ระบบสืบค้นข้อมูลการท่องเที่ยวบนเว็บบล็อคด้วยการเรียนรู้เชิงลึก

SEARCHING SYSTEM ON TRAVELING WEB BLOG WITH DEEP LEARNING

นิรชิต ศรีประดู่

บัณฑิต สีดาว

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์บัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2562

ระบบสืบค้นข้อมูลการท่องเที่ยวบนเว็บบล็อคด้วยการเรียนรู้เชิงลึก

นายนิรชิต	ศรีประดู่	59010744
นายบัณฑิต	สีดาว	59010759
ผศ.บัณฑิต	พัสยา	อาจารย์ที่ปรึกษา
รศ. ดร.เกียรติกูล	เจียรนัยธนะกิจ	อาจารย์ที่ปรึกษาร่วม

ปีการศึกษา 2562

บทคัดย่อ

การใช้ชีวิตในปัจจุบันนั้นไม่ว่าจะพบเจอปัญหาหรือความต้องการรู้อะไรบางอย่างนั้นถูกแก้ไขและตอบสนองได้อย่างง่ายดายจากข้อมูลมากมายที่อยู่บนโลกของอินเตอร์เน็ต ปัญหาของมนุษย์เริ่มเปลี่ยนไปจากการต้องพยายามแก้ปัญหาหรือการลองผิดลองถูกด้วยตัวเองเป็นการหาข้อมูลอย่างไรให้ได้คำตอบของปัญหาโดยเร็วที่สุดจากอินเตอร์เน็ต ในช่วงเวลาตลอดหลายปีที่ผ่านมาประเทศไทยมีเว็บไซต์ที่เป็นแหล่งรวบรวมคำตอบของปัญหามากมายหรือกล่าวให้ถูกคือเป็นเว็บไซต์ที่เป็นแหล่งชุมชนเพื่อการพูดคุยแลกเปลี่ยนจากผู้ใช้งาน ซึ่งเว็บไซต์นั้นก็คือพันทิพดอทคอม (Pantip.com) พันทิพดอทคอมเป็นเว็บไซต์ที่มีข้อมูลเรื่องต่างๆมากมายจากการที่ผู้ใช้เขียนขึ้นมา ข้อมูลถูกแบ่งออกเป็นประเภทๆเพื่อการเข้าค้นหาข้อมูลต่างๆจะได้เป็นเรื่องง่าย แต่ทั้งนี้นั้นการจัดการข้อมูลเช่นนี้ก็ยังไม่พอสำหรับการค้นหาที่เฉพาะเจาะจงจากข้อมูลนับล้านนั่นทำให้เว็บไซต์ต้องมีฟังก์ชั่นการค้นหา แต่การทำงานของฟังก์ชั่นนั้นก็ไม่ได้ตอบโจทย์การใช้งานมากเท่าที่ควร ดังนั้นจะเป็นการดีถ้าหากระบบค้นหานั้นทำงานได้เป็นอย่างดีมากขึ้น เพื่อการนั้นการนำเทคโนโลยีที่เป็นที่รู้จักกันดีในปัจจุบันอย่างการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์หรือเอ็มแอล (machine learning) เข้ามาพัฒนาร่วมกับระบบค้นหาจะเป็นการแก้ปัญหาได้เป็นอย่างดี

งานวิจัยชิ้นนี้เป็นการสร้างระบบจำลองและเป็นต้นแบบของการนำเทคโนโลยี 2 ส่วนคือเทคโนโลยีการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์และระบบการค้นหา(elastic search) เข้ามาพัฒนาร่วมกันเพื่อตอบสนองต่อคำค้นที่เป็นคำพูดภาษามนุษย์มากกว่าเป็นคำค้นปรกติ โดยเทคโนโลยีเอ็มแอลจะทำหน้าที่ในการหาจุดประสงค์ที่แท้จริงของค้นที่ถูกป้อนเข้าสู่ระบบโดยผู้ใช้และส่งผลลัพธ์นั้นให้ทางระบบค้นหาเพื่อค้นหาบทความที่เกี่ยวข้องในระบบ ข้อมูลที่ถูกใช้ในระบบมาจากข้อมูลของเว็บไซต์พันทิพดอทคอมเพราะด้วยตัวข้อมูลและภาษาของข้อมูลนั้นตรงตามความต้องการและเพื่อเป็นการเปรียบเทียบการทำของระบบค้นหาทั้งสองแบบด้วย

Searching System on Traveling Web Blog with Deep Learning

Mr. Nirachit	Sripradu	59010744
Mr. Bundit	Seedao	59010759
Asst.Bundit	Patsaya	Advisor
Assoc.Prof.Dr. Kietikul	Jearanaitanakij	Co-Advisor

Academic Year 2019

ABSTRACT

Nowadays the way to troubleshooting problem or find some knowledge is easy to response it, there are a huge information on the internet so instead of figure out themselves problem they have to search too quickly to find it, as it can be new problem. In the past decade, in Thailand, there is website contain information, sentence of user and things, it is like a Web Blog and its name is “Pantip.com”. Website collect information from user then manage it to each tag that make easy to find some and there is some function of web called “search” have implemented to make people get more short to their want but this function of Pantip.com has not good enough for its work for some situation so it is good to be improved, there is some technology are mention to in the present called “machine learning” which can make computer understand about human language, I would be worth to implement this to the search function

This research were build to demonstrate and make prototype for co-improvement of machine learning and search function. The machine learning come to find main point of user input which like human language then pass to output to search function to query information from database. Information in the research is based on Pantip.com information in order to compare the old search function of Pantip.com and new one with improved with machine learning

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีได้อันเนื่องมาจากความช่วยเหลือจากอาจารย์ที่ปรึกษาปริญญานิพนธ์ ผศ. บัณฑิต พัสยา และอาจารย์ที่ปรึกษาร่วมปริญญานิพนธ์ รศ. ดร.เกียรติกูล เจียรนัยธนะกิจ ผู้ให้คำแนะนำและเสนอแนะ ให้คำชี้แจงและกำหนดแนวทางที่ถูกต้องแก่คณะผู้จัดทำ อีกทั้งยังช่วยแก้ปัญหาต่างๆ

ขอบคุณเพื่อนๆที่เกี่ยวข้องและอาจารย์ท่านอื่นๆที่มีส่วนเกี่ยวกับข้องทั้งทางตรงและทางอ้อมสำหรับการสนับสนุนให้ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงได้

นิรชิต ศรีประดู่
บัณฑิต สีดาว

สารบัญ

เรื่อง **หน้า**

สารบัญ(ต่อ)

เรื่อง **หน้า**

สารบัญตาราง

ตาราง **หน้า**

สารบัญรูป

รูป **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

ในช่วงสองทศวรรษที่ผ่านมาเรียกได้ว่าเป็นยุคแห่งข้อมูลและการเติบโตขึ้นของฮาร์ดแวร์(Hardware) ด้วยความสามารถที่ฮาร์ดแวร์มีประสิทธิภาพมากขึ้นแต่ขนาดเล็กลงและราคาที่ไม่สูงมากเกินไปทำให้การใช้งานถูกเปลี่ยนจากระดับบริษัทเป็นส่วนใหญ่มาเป็นระดับบุคคลมากขึ้นและรวมถึงการเติบโตขึ้นของสมาร์ทโฟน อุปกรณ์ขนาดเล็กมากความสามารถ ส่งผลทำให้การใช้งานสมาร์ทโฟนของบุคคลกลายเป็นส่วนที่สำคัญในการเพิ่มขึ้นของข้อมูลไม่ว่าจะเป็นรูปภาพ บทความ วีดีโอ หรืออื่นๆ ลงในระบบอินเตอร์เน็ต ด้วยความที่ข้อมูลเพิ่มมากขึ้นการใช้งานจึงมากขึ้นเช่นกันโดยตัวอย่างที่เห็นได้ชัดคืองานเกี่ยวกับการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์ที่เป็นการนำข้อมูลขนาดใหญ่จากหลายๆที่มา หาความสัมพันธ์ของข้อมูลเหล่านั้นแล้วนำมาใช้ในหลายๆด้าน เช่น ด้านธุรกิจ ด้านการงานวิจัย ตัวอย่างงานวิจัยที่จำเป็นต้องใช้ข้อมูลมากมายเข้ามาช่วย เช่น งานด้าน Computer Vision ซึ่งเป็นงานที่ทำเกี่ยวกับการประมวลผลภาพให้คอมพิวเตอร์สามารถแยกแยะวัตถุในภาพให้ได้ หรืองานด้านการประมวลภาษาธรรมชาติ(NLP) เป็นงานที่พยายามให้คอมพิวเตอร์สามารถเข้าใจภาษาจากมนุษย์ได้ จากทั้งหมดที่กล่าวมานั้นทำให้ยุคนี้เป็นยุคแห่งข้อมูลและการเติบโตของเทคโนโลยี

หนึ่งในส่วนสำคัญที่ทำให้การใช้งานสมาร์ทโฟนของบุคคลเพิ่มขึ้นคงปฏิเสธไม่ได้ว่าคือการเข้าถึงข้อมูลที่รวดเร็วจากโลกอินเตอร์เน็ต ซึ่งในประเทศไทยก็เช่นกันมีการใช้งานที่เพิ่มขึ้นอย่างมาก และก็คงปฏิเสธไม่ได้ว่าเป็นเพราะความต้องการการเข้าถึงข้อมูลต่างๆเป็นแน่ หนึ่งในเว็บไซต์ที่มีการใช้งานมากเพื่อพูดคุยและตั้งคำถามเพื่อไขความสงสัยในเรื่องต่างๆคือเว็บไซต์พันทิพ (Pantip) ซึ่งเป็นเว็บไซต์ที่มีการสร้างบทความขึ้นเพื่อใช้ในการแลกเปลี่ยนความเห็นและขอบเขตเรื่องพูดคุยที่กว้างมากๆ ไม่ว่าจะเป็น การท่องเที่ยว อาหาร การเดินทาง ที่พัก ปัญหาเล็กๆน้อย ข้อสงสัยที่ถกเถียงกันในหมู่เพื่อน นั้นทำให้ไม่ต้องสงสัยเลยว่าทำไมเว็บไซต์พันทิพจึงเป็นที่นิยมค่อนข้างมาก แต่ตัวระบบค้นหาที่เป็นส่วนสำคัญในการเข้าถึงบทความต่างๆกลับยังไม่ตอบโจทย์การใช้งานในส่วนของคำค้นที่เป็นรูปประโยคมากเท่าที่ควรนัก

ด้วยการเติบโตของงานวิจัยเพื่อให้คอมพิวเตอร์สามารถเข้าใจภาษาธรรมชาติได้ ทำให้การทำงานของเทคโนโลยีหลายๆอย่างสามารถตอบสนองต่อคำพูดหรือภาษามนุษย์ได้มากขึ้นจากที่อาจจะตอบสนองได้เพียงแค่คำหนึ่งคำ และหนึ่งในงานวิจัยที่ได้ยอมรับในปัจจุบันคืองานวิจัยของบริษัทกูเกิล(Google) ชื่อว่า BERT ที่สามารถตอบสนองต่อภาษาธรรมชาติได้อย่างดีในงานหลายๆด้าน เช่นการจำแนกประเภทประโยค หรือการแปลภาษา ด้วยความสามารถที่กล่าวมาทำให้ BERT เป็นงานวิจัยที่ดีที่สุดในปัจจุบัน

จากปัญหาเรื่องระบบการค้นหาข้อมูลของเว็บไซต์พันทิพและความสามารถของเทคโนโลยีการประมวลภาษาธรรมชาติที่กล่าวมา ผู้จัดทำเล็งเห็นว่าการนำเทคโนโลยี BERT เข้ามามีส่วนในระบบค้นหาของเว็บไซต์พันทิพเพื่อให้สามารถตอบสนองต่อคำค้นที่เป็นประโยคได้ จะส่งผลให้การค้นหาข้อมูลง่ายและตรงตามความต้องการมากยิ่งขึ้น

1.2 วัตถุประสงค์ของปริญญานิพนธ์

1) เพื่อศึกษาเทคโนโลยีการเรียนรู้เชิงลึก
2) เพื่อประยุกต์ใช้เทคโนโลยีการเรียนรู้เชิงลึกในการค้นหาข้อมูลที่มีประสิทธิภาพมากขึ้น
3) เพื่อสร้างแนวทางในการสร้างระบบค้นหาข้อมูลด้วยเทคโนโลยีการเรียนรู้เชิงลึก

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1) ระบบสามารถตอบสนองต่อคำค้นหาที่เป็นภาษาธรรมชาติได้ดี
2) เข้าใจหลักการทำงานของ BERT
3) เข้าใจการนำเทคโนโลยี BERT เข้ามาประยุกต์ใช้งาน
4) เข้าใจการพัฒนาเว็บแอพลิเคชั่นโดยใช้ React GraphQL และ Django

1.4 ขอบเขตของปริญญานิพนธ์

1) ระบบสามารถค้นหาข้อมูลที่เกี่ยวกับการท่องเที่ยวในกรุงเทพฯเท่านั้น
2) บทความที่ถูกใช้ในระบบเป็นเพียงบทความที่มีอยู่ในระบบของเว็บไซต์พันทิพ

1.5 ข้อจำกัดของปริญญานิพนธ์

1) ระบบสามารถตอบสนองต่อความต้องการค้นหาบทความที่มีความเกี่ยวข้องกับแท็กที่ถูกกำหนดเพียง 6 แท็กเท่านั้นซึ่งประกอบไปด้วย crime disease hotel restaurant travel location

1.6 ตารางการดำเนินงาน

ตาราง 1.1 ระยะเวลาการดำเนินในภาคเรียนที่ 1

รายการ	เดือน															
	ส.ค.				ก.ย.				ต.ค.				พ.ย.			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. ค้นหาและเสนอหัวข้อปริญญานิพนธ์ให้กับอาจารย์ที่ปรึกษา																
2. ศึกษา Deep Learning																
3. ศึกษา BERT																
4. ศึกษาการนำ BERT มาใช้งาน																
5. ออกแบบโครงสร้างระบบ																
6. ออกแบบหน้าเว็บแสดงผล ฐานข้อมูล และส่วนการติดต่อของ API																
7. พัฒนาโปรแกรม																

ตาราง 1.2 ระยะเวลาการดำเนินการในภาคเรียนที่ 2

รายการ	เดือน															
	ม.ค.				ก.พ.				มี.ค.				เม.ย			
1. ศึกษาและออกแบบ Database และ API																
2. ศึกษาและทดลอง ElasticSearch																
3. จัดทำระบบ back-end เกี่ยวกับบริการค้นหา																
4. ทำระบบบริการโมเดลผ่าน API																
5. ทำ fronet-end และเชื่อมต่อกับ back-end																
6. ศึกษาและดีพล็อยระบบเพื่อใช้งานจริง																
7. ศึกษา NLP ในการตัดประโยคและ NER																
8. ฝึกสอนและปรับปรุงโมเดลเพิ่มเติม																
9. ศึกษาและจัดทำระบบ ETL																

ตาราง 1.3 ระยะเวลาการดำเนินการในภาคเรียนที่ 2(ต่อ)

รายการ	เดือน															
	ม.ค.				ก.พ.				มี.ค.				เม.ย			
10. จัดทำเล่มรายงาน																

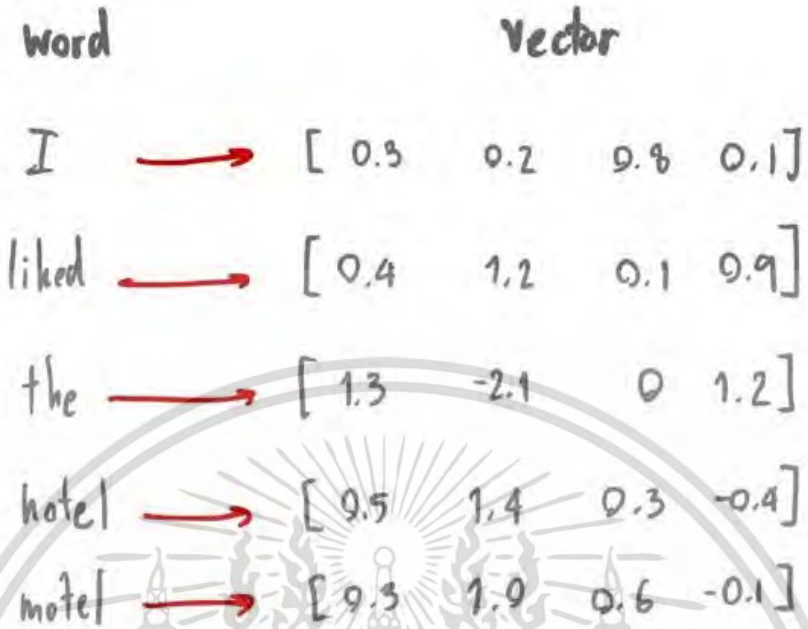
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 Natural Language Processing

การประมวลภาษาธรรมชาติหรือ Natural Language Processing (NLP) เป็นสาขาหนึ่งของเทคโนโลยีปัญญาประดิษฐ์(Artificial Intelligence หรือ AI) โดยจุดประสงค์หลักของการศึกษาก็เพื่อที่จะทำให้คอมพิวเตอร์สามารถเข้าใจ เรียนรู้ หรือสามารถวิเคราะห์ภาษาธรรมชาติได้ การประมวลภาษาธรรมชาตินับเป็นหนึ่งในสาขาที่เป็นปัญหาที่ซับซ้อนยากต่อการแก้ไขเนื่องจากการทำให้คอมพิวเตอร์เรียนรู้ได้เหมือนมนุษย์ที่อ้างอิงจากโครงสร้างของภาษานั้นทำได้ค่อนข้างยาก ดังนั้นศาสตร์การศึกษาในด้านนี้นั้นจะเน้นไปที่การสอนให้คอมพิวเตอร์เรียนรู้ภาษาธรรมชาติในแบบของคอมพิวเตอร์โดยไม่ได้อ้างอิงจากรูปแบบที่มนุษย์เรียนรู้

2.2 Word Embedding

หนึ่งในความพยายามเกี่ยวกับ NLP คือทำให้คอมพิวเตอร์สามารถเข้าใจภาษาธรรมชาติได้แต่ในการประมวลผลโดยคอมพิวเตอร์จำเป็นต้องแปลงข้อมูลอยู่ในรูปแบบที่คอมพิวเตอร์สามารถเข้าใจได้ ดังนั้นการที่จะทำให้คอมพิวเตอร์สามารถเรียนรู้ภาษาได้จึงต้องทำการแปลงคำให้อยู่ในรูปแบบเวกเตอร์(Vector) หรือเรียกกระบวนการนี้ว่า Word Embedding ซึ่งเป็นการนำเสนอคำในรูปแบบตัวเลขที่คอมพิวเตอร์สามารถนำไปใช้ได้ แต่ในการแปลงคำหนึ่งคำนั้นไม่สามารถถูกนำเสนอได้โดยเวกเตอร์เพียงหนึ่งค่า ดังนั้น คำหนึ่งคำจะถูกแทนที่ด้วยชุดของเวกเตอร์



รูป 2.1 รูปแบบการนำเสนอคำที่นำเข้าคอมพิวเตอร์

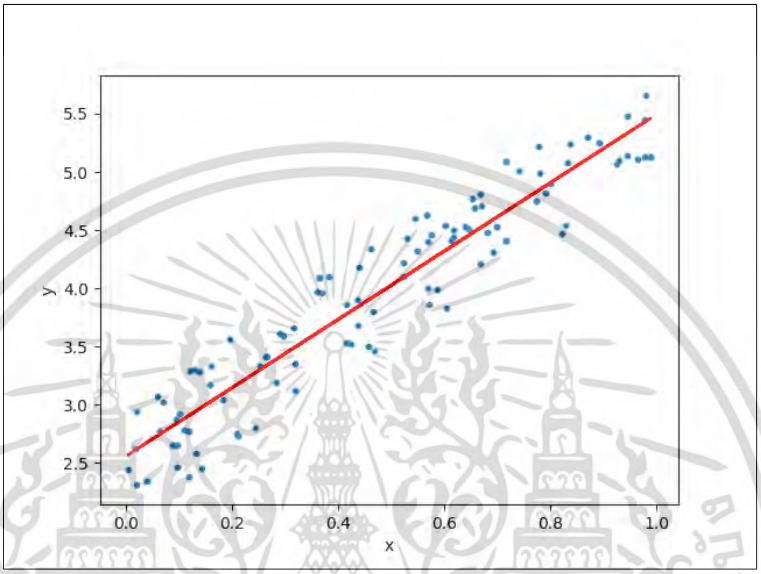
2.3 Machine Learning

การเรียนรู้ของเครื่องหรือ Machine Learning เป็นวิธีการที่ทำให้คอมพิวเตอร์สามารถเกิดการเรียนรู้รูปแบบ โครงสร้าง หรือความสัมพันธ์ของข้อมูลที่มีขนาดใหญ่ได้ด้วยตัวเอง โดยอาศัยการคำนวณทางคณิตศาสตร์ที่มีความซับซ้อน การทำให้เครื่องเกิดการเรียนรู้ด้วยตัวเองนั้นเรียกว่าการสอน(Training) คือการนำข้อมูลป้อนให้เครื่อง ทำให้เครื่องมีการปรับปรุงตัวเองเพื่อให้เหมาะสมกับชุดข้อมูลประเภทนั้นๆ ซึ่งรูปแบบการสอนแบ่งออกได้เป็น 2 ประเภท คือ การสอนแบบมีโครงสร้าง(Supervised Learning) และการสอนแบบไม่มีโครงสร้าง(Unsupervised Learning) โดยการสอนแบบมีโครงสร้างคือผู้สอนรู้ว่าโครงสร้างของข้อมูลเป็นแบบไหน มีรูปแบบอย่างไรบ้าง ส่วนการสอนแบบไม่มีโครงสร้างคือผู้สอนไม่สามารถรู้ถึงความซับซ้อนหรือรูปแบบโครงสร้าง

การสอนแบบมีโครงสร้างนั้นผู้สอนจะต้องทำการเตรียมข้อมูลโดยจัดข้อมูลแบ่งแยกออกเป็นแต่ละหมวดหมู่ชัดเจน(Labeling) ทำให้การเรียนรู้แบบวิธีนี้มีความยากในส่วนเตรียมข้อมูลค่อนข้างมากเพื่อให้ได้ข้อมูลที่เพียงพอต่อการสอน มีการนำการเรียนรู้ในรูปแบบนี้ไปใช้ในการแก้ปัญหาการแยกประเภทวัตถุหรือข้อมูล(Classification) ข้อมูลในปัญหาประเภทนี้เป็นแบบข้อมูลไม่ต้องเนื่อง เช่น การแยกประเภทของสินค้าจากข้อมูลของสินค้า การคาดเดาหุ้นว่าอยู่ในกลุ่มที่ควรซื้อหรือขายจากข้อมูลของตลาดในช่วงเวลาที่ผ่านมาเป็นต้น และการทำนายข้อมูล(Regression) เป็นปัญหาที่ข้อมูลอยู่ในรูปแบบที่มีค่าต่อเนื่อง เช่น ค่าเงิน เป็นต้น

จากรูปที่ 2.2 เป็นวิธีการแก้ปัญหาโดยใช้การเรียนรู้ของเครื่องแบบการสอนแบบมีโครงสร้าง ตำแหน่งจุดในรูปบ่งบอกถึงข้อมูลจริงที่นำมาแสดงบนกราฟ ส่วนเส้นตรงบ่งบอกถึงรูปแบบการนำเสนอข้อมูลของคอมพิวเตอร์ รูปแสดงให้เห็นการแก้ปัญหาโดยการเรียนรู้ของเครื่องคือการพยายาม

นำเสนอข้อมูลให้อยู่ในรูปแบบของสมการทางคณิตศาสตร์ จะเห็นได้ว่าคอมพิวเตอร์จะสามารถทำนายค่าของ x ที่ ตำแหน่งถัดจาก 1 ได้ ค่อนข้างแม่นยำ แต่ในความเป็นจริงการนำเสนอข้อมูลหรือการเลือกสมการทางคณิตศาสตร์มาเพื่ออธิบายเป็นส่วนที่ยากและสำคัญที่สุดในการแก้ปัญหา



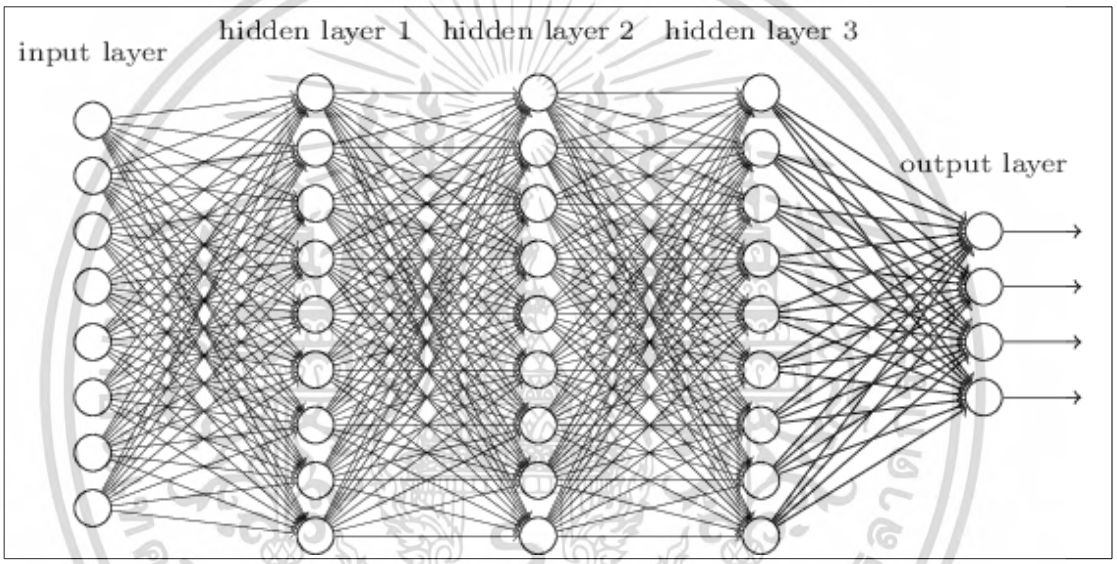
รูป 2.2 ตัวอย่างการแก้ปัญหาแบบ Regression

การสอนแบบไม่มีโครงสร้างเป็นปัญหาที่ข้อมูลมีขนาดใหญ่ซับซ้อน โครงสร้างไม่ชัดเจนหรือยากต่อการจัดหมวดหมู่ ข้อมูลเหล่านี้เป็นข้อมูลที่จำเป็นต้องอาศัยการอ้างอิงจากโครงสร้างมากกว่าหนึ่งโครงสร้างจึงจะสามารถระบุเป็นตัวตนของข้อมูลนั้นๆได้ ยกตัวอย่างเช่น ในการบอกว่ารูปภาพสุนัขหนึ่งรูปเป็นรูปสุนัขจำเป็นต้องบอกว่าในรูปมีส่วนของแขน ขา ปากที่ยื่น จมูก หาง ตา หู หรือรวมทั้งอื่นๆอีกมากมาย ส่งผลให้การเรียนรู้ของเครื่องที่สามารถแก้ปัญหาได้โดยการนำข้อมูลจากโครงสร้างเพียงไม่กี่ส่วนมาเพื่อแก้ปัญหานั้นไม่สามารถนำมาแก้ปัญหาในระดับนี้ได้ จึงมีการพัฒนาการเรียนรู้ที่เพิ่มความสามารถขึ้นมาเรียกว่า การเรียนรู้เชิงลึก(Deep Learning) ซึ่งเป็นการนำการเรียนรู้ของเครื่องหลายๆ

2.4 Deep learning

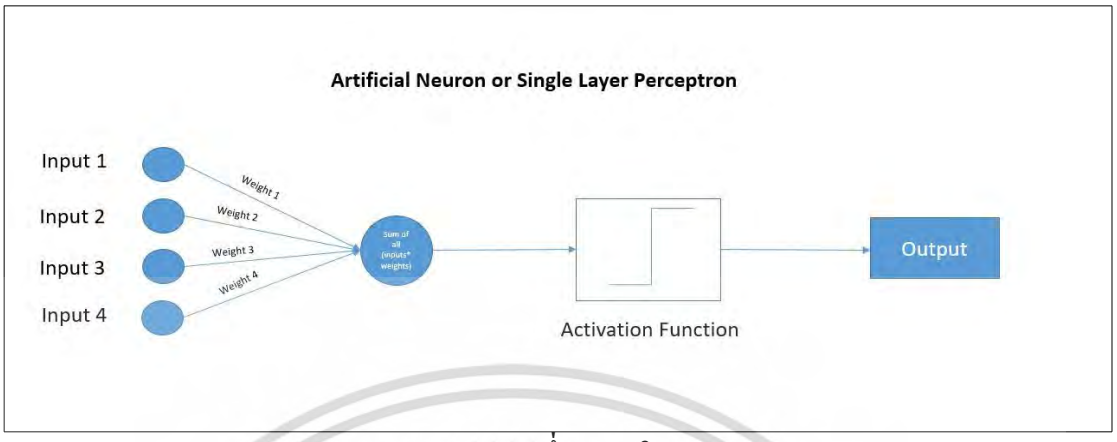
การเรียนรู้เชิงลึกหรือ Deep Learning เป็นการพัฒนาอีกขั้นของการเรียนรู้ของเครื่องเพื่อที่จะแก้ปัญหาที่มีความซับซ้อนมากขึ้น โดยการเรียนรู้เชิงลึกนั้นเป็นการนำการเรียนรู้ของเครื่องจำนวนมากมารวมกัน ส่งผลให้การเรียนรู้เชิงลึกสามารถจำแนกรูปแบบโครงสร้าง ความสัมพันธ์ข้อมูลได้มากกว่า จากรูป 2.3 จะเห็นได้ว่าโครงสร้างของการเรียนรู้เชิงลึกนั้นประกอบด้วย 3 ส่วน คือ ส่วนรับข้อมูลเข้า(Input Layer) ส่วนที่มีการปรับปรุงตัวเองและทำหน้าที่แยกแยะโครงสร้างข้อมูล(Hidden Layer) และส่วนผลลัพธ์(Output Layer) หน้าที่ของแต่ละส่วนจะแตกต่างกันออกไป

โดยส่วนที่สำคัญที่สุดจะเป็นส่วนที่อยู่ตรงกลางของโครงสร้างคือส่วนที่มีการรับปรุงตัวเอง ในส่วนนี้โครงสร้างจะถูกแบ่งออกเป็นหลายชั้นขึ้นอยู่กับความซับซ้อนของปัญหาที่แก้ ซึ่งในแต่ละชั้นเรียกได้ว่าเป็นส่วนของการแยกแยะโครงสร้างข้อมูลก็ได้ กล่าวคือ ปัญหาในการแยกแยะรูปภาพสัตว์ว่าเป็นประเภทไหนหรือพันธุ์ไหน ส่วนตรงกลางนี้จะทำหน้าที่แยกแยะข้อมูลในแต่ละชั้นต่างกันไป เช่น ชั้นแรก ระบุว่ารูปแบบขาเป็นแบบไหน ชั้นที่สองระบุว่าใบหูเป็นแบบไหน เป็นต้น นั้นส่งผลให้การเรียนรู้เชิงลึกสามารถนำโครงสร้างของข้อมูลที่มากกว่าหนึ่งโครงสร้างมาแก้ปัญหาได้ ส่วนอีกปัญหาที่มีความซับซ้อนและสามารถนำไปใช้ได้อย่างกว้างขวางในอนาคตที่จำเป็นต้องใช้การเรียนรู้เชิงลึกเข้ามาแก้อีกปัญหาคือ การประมวลภาษาธรรมชาติ (Natural Language Processing)



รูป 2.3 แสดงการทำงานของ Deep Learning

การเรียนรู้เชิงลึกเปรียบเสมือนการจำลองการทำงานของสมองมนุษย์เพื่อให้สามารถทำงานที่ซับซ้อนได้ โครงสร้างประกอบด้วยเครือข่ายของกระแสประสาทจำนวนมากแบ่งออกเป็นชั้นๆ ซึ่งแต่ละหน่วยของกระแสประสาท(Perceptron) จะหมายถึงการเรียนรู้ของเครื่อง



รูป 2.4 การทำงานของหนึ่งหน่วยในกระแสประสาท

ลำดับการทำงานของการเรียนรู้เชิงลึกคือข้อมูลที่ถูกส่งเข้าไปยังเครื่องจะผ่านไปยังแต่ละหน่วยของกระแสประสาท (Perceptron) ซึ่งในแต่ละหน่วยจะประกอบด้วยสมการทางคณิตศาสตร์(Activation Function) และเส้นประสาทที่เชื่อมโยงกับแต่ละหน่วยของกระแสประสาทจะมีค่าประจำตัว(Weight) ที่จะเกิดการปรับตัวเพื่อให้เหมาะสมกับการแก้ปัญหานั้นๆได้ ซึ่งการปรับตัวเกิดจากการคำนวณผลลัพธ์ที่ได้จากสมการทางคณิตศาสตร์ที่ผ่านการคำนวณเทียบกับผลลัพธ์ที่ควรได้ นั้นส่งผลให้เครื่องมีการปรับตัวไปในทางที่ถูกต้องได้และข้อมูลที่จำเป็นต้องใช้ในการเรียนรู้จึงต้องประกอบด้วยสองส่วน คือส่วนที่เป็นข้อมูลที่ทำให้คอมพิวเตอร์ทำนายผลลัพธ์(Input) และส่วนที่ทำให้คอมพิวเตอร์มีการปรับปรุงตัวเอง(Output) ที่ถูกในกระบวนการเทียบค่าความผิดพลาด

กระบวนการปรับปรุงตัวเองเป็นส่วนที่สำคัญที่สุดของการเรียนรู้ด้วยตัวเองหรือการเรียนรู้เชิงลึก หลักการทำงานของกระบวนการนี้คือเพื่อที่จะลดค่าความผิดพลาดและเพิ่มความแม่นยำในการคาดเดาค่าผลลัพธ์ การทำกระบวนการนี้ประกอบด้วยค่าอย่างน้อย 2 ค่า คือผลลัพธ์จริงกับผลลัพธ์ที่ได้จากการเรียนรู้ โดยทำการเปรียบเทียบผลลัพธ์ทั้งสองเพื่อหาแนวทางในการปรับลดค่าผิดพลาด ซึ่งในการเรียนรู้เชิงลึกนั้นจะมีกระบวนการที่เรียกว่า Backward Propagation of Errors หรือ Backpropagation เป็นกระบวนการที่จะปรับค่าน้ำหนักของเส้นประสาทแต่ละเส้น หรือที่เรียกว่าการปรับปรุงตัวเอง กระบวนการนี้จะทำตามลำดับมาจากส่วนปลายของโครงสร้างเรื่อยๆมาที่ส่วนต้น แต่การปรับอาจจะไม่ได้เป็นไปในทางเดียวกันหมดขึ้นอยู่กับรูปแบบหรือหน่วยกระแสประสาทที่เชื่อต่ออยู่

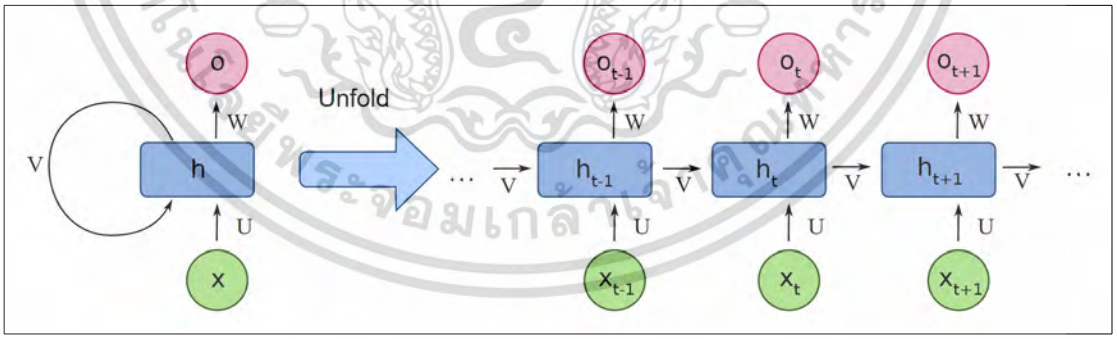
2.4.1 RNN

Recurrent Neural Network (RNN) หรือ อาร์เอ็นเอ็น การเรียนรู้เชิงลึกประเภทหนึ่งที่ถูกนำมาใช้แก้ปัญหาที่มีลักษณะเป็นลำดับ อาร์เอ็นเอ็นถูกนำเข้ามาแก้ไขปัญหาด้านการประมวลภาษาธรรมชาติเป็นหลัก โดยวิธีการที่ทำให้อาร์เอ็นเอ็นเกิดการเรียนรู้ด้วยตนเอง (Training) คือเครือข่ายประสาทเทียมจะทำการรับข้อมูล (X) เข้าเป็นลำดับ(ตามคำในประโยค) ผ่านส่วนที่เป็นการเรียนรู้ด้วยตัวเอง(h) และมีการปรับค่าน้ำหนัก(Backpropagation Throught Time) ทุกค่าที่อยู่ในเครือข่ายประสาทเทียมทุกๆครั้งที่จบการทำงาน 1 รอบ โดย 1 รอบการทำงานคือ รับข้อมูล(1 คำจากในประโยค) ผ่านเข้าในเครือข่าย แล้วทำการปรับปรุงตัวเอง

ในรูป 2.5 ได้แสดงวิธีการทำงานของอาร์เอ็นเอ็นไว้ คือ ข้อมูล(X) จะผ่านการคำนวณโดยน้ำหนักตามค่าประจำหน่วยของกระแสประสาท(U) แล้วได้ผลลัพธ์ส่งเข้าไปที่สเตท(State) ปัจจุบัน(h) ซึ่งเป็นเครือข่ายจำนวนมากประกอบด้วยส่วนที่เป็นฟังก์ชั่นกระตุ้น(Activation Function)หรือหน่วยกระแสประสาทที่กล่าวไปก่อนหน้านี้ และน้ำหนักประจำเส้นประสาทที่เชื่อมโยงแต่ละเส้นประสาท ในส่วนสเตทจะทำการคำนวณค่าที่ได้รับจากส่วนก่อนหน้าและค่าที่ได้จากสเตทก่อนหน้าคำนวณกับค่าน้ำหนักของสเตทปัจจุบันแสดงได้สมการดังนี้

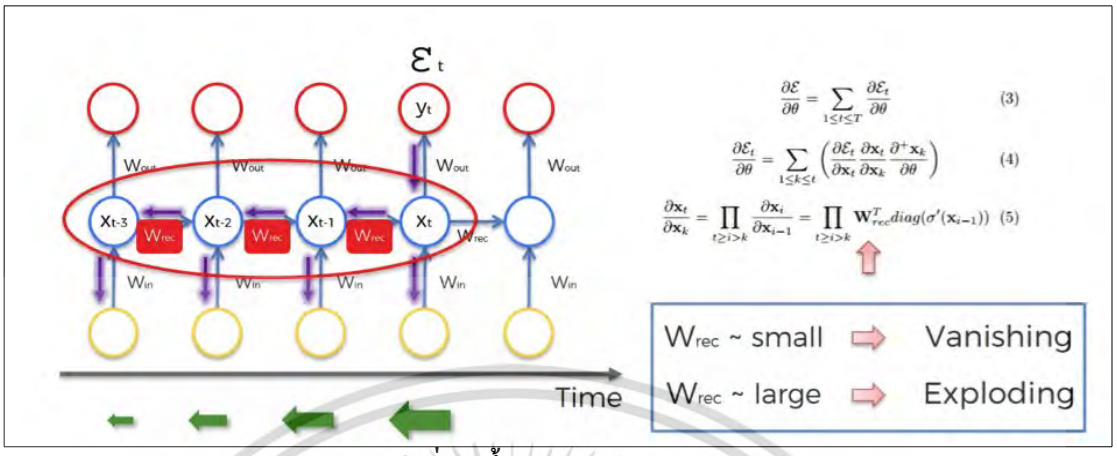
$$h_t = \tanh\left(V\, h_{(t-1)} + U\, X_t\right)$$

ในที่นี้จะทำการกำหนดให้ฟังก์ชั่นกระตุ้นเป็นฟังก์ชั่นไฮเพอโบลิก(Hyperbolic function) เพื่อลดความสับสนจึงเลือกยกตัวอย่างขึ้นมาหนึ่งฟังก์ชั่น แต่ในความเป็นจริงสามารถเลือกใช้ฟังก์ชั่นอื่นได้ และส่วนสุดท้ายคือ ผลลัพธ์ที่ได้จากการนำค่าที่สเตทปัจจุบันไปคำนวณกับค่า W



รูป 2.5 โครงสร้างและการทำงานของ RNN

ในกระบวนการคำนวณค่าความผิดพลาดของหลักการเรียนรู้เชิงลึก อาร์เอ็นเอ็นได้ปรับลดค่าน้ำหนักใน U และ V โดยใช้หลักการเคลื่อนลงตามความชัน(Gradient Descent) ที่จะทำการคำนวณโดยการหาอนุพันธ์(Derivative) ซึ่งในการคำที่ตำแหน่งสเตทปัจจุบันนั้นคือค่าที่ได้จากผลรวมของความผิดพลาดในสเตทที่ผ่านมาทั้งหมดด้วย ค่าความผิดพลาดเหล่านี้จะถูกนำไปคำนวณกับน้ำหนัก U และ V ซึ่งถ้าค่ามีค่าน้อยจะทำข้อมูลน้ำหนักบางตำแหน่งหายไปเรียกว่า Vanishing หรือถ้าค่ามากไปจะเรียกว่า Exploding ดังแสดงในรูป 2.6



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \quad (3)$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \quad (4)$$

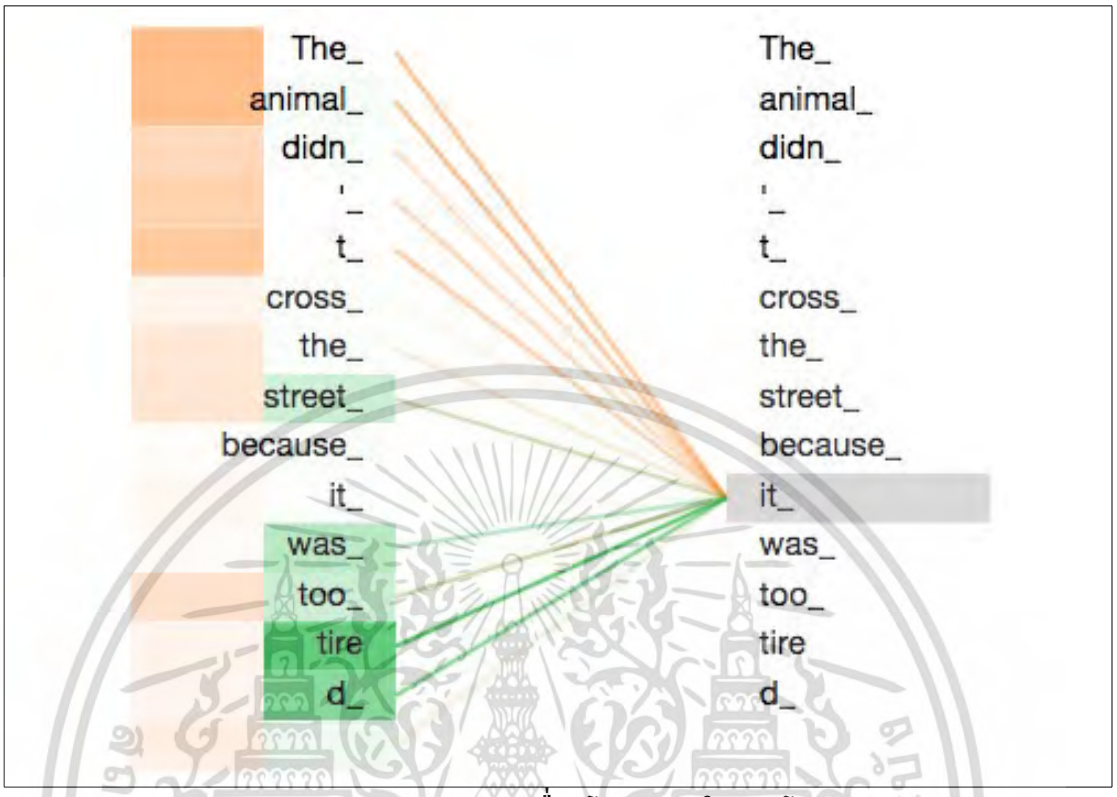
$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \quad (5)$$

รูป 2.6 ปัญหาที่เกิดขึ้นการปรับปรุงตัวเองของ RNN

2.4.2 Transformer and Attention

Transformer หรือ ทรานฟอร์เมอร์ โมเดลการเรียนรู้เชิงลึกที่ถูกพัฒนาขึ้นมาเพื่อแก้ปัญหาด้าน NLP ซึ่งจะมีรูปแบบโครงสร้างและการสอนให้เรียนรู้ที่ต่างกัน โดยโครงสร้างจะถูกแบ่งออกเป็น 2 ส่วน คือ Encoder และ Decoder ในแต่ละส่วนจะมีการซ้อนทับกันมากกว่า 1 ชั้น โดยการทำงานในส่วนการเข้ารหัส(encoder) จะเป็นการหาความสัมพันธ์ระหว่างคำในประโยค และส่วนถอดรหัส(decoder) จะทำหน้าที่แปลงข้อมูลที่ได้รับจากส่วนการเข้ารหัสออกมา

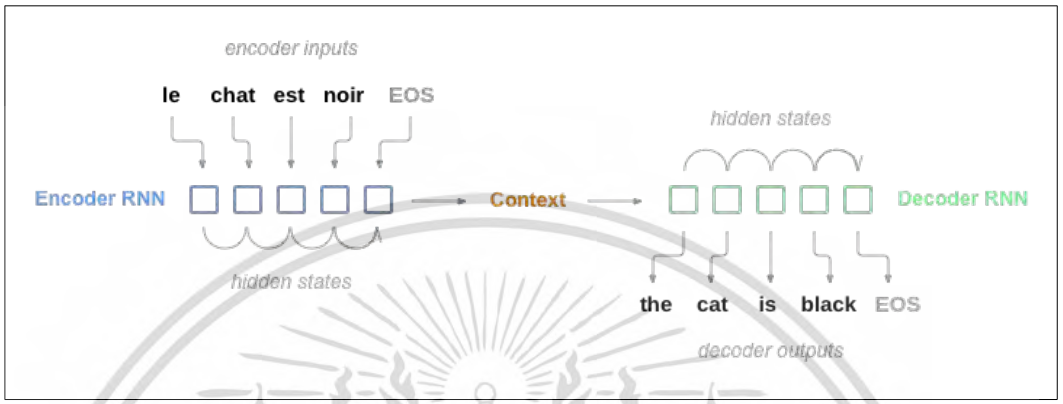
ส่วนประกอบที่สำคัญของทรานฟอร์เมอร์คือ แอทเทนชั่น(Attention) เป็นอัลกอริทึมที่บริษัทกูเกิลพัฒนาขึ้นมาเพื่อให้คอมพิวเตอร์สามารถให้ความสนใจกับคำที่เป็นใจความสำคัญของประโยคได้ ทำให้คอมพิวเตอร์สามารถระบุความเชื่อมโยงของคำแต่ละคำในประโยคได้ ด้วยอัลกอริทึมนี้ทำให้คอมพิวเตอร์สามารถเรียนรู้ได้คล้ายมนุษย์มากขึ้นคือจากเดิมที่เป็นการพิจารณาคำจากบริบทเพียงด้านเดียวคือข้อมูลผ่านเข้าเครือข่ายก่อนหน้า เปลี่ยนมาเป็นการเรียนรู้ที่เป็นแบบคิดตามบริบทโดยรอบทั้งด้านซ้ายและขวา เรียกว่า Bidirectional



รูป 2.7 ตัวอย่างความเชื่อมโยงของคำในประโยค

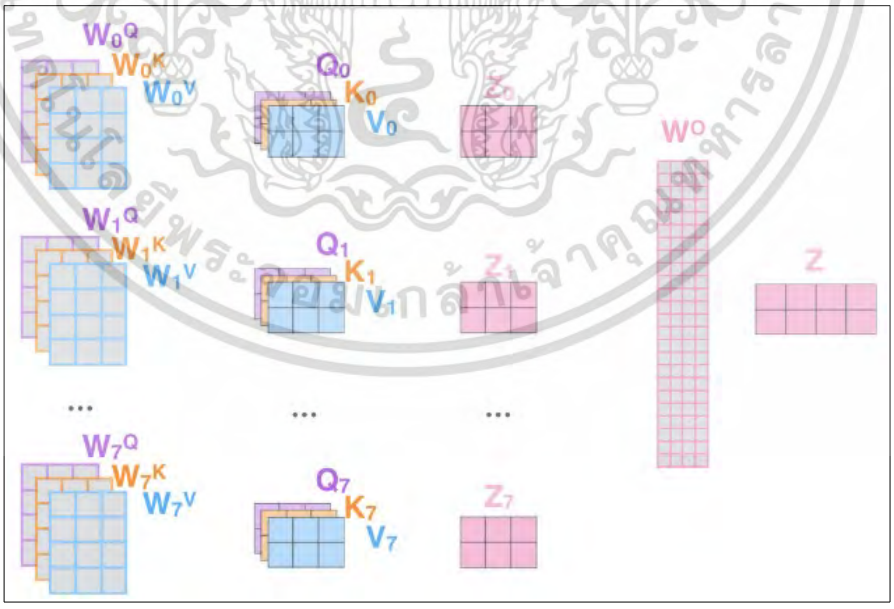
รูปแบบการเรียนรู้ของโมเดลจะต่างไปจากอาร์เอ็นเอ็น คือ อาร์เอ็นเอ็นจะรับข้อมูลเข้าเป็นลำดับ แต่การรับข้อมูลเข้าของทรานฟอร์เมอร์จะรับข้อมูลเข้าพร้อมกันทั้งหมดและใช้แอทเทนชั่นเข้ามามีส่วนในการสร้างความสัมพันธ์ของคำในประโยค โครงสร้างทรานฟอร์เมอร์ทั้งส่วนการเข้ารหัสและถอดรหัสจะมีส่วนประกอบที่เป็นการใช้งานของแอทเทนชั่นทั้งคู่ ส่วนการแปลงข้อมูลขาออกจะยังแปลงออกเป็นลำดับเหมือนอาร์เอ็นเอ็นโดยใช้ข้อมูลขาออกตัวก่อนหน้ามามีส่วนในการแปลงข้อมูลด้วย การสอนรูปแบบนี้ทำให้ทรานฟอร์เมอร์แก้จุดอ่อนของอาร์เอ็นเอ็นที่ไม่สามารถงานแบบขนานได้ โครงสร้างของทรานฟอร์เมอร์ประกอบด้วยชั้นของการเข้ารหัสมากกว่าหนึ่งชั้น และชั้นของการถอดรหัสมากกว่าหนึ่งชั้น ด้วยโครงสร้างแบบนี้ทำให้ทรานฟอร์เมอร์สามารถแก้ปัญหาของอาร์เอ็นเอ็น คือ vanishing และ exploding ได้ เพราะว่าการคำนวณค่าความผิดพลาดและการปรับปรุงตัวเองถูกแยกออกเป็นชั้นๆต่างจากอาร์เอ็นเอ็นที่เป็นเครือข่ายเชื่อมโยงกันทั้งหมด ในโครงสร้างที่มีการแบ่งตัวออกเป็นชั้นๆทำให้แต่ละชั้นจำเป็นต้องส่งค่าผ่านแต่ละชั้น ซึ่งรูปแบบหรือโครงสร้างของข้อมูลจะอยู่ในรูปแบบเดียวกับข้อมูลที่รับเข้ามา เช่น ปัญหาด้านการประมวลภาษาธรรมชาติที่ข้อมูลเข้าคอมพิวเตอร์เป็นเวกเตอร์ 1 มิติ ข้อมูลที่ส่งผ่านในระหว่างชั้นของแต่ละชั้นก็จะเป็นเวกเตอร์ 1 มิติ และระหว่างส่วนการเข้าหรัสและส่วนถอดรหัสนั้นจะมีการส่งค่าจากเพียงชั้นสุดท้ายของส่วนการเข้ารหัสไปให้ส่วนการถอดรหัส และส่วนการถอดรหัสจะยังทำการเรียนรู้ในแบบเดียวกับอาร์เอ็นเอ็นคือการเรียนรู้ที่จะคาดเดาคำต่อไปโดยใช้ข้อมูลที่ทำนาย

ออกมาก่อนหน้าส่งเข้าส่วนการถอดรหัสและข้อมูลที่ได้จากส่วนการเข้ารหัสประกอบกันสองส่วนในการทำนายคำถัดไป



รูป 2.8 โครงสร้างและการทำงานของ Transformer

มัลติเฮด(Multi-Head) หนึ่งในส่วนที่มีความจำเป็นในการสร้างรูปแบบของผลลัพธ์ที่หลากหลายที่ถูกนำมาใช้โครงสร้างแอทเทนชั่น ในการคำความสัมพันธ์ของคำจากอัลกอริทึมของแอทเทนชั่นนั้นจะมีการส่งข้อมูลเข้าไปในเครือข่ายที่ต่างกัน(Head) มากกว่า 1 ชุด(Multi) แล้วนำผลลัพธ์ที่ได้จากแต่ละเครือข่ายมาผ่านการคำนวณจึงจะได้ผลลัพธ์ที่ถูกนำไปใช้ในชั้นต่อๆไป



รูป 2.9 แสดงการสร้าง multi-head

2.4.3 BERT

BERT หรือ Bidirectional Encoder Representations from Transformers โมเดลการเรียนรู้เชิงลึกที่ถูกพัฒนาโดยบริษัทกูเกิล ถูกพัฒนาขึ้นมาเพื่อนำมาใช้แก้ไขปัญหาด้านการประมวลภาษาธรรมชาติโดยเฉพาะโครงสร้างของ BERT ประกอบด้วยเพียงโครงสร้างของส่วนที่เป็นการเข้ารหัสจากทรานฟอร์เมอร์ที่ซ้อนทับกันมากกว่า 1 ชั้น ตามหลักการของการเรียนรู้เชิงลึก

BERT ถูกยอมรับว่าเป็นโมเดลที่มีประสิทธิภาพมากที่สุดในปัจจุบันเนื่องจากรูปแบบวิธีการสอนให้โมเดลเกิดการเรียนรู้นั้นเรียกว่า Masked Language Modeling (MLM) ซึ่งต่างจากโมเดลอื่นๆที่ทำการสอนโดยให้ทำนายคำในลำดับถัดจากผลลัพธ์ปัจจุบัน รูปแบบ MLM คือการปิดคำในประโยคแล้วให้โมเดลพยายามทำการทำนายคำที่ถูกปิดไว้โดยคาดเดาจากบริบทรอบข้างและการสอนที่เป็นส่วนสำคัญในการนำไปใช้ในงานหลายๆด้านคือการสอนโดยวิธีที่เรียกว่า Next Sentence Prediction หรือการพยายามตอบให้ได้ว่าสองประโยคที่ถูกนำเข้าโมเดลเป็นประโยคที่ต่อเนื่องกันหรือไม่ จากการสอนด้วยสองวิธีนี้ทำให้โมเดลนี้กลายเป็นโมเดลที่มีความแม่นยำมากที่สุดในปัจจุบัน

BERT ถูกสร้างขึ้นด้วยโมเดลสองขนาด คือขนาดเล็ก(BERT-Base) กับขนาดใหญ่(BERT-Large) โดยในการสร้างขนาดเล็กมาก็เพื่อที่จะเปรียบเทียบประสิทธิภาพของโมเดลที่มีอยู่ก่อนหน้าอย่าง OpenAI ที่ถูกพัฒนาโดยใช้โครงสร้างของส่วนการถอดรหัสจากทรานฟอร์เมอร์ BERT สามารถนำไปใช้ในงานหลายด้านเช่นการจัดหมวดหมู่ประเภทประโยค(Classification Task) การถามตอบ(QA Task) และอื่นๆ

2.5 React

รีแอคท์ หรือ React เป็น Javascript Library ที่ถูกเริ่มพัฒนาโดยบริษัท Facebook มีคำสั่งสามารถสร้างส่วนติดต่อผู้ใช้ การจัดการข้อมูล โดยในการสร้างส่วนติดต่อผู้ใช้นั้นรีแอคท์จะทำการสร้าง Virtual DOM (Virtual Document Object Model) ที่เป็นตัวแทนของ DOM (Document Object Model) โดยหากมีการปรับปรุงส่วนติดต่อผู้ใช้รีแอคท์จะทำการแก้ที่ Virtual Dom แล้วทำการส่งส่วนที่ถูกแก้ไขกลับไปยัง DOM ทำให้ส่วนติดต่อผู้ใช้มีการเปลี่ยนแปลงที่รวดเร็ว ซึ่งโดยปกติแล้วรีแอคท์จะทำการอัพเดตและประมวลผลเพื่อแสดงใน Component นั้นตามโครงสร้างข้อมูล(Data Model) โดยที่เปลี่ยนแปลงไปอย่างอัตโนมัติ

2.6 GraphQL

GraphQL (Graph Query Language) เป็นภาษาสำหรับการคำค้นและภาษาจัดการข้อมูล (Data manipulation language) ที่ถูกเริ่มพัฒนาโดยบริษัท Facebook ทำงานเป็น API (Application Programming Interface) ซึ่งในการพัฒนาเว็บไซต์นั้น จะเป็นส่วนติดต่อสื่อสารระหว่าง Web Client (Front-end) กับ Web Application (Back-end) ในโครงสร้างแบบ 3-Tier

GraphQL นั้นจะมีโครงสร้างสำหรับการค้น ซึ่งยืดหยุ่นกว่าวิธี REST-ful โดยที่ในการค้น 1 โครงสร้างข้อมูล(Data Model) สามารถกำหนดได้ว่าต้องการตอบรับข้อมูลส่วนไหนกลับมา สามารถค้นโครงสร้างข้อมูลที่ซ้อนกันได้และสามารถค้นโครงสร้างข้อมูลได้มากกว่า 1 โครงสร้าง ทำให้สามารถดึงข้อมูลที่ต้องการได้ใน 1 ครั้ง ซึ่งการทำงานเหมือนเป็น getway สำหรับการค้นข้อมูล นอกจากนี้ยังสามารถทำบริการ Subscribe ซึ่งเป็นการส่ง/รับข้อมูลแบบทันที(Real-Time)

2.7 Django

Django เป็น web-framwork บนภาษา python ซึ่งถูกพัฒนาตามแนวคิด MTV (Model Template View) ซึ่งได้เตรียมคำสั่งการทำงานเบื้องหลังไว้เกือบสมบูรณ์แล้ว ทั้งการจัดการข้อมูล การประมวลผลในระบบ การแสดงผล การจัดการยูอาร์แอล(URL) และการส่งข้อมูล การจัดการแคช(Cache) ทำให้นักพัฒนานั้นสามารถ เน้นไปยังการออกแบบระบบ และเขียนคำสั่งน้อย

2.8 MongoDB

MongoDB เป็น document-database ที่จัดอยู่ในกลุ่ม ภาษาค้นข้อมูลแบบไม่มีโครงสร้าง (Non-Structured Query Language)

document-database เป็นการจัดเก็บข้อมูลแบบ key-value ที่สามารถทำการซ้อนกัน โดยในแต่ละ document นั้นจะมีโครงสร้าง(Schema) เป็นของตนเอง ซึ่งจะมี unqiue-key เป็นส่วนระบุตัวตน ซึ่งโครงสร้างนั้นจะมีความคล้ายคลึงกับฟอร์แมตการสื่อสารข้อมูลทั่วไปอย่างเช่น JSON หรือ XML

Non-SQL นั้นง่ายต่อการพัฒนาทั้งการจัดเก็บและการค้น เพราะไม่ต้องกำหนดโครงสร้างฐานข้อมูลตั้งแต่ต้นเหมือน SQL มีข้อดีที่สามารถทำการขยายได้ดีในแนวนอน มีประสิทธิภาพสูงสำหรับโมเดลบางโมเดล เช่น งานเอกสาร งานผลการทำงานของระบบ(Log) และ Non-SQL นั้นให้ประสิทธิภาพการอ่านข้อมูลแบบง่ายได้รวดเร็วกว่า SQL

2.9 Elastic Search

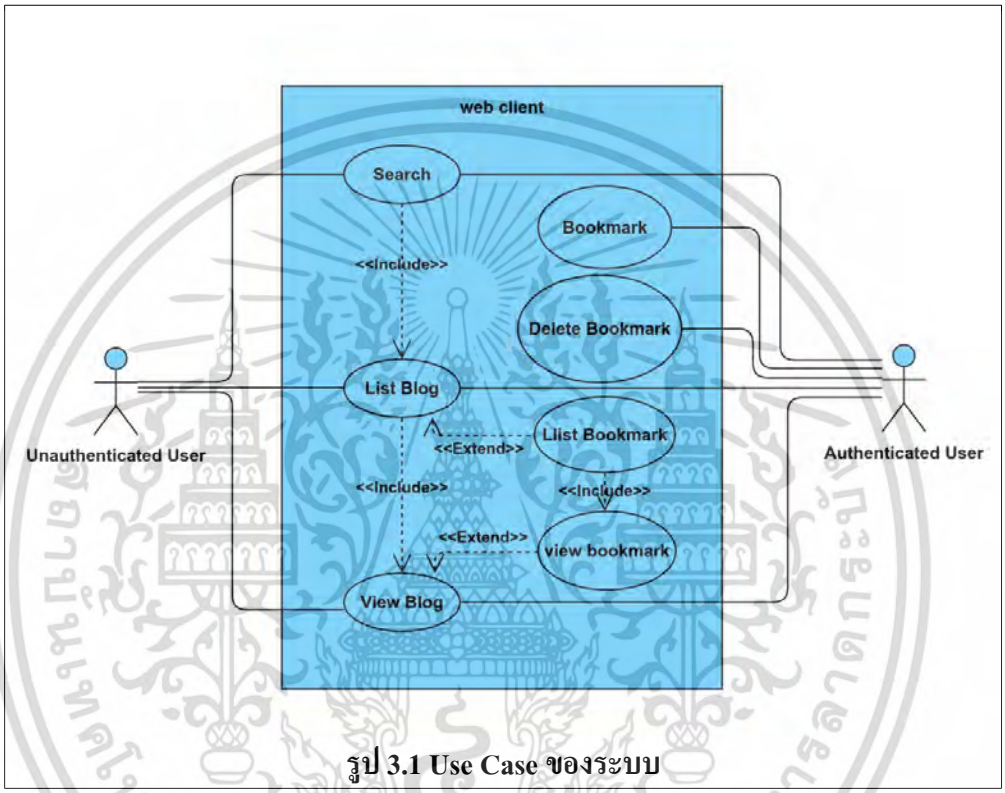
ElasticSearch เป็น application search engine ที่ค้นข้อมูลได้เกือบทันทีทันใด ซึ่งพื้นฐานของระบบเป็นการเก็บข้อมูลและการสืบค้นแบบกระจาย ทำให้สามารถปรับขนาด และสำรองระบบได้ และมีเอชทีทีพีและเอพีไอติดมาให้ทำให้สามารถใช้งานได้สะดวก

มีความสามารถครบจงวรในโดเมนของการค้นข้อมูล แต่ตั้งการจัดเก็บข้อมูล การจัดการแบ่งข้อมูลเพื่อทำ index และการคืนผลลัพธ์จากคำ ในการจัดเก็บเอกสาร สามารถรับชนิดของ field ของเอกสารได้หลายชนิด และไม่จำเป็นที่จะต้องกำหนดชนิดของเอกสารตั้งแต่แรก ตัว elasticsearch นั้นจะทำการสร้างและผูกไว้ให้เอง ซึ่งเมื่อนำเข้าเอกสาร elasticsearch จะทำกระบวนการสร้าง inverted index ให้โดยอัตโนมัติ โดยเราสามารถเลือกอัลกอริทึมมาตฐานในการทำ index หรือสร้างขึ้นมาใหม่ ปรับแต่งให้เหมาะสมกับแต่ละงานได้

ในการบวนการสิบค้นข้อมูลนั้นมีเอชทีทีพีและเอพีไอติดมาให้ ทำให้ไม่จำเป็นที่ต้องมีเอสดีเค(Software Development Kids ; SDK) ของแต่ละภาษาออกมา ซึ่งจากตัวเอพีไอนี้สามารถสร้างคำค้นที่ซับซ้อนได้ ให้เปรียบเทียบใกล้เคียงกำภาษาคำค้น SQL ซึ่งสามารถให้ผลค้นที่เวลาทันทีทันใดได้

บทที่ 3
การออกแบบและพัฒนา

3.1 Use Case ของระบบ



รูป 3.1 Use Case ของระบบ

ตาราง 3.1 Use Case การค้นหาบทความ

ID	Use case
Title	การค้นหาบทความ
Description	ค้นหาบทความ
Actor	Unauthenticated User, Authenticated User
Pre-conditions	-
Post-conditions	แสดงชุดของบทความที่เกี่ยวข้อง
Main success scenario	1. ผู้ใช้เลือกใช้งานฟังก์ชั่นการค้นหา 2. ผู้ใช้ใส่คำค้นหาที่เป็นคำพูดภาษาไทย 3. ระบบแสดงผลชุดของบทความที่เกี่ยวข้อง

ตาราง 3.2 Use Case การเรียกดูบทความ

ID	Use case
Title	การเรียกดูบทความ
Description	เรียกดูบทความหนึ่งๆ
Actor(s)	Unauthenticated User, Authenticated User
Pre-conditions	ระบบแสดงชุดของบทความ
Post-conditions	ระบบแสดงบทความหนึ่งๆ
Main success scenario	1. ผู้ใช้เลือกดูบทความที่เฉพาะเจาะจง 2. ระบบแสดงรายละเอียดของบทความ

ตาราง 3.3 Use Case การบันทึกบทความ

ID	Use case
Title	การเรียกดูบทความ
Description	การบันทึกบทความ
Actor(s)	Authenticated User
Pre-conditions	ระบบแสดงชุดของบทความ
Post-conditions	ระบบทำการบันทึกบทความสำหรับผู้ใช้
Main success scenario	1. ผู้ใช้ค้นหาชุดของข้อมูล 2. ผู้ใช้เลือกดูบทความที่เฉพาะเจาะจง 3. ผู้ใช้เลือกไอคอนที่เป็นปุ่มกดบันทึกบทความ

ตาราง 3.4 Use Case การลบบทความ

ID	Use case
Title	การเรียกดูบทความ
Description	ผู้ใช้ลบบทความที่ถูกบันทึกไว้
Actor(s)	Authenticated User
Pre-conditions	ระบบแสดงชุดของบทความที่เคยถูกบันทึกโดยผู้ใช้
Post-conditions	ระบบทำการลบบทความที่เคยถูกบันทึกไว้ถูกต้อง
Main success scenario	1. ผู้ใช้เลือกดูบทความที่บันทึกไว้ 2. ระบบแสดงรายละเอียดของบทความพร้อมกับไอคอมลบบทความ 3. ผู้ใช้กดเลือกไอคอมเพื่อทำการลบบทความที่ถูกบันทึก

ตาราง 3.5 Use Case เรียกดูชุดของบันทึกบทความ

ID	Use case
Title	การเรียกดูบันทึกบทความ
Description	ผู้ใช้เรียกดูบทความที่ถูกบันทึกไว้ทั้งหมด
Actor(s)	Authenticated User
Pre-conditions	-
Post-conditions	ระบบแสดงชุดของบทความที่ถูกบันทึกสำหรับผู้ใช้
Main success scenario	1. ผู้ใช้เลือกเมนูไปสู่หน้าที่แสดงบทความที่ถูกบันทึก 2. ระบบแสดงรายละเอียดของบทความ

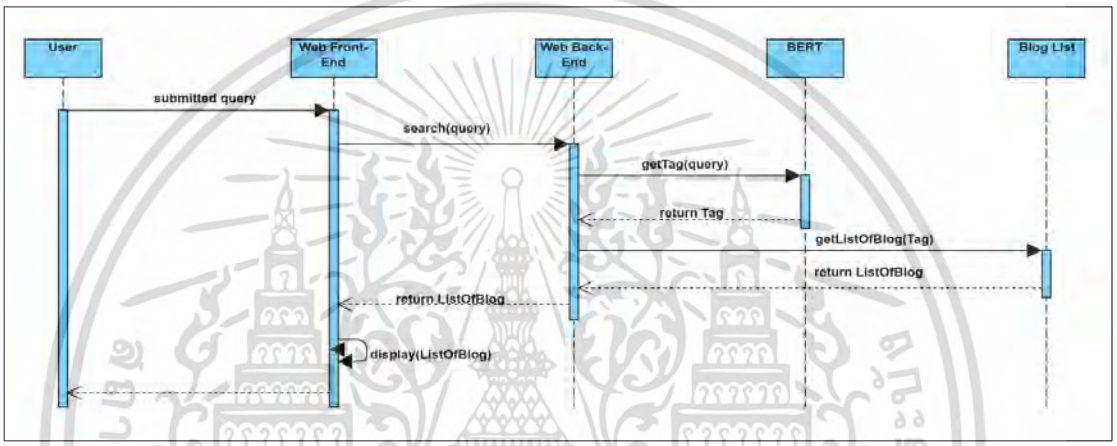
ตาราง 3.6 Use Case การเรียกดูบันทึกบทความ

ID	Use case
Title	การเรียกดูบทความที่ถูกบันทึกไว้
Description	เรียกดูบทความหนึ่งๆ
Actor(s)	Authenticated User
Pre-conditions	ระบบแสดงชุดของบทความที่ถูกบันทึก
Post-conditions	ระบบแสดงบทความหนึ่งๆ

ตาราง 3.7 Use Case การเรียกดูบันทึกบทความ(ต่อ)

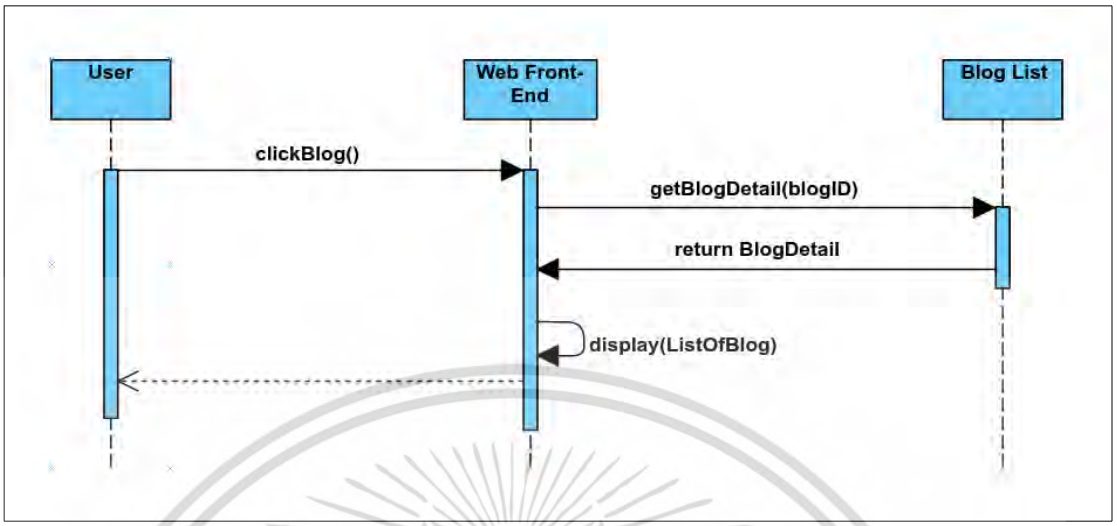
ID	Use case
Main success scenario	1. ผู้ใช้เลือกดูบทความที่เฉพาะเจาะจง 2. ระบบแสดงรายละเอียดของบทความ

3.2 sequence diagram



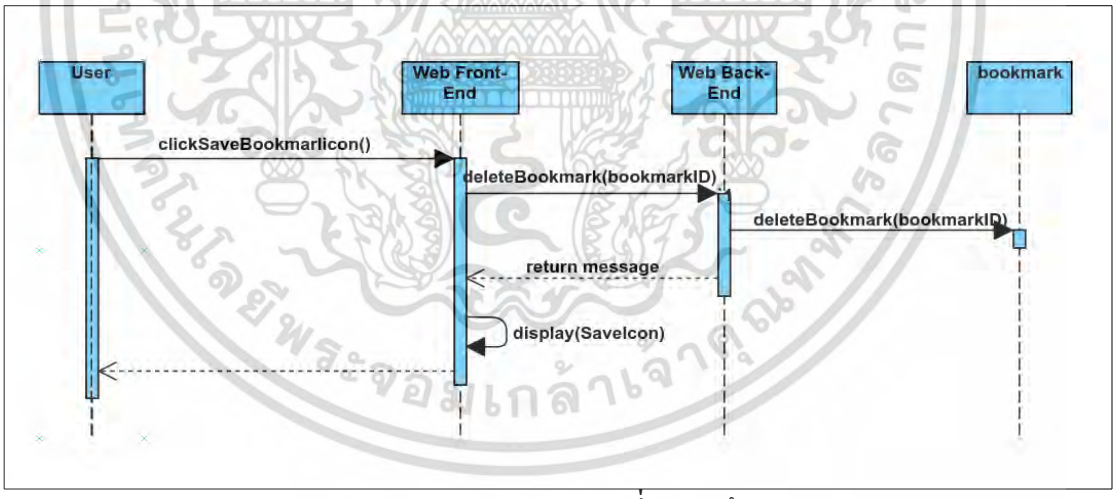
รูป 3.2 การทำงานของการค้นหา

ผู้ใช้งานทำการเรียกใช้ฟังก์ชั่นการค้นหาบทความ เว็บบราวเซอร์เรียกขอข้อมูลจากผู้ใช้แล้วทำการส่งข้อมูลไปยังแอพลิเคชั่นเซิร์ฟเวอร์ ระบบแอพลิเคชั่นเซิร์ฟเวอร์ทำการส่งคำค้นหาไปประมวลผลที่โมเดลการเรียนรู้ด้วยตัวเองที่ชื่อว่า BERT เพื่อวิเคราะห์เปลี่ยนคำค้นหาให้อยู่ในรูปที่สืบค้นในดาต้าเบสได้จากนั้นนำคำค้นหาที่ได้จากโมเดลไปค้นหาบทความที่เกี่ยวข้องแล้วส่งกลับแสดงผลที่เว็บบราวเซอร์



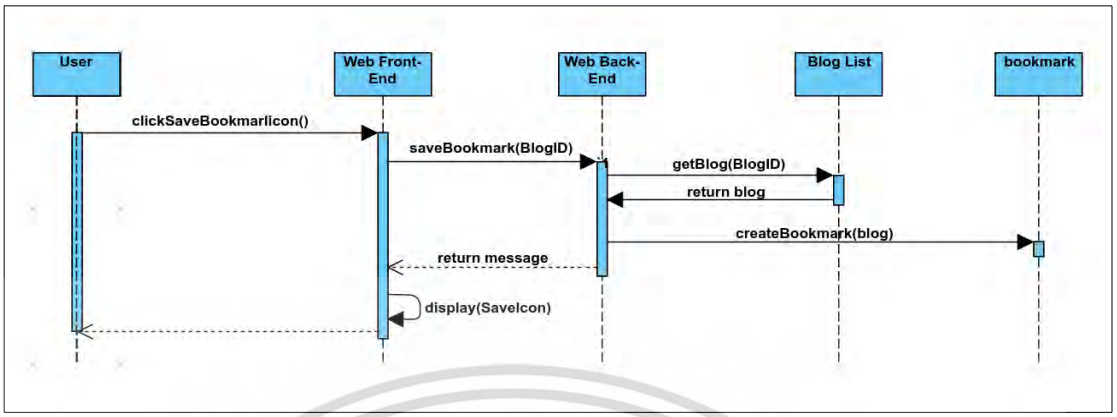
รูป 3.3 การทำงานของการเรียกดูบทความ

ผู้ใช้งานกดเลือกบทความหนึ่งๆจากชุดของบทความ เว็บบราวเซอร์จะทำการระบุข้อมูลของบทความนั้นๆส่งไปติดต่อขอข้อมูลเพิ่มเติมจากแอพลิเคชั่นเซิร์ฟเวอร์แล้วนำมาแสดงผลบนหน้าเว็บบราวเซอร์



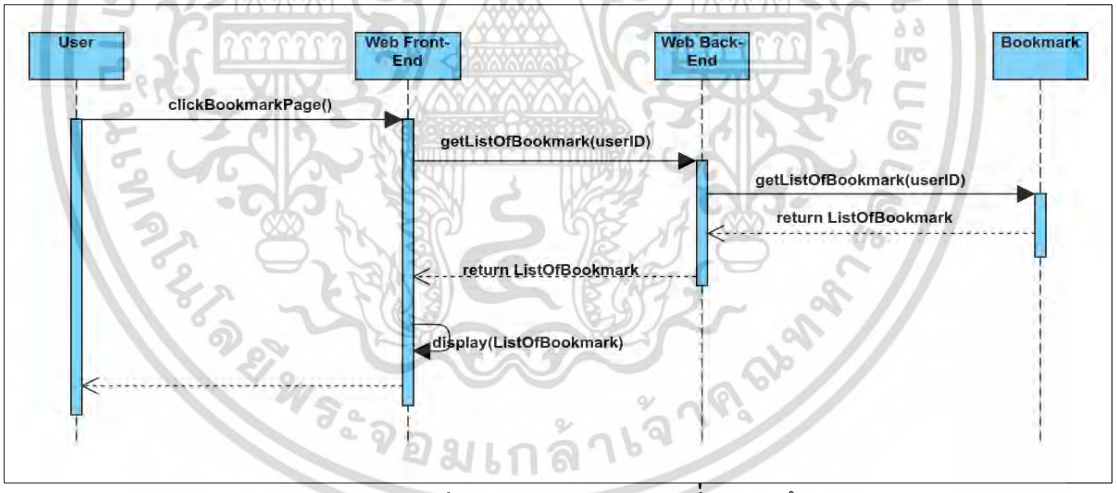
รูป 3.4 การลบบทความที่ถูกบันทึก

ผู้ใช้ที่มีข้อมูลในระบบทำการเลือกแสดงบทความหนึ่งๆที่เคยทำการบันทึกไว้ กดเลือกไอคอนที่ส่งคำสั่งไปยังแอพลิเคชั่นเซิร์ฟเวอร์ให้ทำการลบบทความนั้นๆออกจากบันทึก หลังจากนั้นเว็บไซต์ที่ติดต่อกับผู้ใช้แสดงผลไอคอนใหม่



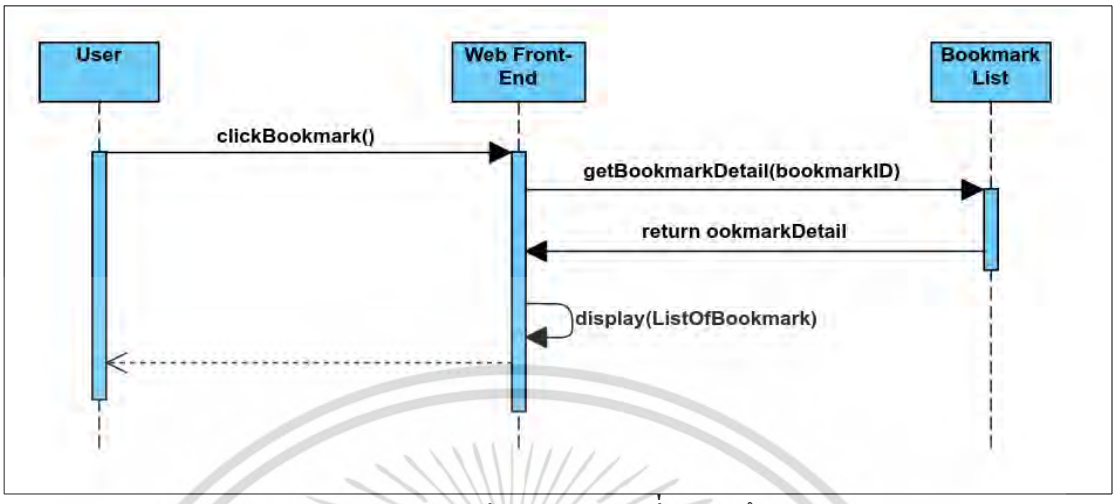
รูป 3.5 การบันทึกบทความ

ผู้ใช้กดเลือกไอคอนบันทึกบทความที่ส่งไอดีของบทความนั้นๆไปยังแอพลิเคชั่นเซิร์ฟเวอร์ผ่านฟังก์ชั่นการบันทึกบทความ จากนั้นแอพลิเคชั่นทำการค้นหาบทความนั้นๆแล้วทำการส่งข้อมูลของบทความไปบันทึกไว้สำหรับผู้ใช้ หลังจากทำการบันทึกเสร็จ ระบบแสดงผลไอคอนบันทึกบทความใหม่



รูป 3.6 การเรียกดูชุดของบทความที่ถูกบันทึก

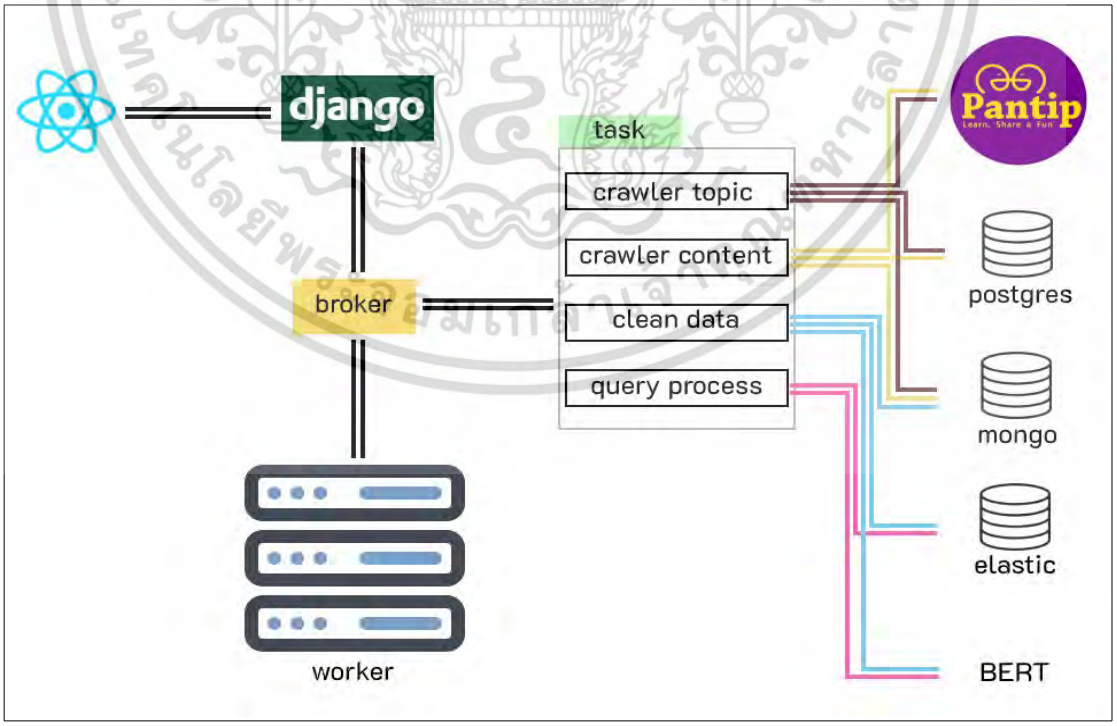
ผู้ใช้เรียกงานฟังก์ชั่นการเรียกดูบทความที่ถูกบันทึกผ่านหน้าเว็บจากนั้นเว็บไซต์ทำการส่งข้อมูลผู้ใช้ไปยังแอพลิเคชั่นเซิร์ฟเวอร์เพื่อเรียกดูบทความที่ถูกบันทึกทั้งหมด แล้วทำการส่งกลับไปยังหน้าเว็บเพื่อแสดงผล



รูป 3.7 การเรียกดูบทความที่ถูกบันทึก

ผู้ใช้งานกดเลือกบทความหนึ่งๆจากชุดของบทความที่ถูกบันทึก เว็บบราวเซอร์จะทำการระบุข้อมูลของบทความนั้นๆส่งไปติดต่อขอข้อมูลเพิ่มเติมจากแอพลิเคชั่นเซิร์ฟเวอร์แล้วนำมาแสดงผลบนหน้าเว็บบราวเซอร์

3.3 ภาพรวมระบบ



รูป 3.8 ภาพรวมการทำงานของระบบ

ระบบแบ่งออกเป็น 2 ส่วนใหญ่ คือ ส่วนแสดงผลบนเว็บบราวเซอร์ติดต่อกับผู้ใช้โดยตรงถูกพัฒนาด้วยไลบรารีภาษา Javascript ที่เรียกว่า React และส่วนระบบค้นหาข้อมูลบทความเพียงบางส่วนของเว็บไซต์ Pantip ซึ่งถูกพัฒนาขึ้นร่วมกันจาก 2 เทคโนโลยี คือ Elastic Search และโมเดลการเรียนรู้เชิงลึกที่เรียกว่า BERT

3.3.1 องค์ประกอบของระบบ

3.3.1.1 Front-End

ถูกพัฒนาขึ้นด้วยไลบรารีของจาวาสคริปต์ที่ชื่อว่า รีแอ็ค(React) ทำหน้าที่ในการแสดงผลลัพธ์เกี่ยวกับผลลัพธ์จากการค้นหาบทความและลิ้งค์ไปยังเว็บไซต์พันทิพ(Pantip) ที่เป็นข้อมูลทั้งหมดเกี่ยวกับบทความนั้นๆ

3.3.1.2 API

ถูกพัฒนาขึ้นโดยเฟรมเวิร์คของภาษาไพทอน(python) ทำหน้าที่ในการเชื่อมโยงระหว่างส่วนเว็บบราว์เซอร์และส่วนระบบค้นหา

3.3.1.3 Back-End

ถูกพัฒนาขึ้นมาจากหลายส่วน โดยมีโบรกเกอร์(Broker) ทำหน้าที่เชื่อมต่อประสานงานทุกส่วนเชื่อมโยงเข้าด้วยกัน โบรกเกอร์มีหน้าที่ในการเรียกงานจากทาสก์(Task) แล้วส่งทาสก์ไปทำงานบนเวิร์กเกอร์(Worker) ซึ่งเวิร์กเกอร์เปรียบเสมือนคอมพิวเตอร์ที่รอทำงานตามที่ถูกสั่งเข้ามาในที่นี้คือทาสก์ที่ได้รับจากโบรกเกอร์

ทาสก์ของระบบประด้วย 4 ทาสก์

1) crawler topic เป็นทาสก์ที่ทำหน้าที่ในการดึงข้อมูลกระทู้ภายใต้ขอบเขตของป้าย(Label) ที่ใช้ในการเทรนโมเดลการเรียนรู้เชิงลึกด้วยตัวเองจากเว็บไซต์พันทิพในส่วนของข้อมูลเบื้องต้นและข้อมูลที่จำเป็นในการระบุตัวตนเพื่อเข้าถึงกระทู้นั้นๆของเช่น topic title, url, topic id แล้วทำการเก็บข้อมูลในรูปแบบที่ไม่เป็นตารางในฐานข้อมูล(MongoDB)

2) crawler content เป็นทาสก์ที่ทำงานเพิ่มเติมจากส่วนของทาสก์แรกคือจะทำการดึงข้อมูลในส่วนของข้อมูลที่เป็นเนื้อหาของกระทู้(ทำการดึงข้อมูลเพียงแค่ส่วนที่ผู้ตั้งกระทู้สร้างขึ้นเท่านั้น ไม่รวมถึงคอมเมนต์ภายใต้กระทู้) โดยจะอ้างอิงกระทู้ที่จะทำการดึงข้อมูลจากข้อมูลที่ได้มาจากทาสก์แรก แล้วเก็บข้อมูลในรูปแบบที่ไม่เป็นตารางในฐานข้อมูลเช่นเดียวกับทาสก์แรก

3) clean data หรือคือกระบวนการเตรียมข้อมูล(Pre-Processing data) ทาสก์จะทำการดึงข้อมูลที่ถูกเก็บไว้จากทาสก์แรกและทาสก์ที่สองมาทำการประมวลผล จัดรูปแบบข้อมูลให้อยู่ในรูปแบบที่เหมาะสมแล้วทำการส่งข้อมูลเหล่านั้นเข้าโมเดลการเรียนรู้เชิงลึก โมเดลจะทำติดป้ายให้กับแต่ละบทความที่ถูกผ่านเข้าไปแล้วส่งข้อมูลทั้งบทความและป้ายออกไปเก็บไว้ที่ฐานข้อมูลที่ชื่อว่า Elastic ซึ่งฐานข้อมูลนี้จะทำหน้าที่ในการเก็บข้อมูลบทความที่ผ่านโมเดลการเรียนรู้เชิงลึกมาแล้วและทำถูกใช้ในระบบค้นข้อมูลบทความ

4) query process ทาสก์ที่ทำหน้าที่นำคำค้นหาจากผู้ใช้ส่งเข้าโมเดลการเรียนรู้เชิงลึกเพื่อจัดรูปแบบคำค้นหาให้อยู่ในรูปแบบที่เหมาะสมจากที่เป็นภาษาธรรมชาติ แล้วจะทำการนำป้ายที่ได้จากโมเดลออกไปใช้กับฐานข้อมูล Elastic เพื่อค้นหาบทความที่เกี่ยวข้องกับคำค้นของผู้ใช้แล้วส่งกลับไปแสดงผลที่บราวเซอร์

3.4 การรวมรวมข้อมูล

3.4.1 การดึงข้อมูล

3.4.1.1 รูปแบบข้อมูลที่ใช้งาน

ระบบนี้จะใช้ข้อมูลการท่องเที่ยวจากเว็บไซต์ pantip.com ซึ่งจะเป็นเว็บไซต์ที่มีโครงสร้างเป็นชนิดเว็บบอร์ด ซึ่งจะมีกลุ่มของหัวเรื่อง(forums) ที่จะมีสมาชิกเป็นหัวเรื่อง(topic) โดยในแต่ละกลุ่มของหัวเรื่องนั้นจะมีเนื้อหาเป็นไปในแนวทางเดียวกัน ในแต่ละหัวเรื่องจะมีสมาชิกเป็นบทความ (blog) ซึ่งจะประกอบไปด้วย ผู้สร้าง, เนื้อหา, เวลาที่สร้าง เป็นต้น ทั้งนี้ยังมีระบบที่สร้างออกมาเพิ่มเติม เช่น ระบบแท็ก(tag) ที่จะเป็นสมาชิกของหัวเรื่อง ซึ่งสามารถระบุถึงแนวทางของหัวเรื่องได้ละเอียดกว่ากลุ่มของหัวเรื่อง

ตาราง 3.8 แท็กข้อมูลที่ถูกนำมาใช้

แท็กที่ถูกใช้งานในระบบ	แท็กของเว็บไซต์ pantip
crime	อาวุธยุทโธปกรณ์
disease	โรคติดเชื้อไวรัสโคโรนาสายพันธุ์ใหม่_2019_(COVID-19)
	ไวรัสอีโบลา
	ไข้หวัด
	โรคติดต่อ

ตาราง 3.9 แท็กข้อมูลที่ถูกนำมาใช้(ต่อ)

แท็กที่ถูกใช้งานในระบบ	แท็กของเว็บไซต์ pantip
hotel	ที่พัก
	โรงแรมรีสอร์ท
	อพาร์ตเมนต์
	บ้าน
	Airbnb
	Agoda
location	กิจกรรมท่องเที่ยว
	สถานที่ท่องเที่ยวในประเทศ
	เที่ยวภูเขา
	เที่ยวทะเล
	เดินป่า
	สถานที่ถ่ายรูป
restaurant	อาหารไทย
	อาหารปิ้งย่าง
	อาหารซีฟู้ด
	อาหารญี่ปุ่น
	อาหารคาว
	อาหารจีน
travel	รถโดยสารประจำทาง
	นักท่องเที่ยว
	แผนการเดินทางและท่องเที่ยว
	รถโดยสาร
	ระบบขนส่งมวลชน

จากตารางจะเห็นได้ว่ามีการจัดหมวดหมู่แท็กจากเว็บไซต์ให้สอดคล้องกับแท็กของระบบทั้ง 6 แท็ก ทั้งนี้เพื่อทำให้มีข้อมูลในแต่ละแท็กมากเพียงพอที่จะนำมาใช้ในการฝึกสอนโมเดลในงานวิจัยชิ้น โดยการจัดกลุ่มทำโดยหาความหมายใกล้เคียงกันของความหมายมากที่สุด

3.4.1.2 การดึงข้อมูล

ในการดึงข้อมูลเพื่อมาใช้ในระบบ ถูกแบ่งออกเป็น 2 ส่วน คือการดึงข้อมูลหัวเรื่องแท็กต่างๆ จากเว็บไซต์ pantip.com เพื่อเก็บข้อมูลที่สามารถสื่อถึงเนื้อหาของระบบโดยส่วนนี้สามารถดึงได้จากเอพีไอ(API) ของทางเว็บไซต์ได้ ซึ่งตัวระบบนี้อ้างอิงตามโมเดลอีทีแอล(Extract, transform, load ; ETL) และอีกส่วนจะทำการนำเอาข้อมูลที่ได้จากในส่วนแรกมาเพื่อใช้ดึงข้อมูลเนื้อหาผ่านหน้าเว็บเนื่องจากไม่มีเอพีไอที่เปิดให้ใช้งาน แล้วจึงนำข้อมูลจากส่วนนี้ไปเข้าคิวแล้วส่งไปยังการทำงานในหัวข้อการเตรียมข้อมูลต่อไป ซึ่งคิวนี้จะเก็บหมายเลขอ้างอิงของหัวเรื่อง เพื่อนำไป transform ในรูปแบบต่างๆ ที่สามารถกระจายงานแล้วประมวลผลพร้อมกันได้ เพื่อให้ไม่สูญเสียประสิทธิภาพในเรื่องของเวลาที่ใช้ และทรัพยากรที่เหลือ โดยระบบที่ทำการควบคุมส่วนการดึงข้อมูลทั้งหมดจะทำงานเป็น batch ซึ่งจะดึงทุกแท็กตามระยะเวลาที่กำหนด เพื่อให้ได้ข้อมูลที่ทันสมัยเสมอ

3.4.2 การเตรียมข้อมูล

3.4.2.1 การดึงข้อมูลที่สนใจจากภาษา HTML

เนื่องจากเว็บไซต์ pantip.com ไม่ได้มี api ในการดึงข้อมูลหัวเรื่อง (topic) จึงจำเป็นต้องทำการดึงข้อมูลผ่านทางหน้าเว็บ โดยใช้ไลบรารีของภาษาไพทอนที่ชื่อว่า "beautifulsoup" แล้วจึงทำการเข้าถึงข้อมูลเอชทีเอ็มแอลที่ต้องการเพื่อนำมาใช้

3.4.2.2 การทำความสะอาดข้อมูล

ในเบื้องต้นนั้นจะทำการตัดอักขระที่ไม่ใช้ตัวอักษรภาษาอังกฤษ ตัวเลขภาษาอังกฤษ ตัวอักษรภาษาไทย ออกไป เช่น สัญลักษณ์ อิโมจิคอน ถึงแม้บางประโยคจะมีความหมายต่อข้อความ หรือประโยคนั้นๆ

3.4.2.3 การตัดคำภาษาไทย

การตัดคคำในข้อความภาษาไทยนั้น ได้ใช้เทคนิกการตัดคำแบบเหมือนกันมากที่สุด (maximal matching algorithm) คือตัดคำทุกรูปแบบที่เป็ยไปได้ แล้วเลือกรูปแบบที่มีจำนวนคำน้อยที่สุดออกมา โดยใช้ไลบารี PyThaiNLP ในการตัดคำ

3.4.2.4 การตัดประโยคภาษาไทย

การตัดประโยคภาษาไทย เริ่มต้นโดยการตัดข้อความใดๆ ให้เป็นลำดับของคำก่อน แล้วจึงนำแต่ละคำมาหาชนิดของคำ โดยมีแนวคิดที่ว่า ประโยคในภาษาไทยนั้น จะประกอบไปด้วย ภาคประธานกับภาคแสดง โดยภาคประธานนั้นจะแสดงไปที่ใครทำอะไร ซึ่งจะเป็นคำนานสรรพนามเป็นหลัก ภาคแสดงนั้นจะเป็นทำแบบไหน ซึ่งใช้คำกริยา โดยตัวอัลกอรึทึมจะเริ่มทำการนำเข้าคำ และตรวจสอบชนิดของคำว่าเป็นคำใด ไปเรื่อยๆ โดยมีเงื่อนไขคือจะพบว่าเป็นประโยคก็ต่อเมื่อ ประกอบด้วยทั้งภาคประธานและภาคแสดงครบแล้ว และคำต่อไปนั้น เป็นภาคประธานด้วย หรือว่าหากพบคำบุพบทด้วยก็เช่นกัน โดยใช้ไลบารี PyThaiNLP ในการหาชนิดของคำ

3.4.3 การเตรียมคำตอบของข้อมูล

ในการเตรียมข้อมูลคำตอบ ทำได้โดยการตัดเนื้อหาที่ได้จากเว็บไซต์พันทิพให้อยู่ในรูปประโยคเดียวจากบทความยาวแล้วทำการติดคำตอบให้แต่ละประโยคโดยคำตอบได้มากจากเนื้อหาของบทความในส่วนของแท็ก ซึ่งในบทความนั้นอาจจะประกอบด้วยเพียงหนึ่งแท็กหรือมากกว่าหนึ่ง ดังนั้นทางผู้จักทำจึงทำการเลือกว่าถ้าหากเป็นแบบคำตอบเดียวก็สามารถนำมาเป็นคำตอบของหัวเรื่องนั้นได้เลย แต่ถ้าหากเป็บแบบหลายคำตอบก็จะทำการเลือกแท็กลำดับแรกมาเป็นคำตอบ

3.5 ระบบค้นหาข้อมูล

3.2.1 โครงสร้างการเก็บข้อมูล

ตาราง 3.10 โครงสร้างการจัดเก็บเอกสารเพื่อใช้ในการสืบค้น

Attribute	Type	Analyzer	Filter	Explian
name	text	shingle	filter_shingle	หัวข้อ
message	text	shingle	filter_shingle	เนื้อหา
tag	text	keyword	-	แท็ก
user	text	keyword	-	Author Username
created_at	date	-	-	เวลาที่สร้างหัวเรื่อง

ตารางแสดงให้เห็นถึงโครงสร้างและความสัมพันธ์ของข้อมูลที่ถูกนำมาใช้ในส่วนของระบบค้นหา

3.5.1 ตัวตัดคำ (tokenizer)

ข้อความหรือคำค้น สามารถตัดเป็นคำ หลายๆ คำได้ ซึ่งคำในที่นี้หมายถึง token ในภาษาไทยนั้น ได้มีผู้พัฒนาตัวตัดคำไว้แล้วในชื่อ ICU Tokenizer ซึ่งสามารถตัดคำภาษาไทย, ลาว, จีน, ญี่ปุ่น และเกาหลีได้

โดย token เป็นหน่วยที่เล็กที่สุดของเอกสารนั้นๆ จะถูกนำไปใช้ในการทำ inverted index เพื่อเป็นฐานข้อมูลในกระบวนการสืบค้นต่อไป

3.5.2 ตัวกรอง (filter)

ในบางกรณีคำหรือโทเคนเดี่ยวๆนั้นอาจจะไม่สามารถให้ผลการสืบค้นที่ดี เช่น คำบางคำนั้นมีความถี่กับคำติดกันรอบข้างสูง หรือคำบางคำ มักจะพบกับคำบางคำในระยะไม่เกิน x คำ เป็นต้น filter เป็นเครื่องมือที่ใช้ในการสร้างกฎเพิ่มเติมของตัวตัดคำ (tokenizer) เพื่อเป็นตัวโทเคนใหม่ โดยจะถูกนำไปใช้ในการทำ inverted index เช่นกัน

3.5.3 การค้นหา (search)

การค้นหานั้นจะเริ่มต้นจากการนำคำค้นมาทำการตัดคำ โดยกฎต่างๆ แล้วนำโทเคนไปค้นหาใน inverted index นั้นๆ โดยใช้ TF-IDF(term frequency–inverse document frequency) ซึ่งเป็นการคำนวณระหว่างผลคูณของความถี่ของโทเคนนั้นๆ ในเอกสารต่างๆ กับจำนวนเอกสารต่างๆ ที่มีโทเคนนั้นอยู่ด้วยมาคำนวณหาผลลัพธ์ว่าเอกสารไหนเป็นผลการค้นหาของคำๆ นั้น

3.2.5 การจัดอันดับ (ranking)

ในการจัดลำดับความสำคัญของเอกสารทำด้วยวิธี

1) ให้น้ำหนักกับโทเคนใดๆ มากกว่าปกติ โดยจะให้นำหนักกับโทเคนในประวัติการของคำค้นที่เคยค้นในระยะเวลาเดียวกัน หรือประวัติการกดลิ้งที่เกี่ยวข้องกับคำค้นนั้นๆ ที่ผ่านมา
2) การให้น้ำหนักกับเอกสารใดๆ มากกว่าปกติ โดยพิจารณาจากจำนวนผู้อ่านเอกสารนั้นๆ หรือระยะเวลาที่เอกสารนั้นถูกสร้างขึ้นมาเป็นต้น

3.6 การเตรียม BERT pretrain weight

BERT (Bidirectional Encoder Representations from Transformers) ภาษาไทยนั้น ได้มีผู้พัฒนาไว้แล้วโดยคุณ u41ppp ในชื่อโครงการ BERT-th (BERT pre-training in Thai language) ซึ่งได้เตรียมทั้งส่วนของโมเดล bert ภาษาไทย ข้อมูลต่างๆ ที่สามารถสร้าง weight pre-trained ของ bert ภาษาไทยได้ ตลอดจนไปถึงตัวอย่างการจำแนกประโยคต่างๆ ซึ่งวข้อมูลที่ใช้ในการเตรียม pretrain bert มีดังนี้

1) คลังประโยคภาษาไทย การหาข้อมูลของประโยคภาษาไทยจำนวนมาก โดยเตรียมจากการนำบทความทั้งหมดใน wikipedia ภาษาไทย มาทำการเปลี่ยนรูปแบบเอกสาร โดยการใส่แท็ก xml แล้วตัดมาแต่ละย่อหน้าของแต่ละบทความ นำมาตัดเป็นประโยคให้ได้ 1 ประโยคต่อ 1 บรรทัด
2) รหัสคำภาษาไทย (Tokenization) โมเดล(BERT) นั้นไม่ได้สนใจในการตัดคำแต่จะสนใจในหน่วยย่อยของคำ(subword) ซึ่งในที่นี้จะใช้หลักการจำคู่ของตัวอักษร(byte-pair encoding) คือการหาความถี่ของการพบตัวอักษรตั้งแต่สองตัวขึ้นไปที่ติดกัน(unigram, trigram) ที่พบมากที่สุด ซึ่งนับว่าเป็นข้อดีที่ไม่ขึ้นอยู่กับภาษาใดๆ โดยจะเรียงลำดับความถี่ของการพบจากมากที่สุดไปน้อยที่สุด แล้วสร้าง token เป็นหมายเลขเรียงลำดับตั้งแต่ 0, 1, 2, …, N ต่อท้ายของแต่ละหน่วยย่อยของคำ

การกำหนดค่าพารามิเตอร์จากผู้พัฒนาโมเดล(BERT) ภาษาไทยนั้น ถูกกำหนดไว้ดังนี้

1) ความยาวของโยคสูงสุด 128 หน่วยคำย่อย
2) ขนาดของแบท(Batch) 32
3) เปอร์เซ็นการทำนายคำในแต่ละประโยค 20%

3.7 การปรับแต่ง BERT pretrain weight

การปรับแต่น้ำหนักของโมเดล (Fine-tuning model) ของโมเดล(BERT) นั้น จะเป็นการเปลี่ยนแปลงเพียงแค่ใน layer 7-12 ให้เปลี่ยนไปให้เหมาะกับประเภทของงาน โดยใช้ข้อมูลในการปรับน้ำหนักเป็นคลังประโยคภาษาไทยพร้อมคำตอบแบบหลายคำตอบ dataset นั้นจะเป็นประโยคคำค้นต่างๆ พร้อมคำตอบแบบหลายคำตอบที่เกี่ยวข้องกับคำค้น เช่น

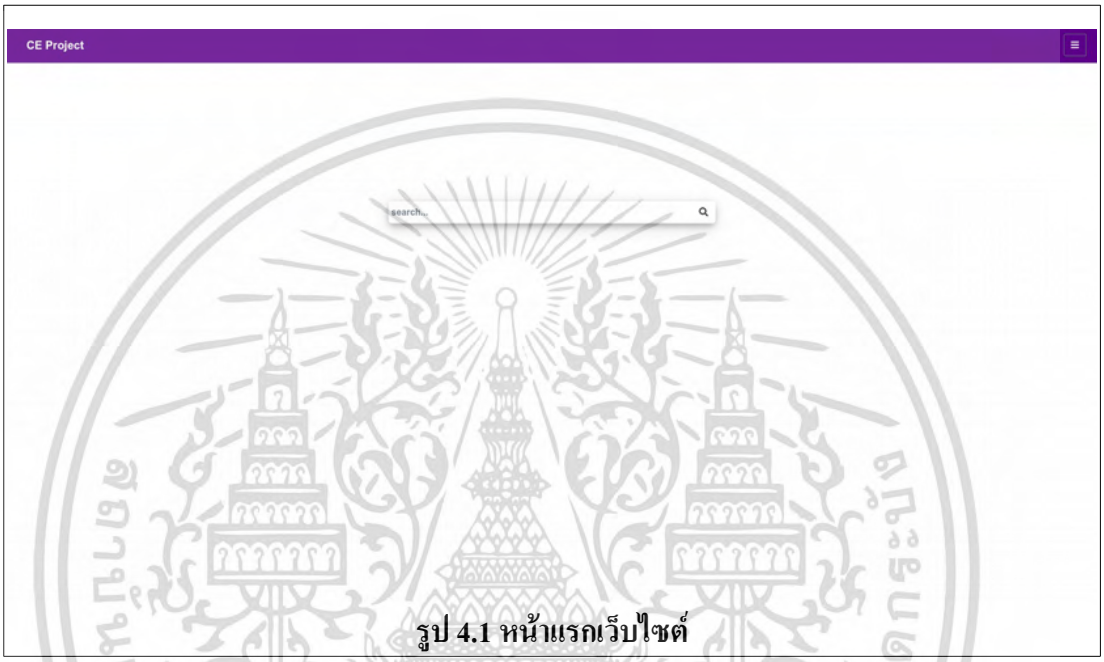
ตาราง 3.11 ตารางแสดงรูปแบบข้อมูลที่นำเข้าโมเดล

คำถาม	คำตอบ
การเดินทางไปวัดภูเขาทอง	การเดินทาง, วัดภูเขาทอง
วัดภูเขาทองปิดกี่โมง	วัดภูเขาทอง, เวลาเปิดปิด

ในการพัฒนางานชิ้นนี้ทางผู้พัฒนาจะใช้พารามิเตอร์เหมือนกับการเตรียม pre-train bert model แต่ในการ classify นั้น ทางผู้พัฒนาจะใช้โมเดล DNN มาใช้เพื่อรับผลลัพธ์ที่เป็นเวกเตอร์จาก BERT มาทำการหา weight กับคำตอบของเรา และใช้ฟังก์ชั่นซิกมอยด์(Sigmoid Function) เพื่อหาความน่าจะเป็นของคำตอบ โดยการใช้ theshold มาพิจารณาว่าเป็นคำตอบหรือไม่

บทที่ 4
การดำเนินการและผลการทดลอง

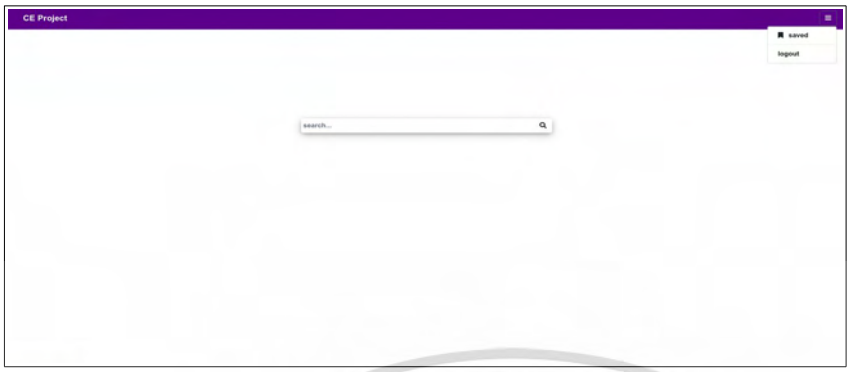
4.1 เว็บไซต์



รูป 4.1 หน้าแรกเว็บไซต์

หน้าแรกเว็บไซต์จะแสดงหน้าว่างเปล่าพร้อมกับช่องค้นหาข้อมูลของระบบ ที่ไอคอนมุมซ้ายเป็นการลิ้งค์มายังหน้าแรกและไอคอนที่มุมขวาเป็นเมนูสำหรับผู้ใช้งานมีรายละเอียดการแสดงผลตามรูป 4.2

หากทำการค้นหาข้อมูลในช่องค้นหา ระบบจะแสดงผลตามรูป 4.3 ต่อไป

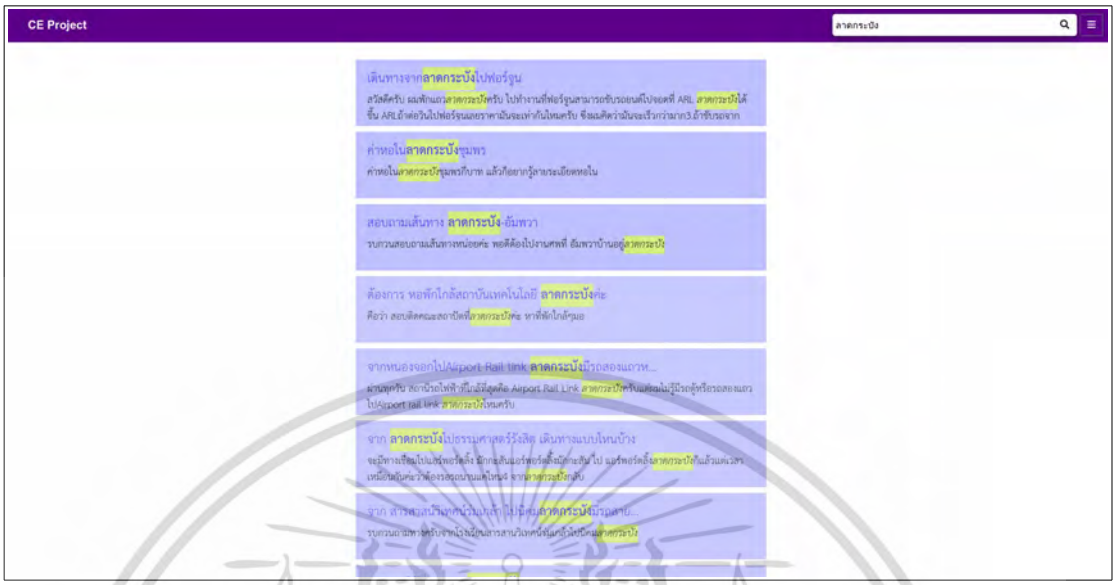


ก)

ข)

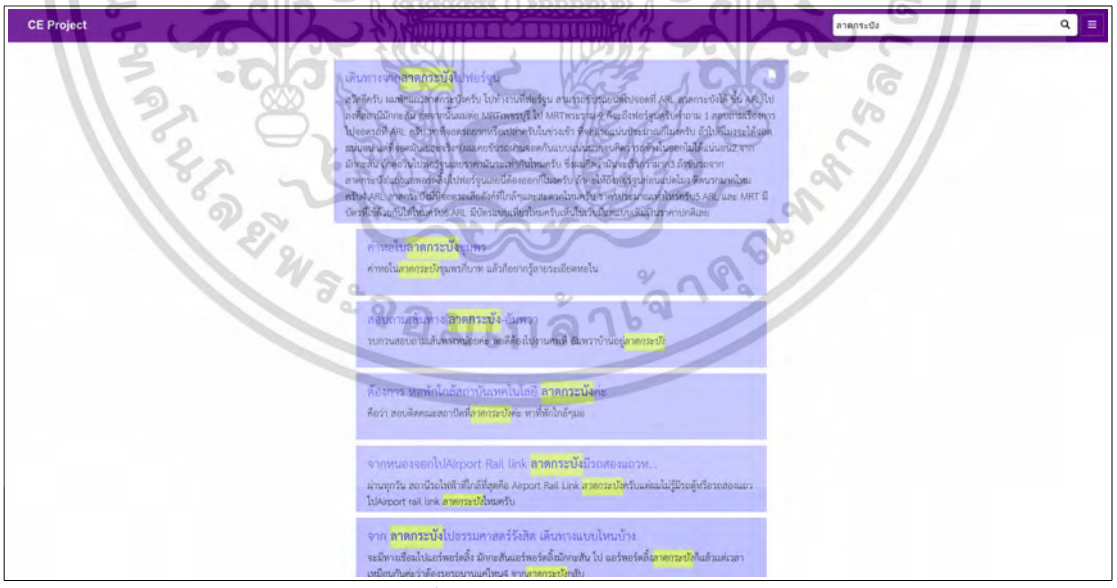
รูป 4.*2* หน้าแรกเว็บไซต์สำหรับผู้ใช้งานทั้งสองประเภท

หน้าแรกเว็บไซต์สำหรับผู้ใช้งานทั้งสองประเภทแตกต่างกันตรงไอคอนเมนูด้านมุมขวาโดย รูป ก) เป็นการแสดงผลสำหรับผู้ใช้งานที่ผ่านการเข้าสู่ระบบ และรูป ข) เป็นการแสดงผลสำหรับผู้ใช้งานที่ไม่ผ่านการเข้าสู่ระบบ



รูป 4.3 เว็บไซต์แสดงผลลัพธ์การค้นหาข้อมูล

เว็บไซต์แสดงผลลัพธ์การค้นหาข้อมูลแบบค้นหาสำเร็จปรกติโดยแสดงชุดของบทความในแนวตั้ง ในแต่ละหัวเรื่องบทความจะมีการแสดงผลให้โดดเด่นคำที่ตรงกับความต้องการของผู้ที่ผ่านการคำนวณจากการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์ และมีช่องค้นหาที่มุมบนด้านขวา



รูป 4.4 เว็บไซต์แสดงข้อมูลรายละเอียดของบทความ

เมื่อผู้ใช้กดบริเวณเนื้อหาของบทความตามรูปที่ 4.3 ระบบจะแสดงบทความรายละเอียดของบทความนั้นๆมากขึ้น และหากเป็นผู้ใช้งานที่ผ่านการเข้าสู่ระบบจะมีไอคอนบริเวณมุมบนขวาเพื่อให้กดบันทึกบทความนั้นๆได้

รูป 4.5 เว็บไซต์แสดงผลลัพธ์แบบไม่มีบทความที่ตรงกับความต้องการ

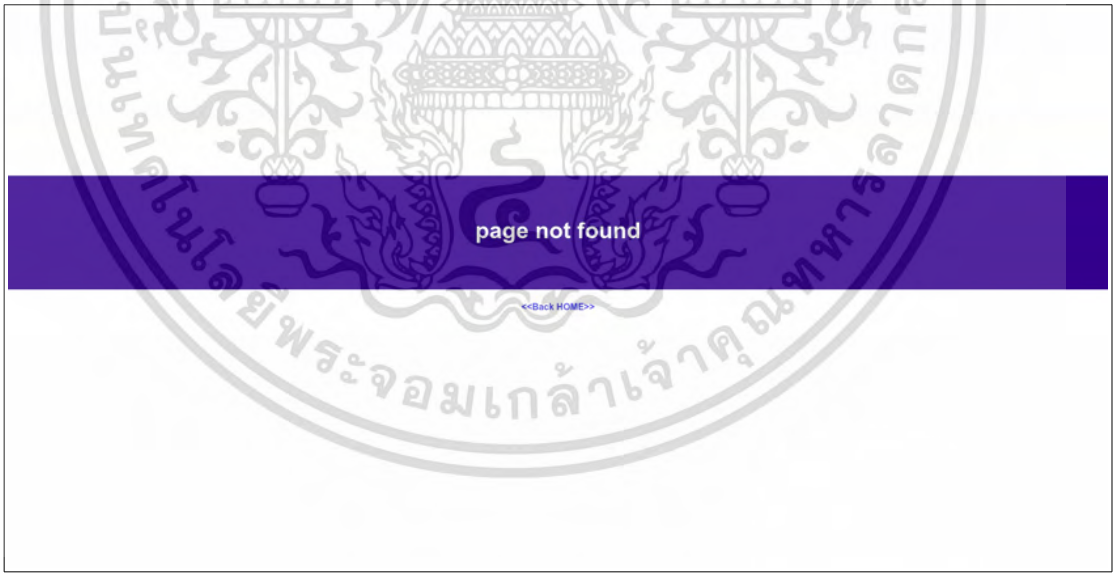
เว็บไซต์แสดงผลลัพธ์การค้นหาแบบไม่ปรกติคือไม่มีบทความที่ตรงกับความต้องการของผู้ใช้งานโดยแสดงข้อความ “no blog relate to your query… please try again :)”

รูป 4.6 เว็บไซต์แสดงผลหน้าเพื่อการลงทะเบียนเข้าสู่ระบบ

การเข้าสู่ระบบต้องการข้อมูล 2 อย่าง คือ ชื่อผู้ใช้งาน(Username) และรหัสผ่าน(Password) โดยกดปุ่ม login เพื่อตรวจสอบข้อมูลหากข้อมูลถูกต้องระบบจะแสดงผลลัพธ์ตามรูป 4.7

รูป 4.7 เว็บไซต์แสดงผลชุดของบทความที่ถูกบันทึก

เว็บไซต์แสดงผลหน้าบทความที่ถูกบันทึกโดยผู้ใช้งาน โดยมีไอคอนที่มุมบนขวาเป็นปุ่มสำหรับการกดลบบทความที่ไม่ต้องการบันทึกต่อไปได้



รูป 4.8 เว็บไซต์แสดงผลจากการทำงานไม่ปรกติของระบบ

หากระบบเกิดความผิดพลาดจากการสื่อสารระหว่างหน้าเว็บกับแอพลิเคชั่นเซิร์ฟเวอร์หรือผู้ใช้พยายามเข้าถึงยูอาร์ไอ(URI) ที่ไม่เหมาะสมในระบบ เว็บไซต์จะทำการแสดงผลตามรูป 4.8

4.2 การรวบรวมข้อมูล

4.1.1 ผลการทำความสะอาดข้อมูล

ตาราง 4.1 ตัวอย่างการตัดอิโมติคอนออกจากข้อความ

ประโยคตั้งต้น	ลูกชินปลามหาเทพ เจิเฮียง อร่อย มากกกก
ผลลัพธ์	ลูกชิ้นปลามหาเทพ เจ้เฮียง อร่อย มากกกก

ตาราง 4.2 ตัวอย่างการตัดสัญลักษณ์ออกจากข้อความ

ประโยคตั้งต้น	Autumn and I ♡ ชีวิตฉันกับวันใบไม้ร่วง @ นิวซีแลนด์ ...By Pla Gallery
ผลลัพธ์	autumn and I ชีวิตฉันกับวันใบไม้ร่วง นิวซีแลนด์ by pla gallery

4.1.2 การตัดคำและประโยคภาษาไทย

ตาราง 4.3 ตัวอย่างที่ 1 การตัดคำออกจากบทความ

บทความ	สมัยเมื่อ 10 กว่าปีที่แล้วเคยไปทาน ภัตตาคาร ศรีหยก ถศรีนครินทร์ แล้วเจ้าของร้านบอกว่าจะย้ายไปที่ชลบุรี แล้วก็ติดต่อเจ้าของร้านไม่ได้อีก เพื่อน ๆ พอทราบไหมครับว่า ภัตตาคารศรีหยกย้ายไปที่ไหนในชลบุรีครับ คิดถึง หอยจ๊อ และ เป็ดปักกิ่ง มาก ๆ ขอบคุณทุกท่านที่ช่วยเหลือนะครับ
ผลลัพธ์	'สมัย', 'เมื่อ', '10', 'กว่า', 'ปี', 'ที่แล้ว', 'เคย', 'ไป', 'ทาน', 'ภัตตาคาร', 'ศรี', 'หยก', 'ถ', 'ศรีนครินทร์', 'แล้ว', 'เจ้าของร้าน', 'บอก', 'ว่า', 'จะ', 'ย้าย', 'ไป', 'ที่', 'ชลบุรี', 'แล้วก็', 'ติดต่อ', 'เจ้าของร้าน', 'ไม่', 'ได้', 'อีก', 'เพื่อน ๆ', 'พอ', 'ทราบ', 'ไหม', 'ครับ', 'ว่า', 'ภัตตาคาร', 'ศรี', 'หยก', 'ย้าย', 'ไป', 'ที่ไหน', 'ใน', 'ชลบุรี', 'ครับ', 'คิดถึง', 'หอย', 'จ๊อ', 'และ', 'เป็ด', 'ปักกิ่ง', 'มาก', 'ๆ', 'ขอบคุณ', 'ทุกท่าน', 'ที่', 'ช่วยเหลือ', 'นะ', 'ครับ'

ตาราง 4.4 ตัวอย่างที่ 1 การตัดประโยคออกจากบทความ

บทความ	สมัยเมื่อ 10 กว่าปีที่แล้วเคยไปทาน ภัตตาคาร ศรีหยก ถศรีนครินทร์ แล้วเจ้าของร้านบอกว่าจะย้ายไปที่ชลบุรี แล้วก็ติดต่อเจ้าของร้านไม่ได้อีก เพื่อน ๆ พอทราบไหมครับว่า ภัตตาคารศรีหยกย้ายไปที่ไหนในชลบุรีครับ คิดถึง หอยจ๊อ และ เป็ดปักกิ่ง มาก ๆ ขอบคุณทุกท่านที่ช่วยเหลือนะครับ
ผลลัพธ์	สมัยเมื่อ 10 กว่าปีที่แล้วเคยไปทาน ภัตตาคาร ศรีหยก ถศรีนครินทร์ แล้วเจ้าของร้านบอกว่าจะย้ายไปที่ชลบุรี แล้วก็ติดต่อเจ้าของร้านไม่ได้อีก
	เพื่อน ๆ พอทราบไหมครับว่า ภัตตาคารศรีหยกย้ายไป
	ที่ไหนในชลบุรีครับ คิดถึง หอยจ๊อ และ เป็ดปักกิ่ง มาก ๆ ขอบคุณทุกท่านที่ช่วยเหลือนะ

ซึ่งจะเห็นได้ว่าสามารถตัดคำในภาษาไทยได้อย่างถูกต้อง เนื่องจากระดับภาษาที่ใช้เป็นภาษาทั่วไป ที่มีการสะกดคำได้อย่างถูกต้องอยู่แล้ว ส่วนการตัดประโยคนั้น มีการตัดที่ผิดพลาด 1 จุด อันเนื่องจากการระบุชนิดของคำนั้น ไม่ได้ดูบริบทของคำโดยรอบ

ตาราง 4.5 ตัวอย่างที่ 2 การตัดคำออกจากบทความ

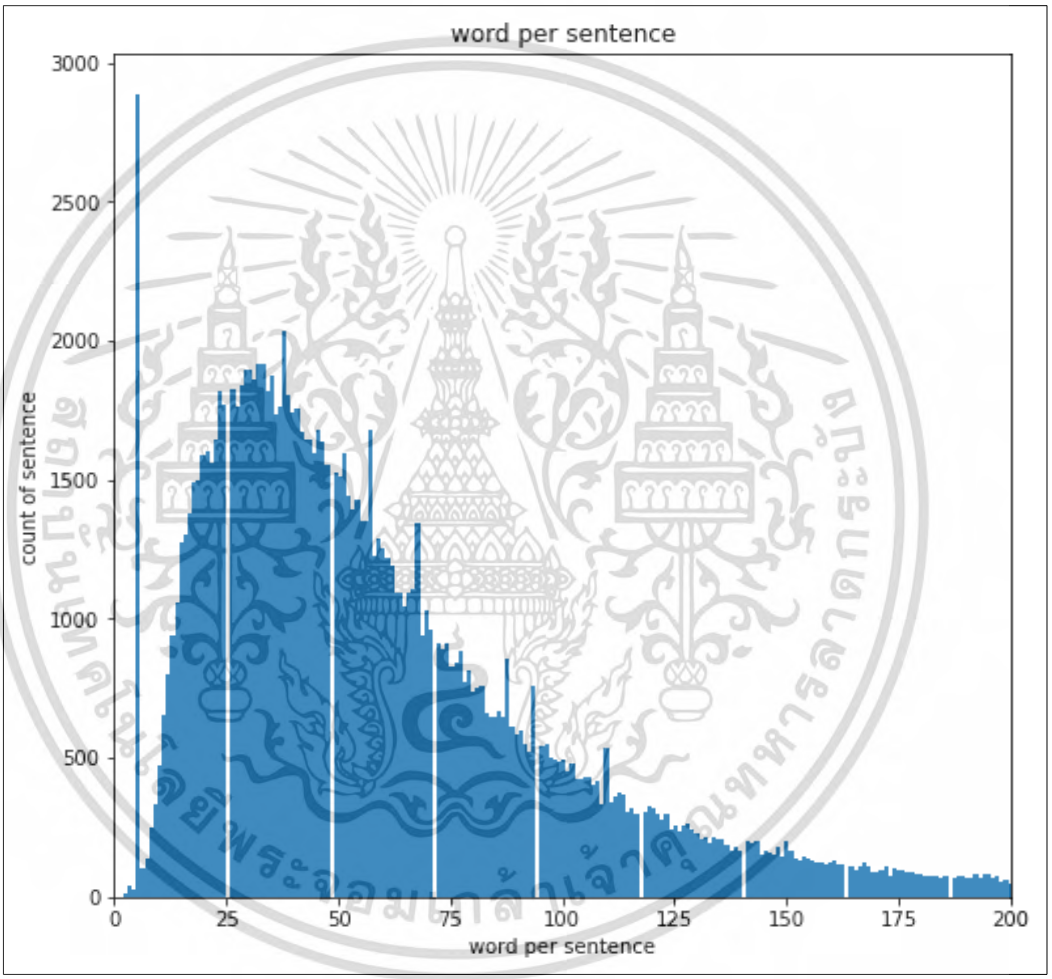
บทความ	พอดีไปเกาะล้านครั้งแรกแล้วหาที่พักสำหรับ 6 คนอยู่ค่ะถ้าใครพอมีประสบการณ์ อยากแนะนำหาด โรงแรม หรืออะไรก็ได้บอกได้เลยนะคะ ขอบคุณล่วงหน้านะคะะะะะ
ผลลัพธ์	'พอดี', 'ไป', 'เกาะ', 'ล้าน', 'ครั้งแรก', 'แล้ว', 'หา', 'ที่พัก', 'สำหรับ', '6', 'คน', 'อยู่', 'ค่ะ', 'ถ้า', 'ใคร', 'พอ', 'มีประสบการณ์', 'อยาก', 'แนะนำ', 'หาด', 'โรงแรม', 'หรือ', 'อะไร', 'ก็ได้', 'บอก', 'ได้', 'เลย', 'นะคะ', 'ขอบคุณ', 'ล่วงหน้า', 'นะคะ', 'ะะะะะ'

ตาราง 4.6 ตัวอย่างที่ 2 การตัดประโยคออกจากบทความ

บทความ	พอดีไปเกาะล้านครั้งแรกแล้วหาที่พักสำหรับ 6 คนอยู่ค่ะถ้าใครพอมีประสบการณ์ อยากแนะนำหาด โรงแรม หรืออะไรก็ได้บอกได้เลยนะคะ ขอบคุณล่วงหน้านะคะะะะะ
ผลลัพธ์	'พอดีไปเกาะล้านครั้งแรกแล้วหาที่พักสำหรับ 6 คนอยู่ค่ะถ้า'
	'ใครพอมีประสบการณ์ อยากแนะนำหาด โรงแรม หรือ'
	'อะไรก็ได้บอกได้เลยนะคะ ขอบคุณล่วงหน้านะคะะะะะ'

ซึ่งจะเห็นได้ว่าสามารถตัดคำในภาษาไทยส่วนใหญ่ได้อย่างถูกต้อง ส่วนการตัดประโยคนั้น เช่นเดียวกับตัวอย่างที่แล้ว ที่การะบุชนิดของคำนั้น ไม่ได้ดูบริบทของคำโดยรอบ คำบุพบท (“หรือ”) นั้นในบริบทนี้ไม่สามารถตัดประโยคนี้ได้

การเตรียมประโยคนั้นจำเป็นที่จะต้องรักษาจำนวนของคำในแต่ละประโยคไม่ได้ยาวเกินที่กำหนดไว้ เนื่องจากข้อกำจัดในตัวของ deep learning ที่มีการกำหนดโหนดของขาเข้าไว้คงที่



รูป 4.9 กราฟแสดงถึงจำนวนคำในแต่ละประโยค

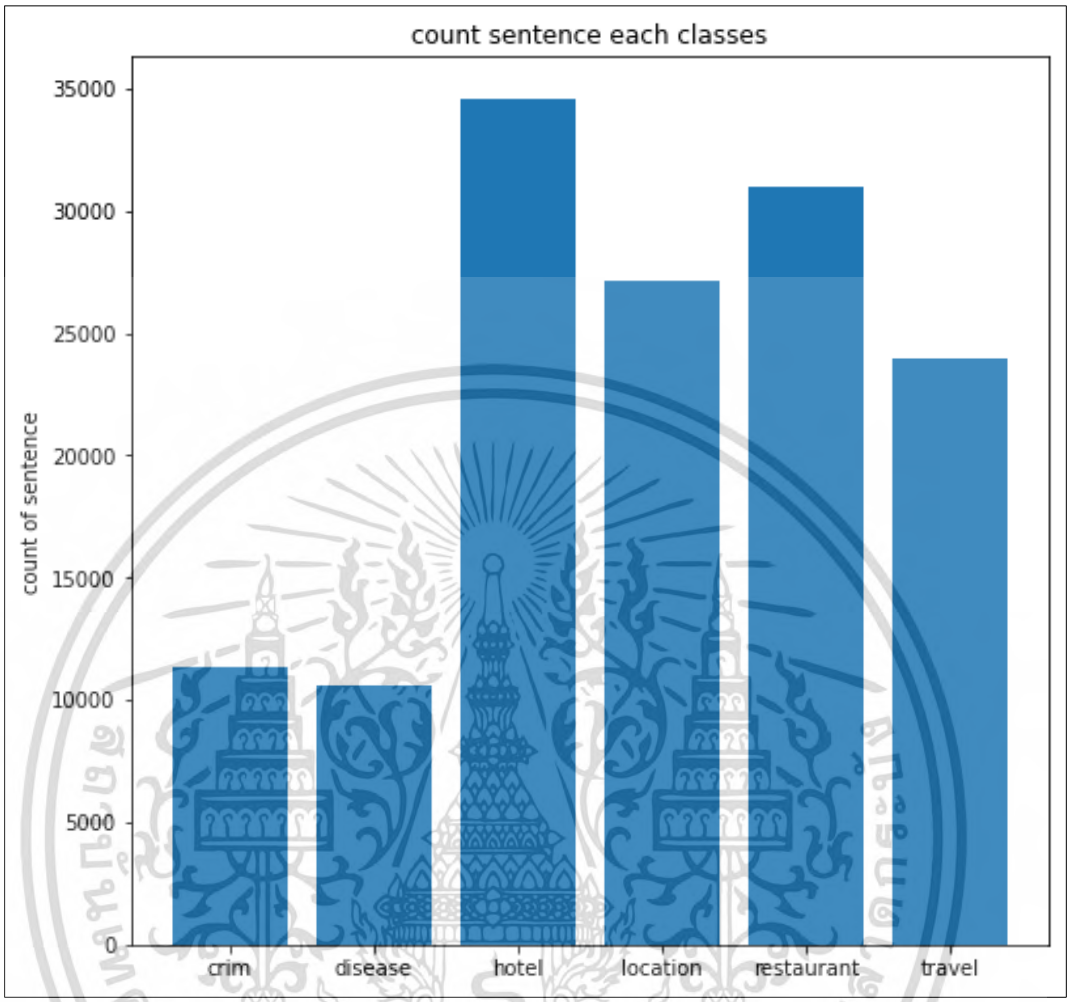
4.1.3 การแบ่งชุดข้อมูล

การแบ่งชุดข้อมูลสำหรับการฝึกสอน และการทดสอบนั้น ในอัลกอริทึมการเรียนรู้ด้วยตัวเองส่วนใหญ่นั้น จำเป็นที่จะต้องมีการรักษาสัดส่วนของจำนวนชุดข้อมูลในแต่ละคำตอบต่างๆ ไม่ให้ต่างกันมากเกินไป เพื่อไม่ให้โมเดลนั้นเรียนรู้แล้วมีแนวโน้มไปทางคำตอบที่มีสัดส่วนใหญ่

โดยได้ทำการแบ่งสัดส่วนของชุดข้อมูลฝึกสอนกับชุดข้อมูลทดสอบเป็นอัตราส่วน 70:30 ตามลำดับ

ตาราง 4.7 จำนวนข้อมูลที่ถูกนำเข้ามาเพื่อฝึกสอนโมเดล

label	amount	probability
crime	3407	0.08
disease	3173	0.07
hotel	10383	0.24
restaurant	8147	0.20
travel	9301	0.22
location	7181	0.17
total	41592	1



รูป 4.10 แสดงจำนวนประโยคของแต่ละคลาส

ตาราง 4.8 ตัวอย่างประโยคในแต่ละคำตอบต่างๆ ในชุดข้อมูลเรียนรู้

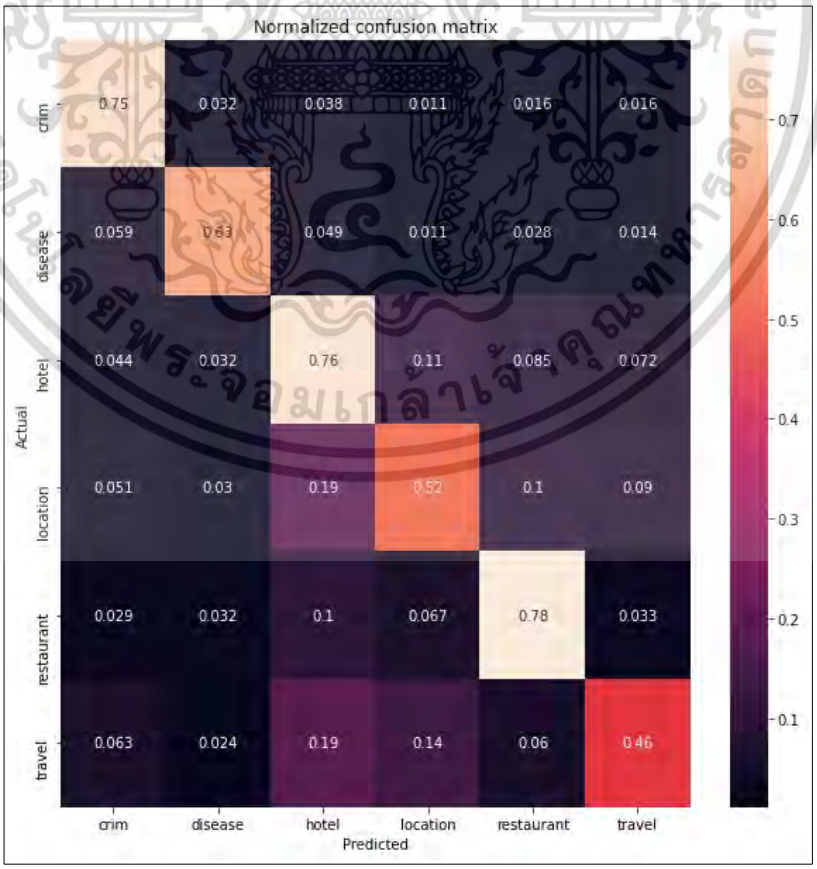
classes	sentence
crime	เพราะมันเป็นการทำลายมนุษยชาติ3 มุมมองสุดท้าย อา...
	ก็เป็นเรื่องชั่ว พอๆ กัน
disease	เหงื่อถือเป็นสารคัดหลั่งไหมคะ
	โควิด 19 จะส่งผลระดับไหนต่อการเที่ยวสงกรานต์?
hotel	พอจะมีที่พักรายเดือนราคาไม่สูงมาแนะนำไหมคะ ในอ...
	ที่สะดวกเลี้ยงแมวไทยสองตัวได้ด้วย ขอบคุณ

ตาราง 4.9 ตัวอย่างประโยคในแต่ละคำตอบต่างๆ ในชุดข้อมูลเรียนรู้(ต่อ)

classes	sentence
location	ที่แนะนำแบบไปแล้วประทับใจไหมคะอีกอย่างเคยไปช่ว…
	ว่าจะไปเดินเก็บภาพแสงเช้าเย็นด้วยอ่ะครับ มีจุดไหน
restaurant	เราไปถึงร้านกับเพื่อน 2 คน เลือกประเภทบุฟเฟ่...
	ที่อยากทานปลาแซลม่อนหรือเมนูอื่นๆเพิ่ม
travel	จากสนามเสือป่า สำนักพระราชวัง มาที่กรมชลประทา...
	สาย 1 ถ้าเปลี่ยนเส้นทางแบบนี้ตลอดดีไหมครับ คนท...

4.1.4 การทดสอบประสิทธิภาพของโมเดล

การทำนายชุดข้อมูลทดสอบโดยโมเดลที่ฝึกสอนด้วยชุดข้อมูลฝึกสอนนั้นมีคะแนนความแม่นยำเฉลี่ยที่ 0.69% (precision average) มีคะแนนความถูกต้องเฉลี่ยที่ 0.65% (recall average) มีคะแนน F1 เฉลี่ย ที่ 66% (F1 score average)



รูป 4.11 confusion matrix ของการทดสอบชุดข้อมูลชุดทดสอบ

จากตาราง 4.7 จะเห็นได้ว่า แม้ว่าสัดส่วนของคลาส crime จะมีสัดส่วนประโยคต่อทั้งหมดอยู่ที่ 8% เท่านั้น แต่พบว่ายังสามารถทำนายกลุ่มได้ดีอยู่ ซึ่งแสดงให้เห็นว่าโมเดลกับข้อมูลชุดทดสอบยังสามารถสร้างน้ำหนักเฉพาะคลาสได้ดีอยู่

อัตราการทำนายผิดของชุดคำตอบในคลาส hotel และคลาส location มีความผิดพลาดสูงกว่าคลาสอื่นๆอยู่มาก อันเนื่องมาจากความคลุมเคลือของระหว่าง 2 คลาสนี้ ที่เกิดจากขั้นตอนการสร้างแท็กคำตอบอัตโนมัติโดยใช้แท็กที่มีอยู่แล้ว เพราะว่ากระทู้โดยส่วนใหญ่นั้น เวลาพูดถึงสถานที่ท่องเที่ยว (location) นั้น จะพบประโยคที่เกี่ยวกับที่พัก (hotel) ด้วยเสมอ

การทำนายว่าจะมีโอกาสของผลลัพธ์ของการทำนายว่าเป็นแบบใด ในกรณีการทำนาย 2 คลาส (ในกรณีที่มีหลายคลาสนั้น ก็ใช้การจับคู่ไม่ซ้ำทุกความเป็นไปได้) นั้น มีโอกาสความเป็นไปได้ 4 แบบ คือ

1) True Positive จำนวนเอกสารที่เป็นคลาสนั้น แล้วทำนายว่าเป็นคลาสนั้น

2) False Positive จำนวนเอกสารที่ไม่ได้เป็นคลาสนั้น แต่ทำนายว่าเป็นคลาสนั้น

3) False Negative จำนวนเอกสารที่เป็นคลาสนั้น แล้วทำนายว่าไม่ได้เป็นคลาสนั้น

4) True Negative จำนวนเอกสารไม่ได้เป็นคลาสนั้น แล้วทำนายว่าไม่ได้เป็นคลาสนั้น

ซึ่ง

$$Percision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

หมายความว่า ยอมรับได้กับโอกาสของการปรากฏของเอกสารที่ไม่เกี่ยวของของคลาสนั้น

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

หมายความว่า ยอมรับได้กับโอกาสของการที่เอกสารนั้นอาจจะไม่ได้ปรากฏ ทั้งที่เอกสารนั้นควรที่จะปรากฏขึ้น

$$F1 = \frac{2}{(\frac{1}{Percision} + \frac{1}{Recall})}$$

เป็นค่าเฉลี่ยระหว่าง Percision และ Recall

ตาราง 4.10 การวัดผลชุดทดสอบด้วย precision, recall และ f1-score และคะแนนเฉลี่ย

label	precision	Recall	f1-score	support
crime	0.75	0.75	0.75	0.08
disease	0.81	0.63	0.71	0.07
hotel	0.57	0.76	0.65	0.24
location	0.61	0.52	0.56	0.20
restaurant	0.73	0.78	0..75	0.22
travel	0.67	0.46	0.54	0.17
accuracy			0.66	41592
Macro avg	0.69	0.65	0.66	41592
Weightrf avg	0.66	0.66	0.65	41592

จากตาราง 4.10 พบว่า ค่า precision กับ recall นั้นใกล้เคียงกันมาก แสดงถึงประสิทธิภาพในการทำนายว่าจะมีโอกาสเกิดเหตุการทำนายว่าเป็นคลาส A ทั้งที่คำตอบจริงๆ เป็นคลาส A หรือว่าจะทำนายว่าเป็นคลาส B ทั้งที่คำตอบจริงๆเป็นคลาส A เท่าๆกัน ซึ่งในการค้นหาค่า recall ที่น้อยนั้นเป็นสิ่งที่ยอมรับไม่ได้ เนื่องจากหากค้นหาสิ่งที่เกี่ยวข้องกับ เซต A แต่พบเซท B นั้นถูกยอมรับไม่ได้ น้อยกว่าหาเซต A แต่ไม่เจอเซต A

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 บทสรุป

ภาพรวมของระบบที่ถูกพัฒนาโดยการร่วมกันระหว่างโมเดลการเรียนรู้และระบบการค้นหาข้อมูลจากอีลาสติกเสิร์ชนั้นสามารถทำงานเชื่อมต่อกันได้ดีในระดับยอมรับได้ เพียงแต่การนำข้อมูลจากระบบโมเดลการเรียนรู้ด้วยตัวเองเพื่อนำไปใช้ในระบบค้นหานั้นไม่สามารถนำข้อมูลที่เหมาะสมไปใช้ได้เพราะระบบโมเดลการเรียนรู้ส่งข้อมูลที่เป็นผลสำเร็จมากเกินไปคือได้เป็นคำตอบที่เป็นภาษาธรรมชาติเลย แต่ระบบค้นหานั้นต้องการข้อมูลที่เป็นระบบเวกเตอร์เพื่อนำไปใช้ นั้นส่งผลให้ระบบอาจจะไม่สามารถทำงานได้อย่างเต็มประสิทธิภาพเท่าที่ควร

ในส่วนของการข้อมูลที่มีอยู่ในระบบทั้งที่ถูกใช้ในการฝึกสอนโมเดลและระบบข้อมูลนั้นมีปริมาณที่มากพอสมควรแต่รูปแบบของประโยคไม่ได้หลากหลายครอบคลุมความหมายของภาษามนุษย์มากเท่าที่ควร ส่งผลให้ทั้งระบบค้นหาที่อาจไม่สามารถค้นหาบทความที่สอดคล้องกับความต้องการที่สื่อถึงความเชื่อมโยงกันระหว่างคำมากกว่าหนึ่งคำเช่น ผู้ใช้อาจต้องการค้นหาบทความที่เกี่ยวข้องกับ ที่พักในกทมที่มีรถเมล์สายแปดผ่าน เนื่องจากระบบไม่มีบทความที่ครอบคลุมความต้องการเหล่านี้จึงอาจจะมีผลลัพธ์ที่ไม่ตรงความต้องการมากนักและอีกส่วนคือโมเดลการเรียนรู้ด้วยตัวเองของคอมพิวเตอร์ที่ส่งผลโดยตรงจากการข้อมูลที่ไม่ครอบคลุมนี้ทำให้ไม่สามารถเข้ารูปแบบของบางประโยคได้

5.2 ปัญหาอุปสรรคและแนวทางแก้ไข

1) การตัดบทความมาเป็นประโยคในภาษาไทยนั้น ไม่สามารถตัดได้อย่างมีประสิทธิภาพในหากหลายกรณี เช่น
   - **ก)** ประโยคนั้นประกอบไปด้วยคำใหม่ ที่ไม่เคยอยู่ในพจนานุกรมของอัลกอริทึมในการตัดคำ นั้นแก้ได้โดยการเพิ่มคำนั้นๆ ในพจนานุกรม
   - **ข)** การสะกดคำผิด ทำให้หากใช้บางอัลกอริทึม ทำให้กลายเป็นคำใหม่ไป ซึ่งอาจจะแก้ได้โดยการแก้คำผิดจากตำแหน่งของ keyboard แต่ถ้าหากเป็นภาษาวิบัตินั้น ยังไม่สามารถหาวิธีแก้ไขที่มีประสิทธิภาพได้

2) การสร้างข้อมูลฝึกสอนโดยการติดคำตอบจากแท็กให้อัตโนมัตินั้น ไม่สามารถให้ข้อมูลประโยคนั้นถูกต้องในบางกรณี ที่เอกสารนั้นสามารถจัดอยู่ได้หลายหมวดหมู่ ซึ่งวิธีแก้ไขนั้นเปลี่ยนโมเดลเป็นการให้คำตอบแบบหลายคำตอบ (multi-label)

3) ตัว pretain-model ของโมเดล BERT ภาษาไทยนั้น ทางผู้จัดทำไม่มีทรัพยากรในการสร้างขึ้นมาใช้เอง ทำให้ไม่สามารถรับรู้ผลลัพธ์ได้ว่า จะเพิ่มประสิทธิภาพของโมเดลได้ดีหรือไม่

5.3 แนวทางในการพัฒนา

1) ในงานนี้ มีการใช้ข้อมูลจากแหล่งเดียว ทำให้ขาดความหลากหลายของข้อมูล ซึ่งสามารถนำเข้าข้อมูลจากหลายๆ แหล่งได้

2) การหาเทคนิกใหม่ๆ ในการให้คำตอบกับชุดข้อมูลฝึกสอน

3) การใช้การเรียนรู้เชิงลึก โครงสร้างใหม่ๆ ที่อาจจะทำให้สามารถแยกประโยคได้มีประสิทธิภาพขึ้น

บรรณานุกรม

Ashish, V., Noam, S. and Niki P., 2017. “Attention Is All You Need.”, google research and google brain

Jacob, D., Ming-Wei, C., Kenton, Lee. and Kristina Toutanova. 2018. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”, Google AI Language.

Alammar, Jay. 2018. “Visualizing A Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention).”, *[Online].Available : https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/*

Alammar, Jay. 2018. “The Illustrated Transformer.”, [Online].Available : https://jalammar.github.io/illustrated-transformer/

Alammar, Jay. 2018. “The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning).”, [Online].Available : https://jalammar.github.io/illustrated-bert/

Javaid Nabi. 2019. “Building a Multi-label Text Classifier using BERT and TensorFlow.”, [Online].Available : https://towardsdatascience.com/building-a-multi-label-text-classifier-using-bert-and-tensorflow-f188e0ecdc5d

u41ppp. 2018. “All You Need Is Attention … แค่ใส่ใจกันเท่านั้นก็พอ". [Online].Available : https://medium.com/@u41ppp/all-you-need-is- attention-แค่ใส่ใจกันเท่านั้นก็พอ-f473f112db12

u41ppp. 2018. BERT-th. [Online].Available : https://github.com/ThAIKeras/bert/