



การวิเคราะห์ข้อความภาษาไทยเกี่ยวกับการตั้งครรภ์ด้วยวิธีการสร้างแบบจำลองหัวข้อ
(TOPIC MODELING)

THE ANALYSIS OF THAI LANGUAGE ON PREGNANCY PROBLEMS DOMAIN:USING
LATENT DIRICHLET ALLOCATION FOR TOPIC MODELING

ปิยวรรณ ทองพลอย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การวิเคราะห์ข้อความภาษาไทยเกี่ยวกับการตั้งครุฑด้วยวิธีการสร้างแบบจำลองหัวข้อ
(TOPIC MODELING)



สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

THE ANALYSIS OF THAI LANGUAGE ON PREGNANCY PROBLEMS DOMAIN:USING
LATENT DIRICHLET ALLOCATION FOR TOPIC MODELING



A Master's Project Submitted in Partial Fulfillment of the Requirements
for the Degree of MASTER OF SCIENCE
(Data Science)

Faculty of Science, Srinakharinwirot University

2020

Copyright of Srinakharinwirot University

สารนิพนธ์
เรื่อง
การวิเคราะห์ข้อความภาษาไทยเกี่ยวกับการตั้งครรภ์ด้วยวิธีการสร้างแบบจำลองหัวข้อ
(TOPIC MODELING)
ของ
ปิยวรรณ ทองพลอย

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)
คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก	 ประธาน
(อาจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)	(ผู้ช่วยศาสตราจารย์ ดร.อัศวินทร์ ไพบุญย์พานิช)
..... ที่ปรึกษาร่วม	 กรรมการ
(อาจารย์ ดร.รัตนชัยนันท์ ธรรมสุจริต)	(ผู้ช่วยศาสตราจารย์ ดร.นุรีย์ วิวัฒน์วัฒนา)

ชื่อเรื่อง	การวิเคราะห์ข้อความภาษาไทยเกี่ยวกับการตั้งครรภ์ด้วยวิธีการสร้างแบบจำลองหัวข้อ (TOPIC MODELING)
ผู้วิจัย	ปิยวรรณ ทองพลอย
ปริญญา	วิทยาศาสตร์มหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ
อาจารย์ที่ปรึกษาร่วม	อาจารย์ ดร. รัตน์ชัยนันท์ ธรรมสุจริต

ปัจจุบันมีผู้ใช้งานอินเทอร์เน็ตจำนวนมากเข้าปรึกษาปัญหาเกี่ยวกับแพทย์ผ่านชุมชนออนไลน์ (Community question answering: CQA) ซึ่งเป็นงานที่หนักสำหรับแพทย์ที่ต้องตอบคำถามให้ทัน ซึ่งปัจจุบันมีการนำระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) มาประยุกต์ใช้ในการให้ข้อมูล ถามตอบปัญหา แต่การพัฒนาหุ่นยนต์ได้ตอบการสนทนา (Chatbot) นั้นมีข้อจำกัด เช่น มีราคาแพง และเป็นงานที่ทำหายในการที่จะทำให้คอมพิวเตอร์เข้าใจถึงภาษามนุษย์จากข้อความในเอกสารที่เป็นภาษาไทย งานวิจัยนี้จึงมีวัตถุประสงค์ในการวิเคราะห์หาคำที่มีนัยสำคัญและการจำแนกหัวข้อในข้อความที่มีความคล้ายคลึงกัน เพื่อสร้างแบบจำลองที่สามารถนำไปพัฒนาระบบได้ตอบอัตโนมัติ (Chatbot) ให้สามารถโต้ตอบกับผู้ใช้งานได้ตรงประเด็นมากขึ้น โดยใช้วิธีการสร้างแบบจำลองหัวข้อ (Topic modeling) และการจำแนกกลุ่มข้อความ (Clustering) โดยงานวิจัยนี้ได้ประยุกต์ใช้เทคนิค การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการวิเคราะห์หาคำที่มีนัยสำคัญและจำแนกหัวข้อในข้อความเกี่ยวกับเรื่องของการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิด ในส่วนของการประเมินประสิทธิภาพ ได้ใช้การวัดผลแบบ extrinsic evaluation โดยใช้เทคนิค K-means ในการจัดกลุ่มหัวข้อและประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient

คำสำคัญ : การจัดสรรดีรีเคลแฝง, แบบจำลองหัวข้อ, การจำแนกกลุ่มข้อความ

Title	THE ANALYSIS OF THAI LANGUAGE ON PREGNANCY PROBLEMS DOMAIN:USING LATENT DIRICHLET ALLOCATION FOR TOPIC MODELING
Author	PIYAWAN THONGPLOY
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Professor Dr. Sirisup Laohakiat
Co Advisor	Professor Dr. Ratchainant Thammasudjarit

Nowadays, there are many messages used to discuss problems with doctors through online communities. It is a difficult task for doctors to answer questions in a timely manner. Currently, chatbot systems are applied to provide information, ask and answer questions, but the development of chatbot systems is limited. For example, it is expensive and a challenging task to make computers understand human language from text in documents. The objective of this research is to analyze significant words and topics in the text, to create a model that can be used to develop an automated response system (Chatbot) , to be able to interact with users more relevantly by using topic modeling and clustering methods. The technique used in this research is Latent Dirichlet Allocation in the Analysis of Significant Words and Topics in Pregnancy sexual relations and birth control Texts. Also, as a part of the performance assessment, the K-means technique was used for clustering topics and to assess the cluster efficiency using Silhouette Coefficient.

Keyword : Topic modeling, Clustering, Latent Dirichlet Allocation

กิตติกรรมประกาศ

ขอขอบพระคุณอาจารย์ที่ปรึกษางานวิจัยนี้ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ อันเป็นประโยชน์อย่างยิ่งในการทำวิจัย ตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่างๆ ที่เกิดขึ้นระหว่างการดำเนินการวิจัย

ขอขอบพระคุณคณาจารย์ คณะวิทยาศาสตร์ สาขาเทคโนโลยีสารสนเทศทุกท่านที่ได้อบรมและสั่งสอนให้วิชาความรู้เพื่อนำไปต่อยอดและให้คำแนะนำให้งานวิจัยนี้ถูกต้องสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ เจ้าหน้าที่ที่เกี่ยวข้องที่กรุณาให้คำแนะนำและประสานงานให้สำเร็จด้วยดี

ขอขอบคุณเพื่อนๆ นักศึกษาระดับปริญญาโท หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ช่วยให้คำแนะนำอันเป็นประโยชน์และช่วยเหลือเกื้อกูลเพื่อให้งานวิจัยนี้ลุล่วงไปได้ด้วยดี

สุดท้ายนี้ ผู้วิจัยขอขอบพระคุณครอบครัวที่เป็นกำลังใจที่สำคัญ และสนับสนุนเรื่องการศึกษาในทุกๆ ด้านจนกระทั่งทุกอย่างประสบความสำเร็จและลุล่วงมาจนถึงวันนี้ ขอขอบพระคุณอย่างสูง

ปิยวรรณ ทองพลอย

สารบัญ

หน้า

บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ	ช
สารบัญตาราง.....	ณ
สารบัญรูปภาพ	ญ
บทที่ 1 บทนำ.....	1
ภูมิหลัง	1
ความมุ่งหมายของงานวิจัย.....	2
ขอบเขตของการวิจัย	2
กรอบแนวคิดในงานวิจัย	3
สมมติฐานในการวิจัย.....	3
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
2.1 ทฤษฎีที่เกี่ยวข้อง	4
2.2 งานวิจัยที่เกี่ยวข้อง	8
บทที่ 3 วิธีดำเนินการวิจัย.....	15
3.1 การออกแบบโครงสร้างงานวิจัย	15
3.2 เครื่องมือที่ใช้ในงานวิจัย.....	17
3.3 ชุดข้อมูล (Dataset)	17
3.4 แบบจำลอง (Model).....	20
3.5 การวัดประสิทธิภาพ.....	23

บทที่ 4 ผลการดำเนินงานวิจัย	25
ผลลัพธ์การทดลองการจัดกลุ่มด้วยอัลกิริทึม K-mean และประเมินประสิทธิภาพการจำแนก กลุ่มด้วย Silhouette Coefficient	25
ผลลัพธ์ของการจำแนกกลุ่มข้อความด้วยเทคนิคการจัดสรรดีรีเคิลแฟง (Latent Dirichlet Allocation)	27
บทที่ 5 สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ	32
สรุปผลการวิจัย	32
อภิปรายผลการวิจัย	33
ข้อเสนอแนะ	35
บรรณานุกรม	36
ประวัติผู้เขียน	38

สารบัญตาราง

หน้า

ตาราง 1 ตัวอย่างการเปรียบเทียบระหว่าง TBLMSQATM กับ LMIR	11
ตาราง 2 เป็นตัวอย่าง Keyword ในหัวข้อของ LDA กับ 5 หัวข้อ	12
ตาราง 3 แสดงตัวอย่างจากการศึกษา เรื่องราวใหม่ (New story) และเป็นการจับคู่ที่ดีที่สุด	13
ตาราง 4 การประเมินประสิทธิภาพการจำแนกกลุ่มหัวข้อ	26
ตาราง 5 ผลการวิเคราะห์หาค่าที่มีนัยสำคัญและหัวข้อของแบบจำลอง	28
ตาราง 6 ตารางผลการเปรียบเทียบสมมุติฐานกับผลการทำนายของแบบจำลอง	34



สารบัญรูปภาพ

หน้า

ภาพประกอบ 1 โครงสร้างงานวิจัย	16
ภาพประกอบ 2 เว็บไซต์ honestdocs ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรภ์	18
ภาพประกอบ 3 ตัวอย่างข้อมูล	18
ภาพประกอบ 4 แสดงตัวอย่างโค้ดนำเข้าโมดูลสำคัญที่ใช้สำหรับสร้างแบบจำลอง	19
ภาพประกอบ 5 แสดงตัวอย่างโค้ดนำเข้าไฟล์ข้อมูลและข้อมูลที่ใช้สำหรับสร้างแบบจำลอง	19
ภาพประกอบ 6 กระบวนการเตรียมข้อมูลและผลลัพธ์	19
ภาพประกอบ 7 กระบวนการสกัดใจความสำคัญของข้อความและผลลัพธ์	20
ภาพประกอบ 8 การจับกลุ่มคำ (Word Cloud)	21
ภาพประกอบ 9 ผลลัพธ์จำนวนคำในข้อมูล	21
ภาพประกอบ 10 ถูคำ (Bag of word)	21
ภาพประกอบ 11 กระบวนการลดมิติของข้อมูลและผลลัพธ์	22
ภาพประกอบ 12 กระบวนการวัดประสิทธิภาพของแบบจำลอง	23
ภาพประกอบ 13 กราฟ Silhouette score ในแต่ละหัวข้อ	26
ภาพประกอบ 14 ภาพผลลัพธ์การทำวิเคราะห์หัวข้อจากแบบจำลอง	27
ภาพประกอบ 15 ผลการทำนายข้อความว่าเป็นข้อที่ 1	29
ภาพประกอบ 16 ผลการทำนายข้อความว่าเป็นข้อที่ 2	29
ภาพประกอบ 17 ผลการทำนายข้อความว่าเป็นข้อที่ 3	30
ภาพประกอบ 18 ผลการทำนายข้อความว่าเป็นข้อที่ 4	31
ภาพประกอบ 19 ภาพผลการทำนายจากแบบจำลอง	33

บทที่ 1

บทนำ

ภูมิหลัง

มีผู้ใช้งานอินเทอร์เน็ตจำนวนมากเข้าปรึกษาปัญหาสุขภาพกับแพทย์ผ่านชุมชนออนไลน์ (Community question answering: CQA) ในการปรึกษาเกี่ยวกับเรื่องการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิด การขอคำปรึกษาผ่านชุมชนออนไลน์นั้นเป็นช่องทางที่สะดวกต่อการใช้งาน เพราะสามารถปิดบังตัวตนและสามารถเข้าปรึกษาแพทย์ได้ตลอดเวลา แต่ผู้ใช้งานต้องรอการตอบกลับจากแพทย์ เพราะแพทย์ไม่สามารถให้คำปรึกษาได้ทันทีเมื่อผู้ใช้งานขอคำปรึกษา (Availability) และหากมีปริมาณคำปรึกษามาก แพทย์ไม่สามารถให้คำปรึกษาได้ทัน (Scalability) ซึ่งจากข้อจำกัดเหล่านี้ ปัจจุบันได้มีการนำระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) มาประยุกต์ใช้ในการให้ข้อมูล ถามตอบปัญหา การวิเคราะห์ข้อมูล หรือการเก็บข้อมูลจากผู้ใช้งาน ในด้านธุรกิจต่างๆ ทั้งในด้านการบริการลูกค้า ด้านการศึกษา ด้านการบริหารทรัพยากรบุคคล ด้านการเงิน และด้านสุขภาพ (1, 4, 10)

ระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) ถูกพัฒนาขึ้นมาให้มีบทบาทในการตอบกลับการสนทนาผ่านตัวอักษรแบบอัตโนมัติผ่านแพลตฟอร์มต่างๆ ซึ่งกลายมาเป็นที่นิยมในกลุ่มของธุรกิจหลากหลายด้าน โดยระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) นั้นมี 2 ประเภทหลัก คือ แชทบอทที่มีการกำหนดการถามตอบชัดเจน (Rule based Chatbot) และ แชทบอทเพื่อการสนทนา (Conversation Chatbot) ที่มีการนำเทคโนโลยีปัญญาประดิษฐ์ หรือ Machine learning (15) เข้ามาประยุกต์ใช้ให้ระบบสามารถสนทนาที่มีการตอบกลับคู่สนทนาได้เสมือนมนุษย์จริงและสามารถตอบกลับได้ทันที แต่การพัฒนาระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) นั้นมีข้อจำกัด เช่น มีราคาแพง ไม่เข้าใจภาษาท้องถิ่นของผู้ใช้งาน การสะกดคำผิดหรือการใช้คำแสลงบางคำอาจทำให้การตอบไม่ถูกต้อง (11) ซึ่งการประมวลผลภาษาธรรมชาติ หรือ Natural language processing (NLP) ยังเป็นงานที่ท้าทายในการที่จะทำให้คอมพิวเตอร์เข้าใจถึงภาษามนุษย์จากข้อความในเอกสาร โดยวิธีการสร้างแบบจำลองหัวข้อ (Topic modeling) โดยเทคนิคการจัดสรรตัวเคลแฝง (Latent Dirichlet Allocation หรือ LDA) เป็นเทคนิคที่นิยมอย่างแพร่หลายในการประมวลผลภาษาธรรมชาติ เพื่อค้นหาหัวข้อและคำที่มีนัยสำคัญที่ถูกซ่อนในเอกสาร (3, 7)

จากการศึกษางานวิจัยมีการนำ การจัดสรรตัวเคลแฝง (Latent Dirichlet Allocation หรือ LDA) ในการประยุกต์ใช้กับระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) ได้นำมาจำแนกข้อความจากการรับข้อมูลจากผู้ใช้งานหาคำที่มีนัยสำคัญ (13) และ การแปลงข้อความเป็นเวกเตอร์ (9)

และได้ถูกมาพัฒนาในหลายๆ ด้าน ตัวอย่างเช่น ในด้านการดูแลสุขภาพ (Healthcare) ได้นำมาประยุกต์ใช้ในการ วิเคราะห์ฟอร์มการตั้งครรถ์ออนไลน์เพื่อต้องการเข้าใจผู้หญิงตั้งครรถ์ให้มากขึ้น (16) ในงานด้านการจำแนกเอกสารเพื่อพัฒนาระบบการจัดประเภทเอกสารวิจัย (8) การนำมาหาข้อเท็จจริงในข่าว โดยการนำมาเปรียบเทียบข้อเท็จจริงระหว่างองค์การอนามัยโลก (WHO) และจากข่าวในช่วงเวลานั้นๆ (6) สรุปเอกสารข่าวออนไลน์หลายฉบับเพื่อให้ผู้อ่านเข้าใจเจตนาของเอกสารข่าวออนไลน์ (14) มีการประยุกต์ใช้ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) กับในหลายๆ โดเมน ซึ่งนำมาใช้กับการวิเคราะห์ข้อความจำนวนมาก

ในงานวิจัยนี้จึงได้ศึกษาการสร้างแบบจำลองหัวข้อ (Topic Modeling) โดยการใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) เพื่อจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน เพื่อพัฒนาแบบจำลองที่สามารถนำไปพัฒนาระบบโต้ตอบอัตโนมัติ (Chatbot) ให้สามารถโต้ตอบกับผู้ใช้งานได้ตรงประเด็นมากขึ้น ในเรื่องของการตั้งครรถ์ การมีเพศสัมพันธ์ และการคุมกำเนิดในข้อความภาษาไทย โดยมีการวัดประสิทธิภาพของแบบจำลองหัวข้อวัดประสิทธิภาพได้ใช้การวัดผลแบบ Extrinsic evaluation ด้วยการจำแนกกลุ่ม (Clustering) ของข้อความโดยใช้ K-mean โดยประเมินประสิทธิภาพด้วย Silhouette Coefficient

ความมุ่งหมายของงานวิจัย

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ ดังนี้

1. เพื่อแบ่งกลุ่มของข้อความคำปรึกษาที่มีความใกล้เคียงกัน ในคำปรึกษาปัญหาเกี่ยวกับการตั้งครรถ์ การมีเพศสัมพันธ์ และการคุมกำเนิด โดยการใช้เทคนิค Machine learning แบบ Unsupervised Learning
2. ทราบถึงหลักการของอัลกอริทึมการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการวิเคราะห์หาคำที่มีนัยสำคัญและจำแนกกลุ่มคำปรึกษาที่มีความใกล้เคียงกันในหัวข้อเดียวกัน อีกทั้งทราบถึงอัลกอริทึม K-mean ในการจำแนกกลุ่มเพื่อใช้ในการวัดผลแบบ extrinsic evaluation และประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient

ขอบเขตของการวิจัย

ประชากรที่ใช้ในการวิจัย

ข้อมูลเกี่ยวกับข้อความคำปรึกษาปัญหาการตั้งครรถ์ การมีเพศสัมพันธ์ และการคุมกำเนิดภาษาไทย ชุดข้อมูลในช่วงวันที่ 23 มกราคม 2017 ถึง 15 ตุลาคม 2020 จำนวนข้อมูล

ทั้งหมดมี 9,987 ข้อความ จากเว็บไซต์ Honestdocs ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรรค์

กลุ่มตัวอย่างที่ใช้ในการวิจัย

ข้อมูลเกี่ยวกับข้อความคำปรึกษาปัญหาการตั้งครรรค์ การมีเพศสัมพันธ์ และการคุมกำเนิดและวิธีการคุมกำเนิดภาษาไทย ชุดข้อมูลในช่วงวันที่ 23 มกราคม 2017 ถึง 15 ตุลาคม 2020

ตัวแปรที่ศึกษา

ตัวแปรอิสระ เป็นข้อมูลข้อความคำปรึกษา

กรอบแนวคิดในงานวิจัย

งานวิจัยนี้ศึกษาการแบ่งกลุ่มข้อความคำปรึกษาที่มีความใกล้เคียงกัน โดยใช้เทคนิค Machine learning แบบ Unsupervised Learning โดยข้อมูลข้อความคำปรึกษาปัญหาการตั้งครรรค์ การมีเพศสัมพันธ์ และการคุมกำเนิดและวิธีการคุมกำเนิดภาษาไทย ชุดข้อมูลในช่วงวันที่ 23 มกราคม 2017 ถึง 15 ตุลาคม 2020 จำนวนข้อมูลทั้งหมดมี 9,987 ข้อความ จากเว็บไซต์ Honestdocs ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรรค์ ซึ่งข้อมูลถูกเก็บในรูปแบบของตารางสร้างแบบจำลองเพื่อแบ่งกลุ่มข้อความ โดยใช้เครื่องมือ Google Colab เขียนด้วยโปรแกรมภาษาไพทอน (python) โดยใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ในการหาค่าที่มีนัยสำคัญและจำแนกข้อความที่คล้ายคลึงกันให้อยู่ในหัวข้อเดียวกัน และใช้เทคนิค K-mean ในการทำ Clustering เพื่อประเมินประสิทธิภาพในการแบ่งกลุ่มที่ดีที่สุด

สมมติฐานในการวิจัย

แบบจำลองการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) สามารถวิเคราะห์ค่าที่มีนัยสำคัญและจำแนกข้อความคำปรึกษาปัญหาการตั้งครรรค์ การมีเพศสัมพันธ์ และการคุมกำเนิดภาษาไทยที่คล้ายคลึงกันให้อยู่ในหัวข้อเดียวกันได้

โดยผู้วิจัยได้ทำการพิจารณาและแบ่งกลุ่มของหัวข้อเป็น 4 หัวข้อดังนี้

- (1) อาการผิดปกติในขณะตั้งครรรค์
- (2) ความผิดปกติของประจำเดือน
- (3) ความกังวลเมื่อมีเพศสัมพันธ์
- (4) การคุมกำเนิด

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

งานวิจัยนี้มีการจัดกลุ่มของข้อความ โดยงานวิจัยนี้เป็นงานที่เกี่ยวข้องกับการประมวลผลทางภาษา คือ การประมวลภาษาธรรมชาติ (Natural Language Processing หรือ NLP) โดยแบ่งกลุ่มของข้อความโดยใช้ทฤษฎี แบบจำลองหัวข้อ (Topic Modeling) โดยการใช้เทคนิค การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation หรือ LDA) โดยลักษณะของโมเดลเป็น Unsupervised learning และ Probabilistic Model ซึ่งขั้นตอนแรกที่สำคัญของการประมวลข้อมูลที่เป็นข้อความคือขั้นตอนการเตรียมข้อมูล (Text preprocessing) งานวิจัยนี้ได้มีการศึกษาและได้นำเอาทฤษฎีต่างๆ มาประยุกต์ใช้กับงานวิจัย ดังนี้

2.1.1 การประมวลภาษาธรรมชาติ (Natural Language Processing หรือ NLP)

Natural Language Processing หรือ การประมวลภาษาธรรมชาติ เป็นศาสตร์หนึ่งในสาขาวิชาวิทยาการคอมพิวเตอร์ที่เกี่ยวข้องกับปัญญาประดิษฐ์ (Artificial Intelligence) และภาษาศาสตร์คอมพิวเตอร์ (Computational Linguistics) เป็นศาสตร์ที่จะทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์และสามารถโต้ตอบด้วยภาษามนุษย์ โดยข้อความหรือเสียง โดยปัจจุบันมีเทคโนโลยีระบบหุ่นยนต์ได้ตอบการสนทนา (Chatbot) เป็นโปรแกรมคอมพิวเตอร์ที่ถูกพัฒนาขึ้นมาเพื่อการตอบกลับสนทนาอัตโนมัติ ผ่าน Messaging Application เสมือนว่าผู้ใช้งานระบบคุยกับมนุษย์จริงๆ โดยโปรแกรมตอบกลับอัตโนมัติ มักจะใช้ในระบบในการบริการหรือการเก็บข้อมูล โดยมีระบบฐานข้อมูลในการเก็บบันทึกการโต้ตอบ นำข้อมูลเหล่านั้นมาตรวจจับคำสำคัญจากข้อมูลเพื่อประมวลหาคำตอบส่งกลับไปยังคู่สนทนา โดยบางระบบหุ่นยนต์ได้ตอบการสนทนา โดยแชทบอทสามารถพัฒนาได้ในรูปแบบ Rule based Chatbot หรือการพัฒนาแบบใช้ปัญญาประดิษฐ์ (AI) มีการใช้ NLP เพื่อช่วยให้ระบบหุ่นยนต์ได้ตอบการสนทนาสามารถพูดคุยอย่างเป็นธรรมชาติ โดยปัจจุบันระบบหุ่นยนต์ได้ตอบการสนทนามีบทบาทต่อภาคธุรกิจและบริการต่างๆ ระบบหุ่นยนต์ได้ตอบการสนทนา สามารถให้ข้อมูลข่าวสารต่างๆ แก่ลูกค้า ช่วยลดการทำงานของมนุษย์ได้ และสามารถเข้าถึงลูกค้าได้มากขึ้น

2.1.2 แบบจำลองหัวข้อ (Topic Modeling)

แบบจำลองหัวข้อ (Topic Modeling) เป็นเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เป็นการเรียนรู้ของเครื่องแบบ Unsupervised ใช้ในการวิเคราะห์ข้อมูลตัวอักษรเพื่อกำหนดกลุ่มคำสำหรับชุดเอกสาร (Document) เป็นการจำลองหัวข้อแบบการกระจายตัวของ

ข้อมูล เพื่อนำไปใช้ในการจัดกลุ่มข้อมูล โดยเอกสารแต่ละเอกสารจะประกอบด้วยคำรวมกันหลายๆ หัวข้อแต่ละหัวข้อมีการแจกแจงความน่าจะเป็นของคำนั้นๆ ที่เกิดขึ้นในแต่ละหัวข้อ

2.1.3 การจัดสรรตรีเคิลแฝง (Latent Dirichlet Allocation หรือ LDA)

การจัดสรรตรีเคิลแฝง (Latent Dirichlet Allocation หรือ LDA) เป็นแบบจำลองความน่าจะเป็นสำหรับการทำ แบบจำลองหัวข้อ (Topic Modeling) เป็นการหาหัวข้อ (topic) ที่ซ่อนอยู่ในเอกสาร โดยมีหัวข้อที่แฝงอยู่ในเอกสาร ถูกสร้างจากแนวคิดที่ว่า ในเอกสารนั้นจะประกอบไปด้วยหัวข้อที่ถูกรวมกันแบบไม่มีรูปแบบและกระจายในกลุ่มของคำศัพท์ของเอกสารนั้น โดยการทำ LDA จะทำการหาค่าความน่าจะเป็นของความสัมพันธ์ของหัวข้อของแต่ละประโยค และค่าความน่าจะเป็นของคำในแต่ละหัวข้อ ลักษณะของเทคนิคนี้เป็นแบบ การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ ใช้หลักความน่าจะเป็นในการหาพารามิเตอร์ (Probabilistic Model) หลักการทำงานของ LDA ได้แก่

(1) เลือกพารามิเตอร์ โดยกำหนดจำนวนหัวข้อ ตัวอย่างเช่น Clustering algorithms เช่น K-mean ที่เลือกจำนวน Cluster หรือค่า k

(2) LDA จะทำการอ่านทุกคำของแต่ละเอกสารและกำหนดหัวข้อ โดยผลลัพธ์ที่ได้คือ กลุ่มคำของหัวข้อ (Topic Representation) ที่อยู่รวมกันใน ถูคำ (bag of words) และ หัวข้อของแต่ละเอกสาร

จำนวนหัวข้อที่ดีที่สุดมีวิธี คือการหา Perplexity และ Log-Likelihood function

วิธีแรก Perplexity หาจำนวน k ที่เหมาะสม และประเมินประสิทธิภาพของโมเดลในการอธิบายข้อมูล ค่า Perplexity ต่ำมาก ก็หมายความว่าโมเดลนั้นมีประสิทธิภาพดีมาก วิธีการหาแบบนี้เมื่อมีการกำหนดค่า k จากการ LDA ได้กลุ่มของคำ นำคำเหล่านั้นที่ได้กำหนดความเชื่อมโยงกับแต่ละหัวข้อเปรียบเทียบกับกระจายของคำในเอกสาร เพื่อหา cluster ของหัวข้อของเอกสารนั้นๆ สามารถอธิบายเป็นสมการได้ดังนี้

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^M \log [p(w_d)]}{\sum_{d=1}^M N_d}\right)$$

วิธีที่สอง Log-Likelihood function การหาค่าความใกล้เคียง โดยผลลัพธ์ที่ได้หากมีค่า สูง แสดงว่ามีความใกล้เคียงกันมาก

$$L(\theta, m|D) = p_{\theta, m}(d) = P_{\theta, m}(D = d)$$

$$L(\theta, m|D) = p(D|\theta, m)$$

2.1.4 ขั้นตอนการเตรียมข้อมูลภาษาไทย (Thai text preprocessing)

ทำความสะอาดข้อความ (Clean text) นำข้อความมาทำความสะอาดโดยการตัดเครื่องหมายพิเศษต่างๆ ที่ไม่จำเป็น

ตัดคำ (Tokenize) สำหรับภาษาไทยมี library ในการตัดคำภาษาไทย เช่น PyThaiNLP มีการรองรับหลายอัลกอริทึม ได้แก่ newmm, longest, multi_cut, pyicu, deepcut, attacut, Thai Character Clusters (TCCs) และ Enhanced Thai Character Cluster (ETCC)

การตัด Stop words หรือคำที่ไม่ได้สื่อความหมายและพบบ่อย เช่น การ, ความ, ที่, ซึ่ง, และอื่น เป็นต้น การตัดคำจะช่วยให้คลังคำศัพท์ (Bag of words) เก็บคำศัพท์ไม่เยอะเกินความจำเป็น โดยเฉพาะเมื่อเรามีชุดข้อมูลที่มีขนาดใหญ่ แต่การตัดคำบางส่วนก็อาจทำให้ความแม่นยำของแบบจำลองลดลง และการแปลงคำศัพท์ให้อยู่ในรากศัพท์ เพื่อลดจำนวนคำศัพท์ลง

การสร้างคลังคำศัพท์จากข้อความทั้งหมด (bag of words) เป็นการแปลงคำที่มีลักษณะเฉพาะเป็นรหัส มีการสกัดคุณลักษณะ 2 แบบ คือ การสกัดคุณลักษณะ (Feature) จากการนับคำ โดยนับจำนวนครั้งที่แต่ละคำจากคำทั้งหมดปรากฏในแต่ละข้อความ และการสกัดคุณลักษณะ ด้วย TF-IDF คือการคำนวณค่า TF-IDF ของแต่ละคำ TF-IDF (term frequency-inverse document frequency) นั่นคือ การนับจำนวนครั้งที่แต่ละคำปรากฏในแต่ละข้อความ (Term frequency)หารด้วยคำทั้งหมดในข้อความนั้น และนำมาคูณกับจำนวนของข้อความทั้งหมดหารด้วยจำนวนข้อความที่แต่ละคำปรากฏ และ เทค log โดยคำที่ปรากฏในข้อความมากๆ ค่า IDF จะมีค่ามากตาม แต่ถ้าคำนั้นปรากฏอยู่ในทุกข้อความก็จะให้ค่า IDF ที่ต่ำ

2.1.5 การวิเคราะห์การจัดกลุ่มข้อมูล (Clustering Analysis)

การจัดกลุ่มข้อมูล หรือ Clustering Analysis เป็น Machine Learning ประเภท การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยเป็นวิธีที่เป็นการจำแนกข้อมูลที่มีความคล้ายกันให้อยู่ในกลุ่มเดียวกัน เทคนิคนี้ไม่ได้ใช้เพื่อหาความแม่นยำ แต่เป็นการหาความสัมพันธ์ของข้อมูลอีกรูปแบบหนึ่ง อัลกอริทึมที่ใช้ในการจัดกลุ่ม แบ่งเป็น 2 ประเภท ได้แก่

- (1) ประเภทการแบ่งกลุ่มชัดเจน (Hard Clustering) เทคนิคนี้มีการแบ่งกลุ่มข้อมูล แต่ละข้อมูลไปอยู่ในกลุ่มใดกลุ่มหนึ่งเท่านั้น เทคนิคประเภทนี้ ได้แก่ เทคนิคการจัดกลุ่มข้อมูลแบบเคมีน (K-means clustering algorithm)
- (2) ประเภทการแบ่งกลุ่มแบบไม่ชัดเจน (Soft Clustering) เทคนิคนี้มีการแบ่งข้อมูล แต่ละข้อมูลที่สามารถอยู่ได้ในหลายๆ กลุ่ม เทคนิคประเภทนี้ ได้แก่ เทคนิคการจัดกลุ่มแบบข้อมูลแบบลำดับขั้น (Hierarchical Clustering methods)

โดยการจัดกลุ่มข้อมูลแบบเคมีน (K-means clustering Algorithm) เป็นวิธีการที่ถูกนำมาใช้บ่อยมากที่สุด เนื่องจากมีขั้นตอนในการทำงานไม่ซับซ้อน และเข้าใจง่าย โดยมีขั้นตอนการทำงาน ดังนี้

- ขั้นตอนที่ 1 กำหนดค่า K หรือกำหนดจำนวนกลุ่มข้อมูลที่ต้องการ
- ขั้นตอนที่ 2 สุ่มวางตำแหน่งจุดศูนย์กลาง (Centroid) ของข้อมูลแต่ละกลุ่ม
- ขั้นตอนที่ 3 มีการคำนวณระยะทางจากจุดศูนย์กลางกับข้อมูลในชุดข้อมูล จากนั้นแต่ละข้อมูลจะถูกจัดอยู่ในกลุ่มของศูนย์กลางที่มีระยะทางใกล้ที่สุดเท่านั้น
- ขั้นตอนที่ 4 หลังจากทำการจัดกลุ่มข้อมูลใหม่ และทำการคำนวณค่าเฉลี่ยของข้อมูลในกลุ่ม ทำการกำหนดศูนย์กลางของข้อมูลใหม่
- ขั้นตอนที่ 5 ทำการซ้ำในขั้นขั้นตอน 3 และ 4 จนกระทั่งค่าจุดศูนย์กลางของกลุ่มข้อมูลใหม่ไม่เปลี่ยนแปลงหรือเปลี่ยนแปลงเพียงเล็กน้อย

2.2 งานวิจัยที่เกี่ยวข้อง

ในส่วนนี้จะมุ่งเน้นการอธิบายการประยุกต์ใช้ การจัดสรรตัวรีเคลแฝง (Latent Dirichlet Allocation) ในการวิเคราะห์หาค่าที่มีนัยสำคัญและหัวข้อที่ถูกซ่อนอยู่ในเอกสาร ซึ่งการจัดสรรตัวรีเคลแฝง (Latent Dirichlet Allocation) (3) เป็นแบบจำลองในการสร้างคลังข้อมูลด้วยความน่าจะเป็น แนวคิดพื้นฐานคือเอกสารประกอบด้วยผลรวมของหัวข้อแฝงแบบสุ่ม ซึ่งแต่ละหัวข้อมีลักษณะเฉพาะที่ถูกแจกแจงด้วยค่าต่างๆ โดยจากการศึกษางานวิจัยมีงานวิจัยที่เกี่ยวข้อง ดังนี้

งานวิจัยของ H. Sheikha (6) ได้เสนอการทำเหมืองข้อมูลเกี่ยวกับ covid-19 โดยมีการใช้อัลกอริทึม Latent Semantic Analysis (LSA) และ Latent Dirichlet Allocation (LDA) ในการลดมิติของข้อมูล ซึ่งจะถูกจัดกลุ่มโดยใช้อัลกอริทึม HDBSCAN และ K-Means โดยผลสรุปพบว่าจากรายงานไม่มีความสัมพันธ์ของข้อมูลโซเชียลมีเดียกับข้อมูล WHO โดยตรง แต่มีความสัมพันธ์กับการพาดหัวข่าวโดยการเปรียบเทียบค่าที่ใช้บ่อยของกลุ่มคำในวันเดียวกัน โดย HDBSCAN และ การจัดสรรตัวรีเคลแฝง (Latent Dirichlet Allocation) ให้ผลลัพธ์ที่ดีที่สุด

Kim, S.-W., & Gil, J.-M. (8) ได้เสนอระบบการจำแนกเอกสารการวิจัยที่มีความคล้ายคลึงกัน โดยใช้เทคนิคการจัดสรรตัวรีเคลแฝง (Latent Dirichlet Allocation) ในการดึงค่าที่มีนัยสำคัญจากบทคัดย่อ (Abstract) ของแต่ละบทความและตามหัวข้อ จากนั้นใช้อัลกอริทึมการจัดกลุ่ม โดยใช้ K-Means เพื่อใช้จำแนกบทความทั้งหมดที่เป็นรายงานการวิจัยที่มีหัวข้อคล้ายคลึงกัน โดยยึดตามค่า Term frequency-inverse document frequency (TF-IDF) ของในแต่ละบทความ โดยการประเมินผลจากค่า F-Score สูงสุดคือการทำ TFIDF-LDA_30 โดยมี 15 คำสำคัญ และ 15 หัวข้อ

Wexler, A., Davoudi, A., Weissenbacher, D., Choi, R., Oconnor, K., Cummings, H., & Gonzalez, G. (16) ได้เสนอการวิเคราะห์ฟอรัมการตั้งกระทู้ออนไลน์ เพื่อให้เข้าใจวิธีการหาข้อมูลของหญิงตั้งครรภ์ในชุมชนออนไลน์ โดยใช้การสร้างแบบจำลองหัวข้อ การจัดสรรตัวรีเคลแฝง (Latent Dirichlet Allocation) โดยอัลกอริทึมนี้สามารถค้นพบ ถุงคำ (Bag of words) ที่มีความน่าจะเป็นสูงที่จะปรากฏร่วมกัน โดยผลลัพธ์ในการจัดหมวดหมู่ ได้หัวข้อที่ใหญ่ที่สุด ได้แก่ สุขภาพมารดา (45%) หัวข้อเกี่ยวกับทารก (29%) และ คน/ความสัมพันธ์ (10%) และในการจัดหมวดหมู่ตามไตรมาสการตั้งครรภ์ ผลลัพธ์ในไตรมาสแรกหญิงตั้งครรภ์มีความกังวลในเรื่องการแท้งบุตร ไตรมาสที่สามมีความกังวลในการคลอดบุตร และกิจวัตรการนอนของทารกในช่วงหลังคลอด

Twinandilla, S., Adhy, S., Surarso, B., & Kusumaningrum, R. (14) ได้เสนอการสรุปเอกสารหลายฉบับเพื่อให้ผู้อ่านสามารถเข้าใจเจตนาของเอกสารข่าวออนไลน์ได้อย่างง่ายดาย

โดยการใช้เทคนิคการสร้างแบบจำลองหัวข้อ การจัดสรรดีรีเคลแฟง (Latent Dirichlet Allocation) ในการสรุปประโยคที่สำคัญในเอกสารหลายๆฉบับในภาพรวม โดยไม่ต้องจัดกลุ่มตามหัวข้อ และได้เสนอวิธีการสรุปแบบใหม่คือการรวม K-Means Clustering และ LDA - Significance Sentences เข้าด้วยกัน

Lim, David S. (9) ได้ทำการพัฒนาระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) สุขภาพจิต ชื่อว่า Woebot สำหรับคาดเดาผู้ใช้งานในเรื่องความรุนแรงของภาวะซึมเศร้า โดยป้อนข้อมูลไปยังตัวแยกประเภท (Classifier) 3 ประเภท ได้แก่ ข้อมูลการสำรวจพื้นฐาน หรือ เวกเตอร์ TFIDF จากข้อความทั้งหมด หรือเวกเตอร์การหาความน่าจะเป็นจากแบบจำลอง LDA โดยงานวิจัยนี้พบว่าหัวข้อที่สร้างโดย LDA มีประสิทธิภาพที่ดีสำหรับแสดงลักษณะอารมณ์จากข้อความ ซึ่งสามารถใช้เพื่อทำนายผลภาวะซึมเศร้า

Y. B. Touimi, A. Hadioui, N. E. Faddouli, and S. Bennani (13) บทความนี้กล่าวถึงระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) ในเชิงความหมาย สำหรับการวิเคราะห์ แลกเปลี่ยน แบ่งปันความรู้และประสบการณ์ของผู้เรียน โดยงานวิจัยนี้ได้รับข้อความจากผู้ใช้งานด้วยระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) จำแนกคำที่มีนัยสำคัญด้วยแบบจำลอง LDA ทำการ mapping ไปยังโดเมน ontology ของ MOOC การประยุกต์ใช้แบบจำลองความน่าจะเป็นของ LDA ทำให้สามารถดึงความรู้ที่เกี่ยวข้องตามที่ผู้เรียนร้องขอได้

Tian, Kochhar (12) แหล่งชุมชนในการถามตอบคำถาม (Community Question Answering: CQA) กำลังกลายเป็นแหล่งข้อมูลที่สำคัญมาก ซึ่งผู้ใช้สามารถแบ่งปันความรู้ในหัวข้อต่างๆ ระหว่างกันได้ แม้ว่าแพลตฟอร์มสามารถช่วยเหลือผู้ใช้งาน แต่ก็ยังก่อให้เกิดความท้าทายมากมายกับขนาดของชุมชนที่เติบโตขึ้นเรื่อยๆ จำนวนคำถามที่โพสต์ทุกวันเป็นสาเหตุให้เกิดปัญหาในการค้นหาคำถามแก่ผู้ใช้ที่เหมาะสม ในบทความนี้ได้เสนอแนวทางในการทำนายคำตอบที่ดีที่สุดสำหรับคำถามใหม่บน CQA แนวทางของงานวิจัยนี้ได้พิจารณาทั้งความสนใจของผู้ใช้และความเชี่ยวชาญของผู้ใช้ที่เกี่ยวข้องกับหัวข้อของคำถามที่กำหนด ความสนใจของผู้ใช้ในหัวข้อต่างๆ สามารถเรียนรู้ได้จากการใช้การสร้างแบบจำลองหัวข้อ (Topic modeling) กับคำถามก่อนหน้านี้ที่ตอบโดยผู้ใช้ ในขณะที่ความเชี่ยวชาญของผู้ใช้จะได้รับการเรียนรู้โดยใช้ประโยชน์จากกลไกการลงคะแนนร่วมกันใน CQA เราได้ใช้โมเดลของเรากับชุดข้อมูลที่ดึงมาจาก Stack Overflow ซึ่งเป็นหนึ่งใน CQA ที่ใหญ่ที่สุด ผลการศึกษาแสดงให้เห็นว่าแนวทางของเรามีประสิทธิภาพดีกว่าแนวทางที่ใช้ TF-IDF โดยงานวิจัยนี้ประกอบด้วย 4 ส่วน ได้แก่ การจัดเตรียมข้อมูล (Data preprocessing) การสร้างโปรไฟล์ผู้ใช้งาน (User profile building) ความสนใจ

เฉพาะของผู้ใช้และการเรียนรู้แบบเชี่ยวชาญ (User topical interest and Expertise learning) และการสร้างแบบจำลองการจัดอันดับ (Ranking model building.) โดยเริ่มต้นจากการสร้างโปรไฟล์ของผู้ใช้งานจากคำถามที่เคยถามก่อนหน้านี้ ประกอบด้วยข้อความและข้อมูลการลงคะแนน (Voting information) ซึ่งจะใช้ในการเรียนรู้หัวข้อตลอดจนความสนใจของผู้ใช้และความเชี่ยวชาญของผู้ใช้ในแต่ละหัวข้อที่เรียนรู้ งานวิจัยนี้ได้เสนอแบบจำลองการจัดอันดับ เพื่อที่จะคำนวณหาความน่าจะเป็นที่ผู้ใช้จะเป็นผู้ที่ตอบคำถามได้ดีที่สุด ผู้ใช้มีความสนใจและเชี่ยวชาญในหัวข้อต่างๆ จะถูกบันทึกลงในแบบจำลองการจัดอันดับ โดยหลักของงานวิจัยนี้คือการจัดอันดับผู้ใช้ โดยใช้ค่าความน่าจะเป็นและค้นหาผู้ใช้ TOP-K ที่มีโอกาสสูงในการให้คำตอบที่ดีที่สุด

Zhang, Wu (17) บทความนี้ได้เสนอวิธี supervised question-answer topic modeling approach สำหรับการดึงข้อมูลใน community question answering (CQA) โดยมีการตั้งสมมติฐานพื้นฐานคือ แม้ว่าคำถามและคำตอบจะมีความแตกต่างกันในหลายแง่มุม แต่พวกเขามีการแบ่งปันปัจจัยแฝงบางอย่างที่แสดงให้เห็นถึงหัวข้อที่ผู้ถามและผู้ตอบต้องการ ยิ่งไปกว่านั้นยังมีตัวบ่งชี้คุณภาพของคำตอบ จำลองวิธีที่ผู้ใช้งานคนอื่นประเมินคำตอบที่เกี่ยวข้องกับคำถามในคู่ของคำถาม-คำตอบ กำหนดให้ผู้ถาม และผู้ตอบตามหัวข้อทั่วไปในแบบจำลองได้ดีเพียงใด ด้วยคำแนะนำตัวอย่างของคุณภาพ คือคำถาม และคำตอบจะถูกจับคู่กับ latent space หรือ topic space และความคล้ายคลึงของคำถามสามารถวัดได้ด้วยความช่วยเหลือของข้อมูลในคำตอบที่ถูกย่อใน space ด้วย topic space และได้เพิ่มเติมการจับคู่คำถามใน Latent space ให้เป็น LMIR แบบดั้งเดิมและเสนอรูปแบบ Topic-based language model และยังสามารถใช้ข้อมูลจาก Metadata มากมายที่เกี่ยวข้องกับคำถาม และเรียนรู้คุณภาพของคำตอบโดยอัตโนมัติ โดยการใช้ คำตอบที่ดีที่สุด เป็นมาตรฐานและดึงคุณลักษณะต่างๆ จากทั้งข้อความคำตอบ และข้อมูล Metadata ที่เกี่ยวข้อง ด้วยข้อมูลการฝึกฝน และคุณลักษณะ เรียนรู้ฟังก์ชันคะแนนเพื่อสร้างคำตอบที่มีคุณภาพ และปล่อยให้ควบคุมดูแลการเรียนรู้ของหัวข้อ แบบจำลองนี้เป็นนอกจากจะนำข้อมูลขนาดใหญ่มาใช้ได้เป็นประโยชน์ได้เท่านั้น ยังสามารถช่วยประหยัดแรงงานของผู้สังเกตอีกด้วย บทความนี้ได้ทำการทดลองกับข้อมูลขนาดใหญ่จาก Yahoo! Answers และ Baidu Knows ผลการทดลองแสดงให้เห็นว่าแบบจำลองที่นำเสนอสามารถมีประสิทธิภาพเหนือกว่าโมเดลการดึงข้อมูลที่ทันสมัยใน CQA อย่างมีนัยสำคัญ

จากตารางที่ 1 ตัวอย่างการเปรียบเทียบระหว่าง TBLMSQATM กับ LMIR คำถามคือ “What is the eye doctor looking for when he/she dilates my eyes?” ซึ่งเกี่ยวข้องกับข้อความ (Relevant) “what does a doctor look for when they dilate your eyes” โดยคำว่า แพทย์ ในข้อความไม่ตรงกับ ดร. และคำถามที่ไม่ตรงประเด็น “When doctors dilate your eyes, what do they use?” โดย LMIR วัดความคล้ายคลึงกันของคำถามโดยพิจารณาจากคำที่พบบ่อยเท่านั้น ดังนั้นจึงจัดอันดับรายการที่ไม่เกี่ยวข้องสูงกว่าที่เกี่ยวข้อง ส่วนผลของงานโมเดลของงานวิจัยนี้คือ TBLMSQATM สามารถจัดกลุ่มคำที่เกี่ยวข้องกับความหมาย เช่น แพทย์ ดร. พยาบาล อายุรกรรม และ มะเร็ง เป็นต้น จัดรวมกันใน Latent topic space และสามารถรู้ว่ามี ความคล้ายคลึงระหว่างคำถามที่เกี่ยวข้องได้สำเร็จ

ตาราง 1 ตัวอย่างการเปรียบเทียบระหว่าง TBLMSQATM กับ LMIR

Query	what does a doctor look for when they dilate your eyes	
Label	Relevant	Irrelevant
Question	What is the eye dr. looking for when he/she dilates my eyes?	When doctors dilate your eyes, what do they use?
LMIR	Rank3	Rank2
TBLMSQATM	Rank2	Rank3

Bahng and Lee (2) ในบทความนี้ได้้นำ Latent Dirichlet Allocation (LDA) มาประยุกต์ใช้กับคำถามที่ผู้ป่วยการได้ยินหรือมีความเสี่ยงต่อการสูญเสียการได้ยิน แลกเปลี่ยนด้านสุขภาพทางอินเทอร์เน็ต ในแหล่งสังคม question-and-answer (Q&A) เพื่อระบุการรับรู้ ความกังวล และความต้องการของผู้ป่วยเกี่ยวกับการสูญเสียการได้ยิน ผลการวิจัยได้เปิดเผยหัวข้อ 21 หัวข้อที่เป็นตัวแทน มีความหมายและส่วนใหญ่สอดคล้องกับหัวข้อย่อยที่ได้กำหนดขึ้น ในงานวิจัยด้านวิทยาศาสตร์การได้ยิน หัวข้อแฝง (Latent topics) แบ่งออกเป็น 5 หัวข้อ ได้แก่ การสูญเสียการได้ยินแบบกะทันหัน หูอื้อ สูญเสียการได้ยินจากเสียง เครื่องช่วยฟัง เวียนศีรษะ ความอยากรู้เกี่ยวกับการสูญเสียการได้ยิน หูชั้นกลางอักเสบ และ โรคแทรกซ้อน การวิเคราะห์หัวข้อคำถามของ

ผู้ป่วยในหัวข้อการได้ยิน ช่วยให้เข้าใจมุมมองของผู้ป่วยมากขึ้น และนำไปสู่การพัฒนาการที่ดีขึ้นของการช่วยเหลือโดยมีผู้ป่วยเป็นศูนย์กลาง

ตาราง 2 เป็นตัวอย่าง Keyword ในหัวข้อของ LDA กับ 5 หัวข้อ

Topic No.	Topic Label	Keywords	Theme Category
1	Tinnitus Symptoms	ringing, buzzing, tinnitus, sound, hear, constant, continue, hissing, severe, ear	Tinnitus
2	Loud Music	loud, music, earphone, listen, eardrum, affect, damage, cause, headset, longtime	Noise-induced Hearing Loss
3	Steroids for Sudden Deafness	steroid, injection, treatment, drug, pill, prescription, shot, sudden, deafness, effect	Sudden Hearing Loss
4	Symptoms of Otitis Media	symptoms, otitis media, pain, headache, sore throat, infection, deafen, today, morning	Otitis Media
5	Hearing Test Results	diagnosis, treatment, cure, curious, communication, hospital, learning, disability, quality, doubt	General Inquiry

ผลการแยกหัวข้อโดย LDA เกี่ยวกับคำถามแรกเผยให้เห็นมิติที่กันกับที่ระบุไว้ในเอกสารเกี่ยวกับการสูญเสียการได้ยิน โดย 21 หัวข้อเกิดจากการคำนวณจาก LDA แสดงให้เห็นใน ตารางที่ 2 มีรายการของคำที่ประกอบเป็นหัวข้อเฉพาะ มีหัวข้อที่ถูกกำหนดไว้ และธีมของข้อมูล พบว่าธีมที่ได้นั้นคล้ายกับการแบ่งประเภทของการวินิจฉัยการสูญเสียการได้ยินและการรักษา แต่ยังคงครอบคลุมประเด็นในหัวข้อที่เกี่ยวข้องในหัวข้อสูญเสียการได้ยิน

Dinakar, Jones (5) งานวิจัยนี้กล่าวถึงการแก้ไขปัญหการกลั่นแกล้งในวัยรุ่นบนเครือข่ายสังคมทางอินเทอร์เน็ต ประกอบด้วย 2 ส่วน ได้แก่ การตรวจจับการกลั่นแกล้งได้โดยอัตโนมัติและกลยุทธ์ในการโต้ตอบที่กระตุ้นให้เกิดการไตร่ตรองและการสนับสนุนทางอารมณ์ โดยสิ่งสำคัญทำให้วัยรุ่นที่กำลังตกอยู่ในสถานการณ์ที่โดนกลั่นแกล้ง ว่าพวกเขานั้นไม่ได้ต่อสู้เพียงลำพัง โดยในงานวิจัยนี้ได้ทดสอบกับคลัง (Corpus) เรื่องราวจากวัยรุ่นที่มีปัญหาจากเครือข่ายสังคม (Social network) 5,500 เรื่อง งานวิจัยนี้ได้นำ LDA และ ใช้คนในการตีความผลลัพธ์ที่ได้ โดยการใช้หลักการภาษาศาสตร์สังคม เพื่อดึงประเด็นที่มีนัยสำคัญในเรื่องราวและใช้เพื่อจับคู่กับเรื่องราวใหม่ที่คล้ายกัน โดยผลประเมินโดยผู้ใช้งานเกี่ยวกับการจับคู่เรื่องราวแสดงให้เห็นว่าการดึงข้อมูลตามนัยสำคัญช่วยให้ค้นหาเรื่องราวที่เกี่ยวข้องได้ดีขึ้นและเรื่องราวที่มีประสิทธิภาพสำหรับแอปพลิเคชันนี้ดีกว่าวิธีการทั่วไป

ในตารางที่ 3 แสดงตัวอย่างจากการศึกษา เรื่องราวใหม่ (New story) และเป็นการจับคู่ที่ดีที่สุดโดยจากการวัด KL divergence แบบไม่สมมาตร จากเรื่องราวเป็นการกลั่นแกล้งที่เกี่ยวข้องกับลักษณะทางกายภาพของผู้หญิงตัวม ในคลังข้อมูลของ “High school & college drama” ที่เกี่ยวข้องอย่างยิ่ง

ตาราง 3 แสดงตัวอย่างจากการศึกษา เรื่องราวใหม่ (New story) และเป็นการจับคู่ที่ดีที่สุด

New story	Matched story
I like wearing makeup. I dont want to be teased for my weight. Some girls hurt me that no one will like me at all. I don't want to be called names about my looks!	Growing up I havent been the skinnest girls. When I first started Middle School people would always make fun of my wieght. Because of that I would always tell my self that I was fat and ugly and i really took it to heart. And it still follows me.

จากงานวิจัยที่ได้กล่าวมานั้นเป็น การนำเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มาใช้ในการทำนายหัวข้อ ลมิตีของข้อมูลและการหาค่าที่มีนัยสำคัญในเอกสารที่มีจำนวนมาก มีการนำไปประยุกต์ใช้กับการจัดกลุ่มเพื่อเพิ่มประสิทธิภาพในการจำแนกหมวดหมู่หัวข้อ ในด้านหาข้อเท็จจริงของข่าวสาร การจำแนกเอกสาร และการจัดหมวดหมู่ของเอกสาร รวม

ไปถึงการนำการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มาประยุกต์ใช้กับ ระบบหุ่นยนต์
โต้ตอบการสนทนา (Chatbot) ในการวิเคราะห์หาคำที่มีนัยสำคัญในข้อความ และแหล่งชุมชนใน
การถามตอบคำถาม (Community Question Answering: CQA) เป็นแหล่งข้อมูลที่สำคัญและมี
ขนาดใหญ่ ในการแบ่งปันข้อมูลต่างๆ วมกัน อาทิ StackOverflow, เว็บไซต์ในการแลกเปลี่ยน
ปัญหาสุขภาพต่างๆ, Yahoo! Answers Baidu Knows และเครือข่ายสังคมต่างๆ เช่น ทวิตเตอร์
(Twitter) และเฟสบุ๊ก (Facebook) โดยโดเมนเกี่ยวกับ การกลั่นแกล้ง ความวิตกกังวล และปัญหา
สุขภาพ ในแต่ละงานวิจัยได้มีการจัดกลุ่มของคำถามออกมาเป็นกลุ่มๆ และจับคู่ข้อความเก่าและ
ข้อความใหม่ การหาหัวข้อในข้อความ และมุ่งเน้นพัฒนาการถามตอบคำถามให้ดีขึ้นเพื่อ
ประโยชน์ต่อผู้ที่ต้องการรับความรู้



บทที่ 3

วิธีดำเนินการวิจัย

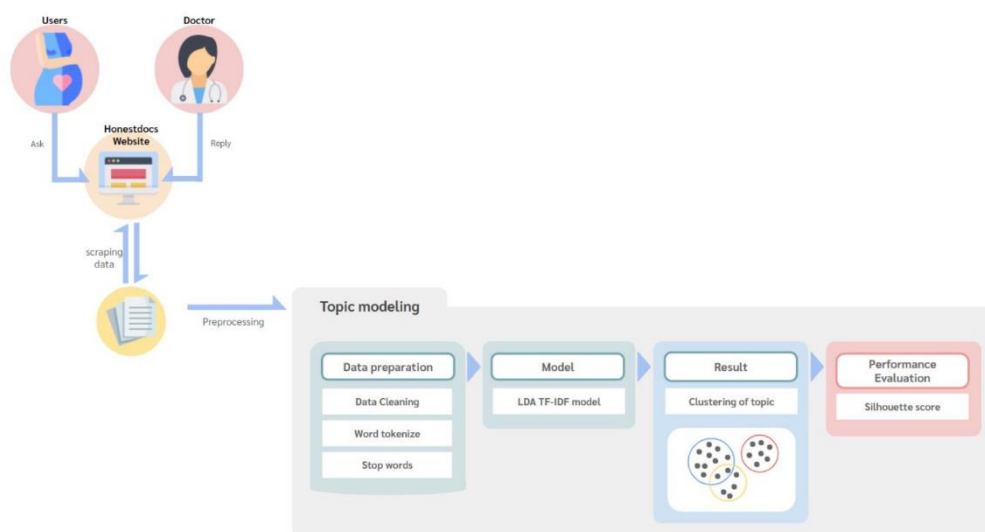
ในบทนี้จะกล่าวถึง ขั้นตอนในการวิจัยการพัฒนาการสร้างแบบจำลองหัวข้อ (Topic Modeling) การวิเคราะห์คำที่มีนัยสำคัญ และการจำแนกหัวข้อในคำปรึกษาปัญหาเกี่ยวกับการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดที่เป็นข้อความภาษาไทย มีขั้นตอนการดำเนินงานวิจัยดังนี้

1. การออกแบบโครงสร้างงานวิจัย
2. การเตรียมชุดข้อมูล (Dataset)
3. การสร้างแบบจำลอง (Model)
4. การประเมินผล (Evaluation)

3.1 การออกแบบโครงสร้างงานวิจัย

การออกแบบโครงสร้างงานวิจัย ประกอบไปด้วยกระบวนการหลัก ได้แก่

- 3.1.1 การดึงข้อมูล (Scraping data)
- 3.1.2 การเตรียมข้อมูล (Data preparation)
- 3.1.3 แบบจำลอง (Model)
- 3.1.4 ผลงานวิจัย (Result)
- 3.1.5 การประเมินประสิทธิภาพของโมเดล (Performance Evaluation)



ภาพประกอบ 1 โครงสร้างงานวิจัย

ตามภาพประกอบ 1 โดยกระบวนการแรกของการดำเนินงานวิจัยคือการดึงข้อมูล (Scraping data) จากเว็บไซต์ที่มีผู้ใช้งานที่มีปัญหาในหมวดหมู่การตั้งครรภ์เข้ามาปรึกษากับทางทีมแพทย์ผู้เชี่ยวชาญ ผ่านเว็บไซต์ Honestdocs เมื่อทำการดึงข้อมูลแล้ว ข้อมูลจะอยู่ในรูปแบบของไฟล์ Comma Separated Values (CSV) หรือ Text File นำไฟล์ที่ได้มาเข้าสู่กระบวนการจัดเตรียมข้อมูล (Data preparation) โดยมีขั้นตอน ดังนี้

- (1) การทำความสะอาดข้อมูล (Data cleaning)
- (2) ทำการตัดคำ (Word tokenize)
- (3) ตัดกลุ่มคำที่ไม่สื่อความหมาย (stop words)

ต่อมาเข้าสู่กระบวนการนำข้อมูลมาประมวลผลในโมเดล Latent Dirichlet Allocation (LDA) ผลลัพธ์ที่ได้จะเป็นการจัดกลุ่มของหัวข้อคำถาม (Clustering of topic) และประเมินประสิทธิภาพ (Performance evaluation) โดยการใช้ Silhouette score

3.2 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้ในงานวิจัยการวิเคราะห์แบบจำลองหัวข้อเกี่ยวกับการตั้งครรภ์นั้น ได้แก่ Google Colab และ โปรแกรมภาษา Python

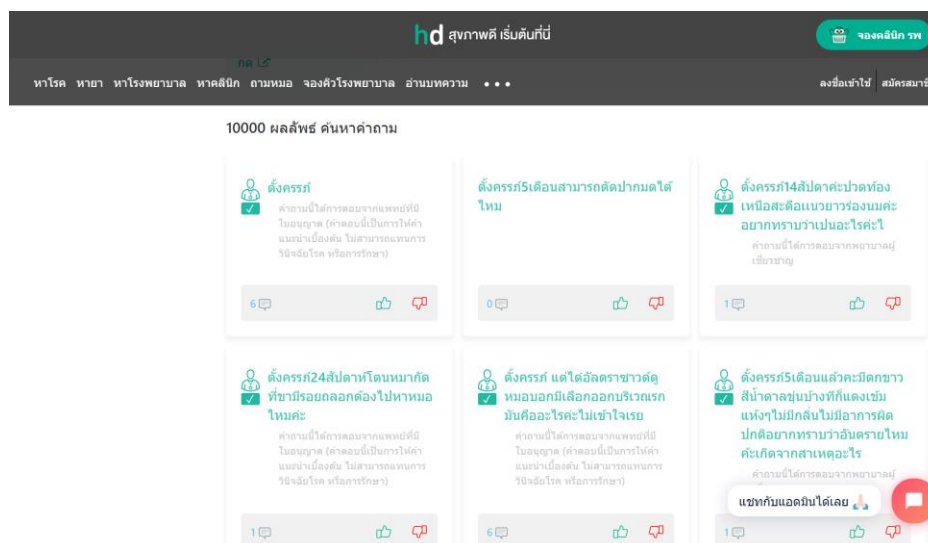
1. Google Colab (Google Colaboratory) เป็นบริการ Software as a service เป็นโปรแกรม Jupyter Notebook บน Cloud จาก Google ไม่จำเป็นต้องติดตั้งโปรแกรมอื่นๆ ก่อนใช้งาน มี GPU และ TPU ให้ได้ใช้ฟรี ทีละ 12 ชั่วโมง ในการรันภาษา Python เพื่อใช้ในการทำงานด้าน Data science

2. โปรแกรมภาษา Python เป็นภาษาทางคอมพิวเตอร์ชนิดหนึ่งที่ใช้ในการเขียนโปรแกรม โดย Python ถูกออกแบบมาให้เน้นการอ่าน code ได้ง่ายและให้มีความคล้ายคลึงกับภาษาอังกฤษ มีรูปแบบการเขียน Syntax ที่ช่วยให้นักพัฒนาสามารถเขียนโปรแกรมได้ง่ายด้วยจำนวนบรรทัดการเขียนน้อยกว่าภาษาทางคอมพิวเตอร์ภาษาอื่นๆ และทำงานในรูปแบบ object-oriented

3.3 ชุดข้อมูล (Dataset)

3.2.1 การเก็บรวบรวมข้อมูล

ผู้วิจัยนำเข้าข้อมูลเกี่ยวกับคำปรึกษาปัญหาการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดภาษาไทยสำหรับการทำวิจัยในครั้งนี้ โดยใช้ชุดข้อมูลในช่วงวันที่ 23 มกราคม 2017 ถึง 15 ตุลาคม 2020 จำนวน ข้อมูลทั้งหมดมี 9,987 ข้อความ ในการเก็บรวบรวมข้อมูลได้ใช้ Selenium ในการทำ Web Scraping ข้อมูล ดึงข้อมูลจากเว็บไซต์ Honestdocs จากภาพประกอบที่ 2 ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรภ์ บันทึกข้อมูลในรูปแบบของ ไฟล์นามสกุล csv



ภาพประกอบ 2 เว็บไซต์ honestdocs ในหมวดหมู่คำปรึกษาปัญหาของการตั้งครรภ์

QA_pregnancy_v1						
File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago						
100% Arial 10						
E44 - โรงพยาบาลสุรินทร์ผาครกอยู่แล้วละ ไม่วางเป็นอายุครรภ์เท่าไรก็ตาม						
	A	B	C	D	E	F
167	ถ้าหญิงที่ตั้งครรภ์ มีภาวะปวดเป็น ๑๒ ชั่วโมงปรึกษาหาอาการอะไร	https://hd.co.th/ask/53445	April 06, 2017 11:31	HonestDocs User	หาอาการได้เหมือนกันคุณแม่ทั่วไปครับ ไม่มีข้อห้าม	
168	ทำอย่างไรให้ลูกแข็งแรง	https://hd.co.th/ask/22335	April 06, 2017 21:33	HonestDocs User	หมั่นออกกำลังกายสม่ำเสมอ รับประทานอาหารที่มีประโยชน์ งดสูบบุหรี่ งดดื่มแอลกอฮอล์	
169	ผู้หญิงจะท้องได้ตั้งแต่อายุเท่าไร เพราะตอนนี้ฉันเพิ่ง 30 ปี	https://hd.co.th/ask/42434	February 25, 2017 16:4	HonestDocs User	ช่วงเวลาที่ผู้หญิงสามารถตั้งครรภ์ได้คือ 20-35 ปีค่ะ แต่	
170	คุณแม่ตั้งครรภ์ทานยาคุมกำเนิดได้ไหม	https://hd.co.th/ask/28901	March 05, 2017 13:37	HonestDocs User	คุณแม่ตั้งครรภ์ไม่ควรทานยาคุมกำเนิด เพราะอาจมีผลต่อทารกในครรภ์ได้	
171	อยากทราบค่าคลอดในโรงพยาบาลแล้วได้ไหม แล้วอีก	https://hd.co.th/ask/15584	February 07, 2017 11:1	HonestDocs User	แนะนำคุณแม่ที่ตั้งครรภ์ให้ไปพบสูติศาสตร์ก่อนคลอด	
172	งาทำอย่างไรกับเชื้อที่อยู่ในช่องคลอดแล้วตั้งครรภ์	https://hd.co.th/ask/31828	February 07, 2017 13:5	HonestDocs User	แล้วแต่ขนาดค่ะ ถ้าขนาดเล็กอาจไม่ส่งผลต่อการตั้งครรภ์	
173	ตั้งครรภ์ตอนอายุ 39 ปี เสี่ยงต่อการคลอดยากไหม	https://hd.co.th/ask/34950	February 05, 2017 22:2	HonestDocs User	การตั้งครรภ์ในช่วงอายุที่มากกว่า 35 ปี อาจเสี่ยงต่อ	
174	ถ้าคุณแม่ตั้งครรภ์แต่ยังไม่ถึงกำหนดคลอดแล้วคลอดไหม	https://hd.co.th/ask/51659	February 03, 2017 16:4	HonestDocs User	ก็อาจคลอดก่อนกำหนดได้ค่ะ ถ้าคุณแม่ตั้งครรภ์ก่อนกำหนด	
175	ตรวจสวนครั้งแรกแล้ว ยังแบบนี้ท้องไหม	https://hd.co.th/ask/22520	November 01, 2017 13:	HonestDocs User 876888	รูปและบันทึกของแพทย์เป็นสิ่งจำเป็นสำหรับแพทย์ผู้ตรวจ	
176				HonestDocs User 874704		

ภาพประกอบ 3 ตัวอย่างข้อมูล

จากภาพประกอบที่ 3 ประกอบด้วยข้อมูล ข้อความคำปรึกษา ลิงค์เชื่อมโยงไปยังเว็บไซต์ คำปรึกษา วันที่ในการที่ผู้ใช้ขอคำปรึกษากับแพทย์ รหัสผู้ใช้งาน และคำตอบจากแพทย์ โดยมีข้อมูลทั้งหมดมี 9,987 ข้อความ ซึ่งบางคำปรึกษาจากผู้ใช้งานไม่มีคำตอบจากแพทย์ผู้เชี่ยวชาญ

3.2.2 การจัดเตรียมข้อมูล (Data preparation)

ในการจัดเตรียมข้อมูลได้ใช้ ภาษาไพทอน (Python) ในการทำความสะอาดข้อมูล (Text cleaning) ตัดประโยคที่ไม่เกี่ยวข้องกับข้อความออกจากข้อความที่เกิดจากการดึงข้อมูลและตัดตัวอักษรพิเศษ (Special characters) ในขั้นตอนแรกเริ่มต้นด้วยการนำเข้าโมดูลสำคัญสำหรับการสร้างแบบจำลอง ดังภาพที่ 4 ต่อมานำเข้าไฟล์ด้วยตัวอย่างที่ 5

```
import requests
from bs4 import BeautifulSoup
from time import time, sleep
from random import randint

# from googletrans import Translator

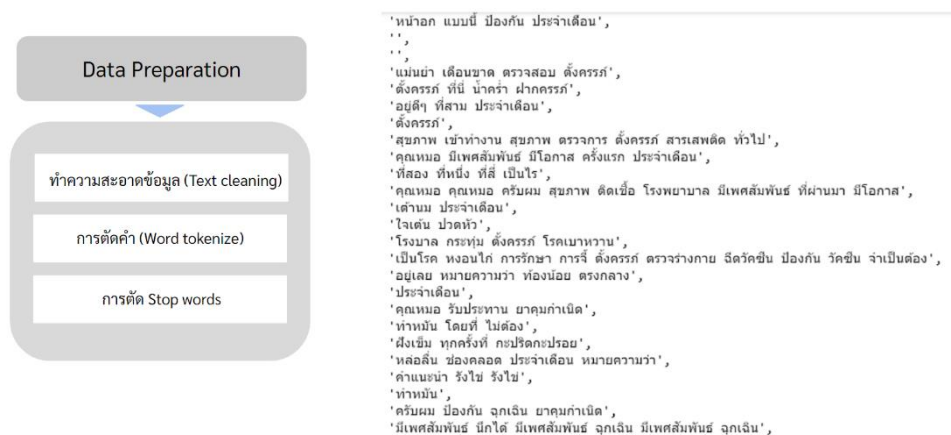
from tqdm import tqdm
from tqdm import tqdm_notebook
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from nltk import sent_tokenize, word_tokenize
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from nltk.sentiment.util import mark_negation
```

ภาพประกอบ 4 แสดงตัวอย่างโค้ดนำเข้าโมดูลสำคัญที่ใช้สำหรับสร้างแบบจำลอง

	question	href	date_ask	user	answer
0	ตั้งครรภ์	https://hd.co.th/ask/27734160	October 01, 2019 17:30	\nHonestDocs User 1082087ก	เราได้รับคำถามของคุณแล้ว รอทีมแพทย์มาตอบคำถามส...
1	ตั้งครรภ์5เดือนสามารถคลอดได้ไหม	https://hd.co.th/ask/51006398	June 30, 2018 14:49	\nHonestDocs User 912491ก	NaN
2	ตั้งครรภ์14สัปดาห์ปวดท้องเหมือนจะคลอดแล้ว...	https://hd.co.th/ask/4362714	August 31, 2018 15:07	\nHonestDocs User 929536ก	NaN
3	ตั้งครรภ์24สัปดาห์โดนหมากัดที่ขาบริเวณข้อศอก...	https://hd.co.th/ask/33971594	November 07, 2018 18:58	\nHonestDocs User 954767ก	รักซิ่นพิษสุนัขบ้าและอัมพาตในไขสันหลังสามารถไ...
4	ตั้งครรภ์ แลได้ฉีดสารชาวดูเหมือนออกน้เลือดออก...	https://hd.co.th/ask/48478898	August 06, 2018 1:35	\nHonestDocs User 921834ก	รบกวนขอข้อมูลเพิ่มเติมด้วยครับ คุณแม่อายุเท่า...

ภาพประกอบ 5 แสดงตัวอย่างโค้ดนำเข้าไฟล์ข้อมูลและข้อมูลที่ใช้สำหรับสร้างแบบจำลอง



ภาพประกอบ 6 กระบวนการเตรียมข้อมูลและผลลัพธ์

จากภาพประกอบที่ 6 คือกระบวนการเตรียมข้อมูล การทำความสะอาดข้อมูล การตัดคำ (Word tokenize) นำข้อความทั้งหมดเข้าสู่กระบวนการตัดคำ โดยใช้ชุดคำสั่ง newmm และตัด Stop words ในการกำจัดคำที่ไม่จำเป็นหรือไม่เกี่ยวข้องออกจากข้อความ โดยใช้ชุดคำสั่งของ PythaiNLP โดยหลังจากการ เตรียมข้อมูล ข้อมูลคงเหลือ 8,248 ข้อความ

3.4 แบบจำลอง (Model)

3.4.1 การสกัดใจความสำคัญของข้อความ (Vectorizer)

จากภาพประกอบที่ 7 เป็นกระบวนการสกัดใจความสำคัญจากข้อความ โดยกระบวนการนี้ใช้ชุดโปรแกรม Sklearn ชื่อว่า Tfidfvectorizer ในการสกัดคุณลักษณะ (Feature extraction) เพื่อหาความถี่ของคำที่มีลักษณะเฉพาะในแต่ละเอกสาร ช่วยในการเลือกคำที่มีนัยสำคัญในเอกสาร ผลลัพธ์ที่ได้จากการสกัดคุณลักษณะจากข้อความทั้งหมด 8,248 ข้อความ โดยมีการกำหนดตัวแปร $\text{max_df} = 0.5$ คือละเว้นคำที่ปรากฏในเอกสารมากกว่า 50% และ $\text{min_df} = 10$ คือละเว้นคำที่ปรากฏในเอกสารน้อยกว่า 10 ฉบับ โดยมีการนับจำนวนคำที่มีนัยสำคัญในแต่ละเอกสารทั้งหมดและนับจำนวนคำที่มีลักษณะเฉพาะ (Unique Words)



ภาพประกอบ 7 กระบวนการสกัดใจความสำคัญของข้อความและผลลัพธ์

3.4.2 ถู่งคำ (Bag of word)

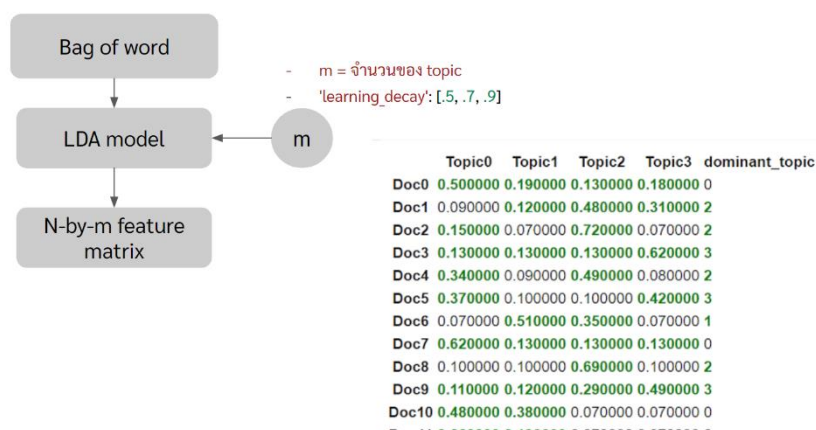
เมื่อเข้าสู่กระบวนการ Tfidfvectorizer จะได้ถู่งคำ (Bag of word) ที่มีการนำความถี่ของคำที่มีลักษณะเฉพาะ (Unique Words) ที่ปรากฏในแต่ละเอกสาร โดยจำนวนคุณลักษณะ (Feature) จะขึ้นอยู่กับจำนวนของคำที่มีลักษณะเฉพาะ (Unique Words) โดยตัวอย่างของ ถู่งคำ (Bag of word) ประกอบด้วย รหัสของเอกสาร รหัสคำที่ปรากฏในเอกสาร และความถี่ของการ

ได้แก่ ประจำเดือน, ตั้งครรภ์, มีเพศสัมพันธ์, มีโอกาส, จุกเงิน, คุณหมอม, ฤกษ์งาม, บ๊องกัน, ยาคุมกำเนิด และครั้งแรก เป็นต้น โดยเมื่อพิจารณาจากการจับกลุ่มคำ จากภาพที่ 1 ผู้วิจัยได้ทำการพิจารณาและแบ่งกลุ่มของหัวข้อเป็น 4 หัวข้อดังนี้

- (1) อาการผิดปกติในขณะตั้งครรภ์
- (2) ความผิดปกติของประจำเดือน
- (3) ความกังวลเมื่อมีเพศสัมพันธ์
- (4) การคุมกำเนิด

3.4.3 แบบจำลองการจัดสรรตรีเคลแฝง (Latent Dirichlet Allocation)

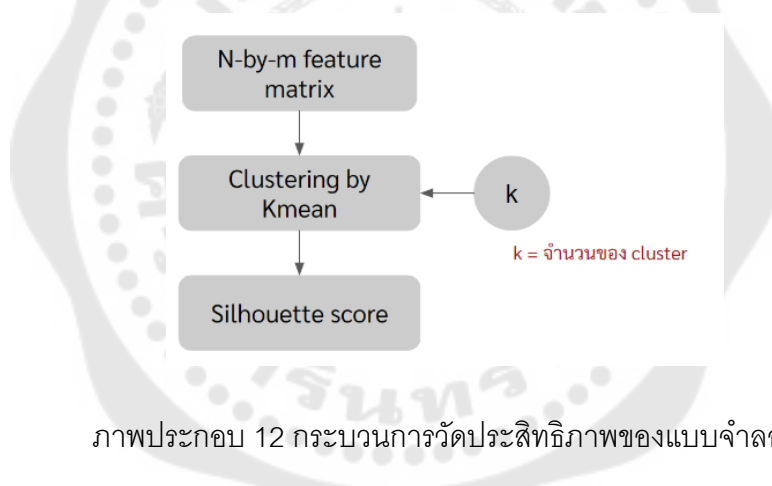
เทคนิคการจัดสรรตรีเคลแฝง (Latent Dirichlet Allocation) ใช้เทคนิคนี้เพื่อต้องการจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน โดยการลดมิติของข้อมูล (Dimensionality-reduction) แนวคิดของเทคนิคนี้คือการกำหนดค่า m (จำนวนหัวข้อ) ไปยังในแต่ละ N (เอกสาร) สามารถหาเวกเตอร์ความยาวของค่า m ซึ่งแทนค่าด้วยการ กระจายความน่าจะเป็นของหัวข้อในแต่ละเอกสาร เพิ่มเวกเตอร์ไปยังเอกสารทั้งหมดจะได้ เมทริกซ์คุณลักษณะ N -by- m และมีการ Label หัวข้อในแต่ละเอกสาร ดังในภาพประกอบที่ 11



ภาพประกอบ 11 กระบวนการลดมิติของข้อมูลและผลลัพธ์

3.5 การวัดประสิทธิภาพ

แนวคิดในการประเมินประสิทธิภาพของงานวิจัยนี้ เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) จากภาพประกอบที่ 12 เป็นการสกัด feature ออกมาจาก bag of words (TF-IDF) โดยแต่ละ feature คือ คำที่กระจายตัวอยู่ในบางหัวข้อที่เป็นนามธรรม ซึ่ง เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มีการวัดผลแบบ Intrinsic evaluation คือการใช้ metric ที่อยู่ในแบบจำลองเอง โดย เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) มีการวัดผลโดยใช้ Perplexity หรือ ค่าความสับสนเพื่อดูการกระจายตัวของความน่าจะเป็นว่ามั่นคงมากน้อยเพียงใด แต่ Intrinsic Evaluation จะมีข้อเสียที่ว่า ในความจริงแบบจำลองที่ได้นั้น มีประสิทธิภาพที่ดีกับงานจริงหรือไม่ เพราะว่าแบบจำลองถูกตั้งสมมุติฐานว่า ข้อมูลจริงหรือข้อมูลทดสอบ จะเหมือนกับชุดข้อมูลฝึกฝน ดังนั้นเราจึงใช้ K-mean ที่เป็นการวัดผลแบบ Extrinsic evaluation ด้วยการดูว่า vector ที่ได้จาก LDA มันให้ผลลัพธ์ได้ดีแค่ไหนใน downstream task



การประเมินประสิทธิภาพในการจำแนกกลุ่มข้อความ (Clustering) ใช้การประเมินประสิทธิภาพโดยการ วัดค่าในกลุ่มเดียวกันของสมาชิกในกลุ่มนั้นๆ และค่าความเหมือนระหว่างกลุ่มมากน้อยเพียงใด การคำนวณค่าได้ใช้ Silhouette Coefficient โดยชุดโปรแกรม Sklearn ที่ชื่อว่า silhouette_score ดังสมการ

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

โดย a คือ ระยะทางเฉลี่ยระหว่างจุดในกลุ่มเดียวกัน

b คือ ระยะทางเฉลี่ยระหว่างจุดระหว่างกลุ่ม

i คือ จุดของข้อมูล

ค่า Silhouette score $S(i)$ มีค่าอยู่ในช่วง $[-1, 1]$ ถ้ามีค่าใกล้ 1 มากคือดี



บทที่ 4

ผลการดำเนินงานวิจัย

งานวิจัยนี้ได้ทำการวิเคราะห์หาค่าที่มีนัยสำคัญและการจำแนกกลุ่มหัวข้อในคำปรึกษาปัญหาการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดในข้อความภาษาไทย โดยใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) พัฒนาด้วยภาษาไพทอน (Python) และเครื่องมือ Google Colab มีการรวบรวมคำปรึกษาทั้งหมด 8,248 ข้อความ โดยมีการประเมินประสิทธิภาพของการจำแนกกลุ่มข้อมูล โดยใช้ K-means และวัดผลด้วย Silhouette Coefficient และผลการวิเคราะห์ค่าที่มีนัยสำคัญและหัวข้อในข้อความจากแบบจำลอง

1. ผลลัพธ์การทดลองการจัดกลุ่มด้วยอัลกอริทึม K-mean และประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient
2. ผลลัพธ์ของการจำแนกกลุ่มข้อความด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation)

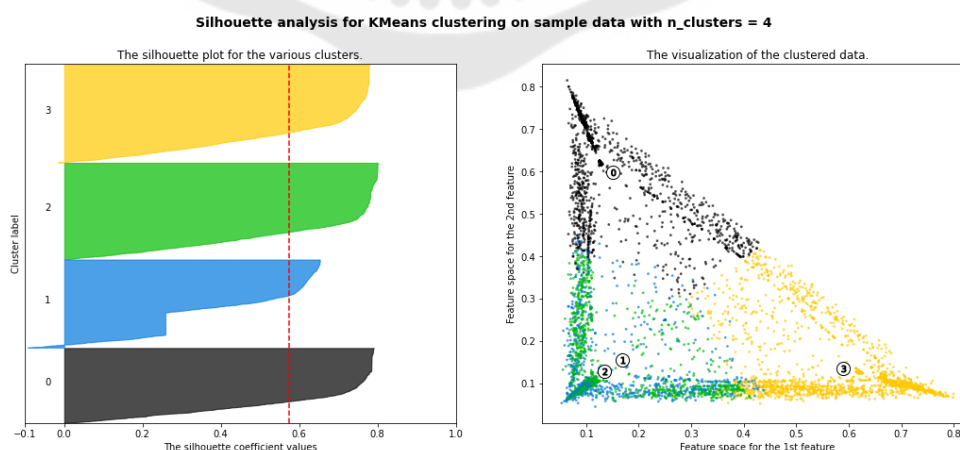
ผลลัพธ์การทดลองการจัดกลุ่มด้วยอัลกอริทึม K-mean และประเมินประสิทธิภาพการจำแนกกลุ่มด้วย Silhouette Coefficient

การจำแนกกลุ่มหัวข้อ โดยใช้ K-means ในการจำแนกกลุ่ม (Cluster) และประเมินประสิทธิภาพโดยใช้ Silhouette score ในการหาค่าจำนวนการจำแนกกลุ่มหัวข้อที่เหมาะสม ได้ผลลัพธ์ดังนี้

ตาราง 4 การประเมินประสิทธิภาพการจำแนกกลุ่มหัวข้อ

จำนวนคลัสเตอร์	ผลการประเมินประสิทธิภาพ
4	0.5397
5	0.4984
6	0.5174
7	0.410
8	0.4838
9	0.4424
10	0.4893

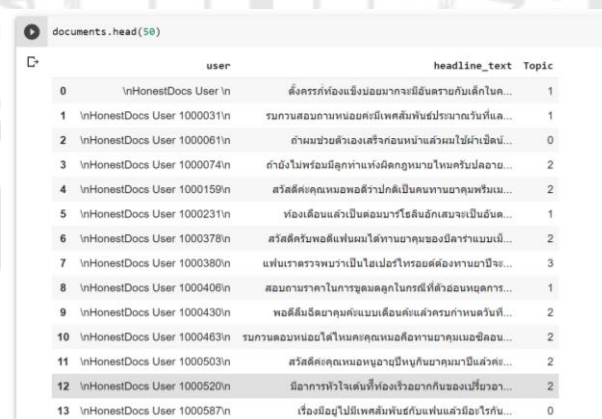
จากการประเมินประสิทธิภาพการจำแนกกลุ่มหัวข้อใน ตารางที่ 4 ผลลัพธ์ของการจำแนกกลุ่มหัวข้อ โดยใช้ K-means โดยการเปรียบเทียบการจำแนกคลัสเตอร์ 2 ถึง 10 คลัสเตอร์ พบว่าการจำแนกกลุ่มข้อมูล เป็น 4 หัวข้อนั้น ค่า Silhouette score มีค่า 0.5397 มีค่าเฉลี่ยในการแบ่งคลัสเตอร์ที่สูงที่สุด เมื่อเปรียบเทียบกับคลัสเตอร์อื่นๆ



ภาพประกอบ 13 กราฟ Silhouette score ในแต่ละหัวข้อ

จากภาพที่ 13 แสดงกราฟ Silhouette วัดค่าของความใกล้เคียงกันของจุดในคลัสเตอร์เดียวกันและวัดกับจุดในคลัสเตอร์อื่นๆ โดยจากกราฟ Silhouette วัดค่าของความใกล้เคียงกันของจุดในคลัสเตอร์เดียวกันและวัดกับจุดในคลัสเตอร์อื่นๆ โดยจากภาพในแกน x คือ ค่า Silhouette ในแกน y คือ จำนวนคลัสเตอร์ ซึ่งมีกลุ่มข้อมูล 4 คลัสเตอร์ โดยขนาดความกว้างในแนวแกน y บ่งชี้ถึงจำนวนของข้อความที่ประกอบอยู่ในหัวข้อนั้นๆ และเส้นปะสีแดง คือ ค่าเฉลี่ยของค่า Silhouette score โดยเมื่อเปรียบเทียบจะเห็นว่า หัวข้อที่ 0, 3 และ 4 ประกอบด้วยข้อความใกล้เคียงกัน และมีค่า Silhouette score ที่ใกล้เคียงกัน และไม่มีค่าที่ปะปนไปยังกลุ่มหัวข้ออื่นๆ หรือมีการปะปนเพียงเล็กน้อย และในหัวข้อที่ 1 มีการปะปนของกลุ่มข้อความอยู่เพียงเล็กน้อย โดยดูจากค่าที่ติดลบของ Silhouette score และในหัวข้อที่ 1 ยังมีกลุ่มค่าที่ต่ำกว่าค่าเฉลี่ย ซึ่งกราฟชี้ให้เห็นว่าการจับกลุ่มของข้อมูลนั้นในแต่ละคลัสเตอร์มีการจับกลุ่มที่ดีแต่ก็มีการปะปนของข้อความเล็กน้อย

ผลลัพธ์ของการจำแนกกลุ่มข้อความด้วยเทคนิคการจัดสรรDirichlet (Latent Dirichlet Allocation)



	user	headline_text	Topic
0	\nHonestDocs User \n	ตั้งรกรักห้องนั่งเล่นมากรับอีกสายกันเล็กในค...	1
1	\nHonestDocs User 1000031\n	รบกวนสอบถามขอทราบพิเศษที่รับทราบกันที่แล...	1
2	\nHonestDocs User 1000061\n	ถ้าผมช่วยตัวเองเสร็จก่อนหน้าแล้วผมไม่เข้าเชิ...	0
3	\nHonestDocs User 1000074\n	ถ้ายังไม่พร้อมมีลูกทำแท้งมีผลกฎหมายไหมครับปลอ...	2
4	\nHonestDocs User 1000159\n	สวัสดิ์ศฤงคารหมอศิริราชปดเป็นคณพชานคุณพริ...	2
5	\nHonestDocs User 1000231\n	ห้องเดือนแล้วเป็นสอมนารโธดิษอีกเสมาจะเป็นอันค...	1
6	\nHonestDocs User 1000378\n	สวัสดิ์ศฤงคารหมอศิริราชปดเป็นคณพชานคุณพริ...	2
7	\nHonestDocs User 1000380\n	แท้งแล้วตรวจพบว่าเป็นไฮโปอีโรหรือต้องทานยาบิ...	3
8	\nHonestDocs User 1000406\n	สอบถามราคาในการซ่อมรถในกรณีใส่ล้อรถหลุดการ...	1
9	\nHonestDocs User 1000430\n	พอดีมีเรื่องอยากปรึกษาแบบละเอียดแล้วครับท่านที่...	2
10	\nHonestDocs User 1000463\n	รบกวนสอบถามขอทราบพิเศษที่รับทราบกันที่แล...	2
11	\nHonestDocs User 1000503\n	สวัสดิ์ศฤงคารหมอศิริราชปดเป็นคณพชานคุณพริ...	2
12	\nHonestDocs User 1000520\n	มีอาการหัวใจเต้นที่ห้องเช่าอยากกินของเบิ...	2
13	\nHonestDocs User 1000587\n	เรื่องมีอยู่ไปมีเพศสัมพันธ์กับแฟนแล้วมีอะไร...	0

ภาพประกอบ 14 ภาพผลลัพธ์การทำวิเคราะห์หัวข้อจากแบบจำลอง

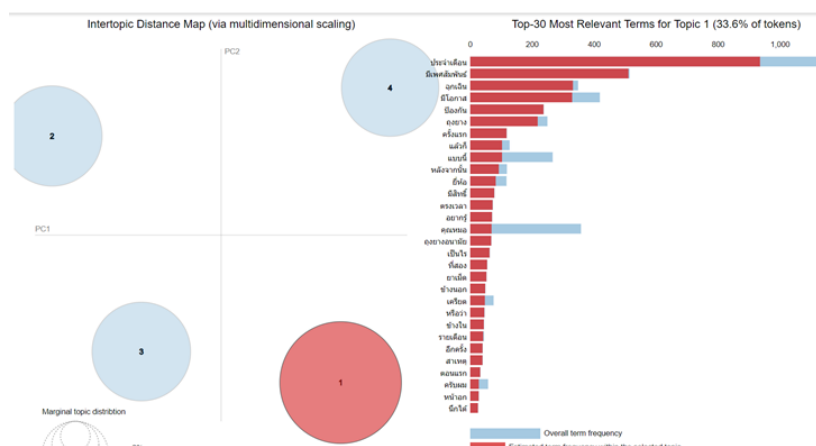
ตาราง 5 ผลการวิเคราะห์หาคำที่มีนัยสำคัญและหัวข้อของแบบจำลอง

ลำดับหัวข้อ	10 อันดับ คำที่ปรากฏในหัวข้อ	จำนวนข้อความ ที่ถูกทำนาย (ข้อความ)
1	('ประจำเดือน', 1100.144), ('มีเพศสัมพันธ์', 600.697), ('ฉุกเฉิน', 389.811), ('มี โอกาส', 386.944), ('ป้องกัน', 278.269), ('ถุงยาง', 256.189), ('ครั้งแรก', 138.034), (' แล้วก็', 120.891), ('แบบนี้', 120.824), ('หลังจากนั้น', 108.507)	3156
2	('ตั้งครรภ์', 512.758), ('คุมกำเนิด', 102.379), ('ค่าใช้จ่าย', 86.645), ('ที่ผ่านมา', 81.052), ('รอบเดือน', 74.40), ('อันตราย', 72.434), ('สีน้ำตาล', 71.968), ('คุณหมอ', 63.779), ('ประจำเดือน', 61.396), ('ฝากครรภ์', 60.656)	1681
3	('ยาคุมกำเนิด', 213.260), ('คุณหมอ', 180.215), ('ประจำเดือน', 140.011), ('ตั้งครรภ์', 106.0836), ('แบบนี้', 105.248), ('ปวดท้อง', 82.235), ('โรงพยาบาล', 75.392), (' กลับมา', 56.274), ('ฮอร์โมน', 55.680), ('รับประทาน', 50.166)	2079
4	('ทำหมัน', 124.553), ('ช่องคลอด', 121.687), ('เข้าไป', 114.509), ('สอดใส่', 110.712), ('อวัยวะเพศ', 101.148), ('พันธุ์', 80.968), ('มีโอกาส', 79.538), ('หลุดลื่น', 78.555), ('น้ำอสุจิ', 69.058), ('ท้องน้อย', 67.856)	1332

ผลการทำนายของแบบจำลองได้วิเคราะห์คำที่มีนัยสำคัญและหัวข้อ ในตารางที่ 5 พบว่าแบบจำลองหัวข้อได้จำแนกหัวข้อให้กับข้อความทั้งหมด 8,248 ข้อความ โดยในคอลัมน์ที่ 2 เป็น 10 อันดับของคำที่ปรากฏในหัวข้อในหัวข้อนั้นๆ ซึ่งแต่ละหัวข้อมีการวิเคราะห์ชุดของคำที่ประกอบในหัวข้อและคำหลักแต่ละคำมีการถ่วงน้ำหนัก (weightage) กำกับให้กับแต่ละหัวข้อ ซึ่งเมื่อวิเคราะห์ผลลัพธ์ที่ได้ พบว่า

การทำนายว่าเป็นหัวข้อที่ 1 จำนวน 3156 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 1 มีตัวอย่าง ดังนี้

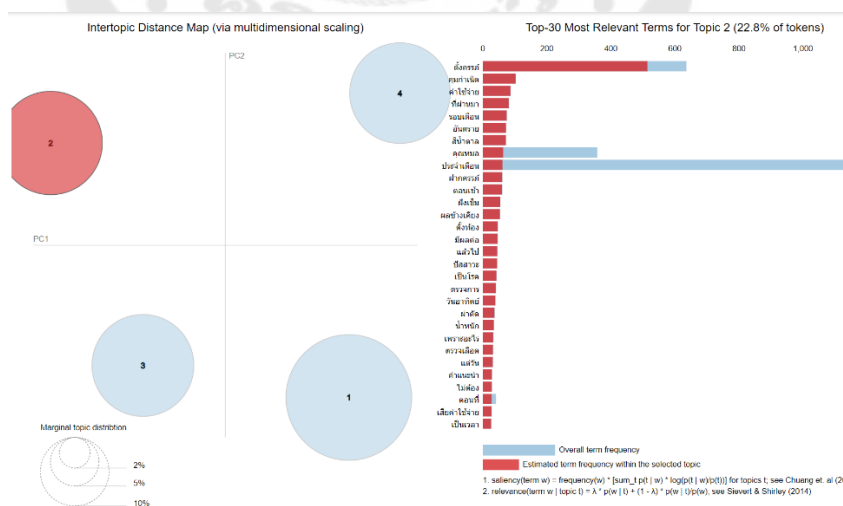
- กินยาคุมประจำเดือนไม่มาตรวจรอบภายในอาทิตย์กว่าๆ แล้วไม่ท้องสามารถเริ่มทานยาคุมแผงต่อไปได้มั้ยคะหรือต้องรอประจำเดือนมาก่อน
- หมอคะคือหนูมีอะไรกับแฟนเมื่อวานไม่ได้ป้องกันและไม่แน่ใจว่าแฟนดึงออกทันมั้ย ถ้าจะกินยาคุมแบบเม็ดวันนี้จะกินทันไหมคะ
- หมอคะหนูสงสัยคะคือหนูกินยาคุมแบบเม็ดเช้าแผงที่แล้วแต่มีเพศสัมพันธ์กับแฟนวันเดียวแล้วหลังใน อยากทราบว่าโอกาสในการตั้งครรภ์มีมากน้อยแค่ไหนคะ



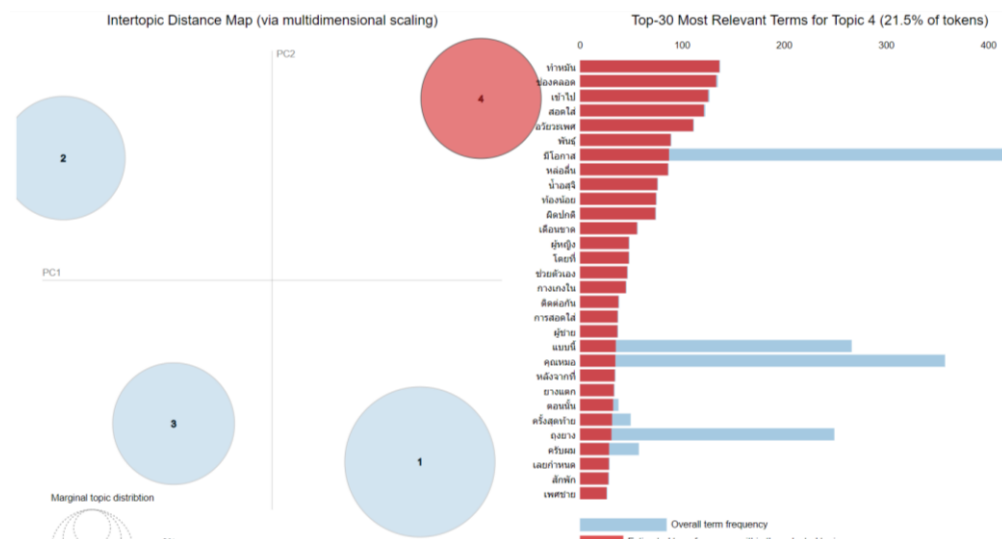
ภาพประกอบ 15 ผลการทำนายข้อความว่าเป็นข้อที่ 1

การทำนายว่าเป็นหัวข้อที่ 2 จำนวน 1681 ข้อความ โดยมีตัวอย่างของข้อความคำปรึกษาในหัวข้อที่ 2 มีตัวอย่าง ดังนี้

- เล่นสล็อตในตอนที่ตั้งครรภ์อยู่ได้ไหมคะ จะทำให้แท้งไหมคะ
- คลอดลูกได้เดือนแล้วคะ ลืมทานยาแบบให้หมอบุตรเดลิตอนเวลาทานตอนกลางคืนคะ ตอนเช้ามีเพศสัมพันธ์จะท้องไหมคะ ถ้าลืมเราต้องทำอย่างไรกังวลมากเลยคะ
- ตั้งครรภ์อยู่ตรวจเลือดเจอพาหะกามโรคจะมีผลต่อเด็กในครรภ์ไหมคะ



ภาพประกอบ 16 ผลการทำนายข้อความว่าเป็นข้อที่ 2



ภาพประกอบ 18 ผลการทำนายข้อความว่าเป็นข้อที่ 4

ภาพประกอบที่ 15-18 เป็นกราฟที่แสดงการ Visualize โดยใช้ library pyLDAvis เป็นการแสดงข้อมูลจากผลลัพธ์ที่ได้จากแบบจำลอง LDA (ซึ่งในกรณีคือ 4 มิติ) โดยการทำให้เหลือเพียง 2 มิติ เพื่อจุดประสงค์ในการ Visualize ภาพ มีวงกลม 4 วงและจุดศูนย์กลางของแต่ละวงกลมแสดงถึงตำแหน่งของหัวข้อในพื้นที่ที่เฟเจอร์แฝง ระยะห่างระหว่างหัวข้อแสดงให้เห็นว่ามีความคล้ายคลึงกับหัวข้ออย่างไรและพื้นที่ของวงกลมเป็นสัดส่วนกับจำนวนเอกสารที่แสดงในแต่ละหัวข้อ แผนภูมิแท่งด้านขวาจะแสดงคำศัพท์ในหัวข้อโดยลดลำดับความเกี่ยวข้อง แต่ละแถบจะระบุความถี่ของคำศัพท์ สีน้ำเงินแสดงถึงความถี่ของคำศัพท์โดยรวมภายในคลังเอกสาร และสีแดงแสดงถึงความถี่ของคำศัพท์เฉพาะในหัวข้อนั้นๆ

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยนี้เพื่อศึกษาวิธีการแบ่งกลุ่มข้อความ ซึ่งใช้ข้อมูลเกี่ยวกับคำปรึกษาปัญหาเกี่ยวกับการตั้งครรภ์ การมีเพศสัมพันธ์ และการคุมกำเนิดภาษาไทย โดยใช้เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลอง โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

สรุปผลการวิจัย

ปัจจุบันชุมชนออนไลน์เป็นพื้นที่หนึ่ง ที่ทำให้คนจำนวนมากสามารถเข้าปรึกษาแพทย์ผู้เชี่ยวชาญได้สะดวก แต่ผู้ใช้งานต้องรอการตอบกลับจากแพทย์ เพราะแพทย์ไม่สามารถให้คำปรึกษาได้ทันทีเมื่อผู้ใช้งานขอคำปรึกษา (Availability) และหากมีปริมาณคำปรึกษามาก แพทย์ไม่สามารถให้คำปรึกษาได้ทัน (Scalability) ซึ่งจากข้อจำกัดเหล่านี้ ระบบโต้ตอบอัตโนมัติสามารถแก้ไขได้ แต่ว่าระบบโต้ตอบอัตโนมัติยังคงมีความท้าทายในด้านภาษาธรรมชาติ ที่ต้องทำให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์

งานวิจัยนี้เป็นการศึกษาวิธีการวิเคราะห์หาค่าที่มีนัยสำคัญและหัวข้อในข้อความ และทำการจัดกลุ่มข้อความที่มีลักษณะคล้ายคลึงกันอยู่ในกลุ่มเดียวกัน ด้วยเทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) ซึ่งให้เห็นว่าผู้ใช้งานอินเทอร์เน็ตจำนวนมากที่เข้าปรึกษาปัญหากับแพทย์ผ่านชุมชนออนไลน์นั้น ส่วนมากมีความกังวลในเรื่องเรื่อง เกี่ยวกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือนและการคุมกำเนิด โดยงานวิจัยนี้ได้ทำการแบ่งกลุ่มข้อมูลโดยใช้ K-means ในการวัดประสิทธิภาพในการจัดกลุ่มของ เทคนิคการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation) โดยค่า Silhouette Coefficient มีการจัดกลุ่มข้อมูลที่มี 4 กลุ่ม มีค่าอยู่ที่ 0.5397 ซึ่ง มีค่าที่สูงและข้อมูลมีการจับกลุ่มได้ดีเมื่อเทียบกับการจัดกลุ่มด้วยจำนวนอื่นๆ

อภิปรายผลการวิจัย

การทดลองนี้ใช้ Colab สำหรับสร้างแบบจำลองการวิเคราะห์คำที่มีนัยสำคัญและหัวข้อจากข้อความ มีจำนวนข้อความทั้งสิ้น ทั้งหมด 8,248 ข้อความ ซึ่งผู้วิจัยได้เปรียบเทียบคำสำคัญในแต่ละหัวข้อกับหัวข้อที่ผู้วิจัยตั้งสมมุติฐานไว้เบื้องต้น โดยสมมุติฐานที่ตั้งไว้เบื้องต้น ได้แก่ (1) อาการผิดปกติในขณะตั้งครรภ์ (2) ความผิดปกติของประจำเดือน (3) ความกังวลเมื่อมีเพศสัมพันธ์ (4) การคุมกำเนิด

7745	3619	ท้องได้เดือนราคาเท่าไรคือมีโรคประจำตัวเป็นไทรอยเป็นพิษตอนมีลูกกับแฟนไม่รู้ว่าท้องเพิ่งตรวจเลยรู้ว่าท้อง	1
7746	3620	มีภาวะน้ำหนักน้อยตั้งครรถ์ได้วิธอันตรงมากไหมคะหนูเป็นผู้ป่วยนอกค่าใช้จ่ายในการตรวจหาโรค	1
7747	3621	มีตุ่มเล็กๆเกาะกันคล้ายดอกกะหล่ำคะเหมือนเป็นตุ่มหนองในบริเวณปากช่องคลอดเกือบถึงทวารแล้วกำลังตั้งท้องอยาก	3
7748	3622	ประจำเดือนครั้งสุดท้ายของแฟนวันที่ขึ้นวาคมแล้วมีเพศสัมพันธ์กับแฟนที่มกราคมตอนนั้นวันที่มกราคมแล้วประจำเดือน	2
7749	3623	สวัสดิ์ศุภคุณหมอคือตอนนี้ประจำเดือนมามากกว่าปกติคือมาบ่อยมานานเหตุเพราะฝังยาคุมกำเนิดแต่ตอนนี้ออกดอกแล้ว	2
7750	3624	กินยาคุมแบบเม็ดละตอนนั้นกินหมดแล้วแต่กินไม่ตรงเวลาบ้างบางทีก็ลืมกินต้องมากินพร้อมกันกับอีกเม็ดบ้างอยากทราบ	2
7751	3625	ตั้งครรถ์เดือนไหนคะจะได้บ้างคะคำถามนี้ได้รับการตอบจากพยาบาลผู้เชี่ยวชาญ	1
7752	3626	ประจำเดือนไม่มาแต่ตรวจแล้วคือไม่ท้องกินยาคุมจนหมดแผงแล้วยังไม่มาแต่เคยมาตอนกินยาคุมประมาณเม็ดที่หลังจาก	0
7753	3627	พอดีว่ามีเรื่องกังวลนิดหน่อยคือพอดีว่าได้ทำการสำเร็จความใคร่ให้แฟนในท่ายืนในวันที่ประมาณเย็นว่าแต่ไม่แน่ใจว่า	2
7754	3628	ใช้มือช่วยแฟนแคะแคะมดมือคะแล้วแฟนพาไปใส่เราแล้วมันค้างที่สะดือมันจะมีผลอะไรไหมคะหรือไม่มีห่าอะไรให้	0
7755	3629	ถ้ามีเพศสัมพันธ์ครั้งติดกันเร็วเกินไปจน	0
7756	3630	ฉีดยาคุมแบบเดือนแต่เคยกำหนดมาวันแล้วผลออกไปมีอะไรกับแฟนจะเป็นอะไรไหมคะ	1
7757	3631	อยากทราบว่ายาคุมฉุกเฉินมีประสิทธิภาพที่เปอร์เซ็นต์ครับแล้วหลังจากนั้นหลังจากนั้นหลังจากนั้นหลังจากนั้นหลังจากนั้นหลังจากนั้น	3
7758	3632	เพิ่งกินยาคุมครั้งแรกหลังประจำเดือนหยุดไปวันรวมแล้ววันครึ่งเพิ่งได้กินเม็ดแรกพริบแต่ยังไม่ได้อะไรกับแฟนยาคุมจะ	2
7759	3633	ฝังยาคุมมาจะสามปีแล้วคะมีเพศสัมพันธ์ไม่ป้องกันหลังในมีสิทธ์ท้องน้อยคะฝังยาคุมมีเพศสัมพันธ์กับหลายคน	1
7760	3634	เพิ่งกินยาคุมครั้งแรกหลังจากมีเพศสัมพันธ์กับแฟนมีเสิร์กไปซื้อยาคุมแบบเม็ดมากินซึ่งวันที่มีอะไรกับแฟนมันเป็นวันพุธก็	2
7761	3635	ตรวจไวรัสตับตรวจมีสาระเสฟติดตั้งครรถ์ค่าใช้จ่ายเท่าไรคำถามนี้ได้รับการตอบจากพยาบาลผู้เชี่ยวชาญ	1
7762	3636	คะอยากแจ้งหัวข้อยามาทำสองครั้งแล้วคะแต่ไม่หลุดคะ	0
7763	3637	แฟนผมอาหารเป็นพิษหมอเอาเลือดไปตรวจหลังจากนั้นหมอบอกว่าประจำเดือนมาครั้งล่าสุดเมื่อไหร่มีเกี่ยวกับกับปัจจัย	1
7764	3638	สวัสดิ์ศุภคุณกรยาหม่วยเป็นโรคซึมเศร้ารักษาอาการต่อเนื่องมาได้อปีแล้วและตอนนี้กำลังตั้งครรถ์ได้เดือนแล้วครับและนิ่ว	1
7765	3639	มีเพศสัมพันธ์กับแฟนแต่หลังนอกแต่ถุงยางหลุดตอนเอาออกมาพอดีถุงยางไม่ได้หลุดไปในช่องคลอดมีโอกาท้องไหม	0
7766	3640	เดือนพฤศจิกายนเป็นประจำเดือนวันแล้วมีเพศสัมพันธ์ตอนเป็นประจำเดือนพอดีตอนนั้นมาครบกำหนดประจำเดือนมาเป็น	2
7767	3641	เปลี่ยนยาคุมแต่ก่อนทานยาคุมยี่ห้อแอนนาเม็ดเลื่อยาคุมเปลี่ยนยี่ห้ออื่นนี้เม็ดคือประจำเดือนมาวันที่คือเมื่อยาที่แล้วหาบย	2
7768	3642	มีโรคแฟนแต่แฟนหลังในมีสิทธ์ท้องไหมคะเพิ่งคลอดลูกได้เดือนกว่าแต่กินยาคุม	0

ภาพประกอบ 19 ภาพผลการทำนายจากแบบจำลอง

จากภาพประกอบที่ 19 เป็นตัวอย่างของคำปรึกษาที่ผู้ใช้งานอินเตอร์เน็ตปรึกษากับแพทย์ผู้เชี่ยวชาญผ่านเว็บไซต์และผลการทำนายหัวข้อในแต่ละข้อความจากแบบจำลอง

ตาราง 6 ตารางผลการเปรียบเทียบสมมุติฐานกับผลการทำนายของแบบจำลอง

สมมุติฐาน	1	2	3	4
อาการผิดปกติในขณะตั้งครรภ์		/	/	
ความผิดปกติของประจำเดือน		/	/	
ความกังวลเมื่อมีเพศสัมพันธ์	/			/
การคุมกำเนิด	/		/	/

โดยจากการตั้งสมมุติฐาน และวิเคราะห์ค่าที่มีนัยสำคัญในแต่ละหัวข้อ พบว่า

หัวข้อที่ 1 มีความเกี่ยวข้องกับความกังวลเมื่อมีเพศสัมพันธ์และการคุมกำเนิด โดยมีค่าที่เกี่ยวข้องได้แก่ ประจำเดือน มีเพศสัมพันธ์ บ่อยกัน อ้วนง่าย และเครียด

หัวข้อที่ 2 มีความเกี่ยวข้องกับ อาการผิดปกติในขณะตั้งครรภ์และความผิดปกติของประจำเดือน โดยมีค่าที่เกี่ยวข้อง ได้แก่ มีคว่ำตั้งครรภ์ คุมกำเนิด รอบเดือน อันตราย และสีน้ำตาล

หัวข้อที่ 3 มีความเกี่ยวข้องกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือน และการคุมกำเนิด โดยมีค่าที่เกี่ยวข้อง ได้แก่ ยาคุมกำเนิด ประจำเดือน ตั้งครรภ์ และรับประทาน

หัวข้อที่ 4 มีความเกี่ยวข้องกับ ความกังวลหลังมีเพศสัมพันธ์,ความผิดปกติของประจำเดือน มีค่าที่เกี่ยวข้อง การทำหมัน สอดใส่ ประจำเดือน และผิดปกติ

เมื่อพิจารณาพบว่าการตั้งสมมุติฐานเบื้องต้นและผลการทำนาย พบว่า ผู้ใช้งานมีการเข้าปรึกษาปัญหาเกี่ยวกับ อาการผิดปกติในการตั้งครรภ์ ความผิดปกติของประจำเดือนและการคุมกำเนิด ซึ่งผลการทำนายว่าเป็นหัวข้อที่ 3 จำนวน 3020 ข้อความ มากที่สุด นั้นหมายความว่า ผู้ใช้งานส่วนมากที่เข้ามาปรึกษากับแพทย์นั้นมีความกังวลเรื่อง การทานยาคุม การคุมกำเนิด ความผิดปกติของประจำเดือน อาจจะเป็นจากการทานยาคุม หรือความกังวลหลังการมีเพศสัมพันธ์นั่นเอง และการพิจารณาพบว่าการตั้งสมมุติฐานเบื้องต้นไม่ได้กล่าวถึง การทำหมัน สิทธิบัตรทองหรือค่าใช้จ่าย ซึ่งการทำ วิธีการวิเคราะห์หาค่าที่มีนัยสำคัญและหัวข้อในข้อความ

ด้วยเทคนิคการจัดสรรดีรีเคลแฟง (Latent Dirichlet Allocation) ทำให้เห็นคำที่แฝงอยู่ในข้อความมากกว่าผู้วิจัยได้ตั้งสมมุติฐานไว้

ข้อเสนอแนะ

1. งานวิจัยนี้มีการใช้เทคนิค เพียงเทคนิคเดียว ดังนั้นอาจมีเทคนิคอื่นที่มีความสามารถในการวิเคราะห์คำที่มีนัยสำคัญและหัวข้อที่มีประสิทธิภาพมากกว่า
2. งานวิจัยนี้ในขั้นตอนการเตรียมข้อมูลไม่ได้แก้คำผิด เมื่อเข้ากระบวนการ การตัดคำ คำเหล่านั้นอาจถูกตัดทิ้งไป ดังนั้นหากแก้ไขคำผิดอาจช่วยหาคำที่มีนัยสำคัญมากขึ้นหรือการวิเคราะห์หัวข้อมีความแม่นยำมากขึ้น
3. งานวิจัยไม่ได้มีผู้เชี่ยวชาญจากแพทย์ช่วยในการตีความผลลัพธ์ที่ได้ เพื่อวัดความแม่นยำของการทำนายของแบบจำลอง หากมีผู้เชี่ยวชาญก็จะช่วยให้การพัฒนาแบบจำลองดีขึ้นหรือการประเมินประสิทธิภาพโมเดลมีประสิทธิภาพมากขึ้น
4. แบบจำลองนี้สามารถนำไปปรับและประยุกต์ใช้กับระบบหุ่นยนต์โต้ตอบการสนทนา (Chatbot) และ การ กำหนด ชุด คำ ตอบ ในการ วิจัย ครั้ง ต่อไป

บรรณานุกรม

1. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J Med Internet Res*. 2021;23(1):e17828.
2. Bahng J, Lee CH. Topic Modeling for Analyzing Patients' Perceptions and Concerns of Hearing Loss on Social Q&A Sites: Incorporating Patients' Perspective. *Int J Environ Res Public Health*. 2020;17(17).
3. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *the Journal of machine Learning research*. 2003;3:993-1022.
4. Chaves A, Gerosa M. How Should My Chatbot Interact? A Survey on Social Characteristics in Human Chatbot Interaction Design. *International Journal of Human Computer Interaction*. 2021;37:729 - 58.
5. Dinakar K, Jones B, Lieberman H, Picard R, Rose C, Thoman M, et al., editors. You too?! mixed-initiative lida story matching to help teens in distress. *Sixth International AAAI Conference on Weblogs and Social Media*; 2012.
6. H. S. Text mining Twitter social media for Covid-19: Comparing latent semantic analysis and latent Dirichlet allocation. 2020.
7. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*. 2019;78(11):15169-211.
8. Kim S-W, Gil J-M. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*. 2019;9(1):30.
9. Lim DS. Predicting outcomes in online chatbot-mediated therapy. 2017.
10. Maroengsit W, Piyakulpinyo T, Phonyiam K, Pongnumkul S, Chaovalit P, Theeramunkong T. A Survey on Evaluation Methods for Chatbots. *Proceedings of the 2019 7th International Conference on Information and Education Technology*; Aizu-Wakamatsu, Japan: Association for Computing Machinery; 2019. p. 111–9.

11. Raval H. Limitations of Existing Chatbot with Analytical Survey to Enhance the Functionality Using Emerging Technology. *International Journal of Research and Analytical Reviews (IJRAR)*. 2020;7(2).
12. Tian Y, Kochhar PS, Lim E-P, Zhu F, Lo D, editors. Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting. *International Conference on Social Informatics*; 2013: Springer.
13. Touimi YB, Hadioui A, Faddouli NE, Bennani S. Intelligent Chatbot-LDA Recommender System. *International Journal of Emerging Technologies in Learning (IJET)*. 2020;15(20).
14. Twinandilla S, Adhy S, Surarso B, Kusumaningrum R. Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) – Significance Sentences. *Procedia Computer Science*. 2018;135:663-70.
15. Vamsi GK, Rasool A, Hajela G, editors. Chatbot: A Deep Neural Network Based Human to Machine Conversation Model. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2020 1-3 July 2020.
16. Wexler A, Davoudi A, Weissenbacher D, Choi R, Oconnor K, Cummings H, et al. Pregnancy and health in the age of the Internet: A content analysis of online “birth club” forums. *PLOS ONE*. 2020;15:e0230947.
17. Zhang K, Wu W, Wu H, Li Z, Zhou M, editors. Question retrieval with high quality answers in community question answering. *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*; 2014.

ประวัติผู้เขียน

ชื่อ-สกุล

ปิยวรรณ ทองพลอย

วัน เดือน ปี เกิด

23 มีนาคม 2538

สถานที่เกิด

ปทุมธานี

วุฒิการศึกษา

พ.ศ. 2560

วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

จาก มหาวิทยาลัยแม่ฟ้าหลวง

พ.ศ. 2562

วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล

จาก มหาวิทยาลัยศรีนครินทรวิโรฒ

