# ระบบแนะนำวารสารวิชาการให้กับผู้เขียนบทความ โดยใช้ข้อมูลภาษาไทยและภาษาอังกฤษจากบทความ

นายนิธิรันดร์ นุ่มนนท์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

# Journal Recommendation System for Author Using Thai and English information from Manuscript

Mr. Nithirun Numnonda



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2021

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | ระบบแนะนำวารสารวิชาการให้กับผู้เขียนบทความ โดยใช้ข้อมูลภาษาไทยและภาษาอังกฤษจากบทความ |
| โดย | นายนิธิรันดร์ นุ่มนนท์ |
| สาขาวิชา | วิทยาศาสตร์คอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | อาจารย์ ดร.เนื่องวงศ์ ทวยเจริญ |

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

---------- คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

---------- ประธานกรรมการ
(ศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)

---------- อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(อาจารย์ ดร.เนื่องวงศ์ ทวยเจริญ)

---------- กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.ณรงค์เดช กีรติพรานนท์)

นิธิรันดร์ นุ่มนนท์ : ระบบแนะนำวารสารวิชาการให้กับผู้เขียนบทความ โดยใช้ข้อมูลภาษาไทยและภาษาอังกฤษจากบทความ . ( Journal Recommendation System for Author Using Thai and English information from Manuscript ) อ.ที่ปรึกษาหลัก : อ. ดร.เนื่องวงศ์ ทวยเจริญ

ในปัจจุบันมีวารสารทางด้านวิชาการอยู่เป็นจำนวนมากหลากหลายประเภท ส่งผลให้ผู้เขียนบทความ ต้องใช้เวลามากไปกับการค้นหาคัดเลือกวารสารทางด้านวิชาการที่เหมาะสมกับเนื้อหาของแต่ละบทความของผู้เขียน ก่อนจะส่งบทความให้ทางบรรณาธิการวารสารทำการพิจารณารับบทความในลำดับถัดไป เนื่องจากทางบรรณาธิการได้รับบทความจำนวนมาก จึงทำให้ใช้เวลามากในการพิจารณาบทความ งานวิจัยฉบับนี้จึงเล็งเห็นว่าการนำระบบแนะนำเข้ามาช่วยวิเคราะห์เพื่อแนะนำวารสารที่เหมาะสมกับบทความนั้นจะทำให้กระบวนการตัดสินใจในการส่งบทความเพื่อตีพิมพ์มีประสิทธิภาพยิ่งขึ้น โดยจะใช้ข้อมูลจาก Thai Journals Online (ThaiJO) ซึ่งจะใช้ข้อมูลจากบทความภาษาไทยและบทความภาษาอังกฤษในการวิเคราะห์ในงานวิจัยนี้ โดยในงานวิจัยนี้รวมการศึกษาข้อมูลที่ใช้ การทำความสะอาดข้อมูล และการทำแบบจำลองสำหรับระบบแนะนำ โดยจะทำแบบจำลองจากการคำนวณหาความสำคัญจากข้อความด้วยเทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) และการวิเคราะห์ความคล้ายคลึงระหว่างบทความและวารสารโดยใช้ Cosine Similarity แล้วจึงจัดอันดับค่าคะแนนเพื่อแนะนำบทความที่เหมาะสม จากผลการทดลองในงานวิจัยนี้การตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold cross-validation) พบว่าเมื่อเรานำข้อมูลคำสำคัญและบทคัดย่อจากทั้งภาษาไทยและภาษาอังกฤษมารวมกัน ระบบสามารถแนะนำออกมาได้ค่าความแม่นยำที่วัดด้วย Hit Rate ได้ค่าสูงสุดที่ 0.87965 ซึ่งมากกว่าแบบจำลองที่ใช้ข้อมูลภาษาอังกฤษอย่างเดียว (0.84948) หรือ แบบจำลองที่ใช้ข้อมูลภาษาไทยอย่างเดียว (0.80383) และได้ค่าความแม่นยำที่สูงกว่าการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ และการทดลองแบบจำลองในลักษณะความคล้ายระหว่างบทความ

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

# # 6270142321 : MAJOR COMPUTER SCIENCE

KEYWORD: Journal recommendation system, TF-IDF, Cosine similarity, Thai and English articles

Nithirun Numnonda : Journal Recommendation System for Author Using Thai and English information from Manuscript . Advisor: NUENGWONG TUAYCHAROEN

There are thousands of academic journals in various fields of study. An article author must spend significant time searching and selecting a journal suitable for the article's content before submitting it to a journal for consideration. Since many articles are submitted to a journal at a time, it would take time for an editor to review, submit it to reviewers, and inform the results back to the author. Therefore, this research introduced a recommendation system to help the author choose an appropriate journal more effectively, based on TCI Thai Journals Online Database (ThaiJO). Data from Thai and English articles were used for analysis in this research. Our work involved studying the applied data, cleaning the data, and modeling, which includes calculating the importance of text by Term Frequency - Inverse Document Frequency (TF-IDF), calculating similarity scores between articles and journals using Cosine Similarity and then ranking the scores to recommend the most suitable journal. The experiment with 10-fold cross-validation shows that when we combine Thai and English keywords and abstract data, the accuracy in the form of hit rate is improved to 0.87965 from applying only English (0.84948) or Thai data (0.80383) and the accuracy of 10-fold cross-validation is better than the accuracy from 5-fold cross-validation and modeling using cosine similarity between research article.



Field of Study: Computer Science Student's Signature .....

Academic Year: 2021 Advisor's Signature .....

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีจากแรงสนับสนุน คำแนะนำ ความช่วยเหลือ และกำลังใจจากบุคคลหลายฝ่าย ผู้วิจัยจึงใคร่ขอใช้เนื้อหาในส่วนของกิตติกรรมประกาศเพื่อขอขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

ข้าพเจ้าขอขอบคุณ อาจารย์ ดร.เนื่องวงศ์ ทวยเจริญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ของข้าพเจ้า และคุณสภา จรรยาชัชวาลที่คอยให้คำแนะนำในการทำวิจัยเป็นอย่างดี จนทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วง

ข้าพเจ้าขอขอบคุณ คณะกรรมการการสอบวิทยานิพนธ์ศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล ผู้ช่วยศาสตราจารย์ ดร. ณรงค์เดช กีรติพรานนท์ ที่ได้สละเวลาในการให้ข้อเสนอแนะนำสิ่งที่ควรเพิ่มเติมลงในวิทยานิพพนธ์ฉบับนี้จบเสร็จสมบูรณ์

ข้าพเจ้าขอขอบคุณ เพื่อน ๆ ทุกคน แฟนและครอบครัว ที่สนับสนุนและให้กำลังใจเป็นอย่างมากในการศึกษาต่อระดับปริญญาโทในครั้งนี้

ข้าพเจ้าขอขอบคุณ อาจารย์ทุกท่าน ที่ได้สั่งสอนข้าพเจ้ามา ซึ่งความรู้แนวคิด และกระบวนการต่าง ๆ ที่ได้รับมา ล้วนมีประโยชน์ต่อวิทยานิพนธ์ฉบับนี้ทั้งสิ้น

ขอขอบพระคุณ Thai Journals Online (ThaiJO) ที่รวบรวมชุดข้อมูลวารสารทางด้านวิชาการที่เป็นประโยชน์อย่างยิ่งสำหรับการวิจัย

สุดท้ายนี้ข้าพเจ้าขอขอบคุณบิดา มารดาที่คอยอยู่เคียงข้างข้าพเจ้า และเป็นกำลังใจให้ข้าพเจ้าเสมอมา จนทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้ได้ลุล่วง และสำเร็จการศึกษาได้ในที่สุด



นิธิรันดร์ นุ่มนนท์

# สารบัญ

หน้า

# บทที่ 1
# บทนำ

## 1.1 ที่มาและความสำคัญ

ในปัจจุบันมีวารสารทางด้านวิชาการอยู่เป็นจำนวนมากหลากหลายประเภท ส่งผลให้ผู้เขียนบทความต่าง ๆ ต้องใช้เวลามากในการค้นหาคัดเลือกวารสารทางด้านวิชาการที่เหมาะสมกับเนื้อหาของแต่ละบทความของผู้เขียน ก่อนจะส่งบทความให้ทางบรรณาธิการวารสารทำการพิจารณารับบทความในลำดับถัดไป ซึ่ง Scopus [1] จัดทำดัชนีเนื้อหาจากมากกว่า 25,000 ชื่อที่มีการใช้งานและผู้เผยแพร่ 7,000 ราย ทั้งหมดผ่านการตรวจสอบและคัดเลือกอย่างเข้มงวดโดยคณะกรรมการตรวจสอบอิสระ ผู้ใช้สามารถเข้าถึงหนังสือหลายพันเรื่อง บัญชีผู้แต่งนับล้าน และข้อมูลอ้างอิง 1.7 พันล้านรายการ ในขณะที่ Thai Journals Online (ThaiJO) [2] เป็นแหล่งรวมวารสารวิชาการทุกสาขาวิชาที่ผลิตในประเทศไทย ทั้งสาขาวิทยาศาสตร์/เทคโนโลยี และมนุษยศาสตร์และสังคมศาสตร์ ซึ่งรวบรวมวารสารจำนวนมากกว่า 1,013 วารสารและมีจำนวนมากกว่า 18,542 เล่ม และมีบทความจำนวนมากกว่า 192,044 บทความ เนื่องจากมีบทความจำนวนมาก จึงทำให้ใช้เวลาในกระบวนการส่งบทความและกระบวนการพิจารณาของบรรณาธิการวารสาร หากถูกปฏิเสธการตีพิมพ์จะยิ่งทำให้ใช้เวลามากยิ่งขึ้น ประกอบกับการที่ในวารสารบางสาขามีเนื้อหาที่เป็นไปในทิศทางเดียวกัน แต่มีเนื้อหารายละเอียดแตกต่างกัน ยกตัวอย่างเช่น วารสารข่ายงานวิศวกรรมอุตสาหการไทย และวารสารสมาคมวิศวกรรมเกษตรแห่งประเทศไทย มีเนื้อหาไปในด้านวิศวกรรมศาสตร์เช่นเดียวกัน แต่มีความแตกต่างกันด้านเนื้อหาต่าง ๆ จึงทำให้เป็นความท้าทายของผู้เขียนบทความที่จะเลือกวารสารที่เหมาะสมกับบทความได้

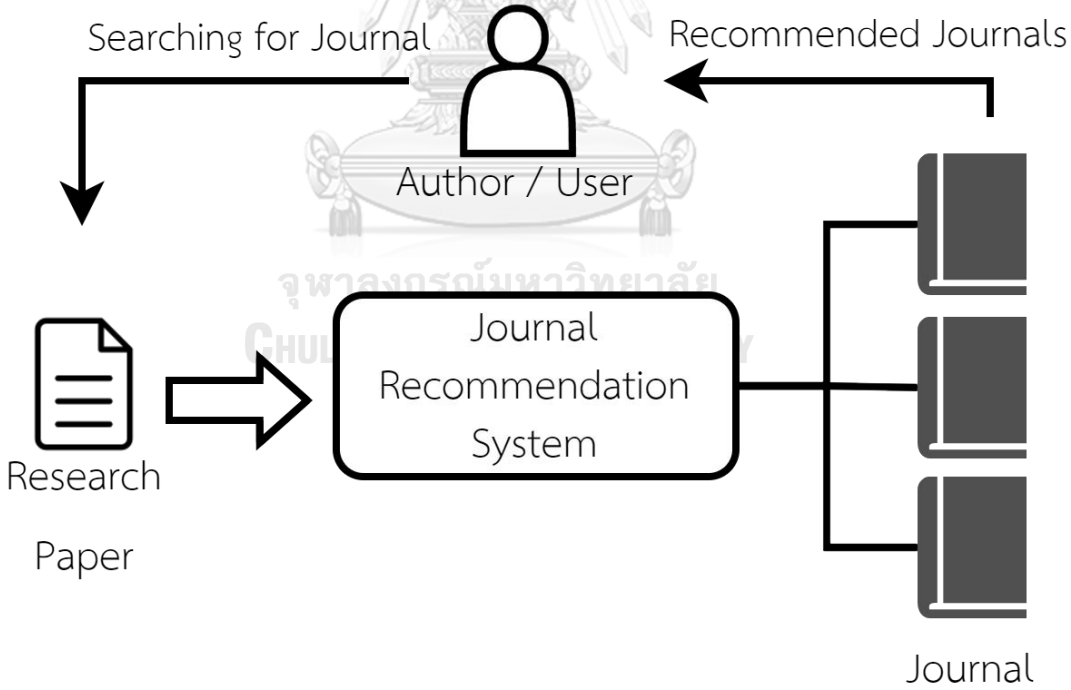


เพื่อความสะดวกในการคัดกรองเบื้องต้น จึงได้มีการนำระบบแนะนำ (Recommendation System) มาใช้ในงานวิจัยนี้ ซึ่งระบบแนะนำได้ถูกใช้อย่างแพร่หลายในปัจจุบัน ไม่ว่าจะเป็นเว็บบริการ สื่อส่งผ่านสัญญาณต่อเนื่อง (Streaming Media), การพาณิชย์อิเล็กทรอนิกส์ (E-Commerce), หรือแม้กระทั่ง สื่อสังคม (Social Media) ต่าง ๆ ล้วนมีระบบแนะนำเข้ามาช่วยคัดกรองเนื้อหาให้ผู้ใช้บริการได้พบกับสิ่งที่ตรงกับความสนใจของตนเองมากที่สุดก่อนเป็นลำดับต้น ๆ

ระบบแนะนำดังกล่าวจะนำข้อมูลที่มีอยู่ในระบบมาทำการวิเคราะห์ เพื่อให้สามารถแนะนำผลลัพธ์ต่าง ๆ ได้ตรงกับความตั้งใจของระบบนั้น ๆ ยิ่งมีข้อมูลในระบบเป็นจำนวนมากเท่าใด ยิ่งทำให้สามารถประมวลผลทำนายผลลัพธ์การแนะนำได้ดียิ่งขึ้นเท่านั้น ซึ่งระบบแนะนำมีการใช้ขั้นตอนวิธีและวิธีการวิเคราะห์ที่หลากหลาย ทั้งการใช้ข้อมูลจากตัวอักษรจากทั้งบทความ ใช้ชื่อบทความ บทคัดย่อ คำสำคัญของบทความ บทความที่เคยตีพิมพ์ของเจ้าของบทความ รวมถึง บทความที่อ้างอิง

และถูกอ้างอิงจากบทความ ในขณะที่การวิเคราะห์และพัฒนาแบบจำลองมีวิธีการที่หลากหลาย ไม่ว่าจะเป็น สามารถคำนวณหาความสำคัญจากข้อความด้วยเทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) หรือบางระบบใช้ความถี่ของคำ (Term Frequency) แทน ในบางระบบแนะนำได้มีการใช้การจัดกลุ่ม (Clustering) เข้ามาช่วย มีการคำนวณค่าความคล้ายด้วย Cosine Similarity แล้วจัดอันดับค่าคะแนนออกมาแนะนำ และในบางระบบได้นำค่าคะแนนความคล้ายไปใช้ในขั้นตอนวิธีในการปรับค่าน้ำหนักหรือขั้นตอนวิธีต่าง ๆ ในการพัฒนาระบบแนะนำให้ดียิ่งขึ้น

งานวิจัยฉบับนี้จึงเล็งเห็นว่าการนำระบบแนะนำเข้ามาช่วยวิเคราะห์เพื่อแนะนำวารสารที่เหมาะสมกับบทความนั้น จะทำให้การคัดเลือกวารสารมีประสิทธิภาพต่อผู้เขียนบทความ ทางผู้วิจัยจึงนำข้อมูลวารสารและบทความต่าง ๆ ที่มีอยู่ มาทำการวิเคราะห์และจัดสร้างระบบแนะนำ โดยจะใช้ข้อมูลจากบทความภาษาไทยและบทความภาษาอังกฤษในการวิเคราะห์และพัฒนาแบบจำลองขึ้น จากนั้นทำการทดลองเปรียบเทียบระหว่างแบบจำลองแต่ละภาษา นอกจากนั้น การทำระบบแนะนำวารสารวิชาการไทยจะยังช่วยส่งเสริมวารสารวิชาการไทยมีเครื่องมืออำนวยความสะดวกและส่งเสริมให้ผู้เขียนตีพิมพ์ในวารสารวิชาการไทยมากยิ่งขึ้น




*รูปที่ 1 แสดงภาพรวมของระบบแนะนำวารสารวิชาการให้กับผู้เขียนบทความ*

## 1.2 วัตถุประสงค์

งานวิจัยนี้นำเสนอแนวทางการสร้างระบบแนะนำวารสารทางวิชาการที่เหมาะสมกับบทความประเภทต่าง ๆ เพื่อให้ผู้เขียนบทความสามารถค้นหาและคัดเลือกวารสารที่เหมาะสมกับบทความของตนเองได้อย่างรวดเร็วและมีประสิทธิภาพ

## 1.3 ขอบเขตของงาน

งานวิจัยนี้นำเสนอแนวทางการสร้างระบบแนะนำวารสารทางด้านวิชาการที่เหมาะสมกับบทความ โดยนำข้อมูลที่มีคุณลักษณะดังต่อไปนี้มาใช้ในการพัฒนาระบบ

1. ข้อมูลบทความและวารสารทางวิชาการที่ใช้ในงานวิจัยนี้ มาจากข้อมูลวารสารวิชาการไทยที่เผยแพร่ในเว็บ ThaiJO
2. ข้อมูลบทความและวารสารทางวิชาการที่ใช้ในงานวิจัยนี้ เป็นข้อมูลจากบทความภาษาไทยและบทความภาษาอังกฤษ
3. ข้อมูลบทความและวารสารทางวิชาการที่นำมาวิเคราะห์จะคัดมาเฉพาะข้อมูลที่มีบทคัดย่อและมีคำสำคัญเท่านั้น
4. ข้อมูลบทความและวารสารทางวิชาการที่นำมาวิเคราะห์เป็นข้อมูลจากวารสารที่เกี่ยวข้องกับวิศวกรรมศาสตร์และเทคโนโลยีเท่านั้น



## 1.4 ประโยชน์ที่ได้รับ

1. ผู้วิจัยสามารถสร้างแบบจำลองแนะนำวารสารทางด้านวิชาการที่เหมาะสมกับบทความจากข้อมูลคำสำคัญและบทคัดย่อ และผู้ที่สนใจสามารถนำข้อมูลคำสำคัญและบทคัดย่อไปหาวารสารทางด้านวิชาการที่เหมาะสมกับบทความเบื้องต้นได้ด้วยแบบจำลองนี้
2. ผู้วิจัยสามารถสร้างแบบจำลองแนะนำวารสารทางด้านวิชาการจากข้อมูลทั้งภาษาไทยและภาษาอังกฤษ และผู้ที่สนใจสามารถนำข้อมูลทั้งภาษาไทยและภาษาอังกฤษไปหาวารสารทางด้านวิชาการที่เหมาะสมกับบทความเบื้องต้นได้ด้วยแบบจำลองนี้
3. ผู้วิจัยและผู้อ่านสามารถนำข้อมูลวารสารทางวิชาการเพิ่มเติมมาใช้เพื่อเพิ่มประสิทธิภาพของแบบจำลองแนะนำ
4. ผู้วิจัยและผู้อ่านสามารถนำกรอบงานวิจัยนี้ไปประยุกต์กับงานอื่น ๆ เช่นนำไปใช้ในระบบแนะนำที่มีการใช้ข้อมูลหลายภาษา

5. ผู้วิจัยส่งเสริมให้วารสารไทยมีระบบแนะนำและสิ่งอำนวยความสะดวกเพื่อสนับสนุนให้มีงานวิจัยที่ตพิมพ์เผยแพร่ในวารสารวิชาการภาษาไทยมากยิ่งขึ้น

## 1.5 ขั้นตอนการดำเนินงาน

1. ศึกษาทฤษฎีและงานที่เกี่ยวข้อง
2. เก็บข้อมูลและศึกษาข้อมูลที่ใช้ในงานวิจัย
3. เตรียมข้อมูล
4. ออกแบบแบบจำลองสำหรับการแนะนำข้อมูล
5. ทดสอบแบบจำลองและปรับปรุงแบบจำลอง
6. สรุปผลและเขียนวิทยานิพนธ์

## 1.6 ผลงานวิจัยที่ได้รับการตีพิมพ์

“Journal Recommendation System for Author Using Thai and English Information from Manuscript” โดย นิธิรันดร์ นุ่มนนท์, สภา จรรยาชัชวาล และ เนื่องวงศ์ ทวยเจริญ ได้ตีพิมพ์และนำเสนอใน งานประชุมวิชาการ “the 18th International Conference on Computing and Information Technology (IC2IT 2022)” จัดขึ้น ณ ประเทศไทย ระหว่างวันที่ 19 ถึง 20 พฤษภาคม 2565

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้องเป็นความรู้พื้นฐานที่เกี่ยวข้องกับงานวิจัยนี้ แบ่งออกเป็น 4 หัวข้อดังนี้

1. การทำความสะอาดข้อความ
2. การตัดคำภาษาไทย
3. การคำนวณค่าคำที่สำคัญในเอกสาร
4. การวิเคราะห์ความคล้ายคลึงของเอกสาร
5. การประเมินความถูกต้องของแบบจำลองด้วยการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ (K-fold Cross Validation)

## 2.1 การทำความสะอาดข้อความ

ข้อมูลข้อความที่เราได้มาจากบทความ บางครั้งจะมีตัวอักขระพิเศษ หรือมีเครื่องหมายต่าง ๆ ที่เราไม่ต้องการ ยกตัวอย่างเช่น =+!?.*/ ซึ่งตัวอักษรเหล่านี้จะส่งผลต่อการตัดคำ รวมถึงการประมวลผลภาษา เพื่อเพิ่มความแม่นยำของการวิเคราะห์จึงจะต้องทำการทำความสะอาดข้อความขั้นต้น [3] ก่อนนำไปประมวลผลแบบจำลอง โดยวิธีการทำความสะอาดข้อความทั้งการตัดคำอักขระที่ไม่จำเป็นหรือไม่สำคัญ รวมถึงการลดรูปของคำ เพื่อทำให้สามารถนำคำที่มีความหมายในลักษณะเดียวกันไปประมวลให้แม่นยำยิ่งขึ้น



**ตัวอย่างข้อความก่อนทำความสะอาดข้อความ:**
ตั้งแต่จะทำโดยข้อมูล/เตรียมข้อมูล* + สร้างแบบจำลอง = การทำงานวิจัย?
**ตัวอย่างข้อความหลังทำความสะอาดข้อความด้วยการตัดคำ อักขระที่ไม่จำเป็นหรือไม่สำคัญ:**
ทำข้อมูล เตรียมข้อมูล สร้างแบบจำลอง ทำงานวิจัย
*(ตัวอักษรที่ถูกตัดไปได้แก่ "/ * = ?" และคำที่ไม่สำคัญถูกตัดได้แก่ "ตั้งแต่ จะ การ")*

*รูปที่ 2 แสดงตัวอย่างการทำความสะอาดข้อความ*

## 2.2.การตัดคำภาษาไทย

การตัดคำ (Word Segmentation) [4] เป็นการแบ่งข้อความเพื่อหาขอบเขตของหน่วยคำ ซึ่งเป็นขั้นตอนพื้นฐานของการประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP)

โดยในภาษาไทยมีการเขียนติด ๆ กันโดยไม่ได้มีช่องว่างเว้นวรรค ยกเว้นการเว้นวรรคเป็นระยะเพื่อให้ผู้อ่านได้หยุดพักและเข้าใจความหมายเป็นตอน ๆ ทำให้มีความซับซ้อนมากกว่าภาษาอังกฤษที่มีการเว้นวรรคเพื่อแบ่งแยกคำ ในการอำนวยความสะดวกในการตัดคำภาษาไทยมีไลบารี PyThaiNLP [5] ซึ่งเป็นไลบารีภาษาไพทอนสำหรับประมวลผลภาษาธรรมชาติ โดยเน้นภาษาไทย โดยมีความสามารถพื้นฐานสำหรับการประมวลผลภาษาไทย ซึ่งประกอบด้วยการตัดคำภาษาไทย, ตรวจสอบการสะกดคำ

**ข้อความต้นฉบับ:**
โอเคบ่เรารักภาษาถิ่น
**ตัวอย่างการตัดคำภาษาไทยด้วย PyThaiNLP:**
['โอเค', 'บ่', 'เรา', 'รัก', 'ภาษาถิ่น']

*รูปที่ 3 แสดงตัวอย่างการตัดคำภาษาไทยด้วย PyThaiNLP*

**คำต้นฉบับ:**
ศาธารณสุบ
**ตัวอย่างการตรวจสอบการสะกดคำภาษาไทยด้วย PyThaiNLP:**
['สาธารณสุข']

*รูปที่ 4 แสดงตัวอย่างการตรวจสอบการสะกดคำภาษาไทยด้วย PyThaiNLP*



### 2.3 การคำนวณค่าคำที่สำคัญในเอกสาร

เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) [6] เป็นหนึ่งในวิธีหาคำที่สำคัญในเอกสาร โดยดูจากเนื้อหาของเอกสารทั้งหมด เนื่องจากความถี่ของคำเพียงอย่างเดียวไม่ได้บ่งบอกความสำคัญของคำในเอกสารได้ เพราะคำที่มีมากในเอกสารแต่อาจไม่ได้มีความสำคัญในเอกสาร เช่น คำหยุด (Stop Word) เป็นต้น การใช้เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร จะทำให้ได้คำที่สำคัญจากในเอกสาร โดยประกอบไปด้วยสองส่วน ความถี่ของคำ (Term Frequency) และ ส่วนกลับความถี่ของเอกสาร (Inverse Document Frequency)

ในส่วนของความถี่ของคำ เนื่องจากว่าแต่ละเอกสาร อาจจะมีความยาวที่แตกต่างกัน ดังนั้น การคำนวณความถี่ของคำ จึงมักจะหารด้วยความยาวหรือจำนวนคำของเอกสารนั้น โดยคำนวณได้จากสูตรต่อไปนี้

$$TF_{(term,document)} = \frac{f_{(term,document)}}{\sum_{term' \in document} f_{(term',document)}} \quad \text{- (1)}$$

ในส่วนของส่วนกลับความถี่ของเอกสารใช้สำหรับวัดความสำคัญของคำในเอกสารทั้งหมด โดยถ้าเป็นคำที่พบได้ในหลายเอกสาร คำนั้น ๆ ย่อมเป็นคำทั่วไปที่มีความสำคัญลดลง โดยคำนวณได้จากสูตรต่อไปนี้

$$IDF_{(term,allDocument)} = \log \frac{N}{df(t)} \quad \text{- (2)}$$

สุดท้าย ในการคำนวณหาค่าความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร ทำได้โดยนำค่าความถี่ของคำและค่าส่วนกลับความถี่ของเอกสารมาคำนวณร่วมกัน โดยคำนวณได้จากสูตรต่อไปนี้

$$TF\ IDF = TF_{(term,document)} \times IDF_{(term,allDocument)} \quad \text{- (3)}$$

## 2.4 การวิเคราะห์ความคล้ายคลึงของเอกสาร



การวิเคราะห์ความคล้ายคลึงของเอกสารนั้นมีวิธีการทางสถิติเพื่อใช้วัดค่าความคล้ายคลึงมีหลายวิธีเช่น Jaccard Similarity และ Cosine Similarity [7] โดยในงานวิจัยนี้เลือกใช้ Cosine Similarity [8] ซึ่งเป็นการคำนวณจากมุมโคไซน์ระหว่างสองเวกเตอร์ ถ้าค่าโคไซน์เข้าใกล้ 1 หรือเท่ากับ 1 หมายความว่าทั้งสองเวกเตอร์มีความคล้ายคลึงกันมากหรือเป็นเวกเตอร์เดียวกัน

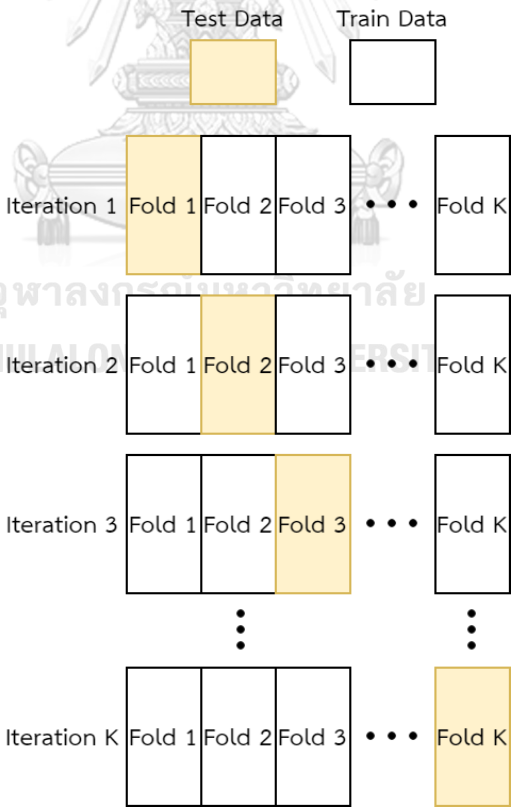
การวัดค่าความคล้ายคลึงของโคไซน์สามารถคำนวณโดยใช้สูตรต่อไปนี้

$$\text{Cosine}(\bar{v}, \bar{w}) = \frac{\bar{v} \cdot \bar{w}}{|\bar{v}||\bar{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \quad \text{- (4)}$$

ซึ่งค่าความคล้ายนี้จะนำไปใช้ในการแนะนำบทความที่เหมาะสมสำหรับการแนะนำมากที่สุด โดยการจัดอันดับค่าคะแนนของความคล้ายระหว่างบทความกับวารสารเป้าหมาย

## 2.5 การประเมินความถูกต้องของแบบจำลองด้วย การตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ (K-fold Cross Validation)

การตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ (K-fold Cross-Validation) [9] เป็นการตรวจสอบข้อมูลโดยการแบ่งข้อมูลออกเป็น K ส่วน โดยแต่ละส่วนมีจำนวนข้อมูลที่เท่ากัน จากนั้นข้อมูลหนึ่งส่วนจะใช้ในการทดสอบประสิทธิภาพของแบบจำลอง ซึ่งจะทำการทดสอบวนซ้ำจนกว่าจะครบจำนวน K ที่แบ่งไว้ แล้วจึงนำค่าความแม่นยำมาหาค่าเฉลี่ย จึงได้เป็นค่าความแม่นยำที่ได้จากการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ โดยการประเมินความถูกต้องของแบบจำลองด้วยการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ มีความเชื่อถือได้มากกว่าการทดลองเพียงรอบเดียว




*รูปที่ 5 แสดงการประเมินความถูกต้องของแบบจำลองด้วยการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ*

## 2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีความเกี่ยวข้องกับงานวิจัยหัวข้อต่าง ๆ ซึ่งแสดงถึงวิธีการที่น่านำมาต่อยอด รวมถึงปัญหาข้อจำกัดต่าง ๆ โดยจะแบ่งงานวิจัยที่เกี่ยวข้องเป็นสองกลุ่มคือกลุ่มที่เกี่ยวข้องกับระบบแนะนำ และกลุ่มที่เกี่ยวข้องกับระบบแนะนำบทความและวารสารวิชาการ

### 2.6.1 งานวิจัยที่เกี่ยวข้องกับระบบแนะนำ

งานวิจัยในกลุ่มนี้จะเกี่ยวข้องกับระบบแนะนำ ซึ่งไม่ได้เกี่ยวข้องกับระบบแนะนำบทความและวารสารวิชาการโดยตรง ซึ่งผู้วิจัยนำมาต่อยอดรวมกับงานวิจัยนี้ โดยงานวิจัยที่เกี่ยวข้องมีดังนี้

*2.6.1.1 Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification*

งานวิจัยของ Jan Deriu [10] ทำแบบจำลองวิเคราะห์อารมณ์ของข้อความ (Sentiment Analysis) โดยนำข้อมูลจากหลากหลายภาษามาใช้ในแบบจำลอง โดยเลือกใช้ภาษาอังกฤษ, ภาษาฝรั่งเศส, ภาษาเยอรมัน และภาษาอิตาลี โดยทำแบบจำลองเปรียบเทียบกันในหลายรูปแบบ ได้แก่ Random forest, Single Lauguage CNN, Multi Language CNN, Fully Multi Language CNN, SemEval benchmark และ การแปลภาษา (Translate) โดยนำขั้นตอนวิธีต่าง ๆ มาใช้กับข้อมูล Twitter กว่า 300 ล้านข้อความ ซึ่งเมื่อนำมาทำการทดลองแล้วพบว่า Single Language CNN หรือใช้ข้อมูลจากภาษานั้น ๆ ภาษาเดียวมาทำแบบจำลองมีความแม่นยำที่สุด ถึงแม้ว่าการใช้ข้อมูลจากภาษาเดียวทำให้เกิดความแม่นยำกว่า แต่การนำภาษาหลากหลายภาษามาทำแบบจำลองทำให้แบบจำลองสามารถรองรับข้อความจากหลายภาษาได้มากขึ้น



*2.6.1.2 Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data*

งานวิจัยของ Alexandra Balahur [11] ทำแบบจำลองวิเคราะห์อารมณ์ของข้อความ (Sentiment Analysis) โดยนำข้อมูล Twitter จากหลายภาษามาใช้ในการปรับปรุงแบบจำลอง ซึ่งใช้ข้อมูลจากภาษาอังกฤษ, ภาษาอิตาลี, ภาษาฝรั่งเศส, ภาษาเยอรมัน และภาษาสเปน โดยนำข้อมูลภาษาต่าง ๆ มาทำการแปลภาษาให้เป็นภาษาต่าง ๆ และทำการเปรียบเทียบความแม่นยำของแบบจำลอง ในขั้นตอนการทำงานได้มีการตัดคำที่ไม่สำคัญและเครื่องหมายวรรคตอนต่าง ๆ เพื่อทำความสะอาดข้อมูลเบื้องต้น จากนั้นทำแบบจำลองสำหรับภาษาอังกฤษ และในลำดับถัดไปได้ใช้ Google Machine Translate System หรือระบบแปลภาษาของ Google มาทำการแปลข้อมูลจาก

ภาษาอังกฤษเป็นภาษาอื่น ๆ และเมื่อทำการทดลองพบว่าแบบจำลองวิเคราะห์อารมณ์ของข้อความสามารถทำงานได้ดีในแต่ละภาษาที่ได้ทำการแปลไป จากงานวิจัยของ Alexandra Balahur ที่แสดงให้เห็นว่าการนำหลายภาษามาใช้สามารถทำให้แบบจำลองทำงานได้อย่างมีประสิทธิภาพยิ่งขึ้น

*2.6.1.3 Movie Recommendation System using Cosine Similarity and KNN*

งานวิจัยของ Ramni Harbir Singh [12] สร้างแบบจำลองของระบบแนะนำภาพยนตร์โดยดูข้อมูลส่วนของประเภทของภาพยนตร์และความนิยมของภาพยนตร์มาสร้างแบบจำลองจากเนื้อหา (Content Based) โดยในการทำแบบจำลองนั้นได้ใช้ Cosine Similarity ในการหาความคล้ายระหว่างบทความ ร่วมกับการใช้ K-Nearest Neighbors (KNN) เพื่อแนะนำภาพยนตร์ที่เกี่ยวข้องมากที่สุดให้กับผู้ใช้

### 2.6.2 งานวิจัยที่เกี่ยวข้องกับระบบแนะนำบทความและวารสารวิชาการ

งานวิจัยในกลุ่มนี้จะเกี่ยวข้องกับระบบแนะนำบทความทางวิชาการหรือระบบแนะนำวารสารวิชาการ ซึ่งได้แสดงความแตกต่างระหว่างงานวิจัยไว้ในตารางที่ 1 โดยงานวิจัยที่เกี่ยวข้องมีดังนี้

*2.6.2.1 A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems*

งานวิจัยของ Maake Benard Magara [13] ทำแบบจำลองสำหรับการแนะนำบทความทางด้านวิชาการจากการนำความถี่ของคำในข้อความทั้งหมดบทความภาษาอังกฤษมา แล้วเพื่อลดคำที่ไม่สำคัญออก จึงทำการคำนวณค่าความสำคัญด้วย TF-IDF และจึงนำเอกสารไปวัดค่าความคล้ายในงานวิจัยนี้นำตัววัดค่าความคล้ายหลายตัวมาใช้ในการเปรียบเทียบได้แก่ Cosine Similarity, Euclidean distance, Jaccard coefficient, Pearson correlation coefficient เพื่อนำเอาตัววัดค่าความคล้ายที่เหมาะสมมาใช้

ในการทดลองขั้นต้นได้ใช้ข้อมูลบทความทางด้านปัญญาประดิษฐ์ (Artificial Intelligent : AI) จำนวน 220 บทความมาทำการทดลองเพื่อหาค่า TF-IDF ประกอบกับ Cosine Similarity และในการประเมินผล (Evaluation) ได้ใช้ข้อมูลบทความจากคณะวิทยาศาสตร์คอมพิวเตอร์มหาวิทยาลัย Ghent University มาแบ่งข้อมูลเป็น 70% สำหรับการทำแบบจำลองและอีก 30% สำหรับการทดสอบ

โดยเป้าหมายของการทดลองคือนำขั้นตอนวิธีสามแบบที่จะมาพัฒนาระบบแนะนำบทความทางด้านวิชาการ ได้แก่ Random Forest, Recursive Partitioning และ Boosted tree นำมาเปรียบเทียบความแม่นยำและประสิทธิภาพ ซึ่งจากผลการทดลองงานวิจัยนี้เลือกใช้ขั้นตอนวิธี Recursive Partitioning ในการพัฒนาแบบจำลองเนื่องจากมีค่าความแม่นยำที่สูง 80.73% และใช้เวลาในการประมวลผลที่ 2.354628 วินาที ในขณะที่ Boosted tree มีค่าความแม่นยำที่สูงกว่าอยู่ที่ 83.2% แต่ใช้เวลามากถึง 41.35908 วินาทีในการประมวลผล และ Random Forest มีค่าความแม่นยำที่ต่ำกว่าอยู่ที่ 80.38% และใช้เวลามากถึง 39.83342 วินาทีในการประมวลผล

ข้อแตกต่างของงานวิจัยของ Maake Benard Magara กับงานวิจัยนี้คือ งานวิจัยของ Maake Benard Magara ใช้ข้อมูลทั้งบทความ ไม่ได้เลือกมาเฉพาะบทคัดย่อหรือคำสำคัญ เนื่องจากงานวิจัยนี้เลือกมาเฉพาะบางส่วนเพราะว่าข้อมูลวารสารทางด้านวิชาการที่มีส่วนใหญ่มีเฉพาะส่วนของบทคัดย่อและคำสำคัญเท่านั้นที่มีทั้งภาษาไทยและภาษาอังกฤษ และข้อมูลวารสารที่ใช้ในงานวิจัยของ Maake Benard Magara ยังเป็นกลุ่มวารสารจากสาขาวิทยาศาสตร์คอมพิวเตอร์เท่านั้น นอกจากนั้นยังเป็นบทความภาษาอังกฤษทั้งหมด ทั้งนี้งานวิจัยของ Maake Benard Magara ยังได้นำส่วนของขั้นตอนวิธีมาพัฒนาระบบแนะนำให้ดียิ่งขึ้น

*2.6.2.2 Personalized Academic Research Paper Recommendation System*

งานวิจัยของ Joonseok Lee [14] ทำแบบจำลองแนะนำบทความทางด้านวิชาการให้กับผู้ที่ต้องการอ่านบทความ โดยใช้ข้อมูลจากบทความที่เจ้าของบทความเขียน โดยทำ Web Crawler ดึงข้อมูลจากสอง เว็บไซต์รวบรวมบทความทางด้านวิชาการ ได้แก่ IEEE Xplore และ ACM Digital Library นำส่วนของ ชื่อเรื่อง ,คำสำคัญ และบทคัดย่อมาแทนกลุ่มคำที่จะแสดงถึงบทความนั้น ๆ

ซึ่งในการทำความสะอาดข้อมูลนั้น งานวิจัยนี้เลือกจะตัดคำเพื่อให้อยู่ในรูปแบบเดียวกันโดยตัดจาก การตัดตัวอักษรที่ลงท้าย “-ed”, “-ly” และ “-ing” ออกจากคำ และเลือกตัดคำไม่สำคัญที่ทางผู้วิจัยเลือกตัดอีก 140 คำ โดยในการสร้างแบบจำลองนั้น ทาง Joonseok Lee ได้กำหนดลักษณะเฉพาะของผู้ใช้ (User Metric) ขึ้นมา โดยกำหนดโดยตั้งหลักการว่าเจ้าของบทความจะชอบบทความที่เจ้าของบทความเขียน และใส่เป็นระดับความชอบ ซึ่งปัญหาและข้อจำกัดของการกำหนดแบบนี้คือจะไม่มีข้อมูลที่ไม่ชอบบทความเป้าหมาย นอกจากนั้นยังมีการคำนวณความคล้ายระหว่างทุกบทความกับบทความของผู้ใช้ ในลักษณะของ k-Nearest Neighbors (kNN) โดยในสถานการณ์จริงบางงานวิจัยมีงานวิจัยมากกว่าหนึ่ง จึงทำการจัดกลุ่ม (Clustering) กลุ่มของงานวิจัยในลักษณะแบบเดียวกับ K-Means แล้วจึงคำนวณคะแนนจากระยะห่างระหว่างงานวิจัยเป้าหมาย และจุดศูนย์กลางของ centroid

ในการประเมินผลนั้น ได้ใช้งานวิจัยจากสามกลุ่มงานได้แก่ Machine learning, Database,

และ Human-Computer Interaction โดยทำการแนะนำ 10 บทความจากบทความทั้งหมด 10,386 บทความให้กับแต่ละกลุ่มงาน ซึ่งระบบแนะนำได้แนะนำกลุ่มงานที่ถูกต้องคิดเป็น 89 % นอกจากนั้นยังให้กลุ่มผู้ทดลองจากสามกลุ่มงาน ได้ให้คะแนนบทความที่แนะนำว่ามีความเกี่ยวข้องกับงานที่เคยเขียนหรือไม่ พบว่ากลุ่มผู้ทดลองพึงพอใจกับผลลัพธ์ของการแนะนำและให้คะแนนมากกว่า 5 ทั้งหมด

ข้อแตกต่างของงานวิจัยของ Joonseok Lee กับงานวิจัยนี้คือ งานวิจัยของ Joonseok Lee ใช้ข้อมูลบทความภาษาอังกฤษและเลือกลดรูปคำโดยไม่ได้ยึดตามหลักพจนานุกรม รวมถึงวิธีการทำค่อนข้างแตกต่างกันเนื่องจากใช้คะแนนความชอบของเจ้าของคนเขียนบทความอิงตามงานบทความที่เคยเขียนไป รวมถึงมีการจัดกลุ่ม (Clustering) ด้วย K-Means นำมาสร้างระบบแนะนำ นอกจากนั้นงานวิจัยของ Joonseok Lee ยังเล็งที่จะมองในมุมมองของการเลือกบทความ แต่งานวิจัยนี้แตกต่างที่จะมองในมุมมองการเลือกวารสารที่เกี่ยวข้อง

*2.6.2.3 Scholarly Paper Recommendation via User's Recent Research Interests*

งานวิจัยของ Kazunari Sugiyama [15] ทำแบบจำลองสำหรับการแนะนำบทความทางด้านวิชาการ โดยใช้ข้อมูลบทความจากในอดีตของนักวิจัยแต่ละคน ประกอบกับ บทความที่อ้างถึงบทความเป้าหมาย และ บทความที่ถูกอ้างอิงโดยบทความเป้าหมาย เพื่อมาประกอบในการทำแบบจำลอง

โดยขั้นแรกของการทำคือสร้างลักษณะของผู้ใช้ (User Profile) โดยแบ่งบทความออกเป็นสองประเภท คือ บทความนักวิจัยอ่อนอาวุโส (Junior) และ บทความนักวิจัยอาวุโส (Senior) ซึ่งนักวิจัยอ่อนอาวุโสจะมีเพียงบทความเดียว ไม่มีข้อมูลบทความที่เคยตีพิมพ์ก่อนหน้า ในขณะที่นักวิจัยอาวุโส จะมีข้อมูลบทความที่เคยถูกตีพิมพ์ก่อนหน้าหลายบทความ ดังนั้นในลักษณะของผู้ใช้ของนักวิจัยอ่อนอาวุโสจะใช้เฉพาะข้อมูลบทความที่ถูกอ้างอิงโดยบทความเป้าหมาย ในขณะที่ลักษณะของผู้ใช้ของนักวิจัยอาวุโสจะใช้ข้อมูลบทความที่เคยถูกตีพิมพ์ก่อนหน้าของนักวิจัยนั้น และบทความโดยคำนวณจากความถี่ของคำ (Term Frequency) แทนการใช้เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) และคำนวณค่าความคล้ายด้วย Cosine Similarity นอกจากนั้นยังมีการปรับค่าน้ำหนักตามปีที่ออกตีพิมพ์ของทั้งบทความที่อ้างอิงและบทความเป้าหมาย โดยหากยิ่งปีที่ตีพิมพ์ของบทความอ้างอิงใกล้เคียงกับบทความเป้าหมาย ค่าน้ำหนักของความคล้ายจะถูกปรับให้สูงยิ่งขึ้นตามไปด้วย จากนั้นแนะนำ

บทความโดยจัดอันดับคะแนนออกมาเป็นกลุ่มอันดับสูงสุด n อันดับ แล้วแนะนำบทความที่มีค่าคะแนนที่สูงที่สุด n อันดับนั้น

ในการทดลองใช้ข้อมูลจากนักวิจัยอ่อนอาวุโส 15 คน และนักวิจัยอาวุโส 13 คนที่ได้ตีพิมพ์ใน DBLP นำมาเป็นบทความอ้างอิง และนำบทความเป้าหมายจาก ACL Anthology Reference Corpus จำนวน 597 บทความแล้วให้นักวิจัยเลือกบทความที่เกี่ยวข้องกับความสนใจของนักวิจัย แล้วทำการทดลองเปรียบเทียบผลลัพธ์ระหว่างการแนะนำบทความทั้งแบบ ใช้แค่บทความที่เคยตีพิมพ์ ใช้บทความที่อ้างอิงถึงและถูกอ้างอิง ซึ่งผลลัพธ์จากการทดลองทำให้เห็นว่าเมื่อนำข้อมูลบทความที่อ้างอิงและถูกอ้างอิงมาช่วยในการแนะนำให้ผลลัพธ์ที่แม่นยำกว่านำข้อมูลจากบทความที่เคยตีพิมพ์อย่างเดียวมาทำการแนะนำ

ข้อแตกต่างของงานวิจัยของ Kazunari Sugiyama กับงานวิจัยนี้คือ งานวิจัยของ Kazunari Sugiyama เน้นใช้ข้อมูลบทความอ้างอิงมาสร้างลักษณะของผู้ใช้ และในการหาความสำคัญของคำนั้นเลือกจะใช้ความถี่ของคำ (Term Frequency) แทน เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) ทั้งนี้งานวิจัยของ Kazunari Sugiyama ยังได้นำปีที่ตีพิมพ์ของบทความมาปรับค่าน้ำหนักของคะแนนมาพัฒนาระบบแนะนำอีกด้วย

*2.6.2.4 Journal Recommendation System Using Content-Based Filtering*

งานวิจัยของ Sonal Jain [16] ได้ทำแบบจำลองสำหรับแนะนำวารสารทางด้านวิชาการ โดยใช้ข้อมูลในส่วนของชื่อบทความ เนื้อหาในส่วนบทคัดย่อและคำสำคัญ ซึ่งนำข้อมูลมาจากสำนักพิมพ์ที่มีชื่อเสียง เช่น Elsevier, Springer, IEEE, ACM, และ Inderscience

ขั้นตอนการทำงานของงานวิจัยนี้เริ่มจากเก็บข้อมูลนำเข้าจากผู้ใช้งานผ่าน Web Portal โดยให้ผู้ใช้กรอกข้อมูลในส่วนของชื่อบทความ บทคัดย่อและคำสำคัญของบทความ จากนั้นจึงนำข้อมูลมาประมวลผลเบื้องต้นด้วยการทำความสะอาดข้อมูลตัดคำหยุดและเครื่องหมายวรรคตอน ขั้นตอนต่อมาได้ทำการคำนวณค่าความสำคัญด้วย TF-IDF โดยใช้ TfidfVectorizer จากไลบรารี่ของ sklearn แล้วแปลงชุดคำและความถี่เป็นส่วนประกอบเชิงเส้นของคำโดยการใช้ Latent Semantic Analysis (LSA) ด้วยการคำนวณค่า eigen values โดยใช้ Single-Value Decomposition (SVD) แล้วจากนั้นหาค่าความคล้ายระหว่างวารสารทางด้านวิชาการกับข้อมูลที่ผู้ใช้กรอกเข้ามาด้วย Euclidean distance แล้วจึงแนะนำ 5 วารสารที่มีระยะห่าง Euclidean distance น้อยที่สุด

ข้อแตกต่างของงานวิจัยของ Sonal Jain กับงานวิจัยนี้คือ ในงานวิจัยของ Sonal Jain ใช้ Latent Semantic Analysis ด้วยการใช้ Single-Value Decomposition คำนวณค่า eigen values แล้วจึงไปหาความคล้ายด้วย Euclidean distance ในขณะที่งานวิจัยนี้ใช้ Cosine

Similarity ในการคำนวณค่าความคล้าย และนอกนจากนั้นงานวิจัยนี้ได้มีนำข้อมูลไปทดลองเพื่อประเมินผลความแม่นยำของแบบจำลอง แต่งานวิจัยของ Sonal Jain ไม่มีการกล่าวถึงการประเมินผลความถูกต้องของแบบจำลอง และอีกข้อแตกต่างคืองานวิจัยนี้ใช้ข้อมูลทั้งจากภาษาไทยและภาษาอังกฤษ

*2.6.2.5 ความแตกต่างระหว่างงานวิจัยที่เกี่ยวข้องกับระบบแนะนำบทความ/วารสารวิชาการและงานวิจัยนี้*

ในส่วนของข้อมูลที่นำมาใช้ งานวิจัยนี้จะแตกต่างจากงานอื่นที่ใช้ข้อมูลสองภาษาทั้งภาษาไทยและภาษาอังกฤษ และใช้ข้อมูลแค่บางส่วนของบทความได้แก่ส่วนของคำสำคัญและบทคัดย่อเท่านั้น ในส่วนของการเตรียมข้อมูลจะค่อนข้างคล้ายงานวิจัยที่ [14] และงานวิจัย [16] ที่จะมีการตัดคำ ทำการลดรูปคำ และนำคำหยุดออก ซึ่งจะทำทั้งภาษาไทยและภาษาอังกฤษ ในการทำแบบจำลองงานวิจัยนี้หาค่า TF-IDF แล้วคำนวณค่าความคล้าย Cosine Similarity ระหว่างบทความและวารสารทั้งภาษาไทยและภาษาอังกฤษคล้ายกับงานวิจัยที่ [13] แต่แตกต่างที่จะหาความคล้ายระหว่างวารสารกับบทความแทนที่จะความคล้ายระหว่างบทความกับบทความคล้ายกับงานวิจัยที่ [16] จากนั้นจัดอันดับคะแนนออกมาเป็นกลุ่มอันดับสูงสุด 5 อันดับ แล้วแนะนำวารสารที่มีค่าคะแนนที่สูงที่สุด 5 อันดับนั้น ซึ่งคล้ายกับงานวิจัยที่ [15] และงานวิจัย [16]

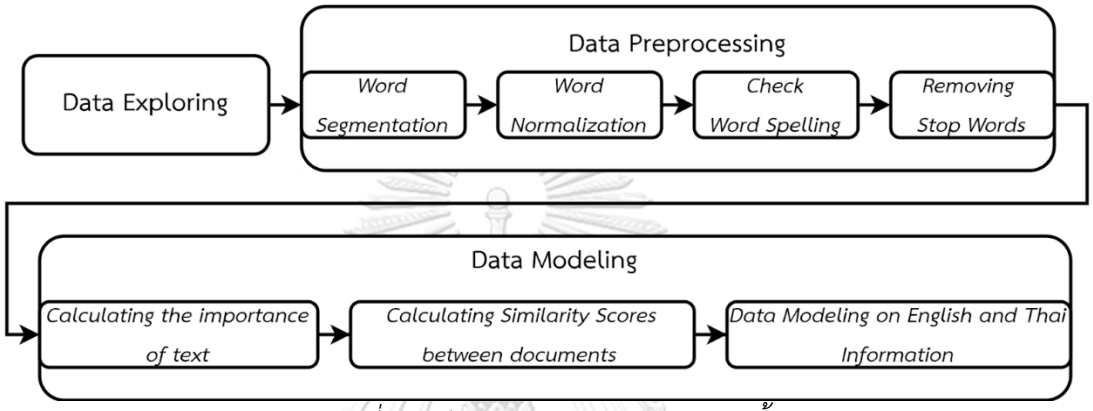
*ตาราง 1 ความแตกต่างระหว่างงานวิจัยที่เกี่ยวข้องกับระบบแนะนำบทความ/วารสารวิชาการและงานวิจัยนี้*

| งานวิจัย | A Comparative Analysis of Text Similarity Measures and Algorithms in Research Paper Recommender Systems | Personalized Academic Research Paper Recommendation System | Scholarly Paper Recommendation via User's Recent Research Interests | Journal Recommendation System Using Content-Based Filtering | Journal Recommendation System based on Multi-Language (งานวิจัยนี้) |
|---|---|---|---|---|---|
| **ข้อมูลที่ใช้** | คำในข้อความทั้งหมดในบทความภาษาอังกฤษ | ส่วนของชื่อเรื่อง,คำสำคัญและบทคัดย่อจากบทความภาษาอังกฤษ | ข้อมูลบทความจากในอดีตของนักวิจัยแต่ละคน และข้อมูลบทความที่อ้างอิงและถูกอ้างอิงจากบทความ | ส่วนของชื่อเรื่อง,คำสำคัญและบทคัดย่อจากวารสารภาษาอังกฤษ | ส่วนของคำสำคัญและบทคัดย่อจากวารสารภาษาอังกฤษและวารสารภาษาไทย |
| **การทำความสะอาดข้อมูลและการเตรียมข้อมูลขั้นต้น** | ไม่มีกล่าวถึง | การตัดตัวอักษรที่ลงท้าย "-ed", "-ly" และ "-ing" ออกจากคำ และเลือกตัดคำไม่สำคัญที่ทางผู้วิจัยเลือกตัดอีก 140 คำ | แบ่งงานวิจัยออกเป็นบทความนักวิจัยอ่อนอาวุโส (Junior) และบทความนักวิจัยอาวุโส (Senior) | ตัดเครื่องหมายเว้นวรรคตอนและคำหยุด | ตัดคำภาษาอังกฤษและภาษาไทย ทำการลดรูปคำและกำจัดคำหยุด |
| **ลักษณะที่มาใช้ในแบบจำลอง (Feature)** | หาค่า TF-IDF แล้วคำนวณค่าความคล้าย Cosine Similarity ระหว่างบทความ | กำหนดลักษณะเฉพาะของผู้ใช้ ใส่เป็นระดับความชอบและจัดกลุ่มของงานวิจัยในลักษณะแบบเดียวกับ K-Means | หาค่า TF แล้วคำนวณค่าความคล้ายด้วย Cosine Similarity นอกจากนั้นยังมีการปรับค่าน้ำหนักตามปีที่ออกตีพิมพ์ของทั้งบทความที่อ้างอิงและบทความเป้าหมาย | หาค่า TF-IDF ใช้ Latent Semantic Analysis ด้วยการใช้ Single-Value Decomposition คำนวณค่า eigen values แล้วจึงไปหาความคล้ายด้วย Euclidean distance | หาค่า TF-IDF แล้วคำนวณค่าความคล้าย Cosine Similarity ระหว่างบทความและวารสารทั้งภาษาไทยและภาษาอังกฤษ |

| | | | | | |
|---|---|---|---|---|---|
| **แบบจำลองที่นำเสนอและการทดลอง** | นำค่าคะแนนความคล้ายไปใช้ขั้นตอนวิธี Random Forest, Recursive Partitioning และ Boosted tree นำมาเปรียบเทียบความแม่นยำและประสิทธิภาพ | คำนวณคะแนนจากระยะห่างระหว่างงานวิจัยเป้าหมาย และจุดศูนย์กลางของ centroid แล้วทำการแนะนำ 10 บทความให้กับแต่ละกลุ่มงาน | จัดอันดับคะแนนออกมาเป็นกลุ่มอันดับสูงสุด n อันดับ แล้วแนะนำบทความที่มีค่าคะแนนที่สูงที่สุด n อันดับนั้น | จัดอันดับคะแนนออกมาเป็นกลุ่มอันดับสูงสุด 5 อันดับ แล้วแนะนำวารสารที่มีค่าระยะห่างที่น้อยที่สุด 5 อันดับนั้น | จัดอันดับคะแนนออกมาเป็นกลุ่มอันดับสูงสุด 5 อันดับ แล้วแนะนำวารสารที่มีค่าคะแนนที่สูงที่สุด 5 อันดับนั้น |
| **การประเมินผลความถูกต้องของแบบจำลองและผลของการประเมินผล** | นำขั้นตอนวิธีสามแบบนำมาเปรียบเทียบความแม่นยำ โดยขั้นตอนวิธี Recursive Partitioning มีค่าความแม่นยำที่สูง 80.73% และใช้เวลาในการประมวลผลที่ 2.354628 วินาที | ทำการแนะนำ 10 บทความจากบทความทั้งหมด 10,386 บทความให้กับแต่ละกลุ่มงานจากสามกลุ่มงาน ซึ่งระบบแนะนำได้แนะนำกลุ่มงานที่ถูกต้องคิดเป็น 89 % | ในการทดลองใช้ข้อมูลจากนักวิจัยอ่อนอาวุโส 15 คน และนักวิจัยอาวุโส 13 คน โดยได้ค่าความถูกต้องสูงสุดอยู่ที่ 73.9% | ไม่มีการกล่าวถึงการประเมินความถูกต้องของแบบจำลอง แต่มีการกล่าวถึงวารสารที่มีค่าระยะห่างน้อยที่สุด 5 อันดับ ที่ 0.319759, 0.384592, 0.390679, 0.403017, 0.505765 | ผู้วิจัยเลือกใช้ K-fold Cross-Validation ประกอบกับค่า Hit Rate ซึ่งค่าความถูกต้องสูงสุดอยู่ที่ 0.87965 |
| **ผลลัพธ์ของระบบแนะนำ** | แนะนำบทความที่เกี่ยวข้องกับบทความของผู้เขียน | แนะนำบทความที่เกี่ยวข้องกับบทความของผู้เขียน | แนะนำบทความที่เกี่ยวข้องกับบทความของผู้เขียน | แนะนำวารสารที่เกี่ยวข้องกับบทความ | แนะนำวารสารที่เกี่ยวข้องกับบทความ |

# บทที่ 3
# วิธีดำเนินการ

ในงานวิจัยนี้มีวิธีการดำเนินการหลัก ดังนี้ ศึกษาข้อมูลที่นำมาใช้ (Data Explore), เตรียมข้อมูล (Data Preprocessing) และทำแบบจำลองสำหรับระบบแนะนำ (Data Modeling) ซึ่งจะเป็นไปตามรูปที่ 5




*รูปที่ 6 วิธีการดำเนินการของงานวิจัยนี้*

## 3.1 ศึกษาข้อมูลที่นำมาใช้ (Data Exploring)

ข้อมูลที่นำมาใช้วิเคราะห์ในงานวิจัยนี้ เป็นข้อมูลวารสารทางด้านวิชาการที่เผยแพร่ใน Thai Journals Online (ThaiJO) ซึ่งเป็นระบบฐานข้อมูลที่รวบรวมแหล่งวารสารวิชาการที่ผลิตในประเทศไทยทุกสาขาวิชา โดยมีทั้งวารสารที่เป็นภาษาไทยและ ภาษาอังกฤษ ซึ่งบางวารสารมีข้อมูลเฉพาะภาษาไทย และบางวารสารมีข้อมูลเฉพาะภาษาอังกฤษ

### 3.1.1 การคัดเลือกวารสารที่นำมาใช้ในงานวิจัย

วารสารที่นำมาใช้วิเคราะห์ในงานวิจัยนี้มีจำนวนบทความ 3282 บทความ จากจำนวนวารสาร 17 วารสาร ซึ่งเป็นวารสารที่เกี่ยวกับวิศวกรรมศาสตร์ และเทคโนโลยี เราคัดเลือกวารสารโดยใช้เกณฑ์คือใช้วารสารเฉพาะในกลุ่ม TCI กลุ่มที่ 1 และ กลุ่มที่ 2 เท่านั้น ซึ่งจากบทความทั้งหมด 3282 บทความ สามารถแบ่งเป็นบทความจำนวน 1283 ที่มีทั้งข้อมูลภาษาอังกฤษและภาษาไทย และบทความจำนวน 2049 ที่มีเฉพาะข้อมูลภาษาอังกฤษ

*ตารางที่ 2 แสดงจำนวนบทความและกลุ่ม TCI ของวารสารที่นำมาใช้ในงานวิจัย*

| **ชื่อวารสาร (ภาษาอังกฤษ)** | **TCI กลุ่มที่** | **จำนวนบทความ** |
|---|---|---|
| The Journal of Industrial Technology | 1 | 343 |
| Information Technology Journal | 2 | 343 |
| Environment and Natural Resources Journal | 1 | 308 |
| ECTI Transactions on Computer and Information Technology | 1 | 289 |
| UBU Engineering Journal | 1 | 252 |
| Naresuan University Engineering Journal | 1 | 229 |
| Journal of Industrial Technology Ubon Ratchathani Rajabhat University | 1 | 214 |
| Ladkrabang Engineering Journal | 1 | 202 |
| Journal of Renewable Energy and Smart Grid Technology | 1 | 193 |
| TNI Journal of Engineering and Technology | 1 | 176 |
| Thai Environmental Engineering Journal | 2 | 133 |
| Journal of Project in Computer Science and Information Technology | 2 | 125 |
| Thai Society of Agricultural Engineering Journal | 2 | 125 |
| Journal of Energy and Environment Technology of Graduate School Siam Technology College | 2 | 107 |
| Thai Industrial Engineering Network Journal | 2 | 104 |
| Farm Engineering and Automation Technology Journal | 2 | 87 |
| Journal of Applied Statistics and Information Technology | 2 | 52 |

3.1.2 การศึกษาข้อมูลวารสารโดยใช้ Word Cloud

โดยในช่วงการศึกษาข้อมูลได้ทำการสร้าง Word Cloud จากคำที่มีในเอกสาร เพื่อที่จะศึกษาลักษณะของข้อความที่ปรากฏเอกสารในเบื้องต้นในลักษณะความถี่ของคำ [17] ซึ่งจากการศึกษาข้อมูลด้วย Word Cloud ทำให้เห็นว่ามีคำบางคำที่สะกดไม่ถูกต้อง บางคำมีรากศัพท์ หรือบางคำที่เป็นเพียงคำบุพบทหรือคำหยุดที่ไม่สื่อความหมายมากนัก ทำให้ผู้วิจัยจำเป็นต้องทำการเตรียมข้อมูลในลำดับถัดไป


*รูปที่ 7 ตัวอย่าง Word Cloud ของเอกสารบางส่วนในการศึกษาข้อมูลเบื้องต้น*



## 3.2 การเตรียมข้อมูล (Data Preprocessing)

จากข้อมูลที่ได้มา ทางผู้วิจัยได้ทำการเลือกข้อมูลที่จะใช้ในแต่ละบทความได้แก่ ข้อมูลเลขบทความ (ID ของบทความ), ข้อมูลเลขวารสาร (ID ของวารสาร), ข้อมูลบทคัดย่อ (Abstract), ข้อมูลคำสำคัญ (Keyword) โดยข้อมูลภาษาไทยและภาษาอังกฤษในแต่ละบทความจะแบ่งออกมาประมวลผลในแต่ละภาษา โดยขั้นตอนในการเตรียมข้อมูลทั้งสองภาษามีวิธีการแตกต่างกันเล็กน้อย

3.2.1 การตัดคำ (Word Segmentation)

เป็นขั้นตอนการตัดคำออกจากประโยค ซึ่งในงานวิจัยนี้ ได้มีกระบวนการดังนี้ ในข้อมูลภาษาอังกฤษ การตัดคำออกจากประโยคสามารถทำได้ด้วยการแบ่งคำโดยช่องว่าง เพราะรูปประโยคของภาษาอังกฤษนั้นต้องทำการเว้นวรรคทุกคำ ส่วนในข้อมูลภาษาไทยจะไม่สามารถตัดคำด้วย

ช่องว่างได้เช่นเดียวกันกับภาษาอังกฤษ ทำให้การตัดคำในข้อมูลภาษาไทยมีความซับซ้อนมากกว่า ดังนั้นทางผู้วิจัยจึงใช้ PyThaiNLP ซึ่งเป็นไลบารีสำหรับการประมวลผลภาษาธรรมชาติในภาษาไทย

3.2.2 การลดรูปคำ (Normalization)

ทั้งข้อมูลภาษาไทยและภาษาอังกฤษ จะมีบางคำที่มีความหมายเหมือนหรือใกล้เคียงกัน มีรากศัพท์เหมือนกัน แต่ใช้คำแตกต่างกันตามรูปประโยคและบริบท โดยในภาษาอังกฤษจะมีการผันคำกริยาให้ถูกต้องตามหลักไวยากรณ์ ยกตัวอย่างเช่น การใช้ ran แทน run, ate แทน eat, deleted แทน delete เมื่อประโยคเหล่านี้เล่าถึงเหตุการณ์ในอดีต หรือการใช้คำที่มีรากศัพท์เดียวกันแต่ใช้เป็นคำนาม กริยา หรือคำบุพบทต่าง ๆ เช่น prepare, preparation, prepared เป็นต้น ดังนั้นในขั้นตอนนี้จึงจำเป็นจะต้องทำให้คำศัพท์เหล่านี้กลับมาเป็นรากศัพท์ของคำโดยผู้วิจัยใช้ nltk ซึ่งเป็นไลบารีสำหรับการประมวลผลภาษาธรรมชาติ เพื่อลดรูปของคำภาษาอังกฤษให้เป็นลักษณะเดียวกัน ส่วนในภาษาไทยนั้นทางผู้วิจัยใช้ PyThaiNLP ซึ่งเป็นไลบารีสำหรับการประมวลผลภาษาธรรมชาติในภาษาไทย

3.2.3 การตรวจสอบการสะกดคำผิด

ข้อมูลบางส่วนในบทความมีทั้งคำที่เขียนสะกดได้ถูกต้องแล้ว และมีบางคำที่ยังสะกดคำไม่ถูก รวมถึงมีบางประโยคที่หลังจากถูกตัดคำนั้น แล้วมีคำที่ถูกตัดออกมาไม่ถูกต้องเท่าที่ควรเท่สให้เกิดเป็นคำที่สะกด เพื่อทำให้คำเหล่านี้ที่สะกดไม่ถูกไม่ว่าจากสาเหตุใดก็ตาม มีจำนวนที่ลดน้อยลง ผู้วิจัยจึงทำการใช้คำสั่งในการตรวจสอบการสะกดคำของไลบารีที่นำมาใช้ โดยใช้ nltk ในการตรวจสอบคำภาษาอังกฤษ และใช้ PyThaiNLP ในการตรวจสอบคำภาษาไทย



3.2.4 การกำจัดคำหยุด (Stopword)

คำบุพบทหรือคำเชื่อมประโยคต่าง ๆ เป็นคำที่ต้องมีอยู่ในประโยคแทบทุกบทความ ทำให้ในการจะหาคำสำคัญในแต่ละบทความ ควรจะตัดคำเหล่านี้ออกจากข้อมูล เพื่อลดภาระในการประมวลผล และเพิ่มความถูกต้องในขั้นตอนประมวลผล

ในภาษาอังกฤษมีคำหยุดที่นิยามไว้ในไลบารี nltk แล้ว จึงสามารถนำฟังก์ชั่นจาก nltk เข้ามาใช้ได้เลย ส่วนในข้อมูลภาษาไทย ทางผู้วิจัยนำข้อมูลที่เผยแพร่บน Stopwords ISO [18] ซึ่งรวบรวมคำหยุดในหลากหลายภาษารวมถึงภาษาไทย มาเป็นกลุ่มคำศัพท์ที่จะตัดออกจากข้อมูล

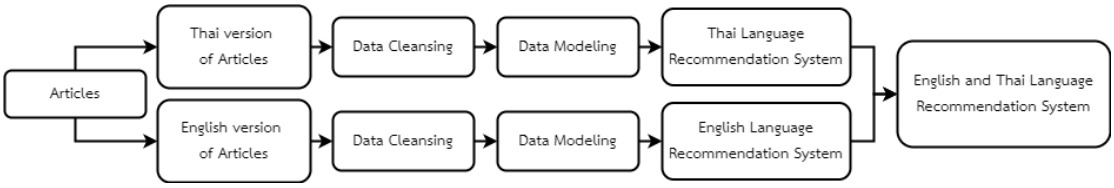
```
from autocorrect import Speller
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
#Init Spell
spell = Speller()
# Init the Wordnet Lemmatizer
lemmatizer = WordNetLemmatizer()

# init stop word
stop_words = set(stopwords.words('english'))

def autoCrrectWord(listWord):
  tokens=[]
  if listWord == None:
    return []
  for word in listWord:
    new_word=""
    for char in word:
      if char.isalpha() and char != " ":
        new_word = new_word+char
    new_word = spell(new_word).lower()
    new_word = lemmatizer.lemmatize(new_word)
    if new_word not in stop_words and new_word!="":
      tokens.append(new_word)
  return tokens
journal_df2['en_keywords_tokens'] = journal_df2.apply(lambda row:autoCrrectWord(row.en_keywords_tokens), axis=1)
journal_df2['en_abstract_tokens'] = journal_df2.apply(lambda row:autoCrrectWord(row.en_abstract_tokens), axis=1)
journal_df2['en_title_tokens'] = journal_df2.apply(lambda row:autoCrrectWord(row.en_title_tokens), axis=1)
```

*รูปที่ 8 ตัวอย่างบางส่วนของกระบวนการเตรียมข้อมูล*

### 3.3 แบบจำลองที่นำเสนอ (Data Modeling)

ในการแนะนำวารสารทางด้านวิชาการที่เหมาะสมกับบทความนั้น ทางผู้วิจัยได้นำแต่ละบทความมาหาความสำคัญของคำ และนำค่าคะแนนมาหาความคล้ายระหว่างแต่ละบทความกับวารสารต่าง ๆ และแนะนำวารสารที่เหมาะสมให้กับบทความ โดยในกระบวนการเหล่านี้ทางผู้วิจัยได้ทดลองประมวลผลข้อมูลภาษาไทยและข้อมูลภาษาอังกฤษแยกจากกัน จากนั้นจึงนำแบบจำลองของข้อมูลภาษาไทยและข้อมูลภาษาอังกฤษมารวมกันเพื่อให้เกิดประสิทธิภาพที่ดียิ่งขึ้น (ดังรูปที่ 3.2)

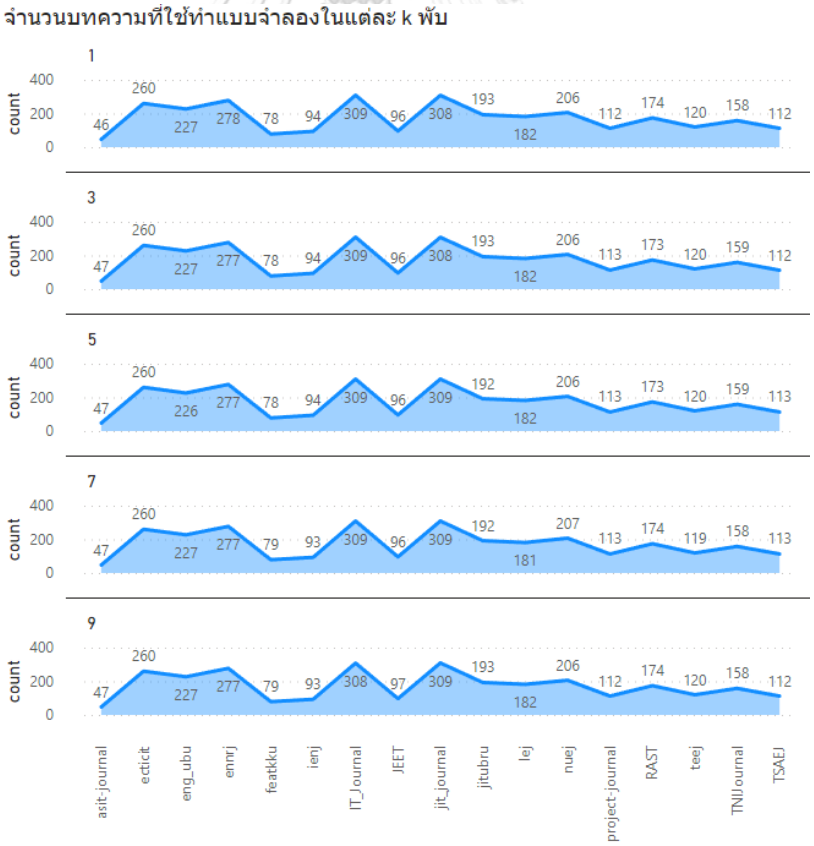
โดยทั้งสองภาษามีขั้นตอนประมวลผลคือคำนวณค่าความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency: TF-IDF) และ คำนวณความคล้ายของแต่ละบทความ (Cosine Similarity)



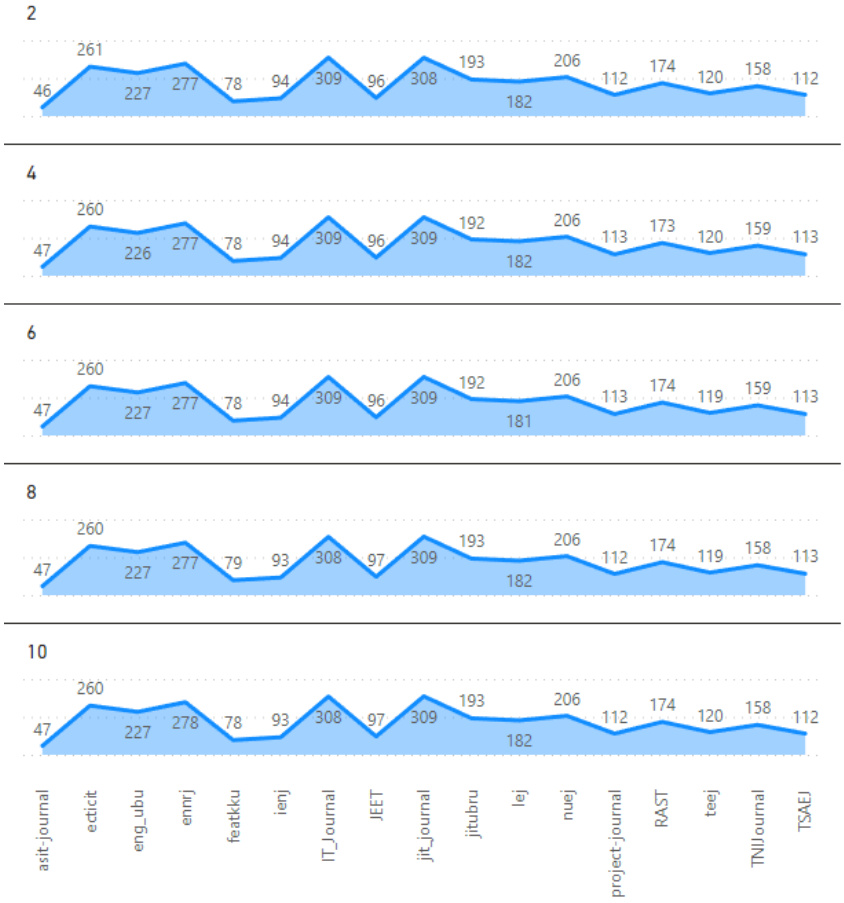

*รูปที่ 9 แสดงวิธีดำเนินการสร้างแบบจำลองสำหรับระบบแนะนำจากภาษาไทยและภาษาอังกฤษ*

### 3.3.1 จัดการกระจายของข้อมูล

ข้อมูลแต่ละวารสารที่ใช้ในแบบจำลองมีการกระจายตัวในข้อมูลในระดับหนึ่ง แต่หากไม่มีการจัดการกระจายของข้อมูล เมื่อนำข้อมูลไปทำแบบจำลองและตรวจสอบด้วยการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับจะทำให้ระบบมีความแม่นยำที่ไม่เที่ยงตรงในแต่ละไขว้ได้ ดังนั้นจึงจะต้องทำการจัดการกระจายของข้อมูล โดยนำข้อมูลวารสารแต่ละฉบับมาแบ่งเป็นจำนวนที่เท่า ๆ กันแล้วนำข้อมูลที่แบ่งไว้ในแต่ละวารสารมารวมกันในจำนวนที่เท่า ๆ กัน ซึ่งผลลัพธ์การกระจายของข้อมูลเป็นดังรูปที่ 10 - 11




*รูปที่ 10 แสดงการกระจายของข้อมูลโดยแสดงจำนวนบทความที่ใช้ทำแบบจำลองในแต่ละ k พับ (หน้าที่ 1)*






*รูปที่ 11 แสดงการกระจายของข้อมูลโดยแสดงจำนวนบทความที่ใช้ทำแบบจำลองในแต่ละ k พับ (หน้าที่ 2)*

### 3.3.2 คำนวณค่าคำที่สำคัญในเอกสาร

บทความจะมีคำสำคัญที่ระบุไว้ในแต่ละบทความ ซึ่งถูกกำหนดโดยผู้เขียนบทความ ในวารสารประเภทเดียวกันมักจะมีคำสำคัญในลักษณะเดียวกันอยู่ แต่เนื่องด้วยคำสำคัญที่ผู้เขียนบทความใส่มาอาจจะยังไม่ถูกต้องหรือครอบคลุมคำสำคัญของบทความต่าง ๆ ได้ ดังนั้นเพื่อจะให้ได้คำสำคัญที่มีค่าความสำคัญที่ดียิ่งขึ้นจึงต้องมาใช้เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร โดยในงานวิจัยนี้นำข้อความในส่วนของบทคัดย่อและคำสำคัญในแต่ละบทความที่อยู่ในวารสารเดียวกันเข้าด้วยกัน แล้วจากนั้นจึงนำไปคำนวณค่าความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร

### 3.3.3 คำนวณความคล้ายของแต่ละบทความ (Cosine Similarity)

เพื่อที่จะหาว่าบทความหนึ่งเหมาะสมกับวารสารไหน ทางผู้วิจัยใช้ข้อสมมติฐานที่ว่า วารสารหนึ่งย่อมประกอบไปด้วยบทความที่มีความคล้ายกัน ทางผู้วิจัยจึงนำบทความที่ต้องการแนะนำวารสารไปคำนวณค่า Cosine Similaritiy ร่วมกับวารสารต่าง ๆ และทำการแนะนำวารสารในลำดับถัดไป จัดอันดับค่าคะแนนของความคล้ายระหว่างบทความกับวารสาร แล้วแนะนำวารสารที่มีค่าคะแนนที่สูงที่สุดออกมา โดยจะแนะนำวารสารมาเป็น 5 อันดับ เนื่องจากว่าบทความหนึ่งนั้นอาจผ่านการคัดเลือกให้ตีพิมพ์ในวารสารที่ใกล้เคียงกันได้มากกว่า 1 วารสาร ยกตัวอย่างเช่น บทความวิศวกรรมศาสตร์ของนักวิจัยกลุ่มหนึ่งที่อยู่ในวารสารวิศวกรรมศาสตร์ของมหาวิทยาลัย A อาจผ่านการคัดเลือกให้ตีพิมพ์ในวารสารวิศวกรรมศาสตร์ของมหาวิทยาลัย B ได้เช่นเดียวกัน โดยเอกลักษณ์ของวารสารนั้นทำการอิงข้อมูลประกอบจากบทความในวารสารนั้น ๆ ซึ่งจะใช้บทความจำนวนหนึ่งในวารสาร โดยจะแบ่งตามหลักการ K-Fold Cross Validation ที่จะนำบทความส่วนหนึ่งมาใช้ในการทำแบบจำลอง และใช้บทความอีกส่วนหนึ่งในการทำการทดสอบแบบจำลอง เพื่อพัฒนาให้วารสารมีเอกลักษณ์ของวารสารได้มีประสิทธิภาพที่เพียงพอ

### 3.3.4 แบบจำลองจากข้อมูลภาษาอังกฤษและภาษาไทย

เพื่อที่จะพัฒนาให้แบบจำลองมีความแม่นยำมากขึ้น รวมถึงเพื่อให้แบบจำลองสามารถทำงานรองรับทั้งภาษาไทยและภาษาอังกฤษได้อย่างมีประสิทธิภาพยิ่งขึ้น [11] เราได้ทำการปรับค่าน้ำหนักของค่าคะแนนความคล้ายโดยใช้ค่าคะแนนจากทั้งสองภาษา โดยหากมีค่าคะแนนจากภาษาใดภาษานึงเป็น 0 เราจะใช้ค่าคะแนนจากอีกภาษาเป็นหลักแทน แต่หากค่าคะแนนในทั้งสองภาษามากกว่า 0 เราจะปรับค่าน้ำหนักของคะแนนโดยอิงจากค่าเฉลี่ยของทั้งสองภาษา

# บทที่ 4

# การทดลอง

ในบทนี้เพื่อที่จะเปรียบเทียบการใช้ข้อมูลต่าง ๆ ในการสร้างแบบจำลอง ทางผู้วิจัยจะอธิบายการตั้งค่าทดลองในงานวิจัย อธิบายตัวชี้วัดการประเมินแบบจำลอง และอธิบายผลลัพธ์การทดลองในแบบต่าง ๆ

## 4.1 การตั้งค่าการทดลอง

ทางผู้วิจัยจะนำแบบจำลองภาษาอังกฤษ แบบจำลองภาษาไทย และแบบจำลองสองภาษาร่วมกันมาทดลองเปรียบเทียบผลลัพธ์ ซึ่งในแบบจำลองแต่ละภาษาจะทำการทดลองทั้งการใช้ข้อมูลคำสำคัญ ข้อมูลบทคัดย่อ และข้อมูลคำสำคัญและบทคัดย่อร่วมกัน

## 4.2 ตัวชี้วัดการประเมินแบบจำลอง

ในการตรวจสอบความถูกต้องของแบบจำลอง ทางผู้วิจัยเลือกใช้การตรวจสอบความสมเหตุสมผลแบบไขว้จำนวนเคพับ (K-fold Cross-Validation) ซึ่งเป็นการตรวจสอบข้อมูลโดยการแบ่งข้อมูลออกเป็น K ส่วน โดยจะทำการทดสอบวนซ้ำจนกว่าจะครบจำนวน K ที่แบ่งไว้ ทั้งนี้ผู้วิจัยเลือกใช้ K เป็น 5 และ K เป็น 10 ในการทดลองนี้ ซึ่งเป็นค่าทางเลือกที่พบเห็นที่สุดระหว่าง K เป็น 5 และ K เป็น 10 [19] เนื่องจากค่าทั้งสองได้ผ่านการทดลองแล้วว่าไม่ได้รับผลกระทบจากอคติมากเกินไป [20] ประกอบกับการเลือกใช้ Hit Rate [21] ในการตรวจสอบแต่ละรอบ ซึ่งเป็นการตรวจสอบโดยหากว่ามีวารสารที่ได้รับการตีพิมพ์จริงอยู่ใน 5 อันดับที่แบบจำลองแนะนำจะนับเป็น 1 Hit แล้วนำค่า Hit ไปหารด้วยจำนวนข้อมูลที่ใช้ทดสอบทั้งหมดจะออกเป็นค่า Hit Rate ซึ่งค่า Hit Rate ยิ่งใกล้เคียง 1 จะถือว่าแบบจำลองยิ่งมีความแม่นยำมากขึ้น ซึ่งทางผู้วิจัยตั้งเป้าก่อนการทดลองให้แบบจำลองมีค่าความถูกต้องมากกว่า 0.8 หรือ จึงถือว่าถูกต้องเพียงพอ เนื่องจากงานวิจัยที่เกี่ยวข้องกับระบบแนะนำบทความ/วารสารวิชาการ [13, 14, 15] มีความแม่นยำ 0.8073, 0.89 และ 0.739 ซึ่งค่าเฉลี่ยคือ 0.8121 จึงตั้งว่าความแม่นยำ 0.8 เป็นเป้าหมายที่สมเหตุสมผล

## 4.3 การทดลองแบบจำลองกับข้อมูลภาษาอังกฤษ

ในการทดลองเมื่อเรานำข้อมูลคำสำคัญเพียงอย่างเดียวมาทำแบบจำลองแล้วแนะนำวารสารออกมาทำการทดลองการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ (5-fold cross-validation) มีบทความที่ได้รับการแนะนำวารสารตรงกับข้อมูลวารสารที่บทความนั้นตีพิมพ์ใน 5

อันดับแรกอยู่ที่ 530.2 บทความจากทั้งหมด 656.4 บทความหรือคิดเป็นค่าความแม่นยำเฉลี่ยอยู่ที่ 0.80774 แต่เมื่อเรานำข้อมูลบทคัดย่อเพียงอย่างเดียวมาทำแบบจำลองได้ค่าความแม่นยำที่ลดลงเหลือ 0.75107 หรือ 493 บทความจากทั้งหมด 656.4 บทความ และเมื่อเรารวมข้อมูลจากทั้งบทคัดย่อและคำสำคัญและทดลองปรับน้ำหนักของคะแนน 3 แบบ โดยแบบแรกให้ค่าคะแนนของคำสำคัญเท่ากับค่าคะแนนของบทคัดย่อ ซึ่งผลการทดลองแบบแรกนั้นได้ค่าความแม่นยำมากกว่าใช้คำสำคัญหรือบทคัดย่ออย่างเดียวอยู่ที่ 0.82785 หรือ 543.8 บทความจาก 656.4 บทความ การทดลองปรับน้ำหนักของคะแนนแบบที่ 2 เมื่อปรับให้ค่าคะแนนของคำสำคัญมากกว่าบทคัดย่อในอัตราส่วน 10 ต่อ 1 ให้ค่าแม่นยำที่ 0.84369 หรือ 553.8 บทความจาก 656.4 บทความ และการทดลองปรับน้ำหนักของคะแนนแบบที่ 3 เมื่อปรับให้ค่าคะแนนของคำสำคัญมากกว่าบทคัดย่อในอัตราส่วน 5 ต่อ 1 ซึ่งให้ผลลัพธ์ที่ดีที่สุดเมื่อเปรียบเทียบการปรับน้ำหนักของคะแนนทั้ง 3 แบบ โดยให้ความสำคัญกับคำสำคัญมากกว่าบทคัดย่อจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นมาที่ 0.84673 หรือ 555.8 บทความ จาก 656.4 บทความ

เมื่อเราทำการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold cross-validation) ผลลัพธ์การทดลองให้ผลลัพธ์ที่ดีกว่าการทดลอง 5 โดยเมื่อใช้ข้อมูลคำสำคัญเพียงอย่างเดียวได้ค่าความแม่นยำเฉลี่ยอยู่ที่ 0.80683 ซึ่งคิดเป็นแนะนำถูกต้องเฉลี่ย 264.8 บทความจากทั้งหมด 328.2 บทความ เช่นเดียวกับการทดลอง 5 พับ นำข้อมูลบทคัดย่อเพียงอย่างเดียวมาทำแบบจำลองได้ค่าความแม่นยำที่ลดลงเหลือ 0.75990 หรือ 249.4 บทความจากทั้งหมด 328.2 บทความ และเมื่อเรารวมข้อมูลจากทั้งบทคัดย่อและคำสำคัญและปรับน้ำหนักของคะแนนจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นเป็นแนะนำถูกต้องเฉลี่ย 278.8 บทความจากทั้งหมด 328.2 บทความ หรือคิดเป็นค่าความแม่นยำเฉลี่ยอยู่ที่ 0.84948



### 4.4 การทดลองแบบจำลองกับข้อมูลภาษาไทย

ในการทดลองกับข้อมูลภาษาไทยได้ค่าความแม่นยำที่ลดลงจากแบบจำลองภาษาอังกฤษ โดยในการทดลอง 5 พับแบบจำลองจากข้อมูลคำสำคัญภาษาไทยเพียงอย่างเดียวมีบทความที่ได้รับการแนะนำวารสารตรงกับข้อมูลวารสารที่บทความนั้นตีพิมพ์ใน 5 อันดับแรกอยู่ที่ 179.4 บทความจากทั้งหมด 240.6 บทความ หรือคิดเป็นค่าความแม่นยำเฉลี่ยอยู่ที่ 0.74562 ในขณะที่เมื่อเรานำข้อมูลบทคัดย่อเพียงอย่างเดียวมาทำแบบจำลองได้ค่าความแม่นยำที่เพิ่มขึ้นเป็น 0.74812 หรือ 180 บทความจากทั้งหมด 240.6 บทความ และเมื่อเรารวมข้อมูลจากทั้งบทคัดย่อและคำสำคัญและปรับน้ำหนักของคะแนนในอัตราส่วน 5 ต่อ 1 เช่นเดียวกับแบบจำลองกับข้อมูลภาษาอังกฤษโดยให้

ความสำคัญกับคำสำคัญมากกว่าบทคัดย่อในลักษณะเดียวกับแบบจำลองภาษาอังกฤษจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นมาที่ 0.78970 หรือ 190 บทความ จาก 240.6 บทความ

ในการทดลอง 10 พับ ความแม่นยำในข้อมูลแต่ละส่วนเป็นไปในทิศทางเดียวกับการทดลอง 5 พับแต่มีผลลัพธ์ที่ดีกว่า โดยเมื่อใช้ข้อมูลคำสำคัญเพียงอย่างเดียวแนะนำบทความ 5 อันดับแรกถูกต้อง 92.2 บทความจากทั้งหมด 120.3 บทความ หรือคิดเป็นค่าความแม่นยำที่ 0.76636 ในขณะที่เมื่อเรานำข้อมูลบทคัดย่อเพียงอย่างเดียวมาทำแบบจำลองได้ค่าความแม่นยำที่ลดลงเล็กน้อยโดยแนะนำถูกต้อง 90.9 บทความจากทั้งหมด 120.3 บทความหรือคิดเป็น 0.75563 และเมื่อเรารวมข้อมูลจากทั้งบทคัดย่อและคำสำคัญจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นมาที่ 0.80383 หรือ 96.7 บทความ จาก 120.3 บทความ

## 4.5 การทดลองแบบจำลองกับข้อมูลภาษาอังกฤษและภาษาไทย

ในการสร้างแบบจำลองกับข้อมูลภาษาไทยและภาษาอังกฤษ เราได้ทำการสร้างฟังก์ชั่นในการปรับค่าน้ำหนักอิงตามค่าคะแนนจากข้อมูลภาษาไทยและภาษาอังกฤษ ซึ่งผลลัพธ์ของแบบจำลองจากสองภาษาเพิ่มขึ้นจากใช้ภาษาอังกฤษอย่างเดียวหรือภาษาไทยอย่างเดียว โดยผลการทดลอง 5 พับด้วยแบบจำลองจากข้อมูลคำสำคัญเพียงอย่างเดียว มีบทความที่ได้รับการแนะนำวารสารตรงกับข้อมูลวารสารที่บทความนั้นตีพิมพ์ใน 5 อันดับแรกอยู่ที่ 549.6 บทความจากทั้งหมด 656.4 บทความหรือคิดเป็นค่าความแม่นยำเฉลี่ยอยู่ที่ 0.83730 ในขณะที่เมื่อเรานำข้อมูลบทคัดย่อเพียงอย่างเดียวมาทำแบบจำลองได้ค่าความแม่นยำเป็น 0.78001 หรือ 512 บทความจากทั้งหมด 656.4 บทความ และเมื่อเรารวมข้อมูลจากทั้งบทคัดย่อและคำสำคัญและปรับค่าน้ำหนักของคะแนนจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นมาที่ 0.87752 หรือ 576 บทความ จาก 656.4 บทความ

ในขณะที่ผลลัพธ์การทดลอง 10 พับมีผลลัพธ์ในลักษณะเดียวกันกับการทดลอง 5 พับแต่มีผลลัพธ์ที่ดีกว่า โดยเมื่อใช้ข้อมูลคำสำคัญเพียงอย่างเดียวความแม่นยำอยู่ที่ 0.83730 หรือ 274.8 บทความจาก 328.2 บทความ, ใช้ข้อมูลบทคัดย่อเพียงอย่างเดียวความแม่นยำอยู่ที่ 0.78366 หรือ 257.2 บทความจาก 328.2 บทความ และเมื่อใช้ข้อมูลจากทั้งบทคัดย่อและคำสำคัญจึงได้ความแม่นยำของแบบจำลองเพิ่มขึ้นมาที่ 0.87965 หรือ 288.7 บทความจาก 328.2 บทความ

*ตาราง 3 เปรียบเทียบผลลัพธ์การทดลองแบบจำลองระหว่างข้อมูลในแต่ละภาษา ในการทดลองการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold cross-validation)*

| **แบบจำลองจากข้อมูล** | **ข้อมูลในส่วน** | **จำนวนบทความที่ถูกต้อง** | **ค่าความแม่นยำ** |
|---|---|---|---|
| ข้อมูลภาษาอังกฤษ | คำสำคัญ | 264.8 บทความจาก 328.2 | 0.80683 |
| | บทคัดย่อ | 249.4 บทความจาก 328.2 | 0.75990 |
| | คำสำคัญและบทคัดย่อ | 278.8 บทความจาก 328.2 | 0.84948 |
| ข้อมูลภาษาไทย | คำสำคัญ | 92.2 บทความจาก 120.3 | 0.76636 |
| | บทคัดย่อ | 90.9 บทความจาก 120.3 | 0.75563 |
| | คำสำคัญและบทคัดย่อ | 96.7 บทความจาก 120.3 | 0.80383 |
| ข้อมูลภาษาอังกฤษและภาษาไทย | คำสำคัญ | 274.8บทความจาก 328.2 | 0.83730 |
| | บทคัดย่อ | 257.2บทความจาก 328.2 | 0.78366 |
| | คำสำคัญและบทคัดย่อ | 288.7บทความจาก 328.2 | 0.87965 |



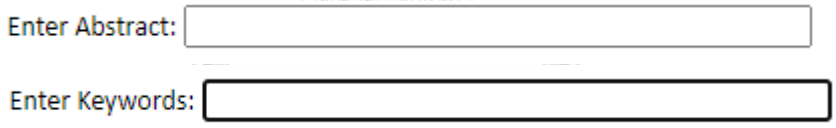
## 4.6 การทดลองแบบจำลองในลักษณะหาความคล้ายระหว่างบทความ

จากที่เราหาค่าความคล้ายระหว่างบทความของผู้เขียนกับวารสารในการทำการทดลองอาจจะทำให้เกิดการสูญหายของเอกลักษณ์ของวารสารไปได้หากข้อมูลบทความในวารสารไม่ได้ครอบคลุมเนื้อหาวารสารอย่างเป็นสัดส่วนและมากเพียงพอมากพอ ผู้วิจัยจึงได้ทดลองเปลี่ยนจากการหาค่าความคล้ายระหว่างบทความของผู้เขียนกับวารสาร เป็นการหาค่าความคล้ายระหว่างบทความของผู้เขียนและแต่ละบทความในวารสารแทน แล้วจากนั้นทำการตรวจสอบว่าบทความที่แนะนำอยู่ในวารสารไหนแล้วจึงแนะนำวารสารที่มีบทความที่มีความคล้ายใกล้เคียงกับบทความของผู้เขียนนั้น ซึ่งหลังจากทดลองพบว่าค่าความแม่นยำน้อยกว่าแบบจำลองโดยใช้ค่าความคล้ายระหว่างบทความกับวารสาร โดยค่าความแม่นยำของการใช้แบบจำลองในการทดลอง 5 พับอยู่ที่ 0.63757 ในแบบจำลองที่ใช้ข้อมูลภาษาอังกฤษ, 0.74733 ในแบบจำลองที่ใช้ข้อมูลภาษาไทย, และ 0.71819 ในแบบจำลอง

ที่ใช้ข้อมูลภาษาอังกฤษและภาษาไทยร่วมกัน และในการทดลอง 10 พับ ค่าความแม่นยำของการใช้แบบจำลองรอบอยู่ที่ 0.67883 ในแบบจำลองที่ใช้ข้อมูลภาษาอังกฤษ, 0.76223 ในแบบจำลองที่ใช้ข้อมูลภาษาไทย, และ 0.73034 ในแบบจำลองที่ใช้ข้อมูลภาษาอังกฤษและภาษาไทยร่วมกัน

## 4.7 การจัดทำการใช้งานระบบแนะนำเบื้องต้น

จากแบบจำลองที่ได้จัดทำเพื่อทำการทดลอง เราได้นำมาจัดทำเป็นระบบแนะนำเบื้องต้นโดยให้ผู้ใช้งานกรอกข้อมูลบทคัดย่อและคำสำคัญของบทความที่ต้องการให้แนะนำวารสารทางวิชาการที่เหมาะสมกับบทความ ระบบจะนำข้อมูลที่ได้ไปประมวลผลโดยในการประมวลผลนั้นจะผ่านกระบวนการในลักษณะเดียวกับบทความที่นำมาฝึกแบบจำลองนั้น กล่าวคือ ระบบจะนำข้อมูลที่ได้ไปตัดคำ ลดรูปคำ ตรวจสอบการสะกดผิดและกำจัดคำหยุด จากนั้นจะแนะนำวารสารที่เหมาะสมออกมาพร้อมทั้งแสดงให้เห็นค่าคะแนนของบทความที่แนะนำและประเภทของวารสาร ทั้งนี้ระบบยังสามารถทำงานและแนะนำได้เฉพาะข้อมูลวารสารที่มีอยู่ในระบบ หากมีวารสารมาเพิ่มเติมจำเป็นจะต้องทำแบบจำลองเพิ่มเติมเพื่อให้ระบบแนะนำสามารถทำงานได้ดียิ่งขึ้น ซึ่งระบบแนะนำเบื้องต้นนี้เป็นส่วนนึงในแนวทางงานในอนาคตที่จะสามารถนำระบบนี้ไปพัฒนาเพื่อให้ผู้ใช้สามารถเข้ามาใช้ได้อย่างสะดวกและมีประโยชน์มากขึ้น






*รูปที่ 12 ตัวอย่างหน้าจอระบบแนะนำเบื้องต้นในส่วนของการรับข้อมูลจากผู้ใช้งาน*

| | index | score | journal_id | path | journal_name | journal_about |
|---|---|---|---|---|---|---|
| 0 | 11 | 0.246660 | 563 | project-journal | Journal of Project in Computer Science and Information Technology | คอมพิวเตอร์ |
| 1 | 4 | 0.175101 | 264 | IT_Journal | Information Technology Journal | คอมพิวเตอร์ |
| 2 | 5 | 0.166867 | 274 | jitubru | Journal of Industrial Technology Ubon Ratchathani Rajabhat University | อุตสาหกรรม |
| 3 | 3 | 0.146045 | 256 | ecticit | ECTI Transactions on Computer and Information Technology | คอมพิวเตอร์ |
| 4 | 9 | 0.142996 | 417 | TNIJournal | TNI Journal of Engineering and Technology | รวม |

*รูปที่ 13 ตัวอย่างหน้าจอระบบแนะนำเบื้องต้นในส่วนของการแสดงผลลัพธ์แนะนำวารสารทางวิชาการที่เหมาะสมกับบทความจากผู้ใช้งาน*

# บทที่ 5

# สรุปผล

งานวิจัยนี้นำเสนอแบบจำลองสำหรับการแนะนำวารสารทางวิชาการให้กับผู้เขียนบทความที่ต้องการตีพิมพ์ในวารสารวิชการ โดยใช้ข้อมูลที่มีบน TCI Thai Journals Online Database (tci-thaijo.org) ซึ่งงานวิจัยนี้ได้ใช้ข้อมูลวารสารวิศวกรรมศาสตร์และเทคโนโลยี ทั้งข้อมูลภาษาอังกฤษและภาษาไทย โดยนำส่วนของคำสำคัญและบทคัดย่อมาใช้

โดยงานวิจัยนี้ได้ทำการทดสอบแบบจำลองบนข้อมูลวารสาร โดยเปรียบเทียบแบบจำลองที่ใช้คำสำคัญแบบจำลองที่ใช้บทคัดย่อและนำข้อมูลทั้งสองส่วนมารวมกัน และได้ทำการทดสอบแบบจำลองที่ใช้ข้อมูลภาษาอังกฤษอย่างเดียว แบบจำลองที่ใช้ข้อมูลภาษาไทยอย่างเดียว และแบบจำลองที่ใช้ข้อมูลภาษาไทยและภาษาอังกฤษมารวมกัน ซึ่งจากการทดลอง เมื่อเราใช้ข้อมูลจากข้อมูลทั้งสองภาษามารวมกันส่งผลให้ได้แบบจำลองที่มีความแม่นยำขึ้น โดยจากการทดลองด้วยการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ เมื่อทดลองด้วยแบบจำลองที่ใช้แค่ภาษาอังกฤษอย่างเดียวได้ค่าความแม่นยำอยู่ที่ 0.84948 และแบบจำลองใช้ภาษาไทยอย่างเดียวมีค่าความแม่นยำอยู่ที่ 0.80383 แต่เมื่อเรานำสองภาษามารวมกันทำให้ได้ค่าความแม่นยำอยู่ที่ 0.87965 ซึ่งสูงกว่าทั้งโมเดลภาษาอังกฤษอย่างเดียวและโมเดลภาษาไทยอย่างเดียว และค่าความแม่นยำของการทดลองการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ ได้ค่าความแม่นยำที่สูงกว่าการทดลองจำนวน 5 พับและการทดลองแบบจำลองในลักษณะความคล้ายระหว่างบทความ

## 5.1 อุปสรรคในงานวิจัย

งานวิจัยนี้มีอุปสรรคในการเลือกข้อมูล เนื่องจากเราต้องการแนะนำวารสารที่เหมาะสม แต่วารสารบางฉบับเป็นประเภททั่วไป ซึ่งยอมรับบทความในสาขาที่หลากหลายมาก ยกตัวอย่างเช่น วารสารวิศวกรรมศาสตร์บางวารสารยอมรับบทความวิศวกรรมศาสตร์ทั้งสาขาวิศวกรรมไฟฟ้า, วิศวกรรมโยธา และ สาขาวิศวกรรมการเกษตร เป็นต้น ทำให้สูญเสียอัตลักษณ์ของวารสาร ผู้วิจัยจึงจำเป็นต้องเลือกวารสารที่เหมาะสมสำหรับการสร้างแบบจำลอง

นอกจากนั้นยังมีอุปสรรคเกี่ยวกับแบบจำลองในภาษาไทยอยู่บ้าง ถึงแม้จะมีไลบารีตัดคำภาษาไทยที่ดีและสะดวกกว่าสมัยก่อนแล้ว แต่ยังมีบางคำที่ยังตัดคำออกมาไม่ถูกต้องมากนัก ทำให้แบบจำลองภาษาไทยยังไม่แม่นยำเท่าที่ควร

## 5.2 แนวทางงานในอนาคต

สำหรับงานในอนาคต ผู้วิจัยวางแผนที่จะใช้ระบบแนะนำวารสารกับผู้แต่งบทความและวัดความพึงพอใจและความมั่นใจของผู้แต่ง โดยปรับปรุงต่อยอดจากระบบแนะนำเบื้องต้นให้เป็นแอปพลิเคชันเพื่อให้ง่ายและสะดวกกับผู้ใช้งาน

นอกจากนั้นเราจะพิจารณาใช้ข้อมูลวารสารต่าง ๆ ที่มีจำนวนเพิ่มขึ้นและขั้นตอนวิธีอื่น ๆ ในการพัฒนาแบบจำลองให้มีความแม่นยำยิ่งขึ้นและรองรับการแนะนำวารสารในประเภทอื่น ๆ นอกจากวารสารวิศวกรรมศาสตร์ให้มากขึ้น โดยเล็งจะพัฒนาปรับปรุงในส่วนแบบจำลองภาษาไทยที่ยังมีความแม่นยำน้อยให้มีการพัฒนาความแม่นยำเพิ่มขึ้นอีกเช่นกัน

# ภาคผนวก

*ตารางที่ 4 เปรียบเทียบผลลัพธ์การทดลองแบบจำลองระหว่างข้อมูลในแต่ละภาษา ในการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ (5-fold cross-validation)*

| **แบบจำลองจากข้อมูล** | **ข้อมูลในส่วน** | **จำนวนบทความที่ถูกต้อง** | **ค่าความแม่นยำ** |
|---|---|---|---|
| ข้อมูลภาษาอังกฤษ | คำสำคัญ | 530.2 บทความจาก 656.4 | 0.80774 |
| | บทคัดย่อ | 493.0 บทความจาก 656.4 | 0.75107 |
| | คำสำคัญและบทคัดย่อ | 555.8 บทความจาก 656.4 | 0.84673 |
| ข้อมูลภาษาไทย | คำสำคัญ | 179.4 บทความจาก 240.6 | 0.74562 |
| | บทคัดย่อ | 180.0 บทความจาก 240.6 | 0.74812 |
| | คำสำคัญและบทคัดย่อ | 190.0 บทความจาก 240.6 | 0.78970 |
| ข้อมูลภาษาอังกฤษและภาษาไทย | คำสำคัญ | 549.6 บทความจาก 656.4 | 0.83730 |
| | บทคัดย่อ | 512.0 บทความจาก 656.4 | 0.78001 |
| | คำสำคัญและบทคัดย่อ | 576.0 บทความจาก 656.4 | 0.87752 |

*ตารางที่ 5 เปรียบเทียบผลลัพธ์การทดลองการปรับน้ำหนักของคะแนนแบบจำลองจากข้อมูลภาษาอังกฤษ ในการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ (5-fold cross-validation)*

| **อัตราส่วนค่าน้ำหนักของคำสำคัญต่อค่าน้ำหนักของบทคัดย่อ** | **จำนวนบทความที่ถูกต้อง** | **ค่าความแม่นยำ** |
|---|---|---|
| อัตราส่วน 1 ต่อ 1 | 543.4 บทความจาก 656.4 | 0.82785 |
| อัตราส่วน 10 ต่อ 1 | 555.3 บทความจาก 656.4 | 0.84369 |
| อัตราส่วน 5 ต่อ 1 | 555.8 บทความจาก 656.4 | 0.84673 |

*ตารางที่ 6 เปรียบเทียบผลลัพธ์การทดลองการปรับน้ำหนักของคะแนนแบบจำลองจากข้อมูลภาษาอังกฤษและภาษาไทย ในการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ (5-fold cross-validation)*

| **อัตราส่วนค่าน้ำหนักของคำสำคัญต่อค่าน้ำหนักของบทคัดย่อ** | **จำนวนบทความที่ถูกต้อง** | **ค่าความแม่นยำ** |
|---|---|---|
| อัตราส่วน 1 ต่อ 1 | 565.6 บทความจาก 656.4 | 0.86167 |
| อัตราส่วน 10 ต่อ 1 | 573.2 บทความจาก 656.4 | 0.87325 |
| อัตราส่วน 5 ต่อ 1 | 576.0 บทความจาก 656.4 | 0.87751 |

*ตาราง 7 เปรียบเทียบผลลัพธ์การทดลองแบบจำลองโดยใช้ข้อมูลบทคัดย่อและคำสำคัญร่วมกันแจกแจงตามแต่ละรอบของการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ (5-fold Cross Validation)*

| **ข้อมูล** | **รอบของ K-fold** | **จำนวนบทความที่ถูกต้อง** | **ความแม่นยำของแบบจำลอง** |
|---|---|---|---|
| ข้อมูลบทความภาษาอังกฤษ | 1 | 549 บทความจาก 657 บทความ | 0.83561 |
| | 2 | 570 บทความจาก 657 บทความ | 0.86758 |
| | 3 | 546 บทความจาก 656 บทความ | 0.83232 |
| | 4 | 553 บทความจาก 656 บทความ | 0.84299 |
| | 5 | 561 บทความจาก 656 บทความ | 0.85518 |
| | **ค่าเฉลี่ย** | **555.8 บทความจาก 656.4 บทความ** | **0.84674** |
| ข้อมูลบทความภาษาไทย | 1 | 191 บทความจาก 241บทความ | 0.79253 |
| | 2 | 187 บทความจาก 241 บทความ | 0.77593 |
| | 3 | 190 บทความจาก 241 บทความ | 0.78838 |
| | 4 | 189 บทความจาก 240 บทความ | 0.78750 |
| | 5 | 193 บทความจาก 240 บทความ | 0.80417 |
| | **ค่าเฉลี่ย** | **190 บทความจาก 240.6 บทความ** | **0.78970** |
| ข้อมูลบทความภาษาอังกฤษและภาษาไทย | 1 | 569 บทความจาก 657 บทความ | 0.86606 |
| | 2 | 577 บทความจาก 657 บทความ | 0.87823 |
| | 3 | 576 บทความจาก 656 บทความ | 0.87805 |
| | 4 | 578 บทความจาก 656 บทความ | 0.88120 |

| | | | |
|---|---|---|---|
| | 5 | 580 บทความจาก 656 บทความ | 0.88415 |
| | **ค่าเฉลี่ย** | **576.0 บทความจาก 656.4 บทความ** | **0.87752** |

*ตารางที่ 8 เปรียบเทียบผลลัพธ์การทดลองแบบจำลองจากข้อมูลบทความภาษาอังกฤษโดยใช้ข้อมูลบทคัดย่อและคำสำคัญร่วมกันแจกแจงตามแต่ละรอบของการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold Cross Validation)*

| **ข้อมูล** | **รอบของ K-fold** | **จำนวนบทความที่ถูกต้อง** | **ความแม่นยำของแบบจำลอง** |
|---|---|---|---|
| ข้อมูลบทความภาษาอังกฤษ | 1 | 274 บทความจาก 329 บทความ | 0.83283 |
| | 2 | 278 บทความจาก 329 บทความ | 0.84498 |
| | 3 | 283 บทความจาก 328 บทความ | 0.86280 |
| | 4 | 289 บทความจาก 328 บทความ | 0.88120 |
| | 5 | 270 บทความจาก 328 บทความ | 0.82317 |
| | 6 | 273 บทความจาก 328 บทความ | 0.83232 |
| | 7 | 274 บทความจาก 328 บทความ | 0.83536 |
| | 8 | 289 บทความจาก 328 บทความ | 0.88110 |
| | 9 | 275 บทความจาก 328 บทความ | 0.83841 |
| | 10 | 283 บทความจาก 328 บทความ | 0.86280 |
| | **ค่าเฉลี่ย** | **278.8 บทความจาก 328.2 บทความ** | **0.84949** |

*ตารางที่ 9 เปรียบเทียบผลลัพธ์การแบบจำลองจากข้อมูลบทความภาษาไทยโดยใช้ข้อมูลบทคัดย่อและคำสำคัญร่วมกันแจกแจงตามแต่ละรอบของการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold Cross Validation)*

| **ข้อมูล** | **รอบของ K-fold** | **จำนวนบทความที่ถูกต้อง** | **ความแม่นยำของแบบจำลอง** |
|---|---|---|---|
| ข้อมูลบทความภาษาไทย | 1 | 96 บทความจาก 121 บทความ | 0.79339 |
| | 2 | 100 บทความจาก 121 บทความ | 0.82645 |
| | 3 | 95 บทความจาก 121 บทความ | 0.78512 |
| | 4 | 95 บทความจาก 120 บทความ | 0.79167 |
| | 5 | 94 บทความจาก 120 บทความ | 0.78333 |
| | 6 | 94 บทความจาก 120 บทความ | 0.78333 |
| | 7 | 96 บทความจาก 120 บทความ | 0.80000 |
| | 8 | 99 บทความจาก 120 บทความ | 0.82500 |
| | 9 | 99 บทความจาก 120 บทความ | 0.82500 |
| | 10 | 99 บทความจาก 120 บทความ | 0.82500 |
| | **ค่าเฉลี่ย** | **96.7 บทความจาก 120.3 บทความ** | **0.80383** |

*ตารางที่ 10 เปรียบเทียบผลลัพธ์การแบบจำลองจากข้อมูลบทความภาษาอังกฤษและภาษาไทยโดยใช้ข้อมูลบทคัดย่อและคำสำคัญร่วมกันแจกแจงตามแต่ละรอบของการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold Cross Validation)*

| **ข้อมูล** | **รอบของ K-fold** | **จำนวนบทความที่ถูกต้อง** | **ความแม่นยำของแบบจำลอง** |
|---|---|---|---|
| ข้อมูลบทความภาษาไทย | 1 | 287 บทความจาก 329 บทความ | 0.87234 |
| | 2 | 284 บทความจาก 329 บทความ | 0.86322 |
| | 3 | 289 บทความจาก 328 บทความ | 0.88120 |
| | 4 | 292 บทความจาก 328 บทความ | 0.89024 |
| | 5 | 286 บทความจาก 328 บทความ | 0.87195 |
| | 6 | 288 บทความจาก 328 บทความ | 0.87805 |
| | 7 | 289 บทความจาก 328 บทความ | 0.88109 |
| | 8 | 293 บทความจาก 328 บทความ | 0.89329 |
| | 9 | 285 บทความจาก 328 บทความ | 0.86890 |
| | 10 | 294 บทความจาก 328 บทความ | 0.89363 |
| | **ค่าเฉลี่ย** | **288.7 บทความจาก 328.2 บทความ** | **0.87965** |

# บรรณานุกรม


1. Scopus 2021 [cited 2021 10 October]. Available from: https://www.elsevier.com/solutions/scopus/how-scopus-works/content.

2. ThaiJo 2021 [Available from: https://www.tci-thaijo.org/.

3. Haddi E, Liu X, Shi Y. The Role of Text Pre-processing in Sentiment Analysis. Procedia Computer Science. 2013;17:26-32.

4. Sornlertlamvanich V. Word segmentation for Thai in machine translation system. Machine Translation. 1993.

5. PyThaiNLP [Available from: https://pythainlp.github.io/docs/2.3/.

6. Robertson S. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. Journal of Documentation - J DOC. 2004;60:503-20.

7. Pavan M, Mizzaro S, Scagnetto I. Content-Based Similarity of Twitter Users2015.

8. Li B, Han L. Distance Weighted Cosine Similarity Measure for Text Classification. Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning --- IDEAL 2013 - Volume 8206; Hefei, China: Springer-Verlag; 2013. p. 611–8.

9. Moreno-Torres JG, Saez JA, Herrera F. Study on the Impact of Partition-Induced Dataset Shift on k-Fold Cross-Validation. IEEE Transactions on Neural Networks and Learning Systems. 2012;23(8):1304-12.

10. Deriu JaL, Aurelien and De Luca, Valeria and Severyn, Aliaksei and Müller, Simon and Cieliebak, Mark and Hofmann, Thomas and Jaggi, Martin. Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification. arXiv. 2017.

11. Balahur A, Turchi M, editors. Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data2013 sep; Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA.

12. Singh R, Maurya S, Tripathi T, Narula T, Srivastav G. Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology. 2020;9:2249-8958.

13. Magara MB, Ojo SO, Zuva T, editors. A comparative analysis of text similarity

measures and algorithms in research paper recommender systems. 2018 Conference on Information Communications Technology and Society (ICTAS); 2018 8-9 March 2018.

14. Lee J, Lee K, Kim J. Personalized Academic Research Paper Recommendation System. 2013.

15. Sugiyama K, Kan M-Y. Scholarly paper recommendation via user's recent research interests. Proceedings of the 10th annual joint conference on Digital libraries; Gold Coast, Queensland, Australia: Association for Computing Machinery; 2010. p. 29–38.

16. Jain S, Khangarot H, Singh S, editors. Journal Recommendation System Using Content-Based Filtering2019; Singapore: Springer Singapore.

17. Heimerl F, Lohmann S, Lange S, Ertl T, editors. Word Cloud Explorer: Text Analytics Based on Word Clouds. 2014 47th Hawaii International Conference on System Sciences; 2014 6-9 Jan. 2014.

18. Stopwords ISO [Available from: https://github.com/stopwords-iso/stopwords-iso.

19. Kuhn M, Johnson K. Over-Fitting and Model Tuning. Applied Predictive Modeling. New York, NY: Springer New York; 2013. p. 61-92.

20. James G, Witten D, Hastie T, Tibshirani R. Resampling Methods. An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer New York; 2013. p. 175-201.

21. Deshpande M, Karypis G. Item-based top-N recommendation algorithms. ACM Trans Inf Syst. 2004;22(1):143–77.

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ-สกุล** | นิธิรันดร์ นุ่มนนท์ |
| **วัน เดือน ปี เกิด** | 16 สิงหาคม 2538 |
| **สถานที่เกิด** | ขอนแก่น |
| **วุฒิการศึกษา** | ปริญญาตรี คณะวิศวกรรมศาสตร์ สาขาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย |
| **ที่อยู่ปัจจุบัน** | 273/96 หมู่ 23 ถนนศรีจันทร์ ตำบลบ้านเป็ด อำเภอเมือง จังหวัดขอนแก่น |
| **ผลงานตีพิมพ์** | Numnonda N, Chanyachatchawan S, Tuaycharoen N. Journal Recommendation System for Author Using Thai and English Information from Manuscript. Cham: Springer International Publishing; 2022. p. 142-51. |