ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

นายอัษฎาวุธ ชนะกิจการโชค

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2559

Automated Webpage Categorization System based on Word Statistics

Mr. Adsadawut Chanakitkarnchok

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2016

หัวข้อวิทยานิพนธ์	ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ
โดย	นายอัษฎาวุธ ชนะกิจการโชค
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร. กุลธิดา โรจน์วิบูลย์ชัย

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร. สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ณัฐวุฒิ หนูไพโรจน์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(รองศาสตราจารย์ ดร. กุลธิดา โรจน์วิบูลย์ชัย)

.....กรรมการ
(ดร. พีรพล เวทีกูล)

.....กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร. กฤษณะ ไวยมัย)

อัษฎาวุธ ชนะกิจการโชค : ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ (Automated Webpage Categorization System based on Word Statistics) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร. กุลธิดา โรจน์วิบูลย์ชัย, 61 หน้า.

เมื่อยุคข้อมูลข่าวสารมาถึง อินเทอร์เน็ตถือเป็นปัจจัยในการดำรงชีวิตอีกอย่างหนึ่ง เพราะทั้งข้อมูลข่าวสารรวมไปถึงบริการต่าง ๆ สามารถเข้าถึงได้ผ่านทางอินเทอร์เน็ต ในขณะที่ข้อมูลข่าวสารที่มีอยู่ในอินเทอร์เน็ตมีปริมาณเพิ่มขึ้นอย่างมากในแต่ละปี และการวิเคราะห์ข้อมูลขนาดใหญ่เริ่มเป็นที่จับตามองของทุกคน มีบริษัทยักษ์ใหญ่มากมาย อาทิ กูเกิล เฟสบุ๊ค อเมซอน และเน็ตฟลิกซ์ ต่างก็กำลังสนใจการนำข้อมูลที่มีอยู่มาวิเคราะห์เพื่อปรับปรุงการให้บริการให้ดียิ่งขึ้น นอกจากนี้จำนวนเว็บไซต์ที่ได้ทำการจดทะเบียนแล้วยังมีปริมาณเพิ่มขึ้นอย่างน่าเหลือเชื่อถึงหนึ่งพันล้านเว็บไซต์ เว็บไซต์หนึ่งเว็บประกอบไปด้วยเว็บเพจมากมายที่สามารถกล่าวถึงหัวข้อหลายประเภทได้ การจัดประเภทเว็บเพจนั้นจำเป็นต้องมีระบบที่สามารถใช้ในการจัดประเภทเว็บเพจได้ก่อน โดยที่ระบบจัดประเภทเว็บเพจสามารถนำไปใช้ในการคัดกรองเว็บไซต์ที่ไม่เหมาะสม ระบุความสนใจของผู้ใช้งาน และยังสามารถติดฉลากให้กับเนื้อหาแบบอัตโนมัติได้อีกด้วย ปัจจุบันมีส่วนต่อประสานโปรแกรมประยุกต์ในเชิงพาณิชย์ที่ใช้ในการจัดประเภทเว็บเพจที่สามารถจัดประเภทเว็บเพจเป็นหมวดหมู่ได้หลากหลาย ส่วนต่อประสานโปรแกรมประยุกต์เหล่านี้สามารถแบ่งออกได้เป็น 2 ชนิด ส่วนต่อประสานโปรแกรมประยุกต์กลุ่มแรกจัดประเภทจากหน้าเว็บเพจแรกของหน้าเว็บไซต์ที่เรียกว่าโฮมเพจเท่านั้น ส่วนต่อประสานโปรแกรมประยุกต์กลุ่มที่สองที่จัดประเภทเนื้อหาในหน้าเว็บเพจโดยเฉพาะ เนื่องจากส่วนต่อประสานโปรแกรมประยุกต์ในกลุ่มที่สองจัดประเภทเนื้อหาโดยมีพื้นฐานจากเนื้อหาที่อยู่ภายใน ไม่ใช่แค่หน้าโฮมเพจเท่านั้น ส่วนต่อประสานโปรแกรมประยุกต์ในในกลุ่มที่สองมีแนวโน้มที่จะให้ความแม่นยำมากกว่า จากการศึกษาที่ผ่านมาพบว่า ส่วนต่อประสานโปรแกรมประยุกต์ที่มีอยู่ในท้องตลาดไม่สามารถจัดประเภทเว็บเพจโดยการพิจารณาเนื้อหาที่เป็นภาษาไทยได้ และในงานวิจัยที่ผ่านมาไม่ได้พิจารณาเรื่องความเร็วในการประมวลผลและการทำเป็นระบบอัตโนมัติที่รองรับการวิเคราะห์เว็บเพจจากการจราจรทางเครือข่ายอินเทอร์เน็ตจริงได้ จากปัญหาที่กล่าวมาในข้างต้น งานวิจัยนี้จึงได้เสนอระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ ระบบนี้จะทำการประมวลผลข้อมูลจากยูอาร์แอลดิบและทำการจัดประเภทเนื้อหา ยิ่งไปกว่านั้นระบบนี้สามารถรองรับทั้งภาษาไทยและภาษาอังกฤษและรองรับการจัดประเภทเว็บเพจที่มีปริมาณมากได้ ระบบนี้ประกอบไปด้วยระบบย่อย 2 ระบบ ระบบย่อยแรกคือระบบสกัดคำสำคัญอัตโนมัติ ระบบย่อยนี้ใช้สำหรับสกัดคำสำคัญในเนื้อหาของเว็บเพจเพื่อที่จะนำไปใช้สร้างพจนานุกรม ซึ่งพจนานุกรมนี้จะถูกใช้ในระบบย่อยที่สอง ระบบย่อยที่สองคือระบบจัดประเภทเว็บเพจ ระบบย่อยนี้จะทำการประมวลผลข้อมูลดิบและทำการจัดประเภทเนื้อหาไปยังหมวดหมู่ที่เหมาะสม ผลลัพธ์ของระบบมีค่าประสิทธิภาพโดยรวมมากถึงร้อยละ 99 และใช้เวลาในการประมวลผลโดยรวมร้อยกว่าอัลกอริทึมอื่น ยิ่งไปกว่านั้นงานวิจัยยังแสดงให้เห็นว่าระบบนี้สามารถจัดประเภทเว็บเพจด้วยวิธีการที่ง่ายแต่ได้ผลดี

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ปีการศึกษา 2559

5870279421 : MAJOR COMPUTER ENGINEERING
KEYWORDS: BIGDATA / URL CATEGORIZATION / TEXT CLASSIFICATION / WEBPAGE CATEGORIZATION SYSTEM

ADSADAWUT CHANAKITKARNCHOK: Automated Webpage Categorization System based on Word Statistics. ADVISOR: ASSOC. PROF. KULTIDA ROJVIBOONCHAI, 61 pp.

Since the information era has come, the Internet has become one of our living factors because every information and services can be accessed though the Internet. While the information in the Internet has been dramatically increasing among a year and the trend of big data has already been kept eyes on from everyone. A number of big companies such as Google, Facebook, Amazon, and Netflix are also interested in analyzing their data for improving their services. Additionally, the number of registered websites are incredibly increasing up to one billion websites. A website can contain many webpages which can be talked about different topics. To classify each webpage, webpage categorization system is needed. The webpage categorization system can be used for filtering inappropriate websites, identifying user interests, and also automatically labeling contents. There are various commercial webpage categorization APIs which can classify webpages into a number of categories. They can be grouped into 2 types. The first group is to classify only the first webpage, so-called “home page”, of the website. The second group is to classify the contents inside the particular webpage. Since the second group classifies the webpage based on the contents inside, not just the home page, the second group tends to achieve more accuracy. To the best of our knowledge, the existing commercial webpage categorization APIs have not been able to categorize the webpage by considering Thai contents and the previous researches have not considered the computation time and the automated system which can preserve the real Internet traffic. From the above-mentioned problems, this research proposes an automated webpage categorization system based on word statistics. This system will preprocess data from the raw URLs and then categorize the contents. Furthermore, it can support both Thai and English languages and also preserve the high Internet traffic volume. This system has 2 sub-systems. The first sub-system is automatic keyword extraction system. It is used to extract the keywords in the content of categorized webpage for creating the dictionary. The dictionary is used in the second sub-system. The second sub-system is webpage categorization system. It is used to preprocess raw data and categorize the content into the appropriate category. The result of the system can yield the F-Measure up to 0.99 and spend less overall computation time than other existing algorithms. Moreover, this research show that this system can categorize webpages with the simple but powerful technique.

Department:	Computer Engineering	Student's Signature	.....
Field of Study:	Computer Engineering	Advisor's Signature	.....
Academic Year:	2016		

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จากรองศาสตราจารย์ ดร. กุลธิดา โรจน์วิบูลย์ชัย อาจารย์ที่ปรึกษาวิทยานิพนธ์ อาจารย์ได้ให้คำปรึกษาและข้อคิดเห็นต่าง ๆ สำหรับพัฒนางานวิจัย อีกทั้งยังให้คำแนะนำเพื่อช่วยแก้ปัญหาที่เกิดขึ้นระหว่างดำเนินงานวิจัยอีกด้วย

ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ ได้แก่ ผู้ช่วยศาสตราจารย์ ดร.ณัฐวุฒิ หนูไพโรจน์ ดร. พีรพล เวทีกูล และรองศาสตราจารย์ ดร. กฤษณะ ไวยมัย ที่ได้ให้คำแนะนำซึ่งเป็นประโยชน์ต่อวิทยานิพนธ์ฉบับนี้

ขอขอบคุณทุนอุดหนุนการศึกษาอัจฉริยะคืนรัง จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

งานวิจัยชิ้นนี้เกิดจากความร่วมมือในการวิจัยร่วมในโครงการพัฒนาระบบวิเคราะห์ข้อมูลเครือข่ายโทรศัพท์เคลื่อนที่เพื่อการแบ่งกลุ่มผู้ใช้บริการกับบริษัทแอดวานซ์ อินโฟร์ เซอร์วิส

ขอขอบคุณสมาชิกทุกคนในห้องปฏิบัติการโดยเฉพาะ ดร. กุลิสร์ ณ นคร ที่ให้ความคิดเห็นและข้อเสนอแนะสำหรับการทำวิจัยตลอดระยะเวลา 2 ปีที่ผ่านมา

สุดท้ายนี้ขอขอบคุณคุณพ่อ คุณแม่และครอบครัวที่เป็นกำลังใจให้ตลอดระยะเวลาที่ทำการวิจัยจนกระทั่งสำเร็จการศึกษา

สารบัญ

หน้า

หน้า

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อินเทอร์เน็ต (Internet) ถือเป็นอีกปัจจัยสำคัญในการดำรงชีวิตอีกปัจจัยหนึ่งสำหรับมนุษย์ในยุคข้อมูลข่าวสาร โดยเราสามารถเข้าถึงเนื้อหาและบริการต่าง ๆ ได้จากที่ไหนเมื่อไหร่ก็ได้ผ่านอินเทอร์เน็ต ไม่ว่าจะเป็นการติดต่อสื่อสาร ความบันเทิง ข่าว รวมไปถึงการค้าขายและบริการอื่น ๆ อีกมากมาย ข้อมูลปริมาณมหาศาลที่ถูกสร้างขึ้นในโลกของอินเทอร์เน็ตในแต่ละวินาทีมีปริมาณมากกว่า 44 เทระไบต์ [1] ซึ่งในยุคแห่งข้อมูลข่าวสารการนำข้อมูลเหล่านี้มาวิเคราะห์เพื่อหาสิ่งที่มีประโยชน์นั้นสามารถสร้างมูลค่าได้เป็นอย่างมาก เนื่องจากผู้ให้บริการต่าง ๆ ดังแสดงในภาพที่ 1 ไม่ว่าจะเป็นบริษัทยักษ์ใหญ่อย่างบริษัทกูเกิล (Google) ทำการเก็บข้อมูลสถิติการสืบค้นข้อมูลบริษัทเฟสบุ๊ค (Facebook) ทำการเก็บข้อมูลการใช้งานผู้ใช้บริการเน็ตฟลิกซ์ (Netflix) ทำการเก็บข้อมูลสถิติการค้นหาและเข้าชมภาพยนตร์ บริษัทอเมซอน (Amazon) ทำการเก็บข้อมูลสถิติการค้นหาและซื้อสิ้นค้า รวมไปถึงผู้ให้บริการเครือข่ายอินเทอร์เน็ตที่ได้ทำการเก็บข้อมูลการใช้งานของผู้ใช้บริการไว้เพื่อนำมาวิเคราะห์สำหรับการปรับปรุงการบริการหรือทำการตลาดให้ตอบโจทย์ของผู้ใช้บริการอย่างตรงจุด ในส่วนของผู้ให้บริการเครือข่ายอินเทอร์เน็ตในประเทศไทยนั้นได้มีการเก็บข้อมูลการเข้าใช้งานอินเทอร์เน็ตของผู้ใช้บริการตามพระราชบัญญัติคอมพิวเตอร์ ซึ่งผู้ให้บริการอินเทอร์เน็ตสามารถนำข้อมูลเหล่านั้นไปใช้เพื่อพัฒนาการบริการได้ เช่น การเก็บแคชของเว็บเพจที่มีการเข้าใช้งานเป็นจำนวนมากเพื่อเพิ่มความเร็วในการเข้าถึงเว็บเพจนั้น ๆ ของผู้ใช้บริการ หรือแม้แต่การหาความสนใจของกลุ่มผู้ใช้บริการจากข้อมูลการใช้งานเพื่อเสนอโปรโมชันที่ตรงกับความต้องการของผู้ใช้บริการได้อย่างถูกต้องแม่นยำ

ภาพที่ 1 แสดงตัวอย่างบริษัทที่ทำการเก็บข้อมูลการใช้งานของผู้ใช้บริการ

ระบบจัดประเภทเว็บเพจเป็นระบบออกแบบมาเพื่อใช้จัดประเภทเว็บเพจจากข้อมูลยูอาร์แอล (URL : Uniform Resource Locator) โดยระบบจัดประเภทเว็บเพจนี้สามารถนำใช้ในการคัดกรองเว็บเพจที่ไม่เหมาะสม เช่น เว็บเพจที่เกี่ยวข้องการพนัน หรือเว็บเพจที่เกี่ยวกับสื่อลามกอนาจาร เป็นต้น ระบบจัดประเภทเว็บเพจสามารถนำไปใช้ในการจัดหมวดหมู่หรือห้องสำหรับเว็บจำพวกเว็บบอร์ดหรือเว็บข่าวต่างๆ นอกจากนั้นระบบจัดประเภทเว็บเพจยังสามารถใช้หาความสนใจของผู้ใช้งาน จากการเก็บข้อมูลการเข้าอินเทอร์เน็ตของผู้ใช้งาน แล้วทำการจัดประเภทยูอาร์แอลที่ทำการเก็บได้ เราจะสามารถทราบถึงประเภทของยูอาร์แอลที่ผู้ใช้งานกำลังสนใจเข้าชมอยู่ และสามารถรู้ความสนใจของผู้ใช้งานได้

จากการที่ผู้ให้บริการเครือข่ายอินเทอร์เน็ตรายหนึ่งต้องการปรับปรุงการให้บริการผู้ใช้บริการภายในเครือข่าย จึงเกิดเป็นการทำวิจัยร่วมระหว่างมหาวิทยาลัยและผู้ให้บริการเครือข่ายขึ้น โดยผู้ให้บริการต้องการนำข้อมูลการใช้งานของผู้ใช้บริการวิเคราะห์เพื่อแบ่งกลุ่มผู้ใช้งานออกเป็นกลุ่มย่อยเพื่อปรับปรุงการบริการให้ตอบสนองความต้องการของผู้ใช้บริการมากยิ่งขึ้น โดยข้อมูลจากทางผู้ให้บริการมีด้วยกันหลายอย่าง อาทิ ข้อมูลการใช้งานโทรศัพท์ ข้อมูลการใช้อินเทอร์เน็ตผ่านอุปกรณ์สื่อสารแบบพกพา และข้อมูลตำแหน่งเสาสัญญาณ ข้อมูลเหล่านี้จะถูกนำมาวิเคราะห์และแบ่งออกเป็นหัวข้อย่อย 3 หัวข้อ ได้แก่ ข้อมูลการใช้งานโทรศัพท์ (Call Detail Recode: CDR) ข้อมูลการเคลื่อนที่ (User Mobility) และความสนใจของผู้ใช้งาน (User Interest) ในหัวข้อความสนใจของผู้ใช้งานเป็นหัวข้อที่ตรงกับสิ่งที่ผู้วิจัยกำลังสนใจอยู่ ผู้วิจัยจึงได้เข้าร่วมในโครงการเพื่อทำระบบสำหรับวิเคราะห์หาสนใจของผู้ใช้งาน โดยในหัวข้อความสนใจนี้ใช้ข้อมูลการใช้อินเทอร์เน็ตผ่านอุปกรณ์สื่อสารแบบพกพามาใช้ในการวิเคราะห์ ซึ่งทางผู้วิจัยมีพื้นฐานความรู้ทางด้านเครือข่ายจึงคิดว่าสามารถนำข้อมูลการใช้อินเทอร์เน็ตมาวิเคราะห์เพื่อให้เกิดประโยชน์ได้

ข้อมูลการใช้งานอินเทอร์เน็ตผ่านอุปกรณ์สื่อสารแบบพกพาที่น่าสนใจและสามารถนำมาใช้ในการหาความสนใจของผู้ใช้งานได้คือข้อมูลการเข้าใช้งานเว็บไซต์ (Website) ผ่านเว็บบราวเซอร์ (Web Browser) บนอุปกรณ์สื่อสารแบบพกพา โดยข้อมูลดังกล่าวเรียกว่ายูอาร์แอล โดยยูอาร์แอลเป็นที่อยู่ที่ใช้สำหรับระบุถึงเว็บไซต์ที่ต้องการเข้าถึง ซึ่งปัจจุบันมีปริมาณเว็บไซต์มากถึง 1 พันล้านเว็บไซต์ที่ได้ทำการจดทะเบียนแล้วและมีแนวโน้มที่จะมีปริมาณเพิ่มขึ้นอย่างต่อเนื่อง เว็บไซต์เหล่านี้ถูกสร้างขึ้นเพื่อตอบสนองความต้องการของผู้ใช้บริการอินเทอร์เน็ตในด้านของการสืบค้นข้อมูล ซื้อขายสินค้าออนไลน์ ติดต่อสื่อสารผ่านสังคมออนไลน์ รวมไปถึงการเข้าถึงความบันเทิงและบริการอื่นอีกมากมาย ในเว็บไซต์แต่ละเว็บไซต์ประกอบไปด้วยหน้าเว็บเพจต่างๆ ที่มีเนื้อหาเป็นข้อความ รูปภาพ หรือวิดีโอ โดยเนื้อหาเหล่านี้สามารถนำมาเป็นสิ่งที่บ่งชี้ถึงความสนใจของผู้ใช้บริการได้ เนื้อหาภายในเว็บเพจเป็นสิ่งที่ดึงดูดผู้ใช้บริการให้เกิดความสนใจเข้ามาใช้บริการ โดยการวิเคราะห์เว็บเพจเพื่อหาความสนใจของผู้ใช้บริการจะใช้วิธีการจัดประเภทของเว็บเพจแต่ละเว็บเพจ และใช้ประเภทที่ได้มาบ่งบอก

ความสนใจของผู้ใช้บริการ แต่ไม่ใช่ทุกเว็บเพจจะประกอบไปด้วยข้อความ รูปภาพ และวิดีโอครบทั้งสามอย่าง บางเว็บเพจอาจมีเพียงข้อความกับรูปภาพ ข้อความกับวิดีโอ หรือข้อความอย่างเดียวก็ได้ แต่โดยทั่วไปแล้ว เว็บเพจส่วนมากจะมีข้อความเป็นส่วนประกอบของเว็บเพจเสมอ ดังนั้นการนำข้อความมาพิจารณาเพื่อจัดประเภทเว็บเพจสามารถครอบคลุมเว็บเพจส่วนใหญ่ได้

ในปัจจุบันมีส่วนต่อประสานโปรแกรมประยุกต์ (API : Application Programming Interface) ที่สามารถจัดประเภทเว็บเพจได้มากมาย ดังที่แสดงตัวอย่างในภาพที่ 2 ยกตัวอย่างเช่น SimilarWeb [2] เป็นผู้ให้บริการการวัดผลเว็บ (Web Measurement) รายหนึ่งที่ทำการวัดผลเว็บไซต์และแอปพลิเคชันบนอุปกรณ์สื่อสารแบบพกพาทั้งระบบปฏิบัติการแอนดรอยด์ (Android) และ ไอโอเอส (iOS) Bluecoat Webpulse [3] เป็นระบบป้องกันภัยคุกคามแบบร่วมมือกันระหว่างผู้ใช้ โดยทำการเก็บข้อมูลการใช้งานเว็บผ่านอุปกรณ์ของบลูโค้ท แล้วทำการส่งมาวิเคราะห์ที่ศูนย์กลาง ทั้งนี้ยังสามารถให้ผู้ใช้งานช่วยกันรายงานข้อผิดพลาดในจัดหมวดหมู่อีกด้วย AlchemyAPI [4] เป็นส่วนต่อประสานโปรแกรมประยุกต์ที่สามารถจัดหมวดหมู่ของเว็บไซต์โดยการดูเนื้อหาภายในเว็บไซต์ด้วยระบบวัตสัน (Watson) และ AYLIEN [5] เป็นอีกหนึ่งส่วนต่อประสานโปรแกรมประยุกต์ที่สามารถจัดหมวดหมู่ของเว็บเพจได้ แต่ส่วนต่อประสานโปรแกรมประยุกต์เหล่านี้ไม่สามารถจัดประเภทเว็บเพจที่มีเนื้อหาภายในเว็บเพจเป็นภาษาไทยได้ จากข้อมูลการเข้าใช้งานอินเทอร์เน็ตนั้นมีเว็บเพจที่มีเนื้อหาเป็นภาษาไทยมากถึงร้อยละ 30 การไม่มีระบบที่สามารถจัดประเภทเว็บเพจที่มีเนื้อหาที่เป็นภาษาไทยได้ทั้งๆ ที่เป็นการวิเคราะห์หาความสนใจของผู้ใช้งานที่เป็นคนไทยอาจส่งผลให้ผลการวิเคราะห์คลาดเคลื่อนได้ ดังนั้นระบบจัดประเภทเว็บเพจที่สามารถพิจารณาเนื้อหาที่เป็นภาษาไทยได้จึงเป็นสิ่งจำเป็น

ภาพที่ 2 แสดงตัวอย่างบริการส่วนต่อประสานโปรแกรมประยุกต์จัดประเภทเว็บไซต์

งานวิจัยนี้มีวัตถุประสงค์เพื่อเสนอระบบจัดประเภทเว็บเพจที่สามารถจัดประเภทเว็บเพจที่มีเนื้อหาเป็นทั้งภาษาไทยและภาษาอังกฤษได้ โดยระบบนี้จะพิจารณาความเร็วในการประมวลเป็นสำคัญ เพื่อรองรับการวิเคราะห์ข้อมูลการใช้งานอินเทอร์เน็ตที่มีปริมาณมากได้ โดยที่ผลการวิเคราะห์ต้องมีความแม่นยำที่มากกว่าร้อยละ 50

วิธีการวิเคราะห์เพื่อจัดประเภทเว็บเพจในอดีตได้ถูกพัฒนาขึ้นเป็นจำนวนมาก ซึ่งสามารถแบ่งออกได้เป็น 2 แบบ คือ การจัดประเภทโดยสนใจเนื้อหาภายในเว็บเพจ และการจัดประเภทโดยสนใจ

เพียงชื่อโดเมนเท่านั้น โดยการจัดประเภทโดยสนใจเนื้อหาภายในเว็บเพจจะทำให้ทราบหมวดหมู่ของเว็บเพจในแต่ละหน้า ซึ่งต่างจากการจัดประเภทโดยสนใจเพียงชื่อโดเมนที่จะทำการจัดหมวดหมู่โดยสนใจเพียงแค่หน้าเว็บเพจแรกของเว็บไซต์หรือวิเคราะห์จากชื่อโดเมนเพียงอย่างเดียว

จากการศึกษางานวิจัยที่ผ่านมานั้นสามารถแบ่งกระบวนการวิเคราะห์ข้อมูลได้เป็น 3 ประเภท ได้แก่ กระบวนการเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithm) กระบวนการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning Algorithm) และกระบวนการทางสถิติ (Statistical Algorithm) หรือ กระบวนการทางความหมาย (Semantic Algorithm)

กระบวนการเรียนรู้แบบมีผู้สอนเป็นการนำข้อมูลที่มีอยู่แล้วในอดีตมาทำการสร้างโมเดลเพื่อนำไปใช้ในการทำนายหรือคาดการณ์สิ่งที่กำลังจะเกิดขึ้นในอนาคต กระบวนการเรียนรู้แบบนี้สามารถสร้างมาจากสมการทางคณิตศาสตร์หรือกฎเกณฑ์ต่าง ๆ ได้ โดยมีข้อจำกัดอยู่ที่ข้อมูลที่นำมาเรียนรู้หากมีปริมาณไม่มากพอหรือไม่มีประสิทธิภาพจะส่งผลให้โมเดลที่สร้างขึ้นไม่มีประสิทธิภาพตามไปด้วย

กระบวนการเรียนรู้แบบไม่มีผู้สอนเป็นอีกหนึ่งกระบวนการวิเคราะห์ข้อมูล โดยจะนำข้อมูลที่มีอยู่มาทำการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างข้อมูลและทำการจัดกลุ่มข้อมูลที่มีอยู่ออกเป็นกลุ่มย่อย กระบวนเรียนรู้แบบไม่มีผู้สอนนั้นหลังจากทำการแบ่งกลุ่มแล้ว จะไม่ทราบความหมายของแต่ละกลุ่มว่าหมายถึงอะไรจนกว่าจะทำการแปลความหมายด้วยตนเอง

กระบวนการทางสถิติหรือกระบวนการทางความหมายเป็นกระบวนการที่อาศัยวิธีการทางสถิติมาทำการวิเคราะห์ข้อมูลที่มีอยู่แล้ว อาจมีการพิจารณาความหมายของข้อมูลเพื่อหาความสัมพันธ์ของข้อมูลประกอบด้วย

จากกระบวนการต่าง ๆ ที่ได้กล่าวมาในข้างต้นนั้นเป็นการนำข้อมูลที่ผ่านการทำความสะอาดหรือข้อมูลที่ได้ถูกจัดเตรียมไว้ก่อนแล้วมาทำการวิเคราะห์ แต่การวิเคราะห์เพื่อจัดประเภทเว็บเพจจากยูอาร์แอลจริงนั้น มีสิ่งที่ต้องพิจารณามากกว่านั้น เช่น การคัดกรองยูอาร์แอลที่ไม่เกี่ยวข้อง วิธีการร้องขอข้อมูลหน้าเว็บที่ใช้เวลาในการร้องขอน้อย แต่ได้ข้อมูลมากเพียงพอ และการทำความสะอาดข้อมูลดิบที่ได้จากการร้องขอข้อมูลหน้าเว็บเพจ เป็นต้น ซึ่งกระบวนการเหล่านี้เป็นสิ่งที่จะช่วยลดระยะเวลาที่ไม่จำเป็น รวมไปถึงสามารถเพิ่มประสิทธิภาพโดยรวมของการจัดประเภทเว็บเพจอีกด้วย งานวิจัยนี้ได้นำเสนอระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ ระบบแบ่งออกเป็น 2 ส่วน ได้แก่ ระบบสกัดคำสำคัญเป็นระบบที่ใช้สำหรับการสกัดคำสำคัญจากข้อความในเว็บเพจ เพื่อนำมาสร้างเป็นพจนานุกรมคำศัพท์สำหรับใช้ในระบบจัดประเภทเว็บเพจ และระบบจัดประเภทเว็บเพจเป็นระบบที่ทำการจัดประเภทเว็บเพจจากการวิเคราะห์เนื้อหาที่เป็นข้อความภายในเว็บเพจนั้น โดยระบบสกัดคำสำคัญเป็นระบบที่ใช้กระบวนการเรียนรู้แบบมีผู้สอนเพื่อให้ได้พจนานุกรมที่เกิดจากคำในเว็บเพจประเภทนั้นอย่างแท้จริง ในส่วนของระบบจัดประเภทเว็บเพจจะ

ทำงานด้วยกระบวนการทางสถิติ โดยการหาความถี่ของคำที่ปรากฏในเว็บเพจมาตรวจสอบประเภทของคำกับพจนานุกรมที่ได้ทำขึ้นมา

1.2 วัตถุประสงค์ของการวิจัย

1) เพื่อพัฒนาระบบจัดประเภทเว็บเพจโดยการพิจารณาจากเนื้อหาภายในเว็บเพจ

2) เพื่อพัฒนาระบบสกัดคำสำคัญแบบอัตโนมัติสำหรับการจัดทำพจนานุกรม

3) เพื่อสร้างระบบที่ใช้สำหรับหาความสนใจของผู้ใช้งานอินเทอร์เน็ต

1.3 ขอบเขตการวิจัย

1) พัฒนาและศึกษาวิธีจัดประเภทเว็บเพจเพื่อให้สามารถรองรับข้อมูลการจราจรทางเครือข่ายจริงได้ โดยอาศัยเนื้อหาภายในเว็บเพจที่เป็นข้อความมาวิเคราะห์

2) ระบบสามารถรองรับเนื้อหาได้ทั้งภาษาไทยและภาษาอังกฤษ

3) ระบบพิจารณาประเภทตามห้องในกระทู้เว็บพันทิป

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้ระบบสำหรับสกัดคำสำคัญจากเนื้อหาในหน้าเว็บเพจ ซึ่งสามารถนำมาใช้ในการสร้างพจนานุกรมคำ

2) ได้พจนานุกรมคำที่ใช้สำหรับจัดประเภทเว็บเพจที่สามารถรองรับได้ทั้งภาษาไทยและภาษาอังกฤษ

3) ได้ระบบสำหรับจัดประเภทเว็บเพจที่สามารถรองรับข้อมูลจากการจราจรทางอินเทอร์เน็ตจริงได้

4) ได้ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

1.5 วิธีดำเนินการวิจัย

1) ศึกษางานวิจัยที่เกี่ยวข้องกับวิธีการจัดประเภทเว็บเพจ

2) พัฒนาระบบสกัดคำสำคัญ

2.1) ศึกษางานวิจัยที่เกี่ยวข้องกับการสกัดคำสำคัญ

2.2) พัฒนาระบบสกัดคำสำคัญจากเนื้อหาในเว็บเพจ

2.3) ทำการหาเว็บเพจที่ทราบประเภทมาใช้ในการสร้างพจนานุกรม

2.4) ทำการทดลองและวัดประสิทธิภาพของระบบสกัดคำสำคัญ

2.5) ปรับปรุงและพัฒนาพจนานุกรมที่ได้จากระบบสกัดคำสำคัญ

2.6) สรุปและอภิปรายผลการทดลอง

3) พัฒนาระบบจัดประเภทเว็บเพจ

3.1) ศึกษาความเป็นไปได้ในการจัดประเภทเว็บเพจโดยการใช้พจนานุกรม

3.2) สำรวจหาพจนานุกรมคำที่มีอยู่แล้วมาทดสอบ

3.3) พัฒนาอัลกอริทึมสำหรับจัดประเภทเว็บเพจ

3.4) ทำการหาเว็บเพจที่ทราบหมวดหมู่มาทดสอบระบบที่ได้พัฒนาขึ้น

3.5) ทำการทดลองเพื่อวัดประสิทธิภาพของระบบจัดประเภทเว็บเพจ

3.6) สรุปและอภิปรายผลการทดลอง

3.7) ปรับปรุงระบบและพจนานุกรมคำ

3.8) ทำการทดลองเพื่อวัดประสิทธิภาพของระบบจัดประเภทเว็บเพจที่ถูกปรับปรุง

3.9) สรุปและอภิปรายผลการทดลอง

4) ทำการทดลองเพื่อวัดประสิทธิภาพระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

5) สรุปและอภิปรายผลการทดลอง

1.6 ผลงานตีพิมพ์

1) บทความชื่อ “Autonomous website categorization with pre-defined dictionary” โดย Adsadawut Chanakitkarnchok, Kulit Na Nakorn และ Kultida Rojviboonchai ตีพิมพ์ และนำเสนอในงานประชุมวิชาการชื่อ “2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2016)” [6]

2) บทความชื่อ “Automatic Keyword Extraction System for Thai Website Categorization System” โดย Adsadawut Chanakitkarnchok, Kulit Na Nakorn และ Kultida Rojviboonchai ตีพิมพ์ และนำเสนอในงานประชุมวิชาการชื่อ “2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2017)” [7]

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำเป็นระบบการทำงานที่จะทำการวิเคราะห์เว็บเพจจากเนื้อหาภายในเว็บเพจตามยูอาร์แอลที่ได้รับมาเป็นข้อมูลนำเข้า แล้วตอบประเภทของเว็บเพจนั้นคืนกลับไป ด้วยการพิจารณาปริมาณคำสำคัญที่ปรากฏในเนื้อหาของเว็บเพจนั้น เนื่องจากจำนวนเว็บไซต์ที่จดทะเบียนในปัจจุบันมีปริมาณมากกว่าหนึ่งพันล้านเว็บ การใช้งานอินเทอร์เน็ตในแต่ละครั้งมีการร้องขอข้อมูลยูอาร์แอลมากมาย ไม่ว่าจะเป็นข้อมูลหน้าเว็บ โฆษณาส่วนต่อประสานโปรแกรมประยุกต์ รูปภาพ วิดีโอ หรืออื่นๆ การจัดประเภทเว็บเพจจากข้อมูลจริงจำเป็นต้องเข้าใจลักษณะของข้อมูลก่อน ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ยูอาร์แอล (Uniform Resource Locator: URL)

ยูอาร์แอลที่เป็นข้อมูลนำเข้านั้นหมายถึงที่อยู่ที่ใช้บ่งบอกถึงเว็บเพจที่ต้องการเข้าถึง โดยยูอาร์แอลแบบสัมบูรณ์หรือยูอาร์แอลที่มีองค์ประกอบครบถ้วนประกอบไปด้วย

- โปรโทคอล (Protocol) ที่ใช้ในการเข้าถึงข้อมูล เช่น เอชทีทีพี (Hyper Text Transfer Protocol: HTTP) หรือเอชทีทีพีเอส (Hyper Text Transfer Protocol over SSL: HTTPS)
- ชื่อโดเมน (Domain Name) เป็นส่วนที่ใช้บ่งบอกถึงชื่อของเว็บไซต์นั้น
- พอร์ต (Port) เป็นหมายเลขที่บ่งบอกถึงช่องทางบริการที่เครื่องแม่ข่าย (Server) เปิดใช้งานอยู่
- ที่อยู่แฟ้มข้อมูล (Path) เป็นตำแหน่งที่อยู่ของแฟ้มข้อมูลที่ต้องการเข้าถึง
- ชื่อแฟ้มข้อมูล (Filename) เป็นชื่อของเอกสารที่ต้องการเข้าถึง
- สกุลแฟ้มข้อมูล (Filename Extension) เป็นชื่อสกุลของเอกสารที่ต้องการเข้าถึง
- คำร้องขอ (Query) เป็นข้อความที่ระบุถึงสิ่งที่ต้องการร้องขอจากเครื่องแม่ข่าย

ภาพที่ 3 แสดงองค์ประกอบของยูอาร์แอล

2.1.2 ป้ายระบุเอชทีเอ็มแอล (Hypertext Markup Language Tag)

ป้ายระบุเอชทีเอ็มแอลเป็นลักษณะเฉพาะของโครงสร้างภาษาเอชทีเอ็มแอลที่ใช้ในการระบุคำสั่งลงในเครื่องหมายน้อยกว่า (<) และเครื่องหมายมากกว่า (>) เพื่อให้ทราบว่า เป็นคำสั่งให้ทำงานตามที่ได้ระบุไว้ โดยป้ายระบุเอชทีเอ็มแอลสามารถแบ่งได้ 2 แบบ คือ

- ป้ายระบุเอชทีเอ็มแอลแบบเดี่ยว

ป้ายระบุเอชทีเอ็มแอลแบบเดี่ยวเป็นป้ายระบุเอชทีเอ็มแอลที่ไม่ต้องการป้ายระบุเอชทีเอ็มแอลปิด เช่น <b>, <i>,
 เป็นต้น

- ป้ายระบุเอชทีเอ็มแอลแบบเปิดปิด

ป้ายระบุเอชทีเอ็มแอลแบบเปิดปิดเป็นป้ายระบุเอชทีเอ็มแอลที่ประกอบด้วยป้ายระบุเอชทีเอ็มแอลแบบเปิดและป้ายระบุเอชทีเอ็มแอลแบบปิด โดยป้ายระบุเอชทีเอ็มแอลแบบปิดจะมีเครื่องหมายทับนำหน้าคำสั่งในป้ายระบุเอชทีเอ็มแอล เช่น <html>...</html>, <body>...</body> เป็นต้น โดยคำสั่งที่อยู่ภายในป้ายระบุเอชทีเอ็มแอลแบบเปิดและคำสั่งที่อยู่ภายในป้ายระบุเอชทีเอ็มแอลแบบปิดต้องเป็นคำสั่งเดียวกัน

```
▼<li class="submenu-room-item">
  ▼<a href="/forum/food" title="ร้านอาหาร สูตรอาหาร อาหารคาว อาหารหวาน เบเกอรี่ ไอศกรี
    ▶<em class="iconwrap">…</em>
     <span class="title">กินครัว</span>
     <br>
     <span class="desc">ร้านอาหาร สูตรอาหาร อาหารคาว อาหารหวาน เบเกอรี่ ไอศกรีม</span>
    </a>
  </li>
```

ภาพที่ 4 แสดงตัวอย่างป้ายระบุเอชทีเอ็มแอล

จากภาพที่ 4 แสดงตัวอย่างจริงของการใช้งานป้ายระบุแอชทีเอ็มแอล โดยมีป้ายระบุ “
...” “<a>...” “...” และ “...” เป็นป้ายระบุเอชทีเอ็มแอลแบบเปิดปิด และป้ายระบุ “
” เป็นป้ายระบุเอชทีเอ็มแอลแบบเดี่ยว

2.1.3 รหัสสถานะเอชทีทีพี (HTTP Status Code)

รหัสสถานะเอชทีทีพี คือ ตัวเลขมาตรฐานที่ได้รับมาจากการตอบกลับของเว็บไซต์ที่ทำงานอยู่บนเครื่องแม่ข่ายต่าง ๆ ที่ใช้บ่งบอกถึงสิ่งที่เกิดขึ้นกับทางเว็บไซต์เมื่อได้รับการร้องขอจากผู้ร้องขอ เราสามารถแปลรหัสสถานะเอชทีทีพีให้อยู่ในรูปแบบที่มนุษย์สามารถเข้าใจได้ โดยแบ่งออกเป็น 5 กลุ่มใหญ่ๆ ดังนี้

- รหัส 1XX คือ รหัสแสดงข้อมูลสถานะทั่วไป
- รหัส 2XX คือ การร้องขอสำเร็จ

- รหัส 3XX คือ การเปลี่ยนแปลงเส้นทาง
- รหัส 4XX คือ ความผิดพลาดที่เกิดจากเครื่องลูกข่าย
- รหัส 5XX คือ ความผิดพลาดที่เกิดจากเครื่องแม่ข่าย

โดยในภาพที่ 5 แสดงรายละเอียดของรหัสสถานะเอชทีทีพีแบบแจกแจงความหมายของรหัสสถานะในแต่ละกลุ่ม โดยรหัสสถานะเอชทีทีพีที่พบเจอบ่อย ๆ เช่น รหัสสถานะเอชทีทีพี 200 ที่หมายถึงการร้องขอสำเร็จส่งผลให้ผู้ใช้งานสามารถเข้าใช้งานเว็บไซต์ที่ร้องขอได้ รหัสสถานะเอชทีทีพี 404 หมายถึง ไม่พบหน้าเว็บที่ทำการร้องขอโดยปัญหานี้มีสาเหตุอยู่ที่เครื่องลูกข่าย หรือรหัสสถานะเอชทีทีพี 500 หมายถึง เครื่องแม่ข่ายที่ทำการร้องขอข้อมูลไปนั้นมีปัญหาอยู่ เป็นต้น

Cheatography

HTTP Status Codes Cheat Sheet
by kstep via cheatography.com/424/cs/199/

1xx: HTTP Informational Codes

100	Continue
101	Switching Protocols
102	Processing WebDAV
103	Checkpoint draft POST PUT
122	Request-URI too long IE7

2xx: HTTP Successful Codes

200	OK
201	Created
202	Accepted
203	Non-Authoritative Information 1.1
204	No Content
205	Reset Content
206	Partial Content
207	Multi-Status WebDAV 4918
208	Already Reported WebDAV 5842
226	IM Used 3229 GET

3xx: HTTP Redirection Codes

300	Multiple Choices
301	Moved Permanently
302	Found
303	See Other 1.1
304	Not Modified
305	Use Proxy 1.1
306	Switch Proxy unused
307	Temporary Redirect 1.1
308	Permanent Redirect 7538

307 and 308 are similar to 302 and 301, but the new request method after redirect must be the same, as on initial request.

4xx: HTTP Client Error Code

400	Bad Request
401	Unauthorized
402	Payment Required res
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
418	I'm a teapot 2324
422	Unprocessable Entity WebDAV 4918
423	Locked WebDAV 4918
424	Failed Dependency WebDAV 4918
425	Unordered Collection 3648
426	Upgrade Required 2817
428	Precondition Required draft
429	Too Many Requests draft
431	Request Header Fields Too Large draft
444	No Response nginx
449	Retry With MS
450	Blocked By Windows Parental Controls MS
451	Unavailable For Legal Reasons draft
499	Client Closed Request nginx

5xx: HTTP Server Error Codes

500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported
506	Variant Also Negotiates 2295
507	Insufficient Storage WebDAV 4918
508	Loop Detected WebDAV 5842
509	Bandwidth Limit Exceeded nostd
510	Not Extended 2774
511	Network Authentication Required draft
598	Network read timeout error nostd
599	Network connect timeout error nostd

HTTP Code Comments

WebDAV	WebDAV extension
1.1	HTTP/1.1
GET, POST, PUT, POST	For these methods only
IE	IE extension
MS	MS extension
nginx	nginx extension
2518, 2817, 2295, 2774, 3229, 4918, 5842	RFC number
draft	Proposed draft
nostd	Non standard extension
res	Reserved for future use
unused	No more in use, deprecated

Wikipedia was used to produce all HTTP codes content:
http://en.wikipedia.org/wiki/HTTP_status

By kstep
cheatography.com/kstep/

Published 24th January, 2012.
Last updated 1st July, 2016.
Page 1 of 1.

Sponsored by CrosswordCheats.com
Learn to solve cryptic crosswords!
http://crosswordcheats.com

ภาพที่ 5 แสดงรหัสสถานะเอชทีทีพี [8]

2.1.4 การตัดคำ (Word Segmentation)

การตัดคำโดยทั่วไปแล้วจำเป็นต้องมีความเข้าใจถึงโครงสร้างและหลักไวยกรณ์ของภาษาที่ต้องการตัดคำก่อน เนื่องจากการไม่เข้าใจลักษณะโครงสร้างของภาษาหรือไวยกรณ์ของภาษานั้นจะทำให้ไม่ทราบถึงวิธีการนำคำมาสร้างรูปประโยค โดยระบบจัดประเภทเว็บเพจในงานวิจัยชิ้นนี้จะ

พิจารณาเฉพาะเนื้อหาที่เป็นภาษาไทยและภาษาอังกฤษเท่านั้น การตัดคำของทั้งสองภาษามีวิธีการที่แตกต่างกันอย่างสิ้นเชิง สำหรับการตัดคำในข้อความที่เป็นภาษาอังกฤษนั้น เนื่องจากรูปแบบโครงสร้างการสร้างรูปประโยคของภาษาอังกฤษจะทำการนำคำมาต่อกันโดยมีช่องว่างคั่นกลางระหว่างแต่ละคำส่งผลให้การตัดคำในภาษาอังกฤษสามารถดำเนินการตัดคำได้โดยใช้ช่องว่างในการแบ่งประโยคออกเป็นคำ แต่สำหรับการตัดคำในภาษาไทยนั้นแตกต่างจากการตัดคำในภาษาอังกฤษ เพราะการเรียงรูปประโยคในภาษาไทยจะทำการนำคำแต่ละคำมาเรียงติดกันตามบริบทของคำจนเกิดเป็นรูปประโยคขึ้นส่งผลให้ไม่สามารถใช้ช่องว่างในการแบ่งประโยคออกเป็นคำได้เช่นเดียวกับการตัดคำในภาษาอังกฤษ โดยวิธีการตัดคำสามารถแบ่งออกได้เป็น 5 วิธี ดังนี้

- การตัดคำโดยใช้คำที่มีความยาวที่สุด (Longest Matching)

  วิธีการตัดคำโดยใช้คำที่มีความยาวที่สุดจะทำการตัดคำโดยไล่ดูจากตัวอักษรทางซ้ายสุดก่อนไล่ไปทางขวาทีละตัวอักษรจนกลายเป็นคำที่พบในพจนานุกรมโดยจะทำการเพิ่มตัวอักษรต่อไปเรื่อยๆ หากคำนั้นยังพบในพจนานุกรมอยู่ เช่น “ตากลม” จะได้เป็น “ตาก” และ “ลม” เนื่องจากคำว่า “ตาก” ยาวกว่าคำว่า “ตา” เป็นต้น

- การตัดคำโดยใช้คำที่มีความยาวน้อยที่สุด (Shortest Matching)

  วิธีการตัดคำโดยใช้คำที่มีความยาวน้อยที่สุดเป็นวิธีการตัดคำที่ตรงกันข้ามกับวิธีการตัดคำโดยใช้คำที่มีความยาวมากที่สุด โดยการตัดคำด้วยวิธีนี้จะไล่ดูตัวอักษรจากซ้ายไปขวาหากสามารถจัดเป็นคำและพบในพจนานุกรมแล้วก็จะเริ่มคำใหม่ทันที เช่น “ตากลม” จะได้เป็น “ตา” และ “กลม” โดยวิธีการตัดคำโดยใช้คำที่มีความยาวน้อยที่สุดนี้จะได้ผลลัพธ์ที่มีจำนวนคำมากที่สุด

- การตัดคำโดยอาศัยความสอดคล้องของคำมากที่สุด (Maximum Matching)

  วิธีการตัดคำโดยอาศัยความสอดคล้องของคำมากที่สุดจะทำการตัดคำที่เป็นไปได้ทุกรูปแบบออก จากนั้นจะทำการเลือกรูปแบบการตัดคำที่ให้ผลลัพธ์ออกมาเป็นจำนวนคำที่น้อยที่สุด หากผลลัพธ์ของการตัดคำออกมาด้วยจำนวนคำที่เท่ากันจะทำการใช้วิธีการตัดคำโดยใช้คำที่มีความยาวมากที่สุดเข้ามาช่วย เช่น “ตามหามเหสี” จะสามารถตัดคำออกได้ 2 รูปแบบ คือ “ตาม หา มเหสี” กับ “ตาม หาม เห สี” โดยวิธีการตัดคำโดยอาศัยความสอดคล้องของคำมากที่สุดจะเลือกผลลัพธ์ที่มีจำนวนคำน้อยที่สุดก็คือ “ตาม หา มเหสี”

- การตัดคำโดยใช้ความน่าจะเป็นทางสถิติ (Probabilistic Matching)

  วิธีการตัดคำโดยใช้ความน่าจะเป็นทางสถิติเป็นวิธีการตัดคำที่จะใช้ข้อมูลทางสถิติของการเกิดคำมาช่วยในการตัดคำ โดยการตัดคำแล้วดูค่าสถิติของการเกิดคำนั้น หากคำไหนมีค่าสถิติในการเกิดมากกว่าก็จะตัดคำออกมาในรูปแบบนั้น เช่น “ก็อด” จะสามารถตัดคำได้เป็น “ก็” กับ “อด” เนื่องจากทั้งสองคำนี้มีค่าสถิติของการเกิดมากกว่าคำว่า “ก็อด” นั่นเอง

- การตัดคำโดยใช้คุณลักษณะ (Rule-Based Matching or Feature-Based Matching)

วิธีการตัดคำโดยใช้คุณลักษณะของคำจะทำการตัดคำโดยพิจารณาจากบริบทและการเกิดร่วมกันของคำมาช่วยในการการตัดสินใจตัดคำ เช่น “ตากลม” หากพบคำว่า “โต” ในบริบทก็จะทำการตัดคำออกมาเป็นคำว่า “ตา” และคำว่า “กลม”

- การตัดคำโดยใช้วิธีย้อนรอย (Back Tracking)

วิธีการตัดคำโดยใช้วิธีย้อนรอยนั้นเป็นวิธีการที่ใช้เมื่อเกิดการตัดคำแล้วประโยคส่วนที่เหลือนั้นไม่สามารถตัดคำได้ เนื่องจากไม่พบคำจากอักขระถัดไปในพจนานุกรม จะทำการย้อนกลับไปเลือกรูปแบบของการตัดคำของคำก่อนหน้าใหม่ เช่น “ตามาหา” หากใช้วิธีการตัดคำโดยใช้คำที่มีความยาวที่สุดจะได้คำแรกเป็น “ตาม” และเหลือ “ามหา” ซึ่งไม่สามารถตัดคำต่อได้ จึงทำการย้อนรอยกลับไปยังคำก่อนหน้าคือคำว่า “ตาม” สามารถตัดคำเป็นคำว่า “ตา” ได้ ส่วนที่เหลือก็จะเป็น “มาหา” ซึ่งสามารถทำการตัดคำต่อไปเป็น “มา” และ “หา”

จากวิธีการตัดคำต่างๆ ที่ได้กล่าวมาในข้างต้น ปัจจุบันก็ได้มีผู้พัฒนาเครื่องมือที่ใช้สำหรับตัดคำภาษาไทยเป็นจำนวนมาก เช่น

- PyThaiNLP [9]

ใช้วิธีการตัดคำโดยอาศัยความสอดคล้องของคำมากที่สุด

- Swath (Smart Word Analysis for Thai) [10]

ใช้วิธีการตัดคำโดยใช้คำที่มีความยาวที่สุดและอาศัยความสอดคล้องของคำมากที่สุด

- Inspica [11]

ใช้วิธีการตัดคำโดยใช้คุณลักษณะ

2.1.5 วิธีการถ่วงน้ำหนัก (Weighting Schemes)

วิธีการถ่วงน้ำหนักเป็นวิธีการที่ถูกนำมาใช้เพื่อเพิ่มความแม่นยำให้กับการวิเคราะห์ข้อมูล โดยการถ่วงน้ำหนักของคำที่พบด้วยวิธีการคำนวณที่แตกต่างกันไปในแต่ละวิธีนั้น ขึ้นอยู่กับลักษณะของข้อมูลและวัตถุประสงค์ในการดำเนินงาน สามารถแบ่งได้ 6 วิธี ดังนี้

- Boolean Weighting

เป็นวิธีการนับความถี่และถ่วงน้ำหนักแบบง่ายที่สุดคือการนับว่ามีหรือไม่มีคำนั้น

- Term Frequency (TF) Weighting

เป็นวิธีการนับความถี่และถ่วงน้ำหนักของคำตามความถี่ของคำ

- Term Frequency-Inverse Document Frequency (TFxIDF) Weighting

เป็นวิธีการนับความถี่ของคำโดยถ่วงน้ำหนักความสำคัญของคำนั้นในเอกสาร

- Term Frequency-Cosine (TFC) Weighting

เป็นวิธีการนับความถี่ของคำและถ่วงน้ำหนักที่คล้ายกับ TFxIDF แต่วิธีการนี้จะมีการคำนึงความยาวของเอกสารด้วย

- Logarithm Term-Cosine (LTC) Weighting

เป็นวิธีการนับความถี่ของคำและถ่วงน้ำหนักโดยใช้ลอการิทึมเพื่อลดผลกระทบของปริมาณคำที่มีความแตกต่างมาก

- Entropy Weighting

เป็นวิธีการนับความถี่ของคำและถ่วงน้ำหนักจากการกระจายตัวของคำในแต่ละเอกสาร ถ้าข้อมูลมีการกระจายตัวกันมากค่าเอนโทรปีก็จะมีค่าสูง ในทางตรงกันข้ามถ้าข้อมูลมีความคล้ายคลึงกันมากค่าเอนโทรปีก็จะมีค่าต่ำ

2.1.6 อัลกอริทึมในการวิเคราะห์ข้อความ (Text Analysis Algorithm)

ตลอดระยะเวลาที่ผ่านมา อัลกอริทึมสำหรับการจัดประเภทเว็บเพจได้ถูกพัฒนาขึ้นเป็นจำนวนมาก จากกระบวนการวิเคราะห์ข้อมูลทั้ง 3 กระบวนการที่กล่าวมาในข้างต้น กระบวนการเรียนรู้แบบมีผู้สอนเป็นกระบวนการเรียนรู้ที่ได้รับความนิยมในการนำมาใช้ในการวิเคราะห์ข้อความมากที่สุด โดยอัลกอริทึมที่เป็นที่นิยมนำมาใช้ในการจัดประเภทข้อความในกระบวนการเรียนรู้แบบมีผู้สอน มีชื่อว่า นาอีฟเบย์ (Naïve Bayes) และอัลกอริทึมเวิร์ดทูเวค

- นาอีฟเบย์ (Naïve Bayes) [12]

อัลกอริทึมนาอีฟเบย์เป็นหนึ่งในกระบวนการเรียนรู้แบบมีผู้สอน โดยมีพื้นฐานมาจากกฎของเบย์ที่กล่าวถึงความน่าจะเป็นที่จะเกิดเหตุการณ์ขึ้นเมื่อมีเงื่อนไขที่อาจจะส่งผลให้เกิดเหตุการณ์นั้นและสมมติฐานที่ทำให้เกิดเหตุการณ์แต่ละเหตุการณ์มีความเป็นอิสระต่อกัน โดยกระบวนการเรียนรู้แบบมีผู้สอนของนาอีฟเบย์เป็นกระบวนการเรียนรู้ที่นิยมใช้ในงานวิจัยส่วนมาก เนื่องจากเป็นกระบวนการเรียนรู้ที่ซับซ้อนไม่มากและทำงานได้อย่างมีประสิทธิภาพ โดยเฉพาะในงานวิจัยด้านการวิเคราะห์ข้อความ

- เวิร์ดทูเวค (Word2Vec) [13]

เวิร์ดทูเวคเป็นกระบวนการวิเคราะห์ใหม่ที่ได้รับความนิยมในการนำมาใช้วิเคราะห์ข้อความในปัจจุบัน เวิร์ดทูเวคได้ถูกพัฒนาโดยนักวิจัยจากบริษัทกูเกิ้ล โดยมีแนวคิดพื้นฐานมาจากการทำโครงข่ายประสาทเทียม (Artificial Neural Network) แต่ทำการลดจำนวนชั้นการประมวลผลเหลือเพียง 2 ชั้นเพื่อลดเวลาในการประมวลให้เป็นแบบเส้นตรงแทน ประกอบด้วย 2 โมเดล คือ ถุงของคำที่ต้องเนื่องกัน (Continuous Bag of Word: CBOW) เหมาะสำหรับการจัดการกับประโยค และ Skip-gram Model เหมาะสำหรับการจัดการกับ

คำ แต่โมเดลแบบ CBOW จะเรียนรู้ได้เร็วกว่าแบบ Skip-gram Model หลักการทำงานของเวิร์ดทูเวคคือการเปลี่ยนข้อความให้กลายเป็นเวกเตอร์ แล้วทำการเปรียบเทียบหาความคล้ายระหว่างเวกเตอร์แทน

2.2 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่ผ่านมาได้มีการวิจัยเกี่ยวกับการจัดประเภทเว็บเพจไว้มากมาย โดยสามารถจัดกลุ่มด้วยกระบวนการเรียนรู้ที่ใช้ในการจัดประเภทเว็บเพจออกได้เป็น 3 กลุ่ม คือ กลุ่มที่ใช้กระบวนการเรียนรู้แบบมีผู้สอนในการวิเคราะห์ กลุ่มที่ใช้กระบวนการเรียนรู้แบบไม่มีผู้สอนในการวิเคราะห์ และกลุ่มที่ใช้กระบวนการเรียนรู้ทางสถิติหรือความหมายของคำ (Semantic Algorithm) เป็นเครื่องมือในการวิเคราะห์

ตัวอย่างงานวิจัยในส่วนของกระบวนการเรียนรู้แบบมีผู้สอน [14] ได้เสนออัลกอริทึมที่ใช้เพียงแค่ URL มาใช้ในการจัดประเภทเว็บเพจด้วยการรับข้อมูลนำเข้ามาเป็น URL จากนั้นทำการแยกคำใน URL ออกจากกันด้วยเครื่องหมายจุลภาค “.” และทับ “/” แล้วมาผ่านวิธี n-gram โดยได้ทำการเปรียบเทียบกับอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และพจนานุกรมที่ถูกกำหนดไว้ก่อนแล้ว ซึ่งได้วัดประสิทธิภาพกับชุดข้อมูลจาก Open Directory Project (ODP) และชุดข้อมูลจาก 4 มหาวิทยาลัย โดยได้ความแม่นยำประมาณร้อยละ 85 โดยข้อจำกัดของงานนี้คือ ยูอาร์แอลที่นำมาใช้ต้องเป็นยูอาร์แอลที่ใช้คำที่มีความหมายมาตั้ง หากใช้คำที่ไม่มีในพจนานุกรมก็จะไม่สามารถจัดประเภทได้ เช่น “pantip.com” ไม่สามารถจัดประเภทได้เพราะคำว่า “pantip” ไม่อยู่ในพจนานุกรม

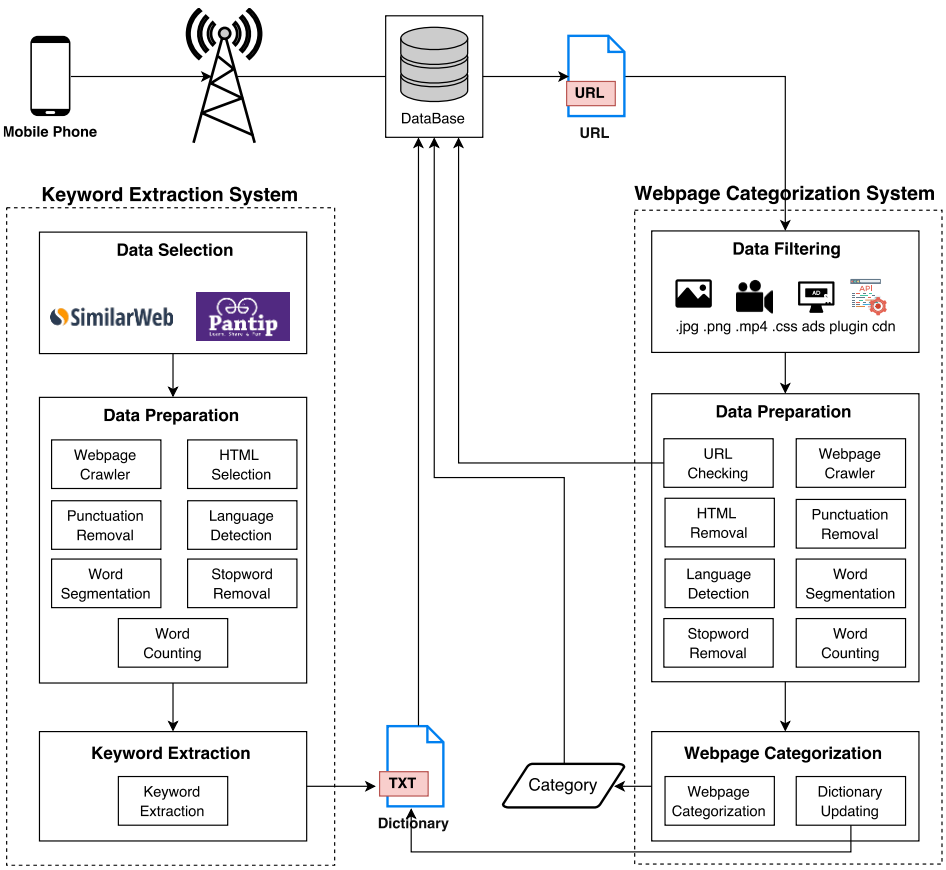
งานวิจัยที่ใช้กระบวนการเรียนรู้แบบไม่มีผู้สอน [15] ได้ทำการพัฒนาอัลกอริทึมเบย์เซียน (Bayesian Algorithm) และการตัดสินใจแบบต้นไม้ (Decision Tree) สำหรับการจัดประเภทข้อความ โดยการเปรียบเทียบประสิทธิภาพของอัลกอริทึมทั้งสองด้วยชุดข้อมูลทดสอบของรอยเตอร์ส (Reuters) และมักค์สาม (MUC-3) จากผลการทดลองของงานวิจัยนี้เขาพบว่าชุดข้อมูลทดสอบของมักค์สามมีความซับซ้อนมากกว่าชุดข้อมูลทดสอบของรอยเตอร์สจึงส่งผลให้ความแม่นยำในการทดสอบกับชุดข้อมูลมักค์สามมีค่าน้อยว่าชุดข้อมูลทดสอบรอยเตอร์ส โดยงานวิจัยนี้ให้ข้อสรุปว่าอัลกอริทึมเบย์เซียนจะทำงานได้ดีกับปริมาณฟีเจอร์น้อย ๆ ในขณะที่อัลกอริทึมการตัดสินใจแบบต้นไม้จะเหมาะกับข้อมูลที่มีปริมาณฟีเจอร์เยอะ

สำหรับงานวิจัยที่ใช้กระบวนการเรียนรู้ทางสถิติหรืออัลกอริทึมที่พิจารณาความหมายของคำ [16] ได้เสนออัลกอริทึมที่ใช้สำหรับพิจารณาคำที่มีความกำกวม (Word Sense Disambiguation Algorithm : WSD) โดยมีเป้าหมายในการจัดประเภทและคัดกรองเว็บไซต์ที่ไม่เหมาะสมด้วยการผสมผสานจุดเด่นของ YAGO2s และ DS-Onto เข้าด้วยกัน โดยพิจารณาเนื้อหาในเว็บไซต์ โดยการ

ตรวจสอบกับฐานข้อมูล DS-Onto ก่อน หากไม่สามารถจัดประเภทได้หรือจัดได้มากกว่าหนึ่งประเภท จะใช้ฐานข้อมูล YAGO2s ร่วมกับอัลกอริทึม WSD เพื่อจัดประเภทให้ได้เพียงแค่ 1 หมวดหมู่เท่านั้น โดยผลการทดลองพบว่า ความแม่นยำเมื่อใช้ฐานข้อมูล DS-Onto และ YAGO2s ร่วมกับ WSD ได้ความแม่นยำสูงถึงร้อยละ 93 หากไม่ใช้อัลกอริทึม WSD ร่วมด้วยจะมีความแม่นยำเหลือเพียงร้อยละ 71

บทที่ 3
ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

3.1 ภาพรวมของระบบ



ภาพที่ 6 แสดงสถาปัตยกรรมของระบบ

ในหัวข้อที่ 3.1 จะอธิบายขั้นตอนการทำงานและสถาปัตยกรรมของระบบที่ถูกออกแบบมาเพื่อสนับสนุนขั้นตอนการทำงานทั้งหมด ในภาพที่ 6 แสดงสถาปัตยกรรมของระบบ โดยระบบสามารถแบ่งออกเป็น 2 ส่วน คือ ระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ ในส่วนของระบบสกัดคำสำคัญเป็นระบบที่ใช้สำหรับการจัดทำพจนานุกรมเพื่อใช้ในการจัดประเภทเว็บเพจ ข้อมูลพจนานุกรมจะถูกเก็บเป็นเอกสารและมีการสำรองข้อมูลไว้ในฐานข้อมูลด้วย โดยข้อมูลการใช้การอินเทอร์เน็ตจากอุปกรณ์สื่อสารแบบพกพาที่สื่อสารผ่านเสาสัญญาณก็จะถูกเก็บอยู่ในฐานข้อมูลตาม พรบ. คอมพิวเตอร์ ซึ่งข้อมูลยูอาร์แอลนี้จะถูกนำมาวิเคราะห์เพื่อจัดประเภทเว็บเพจในระบบจัดประเภทเว็บเพจต่อไป

3.1.1 ขั้นตอนการทำงานระบบ

ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำสามารถแบ่งออกได้เป็น 2 ระบบย่อย คือ ระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ ระบบทั้งสองมีโมดูลที่จำเป็นในการดำเนินงานรวมกันทั้งสิ้น 14 โมดูล โดยโมดูลที่ทำงานเหมือนกันทั้งระบบจัดประเภทเว็บเพจและระบบสกัดคำสำคัญจะถูกนับเป็น 1 โมดูล ซึ่งโมดูลแต่ละโมดูลมีรายละเอียดดังนี้

1) โมดูลเลือกข้อมูล (Data Selection Module)

โมดูลเลือกข้อมูลเป็นโมดูลที่ถูกใช้ในระบบสกัดคำสำคัญ โมดูลนี้ทำหน้าที่ในการเลือกข้อมูลที่จะนำมาใช้ในการสร้างพจนานุกรม จากสมมติฐานที่ว่า เนื้อหาภายในเว็บเพจแต่ละเว็บต้องมีคำสำคัญที่สามารถบ่งบอกถึงหมวดหมู่ของเว็บเพจเว็บนั้นได้ ดังนั้นการนำเว็บเพจที่ถูกระบุหมวดหมู่แล้วมาทำการสกัดคำสำคัญย่อมต้องได้คำสำคัญที่มักจะถูกนำมาใช้บ่อย ๆ เมื่อมีการกล่าวถึงหมวดหมู่นั้น ๆ โดยระบบจัดประเภทเว็บเพจจะทำการพิจารณาเว็บเพจที่มีเนื้อหาเป็นภาษาไทยและภาษาอังกฤษ ดังนั้นข้อมูลที่จะถูกนำมาใช้ในการสร้างพจนานุกรมจึงจำเป็นต้องมีทั้ง 2 ภาษา คือ ภาษาไทยและภาษาอังกฤษ โดยข้อมูลที่จะนำมาใช้ในการสร้างพจนานุกรมสำหรับภาษาอังกฤษมาจากเว็บไซต์ยอดนิยม 50 อันดับแรกของคนทั่วโลกที่ถูกจัดอันดับโดย SimilarWeb เฉพาะเว็บไซต์ที่มีเนื้อหาเป็นภาษาอังกฤษเท่านั้น สำหรับข้อมูลที่จะถูกนำมาใช้ในการสร้างพจนานุกรมสำหรับภาษาไทยเป็นข้อมูลที่มาจากเว็บบอร์ดยอดนิยมของคนไทยที่มีชื่อว่า พันทิป (Pantip.com)

2) โมดูลคัดกรองยูอาร์แอล (Data Filtering Module)

โมดูลคัดกรองยูอาร์แอลเป็นโมดูลที่ถูกใช้ในระบบจัดประเภทเว็บเพจ โมดูลนี้ทำหน้าที่คัดกรองยูอาร์แอลที่ไม่สามารถนำใช้ในการจัดหมวดหมู่ได้ เนื่องจากยูอาร์แอลแบบสัมบูรณ์นั้นเป็นที่อยู่ที่ระบุถึงสิ่งที่ต้องการร้องขอจากเครื่องแม่ข่าย โดยสิ่งที่ร้องขอเหล่านั้นสามารถเป็นได้หลายประเภท เช่น แฟ้มข้อมูล รูปภาพ วิดีโอ จาวาสคริปต์ (JavaScript) สไตล์ชีต (Cascading Style Sheet) หรือรูปแบบตัวอักษร (Font) เป็นต้น โดยปกติแล้วการเข้าใช้งานเว็บหนึ่งเว็บนั้นมีการร้องขอข้อมูลไปยังเครื่องแม่ข่ายมากกว่าหนึ่งครั้ง การร้องขอนั้นประกอบไปด้วยหน้าเว็บเพจ จาวาสคริปต์ และสไตล์ชีต ในการร้องขอแต่ละครั้งนั้น สิ่งที่ระบบจะนำมาใช้ในการวิเคราะห์มีเพียงแค่หน้าเว็บเพจเท่านั้น เนื่องจากในระบบจัดประเภทเว็บเพจทำการวิเคราะห์เว็บเพจจากข้อความภายในเว็บเพจ ดังนั้นเอกสารอื่น ๆ ที่ไม่ใช่เอกสารที่สามารถเข้าถึงเนื้อหาที่เป็นข้อความจะไม่ถูกนำมาพิจารณาในงานวิจัยชิ้นนี้ เอกสารเหล่านั้นสามารถตรวจสอบได้โดยการดูชื่อสกุลของแฟ้มข้อมูลจากที่อยู่แฟ้มข้อมูลในยูอาร์แอลแบบสัมบูรณ์ โดยในโมดูลคัดกรองยูอาร์แอลนั้นระบบจะทำการกรองยูอาร์แอลสัมบูรณ์ที่ไม่เป็นไปตามเงื่อนไขออก ระบบจะไม่ต้องเสียเวลาในการประมวลผลข้อมูลที่ไม่จำเป็นและสามารถเพิ่มประสิทธิภาพในการทำงานโดยรวมได้ โดยระบบสามารถเก็บรายการยูอาร์แอลเหล่านี้ไว้เพื่อใช้กรองยูอาร์แอลก่อนที่จะเก็บลงฐานข้อมูลได้ด้วย

3) โมดูลตรวจสอบยูอาร์แอล (URL Checking Module)

โมดูลตรวจสอบยูอาร์แอลเป็นโมดูลที่ถูกใช้ในระบบจัดประเภทเว็บเพจ โมดูลนี้จะทำหน้าที่ตรวจสอบยูอาร์แอลที่เข้ามาในระบบหลังจากผ่านการกรองแล้วว่าเคยจัดประเภทไปแล้วหรือยังไม่เคยได้รับการจัดประเภท หากยูอาร์แอลที่เข้ามาเคยจัดประเภทไปแล้ว ระบบจะทำการตรวจสอบวันที่อัพเดตล่าสุด หากระยะเวลาการอัพเดตล่าสุดเกินระยะเวลาในการแคชชิ่ง (Caching) ระบบจะทำการจัดประเภทให้กับยูอาร์แอลนั้นใหม่ เนื่องจากข้อมูลในหน้าเว็บเพจอาจจะมีการเปลี่ยนแปลงไปแล้ว สำหรับยูอาร์แอลที่เคยได้รับการจัดประเภทแล้วและมีระยะเวลาในการอัพเดตน้อยกว่าระยะเวลาในการแคชชิ่ง ระบบจะทำการคืนค่าหมวดหมู่ของยูอาร์แอลนั้นที่เคยได้รับการจัดประเภทไว้แล้วกลับไป

4) โมดูลร้องขอข้อมูลเว็บเพจ (Webpage Crawler Module)

โมดูลร้องขอข้อมูลเว็บเพจเป็นโมดูลที่ถูกใช้ทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้จะทำหน้าที่ดึงข้อมูลหน้าเว็บเพจ โดยการร้องขอหน้าเว็บเพจจากยูอาร์แอลที่ผ่านการคัดกรองและตรวจสอบเป็นที่เรียบร้อยแล้ว ก่อนเริ่มการร้องขอข้อมูลหน้าเว็บระบบจะทำการตรวจสอบรูปแบบยูอาร์แอลอีกครั้งด้วยการตรวจสอบส่วนโปรโทคอลของยูอาร์แอลว่ามีการระบุโปรแกรมเป็นเอชทีทีพีหรือไม่ หากไม่ได้มีการระบุระบบจะทำการเติมโปรโทคอลเอชทีทีพีให้ การดึงข้อมูลหน้าเว็บเพจนั้นเราสนใจเฉพาะเว็บเพจที่สามารถเข้าใช้งานได้หมายถึงเว็บเพจที่เครื่องแม่ข่ายทำการตอบกลับหน้าเว็บเพจมาให้ การตรวจสอบว่าเว็บเพจที่ร้องขอไปนั้นสามารถใช้งานได้นั้นสามารถดูได้จากรหัสสถานะเอชทีทีพี (HTTP Status Code) ของการร้องขอหน้าเว็บเพจ โดยรหัสสถานะเอชทีทีพีจะบอกถึงสถานะของการร้องขอข้อมูลหน้าเว็บเพจ หากรหัสสถานะเอชทีทีพีมีค่าเป็น 200 จะถือว่าข้อมูลเว็บเพจที่ได้จากการร้องขอยูอาร์แอลนั้นสามารถนำไปใช้งานได้ ในทางกลับกัน หากรหัสสถานะเอชทีทีพีไม่ได้มีค่าเป็น 200 จะถือว่า ยูอาร์แอล ดังกล่าวไม่สามารถใช้งานได้

การดึงข้อมูลหน้าเว็บเพจโดยทั่วไปแล้วจะมีบางเว็บเพจที่เครื่องแม่ข่ายไม่ดำเนินการตอบรหัสสถานะเอชทีทีพีกลับมา ซึ่งส่งผลให้ต้องเสียระยะเวลาในการรอคอยการตอบกลับโดยเปล่าประโยชน์ ดังนั้นการดึงข้อมูลหน้าเว็บเพจจึงจำเป็นต้องมีการระบุขอบเขตระยะเวลาที่สามารถรอการดำเนินการตอบกลับของเครื่องแม่ข่ายไว้ด้วย ทางผู้วิจัยได้เลือกระยะเวลาในการรอการดำเนินการตอบกลับจากเครื่องแม่ข่ายเป็นระยะเวลา 2 วินาที เนื่องจากการดึงข้อมูลหน้าเว็บเพจจากเครื่องแม่ข่ายโดยปกติจะใช้เวลาไม่ถึง 1 วินาที แต่การจราจรทางเครือข่ายอาจจะมีความคับคั่งอยู่บ้างในชั่วขณะนั้น ซึ่งส่งผลให้เกิดการหน่วงของเวลาขึ้น ทางผู้วิจัยเล็งเห็นว่าควรทำการเพิ่มระยะเวลาในการรอเป็น 2 วินาที เพื่อรองรับปัญหาที่ได้กล่าวมาในข้างต้น

สำหรับยูอาร์แอลที่ทำการร้องขอข้อมูลไปยังเว็บเพจที่มีการเข้ารหัสข้อมูลภายในเว็บเพจ ทางผู้วิจัยจะไม่พิจารณายูอาร์แอลนั้น โดยโมดูลร้องขอข้อมูลเว็บเพจจะได้ข้อมูลหน้าเว็บเพจที่อยู่ใน

รูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสาร (Document Object Model) ซึ่งเป็นโมเดลโครงสร้างข้อมูลของเว็บสำหรับใช้ในเว็บบราวเซอร์ (Web Browser)

5) โมดูลกำจัดป้ายระบุเอชทีเอ็มแอล (HTML Tag Removal Module)

โมดูลกำจัดป้ายระบุเอชทีเอ็มแอลเป็นโมดูลที่ถูกใช้ในระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ในการลบป้ายระบุเอชทีเอ็มที่ไม่จำเป็นออก โดยข้อมูลนำเข้าของโมดูลนี้คือข้อความที่อยู่ในรูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสาร ซึ่งแบบจำลองนี้ประกอบไปด้วยส่วนประกอบ 4 อย่าง คือ ป้ายระบุเอชทีเอ็มแอล จาวาสคริปต์ สไตล์ชีต และเนื้อหาข้อความภายในเว็บ

- ป้ายระบุเอชทีเอ็มแอล

ป้ายระบุเอชทีเอ็มแอล คือ ป้ายระบุที่ใช้กำหนดโครงสร้างของเว็บเพจ โดยประกอบไปด้วยเครื่องหมายน้อยกว่า เครื่องหมายมากกว่า และคำสั่งในภาษาเอชทีเอ็มแอล

- จาวาสคริปต์ (JavaScript)

เป็นภาษาคอมพิวเตอร์ภาษาหนึ่งที่ใช้ในการสร้างและพัฒนาเว็บไซต์ร่วมกับภาษาเอชทีเอ็มแอล โดยการทำงานคือจะทำการแปลความคำสั่งและดำเนินการทำงานไปทีละคำสั่ง (Interpret)

- สไตล์ชีต (Cascading Style Sheet)

เป็นภาษาคอมพิวเตอร์ที่ใช้ในการจัดรูปแบบการแสดงผลของเว็บไซต์ให้ผู้ใช้บริการได้ใช้งาน โดยสิ่งที่สิ่งที่สไตล์ชีตทำได้ ได้แก่ การปรับเปลี่ยนสีสัน การเปลี่ยนรูปแบบตัวอักษร การจัดวางตำแหน่ง และการทำภาพเคลื่อนไหว เป็นต้น

- เนื้อหาภายในเว็บเพจ

เป็นส่วนเนื้อหาที่เป็นข้อความที่อยู่ภายในเว็บเพจ โดยเนื้อหาที่เป็นข้อความเหล่านี้มักจะอยู่ภายในป้ายระบุเอชทีเอ็มแอล

เนื่องจากระบบจัดประเภทเว็บเพจสนใจเพียงแค่ส่วนเนื้อหาภายในเว็บเพจเพียงอย่างเดียวและป้ายระบุเอชทีเอ็มแอลแต่มีข้อมูลบางส่วนในป้ายระบุเอชทีเอ็มแอลที่มีประโยชน์ในการนำมาเป็นส่วนหนึ่งในการวิเคราะห์ ได้แก่

- ป้ายระบุนิยามข้อมูล (metadata)

ป้ายระบุนิยามข้อมูลเป็นป้ายระบุที่ใช้แสดงถึงคำสำคัญที่อยู่ภายในเว็บเพจนั้น โดยนักพัฒนาเว็บจะเป็นผู้ใส่ข้อมูลคำสำคัญเหล่านี้ลงในส่วนเนื้อหาของป้ายระบุนิยามข้อมูลเพื่อผลประโยชน์ในการใช้จัดอันดับเว็บเพจจากโปรแกรมค้นหาต่าง ๆ ซึ่งป้ายระบุนิยามข้อมูลนี้มีวัตถุประสงค์เพื่อบ่งบอกถึงหมวดหมู่หรือประเภทของเนื้อหาภายในหน้าเว็บเพจนั้น แต่เนื่องจากโปรแกรมค้นหาได้ใช้ข้อมูลในส่วนของป้ายระบุนิยามข้อมูลนี้ในการจัดลำดับเว็บเพจที่จะถูกแสดงเมื่อผู้ใช้บริการทั่วไปทำการค้นหา ส่งผลให้เกิดกลุ่มนักพัฒนาที่ต้องการให้หน้า

เว็บเพจของตนเองถูกจัดอันดับขึ้นมาเป็นอันดับแรกเพื่อให้ผู้ใช้บริการเข้ามายังหน้าเว็บเพจของตนจึงทำการใส่ข้อมูลคำสำคัญมากเกินความจำเป็น หรือนักพัฒนาเว็บบางกลุ่มก็ไม่ได้ให้ความสำคัญกับการระบุคำสำคัญลงไปในป้ายระบุนิยามข้อมูล จากปัญหาต่าง ๆ ที่ได้กล่าวมาในข้างต้นเป็นเหตุให้ระบบจัดประเภทเว็บเพจไม่สามารถนำข้อมูลภายในป้ายระบุนิยามข้อมูลเพียงอย่างเดียวมาใช้ในการจัดประเภทเว็บเพจได้

- ป้ายระบุหัวเรื่อง (Title)

ป้ายระบุหัวเรื่องเป็นป้ายระบุที่ใช้ในการบอกหัวเรื่องของหน้าเว็บเพจ โดยป้ายระบุหัวเรื่องควรจะเป็นข้อมูลที่มีประโยชน์มากที่สุดในการนำมาวิเคราะห์เพื่อจัดหมวดหมู่ของเนื้อหา เนื่องจากหัวข้อเป็นสิ่งที่สามารถอธิบายเนื้อหาแบบกระชับภายในประโยคสั้น ๆ แต่ในปัจจุบันนั้นการตั้งหัวเรื่องไม่ใช่การพูดถึงใจความสำคัญของเนื้อหา การตั้งหัวเรื่องกลับกลายเป็นการตั้งชื่อหัวเรื่องที่สามารถดึงดูดผู้ใช้บริการให้เกิดความสนใจเข้าไปอ่านเนื้อหาที่อยู่ภายในเว็บเพจได้มากที่สุดแทน เพราะปริมาณเข้าใช้งานเว็บเพจนั้นส่งผลถึงรายได้ที่เจ้าของเว็บเพจนั้นจะได้รับจากค่าโฆษณา ระบบจัดประเภทเว็บเพจจึงไม่สามารถนำข้อมูลในป้ายระบุหัวเรื่องเพียงอย่างเดียวมาใช้ในการจัดประเภทเว็บเพจได้เช่นเดียวกัน

แม้ว่าทั้งป้ายระบุนิยามข้อมูลและป้ายระบุหัวเรื่องจะไม่สามารถนำใช้ในการจัดประเภทเว็บเพจสำหรับระบบจัดประเภทเว็บเพจแบบเดี่ยว ๆ ได้ แต่เนื่องจากยังมีบางเว็บเพจที่ข้อมูลภายในป้ายระบุนิยามข้อมูลและป้ายระบุหัวเรื่องยังสามารถใช้ประโยชน์ได้ ระบบจัดประเภทเว็บเพจจึงเลือกที่จะทำการเก็บข้อมูลภายในป้ายระบุนิยามข้อมูลและป้ายระบุหัวเรื่องมาใช้ในการพิจารณาร่วมกับข้อมูลอื่น ๆ ภายในเว็บเพจต่อไป โดยข้อมูลภายในป้ายระบุนิยามข้อมูลและป้ายระบุหัวเรื่องจะถูกเก็บอยู่ในรูปแบบของข้อความ โดยมีช่องว่างเป็นตัวคั่นระหว่างข้อมูลในป้ายระบุนิยามข้อมูล ป้ายระบุหัวเรื่องและเนื้อหาส่วนอื่น ๆ ของเว็บเพจ

6) โมดูลเลือกป้ายระบุเอชทีเอ็มแอล (HTML Tag Selection Module)

โมดูลเลือกป้ายระบุเอชทีเอ็มแอลเป็นโมดูลที่ถูกใช้ในระบบสกัดคำสำคัญเท่านั้น โมดูลนี้มีหน้าที่ในการเลือกเนื้อหาที่อยู่ภายในเว็บเพจที่เป็นข้อมูลที่จะนำไปใช้ในการสร้างพจนานุกรม โดยเว็บเพจที่จะถูกนำมาใช้ในการสร้างพจนานุกรมนั้นได้ถูกเลือกในโมดูลเลือกข้อมูลเป็นที่เรียบร้อยแล้ว ซึ่งเว็บเพจเหล่านั้นสามารถทราบลักษณะโครงสร้างของแบบจำลองโครงสร้างข้อมูลเอกสารได้ โดยการร้องขอข้อมูลเว็บเพจมาก่อน จากนั้นทำการดูชื่อของป้ายระบุเอชทีเอ็มแอลที่ใช้เก็บเนื้อหาภายในเว็บเพจที่เราต้องการ ก็จะทราบลักษณะโครงสร้างของแบบจำลองโครงสร้างเอกสารของเว็บเพจนั้นได้ เมื่อทราบลักษณะโครงสร้างของแบบจำลองโครงสร้างข้อมูลเอกสารเป็นที่เรียบร้อยแล้ว ระบบก็จะทำการสกัดเฉพาะเนื้อหาที่อยู่ภายในเว็บเพจนั้นออกมาตามชื่อของป้ายระบุเอชทีเอ็มแอลที่ได้จากการเรียนรู้ลักษณะโครงสร้างของแบบจำลองโครงสร้างข้อมูลเอกสาร

7) โมดูลกำจัดเครื่องหมายวรรคตอน (Punctuation Removal Module)

โมดูลกำจัดเครื่องหมายวรรคตอนเป็นโมดูลที่ถูกใช้ทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ในการแยกคำออกกันโดยใช้เครื่องหมายวรรคตอนเป็นตัวแบ่ง เนื่องจากเครื่องหมายวรรคตอนคือเครื่องหมายหรือสัญลักษณ์ที่ถูกนำมาใช้เพื่อประกอบการเขียนในแต่ละภาษา เครื่องหมายวรรคตอนมีวัตถุประสงค์ในการแบ่งวรรคตอนของคำหรือประโยคในการเขียนอยู่แล้ว และการวิเคราะห์ข้อมูลที่มีประสิทธิภาพนั้นข้อมูลต้องเป็นข้อมูลที่สะอาดปราศจากสิ่งที่ไม่จำเป็น จึงเป็นสาเหตุให้เกิดการทำความสะอาดข้อมูลก่อนนำไปวิเคราะห์ โดยการทำความสะอาดข้อมูลด้วยการกำจัดเครื่องหมายวรรคตอนหมายถึงการแทนที่เครื่องหมายวรรคตอนด้วยช่องว่าง (Space) เพื่อแยกข้อความที่มีเครื่องหมายวรรคตอนออกเป็นคำ จากการที่ระบบจัดประเภทเว็บเพจใช้คำสำคัญในเว็บเพจในการบ่งชี้ถึงหมวดหมู่ ดังนั้นจึงต้องมีการเปรียบเทียบคำสำคัญเหล่านั้นเนื่องจากการนำคำที่มีเครื่องหมายวรรคตอนผสมอยู่ไปเปรียบเทียบกับคำที่ไม่มีเครื่องหมายวรรคตอนผสมอยู่ หากเป็นการเปรียบเทียบแบบต้องมีความเหมือนกันทุกตัวอักษรเท่านั้นจะส่งผลให้คำทั้งสองคำนั้นที่มีเครื่องหมายวรรคตอนผสมอยู่และไม่มีเครื่องหมายวรรคตอนผสมอยู่หรือเป็นเครื่องหมายวรรคตอนเดียวกันจะถูกพิจารณาว่าเป็นคำคนละคำกันทันที เช่น “กิน.” กับ “กิน” เป็นต้น

8) โมดูลตรวจสอบภาษา (Language Detection Module)

โมดูลตรวจสอบภาษาเป็นโมดูลที่ถูกใช้ทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ในการตรวจสอบภาษาของเนื้อหาที่ร้องขอมาได้ว่าเป็นเนื้อหาภาษาไทยหรือภาษาอังกฤษเพื่อที่จะส่งต่อเนื้อหาเข้าไปสู่โมดูลตัดคำได้อย่างถูกต้องตามหลักการของแต่ละภาษา เนื่องจากภาษาแต่ละภาษามีโครงสร้างและหลักไวยากรณ์ที่แตกต่างกันไปตามแต่ละภาษา การพิจารณาภาษาของข้อมูลทางผู้วิจัยได้เลือกใช้ส่วนต่อประสานโปรแกรมประยุกต์ในการตรวจสอบภาษาของบริษัทกูเกิล ที่รองรับการตรวจสอบภาษามากกว่า 55 ภาษา แต่ระบบพิจารณาเพียงแค่ข้อความที่เป็นภาษาอังกฤษและภาษาไทยเท่านั้น เหตุผลที่ระบบเลือกรองรับภาษาอังกฤษและภาษาไทย เพราะภาษาอังกฤษถือว่าเป็นภาษาสากลที่ปัจจุบันทั่วโลกให้การยอมรับสำหรับใช้ในการสื่อสาร ดังนั้นเว็บไซต์ส่วนใหญ่ที่ต้องการให้บริการกับผู้ใช้บริการได้หลายเชื้อชาติจึงได้ทำเนื้อหาภายในเว็บไซต์ให้เป็นภาษาอังกฤษ และจากการสำรวจการเข้าใช้บริการเว็บไซต์ยอดนิยม ของผู้ใช้บริการในประเทศไทยจากผู้ให้บริการโทรคมนาคมรายหนึ่งในประเทศไทยพบว่า ผู้ใช้บริการมีการเข้าใช้บริการเว็บไซต์ที่มีเนื้อหาภาษาไทยไม่น้อยกว่าร้อยละ 30 ของการเข้าใช้บริการเว็บไซต์ทั้งหมด ดังนั้นการพิจารณาเว็บไซต์ที่มีเนื้อหาเป็นภาษาไทยจึงมีความจำเป็น หลังจากการตรวจสอบภาษาเสร็จสิ้นเนื้อหาจะถูกส่งไปตัดคำตามหลักภาษาในโมดูลตัดคำต่อไป โดยการตัดคำของภาษาอังกฤษที่ใช้เพียงแค่การแบ่งช่องว่างซึ่งต่างจากของภาษาไทยที่ไม่สามารถตัดคำโดยใช้เพียงแค่ช่องว่างได้เหมือนกับภาษาอังกฤษเพราะการตัดคำในภาษาไทยนั้นความซับซ้อนมากกว่าการตัดคำแบบ

ภาษาอังกฤษตรงที่ภาษาไทยนำคำแต่ละคำมาต่อติด ๆ จนกันเป็นประโยคทำให้ไม่สามารถตัดคำโดยใช้ช่องว่างได้ ดังนั้นการตรวจสอบภาษาถือว่าเป็นที่สำคัญสิ่งหนึ่ง หากตรวจภาษาผิดพลาดจะส่งผลให้การตัดคำผิดพลาดไปด้วยและจะส่งผลต่อไปยังกระบวนการวิเคราะห์ข้อมูลอีกด้วย

9) โมดูลตัดคำ (Word Segmentation Module)

โมดูลตัดคำเป็นโมดูลที่ถูกใช้งานทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ในการตัดคำเพื่อแยกคำออกจากกันในแต่ละประโยค เนื่องจากระบบพิจารณาทั้งเนื้อหาที่เป็นภาษาไทยและเนื้อหาที่เป็นภาษาอังกฤษ แต่การตัดคำในแต่ละภาษามีความแตกต่างกันไปตามแต่หลักภาษาหรือไวยากรณ์ของภาษานั้น ๆ การตัดคำสำหรับภาษาอังกฤษที่มีหลักการเขียนโดยการเขียนแยกคำด้วยช่องว่างอยู่แล้ว ระบบจึงสามารถตัดคำของภาษาอังกฤษโดยใช้ช่องว่างได้เช่นเดียวกัน สำหรับคำบางคำที่เกิดจากการประกอบคำโดยมีสัญลักษณ์เป็นตัวเชื่อม เมื่อผ่านโมดูลกำจัดเครื่องหมายวรรคตอนแล้วคำประกอบเหล่านั้นจะถูกแยกออกจากกันเป็นคำย่อย ๆ โดยอัตโนมัติ ส่วนการตัดคำสำหรับภาษาไทยเป็นสิ่งสำคัญและมีความยากลำบากในการทำเรื่องหนึ่ง เนื่องจากหลักการเขียนภาษาไทยจะทำการเขียนคำติด ๆ กันจนเป็นประโยคโดยไม่มีการเว้นช่องว่างเหมือนอย่างภาษาอังกฤษ ซึ่งส่งผลให้ไม่สามารถตัดคำโดยใช้ช่องว่างเพียงอย่างเดียวได้ ตลอดระยะเวลาที่ผ่านมาได้มีการพัฒนาวิธีการตัดคำภาษาไทยออกมามากมายเพื่อให้สามารถนำไปใช้ในการเตรียมข้อมูลสำหรับการดำเนินการวิเคราะห์ด้วยกระบวนการเรียนรู้ด้วยเครื่องจักร (Machine Learning) วิธีการตัดคำภาษาไทยสามารถแบ่งออกได้เป็นการตัดคำโดยใช้คำที่มีความยาวที่สุดก่อน (Longest Matching) การตัดคำโดยใช้คุณลักษณะ (Rule-based Matching) การตัดคำโดยอาศัยความสอดคล้องของคำมากที่สุด (Maximum Matching) การตัดคำโดยหาความน่าจะเป็นทางสถิติ (Probabilistic Matching) และการตัดคำโดยใช้วิธีย้อนรอย (Back Tracking) ซึ่งแต่ละวิธีก็มีข้อดีข้อเสียแตกต่างกันออกไป แต่การนำข้อดีของวิธีการหลาย ๆ อย่างมารวมกันเพื่อให้ได้วิธีการที่ดีที่สุดก็ย่อมมีสิ่งที่ต้องเสียไปเช่นกัน เช่น วิธีการตัดคำโดยใช้กฎเกณฑ์มีจุดเด่นในด้านของความเร็ว ส่วนวิธีการตัดคำโดยอาศัยพจนานุกรมมีจุดเด่นด้านความแม่นยำ เมื่อนำ 2 วิธีดังกล่าวมารวมกันจะส่งผลให้ได้วิธีการตัดคำภาษาไทยที่มีความแม่นยำสูงขึ้น แต่ก็แลกมาด้วยความเร็วในการตัดคำที่ลดลง การเลือกวิธีการตัดคำสำหรับภาษาไทยจึงขึ้นอยู่กับวัตถุประสงค์ของงานที่ทำ โดยในงานวิจัยชิ้นนี้ได้เลือกใช้ส่วนต่อประสานโปรแกรมประยุกต์ตัดคำภาษาไทยของ Inspica โดยส่วนต่อประสานโปรแกรมประยุกต์ตัดคำภาษาไทยของ Inspica ได้เลือกใช้วิธีการตัดคำโดยใช้กฎเกณฑ์พื้นฐานที่มีจุดเด่นเรื่องความเร็วร่วมกับการใช้พจนานุกรมมาสนับสนุนการตัดคำให้มีความแม่นยำเพิ่มมากขึ้น แต่แลกมาด้วยความเร็วในการตัดคำที่ลดลงเพราะว่าต้องดำเนินการตรวจสอบคำที่ได้จากการใช้กฎเกณฑ์กับพจนานุกรมอีกครั้ง เนื่องจากระบบที่พัฒนาขึ้นต้องการรองรับการจัดประเภทเว็บเพจจากการจราจรทางเครือข่ายของจริง ดังนั้นความเร็วในการตัดคำจึงเป็นสิ่งที่จำเป็นแต่ก็ต้องมีความ

แม่นยำด้วย ส่วนต่อประสานโปรแกรมประยุกต์ในการตัดคำภาษาไทยของ Inspica จึงได้รับการเลือกมาใช้ในงานวิจัยชิ้นนี้ เมื่อกระบวนการตัดคำของแต่ละภาษาดำเนินการเสร็จสิ้นแล้ว ข้อมูลจะอยู่ในรูปแบบของรายการของคำเพื่อนำไปดำเนินการทำความสะอาดต่อในโมดูลถัดไป

10) โมดูลกำจัดคำทั่วไป (Common Removal Module)

โมดูลกำจัดคำทั่วไปเป็นโมดูลที่ถูกใช้งานทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ลบคำทั่วไปออกจากข้อความ โดยคำทั่วไปหมายถึงคำที่มักจะปรากฏอยู่ในทุก ๆ เอกสาร ซึ่งส่งผลให้มีความสำคัญน้อยเมื่อมีการนำคำทั่วไปเหล่านี้ไปวิเคราะห์ การสร้างรายการของคำทั่วไปนั้นสามารถดำเนินการได้โดยการนับความถี่ของคำที่เกิดขึ้น และดูความถี่ของคำนั้นที่ไปปรากฏบนเอกสารต่าง ๆ หากมีความถี่ในการไปปรากฏบนเอกสารต่าง ๆ ในเอกสารมากกว่าค่าที่ระบุไว้จะถือว่าเป็นคำทั่วไป โดยทั่วไปแล้วคำทั่วไปที่นิยมลบก่อนนำข้อมูลไปวิเคราะห์ คือ คำจำพวกคำสรรพนาม คำสันธาน คำบุพบท คำคุณลักษณะ หรือคำวิเศษณ์ เป็นต้น การทำความสะอาดโดยการกำจัดคำทั่วไปออกจากข้อมูลก่อนจะทำให้ข้อมูลที่จะนำไปวิเคราะห์มีความสอดคล้องกันและผลลัพธ์ที่ได้จากการวิเคราะห์มีความถูกต้องมากยิ่งขึ้น แต่ก่อนการทำความสะอาดคำทั่วไปนั้นสำหรับภาษาอังกฤษจำเป็นที่จะต้องทำการลดรูปตัวอักษรให้มีลักษณะเหมือนกันก่อน โดยการเปลี่ยนตัวอักษรทุกตัวให้เป็นตัวพิมพ์เล็ก เพื่อให้การตรวจสอบคำทั่วไปทำงานได้อย่างเต็มประสิทธิภาพ

11) โมดูลนับคำ (Word Counting Module)

โมดูลนับความถี่ของคำเป็นโมดูลที่ถูกใช้งานทั้งในระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โมดูลนี้มีหน้าที่ในการแปลงข้อมูลให้อยู่ในรูปแบบที่พร้อมสำหรับการนำข้อมูลไปวิเคราะห์ต่อ หลังจากกระบวนการทำความสะอาดข้อมูลเสร็จสิ้น ระบบจะทำการนับความถี่ของคำเพื่อนำไปใช้ในการจัดทำพจนานุกรมในระบบสกัดคำสำคัญและจัดประเภทเว็บเพจในระบบจัดประเภทเว็บเพจ การนับความถี่ของคำโดยทั่วไปจะมีการทำงานร่วมกับการถ่วงน้ำหนักของคำ เพื่อให้ข้อมูลที่มีความสำคัญได้มีน้ำหนักที่มากกว่าข้อมูลที่ไม่สำคัญ วิธีการถ่วงน้ำหนักได้ถูกคิดค้นมากมายตลอดระยะเวลาที่ผ่านมา โดยผู้วิจัยได้เลือกใช้วิธีถ่วงน้ำหนักแบบปกติ เนื่องจากผู้วิจัยเล็งเห็นว่า ข้อมูลในหน้าเว็บเพจทุกส่วนมีความสำคัญใกล้เคียงกัน การถ่วงน้ำหนักให้กับส่วนใดส่วนหนึ่งมากเกินไปอาจจะนำไปสู่ความผิดพลาดในการวิเคราะห์ได้ในอนาคต เมื่อระบบทำการนับความถี่ของคำเสร็จสิ้นแล้วจะทำการเรียงลำดับความถี่ของคำจากมากไปน้อย โดยผลลัพธ์หลังจากการนับความถี่ของคำเสร็จสิ้นจะอยู่ในรูปแบบของรายการของคู่ของคำและความถี่ที่ถูกเรียงลำดับจากมากไปน้อยเรียบร้อยแล้ว

12) โมดูลสกัดคำสำคัญ (Keyword Extraction Module)

โมดูลสกัดคำสำคัญเป็นโมดูลที่ถูกใช้งานในระบบสกัดคำสำคัญ โมดูลนี้มีหน้าที่สกัดคำสำคัญออกมาสร้างเป็นพจนานุกรมไว้ใช้ในการจัดประเภทเว็บเพจในระบบจัดประเภทเว็บเพจ โมดูลสกัดคำสำคัญจะรับรายการของคำและความถี่ของคำนั้นจากทุก ๆ เอกสารมาทำการสร้างเป็นพจนานุกรม

ด้วยการรวมรายการของคำและความถี่ที่อยู่ในหมวดหมู่เดียวกันเป็นรายการเดียว แล้วทำการเรียงลำดับความถี่ของรายการของคำและความถี่ในทุก ๆ หมวดหมู่ จากนั้นจะได้กลุ่มของรายการของคำและความถี่ของหมวดหมู่ทั้งหมดมาใช้ในการสร้างพจนานุกรม โดยการใช้งานระบบจะทำการแปลงข้อมูลกลุ่มของรายการของคำและความถี่ของหมวดหมู่ให้อยู่ในรูปแบบของตัวแปรพจนานุกรม (Dictionary) โดยมีกุญแจหลัก (Key) เป็นคำศัพท์และค่า (Value) เป็นรายการของหมวดหมู่ โดยมีวัตถุประสงค์เพื่อให้การค้นหาคำเป็นไปด้วยความรวดเร็วโดยมีความเร็วเพราะใช้เวลาในการค้นหาเป็น O(1) เนื่องจากกุญแจหลักเป็นการเก็บแบบเข้ารหัสสามารถตรวจสอบได้ด้วยความรวดเร็ว ซึ่งเหมาะกับระบบจัดประเภทเว็บเพจที่ต้องการความรวดเร็วในการจัดประเภทเว็บเพจ เนื่องมาจากปริมาณการใช้งานอินเทอร์เน็ตที่มีปริมาณมาก การใช้เวลาในการวิเคราะห์เพื่อจัดประเภทเพจมากเกินไปจะส่งผลให้ไม่สามารถรองรับการจัดประเภทเว็บเพจได้อย่างทันท่วงที

13) โมดูลจัดประเภทเว็บเพจ (Webpage Categorization Module)

โมดูลจัดประเภทเว็บเพจเป็นโมดูลที่ถูกใช้ในระบบจัดประเภทเว็บเพจ โดยมีหน้าที่ในการวิเคราะห์ข้อมูลที่ผ่านกระบวนการต่าง ๆ ในแต่ละโมดูลที่ผ่านมาเพื่อบอกหมวดหมู่ของยูอาร์แอลที่เป็นข้อมูลนำเข้า การวิเคราะห์ของโมดูลจัดประเภทเว็บเพจจะทำงานด้วยการค้นหาคำศัพท์ที่ได้จากโมดูลนับความถี่มาตรวจสอบกับพจนานุกรมว่าอยู่หมวดหมู่ใดจากนั้นจะทำการเก็บไว้ในรายการหมวดหมู่ที่ได้พร้อมความถี่ไว้ในพจนานุกรมอีกหนึ่งอันเป็นรูปแบบของหมวดหมู่และความถี่ หากไม่เคยมีหมวดหมู่นั้นมาก่อนจะทำการเพิ่มหมวดหมู่ใหม่และใช้ความถี่ของคำนั้น แต่ถ้ามีหมวดหมู่นั้นอยู่แล้วจะทำการรวมความถี่ของหมวดหมู่ของเก่ากับความถี่ของหมวดหมู่ของคำใหม่ เมื่อระบบทำการตรวจสอบจนเสร็จสิ้นครบทุกคำเป็นที่เรียบร้อยแล้วจะทำการเรียงลำดับความถี่ของหมวดหมู่ในรายการพจนานุกรม ผลลัพธ์ของการจัดประเภทเว็บเพจจะเป็นหมวดหมู่ที่มีความถี่สูงที่สุดในรายการ

14) โมดูลอัพเดตพจนานุกรม (Dictionary Updating Module)

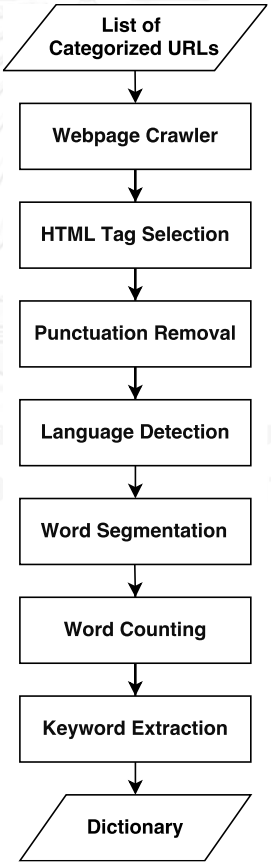
โมดูลอัพเดตพจนานุกรมเป็นโมดูลที่ถูกใช้งานในระบบจัดประเภทเว็บเพจหลังจากเสร็จสิ้นกระบวนการจัดประเภทแล้ว โดยจะทำการนำรายการของคำและความถี่ไปอัพเดตพจนานุกรมที่อยู่ในแฟ้มข้อมูลและฐานข้อมูล เพื่อให้พจนานุกรมมีความทันสมัยอยู่เสมอ

3.1.2 สถาปัตยกรรมของระบบ

ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำประกอบไปด้วย 2 ระบบย่อย คือ ระบบสกัดคำสำคัญและระบบจัดประเภทเว็บเพจ โดยแต่ละระบบมีโมดูลที่ต้องใช้งานและมีขั้นตอนการทำงาน ดังนี้

1) ระบบสกัดคำสำคัญ

ระบบสกัดคำสำคัญเป็นระบบที่ทำขึ้นเพื่อใช้สำหรับสกัดคำสำคัญจากเว็บเพจที่ทราบหมวดหมู่เพื่อนำมาใช้ในการจัดทำพจนานุกรมคำสำหรับการจัดประเภทเว็บเพจ โดยระบบสกัดคำสำคัญมีโมดูลที่จำเป็น คือ โมดูลร้องขอข้อมูลเว็บเพจ โมดูลเลือกป้ายระบุเอชทีเอ็มแอล โมดูลกำจัดเครื่องหมายวรรคตอน โมดูลตรวจสอบภาษา โมดูลตัดคำ โมดูลนับคำ และโมดูลสกัดคำสำคัญ ระบบสกัดคำสำคัญมีแผนภาพการทำงานดังภาพที่ 7



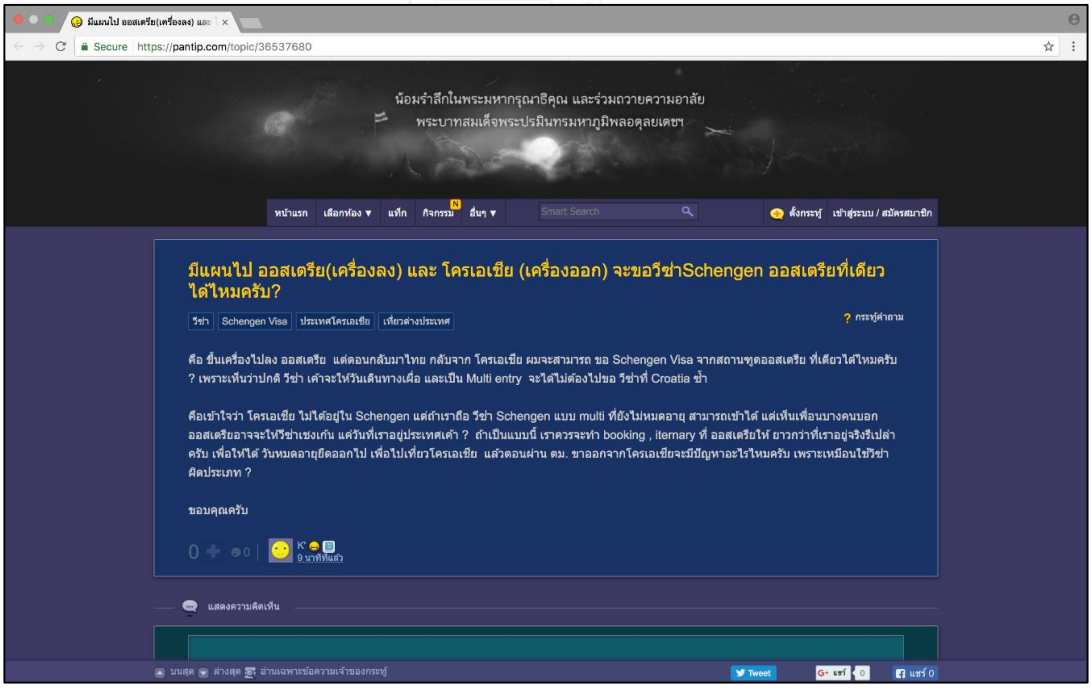
ภาพที่ 7 แสดงแผนภาพการทำงานของระบบสกัดคำสำคัญ

ระบบสกัดคำสำคัญจะทำการรับรายการของยูอาร์แอลที่ทราบหมวดหมู่จากการคัดเลือกมาจาก 50 อันดับเว็บยอดนิยมที่ถูกจัดอันดับโดย SimilarWeb สำหรับภาษาอังกฤษและกระทู้จากเว็บพันทิปเป็นจำนวนห้องละ 10,000 กระทู้ โดยในระบบสกัดคำสำคัญจะทำการยึดหมวดหมู่ของเว็บพันทิปเป็นหลัก และทำการจัดหมวดหมู่ที่ได้จาก SimilarWeb เข้ากับหมวดหมู่ของเว็บพันทิป โดยข้อมูลนำเข้าของระบบจะเป็นรายการของคู่ของยูอาร์แอลกับหมวดหมู่

url	category
pantip.com/topic/35711479	food
pantip.com/topic/36537116	camera
pantip.com/topic/36537680	blueplanet

ภาพที่ 8 แสดงตัวอย่างข้อมูลนำเข้าของระบบสกัดคำสำคัญ

หลังจากที่ได้รับข้อมูลนำเข้ามาแล้วระบบสกัดคำสำคัญจะทำการร้องขอข้อมูลหน้าเว็บเพจจากยูอาร์แอลที่ได้รับมาดังตัวอย่างในภาพที่ 8 โดยข้อมูลหน้าเว็บเพจ (ตัวอย่างหน้าเว็บเพจในภาพที่ 9) ที่ได้รับมาจะอยู่ในรูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสารดังภาพที่ 10



ภาพที่ 9 แสดงตัวอย่างหน้าเว็บเพจ

ภาพที่ 10 แสดงตัวอย่างข้อมูลหน้าเว็บเพจในรูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสาร

หลังจากที่ระบบสกัดคำสำคัญได้ข้อมูลหน้าเว็บเพจที่อยู่ในรูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสารมาแล้ว ระบบจะทำการเลือกข้อมูลเนื้อหาจากป้ายระบุที่ทำการเก็บข้อมูลเนื้อหา ซึ่งในที่นี้คือ ป้ายระบุ “ ” ทำการดึงข้อมูลส่วนหัวเรื่องจาก ป้ายระบุ “<title>...</title>” และป้ายระบุนิยามข้อมูลจาก ป้ายระบุ “<meta name="keywords" content="..." />” โดยจะได้ข้อมูลดังภาพที่ 11

มีแผนไป ออสเตรีย(เครื่องลง) และ โครเอเชีย (เครื่องออก) จะขอวีซ่าSchengen ออสเตรียที่เดียวได้ไหมครับ?

วีซ่า,Schengen Visa,ประเทศโครเอเชีย,เที่ยวต่างประเทศ

คือ ขึ้นเครื่องไปลง ออสเตรีย แต่ตอนกลับมาไทย กลับจาก โครเอเชีย ผมจะสามารถ ขอ Schengen Visa จากสถานฑูตออสเตรีย ที่เดียวได้ไหมครับ ? เพราะเห็นว่าปกติ วีซ่า เค้าจะให้วันเดินทางเผื่อ และเป็น Multi entry จะได้ไม่ต้องไปขอ วีซ่าที่ Croatia ซ้ำ

คือเข้าใจว่า โครเอเชีย ไม่ได้อยู่ใน Schengen แต่ถ้าเราถือ วีซ่า Schengen แบบ multi ที่ยังไม่หมดอายุ สามารถเข้าได้ แต่เห็นเพื่อนบางคนบอก ออสเตรียอาจจะให้วีซ่าเชงเก้น แค่วันที่เราอยู่ประเทศเค้า ? ถ้าเป็นแบบนี้ เราควรจะทำ booking , iternary ที่ ออสเตรียให้ ยาวกว่าที่เราอยู่จริงรีเปล่าครับ เพื่อให้ได้ วันหมดอายุยืดออกไป เพื่อไปเที่ยวโครเอเชีย แล้วตอนผ่าน ตม. ขาออกจากโครเอเชียจะมีปัญหาอะไรไหมครับ เพราะเหมือนใช้วิซ่าผิดประเภท ?

ขอบคุณครับ

ภาพที่ 11 แสดงตัวอย่างเนื้อหาที่จากโมดูลเลือกข้อมูล

เมื่อได้เนื้อหามาแล้วก็มาถึงขั้นตอนในโมดูลกำจัดเครื่องหมายวรรคตอน โดยระบบจะทำการแทนที่เครื่องหมายวรรคตอนทั้งหมดด้วยช่องว่างเพื่อแยกคำออกจากกัน ดังภาพที่ 12

มีแผนไป ออสเตรีย(เครื่องลง) และ โครเอเชีย (เครื่องออก) จะขอวีซ่าSchengen ออสเตรียที่เดียวได้ไหม ครับ?

วีซ่า,Schengen Visa,ประเทศโครเอเชีย,เที่ยวต่างประเทศ

คือ ขึ้นเครื่องไปลง ออสเตรีย แต่ตอนกลับมาไทย กลับจาก โครเอเชีย ผมจะสามารถ ขอ Schengen Visa จากสถานฑูตออสเตรีย ที่เดียวได้ไหมครับ ? เพราะเห็นว่าปกติ วีซ่า เค้าจะให้วันเดินทางเผื่อ และเป็น Multi entry จะได้ไม่ต้องไปขอ วีซ่าที่ Croatia ซ้ำ

คือเข้าใจว่า โครเอเชีย ไม่ได้อยู่ใน Schengen แต่ถ้าเราถือ วีซ่า Schengen แบบ multi ที่ยังไม่หมดอายุ สามารถเข้าได้ แต่เห็นเพื่อนบางคนบอก ออสเตรียอาจจะให้วีซ่าเชงเก้น แค่วันที่เราอยู่ประเทศเค้า ? ถ้าเป็นแบบนี้ เราควรจะทำ booking , iternary ที่ ออสเตรียให้ ยาวกว่าที่เราอยู่จริงรึเปล่าครับ เพื่อให้ได้ วันหมดอายุยืดออกไป เพื่อไปเที่ยวโครเอเชีย แล้วตอนผ่าน ตม. ขาออกจากโครเอเชียจะมีปัญหาอะไรไหมครับ เพราะเหมือนใช้วิซ่าผิดประเภท ?

ขอบคุณครับ

ภาพที่ 12 แสดงตัวอย่างเนื้อหาหลังจากทำการกำจัดเครื่องหมายวรรคตอนแล้ว

หลังจากนั้นระบบจะนำเนื้อหาไปตรวจสอบภาษากับโมดูลตรวจสอบภาษา หากเป็นเนื้อหาภาษาไทยจะนำไปเข้าสู่โมดูลตัดคำภาษาไทยที่ใช้ส่วนต่อประสานโปรแกรมประยุกต์ของ Inspica แต่ถ้าเป็นภาษาอังกฤษจะทำการตัดคำโดยใช้ช่องว่างเลย โดยผลลัพธ์ของการดำเนินการจะได้เนื้อหาที่ถูกแบ่งคำโดยใช้ช่องว่างแบ่งไว้

มี แผน ไป ออสเตรีย เครื่อง ลง และ โครเอเชีย เครื่อง ออก จะ ขอ วีซ่า Schengen ออสเตรีย ที่ เดียว ได้ ไหม ครับ วีซ่า Schengen Visa ประเทศ โครเอเชีย เที่ยว ต่างประเทศ คือ ขึ้น เครื่อง ไป ลง ออสเตรีย แต่ ตอน กลับมา ไทย กลับจาก โครเอเชีย ผม จะ สามารถ ขอ Schengen Visa จาก สถานฑูต ออสเตรีย ที่ เดียว ได้ ไหม ครับ เพราะ เห็น ว่า ปกติ วีซ่า เค้า จะ ให้ วัน เดินทาง เผื่อ และเป็น Multi entry จะ ได้ ไม่ ต้อง ไป ขอ วีซ่า ที่ Croatia ซ้ำ คือ เข้าใจ ว่า โครเอเชีย ไม่ได้ อยู่ใน Schengen แต่ ถ้า เรา ถือ วีซ่า Schengen แบบ multi ที่ ยัง ไม่ หมด อายุ สามารถ เข้า ได้ แต่ เห็น เพื่อน บางคน บอก ออสเตรียอาจจะ ให้ วีซ่า เชงเก้น แค่ วันที่ เรา อยู่ ประเทศ เค้า ถ้า เป็น แบบนี้ เรา ควรจะ ทำ booking iternary ที่ ออสเตรีย ให้ ยาว กว่า ที่ เรา อยู่ จริง รึเปล่า ครับ เพื่อ ให้ ได้ วัน หมด อายุ ยืด ออก ไป เพื่อ ไป เที่ยว โครเอเชีย แล้ว ตอน ผ่าน ตม ขา ออก จาก โครเอเชีย จะ มี ปัญหา อะไร ไหม ครับ เพราะ เหมือน ใช้ วิซ่า ผิด ประเภท ขอบคุณ ครับ

ภาพที่ 13 แสดงตัวอย่างเนื้อหาที่ผ่านโมดูลตัดคำแล้ว

เมื่อข้อมูลอยู่ในรูปแบบของคำที่คั่นด้วยช่องว่างแล้ว ระบบก็จะนำข้อมูลนี้มาทำการนับความถี่ในโมดูลนับความถี่ โดยจะอยู่ในรูปแบบรายการของคำและความถี่

Word	Frequency
วีซ่า	6
โครเอเชีย	6
ที่	6
ออสเตรีย	5
ครับ	5
จะ	6
ได้	5
ไป	5
schengen	5
เรา	4

ภาพที่ 14 แสดงตัวอย่างรายการคู่ของคำและความถี่

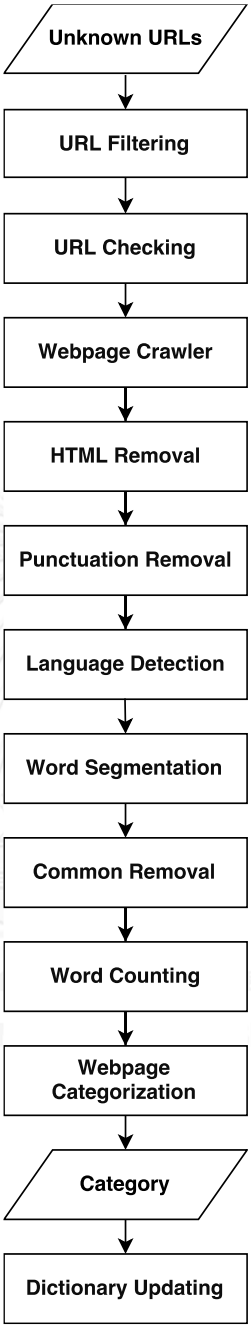
เมื่อทำการนับความถี่ของคำเสร็จสิ้นแล้ว ระบบจะทำการนำรายการของคำและความถี่เพิ่มเข้าไปในพจนานุกรมตามหมวดหมู่ที่ได้มา ดังภาพที่ 15

```
[(travel,[(วีซ่า,6) (โครเอเชีย,6) (ที่,6) (ออสเตรีย,5) (ครับ,5) (จะ,5) ได้,5) ... ]),
...
(food,[(จะ,6) (ไก่,5) (kfc,5) (กิน,5) (ไป,5) (โทร,5) (นาน,5) ... ])]
```

ภาพที่ 15 แสดงตัวอย่างพจนานุกรม

ระบบสกัดคำสำคัญจะดำเนินการจนครบข้อมูลนำเข้าที่ได้รับมา เมื่อเสร็จทั้งกระบวนการก็จะได้พจนานุกรมสำหรับใช้ในระบบจัดประเภทเว็บเพจโดยไม่สนใจว่าคำนั้นจะเป็นภาษาใด ดังนั้นหากเรานำข้อมูลจากภาษาอื่นมาฝึกก็จะสามารถสร้างพจนานุกรมที่สามารถรองรับการจัดประเภทเว็บเพจภาษานั้นได้ด้วย

2) ระบบจัดประเภทเว็บเพจ



ภาพที่ 16 แสดงแผนภาพการทำงานของระบบจัดประเภทเว็บเพจ

ระบบจัดประเภทเว็บเพจเป็นระบบที่ทำขึ้นเพื่อใช้ในการจัดประเภทเว็บเพจที่รองรับการจัดประเภทเว็บเพจที่มีเนื้อหาทั้งภาษาไทยและภาษาอังกฤษและได้รองรับการจัดประเภทเว็บเพจจากข้อมูลการใช้งานจริงของผู้ให้บริการอินเทอร์เน็ต โดยระบบจัดประเภทเว็บเพจมีโมดูลที่จำเป็น คือ โมดูลคัดกรองคัดกรองยูอาร์แอล โมดูลตรวจสอบยูอาร์แอล โมดูลร้องขอเว็บเพจ โมดูลกำจัดป้ายระบุเอชทีเอ็มแอล โมดูลกำจัดเครื่องหมายวรรคตอน โมดูลตรวจสอบภาษา โมดูลตัดคำ โมดูลกำจัดคำ

ทั่วไป โมดูลนับคำ โมดูลจัดประเภทเว็บเพจ และโมดูลอัพเดตพจนานุกรม ระบบจัดประเภทเว็บเพจมีแผนภาพการทำงานดังภาพที่ 16

ระบบจัดประเภทเว็บเพจจะทำการรับยูอาร์แอลที่ต้องการจัดประเภทมาเป็นข้อมูลนำเข้า เมื่อได้รับยูอาร์แอลเข้ามาแล้วจะทำการตรวจสอบส่วนที่อยู่แฟ้มข้อมูลและสกุลแฟ้มข้อมูลของยูอาร์แอลเพื่อคัดกรองยูอาร์แอลที่ไม่ใช้หน้าเว็บเพจออก ตัวอย่างที่อยู่แฟ้มข้อมูลและสกุลแฟ้มข้อมูลที่จะถูกคัดออก เช่น “.jpg” “.png” “.gif” “.mp4” “/img/”, “api.” เป็นต้น

เมื่อผ่านโมดูลคัดกรองยูอาร์แอลแล้ว ระบบจะทำการตรวจสอบยูอาร์แอลที่เข้ามากับฐานข้อมูลว่า มีข้อมูลยูอาร์แอลนี้ถูกจัดประเภทไว้แล้วหรือไม่ หากไม่มีจะทำดำเนินการส่งต่อไปยังโมดูลถัดไป แต่ถ้ามีข้อมูลอยู่แล้วระบบจะทำการตรวจสอบเวลาที่อัพเดตล่าสุด ถ้าเกินเวลาสำหรับการทำแคชชิ่งแล้วจะทำการจัดประเภทให้กับยูอาร์แอลใหม่ เนื่องจากข้อมูลหน้าเว็บเพจนั้นอาจมีการเปลี่ยนแปลงไปแล้ว ในทางตรงกันข้าม เมื่อเวลาที่อัพเดตล่าสุดน้อยกว่าเวลาในการทำแคชชิ่งก็จะทำการคืนหมวดหมู่ที่ถูกจัดไว้แล้วกลับไป

หากยูอาร์แอลที่ได้รับมายังไม่ถูกจัดประเภทหรือระยะเวลาในการอัพเดตล่าสุดเกินเวลาแคชชิ่งไปแล้ว ระบบจัดประเภทเว็บเพจจะทำการจัดประเภทให้ใหม่ โดยทำการร้องขอข้อมูลหน้าเว็บเพจใหม่ โดนข้อมูลจะอยู่ในรูปแบบของแบบจำลองโครงสร้างเอกสารข้อมูลเหมือนกับระบบสกัดคำสำคัญ แต่หลังจากที่ได้ข้อมูลหน้าเว็บเพจมาแล้ว ระบบจะทำการทำความสะอาดป้ายระบุเอชทีเอ็มแอล เหตุที่ระบบไม่ทำการใช้วิธีเดียวกับระบบสกัดคำสำคัญเนื่องจากยูอาร์แอลที่เข้ามายังระบบจัดประเภทเว็บเพจเป็นยูอาร์แอลที่ไม่ทราบโครงสร้างของเว็บเพจ ระบบจัดประเภทเว็บเพจจึงไม่สามารถทำการเลือกดึงเฉพาะเนื้อหาภายในเว็บเพจได้เหมือนกับระบบสกัดคำสำคัญ แต่สามารถดึงข้อมูลส่วนป้ายระบุหัวเรื่องกับป้ายระบุนิยามข้อมูลได้อยู่

การกำจัดป้ายระบุเอชทีเอ็มแอลนั้น ระบบจะทำการลบจาวาสคริปต์ สไตล์ชีต และป้ายระบุเอชทีเอ็มแอลทั้งหมดออกจากระบบ โดยจะได้เนื้อหาที่มีความใกล้เคียงกับโมดูลเลือกข้อมูลของระบบสกัดคำสำคัญ

เมื่อทำการกำจัดป้ายระบุเอชทีเอ็มแอลเสร็จสิ้นแล้ว ระบบจัดประเภทเว็บเพจจะดำเนินการผ่านโมดูลกำจัดเครื่องหมายวรรคตอน โมดูลตรวจสอบภาษา โมดูลตัดคำเหมือนกับระบบสกัดคำสำคัญ

แต่ระบบจัดประเภทเว็บเพจจะมีโมดูลกำจัดคำทั่วไปเพิ่มเข้ามา เนื่องจากคำทั่วไปเป็นคำที่พบได้ทั่วไปในหลายหมวดหมู่ ซึ่งคำเหล่านี้ไม่สามารถนำมาใช้ในการจัดประเภทได้ โดยหลังจากทำการกำจัดคำทั่วไปแล้ว จะเหลือคำดังภาพที่ 17

โครเอเชีย Schengen Schengen Visa โครเอเชีย ต่างประเทศ โครเอเชีย Schengen Visa Croatia โครเอเชีย Schengen Schengen โครเอเชีย โครเอเชีย

ภาพที่ 17 แสดงตัวอย่างเนื้อหาหลังจากกำจัดคำทั่วไปแล้ว

เมื่อผ่านโมดูลกำจัดคำทั่วไปแล้ว เนื้อหาที่เหลืออยู่จะมีแต่คำที่มีความหมายเฉพาะ จากนั้นจะทำการนับความถี่ของคำที่เหลืออยู่ในโมดูลนับคำดังภาพที่ 18

Word	Frequency
โครเอเชีย	6
schengen	5
visa	2
ต่างประเทศ	1
croatia	1

ภาพที่ 18 แสดงตัวอย่างการนับความถี่ของคำในระบบจัดประเภทเว็บเพจ

ต่อมาก็เข้าสู่โมดูลจัดประเภทเว็บเพจ โดยจะทำการตรวจสอบคำเข้ากับพจนานุกรม แล้วทำการสรุปหมวดหมู่สูงสุดออกมาเป็นคำตอบ โดยในที่นี้จะเห็นว่า เหลือแต่คำที่ถ้าให้คนมาดูก็จะรู้ว่าอยู่ในหมวดหมู่ท่องเที่ยว ซึ่งพจนานุกรมที่ทำการฝึกฝนจากข้อมูลในลักษณะเดียวกันก็ย่อมมีคำเหล่านี้ในหมวดหมู่เช่นเดียวกัน โดยในพจนานุกรมจะได้เป็นคำว่า “visa” และ “ต่างประเทศ” อยู่ในหมวดหมู่ท่องเที่ยว ส่วนคำที่เหลือไม่พบในพจนานุกรม

หลังจากนั้นระบบจะทำการคำนวณความน่าจะเป็นของคำตอบโดยการรวมความถี่ของหมวดหมู่แต่ละหมวดหมู่ แล้วทำการหารด้วยจำนวนหมวดหมู่ทั้งหมด หากมีค่ามากกว่าร้อยละ 80 ระบบจะทำการเพิ่มคำลงไปในพจนานุกรม โดยในพจนานุกรมก็จะมีคำว่า “โครเอเชีย” “schengen” และ “croatia” อยู่ด้วย

บทที่ 4
การวัดประสิทธิภาพการทำงานของระบบและผลการทดลอง

การทดสอบเพื่อวัดประสิทธิภาพการทำงานของระบบจัดประเภทเว็บเพจจะแบ่งออกเป็น 3 ส่วน คือ การทดสอบสมมติฐานในการดำเนินงาน การทดสอบระบบสกัดคำสำคัญ การทดสอบระบบจัดประเภทเว็บเพจ การทดสอบระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ และการทดสอบกับส่วนต่อประสานโปรแกรมประยุกต์ที่มีอยู่ในท้องตลาด

4.1 การทดสอบสมมติฐานในการดำเนินงาน

การพัฒนาระบบจัดประเภทเว็บเพจนั้นเริ่มมาจากสมมติฐานที่ว่า ในหน้าเว็บเพจจะประกอบไปด้วยเนื้อหาที่สามารถบ่งบอกถึงหมวดหมู่ของเว็บเพจนั้นได้ ในขั้นแรกของการทำงานทางผู้วิจัยได้เลือกที่จะทำการทดสอบกับภาษาอังกฤษก่อน เนื่องจากการจัดประเภทเว็บเพจภาษาอังกฤษสามารถหาพจนานุกรมคำมาทดสอบและการตัดคำในภาษาอังกฤษก็สามารถกระทำได้โดยง่าย ผู้วิจัยทดสอบสมมติฐานโดยการหาพจนานุกรมคำที่มีการพัฒนาขึ้นโดยผู้อื่นมาทำการทดสอบ และพจนานุกรมคำที่ได้นำมาทดสอบมาจาก [17] ซึ่งประกอบไปด้วยหมวดหมู่และคำศัพท์ที่เกี่ยวข้องมากถึง 660 หมวดหมู่

Entertainment, Recreation, Leisure(121)	Entrepreneurship(312)	Eponyms(48)
Equine therapy(217)	Equinox, Eclipse & Space(215)	ESL, LEP, ELL(347)
Espionage(268)	Ethics(347)	Ethics(353)
Exercise(277)	Explorers(203)	Fables(108)
Fabric and cloth types(133)	Fabrics(151)	Facts(9)
Faith(492)	Family(312)	Fantasy and Imagination(310)
Farming and Agriculture(198)	Fashion(441)	Fashion and clothing(433)
Father's Day(166)	Fencing(177)	Fidel Castro(172)
Finance(303)	Fine Arts(417)	Fire(185)
Fire Prevention & Safety(569)	Firefighters and Safety(569)	Fireworks(190)
First Nation(405)	Fishing(161)	Fitness(232)
Flag Day(191)	Flowers(142)	Flowers and their meanings(38)
Folk Medicine(154)	Folklore(379)	Food and Beverage(457)
Food banks(66)	Football(310)	Football(310)
Force & Gravity(134)	Forensic Anthropology(184)	Fracking(122)
French expressions(38)	French vocabulary words used in English(246)	Friendship(174)
Frosty the Snowman(71)	Gardening(424)	Gardening(415)
GED test(119)	Geography(317)	Geology(279)

ภาพที่ 19 แสดงตัวอย่างหมวดหมู่ของพจนานุกรมที่นำทดสอบ

จากตัวอย่างหมวดหมู่ของพจนานุกรมที่มีอยู่แล้วดังแสดงในภาพที่ 19 นั้นจะเห็นได้ว่า มีปริมาณหมวดหมู่มากเกินความจำเป็นสำหรับการจัดหมวดหมู่เว็บเพจ บางหมวดหมู่เป็นหมวดหมู่ที่ไม่สามารถจัดเว็บเพจลงไปได้เนื่องจากไม่ใช่พจนานุกรมที่ทำขึ้นมาเพื่อใช้ในการจัดประเภทเว็บเพจ เช่น หมวดหมู่ “Facts” เป็นหมวดหมู่ที่กล่าวถึงข้อมูลหรือข้อเท็จจริงซึ่งไม่สามารถนำมาใช้ในการจัด

ประเภทเว็บเพจได้ บางหมวดหมู่ก็มีความซ้ำซ้อนกัน อย่างเช่น “Gardening” กับ “Flowers” เป็นหมวดหมู่ที่กล่าวถึงการทำสวนและดอกไม้ซึ่งมีคำศัพท์ที่อยู่ในหมวดหมู่ทั้งสองหมวดหมู่ซ้ำกันสามารถรวมกันเป็นหมวดหมู่เดียวได้ เป็นต้น นอกจากนั้นยังมีหมวดหมู่ที่มีคำศัพท์ภายในหมวดหมู่ซ้ำซ้อนกันอีกเป็นจำนวนมาก การที่พจนานุกรมมีหมวดหมู่เป็นจำนวนมากสามารถบ่งบอกถึงความละเอียดในการจัดหมวดหมู่ แต่ก็แลกมาด้วยความซ้ำซ้อนของคำศัพท์ในแต่ละหมวดหมู่

Alphalary for Travel and Leisure (363)

A) Abroad, Access, Accommodations, Activities, Addition, Adventure, Affordable, Agency, Airfare, Allure, Ambiance, Amenities, Amount, Ample, Amusement, Appetite, Aquatic, Arrangements, Array, Arts and crafts, Assistance, Assortment, Atmosphere, Attraction, Availability

B) B&B, Backyard, Barbecue, Beach, Bellhop, Beverage, Biking, Boathouse, Boating, Bon voyage, Boutique, Break, Budget, Bug-free, Business class

C) Café, Camper, Campground, Camping, Cancellation, Canoeing, Capacity, Captain, Caravan, Cash, Certification, Challenge, Charter, Chef, Choice, Clientele, Climate, Coach, Coach class, Cocktail hour, Comfort, Comfortable, Contract, Convenience, Costly, Crafts, Credit, Cruise

D) Decadent, Delight, Deluxe, Deposit, Destination, Discounts, Dismay, Dispatch, Distinguish, Diversion, Diversity, Double occupancy, Downtime

E) Earnest, Easy, Energetic, Enjoyable, Enjoyment, Entertainment, Environment, Envision, Equipment, Escape, Event, Exclusive, Excursion, Exercise. expectation, Expedition, Expensive, Experience, Exploration, Extras, Extravagant, Exude

F) Facilities, Fancy, Fanfare, Fare, Fees, First class, Fitness, Flight attendant, Food, Foreign, Free, Free time, Freedom, Friendliness, Function, Furlough, Futon

G) Gastronomy, Gathering, Gear, Get together, Getaway, Gifts, Global, Globetrotter, Golf, Guest, Guide, Gusty

H) Harbor, Hiatus, Hike, Holiday, Honorarium, Hooky, Horseback riding, Hospitality, Host, Hostel, Hostess, Hotel

I) Ideal, Idyll, Impressive, Inn, Instruction, Insurance, Intensive, Interim, International, Island, Itinerary

J) Jaunt, Journey, Journey, Joy, Joyride, Junket

K) Kayaking, Keen, Kid-friendly, Kindly, Kindness, Kinship, Knitting

L) Lake-view, Landmass, Language, Launch, Lazy, Leave, Leisure, Lessons, Liberty, Lifestyle, Limit, Locale, Location, Lodging, Lounge, Luggage, Lull, Luxurious

M) Mandatory, Marina, Massive, Maximum, Meals, Meetings, Memento, Memorable, Minimum, Moderation, Monitor, Mood, Motion, Movement, Music

N) Nice, Nominal, Noteworthy, Noticeable

O) Occasion, Ocean view, Odyssey, Option, Organization, Original, Outdoors, Outing, Outstanding, Overbooking

P) Paddle, Paid vacation, Parade, Park, Participation, Partying, Pause, Payment, Payoff, Peaceful, Pedal boat, Pension, Perambulate, Perks, Picnic, Picturesque, Pizazz, Playground, Playtime, Pleasure, Porter, Promenade, Property, Protection, Public, Purser

Q) Quaint, Quality, Quantity, Query, Quest, Quiet, Quirky

R) Racing, Rate, Reasonable, Recess, Recreation, Recuperation, Refreshment, Refund, Regard, Regatta, Relaxation, Renown, Rental, Reputation, Requisite, Reservation, Reserve, Resort, Restaurant, Retreat, Return ticket, Riparian, Romantic, Round-the-world, Round-trip, Route, Routine, Rowing

S) Sabbatical, Safari, Safety, Sailing, Sanctuary, Sand, Satisfying, Scenic, Secluded, Selection, Setting, Ship, Shore leave, Side-trip, Soothing, Souvenir, Spa, Space, Spacious, Steerage, Steward, Stewardess, Sublime, Successful, Suitcase, Sumptuous, Sunscreen, Sunshine, Swimming pool

T) Tan, Tennis, Tent, Theme park, Time off, Tour, Tourism, Tourist, Tournament, Trail, Trailer, Train, Transfer, Transportation, Travel, Trek, Trip, Tropical, Truancy, Trunk

U) Ubiquitous, Unique, Universal, Updated, Upgrade

V) Vacation, Valuable, Variety, View, Visit, Vista, Volleyball, Voyage

W) Walk, Wander, Water sports, Waterfront, Wayfarer, Weary, Weather, Weekend, Whim, Whirlpool, Wide-ranging, Windsurf, Wireless service, Woo, Workshop, World, World-class, Worldwide

X) Xanadu

Y) Yacht, Year abroad, Yoga, Youth

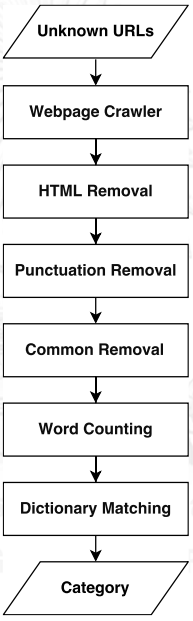
Z) Zeal, Zoological

ภาพที่ 20 แสดงตัวอย่างคำในหมวดหมู่ท่องเที่ยวจากพจนานุกรม

ภาพที่ 20 แสดงตัวอย่างคำศัพท์ในหมวดหมู่ท่องเที่ยวที่ถูกนิยามไว้ในพจนานุกรมนี้ โดยคำศัพท์ที่ถูกนิยามไว้มีทั้งคำศัพท์ที่เกี่ยวข้องกับการท่องเที่ยว เช่น “Abroad” “Agency” “Journey” “Transportation” เป็นต้น คำศัพท์ที่ไม่น่าจะอยู่ในหมวดหมู่การท่องเที่ยว เช่น “Tennis” เป็นคำศัพท์ที่น่าจะอยู่ในหมวดหมู่เกี่ยวกับกีฬามากกว่า และนอกจากนี้ยังมีคำศัพท์ที่เป็นคำทั่วไปที่สามารถพบได้ในหลายหมวดหมู่ เช่น “Easy” “Nice” และ “Unique” เป็นต้น

จากการทดสอบสมมติฐานจึงทำการนำข้อมูลเว็บข่าวมาทำการทดสอบมาทำการทดสอบกับพจนานุกรม โดยระบบจัดประเภทเว็บเพจในช่วงแรกมีแผนผังการทำงานดังภาพที่ 21 กล่าวคือ ระบบจะทำการรับยูอาร์แอลเว็บข่าวมาแล้วทำการร้องขอข้อมูลหน้าเว็บเพจมาในรูปแบบของแบบจำลองโครงสร้างข้อมูลเอกสาร โดยเนื้อหาที่ได้รับมาในขั้นแรกนี้ยังไม่สามารถนำไปใช้ในการวิเคราะห์ได้ เนื่องจากข้อมูลยังเป็นข้อมูลดิบอยู่ การนำข้อมูลดิบไปทำการวิเคราะห์จะส่งผลให้ได้ผลลัพธ์ในการ

วิเคราะห์ที่ไม่เต็มประสิทธิภาพ จึงเป็นที่มาของการทำความสะอาดข้อมูลก่อนการนำข้อมูลไปวิเคราะห์ โดยการทำความสะอาดข้อมูลที่ได้จากการร้องขอข้อมูลหน้าเว็บโดยทั่วไปแล้วมี 3 ส่วน คือ การกำจัดป้ายระบุเอชทีเอ็มแอล การกำจัดเครื่องหมายวรรคตอน และการกำจัดคำทั่วไป สำหรับรายการคำทั่วไปที่นำมาใช้ในระบบนี้เป็นรายการคำทั่วไปสำหรับภาษาอังกฤษที่ใช้กันโดยทั่วไป 100 อันดับคำทั่วไปที่นิยมใช้กัน ซึ่งถูกรวบรวมไว้โดยคลังภาษาอังกฤษของอ็อกฟอร์ด (Oxford English Corpus) ดังตัวอย่างในภาพที่ 22 หลังจากการทำความสะอาดข้อมูลเสร็จสิ้นเป็นที่เรียบร้อยแล้ว ระบบจะทำการนับความถี่ของคำและตรวจสอบคำที่นับได้เหล่านั้นกับพจนานุกรมเพื่อระบุหมวดหมู่ของคำเหล่านั้น และทำการบอกหมวดหมู่ที่มีความถี่มากที่สุด 3 อันดับแรกเป็นคำตอบของระบบ



ภาพที่ 21 แสดงแผนผังการทำงานของระบบจัดประเภทเว็บเพจในช่วงทดสอบสมมติฐาน

จากการทดสอบโดยการนำเว็บข่าวจากสำนักข่าวต่าง ๆ มาเป็นข้อมูลทดสอบพบว่า ระบบจัดประเภทเว็บเพจมีความแม่นยำ (Precision) ในการจัดประเภทเพียงร้อยละ 44 เท่านั้น แต่หากพิจารณาคำตอบของระบบที่คืนค่าเป็นหมวดหมู่ที่มีความถี่มากที่สุด 3 อันดับแรกพบว่า ความแม่นยำในการจัดหมวดหมู่ที่ถูกต้องของระบบจัดประเภทเว็บเพจโดยคำตอบที่ถูกต้องเป็นหนึ่งในสามหมวดหมู่ที่ระบบตอบกลับมามีค่าเพิ่มขึ้นมาเป็นร้อยละ 70 โดยข้อมูลที่นำมาทดสอบเป็นเว็บจากสำนักข่าวจำนวน 4 หมวดหมู่ด้วยกัน ได้แก่ หมวดหมู่การเงิน หมวดหมู่การเมือง หมวดหมู่เทคโนโลยี และหมวดหมู่สุขภาพ

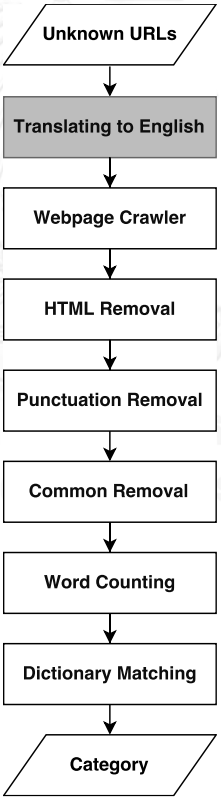
Rank	Word	Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

ภาพที่ 22 แสดงตัวอย่างคำทั่วไป 100 อันดับแรกจากคลังข้อมูลของอ๊อกฟอร์ด [18]

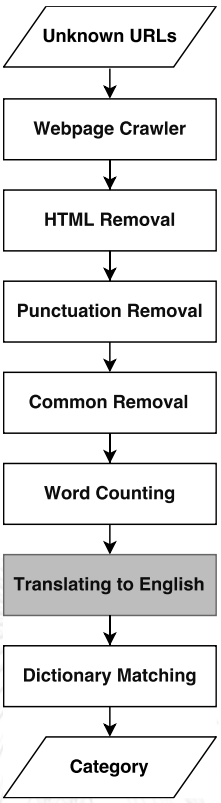
หลังจากทดสอบสมมติฐานการทำระบบจัดประเภทเว็บเพจกับเว็บเพจที่มีเนื้อหาภาษาอังกฤษแล้ว เนื่องจากต้องการทดสอบสมมติฐานนี้กับเว็บเพจที่มีเนื้อหาเป็นภาษาไทยด้วย จึงได้ทำการทดลองหาพจนานุกรมคำภาษาไทยที่อาจจะมีอยู่แล้ว ปรากฏว่าไม่พบพจนานุกรมคำภาษาไทยเลย แต่เนื่องจากการมีพจนานุกรมภาษาอังกฤษอยู่แล้วจึงเกิดความคิดว่า หากนำเว็บเพจที่มีเนื้อหาภาษาไทยมาทำการแปลเป็นภาษาอังกฤษก่อนที่จะนำเนื้อหาในเว็บเพจที่แปลเป็นภาษาอังกฤษเรียบร้อยแล้วไปจัดประเภทในระบบจัดประเภทเว็บเพจควรจะได้ผลลัพธ์ใกล้เคียงกับการจัดประเภทเว็บเพจที่มีเนื้อหาเป็นภาษาอังกฤษ

จากการทดสอบโดยการนำเว็บข่าวที่มีเนื้อหาเป็นภาษาไทยมาทำการทดสอบ 2 รูปแบบ คือ การทดสอบโดยการแปลเว็บที่มีเนื้อเป็นภาษาไทยเป็นภาษาอังกฤษทั้งเว็บเพจ และการทดสอบโดยการตัดคำภาษาไทยก่อนที่จะทำการแปลคำภาษาไทยเป็นภาษาอังกฤษทีละคำเพื่อที่จะนำคำศัพท์ที่แปลแล้วไปตรวจสอบกับพจนานุกรมภาษาอังกฤษที่มีอยู่ การทดสอบครั้งนี้ได้ใช้เว็บข่าวในจำนวน 4 หมวดหมู่เช่นเดียวกับการทดสอบกับเว็บเพจที่มีเนื้อหาภาษาอังกฤษ ได้แก่ หมวดหมู่การเงิน หมวดหมู่การเมือง หมวดหมู่เทคโนโลยี และหมวดหมู่สุขภาพ ผลลัพธ์การทดสอบความแม่นยำพบว่า

การทดสอบโดยการแปลเว็บที่มีเนื้อหาเป็นภาษาไทยเป็นภาษาอังกฤษทั้งเว็บโดยสนใจเพียงแค่หมวดหมู่แรกเท่านั้นได้ค่าความแม่นยำประมาณร้อยละ 40 แต่หากเป็นการทดสอบโดยการตัดคำภาษาไทยก่อนที่จะทำการแปลคำภาษาไทยเป็นภาษาอังกฤษทีละคำกลับให้ค่าความแม่นยำไม่ถึงร้อยละ 10 จากการตรวจสอบพบว่า การตัดคำภาษาไทยแล้วทำการแปลเป็นภาษาอังกฤษทีละคำนั้นเป็นสิ่งที่ผิดพลาด เนื่องจากคำภาษาไทยหนึ่งคำมีความหมายมากกว่า 1 อย่างและสามารถแปลเป็นภาษาอังกฤษได้มากกว่า 1 คำ เช่น “ตา” มีความหมายมากกว่า 1 ความหมาย อาจหมายถึงพ่อของแม่ ส่วนหนึ่งของต้นไม้ตรงที่ใช้สำหรับแตกกิ่ง หรืออวัยวะที่ใช้ในการมองเห็นก็ได้ แต่เมื่อทำการแปลเป็นภาษาอังกฤษแล้วจะกลายเป็นคำว่า “eye” ซึ่งหมายถึง อวัยวะที่ใช้ในการมองเห็นเพียงอย่างเดียว ด้วยเหตุนี้จึงส่งผลให้ค่าความแม่นยำในการทดสอบโดยการตัดคำภาษาไทยก่อนที่จะทำการแปลคำภาษาไทยเป็นภาษาอังกฤษทีละคำมีค่าน้อยกว่าที่ควร



ภาพที่ 23 แสดงแผนภาพการทดสอบโดยการแปลเว็บที่มีเนื้อหาเป็นภาษาไทยเป็นภาษาอังกฤษทั้งเว็บเพจ



ภาพที่ 24 แสดงแผนภาพการทดสอบโดยการตัดคำภาษาไทยก่อนที่จะทำการแปลคำภาษาไทยเป็นภาษาอังกฤษทีละคำ

เมื่อทำการตรวจสอบผลการทดสอบแบบละเอียดพบว่า สาเหตุที่ความแม่นยำในการจัดประเภทน้อยเพราะว่า มีคำทั่วไปอยู่ในพจนานุกรมเยอะมาก สังเกตได้จากการที่คำหนึ่งคำไปปรากฏอยู่ในหมวดหมู่ต่าง ๆ หลายหมวดหมู่ จึงได้ทำการทดสอบการกรองคำทั่วไปโดยแบ่งเป็น 4 ระดับ คือ การทดสอบโดยไม่มีการกรองคำทั่วไปในพจนานุกรม, การทดสอบโดยให้คำหนึ่งคำอยู่ได้แค่หนึ่งหมวดหมู่เท่านั้น การทดสอบโดยให้คำหนึ่งคำอยู่ได้ไม่เกิน 5 หมวดหมู่ และการทดสอบโดยให้คำหนึ่งคำอยู่ได้ไม่เกิน 10 หมวดหมู่ ผลการทดสอบปรากฏว่า การทดสอบโดยไม่มีการกรองคำทั่วไปในพจนานุกรมมีความแม่นยำประมาณร้อยละ 44 เนื่องจากมีคำทั่วไปในพจนานุกรมอยู่เยอะมาก ซึ่งส่งผลให้ผลลัพธ์เกิดความผิดพลาดขึ้น การทดสอบโดยให้คำหนึ่งคำอยู่ได้แค่หนึ่งหมวดหมู่เท่านั้นได้ค่าความแม่นยำประมาณร้อยละ 60 เนื่องจากคำที่เหลืออยู่ในพจนานุกรมจะมีเพียงคำเฉพาะเท่านั้น โดยคำหนึ่งคำจะแสดงถึงหมวดหมู่ใดหมวดหมู่หนึ่งเท่านั้น สำหรับการทดสอบโดยให้คำหนึ่งคำอยู่ได้ไม่เกิน 5 หมวดหมู่ผลปรากฏว่า มีค่าความแม่นยำมากถึงร้อยละ 95 และการทดสอบสุดท้ายคือการทดสอบโดยให้คำหนึ่งคำอยู่ได้ไม่เกิน 10 หมวดหมู่ได้ผลลัพธ์ความแม่นยำอยู่ที่ร้อยละ 80 โดยการทดสอบนี้ได้ทดสอบโดยการนำข้อมูลของชุดข้อมูลรอยเตอร์มาใช้ในการทดสอบ โดยทำการสุ่มตรวจสอบเนื่องจากมีปริมาณข้อมูลสูง

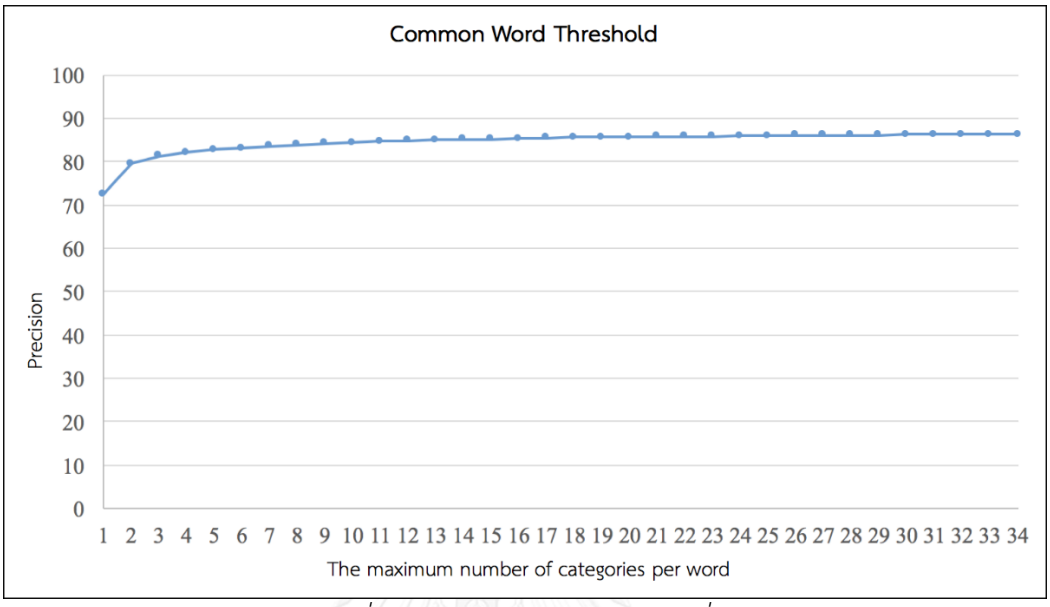
4.2 การทดสอบระบบสกัดคำสำคัญ

หลังจากที่เล็งเห็นถึงความเป็นไปได้ในการทำระบบจัดประเภทเว็บเพจโดยมีสมมติฐานว่า ในหน้าเว็บเพจจะประกอบไปด้วยเนื้อหาที่สามารถบ่งบอกถึงหมวดหมู่ของเว็บเพจนั้นได้ แต่เนื่องจากไม่มีพจนานุกรมที่เหมาะสมที่สามารถนำมาใช้กับระบบจัดประเภทเว็บเพจได้ ผู้วิจัยจึงเกิดแนวคิดที่ว่าหากเชิญผู้เชี่ยวชาญในแต่ละด้านมาทำการจัดทำพจนานุกรมจะทำให้ได้พจนานุกรมที่ประกอบไปด้วยคำที่เกี่ยวข้องกับหมวดหมู่นั้นจริง ๆ แต่ในทางปฏิบัติแล้วมันเป็นสิ่งที่เป็นไปไม่ได้ เนื่องจากคำศัพท์มีการเปลี่ยนแปลงอยู่ตลอดเวลา ไม่ว่าจะเป็นการเปลี่ยนความหมายของคำ เช่น “เท” ในอดีตหมายถึงการรินน้ำลงไป แต่ปัจจุบันมีอีกความหมายคือ การทิ้งให้ผู้อื่นต้องอยู่คนเดียว เป็นต้น หรือการเพิ่มขึ้นของคำ เช่น “AlphaGo” ที่เมื่อก่อนไม่เคยมีคำนี้เกิดขึ้น แต่ในปัจจุบันถูกใช้เป็นชื่อของระบบปัญญาประดิษฐ์ (Artificial Intelligent) สำหรับการเล่นหมากล้อม เป็นต้น จากปัญหาเหล่านี้เป็นเหตุให้แนวคิดในการเชิญผู้เชี่ยวชาญในแต่ละหมวดหมู่มาทำการสร้างพจนานุกรมเป็นเรื่องที่เป็นไปไม่ได้ เพราะเราไม่สามารถให้ผู้เชี่ยวชาญมาทำการเปลี่ยนแปลงพจนานุกรมอยู่ตลอดเวลาได้

ด้วยเหตุนี้จึงเกิดความคิดในการสร้างพจนานุกรมขึ้นมาเอง โดยอาศัยแนวคิดที่ว่า ในเมื่อคำสำคัญที่เป็นตัวแทนของหมวดหมู่มีอยู่ในหน้าเว็บเพจอยู่แล้ว การนำเว็บเพจที่ทราบหมวดหมู่อยู่แล้วมาทำการสกัดเอาคำสำคัญในหน้าเว็บเพจเหล่านั้นออกมาจะทำให้ได้คำสำคัญที่ถูกใช้ในเว็บเพจของหมวดหมู่นั้นอย่างแท้จริงมาใช้ในการสร้างพจนานุกรม โดยในขั้นตอนของการสกัดคำสำคัญยังส่งผลให้เราสามารถระบุได้ด้วยว่า คำ ๆ ใดเป็นคำทั่วไปที่นิยมใช้กันในหลาย ๆ หมวดหมู่จากการเก็บรวบรวมคำต่าง ๆ ที่พบในขั้นตอนการสกัดคำสำคัญของแต่ละหมวดหมู่ จึงเป็นที่มาของระบบสกัดคำสำคัญที่ได้กล่าวถึงไปในข้างต้น โดยได้ทำการทดลองปรับค่าขีดจำกัดของจำนวนหมวดหมู่สูงสุดต่อคำหนึ่งคำตั้งแต่คำหนึ่งคำมีความเกี่ยวข้องได้แค่หมวดหมู่เดียวไปจนถึงคำหนึ่งคำสามารถเกี่ยวข้องได้ทั้ง 34 หมวดหมู่

โดยกราฟในภาพที่ 25 แสดงค่าขีดจำกัด (Threshold) ของคำทั่วไป โดยการเปรียบเทียบระหว่างค่าความแม่นยำกับจำนวนหมวดหมู่สูงสุดต่อคำหนึ่งคำ จากกราฟจะเห็นว่า เมื่อกำหนดเงื่อนไขให้คำหนึ่งคำอยู่ได้เพียงแค่หนึ่งหมวดหมู่เท่านั้นมีค่าเฉลี่ยความแม่นยำอยู่ที่ร้อยละ 72.61 เนื่องจากในพจนานุกรมจะมีเพียงคำที่มีความหมายเฉพาะเจาะจงในแต่ละหมวดหมู่เท่านั้น จากกราฟจะเห็นว่าค่าความแม่นยำหลังจากที่เพิ่มค่าขีดจำกัดจำนวนหมวดหมู่ต่อคำหนึ่งคำเพิ่มขึ้น ค่าความแม่นยำเริ่มเข้าสู่สภาวะสมดุลตั้งแต่ 18 โดยค่าความแม่นยำเริ่มไม่ค่อยมีการพัฒนาเพิ่มขึ้นหรือเพิ่มขึ้นในปริมาณที่น้อย โดยที่ค่าความแม่นยำสำหรับเงื่อนไขที่ค่าขีดจำกัดคำทั่วไปเป็น 18 มีค่าความแม่นยำเฉลี่ยอยู่ที่ร้อยละ 85.20 สำหรับเงื่อนไขค่าขีดจำกัดที่ 34 คือไม่มีการกรองคำทั่วไปออกก่อน ให้ค่า

ความแม่นยำอยู่ที่ร้อยละ 86.30 ซึ่งเป็นปริมาณที่เพิ่มมาจากจุดที่มีเงื่อนไขค่าขีดจำกัดที่ 18 เพียงแค่ร้อยละ 1.10 เท่านั้น



ภาพที่ 25 แสดงค่าขีดจำกัดของคำทั่วไป

ภาพที่ 26 แสดงตัวอย่างของคำทั่วไปที่ได้จากการคัดกรองคำที่อธิบายถังหมวดหมู่มากกว่า 18 หมวดหมู่ จะเห็นคำว่า “มากกกก” ซึ่งเป็นคำที่ไม่อยู่ในพจนานุกรมปกติแต่เป็นคำที่ถูกใช้งานในหลากหลายหมวดหมู่ โดยในการทำความสะอาดข้อมูลปกติแล้ว หากผู้วิจัยทำการระบุรายการคำทั่วไปด้วยตนเองก็จะนึกถึงคำเหล่านี้ในรายการคำทั่วไป

ตรวจสอบ	**ฉะนั้น**	**ถ้าหาก**	**แกง**	**เหล่า**
บุคคล	**บุคลิก**	**มากกกก**	**สอดคล้อง**	**อิทธิพล**
ป้องกัน	**เหล็ก**	**คุ้มครอง**	**เชื่อมั่น**	**สร้าง**
ป่ะ	**พัน**	**เข้ม**	**คณะ**	**ทั้งๆที่**
ไล่	**พับ**	**จุด**	**รอคอย**	**หนึ่ง**
สิ้นสุด	**พัด**	**จุก**	**กี่**	**เนื้อหา**
เกือบ	**พัก**	**เข้า**	**เหม็น**	**สุขภาพ**
ชั่ง	**พัง**	**จุง**	**เงา**	**ระลึก**
ค้าง	**ทราบ**	**แกะ**	**ซื่อสัตย์**	**พบเห็น**
ชั่ว	**ทราย**	**ด่วน**	**ท่าที**	**ขวาง**

ภาพที่ 26 แสดงตัวอย่างของคำทั่วไป

ภาพที่ 27 แสดงปริมาณคำในแต่ละหมวดหมู่ที่มีคำเหมือนกัน โดยคำที่ซ้ำกันในแต่ละหมวดหมู่นั้นหมายถึงคำ ๆ นั้นปรากฏอยู่ในเนื้อหาที่เกี่ยวข้องกับหมวดหมู่ทั้งสอง โดยสองหมวดหมู่ที่มีคำซ้อนทับกันเยอะมากที่สุด คือ หมวดหมู่ศาสนา (Religious) กับหมวดหมู่นักเขียน (Writer) โดยมีคำซ้อนทับกันมากถึง 4,059 คำ ในขณะที่หมวดหมู่ศาสนามีคำในหมวดหมู่นี้ 19,109 คำ ส่วนหมวดหมู่นักเขียนมีคำในหมวดหมู่ 18,634 คำ โดยหมวดหมู่ศาสนาเป็นหมวดหมู่ที่มีคำมากที่สุดในบรรดาหมวดหมู่ทั้งหมด 34 หมวดหมู่

	home	horoscope	isolate	jatujak	klaibann	library	lumpini	mbk	pantip	rajdumnern	ratchada
home	6210										
horoscope	1177	13117									
isolate	374	505	2928								
jatujak	878	1357	298	6274							
klaibann	693	1294	437	662	8327						
library	559	1513	332	601	1129	7046					
lumpini	474	1042	223	595	533	491	4237				
mbk	604	512	334	313	548	348	253	4300			
pantip	645	1359	388	774	773	795	598	502	7453		
rajdumnern	1307	3100	585	1531	1697	1915	981	658	1677	16504	
ratchada	876	1186	351	731	797	606	522	520	643	1531	6864
region	652	1148	281	763	645	635	410	376	693	1387	1058
religious	1096	4313	485	1362	1329	1953	1039	549	1396	3951	1251
siam	376	1219	255	481	625	842	499	250	549	1038	417
siliconvalley	685	533	358	297	629	453	234	978	532	746	485
silom	860	1360	388	680	1156	975	594	549	733	1784	802
sinthorn	942	1221	395	644	989	793	485	611	764	1976	951
social	966	1893	463	911	1110	937	848	793	929	2120	1100
supachalasai	955	1835	488	981	1235	1153	911	532	1212	2454	1261
tvshow	864	1899	509	1012	1163	1239	670	525	1531	2441	892
wahkor	1097	1993	421	990	1300	1538	768	590	945	2528	1159
writer	1160	3495	542	1652	1387	1771	1102	463	1757	3918	1373

	region	religious	siam	siliconvalley	silom	sinthorn	social	supachalasai	tvshow	wahkor	writer
region	5643										
religious	1319	19109									
siam	422	1010	4671								
siliconvalley	306	586	287	5166							
silom	658	1487	714	635	6430						
sinthorn	651	1395	479	586	1486	7635					
social	826	2085	974	672	1601	1212	7212				
supachalasai	928	1972	699	619	1123	1161	1273	14228			
tvshow	860	2073	792	570	1137	1058	1375	1969	12096		
wahkor	855	2125	679	845	1306	1205	1255	1625	1360	10069	
writer	1351	4059	1213	510	1466	1340	1940	2294	2484	2033	18634

art	5500																		
bangrak	278	4504																	
beauty	414	362	5135																
blueplanet	918	566	779	11642															
camera	580	258	378	910	5306														
cartoon	723	516	554	1137	540	11215													
chalermkrung	645	609	753	1253	593	1273	11727												
chalermthai	835	569	711	1542	678	2034	1780	11357											
family	504	861	539	1063	433	823	771	1012	6193										
food	652	525	755	2005	634	860	1003	1106	911	8827									
gallery	1397	886	1106	3842	1573	1649	1855	2230	1515	2328	18279								
greenzone	720	402	552	1271	545	686	659	935	876	1088	1677	6569							
home	602	354	513	949	613	637	642	790	686	836	1282	956							
horoscope	990	1129	832	1778	730	1400	1385	1760	1776	1406	2578	1343							
isolate	282	237	256	501	280	442	462	468	352	335	653	334							
jatujak	511	471	525	1073	433	710	672	847	850	1015	1480	1120							
klaibann	618	471	558	1926	619	874	1132	1227	975	1128	2169	836							
library	835	426	395	1099	425	886	971	1235	765	822	1474	903							
lumpini	334	486	822	650	302	486	501	573	865	692	915	610							
mbk	405	202	267	656	568	478	479	531	400	453	913	359							
pantip	660	541	825	1198	547	997	1087	1404	810	997	1737	879							
rajdumnern	1171	933	845	2248	890	1680	1748	2252	1658	1756	3090	1888							
ratchada	530	409	477	1310	621	716	671	802	796	872	1679	859							
region	557	364	387	1424	475	547	579	707	744	971	2002	844							
religious	1201	976	753	1875	783	1429	1246	1935	1767	1465	2672	1626							
siam	366	1355	381	630	279	524	644	623	778	526	914	482							
siliconvalley	566	192	308	636	532	808	545	694	346	433	889	400							
silom	636	508	516	1118	540	803	835	1018	929	894	1385	896							
sinthorn	529	400	470	1125	504	772	786	996	770	853	1393	859							
social	649	892	566	1256	560	882	903	1122	1410	975	1588	982							
supachalasai	773	672	841	1589	759	1425	1513	1802	1073	1274	2205	1057							
tvshow	790	758	819	1506	653	1488	2221	2319	1141	1206	2216	969							
wahkor	863	491	633	1555	783	1203	1075	1628	999	1194	2075	1441							
writer	1151	1237	948	2141	808	1757	1615	2243	1829	1774	3191	1667							
	art	bangrak	beauty	blueplanet	camera	cartoon	chalermkrung	chalermthai	family	food	gallery	greenzone							

ภาพที่ 27 แสดงจำนวนคำที่ร่วมกันในแต่ละหมวดหมู่

4.3 การทดสอบระบบจัดประเภทเว็บเพจ

4.3.1 สมมติฐานการทดลอง

หลังจากการพิสูจน์สมมติฐานในการดำเนินงานเสร็จสิ้นแล้วและได้ข้อสรุปในการสร้างพจนานุกรมคำขึ้นมาสำหรับใช้ในการจัดประเภทเว็บเพจโดยเฉพาะ จุดประสงค์ในการทดลองครั้งนี้คือเพื่อวัดประสิทธิภาพของระบบจัดประเภทเว็บเพจด้วยอัลกอริทึมแบบต่าง ๆ ได้แก่ อัลกอริทึมพื้นฐานทางสถิติ อัลกอริทึมนาอีฟเบย์ และอัลกอริทึมเวิร์ดทูเวค โดยมีสมมติฐานอยู่ว่า การใช้อัลกอริทึมพื้นฐานทางสถิติมีความซับซ้อนต่ำจะใช้เวลาในการประมวลผลน้อยกว่าการใช้อัลกอริทึมที่มีความซับซ้อนสูงกว่า ซึ่งสามารถพัฒนาอัลกอริทึมนี้ทำให้ได้ค่าวัดประสิทธิภาพ (F-Measure) พอ ๆ กับอัลกอริทึมอื่น การทดสอบสามารถแบ่งออกเป็น 2 ส่วน คือ การวัดความเร็วในการฝึกฝนข้อมูลและทดสอบข้อมูล และการวัดค่าวัดประสิทธิภาพในการจัดประเภทเว็บเพจ

4.3.2 ข้อมูลที่นำมาใช้ในการทดสอบ

การทดลองนี้เป็นการทดสอบระบบจัดประเภทเว็บเพจโดยพิจารณาเว็บเพจที่มีเนื้อหาเป็นภาษาไทยและภาษาอังกฤษ โดยเว็บเพจที่ใช้ทดสอบมาจาก 50 อันดับเว็บยอดนิยมที่ถูกจัดอันดับโดย SimilarWeb ซึ่งจะเลือกมาเฉพาะเว็บเพจที่มีเนื้อหาเป็นภาษาอังกฤษเท่านั้น และสำหรับเว็บเพจที่มีเนื้อหาเป็นภาษาไทยได้เลือกใช้เว็บบอร์ดยอดนิยมของคนไทยที่ชื่อว่า พันทิป มาใช้ในการทดสอบ โดยหมวดหมู่ที่ใช้ในการทดสอบมาจากจำนวนห้องหรือหมวดหมู่ในเว็บพันทิปเป็นจำนวน 34 หมวดหมู่ และมีปริมาณเว็บเพจในแต่ละหมวดหมู่ประมาณ 1,000 เว็บเพจ โดยรวมทั้งเว็บเพจที่เป็นภาษาไทยและภาษาอังกฤษเป็นจำนวนทั้งสิ้น 34,000 เว็บเพจ

4.3.3 มาตรวัดประสิทธิภาพ

ในการทดลองครั้งนี้มีมาตรวัดประสิทธิภาพอยู่ด้วยกัน 2 มาตรวัดสำหรับการวัดประสิทธิภาพของระบบ ได้แก่ มาตรวัดความเร็วในการฝึกฝนและทดสอบระบบ และมาตรวัดค่าประสิทธิภาพโดยรวมของอัลกอริทึมที่ใช้ในระบบ

- มาตรวัดความเร็วในการฝึกฝน

ในการทดสอบความเร็วในการฝึกฝนข้อมูลเพื่อใช้ในการสร้างพจนานุกรมหรือโมเดลเพื่อให้เกิดความยุติธรรมในการทำงาน ทางผู้วิจัยจึงได้ทำการเตรียมข้อมูลให้อยู่ในรูปแบบที่ทุกอัลกอริทึมสามารถนำไปใช้ได้เหมือนกัน โดยการเตรียมข้อมูลนั้นเริ่มตั้งแต่ขั้นตอนการร้องขอข้อมูลเว็บเพจที่จะทำการร้องขอข้อมูลหน้าเว็บเพจมาให้อยู่ในรูปแบบของแบบจำลองโครงสร้างเอกสารข้อมูล แล้วนำข้อมูลนั้นมาผ่านขั้นตอนการทำความสะอาดข้อมูล โดยการทำความสะอาดข้อมูลนั้นจะดำเนินการด้วยโมดูลกำจัดป้ายระบุเอชทีเอ็มแอล โมดูลกำจัด

เครื่องหมายวรรคตอน โมดูลตรวจสอบภาษา โมดูลตัดคำ และโมดูลนับความถี่ของคำ จนได้ข้อมูลออกมาดังภาพที่ 28 เพื่อให้แต่ละอัลกอริทึมได้ข้อมูลเดียวกันและรูปแบบที่เหมือนกัน

$$C_1 = [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)]$$
$$C_2 = [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)]$$
$$C_1 = [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)]$$
$$C_5 = [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)]$$
$$\dots$$

ภาพที่ 28 แสดงตัวอย่างข้อมูลที่ถูกเตรียมไว้

ข้อมูลที่ถูกเตรียมไว้จะถูกเก็บอยู่ในรูปแบบของรายการ (List) ของหมวดหมู่ของเว็บเพจที่นำมาใช้ในการฝึกฝนกับรายการ (List) ของคู่ (Tuple) ของคำกับความถี่ โดยตัวแปรแต่ละตัวมีนิยามดังนี้

C_i หมายถึง หมวดหมู่ของเว็บเพจที่มาจากห้องของเว็บพันทิป

w หมายถึง คำที่พบในเนื้อหาเว็บเพจ

f หมายถึง ความถี่ของคำที่พบ

n หมายถึง จำนวนคำในแต่ละหมวดหมู่

สำหรับขั้นตอนการทำงานในระบบสกัดคำสำคัญเพื่อนำมาใช้ในการทดสอบครั้งนี้จะเริ่มจากการทำงานในโมดูลสกัดคำสำคัญเป็นต้นไป โดยขั้นตอนการทำงานนี้จะทำการรวมรายการของคำกับความถี่ที่อยู่ในหมวดหมู่เดียวกันให้เหลือรายการเดียวโดยเก็บอยู่ในรูปแบบของตัวแปรพจนานุกรมโดยมีกุญแจหลัก (Key) เป็นหมวดหมู่ที่ไม่ซ้ำกันและมีค่า (Value) เป็นรายการของคำและความถี่ที่ถูกเรียงลำดับแล้ว ดังภาพที่ 29

$$\{C_1: [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)],$$
$$C_2: [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)],$$
$$C_3: [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)],$$
$$\dots$$
$$C_m: [(w_1, f_1), (w_2, f_2), \dots, (w_n, f_n)]\}$$

ภาพที่ 29 แสดงตัวอย่างพจนานุกรมที่สร้างขึ้น

- มาตรวัดค่าประสิทธิภาพโดยรวม (F-Measure) ของอัลกอริทึมที่ใช้ในระบบ

มาตรวัดค่าประสิทธิภาพคือการวัดประสิทธิภาพโดยรวมระหว่างค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) โดยมีสมการดังนี้

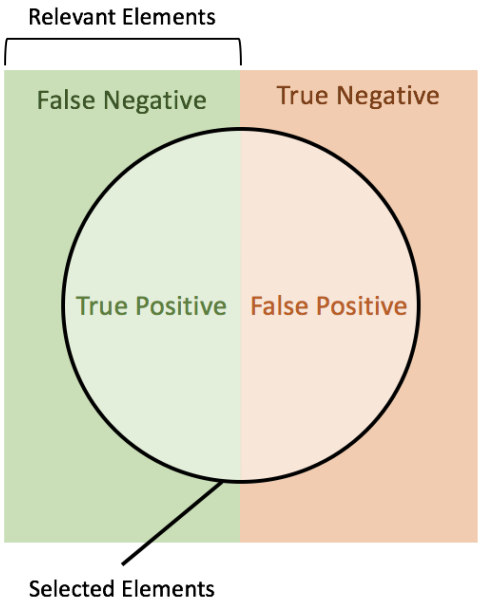
$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}$$ สมการที่ 1

ค่าความแม่นยำ (Precision) คือ จำนวนข้อมูลที่ทำนายถูกจากข้อมูลที่เป็นคลาสที่กำลังพิจารณาอยู่

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$ สมการที่ 2

ค่าความระลึก (Recall) คือ จำนวนข้อมูลที่ทำนายถูกจากข้อมูลทั้งหมดที่เป็นจริง

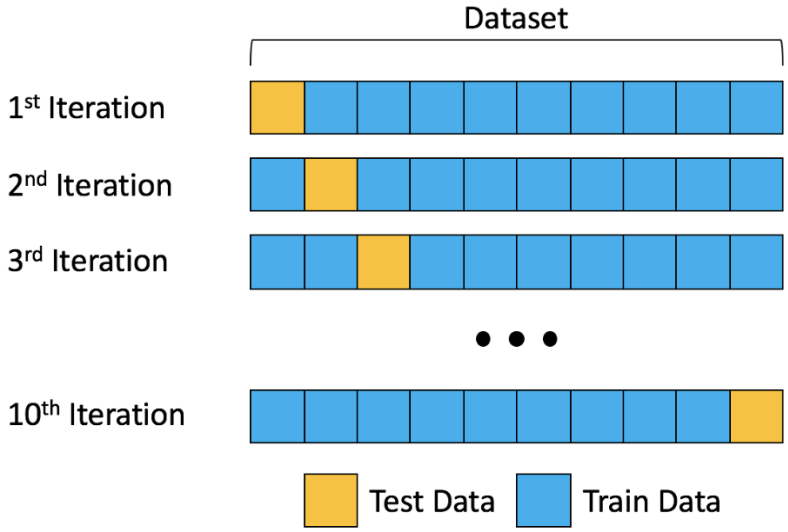
$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$ สมการที่ 2



ภาพที่ 30 แสดงแผนภาพความสัมพันธ์ของการตรวจสอบ

- การสุ่มเลือกข้อมูลแบบการตรวจสอบไขว้ (N-Fold Cross Validation)

การสุ่มเลือกข้อมูลแบบการตรวจสอบไขว้เป็นการนำข้อมูลมาแบ่งออกเป็นส่วน ๆ ตามจำนวน n ที่เลือก โดยในการทดลองนี้ได้เลือกค่า n เท่ากับ 10 หมายความว่า จะทำการแบ่งข้อมูลออกเป็น 10 ส่วนและทำการนำข้อมูล 9 ส่วนไว้ฝึกฝน อีกส่วนหนึ่งที่เหลือไว้ทดสอบ แล้วก็ทำการเปลี่ยนชุดทดสอบสลับไปเรื่อย ๆ จนครบทุกส่วน โดยสุดท้ายจะนำผลลัพธ์ที่ได้มาเฉลี่ยกัน



ภาพที่ 31 แสดงวิธีการสุ่มเลือกข้อมูลแบบการตรวจสอบไขว้ 10 ครั้ง

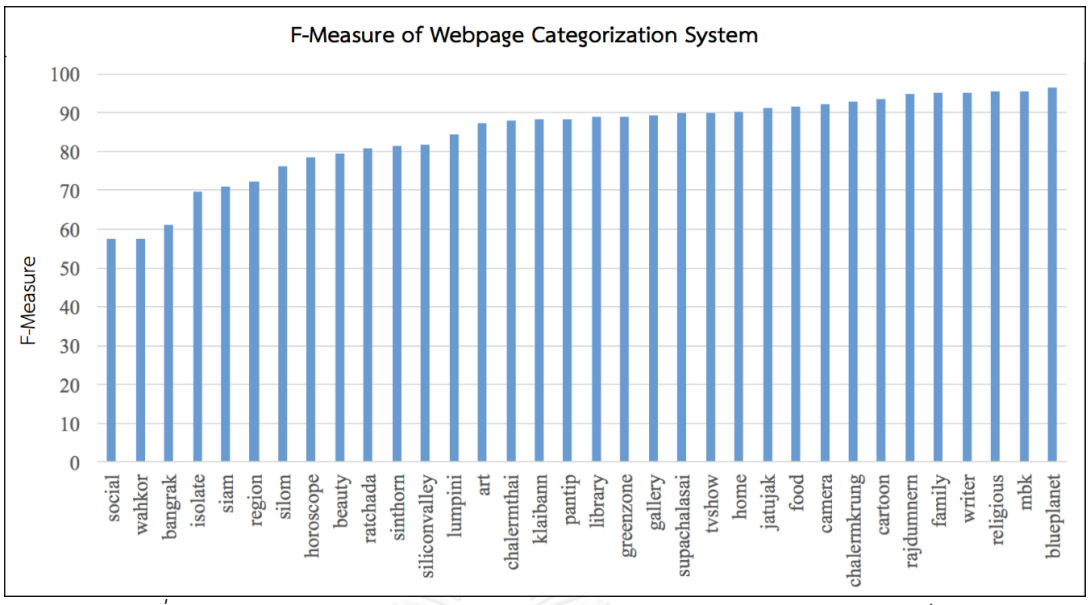
4.3.4 ผลการทดลอง

ตารางที่ 1 แสดงผลการเปรียบเทียบเวลาในการฝึกฝนข้อมูลระหว่างระบบสกัดคำสำคัญ, อัลกอริทึมนาอีฟเบย์, และอัลกอริทึมเวิร์ดทูเวค จากผลการทดสอบพบว่าอัลกอริทึมเวิร์ดทูเวคใช้เวลาในการเทรนนานที่สุด รองลงมาคืออัลกอริทึมนาอีฟเบย์ โดยระบบสกัดคำสำคัญใช้เวลาในการฝึกฝนน้อยที่สุด เนื่องจากมีขั้นตอนการทำงานที่ไม่ซับซ้อน โดยที่อัลกอริทึมนาอีฟเบย์จะมีขั้นตอนการฝึกฝนที่มากกว่า แต่ก็ส่งผลถึงระยะเวลาในการฝึกฝนข้อมูลไม่มาก ต่างกับอัลกอลิทึมเวิร์ดทูเวคที่ใช้ระยะเวลาในการฝึกฝนข้อมูลเยอะมากเพราะต้องทำการแปลงเปลี่ยนแปลงรูปแบบข้อมูลด้วย

ตารางที่ 1 แสดงการเปรียบเทียบระยะเวลาในการฝึกฝนข้อมูลระหว่างระบบสกัดคำสำคัญ, อัลกอริทึมนาอีฟเบย์, และอัลกอริทึมเวิร์ดทูเวค

อัลกอริทึม	ระยะเวลาในการฝึกฝน (วินาที)
ระบบสกัดคำสำคัญ	30.18
อัลกอริทึมนาอีฟเบย์	33.62
อัลกอริทึมเวิร์ดทูเวค	61,965.84

กราฟในภาพที่ 32 แสดงมาตรวัดค่าประสิทธิภาพโดยรวมของระบบจัดประเภทเว็บเพจ โดยแกนนอนแทนหมวดหมู่ทั้ง 34 หมวดหมู่ ส่วนในแกนตั้งแทนค่ามาตรวัดค่าประสิทธิภาพโดยรามของระบบจัดประเภทเว็บเพจ จากกราฟพบว่า หมวดหมู่ “social” เป็นหมวดหมู่ที่เกี่ยวกับกฎหมาย พื้นบ้าน ปัญหาสังคม ปัญหาชีวิต เศรษฐกิจ และการคุ้มครองผู้บริโภค เป็นหมวดหมู่ที่ได้ค่ามาตรวัดค่าประสิทธิภาพโดยรวมจากการจัดประเภทเว็บเพจน้อยที่สุด โดยมีค่ามาตรวัดค่าประสิทธิภาพโดยรวมเพียง 0.58 เนื่องจากเว็บเพจที่อยู่ในหมวดหมู่ social หรือศาลาประชาคมเป็นหมวดหมู่ที่มีการพูดคุยกันถึงเรื่องทั่วไปเป็นส่วนใหญ่จึงส่งผลให้คำสำคัญที่สกัดได้จากเว็บเพจที่อยู่ในหมวดหมู่นี้จะถูกนิยามว่าเป็นคำทั่วไป เพราะว่ามีการใช้คำเหล่านี้ในหมวดหมู่อื่น ๆ อีกหลายหมวดหมู่เช่นเดียวกัน ส่วนอีกหมวดหมู่ที่มีค่ามาตรวัดประสิทธิภาพโดยรวมใกล้เคียงกับหมวดหมู่ social คือหมวดหมู่ “wahkor” หรือหว้ากอ เป็นหมวดหมู่ที่เกี่ยวข้องกับคณิตศาสตร์ วิทยาศาสตร์ ประวัติศาสตร์ ภูมิศาสตร์ และเทคโนโลยี โดยหมวดหมู่ที่มีค่ามาตรวัดประสิทธิภาพโดยรวมสูงที่สุด คือ หมวดหมู่ “blueplanet” เป็นหมวดหมู่ที่เกี่ยวกับการท่องเที่ยว โดยมีค่ามาตรวัดค่าประสิทธิภาพโดยรวมมากถึง 0.96 เหตุผลที่เว็บเพจที่อยู่ในหมวดหมู่ “blueplanet” มีค่ามาตรวัดค่าประสิทธิภาพโดยรวมสูงเพราะว่าคำสำคัญที่ใช้ในหมวดหมู่นี้เป็นคำที่สามารถบ่งบอกได้ถึงหมวดหมู่นี้จริง ๆ โดยที่หมวดหมู่อื่นไม่ค่อยนิยมนำไปใช้กัน เช่น เที่ยว ทริป สนามบิน travel trip airport เป็นต้น และหมวดหมู่ที่มีค่ามาตรวัดค่าประสิทธิภาพโดยรวมรองลงมา คือ หมวดหมู่ “mbk” เป็นหมวดหมู่ที่กล่าวถึงโทรศัพท์มือถือ แทปเล็ต และผู้ให้บริการโทรศัพท์มือถือ โดยมีค่ามาตรวัดค่าประสิทธิภาพโดยรวมอยู่ที่ 0.95 และคำสำคัญที่นิยมใช้ในหมวดหมู่นี้ คือ ais dtac true เอไอเอส ดีเทค ทรู เป็นต้น โดยคำสำคัญในหมวดหมู่นี้ที่มีความโดดเด่นเป็นชื่อของผู้ให้บริการโทรศัพท์มือถือ โดยค่ามาตรวัดค่าประสิทธิภาพโดยรวมของระบบจัดประเภทเว็บเพจมีค่าเท่ากับ 0.85

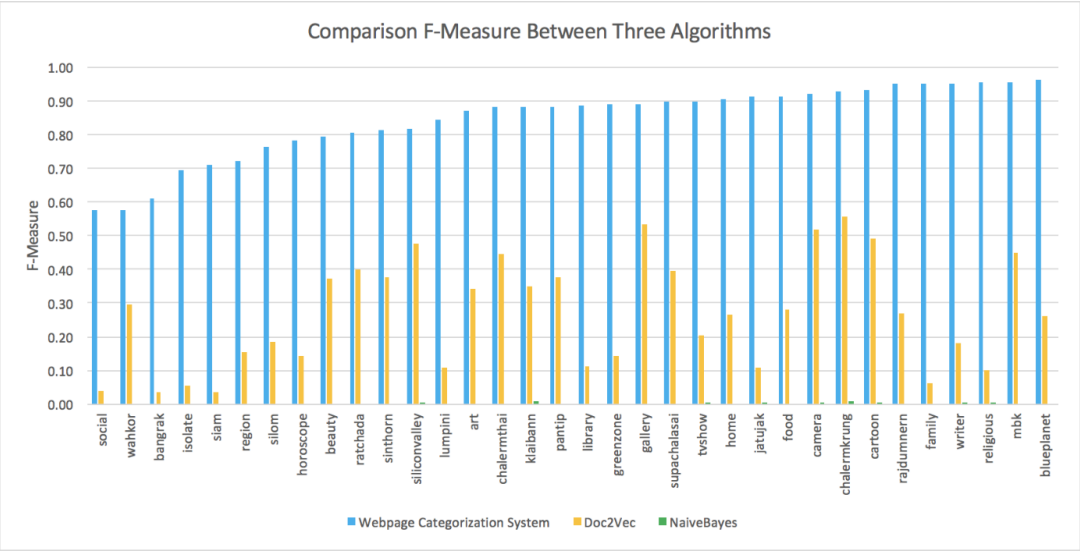


ภาพที่ 32 แสดงค่ามาตรวัดค่าประสิทธิภาพโดยรวมของระบบจัดประเภทเว็บเพจ

Category	Words
Travel	รอยั้วเยี้ย, tripadvisor, เอเจ้น, โฮสเทล, แอร์ไลน์, ...
Food	คาเฟ้, หิวโหย, รีวิววววว, อร่อยยยยย, chef, buffet, ...

ภาพที่ 33 แสดงตัวอย่างคำในหมวดหมู่ท่องเที่ยวและอาหาร

กราฟในภาพที่ 34 แสดงการเปรียบเทียบค่ามาตรวัดประสิทธิภาพระหว่างระบบจัดประเภทเว็บเพจ อัลกอริทึมนาอีฟเบย์ และอัลกอริทึมเวิร์ดทูเวค โดยอัลกอริทึมนาอีฟเบย์ได้ใช้วิธีของเบอร์นูลี่ (Bernoulli) ในการถ่วงน้ำหนักคำ โดยวิธีจะดูเพียงว่ามีคำ ๆ นี้อยู่ในเอกสารที่สนใจหรือไม่ ส่วนอัลกอริทึมเวิร์ดทูเวคได้ใช้คำตอบที่ได้จากการเปรียบเทียบเวคเตอร์ของเอกสารที่มีความคล้ายกับเวคเตอร์ของเอกสารทดสอบมากที่สุดเพียงเวคเตอร์เดียว จากผลการทดลองพบว่า อัลกอริทึมเวิร์ดทูเวคให้ค่าประสิทธิภาพโดยรวมของการจัดประเภทน้อยมาก เนื่องจากใช้คำตอบที่ได้การหาค่าความคล้ายมากที่สุดเพียงคำตอบเดียว ซึ่งสามารถปรับปรุงโดยใช้วิธีหาจำนวนเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) มาช่วยในการเพิ่มประสิทธิภาพได้ ในส่วนของอัลกอริทึมนาอีฟเบย์ได้ใช้ทฤษฎีของเบอร์นูลี่ (Bernoulli) ในการถ่วงน้ำหนักคำ โดยวิธีจะดูเพียงว่ามีคำ ๆ นี้อยู่ในเอกสารที่สนใจหรือไม่ โดยไม่ได้สนใจค่าความถี่ของคำเลยซึ่งอาจส่งผลถึงค่าประสิทธิภาพโดยรวมได้



ภาพที่ 34 แสดงการเปรียบเทียบค่ามาตรวัดค่าประสิทธิภาพโดยรวม

ตารางที่ 2 แสดงผลการเปรียบเทียบเวลาในการทดสอบข้อมูลระหว่างระบบจัดประเภทเว็บเพจ, อัลกอริทึมนาอีฟเบย์, และอัลกอริทึมเวิร์ดทูเวค จากผลการทดสอบพบว่าอัลกอริทึมเวิร์ดนาอีฟเบย์ใช้เวลาในการทดสอบนานที่สุด รองลงมาคืออัลกอริทึมเวิร์ดทูเวค เนื่องจากอัลกอริทึมนาอีฟเบย์และเวิร์ดทูเวคมีวิธีการทำงานที่ซับซ้อนกว่าระบบจัดประเภทเว็บเพจ

ตารางที่ 2 ระยะเวลาในการทดสอบข้อมูลระหว่างระบบสกัดคำสำคัญ, อัลกอริทึมนาอีฟเบย์, และอัลกอริทึมเวิร์ดทูเวค

อัลกอริทึม	**ระยะเวลาในการทดสอบ (วินาที)**
ระบบสกัดคำสำคัญ	12.32
อัลกอริทึมนาอีฟเบย์	323.62
อัลกอริทึมเวิร์ดทูเวค	61.32

4.4 การทดสอบระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

4.4.1 สมมติฐานการทดลอง

การทดลองนี้มีจุดประสงค์เพื่อทำการทดสอบกับข้อมูลชุดอื่นโดยเปรียบระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำกับอัลกอริทึมเบอร์นูลี่นาอีฟเบย์ อัลกอริทึมมัลติโนเมียลนาอีฟเบย์ และเวิร์ดทูเวคแบบพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุดด้วย

4.4.2 ข้อมูลที่นำมาใช้ในการทดสอบ

ข้อมูลที่นำมาใช้ในการทดสอบครั้งนี้มีแหล่งที่มาเดียวกับการทดสอบระบบจัดประเภทเว็บเพจ แต่เป็นการสุ่มข้อมูลคนละชุดมาทำการทดสอบ โดยหมวดหมู่ที่ใช้ในการทดสอบมาจากจำนวนห้องหรือหมวดหมู่ในเว็บพันทิปเป็นจำนวน 34 หมวดหมู่ และมีปริมาณเว็บเพจในแต่ละหมวดหมู่ประมาณ 1,000 เว็บเพจ โดยมีแต่เว็บเพจที่เป็นภาษาไทยเป็นจำนวนทั้งสิ้น 34,000 เว็บเพจ

4.4.3 มาตรวัดประสิทธิภาพ

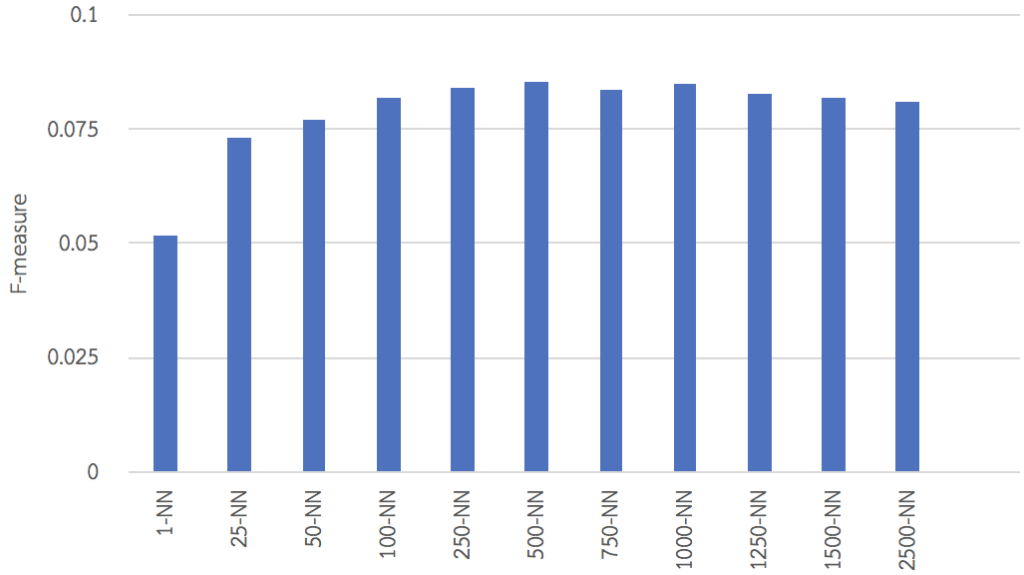
มาตรวัดประสิทธิภาพในการทดลองครั้งมี 2 แบบ คือ มาตรวัดความเร็วในการฝึกฝนและทดสอบระบบ และมาตรวัดค่าประสิทธิภาพโดยรวม ดังที่ได้อธิบายรายละเอียดในการทดสอบระบบจัดประเภทเว็บเพจ

4.4.4 ผลการทดสอบ

สำหรับระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำนั้นใช้ขั้นตอนการดำเนินการเดียวกับการทดสอบก่อนหน้า สำหรับอัลกอริทึมใหม่ที่เพิ่มเข้ามาในการทดสอบครั้งนี้ คือ อัลกอริทึมมัลติโนเมียลนาอีฟเบย์เป็นอัลกอริทึมที่ใช้วิธีการถ่วงน้ำหนักที่พิจารณาความถี่ของคำที่ปรากฏในเอกสารและจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ (TF-IDF : Term Frequency-Inverted Document Frequency) และอัลกอริทึมเวิร์ดทูเวคแบบพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุดใช้วิธีการเปลี่ยนเอกสารให้อยู่ในรูปของเวคเตอร์ จากนั้นจะทำการเปรียบเทียบความคล้ายระหว่างเวคเจตอร์แต่ละเวคเตอร์จากนั้นจะทำการหาจำนวนเวคเตอร์ที่มีความคล้ายมากที่สุดหนึ่งช่วง และทำการตอบหมวดหมู่ของเวคเตอร์ที่มีจำนวนมากที่สุดในช่วงนั้น

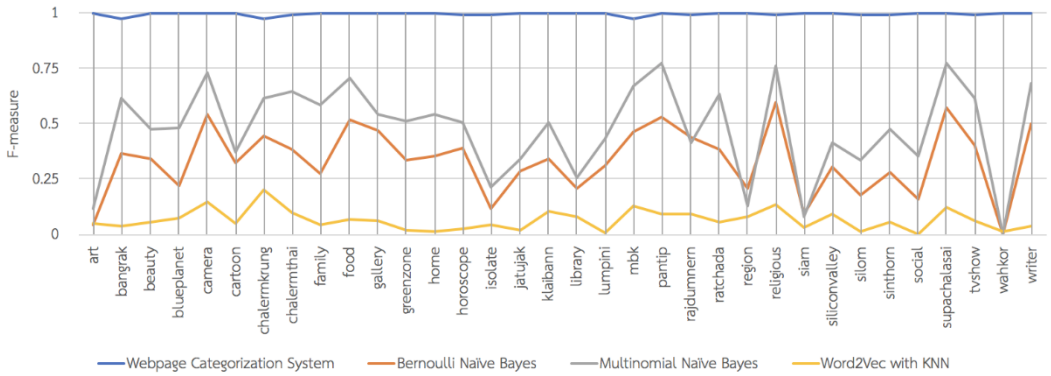
ภาพที่ 35 แสดงค่าประสิทธิภาพโดยรวมอัลกอริทึมเวิร์ดทูเวค โดยมีการปรับค่าจำนวนเพื่อนบ้านที่ใกล้ที่สุดเพื่อหาค่าจำนวนเพื่อนบ้านที่ใกล้ที่สุดที่ให้ค่าประสิทธิภาพโดยรวมของอัลกอริทึมสูงสุดมาใช้ในการเปรียบเทียบ จำนวนเพื่อนบ้านที่ใกล้ที่สุดที่ได้ทำการทดสอบ คือ 1 25 50 100 250 500 750 1000 1250 1500 และ 2500 จากกราฟจะเห็นว่า ที่จำนวนเพื่อนบ้านมีค่าเป็น 500 ให้ค่าประสิทธิภาพโดยรวมของระบบสูงที่สุด โดยมีค่าประสิทธิภาพโดยรวมเป็น 0.085 ซึ่งกราฟที่แสดงใน

ภาพได้ปรับสเกลแกนตั้งเป็น 0 ถึง 0.1 เพื่อให้เห็นถึงความแตกต่างชัดเจนยิ่งขึ้น โดยค่าประสิทธิภาพโดยรวมสูงสุดมีค่าเป็น 1



ภาพที่ 35 แสดงการเปรียบเทียบค่าประสิทธิภาพโดยรวมของอัลกอริทึมเวิร์ดทูเวค เมื่อทำการปรับจำนวนเพื่อนบ้านที่ใกล้ที่สุด

ภาพที่ 36 แสดงการเปรียบเทียบค่าประสิทธิภาพโดยรวมแยกแต่ละหมวดหมู่ระหว่างระบบจัดประเภทเว็บเพจที่พัฒนาขึ้นกับอัลกอริทึมอีก 3 อัลกอริทึม ได้แก่ อัลกอริทึมเบอร์นูลี่ย์นาอีฟเบย์ อัลกอริทึมมัลติโนเมียลนาอีฟเบย์ และอัลกอริทึมเวิร์ดทูเวคแบบพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุด 500 เพื่อนบ้าน จากภาพจะเห็นว่าระบบจัดประเภทเว็บเพจมีค่าประสิทธิภาพโดยรวมสูงที่สุดในทุก ๆ หมวดหมู่ รองลงมาคืออัลกอริทึมมัลติโนเมียลนาอีฟเบย์ อัลกอริทึมเบอร์นูลี่ย์นาอีฟเบย์ และอัลกอริทึมเวิร์ดทูเวคแบบพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุด 500 เพื่อนบ้าน ตามลำดับ



ภาพที่ 36 แสดงการเปรียบเทียบค่าประสิทธิภาพโดยรวมแยกตามหมวดหมู่

เนื่องจากค่าประสิทธิภาพโดยรวมของระบบจัดประเภทเว็บเพจสูงกว่าอัลกอริทึมอื่น ในขณะที่มีวิธีการทำงานคล้ายกับอัลกอริทึมนาอีฟเบย์ ภาพที่ 37 แสดงให้เห็นถึงความแตกต่างระหว่างระบบจัดประเภทเว็บเพจและอัลกอริทึมเบอร์นูลี่ย์นาอีฟเบย์กับอัลกอริทึมมัลติโนเมียลนาอีฟเบย์ การ

เปรียบเทียบนี้มีตัวชี้วัด 2 อย่าง คือ วิธีการถ่วงน้ำหนักขณะฝึกฝนข้อมูลและวิธีการถ่วงน้ำหนักขณะทดสอบข้อมูล อัลกอริทึมเบอร์นูลี่ย์นาอีฟเบย์จะใช้วิธีการถ่วงน้ำหนักแบบพิจารณาว่ามีหรือไม่มีทั้งในขณะฝึกฝนและทดสอบข้อมูล ส่วนอัลกอริทึมมัลติโนเมียลนาอีฟเบย์ใช้วิธีการถ่วงน้ำหนักโดยพิจารณาความถี่ของคำที่ปรากฏในเอกสารและจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ในขณะฝึกฝนข้อมูล แต่ในขณะทดสอบข้อมูลใช้วิธีการถ่วงน้ำหนักแบบพิจารณาแค่ความถี่ของคำ ในขณะที่ระบบจัดประเภทเว็บเพจใช้วิธีการถ่วงน้ำหนักแบบมีหรือไม่มีในขณะฝึกฝนข้อมูลเหมือนกับอัลกอริทึมเบอร์นูลียร์นาอีฟเบย์ แต่ใช้วิธีการถ่วงน้ำหนักแบบพิจารณาความถี่ของคำในขณะทดสอบเหมือนกับอัลกอริทึมมัลติโนเมียลนาอีฟเบย์ สำหรับอัลกอริทึมเบอร์นูลี่ย์นาอีฟเบย์ที่ใช้วิธีการถ่วงน้ำหนักแบบมีหรือไม่มีทั้งในขณะฝึกฝนและทดสอบนั้นมีความยืดหยุ่นต่ำเกินไป การไม่พิจารณาความถี่ของคำที่ปรากฏส่งผลให้คำที่ไม่มีมีความหมายเฉพาะเจาะจงในหมวดหมู่นั้นมีค่าถ่วงน้ำหนักเท่ากับคำที่มีความหมายเฉพาะเจาะจง ในส่วนของอัลกอริทึมมัลติโนเมียลนาอีฟเบย์นั้นใช้วิธีการถ่วงน้ำหนักแบบพิจารณาความถี่ของคำที่ปรากฏในเอกสารและจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ในขณะฝึกฝนข้อมูล เมื่อหมวดหมู่ที่ทำการจัดมีความหลากหลายสูง กล่าวคือหมวดหมู่นั้นประกอบไปด้วยหมวดหมู่ย่อย ๆ อีกหลายหมวดหมู่ ส่งผลให้เมื่อทำการถ่วงน้ำหนักด้วยจำนวนเอกสารที่มีคำนั้นปรากฏอยู่จะทำให้ค่าถ่วงน้ำหนักของแต่ละคำมีค่าน้อยเพราะความหลากหลายของหมวดหมู่ เมื่อนำความถี่ของคำที่ปรากฏขณะทดสอบมาถ่วงน้ำหนักกับค่าที่ได้จากการฝึกฝนแล้วทำให้ค่าถ่วงน้ำหนักโดยรวมมีค่าน้อยกว่าที่ควรจะเป็น จากปัญหาดังกล่าวจึงส่งผลให้อัลกอริทึมเบอร์นูลี่นาอีฟเบย์และอัลกอริทึมมัลติโนเมียลนาอีฟเบย์มีค่าประสิทธิภาพโดยรวมน้อย

	Training Phase	Testing Phase
Bernoulli Naïve Bayes	Binary	Binary
Webpage Categorization System	Binary	Term Frequency
Multinomial Naïve Bayes	Term Frequency-Inverted Document Frequency	Term Frequency

ภาพที่ 37 แสดงตารางเปรียบเทียบวิธีการถ่วงน้ำหนักของระบบจัดประเภทเว็บเพจและอัลกอริทึมนาอีฟเบย์

สำหรับอัลกอริทึมเวิร์ดทูเวคเมื่อพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุดด้วยมีค่าประสิทธิภาพโดยรวมน้อยที่สุด เนื่องจากการทำงานของอัลกอริทึมเวิร์ดทูเวคจะทำการเปลี่ยนเอกสารให้อยู่ในรูปแบบของเวคเตอร์ ด้วยการแยกประโยคแต่ละประโยคในเอกสารมาอยู่ในรูปของเวคเตอร์ แล้วให้เวคเตอร์เหล่านั้นเป็นตัวแทนของเอกสารดังกล่าว ซึ่งแบ่งประโยคในการเขียนของภาษาไทยในอินเทอร์เน็ตนั้นไม่มีหลักการตายตัวที่จะมากำหนดว่า ต้องแบ่งประโยคแบบนี้เท่านั้น หรือต่อให้มีหลักการมากำหนด ผู้ใช้บริการทั่วไปส่วนใหญ่ก็ไม่ได้คำนึงถึงการเขียนให้รูปประโยคถูกต้องตามหลัก ส่งผลให้เมื่อทำการเปรียบเทียบเวคเตอร์แล้วค่าความคล้ายของแต่ละเวคเตอร์จึงมีค่าน้อยมาก

หลังจากการเปรียบเทียบค่าประสิทธิภาพโดยรวมของระบบได้แล้ว ภาพที่ 38 แสดงการเปรียบเทียบระยะเวลาในขณะฝึกฝนและทดสอบของระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำกับอัลกอริทึมทั้งสาม จากภาพจะเห็นว่าระบบที่พัฒนาขึ้นนอกจากมีค่าประสิทธิภาพโดยรวมสูงกว่าอัลกอริทึมอื่นแล้ว ยังคงใช้เวลาในการทดสอบและฝึกฝนน้อยกว่าอัลกอริทึมอื่นอีกด้วย เนื่องจากการใช้วิธีการวิเคราะห์อย่างง่ายโดยพิจารณาแค่ค่าสถิติเท่านั้น สำหรับอัลกอริทึมเวิร์ดทูเวคที่ใช้เวลาในการทดสอบนานกว่าการทดสอบที่แล้ว เนื่องจากต้องพิจารณาจำนวนเพื่อนบ้านที่ใกล้ที่สุดเพิ่มด้วยส่งผลให้ใช้เวลาในการทดสอบเพิ่มขึ้นมากกว่าการทดสอบที่ผ่านมา

	Training Time (s)	Testing Time (s)
Bernoulli Naïve Bayes	286	104
Multinomial Naïve Bayes	264	257
Word2Vec with 500-NN	4826	1200
Webpage Categorization System	93	4

ภาพที่ 38 แสดงการเปรียบเทียบระยะเวลาในขณะฝึกฝนและทดสอบของระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำกับอัลกอริทึมทั้งสาม

ภาพที่ 39 แสดงระยะเวลาในการประมวลผลของแต่ละโมดูลของระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ

Module	Time (s)
Webpage Crawler with JS	3.05
Webpage Crawler	$1.4 * 10^{-1}$
HTML Tag Removal	$6 * 10^{-2}$
Punctuation Removal	$2 * 10^{-4}$
Word Segmentation	$5.6 * 10^{-5}$
Common Word Removal	$3.1 * 10^{-6}$
Word Counting	$1.8 * 10^{-5}$
Webpage Categorization	$6 * 10^{-6}$

ภาพที่ 39 แสดงระยะเวลาในการประมวลผลของแต่ละโมดูล

4.5 การทดสอบกับส่วนต่อประสานโปรแกรมประยุกต์ที่มีอยู่ในท้องตลาด

4.5.1 สมมติฐานการทดลอง

การทดลองนี้มีจุดประสงค์เพื่อเปรียบเทียบความสามารถของระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำกับส่วนต่อประสานโปรแกรมประยุกต์ที่มีอยู่ในท้องตลาด เนื่องจากส่วนต่อประสานโปรแกรมประยุกต์ที่มีอยู่ในท้องตลาดไม่สามารถจัดประเภทเว็บเพจที่มีเนื้อหาเป็นภาษาไทยได้ ดังนั้นเพื่อความยุติธรรม ในการทดสอบนี้จึงจะทำการเปรียบเทียบเฉพาะเว็บเพจที่มีเนื้อหาเป็นภาษาอังกฤษเท่านั้น

4.5.2 ข้อมูลที่นำมาใช้ในการทดสอบ

ข้อมูลที่นำมาใช้ในการทดสอบครั้งนี้เป็นเว็บเพจของสำนักข่าวต่าง ๆ แบ่งออกเป็น 2 ประเภท คือ เว็บเพจของสำนักข่าวสากลและเว็บเพจของสำนักข่าวท้องถิ่นในประเทศไทยที่มีเนื้อหาภายในเว็บเพจเป็นภาษาอังกฤษ โดยเว็บเพจของสำนักข่าวสากลนั้นทางผู้วิจัยได้เลือกสำนักข่าวซีเอ็นเอ็น (CNN) และสำนักข่าวรอยเตอร์ส (Reuters) ส่วนเว็บเพจของสำนักข่าวท้องถิ่นในประเทศไทยที่มีเนื้อหาภายในเว็บเพจเป็นภาษาอังกฤษนั้นได้เลือกสำนักข่าวบางกอกโพส (Bangkok Post) และสำนักข่าวไทยแลนด์เมดิคอลนิวส์ (Thailand Medical News) มาใช้ในการทดสอบครั้งนี้ โดยมีปริมาณเว็บเพจที่ใช้ในการทดสอบทั้งสิ้น 400 เว็บเพจ ทำการจัดประเภทออกเป็น 4 หมวดหมู่ คือ หมวดหมู่การเงิน หมวดหมู่การเมือง หมวดหมู่เทคโนโลยี และหมวดหมู่สุขภาพ

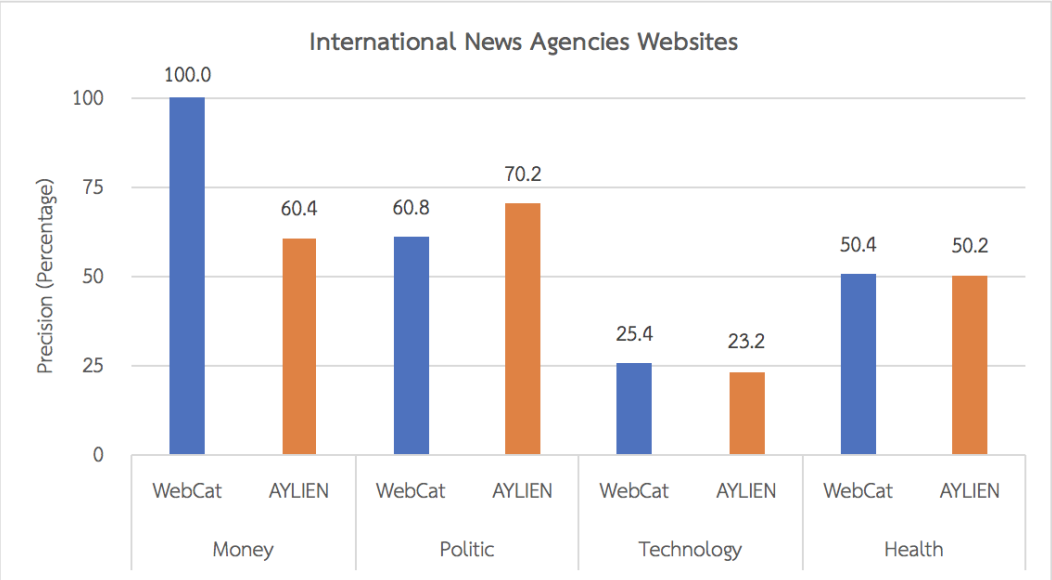
4.5.3 มาตรวัดประสิทธิภาพ

ในการทดสอบนี้จะทำการเปรียบเทียบมาตรวัดประสิทธิภาพ 2 แบบ คือ มาตรวัดความเร็วในการจัดประเภทเว็บเพจ และค่าความแม่นยำในการจัดประเภท ดังที่ได้อธิบายรายละเอียดในการทดสอบระบบจัดประเภทเว็บเพจ

4.5.4 ผลการทดสอบ

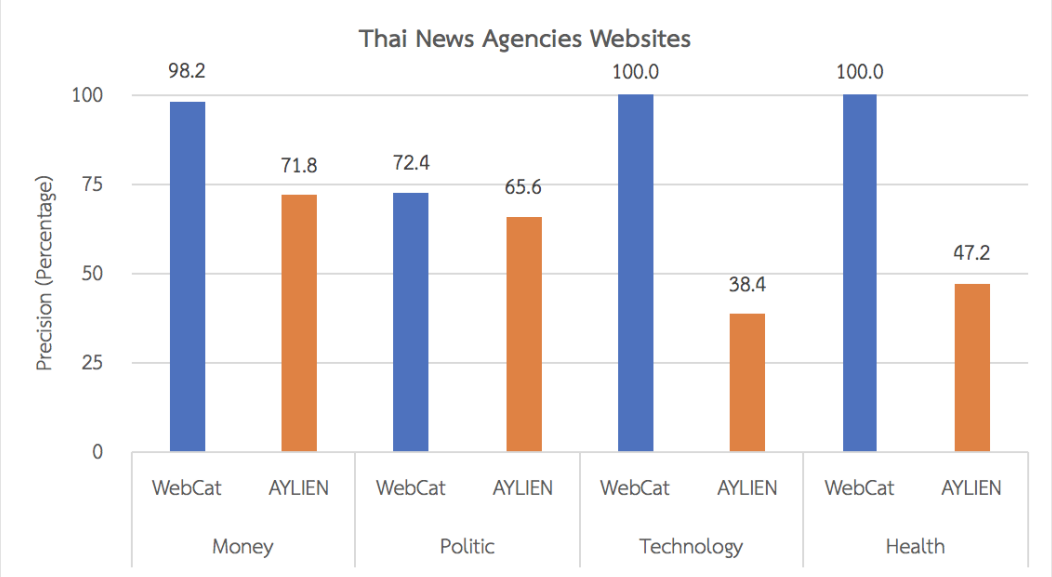
สำหรับส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดที่นำมาเปรียบเทียบด้วยนั้นเป็นส่วนต่อประสานโปรแกรมประยุกต์ที่ใช้วิธีการจัดประเภทโดยคำนึงถึงเนื้อหาภายในเว็บเพจเช่นเดียวกับระบบที่พัฒนาขึ้น การทดสอบครั้งนี้ได้แบ่งข้อมูลเป็น 2 ส่วน คือ เว็บเพจสากลและเว็บเพจท้องถิ่นเพื่อป้องกันการเก็บข้อมูลการจัดประเภทของส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาด

ภาพที่ 40 แสดงการเปรียบเทียบค่าความแม่นยำของการจัดประเภทเว็บเพจจากสำนักข่าวสากล โดยระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำมีค่าความแม่นยำในการจัดประเภทมากกว่าส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดในหมวดหมู่การเงิน หมวดหมู่เทคโนโลยี และหมวดหมู่สุขภาพ ในขณะที่ส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดมีค่าความแม่นยำในการจัดประเภทเว็บเพจในหมวดหมู่การเมืองมากกว่า



ภาพที่ 40 แสดงการเปรียบเทียบค่าความแม่นยำของการจัดประเภทเว็บเพจจากสำนักข่าวสากล

ภาพที่ 41 แสดงการเปรียบเทียบค่าความแม่นยำของการจัดประเภทเว็บเพจจากสำนักข่าวท้องถิ่นในประเทศไทยที่มีเนื้อหาเป็นภาษาอังกฤษ โดยระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำมีค่าความแม่นยำในการจัดประเภทมากกว่าส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดในหมวดหมู่การเงิน หมวดหมู่เทคโนโลยี และหมวดหมู่สุขภาพ โดยค่าความแม่นยำในการจัดประเภทของระบบที่พัฒนาขึ้นมีค่ามากกว่าส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดทุกหมวดหมู่



ภาพที่ 41 แสดงการเปรียบเทียบค่าความแม่นยำของการจัดประเภทเว็บเพจจากสำนักข่าวท้องถิ่นในประเทศไทยที่มีเนื้อหาเป็นภาษาอังกฤษ

การทดสอบมาตรวัดระยะเวลาที่ใช้ในการจัดประเภทของส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาด โดยทั้งระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำและส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดจะทำการรับยูอาร์แอลเป็นข้อมูลนำเข้าทั้งคู่ และทำการวิเคราะห์เพื่อจัดหมวดหมู่ให้กับยูอาร์แอลดังกล่าวตามขั้นตอนของตน จากการทดสอบวัดระยะเวลาที่ใช้ในการจัดประเภทเฉลี่ยต่อ 1 ยูอาร์แอลนั้น ส่วนต่อประสานโปรแกรมประยุกต์ในท้องตลาดใช้เวลาในการจัดประเภทนานกว่าระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำมากถึงร้อยละ 30

บทที่ 5
บทสรุปของงานวิจัยและอภิปรายผลการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำ เป็นระบบที่ใช้สำหรับจัดประเภทเว็บเพจ เป็นระบบที่อาศัยพจนานุกรมคำที่ได้จากการเรียนรู้ในระบบสกัดคำสำคัญซึ่งเป็นระบบย่อยในระบบจัดประเภทเว็บเพจมาช่วยเติมเต็มระบบ หลักการทำงานของระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำจะประกอบไปด้วยระบบย่อย ๆ 2 ระบบ คือ ระบบสกัดคำสำคัญ และระบบจัดประเภทเว็บเพจ

ระบบสกัดคำสำคัญเป็นระบบที่ถือเป็นระบบสำคัญในการดำเนินงานในครั้งนี้ เนื่องจากระบบจัดประเภทเว็บเพจจำเป็นต้องพึ่งพาพจนานุกรมที่ได้จากการเรียนรู้ของระบบสกัดคำสำคัญ โดยระบบสกัดคำสำคัญเป็นระบบที่เป็นกระบวนการเรียนรู้แบบมีผู้สอน คือ จำเป็นต้องมีข้อมูลที่ทราบผลเฉลยแล้วในการสร้างพจนานุกรมคำขึ้นมา โดยมีหลักการทำงานคือการร้องขอข้อมูลหน้าเว็บมาเพื่อสกัดคำสำคัญที่อยู่เนื้อหาของหน้าเว็บเพจนั้นแล้วนำมาใช้ในการจัดทำพจนานุกรมคำ โดยหลังจากการสกัดคำสำคัญเพื่อจัดทำพจนานุกรมคำเสร็จสิ้นแล้ว ทำให้สามารถสร้างรายการคำทั่วไปที่เกิดจากการใช้งานจริงของผู้ใช้บริการได้ โดยการตรวจสอบในพจนานุกรมว่า มีคำใดที่ไปปรากฏอยู่ในหมวดหมู่มากกว่า 18 หมวดหมู่

ระบบจัดประเภทเว็บเพจเป็นระบบที่ใช้ในการจัดประเภทเว็บเพจด้วยการรับยูอาร์แอลที่เป็นที่อยู่ของเว็บเพจมาทำการร้องขอข้อมูลหน้าเว็บเพจและทำการตรวจสอบคำสำคัญที่พบในหน้าเว็บเพจนั้นกับพจนานุกรมคำที่ได้จัดเตรียมไว้ในระบบสกัดคำสำคัญ เพื่อบอกหมวดหมู่นั้นโดยดูจากปริมาณคำสำคัญที่เกี่ยวกับหมวดหมู่ที่เป็นคำตอบมากที่สุด โดยระบบจัดประเภทเว็บเพจมีค่ามาตรวัดค่าประสิทธิภาพโดยรวมของระบบสูงถึง 0.99 ใช้เวลาในการฝึกฝนเรียนรู้ข้อมูลทดสอบน้อยกว่าวิธีอื่นและใช้เวลาในการวิเคราะห์เพื่อจัดหมวดหมู่น้อยกว่าอัลกอริทึมนาอีฟเบย์และเวิร์ดทูเวค

ระบบจัดประเภทเว็บเพจแบบอัตโนมัติที่มีพื้นฐานมาจากสถิติคำมีความสามารถในการแบ่งกระจายงานให้ทำพร้อมกันหลายๆ เครื่องได้ ทำให้สามารถเพิ่มความรวดเร็วในการประมวลได้มากขึ้นอีกด้วย

5.2 ข้อจำกัดของระบบ

5.2.1 ระบบจำเป็นที่จะต้องทำการเรียนรู้เพื่อสร้างพจนานุกรมก่อน

5.2.2 การสร้างพจนานุกรมจำเป็นต้องอาศัยปริมาณข้อมูลฝึกฝนที่มีปริมาณมากและมีประสิทธิภาพ

5.2.3 ถ้าโมดูลร้องขอข้อมูลหน้าเว็บเพจต้องการข้อมูลครบถ้วนจำเป็นต้องใช้เวลาเพิ่มมากขึ้น

5.2.4 การสร้างพจนานุกรมด้วยข้อมูลฝึกฝนประเภทเดียวไม่สามารถใช้ได้กับข้อมูลรูปแบบอื่นได้

5.3 ข้อเสนอแนะ

5.3.1 การเพิ่มประสิทธิภาพในการจัดประเภทเว็บเพจ

แม้ว่าระบบจัดประเภทเว็บเพจจะมีค่าประสิทธิภาพโดยรวมสูง แต่ก็มีเว็บเพจบางหมวดหมู่ที่ยังคงไม่สามารถจัดประเภทได้อย่างเด็ดขาด เนื่องมาจากหมวดหมู่ดังกล่าวเป็นการพูดถึงเรื่องทั่ว ๆ ไป ทำให้ไม่สามารถสกัดเนื้อหาเพื่อหาคำสำคัญที่ใช้ในการระบุหมวดหมู่ได้

จากการพิจารณาหมวดหมู่ดังกล่าวพบว่า สามารถแก้ไขปัญหาได้โดยการนิยามหมวดหมู่เหล่านั้นให้เป็นหมวดหมู่ทั่วไป โดยการดูความน่าจะเป็นที่จะถูกระบุว่าเป็นหมวดหมู่ หากมีค่าความน่าจะเป็นต่ำกว่าค่าขีดจำกัดก็สามารถระบุให้เว็บเพจนั้นอยู่ในหมวดหมู่ทั่วไปได้เลย

การเปลี่ยนแปลงดังกล่าวจะสามารถช่วยให้การจัดประเภทเว็บเพจสำหรับเว็บเพจที่กล่าวถึงเนื้อหาทั่วไปหรือเว็บเพจที่มีเนื้อหาน้อยไปจัดอยู่ในหมวดหมู่ทั่วไปได้ โดยหมวดหมู่ที่กล่าวถึงเนื้อหาทั่วไปก็จะถูกรวมเป็นหมวดหมู่ทั่วไป

5.3.2 การเพิ่มความเร็วในการจัดประเภทเว็บเพจ

การจัดประเภทเว็บเพจใช้วิธีการค้นหาคำในพจนานุกรมแม้ว่าจะใช้เวลาต่อคำเป็น O(1) แต่ว่าต้องทำการตรวจสอบจนครบทุกคำ และเมื่อทำการตรวจสอบคำเสร็จแล้วก็ต้องทำการคำนวณหมวดหมู่อีกรอบ ซึ่งเป็นส่วนที่ใช้เวลาในการประมวลผลนาน

ปัญหาดังกล่าวสามารถแก้ได้ด้วยการทำรายการที่เก็บหมวดหมู่และความถี่เตรียมเอาไว้แล้วทำการเพิ่มความถี่ของหมวดหมู่หลังจากที่ทำการตรวจสอบคำได้เลย โดยจะสามารถลดเวลาในส่วนของการคำนวณความถี่ของหมวดหมู่ไปได้จนเหลือเพียงแค่เวลาในการเรียงลำดับหมวดหมู่เท่านั้น

รายการอ้างอิง

1. Worldometers. *Internet traffic in 1 second*. [cited 2017 April, 7]; Available from: http://www.internetlivestats.com/one-second/#traffic-band.
2. SimilarWeb. *Website Analysis*. [cited 2016 6 June]; Available from: https://www.similarweb.com/.
3. Bluecoat. *WebPulse Site Review*. Available from: https://sitereview.bluecoat.com/sitereview.jsp.
4. IBM. *Natural Language Classifier*. Available from: https://www.ibm.com/watson/alchemy-api.html.
5. AYLIEN. *Text Analysis API*. Available from: http://aylien.com/.
6. Chanakitkarnchok, A., K.N. Nakorn, and K. Rojviboolchai. *Autonomous website categorization with pre-defined dictionary*. in *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 2016.
7. Chanakitkarnchok, A., K.N. Nakorn, and K. Rojviboolchai, *Automatic Keyword Extraction System for Thai Website Categorization System*, in *2017 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. 2017. p. 1-6.
8. kstep, *HTTP Status Codes Cheat Sheet*.
9. P., W., *PyThaiNLP*. Github.
10. Charoenpornsawat, P., *Software: SWATH - Thai Word Segmentation*.
11. inspica, *Text Analysis API*.
12. PACHARAWONGSAKDA, E., โมเดล *Naive Bayes* และการแปลความหมาย. 2014.
13. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.
14. Baykan, E., et al. *Purely url-based topic classification*. in *Proceedings of the 18th international conference on World wide web*. 2009. ACM.

15. Lewis, D.D. and M. Ringuette. *A comparison of two learning algorithms for text categorization*. in *Third annual symposium on document analysis and information retrieval*. 1994.

16. Mahmood, K., et al. *Semantic based highly accurate autonomous decentralized URL classification system for Web filtering*. in *Autonomous Decentralized Systems (ISADS), 2015 IEEE Twelfth International Symposium on*. 2015. IEEE.

17. MyVocabulary. *VOCABULARY WORD LIST*. June 6, 2017]; Available from: https://myvocabulary.com/word-list/.

18. Corpus, O. *The OEC: Facts about the language*. February 2, 2017; Available from: https://web.archive.org/web/20111226085859/http://oxforddictionaries.com/words/the-oec-facts-about-the-language.

ภาคผนวก

ประวัติผู้เขียนวิทยานิพนธ์

นาย อัษฎาวุธ ชนะกิจการโชค ผู้เขียนวิทยานิพนธ์ เกิดเมื่อวันที่ 8 ธันวาคม พ.ศ. 2535 สำเร็จการศึกษาระดับปริญญาวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อ พ.ศ. 2558 ปัจจุบันกำลังศึกษาในหลักสูตรวิศวกรรมศาสตร์บัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย โดยได้รับทุนอุดหนุนการศึกษาอัจฉริยะคืนรัง จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และเป็นหนึ่งในทีมวิจัยในโครงการพัฒนาระบบวิเคราะห์ข้อมูลเครือข่ายโทรศัพท์เคลื่อนที่เพื่อการแบ่งกลุ่มผู้ใช้บริการกับบริษัทแอดวานซ์ อินโฟร์เซอร์วิส