ระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา

นิติรุจน์ อติไชยพิทักษ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตวิทยาลัย สาขาวิชาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2563

AUTOMATIC ARTICLE RECOMMENDATION SYSTEM ON THE WEBSITE OF THE ADVERTISING MEDIA BUSINESS COMPANY

Nitirut Atichaipitak

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate School
Thai-Nichi Institute of Technology
Academic Year 2020

หัวข้อวิทยานิพนธ์	ระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา
โดย	นิติรุจน์ อติไชยพิทักษ์
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษา	ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ ดร. พิชิต สุขเจริญพงษ์)

วันที่ ..... เดือน ..... พ.ศ. .....

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. ธงชัย แก้วกิริยา)

.....กรรมการ

(ดร.ชฎาพร เกตุมณี)

.....กรรมการ

(ดร.ธรรศฎภณ สุระศักดิ์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์

(ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล)

นิติรุจน์ อติไชยพิทักษ์: ระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา. อาจารย์ที่ปรึกษา : ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล, 91 หน้า.

ในปี 2563 สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ (ETDA) ได้เผยแพร่รายงานพฤติกรรมของผู้ใช้อินเทอร์เน็ตในประเทศไทย ผลสำรวจระบุว่า 48.3% ชอบอ่านหนังสือหรือบทความออนไลน์ แม้ว่าจะมีหลายเว็บไซต์ให้เลือก แต่ปัญหาที่ผู้อ่านอาจพบคือมีเว็บไซต์มากเกินไปและมีบทความให้เลือกมากเกินไป เป็นการยากและไม่สะดวกสำหรับผู้อ่านที่จะค้นหาบทความที่อาจสนใจ

ผู้วิจัยทราบปัญหาจึงเริ่มศึกษาค้นคว้าปัจจัยที่เกี่ยวข้องกับการนำข้อมูลจากเว็บไซต์ของบริษัท เพื่อพัฒนาระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ ผู้วิจัยได้เสนอระบบแนะนำบทความอัตโนมัติ 2 ระบบ ได้แก่ ระบบแนะนำหมวดหมู่บทความบนเว็บไซต์สำหรับผู้เขียน และระบบแนะนำบทความสำหรับผู้อ่านบทความ ระบบแนะนำบทความที่นำเสนอได้รับการพัฒนาโดยใช้ 2 เทคนิค ได้แก่ Content Based Filtering และ Collaborative Filtering โดยทั้งสองเทคนิคใช้อัลกอริทึม Cosine Similarity

ผลการวิจัยพบว่า ระบบแนะนำหมวดหมู่บทความบนเว็บไซต์สำหรับผู้เขียน ด้วยเทคนิค Content based filtering และผลการคำนวณหาความคล้ายคลึงของเอกสารโดยใช้สูตร Cosine similarity พบว่ามีประสิทธิภาพความคล้ายคลึงกัน 46% ในขณะที่ระบบแนะนำบทความบนเว็บไซต์สำหรับผู้อ่านบทความ โดยใช้เทคนิค Collaborative Filtering เพื่อค้นหาความสัมพันธ์ของแต่ละรายการหรือบทความเพื่อเปรียบเทียบกัน มีประสิทธิภาพ 86% ซึ่งจากผลลัพธ์ดังกล่าวจะเป็นได้ว่าระบบที่พัฒนาขึ้นสามารถเป็นประโยชน์ต่อการวิจัยและพัฒนาเพื่อประยุกต์ใช้งานระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาได้เป็นอย่างมาก

บัณฑิตวิทยาลัย ลายมือชื่อนักศึกษา .....
สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2563

NITIRUT ATICHAIPITAK : AUTOMATIC ARTICLE RECOMMENDATION SYSTEM ON THE WEBSITE OF THE ADVERTISING MEDIA BUSINESS COMPANY. ADVISOR: DR. NATTAGIT JITEURTRAGOOL, 91 PP.

In 2020, the Electronic Transactions Development Agency (ETDA) has published a report on the behavior of internet users in Thailand. Survey results show that 48.3% prefer to read books or articles online. But the problem that readers may encounter is that there are too many websites and too many articles to choose from. It is difficult and inconvenient for readers to find articles that may interest them.

The researcher realized the problem and therefore began to study and research the factors involved in bringing information from the website of a company. In order to development of automatic article recommendations on the website. The researcher has proposed 2 automated article recommendation systems: Article Category Recommendation for Authors, and Web-Based Article Recommendations for Article Readers. The proposed article recommendation systems were developed using 2 techniques, namely Content Based Filtering and Collaborative Filtering where both techniques are based on the use of Cosine Similarity algorithm

The results of the proposed article category recommendation system on the website for authors using techniques such as content-based filtering, and document similarity determination calculation results using the Cosine similarity formula were found to have 46% similarity efficiency while the proposed article recommendation system on the website for article readers by using Collaborative Filtering technique to find the relationship of each item or article to compare showed efficiency of 86%. The efficiency of the proposed systems shows an opportunity for future application of automatic article recommendation systems on websites of advertising media companies.

Graduate School Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2020

กิตติกรรมประกาศ

การทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี เนื่องด้วยได้รับความช่วยเหลือและกรุณาอย่างสูงจาก ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล อาจารย์ที่ปรึกษางานวิจัย ที่ให้ความรู้ ความเข้าใจ แนะนำแนวทาง และคำปรึกษาต่างๆเพื่อในไปทำการปรับปรุงและแก้ไขข้อบกพร่องต่างๆ ตลอดการวิจัย พร้อมทั้งสนับสนุนเครื่องมือที่ใช้ในการทำวิจัยอย่างเต็มที่ อีกทั้งดูแลเอาใจใส่อย่างดีเยี่ยมตลอดระยะเวลาในการทำวิทยานิพนธ์ฉบับนี้ และยังเป็นรุ่นพี่ในสังกัดวิศวกรรมคอมพิวเตอร์ที่ซึ่งผู้วิจัยนั้นได้สำเร็จการศึกษาเมื่อครั้งเรียนอยู่ชั้นปริญญาตรี คณะวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์ ณ สถาบันเทคโนโลยีไทย-ญี่ปุ่น

อีกทั้งผู้วิจัยตระหนักถึงความตั้งใจจริง และความทุ่มเทของคณะกรรมการทุกท่าน ได้แก่ ผู้ช่วยศาสตราจารย์ ดร. ธงชัย แก้วกิริยา ดร.ชฎาพร เกตุมณี และ ดร.ธรรศฎภณ สุระศักดิ์ ที่กรุณาให้ข้อเสนอแนะและคำให้การที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ฉบับนี้ พร้อมทั้งแนะนำแนวทางในการปรับปรุงและแก้ไขข้อบกพร่องต่างๆด้วยความดูแลเอาใจใส่เป็นอย่างดีเยี่ยม จนสามารถทำให้วิทยานิพนธ์ฉบับนี้เสร็จสิ้นสมบูรณ์ตามขอบเขตที่กำหนดไว้

นอกจากนี้ทางผู้วิจัยขอขอบคุณแรงสนุบสนุนจากครอบครัว อาจารย์ทุกท่านที่ได้ให้ความรู้และให้คำปรึกษาตลอดระยะเวลาการศึกษา

ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะก่อให้เกิดประโยชน์แก่ผู้ที่สนใจ และสามารถนำไปต่อยอดเพื่อประยุกต์ใช้กับองค์ความรู้ที่เกี่ยวข้อง อีกทั้งสามารถก่อให้เกิดแนวคิดเพื่อนำไปใช้ในการศึกษาหรือเป็นแนวทางในการทำการวิจับที่เกี่ยวข้องในอนาคต

นิติรุจน์ อติไชยพิทักษ์

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 สภาวะความเป็นมา แนวทางเหตุผล และปัญหา

ปัจจุบันเว็บไซต์บทความ ข้อมูล ข่าวสาร เข้ามามีบทบาทกับชีวิตประจำวันเป็นอย่างมาก ไม่ว่าจะเป็นการให้ข้อมูล การค้นหาข้อมูล หรือการทำธุรกิจ เว็บไซต์เป็นที่เก็บข้อมูลและรวบรวมความรู้ทุกอย่าง ซึ่งข้อมูลต่างๆได้เพิ่มขึ้นมากจนทำให้ผู้อ่านค้นหาบทความ เนื้อหาที่ตัวเองสนใจได้ยากและเสียเวลามาก ดังนั้นเครื่องมือค้นหาที่ดีจึงต้องสามารถแสดงคำตอบที่ตรงกับใจของผู้อ่านและช่วยผู้อ่าน ค้นหาในกรณีที่ไม่สามารถนึกถึงคีย์เวิร์ดที่แม่นยำได้ เพื่อที่จะแก้ปัญหาเหล่านี้ ระบบแนะนำ (Recommendation System) ได้ถูกคิดค้นและสร้างขึ้นมา ตัวอย่างเช่น ผู้อ่านต้องการอ่านบทความให้ความรู้ที่มีความเกี่ยวข้องกันหรือ ซีรีย์บทความ ก่อนเข้าไปนึกได้ว่าอยากอ่านเรื่องการตลาดออนไลน์ที่เป็นบทความล่าสุดเมื่ออ่านจบแล้วจะพบว่าเว็บไซต์ส่วนใหญ่จะเสนอแนะนำบทความล่าสุดให้อ่านซึ่งไม่ตรงใจกับผู้อ่านบทความเพราะว่าต้องใช้เวลาในการค้นหา บางคนตัดใจไม่เลือกอ่านหรือปิดเว็บไซต์ไม่อ่านไปเลยเพราะหาอ่านบทความที่ชอบไม่เจอ แต่ถ้ามีระบบแนะนำบทความ หน้าที่ของระบบคือ ระบบจะเก็บประวัติในการเลือกอ่านของผู้อ่านไว้ เช่นถ้าผู้อ่านมักจะอ่านบทความเรื่องการตลาด ระบบแนะนำก็จะแสดงรายการบทความทั้งหมดที่เกี่ยวข้องกับบทความการตลาดที่มีอยู่ทั้งหมดในเว็บไซต์ให้คุณโดยคุณไม่จำเป็นต้องจำชื่อเรื่องของบทความได้ และ ไม่จำเป็นต้องเสียเวลาค้นหา

ระบบแนะนำบทความคือ ระบบค้นหาและแนะนำบทความให้ตรงกับความต้องการของผู้อ่าน โดยใช้เทคนิคการค้นบทความร่วมกับข้อมูลผู้อ่าน สามารถทราบได้ว่ามีตัวเลือกอะไรบ้าง ซึ่งนอกจากจะช่วยแนะนำให้ผู้อ่าน ที่มีข้อมูลผู้อ่านอยู่ในระบบ สามารถได้รับข้อมูลได้ตรงตามความสนใจของผู้อ่านแล้ว ระบบแนะนำข้อมูลยังสามารถช่วยให้ผู้อ่านที่ไม่มีความรู้ประสบการณ์ช่วยในการตัดสินใจรับข้อมูลได้ตรงตามความต้องการ โดยทำการวิเคราะห์ความต้องการของผู้อ่าน ซึ่งระบบแนะนำข้อมูลโดยทั่วไปมักใช้คำสำคัญ (Keyword) ที่ผู้อ่านใส่เข้ามาในระบบค้นหาและระบบจะนำคำสำคัญ(Keyword) ไปเปรียบเทียบกับคำสำคัญในเอกสารเพื่อหาเอกสารที่เกี่ยวข้อง แล้วนำรายการเอกสารที่เกี่ยวข้องมาแสดงให้กับผู้อ่าน

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาและพัฒนาระบบแนะนำหมวดหมู่บทความบนเว็บไซต์สำหรับผู้เขียน

1.2.2 เพื่อศึกษาและพัฒนาระบบแนะนำบทความบนเว็บไซต์สำหรับผู้อ่านบทความ

1.3 ขอบเขตของการวิจัย

1.3.1 ขอบเขตด้านเทคนิค

ใช้ข้อมูลผู้อ่านบทความเว็บไซต์ของบริษัท Itreelife ย้อนหลังเป็นระยะเวลา 1 ปี โดยเป็นข้อมูลผู้อ่านบทความทั้งหมด 10,000 User จากผู้อ่านบทความทั่วประเทศไทย จำนวนบทความของเว็บไซต์ทั้งหมด 54 บทความ

1.3.2 ขอบเขตการพัฒนาระบบ

การศึกษาและพัฒนาระบบ 3 ระบบ ได้แก่ ระบบค้นหาคำสำคัญในบทความ, ระบบระบบแนะนำหมวดหมู่บทความสำหรับผู้เขียนบทความ และสุดท้าย คือ ระบบแนะนำบทความสำหรับผู้อ่าน ดังรูปที่ 1.1

รูปที่ 1.1 ภาพรวมการทำงานของระบบ

1.4 ขอบเขตด้านเทคนิค

ระบบค้นหาคำสำคัญในบทความจะใช้เทคนิค การตัดคำ (Word Tokenize) โดยใช้ภาษาไทยแบบอิงพจนานุกรม (Dictionary-base) การกำจัดคำหยุด (Stop word) โดยใช้ ผสมผสานกับการนับความถี่ของคำ (TF-IDF) ที่น่าจะมีความสำคัญในบทความนำมาเป็นคำสำคัญสำหรับใช้ในการแบ่งหมวดหมู่ของบทความสำหรับผู้เขียนบทความ การเปรียบเทียบความคล้ายคลึงของเอกสาร (Cosine similarity) และนำไปใช้ประกอบกับระบบแนะนำเพื่อหาว่าบทความได้ความคล้าสยคลึง (Content base filtering)กับบทความที่ผู้ใช้งานสนใจหรือเคยอ่าน หรือ นำข้อมูลบทความแต่ละบทความนำมาหาความคล้ายคลึงกับที่มีผู้อ่านส่วนใหญ่อ่าน นำมา แนะนำผู้อ่านที่มีประวัติการอ่านที่คล้ายคลึงกัน (Collaborative filtering)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้นำเสนอระบบแบบจำลองการค้นหาเมนูไทยอาหารจากส่วนประกอบ ให้แก่ผู้บริโภค

1.5.2 พัฒนาระบบแนะนำบทความบนเว็บไซต์สำหรับผู้อ่านบทความได้ตรงตามความต้องการ

1.5.3 เพิ่มโอกาสที่ผู้ใช้งานจะให้ความสนใจบทความอื่นๆ บนเว็บไซต์ และใช้เวลาอยู่ในหน้าเว็บไซต์นานยิ่งขึ้น

1.5.4 สามารถนำงานวิจัยนี้ไปต่อยอดพัฒนากับงานวิจัยอื่นได้

1.6 นิยามศัพท์เฉพาะทาง

1.6.1 **คำสำคัญ หรือ คีย์เวิร์ด (Keyword)** คือ คำหรือวลีที่ผู้ใช้ค้นหาสิ่งที่ตัวเองต้องการบนใน บทความ หนังสือ เป็นต้น

1.6.2 **สังคมออนไลน์ (Social Media)** คือ สังคมออนไลน์ที่มีผู้ใช้เป็นผู้สื่อสาร หรือเขียนเล่า เนื้อหา เรื่องราว ประสบการณ์บทความ รูปภาพ และวิดีโอ ที่ผู้ใช้เขียนขึ้นเอง ทำขึ้นเอง หรือพบเจอจากสื่ออื่นๆ แล้วนำมาแบ่งปันให้กับผู้อื่นที่อยู่ในเครือข่าย ของตน ผ่านทางเว็บไซต์ Social Network ที่ให้บริการบนโลกออนไลน์ปัจจุบัน เช่น เฟสบุ๊ค (facebooks) เป็นต้น

1.6.3 **Google Analytics** คือ บริการของ Google ที่จัดทำขึ้นมาเพื่อช่วยให้เจ้าของเว็บไซต์ได้ทราบถึงข้อมูลทางสถิติที่เกี่ยวกับเว็บไซต์ของตัวเองได้แบบ Real-Time โดย Google Analytics สามารถใช้วิเคราะห์ข้อมูลในด้านต่าง ๆ เกี่ยวข้องกับเว็บไซต์ของเรา ไม่ว่าจะเป็น ช่องทางในการเข้าถึงเว็บไซต์ ลักษณะของผู้ใช้งานบนเว็บไซต์ พฤติกรรมของผู้ใช้งาน เป็นต้น

1.6.4 **CSV (Comma-Separated Value)** คือ Text File สำหรับเก็บข้อมูลแบบตาราง โดยใช้จุลภาค (,) แบ่งข้อมูลในแต่ละหลัก (Column) และใช้การเว้นบรรทัดแทนการแบ่งแถว (Row)

1.6.5 **Jupyter Notebook** คือ เครื่องมือหนึ่งที่นิยมมากในด้านของ Data Science ซึ่งงานด้านนี้ต้องทำงานที่เกี่ยวกับการจัดการข้อมูลเป็นจำนวนมากๆ แล้วยังต้องรายงาน งานวิจัยที่วิจัยไว้ ซึ่งตัว Jupyter Notebook ก็ได้ออกแบบมาตรงตามจุดประสงค์ของการใช้งานไม่ว่าจะเป็น การเรียกใช้งาน library พร้อมทั้งเขียน code และดูผลได้เลย Jupyter Notebook นั้นถูกออกแบบมาให้ทำงานและอ่านได้ง่ายกว่าการใช้งานโปรแกรมแบบปกติ

1.6.6 **JSON (JavaScript Object Notation)** เป็นรูปแบบในการแลกเปลี่ยนข้อมูลที่มีขนาดเล็ก และด้วยคุณสมบัติของมันทำให้ง่ายในการที่คอมพิวเตอร์จะสร้างและอ่านข้อมูล

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยชิ้นนี้ผู้วิจัยได้ศึกษาค้นคว้าแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องจากเอกสารทางวิชาการ บทความ หนังสือ และแหล่งสารสนเทศทางอินเตอร์เน็ต ดังหัวข้อต่อไปนี้

2.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติ คือการวิเคราะห์ข้อมูลที่อยู่ ในรูปแบบของภาษามนุษย์ที่ใช้สื่อสารทั่วไปในชีวิตประจำวัน ให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์เหล่านั้นได้ เช่นเดียวกันกับเจ้าของภาษามนุษย์ G. Chowdhury [1] ซึ่งประกอบด้วยหลายขั้น ตั้งแต่ขั้นตอนการจัดเตรียมข้อมูลรวมไปถึง การแปลงข้อมูลให้อยู่ในรูปแบบของตัวเลขที่คอมพิวเตอร์สามารถนำไปใช้ในการประมวลผลต่อได้

Natural Language Processing หรือ การประมวลผลภาษาธรรมชาติ เป็นสาขาหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ AI (Artificial Intelligence) จุดเริ่มต้นโดยใช้หลักการแบบการตั้งกฎ (Rule-Based) เป็นส่วนหนึ่งของภาษาศาสตร์ ใช้การวิเคราะห์ในรูปแบบของการตั้งกฎ (Rule-Based) R. C. Schank and R. P. Abelson [2] เป็นการตั้งกฎขึ้นมาใช้ในการแก้ปัญหา มาประยุกต์ใช้ในงานด้านต่าง ๆ โดยสามารถสร้างได้โดยไม่ต้องอาศัยชุดข้อมูลจำนวนมาก ซึ่งมีข้อเสียในการวิเคราะห์คำใหม่ ๆ ที่เกิดขึ้นมา ทา ให้ระบบไม่รู้จกัคา เหล่านั้น และยงัไม่สามารถจำแนกได้จะต้องมีการเพิ่มคำนั้นและกำหนดประเภทของคำนั้นเสียก่อน จากการใช้งานด้วยการตั้งักฎ ต่อมาได้มีการประยุกต์ใช้การเรียนรู้ของเครื่อง (Machine Learning) ในการสร้างตัวจ าแนกประเภทของข้อความ C. Aone et al. [3] โดยจะทำงานแตกต่าง จากการตั้งักฎ โดยวิเคราะห์คำ บางคำ และทำ การสรุปผลการวิเคราะห์ในทัน ทีเท่านั้น ซึ่งสามารถ สร้างได้จากการรวบรวมข้อมูลและทา การกำหนดประเภทของขอ้ มูลจากน้นั ทา การฝึกสอนโดยใช้ อัลกอริทึมของการเรียนรู้ของเครื่อง

ในปัจจุบันการเรียนรู้ของเครื่องได้มีการพัฒนาไปเป็นการเรียนรู้เชิงลึก (Deep Learning) ซึ่งการเรียนรู้เชิงลึกมีจุดเด่นในการเรียนรู้คุณลักษณะต่าง ๆ ได้อย่างหลายหลาย และเรียนรู้ได้จำนวนมาก E. Haddi et al. [4] โดยสามารถทำการสร้างแบบจำลองในการจำแนกประเภทของข้อความได้อย่างแม่นยำมากยิ่งขึ้น ซึ่งได้มีการนำมาประยุกต์ใช้ทางด้าน Natural Language Processing ในด้านของการสร้างแบบจำลองสำหรับแปลภาษา (Machine Translation) โดยสามารถ แปลได้อย่างถูกต้องและแม่นยำสืบเนื่องด้วยความสามารถทางเทคโนโลยีและอัลกอริทึมที่สามารถ วิเคราะห์ข้อมูล

ได้อย่างมีประสิทธิภาพทำให้มีการพัฒนาการวิเคราะห์ทางด้านภาษาธรรมชาติ เป็นไปอย่างก้าวกระโดด

2.2 การตัดคำ (Word Segmentation)

การตัดคำภาษาไทยคือการแยกคำแต่ละคำของ ภาษาไทยที่เขียนติดกัน ในงานวิจัยนี้ทำการตัดคำภาษาไทยเพื่อนำไปเข้ากระบวนการประมวลผลภาษาธรรมชาติ (Natural language processing) เพื่อทำเป็นคลังข้อมูลสำหรับนำไปใช้เปรียบเทียบหาคำสำคัญหรือ Keyword ส่วนวิธีการตัดคำหลักมีใช้อยู่สามแบบ ได้แก่ วิธีการตัดคำโดยใช้กฎ (Rule-based) วิธีการตัดคำโดยใช้พจนานุกรม (Dictionary) และวิธีการตัดคำโดยใช้คลังข้อความ (Corpus) ปัจจุบันมีวิธีการตัดคำโดยใช้ NLP (Natural Language Processing) ซึ่งเป็นสาขาหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence หรือ AI) มาช่วยลดระยะเวลาและการค้นคำที่ไม่มีในระบบได้ถูกต้องและแม่นยำมากขึ้น

กรองคำที่ไม่เกี่ยวข้องหรือ คำหยุด (Stop word) เป็นการนำคำที่ไม่สำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง หรือหมายถึงคำที่ใช้กันโดยทั่วไปที่ไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง เช่น จึง ซึ่ง นี้ เป็นต้น

2.3 การประเมินค่าความสำคัญของวลี

การประเมินค่าความสำคัญของวลี เป็นการคำนวณค่า ความสำคัญที่แต่ละวลีมีต่อเอกสาร พิจารณาความสำคัญของวลีและตำแหน่งของวลีในเอกสาร การคำนวณค่าความถี่ ใช้เทคนิคการกำหนดค่าความสำคัญให้กับคำในการแทนที่เอกสารเป็นแบบจำลองเวคเตอร์ (Vector-Space Model) คือค่าผลคูณระหว่าง TF (Term Frequency) และ (Inverse Document Frequency) ซึ่งเป็นการเปรียบเทียบความถี่ของวลีในเอกสารกับความถี่ของวลีในเอกสารอื่น ดังสมการนี้

$$\mathrm{TF} \times \mathrm{IDF} = \mathrm{freq}(\mathrm{P}, \mathrm{D}) \times \log_2 \frac{\mathrm{N}}{\mathrm{n_p}} \qquad (2\text{-}1)$$

โดยที่

ตัวแปร D คือ เอกสาร

ตัวแปร P คือวลี ตัวแปร $\mathbf{freq(P, D)}$ คือ ความถี่ที่พบวลี P ในเอกสาร D

ตัวแปร $\mathbf{n_p}$คือ จำนวนเอกสารในฐานข้อมูลที่พบวลี P

ตัวแปร N คือจำนวนเอกสารทั้งหมดในฐานข้อมูล

2.4 ระบบแนะนำบทความ (Recommendation System)

ระบบแนะนำบทความ (Recommender System) Terveen and Hill [5] P. Markellou et al. [6] เป็นระบบที่ช่วยแนะนำข้อมูลที่คาดว่าผู้ใช้จะให้ความสนใจโดยเปรียบเทียบจากคุณลักษณะของผู้ใช้งานเพื่อสร้างคำแนะนำให้แก่ผู้ใช้ โดยระบบนี้สามารถนำไปประยุกต์ใช้ได้กับการแนะนำข้อมูลได้ในหลากหลายประเภท เช่น บทความข่าว หนังสือ ภาพยนตร์ และสินค้าต่าง ๆ เป็นต้น ซึ่งโดยทั่วไปแล้ววิธีการสร้างรายการแนะนำให้กับผู้ใช้มี 3 วิธีดังนี้

2.4.1 การกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

เป็นเทคนิคทางด้านการกรองสารสนเทศ (Information Filtering) ที่ได้รับความนิยมในการนำไปใช้และประสบผลสำเร็จเป็นอย่างดี ซึ่งแนวความคิดพื้นฐานของการแนะนำข้อมูลของวิธีนี้ขึ้นอยู่กับความคิดเห็นของผู้ใช้หลาย ๆ คนในระบบ โดยระบบจะทำการค้นหากลุ่มสมาชิกข้างเคียง (Neighbors) ที่มีความชอบที่เหมือนกันกับผู้ใช้เป้าหมาย โดยข้อมูลจากกลุ่มสมาชิกข้างเคียงนี้จะถูกนำมาวิเคราะห์ในระบบเพื่อแนะนำข้อมูลให้แก่ผู้ใช้เป้าหมายต่อไป ซึ่งการทำงานของการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม มีขั้นตอนการทำงาน 4 ขั้นตอนดังนี้

1) วิธีการคำนวณความคล้ายคลึง (Similarity Computation) ในระบบแนะนำข้อมูลมีการเก็บรวบรวมข้อมูลของผู้ใช้ในรูปแบบของตาราง (Matrix) ซึ่งประกอบด้วยจำนวนผู้ใช้ n คน และจำนวนข้อมูล m รายการ ข้อมูลนี้จะถูกเก็บไว้ในรูปแบบของตาราง ดังนั้นการคำนวณหาค่าความคล้ายคลึงระหว่างผู้ใช้ 2 คน วิธีที่นิยมใช้ในการคำนวณความคล้ายคลึงกันมี 2 วิธี J. S. Breese et al. [7] B. Sarwar et al. [8] คือ Correlation-based และ Cosine-based ซึ่งในงานวิจัยของ L. Jonathan et al. [9] ได้นำค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient : PCC) มาใช้ใน การคำนวณค่าความคล้ายคลึงระหว่างผู้ใช้ 2 คน โดยมีวิธีการคำนวณดังนี้ ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน เป็นวิธีการวัดความคล้ายคลึงกันระหว่างผู้ใช้ 2 คนที่ให้คะแนนกับรายการข้อมูลที่สนใจ ข้อดีของค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน คือ วิธีนี้จะไม่ขึ้นอยู่กับค่าเฉลี่ยของคะแนนรายการข้อมูลที่สนใจ หมายความว่าถ้าผู้ใช้มีการโน้มเอียงในการให้คะแนนกับรายการข้อมูลบางอย่างบ่อย ๆ พฤติกรรมของผู้ใช้นี้จะไม่มีผลต่อการทำงานของระบบ ขั้นแรกของ

การคำนวณประกอบด้วย เซตของรายการข้อมูลที่มีการให้คะแนนไว้ จากผู้ใช้ 2 คน แล้วคำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน ระหว่างผู้ใช้เป้าหมายที่ ระบบจะแนะนำข้อมูลให้ (User a) และผู้ใช้ในระบบ (User u) ดังสมการที่นี้

$$C_{a,u} = \frac{\text{covar}(r_a, r_u)}{\sigma_{r_a}\sigma_{r_u}} \quad (2\text{-}2)$$

โดยที่

r_{a} และ r_{u} คือ คะแนนของรายการข้อมูลที่ได้จาก User a และ User u

สมการ Covariance:

$$\text{covar}(r_a, r_u) = \frac{\sum_{i=1}^{m}(r_{a,i}-\bar{r}_a)(r_{u,i}-\bar{r}_u)}{m} \quad (2\text{-}3)$$

$$\bar{r}_u = \frac{\sum_{i=1}^{m} r_{u,i}}{m} \quad (2\text{-}4)$$

โดยที่

$r_{\text{a,i}}$ และ $r_{\text{u,i}}$ คือ ค่าคะแนนของรายการข้อมูล i ที่ให้โดย User a และ User u

$\bar{r}_{\text{u}}$ คือ ค่าเฉลี่ยคะแนนของรายการข้อมูลของ User u

m คือ จำนวน Co-Rated Items

สมการ Standard Deviation

$$\sigma_{r_u} = \frac{\sqrt{\sum_{i=1}^{m}(r_{u,i}-\bar{r}_u)^2}}{m} \quad (2\text{-}5)$$

จากงานวิจัยของ L. Jonathan et al. [9] ได้กล่าวไว้ว่า การใช้ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน เพียงอย่างเดียวไม่เพียงพอในการคำนวณหาความคล้ายคลึง เนื่องจากจะทราบได้อย่างไรว่าคนที่ให้คะแนนรายการข้อมูลนั้นเป็นคนที่มีความชอบคล้ายกับผู้ใช้เป้าหมายที่ระบบกำลังจะแนะนำข้อมูลให้จริง ดังนั้น ได้แนะนำการคำนวณค่าน้ำหนัก (Weight) เพื่อใช้ในการคำนวณหาค่าน้ำหนักของคนที่มีความชอบเหมือนกับผู้ใช้เป้าหมาย โดยมี การคำนวณดังสมการนี้

$$S_{a,u} = \begin{cases} 1 & if \quad m > 50 \\ \frac{m}{50} & if \quad m \leq 50 \end{cases} \quad (2\text{-}6)$$

โดยที่

m *คือ จำนวน Co-Rated Items*

สุดท้ายการคำนวณความคล้ายคลึงสามารถคำนวณได้จากค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน และค่าน้ำหนัก L. Jonathan et al. [9] ดังสมการนี้

$$\sin(a,u) = S_{a,u}C_{a,u} \quad (2\text{-}7)$$

โดยที่

$S_{a,u}$ คือ ค่าน้ำหนัก

$C_{a,u}$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน ระหว่าง User a และ User u

2) วิธีการเลือกสมาชิกข้างเคียง (Neighbor Selection) หลังจากที่มีการคำนวณความคล้ายคลึงเสร็จแล้ววิธีการถัดมา คือ การเลือกสมาชิกข้างเคียงซึ่งเป็นวิธีการที่จำเป็นอย่างยิ่ง โดยจะทำการเลือกจากเซตย่อยของผู้ใช้จากผู้ใช้ที่อยู่ในระบบทั้งหมดเพื่อนำไปใช้ในการทำนายต่อไป

เทคนิคหลักที่ใช้ในการเลือกสมาชิกข้างเคียง มีอยู่ 2 วิธีคือ Similarity Threshold และ Best K-Neighbor, P. Resnick et al. [10] แต่วิธีที่นิยมใช้คือการตั้งค่า Threshold ขึ้นมาเอง

3) วิธีการทำนาย (Prediction) วิธีการทำนายเป็นการพยากรณ์ค่าความชอบของผู้ใช้ต่อไอเท็ม (Item) ใดไอเท็มหนึ่งโดย พิจารณาจากความชอบและความคล้ายระหว่างไอเท็มนั้นกับไอเท็มอื่น ๆ ซึ่งจะนำเซตของสมาชิก ข้างเคียงที่ถูกเลือกไว้ข้างต้นนำมาคำนวณโดยใช้สมการนี้

$$p_{a,j} = \bar{r} + \frac{\sum_{u=1}^{n} w_{a,u}(r_{u,i} - \bar{r}_u)}{\sum_{u=1}^{n} w_{a,u}} \quad (2\text{-}8)$$

โดยที่

n คือ จำนวนผู้ใช้ที่ถูกเลือกมาเป็นสมาชิกข้างเคียง

$p_{a,i}$ คือ การทำนายสำหรับไอเท็ม i ของ User a

$\bar{r}_u$ คือ ค่าเฉลี่ยของคะแนนของ User u

$r_{u,i}$ คือ คะแนนของรายการข้อมูลของ User u ที่ให้กับไอเท็ม i

$w_{a,u}$ คือ ค่าความคล้ายคลึงระหว่าง User a และ User u

4) วิธีการสร้างรายการแนะนำ (Recommendation) การสร้างรายการแนะนำเป็นขั้นตอนที่ทำหลังจากขั้นตอนการทำนาย โดยนำค่าที่ได้จากการทำนายในแต่ละไอเท็มที่ได้มาเรียงตามลำดับโดยเริ่มตั้งแต่รายการที่มีค่าการทำนายที่มากที่สุด จนถึงรายการที่มีค่าการทำนายต่ำสุด ซึ่งการเลือกจำนวนรายการแนะนำมาแสดงให้กับผู้ใช้นั้น ขึ้นอยู่กับผู้สร้างระบบว่าต้องการแสดงรายการแนะนำกี่รายการ

การกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วมถูกนำมาใช้ในการสร้างระบบแนะนำข้อมูล เช่น ระบบแนะนำข่าว GroupLens P. Resnick et al. [10] ระบบแนะนำเรื่องตลกบนเว็บ Jester K. Goldberg et al. [11] และระบบแนะนำข้อมูลที่ได้รับความนิยมคือ Amazon เป็นต้น นอกจากนี้ยังมีงานวิจัยที่แนะนำวิธีที่ช่วยแก้ไขปัญหาการกรองสารสนเทศแบบพึ่งพาผู้ใช้ ร่วม ได้แก่ งานวิจัยของ M. O'Connor and J. Herlocker [12] ได้นำเสนอวิธีจัดกลุ่มข้อมูลเรตติ้ง (Rating) แล้วนำข้อมูลที่จัดกลุ่มเสร็จแล้วมาใช้ในวิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม ผลการวิจัยแสดงให้เห็นว่าวิธีการจัดกลุ่มข้อมูลจะช่วยปรับปรุงประสิทธิภาพในการแนะนำข้อมูลและเพิ่มความสามารถในการรองรับ

ปริมาณข้อมูลที่เพิ่มขึ้น ต่อมางานวิจัยของ D. Lemire and A. Maclachlan [13] ได้นำเสนออัลกอริธึม Slope One ที่ใช้แนวคิดความแตกต่างของความนิยม (Popularity Differential) โดยใช้ความแตกต่างระหว่างคู่ของไอเท็มมาใช้ในการทำนายคะแนน

สินค้าของผู้ใช้ที่ยังไม่ได้ให้คะแนนสินค้านั้น ซึ่งเมื่อนำอัลกอริธึม Slope One ไปทดลองกับข้อมูลของ Movielens แล้วทำการเปรียบเทียบผลลัพธ์ที่ได้กับการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม ผลการวิจัยพบว่า ผลลัพธ์ที่ได้จากอัลกอริธึม Slope One ให้ความแม่นยำที่ใกล้เคียงกับ Memory-Based Scheme อย่างเพียร์สัน (Pearson) แต่มีความซับซ้อนของอัลกอริธึมน้อยกว่า ในงานวิจัย ของ วงกต ศรีอุไร, พยุง มีสัจ และชูชาติ หฤไชยะศักดิ์ [14] ได้นำเสนอวิธีการแทนค่าข้อมูลที่ขาดหายโดยการวิเคราะห์ค่า ความแปรปรวนของข้อมูล และการแทนค่าข้อมูลด้วยวิธีการหาสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor) ผลการวิจัยแสดงให้เห็นว่าวิธีที่นำเสนอช่วยแก้ไขปัญหาความเบาบางของข้อมูล และช่วยเพิ่มประสิทธิภาพในการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม

2.4.2 การกรองสารสนเทศโดยดูจากเนื้อหา (Content based filtering)

เป็นการกรองสารสนเทศวิธีหนึ่งที่นำมาใช้ในระบบแนะนำข้อมูล ซึ่งปกติจะนิยมใช้ในการค้นคืนสารสนเทศ ที่ให้ความสนใจกับเนื้อหาหรือคุณสมบัติของข้อมูลเป็นหลัก โดยวิธีนี้จะตรวจสอบว่าข้อมูลนั้นมีเนื้อหาตรงกับข้อมูลส่วนบุคคล (Profile) วลัยนุช สกุลนุ้ย [15] ของผู้ใช้หรือไม่ถ้าคล้ายคลึงก็จะนำมาเสนอให้กับผู้ใช้ทันที แต่ถ้าไม่ใช่ จะไม่สนใจ ถึงแม้ว่าข้อมูลนั้นอาจจะมีลักษณะใกล้เคียงกับที่ผู้ใช้ต้องการก็ตาม ดังนั้นวิธีการนี้เป็น การคำนวณหาค่าความคล้ายคลึงระหว่างข้อมูลกับข้อมูลส่วนบุคคลของผู้ใช้โดยการจับคู่ข้อมูล กับข้อมูลส่วนบุคคลของผู้ใช้ เพื่อค้นคืนสารสนเทศที่ผู้ใช้สนใจ ซึ่งวิธีการ กรองสารสนเทศโดยดูจากเนื้อหานี้ถูกนำไปใช้ในการแนะนำข้อมูลในหลากหลายโดเมน เช่น เว็บไซต์ ข่าว R. J. Mooney and L. Roy [16] หนังสือ เพลง และ บทความ C. Haruechaiyasak and C. Damrongrat [17] เป็นต้น โดยทั่วไปแล้วเทคนิคที่นำมาใช้ในการกรองสารสนเทศโดยดูจากเนื้อหามีหลายวิธี เช่น TF/IDF T. Nomponkrang and C. Sanrach [18] เบย์เซียน (Bayesian), วิธีการหาสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor), วิธีการจัดกลุ่ม (Clustering Method) และการตัดสินใจโดยใช้แผนผังต้นไม้ (Decision Tree) เป็นต้น การกรองสารสนเทศโดยดูจากเนื้อหานี้จะไม่ประสบปัญหาการให้เรตติ้งของข้อมูลไม่ทั่วถึง (Sparsity) และปัญหาข้อมูลใหม่ที่ยังไม่มีผู้ใช้ให้เรตติ้ง (First Rater) กล่าวคือ ถ้าข้อมูลนั้นมีความคล้ายคลึงกับข้อมูลส่วนบุคคลของผู้ใช้ข้อมูลนั้นก็มีโอกาสที่จะถูกนำเสนอ ออกมาถึงแม้ว่าจะไม่มีผู้ใช้คนใดเคยให้เรตติ้งก็ตาม แต่ปัญหาหลักของการกรองสารสนเทศโดยดูจากเนื้อหา ก็คือ ข้อมูลที่ไม่ตรงกับข้อมูลส่วนบุคคลของผู้ใช้จะไม่มีโอกาสที่จะถูกนำออกมานำเสนอ ถึงแม้ว่าจะเป็นข้อมูลที่ผู้ใช้สนใจก็ตาม

2.4.3 การกรองสารสนเทศแบบผสม (Hybrid Filtering)

เป็นวิธีการกรองสารสนเทศที่รวมทั้ง 2 วิธีระหว่างการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วมและการกรองสารสนเทศโดยดูจากเนื้อหาเข้าด้วยกัน ซึ่งการกรองสารสนเทศแบบผสมนี้จะช่วยหลีกเลี่ยงข้อจำกัดของการกรองสารสนเทศของทั้ง 2 วิธี ระบบแนะนำข้อมูลที่เป็นแบบผสมสามารถแบ่งได้ 4 รูปแบบ ดังนี้

1) สร้างระบบโดยแยกส่วนของการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วมและการกรองสารสนเทศโดยดูจากเนื้อหาออกจากกัน แต่ใช้วิธีการทำนายข้อมูลให้แก่ผู้ใช้ร่วมกัน

2) สร้างระบบโดยรวมคุณลักษณะบางประการของวิธีการกรองสารสนเทศโดยดูจากเนื้อหา เข้าไปในวิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม

3) สร้างระบบโดยรวมคุณลักษณะบางประการของวิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม เข้าไปในวิธีการกรองสารสนเทศโดยดูจากเนื้อหา

4) สร้างระบบโดยรวมคุณลักษณะของทั้ง 2 วิธี คือ การกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วมและการกรองสารสนเทศโดยดูจากเนื้อหาเข้าด้วยกัน

การกรองสารสนเทศแบบผสมนี้ มีนักวิจัยหลายคนที่ได้ทำวิจัยเพื่อสร้างระบบแนะนำข้อมูล เช่น งานวิจัยของ M. J. Pazzani [19] นำเสนอระบบแนะนำข้อมูลโดยใช้วิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม และใช้วิธีการกรองสารสนเทศโดยดูจากเนื้อหาเพื่อสร้างข้อมูลส่วนบุคคลสำหรับผู้ใช้แต่ละคน ซึ่งการสร้างข้อมูลส่วนบุคคลของวิธีนี้จะไม่ได้มาจากการให้คะแนนกับไอเท็มแต่เป็นการสร้างข้อมูลส่วน บุคคลจากการคำนวณความคล้ายคลึง (Similarity Computation) ระหว่างผู้ใช้ 2 คน แทน ต่อมางานวิจัยของ P. Melville et al. [20] ได้ประยุกต์ใช้การกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วมในการสร้างคะแนนที่ผู้ใช้ให้กับไอเท็มนั้นให้อยู่ในรูปแบบของเวกเตอร์ เพื่อที่จะนำคะแนนเหล่านี้ไปใช้ในการ ทำนายข้อมูลให้แก่ผู้ใช้ในวิธีการกรองสารสนเทศโดยดูจากเนื้อหาต่อไป สำหรับงานวิจัยของ I. M. Soboro [21] ได้นำเสนอวิธีการหาความหมายที่แอบแฝง (Latent Semantic Indexing : LSI) เพื่อนำมาสร้างข้อมูลส่วนบุคคล ซึ่งข้อมูลส่วนบุคคลนี้ถูกแทนด้วยเวกเตอร์ของคำ (Term Vector) ซึ่งผลการวิจัยพบว่าวิธีที่นำเสนอให้ประสิทธิภาพดีกว่าการกรองสารสนเทศโดยดูจากเนื้อหาแบบดั้งเดิม

2.5 งานวิจัยที่เกี่ยวข้องกับระบบแนะนำข้อมูล

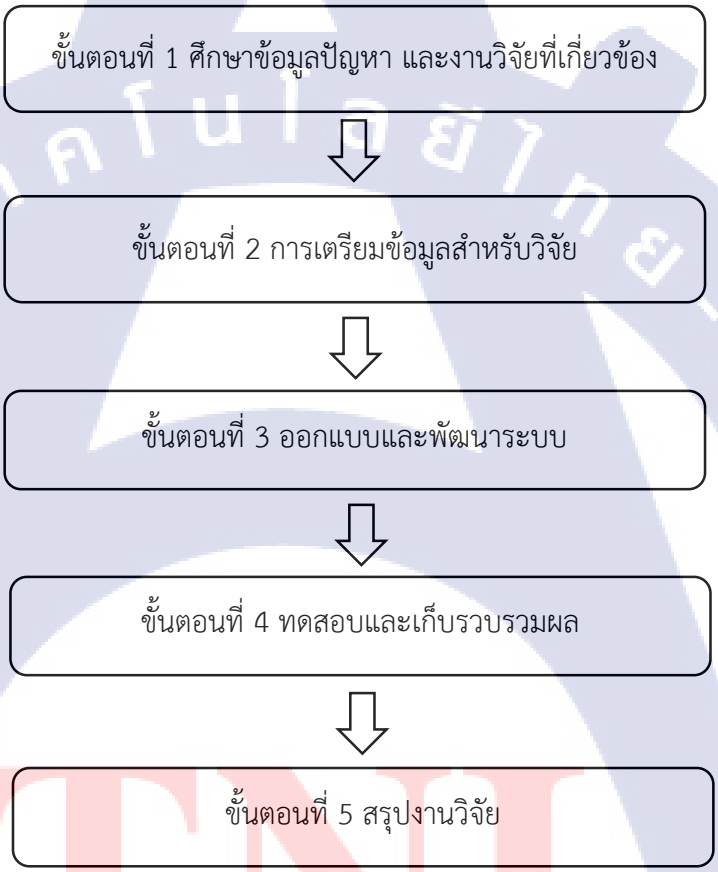
งานวิจัยที่ผ่านมาเป็นการนำเสนองานวิจัยที่เกี่ยวข้องกับระบบแนะนำข้อมูลที่ใช้เทคนิคการกรองข้อมูลทั้ง 3 วิธี อย่างไรก็ตามยังมีงานวิจัยอื่นที่เกี่ยวข้องกับระบบแนะนำข้อมูล แต่มีการนำเทคนิคอื่นนอกเหนือจากเทคนิคดั้งเดิมมาใช้ในการแนะนำ และยังมีงานวิจัยที่เกี่ยวข้องกับการนำเสนอวิธีการแนะนำข้อมูลในรูปแบบต่าง ๆ โดยตัวอย่างงานวิจัยเหล่านี้ เช่น งานวิจัยของ

M. Deshpande and G. Karypis [22] ได้นำเสนออัลกอริธึม Item-Based เพื่อใช้ในการแนะนำข้อมูล อัลกอริธึมนี้จะใช้วิธีการคำนวณความคล้ายคลึงกันของไอเท็มที่ผู้ใช้สนใจ แทนการคำนวณความคล้ายคลึงระหว่างผู้ใช้ในวิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม ซึ่งผลการวิ จัยพบว่าอัลกอริธึม item-based มีผลการทำนายที่ดีกว่าวิธี User-Based ส่วนงานวิจัยของ J. Sobecki and S. L. Szczepan [23] ได้นำเสนอวิธี Artificial Immune System หรือ AIS ร่วมกับวิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม เพื่อสร้างระบบรายงานข่าวโดยใช้ข้อมูลของ Wiki-News การทำงานของระบบให้ผู้ใช้ลงทะเบียนเพื่อเข้าใช้ระบบซึ่งข้อมูลที่ได้จากการลงทะเบียนบางข้อมูล จะเป็นข้อมูลที่ผู้ใช้มีความสนใจอยู่ จากนั้นเข้าสู่ระบบ ผู้ใช้เริ่มอ่านข่าวและมีการให้คะแนนกับ ข่าว นั้น ระบบจะทำการเก็บข้อมูลเหล่านี้แล้วใช้วิธี AIS ในการเลือกกลุ่มของผู้ใช้ทีมีความชอบที่ คล้ายคลึงกัน หลังจากนั้นใช้ วิธีการกรองสารสนเทศแบบพึ่งพาผู้ใช้ร่วม เพื่อสร้างรายการข่าว แนะนำให้กับผู้ใช้ตามความต้องการของแต่ละคน ผลการวิจัยสรุปได้ว่าวิธีการที่นำเสนอนี้ให้ ความถูกต้องของผลการทำนายได้ดีกว่าวิธีการที่ทำนายจากความชอบของผู้ใช้เพียงอย่างเดียว งานวิจัยของ รัตน์ชัยนันท์ ธรรมสุจริต [24] ได้นำเสนอการปรับปรุงเทคนิคการจัดกลุ่มข้อมูลด้วยต้นไม้ลำดับขั้นแบบสมดุล (Balanced Iterative Reducing and Clustering Hierarchies,BIRCH) ร่วมกับการเติมข้อมูลว่างด้วยเทคนิคการเติมค่าจากกลุ่มข้อมูลข้างเคียง (K-Nearest Neighbor Imputation) เพื่อการลดมิติของข้อมูล โดยผลการทดลองกับชุดข้อมูลจาก Movielens แสดงประสิทธิภาพที่ได้จากการผสมผสานระหว่างการจัดกลุ่มข้อมูลกับเทคนิคการคดกรองข้อมูล แบบพึงพาสามารถบรรลุประสิทธิภาพทั้งในด้านความแม่นยำและอัตราการตอบสนองได้ดี งานวิจัยของ พัชราภรณ์ ช่วยเจริญ และ พนิดา ทรงรัมย์ [25] เสนอการขุดค้นความสัมพันธ์หมวดหมู่ของเพจบนเฟสบุ๊ค โดยประยุกต์ใช้อัลกอริทึมเอฟพี-กโรธ (FP-Growth Algorithm) ในการขุดค้นหมวดหมู่ของเพจที่เกิดร่วมกันบ่อย จากนั้นสร้างกฎความสัมพันธ์จากหมวดหมู่ของเพจที่เกิดร่วมกันบ่อย ข้อมูลที่ใช้ในการศึกษาหาความสัมพันธ์หมวดหมู่ของเพจบนเฟสบุ๊คเป็นข้อมูล การกดถูกใจเพจของผู้ใช้จำนวน 1,780 คน ทำการทดลองหา ค่าสนับสนุนขั้นต่ำ (Minimum Support) และค่าความเชื่อมั่นขั้นต่ำ (Minimum Confdence) ที่เหมาะสมในการสร้างกฎซึ่งผลการทดลองแสดงให้เห็นว่าการกำหนดค่าสนับสนุน ขั้นต่ำเท่ากับ 10% ค่าความเชื่อมั่นขั้นต่ำเท่ากับ 50% ให้ค่าความถูกต้องมากที่สุด คือ 93.71% โดยกฎที่สร้างได้ มีจำนวน 7,749 กฎ และสามารถนำมาใช้ได้จริง 892 กฎที่ได้สามารถนำใช้เพื่อเป็นแนวทางในการวางแผน โฆษณาหรือนำเสนอเพจให้ตรงตามความสนใจมากขึ้น งานวิจัยของ วนารัตน์ จุฬพันธ์ทอง และ ไกรศักดิ์ เกษร [26] ระบบแนะนำข้อมูลการท่องเที่ยวเฉพาะบุคคล (Personalized Tourism Recommendation System-PTRS) เป็นเครื่องมือที่ช่วยให้นักท่องเที่ยว ได้รับข้อมูลที่ต้องการค้นหาได้รวดเร็วขึ้น โดยระบบพยายามที่จะแนะนำข้อมูลการท่องเที่ยวให้ตรงกับความสนใจ อย่างไรก็ตามความท้าทายสำหรับผู้วิจัยคือปัญหาโคลด์สตาร์ท (Cold-Start) ซึ่งเกิดจากการขาดข้อมูลที่เพียงพอมาวิเคราะห์หา

ความสนใจของนักท่องเที่ยว ดังนั้นงานวิจัยนี้จึงได้นำเสนอระบบแนะนำข้อมูลการท่องเที่ยว เฉพาะบุคคล โดยใช้ประโยชน์จากข้อมูลบนบริการเครือข่ายสังคมเฟสบุ๊คมาช่วยวิเคราะห์หาความสนใจของผู้ใช้เพื่อแก้ปัญหาข้างต้น โดยมุ่งเน้นไปที่ปัญหาโคล์ดสตาร์ทที่เกิดกับผู้ที่ใช้งานระบบครั้งแรกและภาระของผู้ใช้ในการกรอกข้อมูลความสนใจด้วยตนเองในระบบแบบเดิม ซึ่งจากผลการวิจัยพบว่าการใช้ประโยชน์จากข้อมูลบนเครือข่ายสังคมของผู้ใช้สามารถช่วยลดปัญหาโคลด์สตาร์ทและแนะนำสถานที่ท่องเที่ยวได้อย่างมีประสิทธิภาพบนพื้นฐานความต้องการของผู้ใช้ระบบได้ งานวิจัยของธนพล พุกเส็ง และ คณะ [27] ระบบผู้แนะนำมีบทบาทสำคัญต่อคนในยุคปัจจุบันที่ได้รับข้อมูลข่าวสารเกี่ยวกับสินค้าและบริการต่าง ๆ เป็นจำนวนมากโดยจะช่วยนำเสนอสิ่งที่มีความเหมาะสมแก่ผู้ใช้งานมากที่สุด แต่ระบบผู้แนะนำเองก็มีปัญหาที่ส่งผลกระทบต่อการแนะนำ เช่น ความเบาบางของข้อมูล และการมีผู้ใช้งานรายใหม่ ดังนั้นจึงได้มีการนำเทคนิคต่าง ๆ มาผสมผสานกับระบบผู้แนะนำเพื่อแก้ปัญหาและปรับปรุงประสิทธิภาพในการแนะนำให้ดียิ่งขึ้นสำหรับในบทความนี้ ได้นำเสนอการศึกษางานวิจัยทางด้านระบบผู้แนะนำที่ได้นำความไว้วางใจและผู้เชี่ยวชาญมาเป็นส่วนประกอบตลอดจนการนำทั้งสองเทคนิคมาใช้งานร่วมกันเพื่อพัฒนาระบบผู้แนะนำสำหรับผลการศึกษาได้แสดงให้เห็นว่าเมื่อนำเทคนิคทั้งสองมาใช้งานร่วมกันนั้นจะสามารถแก้ปัญหาระบบผู้แนะนำและปรับปรุงประสิทธิภาพกระบวนการแนะนำได้เป็นอย่างดี

บทที่ 3
วิธีดำเนินการงานวิจัย

ระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาโดยใช้เทคนิคการ Content-Based Filtering และ Collaborative Filtering ซึ่งผู้วิจัยได้แบ่งวิธีการดำเนินงานออกเป็นขั้นตอน ดังรูปที่ 3.1



รูปที่ 3.1 แผนผังวิธีดำเนินการวิจัย

ตารางที่ 3.1 ระยะเวลาการดำเนินการวิจัย

การดำเนินงาน	ระยะเวลา										
	ปี 2563										
	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.	ก.ย.	ต.ค.	พ.ย.
ขั้นตอนที่ 1 ศึกษาข้อมูลปัญหา และงานวิจัยที่เกี่ยวข้อง											
ขั้นตอนที่ 2 การเตรียมข้อมูลสำหรับวิจัย											
ขั้นตอนที่ 3 ออกแบบและพัฒนาระบบ											
ขั้นตอนที่ 4 ทดสอบและเก็บรวบรวมผล											
ขั้นตอนที่ 5 สรุปงานวิจัย											

3.1 ขั้นตอนที่ 1 ศึกษาข้อมูลปัญหา และงานวิจัยที่เกี่ยวข้อง

ในปัจจุบันเว็บไซต์บริษัท Itreelife ไม่มีระบบแนะนำบทความสำหรับผู้อ่าน เมื่อเข้ามาอ่านในเว็บไซต์ หากต้องการอ่านบทความที่มีความเกี่ยวข้องกับบทความที่เคยอ่านมาก่อนหน้านี้ ต้องใช้วิธีค้นหาด้วยตัวเอง ทำให้ต้องใช้เวลาในการค้นหาค่อนข้างนานและไม่ตรงกับความต้องการ จากปัญหาดังกล่าวจึงเป็นแรงจูงใจให้ผู้พัฒนาระบบมีแนวคิดในการพัฒนาระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา เพื่อช่วยแนะนำบทความในเว็บไซต์ให้ตรงกับความต้องการของผู้อ่านแต่ละคนให้มากที่สุด

3.2 ขั้นตอนที่ 2 การเตรียมข้อมูลสำหรับวิจัย

งานวิจัยนี้ผู้วิจัยได้นำข้อมูลเพื่อใช้ในการแนะนำหมวดหมู่บทความอัตโนมัติให้กับผู้เขียนบทความ และ การแนะนำบทความอัตโนมัติให้กับผู้อ่านบทความ โดยขั้นตอนการเตรียมข้อมูลสำหรับวิจัย โดยแบ่งขั้นตอนย่อยสำหรับการเตรียม

3.2.1 การจัดหมวดหมู่และจัดข้อมูลจากเว็บไซต์

การเตรียมข้อมูลผู้วิจัยได้ใช้ข้อมูลทั้งหมดสองส่วน คือ 1) ข้อมูลบทความทั้งหมด 54 บทความจากดาต้าเบสเว็บไซต์ แบ่งออกเป็น 4 หมวดหมู่ ได้แก่ Website and Developer, Social Media, Marketing, News and Activity (Item Profile) 2) ข้อมูลผู้ใช้งานที่เข้ามาอ่านบทความในเว็บไซต์ ซึ่งเป็นข้อมูลย้อนหลังระยะเวลา 1 ปี จาก Google Analytic (User Profile) นำข้อมูล 2 ประเภทนี้ มาประมวลผลหาค่าความคล้ายคลึงของหมวดหมู่บทความกับความต้องการของผู้เขียน และ ความคล้ายคลึงบทความกับความต้องการของผู้อ่านบทความ

3.2.2 การดึงข้อมูลบทความจาก Database นำมาเก็บไว้ที่ไฟล์ CSV

การดึงข้อมูลบทความจาก Database ผู้วิจัยได้ดึงข้อมูลทั้งหมดสองส่วน คือ 1) ข้อมูลบทความทั้งหมด 54 บทความ 2) ข้อมูลผู้ใช้งานที่เข้ามาอ่านบทความในเว็บไซต์ โดยข้อมูลทั้งสองส่วนนี้จะถูกนำมาเก็บไว้ในไฟล์ CSV สำหรับไปใช้ในขั้นตอนต่อไป

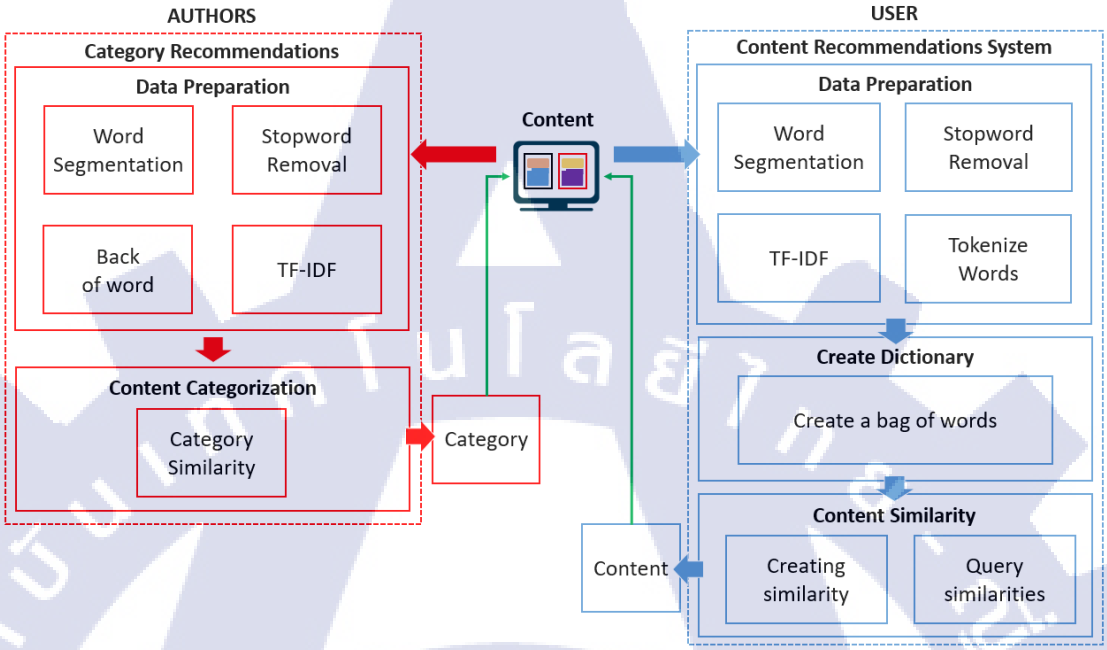
3.2.3 การกรองข้อมูล ในไฟล์ CSV

จากข้อมูลทั้งหมดสองส่วน เนื่องจากข้อมูลที่นำมาทดสอบมีบางส่วนที่ยังไม่สมบูรณ์สำหรับนำไปใช้งาน เช่น ค่าว่าง ข้อมูลที่ไม่ได้ถูกเผยแพร่หรือถูกใช้งาน เป็นต้น เมื่อคัดกรองแล้ว ทำการจัดการข้อมูลให้อยู่ในรูปโครงสร้างข้อมูลที่เหมาะสมสำหรับนำไปใช้งานระบบแนะนำ เช่น ข้อมูลที่อยู่ในรูปแบบบทความ สำหรับนำไปใช้กับระบบแนะนำหมวดหมู่สำหรับผู้ที่เขียนบทความบทเว็บไซต์ และ ระบบแนะนำบทความสำหรับผู้อ่านบนเว็บไซต์เทคนิค การกรองสารสนเทศโดยดูจากเนื้อหา (Content-Based Filtering) ส่วนข้อมูลที่อยู่ในรูปแบบตารางตัวเลข นำไปใช้กับระบบแนะนำบทความสำหรับผู้อ่านบนเว็บไซต์

3.3 ขั้นตอนที่ 3 ออกแบบและพัฒนาระบบ

การวิจัยครั้งนี้เป็นการวิจัยสำหรับระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาผลลัพธ์ที่ได้คือ ระบบแนะนำหมวดหมู่สำหรับผู้ที่เขียนบทความบทเว็บไซต์ และ

ระบบแนะนำบทความสำหรับผู้อ่านบนเว็บไซต์ โดยนำข้อมูลจากหัวข้อที่ 3.2 ขั้นตอนที่ 2 เตรียมข้อมูล มาออกแบบและพัฒนาระบบ มีขั้นตอนดังนี้



รูปที่ 3.2 แนวคิดการทำงานของระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา

3.4 ขั้นตอนที่ 4 ทดสอบและเก็บรวบรวมผล

งานวิจัยนี้ใช้เครื่องมือ Jupyter Notebook สำหรับเขียนโค้ดด้วยภาษา Python สำหรับการทดสอบและเก็บรวบรวมผลโดยใช้กับ 2 ระบบแนะนำดังนี้

3.4.1 ระบบแนะนำหมวดหมู่บทความสำหรับผู้เขียน

รวบรวมนำบทความทั้งหมด 54 บทความ ถูกแบ่งหมวดหมู่ไว้แล้ว นำบทความแบ่งเป็นข้อมูลชุด Train 80% และ แบ่งเป็นข้อมูลชุด Test 20% สำหรับทดสอบระบบแนะนำหมวดหมู่บทความสำหรับผู้เขียน

3.4.2 ระบบแนะนำบทความสำหรับผู้อ่าน

รวบรวมนำบทความทั้งหมด 54 บทความ ถูกแบ่งหมวดหมู่ไว้แล้ว นำบทความแบ่งเป็นข้อมูลชุด Train 80% และ แบ่งเป็นข้อมูลชุด Test 20% สำหรับทดสอบระบบแนะนำบทความสำหรับผู้อ่าน

3.5 ขั้นตอนที่ 5 สรุปงานวิจัย

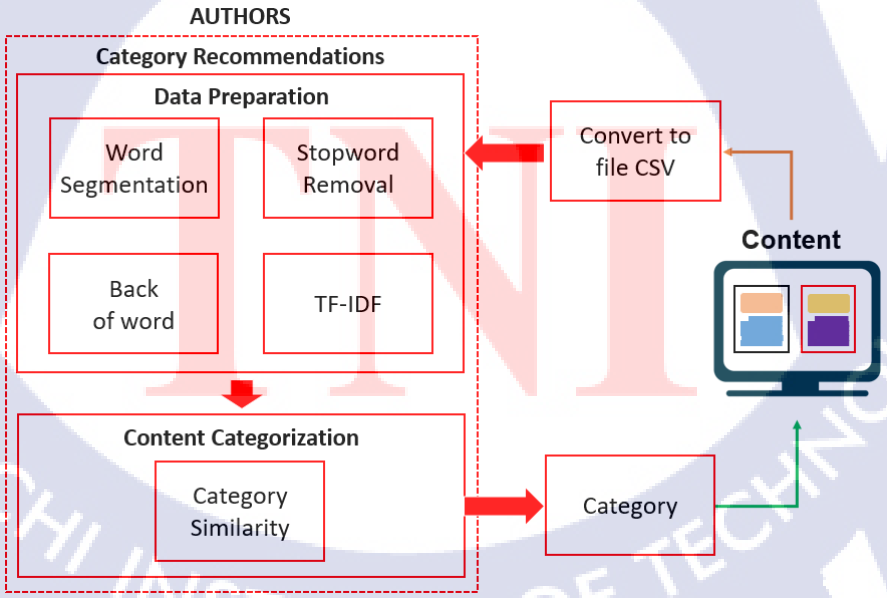
สรุปผลการวิจัยของการศึกษาและพัฒนาระบบแนะนำหมวดหมู่บทความสำหรับผู้เขียนบทความ และ ระบบแนะนำบทความสำหรับผู้อ่านบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา โดยแสดงโมเดลการเรียนรู้ของเครื่อง และแสดงผลเปอร์เซ็นความแม่นยำสำหรับนำไปใช้งานบนเว็บไซต์

บทที่ 4
ผลการวิจัย

การศึกษาวิจัยระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา นี้มีวัตถุประสงค์ในการพัฒนาระบบแนะนำเพื่อการนำไปประยุกต์ใช้กับเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาซึ่งจะประกอบด้วยบทความต่างๆจำนวนมาก โดยงานวิจัยนี้ได้นำ ข้อมูลผู้เข้าใช้งานเว็บและข้อมูลบทความจากเว็บไซต์จำนวน 54 บทความ แบ่งเป็น 4 หมวดหมู่ เพื่อนำมาจัดเตรียมให้อยู่ในรูปแบบสามารถนำมาใช้ในงานวิจัยได้ และออกแบบระบบแนะนำหมวดหมู่บทความอัตโนมัติให้กับผู้เขียนบทความ และการแนะนำบทความอัตโนมัติให้กับผู้อ่านบทความ โดยสามารถแบ่งผลการดำเนินการได้ดังต่อนี้

4.1 ระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนเทคนิค Cosine Similarity

การศึกษาระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียน ในหัวข้อนี้ผู้วิจัยได้รวบ 2 ระบบเข้าไว้ด้วยกัน คือ ระบบหาคำสำคัญในบทความ และ ระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนอธิบายรวมกันเพราะมีขั้นตอนการทำงานที่สัมพันธ์กัน โดยออกแบบแนวคิดการทำงานระบบหาคำสำคัญในบทความ และ ระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนมีภาพรวมการทำงานดังนี้



รูปที่ 4.1 ภาพรวมการทำงานของระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียน

โดยผลลัพธ์ของการพัฒนาระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนเทคนิค Cosine Similarity สามารถแบ่งได้ 4 ส่วน ได้แก่

4.1.1. ผลลัพธ์ของการเตรียมข้อมูล

4.1.2 ผลลัพธ์ของการพัฒนาระบบหาคำสำคัญในบทความ

4.1.3 ผลลัพธ์ของการพัฒนาระบบแนะนำหมวดหมู่บทความ

4.1.4 ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

4.1.1 ผลลัพธ์ของการเตรียมข้อมูล

การดึงข้อมูลบทความจาก Database นำมาเก็บไว้ที่ไฟล์ CSV ด้วยวิธีการดังต่อไปนี้

1) นำข้อมูลมาวิเคราะห์หาความสัมพันธ์ อธิบายถึงข้อมูลที่นำไปใช้ในงานวิจัย เช่น คอลัมน์แต่ละอันคือข้อมูลอะไร ชนิดของข้อมูลเป็นแบบไหน และหาทำความเข้าใจโครงสร้าง ข้อมูลทั้งหมด

2) ทำการคัดเลือกและรวบรวมข้อมูลที่เราสนใจในดาต้าเบส นำมาใช้โดยการ Export ออกมาเป็นไฟล์ CSV เพื่อเข้าสู่กระบวนการคัดรกองข้อมูล

3) นำไฟล์ CSV เข้ากระบวนการคัดกรองข้อมูลด้วยการกำจัดข้อมูลที่ไม่มีความเกี่ยวข้องกับงายวิจัยออกไป เช่น บทความที่ไม่ได้ถูกเผยแพร่ออนไลน์บนเว็บไซต์ ข้อมูลเปล่า หรือการเพิ่มข้อมูลชุดใหม่เพื่อให้ข้อมูลมีความสมบูรณ์มากขึ้น เป็นต้น

4) จัดการข้อมูลให้อยู่ในรูปโครงสร้างข้อมูลที่เหมาะสมสำหรับนำไปใช้งานโดยมีรูปแบบการเก็บข้อมูล ดังนี้

ตารางที่ 4.1 รูปแบบการเก็บข้อมูลในไฟล์ CSV ก่อนนำไปใช้ในระบบหาคำสำคัญในบทความ

หมวดหมู่	บทความ
Website and Developer	**1.แนะนำ UI DESIGN TOOLS ที่น่าจับเมาส์ลอง** แนะนำเครื่องมือในการออกแบบเว็บไซต์กันหน่อย ให้สำหรับนักออกแบบทุกคน ที่กำลังมองหาเครื่องมือใหม่ๆ..... **2. 5 สัญญาณบ่งบอกควรปรับปรุง Web** ยุคนี้คงมีน้อยคนที่ไม่รู้จัก “เว็บไซต์” (Website) เพราะถือเป็นสื่อสำคัญในการนำเสนอข้อมูลข่าวสารต่างๆ ...
Social Media	**1.เตือนภัย เมื่อ Page Facebook ปลิว** ในช่วงที่มีการอัพเดทแอปพลิเคชั่น รวมถึงการประกาศตัวเรื่อง Libra เหมือนว่าอาชญากรทางเฟซบุ๊กก็เริ่มออก..... **2.ใคร ถือ Facebook Page ในมือต้องอ่าน!!** Facebook กำลังเพิ่มสิ่งใหม่ลงไปในฟีดข่าวเพื่อตอบคำถามผู้ใช้ว่า “ทำไมถึงเห็นโพสต์นี้” (Why am I seeing.....
Marketing	**1.แนะนำทำ Content ยังไงให้ปัง !!?** ทำ Content ยังไงให้ปัง !!? เคล็ด (ไม่) ลับจาก ITLก่อนอื่นเราต้องเข้าใจก่อนว่าคำว่า Content จริงๆ ..... **2. 3 วิธีการทำให้คอนเทนต์โดดเด่น** การแข่งขันใน ‘คอนเทนต์มาร์เกตติ้ง’ มีการแย่งชิงพื้นที่และความสนใจที่สูงมากขึ้น บางเพจอาจจะทำ.....
News and Activity	**1. ช็อปช่วยชาติ มีอะไรมาช่วยลดหย่อนภาษีปี 2561 บ้าง?** เข้าใกล้เทศกาลสิ้นปีการยื่นภาษีรายได้บุคคลธรรมดากำลังจะมาถึง หลายๆคนอาจจะสงสัยกันนะคะว่า..... **2. ThaiSalary โปรแกรมคำนวณเงินเดือนออนไลน์** ThaiSalary มีประโยชน์ดีๆของการทำโปรแกรมคำนวณ.....

4.1.2 ผลลัพธ์ของการพัฒนาระบบหาคำสำคัญในบทความ

1) การแบ่งคำภาษาไทย (Word Segmentation) นำข้อมูลที่ได้จากการเตรียมข้อมูลในรูปแบบไฟล์ CSV เข้ากระบวนการแบ่งคำภาษาไทยโดยใช้เครื่อง Jupyter Notebook เขียนโค้ด Python เพื่อติดต่อและใช้งาน Library Pythainlp ในการแบ่งคำด้วยวิธีการดึงข้อมูลจาก Dictionary ของ Library Pythainlp ออกมาเพื่อทำการเปรียบเทียบคำระหว่างข้อมูลทั้งสองแหล่ง ผลลัพธ์ของการแบ่งคำภาษาไทยตามหมวดหมู่บทความของเว็บไซต์ จะได้ดังตัวอย่างจากตารางที่ 4.2 ตัวอย่างการแบ่งคำภาษาไทยของบทความทั้ง 4 หมวดหมู่ ดังนี้

ตารางที่ 4.2 ตัวอย่างการแบ่งคำภาษาไทยของบทความทั้ง 4 หมวดหมู่

หมวดหมู่	บทความ
Website and Developer	"แนะนำ","UI","DESIGN TOOLS","ที่","น่า","จับ","เมาส์","ลอง","แนะนำ","เครื่องมือ","ใน","การ","ออกแบบ","เว็บไซต์","กัน","หน่อย","ให้","สำหรับ","นัก","ออกแบบ","ทุก","คน","ที่","กำลัง"...
Social Media	"เตือน","ภัย","เมื่อ""Page","Facebook","ปลิว","ใน","ช่วง","ที่","มี","การ","อัพเดท","แอปพลิเคชั่น","รวม","ถึง","การ","ประกาศ","ตัว","เรื่อง","Libra","เหมือน","ว่า","อาชญากร","ทาง","เฟซบุ๊ก","ก็","เริ่ม","ออกอาละวาด",.....
Marketing	“แนะนำ”,"ทำ","Content","ยัง","ไง","ให้","ปัง","!""!","?","ทำ","Content","ยัง","ไง","ให้","ปัง","!""!","?","เคล็ด","(","ไม่",")","ลับ","จาก","ITL","ก่อนอื่น","เรา","ต้อง","เข้า","ใจ","ก่อน","ว่า","คำ","ว่า" "Content","จริง","ๆ" .....
News and Activity	"ช็อป","ช่วย","ชาติ","มี","อะไร","มา","ช่วย","ลด","หย่อน","ภาษี","ปี","2561","บ้าง","?","เข้าใกล้","เทศกาล","สิ้น","ปี","การ","ยื่น","ภาษี","รายได้","บุคคล","ธรรมดา","กำลัง","จะ","มา","ถึง","หลาย","ๆ","คน","อาจ","จะ","สงสัย","กัน","นะ","คะ","ว่า".....

2) การกำจัดคำหยุด (Stopword Removal) หลังจากผ่านกระบวนการแบ่งคำภาษาไทย (Word Segmentation) แล้วผู้วิจัยนำข้อมูลเข้ากระบวนการกำจัดคำหยุด (Stopword Removal) เพื่อกำจัดคำที่ไม่มีความสำคัญออกไปจากข้อมูลยกตัวคำหยุดเช่น “ที่”, “มี”, “น่า” เป็นต้น ดังแสดงตารางที่ 4.3

ตารางที่ 4.3 ตัวอย่างข้อความที่ผ่านกระบวนการกำจัดคำหยุด (Stopword Removal)

หมวดหมู่	บทความ
Website and Developer	'แนะนำ', 'UI', 'DESIGN TOOLS', 'เมาส์', 'ลอง', 'แนะนำ', 'เครื่องมือ', 'ออกแบบ', 'เว็บไซต์', 'สำหรับ', 'นักออกแบบ', 'มองหา', 'เครื่อง', 'มือใหม่',.....
Social Media	'เตือนภัย', 'Page', 'Facebook', 'ปลิว', 'อัพเดท', 'แอปพลิเคชั่น', 'ประกาศตัว', 'เรื่อง', 'Libra', 'เหมือนว่า', 'อาชญากร', 'เฟซบุ๊ก', 'อาละวาด',.....
Marketing	'แนะนำ','ทำ', 'Content', 'ปัง', 'ทำ', 'Content', 'ปัง', 'ทำ', 'Content', 'ปัง', 'เคล็ด', 'ลับ', 'ITL', 'ก่อนอื่น',.....
News and Activity	'ช็อป', 'ชาติ', 'ลดหย่อน', 'ภาษี', 'ปี', 'เข้าใกล้', 'เทศกาล', 'สิ้นปี', 'ยื่น', 'ภาษีรายได้', 'บุคคลธรรมดา', 'มาถึง', 'คน', 'สงสัย', 'นะคะ', 'ปี',.....

3) สร้าง Bag of word รวบรวมคำที่ผ่านขั้นตอนคัดกรองข้างต้นมาเก็บรวบรวมไว้ในคลังคำ โดยการแปลงคำที่ไม่ซ้ำกันแกมาเป็น unique ตามตัวอย่างรูปที่ 4.1

'แนะนำ': 1395 , ‘ui': 364 ,‘design tool': 80 ,'เมาส์': 1323 ,'ลอง': 1026

รูปที่ 4.1 ตัวอย่าง การแปลงคำ เป็น unique ID

4) TF-IDF คำสำคัญในบทความของแต่ละหมวดหมู่ มีขั้นตอนในการหาคำสำคัญในเอกสารโดยใช้สูตรต่อไปนี้

$$TF \times IDF = freq(T, D) \times \log_2 \frac{N}{n_p} \quad (4\text{-}1)$$

โดยเริ่มต้นคำนวณหาค่า TF คือ จำนวน คำ = T (Term)" ที่ปรากฏในหนึ่ง เอกสาร = D (Document) โดยใช้สูตร

$$TF = freq(T, D) \quad (4\text{-}2)$$

วิธีการคำนวณ ในแต่ละหมวดหมู่นำส่วนบทความที่ถูกแบ่งคำและกำจัดคำหยุดออก มากำหนดชื่อเอกสารแทนหมวดหมู่ดังนี้

Document 1 = Website and Developer

Document 2 = Social Media

Document 3 = Marketing

Document 4 = News and Activity

วัตถุประสงค์ของการแทนชื่อเอกสารเพราะชื่อหมวดหมู่ค่อนข้างยาวเมื่อยกตัวอยากเปรียบเทียบกับสูตรทำให้เข้าใจยากต่อการทำความเข้าใจ

หลังจากแทนชื่อเอกสารแล้ว จะนับ คำ(Term) = "แนะนำ" ที่ปรากฏในเอกสาร D1 กี่ครั้ง แล้วนำมาหารด้วยจำนวนคำที่มีในเอกสาร D1 ทั้งหมด ก็จะได้ค่า TF สำหรับการคำนวณค่า TF ในตารางที่ 4.4 เป็นการยกตัวอย่างง่ายๆ ด้วยการกำหนดคำทั้งหมดในเอกสาร D1 มีทั้งหมด 9 คำ รวมคำที่ซ้ำกันเข้าด้วยกัน คำว่า "แนะนำ" ปรากฏในเอกสาร D1 สองครั้งจะได้ค่า $TF =$ 2/9 =0.22 ส่วนคำที่เหลือ ปรากฏในเอกสาร D1 หนึ่งครั้ง $TF =$ 1/9 = 0.11 ตามตารางที่ 4.4

ตารางที่ 4.4 ตัวอย่างการคำนวณและผลลัพธ์การหาค่า TF

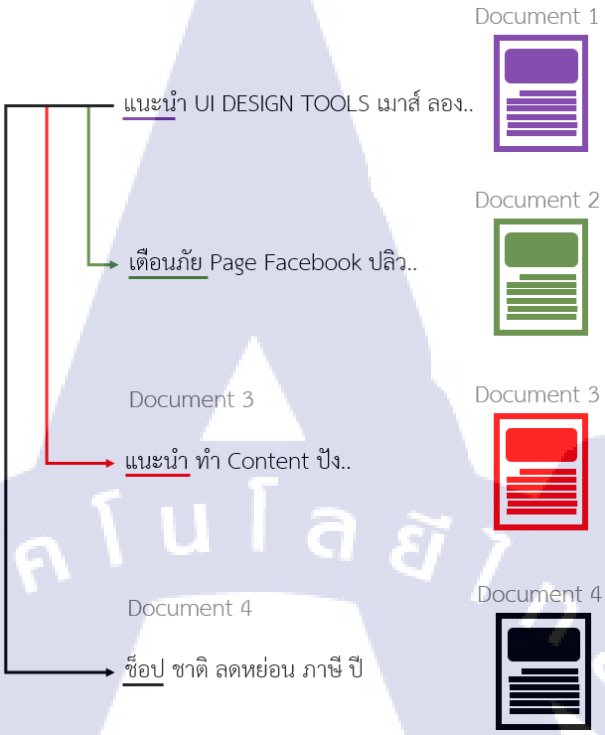
ลำดับ	Word	D1	TF
1	แนะนำ	2/9	0.22
2	UI	1/9	0.11
3	DESIGN TOOLS	1/9	0.11
4	เมาส์	1/9	0.11
5	ลอง	1/9	0.11
6	เครื่องมือ	1/9	0.11
7	ออกแบบ	1/9	0.11
8	เว็บไซต์	1/9	0.11
9	สำหรับ	1/9	0.11

วัตถุประสงค์ของการทำ TF คือ เพื่อหาคำสำคัญในเอกสารนั้นๆโดยถ้าค่า TF มีค่าหรือผลลัพธ์ใกล้เคียงหมายเลขหนึ่งก็จะมีความสำคัญมากที่สุด หลังจากได้ค่า TF จะเป็นการคำนวณหาค่า IDF ดังนี้

การหาค่า IDF คือ การวัดความสำคัญของ คำ(term) ใน เอกสาร(document)ของหมวดหมู่ทั้งหมด โดยใช้สูตร

$$IDF = \log_2 \frac{N}{n_p} \tag{4-3}$$

ทำการตรวจสอบความเหมือนกัน และไม่เหมือนกันของคำในแต่ละ document จากรูปที่ 4.2 เป็นตัวอย่างวิธีการที่ระบบค้นหาคำที่เหมือนกันในเอกสารอื่นโดยการเขาไปสืบค้นที่ละเอกสาร



รูปที่ 4.2 วิธีการตรวจสอบความเหมือน และ ไม่เหมือนกันของคำในทุก Document

หลังเปรียบเทียบแล้ว ถัดไปเป็นวิธีการทำงาน นำจำนวนเอกสารทั้งหมด 4 หมวดหมู่มาหารด้วย จำนวน document ที่มีคำคำนั้นอยู่ แล้วผลลัพธ์ไปหาค่าlog ฐาน 2 ตัวอย่างเช่น คำว่า “แนะนำ” จะพบคำนี้ใน Document 1 เท่านั้นอ้างอิงจากรูปที่ 4-1 โดยมีขั้นตอนในการหาผลลัพธ์ดังนี้ จำนวนเอกสารทั้งหมด 4 เอกสารดังนั้น $N = 4$ จำนวนคำที่เกิดขึ้นในเอกสาร หรือ Document ไหนบ้าง เช่น “แนะนำ” เกิดขึ้นในเอกสารที่หนึ่ง หรือ Document 1 และเอกสารที่สาม หรือ Document 3 สองคำ แต่ในเอกสารอื่นไม่มีคำนี้ปรากฏอยู่ดังนั้นคำว่า “แนะนำ” มีค่า $n_p = 2$ ดังตารางที่ 4.5

ตารางที่ 4.5 ตัวอย่างคำนวณ inverse document frequency สำหรับ Document1

Word	IDF	Result
แนะนำ	Log(4/2)	0.30
UI	Log(4/1)	0.60
DESIGN TOOLS	Log(4/1)	0.60
เมาส์	Log(4/1)	0.60
ลอง	Log(4/1)	0.60
เตือนภัย	Log(4/1)	0.60
Page	Log(4/1)	0.60
Facebook	Log(4/1)	0.60
ปลิว	Log(4/1)	0.60
ทำ	Log(4/1)	0.60
Content	Log(4/1)	0.60
ปัง	Log(4/1)	0.60
ช็อป	Log(4/1)	0.60
ชาติ	Log(4/1)	0.60
ลดหย่อน	Log(4/1)	0.60
ภาษี	Log(4/1)	0.60
ปี	Log(4/1)	0.60

วัตถุประสงค์ของการทำ IDF คือ ใช้สำหรับวัดความสำคัญของ คำ(term) ใน เอกสาร (document) ทั้งหมด คือ ถ้าเจอคำในหลายๆเอกสารคำนั้นจะมีความสำคัญลดลง หลังจากนั้นนำค่า TF และ IDF มาคูณกันเพื่อค่าผลลัพธ์ หรือ TF-IDF Weight ของคำนั้น ดังตารางที่ 4.6 นี้

ตารางที่ 4.6 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF

Word	TF					IDF	TF*IDF			
	D1	D2	D3	D4			D1	D2	D3	D4
แนะนำ	0.22	0	0	0.2		Log(4/2)=0.30	0.066	0	0	0.06
UI	0.11	0	0	0		Log(4/1)=0.60	0.66	0	0	0
DESIGN TOOLS	0.11	0	0	0		Log(4/1)=0.60	0.66	0	0	0
เมาส์	0.11	0	0	0		Log(4/1)=0.60	0.66	0	0	0
ลอง	0.11	0	0	0		Log(4/1)=0.60	0.66	0	0	0
เตือนภัย	0	0.25	0	0		Log(4/1)=0.60	0	0.15	0	0
Page	0	0.25	0	0		Log(4/1)=0.60	0	0.15	0	0
Facebook	0	0.25	0	0		Log(4/1)=0.60	0	0.15	0	0
ปลิว	0	0.25	0	0		Log(4/1)=0.60	0	0.15	0	0
ทำ	0	0	0.33	0		Log(4/1)=0.60	0	0	0.2	0
Content	0	0	0.33	0		Log(4/1)=0.60	0	0	0.2	0
ปัง	0	0	0.33	0		Log(4/1)=0.60	0	0	0.2	0
ช็อป	0	0	0	0		Log(4/1)=0.60	0	0	0	0
ชาติ	0	0	0	0.2		Log(4/1)=0.60	0	0	0	0.12
ลดหย่อน	0	0	0	0.2		Log(4/1)=0.60	0	0	0	0.12
ภาษี	0	0	0	0.2		Log(4/1)=0.60	0	0	0	0.12
ปี	0	0	0	0.2		Log(4/1)=0.60	0	0	0	0.12

วัตถุประสงค์ของการทำ TF-IDF เพื่อ กรองคำที่ปรากฏทุกเอกสารจะมีความสำคัญน้อยกว่าคำที่ปรากฏในเอกสารเดียวจะมีความสำคัญมากกว่า บงบอกได้ว่าเอกสารนั้นเป็นเรื่องเกี่ยวกับอะไร

4.1.3 ผลลัพธ์ของการพัฒนาระบบแนะนำหมวดหมู่บทความ

1) Category Similarity การคำนวณหาความเหมือนของเอกสาร โดยใช้สูตร Cosine Similarity ดังนี้

$$similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4\text{-}4)$$

สำหรับการคำนวณขั้นตอนนี้จะเพิ่มบทความเข้ามาอีกหนึ่งบทความเพื่อทำการเปรียบเทียบความคล้ายคลึงกับหมวดหมู่ต้องการแนะนำให้ผู้เขียนบทความเลือก ดังรูปที่ ตารางที่ 4.7

ตารางที่ 4.7 ตัวอย่างผลลัพธ์การคำนวณ Cosine Similarity

Document	TF-IDF	Cosine Similarity
แนะนำ UI DESIGN TOOLS ที่น่าจับเมาส์ลอง.....	[0.066,0.66,0.66,0.66,0.66,0,0,0,0,0,0,0,0,0,0,0]	0.41
เตือนภัย เมื่อ Page Facebook ปลิว.....	[0,0,0,0,0,0.15,0.15,0.15,0.15,0,0,0,0,0,0,0	0
แนะนำทำ Content ยังไงให้ปัง !!?.....	[0,0,0,0,0,0,0,0,0,0.2,0.2,0.2,0,0,0,0,0]	0
ช็อปช่วยชาติ มีอะไรมาช่วยลดหย่อนภาษีปี 2561บ้าง?.....	[0.06,0,0,0,0,0,0,0,0,0,0,0,0,0.1,0.1,0.1,0.1]	0
4 เหตุผลที่ควรปรับปรุงเว็บไซต์...	[0.06,0,0.06,0,0.06,0,0,0,0,0,0,0,0,0.1,0.1,0.1,0]	1

4.1.4 ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

1) ขั้นตอนถัดมาผู้วิจัยได้นำกระบวนการและการออกแบบแนวคิดการทำงานระบบหาคำสำคัญในบทความและค้นหาแนะนำหมวดหมู่บทความ มาเขียนเป็นโค้ดด้วยภาษา Python เพื่อให้เห็นขั้นตอนและผลลัพธ์ในการแนะนำหมวดหมู่บทความให้กับผู้เขียน

```
with open ('Content_develop_test.csv',encoding='utf8') as f:
    tokens = sent_tokenize(f.read())
    for line in tokens:
        file2_docs.append(line)

print("Number of documents:",len(file2_docs))
for line in file2_docs:
    query_doc = [w.lower() for w in word_tokenize(line)]

    query_doc_bow = dictionary.doc2bow(query_doc)

    query_doc_tf_idf = tf_idf[query_doc_bow]

    print('Comparing Result:', sims[query_doc_tf_idf])
```

```
Number of documents: 1
Comparing Result: [0.41217005 0.02074152 0.08239897 0.00780618]
```

รูปที่ 4.3 ตัวอย่างโค้ดภาษา Python ผลลัพธ์การคำนวณ Cosine Similarity

จากรูปที่ 4.3 มี คะแนนอิงตามหมวดหมู่เอกสาร คือ 1) Website and Developer 2) Social Media 3) Marketing 4) News and Activity โดยคะแนนความคล้ายคลึงของบทความที่คล้ายกับหมวดหมู่ Website and Developer เปอร์เซ็นความคล้ายคลึงอยู่ที่ 41% การแสดงผลข้อมูลแนะนำ Category หรือ หมวดหมู่บทความให้ผู้เขียนนำไปใช้กับบทความเกี่ยวข้องโดยระบบจะแสดงหมวดหมู่บทความที่หนึ่งถึงสี่โดยมีคะแนนความคล้ายคลึงแสดงคู่กับหมวดหมู่เพื่อให้ผู้เขียนบทความนำไปใช้กำหนดหมวดหมู่ให้กับบทความใหม่

4.2 ระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่านบนเว็บไซต์

การศึกษาระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่าน ในหัวข้อนี้ผู้วิจัยได้ออกแบบโดยใช้ 2 เทคนิคดังนี้

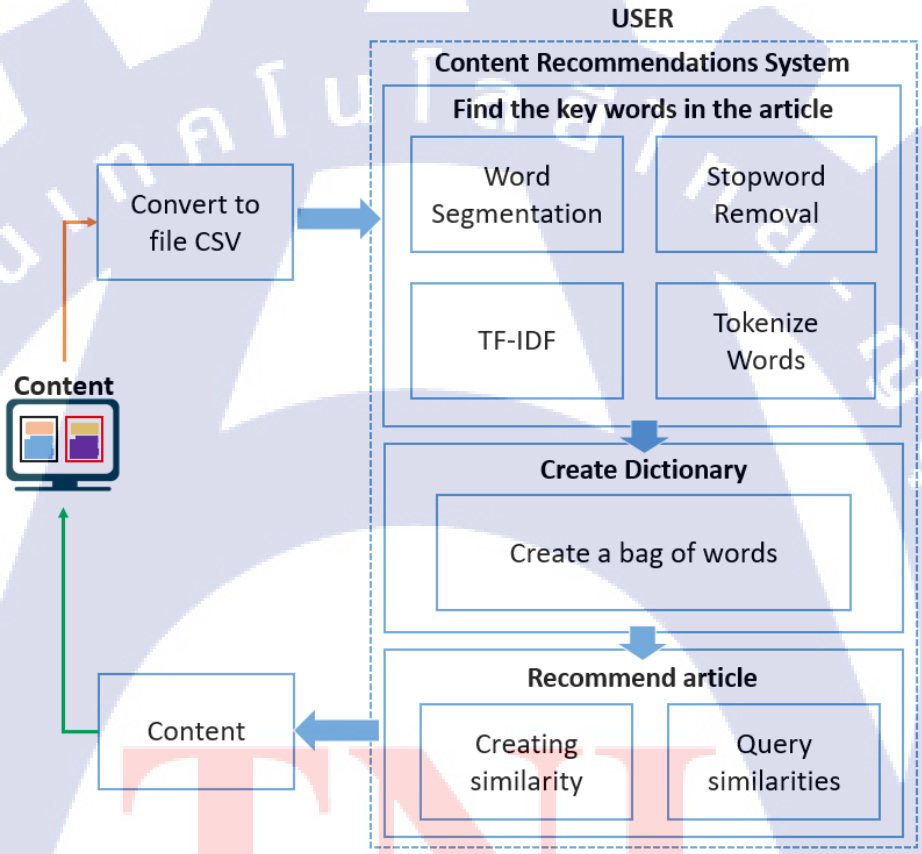
4.2.1 เทคนิค Content-Based Filtering โดยใช้รูปแบบ Cosine similarity

4.2.2 เทคนิค Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering

โดยอธิบายการออกแบบและวิจัย ระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่านบนเว็บไซต์ ทั้งสองเทคนิคดังนี้

4.2.1 เทคนิค Content-Based Filtering โดยใช้รูปแบบ Cosine similarity

การศึกษาเทคนิค Content-Based Filtering โดยใช้รูปแบบ Cosine similarity ในหัวข้อนี้ผู้วิจัยได้อธิบายมีขั้นตอนการทำงานและออกแบบแนวคิดการทำงานมีภาพรวมการทำงานดังนี้



รูปที่ 4.4 ภาพรวมการทำงานเทคนิค Content-Based Filtering โดยใช้รูปแบบ Cosine similality

โดยผลลัพธ์ของการพัฒนาระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่านด้วยเทคนิค Content-Based Filtering โดยใช้รูปแบบ Cosine similarity สามารถแบ่งได้ 4 ส่วน ได้แก่

1) ผลลัพธ์ของการเตรียมข้อมูล ซึ่งเป็นขั้นตอนการดึงข้อมูลและนำมาประมวลผลการนำไปใช้ในการพัฒนาระบบ

2) ผลลัพธ์ของการพัฒนาระบบหาคำสำคัญในบทความ

3) ผลลัพธ์ของการพัฒนาระบบแนะนำบทความ

4) ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

1) ผลลัพธ์ของการเตรียมข้อมูล ซึ่งเป็นขั้นตอนการดึงข้อมูลและนำมาประมวลผลการนำไปใช้ในการพัฒนาระบบ

การดึงข้อมูลบทความจาก Database นำมาเก็บไว้ที่ไฟล์ CSV ด้วยวิธีการดังต่อไปนี้

1.1) นำข้อมูลมาวิเคราะห์หาความสัมพันธ์ อธิบายถึงข้อมูลที่นำไปใช้ในงานวิจัย เช่น คอลัมน์แต่ละอันคือข้อมูลอะไร ชนิดของข้อมูลเป็นแบบไหน และหาทำความเข้าใจโครงสร้างข้อมูลทั้งหมด

1.2) ทำการคัดเลือกและรวบรวมข้อมูลที่เราสนใจในดาต้าเบส นำมาใช้โดยการ Export ออกมาเป็นไฟล์ CSV เพื่อเข้าสู่กระบวนการคัดรกองข้อมูล

1.3) นำไฟล์ CSV เข้ากระบวนการคัดกรองข้อมูลด้วยการกำจัดข้อมูลที่ไม่มีความเกี่ยวข้องกับงายวิจัยออกไป เช่น บทความที่ไม่ได้ถูกเผยแพร่ออนไลน์บนเว็บไซต์ ข้อมูลเปล่า หรือการเพิ่มข้อมูลชุดใหม่เพื่อให้ข้อมูลมีความสมบูรณ์มากขึ้น เป็นต้น

1.4) จัดการข้อมูลให้อยู่ในรูปโครงสร้างข้อมูลที่เหมาะสมสำหรับนำไปใช้งานโดยมีรูปแบบการเก็บข้อมูล ดังนี้

ตารางที่ 4.8 รูปแบการเก็บข้อมูลในไฟล์ CSV ก่อนนำไปใช้ในระบบแนะนำบทความ

Title	Content
10 สิ่งที่น่าสนใจในการออกแบบเว็บไซต์	ในปัจจุบันเราได้เห็นความก้าวหน้าทางเทคโนโลยีที่ช่วยเชื่อมโยงผู้คนได้ดีมากยิ่งขึ้น เว็บไซต์เป็นหนึ่งในเครื่องมือเหล่านั้นซึ่งมีการพัฒนาอยู่ตลอดเวลา และจากการที่เว็บไซต์สามารถเข้าถึงผู้คนได้จากทั่วทุกมุมโลกจึงมีเว็บไซต์เกิดขึ้นจำนวนมาก แล้วเราจะทำอย่างไรให้เว็บไซต์ของเราน่าสนใจและดึงดูดกลุ่มเป้าหมาย...
10 ออกแบบ Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ	หลายๆ ธุรกิจอยากมีเว็บไซต์เป็นของตัวเอง แต่ไม่รู้ว่าควรเริ่มอย่างไร มีแล้วจะตอบโจทย์ธุรกิจของตัวเองไหม ใครที่มีคำถามกับการทำเว็บไซต์แบบนี้ ลองมาดูปัจจัยที่จะช่วยให้เว็บไซต์ออกมาสำเร็จตามเป้าหมายที่วางไว้...
5 สัญญาณบ่งบอกควรปรับปรุง Web	ยุคนี้คงมีน้อยคนที่ไม่รู้จัก “เว็บไซต์” (Website) เพราะถือเป็นสื่อสำคัญในการนำเสนอข้อมูลข่าวสารต่างๆ ผ่านทางอินเตอร์เน็ต ซึ่งมีองค์ประกอบหลายอย่าง โดยเว็บไซต์ จะมีหน้าเว็บเพจหลายหน้าที่เชื่อมโยงเข้ากับลิ้งค์ เพื่อให้สามารถเปิดไปยังหน้าเพจต่างๆ ได้อย่างง่ายดายและถูกจัดเก็บไว้ใน www. (เวิลด์ไวด์เว็บ) โดยส่วนใหญ่มีทั้งเว็บไซต์ที่เปิดให้เข้าชมฟรี...
7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์	Bounce Rate คือ อัตราส่วนของผู้เข้าชมเว็บไซต์เพียงหน้าเดียวแล้วปิดออกไป ซึ่งถ้าค่านี้มีอัตราที่สูงจะถูกจัดอันดับว่าเป็นเว็บไซต์ที่ไม่มีคุณภาพ หรือเนื้อหาไม่น่าสนใจ...

ตารางที่ 4.8 รูปแบการเก็บข้อมูลในไฟล์ CSV ก่อนนำไปใช้ในระบบแนะนำบทความ (ต่อ)

Title	Content
4 เหตุผลที่ควรปรับปรุงเว็บไซต์	จากการศึกษาของ Nielsen Norman Group เรามีเวลาน้อยกว่า 59 วินาที ในการดึงดูดให้ผู้ใช้อยู่บนเว็บไซต์ หากไม่สามารถทำได้เราจะสูญเสียผู้ชมเหล่านั้นไปในทันที...

2) ผลลัพธ์ของการพัฒนาระบบหาคำสำคัญในบทความ

มีขั้นตอนดังนี้

2.1) การแบ่งคำภาษาไทย (Word Segmentation) ขั้นตอนนี้เหมือนการตัดคำระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียน ส่วนที่แตกต่างกันคือการจัดเรียงข้อมูลนำเข้าจะถูกแยกออกเป็น 2 Column คือ Title และ Content และตัดคำตามเรื่องที่วางไว้โดยไม่มีการกำหนดหมวดหมู่ ผลลัพธ์การตัดคำตามตารางที่ 4.9

ตารางที่ 4.9 การแบ่งคำภาษาไทย ชุดข้อมูลสำหรับทำเทคนิค Content-Based Filtering

Title	Content
'10', 'สิ่ง', 'ที่', 'น่าสนใจ', 'ใน', 'การ', 'ออกแบบ', 'เว็บไซต์'	'ใน', 'ปัจจุบัน', 'เรา', 'ได้', 'เห็น', 'ความก้าวหน้า', 'ทางเทคโนโลยี', 'ที่', 'ช่วย', 'เชื่อมโยง', 'ผู้คน', 'ได้ดี', 'มากยิ่งขึ้น', 'เว็บไซต์', 'เป็นหนึ่ง', 'ใน', 'เครื่องมือ', 'เหล่านั้น', 'ซึ่ง', 'มี', 'การพัฒนา', 'อยู่', 'ตลอดเวลา', 'และ', 'จาก', 'การ', 'ที่', 'เว็บไซต์', 'สามารถ', 'เข้าถึง', 'ผู้คน', 'ได้', 'จาก', 'ทั่ว', 'ทุก', 'มุม', 'โลก', 'จึง', 'มี', 'เว็บไซต์', 'เกิดขึ้น', 'จำนวนมาก', 'แล้ว', 'เรา', 'จะ', 'ทำ', 'อย่างไร', 'ให้', 'เว็บไซต์', 'ของ', 'เรา', 'น่าสนใจ', 'และ', 'ดึงดูด', 'กลุ่มเป้าหมาย'

ตารางที่ 4.9 การแบ่งคำภาษาไทย ชุดข้อมูลสำหรับทำเทคนิค Content-Based Filtering (ต่อ)

Title	Content
'10','ออกแบบ', 'check-lists', 'ทำ', 'เว็บไซต์', 'ให้', 'ตอบ', 'โจทย์', 'ธุรกิจ'	'หลาย', 'ๆ', 'ธุรกิจ', 'อยาก', 'มี', 'เว็บไซต์', 'เป็น', 'ของ', 'ตัวเอง', 'แต่', 'ไม่รู้', 'ว่า', 'ควร', 'เริ่ม', 'อย่างไร', 'มี', 'แล้', 'วจะ', 'ตอบ', 'โจทย์', 'ธุรกิจ', 'ของ', 'ตัวเอง', 'ไหม', 'ใคร', 'ที่', 'มี', 'คำถาม', 'กับ', 'การ', 'ทำ', 'เว็บไซต์', 'แบบนี้', 'ลอง', 'มา', 'ดู', 'ปัจจัย', 'ที่ จะ', 'ช่วย', 'ให้', 'เว็บไซต์', 'ออกมา', 'สำเร็จ', 'ตาม', 'เป้าหมาย', 'ที่', 'วาง', 'ไว้'
'5', 'สัญญาณ', 'บ่งบอก', 'ควร', 'ปรับปรุง', 'Web'	'ยุค', 'นี้', 'คง', 'มี', 'น้อย', 'คน', 'ที่', 'ไม่', 'รู้จัก', '“', 'เว็บไซต์', '”', '(', 'website', ')', 'เพราะ', 'ถือเป็น', 'สื่อ', 'สำคัญ', 'ใน', 'การ', 'นำเสนอ', 'ข้อมูลข่าวสาร', 'ต่างๆ', 'ผ่าน', 'ทาง', 'อินเตอร์เน็ต', 'ซึ่ง', 'มี', 'องค์ประกอบ', 'หลายอย่าง', 'โดย', 'เว็บไซต์', 'จะ', 'มีหน้า', 'เว็บเพจ', 'หลาย', 'หน้าที่', 'เชื่อมโยง', 'เข้ากับ', 'ลิ้งค์', 'เพื่อให้', 'สามารถ', 'เปิด', 'ไป', 'ยัง', 'หน้า'...
'7', 'ขั้นตอน', 'ลด', 'ค่า', 'bounce', 'rate', 'บน', 'หน้า', 'เว็บไซต์'	'bounce', 'rate', 'คือ', 'อัตราส่วน', 'ของ', 'ผู้เข้าชม', 'เว็บไซต์', 'เพียง', 'หน้า', 'เดียว', 'แล้ว', 'ปิด', 'ออก', 'ไป', 'ซึ่ง', 'ถ้า', 'ค่า', 'นี้', 'มี', 'อัตรา', 'ที่สูง', 'จะ', 'ถูก', 'จัดอันดับ', 'ว่า', 'เป็น', 'เว็บไซต์', 'ที่', 'ไม่มี', 'คุณภาพ', 'หรือ', 'เนื้อ', 'หาไม่', 'น่าสนใจ'
'4', 'เหตุผล', 'ที่', 'ควร', 'ปรับปรุง', 'เว็บไซต์'	'จาก', 'การศึกษา', 'ของ', 'nielsen', 'norman', 'group', 'เรา', 'มี', 'เวลา', 'น้อยกว่า', '59', 'วินาที', 'ใน', 'การ', 'ดึงดูด', 'ให้', 'ผู้ใช้', 'อยู่', 'บน', 'เว็บไซต์', 'หาก', 'ไม่', 'สามารถ', 'ทำได้', 'เรา', 'จะ', 'สูญเสีย', 'ผู้ชม', 'เหล่านั้น', 'ไป', 'ในทันที'

สำหรับการเขียนโค้ดการแบ่งคำภาษาไทย (Word Segmentation) ของระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่าน ใช้ชุดโค้ดแบบเดียวกับระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนแตกต่างตรงมีข้อมูลภายในไฟล์ ตามตัวอย่างตารางที่ 4.9

2.2) การกำจัดคำหยุด (Stopword Removal) ในขั้นตอนการกำจัดคำหยุดจะเหมือนกับระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนจะแตกต่างตรงที่ข้อมูลที่ใช้จะถูกตัดคำหยุดตามชื่อเรื่องหัวข้อโดยไม่ได้อ้างอิงตามหมวดหมู่ดังตารางที่ 4.10

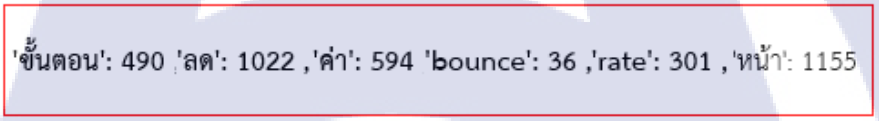
ตารางที่ 4.10 การกำจัดคำหยุด (Stopword Removal)

Title	Content
'น่าสนใจ', 'ออกแบบ', 'เว็บไซต์'	'ความก้าวหน้า', 'ทางเทคโนโลยี', 'เชื่อมโยง', 'ผู้คน', 'ได้ดี', 'มากยิ่งขึ้น', 'เว็บไซต์', 'เป็นหนึ่ง', 'เครื่องมือ', 'การพัฒนา', 'เว็บไซต์', 'เข้าถึง', 'ผู้คน', 'มุม', 'โลก', 'เว็บไซต์', 'เกิดขึ้น', 'จำนวนมาก', 'ทำ', 'เว็บไซต์', 'น่าสนใจ', 'ดึงดูด', 'กลุ่มเป้าหมาย'...
‘ออกแบบ’,'CheckLists', 'ทำ', 'เว็บไซต์', 'ตอบ', 'โจทย์', 'ธุรกิจ'	'ธุรกิจ', 'เว็บไซต์', 'ตัวเอง', 'ไม่รู้', 'แล้ว', 'ตอบ', 'โจทย์', 'ธุรกิจ', 'ตัวเอง', 'ไหม', 'คำถาม', 'ทำ', 'เว็บไซต์', 'แบบนี้', 'ลอง', 'ดู', 'ปัจจัย', 'ที่จะ', 'เว็บไซต์', 'ออกมา', 'สำเร็จ', 'เป้าหมาย', 'วาง'...
'สัญญาณ', 'บ่งบอก', 'ปรับปรุง', 'Web'	'ยุค', 'คน', 'รู้จัก', 'เว็บไซต์', 'Website', 'ถือเป็น', 'สื่อ', 'นำเสนอ', 'ข้อมูลข่าวสาร', 'อินเตอร์เน็ต', 'องค์ประกอบ', 'หลายอย่าง', 'เว็บไซต์', 'มีหน้า', 'เว็บเพจ', 'หน้าที่', 'เชื่อมโยง', 'เข้ากับ', 'ลิ้งค์', 'หน้า', 'เพจ', 'อย่างง่ายดาย', 'จัดเก็บ'...

ตารางที่ 4.10 การกำจัดคำหยุด (Stopword Removal) (ต่อ)

Title	Content
ขั้นตอน', 'ลด', 'ค่า', 'Bounce', 'Rate', 'หน้า', 'เว็บไซต์'	'Bounce', 'Rate', 'อัตราส่วน', 'ผู้เข้าชม', 'เว็บไซต์', 'หน้า', 'ค่า', 'อัตรา', 'ที่สูง', 'จัดอันดับ', 'เว็บไซต์', 'ไม่มี', 'คุณภาพ', 'เนื้อ', 'หาไม่', 'น่าสนใจ'...
'ปรับปรุง', 'เว็บไซต์'	'การศึกษา', 'Nielsen', 'Norman', 'Group', 'เวลา', 'วินาที', 'ดึงดูด', 'ผู้ใช้', 'เว็บไซต์', 'ทำได้', 'สูญเสีย', 'ผู้ชม', 'ในทันที'...

2.3) สร้าง Bag of word รวบรวมคำที่ผ่านขั้นตอนคัดกรองข้างต้นมาเก็บรวบรวมไว้ในคลังคำ โดยการแปลงคำที่ไม่ซ้ำกันแกมาเป็น unique ผู้วิจัยได้ดึงหัวข้อบทความมาแสดงเป็นตัวอย่างตามรูปที่ 4.1



รูปที่ 4.5 ตัวอย่าง การแปลงคำ เป็น unique ID

สำหรับการกำจัดคำหยุด (Stopword Removal) ของระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่าน ใช้ชุดโค้ดแบบเดียวกับระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียนแตกต่างตรงมีข้อมูลภายในไฟล์ ตามตัวอย่างตารางที่ 4-10

2.4) TF-IDF คำสำคัญในบทความของ มีขั้นตอนในการหาคำสำคัญในเอกสารโดยใช้สูตรต่อไปนี้

$$TF \times IDF = freq(T, D) \times \log_2 \frac{N}{n_p} \quad (4\text{-}5)$$

โดยเริ่มต้นคำนวณหาค่า TF คือ จำนวน คำ = T (Term)” ที่ปรากฏในหนึ่ง เอกสาร = D (Document) โดยใช้สูตร

$$TF = freq(T, D) \quad (4\text{-}6)$$

วิธีการคำนวณ ในแต่ละบทความที่ถูกแบ่งคำและกำจัดคำหยุดออก มากำหนดชื่อเอกสารแทนบทความดังนี้

Document 1 = 10 สิ่งที่น่าสนใจในการออกแบบเว็บไซต์

Document 2 = 10 Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ

Document 3 = 5 สัญญาณบ่งบอกควรปรับปรุง Web

Document 4 = 7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์

Document 5 = 4 เหตุผลที่ควรปรับปรุงเว็บไซต์

วัตถุประสงค์ของการแทนชื่อเอกสารเพราะชื่อบทความค่อนข้างยาวเมื่อยกตัวอยากเปรียบเทียบกับสูตรทำให้เข้าใจยากต่อการทำความเข้าใจ

หลังจากแทนชื่อเอกสารแล้ว จะนับ คำ(Term) = “น่าสนใจ” ที่ปรากฏในเอกสาร D1 กี่ครั้งแล้วนำมาหารด้วยจำนวนคำที่มีในเอกสาร D1 ทั้งหมด ก็จะได้ค่า TF สำหรับการคำนวณค่า TF ใน ตารางที่ 4.11 เป็นการยกตัวอย่างง่ายๆ ด้วยการกำหนดคำทั้งหมดในเอกสาร D1 มีทั้งหมด 3,025 คำ รวมคำที่ซ้ำกันเข้าด้วยกัน คำว่า “น่าสนใจ” ปรากฏในเอกสาร D1 รวมหกครั้งจะได้ค่า $TF =$ 6/3,025 =0.001983 ตามตารางที่ 4.11 ส่วนคำอื่นๆ ก็ใช้วิธีการเดียวกันในการหาคำสำคัญดังที่กล่าวมาข้างต้น

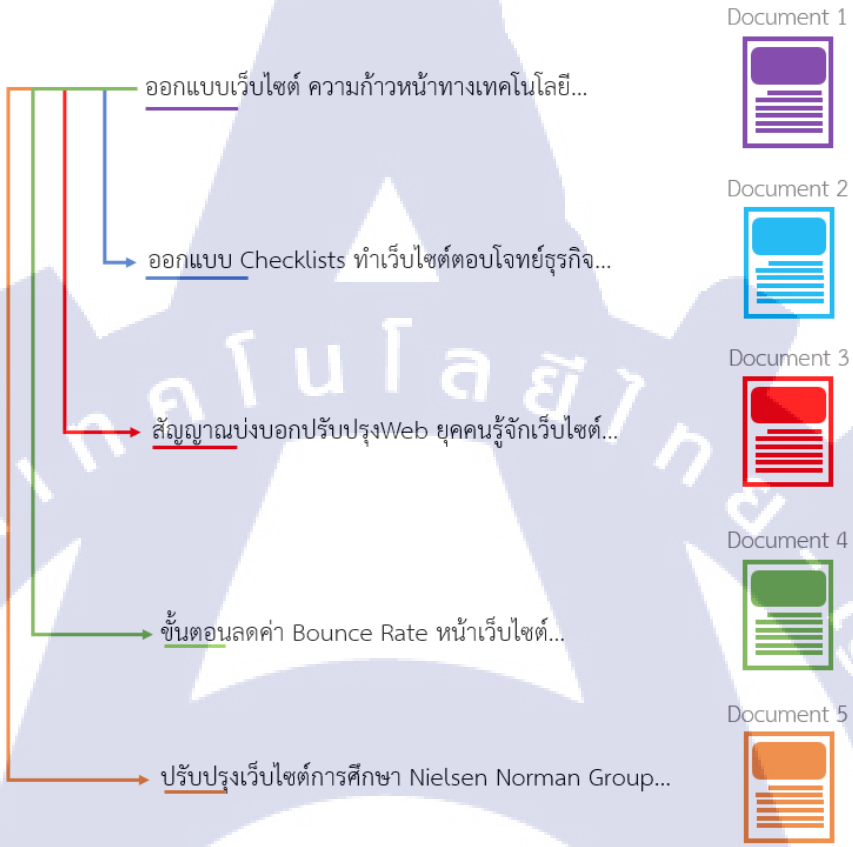
ตารางที่ 4.11 ตัวอย่างการคำนวณและผลลัพธ์การหาค่า TF ของ Document 1

ลำดับ	Word	D1	TF
1	น่าสนใจ	6/290	0.0207
2	ออกแบบ	15/290	0.0517
3	เว็บไซต์	64/290	0.2207
5	ความก้าวหน้า	1/290	0.0034
6	ทางเทคโนโลยี	1/290	0.0034
7	เชื่อมโยง	1/290	0.0034
8	ผู้คน	5/290	0.0172
9	ได้ดี	1/290	0.0034
10	มากยิ่งขึ้น	4/290	0.0138
11	เป็นหนึ่ง	1/290	0.0034
12	เครื่องมือ	2/290	0.0069
13	การพัฒนา	2/290	0.0069
14	เข้าถึง	3/290	0.0103
15	มุม	1/290	0.0034
16	โลก	1/290	0.0034
17	เกิดขึ้น	1/290	0.0034
18	จำนวนมาก	4/290	0.0138
19	ดึงดูด	8/290	0.0276
20	กลุ่มเป้าหมาย	2/290	0.0069

วัตถุประสงค์ของการทำ TF คือ เพื่อหาคำสำคัญในเอกสารนั้นๆโดยถ้าค่า TF มีค่าหรือผลลัพธ์ใกล้เคียงหมายเลขหนึ่งก็จะมีความสำคัญมากที่สุดหลังจากได้ค่า TF จะเป็นการคำนวณหาค่า IDF ดังนี้ การหาค่า IDF คือ การวัดความสำคัญของ คำ (term) ใน เอกสาร (document)ของบทความทั้งหมด โดยใช้สูตร

$$IDF = \log_2 \frac{N}{n_p} \tag{4-7}$$

ทำการตรวจสอบความเหมือนกัน และไม่เหมือนกันของคำในแต่ละ document จากปที่ 4.5 เป็นตัวอย่างวิธีการที่ระบบค้นหาคำที่เหมือนกันในเอกสารอื่นโดยการเขาไปสืบค้นที่ละเอกสาร



รูปที่ 4.6 วิธีการตรวจสอบความเหมือน และ ไม่เหมือนกันของคำในทุก Document

หลังเปรียบเทียบแล้ว ถัดไปเป็นวิธีการทำงาน นำจำนวนเอกสารทั้งหมด 5 บทความมาหารด้วย จำนวน document ที่มีคำคำนั้นอยู่ แล้วผลลัพธ์ไปหาค่าlog ฐาน 2 ตัวอย่างเช่น คำว่า “ออกแบบ” จะพบคำนี้ใน Document 1 เท่านั้นอ้างอิงจากรูปที่ 4.6 โดยมีขั้นตอนในการหาผลลัพธ์ดังนี้

จำนวนเอกสารทั้งหมด 5 เอกสารดังนั้น N = 5 จำนวนคำที่เกิดขึ้นในเอกสาร หรือ Document ไหนบ้าง เช่น “ออกแบบ” เกิดขึ้นในเอกสารที่หนึ่ง หรือ Document 1และ เอกสารที่สอง หรือ Document 2 สองครั้งแต่ในเอกสารอื่นไม่มีคำนี้ปรากฏอยู่ดังนั้นคำว่า “ออกแบบ” มีค่า n_p = 2 ดังตารางที่ 4.12 ตัวอย่างคำนวณ inverse document frequency สำหรับ Document1

ตารางที่ 4.12 ตัวอย่างคำนวณ inverse document frequency สำหรับ Document1

Word	IDF	Result
ออกแบบ	Log(5/2)	0.40
เว็บไซต์	Log(5/5)	0.00
ความก้าวหน้า	Log(5/1)	0.70
ทางเทคโนโลยี	Log(5/1)	0.70
Checklists	Log(5/1)	0.70
ทำ	Log(5/1)	0.70
ตอบ	Log(5/1)	0.70
โจทย์	Log(5/1)	0.70
ธุรกิจ	Log(5/1)	0.70
สัญญาณ	Log(5/1)	0.70
บ่งบอก	Log(5/1)	0.70
ปรับปรุง	Log(5/2)	0.40
Web	Log(5/1)	0.70
ยุค	Log(5/1)	0.70
คน	Log(5/1)	0.70
รู้จัก	Log(5/1)	0.70
ขั้นตอน	Log(5/1)	0.70
ลด	Log(5/1)	0.70
ค่า	Log(5/1)	0.70
Bounce	Log(5/1)	0.70
Rate	Log(5/1)	0.70
หน้า	Log(5/1)	0.70
การศึกษา	Log(5/1)	0.70
Nielsen	Log(5/1)	0.70
Norman	Log(5/1)	0.70
Group	Log(5/1)	0.70

วัตถุประสงค์ของการทำ IDF คือ ใช้สำหรับวัดความสำคัญของ คำ(term) ใน เอกสาร (document) ทั้งหมด คือ ถ้าเจอคำในหลายๆเอกสารคำนั้นจะมีความสำคัญลดลง หลังจากนั้นนำค่า TF และ IDF มาคูณกันเพื่อค่าผลลัพธ์ หรือ TF-IDF Weight ของคำนั้น ดังตารางที่ 4.13-4.16 นี้

ตารางที่ 4.13 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF Document 1

Word	TF					IDF	TF*IDF				
	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
น่าสนใจ	0.0207	0	0	0	0	Log(5/1)=0.69897	0.0145	0.0000	0.0000	0.0000	0.0000
ออกแบบ	0.0517	0.0043	0	0	0	Log(5/2)=0.39794	0.0206	0.0017	0.0000	0.0000	0.0000
เว็บไซต์	0.2207	0.0171	0.0043	0.0152	0.0153	Log(5/5)=0	0.0000	0.0000	0.0000	0.0000	0.0000
ความก้าวหน้า	0.0034	0	0	0	0	Log(5/1)=0.69897	0.0024	0.0000	0.0000	0.0000	0.0000
ทางเทคโนโลยี	0.0034	0	0	0	0	Log(5/1)= 0.69897	0.0024	0.0000	0.0000	0.0000	0.0000
เชื่อมโยง	0.0034	0	0	0	0	Log(5/1)= 0.69897	0.0024	0.0000	0.0000	0.0000	0.0000
ผู้คน	0.0172	0	0	0	0	Log(5/1)= 0.69897	0.0120	0.0000	0.0000	0.0000	0.0000
ได้ดี	0.0034	0	0	0	0	Log(5/1)= 0.69897	0.0024	0.0000	0.0000	0.0000	0.0000
มากยิ่งขึ้น	0.0138	0	0	0	0	Log(5/1)= 0.69897	0.0096	0.0000	0.0000	0.0000	0.0000
เป็นหนึ่ง	0.0034	0	0	0	0	Log(5/1)= 0.69897	0.0024	0.0000	0.0000	0.0000	0.0000
เครื่องมือ	0.0069	0	0	0	0	Log(5/1)= 0.69897	0.0048	0.0000	0.0000	0.0000	0.0000
การพัฒนา	0.0069	0	0	0	0	Log(5/1)= 0.69897	0.0048	0.0000	0.0000	0.0000	0.0000

ตารางที่ 4.14 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF Document 2

Word	TF					IDF	TF*IDF				
	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
CheckLists	0	0.0043	0	0	0	Log(5/1)=0.6990	0.0000	0.0030	0.0000	0.0000	0.0000
ทำ	0	0.0043	0	0	0	Log(5/1)=0.6990	0.0000	0.0030	0.0000	0.0000	0.0000
ตอบ	0	0.0085	0	0	0	Log(5/1)=0.6990	0.0000	0.0059	0.0000	0.0000	0.0000
โจทย์	0	0.0085	0	0	0	Log(5/1)=0.6990	0.0000	0.0059	0.0000	0.0000	0.0000
ธุรกิจ	0	0.0128	0	0	0	Log(5/1)=0.6990	0.0000	0.0089	0.0000	0.0000	0.0000
ตัวเอง	0	0.0085	0	0	0	Log(5/1)=0.6990	0.0000	0.0059	0.0000	0.0000	0.0000
ไม่รู้	0	0.0085	0	0	0	Log(5/1)=0.6990	0.0000	0.0059	0.0000	0.0000	0.0000
แล้ว	0	0.0043	0	0	0	Log(5/1)=0.6990	0.0000	0.0030	0.0000	0.0000	0.0000
คำถาม	0	0.0043	0	0	0	Log(5/1)=0.6990	0.0000	0.0030	0.0000	0.0000	0.0000
แบบนี้	0	0.0043	0	0	0	Log(5/1)=0.6990	0.0000	0.0030	0.0000	0.0000	0.0000

ตารางที่ 4.15 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF Document 3

Word	TF					IDF	TF*IDF				
	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
สัญญาณ	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
บ่งบอก	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
ปรับปรุง	0	0	0.0043	0	0.0076	Log(5/2)=0.3979	0.0000	0.0000	0.0017	0.0000	0.0030
Web	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
ยุค	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
คน	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
รู้จัก	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
Website	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
ถือเป็น	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
สื่อ	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000
นำเสนอ	0	0	0.0043	0	0	Log(5/1)=0.6990	0.0000	0.0000	0.0030	0.0000	0.0000

ตารางที่ 4.16 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF Document 4

Word	TF					IDF	TF*IDF				
	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
ขั้นตอน	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
ลด	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
ค่า	0	0	0	0.0102	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0071	0.0000
Bounce	0	0	0	0.0102	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0071	0.0000
Rate	0	0	0	0.0102	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0071	0.0000
หน้า	0	0	0	0.0102	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0071	0.0000
อัตราส่วน	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
ผู้เข้าชม	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
อัตรา	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
ที่สูง	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000
จัดอันดับ	0	0	0	0.0051	0	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0036	0.0000

ตารางที่ 4.17 ตัวอย่างผลลัพธ์การคำนวณ TF-IDF Document 5

Word	TF					IDF	TF*IDF				
	D1	D2	D3	D4	D5		D1	D2	D3	D4	D5
การศึกษา	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
Nielsen	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
Norman	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
Group	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
เวลา	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
วินาที	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
ดึงดูด	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
ผู้ใช้	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000
ทำได้	0	0	0	0	0.0076	Log(5/1)=0.69897	0.0000	0.0000	0.0000	0.0053	0.0000

3) ผลลัพธ์ของการพัฒนาระบบแนะนำหมวดหมู่บทความ

มีขั้นตอนดังนี้

3.1) Category Similarity การคำนวณหาความเหมือนของเอกสาร โดยใช้สูตร Cosine Similarity ดังนี้

$$similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4\text{-}8)$$

สำหรับการคำนวณขั้นตอนนี้จะนำบทความทั้งห้าบทความ มาเปรียบเทียบความคล้ายคลึงกันโดยแสดงการคำนวณหาผลลัพธ์ได้ดังนี้

ตารางที่ 4.19 ตัวอย่างผลลัพธ์การคำนวณ Cosine Similarity

Document	TF-IDF	Cosine Similarity
10 สิ่งที่น่าสนใจในการออกแบบเว็บไซต์	[0.0207,0.0517,0.2207,0.0034,0.0034,0.0034,0.0172,0.0034,0.0138,0.0034,0.0069,0.0069,0.0000]	1

ตารางที่ 4.19 ตัวอย่างผลลัพธ์การคำนวณ Cosine Similarity (ต่อ)

Document	TF-IDF	Cosine Similarity
10 ออกแบบ Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ	[0.0000,0.0043,0.0171,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0043,0.0043,0.0085,0.0085,0.0128,0.0085,0.0085,0.0043,0.0043,0.0043,0.0000]	0.54
5 สัญญาณบ่งบอกควรปรับปรุง Web	[0.0000,0.0000,0.0043,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0043,0.0000	0.66
7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์	[0.0000,0.0000,0.0152,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.000,0.0051,0.0051,0.0102,0.0102,0.0102,0.0102,0.0051,0.0051,0.0051,0.0051,0.0051,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000,0.0000]	0.52

ตารางที่ 4.19 ตัวอย่างผลลัพธ์การคำนวณ Cosine Similarity (ต่อ)

Document	TF-IDF	Cosine Similarity
4 เหตุผลที่ควรปรับปรุงเว็บไซต์	[0.0000,0.0000,0.0153,0.0000,0.0076,0.0000,0.0076,0.0076,0.0076,0.0076,0.0076,0.0076,0.0076,0.0076,0.0076,0.0076	0.50

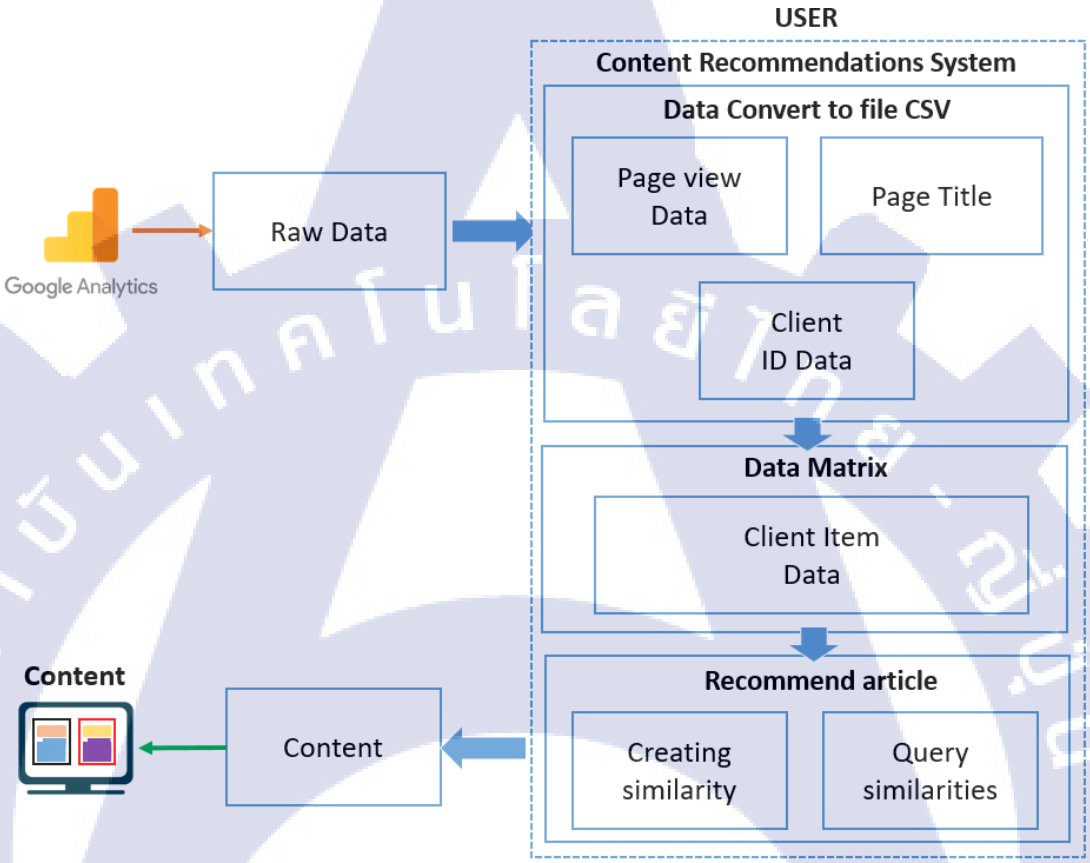
4) ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

มีขั้นตอนดังนี้

หลังจากได้เตรียมชุดข้อมูลไปเปรียบเทียบ Cosine Similarity ผู้วิจัยจะทำการทดสอบความคล้ายคลึงของบทความที่ผู้อ่าน เลือกอ่าน เช่น ผู้อ่านเลือกอ่าน บทความเรื่อง “10 สิ่งที่น่าสนใจในการออกแบบเว็บไซต์” แล้วระบบจะนำบทความนี้ไปเปรียบเทียบความคล้ายคลึงกับบทความที่ถูกเก็บอยู่แล้วในระบบ มานำเสนอให้กับผู้อ่าน ค่าความคล้ายคลึง มี คะแนนอิงตามบทความ ที่ชื่อว่า 1) 10 Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ ค่าความคล้ายคลึง = 0.54% 2) 5 สัญญาณบ่งบอกควรปรับปรุง Web ค่าความคล้ายคลึง = 0.66% 3) 7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์ ค่าความคล้ายคลึง = 0.52% และอันดับสุดท้าย 4) 4 เหตุผลที่ควรปรับปรุงเว็บไซต์ ค่าความคล้ายคลึง = 0.50% การแสดงผลข้อมูลของระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่านนำไปใช้กับบทความที่เกี่ยวข้องโดยระบบจะแสดงบทความที่หนึ่งถึงสี่โดยมีคะแนนความคล้ายคลึงแสดงคู่กับบทความเลือกให้ผู้อ่านได้เลือกอ่านบทความ

4.2.2 เทคนิค Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering

การศึกษาเทคนิค Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering ในหัวข้อนี้ผู้วิจัยได้อธิบายมีขั้นตอนการทำงานและออกแบบแนวคิดการทำงานมีภาพรวมการทำงานดังนี้



รูปที่ 4.7 ภาพรวมการทำงานเทคนิค Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering

โดยผลลัพธ์ของการพัฒนาระบบแนะนำบทความอัตโนมัติสำหรับผู้อ่านด้วย Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering สามารถแบ่งได้ 3 ส่วน ได้แก่

1) ผลลัพธ์ของการเตรียมข้อมูล ซึ่งเป็นขั้นตอนการดึงข้อมูลและนำมาประมวลผลการนำไปใช้ในการพัฒนาระบบ

2) ผลลัพธ์ของการพัฒนาระบบแนะนำบทความ

3) ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

1) ผลลัพธ์ของการเตรียมข้อมูล ซึ่งเป็นขั้นตอนการดึงข้อมูลและนำมาประมวลผล Collaborative Filtering โดยใช้รูปแบบ Item-Based Filtering เป็นการแนะนำโดยหาบทความที่ถูกเปิดดูจาก User กลุ่มเดียวกันนำข้อมูลเหล่านี้มาแนะนำให้กับผู้อ่านเป้าหมายที่คาดว่าจะอ่านบทความจากการแนะนำของระบบ ส่วนขั้นตอนการเตรียมข้อมูล ผู้วิจัยได้สมัครใช้งาน Google Analytic สำหรับเก็บข้อมูลของผู้อ่านบทความเป็นเวลาหนึ่งปี มีข้อมูลจำนวนสองชุดที่เชื่อมโยงถึงกันด้วย Client Id ดังนี้ ข้อมูลชุดที่หนึ่งมี 7 คอลัมน์ 1.Client Id 2.Sessions 3.Avg. Session Duration 4.Bounce Rate 5.Revenue 6.Transactions 7.Goal Conversion Rate และรูปแบบข้อมูลก่อนแปลงให้เป็นไฟล์ CSVตัวอย่างรูปที่ รูปที่ 4.8

	Client Id	Sessions ↓	Avg. Session Duration	Bounce Rate	Revenue	Transactions	Goal Conversion Rate
1.	2024846862.1611671510	4 (0.87%)	00:00:52	50.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
2.	2047815508.1611820575	4 (0.87%)	00:01:53	75.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
3.	92453836.1612025288	4 (0.87%)	00:13:42	50.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
4.	1525547211.1611843956	3 (0.65%)	00:00:00	100.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
5.	428830523.1546930107	3 (0.65%)	00:00:00	100.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
6.	589916507.1611748903	3 (0.65%)	00:00:00	100.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
7.	62503479.1612161482	3 (0.65%)	00:00:00	100.00%	$0.00 (0.00%)	0 (0.00%)	0.00%
8.	1069914339.1611977953	2 (0.43%)	00:00:48	50.00%	$0.00 (0.00%)	0 (0.00%)	0.00%

รูปที่ 4.8 ตัวอย่างชุดข้อมูลที่หนึ่งก่อนแปลงให้เป็นไฟล์ CSV

หลังจากนั้นทำการดึงข้อมูลชุดที่หนึ่งจากแหล่งข้อมูลระบบ Google Analytic ออกมาเป็นไฟล์ Json Raw Data ข้อมูลมี 7 คอลัมน์ ดังนี้ 1.Client Id คือ รหัสประจำตัวผู้ใช้งานเว็บไซต์ 2.Sessions คือ ช่วงเวลาหนึ่งที่ผู้ใช้ 1 คน เข้ามายังเว็บไซต์ 3.Avg. Session Duration คือ ค่าเฉลี่ยนเวลาผู้ใช้ 1 คน เข้ามายังเว็บไซต์ 4.Bounce Rate คือ คือจำนวนผู้เข้าเยี่ยมชมเวบไซต์หน้าใดหน้าหนึ่งหรือหน้า Landing Page แล้วออกไปโดยปราศจากการเยี่ยมชมหน้าอื่นๆ 5.Revenue คือ รายได้ 6.Transactions คือจำนวนครั้งของการสั่งซื้อสินค้าที่เกิดขึ้น 7.Goal Conversion Rate คือ อัตราการที่คนไปถึงจุดที่เราตั้งค่าไว้ หมายความว่า ใน 100 คนจะไปถึง Goals กี่คน นำข้อมูลเหล่านี้ออกมาทั้งหมดโดยไม่ได้คัดกรองหลังจากนั้นทำการแปลงข้อมูลให้อยู่ในรูปแบบไฟล์ CSV จะได้ตัวอย่างข้อมูลตามตารางที่ 4.20

ตารางที่ 4.20 ชุดข้อมูลที่หนึ่งเมื่อเปลี่ยนให้เป็นไฟล์ CSV

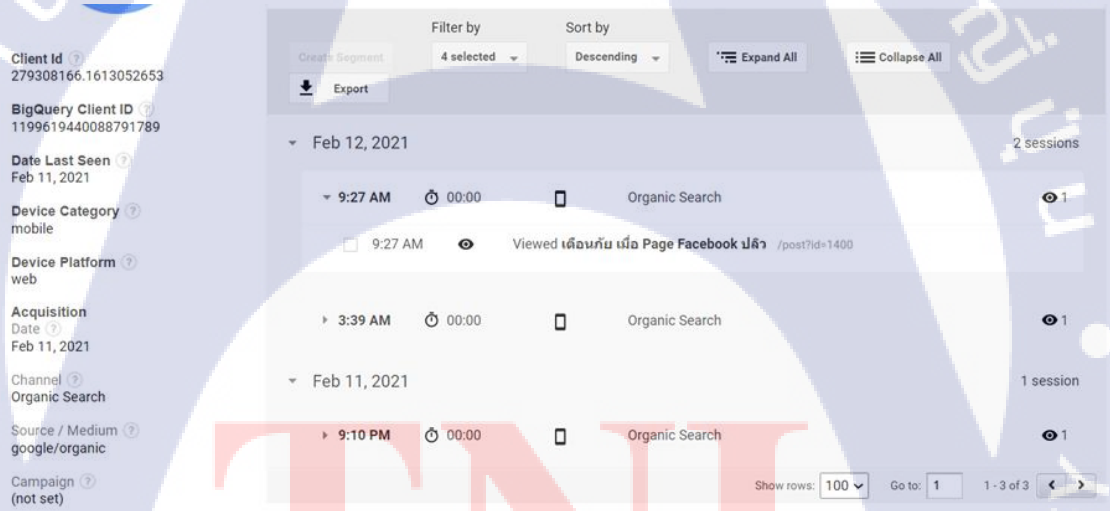
Client Id	Sessions	Avg. Session Duration	Bounce Rate	Revenue	Transactions	Goal Conversion Rate
2047815508.1611820575	5	90.00	80.00%	0.00	0	0.00%
92453836.1612025288	5	752.20	40.00%	0.00	0	0.00%
1424137363.1611891001	4	0.00	100.00%	0.00	0	0.00%
1525547211.1611843956	3	0.00	100.00%	0.00	0	0.00%
428830523.1546930107	3	0.00	100.00%	0.00	0	0.00%
589916507.1611748903	3	0.00	100.00%	0.00	0	0.00%
62503479.1612161482	3	0.00	100.00%	0.00	0	0.00%

ถัดมาผู้วิจัยทำความเข้าใจและคัดกรองข้อมูลเพื่อจะนำไปใช้ในการประมวลผล ในที่นี้ผู้วิจัยเลือกใช้ข้อมูลคอลัมน์ชื่อ Client Id เท่านั้นเพื่อใช้หาค่าความสัมพันธ์กับข้อมูลชุดที่สอง ส่วนคอลัมน์ที่เหลือจะไม่ถูกนำไปใช้งาน

ตารางที่ 4.21 ชุดข้อมูลที่หนึ่งเมื่อถูกคัดรองข้อมูล

Client Id
2047815508.1611820575
92453836.1612025288
1424137363.1611891001
1525547211.1611843956
589916507.1611748903

ข้อมูลชุดที่สองเป็นข้อมูลที่มีความซับซ้อนมากกว่าข้อมูลชุดแรกและมีข้อมูลที่มีรายละเอียดที่มากกว่า ทำการดึงข้อมูลออกมาในรูปแบบไฟล์ Json Raw Data โดยการดึงข้อมูลออกมาทั้งหมดโดยใช้ Client Id ที่เป็นข้อมูลชุดแรกเป็นตัวอ้างอิงในการดึงข้อมูล และไม่มีการคัดกรอง มีรูปแบบก่อนการดึงข้อมูลตามตัวอย่างรูป 4.9



รูปที่ 4.9 ตัวอย่างชุดข้อมูลที่สองก่อนแปลงให้เป็นไฟล์ Json

ถัดมาผู้วิจัยนำข้อมูล Raw Data ออกมาทำการคัดเลือกข้อมูล 3 ข้อมูล ได้แก่ 1. Client id คือ รหัสประจำตัวผู้ใช้งานเว็บไซต์ 2. Page Title คือ ชื่อของบทความ 3. Page view คือ จำนวนครั้งที่เข้าชมบทความ แปลงให้อยู่ในรูปแบบตาราง เก็บอยู่ในไฟล์ CSV เพื่อจะนำข้อมูลไปใช้ในการทำ Data Matrix โดยมีตัวอย่างหลังจากดึงข้อมูลดังตารางที่ 4.22

ตารางที่ 4.22 ตัวอย่างชุดข้อมูลที่สองที่คัดรองแล้ว

Client id	Page Title	Page view
937536369.1535952623	แนะนำ UI DESIGN TOOLS ที่น่าจับเมาส์ลอง	4
366913768.1583302249	5 สัญญาณบ่งบอกควรปรับปรุง Web	1
1471528426.1554338653	เลือกฟอนต์อย่างไร ให้เหมาะกับเว็บไซต์คุณ	2
273457067.1553226147	4 เหตุผลที่ควรปรับปรุงเว็บไซต์	1
725405417.1528369209	11 กฎในการออกแบบประสบการณ์ผู้ใช้ หรือ UX (User Experience)	2
409312877.1564649332	7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์	3
699658708.1567225562	10 Develop Tools ที่ช่วยให้ Frontend ทำงานง่ายขึ้น	2
1667267345.1573466230	10 Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ	0

ถัดมาจะนำข้อมูลชุดที่สองตามตารางที่ 4.22 เข้าสูกระบวนการ Data Matrix โดยข้อมูลที่ใช้ได้แก่ Client id, Page Title, Page view ข้อมูลเหล่านี้เป็นข้อมูลประวัติการเข้าไปอ่านบทความของผู้ใช้งาน (Item Profile) และ ข้อมูลบทความทั้งหมด 54 บทความ สำหรับนำมาสร้างชุดเอกสารโดยไม่มีการแบ่งหมวดหมู่เหมือนกับ ระบบแนะนำหมวดหมู่บทความอัตโนมัติสำหรับผู้เขียน ทำการแปลงให้อยู่ในรูปแบบไฟล์นามสกุล CSV ข้างในจะถูกแบ่งข้อมูลเป็น 54 column เป็น ชื่อบทความ และ 10,000 แถวเป็นรหัสผู้ใช้งาน ส่วนค่า Page viewทำการแปลงค่าให้มีแค่สองค่า ได้แก่ 0 คือ ไม่ได้เปิดดู และ 1 คือ เปิดดู และขอยกตัวอย่างผลลัพธ์ของการเตรียมข้อมูล เพื่อให้เห็นภาพข้อมูลที่สามารถนำไปใช้ในกระบวนการ Collaborative Filtering จำนวน 54 บทความ และรหัสประจำตัวผู้เข้ามาเปิดอ่านบทความ Client Id 12 Column ดังตารางที่ 4.23

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative

Client Id	10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด	เตือนภัย เมื่อ Page Facebook ปลิว	7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์	10 Develop Tools ที่ช่วยให้ Frontend ทำงานง่ายขึ้น	11 กฎในการออกแบบประสบการณ์ผู้ใช้ หรือ UX (User Experience)	ประหยัดเวลา Coding ด้วย 10 Plug-in Free ใน Visual Studio Code
937536369.1535952623	1	1	1	0	0	0
1441092916.15617	0	1	0	0	0	0
1232396967.1485594117	0	0	0	0	1	0
366913768.1583302249	1	1	0	0	0	0
1471528426.1554338653	0	1	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	1	1	0	1	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	10 สิ่งที่น่าสนใจในการออกแบบเว็บไซต์	ทำ Content ยังไงให้ปัง !!?	เคล็ดไม่ลับ หา “คีย์เวิร์ด”	7 ข้อคิดดีๆ ในการสร้างคอนเทนต์บน Youtube	4 เหตุผลที่ควรปรับปรุงเว็บไซต์	ใคร ถือ Facebook Page ในมือต้องอ่าน!!
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	1	0	0	0	0
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	1	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	10 วิธีออกแบบ UX เพื่อสร้างความประทับใจในการใช้งานแอปพลิเคชันครั้งแรก	5 เคล็ดลับ เพิ่มประสิทธิภาพบน Landing Page	5 สัญญาณบ่งบอกควรปรับปรุง Web	เลือกฟอนต์อย่างไร ให้เหมาะกับเว็บไซต์คุณ	แนะนำ UI DESIGN TOOLS ที่น่าจับเมาส์ลอง	3 วิธีการทำให้คอนเทนต์โดดเด่น
937536369.1535952623	0	0	0	0	0	1
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	1	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	1	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	1	0	0	0	0
699658708.1567225562	1	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	5 สิ่งที่น่าสนใจในการตลาดแบบ Inbound 2019	6 เครื่องมือสร้าง CSS Animation ที่จะทำให้เว็บไซต์น่าสนใจขึ้น	10 เครื่องมือสำหรับ Social Media ที่ใช้งานได้ฟรี	สรุปการปรับ METRIC ใหม่ บน Facebook Ads ใน 4 ภาพ	10 Check-Lists ทำเว็บไซต์ให้ตอบโจทย์ธุรกิจ	6 กลยุทธ์สร้างเว็บไซต์ขายสินค้า (E-Commerce)
937536369.1535952623	0	0	0	0	0	1
1441092916.15617	0	0	0	0	0	1
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	1	0	0	0	1
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	1	0	0	0	0
273457067.1553226147	0	0	0	0	0	1
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	1
1667267345.1573466230	0	0	0	1	0	0
510792775.1559622009	0	0	0	0	0	1

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	3 กลยุทธ์การตลาด Word-of-Mouth ที่ช่วยให้ธุรกิจขนาดเล็กเติบโตได้เร็วขึ้น	5 เรื่องยอดนิยมใน Pinterest 2019	7 สิ่งแย่ๆ ที่ไม่ควรทำใน Social Media Marketing	10 คอนเทนต์ดีๆ ที่ช่วยสร้างแรงบันดาลใจให้คนบนโลกโซเชียล	มาทำความรู้จักกับเครื่องมือแสดงผลการค้นหาบน Google (SERP)	4 ICON PACKS แจกให้ฟรีๆ
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	1	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	0	0	1	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	4 เครื่องมือ ที่ช่วยเปลี่ยน Design ให้กลายเป็น Code	5 เครื่องมือสร้าง Flow Chart บน Cloud ใช้งานได้ FREE!	5 วิธีการช่วยส่งเสริมการตลาดในธุรกิจขนาดเล็ก	5 วิธีการสร้างเรื่องราวให้แบรนด์ดึงดูดกลุ่มเป้าหมาย	Auto Chat เทคโนโลยีฝีมือคนไทย	Auto Chat จะช่วยให้ธุรกิจของคุณง่ายขึ้น ได้อย่างไร ? มาดูกัน
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	Digital Thailand Big Bang 2017	Google SEO Trends 2019	Inbound Marketing (การตลาดแบบดึงดูด)	ThaiSalary โปรแกรมคำนวณเงินเดือนออนไลน์	เกมที่ช่วยฝึกทักษะการใช้ path	เครื่องมือ Extension ใน Google Chrome สำหรับนักออกแบบเว็บ
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	เครื่องมือสร้างชุดสี สำหรับ Front-End	เบื่อมั้ย? ที่ต้องคอยตอบคำถามซ้ำๆ	เปลี่ยนภาพสเก็ตในสมุดให้กลายเป็นภาพบน Adobe Illustrator	แชทบอท สร้างให้โดนใจ สไตล์ผม เริ่มต้นยังไง Ep.01	แนวโน้มความนิยม 5 อย่าง สำหรับ Inbound Marketing	ช็อปช่วยชาติ มีอะไรมาช่วยลดหย่อนภาษี ปี 2561 ได้บ้าง?
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

ตารางที่ 4.23 ตัวอย่างการเตรียมข้อมูล ก่อนนำไปเข้ากระบวนการ Collaborative (ต่อ)

Client Id	คุณสมบัติใหม่ Facebook “Why Am I Seeing This Post?”	ชุด UI และ WIREFRAMES แจกให้ใช้ฟรี	มาทดสอบ UI Skill กันเถอะ	สิ่งที่ควรทำและไม่ควรทำในระบบการแจ้งเตือนของแอปพลิเคชัน	อยากทำงานการตลาดออนไลน์ ต้องเรียนจบอะไร (Ep2)	อยากทำงานการตลาดออนไลน์ ต้องเรียนจบอะไร (Ep 1)
937536369.1535952623	0	0	0	0	0	0
1441092916.15617	0	0	0	0	0	0
1232396967.1485594117	0	0	0	0	0	0
366913768.1583302249	0	0	0	0	0	0
1471528426.1554338653	0	0	0	0	0	0
1316748013.1567396114	0	0	0	0	0	0
273457067.1553226147	0	0	0	0	0	0
725405417.1528369209	0	0	0	0	0	0
409312877.1564649332	0	0	0	0	0	0
699658708.1567225562	0	0	0	0	0	0
1667267345.1573466230	0	0	0	0	0	0
510792775.1559622009	0	0	0	0	0	0

2) ผลลัพธ์ของการพัฒนาระบบแนะนำบทความ

ขั้นตอนการทำงานของระบบหาความสัมพันธ์ระหว่างบทความเพื่อแนะนำบทความนั้นให้กับกลุ่มผู้อ่าน

ในขั้นตอนนี้ผู้วิจัยขอคัดเลือกข้อมูลสำหรับนำมาวิจัยจำนวน 4 Client id หรือ 4 User และ บทความจำนวน 4 บทความดังตารางที่ 4-16

เมื่อเตรียมข้อมูลเสร็จแล้วจะทำการค้นหาความสัมพันธ์ระหว่างบทความเพื่อแนะนำ บทความนั้นให้กับกลุ่มผู้อ่าน การค้นหาความสัมพันธ์ระหว่างบทความมีขั้นตอน และใช้สูตรหาความสัมพันธ์ดังนี้

$$similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4\text{-}9)$$

ตารางที่ 4.25 ตัวอย่างข้อมูลหาความสัมพันธ์ของบทความ Collaborative Filtering

User	10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด	เตือนภัย เมื่อ Page Facebook ปลิว	7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์	10 Develop Tools ที่ช่วยให้ Frontend ทำงานง่ายขึ้น
User 1	1	1	1	0
User 2	1	1	0	0
User 3	0	1	0	0
User 4	1	1	0	1

จากตารางที่ 4.16 ผู้วิจัยยกตัวอย่างเพื่อหาความสัมพันธ์ ของทั้ง 4 บทความว่าสองบทความไหนมีความสัมพันธ์กันมากที่สุด ทั้งหมด 3 ครั้ง ด้วยวิธีการคำนวณมีดังนี้

2.1) หาความสัมพันธ์ระหว่าง บทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับ บทความ เตือนภัย เมื่อ Page Facebook ปลิว

A = 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด

B = เตือนภัย เมื่อ Page Facebook ปลิว

A = [1,1,0,1] และ B = [1,1,1,1] และจะได้สมการ A.B ดังนี้

A.B = [1,1,0,1] * [1,1,1,1] หรือ (1*1)+(1*1)+(0*1)+(1*1) = 3

หลังจากนั้นมาหาผลลัพธ์สมการ ||A||||B||

$||A||||B|| = \sqrt{1+1+0+1} * \sqrt{1+1+1+1}$

$||A||||B|| = \sqrt{3} * \sqrt{4} = 3.46$

ผลลัพธ์ $similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{3}{3.46} = 0.86$

ดังนั้น ค่าความสัมพันธ์ ของบทความ ชื่อ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับ เตือนภัย เมื่อ Page Facebook ปลิว มีค่าเท่ากับ 0.86%

2.2) หาความสัมพันธ์ระหว่าง บทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับ 7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์

A = 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด

B = 7 ขั้นตอนลดค่า Bounce Rate บนหน้าเว็บไซต์

A = [1,1,0,1] และ B = [1,0,0,0] และจะได้สมการ A.B ดังนี้

A.B = [1,1,0,1] * [1,0,0,0] หรือ (1*1)+(1*0)+(0*0)+(1*0) = 1

หลังจากนั้นมาหาผลลัพธ์สมการ ||A||||B||

$||A||||B|| = \sqrt{1+1+0+1} * \sqrt{1+0+0+0}$

$||A||||B|| = \sqrt{3} * \sqrt{1} = 1.73$

ผลลัพธ์ $similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{1}{1.73} = 0.57$

2.3) หาความสัมพันธ์ระหว่าง บทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับ 10 Develop Tools ที่ช่วยให้ Frontend ทำงานง่ายขึ้น

A = 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด

B = 10 Develop Tools ที่ช่วยให้ Frontend ทำงานง่ายขึ้น

A = [1,1,0,1] และ B = [0,0,0,1] และจะได้สมการ A.B ดังนี้

A.B = [1,1,0,1] * [0,0,0,1] หรือ (1*0)+(1*0)+(0*0)+(1*1) = 1

หลังจากนั้นมาหาผลลัพธ์สมการ ||A||||B||

$$||A||||B|| = \sqrt{1+1+0+1} * \sqrt{0+0+0+1}$$

$$||A||||B|| = \sqrt{3} * \sqrt{1} = 1.73$$

ผลลัพธ์ $similarity = \cos\theta = \frac{A.B}{||A||||B||} = \frac{1}{1.73} = 0.57$

3) ผลลัพธ์การพัฒนาระบบตัวอย่างเพื่อการประเมินผลลัพธ์

ผลลัพธ์ที่ได้จากการคำนวณการหาค่าความสัมพันธ์ของบทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับบทความ เตือนภัย เมื่อ Page Facebook ปลิว ค่าความสัมพันธ์ เท่ากับ 0.86% ซึ่งมีค่ามากที่สุดเมื่อเทียบกับค่าของบทความอื่นๆ จากผลลัพธ์การคำนวณหาค่าความสัมพันธ์ การคำนวณการหาค่าความสัมพันธ์ของบทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด กับบทความ เตือนภัย เมื่อ Page Facebook ปลิว ค่าความสัมพันธ์ เท่ากับ 0.86% ดังนั้นระบบจะแนะนำบทความชื่อว่า เตือนภัย เมื่อ Page Facebook ปลิว ให้กับบทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด ดังนนั้นผู้ใช้งานที่ทั้งหมดทั้ที่เคยอ่านหรือไม่เคยอ่าน บทความ 10 Extensions และ Plugin ที่ช่วยเพิ่มสีสันในการเขียนโค้ด จะถูกระบบแนะนำให้อ่านบทความ เตือนภัย เมื่อ Page Facebook ปลิว เป็นต้น

บทที่ 5
บทสรุป อภิปราย และข้อเสนอแนะ

จากการศึกษาวิจัยเรื่องระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา ด้วยข้อมูลข้อมูลผู้ใช้งานเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา ซึ่งการวิจัยนี้มีวัตถุประสงค์ตามบทที่ 1 ได้นำเสนอข้อสรุปตามลำดับดังนี้

5.1 สรุปผลการวิจัย

ในการสรุปผลของงานวิจัยชิ้นนี้ ผู้วิจัยได้ทำการสรุปผลตามวัตถุประสงค์ในแต่ละหัวข้อที่ได้กล่าวไปแล้วในบทที่ 1 ดังต่อไปนี้

5.1.1 เพื่อศึกษาและพัฒนาระบบแนะนำหมวดหมู่บทความบนเว็บไซต์สำหรับผู้เขียน โดยใช้เทคนิค Content-Base Filtering การหาค่าความคล้ายคลึงของรายการหมวดหมู่บทความหรือเอกสารที่มีอยู่ กับ ความต้องการหาหมวดหมู่บทความหรือเอกสารในรายการของผู้เขียนบทความ โดยการคำนวณจากสูตร Cosine similarity นั้นพบว่า มีประสิทธิภาพความคล้ายคลึง 41% ซึ่งคะแนนความคล้ายคลึงควรจะอยู่ในช่วง 40%-50% หากมากกว่านั้นจะเป็นการ Plagiarism บทความ เป็นการลอกเลียนแบบบทความ ผลลัพธ์ของการแนะนำหมวดหมู่บทความมีความใกล้เคียงกับความต้องการหาหมวดหมู่บทความของผู้เขียนบทความ และเป็นประโยชน์ต่อผู้ที่ต้องการนำงานวิจัยนี้ไปพัฒนาระบบแนะนำเป็นของตัวเอง

5.1.2 เพื่อศึกษาและพัฒนาระบบแนะนำบทความบนเว็บไซต์สำหรับผู้อ่านบทความ โดยใช้เทคนิค Content base filtering การหาค่าความคล้ายคลึงของบทความหรือเอกสาร กับ ความต้องการหาบทความหรือเอกสารในรายการของผู้ อ่าน ด้วยผลการคำนวณโดยใช้สูตร Cosine similarity นั้นพบว่า มีประสิทธิภาพความคล้ายคลึง 46% ซึ่งคะแนนความคล้ายคลึงควรจะอยู่ในช่วง 40%-50% หากมากกว่านั้นจะเป็นการ Plagiarism บทความ เป็นการลอกเลียนแบบบทความ บทความ ผลลัพธ์ของการแนะนำบทความมีความใกล้เคียงกับความต้องการหาบทความของผู้อ่านบทความ และเป็นประโยชน์ต่อผู้ที่ต้องการนำงานวิจัยนี้ไปพัฒนาระบบแนะนำเป็นของตัวเอง

5.1.3 เพื่อศึกษาและพัฒนาระบบแนะนำบทความบนเว็บไซต์สำหรับผู้อ่านบทความ โดยใช้เทคนิค Collaborative Filtering ในการกรองข้อมูล โดยให้ความสำคัญกับข้อมูลบทความหรือเอกสาร และหาค่าความคล้ายคลึงของรายการบทความหรือเอกสารที่มีอยู่ จาก Profile หรือ ข้อมูลการใช้งานเว็บไซต์ของผู้อ่านคนอื่นๆ นำมาคำนวณโดยใช้สูตร Cosine similarity กับความต้องอ่านบทความที่ผู้อ่านต้องการ มีประสิทธิภาพค่าความสัมพันธ์อยู่ที่ 86% ผลลัพธ์ของการแนะนำบทความ

มีความใกล้เคียงกับความต้องการหาบทความของผู้อ่านบทความ และเป็นประโยชน์ต่อผู้ที่ต้องการนำงานวิจัยนี้ไปพัฒนาระบบแนะนำเป็นของตัวเอง

5.2 ปัญหาและอุปสรรคในการทำวิจัย

5.2.1 ปัญหาในการรวบรวมข้อมูลที่ต้องรวบรวมข้อมูลที่ได้จาก Google Analytic สำหรับข้อมูลการใช้งานที่ไม่ได้อยู่เป็นกลุ่มเดียวกัน จึงต้องทำความเข้าใจการลิงค์ของข้อมูลเพื่อเข้าไปดึงข้อมูลที่ต้องการ จึงทำให้เกิดความล่าช้าในการรวบรวมข้อมูล

5.2.2 ปัญหาในการใช้ภาษา Library Python เนื่องจากบาง Library ต้องการ Version Python ไม่เหมือนกันดังนั้นจึงต้องมีการ ติดตั้ง Environment ที่แตกต่างกัน จึงทำให้เกิดความล่าช้าในขั้นตอนการทำงาน

5.3 ข้อจำกัดของงานวิจัย

งานวิจัยนี้ทำการเก็บและรวบรวมข้อมูลการเข้าใช้งานอ่านบทความของผู้ใช้เว็บไซต์บริษัทโฆษณาซึงมีข้อจำกัดอยู่ที่ Google analytic ให้ข้อมูลได้ 10,000. Column เท่านั้น และ ข้อมูลบทความมีทั้งหมด 54 บทความ ซึ่งเป็นข้อจำกัดในงานวิจัยชิ้นนี้

5.4 ข้อเสนอแนะงานวิจัย

5.4.1 หลังจากใช้งานจริงกับเว็บไซต์แล้วต้องมีการเก็บค่าของบทความของระบบแนะนำว่าเมื่อระบบแนะนำไปแล้ว มีผู้ที่เข้าอ่านบทความที่เกี่ยวข้องมากน้อยเพียงใด เพื่อนำไปปรับปรุง Model ของทั้งสองระบบแนะนำที่ได้กล่าวมาข้างต้น

5.4.2 ทดสอบกับความแม่นยำ กับ Model อื่นเช่น BERT เป็นต้นเพื่อเปรียบเทียบความแม่นยำและเป็นตัวเลือกนำไปใช้งานในอนาคต

5.4.3 ในอนาคตหากต้องการเพิ่มความแม่นยำในการนำเสนอ หมวดหมูบทความและการนำเสนอบทความให้มีหลากหลายมิติมากขึ้น เช่น เวลา สถานที่ อายุ เพศ และอื่นๆ

บรรณานุกรม

บรรณานุกรม

[1] G., Chowdhury, “Natural language processing,” *Annual Review of Information Science and Technology*, vol. 37, no. 1, pp. 51-89, January 2005.

[2] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures,* United Kingdom: Psychology Press, 1977.

[3] C. Aone et al., “Trainable, scalable summarization using robust NLP and machine learning,” *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, ACL '98/COLING '98, United States, August 26-29, 1998, pp. 62-66.

[4] E. Haddi et al., “The role of text pre-processing in sentiment analysis,” *Procedia Computer Science*, vol. 17, no. 5, pp. 26-32, May 2013.

[5] Terveen and Hill, *Beyond Recommender Systems: Helping People Help Each Other*, Addison-Wesley: In HCI In The New Millennium, 2001.

[6] P. Markellou et al., “Personalized E-commerce recommendations,” *IEEE International Conference on e-Business Engineering*, ICEBE 2005, Beijing, China, October 12-18, 2005, pp. 2-6.

[7] J. S. Breese et al., “Empirical analysis of predictive algorithms for collaborative filtering,” *Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence*, UAI'98, July 20-26, 1998, pp. 43–52.

[8] B. Sarwar et al., “Item-based collaborative recommendation algorithms,” *In Proceedings of the 10th international conference on World Wide Web,* Hong Kong, April 10-16, 2001, pp. 285–295.

[9] L. Jonathan et al., “An algorithmic framework for performing collaborative filtering,” *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR 1999, Berkeley, California, United States, August 9-15, 1999, pp. 230–237.

[10] P. Resnick et al., “GroupLens: An open architecture for collaborative filtering of netnews,” *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, CSCW 1994, Chapel Hill, NC, October 16-22, 1994, pp. 175–186.

[11] K. Goldberg et al., “Eigentaste : A constant time collaborative filtering algorithm,” *Information Retrieval Journal*, vol. 4, no. 2, pp. 133-151, July 2001.

[12] M. O’Connor and J. Herlocker, “Clustering items for collaborative filtering,” *In Proceedings of the ACM SIGIR workshop on recommender systems*, vol. 128 no. 2, pp. 257-297, August 1999.

[13] D. Lemire and A. Maclachlan, “Slope one predictors for online rating-based collaborative filtering,” *In SIAM Data Mining, SDM 2005*, Newport Beach, California, April 21-23, 2005, pp. 1-4.

[14] วงกต ศรีอุไร, พยุง มีสัจ และชูชาติ หฤไชยะศักดิ์, “การแทนค่าข้อมูลที่ขาดหายเพื่อแก้ไขปัญหาความเบาบางของข้อมูลแบบพึ่งพาผู้ใช้ร่วม,” *วารสารพระจอมเกล้าลาดกระบัง*, ปีที่ 16, ฉบับที่ 1, หน้า 44–52, เมษายน 2551.

[15] วลัยนุช สกุลนุ้ย, “*วิเคราะห์และพัฒนาระบบแนะนำหนังสือคอมพิวเตอร์ แบบออนไลน์ โดยใช้เทคนิคการกรองแบบอิงเนื้อหา*,” [Online]. Available: http://www.rpu.ac.th/ [Accessed: April 27, 2018].

[16] R. J. Mooney and L. Roy, ”Content-based book recommending using learning for text categorization,” *Submission to Fourth ACM Conference on Digital Libraries, Digital Libraries (cs.DL)*, University of Texas Austin, USA, February 1-7, 1999, pp. 2-8.

[17] C. Haruechaiyasak and C. Damrongrat, “Article recommendation based on a topic model for wikipedia selection for schools,” *Digital Libraries: Universal and Ubiquitous Access to Information*, vol. 5362, no.10 , pp. 339-342, February 2008.

[18] T. Nomponkrang and C. Sanrach, “The comparison of algorithms for thai-sentence classification,” *International Journal of Information and Education*, vol. 6, no. 10, pp. 801-808, May 2016.

[19] M. J. Pazzani, "A Framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, vol. 13, no. 5-6, pp. 393–408, December 1999.

[20] P. Melville et al., "Content-boosted collaborative filtering for improved recommendations," *Proceedings of the Eighteenth National Conference on Artificial Intelligence, (AAAI-2002)*, Edmonton, Canada, July 11-17, 2002, pp. 187-192.

[21] I. M. Soboro, "Combining content and collaboration in text filtering," *In Proceedings of the IJCAI99 Workshop on Machine Learning for Information Filtering, IJCAI99,* Stockholm, Sweden, July 21-27, 1999, pp. 86-91.

[22] M. Deshpande and G. Karypis, "Item-based top-N recommendation algorithms," *ACM Transactions on Information Systems*, vol. 22, no.1, pp. 143-177, January 2004.

[23] J. Sobecki and S.L. Szczepan. "Wiki-news interface agent based on AIS methods," *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, vol. 4496, no. 10, pp. 258-266, June 2007.

[24] รัตน์ชัยนันท์ ธรรมสุจริต, "*การปรับปรุงการจัดกลุ่มข้อมูลด้วยต้นไม้ลำดับขั้นแบบสมดุลสำหรับระบบแนะนำภาพยนตร์ด้วยเทคนิคการคดกรองแบบพึ่งพา*," วิทยานิพนธ์ วท.ม. (เทคโนโลยีสารสนเทศ), สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพฯ, 2551.

[25] พัชราภรณ์ ช่วยเจริญ และ พนิดา ทรงรัมย์, "การขุดค้นความสัมพันธ์หมวดหมู่ของเพจบนเฟสบุ๊ค" *วารสารเทคโนโลยีสารสนเทศ (Information Technology Journal)*, ปีที่ 15, ฉบับที่ 1, หน้า 50-59, มกราคม – มิถุนายน 2562.

[26] วนารัตน์ จุฬพันธ์ทอง และ ไกรศักดิ์ เกษร, "ระบบแนะนำสถานที่ท่องเที่ยวโดยใช้ข้อมูลจากเครือข่ายสังคม," *วารสารวิทยาศาสตร์บูรพา*, ปีที่ 20, ฉบับที่ 1, หน้า 209-226, มกราคม – มิถุนายน 2558.

[27] ธนพล พุกเส็ง และ คณะ, "ระบบผู้แนะนำโดยอาศัยความไว้วางใจร่วมกับผู้เชี่ยวชาญ," *วารสารวิทยาศาสตร์และเทคโนโลยี*, ปีที่ 23, ฉบับที่ 3, หน้า 517-537, กรกฎาคม - กันยายน 2558.

ภาคผนวก

ภาคผนวก ก

โปรแกรมระบบ (System Code)

Category Similarity

```
import gensim
import nltk
from nltk.tokenize import  sent_tokenize
import re
from pythainlp.tokenize import Tokenizer
from pythainlp.corpus.common import thai_words
from pythainlp.corpus import thai_stopwords


file_docs = []

with open ('Content-Separate-Cate.csv',encoding='utf8') as f:
    tokens = sent_tokenize(f.read())
    for line in tokens:
        file_docs.append(line)

print("Number of documents:",len(file_docs))
from pythainlp.tokenize import word_tokenize



#word_tokenize(content,engine="newmm",          keep_whitespace=False
,custom_dict=trie)

gen_docs = [[w.lower() for w in word_tokenize
          (file_docs,engine="newmm", keep_whitespace=False)]
```

```
            for file_docs in file_docs]
print(gen_docs)

dictionary = gensim.corpora.Dictionary(gen_docs)
print(dictionary.token2id)
corpus = [dictionary.doc2bow(gen_doc) for gen_doc in gen_docs]
print(corpus)
import numpy as np
tf_idf = gensim.models.TfidfModel(corpus)

for doc in tf_idf[corpus]:
    [[dictionary[id], np.around(freq, decimals=2)] for id, freq in doc]
    tf_idf
sims                                                                    =
gensim.similarities.Similarity('',tf_idf[corpus],num_features=len(dictionary))

print(sims)
print(type(sims))
file2_docs = []

with open ('Content_develop_test.csv',encoding='utf8') as f:
    tokens = sent_tokenize(f.read())
    for line in tokens:
        file2_docs.append(line)

print("Number of documents:",len(file2_docs))
for line in file2_docs:
```

```
    query_doc = [w.lower() for w in word_tokenize(line)]
   # print(query_doc)
    query_doc_bow = dictionary.doc2bow(query_doc) #update an existing
dictionary and create bag of words
    # perform a similarity query against the corpus
    query_doc_tf_idf = tf_idf[query_doc_bow]
    # print(document_number, document_similarity)
    print('Comparing Result:', sims[query_doc_tf_idf])
import numpy as np


sum_of_sims =(np.sum(sims[query_doc_tf_idf], dtype=np.float32))
print(sum_of_sims)
```

Content based filtering

```
import pandas as pd
metadata = pd.read_csv('test_12.csv', low_memory=False,
error_bad_lines=False, encoding = 'utf-8')
metadata.head(55)

metadata['content'].head()
metadata['title'].head()
print(metadata.dropna(inplace=True))

from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import TfidfVectorizer

metadata['content'] = metadata['content'].fillna('')
tfidf = TfidfVectorizer(stop_words='english')
tfidf_matrix = tfidf.fit_transform(metadata['content'])
print(tfidf_matrix)

doc_term_matrix = tfidf_matrix.todense()
print(doc_term_matrix)
df = pd.DataFrame(doc_term_matrix,
                  columns=tfidf.get_feature_names(),
                  index=metadata.index)

cos=cosine_similarity(df, df)
print(cos)
```

```
def give_recommendation(title, cos=cos):


    indices = pd.Series(metadata.index,
                        index=metadata['title']).drop_duplicates()
    idx = indices[title]
    cos_scores = list(enumerate(cos[idx]))
    print(cos_scores)
    cos_scores = sorted(cos_scores, key=lambda x: x[1], reverse=True)
    print(cos_scores)
    cos_scores = cos_scores[1:5]
    print(cos_scores)
    movie_indices = [i[0] for i in cos_scores]
    return metadata['title'].iloc[movie_indices]
give_recommendation('ทำ Content ยังไงให้ปัง !!?')
```

Collaborative filtering

```
import pandas as pd
from scipy.spatial.distance import cosine
data = pd.read_csv('Data-content-cosine-collaborative-2.csv', low_memory=False, error_bad_lines=False, encoding = 'utf-8')
print(pd.__version__)

data.head(3).ix[:,2:8]
data_germany = data.drop('Client Id', 1)
data_ibs = pd.DataFrame(index=data_germany.columns,columns=data_germany.columns)
print(data_ibs)

import sys
# Lets fill in those empty spaces with cosine similarities
# Loop through the columns
for i in range(0,len(data_ibs.columns)) :
    # Loop through the columns for each column
    for j in range(0,len(data_ibs.columns)) :
      # Fill in placeholder with cosine similarities
      data_ibs.ix[i,j] = 1-cosine(data_germany.ix[:,i],data_germany.ix[:,j])
    print(data_ibs.ix[i,j])

# Create a placeholder items for closes neighbours to an item
data_neighbours = pd.DataFrame(index=data_ibs.columns,columns=range(1,11))
```

```
# Loop through our similarity dataframe and fill in neighbouring item names
for i in range(0,len(data_ibs.columns)):
    data_neighbours.ix[i,:10]                                    =
data_ibs.ix[0:,i].sort_values(ascending=False)[:10].index
print(data_neighbours.ix[i,:10])
# --- End Item Based Recommendations --- #
```

ภาคผนวก ข
การตีพิมพ์ผลงานวิจัย

ได้เข้าร่วมและตีพิมพ์งานวิจัยในงานประชุมวิชาการ TNIAC 2020

TNIAC 2020

IT-3

การพัฒนาระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา

An article recommendation system for advertising media business company website

1st นิติรุจน์ อติไชยทักษ์
Nitirut Atichaipitak
สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย-ญี่ปุ่น
Information Technology Program, Faculty of Information Technology, Thai-Nichi Institute of Technology
Bangkok / Thailand
at.nitirut_st@tni.ac.th

2nd ณัฐกิตติ์ จิตรเอื้อตระกูล
Nattagit Jiteurtragool
สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย-ญี่ปุ่น
Information Technology Program, Faculty of Information Technology, Thai-Nichi Institute of Technology
Bangkok / Thailand
nattagit@tni.ac.th

3rd ประจักษ์ เฉิดโฉม
Prajak Chertchom
สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย-ญี่ปุ่น
Information Technology Program, Faculty of Information Technology, Thai-Nichi Institute of Technology
Bangkok / Thailand
Prajak@tni.ac.th

บทคัดย่อ **— จากข้อมูลสำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์ได้จัดทำ รายงานผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทย ร้อยละ 48.3% ให้ความสำคัญกับการอ่านหนังสือหรือบทความออนไลน์ ดังนั้นผู้ใช้อินเทอร์เน็ตจึงให้ความสำคัญกับการเลือกอ่านหนังสือหรือบทความออนไลน์ โดยเลือกอ่านในสิ่งที่ผู้ใช้อินเทอร์เน็ตสนใจมากขึ้น การเลือกอ่านหนังสือหรือบทความออนไลน์ในเว็บไซต์ใดเว็บไซต์หนึ่งที่มีหนังสือหรือบทความออนไลน์เป็นจำนวนมากมีความยุ่งยากไม่สะดวกและไม่รวดเร็วปัจจุบันมีเทคโนโลยีที่เรียกว่า ระบบแนะนำบทความอัตโนมัติ (Recommendation system) เป็นเครื่องมือที่ช่วยให้ ผู้อ่าน หนังสือหรือบทความออนไลน์บนเว็บไซต์ ได้รับข้อมูลที่ต้องการค้นหาได้รวดเร็วขึ้น โดยระบบพยายามที่จะแนะนำหนังสือหรือบทความออนไลน์ ให้ตรงกับความสนใจของผู้อ่าน ผู้วิจัยจึงมีแนวคิดในการออกแบบระบบเว็บไซต์แนะนำ บทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาโดยการใช้ขั้นตอนวิธีการคัดกรองข้อมูลแบบอิงเนื้อหา กับการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม ดังนั้นบทความวิจัยนี้จะเสนอระบบแนะนำข้อมูลบทความด้านการตลาดออนไลน์เฉพาะบุคคล โดยใช้ประโยชน์จากข้อมูลบน Google Analytic ช่วยวิเคราะห์หาความสนใจของผู้ใช้ ได้อย่างมีประสิทธิภาพบนพื้นฐานความต้องการของผู้ใช้ระบบได้**

คำสำคัญ **— ระบบแนะนำบทความ, วิธีการคัดกรองข้อมูลแบบอิงเนื้อหา, วิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม, Google Analytic**

I. บทนำ

ปัจจุบันเว็บไซต์บทความ ข้อมูล ข่าวสาร เข้ามามีบทบาทกับชีวิตเราเป็นอย่างมาก ไม่ว่าจะเป็นการให้ข้อมูล การค้นหาข้อมูล หรือการทำธุรกิจเว็บไซต์เป็นที่เก็บข้อมูลและรวบรวมความรู้ทุกอย่าง ซึ่งข้อมูลต่างๆได้เพิ่มขึ้นมากจนทำให้ผู้ใช้ยากที่จะหาบทความที่ตัวเองสนใจได้

เครื่องมือค้นหาที่ดีจึงต้องสามารถแสดงคำตอบที่ตรงกับใจของผู้ค้นหาและช่วยผู้ค้นหาในกรณีที่ไม่สามารถนึกถึงคีย์เวิร์ดที่แม่นยำได้ เพื่อที่จะแก้ปัญหาเหล่านี้ ระบบแนะนำ (Recommendation System)ได้ถูกคิดค้นและสร้างขึ้นมา ตัวอย่างเช่น ถ้าคุณต้องการอ่านบทความให้ความรู้ที่มีความเกี่ยวข้องกัน ก่อนเข้าไปนึกได้ว่าอยากอ่านเรื่องการตลาดออนไลน์ที่เป็นบทความล่าสุดเมื่ออ่านจบแล้วจะพบว่าเว็บไซต์ส่วนใหญ่จะเสนอแนะบทความล่าสุดให้อ่านซึ่งไม่ตรงใจกับผู้อ่านบทความเพราะว่าต้องใช้เวลาในการค้นหา บางคนตัดใจไม่เลือกหรือปิดเว็บไซต์ไม่อ่านไปเลยเพราะหาอ่านบทความที่ชอบไม่เจอ แต่ถ้ามีระบบแนะนำบทความ หน้าที่ของระบบคือระบบจะเก็บประวัติในการเลือกอ่านของผู้ใช้งานไว้ เช่นถ้าผู้ใช้งานมักจะอ่านบทความเรื่องการตลาด ระบบแนะนำก็จะแสดงรายการบทความทั้งหมดที่เกี่ยวข้องกับบทความการตลาดที่มีอยู่ทั้งหมดในเว็บไซต์ให้คุณโดยคุณไม่จำเป็นต้องจำชื่อเรื่องของบทความได้ หรือ ไม่จำเป็นต้องเสียเวลาค้นหา

ระบบแนะนำบทความคือ ระบบค้นบทความให้ตรงกับบทความที่ต้องการของผู้ใช้โดยใช้เทคนิคการค้นบทความร่วมกับข้อมูลผู้ใช้สามารถทราบได้ว่ามีตัวเลือกอะไรบ้าง ซึ่งนอกจากจะช่วยแนะนำให้ผู้ใช้ที่มีข้อมูลผู้ใช้ในระบบ สามารถได้รับข้อมูลได้ตรงตามความสนใจของผู้ใช้แล้ว ระบบแนะนำข้อมูลยังสามารถช่วยให้ผู้ใช้ที่ไม่มีความรู้ประสบการณ์ช่วยในการตัดสินใจได้รับข้อมูลได้ตรงตามความต้องการ โดยทำการวิเคราะห์ความต้องการของผู้ใช้ ซึ่งระบบแนะนำข้อมูลโดยทั่วไปมักใช้คำสำคัญ (Keyword) ที่ผู้ใช้ใส่เข้ามาในระบบไปเปรียบเทียบกับคำสำคัญในเอกสารเพื่อหาเอกสารที่เกี่ยวข้อง แล้วนำรายการเอกสารที่เกี่ยวข้องมาแสดงให้กับผู้ใช้

ปัจจุบันได้มีการทำวิจัยและพัฒนาเกี่ยวกับระบบแนะนำสินค้าหรือบริการเพื่อนำเสนอต่อผู้บริโภคมากขึ้น เช่น ระบบนำเสนอภาพยนต์ ระบบนำเสนอการท่องเที่ยว ระบบนำเสนอสินค้าออนไลน์ต่างๆ เช่น [1] งานวิจัยการพัฒนาระบบแนะนำการเลือกสาขาเพื่อศึกษาต่อระดับอาชีวศึกษาโดยใช้เทคนิคการคัดกรองข้อมูลแบบผสมระหว่างการคัดกรองข้อมูลแบบอิงเนื้อหากับการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม งานวิจัยนี้เป็นการวิจัยในเชิงคุณภาพโดยได้ทำการใช้ผู้เชี่ยวชาญและผู้ใช้งานในการประเมิน สรุปได้ว่าระบบที่พัฒนาขึ้นนี้มีความเหมาะสมในระดับดีและสามารนำไปประยุกต์ใช้งานได้อย่างมีความเหมาะสม [2] งานวิจัยและพัฒนาระบบแนะนำร้านอาหารอัตโนมัติบนสมาร์ทโฟนโดยใช้ข้อมูลเชิงตำแหน่งและรายการอาหาร โดยเทคนิคส่วนการคัดกรองข้อมูลแบบอิงเนื้อหา โดยงานวิจัยนี้ กล่าวว่า แอปพลิเคชันบนสมาร์ทโฟนส่วนใหญ่ไม่สามารถ

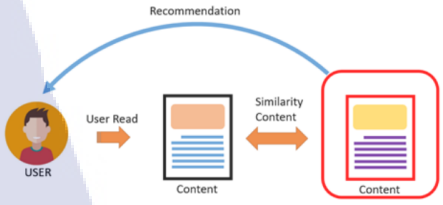
แนะนำรายการอาหารจากรายการอาหารส่วนประกอบและรสชาติได้และไม่สามารถค้นหาร้านอาหารจากรายการอาหารที่ต้องการได้ ดังนั้นผู้วิจัยจึงมีแนวคิดในการพัฒนาระบบเพื่อแก้ไขปัญหาดังกล่าว [3] งานวิจัยและพัฒนาระบบแนะนำแหล่งท่องเที่ยวของจังหวัดมหาสารคาม โดยใช้วิธีการกรองข้อมูลแบบผสมผสาน โดยงานวิจัยนี้ เป็นการวิจัยในเชิงคุณภาพโดยได้ทำการใช้ผู้เชี่ยวชาญและผู้ใช้งานในการประเมิน สรุปได้ว่าระบบที่พัฒนาขึ้นนี้สามารถให้คำแนะนำกับผู้ใช้ได้อย่างมีประสิทธิภาพ ใช้งานที่ง่าย สะดวกรวดเร็ว สามารถนำไปประยุกต์ใช้งานได้อย่างมีความเหมาะสม [4] ระบบให้คำแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน ซึ่งระบบแนะนำข้อมูลยังมีปัญหาผู้ใช้ใหม่และสินค้าใหม่ งานวิจัยนี้เสนอวิธีการแบบผสมผสานแบบใหม่เพื่อพัฒนากรอบความคิดในการแนะนำภาพยนตร์ด้วยวิธีการผสมผสานที่เกี่ยวข้องกับผู้ใช้ใหม่และข้อมูลภาพยนตร์ใหม่ เป็นการเตรียมข้อมูลจากการประมาณค่ากลุ่มเป็นการคำนวนผลรวมทั้งหมดภายในกลุ่มยกกำลังสองเพื่อให้ได้การจัดกลุ่มที่ดีที่สุด การจัดกลุ่มแบบฟัซซีซีมีนเพื่อลดความซับซ้อนของข้อมูลและเพิ่มความเกี่ยวข้องของค่าคะแนนของผู้ใช้กับ ภาพยนตร์ และการหาค่าความใกล้เคียงด้วยวิธีการ Item based การประมวลผลการแนะนำข้อมูล เป็นการหาค่าเพื่อนบ้านที่ใกล้ที่สุดและรวมน้ำหนัก วัดประสิทธิภาพด้วยค่าความแม่นนำ (Precision) และค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย (Mean Absolute Error : MAE) ผลการวิจัยพบว่า การแก้ปัญหา ผู้ใช้ใหม่ ประสิทธิภาพของค่าความแม่นยำร้อยละ 85 ค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย 2.1011 และการ แก้ปัญหาข้อมูลใหม่ประสิทธิภาพของค่าความแม่นยำร้อยละ 87 ค่าคลาดเคลื่อนสมบูรณ์เฉลี่ย 2.0031 โดยสรุป วิธีการแบบผสมผสานที่พัฒนาขึ้น สามารถแนะนำข้อมูลภาพยนตร์ได้อย่างมีประสิทธิภาพ

บทความวิจัยนี้นำเสนอแนวทางการพัฒนาระบบแนะนำข้อมูลบทความด้านการตลาดออนไลน์เฉพาะบุคคล โดยใช้ประโยชน์จากข้อมูลบน Google Analytic ช่วยวิเคราะห์หาความสนใจของผู้ใช้ ได้อย่างมีประสิทธิภาพบนพื้นฐานความต้องการของผู้ใช้ระบบได้ ผ่านช่องทางเว็บไซต์ของบริษัทการตลาดออนไลน์แห่งหนึ่ง จะเป็นการเพิ่มโอกาสที่ข้อมูลจะไปถึงลูกค้าได้รวดเร็ว และมากยิ่งขึ้น ทำให้ผู้ใช้งานสามารถนำข้อมูลเหล่านี้ไปใช้ประโยชน์ได้อย่างรวดเร็ว ระบบแนะนำเป็นเครื่องมือหนึ่งในการทำการตลาดออนไลน์ และถูกใช้เป็นตัวแปรในการเลือกการนำเสนอบทความใหม่ที่ตรงกับความต้องการผู้ใช้งานในแง่การตลาด ทำให้เข้าใจความต้องการจริงของผู้ใช้งานที่เป็นรายบุคคลมากขึ้นซึ่งจะมีความแม่นยำกว่าการนำเสนอแบบภาพรวม

II. ทฤษฎีที่เกี่ยวข้อง

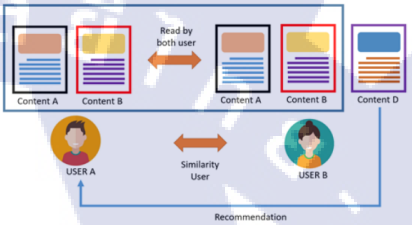
II.I ทฤษฎีระบบแนะนำ (Recommendation System)

วิธีการกรองแบบมีเนื้อหา (Content-based Filtering) เป็นวิธีของการกรองข้อมูลเพื่อหาความคล้ายคลึง การทำงานของเทคนิควิธีนี้ คือ จะ ให้ความสำคัญต่อเนื้อหาของข้อมูลเป็นสำคัญ เพื่อค้นหาข้อมูลที่ผู้ใช้แต่ละคนสนใจ ซึ่งวิธีนี้จะทำการคำนวณหาค่าความคล้ายคลึงกัน (Similarity Measure) ระหว่างข้อมูลหรือสินค้าที่ระบบมีอยู่กับความต้องการในตัวข้อมูลหรือสินค้าที่ผู้ใช้ต้องการ สำหรับการให้คะแนนความชอบต่อข้อมูลยังไม่ทั่วถึงหรือยังไม่ได้ให้คะแนนความชอบแก่ข้อมูลนั้นๆ จะไม่ส่งผลกระทบต่อเทคนิควิธีนี้ [2],[7]



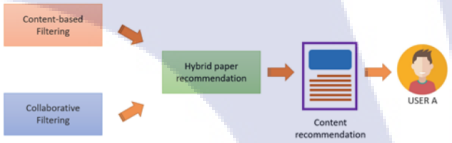
รูปที่ I. วิธีการกรองแบบมีเนื้อหา (Content-based Filtering)

วิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering) เป็นวิธีการที่นิยมใช้มากที่สุด วิธีการนี้จะเป็นการแนะนำข้อมูลให้กับผู้ใช้โดยพิจารณาจากประวัติการค้นหาของผู้ใช้คนอื่น ที่มีพฤติกรรมคล้ายกันซึ่งเคยค้นหาและทำการประเมินหรือให้คำคะแนนความนิยมกับข้อมูลนั้นๆ ไว้ก่อนหน้านี้แล้ว [4],[8]



รูปที่ II. การคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม (Collaborative Filtering)

วิธีการกรองแบบผสม (Hybrid Filtering) ซึ่งเป็นวิธี การที่ผสมผสานระหว่างสองวิธีการข้างต้นเข้าด้วยกัน อาทิ การใช้วิธีการกรองแบบมีเนื้อหารวมกับวิธีการคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม เพื่อแก้ไขปัญหาที่อาจจะเกิด (Cold-start problem) หรือความเบาบางของข้อมูล และช่วยเพิ่มประสิทธิภาพของระบบขึ้นอีกด้วย [4],[9]



รูปที่ III. วิธีการกรองแบบผสม (Hybrid Filtering)

II.II ทฤษฎีการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติ [10] หรือ Natural Language Processing(NLP) คือ กระบวนการรับข้อมูลที่เป็นข้อความต่อเนื่อง กระบวนการวิเคราะห์ระดับคำ วิเคราะห์ชนิดคำ วิเคราะห์โครงสร้างไวยากรณ์และความหมายของข้อความ โดยใช้ฐานความรู้ เพื่อจัดเก็บเข้าคอมพิวเตอร์และเรียกใช้ฐานข้อมูลเพื่อแสดงผลเป็นไปตามเป้าหมายที่ต้องการ

การตัดคำและการกรองคำหยุด (Word Segmentation and StopWord Filtering) การตัดคำภาษาไทยคือการแยกแต่ละคำในประโยคในเอกสารภาษาไทยซึ่งมีการเขียนคำติดกัน การตัดคำนั้นจำเป็นจะต้องมีเพื่อนำไปใช้ประโยชน์ในด้านประมวลผลภาษาธรรมชาติเช่น การแปลภาษา เป็นต้น

วิธีการตัดคำที่นิยมใช้กันมานานได้แก่วิธีการตัดคำโดยใช้กฎ วิธีการตัดคำโดยใช้พจนานุกรม และวิธีการตัดคำโดยใช้คลังข้อความ ปัจจุบันมีวิธีการตัดคำโดยใช้ NLP (Natural Language Processing) ซึ่งเป็นสาขาหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence หรือ AI) มาช่วยลดระยะเวลาและการค้นคำที่ไม่มีในระบบได้ถูกต้องและแม่นยำมากขึ้น ในขณะที่กรองคำที่ไม่เกี่ยวข้องหรือ คำหยุด (Stopword) เป็นการนำคำที่ไม่สำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง หรือหมายถึงคำที่ใช้กันโดยทั่วไปที่ไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง เช่น จึง ซึ่ง นี้ เป็นต้น

การประเมินค่าความสำคัญของวลี (TFIDF Calculate) [5] การประเมินค่าความสำคัญของวลี เป็นการคำนวณค่า ความสำคัญที่แต่ละวลีมีต่อเอกสาร พิจารณาความสำคัญของวลีและตำแหน่งของวลีในเอกสาร การคำนวณค่าความถี่ ใช้เทคนิคการกำหนดค่าความสำคัญให้กับคำในการแทนที่เอกสารเป็นแบบจำลองเวคเตอร์ (Vector-Space Model) คือค่าผลคูณระหว่าง TF (Term Frequency) และ (Inverse Document Frequency) ซึ่งเป็นการเปรียบเทียบความถี่ของวลีในเอกสารกับความถี่ของวลีในเอกสารอื่น ดังสมการ

$$TF * IDF = freq(P,D) * \log_2 \frac{N}{n_p} \quad (1)$$

โดยที่

D คือ เอกสาร

P คือวลี

freq(P,D) คือ ความถี่ที่พบวลี *P* ในเอกสาร *D*

n_p คือ จำนวนเอกสารในฐานข้อมูลที่พบวลี *P*

N คือจำนวนเอกสารทั้งหมดในฐานข้อมูล

III. การพัฒนาระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทของธุรกิจสื่อโฆษณา

ในบทนี้เป็นการเสนอระบบแนะนำข้อมูลบทความด้านการตลาดออนไลน์เฉพาะบุคคล ซึ่งผู้วิจัยได้ออกแบบกระบวนการทำงานของระบบแนะนำบทความอัตโนมัติ โดยมีการนำข้อมูลผู้ใช้งานที่เข้าไปอ่านบทความในเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาแห่งหนึ่ง จำนวน 10,000 Record มาจาก Google Analytics เพื่อสร้างเป็นประวัติการใช้งาน และ กับข้อมูลบทความ มาเก็บในดาต้าเบส และนำข้อมูลบทความในเว็บไซต์ทั้งหมด 54 บทความ มาจัดทำเป็นหมวดหมู่ สำหรับใช้ในการแนะนำบทความให้กับผู้อ่าน ดังที่แสดงในรูปที่ IV

รูปที่ IV ภาพรวมของระบบแนะนำบทความอัตโนมัติ

การพัฒนาระบบแนะนำโดยอ้างอิงจากคำสำคัญและกลุ่มของผู้ใช้งานหรือความชอบของผู้ใช้งานคนนั้น เป็นการพัฒนาโดยใช้เทคนิค Content-Based Filtering หรือ Collaborative Filtering ซึ่งผู้วิจัยพัฒนาโดยใช้โปรแกรมภาษา PHP หรือ Python และใช้ระบบฐานข้อมูลมายเอสคิวแอล (MySql) ในการเก็บข้อมูล

การสร้างแบ่งหมวดหมู่ย่อยของบทความ ด้วย Keyword เพื่อที่จะพัฒนาระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ได้นั้น ผู้วิจัยได้ออกแบบระบบสร้างแบ่งหมวดหมู่ย่อยของบทความด้วย Keyword โดยใช้ข้อมูลจากเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาแห่งหนึ่ง ซึ่งประกอบด้วยบทความทั้งหมด 54 บทความ ซึ่งขั้นตอนการหา Keyword สำหรับแบ่งหมวดหมู่ย่อยของบทความมีการทำงานดังรูปที่ V

รูปที่ V ขั้นตอนการหา Keyword สำหรับแบ่งหมวดหมู่ย่อยของบทความ

กระบวนการแบ่งหมวดหมู่ย่อยของบทความเริ่มจากการนำข้อมูลบทความทั้งหมด 54 บทความ มาเก็บไว้ในรูปแบบไฟล์ txt ทั้งหมด 54 ไฟล์ เพื่อนำเข้ากระบวนการการตัดคำ โดยใช้ชุดคำสั่งของไลบรารี่ PythaiNLP จากนั้นจึงเป็นกระบวนการกรองคำที่ไม่เกี่ยวข้อง (Stopword) และในขั้นตอนสุดท้ายจึงเป็นกระบวนการหาความถี่ของคำหรือการประเมินค่าความสำคัญของวลี (TFIDF Calculate) เพื่อนำมาเป็นคำสำคัญ หรือ keyword สำหรับจัดหมวดหมู่ย่อยของบทความเพื่อใช้สำหรับสืบค้นข้อมูลบทความและพัฒนาระบบแนะนำบทความ

ตัวอย่างบทความที่นำมาจากเว็บไซต์สามารถกำหนดหมวดหมู่หลักทั้งหมด4 หมวด1.SEO/SEM 2.Website 3.Social Media 4.Marketing และนำคำสำคัญของหมวดหมู่ย่อยใส่เข้าไปในหมวดหมู่หลัก เช่น บทความเรื่อง “เตือนภัย เมื่อ Page Facebook ปลิว” คำสำคัญคือคำว่า “Facebook” ก็จะนำคำนี้ไปใส่ในหมวดหมู่หลักชื่อ Social Media เป็นต้น

การสร้างฐานข้อมูลผู้ใช้งาน เพื่อให้คำแนะนำผู้ใช้ได้อย่างถูกต้องจึงต้องมีการสร้างฐานข้อมูลผู้ใช้งาน โดยเป็นการนำข้อมูลผู้ใช้งานเว็บไซต์จากข้อมูลย้อนหลังเพื่อการจัดกลุ่มผู้ใช้งานตามหมวดหมู่ของบทความ โดยมีขั้นตอนดังนี้

นำเข้าข้อมูลผู้ใช้งานเว็บไซต์จากข้อมูลย้อนหลัง ซึ่งเป็นข้อมูลย้อนหลังระยะเวลา 1 ปี จาก Google Analytic โดยมีข้อมูล รหัสผู้ใช้งาน และข้อมูลที่ผู้ใช้งานเข้าไปดูบทความที่ต้องการอ่านและนำข้อมูลผู้ใช้งานเว็บไซต์บริษัท มาเก็บไว้ที่ Database โดยมีโครงสร้างตามตารางที่ I

ตารางที่ I ตารางข้อมูลการอ่านบทความของผู้ใช้เบื้องต้นสำหรับระบบแนะนำบทความอัตโนมัติ

Id	Client Id	Date	Time	Content	URL
1	1154322	12/19/2019	9:31AM	5 Extensions	/post?id
2	2118717	12/19/2019	7:31AM	Facebooks	/post?id

ตารางข้อมูลการอ่านบทความของผู้ใช้เบื้องต้นสำหรับระบบแนะนำบทความอัตโนมัติ

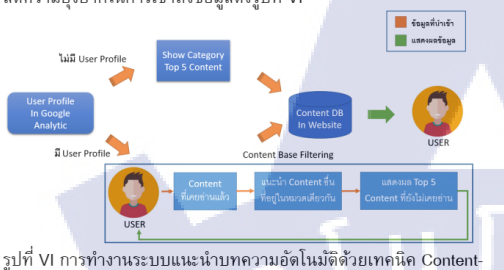
การนำข้อมูลผู้ใช้งานเว็บไซต์ย้อนหลัง มาประมวณผลใน Database เพื่อจัดกลุ่มว่าแต่ละ User ชอบอ่านบทความหมวดหมู่ไหน จัดลำดับบทความแต่ละหมวดหมู่ เป็นต้น สำหรับนำไปใช้งานต่อไป

ระบบแนะนำบทความอัตโนมัติบนเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณาโดยภายในระบบแนะนำบทความจะทำงานที่อ้างอิงจาก User Profile ใน Google Analytic ซึ่งแบ่งได้เป็น 2 กลุ่ม คือ ผู้ใช้งานใหม่และผู้ใช้ที่เคยมีประวัติการเข้าใช้

กรณีผู้ใช้งานที่ไม่มี User Profile ใน Google Analytic มาก่อน หรือเป็นผู้ใช้งานใหม่ ระบบจะทำการดึงข้อมูลจากฐานข้อมูลที่จัดอันดับหนึ่งถึงห้าที่มีคนอ่านมากที่สุดของบทความในแต่ละหมวดหมู่เพื่อแนะนำให้กับผู้ใช้งานกลุ่มนี้ที่ยังไม่มีประวัติการอ่านบทความมาก่อน และทำการเก็บ

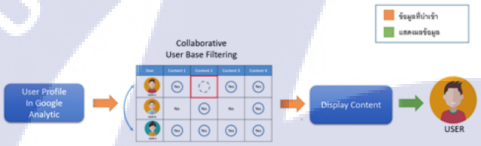
ข้อมูล

กรณีมี User Profile ระบบจะแนะนำบทความตามความต้องการของผู้ใช้งาน เช่น ผู้อ่านต้องการอ่านบทความเกี่ยวข้องกับหมวดหมู่เว็บไซต์ ระบบสามารถแนะนำบทความที่มีความคล้ายคลึงกับความต้องการของผู้อ่านได้ โดยใช้เทคนิค Content-based Filtering ทำให้ผู้ใช้งานเข้าถึงบทความที่ต้องการได้มากยิ่งขึ้น ตอบสนองความต้องการได้ทุกที่ทุกเวลา ลดความยุ่งยากในการเข้าถึงข้อมูลดังรูปที่ VI



รูปที่ VI การทำงานระบบแนะนำบทความอัตโนมัติด้วยเทคนิค Content-based Filtering

ระบบแนะนำบทความอัตโนมัติโดยใช้เทคนิค User-Based Collaborative Filtering ใช้ในกรณีมี User Profile โดยระบบจะนำข้อมูล Profile ของผู้อ่านบทความท่านอื่นมาเปรียบเทียบและแนะนำบทความให้กับผู้อ่านเป้าหมายลดความยุ่งยากในการเข้าถึงข้อมูลดังรูปที่ VII



รูปที่ VII การทำงานระบบแนะนำบทความอัตโนมัติด้วยเทคนิค

IV. แนวทางการประเมินผล

การทดสอบแบบแอลฟ่า (Alpha Test) เป็นการทดสอบเพื่อหาข้อบกพร่องหรือปัญหาที่เกิดขึ้นกับระบบ ดำเนินการทดลองใช้ขั้นต้น โดยผู้วิจัยเอง หลังจากนั้นจึงทำการแก้ไข ระบบให้มีความสมบูรณ์ยิ่งขึ้น

การทดสอบแบบเบต้า (Beta Test) เป็นการทดสอบความสมบูรณ์ของระบบ โดย ผู้วิจัยได้นำระบบไปทำการทดสอบความพึงพอใจต่อคุณภาพของระบบ โดยกลุ่มผู้เชี่ยวชาญ ทางด้านระบบแนะนำหนังสือ จำนวน 10 คน หลังจากนั้นผู้ วิจัยได้ทำการทดสอบ เพื่อหาความพึงพอใจของกลุ่มผู้ใช้จำนวน 30 คน และเมื่อทดสอบความพึงพอใจของกลุ่มผู้ใช้แล้ว ผู้วิจัยได้นำข้อมูลไปแก้ไขปรับปรุงระบบให้ตรงกับความต้องการของกลุ่มผู้ใช้ ให้มีความสมบูรณ์ยิ่งขึ้น หลังจากที่ได้ทำการพัฒนาระบบแล้ว ผู้พัฒนาระบบได้สร้างแบบสอบถามที่นำมาใช้ในการประเมินความพึงพอใจของผู้ใช้งานที่มีต่อระบบแนะนำบทความบนเว็บไซต์นี้ คือ แบบสอบถามเพื่อประเมินความพึงพอใจของระบบที่พัฒนาขึ้น โดยแบ่งการประเมินระบบตามลักษณะการทดสอบระบบออกเป็น 3 ด้าน ดังนี้

ด้าน Function Requirement Test เป็นการประเมินเพื่อดูว่าระบบที่ได้พัฒนาขึ้นนั้น มีความถูกต้องและมีความพึงพอใจตามความต้องการของผู้ใช้มากน้อยเพียงใด

ด้าน Functional Test เป็นการประเมินเพื่อดูว่าระบบที่ได้พัฒนาขึ้นนั้นมีความถูกต้องและมีความพึงพอใจสามารถทำงานได้ตามหน้าที่ที่มีอยู่ในระบบมากน้อยเพียงใด

ด้าน Usability Test เป็นการประเมินเพื่อดูว่าระบบที่ได้พัฒนาขึ้นนั้นมีความง่ายต่อการใช้งานมากน้อยเพียงใดและมีความเหมาะสมในการใช้งานแค่ไหน

เกณฑ์การพิจารณาความพึงพอใจของระบบ วัดจากคะแนนการประเมินผลของ กลุ่มตัวอย่างที่ได้ทำแบบประเมิน โดยต้องมีคะแนนเฉลี่ยการประเมินในระดับดีขึ้นไป จึงจะยอมรับว่าผู้ประเมินมีความพึงพอใจต่อการใช้งานระบบ และสามารถนำระบบไปใช้งานได้จริงตาม ขอบเขตที่กำหนดไว้ เกณฑ์การให้คะแนนแบ่งเป็น 5 ระดับ ดังแสดงในตารางที่ II

ตารางที่ II. แสดงเกณฑ์การพิจารณาความพึงพอใจของระบบ

ระดับเกณฑ์การให้คะแนน		ความหมาย
เชิงคุณภาพ	เชิงปริมาณ	
มากที่สุด	5	ระบบที่พัฒนามีความพึงพอใจมากที่สุด
มาก	4	ระบบที่พัฒนามีความพึงพอใจมาก
ปานกลาง	3	ระบบที่พัฒนามีความพึงพอใจมาก
น้อย	2	ระบบที่พัฒนามีความพึงพอใจน้อย
น้อยที่สุด	1	ระบบที่พัฒนามีความพึงพอใจน้อย

การแปลผลการประเมินโดยกำหนดการแปลความหมายของช่วงคะแนนดังแสดงในตารางที่ III

ตารางที่ III แสดงเกณฑ์การแปลผลการประเมิน

เกณฑ์การให้คะแนน	ความหมาย
4.51 – 5.00	ระบบที่พัฒนามีความพึงพอใจในระดับดีมาก
3.51-4.50	ระบบที่พัฒนามีความพึงพอใจในระดับดี
2.51-3.50	ระบบที่พัฒนามีความพึงพอใจในระดับปานกลาง
1.51-2.50	ระบบที่พัฒนามีความพึงพอใจในระดับน้อย
1.00-1.50	ระบบที่พัฒนามีความพึงพอใจในระดับน้อยที่สุด

V. สรุป

งานวิจัยชิ้นนี้เป็นการนำเสนอแนวคิดและกรอบแนวทางสำหรับการพัฒนาระบบแนะนำบทความอัตโนมัติสำหรับเว็บไซต์ของบริษัทธุรกิจสื่อโฆษณา เพื่อให้ผู้ใช้งานสามารถค้นพบบทความ หรือข้อมูลที่มีความน่าสนใจได้ง่ายมากขึ้น รวมถึงมีการนำเสนอแนวทางการประเมินประสิทธิภาพของระบบแนะนำบทความ เพื่อช่วยเป็นแนวทางสำหรับการพัฒนาและปรับปรุงระบบให้มีประสิทธิภาพมากขึ้น และนอกจากนี้ ด้วยแนวโน้มความสำคัญของเทคโนโลยีในอนาคตที่กำลังจะเกิดขึ้น ผู้วิจัยเล็งเห็นถึงประโยชน์ของระบบแนะนำบทความอัตโนมัติ ที่สามารถนำไปพัฒนาและต่อยอด เพื่อการใช้งานในด้านอื่นๆ อีกมากมาย นอกเหนือไปจากประโยชน์ในธุรกิจออนไลน์ อาทิเช่น ระบบแนะนำภาพยนต์ หนังสือ ร้านอาหาร เป็นต้น

กิตติกรรมประกาศ

ขอขอบคุณคณาจารย์และผู้เชี่ยวชาญที่ให้โอกาสในการนำเสนอแนวคิดในการทำงานวิจัยตามความสนใจของผู้ทำวิจัยในครั้งนี้ ทั้งให้คำปรึกษาทางด้านเทคนิกและ, แนวคิด, และทฤษฎีที่เกี่ยวข้องตลอดระยะเวลาในการทำงานนวิจัยชิ้นนี้จนสามารถกำหนดขอบเขตของงานวิจัยได้อย่างเหมาะสม และสามารถนำไปต่อยอดเพื่อทำเป็นชิ้นงานจริงต่อไป

เอกสารอ้างอิง

[1] ศศิษา หนูเสมียน. (2554). ระบบแนะนำการเลือกสาขาเพื่อศึกษาต่อระดับอาชีวศึกษา โดยเทคนิคการคัดกรองข้อมูลแบบผสมระหว่างการคัด กรองข้อมูลแบบอ้างอิงเนื้อหา กับ การคัดกรองข้อมูลแบบพึ่งพาผู้ใช้ร่วม. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาเทคโนโลยีสารสนเทศ วิทยาศาสตรมหาบัณฑิต มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

[2] ปาลิตา แสงสิริ. (2561). ระบบแนะนำร้านอาหารอัตโนมัติบนสมาร์โฟนโดยใช้ข้อมูลเชิงตำแหน่งและรายการอาหาร โดยเทคนิคส่วนการคัดกรองข้อมูลแบบอิงเนื้อหา. วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาเทคโนโลยีสารสนเทศ วิทยาศาสตรมหาบัณฑิต มหาวิทยาลัยสงขลานครินทร์.

[3] อรนันท์ เชาว์พานิช, ระบบแนะนำแหล่งท่องเที่ยวของจังหวัดมหาสารคามโดยใช้วิธีการกรองข้อมูลแบบผสมผสาน, Vocational Education Innovation and Research Journal, 3(1), 25-31.2019.

[4] ธรรมนูญ ปัญญาทิพย์. (2562). ระบบให้คำแนะนำภาพยนตร์ด้วยวิธีการผสมผสาน. ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม.

[5] นายวัฒนพัชร์ เก้าลิ้ม. (2557). การค้นหาคำสำคัญในเอกสารภาษาไทยโดยใช้เทคนิคการค้นหารูปแบบความสัมพันธ์ในเหมืองข้อมูล ปริญญาวิทยาศาตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ มหาวิทยาลัยศิลปากร.

[6] ทัศนีย์ อุทัยสุรี. (2556). การสกัดคำสำคัญจากบทคัดย่อภาษาอังกฤษ ปริญญาวิทยาศาตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ มหาวิทยาลัยศิลปากร.

[7] KOMPAN, Michal; BIELIKOVÁ, Mária. Content-based news recommendation. In: International conference on electronic commerce and web technologies. Springer, Berlin, Heidelberg, p. 61-72, 2010

[8] CHATURVEDI, Akshay Kumar; PELEJA, Filipa; FREIRE, Ana. Recommender system for news articles using supervised learning. arXiv preprint arXiv:1707.00506, 2017.

[9] SELONEN, Katri, et al. An Adaptive Recommender System for News Delivery. 2017.

[10] มาลินี นาคใหญ่. (2549). การวิเคราะห์ประโยคข้อมูล โดยอิงฐานความรู้ กรณีศึกษาการครองราชย์สมัยกรุงศรีอยุธยา ปริญญาวิทยาศาตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ มหาวิทยาลัยศิลปากร.

Certificate of Participation

This is to certify that

Nitirut Atichaipitak, Nattagit Jiteurtragool, Prajak Chertchom

have presented a paper titled

An Article Recommendation System for Advertising Media Business Company Website

in The 6th Thai-Nichi Institute of Technology Academic Conference (TNIAC 2020)
held at Thai-Nichi Institute of Technology, Bangkok, Thailand, May 21, 2020

Pichit Sukchareonpong

Assoc. Prof. Dr. Pichit Sukchareonpong

Organizing Chair