การประมาณค่าขนาดเอ็นแกรมแบบปิด

A SIZE ESTIMATION OF CLOSED N-GRAM

นครs สาแก้ว

NAKORN SAKEAW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2550

การประมาณค่าขนาดเอ็นแกรมแบบปิด

A SIZE ESTIMATION OF CLOSED N-GRAM

นคร สาแก้ว

NAKORN SAKEAW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2550

A SIZE ESTIMATION OF CLOSED N-GRAM

NAKORN SAKEAW

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2007

หัวข้อวิทยานิพนธ์	การประมาณค่าขนาดเอ็นแกรมแบบปิด
นักศึกษา	นายนคร สาแก้ว
รหัสประจำตัว	45064607
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2550
อาจารย์ที่ปรึกษาวิทยานิพนธ์	รศ.ดร.วีระ บุญจริง

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาสูตรสำหรับการคำนวณจำนวนของเอ็นแกรมแบบปิดจากขนาดของข้อมูลข้อความภาษาไทย โดยวิธีการวิจัยเริ่มจากการหาจำนวนเอ็นแกรมจากแฟ้มข้อความจำนวน 400 แฟ้ม แล้วใช้ 300 แฟ้มแรกในการหาสูตรและส่วนที่เหลือใช้ในการทดสอบสูตรที่ได้ การหาสูตรทำโดยใช้กฎของฮีฟและการวิเคราะห์การถดถอย สำหรับการวิเคราะห์การถดถอยได้ทดสอบตัวแบบบสามตัวแบบ ได้แก่ ตัวแบบเชิงเส้นอย่างง่าย ตัวแบบเอ็กโปเนนเชียลและตัวแบบพาวเวอร์ ด้วยตัววัด R^2 จะได้ว่าตัวแบบพาวเวอร์เป็นตัวแบบที่ดีที่สุด จากนั้นงานวิจัยนี้ได้นำตัวแบบที่ได้ไปพยากรณ์จำนวนเอ็นแกรมแบบปิดโดยวัดความถูกต้องของการพยากรณ์ด้วยตัววัด MAPE ผลปรากฏว่าสูตรที่ได้จากตัวแบบพาวเวอร์ดีกว่าสูตรที่ได้จากการใช้กฎของฮีฟ

Thesis Title	A Size Estimation of Closed N-gram
Student	Mr.Nakorn Sakeaw
Student ID.	45064607
Degree	Master of Science
Program	Computer Science
Year	2007
Thesis Advisor	Assoc.Prof.Dr.Veera Boonjing

ABSTRACT

The purpose of this research is to find closed-form formulas for determining numbers of closed n-gram from size of text files. The research starts with determining closed n-gram with cutoff 2 to 21 from 400 text files. It then uses the first 300 to derive formulas and the rest to test fitness of formulas obtained. Formulas for each cutoff are derived using the well known Heap law and regression analysis. Based on measure of goodness of fit called the coefficient of determination or R^2, the power regression model is the best model described by given data among simple linear, exponential and power regression models. The research uses formulas obtained to predict numbers of closed n-gram and measures prediction errors in terms of mean absolute percentage error or MAPE. The results show that formulas of power regression model are better than of the Heap law.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยดี ผู้วิจัยขอกราบขอบพระคุณ รศ.ดร.วีระ บุญจริง ซึ่งท่านให้ความกรุณาเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ อีกทั้งกรุณาให้คำปรึกษา แนวคิดและแนะนำในการดำเนินงานวิจัยตลอดจนวิธีแก้ปัญหาต่างๆ อันเป็นประโยชน์ต่องานวิจัยนี้เป็นอย่างยิ่ง

ผู้วิจัยขอกราบขอบพระคุณต่อ ผศ.ดร.ศรัณย์ อินทโกสุม ซึ่งท่านให้ความกรุณามาเป็นประธานกรรมการสอบ ผศ.ดร.จีรพร ศรีสวัสดิ์ ซึ่งท่านให้ความกรุณามาเป็นตัวแทนกรรมการจากภาควิชาและ ดร.เฉลิมศักดิ์ เลิศวงศ์เสถียร ซึ่งท่านให้ความกรุณามาเป็นตัวแทนกรรมการจากบุคคลภายนอก ทำให้ได้รับคำปรึกษาและคำแนะนำต่างๆในการสอบวิทยานิพนธ์

ขอขอบคุณอาจารย์อัสนีวัลย์ ทรัพย์ล้น, ดร.สุวัฒน์ เตชะเพชรไพบูลย์, อาจารย์อนุชาติ บุญมากและอาจารย์ประสบโชค มีวงศ์ รวมถึงเพื่อนอาจารย์คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏเพชรบุรี ที่ค่อยช่วยเหลือและเป็นกำลังใจ

ขอขอบคุณเพื่อนๆพี่ๆน้องๆนักศึกษาปริญญาโททุกคน ที่เป็นกำลังใจต่อผู้วิจัยตลอดมา ขอขอบคุณน้องคัมภีร์ เสริมกวินรักษ์ ที่ให้การสนับสนุนข้อมูลและขอขอบคุณเป็นพิเศษสำหรับน้องจิราภรณ์ ทัดรัตน์ ผู้อยู่เบื้องหลังความสำเร็จในการสนับสนุนและช่วยเหลือผู้วิจัยในด้านต่างๆ จนงานวิจัยนี้สำเร็จเสร็จสิ้นด้วยดี

สุดท้ายนี้ขอกราบขอบพระคุณ คุณพ่อกัลยาและคุณแม่ทองสุข สาแก้ว ผู้เปรียบดังพรหมของบุตร ผู้เป็นชีวิตและจิตใจ ผู้คอยอบรมสั่งสอนในทางที่ถูกที่ควร อีกทั้งให้กำลังใจและเป็นทุกสิ่งทุกอย่างสำหรับผู้วิจัย

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

นคร สาแก้ว

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การเตรียมข้อความเพื่อการบีบอัดเป็นวิธีหนึ่งที่ช่วยเพิ่มประสิทธิภาพในการลดขนาดข้อมูล สำหรับการเตรียมข้อความที่อยู่ในรูปแบบสายอักขระอย่างเช่นภาษาไทย ซึ่งมีเอกลักษณ์เฉพาะและยากต่อการจำแนกขอบเขตของคำ วิธีการเตรียมข้อความดังกล่าวจึงต้องสอดคล้องกับลักษณะของข้อมูลดังกล่าว แนวคิดเอ็นแกรม(n-gram) จึงเป็นหนึ่งในวิธีการที่จะแก้ปัญหาลักษณะข้อมูลที่เป็นสายข้อมูลที่ต่อเนื่องกัน การเตรียมข้อมูลด้วยวิธีเอ็นแกรมเป็นวิธีหนึ่งของการเตรียมข้อมูลแบบการแทนวลีด้วยรหัส [1]

การแทนวลีด้วยรหัสใหม่ จะใช้พจนานุกรมเพื่อเก็บความสัมพันธ์ระหว่างรหัสหรือดัชนีที่ใช้แทนและวลีที่ใช้แทน แต่ปัญหาหนึ่งของการสร้างพจนานุกรมด้วยวิธีเอ็นแกรมคือ ความเหมาะสมของจำนวนข้อมูลของเอ็นแกรมที่จะเก็บลงในพจนานุกรมที่ออกแบบรองรับเอาไว้ การกำหนดจำนวนข้อมูลของเอ็นแกรมให้มีมากหรือน้อยเกินไปมีผลต่อความครอบคลุมของข้อมูลที่จะเตรียมรวมไปถึงต้นทุนที่เกิดขึ้นกับขนาดของพจนานุกรมตามมา

งานวิจัยนี้จึงเสนอสูตรสำเร็จการประมาณค่าขนาดเอ็นแกรมแบบปิด เพื่อเป็นความรู้และข้อมูลพื้นฐาน สำหรับการกระบวนการเตรียมข้อมูลเพื่อการบีบอัดต่อไป

1.2 วัตถุประสงค์

ในงานวิจัยนี้เป็นการศึกษาในส่วนของการเตรียมข้อความซึ่งจำเป็นต้องใช้พจนานุกรม ดังนั้น งานวิจัยนี้จึงมุ่งเน้นเพื่อหาจำนวนเอ็นแกรมแบบปิดที่ใช้สำหรับสร้างเป็นพจนานุกรม โดยมีวัตถุประสงค์ในการศึกษา คือ เพื่อพัฒนาสูตรลำดับคำนวนจำนวนเอ็นแกรมแบบปิดจากขนาดของแฟ้มข้อความ

1.3 ขอบเขต

การพัฒนาสูตรงานวิจัยนี้ทำโดยสุ่มแฟ้มข้อความภาษาไทยจำนวน 400 แฟ้มข้อความ แล้วใช้ 300 แฟ้มแรกในการสร้างสูตรโดยใช้กฎของฮีฟและการวิเคราะห์การถดถอย ในการวิเคราะห์การถดถอยเป็นการทดสอบตัวแบบการถดถอยที่เหมาะสมสามตัวแบบได้แก่ ตัวแบบเชิงเส้นอย่างง่าย ตัวแบบเอ็กโปเนนเชียล และตัวแบบแบบพาวเวอร์ เมื่อหาความเหมาะสมแล้วนำตัวแบบที่ได้ไปพยากรณ์โดยใช้แฟ้มข้อความในส่วนที่เหลือที่ได้เตรียมไว้เพื่อหาว่าสูตรใดเหมาะสมที่สุด โดยวัดจากตัววัดความเหมาะสมของการพยากรณ์

1.4 ส่วนประกอบของวิทยานิพนธ์

ส่วนที่เหลือของวิทยานิพนธ์ฉบับนี้แบ่งเนื้อหาออกเป็น 4 บท ดังต่อไปนี้

บทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง โดยแบ่งออกเป็น 3 ส่วน คือ ในส่วนแรกจะเป็นการกล่าวถึงความรู้พื้นฐานเกี่ยวกับเอ็นแกรมและการนำไปประยุกต์ใช้งานในด้านต่างๆ ส่วนถัดมาจะเป็นการกล่าวถึงความรู้พื้นฐานเรื่องการประมาณจำนวนคำสำหรับเอกสารที่ได้มีการพัฒนาสูตรเอาไว้ และในส่วนสุดท้ายเป็นการสรุปในเนื้อหาของบทนี้

บทที่ 3 กล่าวถึงกระบวนการในการประมาณค่าขนาดเอ็นแกรมแบบปิด โดยเริ่มตั้งแต่วิธีการหาสถิติการเกิดเอ็นแกรม การขจัดความถี่ที่ไม่ต้องการออกจากระบบ การขจัดการเหลื่อมกันของแกรมและในส่วนท้ายจะเป็นตัวแบบสมการในการประมาณค่าขนาดเอ็นแกรมแบบปิดที่เลือกใช้ในงานวิจัยนี้ คือ วิธีตามกฎของฮีฟและการวิเคราะห์การถดถอย

บทที่ 4 การประเมินผล โดยนำแฟ้มข้อความที่เหลือที่ได้เตรียมไว้จำนวน 100 แฟ้ม นำมาทดสอบการพยากรณ์กับตัวแบบที่ได้หาความเหมาะสมแล้วเพื่อหาว่าสูตรใดเหมาะสมที่สุด โดยวัดจากตัววัดความเหมาะสมของการพยากรณ์

บทที่ 5 สรุปและข้อเสนอแนะ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับงานวิจัยซึ่งได้พัฒนา โดยจะกล่าวถึงพื้นฐานแนวความคิดเรื่องของเอ็นแกรม ตลอดจนนำความรู้เรื่องของเอ็นแกรมไปประยุกต์ใช้ในงานด้านต่างๆ และความรู้เรื่องการประมาณจำนวนคำโดยใช้กฎของฮีพ (Heaps' Law)

2.1 เอ็นแกรม

ในส่วนนี้จะกล่าวถึงแนวความคิดของเอ็นแกรม ความหมายและการหาเอ็นแกรมตลอดจนการนำไปประยุกต์ใช้ในงานด้านต่างๆ โดยมีรายละเอียดแบ่งเป็นหัวข้อดังนี้

2.1.1 แนวคิดของเอ็นแกรม

เอ็นแกรม [2] คือ ส่วนของสายอักขระที่มีความยาวเป็นเอ็นอักขระ เช่นข้อมูล "faaacbde" เอ็นแกรม หมายถึง ส่วนของข้อมูลที่มีความยาวเอ็นอักขระติดต่อกัน เช่น สี่แกรมอาจจะมีค่า 'faaa' หรือ 'aaac' หรือ 'cbde' เป็นต้น ถ้าหากความยาวของสายอักขระที่พิจารณามีความยาว (หรือเอ็น) เท่ากับ 1 อักขระจะเรียกว่า ยูนิแกรม (unigram) และถ้ามีความยาวเท่ากับ 2, 3 อักขระ จะเรียกว่า ไบแกรม (bigram) และไตรแกรม (trigram) ตามลำดับ เป็นต้น และในการเขียนแทนแกรมเหล่านี้สามารถเขียนในลักษณะตัวเลขกำกับเพื่อบ่งบอกถึงจำนวนอักขระหรือแกรมที่พิจารณาได้

สำหรับการพิจารณาหาค่าเอ็นแกรมนั้นจะแบ่งการพิจารณาเพื่อหาเอ็นแกรม 2 ลักษณะ คือ การพิจารณาเอ็นแกรมของสายข้อมูลโดยมีความเลื่อมกัน (Conjunction) ตัวอย่างเช่น ข้อมูล "faaacbdef" เมื่อต้องการพิจารณาหาค่า 3-แกรม แบบมีความเลื่อมกันจะได้ "faa", "aaa", "aac", "acb" เป็นต้น และอีกวิธีคือ การพิจารณาโดยอิสระแยกจากกัน (Disconjunction) สำหรับในแต่ละแกรมใดๆ ตัวอย่างเช่น (จากข้อมูลข้างต้น) เมื่อต้องการพิจารณาหาค่า 3-แกรม แบบอิสระแยกจากกันจะได้ "faa", "acb", "def" ทั้งนี้ทั้งสองวิธีจะขึ้นอยู่กับวัตถุประสงค์และการนำเอ็นแกรมไปใช้ในงานด้านต่างๆ จากความรู้เรื่องเอ็นแกรมนี้สามารถนำไปประยุกต์ใช้ประโยชน์กับข้อมูลที่มีลักษณะเป็นสายข้อมูลที่เขียนเรียงต่อกัน ซึ่งในปัจจุบันนี้มีหลายภาษาที่มีลักษณะข้อมูลเป็นแบบสายข้อมูลเช่นนี้ เช่นภาษาไทย ภาษาจีน ภาษาญี่ปุ่น เป็นต้น ตลอดจนลักษณะภาษาที่อยู่ในรูปแบบดีเอ็นเอ (DNA) นอกจากนี้ก็ยังสามารถนำไปใช้ได้กับข้อมูลที่มีลักษณะเป็นคำเช่นภาษาอังกฤษได้อีกด้วย

2.1.2 เอ็นแกรมกับการนำไปประยุกต์ใช้

ความรู้เรื่องเอ็นแกรมสามารถนำไปประยุกต์ใช้กับงานด้านต่างๆ มากมาย ซึ่งในหัวข้อนี้จะยกตัวอย่างการนำเอ็นแกรมไปประยุกต์ใช้กับงานด้านต่างๆ ที่เกี่ยวข้องกับข้อความส่วนหนึ่ง

ในงานวิจัย [2] ได้นำความรู้เอ็นแกรมไปประยุกต์ใช้ ในระบบการนำข้อมูลมาสร้างเป็นดัชนี(Text indexing) ซึ่งเรียกว่า วิธีการสร้างดัชนีเอ็นแกรม (n-gram indexing) โดยจะมีกระบวนการค้นหาข้อมูล (full text) ในระบบการสืบค้นสารสนเทศ (Information retrieval) ด้วยวิธีการหาสตริงย่อย (Substring matching) เพื่อสร้างเป็นฐานข้อมูลสำหรับเก็บเอ็นแกรม เนื่องจากการใช้เอ็นแกรมนั้นมีประโยชน์ในเรื่องความอิสระต่อภาษา และเมื่อมีการค้นหาเอกสารไม่จำเป็นต้องใช้คำสำคัญในการสืบค้น การนำเอ็นแกรมมาใช้จะช่วยให้คำสำคัญของเอ็นแกรมนั้นครอบคลุมมากยิ่งขึ้น หลักในการนำเอ็นแกรมมาสร้างเป็นดัชนีสำหรับ จะเลือกแกรมน้อย เนื่องจากแกรมน้อยนั้นจะมีความถี่มากและจำนวนตัวอักขระที่เกิดในแกรมเหล่านี้น้อย จึงทำให้มันครอบคลุมเนื้อหา แต่จะมีข้อเสียคือ คำสำคัญที่เป็นเอ็นแกรมนี้ เมื่อมีการเรียงลำดับใหม่ไม่สามารถนำข้อมูลเหล่านี้มาประกอบให้เป็นคำที่มีความหมายได้

ในงาน [7] ได้นำความรู้เอ็นแกรมไปประยุกต์ใช้ในระบบการจัดกลุ่มข้อความ โดยการหาเอ็นแกรมสำหรับแต่ล่ะเอกสาร ซึ่งเอกสารใดที่มีความถี่ของเอ็นแกรมที่อยู่ในช่วงเดียวกัน ขนาดเอ็นแกรมเดียวกันเอกสารทั้งสองนั้นจัดว่ามีเนื้อหา(Content)เดียวกัน การนำเอ็นแกรมมาประยุกต์ใช้กับงานนี้มีข้อดี คือ สามารถจำแนกเอกสารภาษาใดก็ได้ เนื่องจากข้อดีของเอ็นแกนมคือความอิสระต่อภาษานั้นเอง

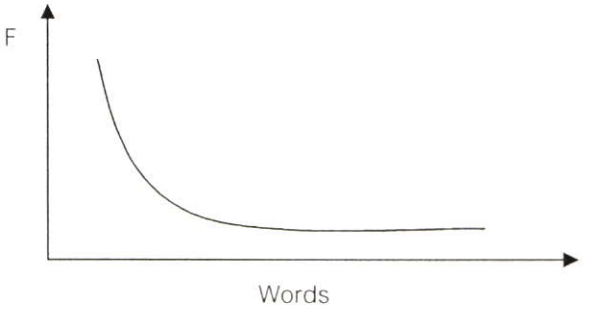
ในงาน [10] ได้นำความรู้เอ็นแกรมมาประยุกต์ใช้สำหรับการสกัดหรือดึงคำออกจากสายอักขระ ในงานวิจัยนี้ได้พัฒนาในระบบภาษาที่มีลักษณะเป็นสายข้อความ เนื่องจากข้อความมีลักษณะติดต่อกันจึงไม่สามารถทราบได้ว่าส่วนใด เป็นส่วนที่เป็นคำที่มีความหมายสำหรับภาษานั้นๆ จึงอาศัยพื้นฐานการทำงานแบบเอ็นแกรมเพื่อสกัดหรือดึงคำออกจากสายข้อมูลนั้น

นอกจากนี้ยังนำความรู้เรื่องเอ็นแกรมมาประยุกต์ใช้สำหรับการพยากรณ์คำที่จะเกิดจากการพิมพ์ของผู้ใช้ [11] โดยปัญหาหนึ่งของการใช้ระบบการสร้างเอกสารแบบหลายภาษาคือ การที่ผู้ใช้พิมพ์อักขระหรือข้อความอยู่บนโหมดการพิมพ์อีกโหมด เช่น ต้องการพิมพ์ข้อความในโหมดการพิมพ์ภาษาไทยแต่ในขณะที่ยังอยู่ในโหมดการพิมพ์ภาษาอังกฤษ ข้อความที่ได้แทนที่จะเป็นอักขระภาษาไทยกลับกลายเป็นอักขระภาษาอังกฤษออกมาแทน ทำให้เสียเวลาในการลบและพิมพ์ข้อความเดิมใหม่อีกครั้ง

2.2 การประมาณจำนวนคำสำหรับเอกสาร

ในส่วนนี้จะกล่าวถึงการประมาณค่าจำนวนคำ โดยอยู่บนพื้นฐานตามกฎของซิฟ (Zipf's Law) และกฎของฮีฟ (Heaps' Law) ซึ่งทั้งสองกฎนี้ตั้งอยู่บนพื้นฐานของกฎยกกำลัง (Power Law) กล่าวคือกฎทั้งสองดังกล่าวจะอยู่ในรูปแบบสมการเลขยกกำลังซึ่งจะกล่าวรายละเอียดต่อไป

กฎของซิฟ [4, 5, 8] ตั้งชื่อตาม George Kingsley Zipf (1902-1950) เป็นกฎที่เกิดจากการสังเกตความถี่ของการเกิดเหตุการณ์ (P) และในแต่ละเหตุการณ์นั้น จะนำมาจัดเรียงลำดับจากมากไปน้อยตามความถี่ของมันซึ่งเรียงว่า ช่วง(Rank)ของความถี่ พบว่าความสัมพันธ์ของเหตุการณ์ดังกล่าวและช่วงความถี่ของมันมีความสัมพันธ์กันแบบฟังก์ชันเลขยกกำลัง ดังรูปที่ 2.1 เป็นการแสดงการกระจายของความถี่ของคำที่ผ่านการเรียงลำดับใหม่



รูปที่ 2.1 การกระจายของความถี่ของคำที่ผ่านการเรียงลำดับใหม่

จากรูปจะได้ว่า

$f \propto 1/r$ เมื่อ f คือ ความถี่ของเหตุการณ์
r คือ ลำดับของคำที่ผ่านการเรียง

ซึ่งมีค่าเท่ากับ $f.r = k$ เมื่อ k คือ ค่าคงที่

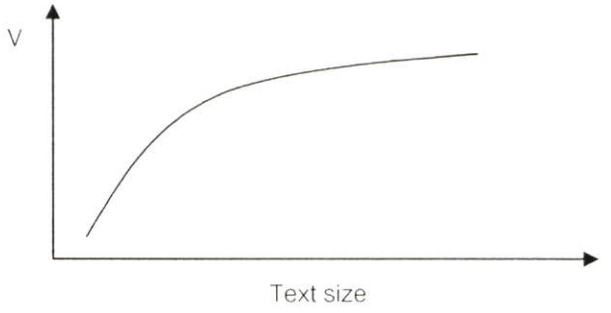
ถ้าพิจารณาความน่าจะเป็นของคำที่จะเกิดในช่วง r แทนด้วย Pr และกำหนดให้ N คือจำนวนของคำที่จะเกิดขึ้นทั้งหมดในเอกสารซึ่งเป็นภาษาอังกฤษ สามารถแสดงความสัมพันธ์ได้ดังต่อไปนี้

$$\begin{aligned} Pr &= f/N \\ &= (k/r).(1/N) \\ &= Ar^{-1} \end{aligned}$$

เมื่อ A คือ ค่าคงที่ซึ่งจากการทดลองมีค่าประมาณ 0.1

กฎของเลขยกกำลังจะมีรูปแบบโดยทั่วไปคือ $Y = kX^{C}$ ซึ่งจะเห็นได้ว่ากฎของซิฟก็เป็นกฎเลขยกกำลังซึ่งมีค่า C = -1 กฎของซิฟสามารถพยายกรณ์ความน่าจะเป็นที่ค่อนข้างแม่นยำสำหรับคำที่เกิดในช่วงที่สูงและต่ำ มากๆ

จากกฎของซิฟนี้ทำให้เกิดการประมาณจำนวนคำขึ้นมาที่เรียกว่ากฎของฮีฟ (Heaps' Law) [12] จากการสังเกตพบว่าเมื่อมีขนาดของเอกสารเพิ่มขึ้น ทำให้จำนวนของคำเพิ่มขึ้นด้วยเช่นกันซึ่งความสัมพันธ์นี้สามารถแสดงดังรูปที่ 2.2



รูปที่ 2.2 การกระจายของคำสำหรับแต่ละขนาดของเอกสาร

โดยทั่วไปจำนวนคำใหม่ที่พบในเอกสารจะเพิ่มขึ้นแบบลอการิทึมกับขนาดเอกสาร ถึงแม้ว่าความยาวของคำจะมีผลกับของเอกสาร ดังนั้นกฎของฮีฟจึงคิดจากความยาวเฉลี่ยของคำและสร้างความสัมพันธ์ระหว่างคำและขนาดของเอกสารขึ้นมาได้ดังนี้

$$V = Kn^{\beta}$$

เมื่อ V คือ จำนวนของคำ

n คือ ขนาดของเอกสาร

K คือ ค่าคงที่จากการทดลองจะมีค่าอยู่ประมาณ 10 – 100

β คือ ค่าคงที่ซึ่งอยู่ระหว่าง 0 ถึง 1 และจากการทดลองจะมีค่าอยู่ประมาณ 0.4 - 0.6

จากสมการกฎของฮีฟ ซึ่งตั้งอยู่บนพื้นฐานกฎเลขยกกำลัง ถูกนำมาใช้ประโยชน์สำหรับการพยากรณ์คำสำหรับแต่ละเอกสาร โดยสามารถพยากรณ์ได้จากสมการดังข้างต้น

2.3 สรุป

สำหรับงานวิจัยนี้ได้นำความรู้เรื่องเอ็นแกรมมาประยุกต์ใช้ เพื่อการเตรียมข้อมูลสำหรับการบีบอัด โดยนำเอ็นแกรมใดๆ มาสร้างเป็นพจนานุกรมเพื่อใช้สำหรับการแปลงและถอดรหัสการแปลง แต่เนื่องจากการนำเอ็นแกรมใดๆ มาสร้างเป็นพจนานุกรมนั้นเกิดปัญหาคือ ไม่สามารถพยากรณ์ได้ว่าควรนำเอ็นแกรมเหล่านั้นมาใช้สำหรับเป็นพจนานุกรมเท่าไร เนื่องจากขนาดของพจนานุกรมสำหรับเอ็นแกรมใดๆ นั้นขึ้นอยู่กับขนาดของเอกสาร ดังนั้นงานวิจัยนี้จึงพัฒนาวิธีการประมาณค่าขนาดเอ็นแกรมแบบปิดสำหรับแต่ละเอกสาร โดยนำความรู้เรื่องเอ็นแกรมและแนวคิดเรื่องการประมาณจำนวนคำจากกฎของฮีฟซึ่งกล่าวมาแล้วข้างต้นมาประยุกต์ใช้ เพื่อหาตัวแบบที่เหมาะสมสำหรับขนาดเอ็นแกรมแบบปิด ด้วยวิธีการวิเคราะห์แบบถดถอย (Regression Analysis) [6, 9] ซึ่งรายละเอียดขั้นตอนวิธีการดังกล่าวในบทถัดไป

บทที่ 3
การประมาณจำนวนของเอ็นแกรมแบบปิด

3.1 เอ็นแกรมและเอ็นแกรมแบบปิด

3.1.1 เอ็นแกรม

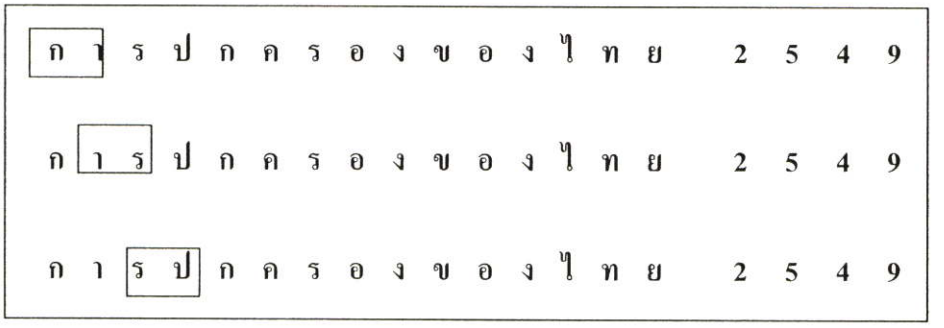
กำหนดให้ $T = \{x_1x_2x_3...x_s\}$ และสมาชิกของ T เรียกว่า อักขระ และ s เป็นจำนวนอักขระภายในแฟ้มข้อมูล โดยที่ $x_i \in$ {ตัวอักษร, ตัวเลข, สัญลักษณ์พิเศษ} *เอ็นแกรมของ* T คือ ส่วนของสายอักขระย่อยที่มีความยาวเป็น n เขียนแทนด้วย $N\text{-}gram = \{x_1x_2...x_n\}$ โดยที่ $N\text{-}gram \subseteq T$

3.1.2 ความถี่ของเอ็นแกรม

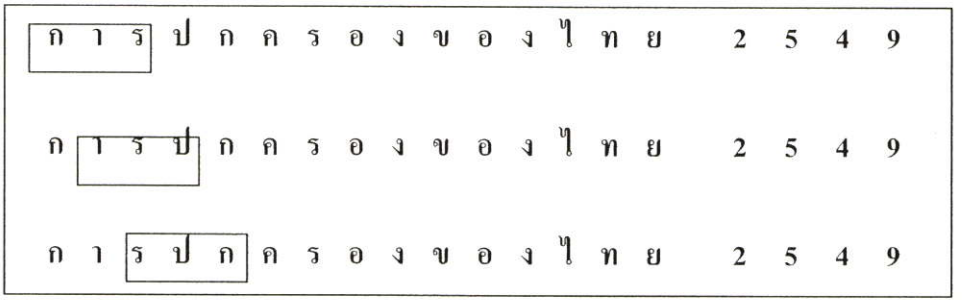
จากแนวคิดเอ็นแกรมข้างต้นได้นำมาใช้หาการเกิดเอ็นแกรมที่เกิดขึ้นซ้ำๆในข้อความจากแฟ้มข้อมูลต้นกำเนิด โดยการสร้างพจนานุกรมเก็บข้อมูลเอ็นแกรมและนับความถี่ของการเกิดเอ็นแกรมดังกล่าว ซึ่งมีขั้นตอนพื้นฐานการทำงานดังต่อไปนี้

1. เริ่มต้นกำหนดให้ค่า file_ptr เป็นตัวเก็บค่าตำแหน่งตัวอักขระในข้อความแฟ้มข้อมูลต้นกำเนิด เมื่อกำหนดให้ file_ptr มีค่าเท่ากับเลขตำแหน่งอักขระตัวแรกในข้อความ
2. อ่านข้อความทีละเอ็นแกรม โดยเริ่มอ่านจากตำแหน่งที่เก็บอยู่ใน file_ptr เป็นจำนวนเอ็นอักขระ (n อักขระ)
3. นำเอ็นแกรมที่ได้จากข้อ 2 ไปค้นหาในพจนานุกรม
   - ถ้าค้นหาพบ ทำข้อ 5
   - ถ้าไม่พบ ทำข้อ 4
4. เพิ่มเอ็นแกรมที่ได้ในพจนานุกรม
5. บวกค่าความถี่ของแกรมเพิ่มหนึ่งหน่วย
6. บวกค่า file_ptr เพิ่มหนึ่งตำแหน่ง (file_ptr = file_ptr + 1)
7. กลับไปทำข้อ 2 จนกระทั่งสิ้นสุดข้อความในแฟ้มข้อมูลต้นกำเนิดและสิ้นสุดกระบวนการ

จากกระบวนการหาความถี่ข้างต้นใช้สำหรับการสร้างพจนานุกรมของแกรมในขนาดที่ค่าเอ็นมีค่าแตกต่างกันไปตั้งแต่ 2 แกรมถึง n แกรม ซึ่งรูปที่ 3.1 แสดงวิธีการอ่านข้อความด้วยวิธีเอ็นแกรม เมื่อค่า n = 2 และรูปที่ 3.2 แสดงวิธีการอ่านข้อความด้วยวิธีเอ็นแกรม เมื่อค่า n = 3 โดยสมมติให้ค่า T = { การปกครองของไทย 2549 } เป็นตัวอย่างข้อความแบบสายอักขระ

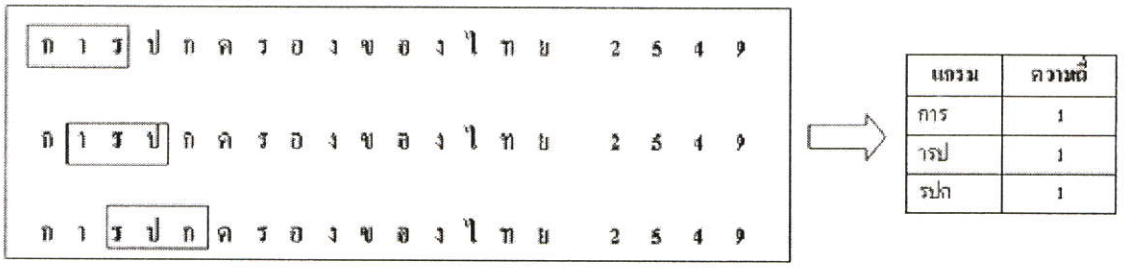


รูปที่ 3.1 ตัวอย่างวิธีการอ่านข้อความด้วยวิธีเอ็นแกรม เมื่อค่า n = 2



รูปที่ 3.2 ตัวอย่างวิธีการอ่านข้อความด้วยวิธีเอ็นแกรม เมื่อค่า n = 3

จากรูปที่ 3.2 เป็นการนำข้อมูลที่อ่านจากข้อความได้ในแต่ละครั้งขนาด 3 แกรม นำไปเก็บลงในพจนานุกรมตามกระบวนการที่ได้กล่าวแล้วข้างต้น ได้มาซึ่งสถิติจำนวนความถี่ของการเกิดแกรมต่างๆรูปแบบกัน ดังรูปที่ 3.3



รูปที่ 3.3 ตัวอย่างวิธีการสร้างพจนานุกรมและการเก็บความถี่การเกิดของแกรมที่ n = 3

หลังจากเสร็จสิ้นกระบวนการหาความถี่ของการเกิดเอ็นแกรม ขั้นตอนถัดมาคือการขจัดแถวข้อมูลเอ็นแกรมใดๆ ที่มีค่าความถี่ต่ำกว่าเกณฑ์ที่กำหนดเอาไว้ เช่น กำหนดเกณฑ์ค่าตัดออกเท่ากับสอง หมายความว่า ค่าจำนวนความถี่ของการเกิดของเอ็นแกรมในพจนานุกรมจะต้องมีค่าตั้งแต่สองขึ้นไป ความถี่ที่ต่ำกว่าเกณฑ์ที่กำหนดใน cutoff จะไม่นำมาพิจารณา

ในการพิจารณาการหาค่าความถี่ตามกำหนดเกณฑ์ค่า cutoff นั้น สามารถแสดงด้วยตัวอย่างที่ 3.1

ตัวอย่างที่ 3.1 การหาสถิติจำนวนแกรมจากข้อความตัวอย่างขนาด 341 ไบต์ ดังรูปที่ 3.4

อย่าคิดที่จะเปลี่ยนใคร
แต่รักได้มั๊ยที่เขาเป็นเขา
คนที่เรารัก อาจดูไม่ดีพอในสายตาคนอื่น
แม้แต่เราเองก็ยังเห็นความไม่ได้เรื่องของเขา
ฉันเจอมัน ความไม่ได้เรื่องน่ะในตัวคนที่ฉันรัก
และฉันก็หลงคิดว่า..ตัวเองดีเด่นซะเต็มประดา
ฉัน..จะเปลี่ยนเขาให้ได้
ฉันจะทำให้เขาเปลี่ยนแปลงตัวเองในทางที่ดีขึ้น
และฉันก็ทำไม่สำเร็จ
เขาก็ยังเป็นเขาคนเดิม

รูปที่ 3.4 ตัวอย่างข้อมูลต้นกำเนิด

จากรูปที่ 3.4 เป็นตัวอย่างข้อมูลเพื่อใช้ทดสอบหาสถิติของการเกิดแกรมในขนาดเอ็นที่ต่างๆ กันโดยแสดงอยู่ในรูปของตารางที่ 3.1 ซึ่งเป็นผลลัพธ์จากการหาสถิติเมื่อกำหนดให้ n = 2 โดยไม่มีการกำหนดค่าตัดออก ส่วนตารางที่ 3.2 ผลลัพธ์ที่ได้อยู่ในรูปกำหนดค่า n=2 ที่ค่า cutoff = 2 และ cutoff = 3 ดังตารางที่ 3.3 ตามลำดับ

ตารางที่ 3.1 แสดงตัวอย่างพจนานุกรมขนาด n = 2 ที่ยังไม่ผ่านกระบวนการตัดออก

ลำดับ	เทรม	ความถี่
1	-..	2
2	-.ด	1
3	-.จ	1
4	-กไ	1
5	-กี้	3
6	-ขา	5
7	-ขอ	1
8	-ขึ	1
9	-คิ	2
10	-คร	1
11	-คน	3
12	-ดว	2
13	-งก	1
14	-งเ	2
15	-งข	1
16	-งน	1
17	-งค	1
18	-งด	1
19	-งต	1
20	-งใ	1
21	-งห	1
22	-จะ	3
23	-จด	1
24	-จอ	1
25	-ฉั	6
26	-ชะ	1
27	-ดห	1
28	-ดั	4
29	-ดู	1
30	-ตี	3
31	-ตว	1
32	-ต่	1
33	-ตา	1
34	-ต่	2
35	-ตา	1
36	-ตั	3
37	-ตั้	1
38	-ที	5
39	-ท่า	2
40	-ทา	1
41	-นใ	1
42	-นเ	3
43	-นท	3
44	-นส	1
45	-นอ	1
46	-นด	1
47	-น่	1
48	-นด	1
49	-นร	1
50	-นก	2
51	-นช	1
52	-น.	1
53	-นจ	1
54	-นแ	1
55	-ปล	4
56	-ปี	1
57	-ปร	1
58	-พอ	1
59	-มิ	2
60	-บ่	4
61	-ม้	1
62	-ม้	2
63	-บป	1
64	-ป	1
65	-ยน	3
66	-ยท	1
67	-ยด	1
68	-ยั	1
69	-ร้	3
70	-รา	2
71	-ร้	2
72	-ระ	1
73	-ริ	1
74	-ลี	3
75	-ละ	2
76	-ลง	2
77	-วา	2
78	-วด	1
79	-ว่	1
80	-วเ	2
81	-สา	1
82	-สำ	1
83	-ห์	1
84	-หล	1
85	-ห้	2
86	-อย	1
87	-อา	1
88	-อใ	1
89	-อื	1
90	-อง	6

ตารางที่ 3.1 (ต่อ) แสดงตัวอย่างพจนานุกรมขนาด n = 2 ที่ยังไม่ผ่านกระบวนการตัดออก

ลำดับ	เทอม	ความถี่
91	-อม	1
92	-ะเ	3
93	-ะฉ	2
94	-ะด	1
95	-ะท	1
96	-ัก	3
97	-ั	1
98	-ัง	1
99	-ัน	7
100	-ัว	3
101	-าค	2
102	-าเ	3
103	-าร	1
104	-าจ	1
105	-าย	1
106	-าม	2
107	-า.	1
108	-าไ	1
109	-าง	1
110	-ำใ	1
111	-ำไ	1
112	-ิเ	1
113	-ิด	2
114	-ี	8
115	-ีพ	1
116	-ีเ	1
117	-ีย	1
118	-ึ	1
119	-ื	3
120	-ุไ	1

ลำดับ	เทอม	ความถี่
121	-เป	4
122	-เบ	5
123	-เร	5
124	-เอ	3
125	-เห	1
126	-เจ	1
127	-เด	1
128	-เต	1
129	-แต	2
130	-แม	1
131	-แล	2
132	-แป	1
133	-ไต	1
134	-ใน	3
135	-ให	2
136	-ได	4
137	-ไม	4
138	-ีน	2
139	-ีย	1
140	-ีห	1
141	-ีม	1
142	-ีท	1
143	-ีจ	1
144	-ีา	2
145	-ีจ	1
146	-ียู	3
147	-ีร	1
148	-ีเ	3
149	-ีด	2
150	-ีน	2

ลำดับ	เทอม	ความถี่
151	-ีไ	2
152	-ีอ	2
153	-ีะ	1
154	-ีฉ	1
155	-ีส	1
156	-ีม	1
157	-ีแ	1
158	-ีเ	3
159	-ีไ	1
160	-ีน	1
161	-ีย	1

ตารางที่ 3.2 แสดงตัวอย่างพจนานุกรมขนาด n = 2 ที่ค่า cutoff = 2

ลำดับ	เทอม	ความถี่
1	-.	2
2	-ก็	3
3	-ขา	5
4	-คิ	2
5	-คน	3
6	-คว	2
7	-งเ	2
8	-จะ	3
9	-ฉั	6
10	-ต้	4
11	-ตี	3
12	-ต่	2
13	-ตั้	3
14	-ที	5
15	-ท่า	2
16	-นเ	3
17	-นท	3
18	-นก	2
19	-ปล	4
20	-ม้	2

ลำดับ	เทอม	ความถี่
21	-ม่	4
22	-มไ	2
23	-ยน	3
24	-ร้	3
25	-รา	2
26	-ร็	2
27	-ล็	3
28	-ละ	2
29	-ลง	2
30	-วา	2
31	-วเ	2
32	-ห้	2
33	-อง	6
34	-ะเ	3
35	-ะฉ	2
36	-ัก	3
37	-ัน	7
38	-ัว	3
39	-าค	2
40	-าเ	3

ลำดับ	เทอม	ความถี่
41	-าม	2
42	-ิด	2
43	-ี	8
44	-ึ	3
45	-เป	4
46	-เบ	5
47	-เร	5
48	-เอ	3
49	-แต	2
50	-แล	2
51	-ใน	3
52	-ให	2
53	-ได	4
54	-ไม	4
55	-ีน	2
56	-ีา	2
57	-ียู	3
58	-ีเ	3
59	-ีด	2
60	-ีน	2

ลำดับ	เทอม	ความถี่
61	-ีไ	2
62	-ีอ	2
63	-ีเ	3

ตารางที่ 3.3 แสดงตัวอย่างพจนานุกรมขนาด n = 2 ที่ค่า cutoff = 3

ลำดับ	เทรม	ความถี่
1	-ก็	3
2	-ขา	5
3	-คน	3
4	-จะ	3
5	-ฉั	6
6	-ด้	4
7	-ดี	3
8	-ต้	3
9	-ที	5
10	-นเ	3
11	-นท	3
12	-ปล	4
13	-ม่	4
14	-ยน	3
15	-ร้	3
16	-ลี	3
17	-อง	6
18	-ะเ	3
19	้ก	3
20	้น	7
21	้ว	3
22	-าเ	3
23	่	8
24	้	3
25	-เป	4
26	-เข	5
27	-เร	5
28	-เอ	3
29	-ใน	3
30	-ได	4
31	-ไม	4
32	่ย	3
33	่เ	3
34	้เ	3

3.1.1 เอ็นแกรมแบบปิด

จากพจนานุกรมข้างต้นจะสังเกตเห็นได้ว่า ข้อมูลส่วนหนึ่งเกิดจากการซ้อนหรือเหลื่อมกัน ในลักษณะแกรมหนึ่งเป็นแกรมย่อยของแกรมอื่นๆ ยิ่งมีขนาดเอ็นใหญ่ก็จะเกิดการเหลื่อมกันของแกรมมากด้วย เหตุนี้จึงเป็นที่มาของการหาแกรมแบบปิดซึ่งเอ็นแกรมแบบปิด $S = \{x_1 x_2 x_3 ... x_N\}$ ของ T คือ เอ็นแกรมที่ไม่มีเอ็นแกรมที่ใหญ่กว่า ที่มีความถี่อย่างน้อยเท่ากันครอบคลุมมัน

ตัวอย่าง เช่น "การปก"มีขนาด n = 5 โดยที่ตัวแกรมเองครองคลุมตั้งแต่แกรมขนาด n = 2 ถึง n = 4 นั้นคือ "กา" , "าร", "รป", "ปก" ซึ่งที่กล่าวมานี้แกรมขนาด n = 2 สำหรับแกรมขนาด n = 3 จะประกอบด้วยชุดของแกรมดังนี้ "การ" , "ารป", "รปก" และสำหรับแกรมขนาด n = 4 จะประกอบด้วยชุดของแกรมดังนี้ "การป" , "ารปก" เป็นต้น

จะเห็นได้ว่าเอ็นแกรมแบบปิดเพียงตัวเดียวสามารถคลอบคลุมจำนวนเอ็นแกรมได้หลายตัว ส่งผลโดยตรงกับจำนวนของเอ็นแกรมเพื่อนำไปใช้สร้างพจนานุกรมในการเตรียมข้อมูล ดังนั้นจึงมีการขจัดปัญหาดังกล่าวด้วยกระบวนการหาเอ็นแกรมแบบปิด ซึ่งมีขั้นตอนดังต่อไปนี้

1. เริ่มต้นกำหนดให้ ค่า n = 2 ,cutoff = 2 และ last_gram = n เมื่อ n คือ จำนวนแกรม, cutoff คือ เกณฑ์ขั้นต่ำในการตัดออกและ last_gram คือ จำนวนแกรมตัวท้ายสุด
2. นำข้อความจากแฟ้มต้นกำเนิดเข้ากระบวนการหาความถี่ขนาด n แล้วสร้างเป็นพจนานุกรมเก็บเอ็นแกรมที่ได้

3. จากข้อ 2 ตัดแกรมที่มีความถี่ไม่ผ่านเกณฑ์ขั้นต่ำที่กำหนดไว้ใน cutoff ออกจากพจนานุกรม
4. ตรวจสอบจำนวนของเอ็นแกรมในพจนานุกรม ถ้าเท่ากับศูนย์ ไปข้อ 6
5. เพิ่มค่าให้ n หนึ่งค่า (n = n + 1)และ last_gram = n หลังจากนั้นไปข้อ 2
6. กำหนดให้ ค่า n = 2
7. นำเอ็นแกรมในแต่ละค่าในพจนานุกรมขนาดแกรม n+1 ค้นหาการเป็นสายอักขระย่อยในพจนานุกรมขนาดแกรม n
8. ถ้ามีการเป็นสายอักขระย่อย ให้นำความถี่ของพจนานุกรมขนาดแกรม n+1 หักลบกับค่าความถี่ในพจนานุกรม n
9. นำเอ็นแกรมในแต่ละค่าในพจนานุกรม n ที่มีความถี่ผ่านเกณฑ์ตาม cutoff และเก็บลงพจนานุกรมผลลัพธ์
10. เพิ่มค่าให้ n หนึ่งค่า (n = n + 1)
11. ไปข้อ 7 จนกระทั่ง n = last_gram สิ้นสุดกระบวนการ

จากกระบวนการหาแกรมแบบปิดข้างต้นสามารถแสดงตัวอย่างการจัดการกับความซ้อนกันของข้อมูลโดยหาเอ็นแกรมแบบปิดโดยสามารถพิจารณาจากตัวอย่างข้อความในตัวอย่างที่ 3.1 เมื่อนำเข้ากระบวนการหาเอ็นแกรมแบบปิดที่ค่าตัดออกมีค่าเท่ากับสาม จะได้พจานานุกรมดังตารางที่ 3.3 เพื่อขจัดแกรมในพจนานุกรมขนาด n แกรมซึ่งอาจจะเป็นแกรมย่อยในพจนานุกรมขนาด (n +1) แกรม โดยใช้วิธีการค้นหาการเป็นสายอักขระย่อยเช่นในตัวอย่างนั่นคือ นำข้อมูลแต่ละตัวในพจนานุกรมเอ็นแกรมขนาด n=2 ที่ผ่านการ cutoff =3 แล้ว นำไปค้นหาในตารางที่ 3.4 ซึ่งเป็นพจนานุกรมขนาดเอ็นแกรมขนาดใหญ่กว่าเพื่อขจัดแกรมในตารางที่ 3.3 ซึ่งแกรมในตาราง 3.4 มีแกรมที่ครอบคลุมอยู่แล้วออกไป ผลลัพธ์จากขั้นตอนนี้ได้พจนานุกรมแกรมแบบปิดขนาด n=2 ที่ค่า cutoff = 3 ดังแสดงในตารางที่ 3.5

ตารางที่ 3.4 แสดงตัวอย่างพจนานุกรมเอ็นแกรมขนาด n = 3 และค่า cutoff = 3

ลำดับ	เทรม	ความถี่
1	-ฉัน	6
2	-ตัว	3
3	-ที่	5
4	-ปลี	3
5	-รัก	3
6	-ลี่	3
7	-ย	3
8	-เปล	3
9	-เขา	5
10	-เอง	3
11	-ได้	4
12	-ไม่	4
13	-ยน	3

ตารางที่ 3.5 แสดงตัวอย่างพจนานุกรมเอ็นแกรมแบบปิดขนาด n = 2 และค่า cutoff = 3

2gram	ความถี่
-ก็	3
-คน	3
-จะ	3
-ดี	3
-นเ	3
-นท	3
-อง	3
-ะเ	3
-าเ	3
-	3
-เร	5
-ใน	3
-เ	3
-เ	3

ในตารางที่ 3.5 เป็นตารางที่สิ้นสุดในการหาเอ็นแกรมแบบปิดที่ขนาด n=2 ที่ค่า cutoff=3 เรียบร้อยแล้ว แต่สำหรับตารางที่ 3.4 ยังต้องมีการทำในรูปแบบเดิมอีกรอบโดยเอ็นแกรมที่ใหญ่กว่านั้นคือตารางที่ 3.6 ผลลัพธ์จากกระบวนการจะได้พจนานุกรมแกรมแบบปิดขนาด n=3 ที่ค่า cutoff = 3 ดังแสดงในตารางที่ 3.7 และกระบวนการจะกระทำวนต่อไปเลื่อยๆ จนสิ้นสุดที่แกรมสุดท้ายดังตารางที่ 3.8 เพราะถัดจากพจนานุกรมขนาด n=7 ที่ cutoff=2 มีค่าเป็นว่าง หรือข้อมูลในพจนานุกรมมีความถี่ต่ำกว่าค่า cutoff ที่กำหนด นั้นเองจึงเป็นที่มาของการสิ้นสุดกระบวนการหาเอ็นแกรมแบบปิด

ตารางที่ 3.6 แสดงตัวอย่างพจนานุกรมเอ็นแกรมขนาด n = 4 และค่า cutoff = 3

ลำดับ	เทอม	ความถี่
1	-ปลี	3
2	-ลีย	3
3	-ียน	3
4	-เปลี	3

ตารางที่ 3.7 แสดงตัวอย่างพจนานุกรมเอ็นแกรมแบบปิดขนาด n = 3 และค่า cutoff = 3

3gram	ความถี่
-ฉัน	6
-ตัว	3
-ที่	5
-รัก	3
-เขา	5
-เอง	3
-ได้	4

ตารางที่ 3.8 แสดงตัวอย่างพจนานุกรมเอ็นแกรมแบบปิดขนาด n = 7 และค่า cutoff = 2

7gram	ความถี่
-เปลียน	3

จากกระบวนการตั้งแต่การหาสถิติการเกิดของแกรมในขนาดเอ็นต่างๆ กัน จนกระทั่งเข้ากระบวนการตัดเอ็นแกรมที่ใดๆ ที่ไม่ผ่านเกณฑ์ขั้นต่ำเพื่อลดต้นทุนของการสร้างพจนานุกรม จนสิ้นสุดที่กระบวนการหาแกรมแบบปิด ผลลัพท์ของจำนวนเอ็นแกรมแบบปิดทั้งหมดที่ได้ค่อนข้างจะครอบคลุมในข้อความของแฟ้มต้นกำเนิดโดยส่วนใหญ่ก็เพื่อใช้สถิติเหล่านี้เป็นฐานข้อมูลตัวอย่างในการนำไปสร้างเป็นสูตรสำเร็จการประมาณค่าขนาดเอ็นแกรมแบบปิด ซึ่งจะกล่าวในหัวข้อถัดไป

3.2 การประมาณค่าขนาดเอ็นแกรมแบบปิดโดยใช้กฎของฮีฟ

จากกระบวนการในหัวข้อที่ 3.1 จะได้ค่าสถิติจำนวนข้อมูลของเอ็นแกรมแบบปิด ซึ่งในงานวิจัยนี้ได้กำหนดช่วงการเก็บข้อมูลโดยใช้ค่า cutoff ตั้งแต่ 2 ถึง 21 เพื่อศึกษาลักษณะความสัมพันธ์ของขนาดแฟ้มข้อมูลจะมีอิทธิพลอย่างไรกับจำนวนเอ็นแกรมแบบปิดที่ได้

เนื่องจากในงานวิจัยนี้ได้พัฒนาเกี่ยวกับการประมาณ ดังนั้นเพื่อให้มีข้อมูลพื้นฐานในวิเคราะห์ถึงการประมาณที่มีกระบวนการอยู่บนพื้นฐานของการประมวลผลข้อความ จึงเลือกกฎของฮีฟที่ได้รับการยอมรับสำหรับการประมาณในการอ้างอิงและเปรียบเทียบกับการประมาณของสูตรสำเร็จที่ได้พัฒนาขึ้น ในการหาค่าสัมประสิทธิ์ของแต่ละ cutoff โดยการใช้กฎของฮีฟมีดังสมการต่อไปนี้

$$V_{r(n)} = Kn^{B}$$

เมื่อ V = จำนวนเอ็นแกรมแบบปิดที่ได้จากการประมาณ

n = ขนาดของแฟ้มข้อมูล

K, B = ตัวแปรอิสระ

ในการหาค่าสัมประสิทธิ์ของตัวแปรอิสระ K, B นั้น ได้จากค่าสถิติของขนาดเอ็นแกรมแบบปิด ซึ่งตัวอย่างการหาในงานวิจัยนี้จะกล่าวในบทถัดไป

3.3 การประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีการถดถอย

การวิเคราะห์การถดถอย เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร [9] โดยมีตัวแบบของการวิเคราะห์การถดถอยเพื่อหาความสัมพันธ์ดังกล่าวว่ามีความสัมพันธ์กันแบบใด ซึ่งจะแบ่งออกเป็น 2 แบบ คือ แบบเส้นตรง (Linear) และ แบบไม่เป็นเส้นตรง (Non-linear) โดยในแต่ละแบบสามารถแบ่งย่อยได้เช่น แบบเส้นตรง แบ่งเป็น การวิเคราะห์การถดถอยแบบเชิงเส้นเชิงเดียว (Simple-linear Regression Analysis) และการวิเคราะห์ถดถอยแบบพหุสัมพันธ์แบบเชิงเส้น (Multi-linear Regression Analysis) สำหรับการวิเคราะห์การถดถอยแบบไม่เป็นเส้นตรง สามารถแบ่งย่อยได้หลายแบบเช่นเดียวกัน เช่น การวิเคราะห์ถดถอยแบบพาวเวอร์ (Power Regression Analysis) และการวิเคราะห์ถดถอยแบบเอ็กโปเนนเชียล (Exponential Regression Analysis) เป็นต้น สำหรับตัวแบบการวิเคราะห์การถดถอยแต่ละแบบซึ่งกล่าวถึงในงานวิจัยนี้ มีตัวแบบดังต่อไปนี้

การวิเคราะห์การถดถอยแบบเชิงเส้นเชิงเดียว : $Y = \alpha + \beta X$

การวิเคราะห์ถดถอยแบบพาวเวอร์ : $Y = \alpha X^{\beta}$

การวิเคราะห์ถดถอยแบบเอ็กโปเนนเชียล : $Y = \alpha e^{\beta X}$

ในงานวิจัยนี้จะทำการวิเคราะห์เพื่อหาค่าคงที่สำหรับประมาณค่าขนาดเอ็นแกรมแบบปิด แต่เนื่องจากในงานวิจัยนี้ต้องทำการวิเคราะห์ข้อมูลจำนวนมากจึงใช้โปรแกรมคอมพิวเตอร์เป็นเครื่องมือช่วยในการวิเคราะห์ ซึ่งแสดงรายละเอียดการวิเคราะห์ค่าตัวแปรต่างๆ ที่ใช้สำหรับการพิจารณาเพื่อหาสมการสำหรับการพยากรณ์เอ็นแกรมแบบปิด ซึ่งจะขอกล่าวในบทถัดไป

บทที่ 4
การประเมินผล

4.1 ข้อมูลที่ใช้ในการทดสอบ

ข้อมูลที่ใช้ในการทดสอบสำหรับงานวิจัยนี้ เป็นข้อมูลตัวอย่างที่สุ่มมาจาก วารสาร นิตยสาร จดหมายราชการ รายงาน หนังสือพิมพ์ บทความ ซึ่งรวมรวบมาจากอินเตอร์ เป็นจำนวนทั้งสิ้น 400 แฟ้มข้อความตัวอย่าง ซึ่งแต่ละตัวอย่างมีขนาดที่แตกต่างกัน ข้อมูลที่เก็บรวบรวมเหล่านี้ล้วนเป็นข้อความภาษาไทยที่มีความทันสมัย และใช้กันอยู่ในปัจจุบัน โดยเนื้อหาภายในแต่ละตัวอย่างจะประกอบด้วย อักษรภาษาไทย สัญลักษณ์ ตัวเลข หรืออักษรภาษาอังกฤษปะปนบ้างเล็กน้อย ตามลักษณะการเขียนภาษาไทย ข้อมูลใช้สำหรับการทดสอบเหล่านี้จะถูกแบ่งออกเป็น 2 ชุด คือ ชุดแรกมีจำนวน 300 แฟ้มข้อความ แฟ้มดังกล่าวเหล่านี้ถูกออกแบบไว้สำหรับการทดสอบหาค่าคงที่ ซึ่งนำไปใช้สำหรับการสร้างสูตรสำเร็จเพื่อประมาณค่าขนาดเอ็นแกรมแบบปิด และชุดที่สองมีจำนวน 100 แฟ้มข้อความ ออกแบบไว้สำหรับการทดสอบการประมาณค่าขนาดเอ็นแกรมแบบปิดของสูตรสำเร็จที่ได้จากข้อมูลชุดแรก

4.2 การทดสอบวิธีประมาณค่าขนาดเอ็นแกรมแบบปิดโดยกฎของฮีฟ

งานวิจัยนี้ได้นำวิธีการประมาณค่าด้วยวิธีตามกฎของฮีฟมาใช้ประกอบการประมาณค่าขนาดเอ็นแกรมแบบปิด แม้ว่าวิธีของฮีฟลอล์จะสนับสนุนการประมาณค่าระหว่างจำนวนคำและขนาดของแฟ้มข้อมูล สำหรับข้อความภาษาอังกฤษหรือข้อความที่มีการแบ่งคำได้อย่างชัดเจน แต่เนื่องจากวิธีดังกล่าวนี้เป็นวิธีที่นิยมนำมาใช้สำหรับการประมาณ สำหรับงานนี้จะเป็นการประมาณค่าขนาดเอ็นแกรมแบบปิดซึ่งมุ่งพัฒนาสำหรับข้อความภาษาไทย ซึ่งมีลักษณะของภาษาที่เป็นสายอักขระดังกล่าวมาแล้วในบทที่ 3 สำหรับการทดสอบวิธีประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีฮีฟลอล์นี้ จะนำข้อมูลจากหัวข้อย่อย 4.1 ชุดที่ 1 มาใช้สำหรับการทดสอบเพื่อให้ได้สูตรสำเร็จตามหลักการหาสูตรสำเร็จของวิธีนี้ โดยข้อมูลชุดนี้จะถูกแบ่งออกเป็น 2 ชุดย่อย เพื่อหาค่าสัมประสิทธิ์ (ตัวแปร K และ B) สำหรับแทนในสมการตามกฎของฮีฟ โดยมีกระบวนการในการหาค่าดังตัวอย่างที่ 4.1 และสรุปค่าสัมประสิทธิ์ที่ได้จากข้อมูลแต่ละชุดย่อยแยกตามค่า cutoff ซึ่งได้แสดงดังตารางที่ 4.3

ตัวอย่างที่ 4.1 กำหนดให้ชุดข้อมูลสำหรับการทดสอบชุดหนึ่งขนาด 300 แฟ้มข้อมูลจะแบ่งออกเป็น 2 ชุดๆละเท่าๆ กัน คือ 150 แฟ้มข้อมูล ดังตารางที่ 4.1 และ 4.2 ตามลำดับ (รายละเอียดของแฟ้มข้อมูลสามารถดูได้จาก ภาคผนวก ก) และคำนวณหาค่าคงที่ K และ B ดังข้างล่างนี้

ตารางที่ 4.1 ข้อมูลตัวอย่าง 150 แฟ้มแรก ที่ cutoff = 2

ชื่อแฟ้ม	ขนาดแฟ้ม	จำนวนเอ็นแกรมแบบปิด
001.txt	28,850	4056
002.txt	14,273	1,923
...	...	...
1095.txt	61,440	6,752
ค่าเฉลี่ย	21,383.31	2,369.98

เมื่อ $V = 2{,}369.98$

$n = 21{,}383.31$

ดังนั้น $2{,}369.98 = K(21{,}383.31)^{B}$

$$K = \frac{2{,}369.98}{21{,}383.31^{B}} \qquad (4.1)$$

ตารางที่ 4.2 ข้อมูลตัวอย่าง 150 แฟ้มที่เหลือ ที่ cutoff = 2

ชื่อแฟ้ม	ขนาดแฟ้ม	จำนวนเอ็นแกรมแบบปิด
051.txt	87,177	6,376
052.txt	66,148	4,632
.	.	.
1100.txt	200,883	15,312
ค่าเฉลี่ย	45,302.27	3,881.25

เมื่อ $V = 3{,}881.25$

$n = 45{,}302.27$

ดังนั้น $3{,}881.25 = K(45{,}302.27)^{B}$

$$K = \frac{3{,}881.25}{45{,}302.27^{B}} \qquad (4.2)$$

แก้สมการเพื่อหาค่า B โดยให้ สมการที่ 4.1 = สมการที่ 4.2

$$\frac{2{,}369.98}{21{,}383.31^{B}} = \frac{3{,}881.25}{45{,}302.27^{B}}$$

$$\frac{45{,}302.27^{B}}{21{,}383.31^{B}} = \frac{3{,}881.25}{2{,}369.98}$$

$$2.118581^{B} = 1.6376706$$

$$\text{Log}_{2.118581}(2.118581)^{B} = \text{Log}_{2.118581}(1.6376706)$$

$$B = 0.657046$$

แทนค่า B ในสมการที่ 4.1 เพื่อหาค่า K

$$K = \frac{2{,}369.98}{21{,}383.31^{B}}$$

$$K = \frac{2{,}369.98}{21{,}383.31^{0.657046}}$$

$$K = 3.386004$$

เพราะฉะนั้น ที่ cutoff = 2 จะได้ค่าสัมประสิทธิ์ในการประมาณ คือ

$$K = 3.386004$$

$$B = 0.657046$$

จากการคำนวณโดยใช้กฎของฮีฟสามารถแสดงผลการคำนวณได้ดังตารางที่ 4.3 ซึ่งแสดงค่าสัมประสิทธิ์สำหรับแต่ละ cutoff สำหรับสมการซึ่งได้จากการคำนวณโดยใช้กฎของฮีฟ

ตารางที่ 4.3 สมการที่ได้จากการใช้กฎของฮีฟสำหรับแต่ละ cutoff

Cutoff	**รูปแบบของสมการ**
2	$V = 3.386 \times n^{0.657}$
3	$V = 1.654 \times n^{0.692}$
4	$V = 0.964 \times n^{0.719}$
5	$V = 0.688 \times n^{0.731}$
6	$V = 0.507 \times n^{0.744}$
7	$V = 0.439 \times n^{0.744}$
8	$V = 0.358 \times n^{0.751}$
9	$V = 0.277 \times n^{0.765}$
10	$V = 0.233 \times n^{0.772}$

ตารางที่ 4.3 (ต่อ) สมการที่ได้จากการใช้กฎของฮีฟสำหรับแต่ละ cutoff

Cutoff	รูปแบบของสมการ
11	$V = 0.210 \times n^{0.774}$
12	$V = 0.177 \times n^{0.782}$
13	$V = 0.147 \times n^{0.792}$
14	$V = 0.128 \times n^{0.798}$
15	$V = 0.111 \times n^{0.806}$
16	$V = 0.098 \times n^{0.812}$
17	$V = 0.083 \times n^{0.822}$
18	$V = 0.077 \times n^{0.824}$
19	$V = 0.070 \times n^{0.829}$
20	$V = 0.063 \times n^{0.832}$
21	$V = 0.059 \times n^{0.836}$

4.3 การทดสอบวิธีประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีการถดถอย

ในส่วนนี้จะกล่าวถึงการทดสอบวิธีการประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีการถดถอยซึ่งในงานวิจัยนี้ได้พัฒนาโปรแกรมและนำข้อมูลที่ได้จากหัวข้อ 4.1 ชุดที่ 1 มาทดสอบด้วยโปรแกรมคอมพิวเตอร์ สำหรับการวิเคราะห์หาความสัมพันธ์เพื่อหาสูตรสำเร็จในการประมาณค่าขนาดเอ็นแกรมแบบปิดใดๆ โดยใช้การวิเคราะห์การถดถอย

4.3.1 การพิจารณาเลือกใช้ตัวแบบความสัมพันธ์สำหรับการวิเคราะห์การถดถอย

การวิเคราะห์การถดถอย เป็นการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร ซึ่งมีหลายตัวแบบได้กล่าวมาแล้วในบทที่ 3 สำหรับหลักการในการพิจารณาว่าตัวแปรแต่ละตัวมีความสัมพันธ์กันมากน้อยเพียงใดจะพิจารณาจากค่า R^2 ที่ได้จากการคำนวณ [9] ซึ่งในงานวิจัยนี้ได้นำโปรแกรมคอมพิวเตอร์มาช่วยสำหรับคำนวณค่าความสัมพันธ์ดังกล่าว โดยมีวัตถุประสงค์เพื่อพิจารณาเลือกตัวแบบสมการสำหรับการวิเคราะห์การถดถอย โดยพิจารณาค่า R^2 ถ้ามีค่ามาก หมายความว่าตัวแปรที่ใช้ในการทดลองสัมพันธ์กันมากด้วย สามารถสรุปได้ดังตารางที่ 4.4

ตารางที่ 4.4 ค่าความสัมพันธ์ของตัวแปร ใน cutoff ใดๆ สำหรับแต่ละตัวแบบสมการถดถอย

Cutoff	R			R^2		
	Linear	Power	Exponential	Linear	Power	Exponential
2	0.935	0.985	0.768	0.847	0.971	0.59
3	0.943	0.987	0.767	0.889	0.975	0.588
4	0.948	0.988	0.762	0.898	0.976	0.581
5	0.95	0.988	0.76	0.902	0.977	0.577
6	0.952	0.988	0.754	0.907	0.976	0.568
7	0.954	0.985	0.744	0.91	0.97	0.554
8	0.955	0.983	0.738	0.913	0.967	0.545
9	0.955	0.99	0.784	0.912	0.98	0.614
10	0.956	0.988	0.779	0.914	0.977	0.607
11	0.957	0.987	0.774	0.915	0.974	0.599
12	0.957	0.987	0.772	0.917	0.974	0.597
13	0.958	0.984	0.764	0.918	0.968	0.584
14	0.959	0.983	0.761	0.919	0.965	0.579
15	0.959	0.98	0.754	0.92	0.96	0.569
16	0.959	0.979	0.752	0.92	0.958	0.566
17	0.96	0.978	0.749	0.921	0.956	0.561
18	0.96	0.975	0.745	0.922	0.951	0.555
19	0.96	0.98	0.754	0.922	0.96	0.568
20	0.961	0.979	0.752	0.924	0.959	0.565
21	0.962	0.976	0.747	0.925	0.953	0.558

จากตารางที่ 4.4 เมื่อพิจารณาค่า R และ R^2 พบว่าค่าดังกล่าวในตัวแบบสมการถดถอยแบบพาวเวอร์ให้ค่าความสัมพันธ์สูงสุด ดังนั้นในงานวิจัยนี้จึงเลือกใช้ตัวแบบสมการถดถอยแบบพาวเวอร์ในการประมาณค่าขนาดเอ็นแกรมแบบปิด

4.3.2 ตัวแบบสมการเพื่อประมาณค่าขนาดเอ็นแกรมแบบปิด

จากหัวข้อที่ผ่านมา งานวิจัยนี้ได้นำข้อมูลชุดที่ 1 จากหัวข้อ 4.1 มาทดลองเพื่อหาค่าสัมประสิทธิ์สำหรับการวิเคราะห์การถดถอยแบบพาวเวอร์ในการประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยโปรแกรมคอมพิวเตอร์ได้ดังตารางที่ 4.5

ตารางที่ 4.5 ค่าสัมประสิทธิ์ของตัวแปรอิสระ α และ β สำหรับแต่ละ cutoff

Cutoff	รูปแบบของสมการ
2	$Y = 0.256 \times X^{0.910}$
3	$Y = 0.147 \times X^{0.930}$
4	$Y = 0.092 \times X^{0.949}$
5	$Y = 0.063 \times X^{0.965}$
6	$Y = 0.042 \times X^{0.987}$
7	$Y = 0.030 \times X^{1.006}$
8	$Y = 0.023 \times X^{1.02}$
9	$Y = 0.016 \times X^{1.041}$
10	$Y = 0.012 \times X^{1.061}$
11	$Y = 0.008 \times X^{1.082}$
12	$Y = 0.006 \times X^{1.101}$
13	$Y = 0.005 \times X^{1.126}$
14	$Y = 0.003 \times X^{1.148}$
15	$Y = 0.002 \times X^{1.169}$
16	$Y = 0.002 \times X^{1.192}$
17	$Y = 0.001 \times X^{1.212}$
18	$Y = 0.001 \times X^{1.232}$
19	$Y = 0.001 \times X^{1.253}$
20	$Y = 0.001 \times X^{1.278}$
21	$Y = 0.0003 \times X^{1.299}$

4.4 ทดสอบค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์

ในงานวิจัยนี้ใช้ค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์ (Mean Absolute Percentage Error-MAPE) สำหรับวัดความแม่นยำของการพยากรณ์ระหว่างสมการที่ได้จากการทดลองด้วยวิธีตามกฎของฮีฟ และ สมการที่ได้จากการวิเคราะห์การถดถอยแบบพาวเวอร์ โดยมีสูตรในการวัดความผิดพลาดของการพยากรณ์ดังนี้

$$MAPE = \frac{\sum(|Y_0 - Y_c| / Y_0) \times 100}{N}$$

เมื่อ Y_0 = ค่าขนาดเอ็นแกรมแบบปิดที่ได้จากข้อมูลต้นกำเนิด

Y_c = ค่าขนาดเอ็นแกรมแบบปิดที่ได้จากการประมาณโดยใช้สูตร

N = ขนาดของแฟ้มข้อมูล

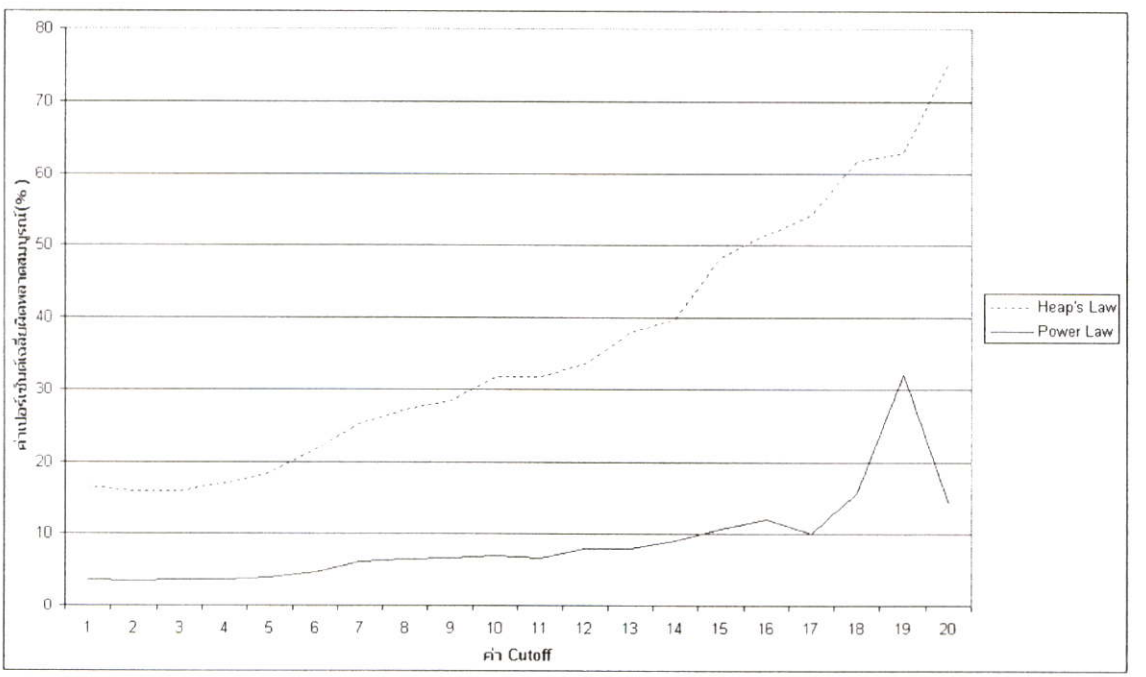
สมการดังกล่าวนี้จะให้ค่าความผิดพลาดของการพยากรณ์ ดังนั้นถ้าค่า *MAPE* มีค่าน้อย แสดงว่าสมการดังกล่าวมีความแม่นยำในการประมาณมาก ในงานวิจัยนี้ได้นำวิธีการทดสอบดังกล่าวมาวัดความแม่นยำในการพยากรณ์ จากการคำนวณค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์สำหรับการใช้สมการทั้งสองสามารถแสดงได้ดังตารางที่ 4.6 และเปรียบเทียบให้เห็นความแตกต่างของค่าดังกล่าวดังกราฟในรูปที่ 4.1

ตารางที่ 4.6 ค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์จากสมการที่ได้จากการทดลองแต่ละ cutoff

Cutoff	MAPE(%)	
	Heaps' Law	Power Regression
2	16.463	3.623
3	15.919	3.435
4	15.800	3.530
5	17.077	3.614
6	18.416	3.860
7	21.696	4.672
8	25.251	6.082
9	27.139	6.436
10	28.379	6.671
11	31.670	6.944

ตารางที่ 4.6 (ต่อ)

Cutoff	MAPE(%)	
	Heaps' Law	Power Regression
12	31.710	6.648
13	33.568	7.927
14	37.863	7.997
15	39.854	9.090
16	48.408	10.645
17	51.281	11.967
18	54.281	10.057
19	61.731	15.757
20	62.839	32.052
21	75.500	14.237



รูปที่ 4.1 กราฟเปรียบเทียบค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์

จากตารางที่ 4.6 และรูปที่ 4.1 แสดงเปอร์เซ็นต์ความผิดพลาดจากการพยากรณ์จากสมการกฎของฮีฟ ซึ่งให้ค่าความผิดพลาดที่สูงกว่านั่นหมายความว่ามีความถูกต้องของการพยากรณ์น้อยกว่าเมื่อเปรียบเทียบกับการใช้สมการที่ได้จากการวิเคราะห์การถดถอยแบบพาวเวอร์

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุป

งานวิจัยนี้เสนอสูตรสำเร็จการประมาณค่าขนาดเอ็นแกรมแบบปิด สำหรับข้อความภาษาไทย เพื่อเป็นข้อมูลสำหรับการกระบวนการเตรียมข้อมูลเพื่อการบีบอัดหรือประโยชน์อื่นๆ ที่เกี่ยวข้องกับการใช้เอ็นแกรมสำหรับข้อความภาษาไทย เนื่องจากภาษาไทยเป็นภาษาที่มีลักษณะข้อความเป็นสายอักขระ ไม่มีสัญลักษณ์คั่นแสดงการสิ้นสุดของคำเหมือนกับภาษาอังกฤษ ด้วยความสำคัญนี้จึงได้พัฒนาสร้างสูตรสำเร็จสำหรับการประมาณค่าขนาดเอ็นแกรมแบบปิด โดยในกระบวนการพัฒนาจะเริ่มจากการสุ่มข้อมูลจากอินเตอร์เน็ตที่มีลักษณะเป็นข้อความภาษาไทย จำนวน 400 แฟ้มข้อความ ซึ่งมีขนาดแตกต่างกันไป โดยแฟ้มข้อความดังกล่าวจะถูกแบ่งออกเป็น 2 ชุดคือ ชุดที่ใช้สำหรับการทดสอบหาความสัมพันธ์ระหว่างขนาดของแฟ้มข้อความและขนาดเอ็นแกรมแบบปิดเพื่อพัฒนาเป็นสูตรสำเร็จที่ใช้สำหรับพยากรณ์ขนาดเอ็นแกรมสำหรับขนาดแฟ้มข้อความใดๆ จำนวน 300 แฟ้มข้อความ และอีกชุดข้อมูลจำนวน 100 แฟ้มข้อความเพื่อใช้ทดสอบความถูกต้องของการพยากรณ์

สำหรับการทดลองเพื่อให้ได้สูตรสำเร็จดังกล่าวโดยนำค่าสถิติการเกิดเอ็นแกรมแบบปิดที่หาได้จากแฟ้มข้อความสำหรับสร้างสูตรแยกตามค่า cutoff เป็นข้อมูลพื้นฐานสำหรับการสร้างสูตรตามกฎของฮีฟและวิธีการวิเคราะห์การถดถอย ซึ่งในวิธีการวิเคราะห์การถดถอยใช้ 3 ตัวแบบคือ ตัวแบบเชิงเส้นอย่างง่าย ตัวแบบเอ็กโปเนนเชียลและตัวแบบพาวเวอร์ เพื่อหาความเหมาะในการถดถอยของตัวแบบจึงใช้โปรแกรมคอมพิวเตอร์เข้ามาช่วยวิเคราะห์ จากค่าสัมประสิทธิ์การตัดสินใจที่ได้บ่งชี้ตัวแบบพาวเวอร์มีเปอร์เซ็นต์ความสัมพันธ์ของขนาดเอกสารกับขนาดเอ็นแกรมแบบปิดสูงกว่าตัวแบบอื่นๆ นั้นแสดงถึงความเหมาะสมในการนำไปสำหรับการพยากรณ์

จากการทดสอบค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์ เพื่อวัดความแม่นยำในการประมาณค่าขนาดเอ็นแกรมแบบปิดทั้งสองวิธีที่เลือกใช้ในงานวิจัยนี้ ผลจากการทดสอบค่าดังกล่าวพบว่าสูตรการประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีการวิเคราะห์การถดถอยแบบพาวเวอร์ให้ค่าเปอร์เซ็นต์เฉลี่ยผิดพลาดสมบูรณ์ต่ำกว่าสูตรการประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีตามกฎของฮีฟ แสดงให้เห็นว่าสูตรการประมาณค่าขนาดเอ็นแกรมแบบปิดด้วยวิธีการวิเคราะห์การถดถอยแบบพาวเวอร์ให้การพยากรณ์ที่แม่นยำกว่านั้นเอง

5.2 ข้อเสนอแนะ

สูตรการคำนวณค่าขนาดเอ็นแกรมแบบปิดจากขนาดของแฟ้มข้อความที่ได้นี้ เป็นสูตรที่สามารถนำไปใช้งานในการสร้างพจนานุกรมเอ็นแกรมแบบปิดสำหรับภาษาใดๆ ที่มีลักษณะการเขียนคำโดยไม่มีช่องว่างแบ่งระหว่างคำ นอกจากนี้สูตรนี้ยังใช้ได้กับสายอักขระต่อเนื่องแบบใดๆ อย่างไรก็ตาม วิธีการวิเคราะห์การถดถอยนี้หากต้องการนำไปใช้ในการประมาณจำนวนคำเพื่อสร้างพจนานุกรมแทนการใช้กฎของฮีฟยังต้องมีการศึกษาในรายละเอียดต่อไป

เอกสารอ้างอิง

[1] Abel J., Teahan W. "Universal Text Preprocessing for Data Compression." **IEEE Trans. On Comput.,** 2005.

[2] Cavnar, William B. and Trenkle, John M. "N-Gram-Based Text Categorization." **Proc. SDAIR-94, 3th Annual Symposium On Document Analysis and Information Retrieval.,** 1994.

[3] C. Y. Suen and et. al. "Categorizing Document Images into Script and Language classes." **International Conference on Advances in Pattern Recognition, Plymouth, England.,** 1998. pp. 297-306.

[4] E. Dellandrea, P. Makris, N. Vincent. "Wavelets and Zipf Law for Audio Signal Analysis." **7th International Symposium on Signal Processing and its Applications, Paris, France.,** vol. 2, 2003. pp. 483-486.

[5] E. Dellandrea, P. Makris, M. Boiron, N. Vincent. "A medical acoustic signal analysis method based on Zipf law." **International Conference on Digital Signal Processing, Santorini, Grece.,** vol. 2, 2002. pp. 615-618.

[6] Kutner Michael H., Nochtsheim Christopher J. **Applied linear regression models.** 4th ED. Singapore : MaGrawHill. 2004.

[7] Peng F. and Schuurmans D. "Combining Naive Bayes and n-Gram Language Models for Text Classication." **25th European Conference on Information Retrieval Research.,** 2003.

[8] R.E. Krichevskii, M.P. Scharova. "Shannon-Hartley Entropy Ratio under Zipf Law." **Proc. IEEE-IMS Workshop : Information Theory and Statistics.,** 1994.

[9] Sen Ashish K., Srivastava Muni. **Regression Analysis : Theory methods, and applications**. New York : Springer-Verlay., 1990.

[10] Somlertlamvanich V., Potipiti T. and Charoenporn T. "Automatic Corpus-Based Thai Word Extraction with the C4.5 Machine Learning Algorithm." **The Proceedings of the 18th International Conference on Computational Linguistics.,** 2000. pp. 802-807.

เอกสารอ้างอิง (ต่อ)

[11] Thanadkran K., Sornlertlamvanich V. and Potipiti T. **"Intelligent key prediction by n-grams and error-correction rules."** [Online]. Available : http://www.tcllab.org/virach/paper/virach/iccpol2001-kp.pdf. 2001.

[12] van Leijenhorst D.C., van der Weide Th.P. "A formal derivation of Heaps Law." **International Journal of Information Sciences.,** 2004.

ภาคผนวก ก

ตารางข้อมูลตัวอย่าง

ตารางที่ ก.1 แสดงรายละเอียดข้อมูลตัวอย่างที่ใช้สร้างสมการ

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=02	Cutoff=03	Cutoff=04	Cutoff=05	Cutoff=06	Cutoff=07	Cutoff=08	Cutoff=09	Cutoff=10	Cutoff=11
1	001.txt	28,950	4,056	2,716	2,084	1,678	1,404	1,213	1,044	955	858	757
2	002.txt	14,273	1,923	1,404	1,046	808	675	594	520	461	405	364
3	003.txt	14,111	1,798	1,297	973	789	652	554	478	417	390	354
4	004.txt	25,522	3,048	2,153	1,610	1,309	1,105	961	850	781	690	623
5	005.txt	11,700	1,477	1,041	785	625	525	471	399	361	317	275
6	006.txt	20,142	2,337	1,611	1,181	990	848	726	630	580	527	492
7	007.txt	23,346	2,838	1,970	1,512	1,242	1,017	889	775	696	620	559
8	008.txt	22,849	2,870	1,997	1,463	1,195	1,004	859	745	679	606	548
9	009.txt	24,437	3,072	2,094	1,599	1,288	1,106	957	834	733	672	618
10	010.txt	28,022	3,673	2,484	1,876	1,506	1,258	1,090	980	873	797	714
11	011.txt	30,427	3,772	2,611	2,008	1,622	1,372	1,195	1,065	945	856	766
12	012.txt	33,402	4,118	2,781	2,122	1,697	1,416	1,251	1,087	954	884	826
13	013.txt	15,480	2,068	1,457	1,132	890	746	639	548	494	429	388
14	014.txt	53,094	6,604	4,474	3,393	2,700	2,252	1,941	1,734	1,561	1,418	1,296
15	015.txt	50,490	6,028	4,120	3,131	2,499	2,117	1,805	1,593	1,422	1,313	1,216
16	016.txt	22,203	2,991	1,992	1,514	1,264	1,056	896	797	695	625	567
17	017.txt	34,715	4,206	2,848	2,205	1,748	1,462	1,303	1,145	1,012	921	830
18	018.txt	24,151	3,059	2,091	1,610	1,346	1,130	947	847	743	679	624
19	019.txt	51,474	6,334	4,292	3,194	2,568	2,184	1,902	1,668	1,509	1,384	1,266
20	020.txt	57,211	7,315	5,002	3,736	3,006	2,511	2,142	1,899	1,719	1,529	1,431
21	021.txt	34,066	4,493	2,969	2,260	1,839	1,565	1,379	1,205	1,049	963	880
22	022.txt	69,764	7,933	5,357	4,094	3,279	2,782	2,376	2,082	1,875	1,743	1,575
23	023.txt	83,481	10,128	6,960	5,221	4,241	3,583	3,112	2,733	2,443	2,191	2,003
24	024.txt	116,454	13,681	9,337	6,932	5,611	4,703	4,065	3,590	3,195	2,861	2,655
25	025.txt	94,278	10,998	7,586	5,652	4,624	3,844	3,337	2,898	2,589	2,374	2,183
26	026.txt	44,712	5,293	3,602	2,703	2,212	1,843	1,585	1,398	1,245	1,115	1,020
27	027.txt	86,285	11,169	7,510	5,718	4,539	3,767	3,242	2,892	2,529	2,293	2,089
28	028.txt	70,363	8,636	5,843	4,379	3,463	2,994	2,566	2,288	2,047	1,861	1,704
29	029.txt	78,815	9,476	6,604	4,962	3,940	3,319	2,886	2,510	2,241	2,025	1,872
30	030.txt	56,774	5,756	3,935	3,014	2,434	2,047	1,798	1,560	1,414	1,261	1,158
31	031.txt	33,886	4,237	2,899	2,168	1,760	1,507	1,327	1,157	1,034	925	836
32	032.txt	22,907	2,948	2,025	1,506	1,225	1,028	882	764	665	597	552
33	033.txt	25,547	3,224	2,216	1,644	1,342	1,140	977	851	776	688	612
34	034.txt	21,078	2,552	1,795	1,362	1,088	919	789	688	606	552	494
35	035.txt	34,043	4,283	2,950	2,223	1,778	1,497	1,281	1,131	1,037	932	853
36	036.txt	43,481	5,239	3,670	2,762	2,219	1,839	1,598	1,402	1,264	1,147	1,068
37	037.txt	38,615	1,672	1,341	1,180	968	936	871	757	771	697	670
38	038.txt	59,280	3,183	2,258	1,955	1,621	1,393	1,414	1,298	1,195	1,107	1,036
39	039.txt	33,992	4,203	2,935	2,171	1,808	1,516	1,318	1,164	1,032	935	856
40	040.txt	29,984	4,049	2,750	2,045	1,638	1,337	1,143	1,042	926	846	753
41	041.txt	44,963	3,350	2,225	2,000	1,669	1,412	1,357	1,187	1,062	969	891
42	042.txt	49,947	1,809	1,419	1,351	1,123	1,005	1,089	969	915	831	765
43	043.txt	63,123	7,345	5,365	4,062	3,284	2,763	2,408	2,125	1,917	1,737	1,580
44	044.txt	37,008	4,515	3,110	2,355	1,924	1,614	1,404	1,249	1,132	1,033	937
45	045.txt	46,891	5,562	3,799	2,880	2,330	1,943	1,701	1,494	1,339	1,220	1,101
46	046.txt	46,919	4,240	3,038	2,353	1,887	1,633	1,439	1,263	1,128	1,031	936
47	047.txt	68,262	6,778	4,730	3,565	2,961	2,520	2,178	1,935	1,740	1,582	1,440
48	048.txt	43,711	3,678	2,727	2,111	1,673	1,455	1,279	1,136	1,017	928	858
49	049.txt	65,965	4,648	3,377	2,591	2,130	1,798	1,612	1,418	1,306	1,200	1,116
50	050.txt	61,866	5,393	3,775	2,848	2,303	1,926	1,682	1,493	1,326	1,203	1,114
51	051.txt	87,177	6,378	4,662	3,673	3,053	2,628	2,275	1,985	1,777	1,631	1,494
52	052.txt	66,146	4,632	3,351	2,580	2,223	1,882	1,692	1,490	1,426	1,312	1,211
53	053.txt	78,294	7,282	5,079	3,889	3,177	2,706	2,384	2,149	1,930	1,748	1,585
54	054.txt	54,082	5,343	3,710	2,785	2,229	1,903	1,683	1,472	1,287	1,183	1,100
55	055.txt	68,662	6,573	4,622	3,512	2,878	2,437	2,074	1,940	1,665	1,537	1,400
56	056.txt	162,564	16,525	11,083	8,466	6,810	5,753	4,979	4,351	3,885	3,525	3,243
57	057.txt	59,059	2,622	2,074	1,866	1,569	1,395	1,266	1,142	1,028	979	913
58	058.txt	52,125	6,265	4,347	3,320	2,664	2,236	1,938	1,710	1,548	1,377	1,272
59	059.txt	44,156	5,793	3,972	3,026	2,433	2,030	1,745	1,513	1,327	1,205	1,098
60	060.txt	60,590	7,472	5,143	3,944	3,151	2,651	2,271	2,000	1,779	1,609	1,474
61	061.txt	190,072	18,897	13,155	10,105	8,271	7,128	6,186	5,461	4,947	4,473	4,160
62	062.txt	85,283	5,760	4,129	3,354	2,799	2,376	2,092	1,856	1,673	1,521	1,448
63	063.txt	99,959	9,297	6,489	4,898	4,001	3,345	2,930	2,586	2,332	2,131	1,962
64	064.txt	63,177	5,482	4,742	3,663	3,089	2,607	2,248	1,955	1,763	1,578	1,450
65	065.txt	65,562	5,359	3,788	2,983	2,486	2,063	1,813	1,618	1,441	1,327	1,211
66	066.txt	48,552	4,735	3,375	2,575	2,072	1,760	1,516	1,331	1,191	1,086	993
67	067.txt	87,064	5,347	4,007	3,184	2,799	2,407	2,114	1,980	1,706	1,538	1,395
68	068.txt	64,897	5,960	4,210	3,166	2,564	2,156	1,888	1,693	1,544	1,400	1,306
69	069.txt	97,384	8,254	5,745	4,497	3,651	3,099	2,733	2,404	2,188	1,992	1,852
70	070.txt	98,899	6,850	5,073	4,143	3,439	2,946	2,608	2,353	2,132	1,926	1,743
71	071.txt	91,445	6,762	4,887	3,789	3,143	2,672	2,366	2,081	1,870	1,732	1,587
72	072.txt	96,051	7,801	5,463	4,135	3,417	2,909	2,599	2,276	2,044	1,856	1,684
73	073.txt	133,913	8,608	7,911	6,193	5,086	4,375	3,811	3,373	3,048	2,771	2,552
74	074.txt	165,753	11,498	8,394	6,479	5,322	4,530	3,953	3,515	3,210	2,941	2,680
75	075.txt	84,736	8,057	5,499	4,119	3,311	2,812	2,452	2,157	1,935	1,756	1,614
76	076.txt	130,508	7,586	5,600	4,643	3,860	3,339	2,923	2,754	2,464	2,226	2,038
77	077.txt	121,413	11,756	8,193	6,269	5,019	4,278	3,686	3,277	2,980	2,746	2,506
78	078.txt	249,511	22,965	15,689	11,822	9,543	8,066	7,047	6,126	5,493	4,976	4,583
79	079.txt	308,574	25,801	17,775	13,744	11,048	9,257	7,957	6,945	6,138	5,526	5,002
80	080.txt	72,280	2,686	2,624	1,948	1,568	1,318	1,125	990	878	777	698
81	081.txt	81,826	2,702	2,397	2,219	2,180	2,639	2,344	2,244	2,125	1,685	1,691
82	082.txt	73,562	4,372	3,316	4,095	3,281	2,898	2,538	2,225	1,975	1,780	1,626
83	083.txt	317,596	8,276	5,904	4,692	3,934	3,423	3,087	2,787	2,556	2,378	2,221
84	084.txt	251,046	17,765	12,510	9,697	7,858	6,700	5,800	5,144	4,648	4,231	3,861
85	085.txt	143,827	8,316	5,990	4,667	3,817	3,262	2,793	2,485	2,290	2,087	1,932
86	086.txt	49,830	4,478	3,273	2,469	2,042	1,730	1,475	1,318	1,218	1,105	1,027
87	087.txt	76,446	8,164	5,603	4,223	3,364	2,851	2,462	2,180	1,953	1,776	1,597
88	088.txt	41,478	3,684	2,561	2,014	1,630	1,414	1,242	1,078	978	904	806
89	089.txt	38,625	3,866	2,703	2,040	1,674	1,414	1,231	1,084	986	896	827
90	090.txt	31,463	3,325	2,359	1,792	1,427	1,211	1,061	947	856	786	709
91	091.txt	47,759	5,341	3,582	2,707	2,228	1,862	1,635	1,433	1,249	1,128	1,033
92	092.txt	44,972	5,364	3,581	2,740	2,167	1,841	1,598	1,419	1,286	1,160	1,067
93	093.txt	65,510	7,729	5,164	3,892	3,064	2,605	2,259	1,998	1,790	1,605	1,448
94	094.txt	113,094	8,651	5,934	4,500	3,580	3,071	2,719	2,427	2,206	1,988	1,805
95	095.txt	109,116	9,767	6,866	5,239	4,249	3,628	3,196	2,827	2,544	2,320	2,115
96	096.txt	143,563	11,375	9,027	6,935	5,701	4,822	4,194	3,709	3,330	3,028	2,787
97	097.txt	120,354	10,415	6,814	5,225	4,249	3,579	3,083	2,732	2,487	2,229	2,033
98	098.txt	115,605	11,132	7,442	5,675	4,649	3,798	3,313	2,923	2,674	2,409	2,210
99	099.txt	128,773	13,405	9,207	7,018	5,736	4,787	4,155	3,633	3,316	2,995	2,749
100	100.txt	249,771	21,435	17,348	13,211	10,826	9,051	7,790	6,817	6,164	5,570	5,098

ตารางที่ ก.1 (ต่อ)

ลำดับ	เพลง	ขนาด(byte)	Cutoff Values									
			Cutoff=12	Cutoff=13	Cutoff=14	Cutoff=15	Cutoff=16	Cutoff=17	Cutoff=18	Cutoff=19	Cutoff=20	Cutoff=21
1	001.txt	28,850	698	620	583	540	516	484	460	441	410	384
2	002.txt	14,273	331	309	282	269	256	243	220	197	182	169
3	003.txt	14,111	298	266	248	229	213	197	183	171	159	152
4	004.txt	25,522	572	535	496	455	416	388	367	356	342	324
5	005.txt	11,700	239	217	198	188	175	163	155	148	139	123
6	006.txt	20,142	437	405	369	344	327	307	288	264	251	240
7	007.txt	23,346	496	458	423	389	368	337	320	313	298	279
8	008.txt	22,849	513	465	440	399	389	360	335	320	311	294
9	009.txt	24,437	548	502	466	440	407	382	359	342	310	293
10	010.txt	28,022	661	606	580	524	471	447	414	416	387	372
11	011.txt	30,427	695	627	582	535	500	475	443	426	404	377
12	012.txt	33,402	782	706	642	603	571	534	506	484	457	436
13	013.txt	15,480	353	319	299	279	264	251	237	222	205	191
14	014.txt	53,094	1,192	1,091	1,017	956	899	847	806	751	729	709
15	015.txt	50,490	1,132	1,034	961	897	852	801	763	731	696	664
16	016.txt	22,203	504	460	440	406	380	358	340	316	301	283
17	017.txt	34,715	775	716	657	632	588	562	535	518	494	468
18	018.txt	24,151	573	522	487	467	449	420	398	381	345	334
19	019.txt	51,474	1,174	1,065	988	922	868	817	760	729	688	660
20	020.txt	57,211	1,313	1,216	1,136	1,066	1,005	936	907	867	830	790
21	021.txt	34,086	810	755	699	621	598	555	526	507	472	453
22	022.txt	69,764	1,487	1,386	1,313	1,216	1,135	1,058	1,017	966	922	886
23	023.txt	83,481	1,843	1,727	1,599	1,507	1,423	1,345	1,282	1,222	1,153	1,111
24	024.txt	116,454	2,416	2,262	2,130	1,969	1,837	1,733	1,652	1,574	1,519	1,447
25	025.txt	94,278	2,005	1,860	1,726	1,625	1,542	1,455	1,378	1,319	1,254	1,206
26	026.txt	44,712	981	910	841	782	723	683	639	606	585	554
27	027.txt	86,265	1,919	1,793	1,687	1,575	1,493	1,397	1,309	1,241	1,179	1,124
28	028.txt	70,363	1,573	1,463	1,378	1,275	1,208	1,151	1,091	1,039	989	956
29	029.txt	76,815	1,743	1,636	1,523	1,415	1,336	1,284	1,193	1,134	1,083	1,034
30	030.txt	56,774	1,089	993	924	867	804	760	717	671	633	611
31	031.txt	33,866	787	695	646	604	563	529	494	470	454	432
32	032.txt	22,907	517	464	430	400	378	352	325	307	287	282
33	033.txt	25,547	557	507	466	436	404	375	359	344	320	308
34	034.txt	21,078	455	425	386	362	334	318	295	283	270	253
35	035.txt	34,043	789	721	662	607	569	533	509	481	454	428
36	036.txt	43,481	972	894	830	781	736	712	677	641	607	576
37	037.txt	39,615	831	581	558	526	494	447	425	410	400	392
38	038.txt	59,280	965	887	839	790	761	717	700	656	610	587
39	039.txt	33,992	777	709	651	606	569	541	515	484	457	440
40	040.txt	29,964	686	639	586	545	510	479	454	423	393	373
41	041.txt	44,963	826	781	738	691	660	625	605	577	542	520
42	042.txt	49,947	711	673	647	610	576	545	526	506	497	473
43	043.txt	83,123	1,448	1,344	1,254	1,173	1,115	1,042	980	934	877	839
44	044.txt	37,008	874	792	751	701	657	613	584	555	516	477
45	045.txt	46,691	1,025	932	874	811	748	709	664	634	607	587
46	046.txt	46,919	887	825	776	724	685	663	626	604	581	560
47	047.txt	68,262	1,323	1,247	1,150	1,078	1,028	976	930	894	842	793
48	048.txt	43,711	776	717	675	631	609	578	543	523	502	484
49	049.txt	65,956	1,023	937	878	821	781	737	700	666	636	612
50	050.txt	61,866	1,022	953	902	856	785	744	709	677	644	617
51	051.txt	87,177	1,370	1,307	1,249	1,158	1,112	1,052	975	924	878	846
52	052.txt	66,148	1,114	1,065	1,004	932	885	856	813	765	740	702
53	053.txt	76,294	1,487	1,381	1,276	1,214	1,149	1,100	1,050	1,005	971	929
54	054.txt	54,082	992	927	851	788	769	728	689	661	627	606
55	055.txt	68,652	1,316	1,226	1,144	1,050	992	939	886	828	793	753
56	056.txt	162,564	3,006	2,823	2,654	2,505	2,331	2,203	2,098	1,997	1,894	1,832
57	057.txt	59,059	870	831	782	739	700	667	639	623	597	570
58	058.txt	52,125	1,171	1,104	994	926	879	831	797	768	728	691
59	059.txt	44,156	1,032	926	862	795	749	711	664	622	579	546
60	060.txt	60,590	1,354	1,252	1,169	1,089	1,016	947	896	862	818	770
61	061.txt	190,072	3,791	3,505	3,262	3,052	2,891	2,734	2,570	2,430	2,328	2,208
62	062.txt	85,283	1,960	1,269	1,199	1,117	1,056	1,009	940	901	872	835
63	063.txt	99,959	1,837	1,707	1,621	1,520	1,450	1,382	1,299	1,233	1,160	1,109
64	064.txt	63,177	1,311	1,226	1,140	1,072	1,025	980	904	955	805	776
65	065.txt	65,682	1,124	1,036	975	917	866	820	779	736	711	692
66	066.txt	49,552	921	876	835	795	736	705	672	643	609	579
67	067.txt	87,064	1,300	1,234	1,162	1,073	1,044	1,007	960	919	865	827
68	068.txt	64,897	1,213	1,116	1,038	968	911	859	800	783	739	705
69	069.txt	97,384	1,712	1,591	1,497	1,403	1,332	1,258	1,204	1,140	1,078	1,049
70	070.txt	98,699	1,602	1,494	1,417	1,352	1,300	1,251	1,186	1,130	1,094	1,053
71	071.txt	91,445	1,480	1,383	1,289	1,244	1,167	1,111	1,052	1,018	982	926
72	072.txt	96,051	1,548	1,449	1,358	1,289	1,214	1,156	1,103	1,057	997	944
73	073.txt	133,913	2,357	2,202	2,076	1,941	1,837	1,758	1,660	1,575	1,517	1,450
74	074.txt	185,753	2,471	2,328	2,191	2,033	1,959	1,864	1,768	1,690	1,597	1,545
75	075.txt	84,736	1,493	1,411	1,316	1,244	1,166	1,094	1,042	984	921	876
76	076.txt	130,508	1,888	1,789	1,691	1,564	1,510	1,439	1,393	1,339	1,302	1,241
77	077.txt	121,413	2,321	2,160	2,042	1,926	1,829	1,736	1,639	1,560	1,480	1,415
78	078.txt	249,511	4,193	3,890	3,651	3,399	3,228	3,050	2,907	2,768	2,635	2,514
79	079.txt	308,574	4,578	4,240	3,905	3,656	3,417	3,199	2,992	2,849	2,705	2,581
80	080.txt	72,280	629	579	536	508	476	453	422	403	374	361
81	081.txt	81,826	1,581	1,519	1,431	1,333	1,222	1,154	1,106	1,049	982	946
82	082.txt	73,552	1,496	1,392	1,306	1,218	1,150	1,097	1,032	992	939	897
83	083.txt	317,596	2,088	1,970	1,877	1,772	1,680	1,614	1,551	1,491	1,449	1,398
84	084.txt	251,046	3,583	3,375	3,189	2,981	2,803	2,651	2,519	2,402	2,314	2,226
85	085.txt	143,827	1,787	1,653	1,552	1,459	1,400	1,317	1,249	1,187	1,143	1,106
86	086.txt	49,830	946	894	833	792	744	708	663	620	583	563
87	087.txt	78,446	1,491	1,404	1,324	1,249	1,180	1,121	1,055	990	949	907
88	088.txt	41,478	741	682	651	608	568	525	506	482	457	438
89	089.txt	38,625	786	730	703	657	621	594	545	527	488	467
90	090.txt	31,463	682	605	569	525	494	468	439	418	399	378
91	091.txt	47,759	970	872	826	771	724	688	647	620	594	563
92	092.txt	44,972	957	894	824	781	709	675	632	600	586	564
93	093.txt	65,510	1,316	1,222	1,153	1,079	1,014	957	893	851	819	774
94	094.txt	113,094	1,672	1,539	1,456	1,370	1,295	1,251	1,183	1,132	1,085	1,040
95	095.txt	109,116	1,969	1,834	1,707	1,605	1,516	1,449	1,389	1,323	1,256	1,207
96	096.txt	143,563	2,550	2,408	2,241	2,103	1,992	1,895	1,795	1,720	1,625	1,565
97	097.txt	120,354	1,874	1,762	1,617	1,513	1,411	1,341	1,269	1,196	1,152	1,109
98	098.txt	115,605	2,044	1,909	1,787	1,682	1,590	1,527	1,450	1,382	1,314	1,253
99	099.txt	128,773	2,566	2,386	2,246	2,111	1,998	1,878	1,791	1,709	1,615	1,548
100	100.txt	249,771	4,669	4,348	4,065	3,818	3,585	3,409	3,264	3,108	2,952	2,807

ตารางที่ ก.1 (ต่อ)

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=02	Cutoff=03	Cutoff=04	Cutoff=05	Cutoff=06	Cutoff=07	Cutoff=08	Cutoff=09	Cutoff=10	Cutoff=11
1	a1.txt	854	80	56	35	29	21	14	12	9	8	6
2	a2.txt	916	74	41	24	20	13	12	11	10	11	8
3	a3.txt	921	112	66	41	28	18	16	11	9	4	1
4	a4.txt	925	87	57	37	25	20	16	14	6	5	4
5	a5.txt	948	117	67	44	28	23	16	11	9	4	2
6	a6.txt	1,051	134	65	48	36	23	16	13	12	10	5
7	a7.txt	1,194	97	55	41	31	24	18	12	6	5	6
8	a8.txt	1,262	142	94	66	52	37	26	18	19	19	11
9	a9.txt	1,737	232	137	95	75	51	42	38	26	22	16
10	a10.txt	1,980	254	160	109	88	67	50	40	34	28	21
11	b1.txt	2,085	215	146	94	79	63	54	42	31	24	18
12	b2.txt	2,149	220	125	99	82	62	46	40	29	24	19
13	b3.txt	2,167	208	156	111	84	65	54	45	37	25	26
14	b4.txt	2,176	252	165	122	92	64	57	43	35	28	17
15	b5.txt	2,198	238	160	111	91	74	60	50	43	40	30
16	b6.txt	2,278	308	187	135	102	72	57	48	42	35	29
17	b7.txt	2,505	320	203	150	116	86	63	56	49	45	34
18	b8.txt	2,798	293	192	148	107	102	85	68	53	39	34
19	b9.txt	2,798	328	214	159	126	104	88	75	63	54	46
20	b10.txt	2,911	377	234	163	125	92	73	61	44	35	30
21	c1.txt	3,088	368	226	163	124	97	80	67	55	45	42
22	c2.txt	3,228	323	228	149	120	99	82	65	54	41	33
23	c3.txt	3,358	386	264	200	146	110	85	67	59	50	47
24	c4.txt	3,360	207	145	98	71	70	62	49	45	37	34
25	c5.txt	3,499	506	322	212	160	120	95	76	61	55	47
26	c6.txt	3,505	476	323	237	181	138	111	95	83	66	57
27	c7.txt	3,678	366	254	202	164	119	97	86	71	61	50
28	c8.txt	3,819	420	288	238	170	140	117	101	83	67	57
29	c9.txt	3,924	506	347	257	197	150	122	108	84	75	63
30	c10.txt	3,991	519	343	234	181	149	116	97	87	74	60
31	d1.txt	4,146	488	330	245	182	145	122	100	81	63	55
32	d2.txt	4,304	460	299	236	184	150	124	104	89	74	63
33	d3.txt	4,436	552	364	275	212	168	132	116	101	93	80
34	d4.txt	4,471	534	347	261	202	152	138	120	108	95	82
35	d5.txt	4,500	465	329	259	187	151	122	111	97	84	70
36	d6.txt	4,689	548	391	293	241	194	149	130	114	103	92
37	d7.txt	4,663	591	409	305	242	207	165	144	128	115	98
38	d8.txt	4,760	623	415	317	240	179	152	127	115	98	77
39	d9.txt	4,863	530	390	291	230	194	150	132	110	99	81
40	d10.txt	5,102	662	434	317	245	204	163	141	127	105	92
41	e1.txt	5,155	744	507	352	274	221	201	164	149	135	109
42	e2.txt	5,222	443	297	232	171	147	120	103	89	78	71
43	e3.txt	5,449	661	451	343	258	207	177	154	131	116	102
44	e4.txt	5,598	573	397	304	238	210	169	152	114	105	93
45	e5.txt	5,686	712	507	361	290	236	198	173	150	133	115
46	e6.txt	5,792	632	459	346	277	218	180	160	143	124	111
47	e7.txt	5,841	767	535	388	299	240	191	166	141	123	109
48	e8.txt	5,968	791	560	412	308	255	197	167	139	120	105
49	e9.txt	5,992	867	588	398	293	241	199	168	139	117	107
50	e10.txt	6,112	752	509	379	284	249	201	172	158	142	119
51	f1.txt	6,190	604	441	325	266	221	190	171	142	124	115
52	f2.txt	6,272	769	549	410	322	276	240	203	165	144	134
53	f3.txt	6,294	849	546	388	296	245	212	177	143	131	113
54	f4.txt	6,318	692	490	371	318	266	217	189	159	138	128
55	f5.txt	6,422	918	600	436	332	266	241	204	167	144	133
56	f6.txt	6,767	684	491	353	283	227	196	174	150	135	122
57	f7.txt	6,793	873	563	424	322	267	248	207	186	154	126
58	f8.txt	6,967	853	638	466	368	300	245	213	191	165	147
59	f9.txt	7,114	763	528	417	322	267	218	186	165	146	133
60	f10.txt	7,137	796	534	414	316	273	238	214	180	162	136
61	g1.txt	7,244	753	530	403	328	273	240	209	179	148	133
62	g2.txt	7,308	897	611	479	357	285	244	211	183	159	150
63	g3.txt	7,327	989	708	492	378	299	243	225	199	180	157
64	g4.txt	7,438	1,006	657	481	364	290	247	216	195	169	152
65	g5.txt	7,527	982	659	494	404	319	269	230	195	172	151
66	g6.txt	7,591	1,002	690	498	391	328	282	245	208	185	166
67	g7.txt	7,610	644	455	351	283	226	203	184	162	141	129
68	g8.txt	7,626	890	628	465	370	312	269	217	196	182	159
69	g9.txt	7,779	826	597	473	373	300	255	218	197	163	135
70	g10.txt	7,889	817	575	463	369	312	274	231	202	184	170
71	h1.txt	8,458	878	642	467	374	324	281	258	221	196	171
72	h2.txt	8,536	981	697	520	429	373	313	273	233	201	178
73	h3.txt	8,611	831	626	458	349	290	235	205	187	162	151
74	h4.txt	8,628	972	649	497	401	354	302	267	218	183	166
75	h5.txt	8,764	864	628	477	392	328	278	245	210	181	159
76	h6.txt	8,783	973	660	494	388	353	288	242	221	190	171
77	h7.txt	8,850	1,050	691	516	415	362	303	266	245	230	205
78	h8.txt	8,853	948	675	513	404	324	283	257	222	202	184
79	h9.txt	8,896	960	689	531	442	355	299	262	229	203	188
80	h10.txt	9,105	1,142	778	568	432	358	306	262	223	196	174
81	i1.txt	9,256	1,004	738	547	427	350	298	254	218	193	180
82	i2.txt	9,298	1,309	890	648	517	412	354	300	268	235	207
83	i3.txt	9,355	1,272	824	620	500	384	339	295	250	221	196
84	i4.txt	9,367	1,194	828	627	492	394	329	293	248	218	203
85	i5.txt	9,571	1,338	862	639	522	428	359	312	261	229	202
86	i6.txt	9,651	957	723	570	450	385	319	293	250	223	207
87	i7.txt	9,905	1,394	947	687	555	466	389	327	289	246	217
88	i8.txt	9,931	1,309	841	644	516	405	340	318	278	234	210
89	i9.txt	10,096	1,084	724	564	447	362	321	279	250	234	215
90	i10.txt	10,226	1,128	773	588	459	394	348	298	262	244	225
91	j1.txt	19,402	2,211	1,486	1,080	900	753	643	559	491	441	414
92	j2.txt	29,526	3,631	2,417	1,806	1,443	1,214	1,045	918	820	727	674
93	j3.txt	38,942	5,305	3,538	2,667	2,103	1,767	1,531	1,324	1,174	1,038	951
94	j4.txt	51,759	5,491	3,792	2,897	2,272	1,946	1,684	1,485	1,347	1,220	1,106
95	j5.txt	61,500	6,755	4,633	3,527	2,901	2,470	2,153	1,875	1,681	1,501	1,393
96	j6.txt	70,928	7,374	5,122	3,948	3,231	2,686	2,317	2,064	1,882	1,713	1,568
97	j7.txt	80,332	8,247	5,739	4,405	3,551	3,008	2,640	2,352	2,107	1,902	1,741
98	j8.txt	90,541	9,272	6,523	4,963	3,951	3,316	2,891	2,542	2,309	2,081	1,917
99	j9.txt	100,267	10,717	7,296	5,591	4,550	3,797	3,285	2,857	2,595	2,350	2,162
100	j10.txt	200,892	15,312	13,117	10,075	8,333	6,991	6,015	5,317	4,807	4,349	3,998

ตารางที่ ก.1 (ต่อ)

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=12	Cutoff=13	Cutoff=14	Cutoff=15	Cutoff=16	Cutoff=17	Cutoff=18	Cutoff=19	Cutoff=20	Cutoff=21
1	a1.txt	854	3	3	3	2	2	1	1	0	0	0
2	a2.txt	916	6	2	2	2	1	1	1	1	1	1
3	a3.txt	921	1	0	0	0	0	0	0	0	0	0
4	a4.txt	925	3	3	1	1	0	0	0	0	0	0
5	a5.txt	948	1	0	0	0	0	0	0	0	0	0
6	a6.txt	1,051	3	3	2	0	0	0	0	0	0	0
7	a7.txt	1,194	6	4	3	3	4	2	2	1	0	0
8	a8.txt	1,262	8	9	8	7	5	5	4	3	2	1
9	a9.txt	1,737	13	10	8	7	4	3	3	3	2	2
10	a10.txt	1,980	17	13	11	8	8	5	5	5	5	5
11	b1.txt	2,085	17	11	10	7	6	5	5	4	4	4
12	b2.txt	2,140	14	15	16	13	15	12	7	5	4	3
13	b3.txt	2,167	26	23	22	17	13	13	13	12	10	9
14	b4.txt	2,176	14	13	10	9	8	8	5	4	4	4
15	b5.txt	2,196	32	26	16	13	13	12	10	7	8	6
16	b6.txt	2,278	20	15	10	8	6	5	5	4	5	6
17	b7.txt	2,505	29	24	18	17	12	9	7	5	5	5
18	b8.txt	2,796	30	25	20	16	15	13	13	9	9	6
19	b9.txt	2,796	38	30	27	25	23	22	16	15	13	14
20	b10.txt	2,911	30	25	21	20	20	19	17	16	11	8
21	c1.txt	3,088	35	32	23	22	20	18	16	13	12	10
22	c2.txt	3,228	29	29	22	19	17	18	17	17	15	13
23	c3.txt	3,356	40	39	36	31	31	31	24	20	16	14
24	c4.txt	3,389	26	22	22	15	12	10	7	8	8	6
25	c5.txt	3,499	41	41	37	31	24	24	21	20	18	17
26	c6.txt	3,505	53	44	38	36	30	27	23	21	18	17
27	c7.txt	3,676	47	39	34	31	32	23	20	19	17	17
28	c8.txt	3,819	53	47	42	39	33	28	25	22	19	19
29	c9.txt	3,924	59	53	42	37	33	30	27	24	21	15
30	c10.txt	3,991	56	49	43	41	40	34	31	28	24	22
31	d1.txt	4,145	50	47	43	41	36	32	32	29	27	24
32	d2.txt	4,304	55	53	46	44	42	39	34	32	29	27
33	d3.txt	4,436	69	63	58	50	46	38	34	34	31	28
34	d4.txt	4,471	68	58	53	45	40	35	30	25	20	18
35	d5.txt	4,500	63	54	47	44	38	38	41	38	34	30
36	d6.txt	4,859	75	65	53	41	36	31	29	26	25	20
37	d7.txt	4,863	87	75	67	59	52	46	40	33	27	22
38	d8.txt	4,780	68	60	54	51	46	41	42	40	34	30
39	d9.txt	4,863	68	56	48	43	41	37	31	31	25	24
40	d10.txt	5,102	85	82	75	68	58	55	48	44	39	41
41	e1.txt	5,155	91	80	71	64	55	49	47	43	38	34
42	e2.txt	5,222	71	62	54	44	38	38	38	35	33	27
43	e3.txt	5,449	93	83	78	72	68	57	52	47	44	41
44	e4.txt	5,598	84	79	74	65	57	53	49	44	37	33
45	e5.txt	5,686	101	94	82	72	62	52	49	46	43	42
46	e6.txt	5,792	94	81	75	68	61	61	53	51	48	46
47	e7.txt	5,841	94	84	78	71	67	64	55	49	47	44
48	e8.txt	5,868	100	89	82	68	63	60	55	53	48	41
49	e9.txt	5,992	98	91	79	71	62	56	55	50	50	42
50	e10.txt	6,112	107	102	89	77	75	66	61	56	55	47
51	f1.txt	6,190	98	84	74	70	71	62	56	48	47	44
52	f2.txt	6,272	121	109	95	77	75	63	59	50	48	43
53	f3.txt	6,294	99	82	80	72	71	65	62	57	51	44
54	f4.txt	6,318	115	106	97	93	83	73	64	55	48	42
55	f5.txt	6,422	125	114	104	97	86	77	75	70	64	59
56	f6.txt	6,787	113	101	92	79	65	60	54	46	39	37
57	f7.txt	6,793	113	105	99	86	81	71	62	53	46	42
58	f8.txt	6,967	130	118	105	96	86	81	75	71	63	59
59	f9.txt	7,114	125	105	96	91	81	75	69	61	56	57
60	f10.txt	7,137	123	120	108	99	84	74	67	59	57	52
61	g1.txt	7,244	116	98	86	79	76	68	69	61	53	50
62	g2.txt	7,306	137	127	109	97	90	83	77	69	68	67
63	g3.txt	7,327	127	122	113	101	91	88	76	69	64	61
64	g4.txt	7,438	137	132	112	102	95	83	78	75	68	66
65	g5.txt	7,527	131	116	104	101	94	90	83	80	72	69
66	g6.txt	7,591	149	128	114	103	91	81	74	67	60	57
67	g7.txt	7,610	118	110	95	87	86	85	78	76	66	63
68	g8.txt	7,626	138	123	112	104	100	82	72	73	69	66
69	g9.txt	7,779	118	103	96	84	76	77	74	64	63	66
70	g10.txt	7,689	150	133	126	112	102	92	87	75	69	59
71	h1.txt	8,458	156	148	136	119	113	101	92	84	80	72
72	h2.txt	8,536	154	139	127	118	110	100	92	84	77	74
73	h3.txt	8,611	140	123	116	113	98	92	83	79	72	64
74	h4.txt	8,628	149	140	130	120	110	95	94	89	84	81
75	h5.txt	8,764	148	135	122	119	105	99	92	89	87	82
76	h6.txt	8,783	163	145	128	123	108	98	87	82	81	76
77	h7.txt	8,850	170	156	138	125	112	105	89	84	81	77
78	h8.txt	8,853	167	150	136	125	121	112	98	94	89	81
79	h9.txt	8,896	176	156	139	128	116	105	96	89	85	79
80	h10.txt	9,105	150	133	120	112	107	97	87	89	84	78
81	i1.txt	9,256	161	148	138	133	123	114	102	96	92	90
82	i2.txt	9,298	184	161	146	134	124	112	102	96	91	83
83	i3.txt	9,355	182	157	149	137	124	114	112	110	108	102
84	i4.txt	9,367	180	170	154	140	130	118	111	105	95	87
85	i5.txt	9,571	186	182	165	155	143	133	122	113	99	90
86	i6.txt	9,651	195	176	160	139	121	113	104	97	89	84
87	i7.txt	9,905	201	186	173	156	144	136	130	129	117	108
88	i8.txt	9,931	179	164	160	143	132	121	111	97	92	83
89	i9.txt	10,096	196	180	159	149	132	121	115	110	93	89
90	i10.txt	10,226	201	183	167	151	135	126	113	104	98	93
91	j1.txt	19,402	374	344	312	290	271	248	234	219	214	205
92	j2.txt	29,526	610	573	532	493	460	432	399	386	351	334
93	j3.txt	38,942	867	779	739	691	641	606	570	537	510	487
94	j4.txt	51,759	1,050	968	902	841	780	738	695	660	633	613
95	j5.txt	61,500	1,290	1,220	1,115	1,050	979	933	883	845	795	760
96	j6.txt	70,928	1,427	1,348	1,256	1,193	1,099	1,040	975	929	881	846
97	j7.txt	80,332	1,803	1,496	1,378	1,272	1,202	1,150	1,095	1,043	999	942
98	j8.txt	90,541	1,774	1,662	1,550	1,478	1,394	1,325	1,258	1,187	1,126	1,068
99	j9.txt	100,267	1,975	1,851	1,747	1,648	1,564	1,473	1,403	1,347	1,258	1,193
100	j10.txt	200,892	3,669	3,416	3,152	2,980	2,819	2,672	2,537	2,435	2,296	2,194

ตารางที่ ก.1 (ต่อ)

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=02	Cutoff=03	Cutoff=04	Cutoff=05	Cutoff=06	Cutoff=07	Cutoff=08	Cutoff=09	Cutoff=10	Cutoff=11
1	1001.txt	803	75	53	35	20	13	4	12	1	8	4
2	1002.txt	876	67	31	17	15	11	5	7	1	5	6
3	1003.txt	905	111	61	38	26	17	16	8	4	0	0
4	1004.txt	904	86	56	31	19	12	9	4	0	0	0
5	1005.txt	909	111	61	38	19	22	12	3	3	0	0
6	1006.txt	1,010	127	62	42	29	16	10	8	11	0	0
7	1007.txt	1,172	88	48	36	26	15	9	10	0	0	0
8	1008.txt	1,217	141	88	56	51	36	21	9	18	10	3
9	1009.txt	1,688	226	128	95	70	48	37	28	22	19	15
10	1010.txt	1,955	245	158	108	81	65	41	31	30	27	12
11	1011.txt	2,046	209	143	86	71	62	50	34	26	17	16
12	1012.txt	2,108	215	116	93	78	59	46	32	28	15	15
13	1013.txt	2,090	200	149	103	74	57	50	45	31	21	25
14	1014.txt	2,141	245	160	118	90	57	54	37	34	20	15
15	1015.txt	2,192	230	153	105	84	68	52	46	36	34	33
16	1016.txt	2,210	302	180	132	95	67	53	44	37	36	27
17	1017.txt	2,424	317	194	145	109	78	59	52	44	43	30
18	1018.txt	2,792	286	191	139	104	99	82	63	50	37	24
19	1019.txt	2,789	326	210	154	125	94	84	69	55	46	39
20	1020.txt	2,876	375	232	156	117	84	69	52	40	26	27
21	1021.txt	3,085	365	219	158	119	92	77	65	51	38	39
22	1022.txt	3,202	317	222	140	114	93	81	55	54	36	32
23	1023.txt	3,281	379	256	191	139	108	82	59	57	48	38
24	1024.txt	3,366	203	139	90	71	63	59	41	39	28	28
25	1025.txt	3,486	496	319	207	157	111	86	69	56	47	46
26	1026.txt	3,445	471	323	233	175	131	104	94	83	58	53
27	1027.txt	3,657	359	248	193	159	110	97	86	64	59	49
28	1028.txt	3,763	411	283	231	164	139	110	92	74	63	56
29	1029.txt	3,905	506	344	254	196	148	115	102	80	67	54
30	1030.txt	3,921	512	334	228	172	148	108	91	78	65	55
31	1031.txt	4,144	483	329	235	181	137	120	94	76	62	53
32	1032.txt	4,229	458	296	231	184	153	122	101	86	65	53
33	1033.txt	4,383	546	354	268	208	160	127	110	99	84	72
34	1034.txt	4,383	529	345	251	197	147	137	112	99	91	73
35	1035.txt	4,467	458	326	257	182	144	119	106	96	76	63
36	1036.txt	4,622	543	388	286	231	193	142	124	108	95	89
37	1037.txt	4,599	585	406	300	232	204	160	137	123	113	91
38	1038.txt	4,680	614	412	316	236	172	144	127	108	93	69
39	1039.txt	4,803	521	387	284	227	177	142	124	110	94	79
40	1040.txt	5,028	655	430	312	236	198	155	140	126	102	91
41	1041.txt	5,080	740	504	342	272	215	196	159	142	135	107
42	1042.txt	5,201	434	289	227	170	143	116	96	89	73	63
43	1043.txt	5,385	675	443	340	249	204	167	145	131	107	100
44	1044.txt	5,533	565	387	299	229	202	164	147	110	98	91
45	1045.txt	5,662	708	501	357	288	232	196	164	140	131	114
46	1046.txt	5,751	630	454	343	276	216	174	155	134	115	109
47	1047.txt	5,751	757	534	385	293	237	184	159	135	115	106
48	1048.txt	5,788	788	557	410	302	251	196	166	131	120	100
49	1049.txt	5,904	885	565	398	290	233	198	166	137	111	98
50	1050.txt	6,055	747	503	370	288	249	197	168	156	132	117
51	1051.txt	6,091	598	441	319	282	216	181	168	134	121	109
52	1052.txt	6,241	761	545	408	319	271	232	196	162	134	130
53	1053.txt	6,264	839	540	381	295	239	207	173	142	129	107
54	1054.txt	6,249	690	479	369	310	261	207	182	152	132	126
55	1055.txt	6,339	916	593	435	329	283	236	202	166	143	125
56	1056.txt	6,672	682	483	349	281	220	193	164	144	132	121
57	1057.txt	6,784	872	565	420	317	281	239	207	178	150	119
58	1058.txt	6,870	850	638	460	361	297	240	205	185	157	138
59	1059.txt	7,060	763	525	408	319	263	218	186	157	138	131
60	1060.txt	7,127	794	525	408	316	269	229	208	179	162	135
61	1061.txt	7,196	746	529	400	319	271	239	208	178	148	132
62	1062.txt	7,278	889	609	470	353	276	243	206	178	150	150
63	1063.txt	7,245	987	699	482	369	290	234	217	191	173	154
64	1064.txt	7,407	1,001	654	473	364	290	239	208	176	167	152
65	1065.txt	7,454	977	655	493	395	316	261	221	194	171	146
66	1066.txt	7,534	999	682	491	386	319	276	239	197	178	156
67	1067.txt	7,570	636	452	347	278	220	197	182	157	136	126
68	1068.txt	7,594	882	626	462	368	305	261	208	187	177	156
69	1069.txt	7,702	821	592	472	366	293	253	215	192	157	130
70	1070.txt	7,854	816	566	456	368	308	272	225	196	180	164
71	1071.txt	8,362	876	636	467	369	321	276	256	213	192	165
72	1072.txt	8,523	973	689	515	423	364	311	265	225	195	171
73	1073.txt	8,550	823	617	450	340	288	230	205	181	157	148
74	1074.txt	8,584	963	648	492	397	351	295	265	215	180	165
75	1075.txt	8,679	859	626	470	388	324	272	235	201	181	152
76	1076.txt	8,695	969	660	491	382	351	280	241	215	187	170
77	1077.txt	8,763	1,044	689	507	405	359	301	256	239	224	199
78	1078.txt	8,834	940	673	505	395	316	275	250	212	193	179
79	1079.txt	8,864	953	686	524	433	348	293	256	227	195	180
80	1080.txt	9,067	1,140	772	566	425	354	298	259	214	193	173
81	1081.txt	9,244	1,000	736	541	420	341	290	251	208	186	179
82	1082.txt	9,240	1,303	884	642	516	405	346	298	265	229	203
83	1083.txt	9,337	1,262	819	620	494	376	336	290	250	216	191
84	1084.txt	9,274	1,191	825	618	483	390	320	285	241	216	199
85	1085.txt	9,489	1,338	858	635	513	424	351	310	261	225	198
86	1086.txt	9,555	954	722	561	446	359	314	293	244	221	198
87	1087.txt	9,871	1,391	942	683	549	461	388	317	282	239	215
88	1088.txt	9,866	1,301	837	639	513	398	332	311	274	232	205
89	1089.txt	10,094	1,081	716	555	439	355	313	275	246	228	214
90	1090.txt	10,178	1,118	768	579	454	391	341	292	253	236	217
91	1091.txt	19,345	2,205	1,464	1,071	895	745	635	555	481	440	409
92	1092.txt	29,513	3,625	2,407	1,803	1,440	1,213	1,045	909	816	720	668
93	1093.txt	38,896	5,304	3,529	2,659	2,101	1,767	1,530	1,319	1,171	1,036	945
94	1094.txt	51,731	5,483	3,784	2,897	2,268	1,940	1,679	1,481	1,344	1,215	1,101
95	1095.txt	61,440	6,752	4,626	3,517	2,900	2,463	2,148	1,874	1,653	1,494	1,383
96	1096.txt	70,852	7,367	5,116	3,939	3,228	2,657	2,314	2,056	1,853	1,705	1,563
97	1097.txt	80,261	8,237	5,735	4,401	3,544	3,005	2,635	2,347	2,102	1,899	1,738
98	1098.txt	90,501	9,264	6,516	4,961	3,945	3,308	2,885	2,534	2,309	2,077	1,911
99	1099.txt	100,261	10,715	7,291	5,585	4,541	3,792	3,278	2,852	2,590	2,341	2,158
100	1100.txt	200,883	15,312	13,109	10,070	8,323	6,989	6,007	5,309	4,804	4,342	3,987

ตารางที่ ก.1 (ต่อ)

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=12	Cutoff=13	Cutoff=14	Cutoff=15	Cutoff=16	Cutoff=17	Cutoff=18	Cutoff=19	Cutoff=20	Cutoff=21
1	1001.txt	803	1	0	0	0	0	0	0	0	0	0
2	1002.txt	876	0	0	0	0	0	0	0	0	0	0
3	1003.txt	905	0	0	0	0	0	0	0	0	0	0
4	1004.txt	904	0	1	0	0	0	0	0	0	0	0
5	1005.txt	909	0	0	0	0	0	0	0	0	0	0
6	1006.txt	1,010	0	0	0	0	0	0	0	0	0	0
7	1007.txt	1,172	4	4	0	0	0	0	0	0	0	0
8	1008.txt	1,217	2	3	4	0	0	5	0	0	0	0
9	1009.txt	1,688	9	8	1	0	0	1	0	0	0	0
10	1010.txt	1,955	16	6	9	5	4	4	5	3	0	0
11	1011.txt	2,046	12	11	9	0	4	2	0	4	0	0
12	1012.txt	2,108	11	11	7	4	6	7	1	0	1	0
13	1013.txt	2,090	26	15	21	12	9	10	8	9	7	4
14	1014.txt	2,141	13	8	6	5	7	4	2	3	0	0
15	1015.txt	2,192	25	21	7	6	8	4	6	6	0	3
16	1016.txt	2,210	11	8	1	1	3	0	0	0	4	0
17	1017.txt	2,424	25	16	12	7	8	5	0	1	3	0
18	1018.txt	2,792	29	24	16	11	5	5	9	6	0	5
19	1019.txt	2,789	38	24	25	16	14	13	12	7	6	12
20	1020.txt	2,876	29	25	16	11	12	12	7	6	7	6
21	1021.txt	3,085	30	29	23	19	19	10	14	10	7	4
22	1022.txt	3,202	27	21	16	11	14	10	8	14	6	3
23	1023.txt	3,281	31	29	34	24	22	29	14	12	8	5
24	1024.txt	3,366	16	14	14	14	7	10	0	0	1	0
25	1025.txt	3,486	32	32	27	30	20	18	20	14	13	11
26	1026.txt	3,445	47	44	31	33	29	20	14	14	14	12
27	1027.txt	3,657	45	35	28	27	25	17	16	10	16	10
28	1028.txt	3,763	49	43	33	30	24	20	20	16	12	18
29	1029.txt	3,905	57	50	33	33	32	25	22	19	13	15
30	1030.txt	3,921	54	41	43	34	34	30	25	22	14	15
31	1031.txt	4,144	41	39	39	40	33	31	24	19	20	17
32	1032.txt	4,229	54	48	42	41	39	33	31	26	23	22
33	1033.txt	4,383	62	60	48	44	39	37	26	25	27	24
34	1034.txt	4,383	59	57	49	39	34	33	28	24	19	13
35	1035.txt	4,467	58	48	37	36	38	37	37	35	27	30
36	1036.txt	4,622	66	63	45	40	33	28	24	21	18	24
37	1037.txt	4,599	78	67	59	54	43	42	36	25	23	14
38	1038.txt	4,680	60	52	47	44	38	34	40	37	28	24
39	1039.txt	4,803	63	46	40	38	37	30	26	30	15	15
40	1040.txt	5,028	76	74	67	61	57	51	40	40	29	39
41	1041.txt	5,080	82	78	70	56	54	46	46	38	35	27
42	1042.txt	5,201	66	53	47	44	29	28	33	33	24	21
43	1043.txt	5,365	85	77	77	63	59	51	47	44	36	33
44	1044.txt	5,533	77	71	64	56	56	53	43	36	32	25
45	1045.txt	5,662	93	88	79	65	54	49	43	38	41	42
46	1046.txt	5,751	90	75	66	59	57	55	43	45	44	43
47	1047.txt	5,751	94	82	76	66	59	63	51	46	43	38
48	1048.txt	5,788	99	84	73	59	59	56	46	53	43	36
49	1049.txt	5,904	89	83	77	65	59	49	47	42	40	39
50	1050.txt	6,055	104	99	80	70	65	61	57	51	47	42
51	1051.txt	6,091	92	81	73	66	62	61	50	40	42	40
52	1052.txt	6,241	114	108	93	72	66	63	57	48	38	39
53	1053.txt	6,264	94	74	78	69	62	57	56	47	47	43
54	1054.txt	6,249	111	105	91	84	78	71	55	45	39	38
55	1055.txt	6,339	118	112	96	95	78	72	68	68	54	53
56	1056.txt	6,672	113	91	82	78	64	54	53	42	36	36
57	1057.txt	6,784	113	104	90	87	73	61	60	52	40	32
58	1058.txt	6,870	120	112	99	95	83	80	71	67	55	56
59	1059.txt	7,060	122	100	92	90	73	67	62	57	50	52
60	1060.txt	7,127	123	120	99	99	75	68	64	58	47	44
61	1061.txt	7,196	109	95	82	72	66	60	67	56	43	48
62	1062.txt	7,278	130	125	99	90	87	73	73	66	61	61
63	1063.txt	7,245	118	116	110	96	90	78	75	59	61	53
64	1064.txt	7,407	134	132	106	94	88	74	71	73	58	55
65	1065.txt	7,454	126	116	96	100	86	84	79	71	68	59
66	1066.txt	7,534	146	125	113	103	83	75	72	63	56	54
67	1067.txt	7,570	117	110	85	84	78	81	70	71	59	53
68	1068.txt	7,594	131	119	109	97	91	72	64	65	63	60
69	1069.txt	7,702	115	98	79	74	72	68	64	60	55	63
70	1070.txt	7,854	147	125	121	107	92	83	83	72	62	51
71	1071.txt	8,362	153	146	132	110	112	93	91	80	79	67
72	1072.txt	8,523	148	134	120	111	107	100	86	81	76	68
73	1073.txt	8,560	131	116	107	113	92	90	74	79	68	60
74	1074.txt	8,584	142	132	124	119	108	95	87	85	76	72
75	1075.txt	8,679	143	128	117	112	102	97	89	86	83	78
76	1076.txt	8,695	157	137	121	123	106	94	81	78	72	68
77	1077.txt	8,763	169	147	129	122	105	96	84	83	78	71
78	1078.txt	8,834	159	146	133	116	119	102	96	85	85	77
79	1079.txt	8,864	169	150	133	124	112	102	91	87	76	76
80	1080.txt	9,067	148	127	110	106	104	94	80	80	80	72
81	1081.txt	9,244	158	143	135	125	116	112	93	88	92	84
82	1082.txt	9,240	176	156	139	131	114	109	97	86	84	80
83	1083.txt	9,337	179	153	139	136	122	105	108	104	103	101
84	1084.txt	9,274	179	164	152	137	128	114	105	102	87	83
85	1085.txt	9,489	179	181	160	150	140	131	119	106	92	83
86	1086.txt	9,555	186	172	157	134	113	111	96	90	85	84
87	1087.txt	9,871	192	177	172	151	139	132	124	126	109	101
88	1088.txt	9,866	177	161	156	136	128	112	101	88	91	80
89	1089.txt	10,094	193	172	156	149	125	113	110	103	87	89
90	1090.txt	10,178	195	176	161	146	127	119	105	99	91	83
91	1091.txt	19,345	367	336	305	289	265	244	231	214	207	201
92	1092.txt	29,513	603	568	529	486	456	428	393	356	345	333
93	1093.txt	38,896	862	770	738	686	632	603	567	531	500	486
94	1094.txt	51,731	1,047	961	893	834	774	731	694	656	628	607
95	1095.txt	61,440	1,281	1,216	1,115	1,042	972	931	879	841	793	758
96	1096.txt	70,852	1,425	1,342	1,248	1,188	1,095	1,036	974	922	875	838
97	1097.txt	80,261	1,595	1,495	1,372	1,267	1,200	1,141	1,087	1,035	996	942
98	1098.txt	90,501	1,773	1,657	1,546	1,469	1,387	1,319	1,255	1,179	1,118	1,065
99	1099.txt	100,261	1,968	1,847	1,745	1,646	1,558	1,465	1,399	1,342	1,258	1,192
100	1100.txt	200,883	3,666	3,413	3,176	2,974	2,813	2,666	2,531	2,426	2,295	2,185

ตารางที่ ก.2 แสดงรายละเอียดข้อมูลตัวอย่างที่ใช้ทดสอบสมการ

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=02	Cutoff=03	Cutoff=04	Cutoff=05	Cutoff=06	Cutoff=07	Cutoff=08	Cutoff=09	Cutoff=10	Cutoff=11
1	ta1.txt	1,150	136	78	52	31	22	13	10	9	11	11
2	ta2.txt	2,709	307	204	159	128	99	76	64	48	45	42
3	ta3.txt	2,433	326	209	149	112	86	71	54	39	37	31
4	ta4.txt	2,232	270	180	126	90	77	57	46	36	32	28
5	ta5.txt	1,388	134	66	42	29	21	15	10	8	7	7
6	ta6.txt	1,356	165	105	73	48	33	26	21	10	8	5
7	ta7.txt	1,341	119	74	49	40	27	20	10	10	9	7
8	ta8.txt	3,550	404	334	244	190	159	126	99	78	69	58
9	ta9.txt	2,705	371	233	167	123	94	81	73	54	44	34
10	ta10.txt	1,350	121	80	60	39	36	34	24	24	15	9
11	tb1.txt	2,111	217	146	94	77	63	54	42	30	24	18
12	tb2.txt	3,703	479	289	236	170	136	106	93	75	65	55
13	tb3.txt	1,575	157	107	73	59	45	30	27	22	20	18
14	tb4.txt	2,600	363	242	157	122	91	70	55	48	39	30
15	tb5.txt	1,692	200	131	93	77	62	56	40	29	21	19
16	tb6.txt	2,983	309	186	139	115	98	82	64	57	50	45
17	tb7.txt	3,820	452	300	218	173	136	117	102	88	71	69
18	tb8.txt	4,379	589	370	279	204	169	132	114	97	79	70
19	tb9.txt	4,218	470	313	222	183	117	98	70	59	52	45
20	tb10.txt	2,558	315	199	145	121	92	76	63	54	43	39
21	tc1.txt	3,308	463	292	201	154	116	84	67	58	48	51
22	tc2.txt	4,329	664	407	293	217	177	141	116	94	80	65
23	tc3.txt	4,236	507	345	245	192	160	131	121	100	87	79
24	tc4.txt	4,324	449	293	228	168	139	119	99	79	66	56
25	tc5.txt	3,727	558	362	254	197	149	116	99	77	69	66
26	tc6.txt	3,784	435	274	202	161	142	116	99	80	73	72
27	tc7.txt	3,046	388	253	199	147	117	97	73	64	58	50
28	tc8.txt	3,182	436	293	215	154	119	99	84	75	61	50
29	tc9.txt	3,059	387	255	184	135	106	86	73	70	58	49
30	tc10.txt	2,920	352	241	176	120	96	77	63	52	50	44
31	td1.txt	4,476	582	387	271	209	172	141	113	92	87	77
32	td2.txt	3,041	358	256	173	140	109	91	78	64	55	47
33	td3.txt	5,584	743	473	376	297	225	185	159	134	120	106
34	td4.txt	2,959	373	216	156	112	85	65	52	44	37	35
35	td5.txt	4,280	560	392	282	208	166	140	118	109	93	79
36	td6.txt	2,753	359	242	183	125	112	91	71	58	45	36
37	td7.txt	2,368	283	196	136	104	81	62	52	47	41	36
38	td8.txt	2,539	308	207	137	116	96	69	56	46	38	32
39	td9.txt	5,391	621	432	314	253	201	169	142	124	113	107
40	td10.txt	4,941	641	389	288	230	196	162	143	123	107	92
41	te1.txt	2,902	401	265	199	146	106	88	71	60	48	39
42	te2.txt	3,515	461	300	220	159	123	96	84	68	60	54
43	te3.txt	5,104	647	446	317	261	212	171	145	132	113	97
44	te4.txt	4,055	513	355	255	194	148	121	105	97	84	73
45	te5.txt	4,760	629	439	326	236	188	166	141	119	101	88
46	te6.txt	4,956	639	437	311	239	191	162	135	121	109	91
47	te7.txt	5,945	805	544	398	295	252	213	183	161	141	122
48	te8.txt	5,778	780	519	370	271	226	186	165	143	123	107
49	te9.txt	5,798	623	440	324	252	209	177	144	133	113	102
50	te10.txt	7,212	722	517	377	317	251	217	191	152	127	113
51	tf1.txt	10,297	880	603	454	382	328	276	242	222	201	182
52	tf2.txt	11,242	980	677	505	437	356	303	257	231	209	192
53	tf3.txt	9,285	1,009	698	563	437	346	298	262	223	201	172
54	tf4.txt	7,874	814	552	419	320	269	233	216	186	174	148
55	tf5.txt	8,519	1,070	733	540	421	354	299	260	233	211	194
56	tf6.txt	16,257	1,927	1,318	1,014	811	656	571	496	431	378	345
57	tf7.txt	15,351	1,705	1,177	907	782	610	517	455	392	352	320
58	tf8.txt	13,438	1,455	1,001	781	613	493	424	383	339	304	277
59	tf9.txt	14,580	1,443	1,036	810	656	556	477	414	370	324	309
60	tf10.txt	16,707	1,922	1,339	1,025	846	708	592	536	467	409	363
61	tg1.txt	10,851	1,234	847	661	530	447	387	335	286	264	232
62	tg2.txt	18,956	1,600	1,134	913	745	633	550	498	438	402	385
63	tg3.txt	15,815	1,753	1,208	922	720	623	524	472	417	380	339
64	tg4.txt	18,655	1,853	1,319	1,015	816	677	577	517	466	423	381
65	tg5.txt	18,959	1,808	1,213	951	781	650	565	481	423	397	374
66	tg6.txt	18,743	1,997	1,357	1,034	846	704	603	523	450	404	380
67	tg7.txt	17,888	1,472	1,049	809	683	575	480	417	373	338	309
68	tg8.txt	16,108	1,905	1,280	958	777	642	583	481	438	394	347
69	tg9.txt	8,531	1,031	713	553	435	363	310	283	224	193	170
70	tg10.txt	14,574	1,745	1,216	916	740	618	534	456	404	374	351
71	th1.txt	37,283	3,183	2,237	1,758	1,457	1,226	1,079	965	871	771	703
72	th2.txt	20,199	2,268	1,571	1,220	1,003	833	721	646	569	512	450
73	th3.txt	21,467	2,222	1,564	1,194	981	797	701	617	551	513	467
74	th4.txt	14,851	1,714	1,234	898	714	610	536	454	394	353	324
75	th5.txt	14,980	1,677	1,167	909	707	570	500	445	403	363	328
76	th6.txt	25,339	2,463	1,695	1,286	1,042	869	751	675	594	550	508
77	th7.txt	21,541	2,351	1,635	1,243	983	824	707	646	571	509	465
78	th8.txt	25,170	2,131	1,707	1,336	1,070	880	776	676	610	555	492
79	th9.txt	14,749	1,711	1,181	872	702	596	510	452	409	355	329
80	th10.txt	53,567	5,732	3,861	2,928	2,348	1,988	1,724	1,553	1,366	1,258	1,150
81	ti1.txt	24,915	2,894	1,931	1,447	1,159	953	839	741	652	595	540
82	ti2.txt	22,224	2,263	1,598	1,242	990	826	708	603	536	498	446
83	ti3.txt	40,688	2,301	2,136	1,737	1,407	1,208	1,082	976	871	792	725
84	ti4.txt	19,179	2,078	1,389	1,082	880	709	623	532	507	444	405
85	ti5.txt	22,763	2,129	1,546	1,226	983	813	715	622	548	494	462
86	ti6.txt	31,231	3,109	2,153	1,653	1,348	1,131	984	874	771	712	652
87	ti7.txt	23,493	2,898	1,980	1,509	1,214	1,011	866	769	677	604	559
88	ti8.txt	26,732	2,732	1,905	1,451	1,168	970	847	766	674	605	565
89	ti9.txt	25,772	2,821	1,898	1,460	1,192	985	874	765	675	606	555
90	ti10.txt	38,058	3,664	2,558	1,948	1,600	1,336	1,162	1,025	954	873	803
91	tj1.txt	30,121	3,146	2,122	1,605	1,272	1,084	948	853	752	697	623
92	tj2.txt	51,358	5,007	3,512	2,661	2,181	1,840	1,582	1,396	1,256	1,148	1,041
93	tj3.txt	48,141	4,956	3,478	2,616	2,110	1,773	1,526	1,341	1,218	1,109	1,025
94	tj4.txt	59,817	5,595	3,870	2,945	2,352	2,023	1,726	1,531	1,378	1,219	1,114
95	tj5.txt	68,307	7,529	5,237	3,948	3,201	2,716	2,341	2,080	1,846	1,685	1,534
96	tj6.txt	105,228	10,245	7,119	5,372	4,317	3,685	3,187	2,850	2,526	2,295	2,109
97	tj7.txt	103,143	10,590	7,288	5,498	4,428	3,734	3,202	2,796	2,552	2,304	2,142
98	tj8.txt	82,095	8,070	5,622	4,242	3,447	2,922	2,517	2,240	1,988	1,822	1,657
99	tj9.txt	92,121	9,929	6,725	5,178	4,133	3,475	3,001	2,657	2,410	2,169	1,995
100	tj10.txt	204,599	15,598	13,264	10,217	8,449	7,113	6,140	5,368	4,822	4,387	4,061

ตารางที่ ก.2 (ต่อ)

ลำดับ	แฟ้ม	ขนาด(byte)	Cutoff Values									
			Cutoff=12	Cutoff=13	Cutoff=14	Cutoff=15	Cutoff=16	Cutoff=17	Cutoff=18	Cutoff=19	Cutoff=20	Cutoff=21
1	ta1.txt	1,150	8	7	3	3	1	1	1	1	2	2
2	ta2.txt	2,709	35	30	25	19	15	14	13	12	9	7
3	ta3.txt	2,433	28	24	20	16	16	14	12	8	6	3
4	ta4.txt	2,232	19	17	14	14	13	12	12	10	10	9
5	ta5.txt	1,388	19	15	13	6	4	3	2	1	1	1
6	ta6.txt	1,356	5	6	4	3	4	3	3	3	2	1
7	ta7.txt	1,341	7	5	5	8	8	5	4	4	3	2
8	ta8.txt	3,550	48	40	33	30	26	24	23	21	19	16
9	ta9.txt	2,705	31	26	21	20	16	13	11	10	8	6
10	ta10.txt	1,350	6	6	6	5	5	5	6	4	3	2
11	tb1.txt	2,111	17	11	10	7	5	5	5	4	4	4
12	tb2.txt	3,703	50	47	39	35	34	31	29	26	25	23
13	tb3.txt	1,575	17	11	9	10	7	6	6	4	4	4
14	tb4.txt	2,600	23	20	17	15	13	9	7	7	7	6
15	tb5.txt	1,692	12	11	11	11	9	9	7	8	5	3
16	tb6.txt	2,983	41	37	33	23	24	18	16	13	14	13
17	tb7.txt	3,820	58	51	40	37	33	29	25	24	24	22
18	tb8.txt	4,379	63	58	50	47	43	37	36	30	24	22
19	tb9.txt	4,218	36	35	31	28	24	19	16	13	10	9
20	tb10.txt	2,556	34	26	23	19	17	14	14	12	9	8
21	tc1.txt	3,308	46	40	37	32	27	25	22	20	14	11
22	tc2.txt	4,329	53	51	50	44	35	27	24	23	21	17
23	tc3.txt	4,236	66	61	53	49	44	41	35	31	31	29
24	tc4.txt	4,324	48	40	31	32	31	28	31	31	26	24
25	tc5.txt	3,727	59	49	41	35	32	29	23	20	17	14
26	tc6.txt	3,794	65	52	44	40	34	32	31	27	21	21
27	tc7.txt	3,046	44	40	41	38	33	25	18	18	17	14
28	tc8.txt	3,182	43	35	30	26	24	18	15	12	9	8
29	tc9.txt	3,059	42	37	33	31	28	22	22	21	16	13
30	tc10.txt	2,920	36	29	26	25	21	21	16	17	16	15
31	td1.txt	4,476	71	63	60	54	48	44	42	37	36	34
32	td2.txt	3,041	41	36	31	28	21	20	18	14	14	12
33	td3.txt	5,584	94	87	81	73	65	58	48	42	40	37
34	td4.txt	2,959	43	40	30	25	19	16	13	12	10	9
35	td5.txt	4,280	73	66	59	53	51	48	40	35	31	28
36	td6.txt	2,753	33	28	24	21	18	17	16	13	12	9
37	td7.txt	2,388	26	23	21	20	18	15	14	13	10	10
38	td8.txt	2,539	31	27	23	19	15	13	12	11	11	11
39	td9.txt	5,391	99	89	79	68	64	60	52	49	46	43
40	td10.txt	4,941	87	76	65	57	51	48	46	43	39	35
41	te1.txt	2,902	32	27	29	28	25	19	16	15	13	11
42	te2.txt	3,515	48	40	34	28	26	22	20	18	17	16
43	te3.txt	5,104	85	75	68	60	59	52	45	42	40	38
44	te4.txt	4,055	60	51	47	44	40	38	33	31	31	28
45	te5.txt	4,780	81	74	69	62	53	51	45	45	36	34
46	te6.txt	4,956	81	76	72	65	61	59	56	52	50	44
47	te7.txt	5,945	103	89	79	77	71	67	65	60	58	55
48	te8.txt	5,778	95	87	79	73	71	64	57	49	47	40
49	te9.txt	5,796	87	80	71	63	57	53	48	48	44	34
50	te10.txt	7,212	98	91	83	77	73	73	71	62	56	54
51	tf1.txt	10,297	170	151	133	123	115	107	99	96	96	87
52	tf2.txt	11,242	172	156	143	126	123	115	103	98	94	88
53	tf3.txt	9,285	144	131	118	105	97	91	86	81	80	79
54	tf4.txt	7,874	131	115	103	91	82	73	69	65	59	55
55	tf5.txt	8,519	172	157	137	121	108	100	93	87	77	73
56	tf6.txt	16,257	326	293	261	256	239	222	209	191	186	173
57	tf7.txt	15,351	306	276	259	225	211	200	187	167	159	150
58	tf8.txt	13,438	242	216	197	177	164	151	140	131	125	122
59	tf9.txt	14,590	286	268	250	232	208	182	174	163	154	148
60	tf10.txt	16,707	351	316	267	276	251	229	216	197	181	176
61	tg1.txt	10,651	198	180	165	148	133	128	117	108	100	89
62	tg2.txt	18,956	344	318	294	271	256	232	218	210	205	191
63	tg3.txt	15,815	297	280	263	243	232	215	201	195	182	165
64	tg4.txt	18,855	354	340	309	281	261	243	228	216	196	181
65	tg5.txt	18,959	336	316	300	275	252	237	220	204	197	188
66	tg6.txt	18,743	344	321	290	261	241	220	206	191	178	162
67	tg7.txt	17,988	279	260	243	240	217	205	198	178	165	163
68	tg8.txt	16,108	324	294	274	255	233	208	203	190	182	174
69	tg9.txt	8,531	148	132	123	109	99	97	85	86	76	67
70	tg10.txt	14,574	314	281	250	229	212	197	178	162	150	140
71	th1.txt	37,283	654	599	557	524	492	460	425	404	390	366
72	th2.txt	20,199	406	373	356	337	313	308	274	266	250	236
73	th3.txt	21,487	428	405	366	332	312	291	278	253	232	226
74	th4.txt	14,851	301	286	266	253	232	211	186	173	166	150
75	th5.txt	14,980	296	276	252	229	211	195	176	162	149	150
76	th6.txt	25,339	467	434	409	376	356	331	308	299	286	269
77	th7.txt	21,541	432	395	372	345	330	308	287	268	253	242
78	th8.txt	25,170	461	431	396	365	352	331	304	286	271	250
79	th9.txt	14,749	287	273	247	227	208	192	185	168	159	154
80	th10.txt	53,567	1,082	1,015	939	861	803	757	714	675	651	624
81	ti1.txt	24,915	497	468	432	398	367	340	318	299	284	268
82	ti2.txt	22,224	406	361	339	318	301	284	276	260	240	230
83	ti3.txt	40,688	679	643	600	572	547	512	479	437	422	395
84	ti4.txt	19,179	372	351	326	298	271	254	239	223	206	188
85	ti5.txt	22,783	424	392	367	346	334	310	289	277	259	234
86	ti6.txt	31,231	587	543	504	472	451	437	408	374	364	348
87	ti7.txt	23,493	515	480	446	414	377	348	332	311	299	277
88	ti8.txt	28,732	522	472	425	390	373	356	340	322	299	294
89	ti9.txt	25,772	501	465	443	400	371	334	321	299	284	273
90	ti10.txt	38,058	739	680	623	564	541	510	485	463	441	428
91	tj1.txt	30,121	571	532	509	471	432	405	389	349	332	320
92	tj2.txt	51,358	975	926	870	818	753	723	680	649	604	578
93	tj3.txt	48,141	957	898	836	782	727	691	652	613	583	556
94	tj4.txt	59,817	1,042	994	942	890	836	803	766	725	684	661
95	tj5.txt	68,307	1,405	1,314	1,241	1,161	1,090	1,033	975	929	884	844
96	tj6.txt	105,228	1,948	1,796	1,692	1,591	1,495	1,422	1,355	1,275	1,220	1,166
97	tj7.txt	103,143	1,989	1,840	1,716	1,612	1,497	1,418	1,343	1,270	1,229	1,192
98	tj8.txt	82,095	1,526	1,414	1,333	1,242	1,170	1,101	1,030	984	946	918
99	tj9.txt	92,121	1,855	1,712	1,615	1,516	1,422	1,345	1,271	1,206	1,140	1,090
100	tj10.txt	204,599	3,745	3,486	3,254	3,010	2,830	2,674	2,551	2,460	2,334	2,223

ประวัติผู้เขียน

ชื่อ – สกุล	นายนคร สาแก้ว
วัน เดือน ปีเกิด	18 เมษายน 2518
ที่อยู่	193/1 หมู่ 7 ตำบลนอกเมือง อำเภอเมือง จังหวัดสุรินทร์ 32000
ประวัติการศึกษา	2541 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันราชภัฏเพชรบุรี
ประวัติการทำงาน	
2538-2539	บริษัทซีซีแอนซี ตำบลในเมือง อำเภอเมือง จังหวัดสุรินทร์
2542-2545	อาจารย์ปฏิบัติการ คณะคอมพิวเตอร์ฯ สถาบันราชภัฏเพชรบุรี
ธันวาคม 2547	อาจารย์พิเศษ ม.ราชภัฏเพชรบุรี
2548-2549	อาจารย์ปฏิบัติการ คณะเทคโนโลยีสารสนเทศ ม.ราชภัฏเพชรบุรี
ปัจจุบัน	หัวหน้างานบริการข้อมูล ส่วนอำนวยการต่อสู้เพื่อเอาชนะปัญหา- ความยากจน(สยจ.) สำนักบริหารการปกครองท้องที่ กรมการปกครอง กระทรวงมหาดไทย