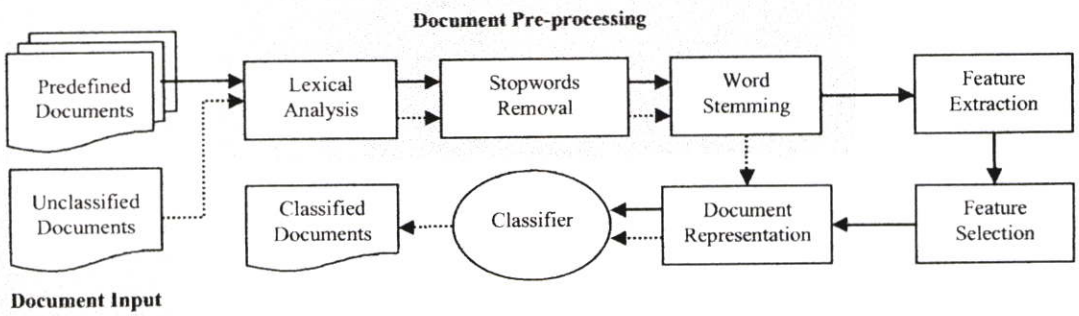
การจำแนกประเภทเอกสาร ประกอบด้วยกระบวนการหลักๆ ดังนี้ คือ การเตรียมเอกสาร การแทนเอกสาร การสร้างตัวจำแนกประเภท (Classifier) ซึ่งในขั้นตอนนี้ ชุดเอกสารที่อยู่ในประเภทต่างๆ จะถูกแบ่งเป็น 2 ส่วน เพื่อใช้ในการเรียนรู้ในการสร้างตัวจำแนกประเภท และอีกชุด

เพื่อใช้ในการทดสอบตัวจำแนกประเภท โดยการวัดประสิทธิภาพในการจำแนกประเภทเอกสารของตัวจำแนกประเภท และกระบวนการสุดท้ายก็คือ นำเอกสารใหม่มาจำแนกประเภทโดยใช้ตัวจำแนกประเภทที่สร้างไว้ ดังแสดงในรูปที่ 2.1



รูปที่ 2.1 การจำแนกประเภทเอกสาร

กระบวนการในการจำแนกประเภทเอกสารแสดงดังรูปที่ 2.2 เส้นทึบแทนการไหลของข้อมูลในขั้นตอนการสร้างตัวจำแนกประเภท โดยเอกสารที่เรารู้คลาสแล้วจะผ่านกระบวนการเตรียมเอกสาร จากนั้นผ่านกระบวนการการสกัดและเลือกฟีเจอร์ และการแทนเอกสาร เพื่อสร้างเป็นตัวจำแนกประเภท ส่วนเส้นประหมายถึงเส้นทางการไหลของข้อมูลในขั้นตอนการจำแนกประเภท โดยเมื่อมีเอกสารที่เรายังไม่ทราบคลาสเข้ามา ก็จะผ่านกระบวนการเตรียมเอกสาร และการแทนเอกสาร เพื่อนำเข้าสู่ตัวจำแนกประเภท ทำให้ทราบประเภทของเอกสารนั้น



รูปที่ 2.2 กระบวนการในการจำแนกประเภทเอกสาร

2.2 การเตรียมเอกสาร (Document Preprocessing)

ขั้นตอนแรกของการจำแนกประเภทเอกสาร (Text Categorization) คือการเตรียมเอกสารให้อยู่ในรูปแบบที่จะนำเข้าสู่ขั้นตอนการจำแนกประเภท ได้ ซึ่งจะมีการดำเนินการเปลี่ยนรูปแบบเอกสาร (Text Transformations หรือ Text Operations) [6] 3 ลักษณะ ดังต่อไปนี้

1. การวิเคราะห์คำ (Lexical Analysis) หรือ โทเคนไนเซชัน (Tokenization): เอกสารจะถูกพิจารณาเป็นสายอักขระ (String) อักขระที่ไม่ใช่ตัวอักษร เช่น ช่องว่าง (Space) บรรทัดใหม่ (New Line) แท็บ (Tab) และเครื่องหมายวรรคตอนต่างๆ จะถูกเอาออก โดยเก็บส่วนที่เหลือไว้ และเปลี่ยนให้เป็นตัวพิมพ์เล็ก (Lower Case) ให้หมด รวมทั้งการเอาแท็ก (Tags) ต่างๆ ของพวกเอกสารที่เป็นเว็บออกด้วย

2. การกำจัด Stopwords (Elimination of Stopwords) คือ การตัดคำที่มีอำนาจในการแบ่งแยกได้น้อยออกไป หรือ Function Word เช่น คำนำหน้า (Articles) เช่น a, an, the คำบุพบท (Prepositions) เช่น in, on, of คำสันธาน (Conjunctions) เช่น and, or, but, if คำสรรพนาม (Pronouns) เช่น I, you, them, it คำต่างๆ ที่เป็นไปได้ เช่น make, thing, similar, all, already, always, nobody โดยรายการคำหยุด สามารถดาวน์โหลดได้ที่ [22]

3. การทำให้อยู่ในรูปแบบของรากคำ (Word Stemming) คือ การแทนที่รูปแปร (Variant) ต่างๆ ของคำ ด้วยรากของคำเพียงคำเดียว รูปแปรประกอบด้วย พหูพจน์ รูปแบบของคำกริยาที่เติมท้ายด้วย “ing” หน่วยคำเติมหลังบุคคลที่สาม หน่วยคำเติมหลังอดีตกาล (Past Tense) เป็นต้น ตัวอย่างเช่น คำว่า connect เป็นรากของคำต่อไปนี้ connects, connected, connecting, connection เป็นต้น ซึ่งคำต่างๆ เหล่านี้มีความหมายที่คล้ายคลึงกัน โดยมาตรฐานที่ใช้ส่วนใหญ่ คือ อัลกอริทึมการสเต็มมิงของพอร์ตเตอร์ (Porter’s Stemming Algorithm) [7]

2.3 การสกัดฟีเจอร์ (Feature Extraction)

การสกัดฟีเจอร์ คือ การเลือกว่าเราจะใช้เทอมอะไรที่เหมาะสมเป็นฟีเจอร์ เพื่อใช้แทนเอกสาร หรือเรียกว่าการทำดัชนี (Indexing) โดยฟีเจอร์ ที่เราใช้ อาจจะเป็นคำเดี่ยว หรือ วลี โดยในการทำดัชนีนั้นมี 2 วิธีด้วยกัน [4] คือ

1. Derived Indexing: ฟีเจอร์จากวิธีนี้จะได้มาจากเนื้อหาของเอกสารโดยตรง โดยฟีเจอร์อาจจะเป็น คำ วลี หรือ n-grams เป็นต้น
2. Assigned Indexing: ฟีเจอร์จากวิธีนี้จะได้มาจากข้อมูลภายนอก เช่น รายการของคำศัพท์ หรือมีการขยายคำที่ใช้เป็นฟีเจอร์ด้วยคำพ้อง (Synonym) จาก WordNet [23] เป็นต้น

โดยฟีเจอร์ที่ใช้กันโดยทั่วไป คือ คำ และวลี โดยมีรายละเอียดดังนี้

2.3.1 คำเดี่ยว (Single Word)

การใช้คำเป็นฟีเจอร์นี้นิยมใช้กันโดยทั่วไป โดยจะเรียกเซตของฟีเจอร์ที่เป็นคำนี้ว่า Bag of Word (BOW) [1] ซึ่งเป็นการใช้คำสำคัญ (Single Word) ที่ได้จากแต่ละเอกสารมาเป็นฟีเจอร์ ใน

เอกสารภาษาอังกฤษซึ่งเป็นภาษาที่เราใช้ในงานวิจัยนี้ สามารถสกัดคำออกมาโดยง่าย เพราะคำในภาษาอังกฤษจะถูกแบ่งแยกด้วยช่องว่างหรือเครื่องหมายต่างๆ ซึ่งต่างจากภาษาไทยที่เราไม่สามารถรู้ขอบเขตของคำอย่างชัดเจนต้องอาศัยพจนานุกรมมาช่วย

2.3.2 วลี (Phrase)

Bag of Phrase (BOP)ใช้วลี (Phrase) ซึ่งคือลำดับของคำ หรือกลุ่มของคำ (Multi-word) เป็นฟีเจอร์ โดยวลีที่ใช้ในการแทนเอกสาร แบ่งได้ 2 ลักษณะ คือ

1. วลีทางภาษา (Syntactic Phrase) วลีในลักษณะนี้ จะถูกต้องตามหลักไวยากรณ์ของภาษา และถูกกำหนดโดยใช้วิธีการทางภาษาธรรมชาติ (NLP) ต่างๆโดย วลีในที่นี้หมายถึง หน่วยของเนื้อหาที่โดยทั่วไปจะใหญ่กว่าคำ แต่เล็กกว่าประโยคเต็ม ตัวอย่างของวลีที่เป็นคำนาม เช่น "*nuclear waste disposal*" "*the dog crossed the street*" และ "*Bill Gates*" ส่วนตัวอย่างของวลีที่เป็นกริยา เช่น "*playing ice hockey*" และ "*went to school*" เป็นต้น

2. วลีทางสถิติ (Statistical Phrase) หมายถึง เซต หรือ ลำดับของคำใดๆ ที่ปรากฏต่อเนื่องกันในเอกสาร โดยมีค่านัยสำคัญทางสถิติ ตัวอย่างวลีประเภทนี้เช่น การสร้าง n-gram ซึ่งคือลำดับของคำใด ๆ ที่ปรากฏต่อเนื่องกันในเอกสาร เช่น มีเนื้อหาคือ "*the dog crossed the street*" หลังจากผ่านการเตรียมข้อมูล และตัด Stopwords ออกแล้ว จะได้เนื้อหาเป็น "*dog cross street*" สร้าง Bi-gram (n=2) จะได้ฟีเจอร์ที่เป็นวลี คือ "*dog cross*" และ "*cross street*"

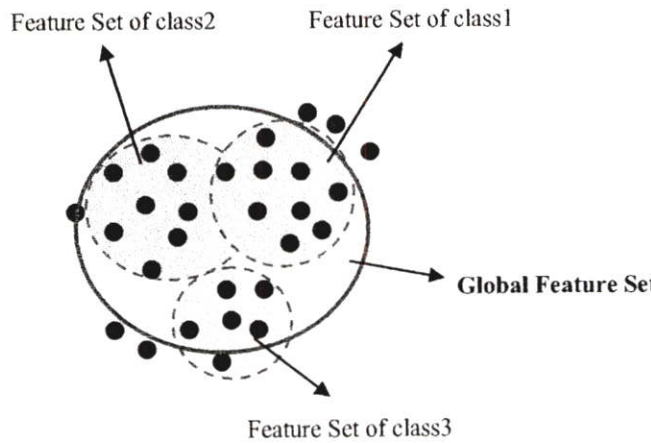
2.4 การเลือกฟีเจอร์ (Feature Selection)

การเลือกฟีเจอร์ มีวัตถุประสงค์คือ ลดจำนวนฟีเจอร์ที่จะใช้แทนเอกสาร เพื่อเพิ่มประสิทธิภาพในการจำแนกประเภทเอกสาร โดยการเลือกฟีเจอร์จะทำโดยเลือกเซตย่อยของฟีเจอร์มาใช้ หรือเรียกได้อีกอย่างหนึ่งว่า Term Space Reduction โดยให้ T คือเซตของเทอมเริ่มต้น การลดฟีเจอร์คือเลือกเซตย่อย T' โดย $|T| \leq |T'|$ [4]

การลดฟีเจอร์มี 2 ระดับ คือ

1. Local Feature Reduction: วิธีนี้จะเลือกเซตย่อยของเทอมภายใต้คลาสต่างๆ โดยเลือกทีละคลาส c_i นั่นคือ ในคลาสต่างๆ จะได้จำนวนฟีเจอร์ไม่เท่ากัน
2. Global Feature Reduction: วิธีนี้จะเลือกเซตย่อยของเทอมภายใต้คลาสทั้งหมด $C = \{c_1, ..., c_{|C|}\}$

รูป 2.3 แสดงการลดฟีเจอร์ทั้ง 2 ระดับ คือ ระดับ Local ที่มีคลาส 3 คลาสด้วยกัน และระดับ Global



รูปที่ 2.3 การเลือกฟีเจอร์ในระดับ Local และ Global

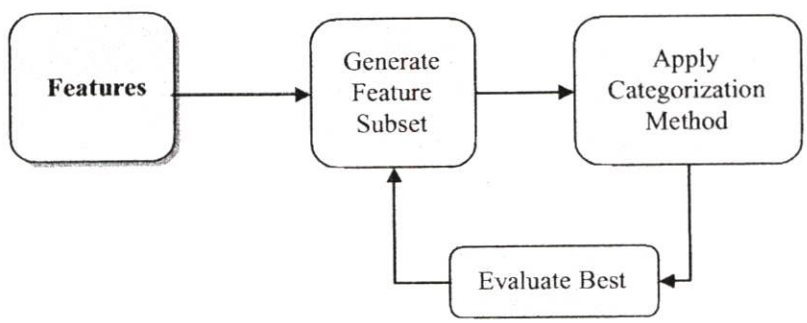
ส่วนวิธีในการเลือกฟีเจอร์นั้นมี 2 แนวทางด้วยกัน คือ

2.4.1 แรพเพอร์ (Wrapper)

ในแนวทางนี้ เซตย่อยของฟีเจอร์ถูกเลือกโดยดูจากความถูกต้องของตัวจำแนกประเภท และนำอัลกอริทึมที่ใช้ในการจำแนกประเภทจริงๆ มาใช้คิดคำนวณ แนวทางนี้เกิดจากข้อเท็จจริงที่ว่า อัลกอริทึมการเรียนรู้ในการการจำแนกประเภทที่ต่างกัน ได้รับผลกระทบต่างกันจากการเลือกฟีเจอร์ [19]

โดยเริ่มจากเซตของเทอมเริ่มต้น เซตของเทอมใหม่ถูกสร้างขึ้นโดยการเพิ่มหรือลบเทอมออก เมื่อเซตของเทอมใหม่ถูกสร้างขึ้นแล้ว ก็ลองสร้างตัวจำแนกประเภทโดยใช้เซตของเทอมใหม่นี้เป็นฟีเจอร์ แล้วทดสอบความถูกต้อง แล้วเลือกใช้เซตของเทอมที่ให้ประสิทธิภาพดีที่สุดเป็นฟีเจอร์ (Optimal Feature Set) การเลือกเซตย่อยที่ดีที่สุดแสดงดังรูป 2.4

แต่โดยทางเทคนิควิธีการนี้ มีความยากในการอิมพลีเมนท์ โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีขนาดใหญ่มากๆ เช่นมีฟีเจอร์ทั้งหมด m ตัว จะมีเซตย่อยที่เป็นไปได้ 2^m เซต โดยเซตย่อยสามารถถูกค้นหาโดยใช้ Forward Selection หรือ Backward Elimination ซึ่งทำโดยใช้ Hill Climbing หรือใช้ Heuristic อื่นๆ



รูปที่ 2.4 การเลือกเซตย่อยของฟีเจอร์ที่ดีที่สุด [5]

2.4.2 การกรอง (Filter)

ในแนวทางนี้ จะเลือกเซตย่อยของฟีเจอร์ โดยทำการกรอง ซึ่งใช้คะแนนเป็นเกณฑ์ ซึ่งคะแนนจะบ่งบอกถึงความสำคัญของเทอม ในการจำแนกประเภทเอกสารมักจะใช้แนวทางนี้ ซึ่งอิมพลีเมนท์ง่ายกว่าแรพเพอร์ โดยฟีเจอร์จะถูกเลือกโดยใช้เกณฑ์ ต่างๆ ดังนี้

1. ความถี่ของเอกสาร (Document Frequency) : วิธีนี้ฟีเจอร์ถูกเลือกโดยดูจากจำนวนเอกสาร โดยเทอมที่ปรากฏในเอกสารส่วนใหญ่จะถูกเก็บไว้ เช่นมีการตั้งค่าจำนวนเอกสารต่ำสุดไว้ แล้วลบเทอมที่ปรากฏแค่เพียงตามจำนวนเอกสารที่ตั้งไว้ออก

2. ใช้ตัววัดจาก Information-theoretic Function (*f*) เช่น

Information Gain Measure: ค่า Information Gain $IG(t,c_i)$ ของเทอม t ในคลาส c_i หาได้ดังนี้ [8]

$$IG(t,c_i)=\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}}P(t,c_i)\cdot\log\frac{P(t,c_i)}{P(t)\cdot P(c_i)} \qquad (2.1)$$

Mutual Information Measure: ค่า Mutual Information $MI(t,c_i)$ ของเทอม t ในคลาส c_i หาได้จากสมการต่อไปนี้ [9]

$$MI(t,c_i)=\log\frac{P(t,c_i)}{P(t)\cdot P(c_i)} \qquad (2.2)$$

Odds Ratio Measure: ค่า Odds Ratio $OR(t,c_i)$ ของเทอม t ในคลาส c_i หาได้จากสมการต่อไปนี้ [10]

$$OR(t,c_i)=\frac{P(t\mid c_i)\cdot(1-P(t\mid\bar{c}_i))}{(1-p(t\mid c_i))\cdot p(t\mid\bar{c}_i)} \qquad (2.3)$$

โดย $P(t,c_i)$ คือ ความน่าจะเป็นที่คลาส c_i และเทอม t เกิดขึ้นร่วมกัน

$P(t\mid c_i)$ คือ ความน่าจะเป็นที่เทอม t จะเกิดเมื่อเกิดคลาส c_i

$P(t\mid\bar{c}_i)$ คือ ความน่าจะเป็นที่เทอม t จะเกิดเมื่อไม่เกิดคลาส c_i

$P(t)$ คือ ความน่าจะเป็นในการเกิดเทอม t

$P(c_i)$ คือ ความน่าจะเป็นของคลาส c_i

ซึ่งฟังก์ชันตัววัดที่กล่าวมาจะเป็นแบบ Local ซึ่งเราสามารถทำแบบ Global ได้โดยนำค่าที่ได้จากแต่ละคลาสมารวมกัน ซึ่งอาจเพิ่มการถ่วงน้ำหนักให้แต่ละคลาส หรือใช้ค่าสูงสุดก็ได้ ดังนี้ [4]

$$f_{sum}(t) = \sum_{i=1}^{|C|} f(t,c_i) \tag{2.4}$$

$$f_{wsum}(t) = \sum_{i=1}^{|C|} P(c_i) f(t,c_i) \tag{2.5}$$

$$f_{\max}(t) = \max_{i=1}^{|C|} f(t,c_i) \tag{2.6}$$

โดย $f(t,c_i)$ คือ ฟังก์ชันของตัววัดที่ใช้คำนวณค่าของเทอม t ในคลาส c_i

$f_{sum}(t)$ คือ ค่าที่ได้จากการรวมค่าจากฟังก์ชัน f ของเทอม t จากทุกคลาส

$f_{wsum}(t)$ คือ ค่าที่ได้จากการรวมค่าจากฟังก์ชัน f ของเทอม t จากทุกคลาสแบบถ่วงน้ำหนัก

$f_{\max}(t)$ คือ ค่าสูงสุดจากฟังก์ชัน f ของเทอม t จากทุกคลาส

2.5 การแทนเอกสาร (Document Representation)

จากฟีเจอร์ที่เราได้จากกระบวนการที่ผ่านมา ต่อไปเราจะทำการแทนเอกสารโดยใช้ฟีเจอร์เหล่านี้ โดยทั่วไปวิธีที่ใช้กันอย่างแพร่หลายคือวิธี Vector Space Model [2] ในโมเดลนี้แต่ละเอกสารจะถูกแทนโดยเวกเตอร์ d โดย d เขียนได้ดังนี้

$$d = \{w_1, w_2, ..., w_m\}$$

โดย w_i (i=1, 2, …, m) คือ ค่าน้ำหนักเทอมของเทอม t_i ที่เป็นฟีเจอร์ และ m คือ จำนวนฟีเจอร์ แต่ละค่าถ่วงน้ำหนักเทอมแทนนัยสำคัญของแต่ละเทอมในเอกสาร ซึ่งในการคำนวณค่าถ่วงน้ำหนักเทอมนี้จะพิจารณาความถี่ในการปรากฏของเทอม ภายในแต่ละเอกสาร และในเซตของเอกสารทั้งหมด วิธีการต่างๆ ในการกำหนดค่าน้ำหนัก มีดังนี้

2.5.1 Boolean Weighting

เป็นวิธีการที่ง่ายที่สุด วิธีนี้สามารถบอกได้เพียงว่ามีฟีเจอร์นั้นปรากฏในเอกสารหรือไม่ คือแทนค่า Weight ด้วย 0 และ 1 โดยจะแทนค่าเป็น 1 ถ้ามีฟีเจอร์นั้นปรากฏในเอกสาร และแทนค่าเป็น 0 เมื่อไม่มีฟีเจอร์นั้นปรากฏในเอกสาร ดังสมการที่ (2.7)

$$w_{ij} = \begin{Bmatrix} 1 & if\ tf_{ij} > 0 \\ 0 & otherwise \end{Bmatrix} \tag{2.7}$$

โดย w_{ij} คือ ค่าน้ำหนักของเทอม i ในเอกสาร j และ

tf_{ij} คือ ความถี่เทอม i ในเอกสาร j

2.5.2 Term Frequency Weighting

วิธีการนี้สามารถบอกความถี่ของฟีเจอร์ที่ปรากฏในเอกสารได้ จึงมีประสิทธิภาพดีกว่าแบบ Boolean Weighting เป็นวิธีการที่หาความถี่ของฟีเจอร์ที่ปรากฏในเอกสารว่ามีฟีเจอร์นั้นปรากฏในเอกสารกี่ครั้ง โดยยังคงใช้ 0 แทนความหมายว่าไม่มีฟีเจอร์นั้นปรากฏในเอกสาร ดังสมการที่ (2.8)

$$w_{ij} = tf_{ij} \tag{2.8}$$

2.5.3 TF-IDF Weighting

วิธีนี้เป็นที่นิยมมากที่สุด โดยการใช้ความถี่ของเทอม (Term Frequency) ร่วมกับความถี่เอกสารผกผัน (Inverse Document Frequency) หรือเรียกย่อๆ ว่า TF-IDF [11] ดังสมการที่ (2.9)

$$w_{ij} = tf_{ij} * \log\left(\frac{N}{n_i}\right) \tag{2.9}$$

โดย N คือจำนวนเอกสารทั้งหมด และ n_i คือ จำนวนเอกสารที่มีเทอม i ปรากฏอยู่

2.5.4 TFC Weighting

วิธีการนี้จะคล้ายกับ TF-IDF Weighting แต่มีการ Normalization เพราะว่าเอกสารมีขนาดไม่เท่ากัน [11] ดังสมการที่ (2.10)

$$w_{ij} = \frac{tf_{ij} * \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum_{k=1}^{M}\left[tf_{kj} * \log\left(\frac{N}{n_k}\right)\right]^2}} \tag{2.10}$$

วิธีการทั้งหมดที่กล่าวมาเป็นวิธีการในการกำหนดค่าน้ำหนักของเทอมที่ใช้เป็นฟีเจอร์ ที่เป็นตัวแทนของเอกสาร

ตัวอย่างการคำนวณค่าน้ำหนัก มีดังนี้ มีเอกสารอยู่ 3 เอกสาร ทำการเตรียมเอกสารแล้วได้เนื้อหา และเทอมดังในตารางที่ 2.1

ตารางที่ 2.1 แสดงตัวอย่างเอกสาร เนื้อหาและเทอมที่ได้หลังการเตรียมเอกสาร

document	text	terms
*d*1	*web web graph*	*web graph*
*d*2	*graph web net graph net*	*graph web net*
*d*3	*page web complex*	*page web complex*

จากตัวอย่างในตาราง 2.1 เทอมที่ได้จากแต่ละเอกสารจะถูกนำมาใช้เป็นฟีเจอร์ สมมติในที่นี้ไม่ได้ทำการเลือกฟีเจอร์ คือใช้ทุกเทอมที่ได้ ดังนั้นเราจะได้เซตของฟีเจอร์ ดังนี้

Features = {*"web", "graph", "net", "page", "complex"*}

ในการให้ค่าน้ำหนักแบบบูลีน เราจะได้เวกเตอร์ของเอกสาร ดังนี้

$$d_1 = \{1, 1, 0, 0, 0\}$$

$$d_2 = \{1, 1, 1, 0, 0\}$$

$$d_3 = \{1, 0, 0, 1, 1\}$$

ส่วนในการให้ค่าน้ำหนักโดยใช้ความถี่ เราจะได้เวกเตอร์ของเอกสารเป็น

$$d_1 = \{2, 1, 0, 0, 0\}$$

$$d_2 = \{1, 2, 2, 0, 0\}$$

$$d_3 = \{1, 0, 0, 1, 1\}$$

ในการให้ค่าน้ำหนักโดยใช้ TF-IDF เราจะได้เวกเตอร์ของเอกสาร ดังนี้

$$d_1 = \{0, 0.4055, 0, 0, 0\}$$

$$d_2 = \{0, 0.8109, 2.1972, 0, 0\}$$

$$d_3 = \{0, 0, 0, 1.0986, 1.0986\}$$

ซึ่งจะสังเกตได้ว่าคำที่ปรากฏทุกเอกสารจะถูกให้ค่าน้ำหนักเท่ากับศูนย์ เพราะไม่สามารถใช้แบ่งแยกเอกสารได้

และในการให้ค่าน้ำหนักโดยใช้ TFC เราจะได้เวกเตอร์ของเอกสารเป็น

$$d_1 = \{0, 1, 0, 0, 0\}$$

$$d_2 = \{0, 0.3462, 0.9381, 0, 0\}$$

$$d_3 = \{0, 0, 0, 0.7071, 0.7071\}$$

2.6 วิธีการจำแนกประเภทเอกสาร (Categorization Method)

หลายๆ วิธีการในการจำแนกประเภทได้ถูกนำมาใช้ในการจำแนกประเภทเอกสาร เช่น Naïve Bayes (NB) Classifier, Decision Tree Classifier, Neural Network และ k-Nearest Neighbor Classifier เป็นต้น

ในงานวิจัยนี้เราได้ใช้ k-Nearest Neighbor Classifier ในการจำแนกประเภทเอกสาร ซึ่งมีรายละเอียดดังต่อไปนี้

การจำแนกประเภทแบบ k-Nearest Neighbor (k-NN) [5] เป็นวิธีการเรียนรู้จากกรณีตัวอย่าง (Instance-based Learning Method) ซึ่งมีการเก็บชุดข้อมูลที่จะใช้เอาไว้ก่อน จนกว่าจะมี

ข้อมูลใหม่เข้ามา จึงจะเริ่มกระบวนการจำแนกประเภท ทำให้มันถูกเรียกว่าเป็นการเรียนรู้แบบขี้เกียจ (Lazy Learning) การจำแนกประเภทแบบ k-NN นี้ ประสิทธิภาพจะขึ้นอยู่กับตัววัดความคล้ายคลึงที่ใช้ (Similarity Measure)

ขั้นตอนแรกในการจำแนกโดยวิธี k-NN มัน จะคำนวณความคล้ายคลึงระหว่างเอกสารใหม่กับเอกสารทุกตัวที่เรารู้ว่าอยู่ประเภทไหนแล้ว และเลือกเอกสารที่มีค่าความคล้ายคลึงสูงสุดมา k ตัว จากนั้นก็จะทำการเลือกประเภทจาก k เอกสารนั้นให้กับเอกสารใหม่ ซึ่งกระบวนการจำแนกประเภทเอกสารแบบ k-NN เขียนเป็นอัลกอริทึมได้ดังรูป 2.5

Algorithm: kNN Text Categorization

Input: D = Vectors of documents, k, new document

Output: Class of new document

For all Document in D

Compute Similarity between

each document and new document

End

Select the k documents that are most similar to new document

Select class of new document from that k documents

รูปที่ 2.5 อัลกอริทึมการจำแนกประเภทเอกสารแบบ kNN

ซึ่งในการเลือกประเภทเอกสารให้เอกสารใหม่จากทั้ง k เอกสาร (kNN) นั้น มี 2 วิธี ที่ใช้กันทั่วไป ดัง 2 สมการต่อไปนี้

$$C(d_i) = \arg\max_{i=1}^{|C|} \sum_{x_j \in kNN} y(x_j, c_i) \tag{2.13}$$

$$C(d_i) = \arg\max_{i=1}^{|C|} \sum_{x_j \in kNN} sim(d_i, x_j) \cdot y(x_j, c_i) \tag{2.14}$$

โดย d_i คือ เอกสารใหม่ที่จะทำการจำแนกประเภท

x_j คือ หนึ่งในเอกสารใกล้เคียงที่เลือกมา k ตัว

$sim(d_i, x_j)$ คือ ความคล้ายคลึงระหว่างเอกสาร d_i กับเอกสาร x_j

$y(x_j, c_i) \in \{0,1\}$ หมายถึงว่าเอกสาร x_j นั้นอยู่ในคลาส c_i หรือไม่ ถ้า มีค่าเท่ากับ 1 แสดงว่าเอกสาร x_j นั้นอยู่ในคลาส c_i แต่ถ้ามีค่าเท่ากับ 0 แสดงว่าเอกสาร x_j ไม่ได้อยู่ในคลาส c_i

สมการที่ (2.13) หมายถึงว่า เอกสารใหม่นั้นจะถูกจำแนกประเภทให้อยู่ในคลาสที่มีจำนวนสมาชิกที่อยู่ใน เอกสารใกล้เคียง k เอกสารนั้นมากที่สุด ในขณะที่สมการที่ (2.14) นั้น หมายถึงว่าเอกสารใหม่นั้นจะถูกจำแนกให้อยู่ในคลาสที่มีผลรวมค่าความคล้ายคลึงของคลาสในเอกสารใกล้เคียง k เอกสารมากที่สุด

จะเห็นได้ว่าการวัดความคล้ายคลึงของเอกสารเป็นส่วนสำคัญที่เราใช้ในการจำแนกประเภทเอกสาร การวัดความคล้ายคลึงของเอกสารที่นิยมใช้ มี 2 วิธีหลักๆ คือ

1. การวัดความคล้ายคลึงโดยใช้ระยะทาง Euclidean

ให้ขนาดของเวกเตอร์ คือจำนวนเทอมเท่ากับ m ความคล้ายคลึงระหว่างเวกเตอร์เอกสาร d_1 และ d_2 หรือ sim(d_1, d_2) โดยใช้การวัดระยะทาง Euclidean [12] มีสมการดังนี้

$$sim(d_1, d_2) = \sqrt{\frac{\sum_{i=1}^{m}(w_{i1} - w_{i2})^2}{m}} \tag{2.11}$$

2. การวัดความคล้ายคลึงโดยใช้สหสัมพันธ์โคไซน์

การวัดสหสัมพันธ์โคไซน์ (Cosine Correlation Measure) [13] เป็นอีกวิธีหนึ่งที่เป็นที่นิยมใช้ในการวัดความคล้ายคลึงของเอกสาร เขียนเป็นสมการได้ดังนี้

$$sim(d_1, d_2) = \frac{\sum_{i=1}^{m} w_{i1} \cdot w_{i2}}{\sqrt{\sum_{i}^{m} {w_{i1}}^2} \cdot \sqrt{\sum_{i}^{m} {w_{i2}}^2}} \tag{2.12}$$

โดย $sim(d_1, d_2)$ คือ ความคล้ายคลึงระหว่างเอกสาร d_1 กับเอกสาร d_2

w_{i1} และ w_{i2} คือ ค่าน้ำหนักของเทอมที่ i ในเอกสาร d_1 กับเอกสาร d_2 ตามลำดับ

m คือ จำนวนฟีเจอร์ที่ใช้แทนเอกสาร

2.7 การจำแนกประเภทเอกสารโดยใช้วลี

การจำแนกประเภทเอกสารโดยใช้วลี มีวัตถุประสงค์เพื่อแก้ปัญหาที่เกิดจากการจำแนกประเภทเอกสารโดยใช้คำ ข้อเสียของการแทนเอกสารโดยการใช้คำ คือ การทำลายความสัมพันธ์เชิงความหมายระหว่างคำไป โดยวลีที่มีความหมายแน่นอน เช่น "White House" หรือ "Bill Gates" เมื่อถูกแทนโดยการใช้คำ วลีเหล่านี้ก็จะถูกแยกคำออก ทำให้ความหมายของพวกมันเสียไป ข้อเสีย

ของการใช้คำอีกประการหนึ่งคำคำหนึ่งอาจมีได้หลายความหมาย ซึ่งถ้าเป็นวลีจะบอกถึงความหมายที่แท้จริงได้มากกว่า เช่น คำว่า "Java" เป็นไปได้ 2 ความหมาย คือ ภาษาโปรแกรมมิง หรือ เกาะของประเทศอินโดนีเซีย โดยถ้าเราใช้วลี "Java Script" จะบอกความหมายที่แท้จริงคือในความหมายแรกได้

วลีที่ใช้ในการแทนเอกสาร แบ่งได้ 2 ลักษณะ คือ วลีทางภาษา (Syntactic Phrase) และ วลีทางสถิติ (Statistical Phrase) โดยในการใช้ฟีเจอร์วลีนั้น ยังมี 2 แนวทางด้วยกัน คือ

1. ใช้ฟีเจอร์ที่เป็นวลีเพียงอย่างเดียวในการแทนเอกสาร
2. ใช้ฟีเจอร์ที่เป็นวลีร่วมกับฟีเจอร์ที่เป็นคำในการแทนเอกสาร เนื่องจาก จากงานวิจัย ที่ศึกษาการใช้ Syntactic Phrase ในการแทนเอกสาร เพียงอย่างเดียว ผลที่ได้กลับทำให้ความถูกต้องในการจำแนกประเภทเอกสารน้อยกว่าการใช้คำเป็นฟีเจอร์เพียงอย่างเดียว

งานวิจัยเรื่อง "*An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task*" ของ Lewis [24] ใช้วลีที่เป็นคำนามที่ปรากฏอย่างน้อย 2 ครั้ง ในการแทนเอกสาร สำหรับการจำแนกประเภทเอกสาร แต่ผลที่ได้กลับแย่ลงกว่าการใช้คำเป็นฟีเจอร์เพียงอย่างเดียว เขาได้ให้เหตุผลไว้ว่า

ก) เทอมที่ได้มีปริมาณมาก

ข) ค่าของฟีเจอร์กระจายไม่เท่ากัน โดยหลายๆ ค่ามีค่าต่ำมาก

ค) ฟีเจอร์มีความซ้ำซ้อนกันสูง

ง) ฟีเจอร์ประกอบด้วย Noise

Lewis ได้พยามยามแก้ปัญหาปัญหาเหล่านี้โดยใช้อัลกอริทึมการจัดกลุ่ม (Clustering) ในการรวมฟีเจอร์ที่เป็นวลี ไปเป็น Meta-features แต่ในการแทนเอกสารแบบใหม่นี้ก็ไม่ได้สร้างผลที่ดีกว่าเท่าใดนัก

ในการศึกษาเปรียบเทียบระหว่างการใช้วลีทางภาษากับวลีทางสถิติ โดย Mitra และกลุ่มได้ทำการทดลองเปรียบเทียบการใช้ระหว่าง Bi-grams กับวลีที่เป็น POS-tag Sequence (เช่น Noun-Noun, Adjective-Noun) ในงาน Information Retrieval ผลที่ได้คือ การใช้วลีทางภาษาดีกว่าเพียงเล็กน้อย ซึ่งสรุปแล้วผลที่ได้น่าจะขึ้นอยู่กับลักษณะการใช้งานและข้อมูลว่าเหมาะกับวลีแบบไหน

แต่จากผลของ Lewis ที่แสดงให้เห็นว่า การใช้วลีเพียงอย่างเดียวนั้น ไม่ได้เพิ่มประสิทธิภาพในการจำแนกเอกสาร จึงมีหลายงานวิจัย ที่ได้แนะนำว่าควรจะเพิ่มความถูกต้องของการจำแนกประเภทเอกสารโดยการใช้วลีทางสถิติร่วมกับการใช้คำเดี่ยวเป็นฟีเจอร์ เช่นงานวิจัยของ Mladeni และ Grobelink รายงานการวิจัยเรื่อง "*Word Sequences as Features in Text Learning*" [25] ได้สร้างฟีเจอร์ใหม่จากลำดับของคำ ตั้งแต่ 2-5 คำ (n-gram, n=5) โดยเลือกใช้เฉพาะฟีเจอร์ที่มีคะแนนสูง โดยในที่นี้คะแนนที่ใช้ คือ ความถี่ของเทอม ข้อมูลที่ใช้คือ Yahoo Text Hierarchy และใช้ Naïve Bayes Classifier ในการจำแนกประเภทเอกสาร พวกเขาแสดงให้เห็นว่าการใช้ลำดับของ

คำถึง 3 (n-gram, n=3) ร่วมกับคำเดี่ยว แทนที่การใช้คำเดี่ยวเพียงอย่างเดียวนั้น สามารถเพิ่มประสิทธิภาพในการจำแนกประเภทได้ โดยลำดับที่ยาวกว่า โดยเฉลี่ยแล้วไม่ส่งผลกระทบต่อประสิทธิภาพในการจำแนกประเภท

2.7.1 การลดฟีเจอร์วลีสำหรับการจำแนกประเภทเอกสาร

เนื่องจากในการใช้ฟีเจอร์ที่เป็นวลียังมีปัญหาในการนำมาใช้อยู่ คือ เซตของฟีเจอร์มีขนาดใหญ่ ซึ่งมีฟีเจอร์ที่มีความสามารถในการจำแนกประเภทต่ำอยู่มาก ดังนั้นจึงได้มีหลายงานวิจัยที่ได้พยามยามสร้างและเลือกฟีเจอร์วลี ที่จะใช้ร่วมกับคำเดี่ยวในการแทนเอกสาร ให้มีจำนวนน้อยลงและมีประสิทธิภาพเพิ่มขึ้น เช่น

งานวิจัยเรื่อง "*The Use of Bi-grams to Enhance Text Categorization*" โดย Chade-Meng Tan และกลุ่ม [26] ได้ทำการเลือกคำเดี่ยว โดยเลือกใช้คำเดี่ยวที่มีความถี่ของเอกสารสูงเป็นอันดับต้นๆ และใช้คำเดี่ยวที่เลือกนั้น มาสร้าง Bi-gram จากคำที่ติดกันในเอกสารที่ผ่านการเตรียมและลบ Stopwords ออกแล้ว โดยคำใดคำหนึ่งจะต้องอยู่ภายในเซตของคำเดี่ยวที่เลือกไว้ และทำการเลือก Bi-grams โดยใช้เกณฑ์ความถี่และค่า Information Gain ในแต่ละคลาส โดยมีการตั้งค่าต่ำสุดของแต่ละค่าไว้ ชุดข้อมูลที่ใช้ คือ เว็บเพจ จาก Yahoo! Science Hierarchy และข้อมูล Reuters-21578 ผลการทดลองจำแนกประเภทเอกสารโดยใช้ Naïve Bayes แสดงให้เห็นว่า วลีที่ถูกเลือกอย่างเหมาะสมรวมกับคำเดี่ยว ให้ผลดีกว่าการใช้คำเดี่ยวเพียงอย่างเดียวเป็นฟีเจอร์

ต่อมา Bekkerman และ Allan ได้รายงานเรื่อง "*Using Bigrams in Text Categorization*" [27] โดยทำการสร้าง Bi-grams จากคำเดี่ยวที่อยู่ในลำดับต้นๆ โดยดูจากค่า Mutual Information และเลือกใช้เฉพาะ Bi-gram ที่ดีกว่าคำที่เป็นส่วนประกอบของมันทั้ง 2 คำ โดยดูจากค่า Mutual Information และเลือก Bi-grams ที่ได้ในอันดับต้นๆ มาใช้

ข้อมูลที่ใช้ในการทดลอง คือ 20 Newsgroups โดยเลือกใช้คำเดี่ยว 5000 คำ และ เลือก Bi-grams 1000 คำ ที่ได้จากการเลือกจากแต่ละคลาสมาใช้เป็นฟีเจอร์ร่วมกัน ผลจากการทดลองโดยใช้ SVM Classifier แสดงให้เห็นว่าการใช้วลีร่วมกับคำให้ผลการจำแนกประเภทที่ดีกว่า การใช้คำเพียงอย่างเดียว

จากงานวิจัยที่ผ่านมา เราสามารถสรุปปัญหาในการใช้วลีเป็นฟีเจอร์ในการแทนเอกสารได้ดังนี้

1. วลีมีความจำเพาะเจาะจงเกินไป คือ มีความหมายเฉพาะมากเกินไป เช่น วลีทางภาษาหรือวลีที่ได้จากเอกสารโดยตรง ทำให้ค่าที่ถูกแทนในเอกสารส่วนใหญ่ มีค่าเป็นศูนย์ ตัวอย่าง เช่น วลี "*k-nearest neighbor categorization algorithm*" จะปรากฏในแต่ละเอกสารน้อยมาก แต่ถ้าเราใช้วลีย่อยเป็น "*k-nearest neighbor*" หรือ "*categorization algorithm*" จะมีการกระจายของค่าในการ

แทนเอกสารดีกว่า เพราะมีโอกาสในการปรากฏมากกว่าในแต่ละเอกสาร และในการสกัดวลีทางภาษายังทำได้ยากด้วย

2. แต่ถ้าวลีนั้นสั้นไป ก็จะเกิดปัญหาคล้ายกับการใช้คำเดี่ยว คือ มีความสามารถในการแบ่งแยกประเภทต่ำ โดยเฉพาะอย่างยิ่งเมื่อประเภทของข้อมูลมีความใกล้เคียงกันมาก เช่น วลี "*drama player voting*" อยู่ในข่าวบันเทิง ส่วนวลี "*football player voting*" อยู่ในข่าวกีฬา ถ้าเราใช้ฟีเจอร์เป็น "*player voting*" ก็ทำให้ประสิทธิภาพในการจำแนกว่าอยู่ในข่าวประเภทไหนลดลง

3. จากปัญหาข้อ 1 กับ 2 ถ้าวลียาวไปก็ไม่ดี สั้นเกินก็ไม่ดี ทางแก้อย่างหนึ่งก็คือ ใช้ทั้งวลีและวลีย่อยเป็นฟีเจอร์ แต่ก็จะเกิดปัญหาความซ้ำซ้อนของวลีขึ้นมาอีก ดังนั้น จึงต้องมีการเลือกว่าควรจะใช้วลีไหนดีในการแทนเอกสาร หรือใช้ทั้ง 3 ตัวอย่างในกรณีตัวอย่างข้อที่ 1 คือใช้ ทั้งวลี "*k-nearest neighbor categorization algorithm*" "*k-nearest neighbor*" และ "*categorization algorithm*" แทนเอกสาร

4. การใช้วลีทางสถิติเป็นการแก้ปัญหาอย่างหนึ่งของการใช้วลีทางภาษา เพราะไม่ต้องคำนึงถึงหลักไวยากรณ์ทางภาษา เช่น การสร้าง n-gram ซึ่งคือ ลำดับของคำใด ๆ ที่ปรากฏต่อเนื่องกันในเอกสาร โดยส่วนใหญ่จะใช้ n=2 เรียกว่า Bi-gram แต่ปัญหาก็คือ วลีที่สร้างได้มีจำนวนมาก ที่เป็นวลีขยะ คือ วลีที่มีอำนาจในการจำแนกประเภทต่ำ และในการสร้าง n-gram นั้นจำเป็นที่จะต้องกำหนดขนาดสูงสุดของวลีไว้ คือ n ทำให้บางครั้งมีวลีที่มีความสามารถในการจำแนกประเภทที่ดี แต่มีขนาดมากกว่า n ก็จะไม่ถูกใช้

ดังนั้น พอจะสรุปได้ว่าฟีเจอร์ที่จะนำมาใช้แทนเอกสารนั้น ควรจะมีคุณสมบัติ 2 ประการ คือ

1. มีอำนาจในการแบ่งแยก (Discriminant Power) สูง คือ วลีนั้นจะต้องไม่มีความหมายแคบเกิน ทำให้มีความถี่น้อย และจะต้องไม่มีความหมายกว้างจนเกินไป
2. มีปริมาณน้อย โดยไม่กระทบต่อความถูกต้องในการจำแนกประเภทเอกสาร ซึ่งต้องอาศัยการเลือกฟีเจอร์ที่มีประสิทธิภาพ

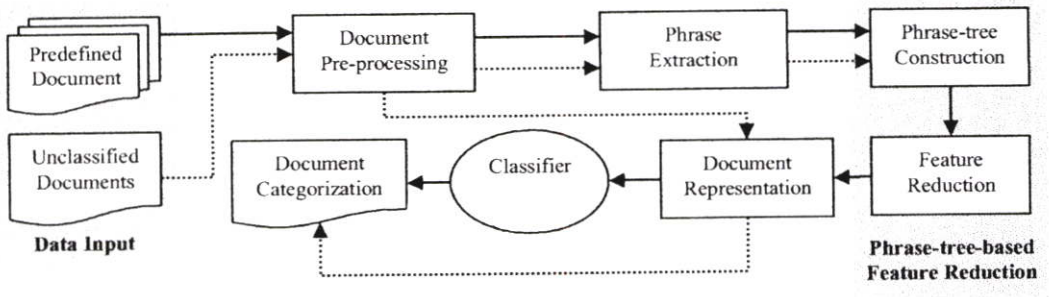
ในงานวิจัยนี้พยายามแก้ปัญหาดังกล่าว โดยนำ ทฤษฎีต่างๆ ที่เกี่ยวข้อง เหล่านี้ มาใช้ในการแก้ปัญหาในการจำแนกประเภทเอกสาร โดยทำการลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร ซึ่งจะกล่าวถึงรายละเอียดในบทที่ 3 ต่อไป

บทที่ 3

การลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร

ในงานวิจัยนี้ได้พยามยามแก้ไขปัญหาจากการใช้วลีเป็นฟีเจอร์ในการแทนเอกสาร โดยนำเสนอวิธีการสร้างฟีเจอร์วลีแบบใหม่ ที่ได้ปริมาณน้อย แต่ยังคงมีประสิทธิภาพในการจำแนกประเภทเอกสาร คือ มีอำนาจในการแบ่งแยกสูง โดยมีส่วนที่สำคัญ คือ เทคนิคในการสกัดวลี การสร้างฟีเจอร์และการเลือกวลีโดยใช้โครงสร้างข้อมูลแบบต้นไม้

กระบวนการของการจำแนกประเภทเอกสาร โดยทำการลดฟีเจอร์โดยใช้ต้นไม้วลี แสดงดังรูป 3.1 โดยเส้นทึบแทนการไหลของข้อมูลในขั้นตอนการสร้างตัวจำแนกประเภท โดยเอกสารที่เรารู้คลาสแล้วจะผ่านกระบวนการเตรียมเอกสาร จากนั้นผ่านกระบวนการการ การลดฟีเจอร์โดยใช้ต้นไม้วลี และการแทนเอกสาร เพื่อสร้างเป็นตัวจำแนกประเภท ส่วนเส้นประหมายถึงเส้นทางการไหลของข้อมูลในขั้นตอนการจำแนกประเภท โดยเมื่อมีเอกสารที่เรายังไม่ทราบคลาสเข้ามา ก็จะผ่านกระบวนการเตรียมเอกสาร และการแทนเอกสาร เพื่อนำเข้าสู่ตัวจำแนกประเภท ทำให้ทราบประเภทของเอกสารนั้น



รูปที่ 3.1 กระบวนการของการจำแนกประเภทเอกสาร โดยทำการลดฟีเจอร์โดยใช้ต้นไม้วลี

3.1 การเตรียมเอกสาร (Data Pre-processing)

ในกระบวนนี้ เอกสารจะถูกเตรียมและแปลงให้อยู่ในรูปแบบที่จะเข้าสู่กระบวนการการสกัดวลีต่อไป โดยมีรายละเอียดดังต่อไปนี้

1. ในกรณีที่เอกสารเป็นเว็บลบส่วนที่เป็นแท็ก (Tag) ของเอกสารที่เป็นเว็บออกก่อน เช่น "*<html>*", "*</html>*", "*
", "<A HREF="...">*"

2. แปลงให้เป็นอักษรตัวพิมพ์เล็กทั้งหมด และลบตัวเลขออก

3. เปลี่ยนเครื่องหมายต่างๆ เช่น , *( ) [ ] { } / \ | = ' ` " " ^ ? < >* และ Stopwords ให้เป็นเครื่องหมาย . (Full Stop Point) เพียงอย่างเดียวเพื่อใช้แบ่งแยกวลีต่อไป แต่จะมีกรณีพิเศษ คือ เครื่องหมาย *- _ & + * ~ # $ %* ซึ่งจะไม่ถูกเปลี่ยน แต่จะถูกลบเครื่องหมายออก เช่น "*black-*

market" → "*black market*", "*black&white game*" → "*black white game*", "*bottle + milk = bottle of milk*" → "*bottle milk . bottle milk*" การที่เราไม่ใช้เครื่องหมาย – _ & + มาแบ่งแยกวลี เพราะทั้งหมดเป็นการบ่งบอก ถึงการเกิดขึ้นร่วมกันของคำ ซึ่งเป็นวลีที่ส่วนมากได้ความหมายอยู่แล้ว ดังนั้น จึงยังไม่ต้องแยกออก

ในการแบ่งแยกวลีแบบนี้ คล้ายกับงานของ Klok ที่ใช้การสกัดวลีโดยใช้เทอมที่อยู่ระหว่าง Stopwords มาเป็นวลี ที่ใช้ในการสืบค้นร่วมกับคำและวลีที่เป็นคำนาม [21]

4. ทำแต่ละคำให้อยู่ในรากศัพท์ (Stemming) เพื่อลดจำจำนวนเทอมที่ได้

3.2 การสกัดวลี (Phrase Extraction)

ในกระบวนการการสกัดวลีนี้ วลีจะถูกสกัดจากเอกสารที่ถูกเตรียมแล้ว โดยมีเครื่องหมาย "." เป็นตัวแบ่งแยกวลี

สมมติมีตัวอย่างเอกสารดังนี้

"*<html> 1. k-Nearest neighbor is an efficient categorization algorithm. </html>*"

หลังจากการทำการเตรียมเอกสาร เราจะได้เอกสารที่เตรียมแล้ว คือ

"*.k-nearest neighbor . effici categor algorithm.*"

เมื่อเข้าสู่กระบวนการสกัดวลีนี้ เราจะได้ 2 วลี คือ "*k-nearest neighbor*" และ "*effici categor algorithm*" ซึ่งในการสกัดวลีแบบนี้เราจะได้คำเดี่ยวมาด้วย เช่น เอกสารเป็น

"*This is categorization method (classification)*"

หลังจากการทำการเตรียมเอกสาร เราจะได้เอกสารที่เตรียมแล้ว คือ

"*categor method .classifi.*"

เมื่อเข้าสู่กระบวนการสกัดวลีนี้ เราจะได้ทั้งวลี และคำเดี่ยว คือ "*categor method*" และ "*classifi*" โดยในที่นี้เราจะรวมคำเดี่ยวที่สกัดได้จากกระบวนนี้เข้าไปด้วย เพื่อใช้ในการสร้างต้นไม้วลีต่อไป

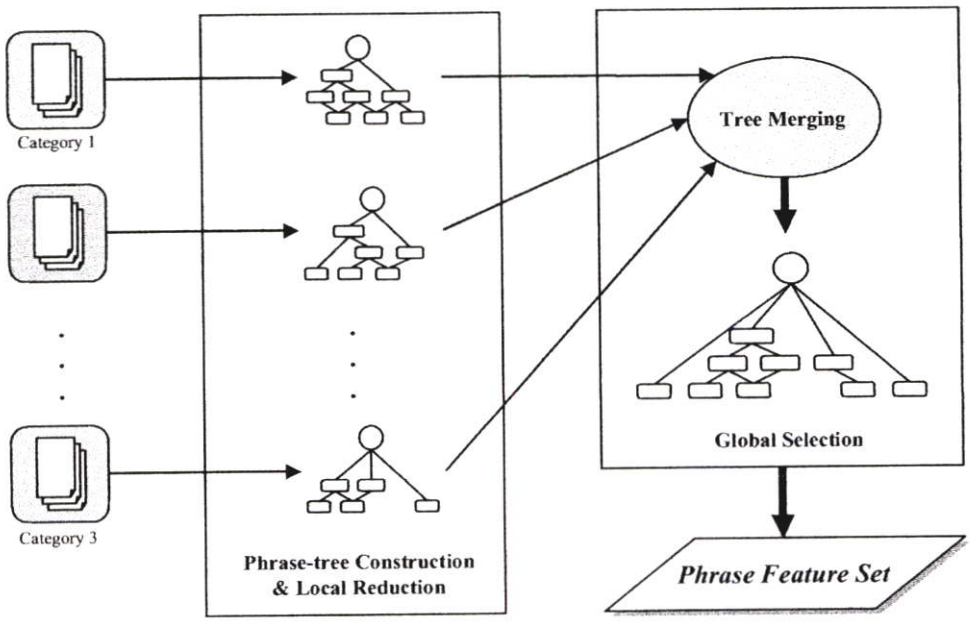
3.3 การลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร

(Phrase-tree-based Feature Reduction for Text Categorization)

ในการสร้างต้นไม้วลีนั้น มีวัตถุประสงค์เพื่อใช้โครงสร้างข้อมูลแบบต้นไม้ มาช่วยในการสร้างฟีเจอร์ โดยโครงสร้างต้นไม้สามารถแสดงถึงความสัมพันธ์ระหว่างวลีและวลีย่อย และคำเดี่ยว ซึ่งช่วยในการเลือกฟีเจอร์ที่ลดความซ้ำซ้อนของวลีได้ โดยเราต้องการฟีเจอร์ที่เป็นตัวแทน

เอกสารที่ดี มีอำนาจในการจำแนกประเภท โดยจะทำการเลือกฟีเจอร์ทั้งในระดับ Local และระดับ Global โดย

ในการทำการลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร ประกอบด้วย 3 ขั้นตอนหลักๆ คือ 1. การสร้างต้นไม้วลี 2. การลดฟีเจอร์ในระดับ Local 3. การลดฟีเจอร์ในระดับ Global แสดงดังรูป 3.2



รูปที่ 3.2 ขั้นตอนการลดฟีเจอร์โดยใช้ต้นไม้วลี

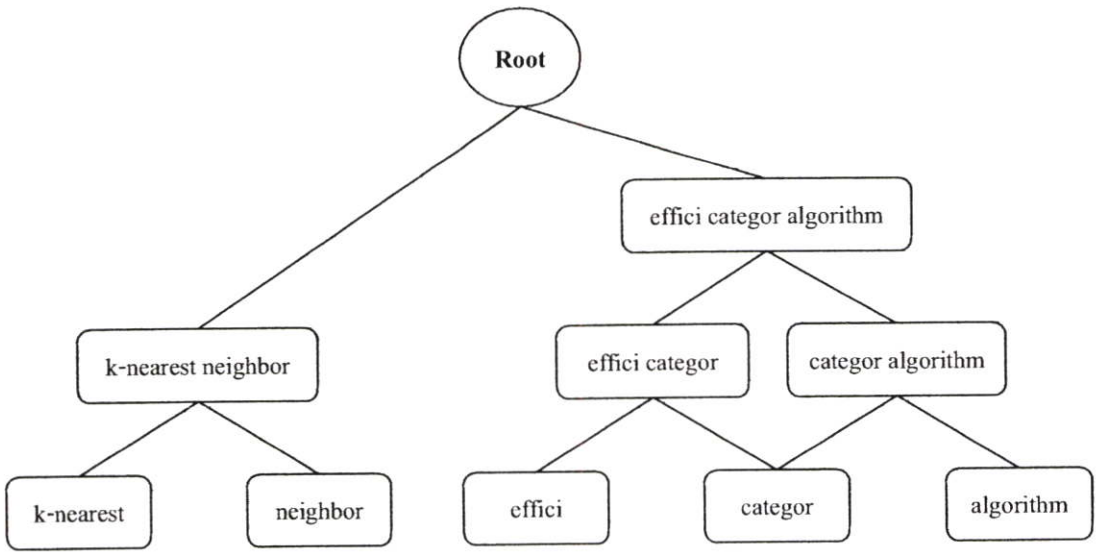
โดยในการทำการลดฟีเจอร์โดยใช้ต้นไม้วลีสำหรับการจำแนกประเภทเอกสาร สามารถเขียนเป็นอัลกอริทึมได้ดังรูปที่ 3.9 ประกอบด้วย 3 ขั้นตอนคือ 1. การสร้างต้นไม้วลี 2. การลดฟีเจอร์ในระดับ Local 3. การลดฟีเจอร์ในระดับ Global และมีรายละเอียดในแต่ละขั้นตอน ดังนี้

3.3.1 การสร้างต้นไม้วลี (Phrase-tree Construction)

ในการสร้างต้นไม้วลี จะเป็นการสร้างต้นไม้วลีของแต่ละคลาส โดยสร้างจากวลีที่สกัดได้จากทุกเอกสารในคลาสนั้นๆ โดยเลือกใช้วลีที่มีค่าความถี่เอกสารมากกว่าค่าที่กำหนดไว้มาใช้

การสร้างจะสร้างทีละวลี ของแต่ละเอกสารในคลาส โดยแทนแต่วลีที่ไม่ซ้ำกันเป็นโหนดของต้นไม้ของคลาส จากนั้น สร้างวลีย่อยจากแต่ละวลี โดยวลีย่อย คือ ลำดับของคำที่ต่อเนื่องกันที่มีขนาดน้อยกว่าของวลี 1 คำ ถ้ายังไม่มีโหนดของวลีย่อยนั้น ก็ทำการเพิ่มโหนดเข้าไปในต้นไม้ของคลาส และสร้างวลีย่อยไปเรื่อยๆ จนสุดท้ายจะได้โหนดใบ คือ คำเดี่ยว และทำทุกวลีที่สกัดได้ของเอกสารจนครบทุกวลี และทุกเอกสารในคลาส

ตัวอย่างเช่น มี 2 วลีที่สกัดได้ คือ "*k-nearest neighbor*" และ "*effici categor algorithm*" นำมาสร้างต้นไม้วลีจะได้ดังรูป 3.3



รูปที่ 3.3 ตัวอย่างต้นไม้วลีที่ได้จากวลี "*k-nearest neighbor*" และ "*effici categor algorithm*"

3.3.2 การลดฟีเจอร์วลีระดับ Local

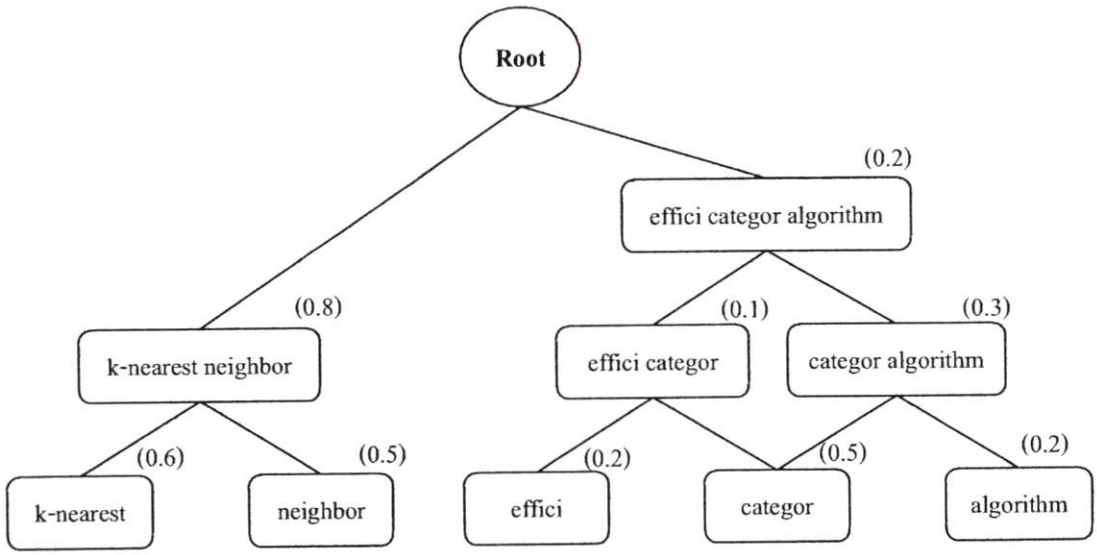
ในระดับ Local นี้ จะเป็นการลดฟีเจอร์จากต้นไม้วลีของแต่ละคลาส โดยทำการคำนวณค่าคะแนนของฟีเจอร์ให้กับแต่ละโหนด ในงานวิจัยนี้ใช้ Odds Ratio Measure [10] โดยค่า Odds Ratio หาได้จากสมการที่ 2.3

นอกจากนี้ยังใช้เกณฑ์จากความถี่เอกสารในการเลือกฟีเจอร์อีกด้วย โดยกำหนดค่าต่ำสุดของความถี่เอกสารไว้

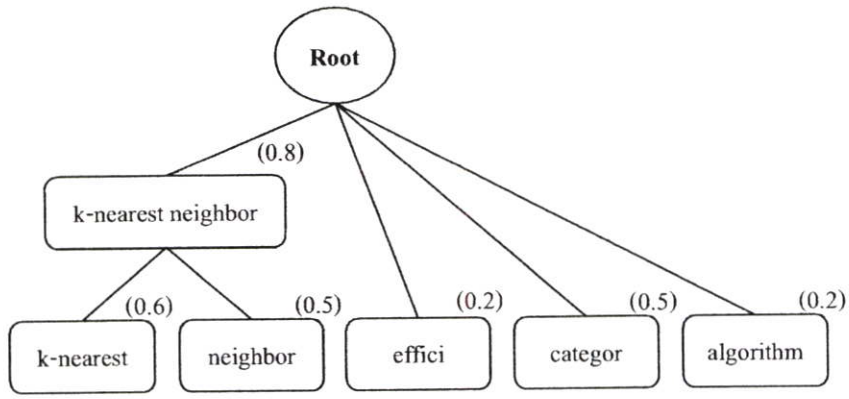
ในการเลือกฟีเจอร์วลีนั้นทำได้โดย

1. คำนวณค่า *OR* ให้กับแต่ละโหนดของต้นไม้วลีแล้วทำการพิจาณาไปทีละโหนด
2. พิจารณาลบโหนดของวลีที่ มีค่าคะแนนน้อยกว่าโหนดใดโหนดหนึ่ง ที่เป็นวลีย่อยของมัน ซึ่งรวมถึงโหนดในระดับต่ำสุดที่เป็น โหนดใบด้วย หรือกล่าวได้ว่า ลบโหนดที่มีค่าคะแนนน้อยกว่าค่าคะแนนสูงสุดของวลีที่อยู่ในต้นไม้ย่อย (Sub-tree) ของมัน

จากตัวอย่างต้นไม้วลี รูปที่ 3.3 สมมติว่าทำการคำนวณคะแนนแล้วได้ผลค่า Odds Ratio ดังรูปที่ 3.4 โดยหลังจากผ่านกระบวนการการลดฟีเจอร์แล้ว จะได้ต้นไม้วลี ดังแสดงในรูปที่ 3.5



รูปที่ 3.4 ตัวอย่างต้นไม้วลีจากรูปที่ 3.3 ที่ให้คะแนนแล้ว

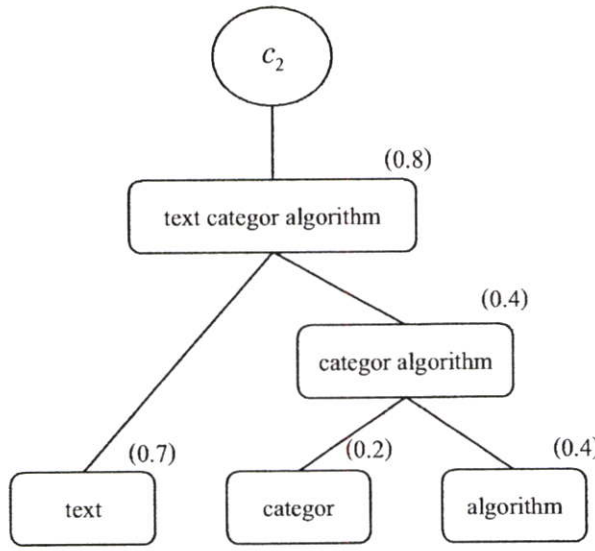


รูปที่ 3.5 ตัวอย่างต้นไม้วลีจากรูป 3.4 ที่ผ่านการเลือกฟีเจอร์แล้ว

3.3.3 การลดฟีเจอร์วลีระดับ Global

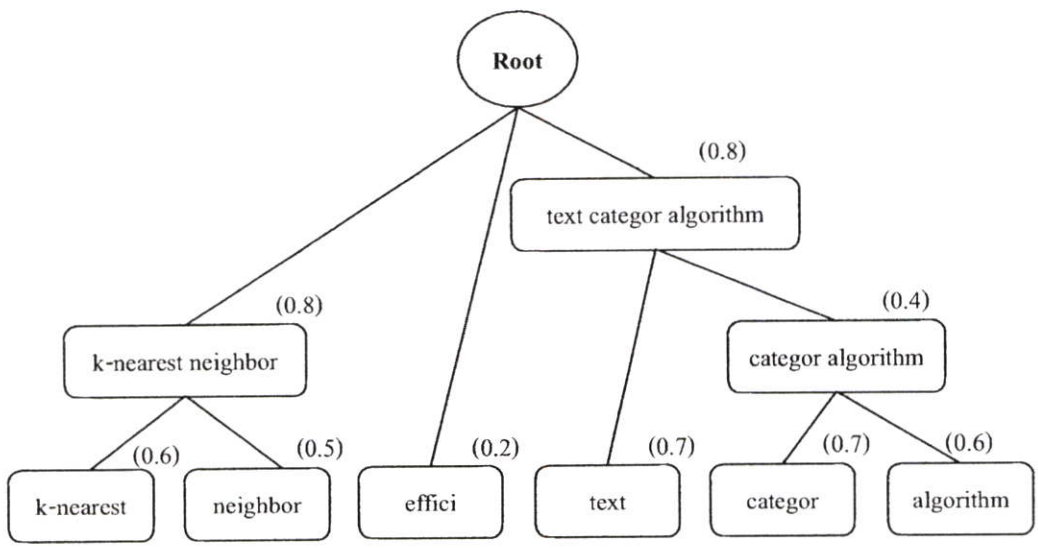
ในการลดฟีเจอร์วลีระดับ Global นั้น ต้นไม้วลีของแต่ละคลาสจะถูกนำมารวมกัน โดยในโหนดที่มีร่วมกัน การให้ค่าคะแนนจะเป็น การรวมค่า Odds Ratio ของโหนดนั้นในแต่ละคลาส เพื่อใช้ในการลดฟีเจอร์แบบ Global ซึ่งเมื่อได้ต้นไม้วลีจากการรวมต้นไม้ของแต่ละคลาสแล้ว ก็ทำกระบวนการลด เช่นเดียวกับในระดับ Local

ตัวอย่างเช่น มีต้นไม้วลีที่ผ่านการเลือกฟีเจอร์ระดับ Local ที่ได้จากคลาส c_1 ดังรูป 3.5 และ c_2 แสดงดังรูป 3.6



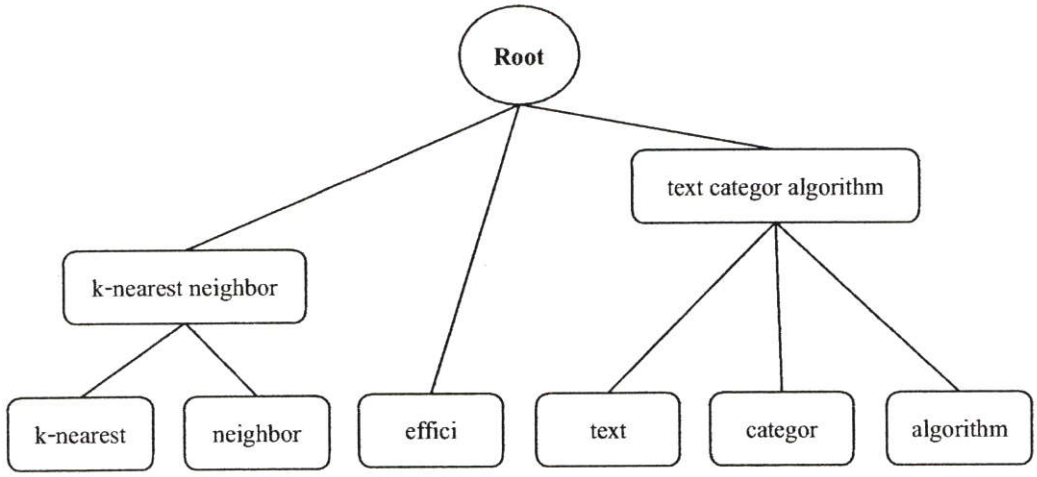
รูปที่ 3.6 ตัวอย่างต้นไม้วลีของคลาส c_2 ที่ผ่านการเลือกฟีเจอร์ระดับ Localแล้ว

เมื่อทำการรวมต้นไม้วลีในรูป 3.5 และ 3.6 แล้วได้ต้นไม้วลีของคลาสทั้งหมด โดยโหนดที่มีร่วมกันจะนำค่าคะแนนมาบวกกัน แสดงดังรูปที่ 3.7



รูปที่ 3.7 ต้นไม้วลีจากการรวมกันของต้นไม้ในรูปที่ 3.5 และ 3.6

เมื่อรวมได้เป็นต้นไม้วลีสำหรับการเลือกวลีในระดับ Global แล้ว ทำการลดฟีเจอร์อีกครั้งด้วยหลักการเดียวกันกับการลดในระดับ Local คือ ลบโหนดที่มีค่าคะแนนน้อยกว่าค่าคะแนนสูงสุดของวลีที่อยู่ในต้นไม้ย่อยของมัน โดยจากรูป 3.7 เมื่อทำการลดฟีเจอร์แล้วได้ต้นไม้วลี แล้วได้ต้นไม้วลี ดังรูปที่ 3.8



รูปที่ 3.8 ต้นไม้วลีที่ได้หลังจากทำการลดฟีเจอร์ จากต้นไม้วลีในรูปที่ 3.7

สุดท้ายจะได้เซตของฟีเจอร์วลี จากวลีในทุกโหนดของต้นไม้วลีที่ได้ ยกเว้นโหนดใบ เพราะเป็นคำเดี่ยว จากตัวอย่าง เราจะได้ เซตของฟีเจอร์เป็น

Phrase_Features = {" *k-nearest neighbor*", "*text categor algorithm*"}

3.4 การแทนเอกสาร (Document Representation)

ในการแทนเอกสาร เราจำเป็นต้องแทนเอกสารด้วย คำเดี่ยว และวลี ร่วมกันเพื่อเพิ่มประสิทธิภาพในการจำแนกประเภทเอกสาร ดังนั้น จากต้นไม้วลีสุดท้ายที่ได้ เราจะทำการเลือกฟีเจอร์ที่เป็นคำเดี่ยวโดยดูที่เฉพาะโหนดใบ ซึ่งสามารถใช้เกณฑ์ต่างๆ ในการเลือกได้

ในงานวิจัยของเราเลือกใช้คำเดี่ยวที่มีค่าคะแนนมากว่าเกณฑ์ที่ตั้งไว้ มาใช้เป็นฟีเจอร์ จากตัวอย่างต้นไม้วลีในรูปที่ 3.7 สมติเราตั้งเกณฑ์คะแนนไว้เป็น 0.3 ดังนั้น เซตของฟีเจอร์วลีที่เป็นคำเดี่ยวที่เราได้คือ

Word_Features = {" *k*-nearest", "*neighbor*", "*text*" , "*categor*", "*algorithm*"}

สุดท้ายนำฟีเจอร์ทั้ง 2 มารวมกัน เพื่อใช้แทนเอกสาร

Features = {" *k*-nearest", "*neighbor*", "*text*" , "*categor*", "*algorithm*",
" *k-nearest neighbor*", "*text categor algorithm*"}

Algorithm: Phrase-tree-based Feature Reduction

Input: D – training set of documents; $d = d_1, d_2, \ldots d_n$

C – set of categories; $C = c_1, c_2, \ldots c_m$

L –set of label of each documents in D; $L(d_i) \in C$

min_f – minimum frequency

min_OR– minimum Odds Ratios Score

Output: F: Set of Features

```
F = ϕ
// Phrase I : Phrase-tree Construction & Local Reduction
For i=1 to m
        P = ϕ
        For each document d_j ∈ D that L(d_j) = c_i
                PH = set of extracted phrases from d that have document frequency > min_f
                P = P ∪ PH
        End
        P_tree_i = construct_phrase_tree(P)              //generate phrase-tree rom P
        P_tree_i = calulate_odds_ratio(P_tree_i)         //calculate OR for each node in tree
        P_tree_i = reduce_phrase_tree(P_tree_i, min_OR)  //Local Reduction
End
// Phrase II & III : Phrase-tree Merging & Global Reduction
M_tree = ϕ
For i=1 to m                                  //merge phrase-tree from each category
        M_tree = merge_tree(M_tree, P_tree_i)
End
M_tree = reduce_phrase_tree(M_tree, min_OR)          //Global Reduction
F = all nodes from M_tree
```

รูปที่ 3.9 อัลกอริทึมการลดฟีเจอร์โดยใช้ต้นไม้วลี

ในบทที่ 4 จะกล่าวถึง การทดลองและผลการทดลองกับชุดข้อมูลที่นำมาทดลองกับกระบวนการทำงานของงานวิจัยนี้ โดยทดลองเปรียบเทียบประสิทธิภาพในการใช้ฟีเจอร์ในแบบต่างๆ

บทที่ 4

การทดลองและผลการทดลอง

ในการพิจารณาประสิทธิภาพของฟีเจอร์ เราจะพิจารณาที่ความความถูกต้องในการจำแนกประเภทเอกสาร เมื่อแทนเอกสารด้วยฟีเจอร์นั้นๆ และพิจารณาที่จำนวนของฟีเจอร์ด้วย ซึ่งควรจะมีจำนวนน้อย แต่ให้ความถูกต้องสูง ในการทดลองได้ทำการเปรียบเทียบประสิทธิภาพของฟีเจอร์ที่ได้จากต้นไม้วลีกับฟีเจอร์แบบอื่นๆ โดยมีรายละเอียดในการทดลองต่อไปนี้

4.1 การวัดความถูกต้องของการจำแนกประเภทเอกสาร

การวัด ความถูกต้องของการจำแนกประเภทเอกสาร ได้ใช้ตัววัดที่ใช้ในงานวิจัยทั่วไป คือ ตัววัดอัตราความถูกต้อง (Accuracy Rate) และตัววัดความถูกต้อง F-measure โดยมีรายละเอียดของแต่ละตัวชี้วัด ดังต่อไปนี้ [13] และ [15]

4.1.1 ตัววัดอัตราความถูกต้อง (Accuracy Rate)

ตัววัดประสิทธิภาพอัตราความถูกต้องในการจำแนกประเภทเอกสาร มีรายละเอียดดังสมการที่ (4.1)

$$R = \left(\frac{Ndc}{Ndt}\right) \times 100 \tag{4.1}$$

โดยที่ Ndc คือ จำนวนเอกสารที่ถูกจำแนกประเภทถูกต้อง

Ndt คือ จำนวนเอกสารที่นำมาทดสอบ

4.1.2 ตัววัดความถูกต้อง F-Measure

ตัววัดความถูกต้อง F-measure ใช้ในการวัดความถูกต้องของการจำแนก ประเภทเอกสาร โดยจะมีการคำนวณหาค่า Recall และ Precision ค่า F-measure จะมีค่าอยู่ระหว่าง 0 ถึง 1 โดยที่ค่านี้จะบอกว่าผลการจำแนก ประเภทเอกสารนั้น มีความถูกต้องมากน้อยเพียงใด ถ้าค่าที่คำนวณได้ มีค่าน้อย แสดงว่ามีความถูกต้องต่ำ ถ้าค่าที่คำนวณได้มีค่าเป็น 1 แสดงว่ามีความถูกต้องสูงสุด การคำนวณหาค่า F-measure $F(i)$ ของ ประเภทเอกสาร ลำดับที่ i จะต้องคำนวณหาค่า Recall และ Precision ซึ่งส่วนที่นำมาหาค่า Recall และ Precision พิจารณาจากรูปที่ 4.1

สมการที่ใช้สำหรับหาค่า Recall และ Precision คือสมการที่ (4.2) และ (4.3) ตามลำดับ

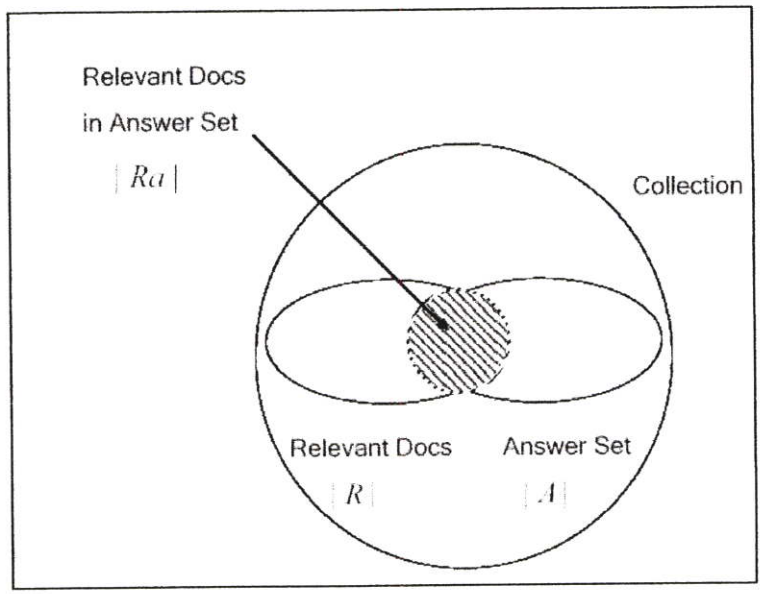
$$Recall(i) = \frac{|Ra_i|}{|R_i|} \tag{4.2}$$

$$Precision(i) = \frac{|Ra_i|}{|A_i|} \tag{4.3}$$

โดยที่ $|Ra_i|$ คือ จำนวนของเอกสารในชุดทดสอบที่อยู่ในประเภทที่ i และเมื่อทำการจำแนกประเภทแล้ว ก็ได้ผลการจำแนกประเภทว่าอยู่ในประเภทที่ i

$|R_i|$ คือ จำนวนของเอกสารในชุดทดสอบที่อยู่ในประเภทที่ i

$|A_i|$ คือ จำนวนของเอกสารที่ได้จากการจำแนกประเภทว่าอยู่ในประเภทที่ i



รูปที่ 4.1 แสดงส่วนที่ใช้วัดค่า Precision และ Recall ในการร้องขอสารสนเทศ [15]

การคำนวณหาค่า $F(i)$ ของ ประเภทเอกสารลำดับที่ i สามารถคำนวณได้จากสมการที่ (4.4)

$$F(i) = \frac{2 \times Recall\ (i) \times Precision\ (i)}{Recall\ (i) + Precision\ (i)} \tag{4.4}$$

การคำนวณค่า F-measure ของทุกๆ ประเภทเอกสาร สามารถคำนวณได้จากสมการที่ (4.5)

$$F = \sum_{i=1}^{Nc} \frac{|R_i|}{Ndt} F(i) \tag{4.5}$$

โดยที่ Nc คือ จำนวนของประเภท

Ndt คือ จำนวนเอกสารที่นำมาทดสอบ

4.2 การออกแบบการทดลอง

4.2.1 การสุ่มตัวอย่างเอกสาร (Document Sampling)

เนื่องจากในบางชุดเอกสารมีเอกสารจำนวนมาก ทำให้มีความซับซ้อนในการคำนวณสูง ดังนั้นจึงจำเป็นต้องลดจำนวนเอกสารลง โดยนำเทคนิคการสุ่มตัวอย่างในทางสถิติมาใช้ในการสุ่มเอกสารตัวอย่างขึ้นมา เพื่อใช้ในการเรียนรู้และทดสอบต่อไป

รศ.สุรินทร์ นิยมางกูร กล่าวในหนังสือเทคนิคการสุ่มตัวอย่าง [20] ว่า ตัวอย่าง (Sample) ที่ดี คือ ตัวอย่างที่สามารถให้รายละเอียดข้อเท็จจริงเกี่ยวกับคุณลักษณะของประชากรได้เป็นอย่างดี ตามหลักวิชาการทางสถิติมีความเชื่อว่า ตัวอย่างที่เลือกมาโดยวิธีการสุ่มจะไม่ทำให้เกิดความเอนเอียงในการเก็บรวบรวมรายละเอียดข้อเท็จจริงเกี่ยวกับคุณลักษณะของประชากร ดังนั้นตัวอย่างที่ดีจึงควรเป็นตัวอย่างที่เลือกมาโดยวิธีการสุ่ม หรือ อาจจะกล่าวได้ว่าตัวอย่างที่ดีก็คือ ตัวอย่างสุ่ม นั่นเอง

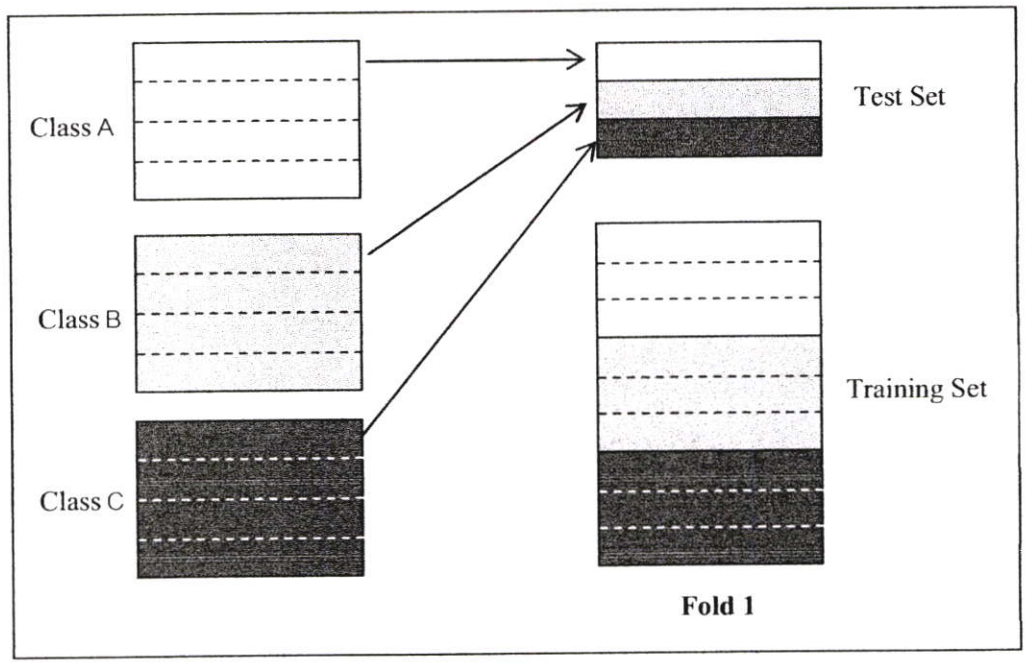
การสุ่มตัวอย่างแบ่งเป็น 2 แบบใหญ่ๆ คือ การสุ่มตัวอย่างที่ไม่เป็นไปตามโอกาสทางสถิติ (Non-Probability Sampling) และ การสุ่มตัวอย่างที่เป็นไปตามโอกาสทางสถิติ (Probability Sampling) โดย การสุ่มตัวอย่างที่คำนึงถึงความน่าจะเป็นในการสุ่ม มี 4 วิธี คือ การสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) การสุ่มอย่างมีระบบ (Systematic Random Sampling) การสุ่มตัวอย่าง แบบเป็นชั้นภูมิ (Stratified Random Sampling) และ การสุ่มแบบแบ่งกลุ่ม (Cluster Sampling)

การสุ่มตัวอย่าง แบบเป็นชั้นภูมิ มีการจัดแบ่งประชากรเป็นกลุ่มหรือชั้นย่อยๆ ก่อน แล้วเลือกสุ่มตัวอย่างตามสัดส่วน (Proportional)ในแต่ละชั้น จากนั้นจึงใช้การสุ่มอย่างง่าย เช่นแบ่งนักศึกษาตามคณะต่างๆ หาขนาดกลุ่มตัวอย่าง จากนั้นเทียบสัดส่วนตามขนาด แล้วจับฉลาก เป็นต้น วัตถุประสงค์หลักของการสุ่มตัวอย่างแบบแบ่งเป็นชั้นภูมิคือ เพื่อให้ได้กลุ่มตัวอย่างที่มีองค์ประกอบของลักษณะต่าง ๆ ใกล้เคียงกับกลุ่มประชากร และให้ได้กลุ่มตัวอย่างที่สามารถตอบสนองวัตถุประสงค์ของการวิจัยได้

ในงานวิจัยนี้เลือกประยุกต์ใช้ การสุ่มตัวอย่างแบบเป็นชั้นภูมิ เพื่อรักษาลักษณะของคำภายในแต่ละประเภทเอกสารไว้ โดยในขั้นแรกจะแบ่งชุดเอกสารตามประเภทเอกสาร และในแต่ละประเภทเอกสารทำการแบ่งชั้นภูมิตามการกระจายตัวของเอกสารในคำต่างๆ ชั้นภูมิในแต่ละประเภทเอกสารของเราในที่นี้ก็คือ คำแต่ละคำ ซึ่งมีจำนวนชั้นภูมิเท่ากับจำนวนคำทั้งหมด ของประเภทเอกสารนั้น เมื่อแบ่งชั้นภูมิแล้วก็ทำการสุ่มเอกสาร จากแต่ละชั้นภูมิ โดยใช้การสุ่มอย่างง่าย จำนวนเอกสารที่จะสุ่มขึ้นมานั้น เป็นไปตามสัดส่วนที่เรากำหนดไว้ เช่น กำหนดไว้ที่ 0.1 ถ้าในคำที่ 1 มีจำนวนเอกสารเท่ากับ 100 จำนวนตัวอย่างที่เราจะสุ่มมาในชั้นภูมิคำนี้จะเท่ากับ 1 เอกสาร

4.2.2 การตรวจสอบไขว้ (Cross-Validation)

การตรวจสอบไขว้ เป็นวิธีการทดสอบการจำแนกประเภทเอกสารที่นักวิจัยนิยมใช้ โดย k-fold Cross Validation หมายถึง ทำการแบ่งชุดเอกสารออกเป็น k ชุด แล้วทำการทดสอบ โดยแต่ละครั้ง ใช้ 1 ส่วนเป็นชุดทดสอบส่วน k-1 ชุดที่เหลือเป็นชุดทดสอบ สับเปลี่ยนชุดทดสอบจนใช้ทุกส่วนเป็นชุดทดสอบ ซึ่งเท่ากับทำการทดลอง k ครั้ง แล้วนำผลที่วัดได้จากทั้ง k ชุดการทดลองมาหาค่าเฉลี่ย ตัวอย่างเช่น การทำ 4-fold Cross validation เริ่มโดยทำการสุ่มแบ่งเอกสารในแต่ละประเภทเอกสารออกเป็น 4 ส่วน แล้วนำ 1 ส่วนจากแต่ละประเภทมารวมเป็นกลุ่มสำหรับทดสอบ แล้วใช้ 3 ส่วนที่เหลือจากแต่ละประเภทมารวมกันเป็นกลุ่มเรียนรู้ แสดงดังรูปที่ 4.2



รูปที่ 4.2 แสดงตัวอย่างการสร้างชุดเอกสารเรียนรู้และทดสอบจากกลุ่มเอกสาร 3 ประเภท A B และ C สำหรับการทำ 4-fold cross validation ในชุดการทดลองลำดับที่ 1 (fold 1)

4.2.3 รูปแบบของฟีเจอร์ที่ใช้ในการทดลอง

ในการทดลองจำแนกประเภทเอกสาร ได้ใช้ฟีเจอร์แบบต่างๆ ดังนี้

- คำเดี่ยว
- คำเดี่ยวและไบ-แกรม โดยสร้างไบ-แกรมจากลำดับของคำในแต่ละวลีที่สกัดได้
- คำเดี่ยวและวลี โดยวลีที่ใช้คือวลีที่สกัดได้
- ฟีเจอร์จากต้นไม้วลี คือวิธีการที่นำเสนอในงานวิจัยนี้ โดยฟีเจอร์จะถูกสร้างและทำการเลือกโดยใช้ต้นไม้วลี

4.3 ชุดเอกสารที่ใช้ในการทดลอง

เอกสารที่ใช้ในการทดลองมีจำนวน 2 ชุดเอกสารใหญ่ คือ

ก) ชุดเอกสาร PDDP [16] เป็นเอกสารเว็บ ซึ่งแบ่งเป็น 3 ชุดย่อย F-series, J-series และ K-series โดยแต่ละเอกสารอยู่ในรูปแบบเอกสาร HTML เป็นชุดเอกสารที่ถูกนำมาใช้ในงานวิจัยทางด้านการจำแนกประเภทเอกสาร

ข) ชุดข้อมูล Reuters-Top10 คัดเลือกมาจาก ชุดข้อมูลข่าว Reuters-21578 [18] 10 กลุ่มแรกที่มีจำนวนข่าวมากที่สุดจาก 135 กลุ่มเอกสาร ข่าว โดยแต่ละเอกสารอยู่ในรูปแบบ XML ซึ่งชุดข้อมูล Reuters-Top10 [19] เป็นชุดข้อมูลที่ใช้ในการทดสอบการจำแนกประเภทเอกสารและการจัดกลุ่มเอกสาร ที่ใช้กันอย่างแพร่หลาย

ชุดเอกสารทั้ง หมดจะผ่านกระบวนการเตรียมเอกสารก่อน ตัวอย่างเอกสาร และการเตรียมเอกสารของเอกสารแต่ละชุด แสดงในภาคผนวก ก

4.4 ชุดเอกสาร PDDP F-series ที่ใช้ในการทดลองและผลการทดลอง

4.4.1 ชุดเอกสาร PDDP F-series ที่ใช้ในการทดลอง

ชุดเอกสาร PDDP F-series มีเอกสารทั้งหมด 98 แบ่งเป็น 4 ประเภท อยู่ในรูปแบบแฟ้มข้อมูล HTML ซึ่งมีรายละเอียดอยู่ในภาคผนวก ก รายละเอียดของแต่ละ ประเภทเอกสารแสดงในตารางที่ 4.1 ในการทดลอง จำแนกประเภทเอกสาร ได้แบ่งชุดเอกสาร PDDP F-series เพื่อใช้ในการเรียนรู้ และใช้ในการทดสอบ สำหรับทำ 4-fold cross validation ได้จำนวนเอกสารสำหรับเรียนรู้และทดสอบในแต่ละกลุ่มการทดลองดังแสดงในตาราง 4.2

4.4.2 ผลการทดลองของชุด เอกสาร PDDP F-series เมื่อใช้ฟีเจอร์ที่ได้จากการเลือกฟีเจอร์โดยใช้ต้นไม้วลี

สร้างต้นไม้วลีและทำการเลือกฟีเจอร์ โดยใช้ต้นไม้วลีโดยกำหนดให้แต่ละเทอมมีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสาร โดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.1 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.2 และ 4.3 ตามลำดับ รายละเอียดของฟีเจอร์ที่ได้แสดงในภาคผนวก ข

ตารางที่ 4.1 แสดงจำนวนเอกสาร F-series แบ่งตามประเภทเอกสาร

ชื่อประเภท	จำนวนเอกสาร
Business & Finance (B)	22
Electronic Communication (E)	23
Labor (L)	26
Manufacturing (M)	27
รวม	**98**

ตารางที่ 4.2 แสดงจำนวนเอกสาร F-series ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชื่อประเภท (ตัวย่อ)	จำนวนเอกสารในชุดการทดลอง (fold) ที่							
	1		2		3		4	
	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ
B	16	6	16	6	17	5	17	5
E	17	6	17	6	17	6	18	5
L	20	6	20	6	19	7	19	7
M	21	6	20	7	20	7	20	7
รวม	74	24	73	25	73	25	74	24
	98		98		98		98	

ตารางที่ 4.3 แสดงจำนวนฟีเจอร์จากต้นไม้วลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP F-series

ชุดที่	จำนวนฟีเจอร์จากต้นไม้วลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	23,469	2,832
2	23,861	2,797
3	22,911	2,738
4	23,415	2,740
ค่าเฉลี่ย	23,414	2,777

ตารางที่ 4.4 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร PDDP F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	100.0	100.0	95.83	100.0	91.67	95.83	87.50	95.83	95.83	100.0
2	97.10	88.00	97.10	96.00	96.00	96.00	96.00	92.00	96.00	97.10
3	96.00	96.00	100.0	100.0	100.0	100.0	96.00	100.0	96.00	100.0
4	92.88	91.67	92.88	87.50	87.50	87.50	87.50	87.50	83.33	92.88
ค่าเฉลี่ย	96.50	93.92	96.45	95.88	93.79	94.83	91.75	93.83	92.79	97.50

ตารางที่ 4.5 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	1.0000	1.0000	0.9580	1.0000	0.9161	0.9580	0.8723	0.9580	0.9580	1.0000
2	0.9710	0.8709	0.9710	0.9600	0.9597	0.9597	0.9597	0.9224	0.9597	0.9710
3	0.9593	0.9600	1.0000	1.0000	1.0000	1.0000	0.9595	1.0000	0.9595	1.0000
4	0.9176	0.9095	0.9176	0.8713	0.8713	0.8713	0.8713	0.8713	0.8171	0.9176
ค่าเฉลี่ย	0.9620	0.9351	0.9617	0.9578	0.9368	0.9473	0.9157	0.9379	0.9236	0.9722

4.4.3 ผลการทดลองของชุดเอกสาร PDDP F-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว

สร้างฟีเจอร์แบบคำเดี่ยว จากทุกคำในทุกเอกสารในชุดเรียนรู้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.6 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.7 และ 4.8 ตามลำดับ

ตารางที่ 4.6 แสดงจำนวนฟีเจอร์แบบคำเดี่ยวที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP F-series

ชุดที่	จำนวนฟีเจอร์แบบคำเดี่ยว	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	5,861	2,676
2	5,948	2,664
3	5,680	2,597
4	5,943	2,678
ค่าเฉลี่ย	5,858	2,654

ตารางที่ 4.7 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร PDDP F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	100.0	100.0	91.67	95.83	91.67	91.67	87.50	87.50	87.50	100.0
2	96.00	96.00	96.00	96.00	92.00	92.00	96.00	96.00	96.00	96.00
3	100.0	100.0	100.0	100.0	100.0	100.0	96.00	96.00	96.00	100.0
4	91.67	87.50	91.67	87.50	91.67	87.50	87.50	87.50	83.33	91.67
ค่าเฉลี่ย	96.92	95.88	94.83	94.83	93.83	92.79	91.75	91.75	90.71	96.92

ตารางที่ 4.8 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	1.0000	1.0000	0.9161	0.9580	0.9161	0.9161	0.8723	0.8723	0.8723	1.0000
2	0.9600	0.9600	0.9600	0.9600	0.9224	0.9224	0.9597	0.9597	0.9597	0.9600
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9595	0.9595	0.9595	1.0000
4	0.9095	0.8713	0.9095	0.8713	0.9095	0.8713	0.8713	0.8713	0.8171	0.9095
ค่าเฉลี่ย	0.9674	0.9578	0.9464	0.9473	0.9370	0.9274	0.9157	0.9157	0.9022	0.9674

4.4.4 ผลการทดลองของชุดเอกสาร PDDP F-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว และไบ-แกรม

สร้างฟีเจอร์ที่เป็นคำเดี่ยว และไบ-แกรม โดยสร้างไบ-แกรมจากวลีที่สกัดได้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสาร โดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.9 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.10 และ 4.11 ตามลำดับ

ตารางที่ 4.9 แสดงจำนวนคำเดี่ยวและไบ-แกรมที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP F-series

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและไบ-แกรม	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	17,750	3,473
2	17,944	3,444
3	17,259	3,379
4	17,762	3,413
ค่าเฉลี่ย	17,679	3,427

ตารางที่ 4.10 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร PDDP F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	100.0	100.0	91.67	100.0	91.67	95.83	87.50	91.67	87.50	100.0
2	96.00	92.00	96.00	96.00	92.00	92.00	96.00	96.00	96.00	96.00
3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.00	100.0
4	91.67	87.50	91.67	87.50	87.50	87.50	87.50	87.50	83.33	91.67
ค่าเฉลี่ย	96.92	94.88	94.83	95.88	92.79	93.83	92.75	93.79	90.71	96.92

ตารางที่ 4.11 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร F-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	1.0000	1.0000	0.9161	1.0000	0.9161	0.9580	0.8723	0.9143	0.8723	1.0000
2	0.9600	0.9164	0.9600	0.9600	0.9224	0.9224	0.9597	0.9597	0.9597	0.9600
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9595	1.0000
4	0.9095	0.8713	0.9095	0.8713	0.8713	0.8713	0.8713	0.8713	0.8171	0.9095
ค่าเฉลี่ย	0.9674	0.9469	0.9464	0.9578	0.9274	0.9379	0.9258	0.9363	0.9022	0.9674

4.4.5ผลการทดลองของชุดเอกสาร PDDP F-series เมื่อใช้ฟีเจอร์คำเดี่ยวและวลี

สร้างฟีเจอร์จากวลีที่สกัดได้และคำเดี่ยว แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.12 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธีเค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.13 และ 4.14 ตามลำดับ

4.4.6 เปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แบบต่าง ของชุดเอกสาร F-series

รูปที่ 4.3 ถึง 4.15 แสดงการเปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แต่ละแบบ โดยเปรียบเทียบทั้งในด้านจำนวนฟีเจอร์ อัตราความถูกต้อง และค่า F-measure

ตารางที่ 4.12 แสดงจำนวนฟีเจอร์คำเดี่ยวและวลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP F-series

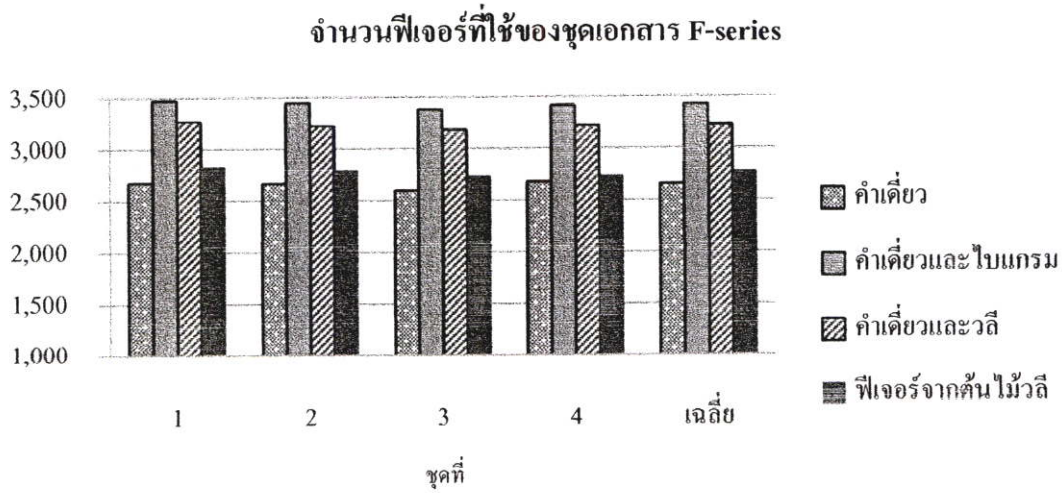
ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและวลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	15,569	3,270
2	15,618	3,222
3	15,159	3,185
4	15,576	3,220
ค่าเฉลี่ย	15,481	3,224

ตารางที่ 4.13 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร PDDP F-series

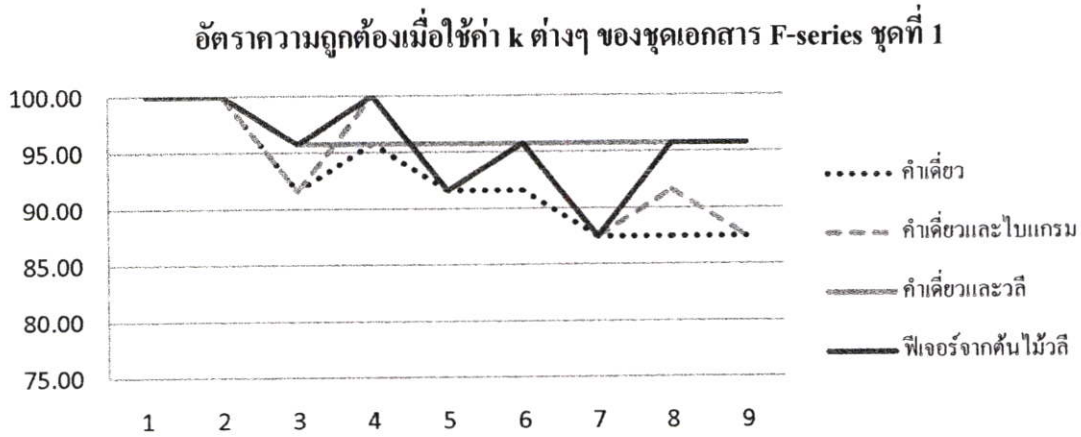
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	100.0	100.0	95.83	95.83	95.83	95.83	95.83	95.83	95.83	100.0
2	96.00	92.00	96.00	96.00	96.00	96.00	96.00	96.00	92.00	96.00
3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.00	100.0
4	87.50	87.50	87.50	87.50	87.50	87.50	87.50	87.50	87.50	87.50
ค่าเฉลี่ย	95.88	94.88	94.83	94.83	94.83	94.83	94.83	94.83	92.83	95.88

ตารางที่ 4.14 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร F-series

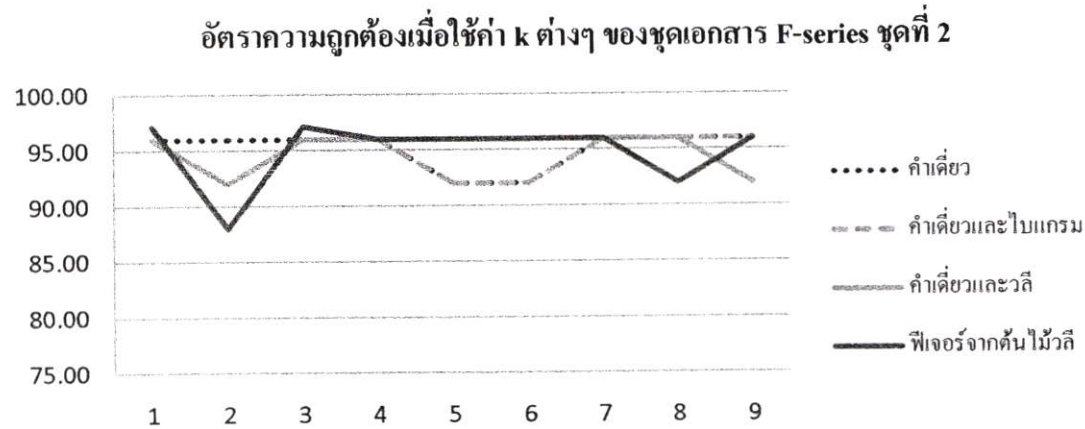
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	1.0000	1.0000	0.9580	0.9580	0.9580	0.9580	0.9580	0.9580	0.9580	1.0000
2	0.9600	0.9164	0.9600	0.9600	0.9600	0.9600	0.9600	0.9600	0.9224	0.9600
3	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9595	1.0000
4	0.8713	0.8713	0.8713	0.8713	0.8713	0.8713	0.8713	0.8713	0.8713	0.8713
ค่าเฉลี่ย	0.9578	0.9469	0.9473	0.9473	0.9473	0.9473	0.9473	0.9473	0.9278	0.9578



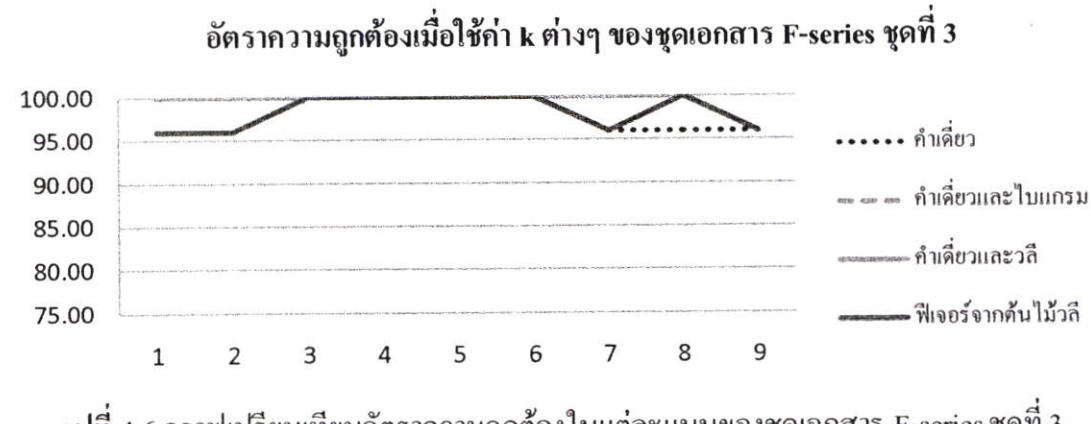
รูปที่ 4.3 แผนภูมิแท่งเปรียบเทียบจำนวนฟีเจอร์ที่ใช้ในแต่ละแบบของชุดเอกสาร F-series



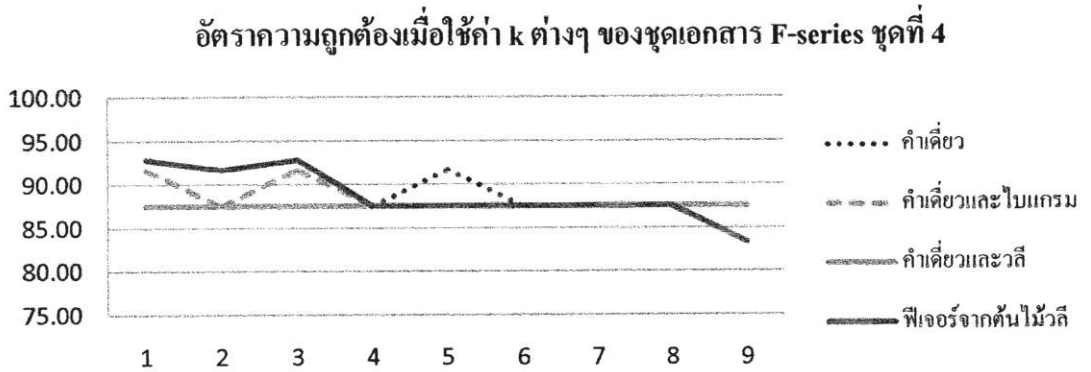
รูปที่ 4.4 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 1



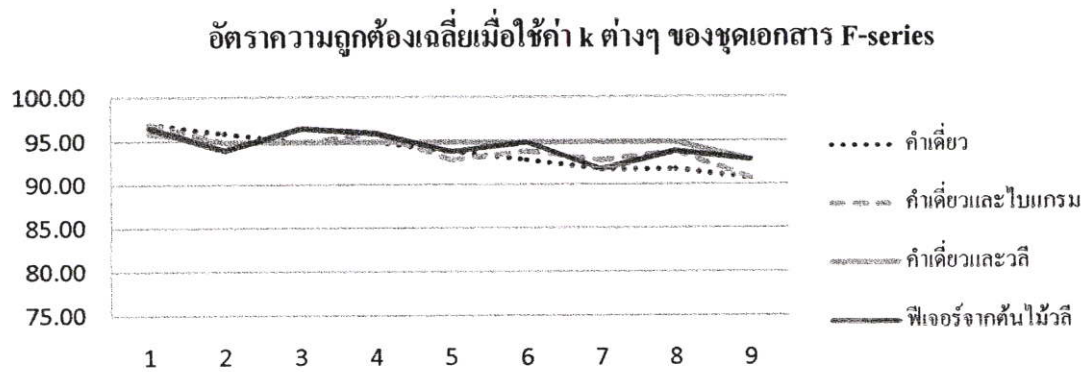
รูปที่ 4.5 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 2



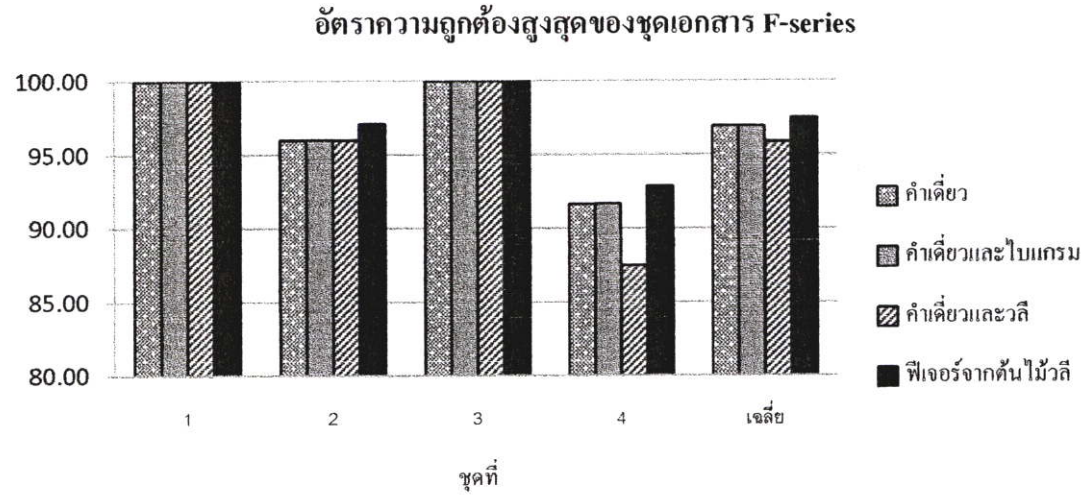
รูปที่ 4.6 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 3



รูปที่ 4.7 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 4

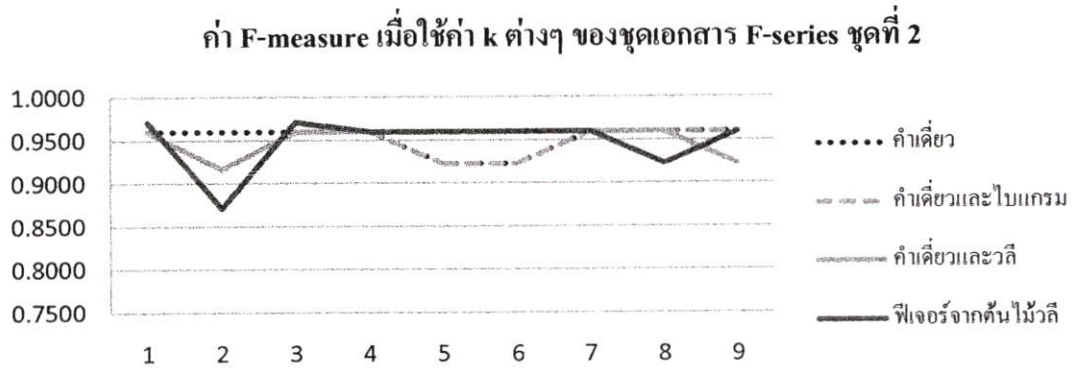


รูปที่ 4.8 กราฟเปรียบเทียบอัตราความถูกต้องเฉลี่ยของทุกชุดการทดลองในแต่ละแบบของชุดเอกสาร F-series

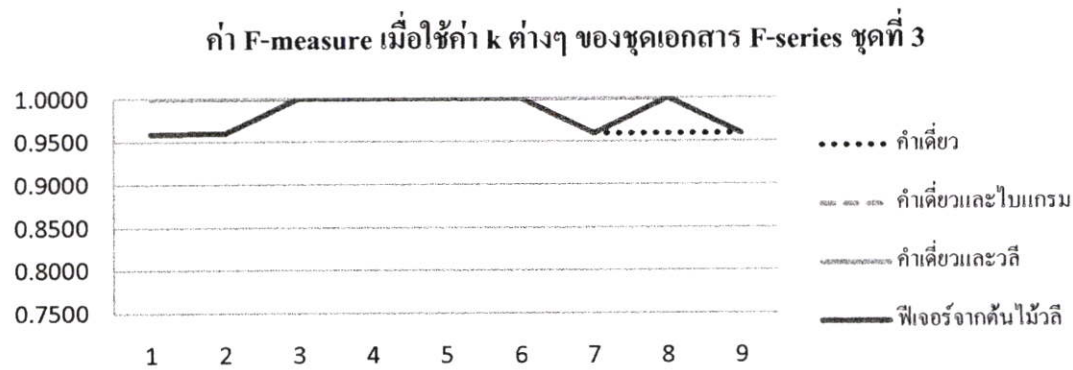


รูปที่ 4.9 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุดในแต่ละแบบของชุดเอกสาร F-series

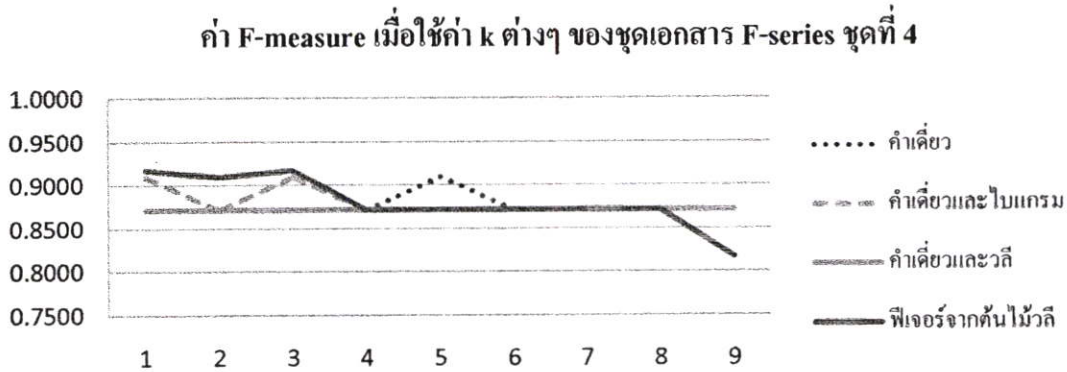
รูปที่ 4.10 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 1



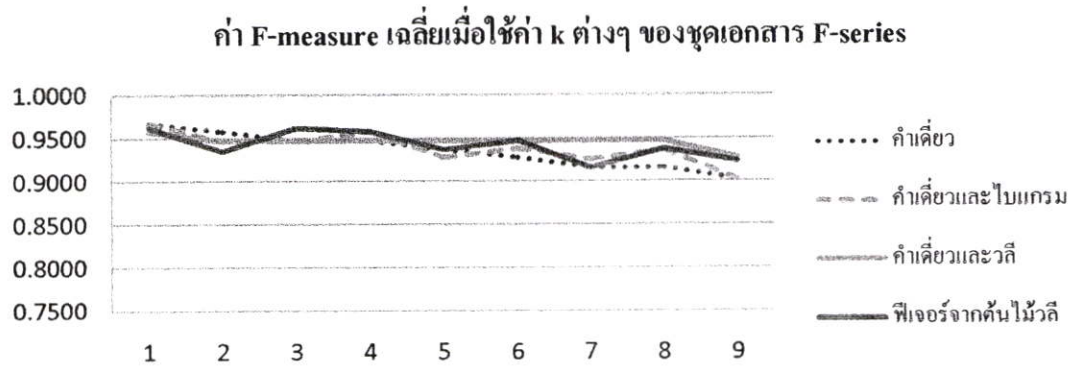
รูปที่ 4.11 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 2



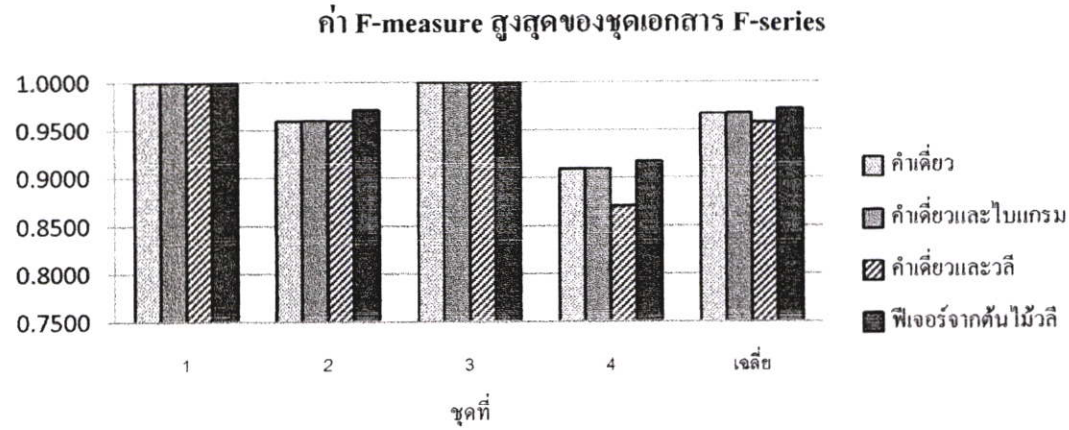
รูปที่ 4.12 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 3



รูปที่ 4.13 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร F-series ชุดที่ 4



รูปที่ 4.14 กราฟเปรียบเทียบ ค่า F-measure เฉลี่ยในแต่ละแบบของชุดเอกสาร F-series



รูปที่ 4.15 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุดในแต่ละแบบของชุดเอกสาร F-series

จากรูปที่ 4.3 จะเห็นได้ว่าฟีเจอร์ที่ได้จากต้นไม้วลีมีจำนวนน้อยกว่าฟีเจอร์แบบอื่นๆ ยกเว้นแบบคำเดี่ยว และเมื่อพิจารณาถึงความถูกต้องในการจำแนกประเภทเอกสาร จากรูปที่ 4.9 และ 4.15

ฟีเจอร์จากต้นไม้วลีให้อัตราความถูกต้องและค่า F-measure เมื่อเฉลี่ยจากทั้ง 4 ชุดสูงกว่าฟีเจอร์แบบอื่นๆ ซึ่งเป็นไปตามเป้าหมายของเราที่ต้องการได้ฟีเจอร์ที่มีประสิทธิภาพ โดยใช้วลีร่วมกับคำเดี่ยวที่ให้ความถูกต้องในการจำแนกประเภทเอกสารสูงแต่มีจำนวนน้อย

4.5 ชุดเอกสาร PDDP J-series ที่ใช้ในการทดลองและผลการทดลอง

4.5.1 ชุดเอกสาร PDDP J-series ที่ใช้ในการทดลอง

ชุดเอกสาร PDDP J-series มีเอกสารทั้งหมด 185 แบ่งเป็น 10 ประเภท ซึ่งในเอกสารชุดนี้ต่างจาก J-series ตรงที่ประเภทของเอกสารจะถูกแบ่งย่อยมากกว่า รายละเอียดของแต่ละประเภทเอกสารแสดงในตารางที่ 4.15

ตารางที่ 4.15 แสดงจำนวนเอกสาร J-series แบ่งตามประเภท

ชื่อประเภท	จำนวนเอกสาร	ชื่อประเภท	จำนวนเอกสาร
Affirmative Action (A)	20	Employee Rights (L)	16
Business Capital (B)	19	Materials Processing (M)	17
Information Systems (C)	19	Personal Management (P)	19
Electronic Commerce (E)	19	Manufacturing Systems (S)	18
Intellectual Property (I)	19	Industrial Partnership (Z)	19

ในการทดลองจำแนกประเภทเอกสารได้แบ่งชุดเอกสาร PDDP J-series เพื่อใช้ในการเรียนรู้ และใช้ในการทดสอบ สำหรับทำ 4-fold cross validation โดยการสุ่มเลือกเอกสารได้จำนวนเอกสารสำหรับเรียนรู้และทดสอบในแต่ละชุดการทดลองดังแสดงในตาราง 4.16 และทดลองเพื่อวัดประสิทธิภาพของฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีในการจำแนกประเภทเอกสารเปรียบเทียบกับฟีเจอร์แบบอื่นๆ โดยทำการทดลองสร้างฟีเจอร์ 4 แบบ สำหรับการจำแนกประเภท คือ 1.ฟีเจอร์จากต้นไม้วลี 2.คำเดี่ยว 3.คำเดี่ยวและไบ-แกรม และ 4.คำเดี่ยวและวลี

4.5.2 ผลการทดลองของชุดเอกสาร PDDP J-series เมื่อใช้ฟีเจอร์ที่ได้จากการเลือกฟีเจอร์โดยใช้ต้นไม้วลี

สร้างต้นไม้วลีและทำการเลือกฟีเจอร์โดยใช้ต้นไม้วลีโดยกำหนดให้แต่ละเทอมมีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.17 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วย

ฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.18 และ 4.19 ตามลำดับ รายละเอียดของฟีเจอร์ที่ได้แสดงในภาคผนวก ข

4.5.3 ผลการทดลองของชุดเอกสาร PDDP J-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว

สร้างฟีเจอร์แบบคำเดี่ยว จากทุกคำในทุกเอกสารในชุดเรียนรู้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.20 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.21 และ 4.22 ตามลำดับ

ตารางที่ 4.16 แสดงจำนวนเอกสาร J-series ที่ใช้ในการเรียนรู้และทดสอบในแต่ละชุดการทดลอง

ชื่อประเภท (ตัวย่อ)	จำนวนเอกสารในชุดการทดลอง (fold) ที่							
	1		2		3		4	
	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ
A	15	5	15	5	15	5	15	5
B	14	5	14	5	14	5	15	4
C	14	5	14	5	15	4	14	5
E	14	5	15	4	14	5	14	5
I	15	4	14	5	14	5	14	5
L	12	4	12	4	12	4	12	4
M	13	4	13	4	13	4	12	5
P	14	5	14	5	14	5	15	4
S	14	4	14	4	13	5	13	5
Z	14	5	14	5	14	5	15	4
รวม	139	46	139	46	138	47	139	46
	185		185		185		185	

ตารางที่ 4.17 แสดงจำนวนฟีเจอร์จากต้นไม้วลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP J-series

ชุดที่	จำนวนฟีเจอร์จากต้นไม้วลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	70,248	2,747
2	63,872	2,543
3	66,730	2,552
4	66,086	2,455
ค่าเฉลี่ย	66,734	2,574

ตารางที่ 4.18 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร PDDP J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	69.57	73.91	71.74	71.74	71.74	76.09	73.91	76.09	69.57	76.09
2	78.26	73.91	84.78	86.96	84.78	89.13	86.96	82.61	80.44	89.13
3	76.60	80.85	78.72	85.11	87.23	89.36	91.49	91.49	91.49	91.49
4	73.91	71.74	78.26	78.26	80.44	80.44	84.78	82.61	86.96	86.96
ค่าเฉลี่ย	74.58	75.10	78.38	80.52	81.05	83.75	84.29	83.20	82.11	85.92

ตารางที่ 4.19 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7015	0.7268	0.7005	0.7117	0.7167	0.7629	0.7358	0.7690	0.6742	0.7690
2	0.7416	0.7223	0.8409	0.8712	0.8460	0.8902	0.8685	0.8204	0.7932	0.8902
3	0.7575	0.7981	0.7864	0.8437	0.8676	0.8827	0.9134	0.9134	0.9134	0.9134
4	0.7342	0.6973	0.7856	0.7822	0.8001	0.7982	0.8457	0.8230	0.8698	0.8698
ค่าเฉลี่ย	0.7337	0.7361	0.7784	0.8022	0.8076	0.8335	0.8409	0.8315	0.8127	0.8606

ตารางที่ 4.20 แสดงจำนวนฟีเจอร์แบบคำเดี่ยวที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP J-series

ชุดที่	จำนวนฟีเจอร์แบบคำเดี่ยว	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	12,568	3,835
2	11,799	3,588
3	11,629	3,532
4	11,164	3,476
ค่าเฉลี่ย	11,790	3,608

ตารางที่ 4.21 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร PDDP J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	65.22	73.91	71.74	73.91	73.91	69.57	71.74	69.57	71.74	73.91
2	80.44	73.91	82.61	78.26	80.44	82.61	80.44	76.09	76.09	82.61
3	74.47	82.98	82.98	85.11	82.98	87.23	85.11	85.11	85.11	87.23
4	73.91	76.09	84.78	71.74	76.09	78.26	80.44	78.26	80.44	84.78
ค่าเฉลี่ย	73.51	76.72	80.53	77.26	78.35	79.42	79.43	77.26	78.34	82.14

ตารางที่ 4.22 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6506	0.7462	0.7069	0.7301	0.7295	0.6941	0.7071	0.6991	0.7138	0.7462
2	0.7888	0.7216	0.8251	0.7811	0.8023	0.8241	0.7958	0.7443	0.7343	0.8251
3	0.7403	0.8241	0.8192	0.8199	0.8002	0.8717	0.8476	0.8476	0.8439	0.8717
4	0.7178	0.7430	0.8492	0.6802	0.7451	0.7670	0.7872	0.7458	0.7872	0.8492
ค่าเฉลี่ย	0.7244	0.7587	0.8001	0.7528	0.7693	0.7892	0.7844	0.7592	0.7698	0.8230

4.5.4 ผลการทดลองของชุดเอกสาร PDDP J-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว และไบ-แกรม

สร้างฟีเจอร์ที่เป็นคำเดี่ยว และไบ-แกรม โดยสร้างไบ-แกรมจากวลีที่สกัดได้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.23 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.24 และ 4.25 ตามลำดับ

ตารางที่ 4.23 แสดงจำนวนคำเดี่ยวและไบ-แกรมที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP J-series

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและไบ-แกรม	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	49,872	5,632
2	45,610	4,957
3	47,220	4,897
4	46,766	4,782
ค่าเฉลี่ย	47,367	5,067

ตารางที่ 4.24 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร PDDP J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	65.22	78.26	67.39	69.57	76.09	67.39	71.74	69.57	69.57	78.26
2	80.44	76.09	80.44	82.61	82.61	80.44	80.44	78.26	78.26	82.61
3	74.47	82.98	82.98	80.85	80.85	87.23	87.23	89.36	89.36	89.36
4	76.09	76.09	82.61	71.74	73.91	78.26	78.26	76.09	78.26	82.61
ค่าเฉลี่ย	74.05	78.35	78.35	76.19	78.37	78.33	79.42	78.32	78.86	83.21

ตารางที่ 4.25 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร J-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6250	0.7788	0.6555	0.6953	0.7527	0.6656	0.7191	0.6977	0.7011	0.7788
2	0.7888	0.7444	0.7826	0.8078	0.8202	0.7861	0.7950	0.7663	0.7592	0.8202
3	0.7455	0.8241	0.8192	0.7807	0.7755	0.8717	0.8747	0.8960	0.8914	0.8960
4	0.7362	0.7430	0.8268	0.6935	0.7097	0.7516	0.7458	0.7333	0.7629	0.8268
ค่าเฉลี่ย	0.7239	0.7726	0.7710	0.7443	0.7645	0.7688	0.7837	0.7733	0.7787	0.8305

4.5.5 ผลการทดลองของชุดเอกสาร PDDP J-series เมื่อใช้ฟีเจอร์คำเดี่ยวและวลี

สร้างฟีเจอร์จากวลีที่สกัดได้และคำเดี่ยว แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.26 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธีเค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.27 และ 4.28 ตามลำดับ

4.5.6 เปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แบบต่าง ของชุดเอกสาร J-series

รูปที่ 4.16 ถึง 4.41 แสดงการเปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แต่ละแบบ โดยเปรียบเทียบทั้งในด้านจำนวนฟีเจอร์ อัตราความถูกต้อง และค่า F-measure

ตารางที่ 4.26 แสดงจำนวนฟีเจอร์คำเดี่ยวและวลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP J-serie

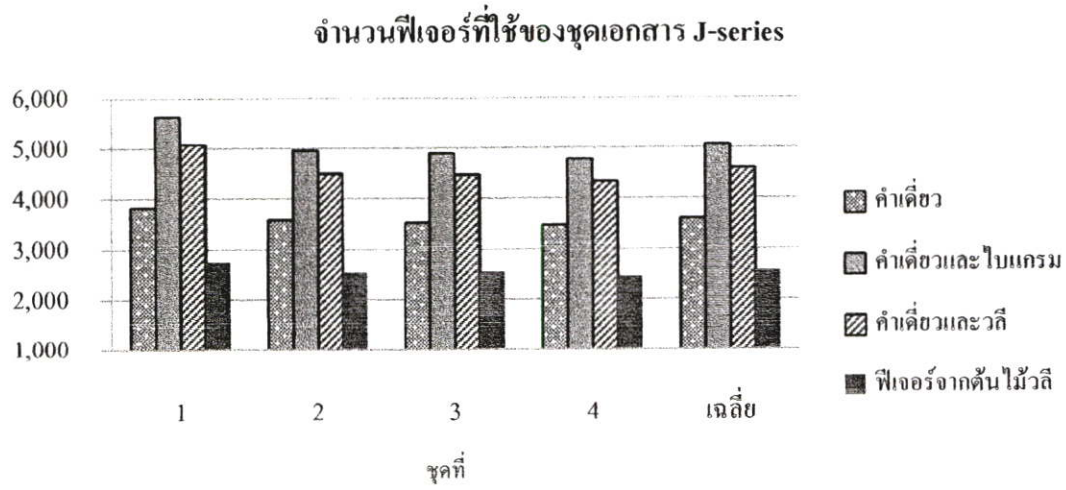
ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและวลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	44,694	5,082
2	40,581	4,504
3	42,333	4,473
4	41,654	4,341
ค่าเฉลี่ย	42,316	4,600

ตารางที่ 4.27 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร PDDP J-series

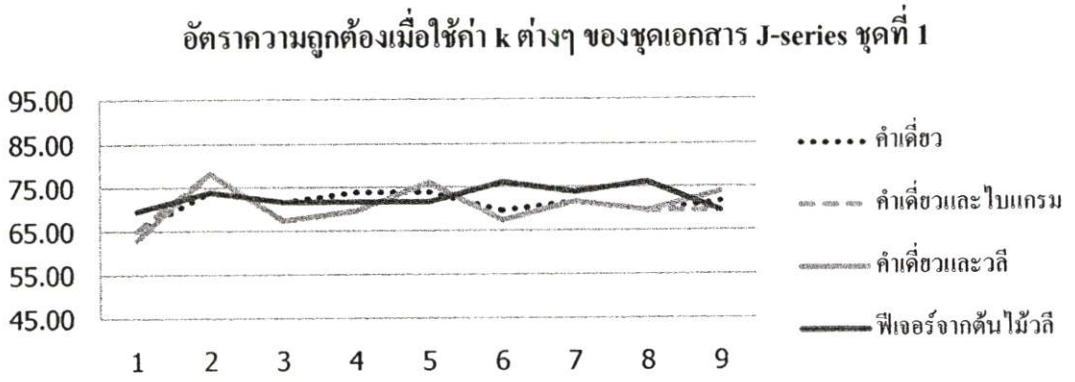
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	63.04	78.26	67.39	69.57	76.09	67.39	71.74	69.57	73.91	78.26
2	80.44	76.09	84.78	78.26	84.78	82.61	82.61	78.26	78.26	84.78
3	76.60	82.98	82.98	82.98	80.85	87.23	89.36	89.36	89.36	89.36
4	73.91	76.09	84.78	71.74	78.26	76.09	78.26	78.26	80.44	84.78
ค่าเฉลี่ย	73.50	78.35	79.98	75.64	80.00	78.33	80.49	78.86	80.49	84.30

ตารางที่ 4.28 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร J-series

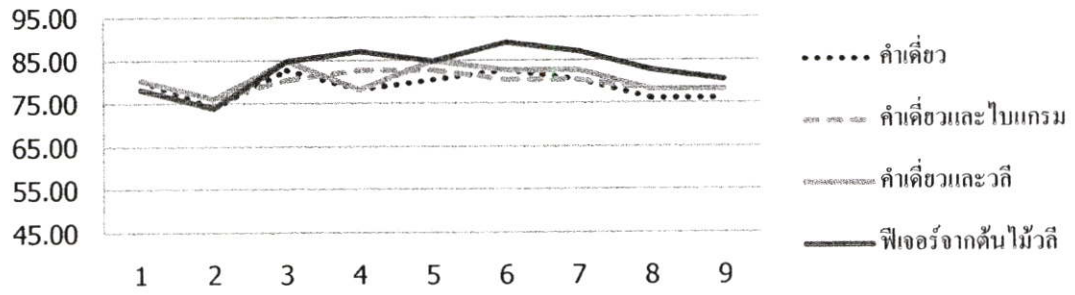
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6670	0.7807	0.7802	0.7312	0.7859	0.7943	0.7599	0.7784	0.7623	0.7943
2	0.7930	0.8870	0.8102	0.8669	0.8752	0.8221	0.7663	0.7663	0.7861	0.8870
3	0.7507	0.6915	0.7269	0.8261	0.7766	0.8062	0.7839	0.8927	0.8927	0.8927
4	0.7459	0.7976	0.7886	0.7969	0.7702	0.7702	0.7519	0.7068	0.7654	0.7976
ค่าเฉลี่ย	0.7391	0.7892	0.7765	0.8053	0.8020	0.7982	0.7655	0.7860	0.8016	0.8429



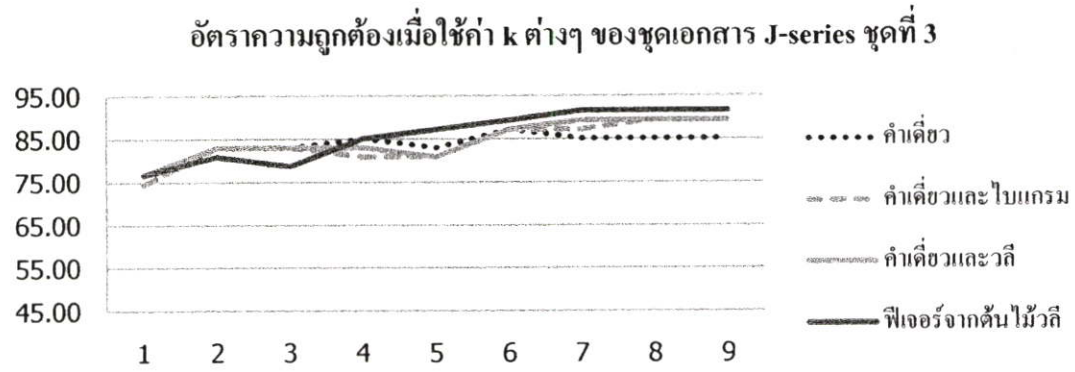
รูปที่ 4.16 แผนภูมิแท่งเปรียบเทียบจำนวนฟีเจอร์ที่ใช้ในแต่ละแบบของชุดเอกสาร J-series



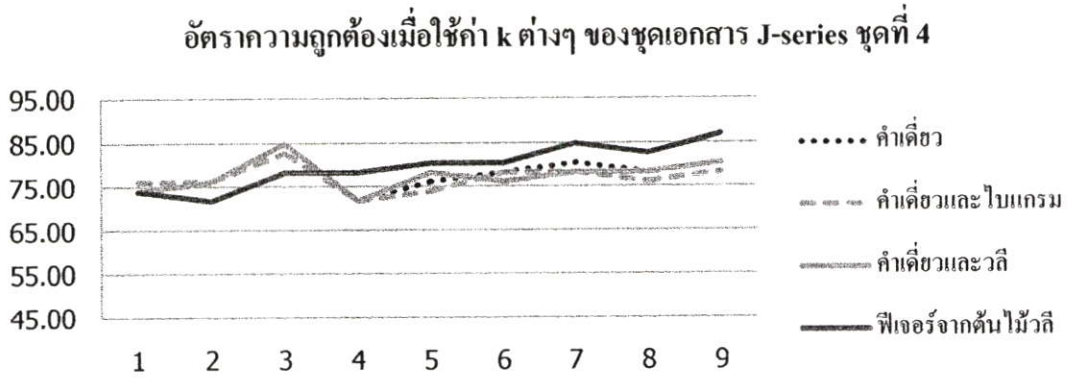
รูปที่ 4.17 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 1



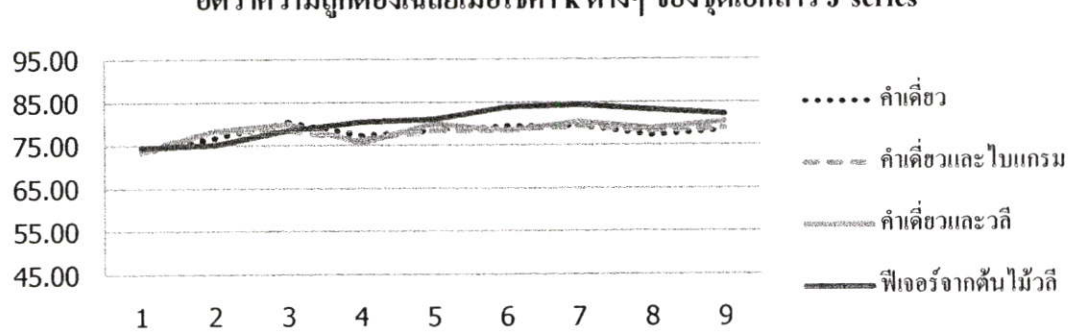
รูปที่ 4.18 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 2



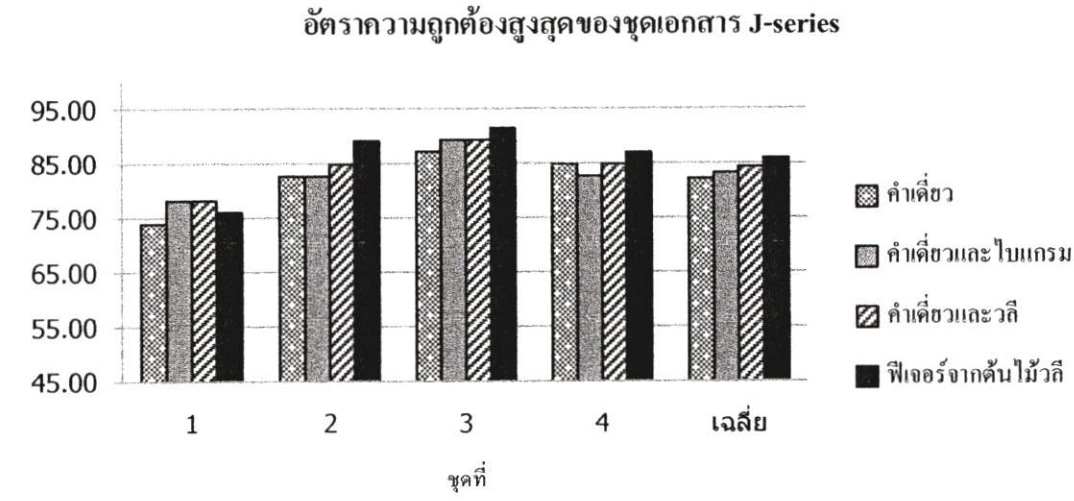
รูปที่ 4.19 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 3



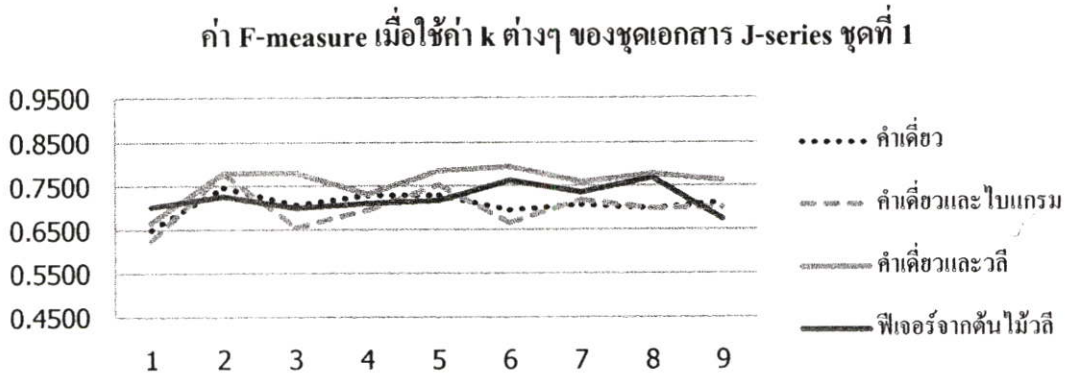
รูปที่ 4.20 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 4



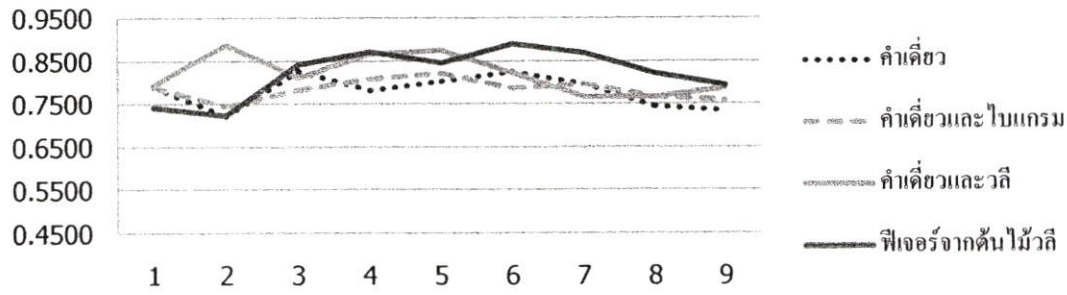
รูปที่ 4.21 กราฟเปรียบเทียบอัตราความถูกต้องเฉลี่ยในแต่ละแบบของชุดเอกสาร J-series



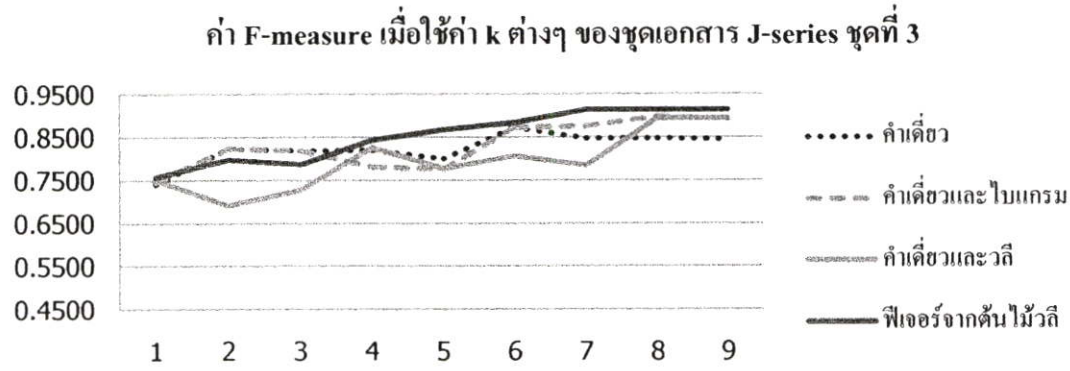
รูปที่ 4.22 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุดในแต่ละแบบของชุดเอกสาร J-series



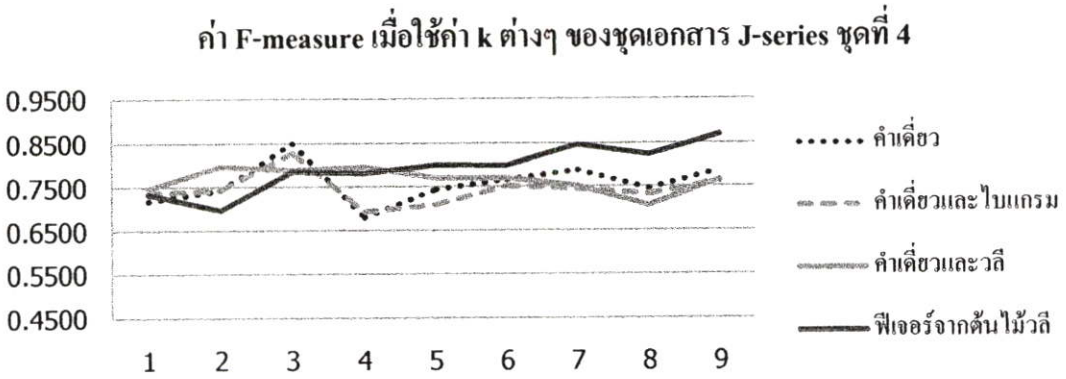
รูปที่ 4.23 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 1



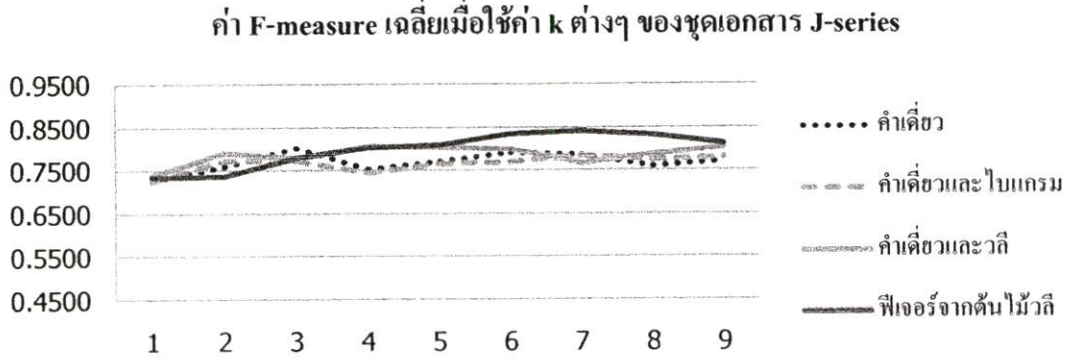
รูปที่ 4.24 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 2



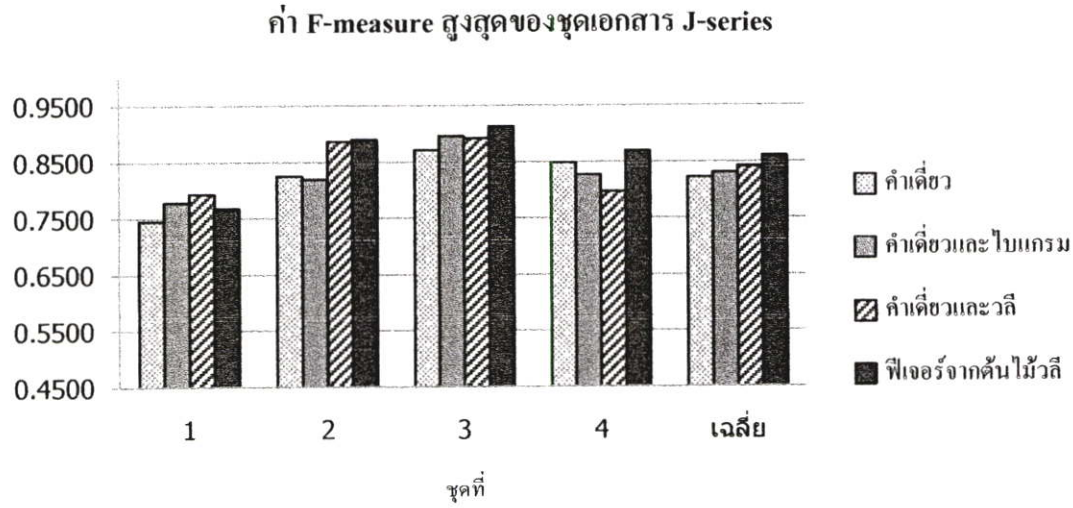
รูปที่ 4.25กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 3



รูปที่ 4.26กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร J-series ชุดที่ 4



รูปที่ 4.27 กราฟเปรียบเทียบ ค่า F-measure เฉลี่ยในแต่ละแบบของชุดเอกสาร J-series



รูปที่ 4.28 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุดในแต่ละแบบของชุดเอกสาร J-series

จากรูปที่ 4.16 จะเห็นได้ว่าฟีเจอร์ที่ได้จากต้นไม้วลีมีจำนวนน้อยกว่าฟีเจอร์แบบอื่นๆ และเมื่อพิจารณาถึงความถูกต้องในการจำแนกประเภทเอกสาร จากรูปที่ 4.22 และ 4.28 ฟีเจอร์จากต้นไม้วลีให้อัตราความถูกต้องและค่า F-measure เฉลี่ยจากทุกชุดการทดลองสูงกว่าฟีเจอร์แบบอื่นๆ ซึ่งเป็นไปตามเป้าหมายของเราที่ต้องการได้ฟีเจอร์ที่มีประสิทธิภาพ โดยใช้วลีร่วมกับคำเดี่ยว ที่ให้ความถูกต้องในการจำแนกประเภทเอกสารสูงแต่มีจำนวนน้อย

4.6 ชุดเอกสาร PDDP K-series ที่ใช้ในการทดลองและผลการทดลอง

4.6.1 ชุดเอกสาร PDDP K-series ที่ใช้ในการทดลอง

ชุดเอกสาร PDDP K-series มีเอกสารทั้งหมด 2,340 แบ่งเป็น 20 ประเภท ทำการสุ่มตัวอย่างแบบชั้นภูมิ โดยใช้อัตราส่วนในการเลือกเอกสารจากแต่ละชั้นภูมิเท่ากับ 0.1 โดยรายละเอียดในการสุ่มตัวอย่างเอกสารแสดงในภาคผนวก ก โดยในขั้นแรกสุ่มตัวอย่างเอกสารได้เอกสารทั้งหมด 796 เอกสาร และจากนั้นเลือกประเภทเอกสารที่มีจำนวนเอกสารน้อยออก คือ ประเภทเอกสาร Entertainment, Art, Cable, Media, Multimedia, Stage และTechnology ทำให้เหลือประเภทเอกสารทั้งหมด 13 ประเภท และเอกสารตัวอย่าง 387 เอกสาร

ในการทดลองจำแนกประเภทเอกสารได้แบ่งชุดเอกสาร PDDP K-series เพื่อใช้ในการเรียนรู้ และใช้ในการทดสอบ สำหรับทำ 4-fold cross validation โดยการสุ่มเลือกเอกสารได้จำนวนเอกสารสำหรับเรียนรู้และทดสอบในแต่ละชุดการทดลองดังแสดงในตาราง 4.29 และทดลองเพื่อวัดประสิทธิภาพของฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีในการจำแนกประเภทเอกสารเปรียบเทียบกับฟีเจอร์แบบอื่นๆ โดยทำการทดลองสร้างฟีเจอร์ 4 แบบ สำหรับการจำแนกประเภท คือ 1.ฟีเจอร์จากต้นไม้วลี 2.คำเดี่ยว 3.คำเดี่ยวและไบ-แกรม และ 4.คำเดี่ยวและวลี

4.6.2 ผลการทดลองของชุดเอกสาร PDDP K-series เมื่อใช้ฟีเจอร์ที่ได้จากการเลือกฟีเจอร์โดยใช้ต้นไม้วลี

สร้างต้นไม้วลีและทำการเลือกฟีเจอร์โดยใช้ต้นไม้วลีโดยกำหนดให้แต่ละเทอมมีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.30 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.31 และ 4.32 ตามลำดับ รายละเอียดของฟีเจอร์ที่ได้แสดงในภาคผนวก ข

4.6.3 ผลการทดลองของชุดเอกสาร PDDP K-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว

สร้างฟีเจอร์แบบคำเดี่ยว จากทุกคำในทุกเอกสารในชุดเรียนรู้ แล้วทำการเลือกฟีเจอร์ที่ มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.33 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.34 และ 4.35 ตามลำดับ

ตารางที่ 4.29 แสดงจำนวนเอกสาร K-series ที่ใช้ในการเรียนรู้ และทดสอบในแต่ละชุดการทดลอง

ชื่อประเภท	จำนวนเอกสารในชุดการทดลอง (fold) ที่							
	1		2		3		4	
	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ
Business	13	5	13	5	14	4	14	4
Culture	9	3	9	3	9	3	9	3
Film	39	13	39	13	39	13	39	13
Industry	7	2	7	2	6	3	7	2
Music	16	5	16	5	15	6	16	5
Online	6	2	6	2	6	2	6	2
People	36	12	36	12	36	12	36	12
Review	22	8	22	8	23	7	23	7
Television	25	8	25	8	24	9	25	8
Variety	7	3	8	2	7	3	8	2
Health	85	28	84	29	85	28	85	28
Politics	11	3	11	3	10	4	10	4
Sports	15	4	14	5	14	5	14	5
รวม	291	96	290	97	288	99	292	95
	387		387		387		387	

ตารางที่ 4.30 แสดงจำนวนฟีเจอร์จากต้นไม้วลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP K-series

ชุดที่	จำนวนฟีเจอร์จากต้นไม้วลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	40,802	2,685
2	41,616	2,745
3	41,945	2,721
4	42,597	2,900
ค่าเฉลี่ย	41,740	2,763

ตารางที่ 4.31 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร PDDP K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	79.17	77.08	76.04	77.08	78.13	83.33	81.25	83.33	86.46	86.46
2	72.17	73.20	75.26	75.26	76.29	76.29	75.26	79.38	81.44	81.44
3	77.32	73.20	74.23	74.23	72.17	74.23	77.32	74.23	74.23	77.32
4	76.29	81.44	82.47	84.54	86.60	85.57	85.57	84.54	84.54	86.60
ค่าเฉลี่ย	76.24	76.23	77.00	77.78	78.29	79.85	79.85	80.37	81.67	82.96

ตารางที่ 4.32 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 เมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7938	0.7694	0.7604	0.7723	0.7749	0.8168	0.7959	0.8164	0.8499	0.8499
2	0.7160	0.7180	0.7564	0.7515	0.7575	0.7545	0.7464	0.7908	0.8132	0.8132
3	0.7715	0.7339	0.7467	0.7424	0.7273	0.7489	0.7734	0.7374	0.7422	0.7734
4	0.7519	0.8162	0.8267	0.8480	0.8692	0.8611	0.8611	0.8396	0.8394	0.8692
ค่าเฉลี่ย	0.7583	0.7594	0.7725	0.7786	0.7822	0.7953	0.7942	0.7961	0.8112	0.8264

ตารางที่ 4.33 แสดงจำนวนฟีเจอร์แบบคำเดี่ยวที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP K-series

ชุดที่	จำนวนฟีเจอร์แบบคำเดี่ยว	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	10,520	3,672
2	10,650	3,618
3	10,867	3,777
4	10,872	3,803
ค่าเฉลี่ย	10,727	3,718

ตารางที่ 4.34 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร PDDP K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	76.04	77.08	73.96	79.17	83.33	84.38	83.33	82.29	83.33	84.38
2	64.95	70.10	72.17	70.10	71.13	74.23	72.17	73.20	75.26	75.26
3	71.13	76.29	79.38	74.23	76.29	80.41	78.35	76.29	75.26	80.41
4	67.01	71.13	76.29	80.41	79.38	81.44	80.41	83.51	82.47	83.51
ค่าเฉลี่ย	69.78	73.65	75.45	75.98	77.53	80.11	78.57	78.82	79.08	80.89

ตารางที่ 4.35 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 เมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7655	0.7595	0.7352	0.7762	0.8156	0.8237	0.8114	0.8018	0.8192	0.8237
2	0.6513	0.6885	0.7145	0.7036	0.7159	0.7415	0.7123	0.7245	0.7372	0.7415
3	0.7032	0.7611	0.7913	0.7418	0.7679	0.8005	0.7833	0.7641	0.7510	0.8005
4	0.6528	0.7114	0.7709	0.8112	0.8067	0.8174	0.8176	0.8404	0.8229	0.8404
ค่าเฉลี่ย	0.6932	0.7301	0.7530	0.7582	0.7765	0.7958	0.7812	0.7827	0.7825	0.8015

4.6.4 ผลการทดลองของชุดเอกสาร PDDP K-series เมื่อใช้ฟีเจอร์แบบคำเดี่ยว และไบ-แกรม

สร้างฟีเจอร์ที่เป็นคำเดี่ยว และไบ-แกรม โดยสร้างไบ-แกรมจากวลีที่สกัดได้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.36 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.37 และ 4.38 ตามลำดับ

ตารางที่ 4.36 แสดงจำนวนคำเดี่ยวและไบ-แกรมที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP K-series

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและไบ-แกรม	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	29,870	4,497
2	30,520	4,430
3	30,902	4,575
4	31,281	4,661
ค่าเฉลี่ย	30,643	4,541

ตารางที่ 4.37 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร PDDP K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	76.04	77.08	75.00	80.21	82.29	83.33	84.38	83.33	83.33	84.38
2	67.01	73.20	72.17	71.13	74.23	73.20	72.17	75.26	73.20	75.26
3	70.10	78.35	78.35	76.29	76.29	80.41	80.41	78.35	77.32	80.41
4	64.95	73.20	80.41	83.51	81.44	82.47	78.35	83.51	83.51	83.51
ค่าเฉลี่ย	69.53	75.46	76.48	77.78	78.56	79.85	78.83	80.11	79.34	80.89

ตารางที่ 4.38 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร K-series

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7633	0.7685	0.7402	0.7857	0.8062	0.8135	0.8226	0.8117	0.8193	0.8226
2	0.6739	0.7155	0.7208	0.7135	0.7440	0.7203	0.7114	0.7428	0.7200	0.7440
3	0.6978	0.7844	0.7810	0.7644	0.7661	0.8065	0.8053	0.7852	0.7730	0.8065
4	0.6325	0.7313	0.8053	0.8391	0.8243	0.8291	0.7970	0.8468	0.8452	0.8468
ค่าเฉลี่ย	0.6919	0.7499	0.7619	0.7757	0.7851	0.7923	0.7841	0.7966	0.7893	0.8050

4.6.5ผลการทดลองของชุดเอกสาร PDDP K-series เมื่อใช้ฟีเจอร์คำเดี่ยวและวลี

สร้างฟีเจอร์จากวลีที่สกัดได้และคำเดี่ยว แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 2 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.39 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธีเค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.40 และ 4.41 ตามลำดับ

ตารางที่ 4.39 แสดงจำนวนฟีเจอร์คำเดี่ยวและวลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร PDDP K-serie

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและวลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	25,063	4,201
2	25,607	4,177
3	25,894	4,280
4	26,213	4,390
ค่าเฉลี่ย	25,694	4,262

ตารางที่ 4.40 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร PDDP K-series

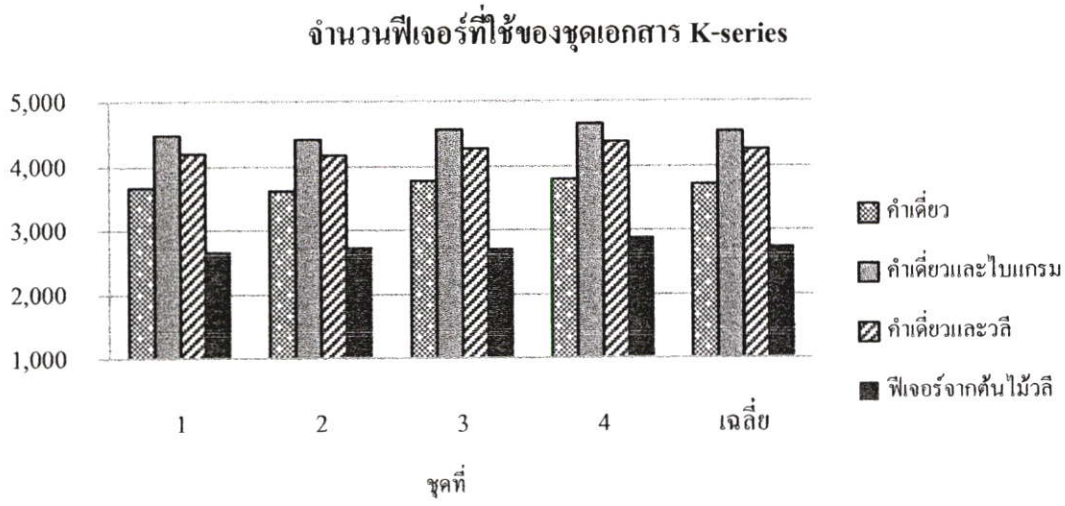
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	77.08	73.96	77.08	81.25	82.29	80.21	80.21	81.25	79.17	82.29
2	68.04	68.04	72.17	77.32	76.29	79.38	79.38	80.41	80.41	80.41
3	63.92	73.20	72.17	72.17	76.29	73.20	76.29	74.23	74.23	76.29
4	64.95	73.20	79.38	80.41	81.44	77.32	81.44	81.44	82.47	82.47
ค่าเฉลี่ย	68.50	72.10	75.20	77.79	79.08	77.53	79.33	79.33	79.07	80.37

ตารางที่ 4.41 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร K-series

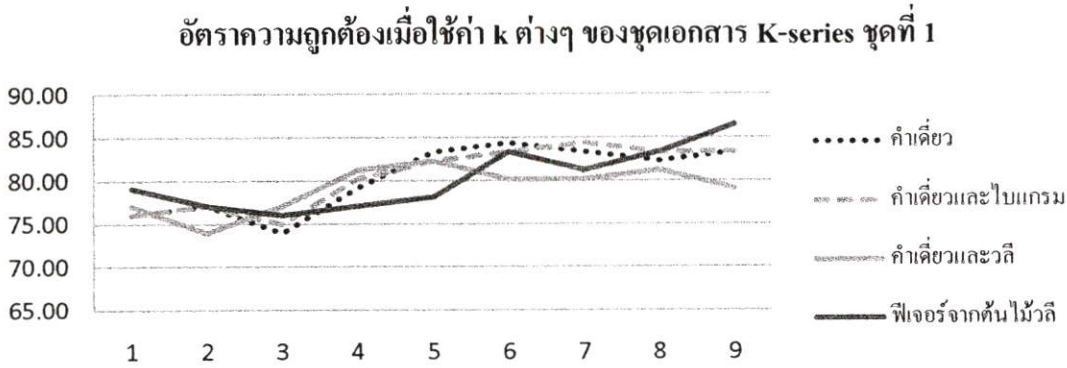
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7633	0.7300	0.7609	0.7973	0.8115	0.7840	0.7850	0.7970	0.7794	0.8115
2	0.6892	0.6718	0.7270	0.7694	0.7680	0.7921	0.7891	0.7974	0.7950	0.7974
3	0.6362	0.7228	0.7235	0.7108	0.7593	0.7303	0.7600	0.7368	0.7441	0.7600
4	0.6364	0.7138	0.7932	0.8028	0.8137	0.7707	0.8100	0.8139	0.8235	0.8235
ค่าเฉลี่ย	0.6813	0.7096	0.7512	0.7701	0.7881	0.7693	0.7860	0.7863	0.7855	0.7981

4.6.6 เปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แบบต่าง ชองชุดเอกสาร K-series

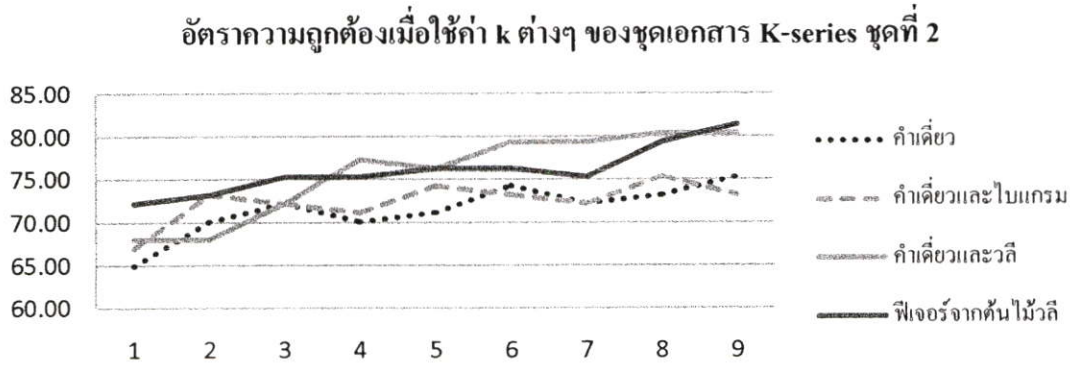
รูปที่ 4.29 ถึง 4.41 แสดงการเปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แต่ละแบบ โดยเปรียบเทียบทั้งในด้านจำนวนฟีเจอร์ อัตราความถูกต้อง และค่า F-measure



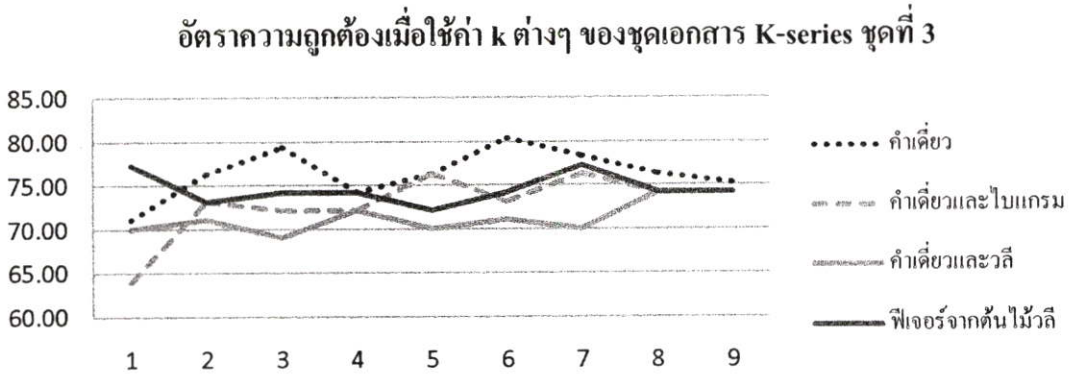
รูปที่ 4.29 แผนภูมิแท่งเปรียบเทียบจำนวนฟีเจอร์ที่ใช้ในแต่ละแบบของชุดเอกสาร K-series



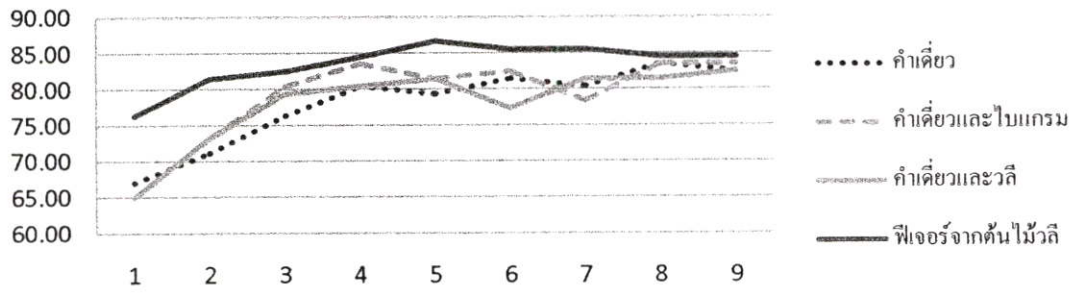
รูปที่ 4.30 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 1



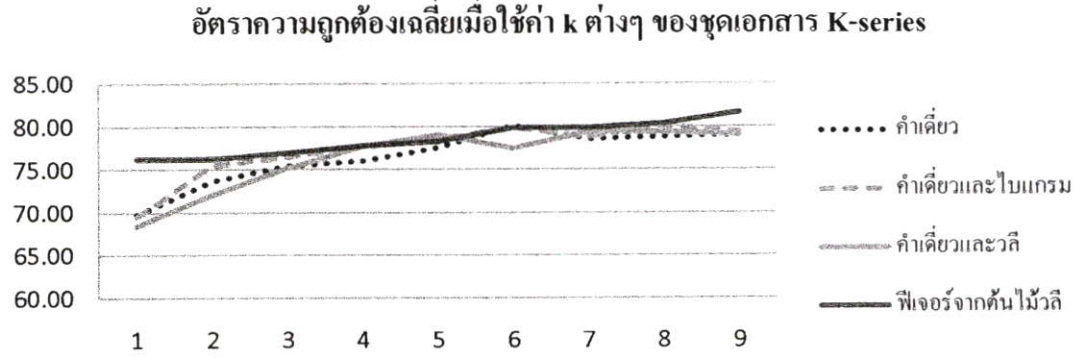
รูปที่ 4.31 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 2



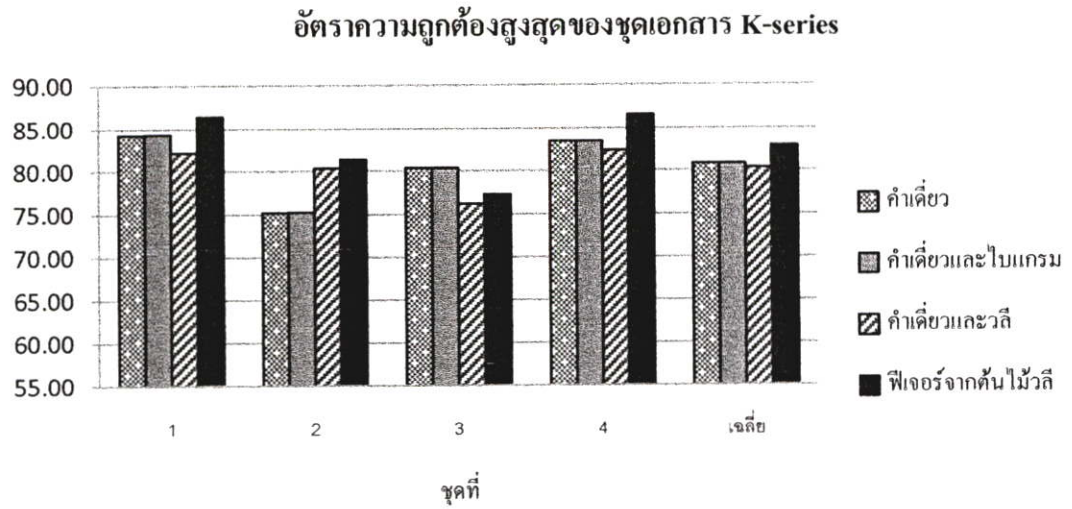
รูปที่ 4.32 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 3



รูปที่ 4.33 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 4

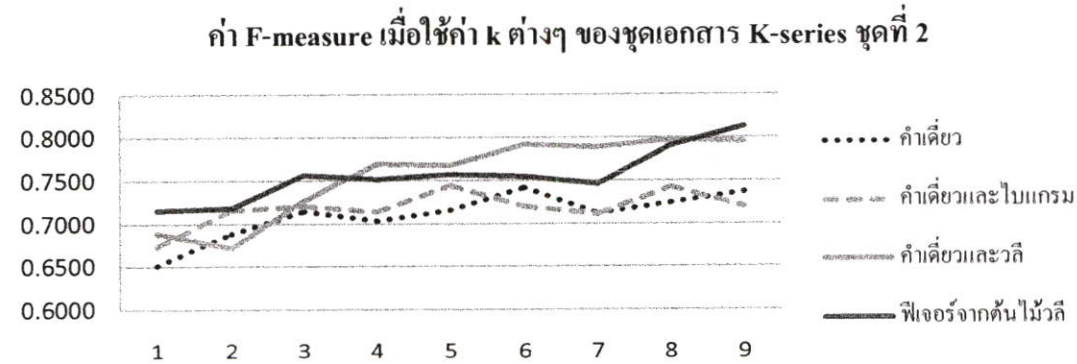


รูปที่ 4.34 กราฟเปรียบเทียบอัตราความถูกต้องเฉลี่ยในแต่ละแบบของชุดเอกสาร K-series

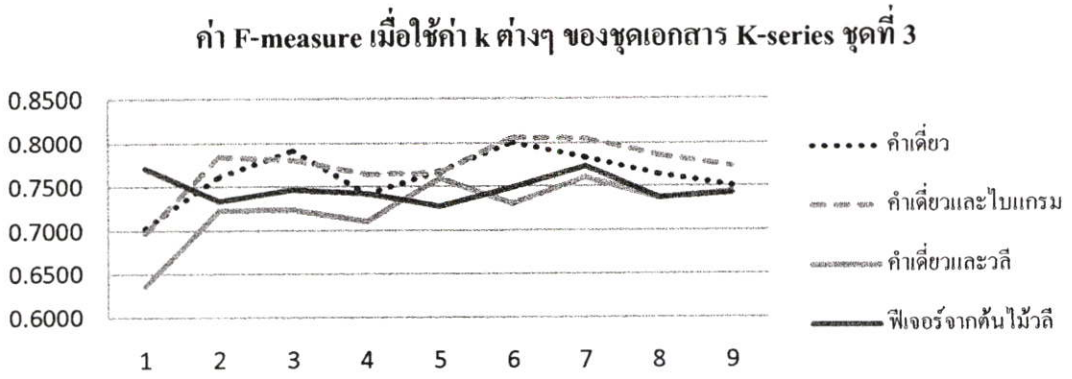


รูปที่ 4.35 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุดในแต่ละแบบของชุดเอกสาร K-series

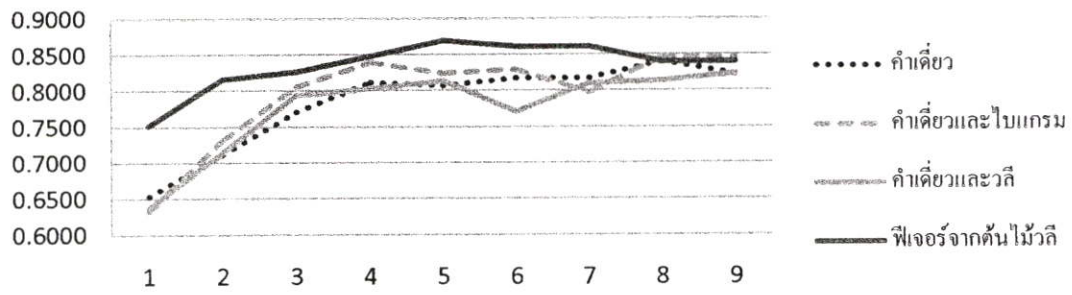
รูปที่ 4.36 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 1



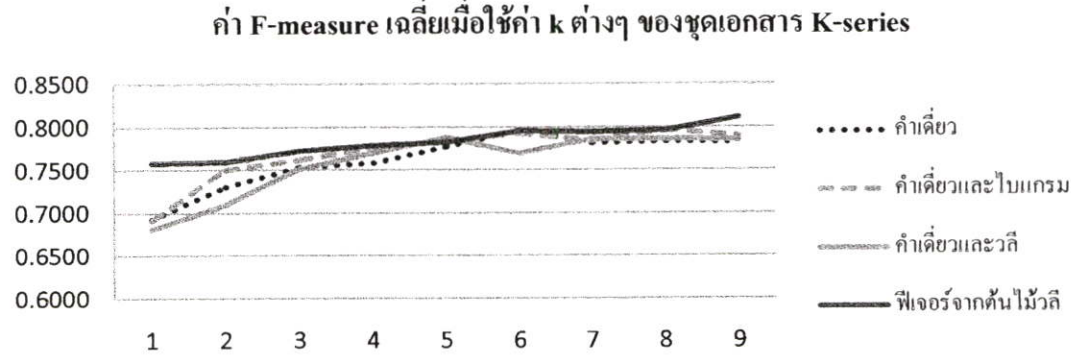
รูปที่ 4.37 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 2



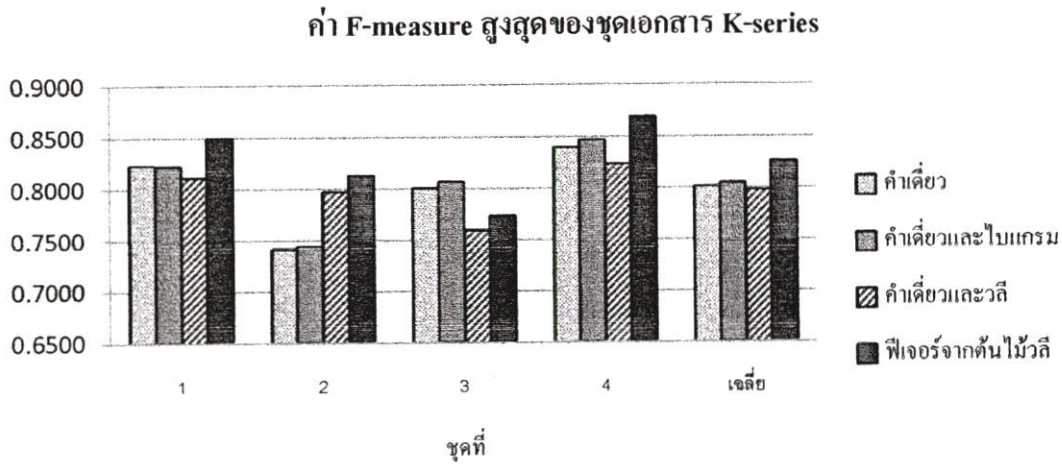
รูปที่ 4.38 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 3



รูปที่ 4.39 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร K-series ชุดที่ 4



รูปที่ 4.40 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร K-series



รูปที่ 4.41 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุดในแต่ละแบบของชุดเอกสาร K-series

จากรูปที่ 4.29 จะเห็นได้ว่าฟีเจอร์ที่ได้จากต้นไม้วลีมีจำนวนน้อยกว่าฟีเจอร์แบบอื่นๆ และเมื่อพิจารณาถึงความถูกต้องในการจำแนกประเภทเอกสาร จากรูปที่ 4.35 และ 4.41 ในทุกชุดการทดลอง ฟีเจอร์จากต้นไม้วลีให้อัตราความถูกต้องและค่า F-measure เฉลี่ยสูงกว่าฟีเจอร์แบบอื่นๆ ซึ่งเป็นไปตามเป้าหมายของเราที่ต้องการได้ฟีเจอร์ที่มีประสิทธิภาพ โดยใช้วลีร่วมกับคำเดี่ยว ที่ให้ความถูกต้องในการจำแนกประเภทเอกสารสูงแต่มีจำนวนน้อย

4.7 ชุดเอกสาร Reuters-top10ที่ใช้ในการทดลองและผลการทดลอง

4.7.1 ชุดเอกสาร Reuters-top10ที่ใช้ในการทดลอง

ชุดข้อมูล Reuters-Top10 คัดเลือกมาจาก ชุดข้อมูลข่าว Reuters-21578 [18] ซึ่ง David D. Lewis และคณะได้รวบรวมไว้ ซึ่งมีประเภทเอกสารอยู่ 135 ประเภทเอกสาร จำนวน 21,578 เอกสารข่าว โดย Reuters-Top10 คัดเลือกเอาเฉพาะเอกสารข่าว 10 กลุ่มแรกที่มีจำนวนข่าวมากที่สุดจาก 135 กลุ่มเอกสารข่าว ซึ่งชุดข้อมูล Reuters-Top10 [19] เป็นชุดข้อมูลที่ใช้ในการทดสอบการจำแนกประเภทเอกสารและการจัดกลุ่มเอกสาร ที่ใช้กันอย่างแพร่หลาย ลักษณะการเก็บอยู่ในรูปแบบแฟ้มข้อมูล XML มี 10 กลุ่มเอกสาร จำนวน 2,770 เอกสาร ทำการสุ่มตัวอย่างแบบชั้นภูมิ โดยใช้อัตราส่วนในการเลือกเอกสารจากแต่ละชั้นภูมิเท่ากับ 0.1 โดยรายละเอียดในการสุ่มตัวอย่างเอกสารแสดงในภาคผนวก ก โดยได้เอกสารตัวอย่างทั้งหมด 504

ในการทดลองจำแนกประเภทเอกสารได้แบ่งชุดเอกสาร Reuters-top10 เพื่อใช้ในการเรียนรู้ และใช้ในการทดสอบ สำหรับทำ 4-fold cross validation โดยการสุ่มเลือกเอกสารได้จำนวนเอกสารสำหรับเรียนรู้และทดสอบในแต่ละชุดการทดลองดังแสดงในตาราง 4.42 และทดลองเพื่อวัดประสิทธิภาพของฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีในการจำแนกประเภทเอกสาร

เปรียบเทียบกับฟีเจอร์แบบอื่นๆ โดยทำการทดลองสร้างฟีเจอร์ 4 แบบ สำหรับการจำแนกประเภท คือ 1.ฟีเจอร์จากต้นไม้วลี 2.คำเดี่ยว 3.คำเดี่ยวและไบ-แกรม และ 4.คำเดี่ยวและวลี

ตารางที่ 4.42 แสดงจำนวนเอกสาร Reuters-top10ที่ใช้ในการเรียนรู้ และทดสอบในแต่ละชุดการทดลอง

ชื่อประเภท	จำนวนเอกสารในชุดการทดลอง (fold) ที่							
	1		2		3		4	
	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ	เรียนรู้	ทดสอบ
acq (a)	102	33	101	34	101	34	101	34
corn (c)	9	3	9	3	9	3	9	3
crude (d)	28	10	28	10	29	9	29	9
earn (e)	127	42	127	42	127	42	126	43
grain (g)	24	7	23	8	23	8	23	8
interest (i)	19	6	19	6	18	7	19	6
money-fx (m)	27	10	28	9	28	9	28	9
ship (s)	14	5	15	4	14	5	14	5
trade (t)	18	6	18	6	18	6	18	6
wheat (w)	11	3	10	4	10	4	11	3
รวม	379	125	378	126	377	127	378	126
	504		504		504		504	

4.7.2 ผลการทดลองของชุดเอกสาร Reuters-top10เมื่อใช้ฟีเจอร์ที่ได้จากการเลือกฟีเจอร์โดยใช้ต้นไม้วลี

สร้างต้นไม้วลีและทำการเลือกฟีเจอร์โดยใช้ต้นไม้วลีโดยกำหนดให้แต่ละเทอมมีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.43 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.44 และ 4.45 ตามลำดับ รายละเอียดของฟีเจอร์ที่ได้แสดงในภาคผนวก ข

4.7.3 ผลการทดลองของชุดเอกสาร Reuters-top10 เมื่อใช้ฟีเจอร์แบบคำเดี่ยว

สร้างฟีเจอร์แบบคำเดี่ยว จากทุกคำในทุกเอกสารในชุดเรียนรู้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.46 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.47 และ 4.48 ตามลำดับ

ตารางที่ 4.43 แสดงจำนวนฟีเจอร์จากต้นไม้วลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร Reuters-top10

ชุดที่	จำนวนฟีเจอร์จากต้นไม้วลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	15,781	2,095
2	15,788	2,036
3	15,610	2,072
4	15,607	2,195
ค่าเฉลี่ย	15,638	2,100

ตารางที่ 4.44 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	70.40	76.80	72.80	77.60	80.00	79.20	80.00	76.80	80.00	80.00
2	77.78	82.54	80.16	80.16	81.75	82.54	83.33	82.54	84.92	84.92
3	74.02	76.38	74.80	77.17	79.53	79.53	81.10	79.53	78.74	81.10
4	72.22	71.43	74.60	73.81	73.81	76.19	75.40	76.19	76.19	76.19
ค่าเฉลี่ย	73.60	76.79	75.59	77.18	78.77	79.36	79.96	78.76	79.96	80.55

ตารางที่ 4.45 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์จากต้นไม้วลี ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.7013	0.7597	0.7226	0.7619	0.7840	0.7826	0.7804	0.7498	0.7877	0.7877
2	0.7764	0.8111	0.7941	0.7984	0.8149	0.8084	0.8226	0.8140	0.8353	0.8353
3	0.7336	0.7555	0.7344	0.7651	0.7887	0.7848	0.8041	0.7886	0.7731	0.8041
4	0.7115	0.7006	0.7408	0.7285	0.7244	0.7556	0.7471	0.7545	0.7520	0.7556
ค่าเฉลี่ย	0.7307	0.7567	0.7480	0.7635	0.7780	0.7828	0.7885	0.7767	0.7870	0.7956

ตารางที่ 4.46 แสดงจำนวนฟีเจอร์แบบคำเดี่ยวที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร Reuters-top10

ชุดที่	จำนวนฟีเจอร์แบบคำเดี่ยว	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	4,103	1,468
2	4,135	1,459
3	4,114	1,438
4	4,084	1,498
ค่าเฉลี่ย	4,109	1,466

ตารางที่ 4.47 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	65.60	71.20	66.40	72.00	69.60	71.20	71.20	73.60	73.60	73.60
2	73.02	81.75	76.19	80.95	78.57	80.95	80.95	80.16	77.78	81.75
3	71.65	73.23	72.44	74.80	75.59	77.17	76.38	76.38	76.38	77.17
4	69.05	70.64	73.81	72.22	72.22	73.02	70.64	72.22	73.02	73.81
ค่าเฉลี่ย	69.83	74.20	72.21	74.99	74.00	75.58	74.79	75.59	75.19	76.58

ตารางที่ 4.48 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยว ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6480	0.6975	0.6513	0.7021	0.6783	0.6896	0.6873	0.7162	0.7134	0.7162
2	0.7262	0.8062	0.7525	0.7965	0.7742	0.7973	0.7927	0.7869	0.7571	0.8062
3	0.7031	0.7049	0.7001	0.7284	0.7363	0.7476	0.7432	0.7432	0.7442	0.7476
4	0.6742	0.6852	0.7250	0.7137	0.6959	0.7216	0.6907	0.7101	0.7137	0.7250
ค่าเฉลี่ย	0.6879	0.7234	0.7072	0.7352	0.7212	0.7390	0.7285	0.7391	0.7321	0.7488

4.7.4 ผลการทดลองของชุดเอกสาร Reuters-top10เมื่อใช้ฟีเจอร์แบบคำเดี่ยว และ ไบ-แกรม

สร้างฟีเจอร์ที่เป็นคำเดี่ยว และไบ-แกรม โดยสร้างไบ-แกรมจากวลีที่สกัดได้ แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.49 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธี เค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.50 และ 4.51 ตามลำดับ

ตารางที่ 4.49 แสดงจำนวนคำเดี่ยวและไบ-แกรมที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร Reuters-top10

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและไบ-แกรม	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	11,043	3,015
2	11,123	2,976
3	10,966	3,098
4	10,975	3,150
ค่าเฉลี่ย	11,027	3,105

ตารางที่ 4.50 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	68.00	71.20	71.20	73.60	72.00	71.20	72.00	74.40	72.00	74.40
2	76.19	81.75	81.75	81.75	80.95	82.54	83.33	83.33	79.37	83.33
3	74.02	74.80	77.95	76.38	77.95	77.95	78.74	80.32	77.95	80.32
4	72.22	73.81	74.60	75.40	72.22	73.81	72.22	76.98	75.40	76.98
ค่าเฉลี่ย	72.61	75.39	76.38	76.78	75.78	76.38	76.57	78.76	76.18	78.76

ตารางที่ 4.51 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการเมื่อใช้ฟีเจอร์แบบคำเดี่ยวและไบ-แกรม ของชุดเอกสาร Reuters-top10

ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6720	0.6949	0.6928	0.7130	0.6932	0.6895	0.6934	0.7246	0.6906	0.7246
2	0.7608	0.8070	0.8074	0.8086	0.7947	0.8111	0.8177	0.8156	0.7738	0.8177
3	0.7264	0.7254	0.7609	0.7371	0.7543	0.7547	0.7684	0.7827	0.7571	0.7827
4	0.7064	0.7280	0.7196	0.7362	0.6941	0.7244	0.7018	0.7526	0.7301	0.7526
ค่าเฉลี่ย	0.7164	0.7388	0.7451	0.7487	0.7341	0.7449	0.7453	0.7689	0.7379	0.7694

4.7.5ผลการทดลองของชุดเอกสาร Reuters-top10เมื่อใช้ฟีเจอร์คำเดี่ยวและวลี

สร้างฟีเจอร์จากวลีที่สกัดได้และคำเดี่ยว แล้วทำการเลือกฟีเจอร์ที่มีค่าคะแนน (Odds Ratio) มากกว่า 0 และปรากฏในชุดเอกสารทั้งหมดมากกว่า 1 เอกสาร และทำการทดลองจำแนกประเภทเอกสารโดยในแต่ละชุดการทดลองได้จำนวนฟีเจอร์จากเอกสารชุดเรียนรู้ ดังแสดงในตารางที่ 4.52 จากนั้นทำการจำแนกประเภทเอกสารชุดทดสอบ โดยการแทนเอกสารด้วยฟีเจอร์ดังกล่าว ด้วยวิธีเค-เนียเรสเนเบอร์ (k-Nearest Neighbor) ค่า k เท่ากับ 1 ถึง 9 แล้วทำการวัดอัตราความถูกต้องและค่าความถูกต้อง F-measure ดังแสดงในตารางที่ 4.53 และ 4.54 ตามลำดับ

ตารางที่ 4.52 แสดงจำนวนฟีเจอร์คำเดี่ยวและวลี ที่ได้จากชุดเอกสารเรียนรู้ ในแต่ละชุดการทดลองของชุดเอกสาร Reuters-top10

ชุดที่	จำนวนฟีเจอร์คำเดี่ยวและวลี	
	ทั้งหมด	ฟีเจอร์ที่ถูกเลือก
1	9,327	1,882
2	9,424	1,874
3	9,304	1,868
4	9,318	1,965
ค่าเฉลี่ย	9,343	1,897

ตารางที่ 4.53 แสดงอัตราความถูกต้องในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 ในแต่ละชุดการทดลองเมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร Reuters-top10

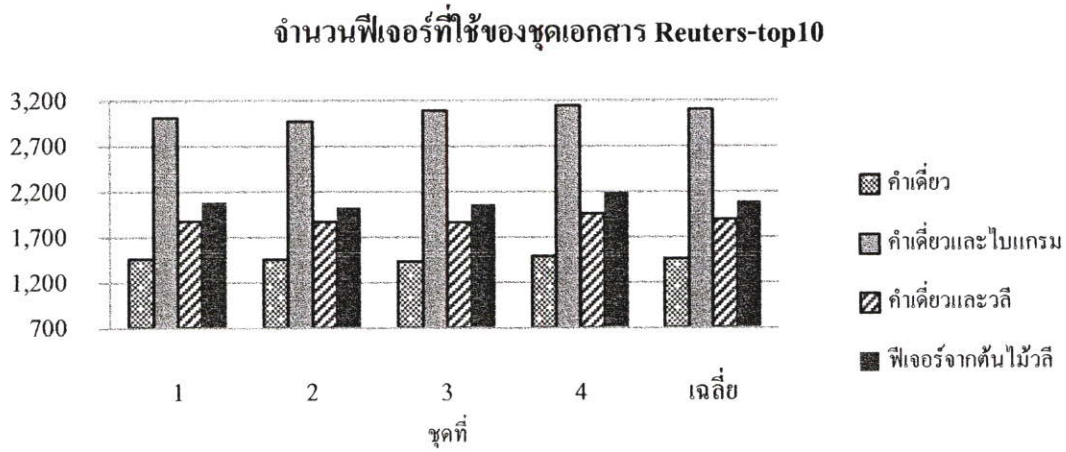
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	64.80	70.40	68.80	69.60	68.00	69.60	72.00	72.80	72.00	72.80
2	76.98	79.37	76.98	79.37	78.57	77.78	81.75	78.57	79.37	81.75
3	70.87	74.80	72.44	74.02	76.38	77.17	77.17	78.74	77.95	78.74
4	70.64	70.64	72.22	73.81	73.81	73.81	71.43	72.22	73.81	73.81
ค่าเฉลี่ย	70.82	73.80	72.61	74.20	74.19	74.59	75.59	75.58	75.78	76.77

ตารางที่ 4.54 แสดงค่าความถูกต้อง F-measure ในการจำแนกประเภทเอกสารวิธีเค-เนียเรสเนเบอร์ ค่า k = 1 ถึง 9 เมื่อใช้ฟีเจอร์คำเดี่ยวและวลี ของชุดเอกสาร Reuters-top10

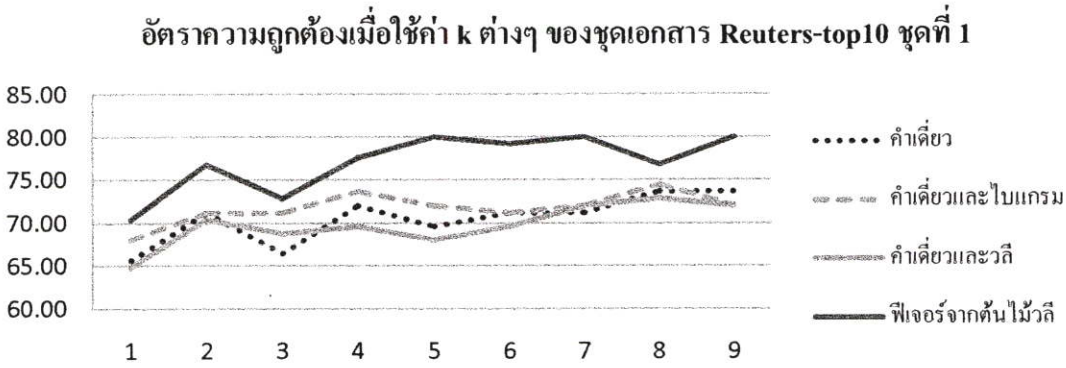
ชุดที่	k									ค่าสูงสุด
	1	2	3	4	5	6	7	8	9	
1	0.6393	0.6874	0.6770	0.6837	0.6633	0.6752	0.6951	0.7074	0.7049	0.7074
2	0.7639	0.7814	0.7576	0.7808	0.7742	0.7646	0.8068	0.7743	0.7786	0.8068
3	0.6875	0.7254	0.7024	0.7165	0.7452	0.7486	0.7544	0.7697	0.7609	0.7697
4	0.6966	0.6931	0.7020	0.7253	0.7175	0.7291	0.7065	0.7149	0.7253	0.7291
ค่าเฉลี่ย	0.6968	0.7218	0.7098	0.7266	0.7250	0.7294	0.7407	0.7416	0.7424	0.7533

4.7.6 เปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แบบต่าง ชองชุดเอกสาร Reuters-top10

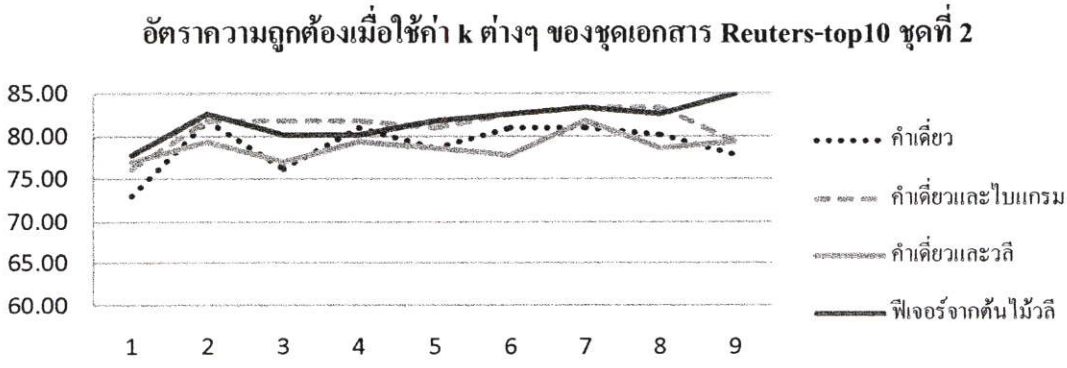
รูปที่ 4.42 ถึง 4.54 แสดงการเปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แต่ละแบบ โดยเปรียบเทียบทั้งในด้านจำนวนฟีเจอร์ อัตราความถูกต้อง และค่า F-measure



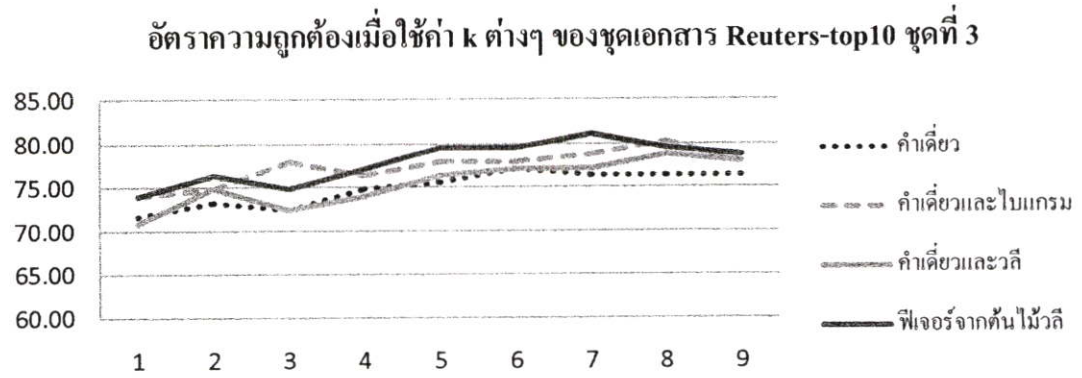
รูปที่ 4.42 แผนภูมิแท่งเปรียบเทียบจำนวนฟีเจอร์ที่ใช้ในแต่ละแบบของชุดเอกสาร Reuters-top10



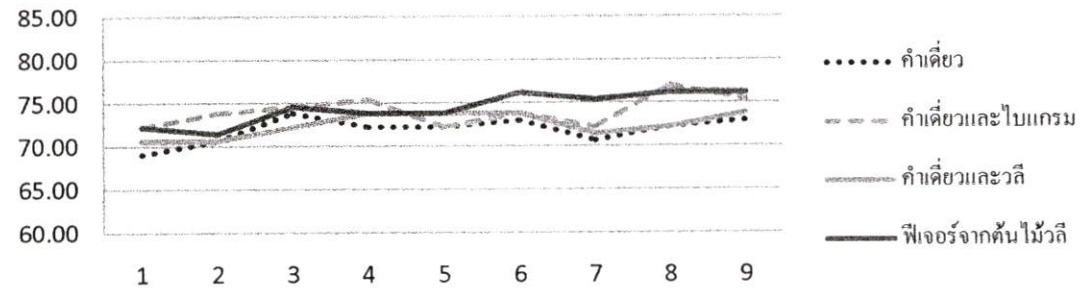
รูปที่ 4.43 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 1



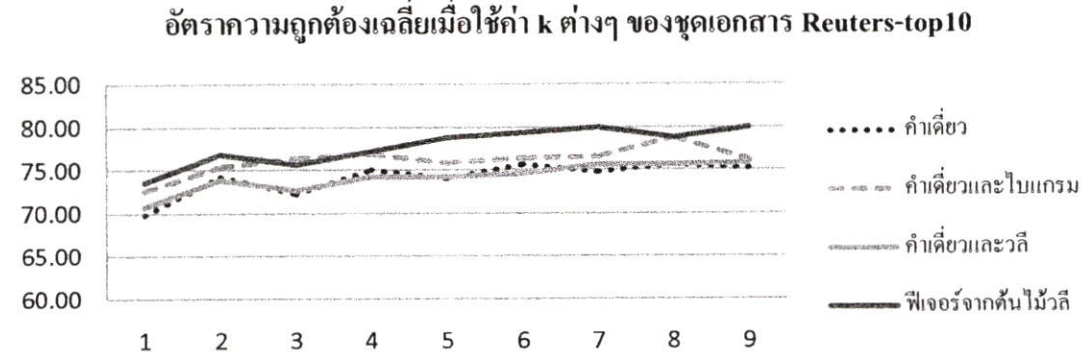
รูปที่ 4.44 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 2



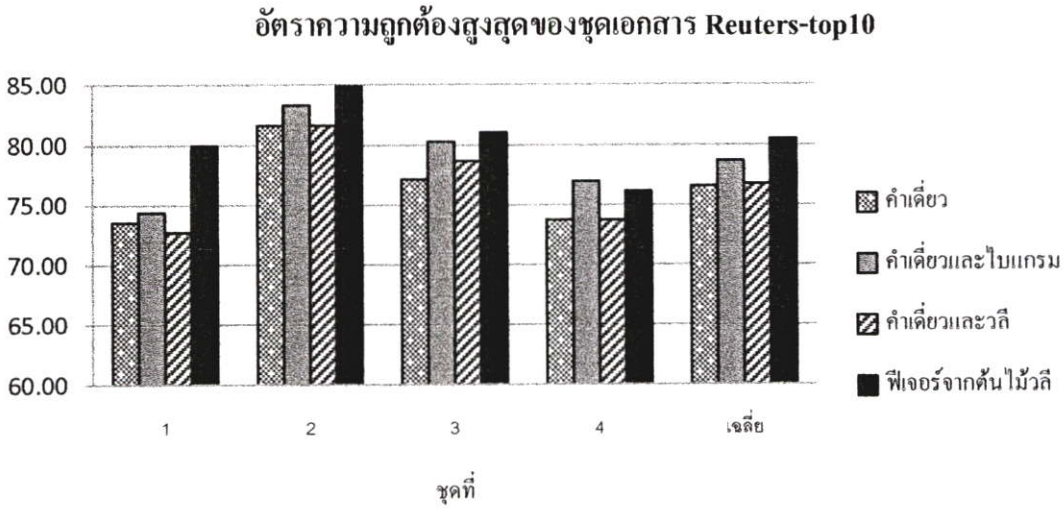
รูปที่ 4.45 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 3



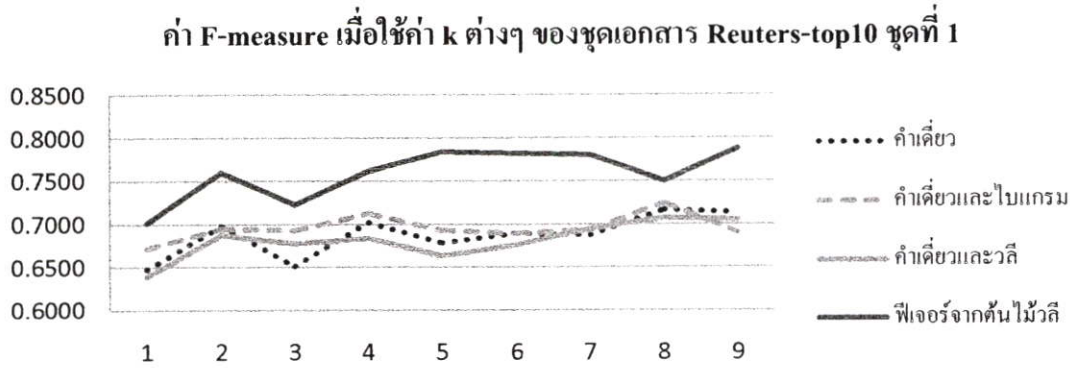
รูปที่ 4.46 กราฟเปรียบเทียบอัตราความถูกต้องในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 4



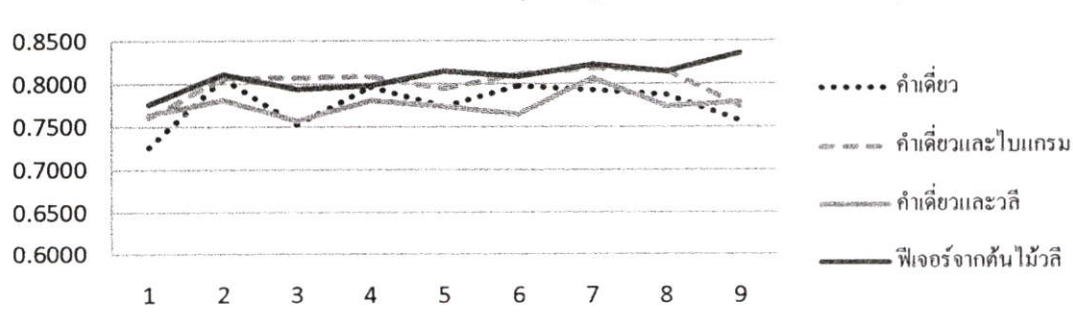
รูปที่ 4.47 กราฟเปรียบเทียบอัตราความถูกต้องเฉลี่ยในแต่ละแบบของชุดเอกสาร Reuters-top10



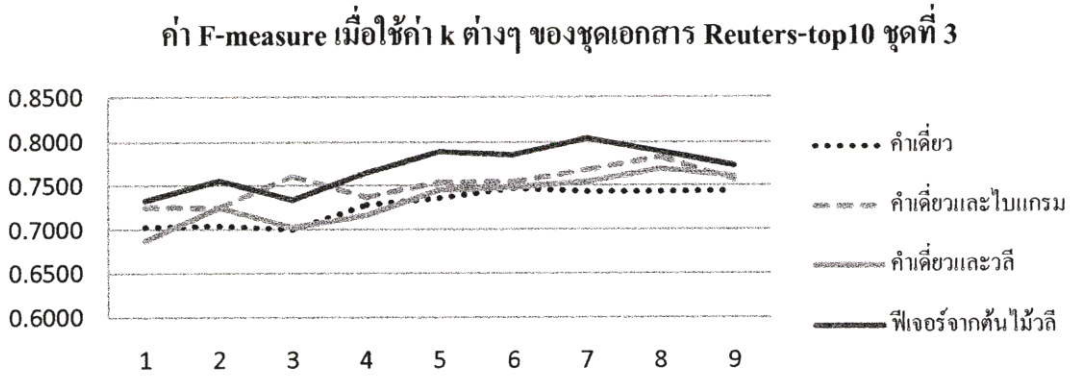
รูปที่ 4.48 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุดในแต่ละแบบของชุดเอกสาร Reuters-top10



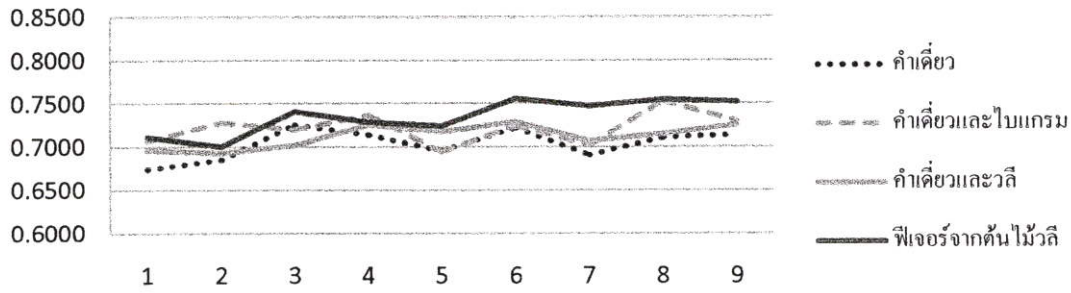
รูปที่ 4.49 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 1



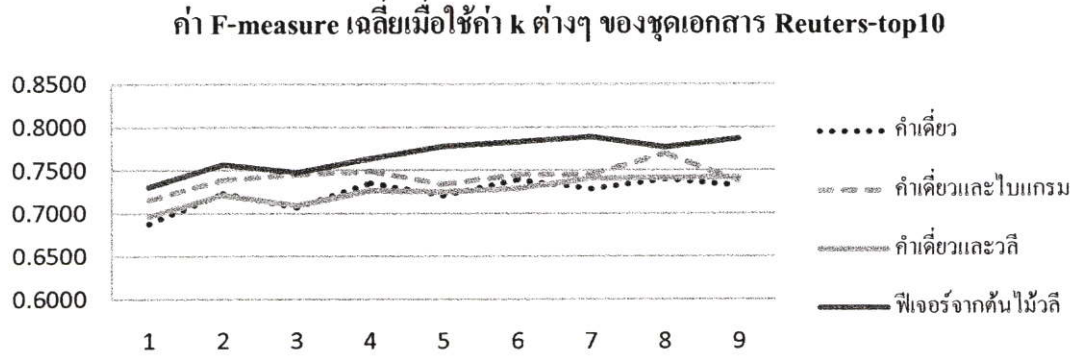
รูปที่ 4.50 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 2



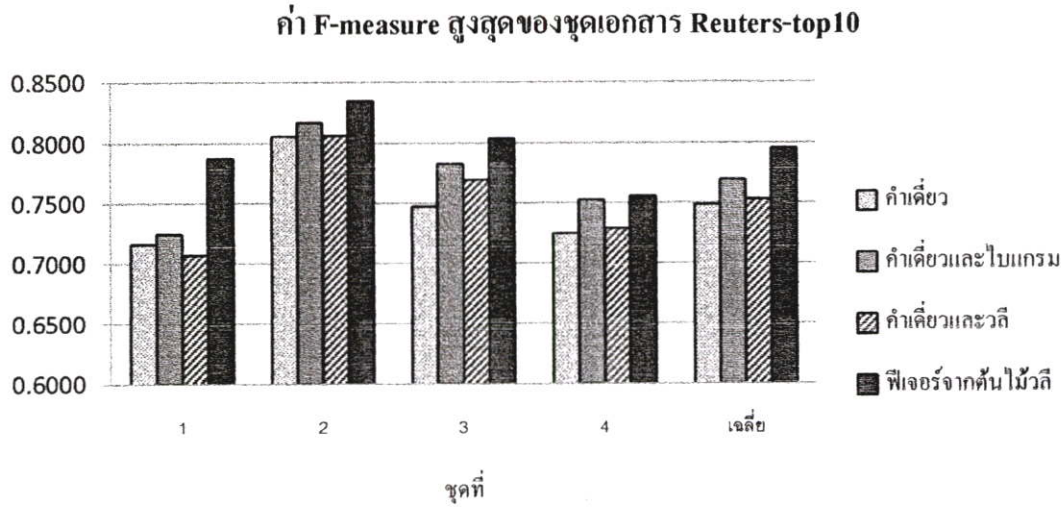
รูปที่ 4.51 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 3



รูปที่ 4.52 กราฟเปรียบเทียบ ค่า F-measure ในแต่ละแบบของชุดเอกสาร Reuters-top10 ชุดที่ 4



รูปที่ 4.53 กราฟเปรียบเทียบ ค่า F-measure เฉลี่ยในแต่ละแบบของชุดเอกสาร Reuters-top10



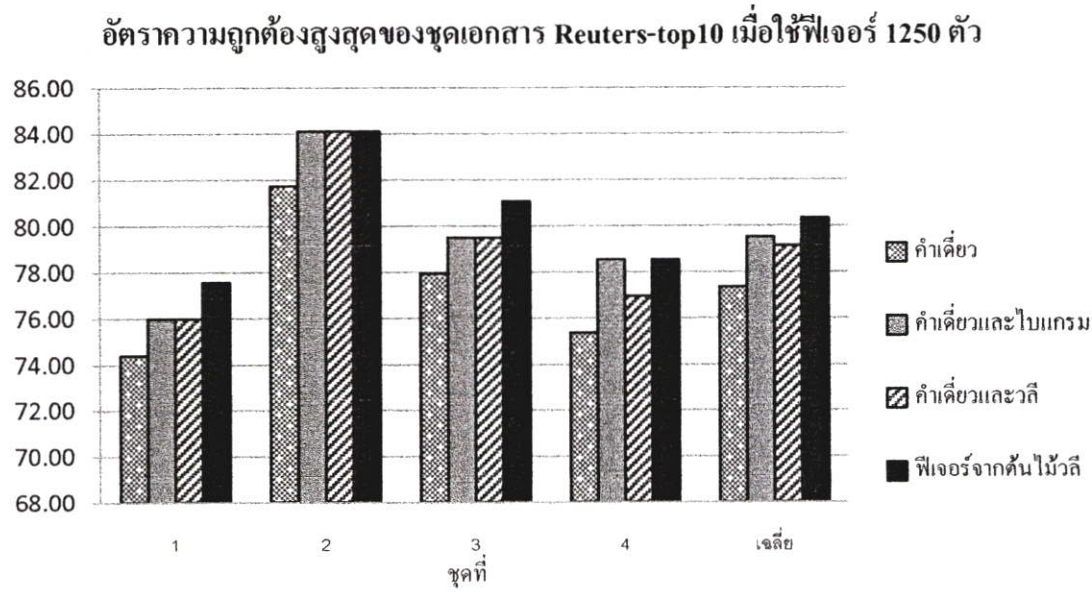
รูปที่ 4.54 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุดในแต่ละแบบของชุดเอกสาร Reuters-top10

จากรูปที่ 4.42 จะเห็นได้ว่าฟีเจอร์ที่ได้จากต้นไม้วลีมีจำนวนมากกว่าฟีเจอร์แบบอื่นๆ และเมื่อพิจารณาถึงความถูกต้องในการจำแนกประเภทเอกสาร จากรูปที่ 4.48 และ 4.54 ในทุกชุดการทดลองยกเว้นชุดที่ 4 นั้น ฟีเจอร์จากต้นไม้วลีให้อัตราความถูกต้องและค่า F-measure โดยเฉลี่ยสูงกว่าฟีเจอร์แบบอื่นๆ ซึ่งเป็นไปตามเป้าหมายของเราที่ต้องการได้ฟีเจอร์ที่มีประสิทธิภาพ โดยใช้วลีร่วมกับคำเดี่ยว ที่ให้ความถูกต้องในการจำแนกประเภทเอกสารสูงแต่มีจำนวนน้อย

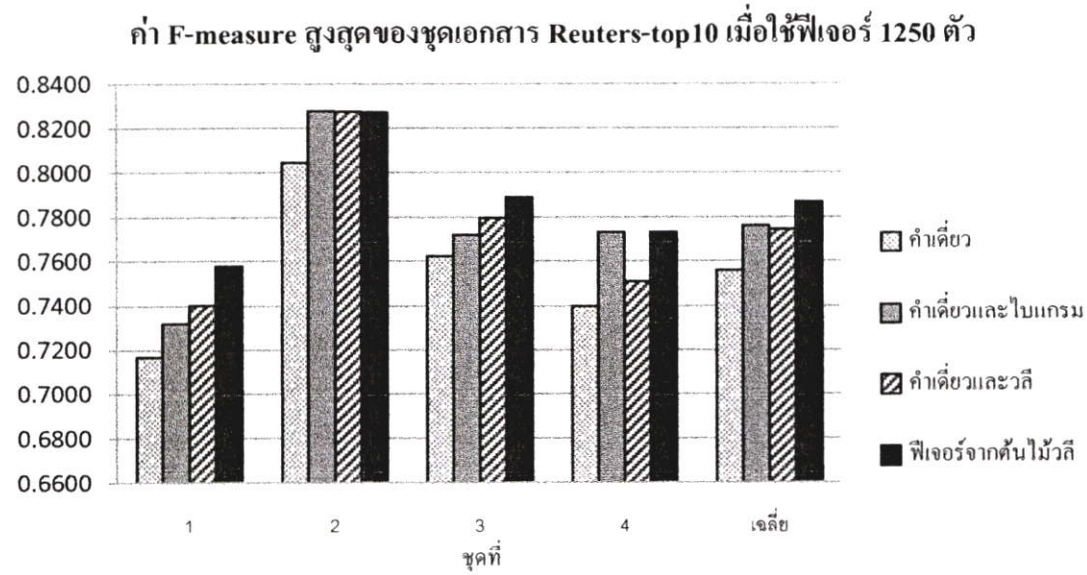
4.7.7 เปรียบเทียบผลการทดลองเมื่อใช้ฟีเจอร์แบบต่าง แบบจำกัดจำนวน ของชุดเอกสาร Reuters-top10

จากผลการทดลองข้างต้น จะเห็นได้ว่าฟีเจอร์จากต้นไม้วลีนั้นทำให้จำแนกประเภทเอกสารได้ความถูกต้องสูงกว่าแบบอื่นๆ แต่ว่ามีจำนวนมากกว่าฟีเจอร์แบบอื่นๆ ยกเว้นฟีเจอร์แบบคำเดี่ยวและไบแกรม ดังนั้น จึงทำการทดลองเอกสารตามแบบข้างต้น แต่เมื่อได้ฟีเจอร์มาแล้วจะทำการจำกัดจำนวนที่จะใช้ฟีเจอร์แต่ละแบบ โดยนำฟีเจอร์มาเรียงตามค่า Odds-Ratio แล้วเลือกฟีเจอร์ที่ให้ค่าคะแนนสูงในลำดับต้นๆ ตามจำนวนที่ต้องการมาใช้ โดยรูปที่ 4.55 ถึงรูปที่ 4.62 แสดงค่าความถูกต้องสูงสุด และค่า F-measure สูงสุดเมื่อใช้จำนวนฟีเจอร์เท่ากับ 1250, 1500, 1750 และ 2000 ตามลำดับ โดยในการเลือกฟีเจอร์ยังคงใช้เกณฑ์ความถี่เอกสารต่ำสุด คือ 1 และค่า Odds-Ratio มากว่า 0 ผลการเปรียบเทียบเมื่อใช้จำนวนฟีเจอร์ต่างๆ กัน ในแต่ละชุดการทดลอง แสดงดังกราฟในรูปที่ 4.63 ถึง 4.70

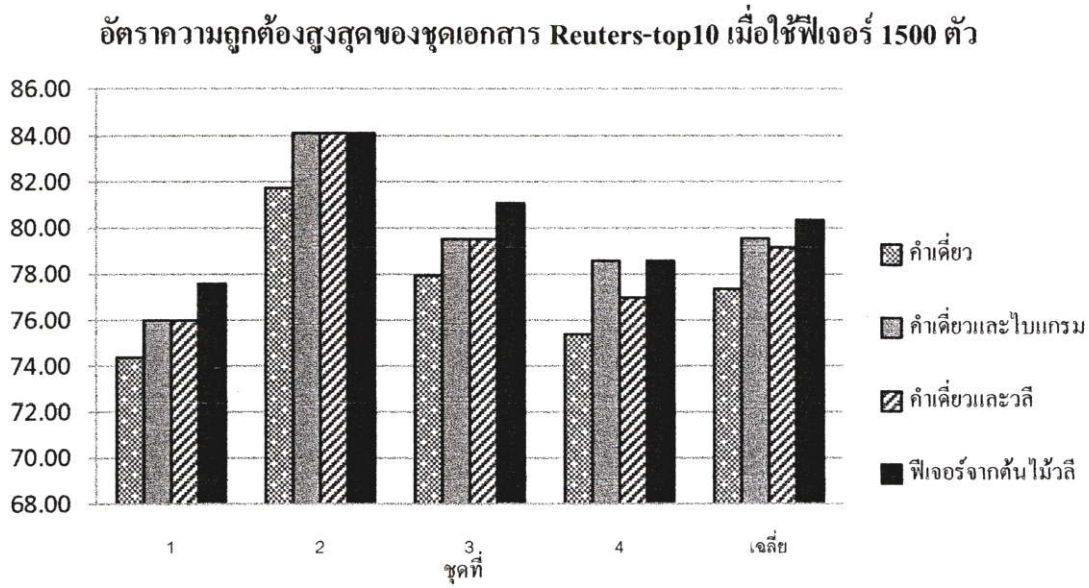
รูปที่ 4.71 และ 4.72 แสดงแผนภูมิแท่งเปรียบเทียบค่าอัตราความถูกต้องสูงสุดเฉลี่ย และค่า F-measure สูงสุดเฉลี่ยจากทุกชุดการทดลองเมื่อใช้ฟีเจอร์ต่างๆ ตามลำดับ



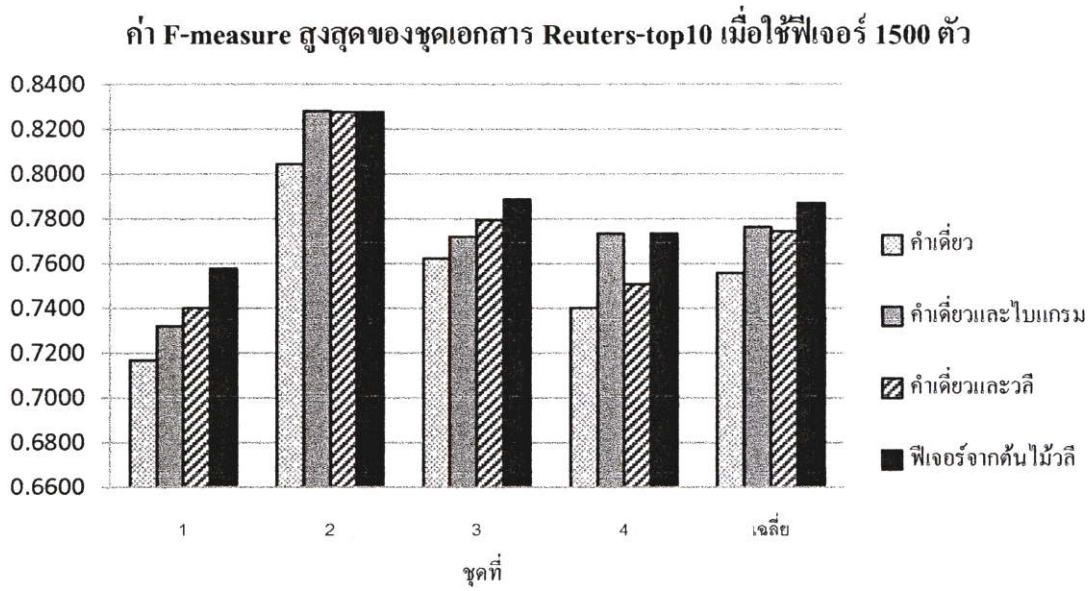
รูปที่ 4.55 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1250 ตัว ของชุดเอกสาร Reuters-top10



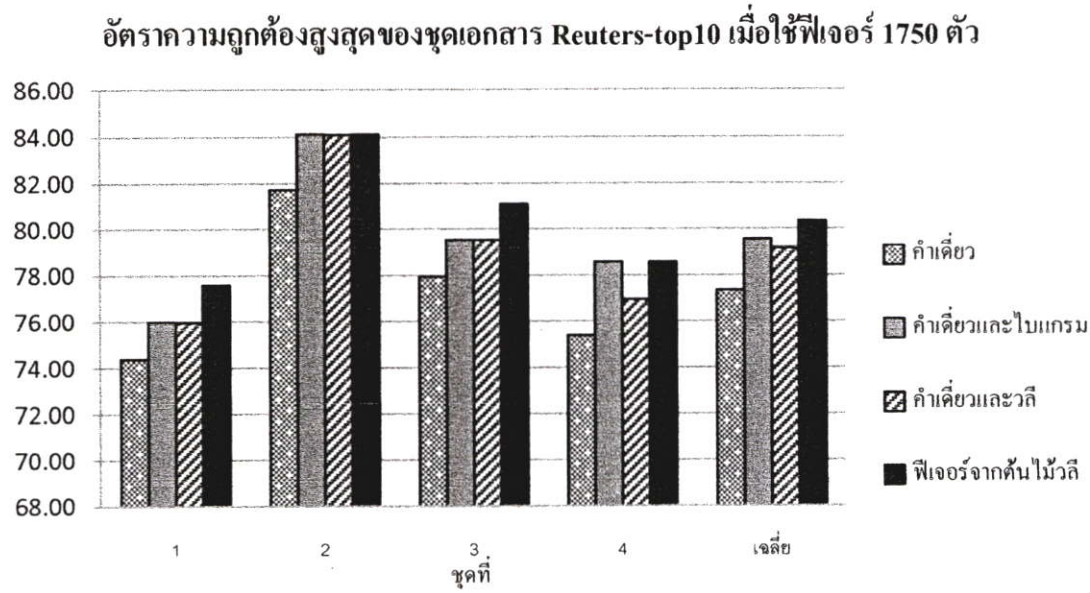
รูปที่ 4.56 แผนภูมิแท่งเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1250 ตัว ของชุดเอกสาร Reuters-top10



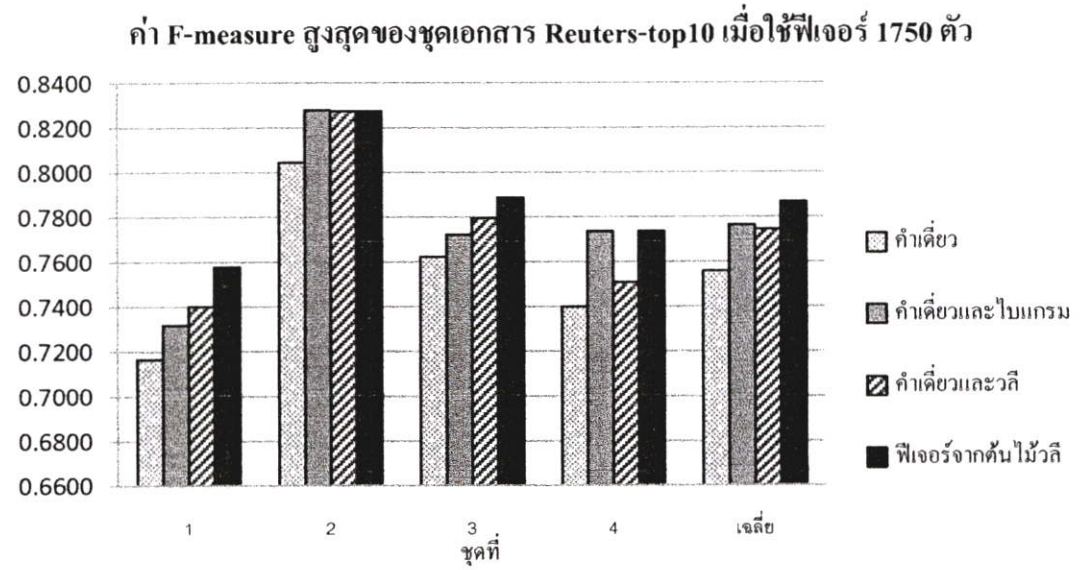
รูปที่ 4.57 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1500 ตัว ของชุดเอกสาร Reuters-top10



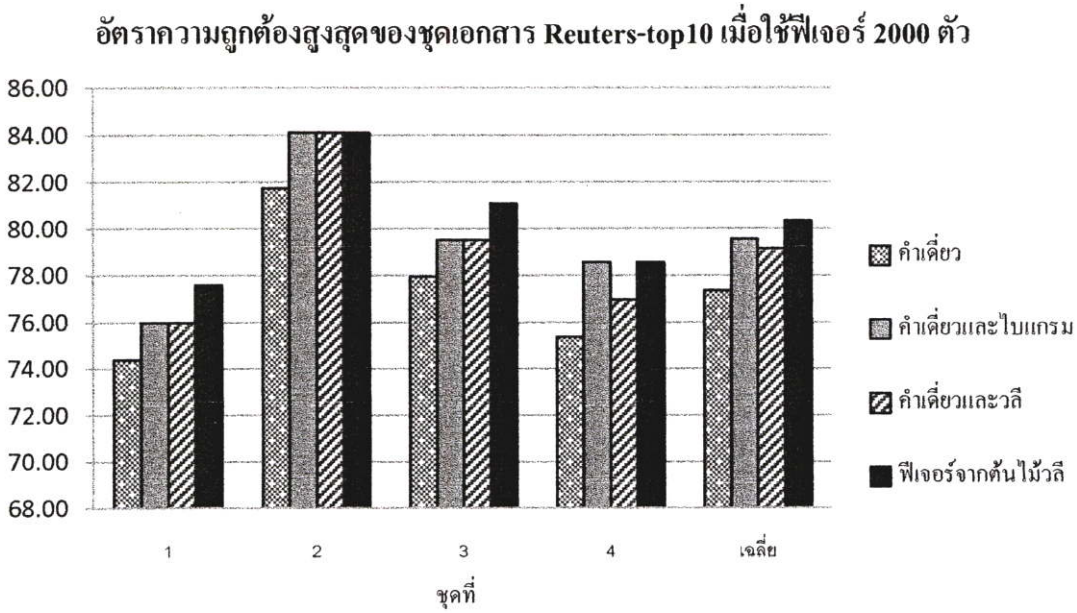
รูปที่ 4.58 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1500 ตัว ของชุดเอกสาร Reuters-top10



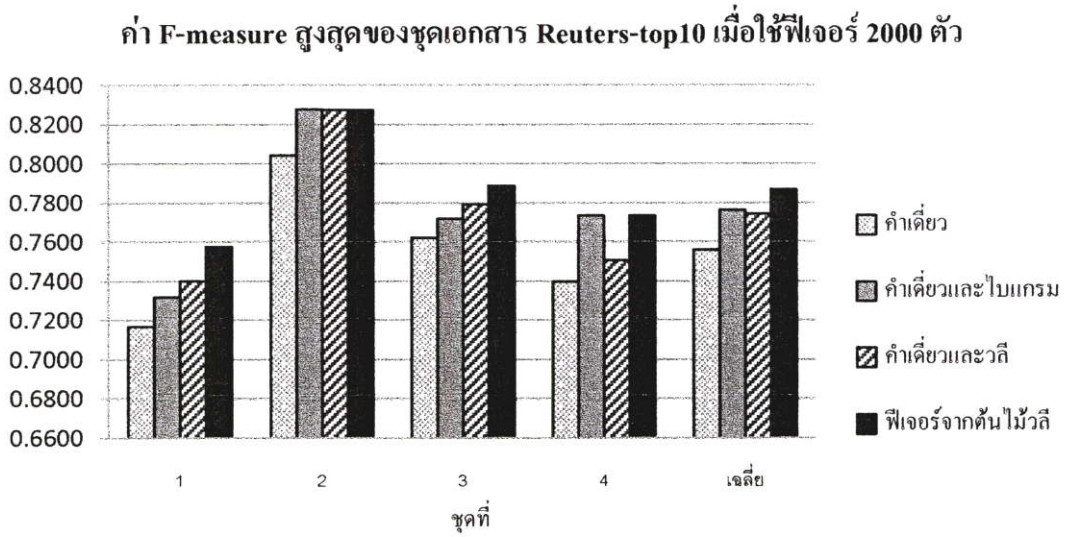
รูปที่ 4.59 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1750 ตัว ของชุดเอกสาร Reuters-top10



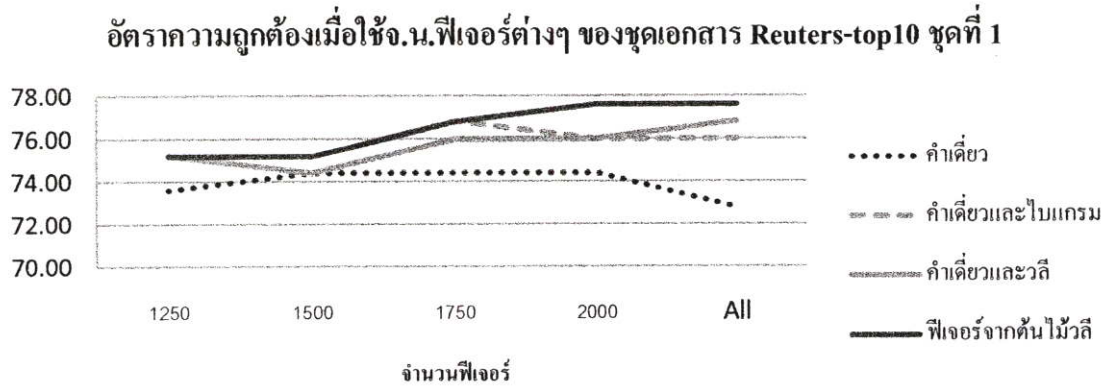
รูปที่ 4.60 แผนภูมิแท่งเปรียบเทียบ ค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 1750 ตัว ของชุดเอกสาร Reuters-top10



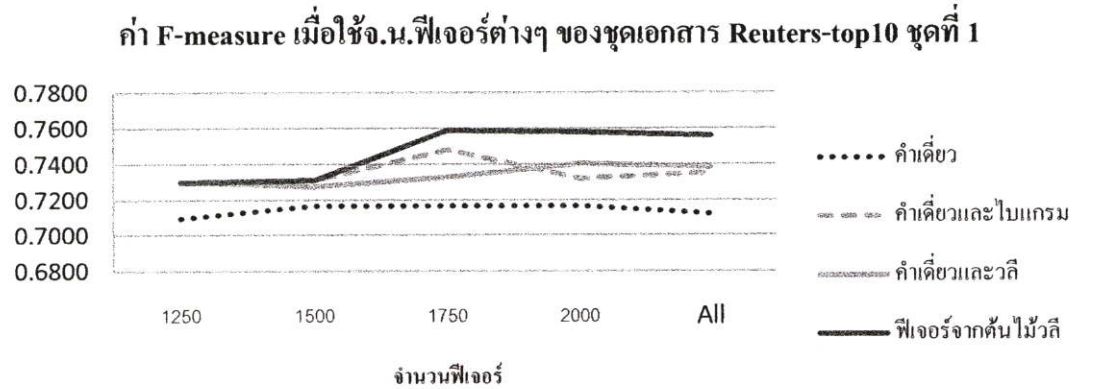
รูปที่ 4.61 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 2000 ตัว ของชุดเอกสาร Reuters-top10



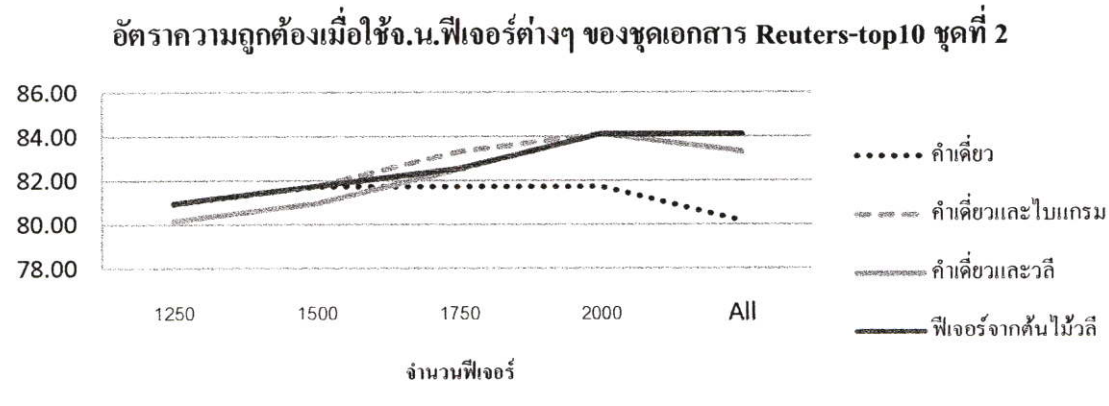
รูปที่ 4.62 แผนภูมิแท่งเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์แต่ละแบบ จำนวน 2000 ตัว ของชุดเอกสาร Reuters-top10



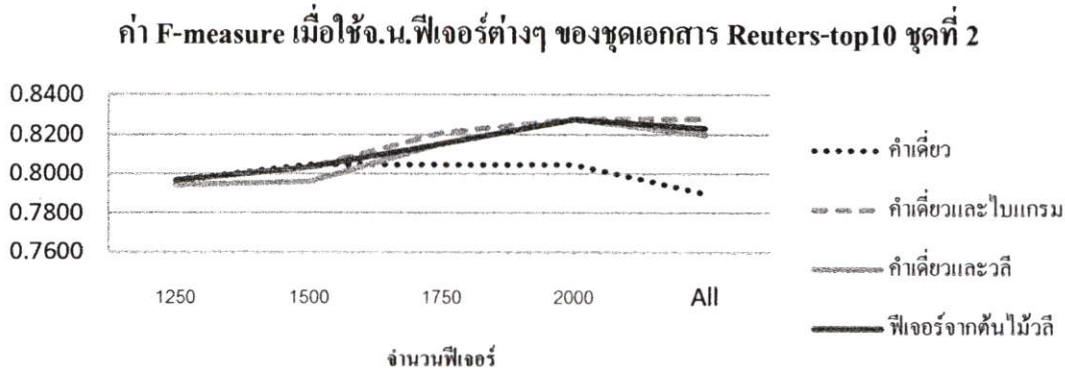
รูปที่ 4.63 กราฟเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 1



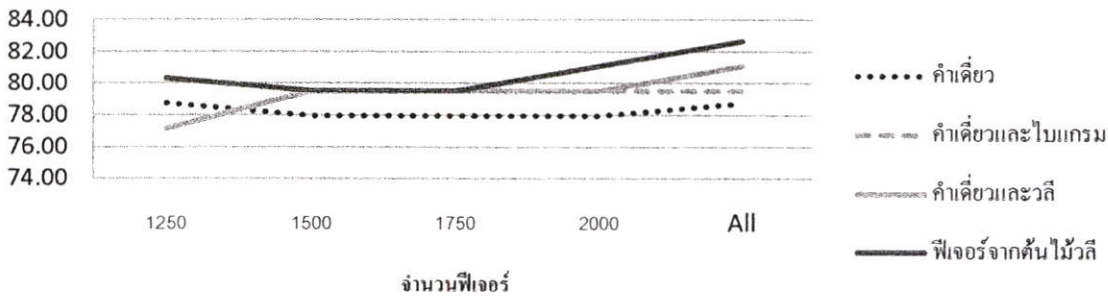
รูปที่ 4.64 กราฟเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 1



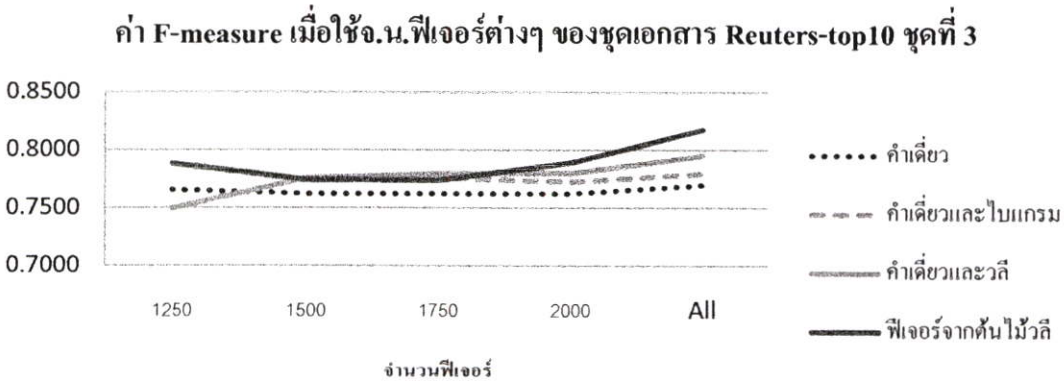
รูปที่ 4.65 กราฟเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 2



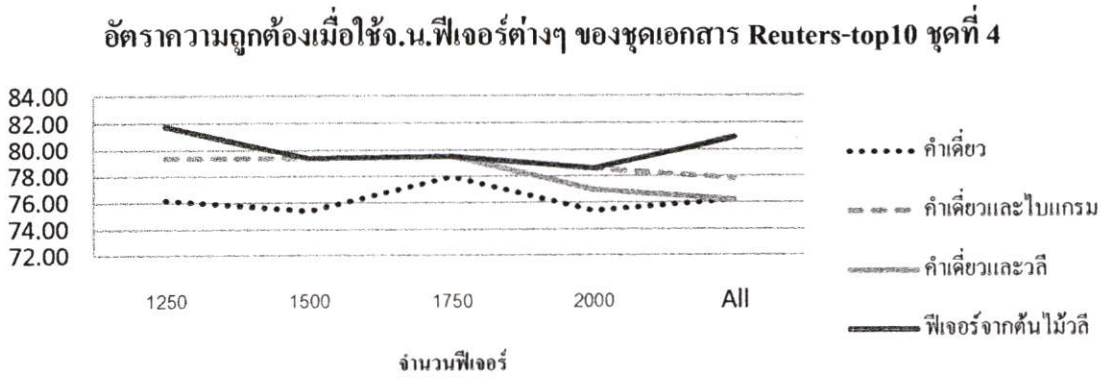
รูปที่ 4.66 กราฟเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 2



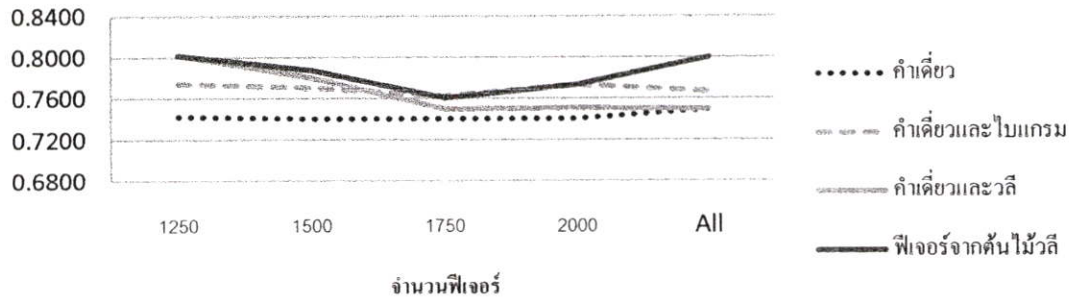
รูปที่ 4.67 กราฟเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 3



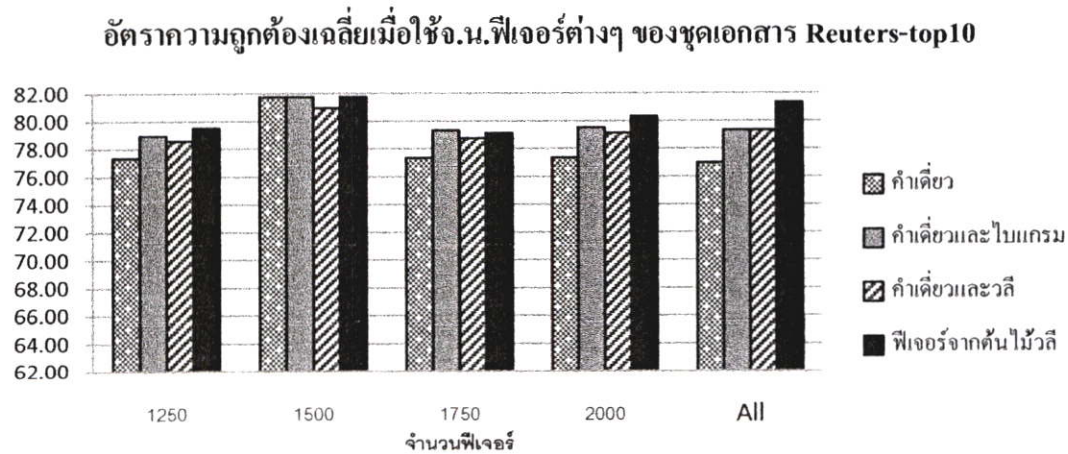
รูปที่ 4.68 กราฟเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 3



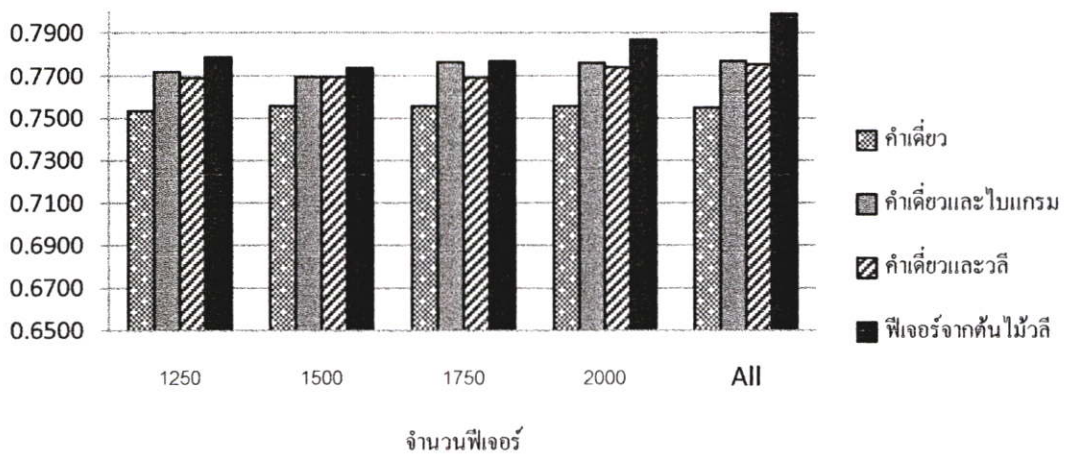
รูปที่ 4.69 กราฟเปรียบเทียบอัตราความถูกต้องสูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 4



รูปที่ 4.70 กราฟเปรียบเทียบค่า F-measure สูงสุด เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10 ชุดที่ 4



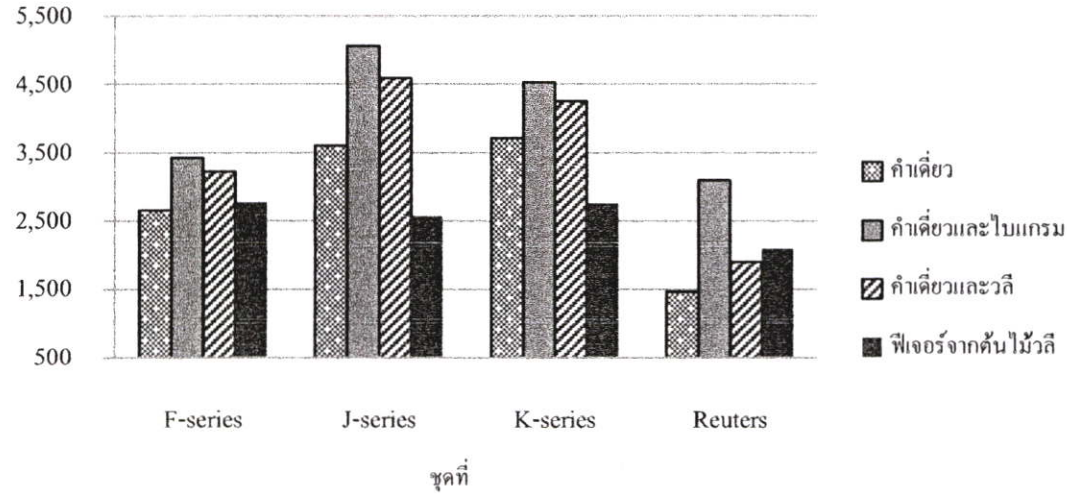
รูปที่ 4.71 แผนภูมิแท่งเปรียบเทียบค่าอัตราความถูกต้องเฉลี่ยเมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10



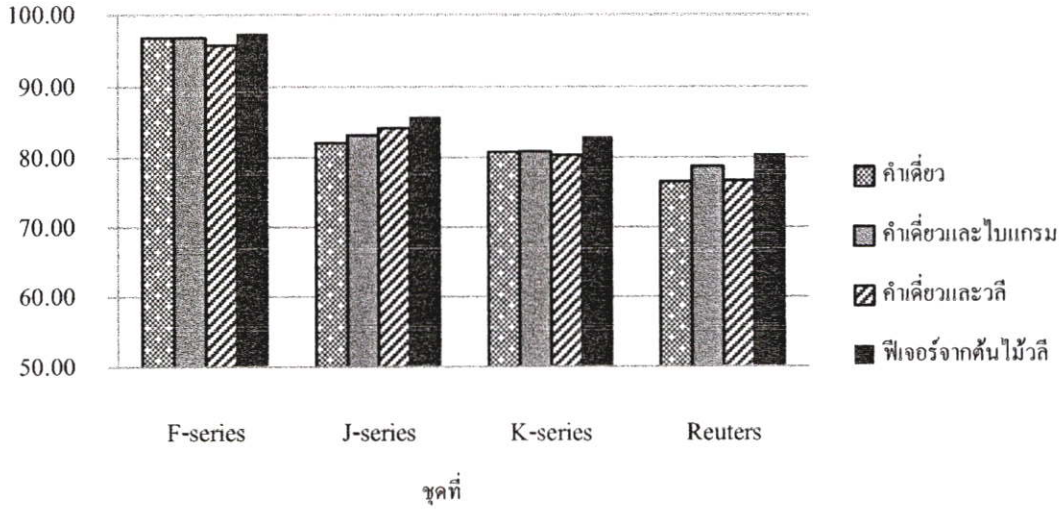
รูปที่ 4.72 แผนภูมิแท่งเปรียบเทียบค่า F-measure เฉลี่ย เมื่อใช้ฟีเจอร์ในแต่ละแบบ โดยมีการจำกัดจำนวนของฟีเจอร์ ของชุดเอกสาร Reuters-top10

4.8 การเปรียบเทียบผลการทดลองทุกชุดข้อมูล

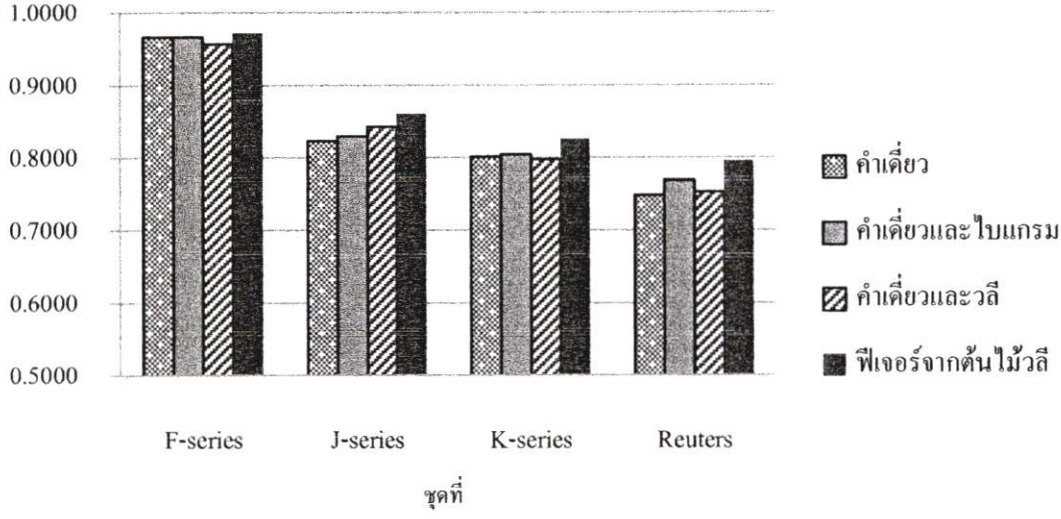
จากผลการทดลองในชุดข้อมูล PDDP ทั้ง F-series, J-series และ K-series การลดฟีเจอร์โดยใช้ต้นไม้วลีนั้นสามารถลดจำนวนของฟีเจอร์ และเพิ่มความถูกต้องได้ แต่ในข้อมูล Reuters-top10 จำนวนฟีเจอร์ยังคงมากกว่าฟีเจอร์แบบอื่นๆ แต่สามารถเพิ่มความถูกต้องในการจำแนกประเภทเอกสารได้ ดังแสดงในรูปที่ 4.55, 4.56 และ 4.57 ซึ่งแสดงผลการเปรียบเทียบจำนวนฟีเจอร์ อัตราความถูกต้อง และค่า F-measure เมื่อใช้ฟีเจอร์แบบต่างๆ ของทุกชุดข้อมูลตามลำดับ



รูปที่ 4.73 แผนภูมิแท่งเปรียบเทียบจำนวนฟีเจอร์แต่ละแบบของทุกชุดข้อมูล



รูปที่ 4.74 แผนภูมิแท่งเปรียบเทียบอัตราความถูกต้องเมื่อใช้ฟีเจอร์แต่ละแบบของทุกชุดข้อมูล



รูปที่ 4.75 แผนภูมิแท่งเปรียบเทียบค่า F-measure เมื่อใช้ฟีเจอร์แต่ละแบบของทุกชุดข้อมูล

พิจารณาจำนวนฟีเจอร์จากรูปที่ 4.73 จะเห็นได้ว่าในชุดข้อมูล J-series และ K-series ฟีเจอร์ที่ได้จากต้นไม้วลีมีปริมาณน้อยกว่าฟีเจอร์แบบคำเดี่ยว ทั้งนี้เนื่องจากเราใช้เกณฑ์ความถี่เอกสารในการเลือกวลีที่สกัดได้ ก่อนนำมาสร้างต้นไม้วลี ทำให้บางวลีไม่ได้ถูกเลือก และทำให้ปริมาณคำเดี่ยวลดลง ทำให้มีปริมาณน้อยกว่าฟีเจอร์แบบคำเดี่ยว

จากการทดลองทั้ง 4 ชุดข้อมูล จะเห็นได้ว่าฟีเจอร์จากต้นไม้วลีมีประสิทธิภาพ คือ ให้ความถูกต้องในการจำแนกประเภทเอกสารมากกว่าฟีเจอร์แบบอื่นๆ และโดยส่วนใหญ่แล้วได้จำนวนที่น้อยว่าฟีเจอร์แบบอื่นๆ ด้วย

ในบทที่ 5 จะกล่าวถึงผลสรุปของงานวิจัยและข้อเสนอแนะ ซึ่งจะเป็นประโยชน์ในการพัฒนาการทำวิจัยต่อไป

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

การจำแนกประเภทเอกสารนั้น คือ การจัดประเภทเอกสารให้กับเอกสารใหม่ โดยใช้ความรู้จากเอกสารเดิมที่เราทราบประเภทอยู่แล้วมาใช้ การจำแนกประเภทเอกสารให้มีประสิทธิภาพ คือ มีความถูกต้องสูง และใช้เวลาไม่นาน ปัจจัยที่สำคัญก็คือ ฟีเจอร์ที่ใช้แทนเอกสาร ซึ่งจะต้องเป็นตัวแทนเอกสารได้ดี คือ สื่อถึงความหมายของประเภทเอกสารต่างๆ ได้ และจะต้องมีจำนวนที่ไม่มากเกิน เพื่อลดเวลาในการประมวลผล

งานวิจัยนี้ได้นำเสนอการลดฟีเจอร์สำหรับการจำแนกประเภทเอกสารโดยใช้ต้นไม้วลี ซึ่งเป็นการใช้โครงสร้างข้อมูลแบบต้นไม้มาช่วยในการสร้างฟีเจอร์จากวลีที่สกัดได้ และช่วยในการเลือกฟีเจอร์โดยดูจากความสัมพันธ์ของฟีเจอร์ภายในต้นไม้ โดยใช้ค่า Odds-Ratio ร่วมด้วย โดยได้ทำการทดลองวัดประสิทธิภาพของฟีเจอร์ที่ได้จากต้นไม้วลีกับฟีเจอร์แบบอื่นๆ โดยดูจากจำนวนและความถูกต้องในการจำแนกประเภทเอกสารที่ได้ โดยใช้ชุดเอกสารตัวอย่างในการทดสอบ ซึ่งเป็นชุดเอกสารที่ถูกใช้ในหลายๆ งานวิจัยทางด้านการจำแนกประเภทเอกสาร

ซึ่งจากผลการทดลองสรุปได้ว่าการใช้ต้นไม้วลีในการลดฟีเจอร์ ทำให้ได้ฟีเจอร์ที่มีประสิทธิภาพคือ มีจำนวนน้อยแต่ยังคงความถูกต้องในการจำแนกประเภทเอกสารไว้ได้ และสามารถนำงานวิจัยนี้ไปประยุกต์ใช้ในงานต่างๆ ได้ เช่น การจำแนกประเภทของจดหมายอิเล็กทรอนิกส์โดยแบ่งเป็นประเภทที่เป็นขยะ และไม่เป็นขยะ ซึ่งจะนำเนื้อหาของแต่ละจดหมายอิเล็กทรอนิกส์มาพิจารณา หรืออาจนำไปประยุกต์ใช้ในการจำแนกหนังสือในห้องสมุดที่มีการแบ่งประเภทไว้แล้ว โดยอาจจะนำส่วนของชื่อเรื่อง คำนำ และสารบัญ มาใช้ในการพิจารณาจำแนกประเภท แทนการพิจารณาในส่วนของเนื้อหาทั้งหมดของหนังสือ

5.2 ข้อเสนอแนะ

จากปัญหาที่จำนวนฟีเจอร์ที่ได้จากการเลือกโดยต้นไม้วลีมีจำนวนมากกว่าแบบอื่นๆ ในชุดเอกสาร Reuters-top10 จึงควรหาวิธีที่จะช่วยลดความซ้ำซ้อนของฟีเจอร์ในต้นไม้วลีให้ได้มากกว่านี้ แต่ยังคงต้องรักษาความถูกต้องในการจำแนกประเภทเอกสารไว้

ในเรื่องของการเลือกฟีเจอร์ในต้นไม้วลีในงานวิจัยนี้ใช้เพียงแค่ค่า Odds Ratio ในการให้คะแนนในแต่ละเทอม ควรหาเกณฑ์อื่นมาใช้ด้วย เช่น อาจจะมีการถ่วงน้ำหนักคำที่มีความสำคัญ คำที่อยู่ในส่วนหัวข้อ เป็นต้น หรืออาจใช้ค่าคะแนนอื่นที่ไม่ใช่ Odds Ratio ก็ได้

บรรณานุกรม

[1] T. Dumis, J. Platt, D. Heckerman, and M. Sahami. "**Inductive learning algorithms and representation for text categorization**". CIKM'98, ACM Info. & KM. Bethesda, US, pp. 148-155, 1998.

[2] G. Salton, A. Wong, and C. S. Yang , "**A Vector Space Model for Automatic Indexing**," Communications of the ACM, vol. 18, nr. 11, pages 613–620, 1975.

[3] F. Sebastiani and Franca Debole. "**Supervised term weighting for automated text categorization**". SAC-03 18th, ACM Applied Computing, pp. 784-788, 2003.

[4] F. Sebastiani. "**Machine learning in automated text categorization**". ACM Computing Surveys, 34(1), pp. 1-47, March 2002.

[5] G. Gonged, H. Wang, D. Bell, Y. Bi, and K. Gree. "**Using kNN model-based approach for automatic text categorization**". European Commission project ICONS, project no. IST-2001-32429.

[6] M. Sholom, T. Zhang and J. Fred, **Text Mining: Predictive Methods for Analyzing Unstructured Information**. MA, USA., Springer, 2004.

[7] M.F. Porter. "**An Algorithm for Suffix Stripping**". Program, 14 no.3, pp. 130-137, 1980.

[8] C.E. Shannon. "**A Mathematical Theory of Communication**", Bell System Technical Journal, 27, pp. 379–423 & 623–656, July & October, 1948.

[9] Y. Y. Yao. **Information-theoretic measures for knowledge discovery and dat mining, in Entropy measures, maximum entropy principle and emerging applications,** Karmeshu (ed.), Springer, pp. 115-136, 2003.

[10] M. F. Caroppreso, S. Matwin, and F. Sebastani. "**A learner-independent evaluation of the usefulness of statistical phrases for automated text categorizatiom**", Text Databases and Document Management: Theory and Practice , Hershey, PA, pp. 78-102, 2001.

[11] G.Salton, and C.Buckley, "**Term-weighting approaches in automatic text retrieval**". Information Processing & Management 24(5): pp. 513–523, 1988.

[12] P.C. Mahalanobis, "**On the generalised distance in statistics**", Proceedings of the National Institute of Science of India 12, pp. 49-55, 1936.

[13] G. Salton, **Automatic Text Processing: the transformation, analysis and retrieval of information by computer**. New York: Addison Wesley Publishing Company. 1989.

[14] Rish, Irina. "**An empirical study of the naive Bayes classifier**". IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 2001.

[15] Baeza-Yates, R. and Ribeiro-Neto, B. **Modern Information Retrieval.** New York : Addison Wesley Publishing Company. 1999.

[16] **PDDP Data Set**, http://ftp.cs.umn.edu/dept/users/boley/PDDPdata/.

[17] Adam Schenker, Mark Last, Horst Bunke and Abraham Kandelm, "**Classification of Web Documents Using a Graph Model**". Proceeding of the 7th ICDAR 2003,IEEE, 2003.

[18] D.D. Lewis, "**Reuters-21578 text categorization test collection distribution 1.0.**", 2004. [Online]. Available : http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[19] [Online]. Available : http://www.reeltwo.com/datasets.html.

[20] สุรินทร์ นิยมางกูร, **เทคนิคการสุ่มตัวอย่าง**, สำนักพิมพ์มหาวิทยาลัยเกษตรศาสตร์, พิมพ์ครั้งที่ 3, กรุงเทพฯ, 2541.

[21] J. Klok, S. Driessen, and M. Brunner, "**Some terms are more interchangeable than others**". Cross-Language Evaluation Forum (CLEF) 2001.

[22] Stopwords List [Online]. Available : http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words.

[23] WordNet [Online]. Available : http://wordnet.princeton.edu/online/.

[24] D.D. Lewis, "**An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task**". Proceeding of SIGIR'92, pp. 37-50, ACM Press, New York, 1992.

[25] D. Mladeni and M. Grobelink, "**Word sequences as features in text learning**". 17th Electro. & Comp. Science Conf., Slovenia, pp. 145-148, 1998.

[26] C. M. Tan, Y.F. Wang and C. D. Lee, "**The use of bigrams to enhance text categorization**". Info. Proc. & Management, 38(4): pp. 529-546, 2002.

[27] R. Bekkerman and J. Allan, "**Using Bigrams in Text Categorization**". CIIR Technical Report IR-408, University of Massachusetts, 2003.

ภาคผนวก

ภาคผนวก ก.
แสดงตัวอย่างชุดเอกสารต่างๆ ที่ใช้ในการทดลอง

ก.1 ชุดเอกสาร PDDP

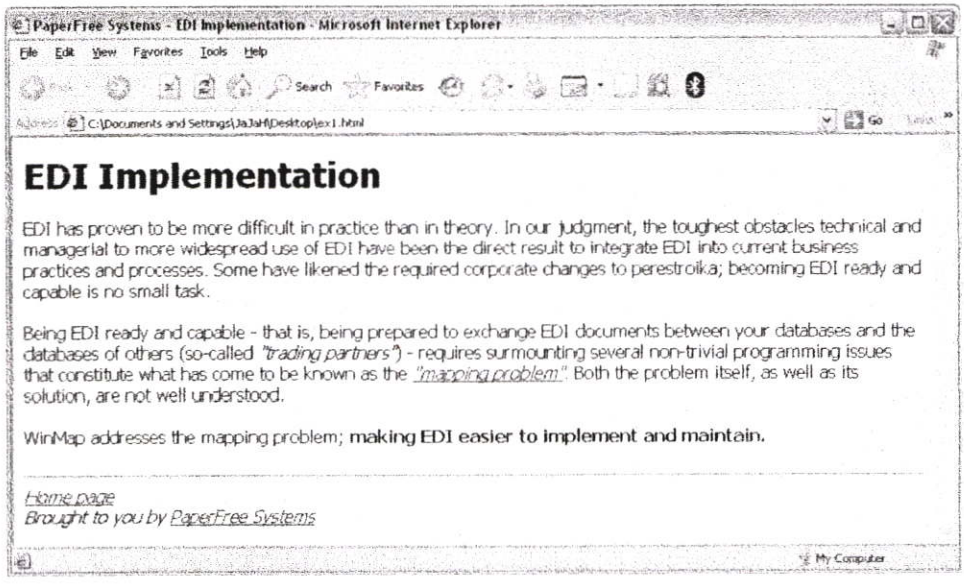
ชุดเอกสาร PDDP [16] เป็นเอกสารเว็บ ซึ่งแบ่งเป็น 3 ชุดย่อย F-series J-series และ K-series โดยแต่ละเอกสารอยู่ในรูปแบบเอกสาร HTML ทั้ง 3 ชุดย่อยมีลักษณะเอกสารเหมือนกัน และมีรายละเอียดและขั้นตอนการเตรียมเอกสารต่างๆ ดังนี้

ก.1.1 ตัวอย่างไฟล์เอกสาร PDDP F-series ในรูปแบบ HTML

รูปแบบของเอกสารก่อนการตัดแท็ก เป็นดังนี้

```
<!-- www.paperfree.com/ediprob.htm -->
<html><head><title>PaperFree Systems - EDI Implementation</title></head>
<body><!-- ediprob.htm  --><h1><B>EDI Implementation</B></h1>
<p>EDI has proven to be more difficult in practice than in theory. In our judgment, the
toughest obstacles technical and managerial  to more widespread use of EDI have
been the direct result to integrate EDI into current business practices and processes. Some
have likened the required corporate changes to perestroika; becoming EDI ready and capable
is no small task.
<p>Being EDI ready and capable - that is, being prepared to exchange EDI documents
between your databases and the databases of others (so-called <i>"trading partners"</i>) -
requires surmounting several non-trivial programming issues that constitute what has
come to be known as the <a href="/edi/chalenge.htm"><i>"mapping problem"</i></a>.
Both the problem itself, as well as its solution, are not well understood.
<p>WinMap addresses the mapping problem; <b>making EDI easier to implement and
maintain.</b> <p><hr>
<i><a href="/edi/index.htm">Home page </A></i>
<br>
<i>Brought to you by <a href="/edi/pfs.htm">PaperFree Systems</A></i>
</BODY>
</html>
```

หน้าตาเว็บเพจของ เอกสารเว็บตัวอย่างนี้ ดังรูป ก.1



รูปที่ ก.1 ตัวอย่างเว็บเพจของเอกสารตัวอย่าง PDDP F-series

ก.1.2 ตัวอย่างไฟล์เอกสาร PDDP F-series ที่ตัดแท็กออกแล้ว

PaperFree Systems - EDI Implementation.

EDI Implementation.

.EDI has proven to be more difficult in practice than in theory. In our judgment, the toughest obstacles technical and managerial to more widespread use of EDI have been the direct result of . to integrate EDIinto current business practices and processes. Some have likened the required corporate changes to perestroika; becoming EDI ready and capable is no small task.

.Being EDI ready and capable - that is, being prepared to exchange EDI documents between your databases and the databases of others (so-called .trading partners".)- requires surmounting several non-trivial programming issues that constitute what has come to be known as the ."mapping problem".

Both the problem itself, as well as its solution, are not well understood.

.WinMap addresses the mapping problem; .making EDI easier to implement and maintain..

.Brought to you by .

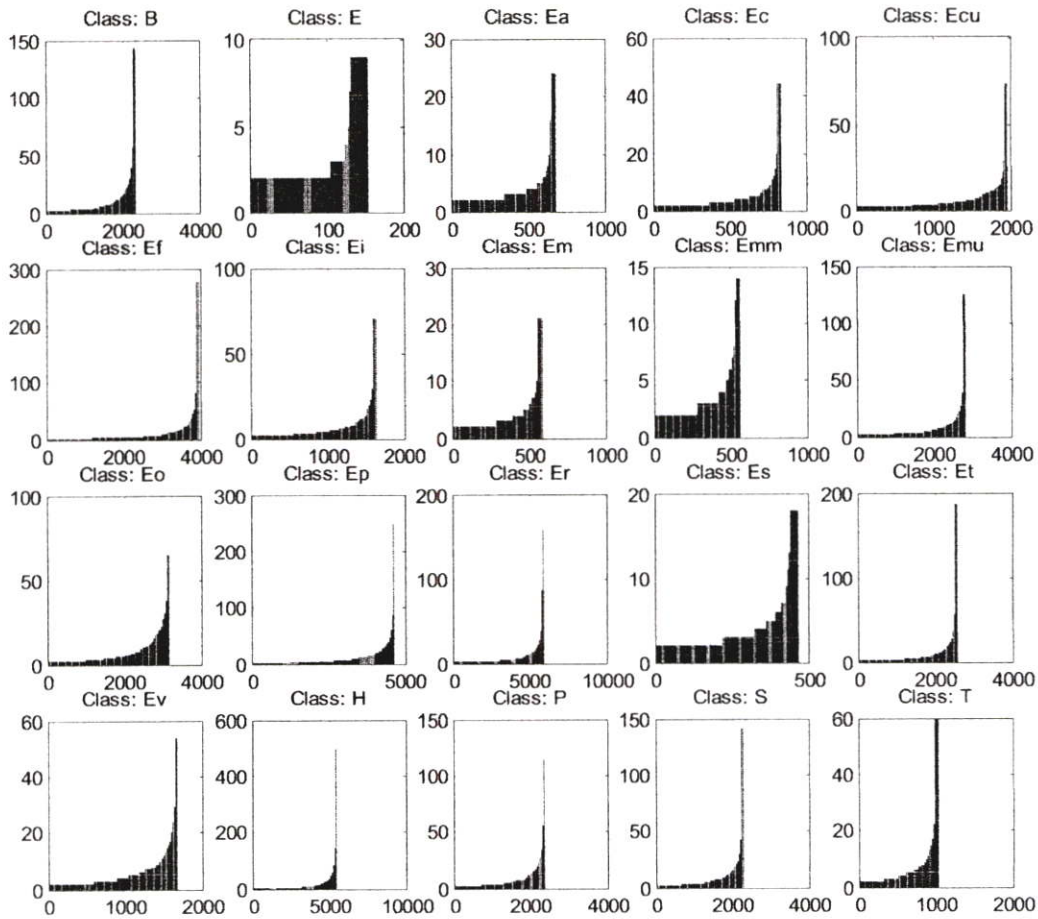
ก.1.3 วลีที่สกัดได้จากตัวอย่างไฟล์เอกสาร PDDP F-series

หลังจากตัดแท็กแล้วเอกสารจะถูกตัด stopwords และทำการสเต็มมิ่ง ได้เป็นเซตของวลีที่สกัดได้จากเอกสารตัวอย่างนี้ ดังนี้

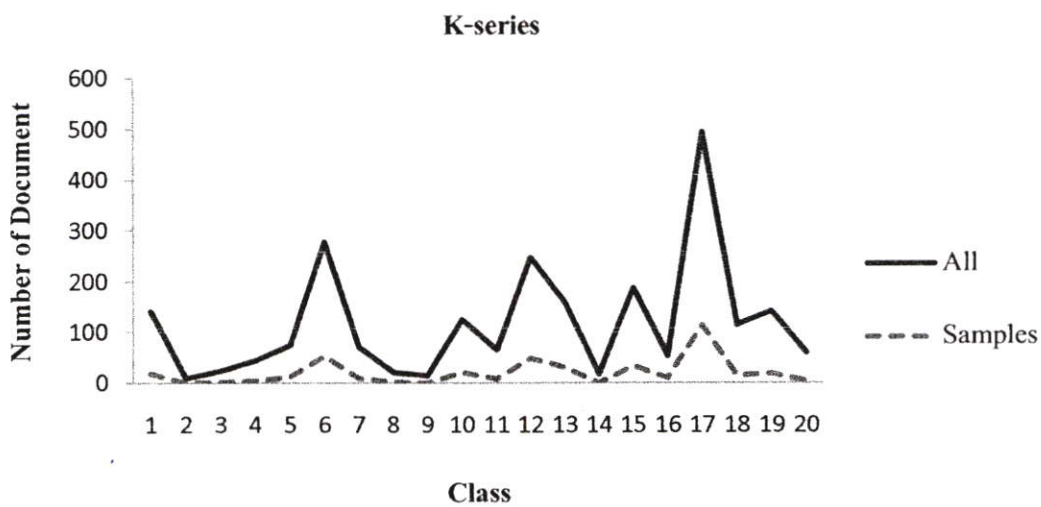
brought	edi implement	map problem	requir corpor chang
cal	edi readi	non trivi	requir surmount
capabl	exchang edi	paperfre system	small task
constitut	implem	perestroika	solution
corpor	integr edi	practic	theori
current busi practic	judgment	prepar	toughest obstacl technic
databas	known	problem	trad partner
difficult	liken	process	understood
direct result	maintain	program issu	widespread
docum	mak edi easier	proven	winmap address
edi	manageri	reluct	

ก.2 การสุ่มตัวอย่างเอกสารของชุดเอกสาร PDDP K-series

ชุดเอกสาร K-series มีเอกสารทั้งหมด 2,340 แบ่งเป็น 20 ประเภท ทำการสุ่มตัวอย่างแบบชั้นภูมิ โดยใช้อัตราส่วนในการเลือกเอกสารจากแต่ละชั้นภูมิเท่ากับ 0.05 รูปที่ ก.2 แสดงแผนภูมิแท่งของชั้นภูมิแต่ละประเภทเอกสาร โดยแกนนอนคือ ชั้นภูมิที่เป็นคำต่างๆ และแกนตั้งคือจำนวนเอกสาร ที่มีคำนั้นๆ อยู่ โดยในที่นี้ได้เรียงลำดับจากชั้นภูมิที่มีจำนวนเอกสารน้อยสุดไปมากสุด หลังการสุ่มแล้ว ได้เอกสารตัวอย่างในแต่ละชั้นภูมิดังแสดงในตารางที่ ก.1 โดยประเภทที่มีเครื่องหมาย * คือประเภทเอกสารที่มีเอกสารจำนวนน้อย ซึ่งเราจะไม่นำมาใช้ในการทดลอง โดยจากเอกสารที่สุ่มได้ในตอนแรก 397 เอกสาร 20 ประเภทก็จะเหลือ 387 เอกสาร 13 ประเภท และสัดส่วนของจำนวนเอกสารแสดงในรูปที่ ก.3



รูปที่ ก.2 แสดงชั้นภูมิคำและจำนวนเอกสารในแต่ละประเภทเอกสารของชุดเอกสาร K-series



รูปที่ ก.3 แสดงสัดส่วนระหว่างจำนวนเอกสารทั้งหมดกับเอกสารตัวอย่างในแต่ละประเภท ของชุดเอกสาร K-series

ตารางที่ ก.1 แสดงจำนวนเอกสารตัวอย่าง K-series แบ่งตามประเภทเอกสาร

ชื่อประเภท	จำนวนเอกสาร	
	ทั้งหมด	ตัวอย่าง
Business (B)	142	18
Entertainment (E)*	9	0
Art (Ea)*	24	1
Cable (Ec)*	44	4
Culture (Ecu)	74	12
Film (Ef)	278	52
Industry (Ei)	70	9
Media (Em)*	21	1
Multimedia (Emm)*	14	0
Music (Emu)	125	21
Online (Eo)	65	8
People (Ep)	248	48
Review (Er)	158	30
Stage (Es)*	18	0
Television (Et)	187	33
Variety (Ev)	54	10
Health (H)	494	113
Politics (P)	114	14
Sports (S)	141	19
Technology (T)*	60	4
รวม	2,340	397

ก.3 ชุดเอกสาร Reuters-top10

ชุดเอกสาร Reuters-top10 มีเอกสารที่อยู่ในรูปแบบ XML มีรายละเอียดและขั้นตอนการเตรียมเอกสารต่างๆ ดังนี้

ก.3.1 ตัวอย่างไฟล์เอกสาร Reuters-top10 ในรูปแบบ XML

รูปแบบของเอกสารก่อนการตัดแท็ก เป็นดังนี้

<?xml version="1.0" ?>
MESSAGE>USDA REPORTS EXPORT SALES ACTIVITY WASHINGTON, April 9 - The U.S. Agriculture Department said private U.S. exporters reported sales of 200,000 tonnes of wheat to Jordan, 300,000 tonnes of soybean meal to Iraq and 100,000 tonnes of corn to Algeria. The wheat for Jordan includes 165,000 tonnes of hard red winter and 35,000 tonnes of soft red winter and is for delivery during the 1987/88 marketing year. The soybean meal sales to Iraq includes 180,000 tonnes for delivery during the 1986/87 season and 120,000 tonnes during the 1987/88 season, the department said. The 100,000 tonnes of corn sales to Algeria are for delivery during the 1986/87 season, it said. The marketing year for wheat begins June 1, corn September 1, and soybean meal October 1. Reuter</MESSAGE>

ก.3.2 ตัวอย่างไฟล์เอกสาร Reuters-top10 ที่ตัดแท็กออกแล้ว

usda reports export sales activity washington, april 9 - the u.s. agriculture department said private u.s. exporters reported sales of 200,000 tonnes of wheat to jordan, 300,000 tonnes of soybean meal to iraq and 100,000 tonnes of corn to algeria. the wheat for jordan includes 165,000 tonnes of hard red winter and 35,000 tonnes of soft red winter and is for delivery during the 1987/88 marketing year. the soybean meal sales to iraq includes 180,000 tonnes for delivery during the 1986/87 season and 120,000 tonnes during the 1987/88 season, the department said. the 100,000 tonnes of corn sales to algeria are for delivery during the 1986/87 season, it said. the marketing year for wheat begins june 1, corn september 1, and soybean meal october 1. reuter

ก.3.3 วลีที่สกัดได้จากตัวอย่างไฟล์เอกสาร Reuters-top10

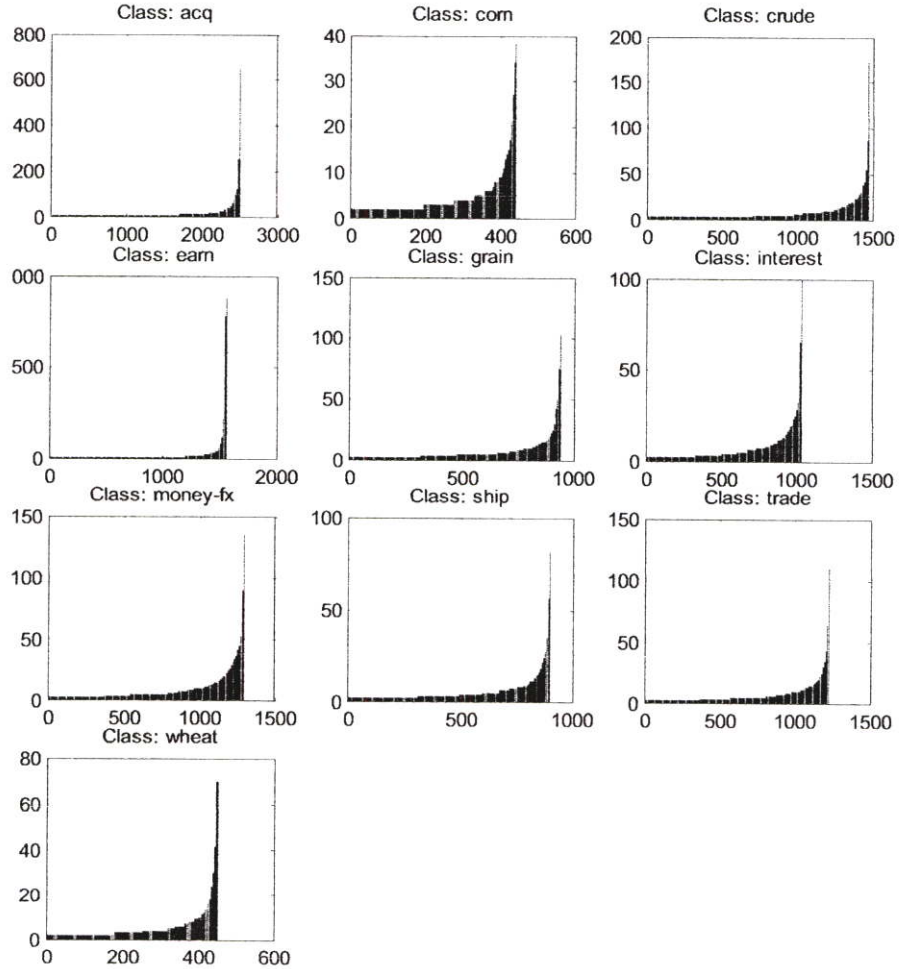
หลังจากตัดแท็กแล้วเอกสารจะถูกตัด stopwords และทำการสเต็มมิ่ง ได้เป็นเซตของวลีที่สกัดได้จากเอกสารตัวอย่างนี้ ดังนี้

agricultur depart said	export report sal	privat	soybean meal sal
algeria	hard red	reuter	tonn
april	iraq	said	usda report export sal activ
corn	iraq includ	season	washington

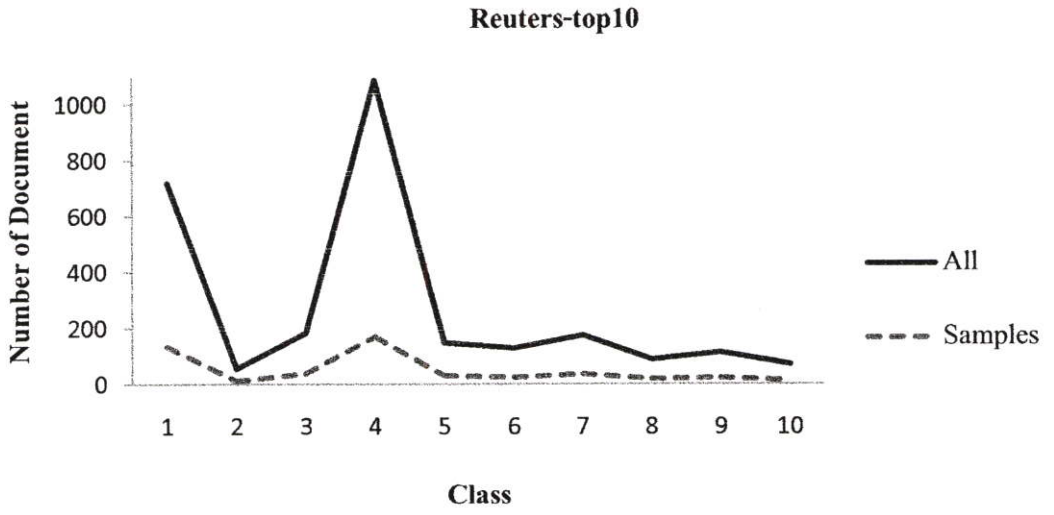
corn sal	jordan	soft red winter	wheat
corn septemb	jordan includ	soybean meal	wheat begin jun
deliveri	market year	soybean meal octob	winter
depart said			

ก.4 การสุ่มตัวอย่างเอกสารของชุดเอกสาร Reuters-top10

ชุดเอกสาร Reuters-top10 มีเอกสารทั้งหมด 2,770 แบ่งเป็น 10 ประเภท ทำการสุ่มตัวอย่างแบบชั้นภูมิ โดยใช้อัตราส่วนในการเลือกเอกสารจากแต่ละชั้นภูมิเท่ากับ 0.1 รูปที่ ก. 4 แสดงแผนภูมิแท่งของชั้นภูมิแต่ละประเภทเอกสาร โดยแกนนอนคือ ชั้นภูมิที่เป็นคำต่างๆ และแกนตั้งคือจำนวนเอกสาร ที่มีคำนั้นๆ อยู่ โดยในที่นี้ได้เรียงลำดับจากชั้นภูมิที่มีจำนวนเอกสารน้อยสุดไปมากสุด หลังการสุ่มแล้ว ได้เอกสารตัวอย่างในแต่ละชั้นภูมิดังแสดงในตารางที่ ก.2 ได้เอกสารตัวอย่างทั้งหมด 504 เอกสารจากทั้ง 10 ประเภท สัดส่วนของจำนวนเอกสารแสดงในรูปที่ ก.5



รูปที่ ก.4 แสดงชั้นภูมิคำและจำนวนเอกสารในแต่ละประเภทเอกสารของชุดเอกสาร Reuters-top10



รูปที่ ก.5 แสดงสัดส่วนระหว่างจำนวนเอกสารทั้งหมดกับเอกสารตัวอย่างในแต่ละประเภทของชุดเอกสาร Reuters-top10

ตารางที่ ก.2 แสดงจำนวนเอกสารของชุดเอกสาร Reuters-top10 แบ่งตามประเภทเอกสาร

ชื่อประเภท	จำนวนเอกสาร	
	ทั้งหมด	ตัวอย่าง
acq	719	135
corn	56	12
crude	184	38
earn	1,085	169
grain	148	31
interest	129	25
money-fx	176	37
ship	89	19
trade	113	24
wheat	71	14
รวม	2,770	504

ก.5 อัลกอริทึม Phrase-tree-based Feature Reduction

Algorithm: Phrase-tree-based Feature Reduction

Input: D – training set of documents; d = $d_1, d_2, \ldots d_n$

C – set of categories; C = $c_1, c_2, \ldots c_m$

L –set of label of each documents in D; $L(d_i) \in C$

min_f – minimum frequency

Output: F: Set of Features

$F = \phi$

// Phrase I : Phrase-tree Construction & Local Reduction

For i=1 to m

$P = \phi$

For each document $d_j \in D$ that $L(d_j) = c_i$

PH = set of extracted phrases from d

$P = P \cup PH$

End

P_tree_i= construct_phrase_tree(P) *//generate phrase-tree rom P*

P_tree_i= calulate_odds_ratio(P_tree_i) *//calculate OR for each node in tree*

P_tree_i= reduce_phrase_tree(P_tree_i, min_f) *//Local Reduction*

End

// Phrase II : Phrase-tree Merging & Global Reduction

$M\_tree = \phi$

For i=1 to m *//merge phrase-tree from each category*

M_tree = merge_tree(M_tree, P_tree_i)

End

M_tree = reduce_phrase_tree(M_tree, min_f) *//Global Reduction*

F = all nodes from M_tree

ก.6 รายการคำหยุด (Stopwords) ที่ใช้ในการทดลอง

a	caption	get	kg	nobody	sm	using
ii	cc	gf	kh	none	sn	uy
about	cd	gg	ki	nonetheless	so	uz
above	cf	gh	km	noone	some	v
according	cg	gi	kn	nor	somehow	va
across	ch	gl	kp	not	someone	vc
actually	ci	gm	kr	nothing	something	ve
ad	ck	gmt	kw	now	sometime	very
adj	cl	gn	ky	nowhere	sometimes	vg
ae	click	go	kz	np	somewhere	vi
af	cm	gov	l	nr	sr	via
after	cn	gp	la	nu	st	vn
afterwards	co	gq	last	nz	still	vu
ag	co.	gr	later	o	stop	w
again	com	gs	latter	of	su	was
against	copy	gt	lb	off	such	wasn
ai	could	gu	lc	often	sv	wasn't
al	couldn	gw	least	om	sy	we
all	couldn't	gy	less	on	sz	we'd
almost	cr	h	let	once	t	we'll
alone	cs	had	let's	one	taking	we're
along	cu	has	li	one's	tc	we've
already	cv	hasn	like	only	td	web
also	cx	hasn't	likely	onto	ten	webpage
although	cy	have	lk	or	text	website
always	cz	haven	ll	org	tf	welcome
am	d	haven't	lr	other	tg	well
among	de	he	ls	others	test	were
amongst	did	he'd	lt	otherwise	th	weren
an	didn	he'll	ltd	our	than	weren't
and	didn't	he's	lu	ours	that	wf
another	dj	help	lv	ourselves	that'll	what
any	dk	hence	ly	out	that's	what'll

anyhow
anyone
anything
anywhere
ao
aq
ar
are
aren
aren't
around
arpa
as
at
au
aw
az
b
ba
bb
bd
be
became
because
become
becomes
becoming
been
before
beforehand
begin
beginning
behind
being
below
dm
do
does
doesn
doesn't
don
don't
down
during
dz
e
each
ec
edu
ee
eg
eh
eight
eighty
either
else
elsewhere
end
ending
enough
er
es
et
etc
even
ever
every
everyone
everything
everywhere
her
here
here's
hereafter
hereby
herein
hereupon
hers
herself
him
himself
his
hk
hm
hn
home
homepage
how
however
hr
ht
htm
html
http
hu
hundred
i
i'd
i'll
i'm
i've
i.e.
id
ie
if
m
ma
made
make
makes
many
maybe
mc
md
me
meantime
meanwhile
mg
mh
microsoft
might
mil
million
miss
mk
ml
mm
mn
mo
more
moreover
most
mostly
mp
mq
mr
mrs
ms
msie
mt
over
overall
own
p
pa
page
pe
per
perhaps
pf
pg
ph
pk
pl
pm
pn
pr
pt
pw
py
q
qa
r
rather
re
recent
recently
reserved
ring
ro
ru
rw
s
sa
same
the
their
them
themselves
then
thence
there
there'll
there's
thereafter
thereby
therefore
therein
thereupon
these
they
they'd
they'll
they're
they've
thirty
this
those
though
thousand
three
through
throughout
thru
thus
tj
tk
tm
tn
to
what's
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
who'd
who'll
who's
whoever
NULL
whole
whom
whomever
whose
why
will
with
within
without
won
won't
would
wouldn

beside
besides
between
beyond
bf
bg
bh
bi
billion
bj
bm
bn
bo
both
br
bs
bt
but
buy
bv
bw
by
bz
c
ca
can
can't
cannot
except
f
few
fi
fifty
find
first
five
fj
fk
fm
fo
for
former
formerly
forty
found
four
fr
free
from
further
fx
g
ga
gb
gd
ge
il
im
in
inc
inc.
indeed
information
instead
int
into
io
iq
ir
is
isn
isn't
it
it's
its
itself
j
je
jm
jo
join
jp
k
ke
mu
much
must
mv
mw
mx
my
myself
mz
n
na
namely
nc
ne
neither
net
netscape
never
nevertheless
new
next
nf
ng
ni
nine
ninety
nl
no
sb
sc
sd
se
seem
seemed
seeming
seems
seven
seventy
several
sg
sh
she
she'd
she'll
she's
should
shouldn
shouldn't
si
since
site
six
sixty
sj
sk
sl
together
too
toward
towards
tp
tr
trillion
tt
tv
tw
twenty
two
tz
u
ua
ug
uk
um
under
unless
unlike
unlikely
until
up
upon
us
use
used
wouldn't
ws
www
x
y
ye
yes
yet
you
you'd
you'll
you're
you've
your
yours
yourself
yourselves
yt
yu
z
za
zm
zr
z

ภาคผนวก ข.

ฟีเจอร์ของชุดเอกสารต่างๆ ที่ได้จากการทดลอง

ข.1 ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีของชุดเอกสาร F-series

ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีจากชุดการทดลองทั้ง 4 ชุด ได้จำนวนฟีเจอร์ทั้งหมด 3,569 ฟีเจอร์ ดังนี้

abbrevi	acknowledg	adopt standard ident	agenc mean
abfd	acorn	advanc	agent
abil	acquir	advanc manufactur	aggreg
abl	acquisi	advanc manufactur system	aggress
abrasion	act	advanc manufactur technologi	agil
absenc	act mean	advantag	agil manufactur
absolut	action	advantag nasa	ago
abstract	action offic	advent	agre
ac	action plan	advers	agreem
ac polici	action program	advert	agricultur
academ	activ	advertis	agricultur servic
accept	activ involv	advic	agricultur worker
access	actual	advis	ahead
accid	ada	advisor	aid
accommod	adam	advoc	aid design
accomod	adapt	aerogel	aid manufactur
accompani	adapt layer	aerospac	aid manufactur engin
accomplish	add	affactweb	aim
accord	addition	affect	air
accordingli	address	affect affirm	alberta
account	adequ	affili	alert
account receiv	adjust	affirm	algorithm
accredit	administ	affirm action	alien
accredit standard	administr	affirm action plan	alleg
accredit standard committe	administr law	affirmative	alli
accru	administr law judg	affirmative ac	alloc
accur	administr procedur	affirmative ac polici	alloi
accuraci	admit	afford	allow
achiev	adop	afl	allow user
acid	adopt	afl cio	alpha
ack	adopt standard	agenc	alphabet

alter
altern
aluminum
amend
america
american
american manufactur
amount
amp
amsant
amus
analog
analogi
analys
analysi
analyt
analyt tool
analyz
anchorag
anecdot
angel
anim
ann
ann arbor
announc
annual
annuiti
ansi
answer
antenna
anti
anticip
antiqu
apart
apc
appar
appeal
appeal rul
appear
appli
appli scienc
applic
applic form
applic procedur
applic protocol
applic system
applications
applications ori
appoint
appreci
appris
approach
appropri
approv
approxim
april
arbitr
arbitrag
arbor
architectur
archiv
area
area includ
argu
argum
aris
arisen
arizona
arizona stat
arizona stat univers
arlington
arm
armi
arrai
arrang
arriv
art
articl
artist
artist work
asc
asid
ask
aspect
assault
assembl
assembli
assembli compon
assembli compon requir
assert
assess
asset
assign
assist
assoc
associ
assum
assur
astm
asynchron
asynchron transfer
asynchron transfer mod
atlanta
atm
atm bas
atm cell
atm forum
atm network
atm switch
atom
atp
attach
attack
attempt
atten
attend
attornei
attornei fee
attornei gener
attract
attribut
audio
audio tap
audiovisu
auditorium
august
austin
authent
author
authorship
auto
autom
autom system
automat
automot
avail
avenu
averag
avoid
awai
awar
award
axi
back
backbon
background
bad
bai
bai area
balanc
ball
ballot
band
bandwidth
bank
banker
bar
bar cod
bargain
bargain agreem
barrel
barrel plat
barrier
bas

bas client
bas network
basi
basic
basic fact
battl
baud
baud rat
bear
began
begun
behalf
behavior
believ
bell
belong
beneficiari
benefit
benefit payment
berkelei
bern
bern conven
best
best manufactur
best manufactur practic
best manufactur practic survei
best practic
better
better repeat
bid
big
bit
black
blank
block
blow
board
boat
bob
bodi
bona
bona fid
bona fid occup
bond
book
booklet
bor
borrow
boston
bother
bottleneck
box
branch
brass
break
bridg
brief
briefli
bright
brilliant
bring
brittl
broad
broadband
broadcast
broadli
brochur
broken
broker
brought
browser
budget
budget plan
buffer
bui
build
build good
build good credit
built
bulk
bulletin
bundl
burden
bureau
burn
burnish
busi
busi administr
busi applic
busi cycle
busi docum
busi oper
busi plan
busi process
busi transac
butt
buyer
byte
cabl
cabl televi
cad
cad model
cad system
cait
cal
calendar
california
california civil
california civil right
california civil right initi
call
caller
caller number
cam
cam system
canada
candid
cap
capabl
capac
capit
capit fund
captur
car
car analysi
carbon
card
career
carefulli
carnegi
carnegi mellon
carnegi mellon univers
carolina
carri
carrier
cas
cas involv
cas suggest
cash
cash flow
cassett
catalog
catalyst
categori
caught
caus
caution
cell
cellular
cellular telephon
censu
center
centr
central
centuri
certain
certain condition
certain minimum
certainli
certif
certif requir
certifi
cfr
cfr part

chain
challeng
chanc
chang
channel
chapter
chapter appli
charact
characterist
charg
charl
chart
cheaper
check
chemic
chemistri
chen
chicago
chief
child
children
choic
choos
chosen
christoph
chronolog
chuck
chuck mclean
cio
cir
circl
circuit
circuit court
circular
circumst
cit
citi
citizen
civ
civil
civil right
civil right act
civil right initi
civil servic
civilian
claim
clarifi
class
classic
classif
classifi
claus
clean
clear
clearanc
clearinghous
clearli
clemson
clemson univers
clerk
client
client incorpor
clinton
clock
clos
club
cnc
coat
cod
cold
collabor
collater
collec
collect
collect bargain
collect bargain agreem
colleg
collin
color
columbia
column
combin
combus
comfort
command
commenc
comment
commerc
commerc integr
commerc integr facil
commerc technologi
commerci
commerci offic
commis
commission
commit
committe
commod
common
common databas
common languag
common law
commonli
commonli refer
commun
commun colleg
commun infrastructur
commun servic
commun softwar
communist
compani
compani establish
compani polici
compani progress
company
company specif
compar
compar feder
compar feder standard
compat
compens
compensatori
compet
competi
competit
competit advantag
competit advantag nasa
competitor
compil
complain
complaint
complet
complet applic
complet stat
complex
compli
complianc
complianc program
complic
compon
compon materi
compon requir
compos
composi
composit
compound
comprehens
compress
compris
compulsori
compulsori licens
compuserv
comput
computer
computer aid
computer aid manufactur
computer aid manufactur engin
conceiv
concentr
concept
conceptu
concern
conclud

conclusion
concurr
concurr engin
condition
conduct
conduct busi
conduct patent
conduct workplac
conduct workplac inspec
conduit
confer
confid
confidenti
configur
confin
conflict
conform
confus
confusingli
confusingli similar
confusion
conges
congress
congression
conjunc
connec
connect
connect industri
connect industri repres
connecticut
consensu
consent
consequ
conserv
consid
consider
consist
consortia
consortium
constant
constitu
constitut
construc
construct
consult
consum
consum credit
consum group
consum nation
consum nation bank
consump
contact
contact fcnb
contact milner
contact milner llnl
contain
contempl
content
contentsbrbr
context
conting
continu
continu improv
contract
contract relat
contractor
contractu
contrari
contrast
contribu
contribut
control
control system
controll
controversi
convei
conven
conveni
conver
convert
convert debt
convinc
cool
cooper
cooper research
coordin
coordin council
copi
copper
copyright
copyright act
copyright law
copyright materi
copyright offic
copyright owner
copyright propos
copyright protec
copyright protect
copyright work
cor
cornell
corner
corp
corpor
correc
correct
correct action
correctli
correspond
corrosion
cosmet
cost
cost sav
costli
council
counsel
counter
counti
countri
coupl
cours
court
court decision
court held
court order
cover
cover public
coverag
cpu
creat
creation
creativ
credit
credit card
credit manag
credit manag tip
crew
criteria
critic
critical
crop
cross
cryogen
crystal
csl
ct
cultur
currenc
current
curriculum
custom
custom produc
custom produc capabl
custom satisfac
custom servic
cut
cut tool
cwa
cwa local
cwatx
cyberspac
cycle
dai
daili

damag
danger
dat
data
data bas
data bit
data gener
data interchang
data link
data manag
data path
data signal
data structur
data transfer
data type
databas
davi
david
day
day haul
dead
deal
dealer
dealt
dearborn
death
debat
debt
debt instrum
dec
decad
decemb
decid
decision
decisionmak
declin
decod
decomposi
decre
dedic
deduc
deduct
deem
deep
defeat
defect
defend
defens
defin
defini
definit
degrad
degre
delai
delawar
deliv
deliveri
demand
demonstr
deni
denial
dental
depart
depend
deploi
deploym
deposi
deposit
depositori
depriv
depth
deriv
describ
descrip
descript
design
design methodologi
design methodologi group
design process
design research
desir
detail
detec
detect
deterior
determin
develop
develop design
develop environ
develop project
develop solution
devic
devic driver
devot
di
diagnost
diamet
diamond
diamond tool
diamond turn
diamond turn machin
dictat
dietitian
differ
differ type
differenti
difficult
difficulti
digit
digit commun
digit equip
digit equip corpor
dimen
dimension
diminish
din
dip
direc
direct
directli
directli involv
director
directori
directori form
dirt
disa
disabl
disabl act
discard
discard polici
discharg
disciplin
disclaim
disclos
disclosur
disclosur docum
discount
discov
discoveri
discretionari
discrimin
discrimin law
discriminatori
discus
discuss
disk
dismiss
dispar
displai
dispos
disput
dissemin
distanc
distinct
distinctli
distinguish
distribu
distribut
district
divers
divid
dividend
division
doc

doctrin
docum
dod
doe
dol
dol mean
dollar
domain
domain logic
domest
domin
door
doubl
dout
download
downsid
draft
draft intern
draft intern standard
drain
dramat
draw
drawn
dream
dri
dril
driv
driven
driver
drop
drop cell
drug
dry
drying
dual
ductil
durabl
durat
duti
dye
dynamic
earli
earli stag
earlier
earn
earth
easi
easier
easili
east
east avenu
echo
ecif
econom
economi
edg
edi
edi implement
edi messag
edi program
edi standard
edi translat
edi user
edi work
edifact
edition
educ
educ foundat
educ plan
eeo
eeoc
effect
effect dat
effici
effort
einet
elect
elect offici
electr
electro
electro opt
electroless
electron
electron commerc
electron commerc integr
electron commerc integr facil
electron commerc technologi
electron data
electron data interchang
electron exchang
electron mail
electropl
elem
elig
elimin
email
embodi
embrac
emerg
emis
emphasi
emploi
employe
employe mean
employe relat
employe train
employer
employer need
employer shall
employm
employm agenc
employm discrimin
employm law
employm offic
employm opportun
employm opportun commis
employm practic
employm relat
employm servic
employm servic activ
enabl
enact
encapsul
encod
encount
encourag
encourag stat
encryption
energi
energi act
energi oak
energi oak ridg
energi system
enforc
engag
engin
engin applic
engin depart
engin design
engin educ
engin function
engin laboratori
engin societi
engin softwar
engin support
engin tool
engin tool kit
engin toolkit
enhanc
enroll
ensur
enter
enterpr
enterpr integr
enterpr integr model
enterpr mod
enterpr represent
entir
entiti
entitl
entrepreneur
entri
envelop
environ

environment
envision
enzym
equal
equal employm
equal employm opportun
equal employm opportun commis
equal opportun
equilibrium
equip
equip corpor
equip leas
equiti
equiti particip
equival
era
ergonom
erosion
error
escap
especi
essenc
essenti
establish
estat
estim
eta
etch
ethernet
ethnic
european
evalu
event
eventu
evid
evolu
evolv
exact
exactli
exagger
examin
exampl
exce
excell
excep
excess
exchang
exchang data
exclud
exclus
exclus right
execu
execut
execut order
exempt
exerc
exercis
exhaus
exhaust
exhibit
exist
exist group
expan
expand
expect
expedit
expens
experi
experienc
experiment
expert
expir
explain
explan
exploit
explor
export
expos
exposur
expres
express
expressli
ext
exten
extend
extend benefit
extens
extent
extern
extra
extrac
extrem
fabric
fac
facil
facilit
facsimil
fact
fact sheet
fact sheet osha
factor
factori
faculti
fail
failur
fair
fairli
fall
fals
famili
famili member
familiar
famou
faq
far
farm
farm labor
farm labor contractor
farmwork
fashion
fast
faster
favor
fax
fcnb
fear
feasibl
featur
feature
februari
feder
feder agenc
feder civilian
feder contract
feder copyright
feder copyright protec
feder financi
feder financi assist
feder fund
feder govern
feder law
feder regist
feder regul
feder standard
federal
fee
feel
fell
feng
ferment
ferri
fiber
fid
fid occup
field
field offic
fight
figur
fil
fil amend
fil fee
film
filter

fin
final
financ
financi
financi assist
finish
fir
firm
fiscal
fiscal year
fit
fix
fixat
flat
flexibl
flood
floor
florida
flow
flow control
fluctuat
fmla
focu
focus
follow
food
foot
footnot
forbid
forc
forecast
foreign
foreign patent
forese
forget
form
formal
format
formul
fort
fort collin
forth
forum
forward
foster
foundat
fourth
fram
framework
franc
franchis
frank
fraud
fre
frequenc
frequenc band
frequent
frequent ask
fri
friend
ftp
fuel
full
full tim
fulli
function
fund
fund sourc
fundament
furnac
furnish
fusion
futur
futur atm
futur atm network
futur contract
futur work
gain
galaxi
gam
gap
garden
gari
gari woitena
gather
gaug
gav
gear
gender
gener
gener issu
gener manag
gener principl
gener provision
gener requir
geograph
geometr
geometri
georg
georgia
georgia tech
giant
giv
given
glass
glob
global
goal
goe
gon
good
good credit
gopher
got
govern
govern agenc
govern laboratori
government
grad
graduat
grant
grant fund
grante
graph
graphic
great
great deal
greater
greater invest
greatest
greatli
greatli exagger
green
green paper
greet
grew
gross
ground
group
grow
growth
guarante
guard
guid
guidelin
hair
half
hand
handbook
handicap
handl
happen
happi
harass
hard
hardwar
harm
haul
hav
hawaii
hazard
head
header
headquart

health
health act
health administr
health car
health plan
health program
health standard
hear
heard
heat
heat treat
heavi
height
held
hexagon
high
high accuraci
high degre
high growth
high perform
high preci
high spe
high speed
high speed data
high speed network
high tech
higher
higher level
highest
highli
highli recommend
highlight
hir
histor
histori
hiv
hol
hold
holder
holidai
hop
hopefulli
hospit
host
hot
hotlin
hour
hous
howard
hug
human
human interven
human resourc
human resourc manag
hundr
hypothet
ibm
iceimt
idaho
idea
ideal
ident
identif
identifi
ieee
igc
ignor
iii
ill
illinoi
illustr
imag
imagin
immedi
immers
impact
impair
implem
implement
implement guidelin
implement plan
implic
import
impos
imposs
improp
improv
improv manufactur
improv product
inadequ
inch
incid
includ
includ addition
includ school
incom
incompat
incomplet
inconveni
incorpor
increas
increasingli
increasingli import
incur
independ
index
indian
indiana
indic
individu
individu emploi
induc
industri
industri associ
industri classif
industri collabor
industri consortium
industri engin
industri group
industri need
industri particip
industri partner
industri partnership
industri repres
industri research
industry
inexpens
info
inform
infrastructur
infrastructur compon
infring
infring suit
ingredi
inherit
initi
injuri
ink
innoc
innov
input
input link
input rat
insert
insight
inspec
inspect
inspector
inspir
instal
install
instanc
institu
institut
instruc
instrum
instrument
insur
intact
integr
integr circuit
integr computer
integr computer aid
integr computer aid

manufactur
integr division
integr facil
integr model
integr technologi
intellectu
intellectu properti
intellectu properti right
intellig
inten
inten discrimin
intend
intens
intent
inter
interac
interact
interchang
interconnec
interconnect
interdisciplinari
interest
interfac
interfac specif
interfer
interim
intern
intern busi
intern commun
intern confer
intern copyright
intern edi
intern revenu
intern revenu cod
intern standard
intern standard statu
intern trad
intern trad administr
internation
internet
internet column
internet sit
internet support
internetwork
interoper
interpret
interrupt
interven
interview
introduc
invalid
inven
invent
inventor
inventori
invest
invest fund
investig
investor
invit
invit industri
invoic
involuntari
involv
involv copyright
ion
iowa
iowa stat
iowa stat univers
ipandc
ipandc llnl
ipo
ipr
iron
island
iso
iso standard
isol
isp
issu
issu involv
issu patent
item
iuliano
jack
jam
jan
januari
jlt
jlt uofrlaw
job
job applic
job bank
job descrip
job function
job open
job opportun
job safeti
job search
job servic
job train
job train partnership
job train partnership act
joe
joe milner
john
johnson
joint
joint labor
joint labor manag
joint labor manag committe
joint project
jointli
jon
journal
judg
judgment
judici
juli
juli parker
jun
jurisdic
jurisdic oper
jurisdic oper complet
jurisdic oper complet stat
jurren
just
justic
justic depart
justif
kansa
kansa technologi
kbyte
keep
kei
kentucki
kept
kevin
kevin jurren
kevin waltz
kind
kit
know
knowledg
known
lab
label
labor
labor contractor
labor manag
labor manag committe
labor market
labor organ
labor regul
labor relat
labor union
laboratori
laboratori invit
laboratori invit industri
laboratori multidisciplinari
laboratori multidisciplinari capabl
lack
lai

lak
lamin
lan
languag
larg
larg number
large
large scal
larger
largest
laser
lat
latest
lath
laundri
law
law judg
law relat
law requir
law technologi
lawrenc
lawrenc livermor
lawrenc livermor nation
lawrenc livermor nation laboratori
lawsuit
lawyer
layer
layout
lbl
lead
leader
leadership
learn
leas
leaseback
leav
lectur
lectur seri
led
lee
left
legal
legal entiti
legal protec
legibl
legisl
legisl histori
legitim
lend
lender
length
letter
level
level packet
leverag
liabil
liabl
liaison
librari
licens
lif
lif cycle
light
lik
likelihood
limit
limit radio
limit scop
lin
lin rat
link
liquid
list
listen
literari
literari work
literatur
litig
littl
liv
livermor
livermor nation
livermor nation laboratori
livermor nation laboratori invit
livermor nation laboratori invit industri
llnl
llnl directori
llnl directori form
llnl disclaim
load
loan
local
local employm
local govern
local law
local telephon
locat
lock
log
logic
long
long term
longer
look
loop
loos
los
loss
lost
lot
loui
louisiana
low
low prioriti
lower
lower cycle
lower level
lowest
machin
machin connect
machin part
machin process
machin tool
machine
madison
magazin
magic
mail
mail address
mail cod
mail list
mail order
mailbox
main
maintain
mainten
major
mak
maker
manag
manag committe
manag effort
manag mean
manag system
manag tip
manner
manual
manufactur
manufactur applic
manufactur associ
manufactur autom
manufactur bas
manufactur compani
manufactur data
manufactur educ
manufactur engin
manufactur engin laboratori
manufactur engin tool
manufactur engin tool kit
manufactur enterpr
manufactur enterpr integr

manufactur equip
manufactur facil
manufactur network
manufactur oper
manufactur part
manufactur practic
manufactur practic survei
manufactur problem
manufactur process
manufactur program
manufactur research
manufactur resourc
manufactur resourc data
manufactur sandia
manufactur sandia nation
manufactur scienc
manufactur simul
manufactur support
manufactur system
manufactur system integr
manufactur technologi
manufacturing
map
map problem
mar
march
margin
marietta
marietta energi
marietta energi system
mark
market
market innov
market maker
marketplac
martin
martin marietta
martin marietta energi
martin marietta energi system
mask
mask work
massiv
match
match fund
materi
materi engin
materi scienc
matrix
matter
matur
maximum
mbp
mclean
mean
mean occup
mean servic
meant
measur
mechan
mechan engin
media
medic
medicin
medium
meet
mel
mel thrust
mellon
mellon univers
member
membership
memori
mental
mention
mer
mer idea
merger
merit
messag
messag specif
met
metal
metal finish
metall
meter
method
methodologi
methodologi group
metk
metrologi
metropolitan
metropolitan area
mexico
mhz
michael
michigan
mid
middl
migrant
mik
mik iuliano
militari
milner
milner llnl
mind
minim
minimum
minimum cost
minneapoli
minor
minut
mississippi
mistak
misunderstand
mit
mix
mobil
mobil commun
mod
model
modem
modern
modern produc
modest
modif
modifi
modifi march
modul
moistur
molecular
molecular model
moment
monei
monica
monitor
monopoli
month
month period
monthli
monthli payment
mortgag
motion
motion pictur
mount
mov
movem
msfw
msid
mtu
multi
multi media
multi servic
multidisciplinari
multidisciplinari capabl
multimedia
multimedia applic
multimedia commun
multipl
multiplex
mural
mural project
murrai
music
mutual

mutual fund
nacha
nam
nasa
nation
nation bank
nation institut
nation laboratori
nation laboratori invit
nation laboratori invit industri
nation law
nation origin
nation research
nation research council
nation scienc
nation scienc foundat
nation technolog
nation technolog univers
nationwid
nativ
nativ american
natl
natur
navi
navig
ncsl
nearest
nearli
necessari
necessarili
necessit
need
neg
neg consequ
negoti
neighbor
neil
neil christoph
nell
netcom
network
network bas
network interfac
network provid
network resourc
network resourc manag
network testb
neural
neutral
newli
newslett
newslett summar
newslett summar court
newslett summar court decision
nickel
nist
nist laboratori
nist manufactur
nist system
nist system laboratori
nistir
nixon
nois
non
non profit
non profit organ
nondiscrimin
nonexclus
nonprofit
normal
north
north american
northwest
notic
notif
notifi
novel
novemb
nuclear
number
numer
numer control
nurs
nyper
nyper public
oak
oak ridg
oak ridg center
oak ridg nation
oak ridg nation laboratori
obes
object
object ori
object orient
oblig
observ
obstacl
obtain
obviou
obvious
occasion
occup
occup safeti
occur
octob
odd
ofccp
offer
offic
offic receiv
offic requir
offici
offset
ohio
oil
old
older
omit
onli
onlin
open
oper
oper complet
oper complet stat
oper cost
opinion
oppon
opportun
opportun commis
oppos
opt
optic
optim
optimist
optimum
option
order
order describ
orderli
ordinari
ordinarili
oregon
organ
organ interest
organiz
ori
orient
origin
origin work
ornl
osha
osha approv
osha conduct
osha conduct workplac
osha conduct workplac inspec
osha inspec
osha program
osha standard
otc
outer
outgo
outlin
outperform

output
output link
output port
outreach
outsid
overhead
overrid
overview
owner
ownership
ownership position
oxford
oxford univers
oxid
pac
packag
packet
packet siz
packet switch
packet switch network
pag
pai
paid
paint
pair
panel
paper
paperfre
paperfre system
paragraph
parallax
parallax educ
parallel
paramet
paramount
parent
park
parker
part
parti
partial
particip
particular
particular fact
particularli
partner
partner agreem
partnership
partnership act
pass
past
patent
patent applic
patent attornei
patent avail
patent law
patent offic
patent protec
patent right
patent search
path
pattern
paul
payload
payment
pc
peddl
pen
penalti
pencil
pend
peopl
perceiv
percent
percentag
percep
perform
perform comput
period
period mean
perman
permis
permiss
permit
perpetu
person
person asset
person characterist
person contact
personnel
personnel manag
perspect
pertain
peys
peys act
phas
philadelphia
phon
phon call
phon compani
phonograph
photocopi
photograph
phras
physical
physical harm
physical locat
physical scienc
pick
pictur
piec
pilot
pilot implement
pin
pip
pip system
pirat
plac
placem
plain
plan
plan applic
plan goal
plan oper
plan oper cost
plan system
plant
plastic
plat
platform
plaza
plu
point
point diamond
polari
polari plat
polic
polic offic
polici
polish
polit
polymer
polytechn
polytechn institut
poor
popul
popular
port
portabl
portfolio
portion
posit
position
posses
possess
possibl
possibl applic
possibl infring
possibli
post
post offic
postscript
potenti
pound

poverti
pow
power
power suppli
practic
practic survei
pre
precaution
preced
preci
precis
precision
precision engin
precision machin
precision manufactur
predict
predominantli
prefer
preferenti
preferenti treatment
preliminari
premis
premium
prepar
prescrib
present
preserv
presid
presid clinton
presid johnson
presid nixon
presidenti
press
press releas
pressur
presump
pretti
prevent
previou
previous
pric
prim
primari
primarili
princip
principl
print
print version
prior
prioriti
privat
privat employer
privat firm
privat investor
privat placem
privat sector
privileg
probabl
problem
problem aris
proce
procedur
proceed
process
process control
process develop
process equip
process involv
process plan
process plan applic
process plan system
process specif
process specif languag
process worker
processor
procur
produc
produc capabl
produc facil
produc manag
produc manag system
produc part
product
product data
product data manag
product lin
product qualiti
profession
professor
profil
profit
profit margin
profit organ
profit shar
program
program budget
program budget plan
program develop
program highlight
program provid
programm
progress
prohibi
prohibit
projec
project
project includ
promin
promot
promotion
prompt
proof
propag
proper
properli
properti
properti law
properti right
propos
proprietari
proprietorship
prospect
protec
protect
protect trademark
protect worker
protocol
prototyp
prov
proven
provid
provid certain
provid equal
provision
provision govern
pub
public
public accommod
public employe
public employm
public law
public librari
public offer
public perform
public school
public servic
publicli
publish
puerto
puerto rico
pump
punish
pur
purchas
purchas option
purchas order
purpos
pursu
pursuant
put
qualif
qualifi
qualifi person
qualiti

qualiti control
qualiti manag
qualiti manag effort
quantiti
quarterli
question
question aris
queue
quick
quickli
quit
quot
quota
rac
racial
racism
rack
rack plat
radio
radio spectrum
rag
rai
rais
rais capit
rang
rapid
rapid access
rapid prototyp
rapid respons
rapid respons manufactur
rapid tool
rapidli
rar
rat
rat fell
ratio
reach
react
reaction
read
read copyright
reader
readi
readili
readili access
real
real estat
real tim
real world
realiti
realiz
realli
realm
ream
reason
reason accommod
reason attornei
reason attornei fee
reason opportun
recall
receipt
receiv
recipi
recogn
recogni
recommend
record
recordkeep
recov
recoveri
recreat
recruit
reduc
refer
referr
refund
refus
regard
regardless
region
region offic
regist
regist trademark
registr
regul
regul concern
regul specif
regular
regularli
regulatori
rehabilit
rehabilit act
reinstat
reject
rel
relat
relat activ
relat board
relationship
releas
relev
reli
reliabl
relianc
relief
reliev
religi
religion
reluct
remain
remedi
rememb
remitt
remot
remot manufactur
remot procedur
remov
remov soil
render
renew
renssela
rent
rental
repair
repeal
repeat
replac
replenish
report
repres
represent
reprint
reproduc
republican
reput
request
request edi
request form
requir
requir elem
requir employer
research
research center
research council
research group
research laboratori
research program
resembl
reserv
resid
resist
resolu
resolv
resourc
resourc data
resourc manag
respect
respond
respond neg
respons
respons manufactur
rest
restor
restric

restrict
result
retail
retain
retali
retir
retrain
return
rev
reveal
revenu
revenu cod
review
revis
revision
revok
revolu
revolution
rhod
rich
richmond
richmond journal
rico
rid
ridg
ridg center
ridg nation
ridg nation laboratori
right
right act
right confer
right initi
rigid
rins
rippl
ris
risk
riski
road
road map
robert
robot
robust
rock
rol
rom
room
root
roughli
round
rout
router
routin
royalti
rrm
rt
rul
run
rural
saf
safeti
sai
said
sal
salari
sampl
sampl issu
san
sandi
sandia
sandia nation
sandia nation laboratori
santa
santa monica
satellit
satisfac
satisfactori
satisfi
sav
sav account
saw
sba
sba field
sba field offic
sbic
scal
scar
scenario
sched
schedul
schem
scholarship
school
scienc
scienc foundat
scientif
scientist
scop
scor
screen
search
searchabl
season
season farmwork
sec
second
secondari
secondari offer
secret
secretari
secretari mean
secretary
secretary treasur
secretary treasur cwa
secretary treasur cwa local
section
sector
secur
secur act
see
seed
seek
seen
segment
segreg
sel
selec
select
self
sell
semant
semiconductor
seminar
send
sender
senior
sens
sensit
sensor
sent
separ
sept
septemb
seq
sequenc
seri
serial
serial commun
serv
server
servic
servic activ
servic agenc
servic mark
servic mean
servic mean servic
servic perform
servic system
set
set forth
set job
set job safeti
settl
settlem

setup
sever
sex
sexual
sexual discrimin
shall
shall appli
shap
shar
shar memori
sharehold
shaw
shaw feng
sheet
sheet osha
shelf
shift
ship
shipment
shop
shop floor
short
short form
short term
short term debt
shorter
shot
show
shown
sic
sick
sid
sign
signal
signal process
signal rt
signatur
signific
significantli
silica
silica aerogel
silicon
silicon vallei
sima
sima program
similar
similar mark
similar right
similarli
simpl
simpl credit
simpl credit manag
simpl credit manag tip
simpler
simplest
simpli
simpli phon
simplifi
simul
simultan
singl
singl physical
single
single point
single point diamond
sit
situ
situat
siz
skill
skin
slic
slightli
slim
slogan
slop
slow
slower
small
small busi
small busi administr
small compani
small firm
smaller
smaller institu
sme
sme educ
sme educ foundat
smok
smooth
social
social secur
social secur act
societi
soft
softwar
softwar applic
softwar tool
softwar vendor
soil
sol
sold
solicit
solid
solubl
solution
solv
somewhat
song
soon
sophist
sort
sought
sound
sound record
sourc
south
south carolina
southern
spac
spe
speak
special
specialist
specif
specif languag
specif technologi
specifi
spectrum
specul
speech
speed
speed data
speed network
spend
spin
sponsor
spread
spur
stabil
staf
staff
stag
stai
stainless
stak
stand
standard
standard committe
standard cover
standard develop
standard ident
standard industri
standard iso
standard number
standard requir
standard requir employer
standard set
standard statu
standard titl
stanford
stanford univers
stanlei
start

stat
stat agenc
stat agenc mean
stat cod
stat employm
stat employm servic
stat govern
stat job
stat law
stat patent
stat plan
stat program
stat unemploy
stat unemploy compens
stat univers
stat university
statem
station
statist
statu
statut
statutori
statutori protec
steal
steam
steel
stem
step
stereolithographi
steven
stewart
stick
stock
stock offer
stor
storag
stori
straightforward
strateg
strategi
streamlin
street
street address
strength
strengthen
stress
strik
strip
striv
strong
strong technologi
strong technologi bas
strongli
struck
structur
student
studi
sub
sub project
subchapt
subdivid
subject
subject matter
submis
submit
subpart
subscrib
subscrip
subsequ
substanc
substanti
substitut
subtl
subtract
success
success stem
successfulli
suddenli
sue
suffici
sugges
suggest
suit
suitabl
sum
summar
summar court
summar court decision
summari
sun
sup
superintend
supervi
supervisor
supp
supplementari
suppli
suppli chain
supplier
support
support disciplin
support servic
suppos
supra
suprem
suprem court
sur
surcharg
surfac
surfac area
surpris
surround
survei
surviv
suscept
suspen
suspend
sustain
sweden
switch
switch fabric
switch network
symbol
system
system develop
system integr
system integr division
system laboratori
system research
system research center
systemat
tabl
tactic
tailor
tak
taken
takeov
talk
tangibl
tank
tap
target
task
tax
tax advantag
tax benefit
tax credit
tb
tcp
teacher
team
tech
techn
technic
technic approach
technic assist
technic resourc
techniqu
technolog
technolog need
technolog univers
technologi
technologi bas
technologi develop

technologi need
technologi program
technologi transfer
technologi transfer offic
tel
telecommun
telecommun industri
telephon
telephon directori
telephon lin
telephon number
telephoni
televi
tell
temperatur
temperatur measur
temporari
temporarili
tend
tent
term
term debt
termin
territori
testb
tetherless
texa
thank
theater
theori
thermal
thick
thing
think
third
thorough
thoroughli
thought
threat
throughput
thrust
ti
tight
tim
time
timet
tip
tir
titanium
titl
titl iii
toc
todai
toler
toll
toll fre
tomorrow
took
tool
tool appropri
tool avail
tool data
tool kit
toolkit
topic
topologi
total
total domain
total qualiti
total qualiti manag
total qualiti manag effort
touch
track
track record
trad
trad administr
trad associ
trad commis
trad develop
trad partner
trad partner agreem
trad secret
trademark
trademark law
trademark offic
trademark right
trademark suit
tradition
traffic
train
train administr
train cours
train institut
train partnership
train partnership act
train program
transac
transfer
transfer mod
transfer offic
translat
translat softwar
transmis
transmit
transmitt
transport
trap
travel
treasur
treasur cwa
treasur cwa local
treasuri
treat
treati
treatment
tree
tremend
trend
tri
troubl
true
trust
try
trying
turbochannel
turn
turn machin
turn machin tool
tutori
twist
type
typewrit
typical
ucrl
ucrl tb
ultim
unabl
unauthor
unawar
unchang
uncontrol
underli
understand
understood
undertak
underwrit
undesir
unemploy
unemploy compens
unequ
unfair
unfair competi
unfortun
unifi
unifi process
unifi process specif
unifi process specif languag
union
uniqu
unit
unit stat
unit stat cod
univers
univers librari

university
unix
unlaw
unlaw employm
unlaw employm practic
unrel
unusu
uofrlaw
updat
updat mar
updat sched
upgrad
upheld
upper
urg
urich
url
usa
usa ornl
usabl
usag
usdoc
user
user data
user network
usual
utah
util
vacat
vacuum
valid
vallei
valu
valuabl
value
vapor
vari
variabl
variat
varieti
variou
vast
vehicl
vendor
ventur
ventur capit
verifi
version
versu
veteran
viabl
vic
victim
video
vietnam
vietnam era
view
vigor
vii
viii
violat
violenc
virgin
virgin island
virginia
virtual
virtual circuit
virtual enterpr
virtual enterpr mod
virtual librari
virtual manufactur
virtual realiti
vision
visit
visual
vocat
voic
volum
voluntari
vot
wag
wagner
wagner peys
wagner peys act
wai
wait
waiver
walk
waltz
want
war
warrant
wash
washington
wast
water
waterloo
wav
wax
weak
weaken
weapon
wear
webmast
week
weekli
weight
welfar
west
western
whit
whit hous
whit paper
wholli
wid
wide
wider
widespread
wil
window
wir
wireless
wireless commun
wireless network
wisconsin
wish
wit
withdrew
withhold
woitena
woman
women
wong
word
work
work capit
work dai
work forc
wŏrk plan
work schedul
worker
worker emploi
worker mean
workforc
workplac
workplac inspec
workshop
worksit
workstat
world
world leader
world wid
worldwid
worri
worth
wrap
writ
written
written descrip
year
year ago
year mean
year plan
year plan goal

yield
york
young
youth
zemo
zero
zinc
zip
zip cod

ข.2 ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีของชุดเอกสาร J-series

ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีจากชุดการทดลองทั้ง 4 ชุดได้จำนวนฟีเจอร์ 3,488 ฟีเจอร์ ดังนี้

abandon
abil
abl
abroad
absenc
absent
absolut
absorp
abstrac
abstract
academ
academia
acceler
accept
access
accid
accommod
accompani
accompani measur
accomplish
accord
accordingli
account
account receiv
accur
accuraci
accus
achiev
acknowledg
acoust
acquir
acquisi
act
action
action polici
action program
activ
activist
actual
actuat
ad
adapt
add
addit
addition
address
address includ
adequ
adher
adhes
adhesion
adjust
administ
administr
administr comput
admis
admit
adop
adopt
advanc
advanc composit
advanc comput
advanc manufactur
advanc materi
advanc system
advanc technologi
advantag
advers
advert
advertis
advic
advis
advisor
advisori
adviz
advoc
aeronaut
aerospac
affair
affd
affect
affili
affirm
affirm action
affirm action polici
affirm action program
afford
african
african american
ag
ag discrimin
agenc
agent
aggreg
agil
agre
agreem
agricultur
aid
aid design
aim
air
air forc
airborn
aircraft
airport
alamo
alamo nation

alamo nation laboratori
alert
algebra
algorithm
align
alleg
allegedli
allen
alli
allianc
alloc
alloi
allow
alston
alter
altern
aluminum
amend
america
american
amount
amplifi
analys
analysi
analysi techniqu
analyst
analyt
analyz
anc
andersen
andersen consult
anim
anisotrop
announc
annual
answer
anticip
api
appar
apparatu
apparel
apparel manufactur
appeal
appear
appell
appelle
appli
applic
applic develop
applic form
applic program
applic softwar
applic system
appoint
approach
appropri
approv
approxim
april
aqueou
arbitr
arc
architect
architectur
architectur design
archiv
area
area includ
area network
argu
argum
aris
arkansa
aros
arrai
arrang
art
articl
artifici
artifici intellig
artifici neural
artifici neural network
artist
artist work
asian
asian american
ask
ask question
aspect
assembl
assembli
asser
assert
assess
asset
assign
assist
associ
assum
assump
assur
atlanta
atm
atmospher
atom
attach
attain
attempt
atten
attend
attornei
attornei fee
attract
attribut
audienc
audio
audit
augment
august
authent
author
autom
automat
automobil
automot
autonom
avail
avenu
averag
aviat
avoid
awai
awalt
awar
award
award program
axi
back
background
backup
bad
balanc
band
bank
bank busi
bank loan
banker
bar
barbara
bargain
bargain agreem
bargain unit
barrier
bas
bas approach
bas data
bas design
bas industri
bas manufactur
bas materi
bas system
bas technologi
basi
basic

batch
batteri
beam
bear
began
begun
behalf
behavior
behaviour
belief
believ
bell
benchmark
benefici
benefit
benign
berkelei
berlin
bern
bern conven
bern union
best
best practic
better
bial
bid
big
bilater
bind
bio
biochem
biolog
biolog process
biolog system
biologi
biomass
biomed
biomed applic
biomimet
bioremedi
biotechnologi
bipartisan
bipartisan commis
black
block
board
bodi
bond
book
boost
borrow
boston
boundari
boundari layer
box
brak
braz
breach
break
bridg
brief
bring
broad
broad rang
broadcast
broaden
broader
broadli
brought
brown
browser
bu
budget
bui
build
built
bulk
burden
burn
bush
busi
busi administr
busi capit
busi capit corpor
busi cas
busi develop
busi electron
busi enterpr
busi environ
busi financ
busi innov
busi manag
busi need
busi object
busi owner
busi plan
busi process
busi process engin
busi system
busi transac
button
buyer
cabl
cad
cal
calcul
calendar
calibr
calif
california
call
cam
cambridg
campu
canada
canadian
candid
cap
capabl
capac
capit
capit corpor
capit format
capit fund
capit invest
capit market
captur
car
carbon
card
career
carolina
carri
carrier
cas
cas law
cas studi
cash
cash flow
cast
cast process
cat
catalog
catalysi
catalyst
catalyt
categori
caus
ceil
ceil commis
cell
cellular
cent
center
centr
central
centrifug
centuri
ceram
cert
certain
certainli
certif
certifi

chain
chair
chairman
challeng
chamber
chanc
chang
channel
chapter
charact
character
characterist
charg
charl
chart
check
chemic
chemic engin
chemic process
chemic reaction
chemic vapor
chemic vapor deposi
chemistri
chen
chicago
chief
children
china
chip
choic
choos
chos
chosen
cim
cim model
cim system
cir
circuit
circuit board
circuit judg
circul
circumst
cisba
cit
citi
citizen
civil
civil engin
civil right act
civil servic
civilian
claim
class
classic
classif
classroom
claus
clean
clear
clearli
client
client server
climat
clinton
clos
cluster
coat
cod
coeffici
coher
coincid
cold
collabor
collater
collec
collect
collect bargain
collect bargain agreem
colleg
collision
colloid
color
columbia
com
combin
combus
combus engin
comfort
comm
command
commenc
comment
commerc
commerc depart
commerc requir
commerci
commerci exploit
commis
commit
committe
common
common law
commonli
commun
commun network
commun system
commun technologi
compact
compani
compar
comparison
compat
compel
compens
compet
competi
competit
competit advantag
competitor
compil
complain
complaint
complem
complementari
complet
completion
complex
complex system
compli
complianc
complic
compon
compos
composi
composit
composit materi
composit structur
compound
comprehens
compres
compress
compris
comprom
compulsori
compulsori licens
comput
comput environ
comput system
concentr
concept
conceptu
concern
concern patent
conclud
conclusion
concurr
concurr design
concurr engin
concurr engin environ
condition
conduc
conduct
confer
confid

confidenti
configur
conflict
conform
conform test
confront
cong
congress
congression
congression intent
conjunc
connec
connect
consciou
consent
consequ
conserv
consid
consider
consist
consolid
consortia
consortium
constant
constitu
constitut
constrain
constraint
constru
construc
construc materi
construc process
construct
consult
consult group
consult servic
consum
consum product
consump
contact
contain
contamin
contempl
contend
content
context
conting
continu
continu improv
contract
contract claim
contractor
contractu
contrari
contribu
contribut
control
control technologi
controll
conven
conver
convert
cool
cooper
cooper research
cooper treati
coordin
copi
copper
copyright
copyright act
copyright conven
copyright law
copyright offic
copyright owner
copyright phillip
copyright phillip publish
copyright protec
copyright relat
copyright work
cor
corp
corpor
correc
correct
correl
correspond
corrosion
cost
cost effect
cost sav
cost shar
costli
council
counsel
count
counter
counti
countri
coupl
cours
court
court conclud
court decision
court grant
court held
coven
cover
coverag
crada
creat
creat job
creation
creativ
credit
credit card
creditor
crimin
crisi
criteria
critic
cross
crystal
crystal growth
crystallin
ctiv
ctiv week
cultur
current
curriculum
custom
custom order
custom servic
cut
cycle
cycle engin
cycle tim
dai
daili
damag
dan
daniel
dat
dat problem
data
data acquisi
data bas
data collec
data exchang
data integr
data interchang
data manag
data model
data secur
data structur
data transfer
databas
databas technologi
davi
davi publish
davi publish compani
david
deadlin
deal

dean
debat
debentur
debit
debt
debug
dec
decad
decemb
decentralis
decid
decis
decision
decision analysi
decision mak
decision support
decision support system
decreas
dedic
deduc
deem
deep
defeat
defect
defend
defens
defer
defin
defini
deform
degrad
degre
delai
deliv
deliveri
demand
democrat
demonstr
deni
denomin
densiti
depart
department
depend
deploi
deploym
deposi
deposit
depreci
depriv
dept
depth
deputi
deregul
deriv
describ
descrip
design
design decision
design environ
design method
design patent
design process
design tool
desir
desk
desktop
despit
destroi
destruct
destruct evalu
detail
detec
detect
determin
develop
develop advanc
develop agreem
develop cycle
develop design
develop method
develop methodologi
develop organ
develop proprietari
develop proprietari research
develop requir
develop tim
devic
devis
diagnost
diagram
dialogu
dictat
die
differ
differ level
differ type
differenti
difficult
diffract
diffusion
diffusion bond
digit
digit equip
digit equip corp
digit signal
digit signal process
dimen
dimension
diminish
diod
dioxid
direc
direct
direct invest
directli
director
directori
disabl
disabl benefit
disadvantag
disappear
discharg
disciplin
disciplinari
disciplinari action
disclaim
disclos
disclosur
discount
discov
discoveri
discret
discretion
discrimin
discriminatori
discus
discuss
diseas
disk
dismiss
dispatch
displai
dispos
disposi
disput
dissemin
dissent
distinct
distinguish
distribu
distribut
distribut databas
distribut environ
distribut manufactur
district
district court
district court grant
divers
diversif
diversifi
division
do
doctor

doctrin
docum
document
doe
dol
dollar
domain
domest
donat
door
doubt
downstream
downward
downward pressur
dozen
draft
drag
dramat
draw
drawn
driv
driven
driver
drop
drug
dsp
dual
due
duk
durabl
durat
duti
dynamic
earli
earlier
earliest
earn
earth
eas
easi
easier
easili
east
econom
econom develop
econom growth
econom impact
economi
ectodai
ed
edg
edi
edit
edition
editor
editor continu
educ
educ servic
edward
eeo
effect
effici
effort
elast
elec
elector
electr
electr power
electro
electro optic
electromagnet
electromechan
electron
electron beam
electron catalog
electron circuit
electron data
electron data interchang
electron devic
electron packag
electron payment
electron product
electron system
electron technologi
elem
elig
elimin
email
email address
embed
embodi
emerg
emerg applic
emis
emori
emphas
emphasi
emphasi ad
empir
emploi
employe
employe benefit
employe receiv
employe right
employe right continu
employe work
employer
employm
employm act
employm contract
employm practic
employm relationship
empow
enabl
enabl technologi
enact
encapsul
encompass
encourag
encryption
end
endheadingbr
endors
energi
energi effici
energi storag
enforc
eng
engag
engin
engin applic
engin center
engin design
engin environ
engin materi
engin program
engin system
engin technologi
english
engr
enhanc
enorm
ensur
ensur equal
entail
enter
enterpr
enterpr integr
enterpr model
enterpr wid
entir
entiti
entitl
entrepreneur
entrepreneuri
entri
environ
environment
environment benign
environment friendli
environment impact
environment protec
environment remedi
environment technologi

envisag
enzym
epoxi
equal
equal employm
equal opportun
equat
equilibrium
equip
equip corp
equip design
equit
equiti
equiti financ
equiti invest
equival
era
ergonom
error
escap
especi
esprit
essenti
establish
estat
estim
etch
ethic
ethnic
europ
european
european industri
european market
evalu
evalu techniqu
event
eventu
evid
evolu
evolutionari
evolv
exam
examin
exampl
exampl includ
exce
exceed
excell
excep
exception
excerpt
excess
exchang
exchang transac
exchang valu
excim
exclud
exclus
exclus right
execu
execu system
execut
execut order
execut summari
exerc
exercis
exist
exist applic
exist busi
exist system
exit
expan
expand
expand busi
expect
expenditur
expens
experi
experienc
experiment
expert
expir
explain
explan
explicitli
exploit
explor
explos
export
expos
exposur
expres
express
express written
express written permis
expressli
exten
exten partnership
exten servic
extend
extens
extent
extern
extra
extrac
extract
extrem
extrusion
fabric
fabric process
fac
facet
facil
facilit
fact
factor
factori
faculti
fail
failur
failur analysi
fair
faith
fall
famili
familiar
famou
far
fashion
fast
faster
fat
fatigu
fault
fault toler
favor
fax
feasibl
featur
februari
fed
feder
feder agenc
feder court
feder fund
feder govern
feder labor
feder law
feder program
fee
feedback
feel
fellowship
femal
fewer
fiber
fiber optic
fibr
fiduciari
field
fifth
figur
figur show
fil

fil dat
fil patent
film
filter
fin
fin particl
final
final product
financ
financ program
financi
financi institu
financi intermediari
financi market
financi servic
financi support
find
finish
finish product
finit
finit elem
fir
firm
fiscal
fiscal year
fit
fix
fix asset
flat
flat panel
flexibl
flexibl manufactur
flight
float
floor
floor control
florida
flow
fluid
fluid dynamic
fluid flow
flux
fly
focal
focu
focus
follow
food
food process
footnot
forc
forecast
foreign
foreign applic
foreign bank
foreign countri
foreign patent
form
formal
formal member
format
formul
forrest
forrest research
forth
fortun
forum
forward
foster
found
foundat
fourth
fraction
fractur
fractur mechan
fram
framework
franc
franchis
francisco
frank
fraud
fred
freedom
french
frequenc
frequent
friction
friendli
fring
fring benefit
frustrat
fuel
fulfill
fuller
fulli
fulli integr
function
function requir
fund
fund sourc
fund transfer
fundament
furnish
furthermor
fusion
futur
fuzzi
ga
gain
gam
gap
gas
gather
gatt
gav
gear
gen
gender
gener
gener object
gener process
genet
genet engin
geograph
geometri
georg
georgia
germani
get
giv
given
glass
glass ceil
glass ceil commis
global
global electron
global electron commerc
global market
go
goal
goe
good
good faith
gorrell
govern
govern agenc
government
governor
grac
grac period
grad
gradient
gradual
graduat
graduat student
grant
grant summari
grant summari judgment
graphic
great
greater
greatest
greatli

greg
grievanc
grievanc procedur
grind
gross
ground
group
groupwar
grow
growth
guarante
guid
guidanc
guidelin
half
hamilton
hand
handicap
handl
handl system
happen
har
hard
hardwar
harmon
hav
hazard
hazard wast
head
headquart
health
health car
hear
heart
heat
heat flux
heat transfer
heat treatment
heavi
heavi metal
held
help
helpdesk
hierarch
high
high densiti
high frequenc
high level
high perform
high perform comput
high power
high prioriti
high puriti
high qualiti
high resolu
high risk
high school
high technologi
high temperatur
high volum
higher
higher educ
highest
highli
highli effici
highli integr
highlight
highwai
hill
hir
hispan
histor
histori
hoc
hold
holder
holograph
hong
hong kong
hop
horizont
hospit
host
hostil
hot
hour
hous
hub
human
human activ
human factor
human resourc
human resourc manag
hundr
hybrid
hydrocarbon
ibm
ic
idea
ideal
ident
identif
identifi
ieee
ignor
iii
iitf
ill
illinoi
illinoi stat
illinoi stat univers
illustr
ilstu
imag
immedi
impact
impair
implant
implem
implement
implement plan
impli
implic
import
import rol
impos
improv
improv flexibl
improv product
improv understand
impur
inadvert
incent
incid
includ
includ analysi
includ applic
includ cad
includ chemic
includ design
includ electron
includ metal
incom
inconsist
incorpor
increas
increasingli
increasingli import
increment
incur
indebted
independ
index
indic
individu
individu compani
individu employe
individu right
induc
industri
industri develop
industri engin
industri firm
industri includ

industri need
industri partner
industri partnership
industri partnership offic
industri product
industri properti
industri standard
inequ
inexpens
infer
influenc
info
inform
informix
infosi
infrar
infrastructur
infring
ing
inher
initi
initi capit
injec
injuri
innov
innov technologi
inorgan
input
inquiri
insid
insight
inspec
instal
install
instanc
institu
institut
instruc
instrum
instrument
insuffici
insul
insur
integr
integr circuit
integr design
integr manufactur system
integr model
integr multi
integr process
integr servic
integr system
integr technologi
integr tool
intel
intellectu
intellectu properti
intellectu properti law
intellectu properti right
intellectualproperti
intellig
intellig materi
inten
intend
intens
intensifi
intent
inter
inter ctiv
inter ctiv week
interac
interact
interchang
interconnec
interest
interest parti
interfac
interfer
intermedi
intermediari
intern
intern bank
intern conven
intern journal
internation
internet
internet commerc
internet commun
internet market
interoper
interpret
interv
interven
interview
introduc
inven
invent
inventor
inventori
inventori control
inventori manag
inventorship
invers
invest
invest strategi
investig
investor
invit
invoic
involv
ion
ipr
irrelev
island
iso
isol
issu
item
jam
jan
januari
japan
japanes
java
jet
jit
job
job creat
job creation
job list
job market
john
johnson
join
joint
journal
judg
judgment
judici
juli
jun
jurisdic
just
justic
justifi
kanban
kansa
keep
kei
ken
kept
kernel
kim
kind
kinemat
kinet
kingdom
know
knowledg
knowledg bas
known
kong
lab
labor

labor law
labor manag
labor organ
labor practic
labor relat
labor relat act
labor relat board
laboratori
lack
laid
lak
lamin
lan
land
languag
lanl
larg
larg scal
larger
largest
laser
laser bas
laser process
lat
latest
launch
law
law firm
law provid
lawrenc
lawsuit
lawyer
layer
layout
layout design
lead
lead edg
lead tim
leader
leadership
leakag
lean
learn
leas
leav
led
lee
left
legaci
legaci system
legal
legal advic
legal protec
legisl
legitim
lend
lender
letter
level
level model
leverag
lewi
lewi research
liabil
liaison
librari
licenc
licens
license
lif
lif cycle
lift
light
light weight
lightweight
likelihood
limit
lin
linear
link
linkag
liquid
list
literari
literari work
literatur
lithographi
litig
littl
liv
livermor
livermor nation
livermor nation laboratori
llp
lo
lo alamo
lo alamo nation
lo alamo nation laboratori
load
loan
local
local govern
locat
logic
logist
logo
london
long
long term
long term capit
longer
longer term
look
loop
loss
lost
lot
lot siz
loui
low
low cost
low level
low temperatur
low volum
lower
lower cost
lower rat
lynx
maarten
machin
machin system
machineri
machinist
macintosh
magazin
magnet
magnitud
mail
mail comment
main
mainfram
mainli
maintain
mainten
major
major bank
major intern
mak
maker
mal
man
manag
manag need
manag practic
manag strategi
manag system
manipul
manner
manual
manufactur
manufactur applic
manufactur compani
manufactur cost
manufactur data

manufactur design
manufactur engin
manufactur enterpr
manufactur exten
manufactur exten partnership
manufactur improv
manufactur industri
manufactur oper
manufactur requir
manufactur softwar
manufactur system engin
manufactur system integr
map
march
margin
mari
marin
maritim
mark
market
market plac
market relat
marketplac
mass
mass produc
mass transfer
massachusett
master
mastercard
mastercard intern
match
match fund
materi
materi develop
materi flow
materi handl
materi process
materi properti
materi research
materi scienc
materi technologi
mathemat
matric
matrix
matrix composit
matter
matur
maxim
maximum
mean
meant
measur
measur techniqu
mechan
mechan engin
mechan properti
mechan system
mechatron
media
medic
medium
medium siz
medium siz manufactur
medium term
meet
melt
mem
member
member countri
membership
membran
memori
men
ment
ment contract
mention
menu
mep
mer
merchant
merg
merit
mesh
messag
met
metabol
metal
metal cast
metal form
metal matrix
metal matrix composit
metall
metallurg
metallurgi
meter
method
methodologi
metr
metric
metrologi
mfg
mi
michael
michigan
mid
middl
middlewar
migrat
mik
militari
mill
min
mind
miner
miner process
miniatur
minim
minimis
minimum
minist
minor
minor group
misconduct
mislead
mission
mitig
mix
mixtur
mobil
mocvd
mod
modal
model
model allow
model bas
model data
model techniqu
modem
moder
modern
modif
modifi
modul
modular
mold
molecul
molecular
monei
monitor
monitor system
monolith
monopoli
month
mortgag
mosaic
motion
motiv
motor
mountain
mov
movem
mrp
multi
multi nation

multichip
multidisciplinari
multifunc
multilater
multilay
multimedia
multin
multipl
music
nam
narrow
nasa
nasa lewi
nat
nation
nation engin
nation institut
nation labor
nation labor relat
nation labor relat act
nation labor relat board
nation laboratori
nation origin
nation scienc
nation scienc foundat
natur
natur languag
navi
navig
near
nearli
necessari
necessarili
need
neglect
neglig
negoti
net
netherland
network
network servic
neural
neural network
neutral
neutron
new
newli
newslett
newspap
newsstand
ng
nich
nickel
nist
nlra
nlrb
nod
nois
non
non destruct
non destruct evalu
non equilibrium
non profit
nondestruct
nondestruct evalu
nondestruct test
nonlinear
nonlinear dynamic
nonprofit
normal
north
north america
north american
not
notic
notif
notifi
notwithstand
novel
novelti
novemb
novo
nrtw
nrwldf
nsf
nuclear
number
numer
numer control
numer method
numer model
oak
object
object orient
oblig
observ
obstacl
obtain
obviou
occasion
occup
occur
ocean
octob
offer
offic
offici
ohio
oil
old
omit
on
ongo
onlin
open
open system
oper
oper condition
oper cost
oper manag
oper system
opinion
opm
opportun
oppos
opposit
optic
optic commun
optic fiber
optim
optim problem
optim techniqu
optimum
option
optoelectron
or
oracl
oracl corp
oral
order
order process
organ
organ repres
organ solvent
organis
organiz
orient
orient model
origin
origin work
outlin
output
output start
outreach
outreach program
outsid
outstand
overcom
overlook
oversea
overview
ow
own
owner

ownership
oxid
ozon
pac
pacif
packag
pag
pag copyright
pai
paid
paint
panel
paper
paper present
paradigm
paragraph
parallel
parallel comput
paramet
pari
pari conven
park
part
parti
partial
particip
particl
particul
particular
particular emphasi
particularli
partner
partnership
partnership offic
pass
passag
passeng
passiv
past
past year
patent
patent applic
patent attornei
patent cooper
patent cooper treati
patent inven
patent issu
patent law
patent offic
patent process
patent protec
patent relat
path
pattern
pattern maker
pattern recogni
paul
payment
pc
pdt
penalti
pend
pennsylvania
peopl
percent
percentag
percentag point
perform
perform comput
perform measur
period
perman
permis
permiss
permit
person
personnel
personnel manag
perspect
persuas
pertain
petroleum
pharmaceut
phas
phenomena
phenomenon
phillip
phillip logo
phillip publish
phillip team
phon
photograph
photon
phras
physic
physical
physical properti
physician
physiolog
physiologi
pilot
pilot project
pirat
plac
placem
plai
plain
plaintiff
plaintiff claim
plan
planar
plant
plant oper
plasma
plastic
platform
plu
plug
point
pointer
polar
polici
policymak
polit
poll
pollut
pollution
polymer
polymer materi
pool
poor
poorli
pop
popul
port
portabl
portfolio
portion
pos
posit
position
possibl
post
potenti
poverti
powder
power
power electron
power gener
practic
practition
pre
preced
precis
precision
preclud
predic
predict
prefer
preferenti
preferenti treatment
preform
prejudic
preliminari

premis
premium
prepar
prerequisit
prescrib
presenc
present
preserv
presid
presid clinton
press
pressur
presum
prevail
preven
prevent
previou
previou section
previous
pric
prim
prim minist
primari
primarili
princip
principl
print
prior
prioriti
privaci
privat
privat network
privat sector
privileg
pro
proactiv
probabl
problem
proce
procedur
proceed
process
process bas
process data
process design
process develop
process engin
process includ
process industri
process manag
process manufactur
process model
process monitor
process plan
process simul
process system
process techniqu
process technologi
process wast
processor
procur
produc
produc activ
produc control
produc environ
produc manag
produc plan
produc process
produc schedul
produc system
product
product announc
product cost
product data
product data model
product design
product lif
product lif cycle
product model
product qualiti
profes
profession
profession servic
professor
profil
profit
program
program languag
program manag
program provid
programm
progress
prohibi
prohibit
project
project cost
project manag
project select
promis
promot
promot electron
promotion
prompt
proof
propag
proper
properli
properti
properti law
properti protec
properti right
propos
proposi
proprietari
proprietari research
propul
prosecu
prospect
protec
protect
protocol
prototyp
prov
proven
provid
provid electron
provid financ
provid insight
provid integr
provid loan
provid user
provision
pto
pub
public
public kei
public polici
public sector
publicli
publish
publish compani
pulp
pur
purchas
purif
puriti
purport
purpos
pursu
pursuant
push
put
qualif
qualifi
qualit
qualiti
qualiti assur
qualiti control
qualiti improv
qualiti manag
quantifi
quantit
quantiti
que

queri
question
quick
quick respons
quickli
quit
quot
quota
rac
racial
racism
radar
radiat
radio
rai
rai smock
railroad
rais
rais capit
ralli
random
rang
rank
rapid
rapid prototyp
rapidli
rapidli chang
rat
ratio
ration
raw
raw materi
reach
reaction
reactiv
reactor
read
readi
readili
real
real estat
real tim
realist
realiti
realiz
realli
reason
receiv
reces
recipi
recogn
recogni
recommend
reconfigur
record
recours
recov
recoveri
recruit
recycl
red
redesign
redistribu
reduc
reduc cost
reduc energi
reengin
refer
referr
refin
reflect
reform
refrain
refus
regard
regardless
regim
region
regist
registr
regul
regular
regularli
regulatori
regulatori framework
reinforc
reinstat
reinvest
reiter
reject
rel
relat
relat act
relat board
relat data
relat technologi
relationship
releas
relev
reli
reliabl
relianc
relief
remain
remand
remark
remedi
remot
remov
remov process
render
renew
rental
reorgan
rep
repair
replac
report
report examin
repositori
repres
represent
reprint
reproduc
reproduc right
republican
request
requir
research
research activ
research area
research center
research group
research institut
research issu
reserv
resid
residu
resign
resin
resist
resolu
resolv
resort
resourc
resourc center
resourc librari
resourc manag
respect
respond
respons
rest
restor
restric
restrict
result
retail
retain
retain job
retali
retir
retriev
retrofit
return

reus
reusabl
revenu
revers
revers discrimin
review
revis
richard
ridg
right
right act
right confer
right continu
right issu
rigor
ris
risk
risk associ
rival
road
robert
robot
robust
rol
roland
rom
room
ross
rossi
rossi alston
rotat
round
rout
routin
royalti
rue
rul
rul bas
run
rural
saf
safer
safeti
sai
said
sal
salari
sampl
san
san francisco
sap
satisfac
satisfi
sav
sba
sba loan
sbir
scal
scan
scatter
schedul
schem
school
scienc
scienc foundat
scientif
scientist
scop
scopi
scopic
scrap
screen
script
sea
seamless
search
sec
second
secondari
secret
secretari
section
section provid
sector
secur
seed
seek
seen
segment
sel
selec
select
self
sell
seller
semant
sematech
semi
semiconductor
semiconductor devic
semiconductor manufactur
seminar
senat
send
senior
senior circuit
senior circuit judg
senior manag
sens
sensit
sensit analysi
sensor
sensor arrai
sensor technologi
sent
separ
septemb
seq
sequenc
seri
serv
server
server architectur
servic
servic compani
servic group
servic mark
servic organ
servic provid
servic sector
session
set
set forth
settl
setup
sever
sex
shall
shap
shar
shear
sheet
shelf
shell
shift
ship
shop
shop floor
shop floor control
short
short term
shorter
show
shown
shut
sign
signal
signal process
signatur
signific
significantli
silicon
similar
similarli

simpl
simpli
simplif
simplifi
simul
simul method
simul model
simul tool
simultan
singl
sinter
sit
situ
situat
sixth
siz
siz manufactur
skill
slight
slow
small
small busi administr
small busi capit
small busi develop
smaller
smart
smart materi
smart structur
sme
smgr
smith
smock
social
societ
societi
soft
softwar
softwar applic
softwar develop
softwar engin
softwar implement
softwar modul
softwar product
softwar system
softwar tool
sol
sold
solder
solid
solid phas
solid stat
solidif
solidif process
solution
solv
solvent
somewhat
sonic
soon
sophist
sort
sought
sound
sourc
south
spac
span
spatial
speak
speci
special
specialist
specialti
specif
specif industri
specif situat
specifi
spectral
spectroscopi
spectrum
specul
speech
speed
spend
spin
spirit
sponsor
spoof
spot
sprai
spread
spreadsheet
sru
stabil
stabl
staf
staff
stag
stai
stand
standard
start
stat
stat court
stat district
stat district court
stat govern
stat law
stat patent
stat personnel
stat personnel manag
stat univers
statem
station
statist
statu
statut
statutori
statutori right
statutori schem
steel
steelwork
step
stereotyp
stev
stev bell
steven
stimul
stochast
stock
stock quot
stop
stor
storag
stori
strain
strateg
strategi
stream
streamlin
street
strength
strengthen
stress
strict
strik
striv
strong
strongli
structur
structur design
structur materi
struggl
student
studi
su
sub
subject
submis
submit
subordin
subsequ
subsidi
subsidiari

substanc
substanti
substitut
substrat
subsystem
success
successfulli
sue
suffer
suffici
suggest
suit
suitabl
sum
summar
summari
summari judg
summari judgment
sun
sun system
supercomput
superconduct
superconductor
superior
superplast
superplast form
supersed
supervi
supervisor
supp
suppli
suppli chain
supplier
support
support develop
support servic
support system
support tool
suprem
sur
surfac
surfac modif
survei
surveill
surviv
sustain
sweet
switch
switzerland
sybas
symposium
synthesi
synthesiz
synthetic
system
system analysi
system analyst
system approach
system bas
system develop
system engin
system group
system includ
system integr
system solution
systemat
tabl
tactic
tailor
tak
taken
talk
tangibl
tap
target
tariff
task
tax
taxat
tcp
teach
teacher
team
teamster
tech
technic
technic assist
technic support
techniqu
technolog
technologi
technologi applic
technologi area
technologi bas
technologi corpor
technologi develop
technologi includ
technologi program
technologi provid
technologi requir
tel
telecom
telecommun
teleoper
telephon
telephon number
televi
tell
temperatur
temporari
tend
tension
ter
term
term capit
termin
termin employe
terrestri
test
testabl
testb
texa
textil
textil worker
thank
them
theoret
theori
therapeut
thermal
thermal analysi
thermal conduct
thermodynam
thermoplast
thermoset
thing
think
thoma
thought
thousand
threat
threaten
throughput
ti
tightli
tim
tim limit
tim scal
tion
tip
tissu
titl
titl vii
tiv
tm
todai
told
toler
took
tool
topic
toronto
total
total number

total qualiti
toxic
tqm
trac
traceabl
track
track record
trad
trad mark
trad relat
trad secret
trademark
trademark applic
trademark offic
tradition
tradition manufactur
traffic
train
transac
transac onlin
transac process
transfer
transfer model
transform
transition
translat
transmis
transport
transport phenomena
travel
treasuri
treat
treati
treatment
treatment process
trend
trial
trip
trp
truck
true
truli
trust
try
tub
turbin
turbul
turn
turnov
tutori
twic
twin
type
typical
ultim
unabl
unauthor
uncertainti
undergradu
underground
underli
underpin
understand
understood
undertak
undertaken
underwai
unfair
unfair labor
unfair labor practic
unfortun
unifi
uniform
union
union disciplin
union effort
union membership
union rul
union secur
union shop
uniqu
unit
unit kingdom
unit oper
unit stat
unit stat court
unit stat district
unit stat district court
unit stat patent
unit steelwork
univers
univers commun
univers copyright
univers copyright conven
unix
unlaw
unpaid
unwant
updat
upgrad
upstream
urban
urg
urgent
url
us
usabl
usag
user
user interfac
usual
util
vacat
vacuum
valid
vallei
valu
valu ad
valu chain
valuabl
valuat
valv
van
vapor
vapor deposi
vari
variabl
variat
varieti
variou
vector
vehicl
veloc
vendor
ventur
ventur capit
verif
verifi
version
vertic
vessel
veteran
vibrat
vic
vic presid
video
view
vii
violat
violet
virtual
virtual enterpr
virtual realiti
visa
vision
vision system
visit
visual
vital
voic
voltag
volum
volum produc
voluntari

voluntarili	weight	withdraw	worldwid
vortic	weld	woman	worth
vot	went	women	worth not
wafer	west	wood	writ
wag	western	word	written
wai	whichev	work	written permis
wait	whit	work capit	wrong
wall	whit mal	work condition	wto
wang	whit men	work environ	www
want	wholesal	work group	year
war	wid	work program	year patent
wareh	wid rang	work relationship	yield
warehous	wid varieti	worker	york
wari	widespread	worker compens	young
washington	wil	worker local	zdnet
wast	william	workflow	zdnet logo
wast dispos	willing	workforc	zdnn
wast stream	win	workplac	zero
water	wind	workshop	ziff
wav	window	workstat	ziff davi
wav process	wipo	world	ziff davi publish
wavelength	wir	world class	ziff davi publish compani
weapon	wireless	world intellectu	
wear	wireless commun	world intellectu properti	
weather	wish	world market	
week	wit	world wid	

ข.3 ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีของชุดเอกสาร K-series

ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีจากชุดการทดลองทั้ง 4 ชุดได้จำนวนฟีเจอร์ 3,731 ฟีเจอร์ ดังนี้

abandon	abnorm	academi award	accid victim
abc	abor	accent	accident
abc ti	abroad	accept	accompani
abdomen	absenc	access	account
abil	abus	access award	accumul
abl	academi	accid	accus

achiev
achiev award
acknowledg
acquir
acquisi
act
act rock
action
activ
actor
actress
acut
ad
adam
adapt
add
addic
addition
address
adelaid
adequ
adher
adjust
administ
administr
admir
admis
admis criteria
admit
adult
adult demograph
adulthood
advanc
advers
advers affect
advertis
advis
advisori
advisori committe
advoc
affair
affect
affili
africa
african
ag
ag associ
ag associ diseas
ag bas
ag bas rat
ag group
ag relat
ag relat deficit
ag year
agenc
agent
agent help
agent help crohn
agent help crohn diseas
aggress
ago
ago averag
agre
agreem
agricultur
ahead
aid
aim
air
air forc
airbag
airbag injuri
airwai
alan
albert
album
album guid
album guid seri
alcohol
alert
alert broaden
alex
aliv
alleg
allegedli
allen
allerg
allerg reaction
allergi
allergi try
allergi try wash
alli
allow
alomar
alongsid
alter
altern
alto
altogeth
alzheim
alzheim brain
alzheim brain damag
alzheim diseas
amaz
amend
america
america onlin
american
american journal
american leagu
american leagu championship
american leagu championship seri
american medic
american medic associ
american societi
ami
amicu
amicu award
ampl
amus
analysi
analysi appear
analyst
analyst said
anderson
andrew
andrew hind
andrew morton
aneurysm
aneurysm risk
angel
angel counti
angel lo
angel lo angel
angel trial
angela
angioplasti
angioplasti arteri
angioplasti arteri clear
anim
ann
ann miller
anni
anniversari
announc
annual
annual meet
answer
anti
antibiot
antibiot avoid
antibiot avoid ulcer
antibodi
anticip
antidepress
antigen
anxieti
anybodi
anymor
apart
apologi
appar
appeal

appear
appl
appl spokeswoman
applic
approach
appropri
approv
approv irradi
approxim
approxim half
archerd
archiv
area
argentin
argu
aris
arkansa
arlington
arm
armi
armi archerd
arrang
arrest
arriv
art
art direc
art scienc
arteri
arteri clear
articl
artist
ashlei
asia
ask
assault
assembl
assess
assign
assist
assist director
assist professor
assn
associ
associ diseas
associ produc
associ professor
atkinson
atkinson comedi
atlant
atlanta
attach
attack
attempt
atten
attend
attitud
attornei
attract
auction
audienc
aug
august
australia
australian
author
author conclud
author not
author sai
author stat
auto
autoimmun
automak
automak seen
automak seen post
automak seen post lukewarm
automak seen post lukewarm profit
avail
averag
averag ag
avoid
avoid drug
avoid ulcer
awai
await
awar
award
award gala
award go
award present
award win
babi
babi born
bacall
back
bacteria
bad
badli
ballad
balloon
baltimor
baltimor oriol
band
bank
bar
barbara
barbra
barrier
bart
bas
bas rat
baseball
baseball gam
baseball playoff
basi
basic
bath
bath seat
bath seat risk
bath seat risk infant
bath seat risk infant liv
bathroom
bathtub
batter
batteri
bean
bear
beat
beatl
beauti
bed
beef
began
begin
begun
behalf
behav
behav badli
behavior
behavior risk
belief
believ
belt
belushi
benefit
benign
benz
bernard
best
best friend
best friend wed
best tim
bestsell
bethesda
better
beverli
beverli hill
bid
big
bigger
biggest
billi
bind
biochemistri

biolog
biologi
biologist
biopsi
birth
birthdai
bishop
bit
biz wir
black
blam
block
blockag
blond
blood
blood pressur
blow
blue
blur
blur deliv
blur deliv mix
blur deliv mix set
bmi
board
bob
bodi
bodi mass
bodyguard
boi
bon
bond
book
boost
born
boss
boston
bought
bovin
bovin spongiform
bow
bowel
bowman
bowman grai
bowman grai school
box
box offic
brad
brain
brain cell
brain damag
brain diseas
brain surgeri
brain surgeri help
brain surgeri help parkinson
brain tissu
branch
brand
break
breakthrough
breast
breast cancer
breast cancer risk
breath
brian
bridg
brief
bring
brit
britain
british
british medic
british medic journal
bro
broad
broadcast
broaden
bronchoscop
bronchoscop spread
bronchoscop spread tuberculosi
brook
brother
brought
brown
browser
buckl
budget
buena
buena vista
bueno
bueno air
bui
build
built
burden
burn
busi
caa
cabl
cabl network
cableac
cableac award
cag
cal
calcul
calendar
calendar item
calif
california
call
calor
calor intak
calor intak prolong
calor intak prolong lif
calor restric
calor restrict
calor restrict diet
calori
cam
camera
camerawork
camp
campaign
campaign fund
campbell
campbell endear
campbell endear perform
canada
canadian
canal
cancer
cancer cell
cancer patient
cancer risk
candid
cann
cap
cap town
capabl
capac
capit
capitol
capitol hill
captur
car
car burden
car cost
car crash
car worker
card
career
caregiv
carei
carlo
carol
carolina
carri
cas
cast
cat
catch
categori
catherin
cathet
cathol

caus
caus breast
caution
cb
cd
cdc
cdc nation
cdc offici
cecer
cedar
cedar sinai
celebr
cell
cell immun
cell receptor
cell slow
cell slow parkinson
cellular
censor
cent
center
centr
central
centuri
ceo
ceremoni
certain
certainli
cessat
chad
chad ogea
chain
chairman
chairman gat
challeng
champion
championship
championship seri
chanc
chang
channel
charact
character
characterist
charg
charl
charli
charlton
charlton heston
charm
chart
chas
chas ami
cheer
chemic
chemotherapi
chest
chicago
chicago hop
chief
chief execut
chief film
chief film critic
chil
child
childhood
children
china
chines
chip
choic
cholesterol
choos
choreographi
chos
chri
chri petrikin
christian
christian soldier
christin
christma
christoph
chronic
chronic ill
chrysler
chrysler corp
church
cigarett
cigarett smok
cinema
circl
circul
cit
citi
civil
clai
claim
class
classic
clean
clear
clearli
cleveland
cleveland indian
clinic
clinic trial
clinton
clinton meat
clinton meat recall
clinton meat recall run
clos
closet
cloud
clout
club
clue
coach
coast
cocain
cod
coffe
coffe pric
cognit
coincid
cold
coli
colicki
colicki babi
colin
collabor
colleagu
colleagu interview
colleagu point
colleagu stat
collec
collect
collect prais
colleg
color
colorado
columbia
columbia univers
column
columnist
com
combat
combin
comedi
comfort
comic
command
commerc
commerci
commerci evad
commis
commit
committe
common
commun
compact
compact disc
compani
compani said
companion

companion dodi
compar
comparison
competi
competit
competitor
complain
complet
complex
compli
complianc
complic
compos
compound
comput
concentr
concep
concept
concern
concert
conclud
condemn
condition
conduct
confer
confidenti
confirm
congress
congression
connecticut
consecut
consensu
consid
consider
consist
consolid
constant
consult
consum
contact
contain
contamin
contemporari
contend
content
contest
continu
contracep
contracept
contract
contrari
contrast
contribu
contribut
control
control group
controversi
conven
convic
convict
convinc
cop
cop land
copi
cor
cord
corner
coronari
coronari arteri
coronari heart
coronari heart diseas
corp
corpor
cosbi
cosmet
cost
costum
costum design
count
counti
countri
coupl
cours
court
cover
coverag
cow
cow link
cox
cpr
cpr lesson
craft
crash
craven
creat
creativ
credibl
credit
credit includ
creutzfeldt
creutzfeldt jakob
creutzfeldt jakob diseas
crimin
criteria
critic
crohn
crohn diseas
cross
crowd
cruis
cult
cultur
cultur dish
cultur showbiz
cultur showbiz event
cup
curb
curb teen
curb teen smok
current
current issu
current studi
cut
cyber
cyber summari
cynthia
cynthia littleton
dad
dai
daili
daili liv
daili varieti
daili varieti chief
daili varieti chief film
daili varieti chief film critic
daili varieti highlight
daili varieti highlight octob
daili varieti senior
daili varieti senior columnist
damag
damag caus
dan
dan cox
danazol
danc
danger
daniel
danni
dario
dark
dat
data
datelin
daughter
davi
david
daytim
dead
deaf
deal
dean
dean goodman
death
death occur

death rat
debat
debut
debut album
dec
decad
decemb
decid
decision
declin
decoi
decontamin
decor
decreas
dedic
dee
deep
defeat
defend
defend world
defend world seri
defiantli
defici
deficit
defin
definit
degen
degre
delai
delai school
delight
deliv
deliv mix
deliv mix set
dell
dell corp
demand
demeanor
dementia
democrat
demograph
demonstr
deni
denni
depart
depend
depres
depress
derek
derek ellei
deriv
dermatologist
describ
deserv
design
despit
destroi
destruc
detail
detect
deterior
determin
detroit
detroit piston
develop
devic
devito
di
diabet
diabet risk
diabet risk identifi
diagnos
diagnosi
diagnost
dialogu
diamond
dian
dian ladd
diana
diana death
diana fatal
diarrhea
dick
die
diet
dietari
differ
difficult
difficulti
digit
dinner
direc
direct
directli
director
director gener
director oliv
director oliv ston
disabl
disappoint
disc
discourag
discov
discoveri
discrep
discus
discuss
diseas
diseas associ
diseas caus
diseas control
diseas occur
diseas relat
diseas transmis
dish
disnei
disord
distant
distribu
distributor
district
disturb
divid
dividend
division
division seri
divorc
dna
dna repair
do
doctor
docum
documentari
dodi
dolbi
dollar
domest
domin
donna
door
dopa
dos
dosag
doubl
doubt
doug
dougla
downward
dr
draft
drama
dramat
draw
dream
dress
drew
drew carei
drift
drink
driv
driver
driver seat
drop
drug

drug abus
drug administr
drug keep
drug keep angioplasti
drug keep angioplasti arteri
drug keep angioplasti arteri clear
drug regimen
drug resist
drug therapi
dry
duchess
durango
durango need
durango need work
durban
dutch
dutch studi
dying
dynamic
ear
earli
earli onset
earlier
earlier relat
earlier relat stori
earn
easi
easier
easili
east
eat
econom
economi
economist
edg
edgar
edinburgh
edit
edition
editor
editori
editori accompani
educ
educ need
edward
effect
effort
egg
eighth
eighth season
elderli
elec
electr
electron
electron commerc
electron edition
elem
elev
elimin
elizabeth
ellei
ellen
emerg
emerg contracep
emi
emmi
emotion
empir
emploi
enabl
enceph
enceph alert
enceph alert broaden
encephalopathi
encount
encourag
end
endear
endear perform
endometri
endometri cancer
energi
enforc
engag
engin
england
england journal
english
enhanc
enjoi
enjoyabl
enrol
ensembl
enter
entertain industri
entertain present
entertain report
entertain tonight
enthusiast
entir
entri
environ
enzym
epic
epidem
epidemiologi
episod
equal
equip
equiti
era
erasmu
erasmu univers
eric
erickson
especi
essenti
establish
estat
estim
estrogen
eunic
eunic simpson
europ
european
evad
evalu
even
event
eventu
evid
evid suggest
evolutionari
evolutionari driver
evolutionari driver seat
evolv
exactli
exam
examin
examin med
examin med school
examin med school admis
examin med school admis criteria
exampl
excell
excess
exchang
exclus
exec
exec produc
execut
execut produc
exerc
exercis
exist
expan
expand
expect
expenditur
expens
experi
experienc

experiment
expert
expert conclud
expert contend
expert sai
expert warn
explain
explan
explor
expos
exposur
expres
express
extend
extent
extra
extract
extrem
eye
fac
fact
factor
factor ti
fai
fail
failur
fair
fall
falter
fam
famili
famili histori
famili member
famili orient
familiar
famou
fan
fantasi
far
fashion
fast
fast track
fast track trad
fat
fatal
fatal human
fatal human disord
father
favor
fax
fda
fda overhaul
fear
fear stir
featur
featur credit
featur film
feb
feder
fee
feel
feet
fell
fellow
felt
femal
fergi
ferguson
fernandez
fertil
fest
festiv
festiv open
fet
fewer
fiction
field
fifth
fifth annual
fight
fight cancer
figur
fil
film
film academi
film critic
film director
film fest
film festiv open
film produc
filmmak
fin
final
financ
financi
find
find suggest
fingerprint
finish
finish stir
finish stir melt
finish stir melt pot
finland
firm
fit
fla
flar
flat
fled
flem
flesh
flexibl
fli
florida
flower
flu
flu epidem
flu shot
flu shot urg
focu
focus
fold
folk
follow
food
football
forc
ford
ford motor
forecast
foreign
foreign box
foreign box offic
foreign total
form
formal
forrest
forth
fortun
forum
forward
foster
found
foundat
founder
fourth
fourth annual
fourth quarter
fox
fractur
franc
francisco
frank
frasier
fred
french
frequent
fresh
fridai
fridai octob
friend
friend wed
fring
front

fruit
fulli
fulvio
fulvio cecer
fun
function
fund
fund rais
funer
funni
futur
futur flu
futur flu epidem
gai
gai men
gain
gala
galleri
gam
gang
gang relat
gang relat mark
gang relat mark shakur
gap
garcia
gari
gari graff
gartner
gartner group
gastrointestin
gat
gather
gav
gen
gen expres
gen link
gender
gener
gener believ
gener hospit
gener motor
gener motor corp
genet
genet engin
genet mutat
genr
genuin
georg
georgia
geriatr
germani
get
gianni
gianni versac
giant
giant said
gift
girl
girlfriend
giv
given
global
global media
go
goal
goe
gold
goldberg
gon
gonna
gonna plai
gonorrhea
gonorrhea infec
gonorrhea infec increas
good
good new
goodman
gor
gordon
gorman
got
got allergi
got allergi try
got allergi try wash
gotham
gotten
govern
grab
grad
gradual
graff
graham
grai
grai school
gram
grand
grand slam
grant
graphic
great
greater
greater risk
greatest
greatli
greek
green
greg
gregori
grew
gross
ground
group
group continu
grow
grown
growth
guarante
guarante impli
guarante impli regard
guarante impli regard inclusion
guard
guest
gui
guid
guid seri
guidelin
guild
guilti
guitarist
gun
gynecolog
gynecologi
habit
hair
hair loss
hair loss possibl
half
half hour
hall
halt
hand
hand heart
hand heart healthi
hander
handl
hang
happen
happi
hard
harder
harm
harpercollin
harri
hart
harvard
harvard medic
harvei
hat
haunt
hav
haye
hazard
head
heal

health benefit
health car
health car burden
health car cost
health car worker
healthi
healthi stat
healthi stat list
healthier
hear
hear aid
heart
heart attack
heart diseas
heart failur
heart healthi
heart muscl
heart muscl cell
heart patient
heart transplant
heart transplant rejec
heat
heavi
heavier
heavili
height
held
helen
hell
helm
helmer
help
help crohn
help crohn diseas
help hand
help hand heart
help hand heart healthi
help parkinson
help prevent
help prevent airbag
help prevent airbag injuri
henri
henri paul
hepat
hercul
heroin
herp
herpesviru
herpesviru link
herv
heston
hid
high
high level
high rat
high risk
high school
high speed
higher
highest
highest rat
highli
highlight
highlight octob
hill
hind
hir
histori
hit
hitchcock
hitter
hiv
hiv infec
hiv posit
hiv posit worker
hmo
hmo perform
hmo perform vari
hmo perform vari wid
hold
holidai
hollywood
hollywood film
hollywood film festiv
holocaust
hom
homer
honor
honore
hontz
hop
hopkin
hopkin univers
hormon
horror
hors
hospit
host
hot
hotel
hotlin
hour
hous
hous panel
hous pass
hous pass fda
hous pass fda overhaul
household
household averag
household rat
hug
hugh
human
human disord
human servic
humor
hundr
hunt
hurrican
hurrican fear
hurrican fear stir
hurt
husband
hyperten
hypothes
ibm
ic
ic storm
icon
ida
idea
ideal
ident
identifi
ignor
iii
ill
illeg
illinoi
imag
imagin
immedi
immers
immun
immun cell
immun practic
impact
impair
implant
impli
impli regard
impli regard inclusion
implic
imporov
import
import produc
impress
improv
in
inaugur
incid
includ
inclusion
incom

increas
increas risk
increasingli
incub
incub period
independ
independ featur
indi
indian
indic
individu
induc
industr
industri
ineffect
infant
infant liv
infant stor
infec
infec increas
infect
infect patient
infecti
infecti agent
inflamm
inflammatori
inflat
influenc
influenza
inform
infrastructur
infusion
inherit
inhibitor
initi
injec
inject
injur
injuri
ink
inocul
inser
insert
insid
insid daili
insid daili varieti
insid daili varieti highlight
insid daili varieti highlight octob
insid edition
insight
insomnia
insomnia rais
insomnia rais health
insomnia rais health car
insomnia rais health car burden
inspir
instanc
institut
instrum
insur
intak
intak prolong
intak prolong lif
intel
intens
interact
interest
interim
interim chief
interim chief execut
intern
intern documentari
intern film
intern film festiv
intern film festiv open
intern medicin
intern sal
internet
internet access
internet commerc
internet sit
interven
interview
intl
intric
intrigu
introduc
invest
investig
investig sai
investig target
investor
involv
involv patient
ion
ion rat
iowa
iron
irradi
isol
issu
itali
italian
item
jack
jack nicholson
jackson
jacqu
jai
jail
jakob
jakob diseas
jam
jama
jan
januari
japanes
jazz
jean
jeff
jenni
jenni hontz
jennif
jeopardi
jerri
jersei
jesu
jewish
jim
jimmi
joan
job
job said
joe
joe pesci
joel
johannesburg
john
john hopkin
john hopkin univers
john wort
johnni
johnson
joi
join
jon
joseph
journal
journal natur
journal pediatr
journal scienc
journalist
judg
judg rul
juli
julio
julio martinez
jump
jun
jungl
jurass
jurass park
juri

juri order
juri order chrysler
just
just wrap
justic
justifi
kathi
kathi lee
kathleen
keep
keep angioplasti
keep angioplasti arteri
keep angioplasti arteri clear
kei
kei adult
ken
kennedi
kept
kevin
kevin smith
kick
kid
kil
kill
killer
kind
king
kiss
kiss plant
kladi
know
knowledg
known
korea
korea second
korea second seoul
label
label requir
labor
laboratori
lack
ladd
ladi
lak
lak buena
lak buena vista
lamb
land
lang
languag
lap
larg
larger
largest
larri
last
lat
latest
latest movi
latex
latex label
latex label requir
latin
latino
laugh
launch
laura
lauren
law
lawmak
lawrenc
lawsuit
lawyer
lax
lead
lead author
lead caus
lead studi
lead studi author
leader
leagu
leagu championship
leagu championship seri
learn
learn word
leav
led
lee
left
leg
legal
legisl
lend
length
lens
leonard
leonard kladi
lesli
lesli lang
lesson
lethal
letter
level
levi
lewi
li
liar
liar liar
lif
lif threaten
lifestyl
lifetim
lifetim achiev
lifetim achiev award
light
likelihood
lin
linda
lineup
link
lisa
list
listen
literatur
literatur priz
littl
littleton
liv
lloyd
lloyd webber
lo
lo angel counti
lo angel lo
lo angel lo angel
lo angel trial
local
locat
lolita
lon
london
london british
long
long liv
long term
long tim
longer
longer period
look
loom
lopez
los
loss
loss possibl
lost
lost world
lot
loui
lov
lover
low
low calor
low calor intak
low calor intak prolong
low calor intak prolong lif

lower
lower breast
lower breast cancer
lower breast cancer risk
lukewarm
lukewarm profit
luncheon
lung
luxuri
lynn
lyric
machin
mad
mad cow
mad cow link
madison
magazin
magic
mail
main
mainli
maintain
majendi
major
mak
maker
mal
malibu
mall
malnutri
man
manag
manhattan
manner
manslaught
manufactur
march
margin
mari
maria
marin
mark
mark shakur
marker
market
marri
martha
martin
martinez
marv
marv albert
marvel
maryland
mass
massachusett
massachusett gener
massachusett gener hospit
match
materi
matern
matter
matthew
maximum
mcbeal
mccarthi
mccartnei
mean
measur
meat
meat product
meat recall
meat recall run
mechan
mechan underli
med
med school
med school admis
med school admis criteria
media
media access
media access award
medic
medic associ
medic center
medic cost
medic devic
medic journal
medic record
medic school
medic treatment
medicin
meet
melani
melod
melt
melt pot
member
membership
memor
memori
men
men behav
men behav badli
menac
menopaus
menstrual
menstrual period
mental
mental health
mention
mer
merced
merced benz
merchand
messag
met
metabol
method
methodist
mexican
mexico
miami
mic
mic genet
michael
michael flem
michael jackson
michell
mid
middl
middl ag
midst
migrat
mik
milan
mild
militari
miller
mind
minim
minnesota
minnesota top
minnesota top healthi
minnesota top healthi stat
minnesota top healthi stat list
minor
minut
mirag
miramax
miramax film
misdemeanor
misdemeanor assault
miss
miss diana
miss link
mississippi
mistakenli
mitchell
mix
mix set
model
moder
modif
modifi

mold
molecul
molecular
molli
mom
moment
mon
mon oct
mon sep
mondai
mondai octob
mondai septemb
monei
monica
monica roman
monitor
month
monti
mood
moon
moor
moral
morbid
morn
morri
mortal
mortal weekli
mortal weekli report
morton
mother
motion
motion pictur
motiv
motor
motor corp
motor function
motown
mountain
mountain view
mov
movem
movi
movi theater
multipl
multipl sclerosi
mundell
murder
murdoch
murdoch said
muscl
muscl cell
muscl tissu
museum
music
musichound
musician
mutant
mutant gen
mutant gen link
mutat
mycobacterium
mysteri
mysteriou
mysteriou second
nab
nam
nanci
narr
narrat
nasdaq
nation
nation academi
nation averag
nation health
nation institut
nation syndicat
nation syndicat journalist
natur
natur medicin
nazi
nba
nbc
near
nearli
nearli ident
nearli undetect
necessari
necessarili
necessarili mean
neck
need
need work
neg
negoti
neil
nelson
neonat
nerv
nerv cell
netherland
network
neurolog
neurologi
neuron
new corp
new coverag
newborn
newcom
newli
newsmagazin
newspap
nic
nicholson
nick
nicol
nielsen
night
nih
nih studi
ninth
noah
nobel
nobel literatur
nobel literatur priz
nobel priz
nomin
non
nonsmok
normal
north
north carolina
northern
not
notabl
notabl quot
notic
nov
novel
novemb
nsaid
nsaid user
number
nurs
nutrition
nypd
nypd blue
obes
obes risk
oblig
obscur
observ
obstetr
obtain
occasion
occup
occur
oct
octob
octob issu
octob wed
octob wed oct
odd
odonnell
offer

offic
offic said
offici
offici sai
offspr
ogea
old
old ag
older
older american
older peopl
oldest
oldest old
oliv
oliv ston
on
onair
oneill
ongo
onlin
onset
ontario
onward
onward christian
onward christian soldier
open
open sept
oper
oper system
opera
opinion
opinion express
opportun
oppos
opposit
oprah
oprah winfrei
option
orchestra
order
order chrysler
organ
orient
origin
oriol
oscar
out
outbreak
outcom
output
output start
outrag
outsid
overhaul
oversea
overwhelm
ow
own
owner
oxid
pac
pacif
pack
pact
pai
paid
pain
paint
pair
palo
palo alto
panel
paparazzi
paparazzi photograph
paper
paper said
parallel
parent
pari
park
parker
parkinson
parkinson diseas
parkinson patient
part
parti
partial
particiap
particip
particl
particular
particularli
partner
partnership
pasadena
pass
pass fda
pass fda overhaul
passion
past
past year
pathogen
patient
patrick
pattern
paul
paul majendi
paul oneill
peacemak
pediatr
pediatrician
pen
pennsylvania
pensacola
pentium
peopl
percent
percent profit
percentag
perdita
perdita durango
perdita durango need
perdita durango need work
perfect
perform
perform vari
perform vari wid
period
permiss
permiss parent
persist
person
person comput
pertussi
pesci
peter
petrikin
philip
phillip
phon
photo
photograph
physical
physician
physiologi
phytoestrogen
phytoestrogen lower
phytoestrogen lower breast
phytoestrogen lower breast cancer
phytoestrogen lower breast cancer risk
piano
pic
pick
pictur
piec
pig
pill
pioneer
piston
pitch
pix
plac

placebo
placem
plai
plan
plant
plastic
playboi
player
playoff
playoff gam
plaza
plc
plead
pleasur
plenti
plot
plu
plung
poet
poetri
poetri dai
point
point repres
polanski
polic
polici
polish
politician
poor
poorli
pop
pop star
pop star michael
pop star michael jackson
popul
popular
portion
portrait
portrayal
pos
posit
posit patient
posit worker
position
possess
possibl
possibli
post
post lukewarm
post lukewarm profit
postseason
pot
potent
potenti
potenti fatal
poultri
pound
poverti
power
practic
prais
pre
predict
predict heart
predict heart transplant
predict heart transplant rejec
predictor
prefer
pregnanc
pregnanc diabet
pregnanc diabet risk
pregnanc diabet risk identifi
pregnanc factor
pregnanc factor ti
pregnanc weight
pregnanc weight gain
pregnant
prematur
premier
premier week
prepar
prescrip
presenc
present
preserv
presid
presid clinton
press
press confer
press releas
pressur
preval
preven
prevent
prevent airbag
prevent airbag injuri
prevent car
prevent malnutri
preview
previou
previou stori
previous
pric
primari
primari goal
primetim
primetim liv
princ
princess
princess diana
princip
principl
print
prion
prion diseas
prion research
prion research win
prion research win nobel
prion research win nobel priz
prior
prison
privat
privat lif
priz
prob
probabl
problem
procedur
proceed
process
prod
produc
produc compani
produc david
produc design
produc jam
product
profession
professor
profit
profound
program
program content
progres
progress
project
prolong
prolong lif
promis
promot
promotion
prompt
properti
propos
prosecutor
prosecutor offic
prospect
protec
protect
protein
protein ti
prov
proven
provid

provid import
provinc
psychiatri
psycholog
psychologi
public
public health
publish
purchas
purpos
pursu
push
put
qualiti
quarter
quarter earn
quartet
queen
quest
questionnair
quickli
quiet
quirki
quit
quot
rac
radic
radio
rai
rai richmond
rais
rais health
rais health car
rais health car burden
randomli
randomli assign
rang
ranger
rank
rap
rapidli
rapidli expand
rar
rat
rat point
rat point repres
ration
reach
react
reaction
read
readi
real
realli
reason
rebound
recal
recall
recall run
receiv
receptor
recipi
recogn
recommend
record
recov
recoveri
red
red meat
reduc
reflect
refus
regain
regard
regard inclusion
regardless
regi
regi kathi
regi kathi lee
regimen
regina
region
regul
regular
regular season
regularli
rejec
reject
rel
relaps
relat
relat death
relat deficit
relat diseas
relat ill
relat mark
relat mark shakur
relat stori
relationship
relax
releas
reliabl
relief
religi
remain
remark
rememb
remind
remov
render
rent
rent collect
rent collect prais
rep
repair
repeat
replac
replic
report
report sai
reportedli
repres
republican
request
requir
requir medic
rerun
rescu
research
research admit
research believ
research conclud
research contend
research institut
research led
research not
research point
research report
research sai
research sought
research stat
research win
research win nobel
research win nobel priz
research writ
resembl
resid
resist
reson
resourc
respect
respond
respons
rest
restric
restric slow
restrict
restrict diet
result
retail
retain
retard
retir
return

reveal
revenu
revenu com
revers
review
review articl
review oct
review sept
reynold
rhythm
rich
richard
richmond
rid
right
right hander
rigor
ris
risk
risk factor
risk identifi
risk infant
risk infant liv
ritz
rival
road
rob
robert
rock
rock band
rock star
rodent
roger
rol
rol ston
roll
roman
roman polanski
romanc
romant
romant comedi
ron
room
ros
rosenberg
rosi
rosi odonnell
roster
rotterdam
round
round pick
routin
rowan
rowan atkinson
rowan atkinson comedi
royal
rug
rul
run
rupert
rupert murdoch
rusedski
rusedski serv
ryan
sad
saf
safer
safeti
safeti expert
safeti rul
sai
sai research
said
said mondai
said studi
said thursdai
said tuesdai
said wednesdai
saint
sal
sal law
salari
salem
salli
sampl
san
san francisco
santa
santa barbara
santa monica
sarah
sarah ferguson
satellit
satir
satisfactori
satisfi
saturdai
saturdai night
sav
saw
scandal
scen
scenario
schedul
school
school admis
school admis criteria
scienc
scientif
scientif evid
scientist
sclerosi
scop
scor
scotland
scott
scott erickson
scream
screen
screen prevent
screen prevent malnutri
screenplai
screenwrit
scrib
script
scripter
se
sean
season
season premier
seat
seat risk
seat risk infant
seat risk infant liv
seatbelt
seatbelt help
seatbelt help prevent
seatbelt help prevent airbag
seatbelt help prevent airbag injuri
seattl
seattl marin
second
second album
second annual
second dos
second quarter
second round
second seoul
second tim
second week
secret
secretari
secretari donna
secur
seduc
seed
seek
seen
seen post
seen post lukewarm
seen post lukewarm profit
seg
seinfeld

select
self
self report
sell
seller
seminar
sen
senat
send
senior
senior columnist
senior investig
sens
sensibl
sensit
sent
sentenc
seoul
sep
separ
sept
septemb
sequel
seri
seri open
serious
serv
server
servic
servic offer
servic provid
session
set
set decor
set design
seventh
sever
sex
sexual
shakur
shalala
sham
shampoo
shap
shar
sharon
sharp
shed
shelter
shin
shirt
shoot
shop
short
shot
shot urg
shoulder
show
showbiz
showbiz event
showcas
shown
sick
sid
sight
sign
signal
signific
significantli
silenc
silicon
silver
similar
simon
simpl
simpli
simpson
sinai
sing
singer
singl
sir
sir paul
sister
sit
sitcom
situat
sixteen
sixth
siz
skill
skin
skin reaction
sky
slain
slam
slat
sleep
sleepless
sleepless children
slightli
slot
slow
slow ag
slow ag relat
slow ag relat deficit
slow parkinson
slowli
small
smaller
smart
smith
smok
smok cessat
smoker
smoker quit
snappi
soap
soar
social
societi
soderbergh
soft
softwar
softwar compani
softwar develop
softwar giant
sol
sold
soldier
solid
solo
solution
somewhat
son
song
soni
soon
sort
sought
soul
soul food
soul man
sound
soundtrack
soup
sourc
sourc clos
sourc said
south
south africa
south african
south carolina
south korea
south korea second
south korea second seoul
southern
spain
span
spanish
spar
spark
spawn
speak

speci
special
specif
spectrum
specul
speech
speed
spend
spent
spielberg
spin
spinal
spinal cord
spirit
spok
spokesman
spokesman said
spokesperson
spokeswoman
spongiform
spongiform encephalopathi
sponsor
sporad
sport psychologi
sportscast
sportscast marv
sportscast marv albert
spot
spread
spread tuberculosi
spring
squar
stabl
staff
stag
stai
stak
stallon
stand
stand ston
standard
star
star michael
star michael jackson
star trek
starrer
start
starter
stat
stat list
statem
statem releas
statem said
station
statist
statu
steadi
step
stephan
stephen
steril
steroid
stev
stev gorman
stev jam
stev job
steven
steven spielberg
stewart
stimul
stir
stir melt
stir melt pot
stock
stockholm
stomach
ston
stop
stop smok
stor
stori bas
storm
straight
strain
strand
strap
strateg
strategi
streak
street
street journal
strength
stress
stretch
strik
strip
stripper
strok
strong
strongli
structur
struggl
student
studi
studi appear
studi author
studi author conclud
studi confirm
studi involv
studi involv patient
studi lead
studi lead author
studi particip
studi provid
studi show
studi suggest
studio
stuff
stunt
style
subject
submis
submit
subscrip
subsequ
substanc
substanc known
succe
success
sudden
suffer
sugar
suggest
suicid
suit
summari
summer
sun
sun said
sun system
sundai
sundai night
sundanc
sunset
super
superior
superman
supernatur
supervis
supervis editor
supervisor
supplem
suppli
support
suprem
suprem court
sur
surfac
surgeon
surgeri
surgeri help
surgeri help parkinson
surgic
surgic remov

surpris
surprisingli
surround
survei
surveill
surviv
surviv colicki
surviv colicki babi
survivor
suscept
suspect
sustain
sweden
swing
switch
sylvester
sylvester stallon
sympathet
symphoni
symposium
symposium host
symptom
syndicat
syndicat journalist
system
tak
taken
tal
talent
talk
taller
tamkin
tap
target
target shelter
task
tax
taylor
team
tear
technic
technicolor
techniqu
technologi
technologi symposium
technologi symposium host
teen
teen smok
telecommun
telephon
televi
televi art
televi art scienc
televi market
televis
televis ceremoni
tell
tend
tennesse
teresa
term
termin
terrif
territori
test
test reveal
testicular
testicular cell
testicular cell slow
testicular cell slow parkinson
testimoni
tetanu
tetanu pertussi
texa
texa ranger
theater
theatr
theatric
theoret
theori
therapi
theresa
theresa tamkin
thermostat
thesp
thing
think
third
thoma
thought
threat
threaten
thriller
thriv
thursdai
thursdai night
thursdai octob
thursdai septemb
ti
ticket
ticket sal
tightli
till
tim
tim greater
tim greater risk
timeslot
tini
tip
tissu
titan
titl
titl charact
tml
tobacco
todai
todd
todd mccarthi
told
told daili
told daili varieti
told report
told reuter
tom
ton
toni
toni scott
tonight
took
took plac
tool
toothless
top
top healthi
top healthi stat
top healthi stat list
toronto
toronto film
toronto film festiv
total
touch
tough
tour
tournam
town
track
track trad
tract
trad
tradition
tragedi
tragic
train
transac
transform
translat
transmis
transmit
transplant
transplant patient
transplant rejec
treat
treat addic
treat hyperten

treat patient
treatment
trek
trend
tri
trial
tribut
trigger
trio
trio finish
trio finish stir
trio finish stir melt
trio finish stir melt pot
troi
troi woefulli
troi woefulli unfocus
troubl
troup
true
true stori
truli
truth
try
try wash
trying
tub
tuberculosi
tue
tue oct
tue sep
tuesdai
tuesdai octob
tuesdai septemb
tumbl
tumor
tun
tunnel
turn
twic
twist
type
typical
ulcer
ultim
unabl
unbridl
unclear
uncontroll
undergo
underli
understand
understood
underw
undetect
unexpect
unfocus
unfortun
union
uniqu
unit
unit stat
univers
univers colleg
univers medic
univers pictur
univers school
unix
unknown
unnecessari
unpredict
unprotect
unusu
up
upcom
upcom art
upcom film
updat
upn
upset
urban
urg
us
user
usual
uterin
vaccin
valu
valuabl
van
vancouv
vari
vari wid
varieti
varieti chief
varieti chief film
varieti chief film critic
varieti cyber
varieti cyber summari
varieti entertain
varieti entertain report
varieti highlight
varieti highlight octob
varieti senior
varieti senior columnist
variou
vast
veget
vehicl
venou
ventur
venu
versac
version
versu
veteran
vic
victim
victor
victori
video
video confer
view
viewer
villag
violat
violenc
violent
violent commerci
violent commerci evad
viral
virgin
virginia
virologi
virtual
viru
virus
vision
visit
vista
visual
vital
vitamin
vocal
vocalist
voic
volum
vot
wai
wait
wak
wal
walk
walker
wall
wall street
wall street journal
walter
want
war
war iii
ward
warn
warner
warner bro

wash	went	wir	worsen
washington	west	wis	wort
watch	western	wisconsin	worth
water	westport	wit	wrap
we	wheel	woefulli	writer
weak	whit	woefulli unfocus	written
wear	whit hous	woman	wrong
wear hear	whoopi	women	wrot
wear hear aid	whoopi goldberg	women report	xofficeoversea
weather	wid	wonder	yanke
webber	widescreen	wonder world	yard
wed	wif	wor	year
wed oct	wil	word	year ago
wednesdai	wild	work	year ago averag
wednesdai octob	william	worker	year contract
wednesdai septemb	wilson	workplac	year period
week	win	workstat	year studi
week issu	win nobel	world largest	york
weekend	win nobel priz	world premier	york citi
weekend box	wind	world seri	york film
weekend box offic	window	world trad	york stock
weekli	winfrei	world war	york yanke
weekli report	wing	world war iii	young
weigh	winner	world wid	young peopl
weight	winston	worldwid	younger
weight gain	winston salem	worn	zero
welcom	winter	worri	

ข.4 ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีของชุดเอกสาร Reuters-Top10

ฟีเจอร์ที่ได้จากการเลือกโดยใช้ต้นไม้วลีจากชุดการทดลองทั้ง 4 ชุดได้จำนวนฟีเจอร์ 3,421 ฟีเจอร์ ดังนี้

abandon	abil	access	accumul export
abbott	abl	accord	achiev
abbott laboratori	abruptli	accordingli	acknowledg respons
abegglen said	absolut sens	account	acquir
abid	accept	accru	acquir total explor australia

acquisi
acreag
acreag bas subject
acreag reduc
act
act quickli
action
activ
activ central
actual damag
ad
ad tom campbell
add
addition
addition mln dlr
addition quota
address excess consump
adelaid
adher
adjust
adjust export pric
administr
administr offici
admit
adop
advanc
advers affect
advers effect
advic
advis
adviz
advoc min
aerospac
affair
affair expert
affect
affili
africa
africa nil nil nil nil
aftermath
agenc
aggres
aggress
aggress seek acquisi
ago
agoni
agre
agre trad deal
agreem
agreem follow month
agreem guarante access
agreem sign
agreem total
agricultur
agricultur depart
agricultur depart gav
agricultur offici
agricultur produc
agricultur report said
agricultur secretari
agricultur secretari carlo
dominguez said
ahead
aid
aim
air
alan wheatlei
alaska
alaska colvill delta
alaskan oil trad continu
albeit
alberta
ald
alert
algeria
ali
alleg
alli
alli accord
alli depend
alli leader
alli signal
alloc
alloc rat
allow
allow bank
allwast
altern
aluminum
alver carlson
alw
amend
america
america attack
america said
america simpli react
american
american citizen liv
american dollar
american farmer
american forc
american helicopt cam
american liv abroad
american peopl
anaheim
analyst
analyst ad
analyst believ mondai wall
street crash
analyst darrell
analyst expect
analyst expect bonn
analyst rang
analyst sai
analyst said
analyst said west germani
stubborn march
anbaa
andov
anger
angri
anhui
anniversari
announc
annual
annual level meet
annual meet
annual sal
annual sharehold
answer
anticip
appeal
appear
appli
applic
appoint
appraoch america
appreci
approach
appropri
approv
apr argentin grain board
preliminari
april
april cheap oil feedstock
april european currenc market
react quietli
april japan littl known ministri
april kuwait
april pest
april qtr
april remark
april stood
april taiwan import
april todai turmoil
april union texa petroleum said
april year
arabia
arbitr
area
argentin
argentin grain board adjust
argentina
argentina design

argentina said
argu
ark
arm embargo
armen estim
arrang
ash
ashland oil
ashraf fouad
asia
asid
ask
ask congression leader
ask gatt
ask presid
assess
asset
assist
assist act
associ
associ profit
assum
assump
atlant
atlanta
atpp block
attack
attack involv
attack kuwaiti connect vessel
attack respond
attempt
atten
attend
attitud
attract
attribut
aubrei
audienc
aug
august
australasia
australia
australian
australian labor union
australian port
australian westpac bank
australian wheat export
author
author common shar
authoris
automat pilot
automobil manufactur
avail
avail opportun
aveng
averag
averag daili reserv
avg shr
avoid
awai foreign investor
awar
backdrop
bader told
badli
bahrain
baker
baker appear
baker cut short
baker said
baker statu
baker weekend blast
baker weekend remark
bal
balanc
ball rol
ban
ban impos
bangkok
bank
bank borrow
bank economist
bank economist said
bank interven
bank meet
bank presid dieter hiss
bank said
banker
banker said
bankruptci
bar
bargain
barlei
barrel
bas
basi
basic commod petrochem
battl
bavadra
bbl
beacon
beacon usual lit
bear
bearish
beef
beer
began
begin
begun escort
behalf
believ
believ econom
bell
benchmark crud west texa
intermedi
benefit
bentsen said
best
best servic
best year
better
beverli hill
bicol
bid
big
biggest
biggest reduc
bill
block
blockad
blockad iran
blown
board
board approv
board declar
boat
bolder step
bolster
bomb threat
bomb threat forc evacu
bond
bond market
bond suffer
bonn
bonn refus
bonu issu
boost
boost output
boost rat
bor
borg warner
borg warner corp
borrow
boston
boston corp
bottom
bought
bpd
bracket
bran pollard wheat
brav
brazil
brazil suspend payment
brazilian loan
breach

bread wheat
bread wheat nil nil nil
break
brief
bring
bring suppli
britain
britain cabl
british
british chancellor
british petroleum
british print
broadli
broker
brother
brought
brunt
brussel
buck
budget
budget deficit
budget deficit rel
budget mln dlr
budget sav propos draft
bueno air
bui
build
build block
builder
built
bullish
bullish outlook
bullish view
bundesbank
bundesbank allot
bundesbank board member clau
koehler
bundesbank board member clau
koehler yesterdai
bundesbank latest liquid allot
bundesbank offici
bundesbank presid karl otto
pochl
burden
bureaucrat
bureaucrat counterpart
burlington
burlington industri
burnham lambert
bushel
busi
businessmen
buyer
cain said
cal
calcul
calgari
calif
california
call
calm
cam
cambridg
campaign
campeau
canada
canada februari trad surplu
canadian claim
canadian dlr
cancell
capabl
capit
capit spend
capit stock
car
career naval offic
cargo
carri
carri awai
carri kuwaiti oil
carryforward
cas
cash
cash crop
casualti
categori paper todai
caught lai min
caus
caution
cdi
ceil
celebr
cellular
cent
center
centr
central
central bank
central bank interven
central banker
central banker rais
central oklahoma
cereal
cereal substitut
certain
certain fish
certainli
chairman
chairman gordon cain
chairman patrick leahi
challeng
chancellor
chang
character
charg
charg off
charg off total
charl
charter
charter soviet tanker
charter tanker
chas
chas awai
chemic
chemic acquisi
chemic busi
chemic engin
chemic execut shar
chemic export
chemic industri
chemic industri biggest custom
chemic maker
chemic manufactur
chicago
chicago manufactur
chief
chief economist
chief execut offic
china
china daili said
china nil nil nil nil
china wheat crop threaten
chines
chip
choic
chok
chokepoint
christma tree
christoph hanson
cincinnati
circumst
cit
citi
citicorp
citru fruit
civil defens offici said
civil disobedi
claim
clash
class
clear
clearli
clercq
clercq told

cleveland
climb
clos
clos link
closur
coars grain
coars grain produc
coast
coastal
cohen
collaps
colo
colombia
com
combin
comment
commerc
commerci
commis
commis authoris
commis cereal manag
commis sourc said
commission willi
commit
committe
committe reject
committe spokesman told reuter
commod
commod chemic
commod chemic busi
commod provid
commod report
common
common secur interest shar
common shar
common stock
common stock outstand
commun
commun corp
communiqu
compani
compani common stock
compani continu turn
compani produc
compani report
compani result
compani said
compani spokesman
compani tonnag
compani total energi produc
compani went public
compar
compar figur
compar quarter
compar week
comparison
compens
compet
competi
competit
competitor
complaint
complet
complet acquisi
complet chok
complet merger
completion
compli
complianc
comprom
comput
concentr
concern
conces
concret
condens
condition
condition continu
confer
confid
confirm
conflict
conflict unduli
confront
congress
congression concern
congressman
conn
connec
consensu
consequ
conserv
consid
consid neglig
consid propos
consider
consist
consolid
consortia trying
consortium
construc
consult
consult firm asia advisori
consum
consumm
consumm politician
contain
content
context
continu
continu export
continu fight
continu talk
contract
contrast
contribut
control
control congress
control declin
conven
conver
convert
convert prefer shar
convert subordin debentur
convinc
cooper
cooper basin
coordin interven
cor busi
corn
corn plantat
corn trader
corp
corp said
corpor
corpor purpos
correct
correspond
cost
cotton
cottonse oil
counti approv
counti loan rat
counti loan rat differenti
countri
countri oust prim minist timoci bavadra renew pressur
countri sugar
coupl
cours
court
cover
cower
crack
crazi eddi
creat
credibl
credibl strategi
credit
credit suiss
crisi
critic
critic charg mpt

crop
crop export registr
crud
crud oil
crud oil post
crush blow
crush wheat
csi
csi said washington
ct
ct ct
ct loss
ct loss ct
ct prior
ct profit
ct profit ct
cubic feet
cumul
cumul export
cumul figur export registr
cur
curb
curb monei suppli growth
currenc
currenc analyst
currenc analyst refus
currenc declin
currenc market
currenc stabil
currenc unit
current
current discuss
current fiscal year
current level
current manag
current oper
current rat
current season
current trad rang
current year
custom
custom design
custom repurchas agreem
cut
cut abroad
cut rat
cycle
dai
dai account
dai averag
dai banker secur
dai drop
dai sama deposit
dai sama deposit ris
dairi
dalla
damag
damag hom
dampen
danger
dash
dat
dat interbank
data
david jon
dead
deadlin
deal
deal agre
dealer
dealer expect volum
dealer said
dean witter reynold
death
debat
debentur
debentur holder
debilit trad war
debt
debt burden
debt burden dom
debtor countri export
decad
decemb
decid
decid factor
decision
declar
declar destin
declin
declin comment
declin earn
defenc
defens
defens depart said just pct
defens secretari caspar weinberg told
deferr
defici payment
deficit
deficit cut
definit
definit agreem
delai
deleg
deliv
deliveri
demand
democraci
demonstr
deni
denot minu figur
denver
depart
depart not
depart offici said
depart said
depart said produc
depend
deposit
deposit account issu
depreci
deputi treasuri secretari richard
darman said
deregul
describ
design
desir
despit
destabil drop
destin
destroi
destroi kei bas
destroyer
destroyer kidd
destroyer pump
destruc
detail
deter
deter iranian attack
deterior
determin
deterr polici
develop
di
diagnost
differ
differenti
difficult
dilemma
dilig
dilut
dilut shr
diminish
dip
diplomat
diplomat mission
diplomat mission oversea
diplomat said
diplomat sourc
diplomat sourc said
direc
direct injec
direct result

director
director gener
disappoint
disast
disclos
disclosur
discontinu
discontinu oper
discount
discount corp
discount corp mcgroarti describ
discount rat
discount rat januari
discov
discoveri
discreet rat adjust
discus
discuss
diseas
dismantl
disput
distribu
distributor
div
diversifi
divest
divid
dividend
division
dlr
dlr bid
dlr cash
dlr compar
dlr ct
dlr gain
dlr loss
dlr loss ct
dlr offer
dlr profit
dlr tax credit
dmp
do
dollar
dollar collaps
dollar crisi
dollar declin
dollar drop
dollar fal
dollar fall
dollar fell
dollar latest
dollar open
dollar open lower
dollar push
dollar slid
dollar trend
dom
dom petroleum
domest
domest industri earn
domest market
domest polici
dominion textil
dosher said
doubt
dow
dow chemic
dow jon industri averag
dramat cheaper
drexel
driv
driver seat
drop
drought
drought account
drug
dry
dry spell
dry spell continu
dump
dump semiconductor
dupont
dupont conoco inc chemic busi
durum wheat
earli
earlier
earlier clash
earn
earn includ
eas
easili
east
east europ
east europ nil
ec cereal manag
econom
econom crisi
econom expan
econom growth
econom malais similar
econom polici
econom polici decision
econom spotlight telecom
econom stimul
econom summit
economi
economi continu
economist
economist believ
economist said
ecu
ecuador
edelman
edmonton
edt
effect
effect jun
effort
elec
electron
elig
elig winter wheat
elimin
embark
embroil
emerg
emerg loan
employe
employe stock ownership plan
employm
enabl
enact
encourag
end
end march
endang
energi
engag
engin
england
enhanc
enjoi
enorm
enquiri
enrol acreag
enroll
entail world war
enter
entertain
entertain market
entir
entitl
entrench
entri
equal
equip
equiti
escal
escap
escort
escort kuwaiti
especi
establish
estim

ethylen
eugen carroll
eugen island block
euromarket
europ
europ nil
europ nil nil nil nil
european
european commis
european commun
european countri
european currenc unit
european market react quietli
european monetari
evacu
evalu
evalu work
eventu
evid
exacerb
exce
excess
exchang
exchang commis
exchang loss
exchang market
exchang rat
exchang rat stabil
exchequ
exchequ nigel lawson
exclud
exclud brazil
exclud gain
execu
execut
execut offic
exist
exlud
exp
expan
expand
expd
expect
expect econom stimulu packag
expect increas
expell
expens
expert told reuter tehran hold
expert worri
expir
expir midnight
explain
exploit
explor
export
export commit
export commit usda
export hold option
export inspec
export said
export subsidi
export tax
express
expt
extend
extens corpor restructur
program
extent
extern
extern trad commission willi
extra
extra chemic
extra liquid
extraordinari
extraordinari item
extrem
extrem danger
fac
facil
fact
faction
faction ignor mpt
facto
factor
factor support
fail
failur
fairli drawn
fal
fall
fallen far
famili
far
farina
farm
farm disast
farm produc
farm subsidi
farmer
fat
favor
favor technic chart signal
favour
fear
feasibl
februari
februari april period
februari louvr accord
fed
fed expect
fed fund
fed interven
feder
feder budget
feder disast relief
feder republ
feder reserv
feder reserv bank
feder reserv chairman paul volcker
feder reserv enter
feder sav
fee
feed
feedstock
feel
fell
felt
fertil
feudal
field
fifth year
fight
figur
figur includ
fiji
fiji see
fil
final
final stag
financ
financ minist
financ minist gerhard
financ offici
financi
financi advisor
financi condition
financi corp
financi market
financi sector
find domest polit prowess
finish
fir
firm
firm said
firm sal
fiscal
fiscal stimulu packag
fiscal year
fish
fish disput
fit
fitzwat said
fitzwat told report

fix
fix alloc rat
fix pric
fix rat
fla
flag
flag kuwaiti tanker
flag kuwaiti tanker sea
flag ship
flag tanker
flag tanker fleet
fledg war
fleet
fleet handl liquid cargo
float
flood
floor
florida
flow
flurri
focu
follow
forc
forc attack
forc baker hand
forecast
forecast australia
foreign
foreign currenc
foreign exchang
foreign exchang market
foreign invest
foreign minist
foreign natur ga sal
foreign purchas
forgotten
form
fort worth
fortun
forward
forward look
fourth
fourth quarter
fourth year
fragil
franc
franc continu
franc said
francais de petrol
frankfurt
freedom
freefall
freight rat
french
french wheat
fresh initi
fresh veget
friction
fridai
fridai averag
friend
friendli
frigat
frigat stark
frontier
frontier insur
frozen boneless beef
frozen hak fillet
frtr
frustrat
fund
fundament
futur
futur belong
ga
ga field
ga properti
gaf
gaf chairman samuel heyman
estim
gaf heyman
gain
gain addition trad right
gam
garo armen
gatt
gaug
gav
gav iran
gener
gener agreem
gener sav
gener weak
german
german bank
german monei
german monei market split
germani
get
giant dow chemic
gibraltar
gillett
giordano said
giv
given
glamor
glenview
global econom
global rat
global recoveri mov
gnp
go
gold
gold reserv
golden period
goldman sach
gon ahead
good
good prospect
gordon
got
govern
govern bond salesman
govern offici
govern said
govern secur market
govern sourc said
govern welcom
gra
grab
grac
grad
gradual market absorp
grain
grain oilse registr
grand
grant
great
greater
greater extent
greek tragedi
greenwich
gross
ground
groundnutse oil
group
group led
group said
group shr
grow
grow consensu
grow world
growth
growth expect
growth orient
guard
gulf
gulf area
gulf follow
gulf oil
gulf polici
gulf reflag
gulf war
gulf war expert

hail
half
half quota
halt
hamper
hand
happen
harbor
harcourt
harcourt brac jovanovich
hard land
hard red spring wheat
hard red winter wheat
hardli
harlan ullman
harri upham
hart scott rodino antitrust
improv act
harvest
hastili cal meet
hav
hbj
head
headquart
health
heavi
heavi import
hebei
hebei wheat aphid
hectar
hectar valu
hectarag plant
hedg
heighten
held
held discus
helicopt destroi
helmet
help
help block financ ministri plan
help boost
hemispher
henan
heng
henlei
henri bellmon
henri schuler
heublein
high
high alert
high margin busi
high qualiti beef
high qualiti hilton beef
high stat
higher
higher cost
higher pric
higher rat
hik
hill
hint
hisanobu ohs
histor experi
histori
hit
hitherto
hoax
hoel
hold
hold compani
hold corp
hold monei market tender todai
holder
hom
hong
hong kong
honor
honour
honour pledg
hop
hormuz
horrif casualti
hostil
hostil takeov attempt
hostil tender offer
hour
hous
hous speaker jim wright said
houston
houston bas cain chemic
houston bas compani said
hug
hug oil platform dot gulf
hurt
hyman
hyman said
hyper inflat
ibm
idaho
idea
identifi
identifi sal
ignor
ill
ill advis
illeg
imbal
imf
immedi
immedi impact
immedi shipment
imo
impact
impair
imperi oil
implem
implement
implicit
import
import plac
import politician
import restric
import shift
import wheat
impos
impos mln dlr
imposi
impress
improv
improv inspec system
improv quota
improv tariff
improv trad condition
improv trad term
incent
inch
inch shell
includ
includ britain
includ gain
includ loan loss provision
includ meat
includ sal
includ secur gain
incom
incom fund
increas
increas demand
increas significantli
increas tariff
increas tension
ind
ind ttl
india
india nil nil nil nil
indiana
indic
indic total
indic total includ report commit
individu
indonesia
industri
industri analyst
industri analyst said

industri nation
industri powerhous
industri said
industri sourc said
industri trend
industrywid
infantri attack result
inflat
inflationari
inflationari fear
inflationari fear boost gold bullion
influenc
influenti polit group
inform
initi
injuri
insist
instabl
institu
institut
insur
insur compani
integr
inten
intend
intend acreag
intent
interest
interest rang
intermedi
intern
intern accord
intern busi
intern econom
intern forc
intern monetari gamesmanship
intern oil produc
intern preced exist
intern studi
intern tanker
intern trad
internation
interven
interven just ahead
interven period
interven pric
interven rat
interven stor
interven tender
interview
invad iran territori
inventori
invest
invest firm
invest purpos
invest secur
invest secur gain
investig
investor
investor asher
investor group
investor group led
investor psychologi
invit
involv
iran
iran iraq war
iran respons
iran rostam
iran sai
iran start
iran war
iran war enemi iraq
iranian
iranian attack
iranian gunboat
iranian man
iranian militari offici said
iranian new agenc irna quot
iranian oil
iranian oil platform
iranian platform
iranian ship
iraq
iraq blam iran
iraqi missil
ireland
irna said
ironi
irv bank
irv bank corp
isl citi
issu
issu capit
itali
item
jam
jam abegglen
jam oneill
jan
januari
januari agreem
januari provid
japan
japan car
japan economi
japan failur
japan lucr oversea telephon busi
japan main offer
japan mov
japan nil
japanes
japanes bank dealer said
japanes bank dealer warn
japanes electron
japanes export
japanes good
japanes insur
japanes market
japanes offer
japanes offici
jardin matheson
jiangsu
jitteri investor
join
joint
joint statem
joint ventur
jonathan lynn
joseph
journalist
jpm
judg
juli
juli vorman
jun
jun fiji todai welcom
jun japan
jun onward
jun oversea shiphold group
jun presid reagan said
jun releas
jun resid
jun santo
jun year
jun year end march
just
justic
justifi
kamal kharrazi
kan
kansa citi
keep
kei
kei backer
kei elem
kei industri sector
kei japan ministri
keiichi udagawa
kentucki bas oil refin
kept
khor fakkan

kil
kil seamen
kill
kind
kingdom
knew
know
known
koehler
kohlberg kravi robert
kokusai denshin denwa
kong
kotc
kraft
kuwait
kuwait back
kuwait oil tanker
kuwait tanker
kuwaiti newspap
kuwaiti oil export said
kuwaiti oil tanker
kuwaiti ship
kuwaiti tanker
kuwaiti tanker convoi
lack
lago
lai
lanston
larg
larg scal
larger
larger wheat set
largest
largest loss
lat
latest
latest attack seen point
latest econom packag
latest exampl
latin american govern
launch
lautoka
law
law pass
lawson
lawson said
lawson told
lawsuit
ldp
lead
lead rat
leader
leahi propos
leas
leather footwear
leav
led
left
leftwich
legal framework
legal framework eas
legisl
lessen mpt author
letter
level
leverag buyout
levi
liabil
liber
liber cdi
libya
licenc
licenc request
lift
light
likelihood
limit
limit access
limit adjust
limit retaliatori strik
lin
linda sieg
linse oil
liquid
list
litig
littl
liv
liv room
lo angel
loan
loan associ
loan loss
loan loss provision
local
local maiz import said
local offic
locat
london
london televi interview tuesdai
long
long plan trip
long run stabil
long term
long term debt
long term loan
long tim
longer
longer term uncertainti
look
loss
loss carryforward
lost
lot
loui
louvr
louvr accord
louvr agreem
louvr currenc accord
low
lower
lower dollar
lower oil pric
lower rat
lower rat prospect
lowest
lubbock yesterdai damag
luck
lucr industri
lukman
luxembourg
machin corp
macro econom
magnitud
main
main disput
main showpiec
mainli
maintain
mainten
maiz
major
major compani execut predict
major econom power
major impetu
major industri
major industri nation
major manag rol
major oil
major plant capac
major produc region
major shift
major stak
major war
mak
maker
manag
manag director
manag director abdul fattah
manhattan
manila
manufactur
manufactur capac creat
march

march opec
march show
margin
marin midland bank
marin midland oneill said
mark
mark flow
mark rat
mark yesterdai
market
market believ
market condition
market develop
market fall
market instabl frighten
market offer glimps
market pric
market psychologi
market soft wheat
market suddenli believ fed
chairman paul volcker
market todai
marri
marri expect
marri felt
marri said
marri told reuter
marshall aid plan
marshall plan
martin
mass
massiv
massiv budget deficit
materi
materiel
matter
matter clos
matur
maxim
maximum
maxwell
mcf
meal
mean
measur
meat
medic profession liabil
medic quadrangl
medicin
medium
medium term
meet
meet need
melbourn
mellon
mellon bank chief economist
member
membership
memori
men
mention
merchant bank
merg
merger
merger agreem
merger propos
merrill
merrill lynch
messeng
metr
metropolitan plc
mexico
mich
michigan
mid
middl east
middl east expert sai
middl eastern oil produc
mideast
mideast gulf
mideast gulf situat
midland bank
midnight
mil
mild
militari
militari action
militari coup
militarili
milwauke
min
min iran
min iran harbor
miner
minimum
minimum export pric
minist
minist fail
minist kakuei tanaka
ministri
ministri action
ministri mov
ministri stanc
minn
minneapoli
minor
minor ministri
minu figur
minu total
missil
missil strik
mission
mix
mln
mln acr
mln barrel
mln barrel report
mln bpd
mln canadian dlr
mln common shar
mln debit
mln dinar
mln dlr
mln dlr budget
mln dlr cash
mln dlr loss
mln dlr plu
mln dlr surplu
mln dlr worth
mln hectar
mln loss
mln mln
mln outstand
mln peso
mln profit
mln riyal issu
mln shar
mln shar outstand
mln stg
mln tonn
mobil hom
moder
modest
modest rat
modest recoveri
modif
mondai
mondai attack
mondai talk
mondai wall street
monei
monei declin
monei mak petrochem busi
monei market
monei market dealer said
monei market oper
monei market rat
monei market rat continu
monei suppli
monetari author bought
monetari polici
monitor
monopol
monopoli

monsanto
month
month ago
month amid
month declin
month includ
monticello
montreal
morgan
morn
morocco
mortgag corp
morton hyman said
moscow
mothball ineffici plant
mount
mouth
mov
mpt
mpt influenc
mtc
mtc spent
mth
mth includ gain
mth includ loss
multilater
nadi
nam
nap
narrow
narrow rang
nation
nation averag loan rat
nation econom summit
nation largest independ oil
nation total
natsuo okada
natur
natur disast
natur ga
natur lower
naval escort
navi began
navi destroyer
navig
near
near record wheat crop
near term
nearli
nearli halv
necessari
need
negoti
negoti merger
neogtiat tabl
new
new confer
newli elect
newli releas annual report
newspap
newspap quot
nich
nicosia
nigel lawson
nigeria
night
nightmar
nikko secur intern
nil
nil nil
nil nil nil nil
non
non accru loan
non accrual
non accrual statu
non expens
non japanes oversea
telecommun firm
non spanish barlei export
non tradition corn farm
normal
normal april
norman robertson
north america
north american
north chicago
northern texa
norwai
not
not compani
not includ bulgur
not total
notabl
notion
novemb
number
nurtur
oaklei said
oat
object
obstructionist
obtain
obviou
obvious
occasion
occur
oceania
oceania nil nil nil nil
oct
oct allwast
oct hug oil platform dot
oct iran
oct kuwait oil export
oct militari expert sai
oct privat export report sal
oct rumor circul
octob
octob shipment
odd
odd man
offer
offer better trad term
offer pric
offer slightli higher rat
offic
offici
offici figur
offici iranian reaction
offici not
offici pric
offici report
offici respons
offici sai
offici said
offici said minist
offici washington sought
offset
ohio
oil
oil compani
oil dril platform
oil equival
oil export
oil field
oil industri
oil industri analyst
oil industri sourc said
oil market
oil minist
oil output
oil platform
oil pric
oil produc
oil rig mil east
oil stock ros
oil tanker
oilse
oilse export pric adjust
oilse product
okada
okla
oklahoma citi
oklahoma clean
old

old on
oljeselskap
opec
opec confer
opec confer presid rilwanu lukman
opec current pric
opec meet
opec member
opec produc
opec produc ceil
opec produc quota drop
opec ris
open
open market transac
oper
oper farm
oper incom
oper loss
oper profit
oper remain divid
oper shr
oper shr ct ct
oper shr loss
oper shr loss ct profit
oper todai
oper vessel
opinion
opportun
oppos
opposi
opposit
option
option origin sal plu actual export
orchestr campaign
order
ordinari
ordinari shar
oreffic
oreffic said
organ
origin
origin bid
orlando
osg see opec quota kei
ottawa
ounc
outbreak
outgo
output
outrag
outstand
outstand common stock
outstand shar
outvot
overproduc
oversea
oversea own
oversea said pct
overthrown
overwhelm major dual own
own
own liberian flag ship
own pct
pac consult
pacif
packag
pact
pact expir
pai
pai april
pai econom
paid
painstakingli slow declin paint
pakistan
palai
panic
paper
paper said abnorm weather condition
par
par valu
parent
pari
pari accord
pari agreem
parti
particip
particular
particularli
partli
partner
partnership
pass
pass legisl
past
pasta
patch
patent
patrol gulf
paul
paul oreffic
payabl
payabl april
payabl juli
payabl jun
payment
payout
payroll
pct
pct annual
pct annual rat
pct capac
pct discount rat
pct duti
pct gain
pct leap
pct ris
pct stak
pct stock dividend
pct tariff
peak
pek
pellet
pend
pension
pension account
pentagon
pentagon announc
peopl
perelman
perform
period
period end april
period includ
permit
permit acreag
permit atpp
persian gulf
pest
petrochem industri
petrochem plant
petrochem relat busi
petroleum
petromin
pfennig
phas
philadelphia
philippin
philippin coconut author said coconut produc
philippin damag agricultur crop
phoenix
pick
piec
pipelin
plac
placem
plai
plan
plant
plant produc

plant util rat approach pct
plastic
platform
plc
pledg
plu
plung
poehl
poehl not
point
point cut
pois
polic said
polici
polici decision
polit
polit analyst said
polit impact
popular
port
portfolio
portug
portug entri
posit
position
positon hedg
possibl
possibl disturb
possibl sal
possibli
post
postal
postal sav
postpon
potent polit forc
potenti
pound
powderi mildew
power
power polit faction
pre summit speech celebr
pre tax profit
precipit
predict
predict lower profit
prefer
prefer dividend
prefer shar
prefer stock
prelud
premium
prepar
present
present level
present liquid monei market condition
preserv
presid
presid reagan
presid reagan said todai
presid reagan speak
press
press confer
press iraq
pressur
presum
pretax
pretax earn
pretax profit
prev
prev week
prevail
prevent
prevent massiv
preventedf
previou
previou week
previous
previous announc
previous announc acquisi
previous announc sal
previous disclos
previous led
pric
pric cut
pric increas
pric tak
prim
primarili
princip
principl
print
prior
prior year
privat
privat center
privat held
privat held cain chemic
privat held compani own
probabl
problem
proce
proceed
process
produc
produc bas
produc loss
produc restraint
product
product area
product pric
profession
profit
profit chemic
profit specialti chemic
program
program crop
progress
prohibit
project
project corn
project stock surpass
prolong dry spell
promin
promin stem
promis
promot
prompt gaf
promptli
propel
properti
propos
propos acquisi
propos offer
prorat
prospect
protec
protect
protect gulf
protect kuwaiti oil tanker
protectionist postur
prov
prov reserv fell
proven
provid
provid consum
provinc
provision
psychologi turn
pty
public
public adjust
public offer
publicli
pull
pump
purchas
purchas author
purchas pric
purpos
pursu
push
push profit margin
put

puzzl
qtly div
qtly div ct ct prior
qtr
qtr feb
qtr mar
qtr march
qtr sept
qtr shr
qtr shr ct ct
qualiti
quantiti
quarter
quarter earn
quarter end
quarter just end
quarter oper
quarter profit
quarterli dividend
question
question presid reagan
tomorrow
quickli
quito
quot
quot todai
quota
radio
radio station
rain
rainfall
rais
rais cash
rais fear
rais hop
rais rat
rang
rapese
rat
rat boost
rat cut
rat fell
rat pul
reach
react
reaction
readi
reaffirm
reaffirm america willing
reaffirm support
reagan
reagan administr
reagan coupl
reagan declar
reagan hint
reagan impos
reagan plan
reagan polici
reagan said
real
real estat
real estat asset
realiz
realli
reappoint
reason
recapit
recapit plan
receiv
reces
reciev
reckon
recogn
recommend
record
record april
record dat
record year
recoveri
red
redress gap world trad imbal
redress global trad imbal
reduc
reduc program
reduc trad
reed
reed shar
refin
reflag
reflect
reflect tighter suppli
refus
regard
region
registr
regret
regul
regulatori
regulatori approv
regulatori control
reinforc
reinstat
reiter advic
reject
reject licenc
rekindl inflat
rel
relat
relat compani
relax
releas
remain
remain shar
remark
remedi
remov
renew pressur
reorgan
repai
repeat past mistak
repeat warn
replac
report
report outstand sal
report quantiti
report said
repres
repres approxim pct
republican parti
repurchas
repurchas agreem
repurchas pact
repurchas pact rat soon
request
requir
research
reserv
reserv ad requir
reserv bank
reshadat oil platform
resid return
resign
resist
resolv
respect
respect market season
respond
respons
rest
restat
restaur
restor faith
restor investor confid
restric
restructur
restructur program
restructur touch
result
result includ
result reflect
result restat
resum
retail
retain

retali
retali continu
retaliatori measur
retir
retreat
return
reuter
rev
revenu
revenu exclud
revers
revers stock split
review
revlon
revlon group
ric
richard
richmond
ridicul
rift
rig
right
ris
risen
risk
rit aid corp
rival
river
riyal
road deliveri
robert
robert giordano
robertson
rol
roll
rom
ronald
room
root caus
ros
rostam offshor oilfield
round
row
royal gold
rpt argentin grain
rul
rul liber democrat parti
rumor
rumour
run
run bal
sabah
saf
sai
said
said approv
said bach secur
said dow
said fred axelgard
said intern tanker market
said iranian
said iranian forc
said japan
said jim mcgroarti
said john dosher
said kotc
said lanston jon
said market impati
said pac john dosher
said retir adm
said robert brusca
said soviet cargo ship bound
said tanker requir decreas
said waseda univers professor
mitsuru
said washington
sail
sal
sama
sama lower
sama offer
samjen
san
san diego
san mateo
sanction
sank
santo
santo group stak
santo said
satellit commun market
saturdai
saudi
saudi arabia
saudi arabian monetari
sav
sav system abil
saw
scal
scal war
scandinavia
scar
sceptic
schedul
schlesing
schuler said
scof
scott
sea isl
sea isl citi
season
season began bas
seat
sec
sec fil
second
second half
second quarter
secret
secretari
secretari immedi
secretari jam baker
section
sector
secur
secur repurchas
seek
seek acquisi
seek control
seen
seiz
sel
sell
semiconductor
semolina
senat
senat agricultur committe
senat democrat
senat panel studi loan rat
senat staff said
senior
senior figur
sens
sent
sentim
sept
septemb
seri
serious
serv
servic
session
set
set asid plan
set custom repurchas
set stock split
settl
settlem
sever drought
sever flood
shandong
shanxi
shar
shar adjust

shar amount
shar cash
shar clos
shar issu
shar tender
sharehold
sharehold group
sharp
sharp drop
sharpli
sharpli higher
shed unrel busi
shell
shelter
shelter set
shift
ship
ship forc
ship sourc said
shipment
shop
shop center
shop closur
short
short dollar
short term
short wai outsid
shortag
shortli
shown
shr
shr ct
shr ct ct
shr dilut
shr dilut ct ct
shr loss
shr loss ct loss
shr loss ct profit
shr loss ct profit ct
shr primari
shr profit
shr profit ct loss
shr profit ct loss ct
shr profit ct profit
shut
sichuan
sickli chemic
sid
sign
signal
signific
signific oil reserv
significantli
silkworm
silkworm missil strik
similar
similarli
simpli
simultan
sit astrid
situat
siz
slap mln dlr
slid
slight
slight review
slow
smack
small
small deposit
small group
smaller volum award
smith barnei
smooth
soft red winter wheat
soft wheat
soi cak
sold
solution
solv
somebodi head
soon
sorghum
sort
sought
sourc
sourc said
south africa
south african maiz
south australian
south korea
south west queensland
southern philippin
southward
soviet
soviet tanker initi
soviet tanker set
soviet union
soybean
soybean cak
soybean oil
soybean total
spain
spanish barlei
spanish market
spark
speak
special
special tender
specialti
specialti busi
specialti product
specif warn
specter
specul
specul capit inflow
specul sel
speech
speech prepar
spell
spend
spent
spik
spin
spirit
spit
split
spok
spokesman
spokesman said
spokeswoman
spokeswoman phylli oaklei
spot
spot market
spot pric
spotlight
spread
spur
srd
stabil
stabil currenc
stabil currenc valu
stabl
staff
stai
stai special long
stak
stamford
stand
stand athwart
standard
standard oil
stapl petrochem commod
stark possibl
start
start cover
stat
stat depart renew
stat depart sai
statem
statem said
station
statoil
steadi

steel
steel product
steer
step
stephen marri
sterl invest bank group
stg
stick
stimul
stimul action
stimul domest demand
stimulu
sto
stock
stock dividend
stock exceed
stock exchang
stock market
stock market reaction todai
stock ownership
stock ownership plan
stock pric
stock pric tumbl
stock sal
stock split
stock valu
stockhold
stoltenberg
stoltenberg said
stoltenberg told
stood
storm
strait
strateg
strateg invest
strategi
stream
streamlin
strength
strengthen
stress
strik
strong
strong ris
stronger
stronger countermeasur
structur
structur remain
studi
stun
style
styren
subject
submit
subsidi
subsidiari
substanti
substanti higher pric
substanti increas
substitut
subtract
sudden wall street retreat
suddenli
suffer
sugar
suit
sullivan
sumitomo bank
summer
summit
sundai
sunflowerse
sunflowerse cak
sunflowerse nil nil
sunflowerse oil
supermarket
superpow
suppli
supplier
support
support kuwait tanker
surfac baker look respons
surplu
surpris
suspend
suva
swamp
swiss
swiss bank corp
swiss franc
sydnei
syracus
system
system corp
taipei
taiwan
taiwan import
taken
takeov
takeov attempt
takeov bid
takeov offer
talk
tanaka
tank trailer
tanker
tanker rat
target
tariff
tarnish
tat
tax
tax credit
tax cut
tax loss
tax loss carryforward
teach
technologi
tehran
tehran radio
telecommun
telecommun industri
telecommun market
telephon
televi interview
televi set
tempo
temporari
temporari reserv
tender
tender expect
tender offer
tension
tepid
term
term debt
term grain suppli agreem
termin
territori water
terror
texa
texa democrat
texa democrat senat lloyd
texa gulf coast
textil
thailand
thing
think
think rat
think tank
third
thou bushel
thought
thousand
threat
threaten
threaten crop
threaten food shortag
threshold
thrust
thursdai
thursdai declar
thwart
ti

tier structur ris feet
tighten
tighten credit
tim
tim earn
tit
tobacco
todai
tokyo
tokyo dealer
tokyo open
told
told report
told reuter
tomorrow
ton
tonn
tonn fob
tonn south african maiz
tonn valu
top yen todai
tornado
toronto
torranc
total
total cie
total explor
total explor area
total explor australia pty
total hold
total nil nil nil nil
total outstand
total reserv
touch
tpn
track
trad
trad ban
trad data
trad data seen
trad deal
trad deficit
trad deficit drag
trad deficit lead
trad disput
trad estim
trad friction
trad friction diplomaci
trad gap
trad gap narrow
trad partner
trad row
trad sanction
trad situat
trad sourc
trad sourc said
trad stand
trad surplu
trader
trader said
transac
transfer
translat
transport
travel
treasuri
treasuri baker
treasuri secretari jam baker
treasuri secretari jam baker cam
treasuri secretari statem help
treasuri sourc said
trigger
trigger larger corn
trim
troubl
trust
try
trying
tuesdai
tulsa
turbul ahead
turf
turkei
turn
turnov
twister
type
uae
uchida
uchida said
ullman said washington
ultim destin
unabl
unchang
unchang rat
understand
underwrit
undisclos
undisclos term
unfavor reaction
unhappi
union
union carbid corp
union texa
union texa oil reserv drop
union texa said
unit
unit arab emir
unit kingdom
unit sal
unit stat
univers
unload arm
unmil ric
unnam
unnam destin
unpaid acreag reduc requir
unravel
unreason
unrel
unrel busi
unsatisfi
unsettl
unsightli mass
unsuccess
unsuccess attempt
unsuccess bid
unusu
unveil
unwelcom
upcom summit
upland cotton
upland domest raw cotton
upturn
upward pressur
urg
usda
usda caution
usda discuss
usda estim australia wheat crop
usda estim european commun crop
usda gav detail breakdown
usda project export
usda report corn sold
usda set
usdaprj
ussr
ussr bui
ussr nil nil nil nil
util
util market
valu
vancouv
veget oil
venezuela
venic
venic econom summit
venic jun
ventur
vessel
vibranc
vibrant economi badli
vic presid helmut schlesing

vienna
view
vigor explor program
violat
violenc
virtual
virul anti american polici
visit
vital
vital interest
vital suppli
volcker
volcker nightmar
volum
vot
wai
wai street
wak
wall street
wall street drop
wall street fear
wall street retreat
waltham
want
war
ward
warlik step
warm water
warn
warrant
warship
wash
washington
washington announc
washington privat center
washington seiz
washita
watch
water
waterburi
watersid worker feder
waterwai
waukesha
wav
weak
weaken
weaker
wednesdai
week
week ago
week dash
week end april
weekend
weight
weinberg said
welcom
wellington
went
west
west berlin
west berlin stat central
west german
west german bundesbank
west german financ minist
gerhard
west german govern
west german offici mondai
west german polici
west germani
west germani withdrew objec
west texa
west texa intermedi
west texa sour
western
western democraci
western diplomat sai iraq start
western world
westhem
wheat
wheat acreag reduc requir
wheat bran
wheat continu
wheat crop
wheat export
wheat midg
wheat product
wheat red mit
whit wheat
whitten
wholli
wholli own subsidiari
wi
wil
wing
winter
wip
wireless plc
wisconsin
withdraw
withdraw right
withdrew
wood
worcest
word
work
worker
world
world bank
world economi
world expan
world market
world stock
world trad
worldwid
worri
wors
worsen
worth
wright
writ
wtc
wti
year
year ago
year ago period
year agreem sign
year earlier
year end march
year exclud
year fall
year iran iraq war
year loss
year low
year meet
year presidenti
year term
year trying
yel
yen
yen strength
yen todai
yen worth
yesterdai
yield
york
york corp said
york law
york said
young
zealand
zurich

ภาคผนวก ค.

ผลงานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์และได้รับการตีพิมพ์

- P.Jitpakdee and W.Kreesuradej, "Dimensionality Reduction of Feature for Text Categorization", Proceeding of The IEASTED International Conference on Advances in Computer Science and Technology (ACST 2007), Phuket, Thailand, April, 2007.

Canadian Secretariat
THE INTERNATIONAL ASSOCIATION OF SCIENCE AND TECHNOLOGY FOR DEVELOPMENT

IASTED

98, Suite #101, 2509 Dieppe Avenue SW, Calgary AB T3E 7J9 Canada
Telephone (403) 288-1175 Fax (403) 247-8851
E-mail: calgary@iasted.com (for conferences) and journals@iasted.com (for publications)
Web Site: http://www.iasted.org

11/12/2006
Parisut Jitpakdee
King Mongkut's Institute of Technology Ladkrabang
Faculty of Information Technology
Ladkrabang
Bangkok
Thailand 10520

Dear Miss Jitpakdee,

Re: 559-120 Dimensionality Reduction of Features for Text Categorization

Congratulations, your paper has been accepted for oral presentation and publication at the IASTED International Conference on Advances in Computer Science and Technology (ACST 2007), to be held from Apr 02, 2007 to Apr 04, 2007, in Phuket, Thailand. We cordially invite you to attend and present your paper at the conference. We also encourage you to register and book your flight as soon as possible, if you have not already done so.

Please complete the following by the registration deadline of Feb 12, 2007.

1. Registration Form and Payment (mandatory)
2. Author Information Form (mandatory)
3. Hotel Reservation Form (if you need assistance in reserving a hotel room)
4. Copyright Form (mandatory)

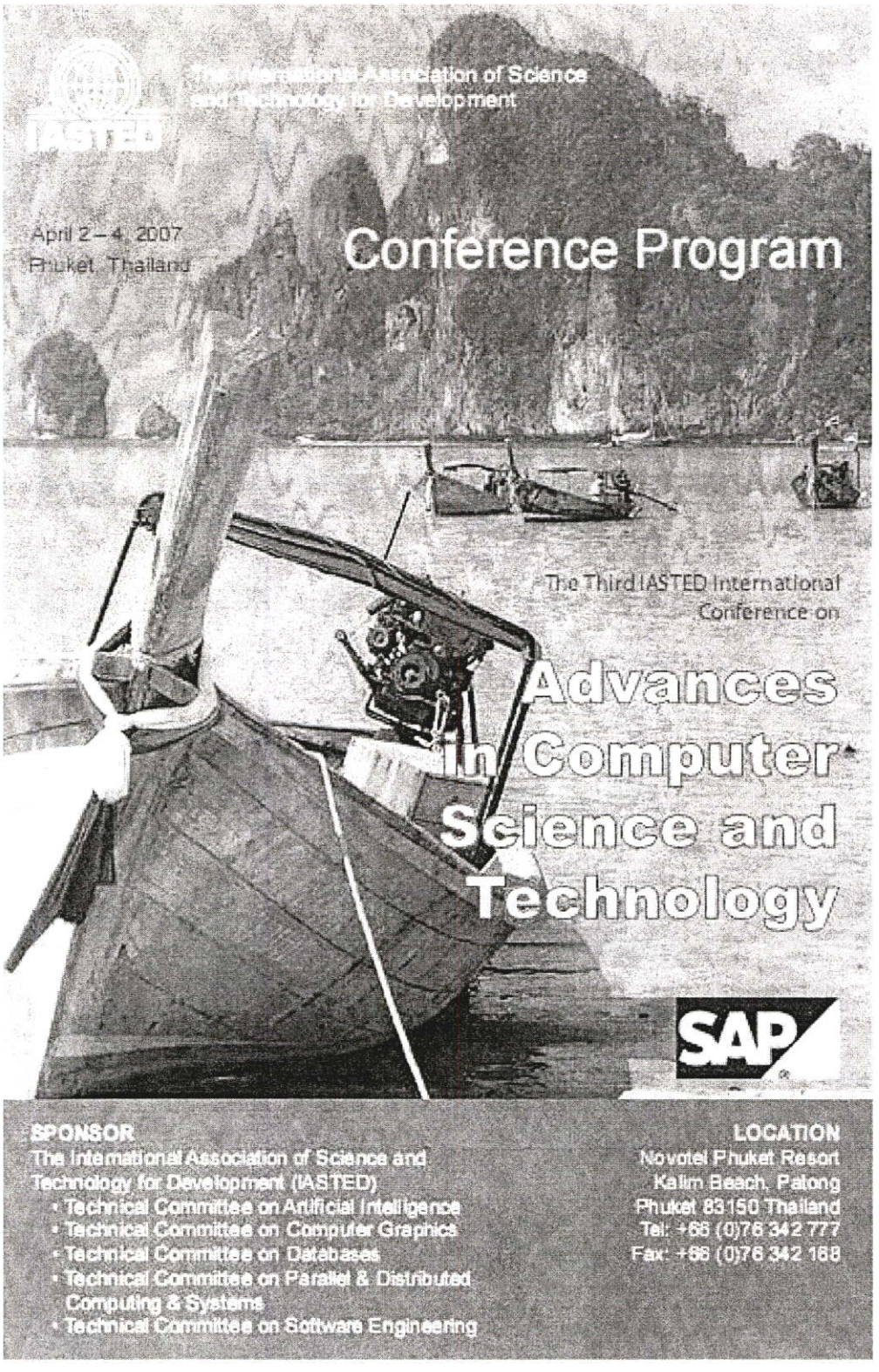
The above materials are available for viewing on our website at http://www.iasted.org/conferences/home-559.html. Please let me know if you have any questions regarding registration.

Once again, congratulations on your ACST 2007 acceptance. We are very excited to be able to include your research and ideas in the conference, and we look forward to seeing you in Phuket, Thailand.

Sincerely,

Amy Christenson
Conference Manager

Please note - For Visa Purposes:
The International Association of Science and Technology for Development (IASTED), is a non-profit organization founded in Zurich, Switzerland in 1977. The purpose of IASTED is to promote economic development through science and technology. After 30 years, IASTED continues to bring top scholars and members of industry together to develop and share new ideas, facilitate cultural exchange and encourage international unity.


The International Association of Science and Technology for Development
IASTED
April 2 – 4, 2007
Phuket, Thailand
Conference Program
The Third IASTED International Conference on
Advances in Computer Science and Technology
SAP
SPONSOR
The International Association of Science and Technology for Development (IASTED)
• Technical Committee on Artificial Intelligence
• Technical Committee on Computer Graphics
• Technical Committee on Databases
• Technical Committee on Parallel & Distributed Computing & Systems
• Technical Committee on Software Engineering
LOCATION
Novotel Phuket Resort
Kalim Beach, Patong
Phuket 83150 Thailand
Tel: +66 (0)76 342 777
Fax: +66 (0)76 342 188

DIMENSIONALITY REDUCTION OF FEATURES FOR TEXT CATEGORIZATION

Parisut Jitpakdee* and Worapoj Kreesuradej†
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
*E-mail: jajah_pj@hotmail.com and † E-mail: worapoj@it.kmitl.ac.th

ABSTRACT
This paper proposes a new technique for dimensionality reduction of features for text categorization. Unlike conventional method, our phrase features are generated based on word sequences of different length (Multi-grams) from phrases extracted from whole documents. Then, we utilize Odds ratio (*OR*) to perform phase feature selection. From preliminary experiments, the proposed techniques show better performance than that of conventional methods.

KEY WORDS
Data mining, text mining, text categorization

1. Introduction

With the rapid growth of online documents available, information organization and retrieval have become a great importance. An automated text classification has been utilized as a method to organize online documents. Text categorization (also known as text classification) is, simply, the automated assignment of natural language texts to predefined categories based on their content [1].
Bag of word (BOW) is the approach most commonly used to represent documents in Text Categorization (TC) [2,3]. In BOW, each document is represented by a vector that contains an importance weighting for every words in the collection. The main drawback of the BOW representation is in destruction of semantic relations between words. Many researches attempt to preserve this information by using phrases as features, the results from the study of David Lewis [4] has shown that phrase-based representation not improve performance of the categorization because it a) contain a high number of terms, b) have an uneven distribution of feature values, c) contain many redundant features, and d) contain a lot of noise.
Several efforts have been made to circumvent the possible problems posed by using phrases and some research results showed that the addition of n-grams to the BOW representation improved performance [10]. But the main problem in using n-grams is the huge potential increase in the number of features. To reduce numbers of features, many papers extract the top-scored features using various feature selection methods [5,6,7]
In addition, unlike conventional method, our phrase features are generated based on word sequences of different length (Multi-grams) from phrases extracted from whole documents. The experiment results show that using Multi-grams and words has better performance than that of using only word features.
This paper is organized as follows: section 2 we introduce dimensionality techniques and odds ratio, section 3 describe our method and our algorithm, section 4 present the experiments and section 5 is conclusion of this paper.

2. Dimensionality Reduction Techniques

Basically, dimensionality reduction techniques can be classified into two approaches: feature generation and feature selection [1].

2.1 Feature Generation (or Feature Induction)

New features, which are not necessarily words, are sought for representation. Usually, the new features are synthesized from the original set of features.
There are two common approaches to feature generation. The first one combines features using disjunction only. In this approach features are grouped into subsets and each such subset is then considered as a new feature. Stemming and word clustering belong to this family of methods. The second approach groups features using only conjunctions, for example, by grouping consequent or close (in proximity) words into phrases. The use of n-grams is a common method in this family.

2.2 Feature Selection (or Feature Reduction)

Feature Selection studies how to select a subset of term (e.g. words) that are used to representing documents. Its purposes include reducing dimensionality, removing irrelevant and redundant features, reducing the amount of data for categorization. Feature Selection works by ranking all the terms and then selecting some percentage. A variety of ranking criteria have been used in text

categorization. One of the simplest methods is selecting terms with medium to high document frequency (*DF*). Other efficient term selection function, widely use, are χ^2 [1], information gain [1], mutual information [1]. In this paper we use Odds Ratio (*OR*) [1] measuring discriminating power for each candidate features.

Odds ratio (*OR*) measures the odds of the word occurring in the positive class normalized by that of negative class. The basic idea is that the distribution of features on the relevant documents is different from the distribution of features on the non relevant documents. It is defined as follows:

$$OR(t_i, c_i) = \log \frac{P(t \mid c_i)[1 - P(t \mid \bar{c}_i)]}{[1 - P(t \mid c_i)]P(t \mid \bar{c}_i)} \qquad (1)$$

Where $P(t \mid c_i)$ is the conditional probability of term t occurring in category c_i, $P(t \mid \bar{c}_i)$ is the conditional probability of term t occurring not in category c_i

3. Our Method

In Our method, first if they are web documents we will remove all tag (such as <html>,
, etc.), then we use each sentence in each document as initially phrase. After that we separate each sentence into phrase by remove stop words and stem every word in phrase with Porter's algorithm [8]. For each distinct phrase in each category, we generate multi-grams (set of word sequences that have varied length from 1 to size of phrase) by using sequence of words in phrases. Our multi-gram features difference from n-grams that they have various sizes and we have not to fix maximum size of gram. Table 1 shows some phrases and its multi-grams obtained from our method.

Table 1
Example Phrase and its Multi-grams

Phrase:	"efficient text categorization algorithm"
4-gram	"efficient text categorization algorithm" (*OR* = 0.5)
3-grams	"efficient text categorization" (0.4), "text categorization algorithm" (0.8)
2-grams	"efficient text" (0.2), "text categorization" (0.8), "categorization algorithm" (0.6)
1-gram	"efficient" (0.4), "text" (0.7), "categorization" (0.8), "algorithm" (0.65)

Then, we calculate Odds Ratio (*OR*) for each gram and we select only gram in each category that has *OR* more than all its sub-grams as features of their category.

Finally, we union all features from every category to be as our Features set.

From example in table 1 with our method, the finally result features set will be {"efficient", "text", "categorization", "algorithm", "text categorization", "text categorization algorithm"}. The algorithm of our method is present in figure 1.

Input: D – training set of documents
C – set of categories
min_f – minimum frequency
Output: Set of Features
Procedure Features Generation

Features = ϕ
For each category $c \in C$
  $P = \phi$
  $W = \phi$
  For each document $d \in D$ and d is in category c
    Wdoc=ϕ
    PH = set of extracted phrases from d
    $P = P \cup PH$
  End
  $MG = \phi$
  For each phrase p in P
    For i=1 to *size*(p)
      $MG = MG \cup$ grams size i from p
    End
  End
  $CF = \phi$
  For each gram *mg* in MG that has frequency >= *min_f*
    Let *sg* = sub-grams of *mg* ; 1-grams to *size*(*mg*)-1 grams
    If *OR*(*mg*,*c*) > max (*OR*(sg,c))
    $CF = CF \cup mg$
  End
  Feature = *Feature* $\cup$ *CF*
End

Figure 1. Feature Generation Algorithm

4. Experiments

4.1 Test Collections

We use web data set, called the J-series (available at http://ftp.cs.umn.edu/dept/users/boley/PDDP data/). The J-series contains 185 documents and ten categories showed in Table 2.

We generate multi-gram from all documents. Table 3 present number of bigrams (from all words) and number of multi-grams (from phrases) from J-series data set but not include 1-gram because we want to compare only phrase features. It shows that our method can reduces number of feature more than 70% of conventional method.

Table 2
J-series Data Set

CATEGORY	LABEL	DOCUMENTS
Affirmative action	AA	20
Business Capital	BC	19
Information Systems	CIS	19
Electronic Commerce	EC	19
Intellectual Property	IP	19
Employee Rights	LER	16
Materials Processing	MP	17
Personal Management	PM	19
Manufacturing Systems	SI	18
Industrial Partnership	ZIPT	19

Table 3
Number of Bigrams from J-series Data Set

Class	Bigram	Multi-gram (not include 1 grams)	Percent of Reduction
AA	132	38	78.79%
BC	660	156	76.36%
CIS	294	86	70.75%
EC	183	82	55.19%
IP	752	137	81.78%
LER	1,193	217	81.81%
MP	1,996	498	75.05%
PM	99	40	59.60%
SI	2,414	635	73.70%
ZIPT	467	190	59.31%
Total (distinct)	7327	2079	71.63%

4.2 Categorization

For categorization, each document is represented as Feature vectors according to the standard vector space model with *tfidf (term frequency-inverse document frequency)* weighting [9], *tfidf* is defined as:

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#Tr(t_k)} \qquad (2)$$

where is $\#(t_k, d_j)$ denotes the number of time term t_k occurs in document d_j, $\#Tr(t_k)$ is the number of documents in which t_k occurs, and $|Tr|$ is the number of documents in training set. The weights resulting from *tfidf* are often normalized by cosine normalization, given by:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_s, d_j))^2}} \qquad (3)$$

where w_{kj} is the weight of term k in document j, and $|T|$ is the number of all terms.
In the vector-space model, the similarity between two documents *di* and *dj* is commonly measured using the cosine function [9], given by

$$S(d_i, d_j) = \frac{\sum_{k=1}^{|T|} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{k=1}^{|T|} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{|T|} w_{kj}^2}}, \qquad (4)$$

In this paper, we use *k*-nearest neighbor (*k*NN) method for categorize documents [9]. Given a test document *d*, the *k*NN classifier finds the *k* nearest document among the training documents, using document-document similarity. The similarity score of each neighbor document to the test document is used as the weight of the categories pre-assigned to the training document. The second step is to estimate the likelihood of each category by summing the weight of the category of the *k* nearest documents as follows :

$$P(C_j|D_x) \approx \sum S(D_x, D_i)\, P(C_j|D_x) \qquad (5)$$

where $P(C_j|D_x) \in \{0,1\}$, if $P(C_j|D_x) = 0$ mean that D_x is not respect to category C_j, and $P(C_j|D_x) = 1$ if D_x is respect to category C_j.
In our experiments, we use k values = 1, 2, 3..., 20 for kNN categorization

4.3 Performance Measure

We applied 4-cross-validation and use average accuracy rate to measure the correctness of categorization where accuracy rate is the ratio of documents classified correctly to the category c_i and other than c_i among all the documents. The result of the experiments show in table 4.

Table 4
Result of Categorization

Feature Type	Accuracy (%)
Words	83.23
Phrases	71.89
Words + Phrases	74.61
Words + Bigrams	84.30
Words + Selected Bigrams	85.38
Multi-grams (Our Method)	86.42

5. Conclusions

This paper proposes a new technique for dimensionality reduction of features for text categorization. From preliminary experiments, the proposed techniques show better performance than that of conventional methods. In the future, extensive theoretical studies and experiments will be presented.

References

[1] F. Sebastiani, Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), March 2002, 1-47

[2] T. Dumis, J. Platt, D. Heckerman, and M. Sahami, Inductive learning algorithms and representation for text categorization. *CIKM'98, ACM Info. & KM.* Bethesda, US, 1998, 148-155.

[3] M. Weiss, C. Apte, F. J. Damerau, et al., Maximizaing text-mining performance. *IEEE Intell. Sys.* 14(2), 1999, 63-69.

[4] D. D. Lewis, An evaluation of phrasal and clustered representations on a text categorization task. *SIGIR'92 15th*, Kobebhavn, DK, 1992, 37-50.

[5] D. Mladeni and M. Grobelink, Word sequences as features in text learning. *17th Electro. & Comp. Science Conf.*, Slovenia, 1998, 145-148.

[6] C. M. Tan, Y.F. Wang and C. D. Lee, The use of bigrams to enhance text categorization. *Info. Proc. & Management*, *38*(4), 2002, 529-546.

[7] R. Bekkerman and J. Allan. Using Bigrams in Text Categorization. *CIIR Technical Report IR-408*, University of Massachusetts, 2003.

[8] M. F. Porter, An algorithm for suffix stripping. *Program, 14*(3), 1980, 130-137.

[9] Y. Yang, An Evaluation of statistical approach to text categorization. *Information Retrieval, 1*(1/2), 1999, 69-90.

[10] M. Yetisgen and W.Pratt, The effect of feature representation on MEDLINE Document Classification. *AMIA '05*, Washington D.C., October, 2005.

ประวัติผู้เขียน

ชื่อ-นามสกุล	นางสาวปริสุทธิ์ จิตต์ภักดี
วัน เดือน ปีเกิด	2 พฤษภาคม 2525 ที่โรงพยาบาลแม่และเด็ก จังหวัดเชียงใหม่
ที่อยู่	41 ถ.ป่าพร้าว ต.ป่าแดด อ.เมือง จ.เชียงใหม่ 50100
ประวัติการศึกษา	ปีพุทธศักราช 2547 สำเร็จการศึกษา วิทยาศาสตรบัณฑิต (วิทยาการ-คอมพิวเตอร์) จากมหาวิทยาลัยเชียงใหม่
ผลงานวิจัย	- Jeerayut Chaijaruwanich, Chalothorn Liamwirat, Parisut Jitpakdee, et al. "DNA Microarray Data Analysis System," The 15th Annual General Meeting of the Thai Society for Biotechnology (TSB), Chiang Mai, February, 2004. - P.Jitpakdee and W.Kreesuradej, "Dimensionality Reduction of Feature for Text Categorization", Proceeding of The IEASTED International Conference on Advances in Computer Science and Technology (ACST 2007), Phuket, Thailand, April, 2007.