



AI PLATFORM THAI CONTENT TAGGING

MS. THITIWORADA AMSA-NGUAN
MS. NAWAKARN LEERATTANACHOTE
MR. PONLAWAT SUPARAT

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)
FACULTY OF ENGINEERING
KING MONGKUT'S UNIVERSITY OF TECHNOLOGY THONBURI
2022

AI Platform Thai Content Tagging

Ms. Thitiworada Amsa-nguan

Ms. Nawakarn Leerattanachote

Mr. Ponlawat Suparat

A Project Submitted in Partial Fulfillment
of the Requirements for
the Degree of Bachelor of Engineering (Computer Engineering)
Faculty of Engineering
King Mongkut's University of Technology Thonburi
2022

Project Committee

Project Advisor

(Asst.Prof. Santitham Prom-on, Ph.D.)

Committee Member

(Asst.Prof. Dr.-Ing Priyakorn Pusawiro)

Committee Member

(Assoc.Prof. Peerapon Siripongwutikorn, Ph.D.)

Committee Member

(Assoc.Prof. Thumrongrat Amornraksa, Ph.D.)

Project Title	AI Platform Thai Content Tagging
Credits	3
Member(s)	Ms. Thitiworada Amsa-nguan Ms. Nawakarn Leerattanachote Mr. Ponlawat Suparat
Project Advisor	Asst.Prof. Santitham Prom-on, Ph.D.
Program	Bachelor of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2022

Abstract

Categorization is the process of grouping related articles under the same category based on commonly understood standards. It facilitates easy information retrieval and exploration of relevant content. With the increasing amount of online content, there is a growing interest among developers in automating categorization to achieve rapid and efficient results, as manual categorization requires additional human resources. One method that developers have shown interest in is automatic categorization using artificial intelligence (AI). AI can analyze a large amount of online content to categorize it quickly and efficiently. Furthermore, it can consistently categorize online content, even from diverse sources.

We have developed three models: Random Forest with TF-IDF, K-Nearest Neighbor with TF-IDF, and LSTM. These models use a dataset obtained by translating online content from Thai to English. The models achieved accuracy scores of 94.56%, 93.11%, 93.78%, respectively. We implemented models in a web application for categorizing Thai online content using AI. The web application was developed to provide a platform for users to search for online content based on their desired categories. Additionally, we developed a web application for system administrators to add and manage categories through the frontend. Based on user evaluations, the web application received a high level of satisfaction regarding usability and accuracy.

Keywords: Automatic categorization / Online Content / Artificial Intelligence / Web Application

หัวข้อปริญญาในพนธ์	เรื่บแอปพลิเคชันสำหรับการจัดหมวดหมู่ Online Content ที่เป็นภาษาไทย ในรูปแบบ AI-Based Platform
หน่วยกิต	3
ผู้เขียน	นางสาววิชิตรา อําส่วน นางสาวนวนานต์ ลีรัตนโชค นายพลาต สุภารัตน์
อาจารย์ที่ปรึกษา	ผศ.ดร.สันติธรรม พรหมอ่อน
หลักสูตร	วิศวกรรมศาสตรบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2565

บทคัดย่อ

การจัดหมวดหมู่ คือ การจัดกลุ่มบทความที่มีเนื้อหาสัมพันธ์กันอยู่ภายใต้กลุ่มเดียวกันตามมาตรฐานที่เข้าใจร่วมกัน เพื่อง่ายต่อการค้นหาข้อมูลและสำรวจเนื้อหาในร่องที่เกี่ยวข้อง ซึ่งในปัจจุบันเริ่มมีนักพัฒนาให้ความสนใจกับการจัดหมวดหมู่แบบอัตโนมัติ เนื่องจากจำนวนเนื้อหาออนไลน์ที่มีจำนวนมากขึ้น ส่งผลให้การจัดหมวดหมู่ย่างรวดเร็วและมีประสิทธิภาพต้องใช้ทรัพยากร่มนุษย์เพิ่มขึ้นเมื่อใช้วิธีการจัดหมวดหมู่ด้วยมนุษย์ ที่นี่ในวิธีการที่นักพัฒนาให้ความสนใจ คือ การจัดหมวดหมู่แบบอัตโนมัติโดยใช้เทคโนโลยีปัญญาประดิษฐ์ เนื่องจากเทคโนโลยีปัญญาประดิษฐ์สามารถวิเคราะห์เนื้อหาออนไลน์ที่มีจำนวนมากเพื่อจัดหมวดหมู่ได้อย่างรวดเร็วและมีประสิทธิภาพ นอกจากนี้ยังสามารถจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ได้อย่างสม่ำเสมอ แม้จะมีแหล่งที่มาจากการหลากหลายแหล่งก็ตาม

คณาจัดทำได้พัฒนาโมเดลทั้งหมด 3 โมเดลด้วยกัน ได้แก่ Random Forest with TF-IDF, K-Nearest Neighbor with TF-IDF และ LSTM โดยใช้ชุดข้อมูลที่เกิดจากการแปลงเนื้อหาออนไลน์ที่เป็นภาษาไทยให้กลายเป็นภาษาอังกฤษ ซึ่งได้ผลทดสอบความแม่นยำอยู่ที่ 94.56%, 93.11%, 93.78% ตามลำดับ และนำโมเดลที่ได้ไปพัฒนาเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทย โดยใช้เทคโนโลยีปัญญาประดิษฐ์ในการจัดหมวดหมู่และพัฒนาเว็บแอปพลิเคชันให้เป็นแพลตฟอร์มสำหรับการค้นหาเนื้อหาออนไลน์ตามหมวดหมู่ที่ผู้ใช้งานต้องการ นอกจากนี้คณาจัดทำได้มีการพัฒนาเว็บแอปพลิเคชันสำหรับผู้จัดการระบบให้สามารถเพิ่มและจัดการหมวดหมู่ได้ด้วยตนเองผ่านทางระบบหน้าบ้าน โดยจากการประเมินเว็บแอปพลิเคชันจากผู้ใช้งานมีระดับความพึงพอใจค่อนข้างมากในด้านการใช้งานรวมถึงความแม่นยำ

คำสำคัญ: การจัดหมวดหมู่แบบอัตโนมัติ / เนื้อหาออนไลน์ / เทคโนโลยีปัญญาประดิษฐ์ / เว็บแอปพลิเคชัน

กิตติกรรมประกาศ

การทำโครงการในครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี ด้วยความช่วยเหลือจาก ผศ.ดร.สันติธรรม พรมอ่อน อาจารย์ที่ปรึกษาโครงการ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่าง ๆ อันเป็นประโยชน์อย่างยิ่งในการทำโครงการ ช่วยแก้ปัญหาต่าง ๆ ที่เกิดขึ้นระหว่างการทำโครงการ อีกทั้งยังดูแลพวกเราเป็นอย่างดี ผู้จัดทำโครงการจึงขอกราบขอบพระคุณเป็นอย่างสูงมา ณ ที่นี่

ขอขอบพระคุณ ผศ.ราชวิชช์ สโรชิกสิต, ดร. จัตุรนต์ หาญสมบูรณ์, รศ.ดร.ธารงรตัน อมรรักษा, ผศ.ดร.บริยกร ปุสวิร และ รศ.ดร.พีรพล ศิริพงศ์ภูมิกร ที่ได้สละเวลาไว้ร่วมเป็นคณะกรรมการตรวจสอบ รวมถึงการให้คำแนะนำและข้อคิดเห็นต่าง ๆ อันเป็นประโยชน์ ต่อการทำโครงการครั้งนี้

ขอบคุณเพื่อน ๆ พี่ ๆ น้อง ๆ ภาควิชาวิศวกรรมคอมพิวเตอร์ที่ให้คำปรึกษาและให้ข้อมูลที่จำเป็นต่อการทำโครงการให้สำเร็จลุล่วง รวมถึงครอบครัวที่กำลังใจและเคยอยู่เคียงข้างกันตลอดระยะเวลาการทำโครงการ

สารบัญ

หน้า

ABSTRACT	ii
บทคัดย่อ	iii
กิตติกรรมประกาศ	iv
สารบัญ	v
สารบัญตาราง	viii
สารบัญรูปภาพ	ix
 บทที่ 1 บทนำ	 1
1.1 ที่มาและความสำคัญ	1
1.2 วัตถุประสงค์	1
1.3 ขอบเขตของโครงงาน	1
1.4 เนื้อหาทางวิศวกรรมที่เป็นต้นฉบับ	1
1.5 ประโยชน์ที่คาดว่าจะได้รับ	2
1.6 ขั้นตอนการทำงานและระยะเวลาการดำเนินงาน	2
1.6.1 ขั้นตอนการทำงาน	2
1.6.2 ระยะเวลาการดำเนินงาน	3
1.7 ผลการดำเนินงาน	4
1.7.1 ผลการดำเนินงานในภาคการศึกษาที่ 1	4
1.7.2 ผลการดำเนินงานในภาคการศึกษาที่ 2	4
 บทที่ 2 ทฤษฎีความรู้และงานที่เกี่ยวข้อง	 5
2.1 ทฤษฎีที่เกี่ยวข้อง	5
2.1.1 Natural Language Processing (NLP)	5
2.1.2 Word Embedding	5
2.1.3 TF-IDF	6
2.1.4 Random Forest	6
2.1.5 K-Nearest Neighbor (KNN)	7
2.1.6 Neural Network (NN)	8
2.1.7 RESTful API	9
2.1.8 Event Driven Architecture (EDA)	10
2.1.9 Dependency Injection (DI)	10
2.1.10 Continuous Integration (CI)	10
2.1.11 Continuous Deployment (CD)	10
2.1.12 Web Scraping	11
2.1.13 Agile	11
2.2 งานวิจัยที่เกี่ยวข้อง	11
2.2.1 Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type	11
2.3 ผลิตภัณฑ์ที่เกี่ยวข้อง	12
2.3.1 Amazon Comprehend	12
2.4 เทคนิคและเทคโนโลยีที่ใช้	13
2.4.1 Development Tools	13
2.4.2 Database Tools	15
2.4.3 DevOps Tool	15

2.4.4 Library Tools	15
2.4.5 Infrastructure Tools	17
2.4.6 Application Testing Tools	17
2.4.7 Project Management	17
บทที่ 3 วิธีการดำเนินงาน	19
3.1 ข้อกำหนดและความต้องการของระบบ	19
3.2 สถาปัตยกรรมระบบ	19
3.2.1 System Overview	19
3.2.2 Content-Tagging	20
3.2.3 Content-Preparing	20
3.2.4 General-Backend	21
3.2.5 Navigation Map	24
3.2.6 Data Science Processing	25
3.3 Use Case Analysis	28
3.3.1 Use Case Diagram	28
3.3.2 Use Case Narrative	28
3.3.2.1 ผู้ใช้งานสมัครสมาชิก	28
3.3.2.2 ผู้ใช้งานเข้าสู่ระบบ	30
3.3.2.3 ตรวจสอบข้อมูลໂປຣີໂພຣີຂອງຕົນເອງ	32
3.3.2.4 สร้างคำร้องขอวิเคราะห์หมวดหมู่ຂອງເນື້ອທາດ້ວຍ Web Link	34
3.3.2.5 สร้างคำร้องขอวิเคราะห์หมวดหมู่ຂອງເນື້ອທາດ້ວຍບທຄວາມ	36
3.3.2.6 สร้างคำร้องขอวิเคราะห์หมวดหมู่ຂອງເນື້ອທາດ້ວຍ Web Link ຜ່ານ Api-Key	38
3.3.2.7 สร้างคำร้องขอวิเคราะห์หมวดหมู่ຂອງເນື້ອທາດ້ວຍບທຄວາມຜ່ານ Api-Key	40
3.3.2.8 ດູກໍາລັງຂອງການວິເຄາະທີ່ໜີ້ອະບທຄວາມຂອງຕົນເອງ	42
3.3.2.9 ຄັ້ນຫາບທຄວາມໂດຍໃຊ້ເລີຂ້າງອັງເຈິ້ງເນື້ອທາ	44
3.3.2.10 ຄັ້ນຫາບທຄວາມໂດຍໃຊ້ເລີຂ້າງອັງເຈິ້ງເນື້ອທາ	46
3.4 ໂຄງສ້າງฐานข้อมูล	47
3.5 User Interface Design	49
3.5.1 User Interface Design Rule	49
3.5.2 User Interface สำหรับผู้ใช้งาน	50
3.5.2.1 หน้า Landing	50
3.5.2.2 หน้า Login	52
3.5.2.3 หน้า Register	53
3.5.2.4 หน้า Home	56
3.5.2.5 หน้า Analytic	58
3.5.2.6 หน้า My Ticket	60
3.5.2.7 หน้าแสดงผลการວິເຄາະທີ່ໜີ້ອະບທຄວາມ	61
3.5.3 User Interface สำหรับຜູ້ຜູ້ແລະຮັບ	61
3.5.3.1 หน้า Home	61
3.5.3.2 หน้า Upload New Model	62
3.6 Class Diagram	63
3.7 การออกแบบการทดลอง	63
3.7.1 User Evaluation	63
3.7.2 Application Evaluation	63
3.7.3 Artificial Intelligence Evaluation	63
บทที่ 4 ผลการดำเนินงาน	65
4.1 การเก็บข้อมูล	65

4.2 การออกแบบโมเดลสำหรับวิเคราะห์หมวดหมู่ของเนื้อหาในบทความต่าง ๆ	66
4.2.1 การเตรียมข้อมูลสำหรับการพัฒนาโมเดล	66
4.2.2 Exploratory Data Analysis	68
4.2.2.1 ชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย	68
4.2.2.2 ชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ	73
4.2.3 การทำ Text Representation ด้วย TF-IDF	78
4.2.4 การเลือก Model ระหว่าง Random Forest KNN และ NN	80
4.2.5 การทดสอบโมเดล	83
4.3 Application	84
4.3.1 User Interface Design	84
4.3.1.1 หน้าแรก	84
4.3.1.2 หน้าเข้าสู่ระบบ	85
4.3.1.3 หน้าหลัก	86
4.3.1.4 หน้าสร้างคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link	86
4.3.1.5 หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ	87
4.3.2 การทดสอบความสามารถของระบบ	88
4.3.3 การทดสอบประสิทธิภาพของระบบ	89
4.4 การประเมินผลและความเที่ยงจากผู้ใช้งาน	89
4.4.1 ผลสำรวจจากการตอบแบบสอบถามของผู้ใช้งาน	93
4.4.2 สรุปผลการสำรวจจากแบบสอบถาม	96
บทที่ 5 บทสรุปและอภิปราย	97
5.1 สรุปผลโครงการ	97
5.2 ปัญหาที่พบและการแก้ไข	97
5.3 ข้อจำกัดและข้อเสนอแนะ	98
5.3.1 ข้อจำกัด	98
5.3.2 ข้อเสนอแนะและแนวทางในการพัฒนา	98
หนังสืออ้างอิง	99
APPENDIX	102
A ชื่อภาคผนวกที่ 1	103

สารบัญตาราง

ตารางที่	หน้า
1.1 ตารางแสดงเวลาดำเนินงานของภาคการเรียนที่ 1/2565	3
1.2 ตารางแสดงเวลาดำเนินงานของภาคการเรียนที่ 2/2565	3
2.1 เปรียบเทียบผลิตภัณฑ์ระหว่าง Amazon Comprehend กับ AI-based Thai Content Tagging Platform	13
3.1 API Document ของ User ใน การสร้างบัญชีสำหรับใช้งานภายในระบบ	30
3.2 API Document ของ User ใน การเข้าสู่ระบบบัญชีผู้ใช้งาน	32
3.3 API Document ของ User ใน การดูรายละเอียดไปรษณีย์ของตนเอง	33
3.4 API Document ของ Ticket ใน การสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link	35
3.5 API Document ของ Ticket ใน การสร้างคำร้องขอวิเคราะห์ด้วยเนื้อหาที่เตรียมไว้	37
3.6 API Document ของ Ticket ใน การสร้างคำร้องขอวิเคราะห์ด้วย Web Link (Api-key)	39
3.7 API Document ของ Ticket ใน การสร้างคำร้องขอวิเคราะห์ด้วยเนื้อหาที่เตรียมไว้ (Api-Key)	41
3.8 API Document ของ Ticket ใน การดูรายการคำร้องขอของตนเอง	43
3.9 API Document ของการแสดงรายการค้นหาด้วยตัวกรองต่าง ๆ	45
3.10 API Document ของการแสดงผลข้อมูลเกี่ยวกับเนื้อหาตามเลขอ้างอิงเนื้อหา	46
3.11 ตารางข้อมูลของสมาชิกในระบบ	47
3.11 ตารางข้อมูลของสมาชิกในระบบ	48
3.12 ตารางข้อมูลคำร้องขอการวิเคราะห์หมวดหมู่ของบทความ	48
3.13 ตารางบทความที่ทำการวิเคราะห์	48
3.14 ตารางหมวดหมู่ที่ใช้งานภายในระบบ	49
3.15 ตารางความสมมติของเนื้อหากับหมวดหมู่ต่าง ๆ	49
4.1 ตารางชุดข้อมูลที่รวมเนื้อหาเป็นภาษาไทย	66
4.2 ตารางชุดข้อมูลที่รวมเนื้อหาเป็นภาษาอังกฤษ	66
4.3 ตารางชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยสำหรับใช้พัฒนาและทดสอบโมเดล	67
4.4 ตารางชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษสำหรับใช้พัฒนาและทดสอบโมเดล	67
4.5 ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาไทยทั้งหมดหลังทำความสะอาดข้อมูล	69
4.6 ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาไทยของแต่ละหมวดหมู่หลังทำความสะอาดข้อมูล	69
4.7 ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษทั้งหมดหลังทำความสะอาดข้อมูล	73
4.8 ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่หลังทำความสะอาดข้อมูล	73
4.9 โครงสร้างของโมเดล TF-IDF ในรูปแบบภาษาไทยและภาษาอังกฤษ	78
4.10 แสดง 3 คำที่มีความสำคัญมากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของแต่ละหมวดหมู่	78
4.11 แสดง 3 คำที่มีความสำคัญมากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่	79
4.12 ค่า Accuracy ของแต่ละโมเดลเมื่อใช้ TF-IDF กับชุดข้อมูลที่เป็นภาษาไทย	80
4.13 ค่า Accuracy ของแต่ละโมเดลเมื่อใช้ TF-IDF กับชุดข้อมูลที่เป็นภาษาอังกฤษ	80
4.14 แสดงค่า Evaluation Metrics ของโมเดล LSTM	81
4.15 แสดงค่า Evaluation Metrics ของโมเดลแต่ละโมเดลสำหรับชุดข้อมูลที่เป็นภาษาไทย	81
4.16 แสดงค่า Evaluation Metrics ของโมเดลแต่ละโมเดลสำหรับชุดข้อมูลที่เป็นภาษาอังกฤษ	81
4.17 ผลการทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายใต้ Domain ที่ระบบรองรับ	83
4.18 ผลการทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายใต้ Domain ที่ระบบรองรับ	83
4.19 แสดงผลการทดสอบเว็บแอปพลิเคชันผู้ใช้งาน	88
4.20 แสดงผลการทดสอบเว็บแอปพลิเคชันผู้ดูแลระบบ	88
4.21 ผลการทดสอบความเร็วในการจัดหมวดหมู่ให้กับเนื้อหาของเว็บแอปพลิเคชัน	89
4.23 ผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ใช้แบบสอบถามในด้านการออกแบบ	96
4.24 ผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ใช้แบบสอบถามในด้านการออกแบบ	96

สารบัญ

รูปที่	หน้า
2.1 ตัวอย่างของ Word Vector ของ คำว่า aardvark, a, at, zebra	6
2.2 Random Forest	6
2.3 K-Nearest Neighbor	7
2.4 ส่วนประกอบของ Neural Network	8
2.5 RESTful API	9
2.6 ผลการทดลองของงานวิจัยจากการใช้ Text Feature ที่แตกต่างกัน	11
2.7 ผลการทดลองของงานวิจัยจากการใช้ Vector Generator ที่แตกต่างกัน	12
2.8 Amazon Comprehend	12
3.1 โครงสร้างทางสถาปัตยกรรมของระบบ	19
3.2 เครื่องมือในโครงสร้างทางสถาปัตยกรรมของระบบ	20
3.3 รายการ Topic ที่ใช้ในการสืบสารในระบบ	21
3.4 การ Monitor การใช้งานทรัพยากร่าง ๆ บน Cloud Run	21
3.5 Sequence Diagram จำลองการส่งประมวลผลการวิเคราะห์หมวดหมู่ของบทความ	22
3.6 Sequence Diagram ของ DevOps Life Cycle	23
3.7 Sequence Diagram ของ Update New AI-Model Process	23
3.8 Navigation Map ของเว็บไซต์สำหรับผู้ใช้งาน	24
3.9 Navigation Map ของเว็บไซต์สำหรับผู้จัดการระบบ	24
3.10 Data Processing	25
3.11 Use Case Diagram	28
3.12 Sequence diagram ของการสมัครสมาชิก	29
3.13 Sequence diagram ของการเข้าสู่ระบบ	31
3.14 Sequence diagram ของการตรวจสอบข้อมูลໂປຣັບເປົດ	33
3.15 Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link	35
3.16 Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบุคลากร	37
3.17 Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link ผ่าน Api-Key	39
3.18 Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบุคลากรผ่าน Api-Key	41
3.19 Sequence diagram ของการดูคำร้องขอการวิเคราะห์หมวดหมู่ของบทความ	43
3.20 Sequence diagram ของการค้นหาบทความโดยใช้ตัวกรองต่าง ๆ	45
3.21 Sequence diagram ของการค้นหาบทความโดยใช้เลขอ้างอิงในเนื้อหา	46
3.22 โครงสร้างฐานข้อมูล	47
3.23 หน้า Landing	50
3.24 Landing page ที่แสดงแต่ละองค์ประกอบ	50
3.25 หน้า Landing ในส่วนของ Overall	51
3.26 หน้า Landing ในส่วนของ Method	51
3.27 หน้า Landing ในส่วนของ Feature	52
3.28 หน้า Login	52
3.29 กรอก Username หรือ Password ผิดพลาด	53
3.30 หน้า Register ขั้นตอนกรอก Email	53
3.31 กรอก Email ผิดพลาด	53
3.32 หน้า Register ขั้นตอนกรอก Password	54
3.33 กรอก Password ไม่ตรงกัน	54
3.34 หน้า Register ขั้นตอนกรอกชื่อและนามสกุล	54
3.35 กรอกชื่อและนามสกุลเป็นภาษาอื่น	55
3.36 หน้า Register ขั้นตอนกรอกข้อมูลของผู้ใช้งาน	55

3.37	กรอก Password ไม่ตรงกัน	55
3.38	หน้า Register เมื่อสมัครสมาชิกสำเร็จ	56
3.39	หน้า Home	56
3.40	หน้า Home ส่วนที่เกี่ยวข้องกับการวิเคราะห์หมวดหมู่ของสมาชิก	57
3.41	หน้า Home ส่วนแสดงข่าวอื่น ๆ เพิ่มเติม	57
3.42	หน้ากรณีที่สมาชิกต้องการวิเคราะห์หมวดหมู่ของบทความด้วย Web Link	58
3.43	กรณีที่กรอก Web Link ไม่ถูกต้อง	58
3.44	หน้ากรณีที่สมาชิกต้องการวิเคราะห์หมวดหมู่ของบทความด้วยเนื้อหาในบทความ	59
3.45	กรณีที่กรอกข้อมูลไม่ครบถ้วน	59
3.46	หน้า My Ticket	60
3.47	หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ	61
3.48	หน้า Home ของผู้ดูแลระบบ	61
3.49	หน้า Upload ไม่เดลไฟล์	62
3.50	Class Diagram ของระบบ	63
4.1	ตัวอย่างการดึงเนื้อหาบทความจากไฟล์ HTML ด้วยการใช้ Python BeautifulSoup	65
4.2	ตัวอย่างการแปลงเนื้อหาจากภาษาไทยเป็นภาษาอังกฤษ	65
4.3	ตัวอย่างชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาไทย	66
4.4	ตัวอย่างชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาอังกฤษ	67
4.5	ตัวอย่างชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยที่ผ่านการเตรียมข้อมูลแล้ว	68
4.6	ตัวอย่างชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษที่ผ่านการเตรียมข้อมูลแล้ว	68
4.7	การกระจายตัวของจำนวนคำในแต่ละชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย	68
4.8	ลักษณะของจำนวนคำในเนื้อหาที่เป็นภาษาไทยของแต่ละหมวดหมู่หลังทำความสะอาดชุดข้อมูล	69
4.9	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่อัชญากรรม	70
4.10	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่กีฬา	70
4.11	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่การเมือง	70
4.12	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่เศรษฐกิจ	71
4.13	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่บันเทิง	71
4.14	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่เทคโนโลยี	71
4.15	100 อันดับคำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของแต่ละหมวดหมู่	72
4.16	การกระจายตัวของจำนวนคำในแต่ละชุดข้อมูลที่มีเนื้อหาเป็นอังกฤษ	73
4.17	ลักษณะของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่หลังทำความสะอาดชุดข้อมูล	74
4.18	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่อัชญากรรม	74
4.19	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่กีฬา	75
4.20	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่การเมือง	75
4.21	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่เศรษฐกิจ	75
4.22	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่บันเทิง	76
4.23	10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่เทคโนโลยี	76
4.24	100 อันดับคำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่	77
4.25	ตัวอย่าง Text Representation ของชุดข้อมูลที่ได้จากการทำ TF-IDF	78
4.26	ผลลัพธ์จากการทำ t-SNE ด้วยข้อมูลที่ได้จากการทำ TF-IDF	79
4.27	กราฟแสดงค่า Accuracy และ Loss ของโมเดล LSTM สำหรับภาษาไทย	81
4.28	กราฟแสดงค่า Accuracy และ Loss ของโมเดล LSTM สำหรับภาษาอังกฤษ	81
4.29	หน้าแรก	84
4.30	หน้าเข้าสู่ระบบ	85
4.31	หน้าเข้าสู่ระบบ กรณีเข้าสู่ระบบผิดพลาด	85

4.32	หน้าหลัก	86
4.33	หน้าสร้างคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link	86
4.34	กรณีส่งคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link สำเร็จ	87
4.35	หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ	87
4.36	รายการ Unit tests ภายใน General Backend	89
4.37	สัดส่วนเพศของผู้ตอบแบบสอบถาม	93
4.38	สัดส่วนอายุของผู้ตอบแบบสอบถาม	93
4.39	สัดส่วนอาชีพของผู้ตอบแบบสอบถาม	94
4.40	สัดส่วนผู้ตอบแบบสอบถามที่เคยค้นหาเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยหมวดหมู่	94
4.41	สัดส่วนผู้ตอบแบบสอบถามที่เคยจัดหรือจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยตนเอง	95
4.42	การให้คะแนนความพึงพอใจต่อเว็บไซต์ในด้านการออกแบบของผู้ตอบแบบสอบถาม	95
4.43	การให้คะแนนความพึงพอใจต่อเว็บไซต์ในด้านประสิทธิภาพของผู้ตอบแบบสอบถาม	96

บทที่ 1 บทนำ

1.1 ที่มาและความสำคัญ

การจัดหมวดหมู่ คือ การจัดกลุ่มบทความที่มีเนื้อหาสัมพันธ์กันอยู่ภายใต้กลุ่มเดียวกันตามมาตรฐานที่เข้าใจร่วมกัน เพื่อจ่ายต่อการค้นหาข้อมูลและสำรวจเนื้อหาในเรื่องที่เกี่ยวข้อง [1] ซึ่งการจัดหมวดหมู่ค่อนข้างมีความสำคัญอย่างมาก เนื่องจากในปัจจุบันมีบทความที่เพิ่มมากขึ้นเรื่อยๆ โดยเฉพาะในโลกออนไลน์ซึ่งเป็นแพลตฟอร์มที่สามารถรับชมบทความได้ตลอดเวลา เพราะฉะนั้นจึงมีการจัดหมวดหมู่ให้กับบทความโดยเฉพาะเนื้อหาออนไลน์เพื่อให้สะดวกต่อผู้ใช้งานหากต้องการค้นหาที่มีความเกี่ยวข้องกันหรือทำให้ผู้อ่านสามารถเข้าใจได้อย่างรวดเร็วว่าเนื้อหาออนไลน์นั้นมีเนื้อหาเกี่ยวข้องกับอะไร โดยการจัดหมวดหมู่มี 2 วิธีหลักด้วยกัน ได้แก่ การจัดหมวดหมู่ด้วยมนุษย์และการจัดหมวดหมู่แบบอัตโนมัติ

ในปัจจุบันเริ่มมีนักพัฒนาให้ความสนใจกับการจัดหมวดหมู่แบบอัตโนมัติ เนื่องจากจำนวนเนื้อหาออนไลน์ที่มีจำนวนมากขึ้น ส่งผลให้การจัดหมวดหมู่อย่างรวดเร็วและมีประสิทธิภาพต้องใช้ทรัพยากรมนุษย์เพิ่มขึ้นเมื่อใช้วิธีการจัดหมวดหมู่ด้วยมนุษย์ [2] ซึ่งการจัดหมวดหมู่แบบอัตโนมัติมีวิธีการที่หลากหลาย แต่นั่นในวิธีการที่นักพัฒนาให้ความสนใจ คือ การจัดหมวดหมู่แบบอัตโนมัติโดยใช้เทคโนโลยีปัญญาประดิษฐ์ หรือ Artificial Intelligence (AI-based tagging) เนื่องจากเทคโนโลยีปัญญาประดิษฐ์สามารถจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ที่มีจำนวนมากเพื่อจัดหมวดหมู่ได้อย่างรวดเร็วและมีประสิทธิภาพ นอกจากนี้ยังสามารถจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ได้อย่างสม่ำเสมอ แม้จะมีแหล่งที่มาหลากหลายหลายแหล่งก็ตาม [3]

คณะผู้จัดทำฯ สนใจที่จะพัฒนาเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทย โดยใช้เทคโนโลยีปัญญาประดิษฐ์ หรือ Artificial Intelligence (AI-based tagging) ในการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์จากนั้นคัดเลือกและจัดทำจะพัฒนาเว็บแอปพลิเคชันให้เป็นแพลตฟอร์ม สำหรับการค้นหาเนื้อหาออนไลน์ตามหมวดหมู่ที่ผู้ใช้งานต้องการ เพื่อที่จะสามารถนำไปต่อรองหรือใช้ประโยชน์ได้จริงและสะดวกมากยิ่งขึ้น

1.2 วัตถุประสงค์

- เพื่อพัฒนาแพลตฟอร์มสำหรับการวิเคราะห์เนื้อหาเพื่อจัดหมวดหมู่ โดยใช้เนื้อหาภายในของ Web Link หรือเนื้อหาของบทความที่อยู่ในรูปแบบของภาษาไทย
- เพื่อพัฒนาแพลตฟอร์มสำหรับค้นหา Web Link หรือเนื้อหาตามหมวดหมู่ที่ต้องการได้ ทำให้ผู้ใช้งานสามารถนำไปประยุกต์ใช้ได้อย่างสะดวกมากยิ่งขึ้น

1.3 ขอบเขตของโครงงาน

พัฒนาระบบจัดหมวดหมู่ของเนื้อหาในบทความต่างๆ โดยใช้เทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence) ในการวิเคราะห์เนื้อหาของ Web Link หรือเนื้อหาของบทความที่อยู่ในรูปแบบของภาษาไทย เพื่อให้ผู้ใช้งานสามารถค้นหาข้อมูลได้อย่างเป็นระบบจากการจัดหมวดหมู่ในเนื้อหาของบทความโดยใช้เทคโนโลยีปัญญาประดิษฐ์ (AI Platform Thai Content Tagging) ซึ่งจะมีการออกแบบ Architecture ที่ใช้ระบบ Pub/Sub ในการสื่อสารระหว่างระบบ เพื่อให้สามารถรองรับการใช้งานในปริมาณมากได้ โดยที่เว็บแอปพลิเคชันของเราจะมี Feature ต่างๆ ดังนี้

- ผู้ใช้งานสามารถค้นหาเนื้อหาที่สนใจได้โดยการใช้คำค้นหาแบบหมวดหมู่ที่จัดเตรียมไว้ โดยมีหมวดหมู่ไม่น้อยกว่า 5 หมวดหมู่
- ผู้ใช้งานสามารถค้นหาเนื้อหาที่สนใจได้โดยการใช้เลขอ้างอิงเนื้อหาสำหรับการค้นหา
- ผู้ใช้งานที่ลงทะเบียนแล้วสามารถส่ง Web Link หรือบทความที่มีมาบนเว็บแอปพลิเคชัน เพื่อทำการจัดหมวดหมู่ของเนื้อหานั้น
- ผู้ใช้งานสามารถสร้างคำร้องขอเพื่อจัดหมวดหมู่ของเนื้อหาผ่านทาง API-Key ได้
- Web Link ที่ส่งเข้ามาประมวลผลจะต้องอยู่ภายใต้โดเมนเดียวกันไม่น้อยกว่า 5 Domain
- ผู้จัดการระบบสามารถเพิ่มและจัดการหมวดหมู่ได้โดยตนเองผ่านหน้า Frontend

1.4 เนื้อหาทางวิศวกรรมที่เป็นต้นฉบับ

โครงการนี้พัฒนาขึ้นจากการใช้ความรู้ด้านการออกแบบฐานข้อมูล การออกแบบระบบการทำงานแบบ Event-Driven และการออกแบบระบบการทำงานแบบ Microservice การพัฒนา โมเดลสำหรับทำงานของหมวดหมู่ของเนื้อหา รวมไปถึงการใช้ความรู้ทางด้านวิศวกรรมซอฟต์แวร์ในการออกแบบ Diagram ต่างๆ เพื่อให้การสื่อสาร และวางแผนการทำงานสามารถทำได้อย่างมีประสิทธิภาพ เพื่อใช้

ในการสร้างแพลตฟอร์มสำหรับการจัดหมวดหมู่ของเนื้อหาออนไลน์ โดยมีการใช้ Artificial Intelligence ในการจัดหมวดหมู่ รวมไปถึงการสร้างระบบการจัดการหลังบ้านเพื่อให้ผู้จัดการระบบสามารถเพิ่ม และจัดการหมวดหมู่ได้ด้วยตนเอง สามารถทำการอัปเดตโน้มถ่ายที่ใช้ในการจัดหมวดหมู่ได้ และสามารถติดตามความต้องการของผู้ใช้งานได้ ผ่านระบบ Backend ที่พัฒนาด้วยภาษา Java และแสดงผลหน้าเว็บด้วยภาษา Javascript ผ่าน NuxtJs Framework

1.5 ประโยชน์ที่คาดว่าจะได้รับ

แพลตฟอร์มนี้สามารถตอบสนองผู้ใช้งานที่ต้องการค้นหาเนื้อหาโดยใช้หมวดหมู่เป็นคำสำคัญในการค้นหา เพื่อค้นหาเนื้อหาต่างๆ จากหลายแหล่งข้อมูลที่ตรงกับความต้องการของผู้ใช้งาน อีกทั้งยังสามารถจำแนกหมวดหมู่ของเนื้อหาจาก Web Link และเนื้อหาในบทความในรูปแบบภาษาไทยได้ โดยสามารถนำไปประยุกต์ใช้ได้อย่างหลากหลาย เช่น การทำระบบจัดหมวดหมู่อัตโนมัติของเว็บบอร์ดต่างๆ หรือแม้กระทั่งการจัดหมวดหมู่ของข้อมูลที่มีจำนวนมหาศาลจากคลังข้อมูลที่เก็บไว้ เพื่อทำให้เกิดความสะดวกต่อการนำเสนอจำนวนมากขึ้น

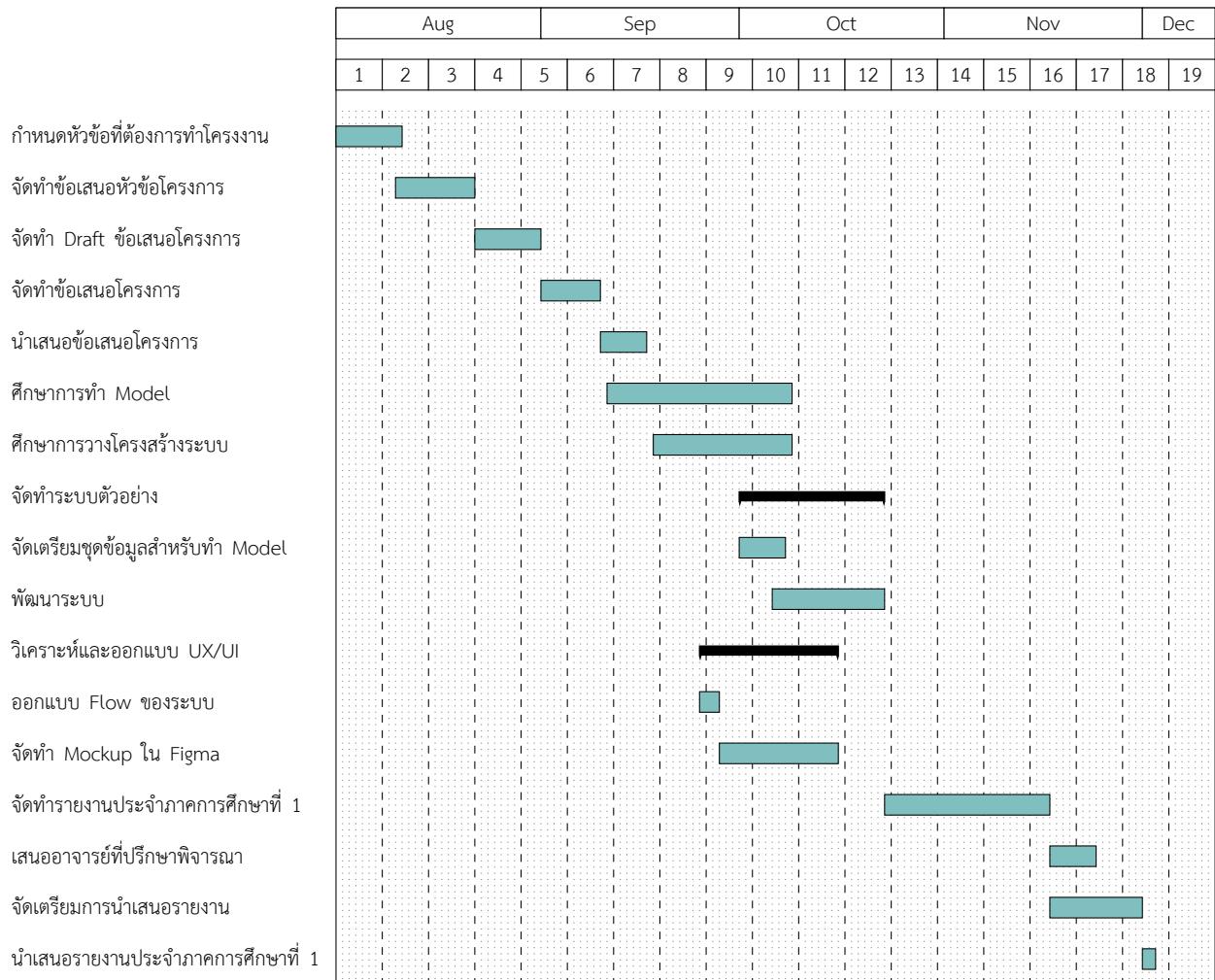
1.6 ขั้นตอนการทำงานและระยะเวลาการดำเนินงาน

1.6.1 ขั้นตอนการทำงาน

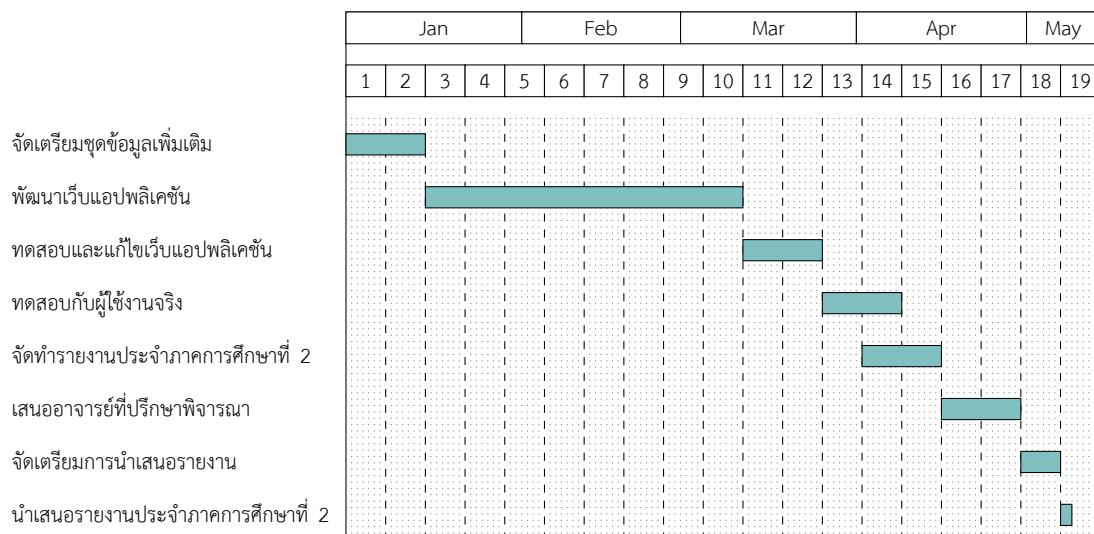
1. กำหนดหัวข้อที่ต้องการทำโครงการ
 - เลือกหัวข้อโครงการที่สนใจและศึกษาหลักการที่เกี่ยวข้องในการทำโครงการนี้
2. จัดทำข้อเสนอหัวข้อโครงการ (Project Idea)
3. จัดทำข้อเสนอโครงการ (Project Proposal)
4. นำเสนอข้อเสนอโครงการ (Proposal Presentation)
5. ศึกษาเทคโนโลยีและซอฟต์แวร์ที่ใช้
 - ศึกษา Artificial Intelligence, Frontend และ Backend
6. ศึกษาการวางแผนสร้างระบบ
7. จัดทำระบบตัวอย่าง
 - เตรียมข้อมูลสำหรับทำ Artificial Intelligence และทำการพัฒนาระบบ
8. วิเคราะห์และออกแบบ UX/UI
 - ออกแบบ Flow ของระบบและ UX/UI ของเว็บแอปพลิเคชัน จากนั้นจัดทำ Mockup ใน Figma
9. จัดทำรายงานประจำภาคการศึกษาที่ 1
 - จัดทำรายงานบทที่ 1 โดยใช้ข้อมูลจากข้อเสนอโครงการ
 - ศึกษาค้นคว้าข้อมูลเพิ่มเติมเพื่อจัดทำรายงานบทที่ 2
 - ออกแบบระบบเพื่อจัดทำรายงานบทที่ 3
 - จัดทำบทที่ 4 บางส่วน
10. นำเสนอรายงานประจำภาคการศึกษาที่ 1
11. จัดเตรียมข้อมูลสำหรับ Train และ Test โน้มถ่ายเพิ่มเติมจากภาคการศึกษาที่ 1
12. พัฒนาเว็บแอปพลิเคชันต่อจากภาคการศึกษาที่ 1
13. ทดสอบและแก้ไขเว็บแอปพลิเคชัน
14. ทดสอบกับผู้ใช้งานจริง
15. จัดทำรายงานฉบับสมบูรณ์
16. นำเสนอโครงการ

1.6.2 ระยะเวลาการดำเนินงาน

ตารางที่ 1.1 ตารางแสดงเวลาดำเนินงานของภาคการเรียนที่ 1/2565



ตารางที่ 1.2 ตารางแสดงเวลาดำเนินงานของภาคการเรียนที่ 2/2565



1.7 ผลการดำเนินงาน

1.7.1 ผลการดำเนินงานในภาคการศึกษาที่ 1

1. รูปเล่มรายงาน 4 บท (บทที่ 1 2 3 และ 4 บางส่วน)
2. ออกแบบโครงสร้างของเว็บแอปพลิเคชัน
 - Functional Requirements
 - แผนผังภาพรวมระบบ
 - Architecture Diagram
 - แผนผังการทำงานของเว็บแอปพลิเคชัน
 - Use Case Diagram
 - API Specification
 - แบบจำลองการทำงานระหว่างผู้ใช้งานกับระบบ
 - Navigation Map
 - User Interface and User Experience Design
3. ออกแบบฐานข้อมูล
 - ER Diagram
 - Data Dictionary
4. สร้างโมเดลสำหรับการจัดหมวดหมู่ 1 – 2 หมวดหมู่
5. จัดทำระบบต่อไปนี้
 - สามารถรับ-ส่งข้อมูลผ่าน Pub/Sub ได้
 - สามารถประมวลผลเนื้อหาและจัดหมวดหมู่ได้ 1 หมวดหมู่

1.7.2 ผลการดำเนินงานในภาคการศึกษาที่ 2

1. รายงานฉบับสมบูรณ์
2. พัฒนาเว็บแอปพลิเคชันและระบบ
3. สร้างโมเดลสำหรับการจัดหมวดหมู่ทุกหมวดหมู่ ซึ่งไม่น้อยกว่า 5 หมวดหมู่
4. เชื่อมต่อเว็บแอปพลิเคชันเข้ากับระบบและฐานข้อมูล โดยมี Artificial Intelligence ประมวลผลร่วม

บทที่ 2 ทฤษฎีความรู้และงานที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 Natural Language Processing (NLP)

Natural language processing (NLP) หรือการประมวลผลภาษาธรรมชาติ เป็นวิทยาการแขนงหนึ่งในหมวดหมู่ของเทคโนโลยีปัญญาประดิษฐ์หรือ Artificial Intelligence ซึ่งช่วยให้คอมพิวเตอร์สามารถเข้าใจ ตลอดจนตีความและใช้งานภาษาปกติที่มนุษย์สื่อสารได้ โดยเทคโนโลยี NLP นี้มีรากฐานจากวิทยาการหลักหลายสาขาด้วยกัน โดยเฉพาะด้านวิทยาการคอมพิวเตอร์ (Computer Science) และภาษาศาสตร์เชิงคำนวน (Computational Linguistics) เพื่อวัตถุประสงค์ในการปิดซ่อนว่าทางการสื่อสารระหว่างมนุษย์และระบบคอมพิวเตอร์ [4]

การทำงานขั้นพื้นฐานของ NLP ประกอบด้วยการทำางดังต่อไปนี้

- Tokenization

การแบ่งคำออกเป็นคำ ๆ อย่างถูกต้องตามหลักภาษา

- Parsing

กระบวนการในการระบุโครงสร้างของข้อความ โดยการวิเคราะห์คำที่เป็นส่วนประกอบ ด้วยหลักไวยากรณ์ของภาษา ผลลัพธ์ที่ได้ออกมาจะเป็นโครงสร้างแบบต้นไม้ เรียกว่า Parse Tree

- Lemmatization/Stemming

การแปลงคำให้อยู่ในรูปแบบเดิม (Lemmatization) เช่น am are is was เปรียบเป็น be และการตัดส่วนขยาย (Stemming) ของคำจะทำการตัดบางส่วนของคำทิ้ง เช่น ในภาษาอังกฤษ จะตัด s es ing หรือ ed

- Part-Of-Speech Tagging

สิ่งที่อธิบายว่าในประโยคหนึ่ง คำ ๆ นั้นถูกใช้อย่างไร สามารถแบ่งได้เป็น 8 อย่างหลัก ๆ ได้แก่ คำนาม คำสรรพนาม คำคุณศัพท์ คำกริยา คำกริยาวิเศษณ์ คำบุพพท คำสันธาน และคำอุทาน

- Language detection

การตรวจสอบภาษาว่าเป็นภาษาอะไร

- Identification of Semantic Relationships

การระบุความสัมพันธ์ของคำต่าง ๆ ในประโยค

2.1.2 Word Embedding

เป็นการแปลงชุดของคำให้กลายเป็นตัวเลข โดยการคำนวณความน่าจะเป็นของคำและบริบทภายในประโยค ซึ่งนำเสนอด้วยแบบของ Vector และได้มีการพัฒนาตัวแบบจำลองสำหรับการทำ Word Embedding เพื่อให้ใช้งานได้อย่างสะดวก เช่น Word2Vec และ Glove

ในขั้นตอนแรกในการสร้างเวกเตอร์ตัวแทนของข้อความนั้น จะต้องทำการรวมและสร้างรายการของคำศัพท์ (Vocabulary) เพื่อพิจารณาจำนวนคำที่เป็นเอกลักษณ์ไม่ซ้ำกับคำอื่น (Unique Word) ที่ปรากฏอยู่ในข้อความที่ทำการพิจารณาทั้งหมดก่อนเป็นอันดับแรก จากนั้นจึงแบ่งคำแต่ละคำภายในข้อความให้อยู่ในลักษณะของเวกเตอร์ของคำ (Word Vector) ซึ่งเวกเตอร์ดังกล่าวจะมีจำนวนมิติเท่ากับจำนวนคำใน Vocabulary และแต่ละมิติจะเป็นตัวแทนของคำนั้น ๆ ดังนั้น Word Vector ของแต่ละคำจะมีค่าของมิติที่เป็นตัวแทนของคำนั้นเป็น 1 และค่าในมิติอื่น ๆ เป็น 0

เนื่องจาก Word Vector เหล่านี้ เป็นตัวแทนของคำที่แตกต่างกันจึงยังไม่ได้สื่อความหมายของคำคำนั้น ขั้นตอนต่อไปหลังจากสร้างเวกเตอร์ตัวแทนของคำจึงเป็นการแปลงเวกเตอร์นี้ให้สามารถใช้งานเป็นตัวแทนความหมายของคำคำนั้นแทน ทั้งนี้ส่วนมากแล้ว เวกเตอร์ที่ใช้แทนความหมายของคำนั้นจะมีจำนวนมิติดลงมาก (คำที่เป็นเอกลักษณ์ในข้อความทั้งหมดที่นำมาพิจารณากจะมีอยู่เป็นจำนวนมาก Word Vector เหล่านี้ จึงมักจะมีจำนวนมิติสูง) [5]

$$w^{aardvark} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^a = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, w^{at} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots w^{zebra} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

รูปที่ 2.1: ตัวอย่างของ Word Vector ของคำว่า aardvark, a, at, zebra

2.1.3 TF-IDF

เป็นหนึ่งในวิธีหาความสำคัญของคำสำคัญในเอกสาร โดยพิจารณาจากค่าน้ำหนักของคำ แล้วนำมาประยุกต์ใช้กับการค้นคืนเอกสาร Information-retrieval หรือ Text mining โดย TF-IDF มาจากผลลัพธ์ของสองค่านั้น คือ TF (Term Frequency) กับ IDF (Inverse Document Frequency) [6]

$$TFIDF = TF * IDF$$

โดย Term Frequency (TF) จะเป็นความถี่ของคำพิเศษ เพื่อหาว่าแต่ละคำนั้นปรากฏมากเท่าไหร่ในเอกสาร ยิ่งมีการใช้คำพิเศษที่ในเอกสารมากเท่าไหร่ การพิจารณาว่าเอกสารนั้นมีเนื้หาเกี่ยวกับอะไรก็จะยิ่งมากขึ้นเท่านั้น ซึ่งค่า TF สามารถคำนวณได้จากการ [7]

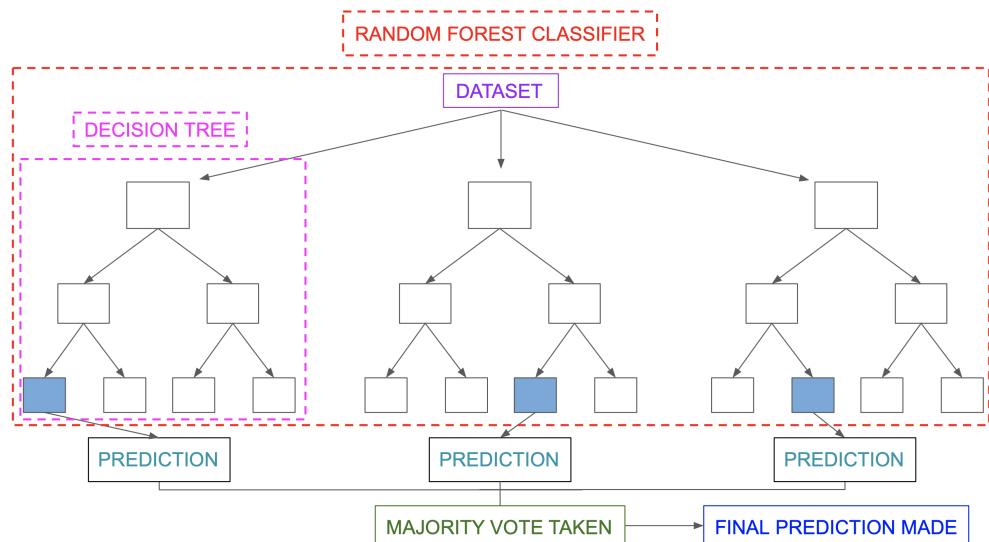
$$TF(word) = \frac{\text{Number of this word in document}}{\text{Number of words in document}}$$

ส่วน Inverse Document Frequency (IDF) เป็นการคำนวณค่าน้ำหนัก (Weight) ความสำคัญของแต่ละคำ โดยคำที่พบเจ้อี้ด้วยบ่อยๆ (ในหลายเอกสาร) จะมีค่า IDF ต่ำ ซึ่งบ่งบอกว่าคำเหล่านั้นจะไม่สามารถตีเสียดูเด่นของเอกสารที่คำเหล่านั้นปรากฏอยู่อ่อนโยนได้ดี ซึ่งค่า IDF สามารถคำนวณได้จากการ [6]

$$IDF(word) = \log \left[\frac{1 + n}{1 + DF(word)} \right] + 1$$

โดยที่ n = จำนวนเอกสารทั้งหมดที่ใช้พิจารณา และ $DF(word)$ = จำนวนเอกสารที่มีคำนั้นปรากฏอยู่

2.1.4 Random Forest



รูปที่ 2.2: Random Forest

คือขั้นตอนวิธีหนึ่งของ Machine Learning ที่นิยมใช้ทั้งกับปัญหาแบบ Regression และ Classification โดย Random Forest เป็นขั้นตอนวิธีพัฒนาต่ออยอดมาจาก Decision Tree ต่างกันที่ Random Forest เป็นการเพิ่มจำนวนต้นไม้เป็นหลายต้น ทำให้ประสิทธิภาพ

การทำงานและพยากรณ์สูงขึ้น Random Forest มีหลักการทำงาน คือ จะแบ่งข้อมูลออกเป็น Decision Tree หลายต้น โดยแต่ละต้นจะได้รับ Feature และข้อมูลที่ไม่เหมือนกันทั้งหมด เพื่อทำให้ได้ต้นไม้ที่มีความหลากหลายและมีความอิสระต่อกันมากขึ้น [8] การทำงานของ Random Forest จะเริ่มต้นจาก

1. ทำการสุ่มเลือก Feature และ Data จากชุดข้อมูลทั้งหมดที่มี
2. สร้าง Decision Tree จากชุดข้อมูลตัวอย่างแต่ละชุดและหาค่าพยากรณ์จากต้นไม้แต่ละต้น
3. เลือกจำนวน Decision Tree ที่ต้องการ จากนั้นทำซ้ำในขั้นตอน 1 และ 2 ในการสร้างต้นไม้
4. หาค่าพยากรณ์ โดยค่าพยากรณ์ที่ได้จะเป็นการที่ Decision Tree แต่ละต้นไม้ห้ามค่าพยากรณ์ของโครงของมัน จากนั้นค่าพยากรณ์สุดท้าย ในกรณีที่ปัญหาเป็นเพื่อ Classification จะใช้วิธี Majority vote โดยค่าพยากรณ์ของ Decision Tree ต้นได้รับค่าผลโหวตมากที่สุดจะถูกเลือกให้เป็นค่าพยากรณ์ของปัญหา แต่ถ้าเป็นปัญหา Regression จะใช้วิธีคำนวณหาค่าเฉลี่ย โดยนำเอาค่าพยากรณ์ของทุก Decision Tree มาคำนวณหาค่าเฉลี่ย เพื่อแสดงเป็นค่าพยากรณ์ของปัญหา

2.1.5 K-Nearest Neighbor (KNN)

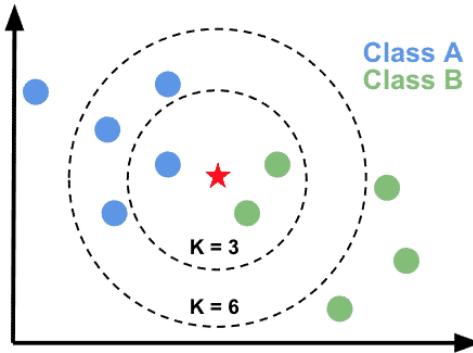
มีหลักการทำงาน คือ จะใช้หลักการเปรียบเทียบความคล้ายคลึงกันของข้อมูลที่สนใจกับข้อมูลอื่นว่ามีความคล้ายคลึงหรืออยู่ใกล้กับข้อมูลมากที่สุด k ตัว จากนั้นจะทำการตัดสินใจว่า คำตอบของข้อมูลที่สนใจนั้นควรเป็นคำตอบเดียวกับข้อมูลที่อยู่ใกล้ที่สุด k ตัวนั้นทั้งนี้ k คือความถี่ของข้อมูลที่อยู่ใกล้กับข้อมูลที่สนใจ [8]

KNN มีขั้นตอนการทำงาน ดังนี้

1. ทำการกำหนดค่า k ซึ่งโดยทั่วไปจะกำหนดให้เป็นเลขคี่ เช่น 3, 5, 7 และ 9 เป็นต้น
2. นำรัศมีที่ต้องการจำแนกมาวัดหาความคล้ายคลึงหรือความต่างกับข้อมูลทั้งหมดในชุดข้อมูล โดยมาตรวัดระยะห่างที่นิยม “ได้แก่ ระยะยุคลิด (Euclidean distance) ตามสมการ

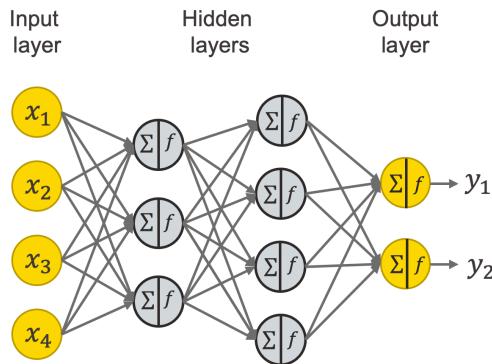
$$dist(p, q) = \sqrt{\sum_{k=1}^n (q_k - p_k)^2}$$

3. เรียงลำดับตามความคล้ายหรือความแตกต่าง จากคล้ายคลึงมากที่สุดไปหาน้อยที่สุด หรือจากแตกต่างน้อยที่สุดไปหามากที่สุด
 4. พิจารณาคำตอบจากจำนวนคลาสคำตอบที่มีมากที่สุดใน k ตัวที่มีความคล้ายมากที่สุด หรือมีความแตกต่างน้อยที่สุด
- ยกตัวอย่าง



รูปที่ 2.3: K-Nearest Neighbor

จากภาพข้างต้น ให้ข้อมูลที่สนใจคือดาวสีแดง วงกลมสีน้ำเงินคือ Class A และวงกลมสีเขียวคือ Class B ถ้าสนใจที่ค่า k เท่ากับ 3 จะได้ว่าข้อมูลที่สนใจจะมีความใกล้เคียงกับ Class B มากกว่า (สัดส่วนความใกล้เคียงของ Class A ต่อ Class B คือ 1:2) ดังนั้นคำตอบของข้อมูลที่สนใจจะเป็น Class B แต่ถ้าสนใจที่ค่า k เท่ากับ 6 ก็จะได้ว่าข้อมูลที่สนใจจะมีความใกล้เคียงกับ Class A มากกว่า (สัดส่วนความใกล้เคียงของ Class A ต่อ Class B คือ 2:1) ดังนั้นคำตอบของข้อมูลที่สนใจจะเป็น Class A



รูปที่ 2.4: ส่วนประกอบของ Neural Network

2.1.6 Neural Network (NN)

โครงข่ายประสาทเทียม เป็นสาขาหนึ่งของปัญญาประดิษฐ์ Artificial Intelligence (AI) เป็นแนวคิดที่ออกแบบระบบโครงข่ายคอมพิวเตอร์ให้เลียนแบบการทำงานของสมองมนุษย์ [8] ส่วนประกอบของ Neural Network ประกอบด้วย 3 ส่วน ได้แก่

1. Input Layer

Layer นี้จะเป็นข้อมูล Input จำนวนของเหตุการณ์ที่มีอยู่กับจำนวนของข้อมูล Input ว่า มีข้อมูลอะไรบ้างที่จะนำเข้ามาคิดใน Model เช่น ถ้าข้อมูลของลูกค้าเป็นข้อมูล Input ที่ประกอบด้วย อายุ เพศ จังหวัดที่อาศัย รวมทั้งสิ้น 4 อย่าง ดังนั้นชั้นข้อมูล Input ก็จะมี 4 โนนด์ ซึ่งอาจจะเรียกปัจจัยที่นำมาวิเคราะห์เหล่านี้ว่าคุณลักษณะ (Feature)

2. Hidden Layer

Layer ที่อยู่ระหว่างกลาง ซึ่งจะมีผลอย่างมากต่อประสิทธิภาพในการเรียนรู้ของ Model ซึ่งใน Hidden Layer นั้นจะมีที่ชั้นก็ได้ และแต่ละชั้นจะมีจำนวนของ Nueral เท่าไหรก็ได้ ซึ่งการเพิ่มขึ้นและจำนวน Neural จะส่งผลต่อการทำงานของ Model ในส่วนของ Hidden Layer มีการทำงานเบรียบสมอื่นส่วนที่เรียบง่ายข้อมูลเชิงลึก หรือ Deep Learning นั่นเอง โดยสิ่งสำคัญใน Hidden Layer อีกประการหนึ่งคือ ทุกด้วยที่ต้องประกอบด้วยฟังก์ชันแบบไม่เป็นเชิงเส้น

3. Output Layer

Layer ที่จะนำเอาข้อมูลจากการคำนวณไปใช้และจำนวนของโนนดใน Layer นี้ขึ้นอยู่กับรูปแบบของข้อมูลออกที่จะเอาไปใช้ ตัวอย่างเช่น ถ้างานที่ทำเป็นสมการถดถอย (Regression) ก็กำหนดให้ Output Layer เป็นแบบ 1 โนนด เพราะต้องการคำตอบเพียงค่าเดียว ถ้าเป็นหลายค่าก็เพิ่มไปตามที่ต้องการ เช่น ในบางงานอาจจะทำนายหาตำแหน่งของภาพในแกน x และ y พร้อมๆ กัน ในการนี้ต้องกำหนดขั้นข้อมูลออกเป็น 2 โนนด เป็นต้น

Neural Network ใช้กระบวนการเรียนรู้ข้อมูลโดยการปรับค่าน้ำหนักเป็นค่าที่เท่ากันทั่วไปที่สุด Neural Network มีการเรียนรู้ 2 แบบ คือ

- Supervised Learning เป็นการเรียนแบบที่มีการตรวจสอบคำตอบ เพื่อให้ Neural Network ปรับตัว ชุดข้อมูลที่ใช้สอน Neural Network จะมีคำตอบป่าวิเคราะห์ตรวจสอบว่า Neural Network ให้คำตอบที่ถูกหรือไม่ ถ้าคำตอบไม่ถูก Neural Network ก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น
 - Unsupervised Learning เป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจสอบว่าถูกหรือผิด Neural Network จะจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูลผลลัพธ์ที่ได้ Neural Network จะสามารถจัดหมวดหมู่ของข้อมูลได้

กระบวนการเรียนรู้ของ Neural Network ถูกพัฒนาขึ้นหลักหลายวิธี เพื่อรับรู้คุณสมบัติในการใช้งานต่าง ๆ วิธีการที่นิยมใช้มากที่สุดคือ Error correction และ Nearest neighbor

Error correction จะเป็น Back propagation ซึ่งมีการเรียนรู้ของ Model เกิดขึ้นเมื่อเราค่าที่ได้จากการคำนวณในของ Forward Propagation มาเทียบกับค่าของข้อมูลอกรที่เกิดขึ้นจริง (Ground Truth) ค่าความผิดพลาดที่เกิดขึ้นเรียกว่า Cost Loss Error หรือ Residual

ดังนั้นกำหนดให้ข้อผิดพลาดของโหนด k (Error : e_k) สามารถคำนวณได้จากค่าความต่างของผลลัพธ์ (y) ของโหนด k ในรอบที่ n แทนด้วยสัญลักษณ์ $y_{k,n}$ และข้อมูล Output ที่เกิดขึ้นจริงของโหนด k แทนด้วยสัญลักษณ์ y_k^* ดังนั้นค่าความผิดพลาดคำนวณได้จากสมการ

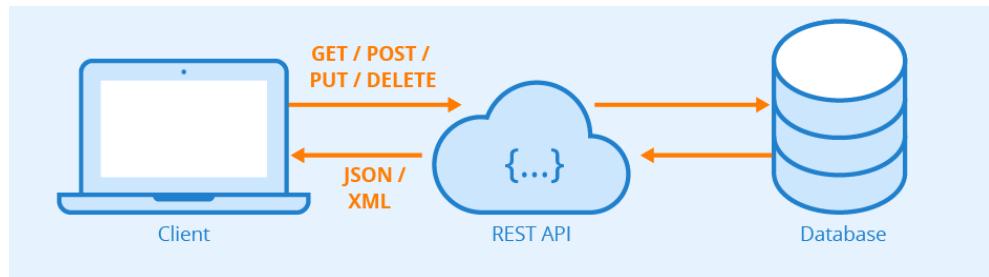
$$e_k = y_{k,n} - y_k^*$$

โดยค่าความผิดพลาด e_k ยิ่งมีค่าใกล้ศูนย์ยิ่งดี Back propagation ทำงานเหมือนสมองคน คือการเรียนรู้จากความผิดพลาด นั่นคือเมื่อรู้ว่าผิดพลาดของโหนด k แล้ว จะนำค่าผิดพลาดนั้นมาคำนวนหากาน้ำหนักใหม่ (w_{new}) ในรอบที่ $n+1$ ของโหนด k ดังสมการ

$$w_{new} = w_{old} - \lambda \frac{\partial E}{\partial w_{old}}$$

โดย λ คือค่าคงที่ในการปรับน้ำหนักซึ่งอาจจะเรียกว่า Step หรือ Learning rate ซึ่งทุก ๆ รอบของการทำงาน จะมีการปรับน้ำหนักใหม่ทุกครั้ง จนกว่าค่าน้ำหนักจะไม่เปลี่ยนแปลงหรือเปลี่ยนแปลงน้อย (Convergence) ถือเป็นค่าน้ำหนักที่เหมาะสมที่สุด (Optimization) จะมีผลทำให้การทำงานหมายผลลัพธ์มีความแม่นยำมากยิ่งขึ้น

2.1.7 RESTful API



รูปที่ 2.5: RESTful API

เป็น Interface ที่ระบบคอมพิวเตอร์ 2 ระบบใช้เพื่อแลกเปลี่ยนข้อมูลผ่านอินเทอร์เน็ตเดียวทั่วโลกด้วย แอปพลิเคชันส่วนใหญ่ต้องสื่อสารกับแอปพลิเคชันภายนอกในอินเทอร์เน็ตและของบุคคลที่สามเพื่อทำงานต่าง ๆ ซึ่งจะอยู่บนมาตรฐานของโปรโตคอล HTTP [9]

ฟังก์ชันพื้นฐานของ RESTful API จะเหมือนกับการท่องอินเทอร์เน็ต โคลอีน์ต์จะติดต่อ กับเซิร์ฟเวอร์โดยใช้ API เมื่อต้องใช้ทรัพยากร นักพัฒนา API อธิบายวิธีการที่โคลอีน์ต์ควรใช้ REST API ในเอกสารประกอบ API ของแอปพลิเคชันเซิร์ฟเวอร์ โดยการเรียกใช้ REST API มีขั้นตอนทั่วไปดังนี้

- โคลอีน์ต์ส่งคำขอไปยังเซิร์ฟเวอร์ โคลอีน์ต์ปฏิบัติตามเอกสารประกอบ API เพื่อจัดรูปแบบคำขอในลักษณะที่เซิร์ฟเวอร์เข้าใจได้
- เชิร์ฟเวอร์รับรองความถูกต้องของโคลอีน์ต์ และยืนยันว่าโคลอีน์ต์มีสิทธิ์ส่งคำขอตั้งกล่าว
- เชิร์ฟเวอร์รับคำขอและประมวลผลเป็นการภายใน
- เชิร์ฟเวอร์ส่งคืนการตอบสนองกลับไปยังโคลอีน์ต์ การตอบสนองมีข้อมูลที่บอกให้ลูกค้าทราบว่าคำขอตั้งกล่าวสำเร็จหรือไม่ การตอบสนองยังรวมถึงข้อมูลใดๆ ที่โคลอีน์ต์ร้องขออีกด้วย

นักพัฒนามักใช้ RESTful API โดยใช้เกณฑ์หรือสิ่งที่ต้องทำกับทรัพยากร โดยวิธีการ HTTP วิธีการ HTTP จะบอกให้เซิร์ฟเวอร์ทราบถึงสิ่งที่ต้องทำกับทรัพยากร โดยวิธีการ HTTP ทั่วไปมี 4 วิธีดังต่อไปนี้

- GET
เพื่อเข้าถึงทรัพยากรที่อยู่ที่ URL ที่ระบุบนเซิร์ฟเวอร์ ซึ่งสามารถแคชคำขอ GET และส่งพารามิเตอร์ในคำขอ RESTful API เพื่อสั่งให้เซิร์ฟเวอร์กรองข้อมูลก่อนส่ง
- POST
เพื่อส่งข้อมูลไปยังเซิร์ฟเวอร์ ซึ่งรวมถึงการแทนข้อมูลพร้อมกับคำขอ การส่งคำขอ POST เดียวกันหลายครั้งมีผลข้างเคียงเหมือนกับการสร้างทรัพยากรเดียวกันหลายครั้ง
- PUT
เพื่ออัปเดตทรัพยากรที่มีอยู่บนเซิร์ฟเวอร์ การส่งคำขอ PUT เดียวกันหลายครั้งในบริการเว็บ RESTful จะให้ผลลัพธ์เหมือนกัน ซึ่งแตกต่างจาก POST

- **DELETE**

เพื่อลบทรัพยากรออก โดยคำขอ DELETE สามารถเปลี่ยนสถานะเชิร์ฟเวอร์ได้ อย่างไรก็ตาม หากผู้ใช้มีการรับรองความถูกต้องที่เหมาะสม คำขอ ก็จะล้มเหลว

2.1.8 Event Driven Architecture (EDA)

เป็นแนวคิดที่ออกแบบระบบโครงข่ายคอมพิวเตอร์ให้สามารถตอบสนองต่อเหตุการณ์ที่เกิดขึ้นได้ โดยการส่งข้อมูลผ่านทางข้อความซึ่งเหตุการณ์ที่เกิดขึ้นนั้นจะถูกส่งไปยัง Event Handler และ Event Handler จะทำการประมวลผลเหตุการณ์นั้น ๆ ต่อไป [10] โดย Event Handler จะมีหน้าที่ดังนี้

- จัดการข้อมูลที่ได้รับจากเหตุการณ์ที่เกิดขึ้น
- ประมวลผลข้อมูลที่ได้รับจากเหตุการณ์ที่เกิดขึ้น
- ส่งข้อมูลไปยัง Event Handler อื่น ๆ ต่อไป

2.1.9 Dependency Injection (DI)

เป็น Design Pattern ที่ใช้ในการเขียนโปรแกรมเพื่อลดความผูกพันของโค้ด โดยการแยกส่วนของการสร้าง Object ออกจาก การใช้งาน Object เพื่อให้การสร้าง Object นั้นสามารถนำไปใช้งานได้หลาย ๆ สถานการณ์ [11] โดยการสร้าง Object จะมีคลาสที่เรียกว่า Container ซึ่งจะสร้าง Object และเก็บไว้ในตัวเอง ซึ่งเมื่อมีการเรียกใช้งาน Object จะทำการส่ง Object ที่ถูกสร้างไว้ให้กับคลาสที่ต้องการใช้งาน โดยไม่จำเป็นต้องสร้าง Object ใหม่ทุกครั้งที่มีการเรียกใช้งาน และสามารถนำ Object ไปใช้งานในสถานการณ์ต่าง ๆ ได้ จึงทำให้สามารถลดความผูกพันของโค้ดลงได้ ซึ่งจะทำให้ได้รับประโยชน์ดังนี้

- ลดความซับซ้อนของโค้ด
- ลดความผูกพันระหว่างโค้ด
- ลดความซ้ำซ้อนของโค้ด
- ทำให้โค้ดมีความยืดหยุ่นมากขึ้น
- ทำให้การทดสอบโค้ดง่ายขึ้น

2.1.10 Continuous Integration (CI)

เป็นกระบวนการที่ใช้ในการพัฒนาซอฟต์แวร์ โดยจะทำการนำ Source Code ของเราผ่านกระบวนการ Testing, Building เพื่อให้ได้เป็นไฟล์ที่พร้อมใช้งานได้ และเมื่อมีการเปลี่ยนแปลง Source Code จะทำการทดสอบและสร้างไฟล์ใหม่อัตโนมัติ [12] ซึ่งจะทำให้ได้รับประโยชน์ดังนี้

- ลดเวลาในการทดสอบและการสร้างไฟล์
- ลดความผิดพลาดในการทดสอบ
- ลดความผิดพลาดในการสร้างไฟล์
- ลดเวลาในการแก้ไขข้อผิดพลาด

2.1.11 Continuous Deployment (CD)

เป็นกระบวนการที่ใช้ในการอัพเดทซอฟต์แวร์ โดยจะทำการนำไฟล์ที่พร้อมใช้งานไปอัพเดทใน Production โดยอัตโนมัติ [12] ซึ่งจะทำให้ได้รับประโยชน์ดังนี้

- ลดเวลาในการอัพเดทซอฟต์แวร์
- ลดความผิดพลาดในการอัพเดทซอฟต์แวร์
- ลดเวลาในการแก้ไขข้อผิดพลาด

2.1.12 Web Scraping

Web Scraping [13] คือ การดึงข้อมูลจากหน้าเว็บไซต์ตามรูปแบบที่กำหนด เพื่อนำข้อมูลไปวิเคราะห์ตามจุดประสงค์ต่าง ๆ เป็นการดึงข้อมูลจาก Element หรือ Tag ของเว็บไซต์โดยเก็บเป็นรูปแบบต่าง ๆ ซึ่งเป็นวิธีการที่มีประโยชน์มากในกระบวนการรวบรวมข้อมูล (Data Collection) เนื่องจากแต่ละเว็บไซต์มีโครงสร้างที่ไม่เหมือนกัน ดังนั้นหากต้องการดึงข้อมูลจะต้องทำ Web Scraping ในกรณีที่ข้อมูลตามโครงสร้างของเว็บไซต์นั้น ๆ เพื่อให้ได้ข้อมูลในส่วนที่ต้องการ การทำ Web Scraping สามารถทำได้ด้วยการเขียนโค้ด หรือกรณีโครงสร้างเว็บไซต์ไม่ซับซ้อนสามารถใช้เครื่องมือช่วยทำ Web Scraping ได้ ซึ่งในบริษัทญี่ปุ่นนี้ได้ทำ Web Scraping ด้วยการเขียนโค้ดโดยใช้ Selenium และ BeautifulSoup เป็นเครื่องมือ

2.1.13 Agile

Agile [13] หมายถึงหลักการและกระบวนการทำงานที่เน้นผลลัพธ์มากกว่าขั้นตอน โดยการทำงานแบบ Agile นั้นเป็นการทำงานเป็นทีมโดยใช้วิธีการทำงานร่วมกัน เพื่อให้ได้ผลลัพธ์ที่ต้องการตามเป้าหมายที่กำหนดไว้ การกำหนดเป้าหมายใน Agile จะเน้นการกำหนดเป้าหมายระยะสั้น ๆ โดยแบ่งโครงการออกเป็นเฟสเล็ก ๆ หรือ Sprint ที่จะช่วยให้สามารถทำงานให้เสร็จสิ้นได้ในระยะเวลาอันสั้น ๆ โดยการทำงานแบบ Agile สามารถมีหลาย Sprint หรือโครงการได้ในเวลาเดียวกัน

ดังนั้นการนำ Agile มาใช้ช่วยเพิ่มความคล่องตัวและความยืดหยุ่นในการพัฒนาโปรเจกต์หรืองาน โดยทำให้สามารถปรับปรุงและปรับเปลี่ยนงานได้อย่างรวดเร็วตามความต้องการของลูกค้าหรือการเปลี่ยนแปลงในสภาพแวดล้อมได้จ่ายชั้น นอกจานี้การทำงานแบบ Agile ยังช่วยสร้างทัศนคติที่ดีในการทำงานร่วมกันในทีม โดยให้พนักงานได้รับการสนับสนุนและการแบ่งปันความรับผิดชอบในการทำงานร่วมกันอย่างเต็มที่ ซึ่งจะช่วยเพิ่มประสิทธิภาพและคุณภาพของงานที่ทำให้สำเร็จได้ตามเป้าหมายที่กำหนดไว้มากขึ้น

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 Machine learning approach to auto-tagging online content for content marketing efficiency:

A comparative analysis between methods and content type

ปัจจุบันนักการตลาดด้านเนื้อหาจะต้องประสบพบเจอกับเนื้อหาที่มีความซับซ้อนมากขึ้น เนื่องจากเนื้อหาที่ไม่มีโครงสร้างที่ชัดเจน ส่งผลให้การวิเคราะห์ การจัดการ และการจำแนกเนื้อหานั้นเป็นไปได้ยากยิ่งขึ้น ดังนั้นจึงมีการเสนอวิธีการจำแนกหมวดหมู่ของเนื้อหาโดยอัตโนมัติ โดยในงานวิจัยนี้ ผู้วิจัยได้ใช้ Machine Learning สำหรับการจัดหมวดหมู่ให้กับเนื้อหาทั้งหมด 3 โมเดลตัวอย่างกัน ได้แก่ Random Forest, K-Nearest Neighbor และ Neural Network ซึ่งจากการวิจัย พบว่า Neural Network มีประสิทธิภาพมากที่สุด โดยได้ค่า F1 อู๊ปที่ 70% สามารถจัดหมวดหมู่ให้กับ Online Content ที่ไม่มีหมวดหมู่ได้ถึง 99.6% และสามารถจัดหมวดหมู่ให้กับ YouTube ที่ไม่มีหมวดหมู่ได้ถึง 96.1% [14]

จากการวิจัยนี้ทำให้ทราบถึงกระบวนการทำการทำ Auto-tagging Online Content ซึ่งคณะผู้จัดทำได้อิงกระบวนการบางส่วนตามงานวิจัยนี้ เนื่องจากลักษณะงานที่มีความคล้ายกัน สิ่งที่คณะผู้จัดทำให้ความสนใจกับงานวิจัยนี้คือ การประเมิน Model การเตรียมข้อมูล และการใช้ Model หลากหลายรูปแบบ โดยงานวิจัยนี้ได้ใช้ F1 Score ค่า Precision และค่า Recall สำหรับการประเมิน Model

ในส่วนของการเตรียมข้อมูล งานวิจัยนี้ได้มีการวิจัยใช้ลักษณะของบทความที่แตกต่างกัน เพื่อที่จะทราบว่าส่วนใดของบทความบ้างที่เหมาะสมกับการนำไปใช้งาน โดยแบ่งลักษณะของบทความเป็นทั้งหมด 3 รูปแบบด้วยกัน ได้แก่ รูปแบบที่ 1 คือ มีเพียงหัวข้อ รูปแบบที่ 2 คือ มีหัวข้อและคำอธิบาย และรูปแบบที่ 3 คือ มีหัวข้อ คำอธิบาย และเนื้อหา ซึ่งจากการวิจัยพบว่า การใช้ส่วนของบทความทั้ง 3 ส่วน ได้แก่ หัวข้อ คำอธิบาย และเนื้อหา จะทำให้ได้ผลลัพธ์ของการทำ Auto-tagging ดีที่สุด ดังนั้นคณะผู้จัดทำจึงเลือกที่จะใช้บทความทั้ง 3 ส่วนโดยอิงจากงานวิจัย

Text feature	Mean F1 Score	Average precision	Average recall
Title only	0.551	0.653	0.549
Title and description	0.426	0.658	0.588
Title, description, and body	0.627	0.666	0.643

รูปที่ 2.6: ผลการทดลองของงานวิจัยจากการใช้ Text Feature ที่แตกต่างกัน

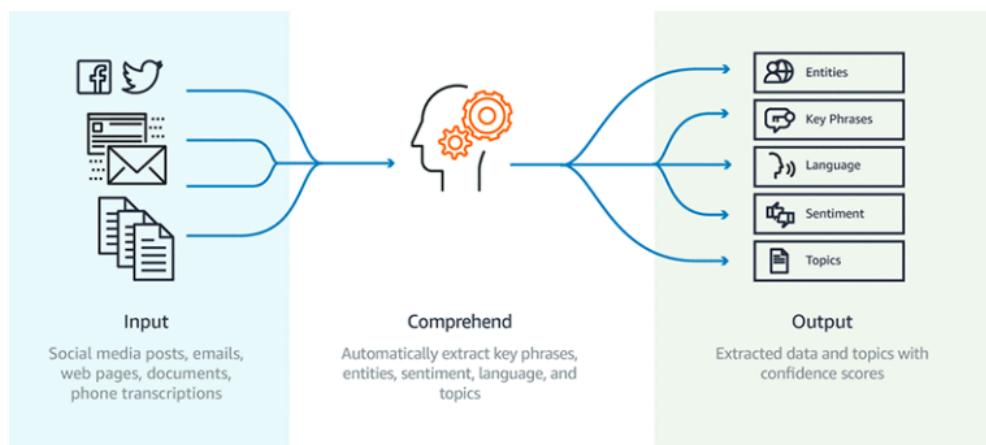
นอกจากลักษณะของบทความ งานวิจัยนี้ได้มีการวิจัยใช้ Algorithm หลากหลายรูปแบบเพื่อเปลี่ยนคำให้เป็น Vector โดยวิจัยทั้งหมด 3 Algorithm ด้วยกัน ได้แก่ TF, TF-IDF และ Doc2Vec ซึ่งจากการวิจัยพบว่า การใช้ TF-IDF จะทำให้ได้ผลลัพธ์ของการทำ Auto-tagging ดีที่สุด ดังนั้นคณะผู้จัดทำจึงเลือกที่จะใช้ TF-IDF สำหรับการแปลงคำเป็น Vector

Feature vector generator	Mean F1 Score	Average precision	Average recall
TF	0.626	0.674	0.610
TF-IDF	0.640	0.667	0.642
Doc2Vec	0.516	0.571	0.514

รูปที่ 2.7: ผลการทดลองของงานวิจัยจากการใช้ Vector Generator ที่แตกต่างกัน

2.3 ผลิตภัณฑ์ที่เกี่ยวข้อง

2.3.1 Amazon Comprehend



รูปที่ 2.8: Amazon Comprehend

เป็นบริการ Natural Language Processing (NLP) ที่จะใช้ Machine Learning มาประมวลผลข้อมูลที่ไม่มีโครงสร้างเพื่อหาข้อมูลเชิงลึกที่มีคุณค่าจากข้อความในเอกสาร สามารถลดความซับซ้อนของการทำงานการประมวลผลข้อมูลโดยการแยกข้อความ วิธีสำคัญ หัวข้อ และอื่น ๆ จากเอกสารเพื่อการจัดเก็บข้อมูลอย่างเป็นระบบ ซึ่ง Amazon Comprehend ก็จะมีกรณีการใช้งานอยู่หลักหลายรูปแบบด้วยกัน [15] เช่น

- ตรวจสอบความรู้สึกของลูกค้าและวิเคราะห์การโต้ตอบของลูกค้า และจัดหมวดหมู่คำขอการสนับสนุนเข้ามาโดยอัตโนมัติ สถิติ สถิติข้อมูลเชิงลึกจากการสำรวจลูกค้าเพื่อปรับปรุงผลิตภัณฑ์
- ให้ความสำคัญกับบริบท โดยการทำให้เครื่องมืออ่านภาษาสามารถจัดทำด้วยนิเวศ เอนทิตี้ และความรู้สึกที่สำคัญได้ ไม่ใช่แค่คำสำคัญเพียงอย่างเดียว
- ระบบอัตโนมัติให้กับการสถิติข้อมูลเชิงลึกจากกลุ่มข้อมูลของข้อสรุปทางกฎหมาย เช่น สัญญาและบันทึกของศาล ยกระดับการรักษาความปลอดภัยให้กับเอกสาร โดยการระบุและแก้ไขข้อมูลที่สามารถระบุตัวบุคคลได้ (PII)
- จำแนกและแยกเนื้อหาออกจากเอกสารบริการทางการเงิน เช่น การเรียกร้องค่าสินไหมทดแทนจากประกันภัย แพคเกจการจำนำของ หรือหากความสมมติฐานระหว่างเหตุการณ์ทางการเงินในบทความเกี่ยวกับการเงิน

ถึงแม้ Amazon Comprehend จะเป็นบริการ Natural Language Processing (NLP) ที่จะใช้ Machine Learning คล้ายกับผลิตภัณฑ์ของเรา แต่ก็มีข้อแตกต่างจากผลิตภัณฑ์ของเราตามตารางดังนี้

ตารางที่ 2.1: เปรียบเทียบผลิตภัณฑ์ระหว่าง Amazon Comprehend กับ AI-based Thai Content Tagging Platform

รายการ	Amazon Comprehend	AI-based Thai Content Tagging Platform
รองรับภาษาไทย	ไม่รองรับภาษาไทย	รองรับภาษาไทย
ความง่ายในการใช้งาน	เหมาะสมสำหรับนักพัฒนา ผู้สร้างเนื้อหา และนักการตลาด	ผู้ใช้งานทั่วไปสามารถใช้ได้
การนำไปประยุกต์ใช้งาน	ผู้ใช้สามารถประยุกต์ได้หลากหลาย	ฟังก์ชันการทำงานจำกัด แต่ต้องมีความเข้าใจในการพัฒนาโปรแกรม โดยนักพัฒนาแพลตฟอร์ม

2.4 เทคนิคและเทคโนโลยีที่ใช้

2.4.1 Development Tools

- JavaScript

เป็นภาษาเขียนโปรแกรมที่ถูกพัฒนาและปฏิบัติตามข้อกำหนดมาตรฐานของ ECMAScript เป็นภาษาระดับสูง คอมไพล์ในขณะที่โปรแกรมรัน (JIT) และเป็นภาษาเขียนโปรแกรมแบบหลายกระบวนการทัศน์ เช่น การเขียนโปรแกรมเชิงขั้นตอน การเขียนโปรแกรมเชิงวัตถุ หรือการเขียนโปรแกรมแบบ Functional ภาษา JavaScript มีไวยากรณ์ที่เหมือนกับภาษา C ใช้งานลื่นไหลเพื่อกำหนดลักษณะของคำสั่ง แตกจากนี้ JavaScript ยังเป็นภาษาที่มีประเภทข้อมูลแบบ Dynamic เป็นภาษาแบบ Prototype-based และ First-class Function [16]

เป็นภาษาโปรแกรมที่นักพัฒนาใช้ในการสร้างหน้าเว็บแบบ Interactive ทำให้เว็บไซต์สามารถตอบสนองกับผู้ใช้งาน ทั้งมี Library ที่ช่วยให้การพัฒนาเว็บไซต์ได้ง่ายขึ้น จำนวนมาก จึงเป็นภาษาที่นิยมในการเขียนเว็บไซต์

- Java

เป็นภาษาเขียนโปรแกรมที่ได้รับความนิยมเป็นอย่างมาก มีการใช้งานอย่างกว้างขวาง เป็นจากมีประสิทธิภาพการทำงานสูง และสามารถรันได้ในทุก Platform โดยไม่ต้องทำการ Compile Code ใหม่ผ่าน JVM (Java Virtual Machine) เนื่องด้วยเป็นภาษาที่มีการพัฒนาต่อเนื่องมาอย่างยาวนาน จึงทำให้มี Library ให้เลือกใช้งานมากมาย โดยรูปแบบการเขียนโปรแกรมภาษา Java จะเขียนโดยใช้หลักการ OOP (Object Oriented Programming) [17]

- Golang

เป็นภาษาเขียนโปรแกรมแบบ Open Source ที่ถูกพัฒนาขึ้นโดยบริษัท Google ในปี 2007 และได้รับความนิยมเพิ่มขึ้นมากขึ้นเรื่อยๆ ในยุคนี้ [18] เนื่องจากประสิทธิภาพในการทำงานที่สูงและใช้ทรัพยากรในการทำงานต่ำ อีกทั้งยังมี Garbage Collector ในการช่วยจัดการกับ Memory นอกจากประสิทธิภาพแล้ว ความง่ายในการเรียนรู้ของภาษา Golang ยังสามารถทำความเข้าใจได้ง่าย ทำให้เป็นภาษาที่ได้รับความนิยมในวงกว้าง

Golang เป็นภาษาที่เหมาะสมสำหรับการทำ Web Development มา กเนื่องจากสามารถใช้สร้างระบบที่รองรับการทำงานใน Scale ใหญ่ที่มี Request จำนวนมากได้ ซึ่ง Golang ถูกออกแบบมาเพื่องานประเภทนี้โดยเฉพาะ ยกตัวอย่างเช่น มี HTTP Package อยู่ใน Standard Library ของ Go โดยที่ไม่ต้องลง Library เพิ่มเติม หรือจะใช้ Framework ต่างๆ ช่วยให้ทำ Web Development ได้ง่ายและสะดวกขึ้น [18]

- Python

เป็นภาษาเขียนโปรแกรมที่ใช้อย่างแพร่หลายในเว็บแอปพลิเคชัน การพัฒนาซอฟต์แวร์ วิทยาศาสตร์ข้อมูล และ แมชชีนเลิร์นинг (ML) นักพัฒนาใช้ Python เนื่องจากมีประสิทธิภาพ เรียนรู้ง่าย และสามารถทำงานบนแพลตฟอร์มต่างๆ ได้มากมาย ทั้งนี้ซอฟต์แวร์ Python สามารถดาวน์โหลดได้ฟรี ผลงานการทำงานร่วมกับระบบทางประดิษฐ์ และเพิ่มความเร็วในการพัฒนา ภาษา Python มีกรอบนิการใช้งานหลายอย่างในการพัฒนาแอปพลิเคชัน ซึ่งรวมถึงตัวอย่างต่อไปนี้ [19]

- การพัฒนาเว็บฝั่งเซิร์ฟเวอร์

การพัฒนาเว็บฝั่งเซิร์ฟเวอร์ประกอบด้วยฟังก์ชัน Backend ที่ซับซ้อน ซึ่งเว็บไซต์ดำเนินการเพื่อแสดงข้อมูลต่อผู้ใช้ ตัวอย่างเช่น เน้นใช้ต้องต่อตอบกับฐานข้อมูล สื่อสารกับเว็บไซต์อื่น และปกป้องข้อมูลเมื่อส่งข้อมูลผ่านเครือข่าย

Python มีประโยชน์สำหรับการเขียนโค้ดฝั่งเซิร์ฟเวอร์ เนื่องจากมี Library จำนวนมากที่ประกอบด้วยตัวอักษรที่เขียนไว้ล่วงหน้าสำหรับฟังก์ชัน Backend ที่ซับซ้อน นักพัฒนาอาจจะใช้ Framework Python ที่หลากหลายซึ่งมีเครื่องมือที่จำเป็นทั้งหมด เพื่อสร้างเว็บแอปพลิเคชันได้เร็วขึ้นและง่ายขึ้น อีกด้วย ตัวอย่างเช่น นักพัฒนาสามารถสร้างโครงสร้างเว็บแอปพลิเคชัน

ได้ภายในไม่กี่วินาที เนื่องจากไม่จำเป็นต้องเขียนขึ้นใหม่ทั้งหมด จากนั้นนักพัฒนาสามารถทดสอบได้โดยใช้เครื่องมือทดสอบของ Framework โดยไม่ต้องพึ่งพาเครื่องมือทดสอบภายนอก

- วิทยาศาสตร์ข้อมูลและแมชชีนเลร์นนิ่ง (ML)

วิทยาศาสตร์ข้อมูลถึงความรู้อันมีคุณค่าจากข้อมูลและแมชชีนเลร์นนิ่ง (ML) จะสอนคอมพิวเตอร์ให้เรียนรู้จากข้อมูลโดยอัตโนมัติและทำนายได้อย่างแม่นยำ นักวิทยาศาสตร์ข้อมูลใช้ Python สำหรับงานด้านวิทยาศาสตร์ข้อมูลต่างๆ ดังต่อไปนี้

- * การแก้ไขและลบข้อมูลที่ไม่ถูกต้อง ซึ่งเรียกว่าการทำความสะอาดข้อมูล
- * การแยกและเลือกคุณสมบัติต่างๆ ของข้อมูล
- * การระบุประเภทข้อมูล ซึ่งเป็นการเพิ่มข้อมูลที่มีความหมายสำหรับข้อมูล
- * การค้นหาสถิติต่างๆ จากข้อมูล
- * การแสดงข้อมูลด้วยภาพโดยใช้แผนภูมิและกราฟ เช่น แผนภูมิเส้น กราฟแท่ง อิสโทแกรม และแผนภูมิวงกลม

นักวิทยาศาสตร์ข้อมูลใช้ไลบรารี Python ML เพื่อฝึกฝนโมเดล ML และสร้างตัวจำแนกที่จำแนกประเภทข้อมูลได้อย่างแม่นยำ บุคลากรในวงการต่างๆ ใช้ตัวจำแนกแบบ Python เพื่อทำงานด้านการจำแนกประเภท เช่น การจำแนกประเภทรูปภาพ ข้อความ และการรับส่งข้อมูลทางเครือข่าย การรู้จำเสียง และการจดจำใบหน้า นักวิทยาศาสตร์ข้อมูลยังใช้ Python สำหรับ Deep Learning ซึ่งเป็นเทคนิค ML ขั้นสูง

- การพัฒนาซอฟต์แวร์

นักพัฒนาซอฟต์แวร์มักใช้ Python สำหรับงานด้านการพัฒนาและการประยุกต์ใช้ซอฟต์แวร์ต่างๆ ดังนี้

- * การติดตามบักในโค้ดของซอฟต์แวร์
- * การสร้างซอฟต์แวร์โดยอัตโนมัติ
- * การดูแลการจัดการโครงการตัวย่อซอฟต์แวร์
- * การพัฒนาต้นแบบซอฟต์แวร์
- * การพัฒนาแอปพลิเคชันบนเดสก์ท็อปโดยใช้ Library ส่วนติดต่อผู้ใช้แบบกราฟิก (Graphical User Interface หรือ GUI)
- * การพัฒนาเกมที่ใช้ข้อมูลแบบง่ายๆ ไปจนถึงวิดีโอยอดน้ำหนักมากขึ้น

- Google Colaboratory

เป็นบริการ Cloud ยกระดับบริการจาก Google Research เป็น IDE ที่อนุญาตให้ผู้ใช้เขียน Source code ในตัวแก้ไขและเรียกใช้จากเบราว์เซอร์ รองรับภาษาเขียนโปรแกรม Python และเน้นงานแมชชีนเลร์นนิ่ง การวิเคราะห์ข้อมูล เป็นต้น Google Colaboratory เป็นบริการ Software as a Service (SaaS) ໂໂສຕ์ໂປຣແກຣມ Jupyter Notebook บน Cloud จาก Google ซึ่งฟังก์ชันที่โดดเด่น ดังนี้ [20]

- แก้ไขและเรียกใช้โค้ดใน Python
- จัดเก็บงานใน Google Drive เพื่อไม่ให้สูญหาย
- แบ่งปัน Notebook กับผู้อื่นได้ (ข้อความ โค้ด ผลลัพธ์ และความคิดเห็น)
- นำเข้า Jupyter Notebook หรือ IPython ได้
- ดาวน์โหลด Colab Notebook ในเครื่องจาก Google Drive ได้

- Figma

เป็นเครื่องมือสำหรับใช้ออกแบบตั้งแต่เว็บไซต์, แอปพลิเคชัน สำหรับ UX/ UI Designer ทั่วโลก หรือใช้สำหรับการแบบโลโก้, artwork ต่างๆ ของสายงาน Graphic Design รวมไปถึงคนทั่วไปที่ใช้ในการออกแบบ Presentation Figma ให้ความสำคัญในเรื่องของการทำงานร่วมกันภายในทีม ทำให้ทีม UX/ UI Designer ทำงานกันได้สะดวกมากขึ้น รวมไปถึงส่งเสริมการทำงานระหว่างทีมที่ช่วยให้ Designer ส่งต่องานกับ Developer ได้ง่ายมากยิ่งขึ้น ซึ่ง Figma จะใช้งานในรูปแบบ browser-based ที่ทุกคนสามารถทำงานพื้นมองกันได้ และมี Features ที่ช่วยให้การส่งต่องานระหว่างทีมทำได้ง่ายขึ้นกว่าเครื่องมือการออกแบบอื่นๆ โดยแบ่ง Feature การทำงานเป็น 4 ด้านดังนี้ [21]

- Collaboration Features

สามารถใช้งานพร้อมกันได้แบบ real-time บนเว็บ โดยไม่ต้องติดตั้งโปรแกรมในเครื่อง และไม่ต้องกด save แม้แต่ครั้งเดียว

- Design Features
มีไฟล์ตัวอย่างออกแบบ ตั้งแต่เริ่มต้นออกแบบจนถึงส่งงานต่อให้ Developer
- Prototyping Features
- Design systems features

2.4.2 Database Tools

- PostgreSQL

เป็น Open Source Object-Relational Database ที่ใช้งานกันอย่างแพร่หลาย มีการพัฒนาให้สามารถรองรับข้อมูลเพิ่มเติมได้ หลายประเภท เช่น UUID, Array, JSON ทำให้สามารถนำไปประยุกต์ใช้ได้หลากหลาย เป็นระบบฐานข้อมูล SQL (Structure Query Language) เช่นเดียวกับ MySQL [22]

2.4.3 DevOps Tool

- Git

ระบบจัดการแก้ไข (Version Control System) ติดตามการเปลี่ยนแปลงของไฟล์ต่าง ๆ ในโปรเจค ช่วยทำให้สามารถทำงานอย่างเป็นระบบ สามารถติดตาม ตรวจสอบ การพัฒนา ดูประวัติการเปลี่ยนแปลงของไฟล์ต่าง ๆ เพื่อดูการเปลี่ยนแปลงเพื่อเกิดปัญหาที่สามารถยกกลับได้ [23]

- Github

เว็บไซต์ที่ให้บริการ Git (Version Control Repository) โดยให้บริการบนออนไลน์แพลตฟอร์ม ที่จะมีไฟล์ที่จะช่วยให้นักพัฒนาคนอื่น ๆ สามารถส่วนร่วมในการทำงานแก้ไขโค้ดที่อยู่ใน Repository ได้ ซึ่งสามารถตั้งค่า Repository เป็น Public ให้ทุกคนสามารถเข้าถึงได้ หรือ Private เพื่อใช้งานเฉพาะกลุ่มได้ [24]

- Docker

แพลตฟอร์มซอฟต์แวร์ที่ช่วยให้นักพัฒนาซอฟต์แวร์สามารถสร้าง ทดสอบ และติดตั้งแอปพลิเคชันผ่านมาตรฐานของคอนเทนเนอร์ซึ่งจะ ช่วยในการจัดการสภาพแวดล้อมของระบบ โค้ด และรันใหม่ ซึ่งเมื่อใช้ Docker จะช่วยให้สามารถติดตั้งแอปพลิเคชันให้เหมาะสมกับทุกสภาพแวดล้อมได้อย่างรวดเร็ว [25]

2.4.4 Library Tools

- Vue

เป็น JS Library ที่รวมเอาข้อดีของ Angular กับ React มารวมกัน มีการประมวลผลที่รวดเร็ว [26] โดยเป็น "Progressive Framework" สำหรับสร้าง User Interface (UI) ซึ่ง Progressive คือการที่ Vue.js ใช้เป็นเหมือนส่วนเพิ่มความสามารถที่เอามาไปใช้งานกับ HTML ซึ่ง Library นี้จะจัดการในส่วนของ View เท่านั้น เพื่อให้สามารถนำ Vue ไปใช้งานร่วมกับ Library อื่น ๆ ได้สะดวก และเหมาะสมสำหรับการทำ Single-Page Applications (SPA) หรือเว็บที่ไม่ต้องเปลี่ยนหน้าบ่อย [27]

- Nuxt.js

เป็น Javascript Framework ที่พัฒนาต่อยอดมาจาก Vue.js ซึ่งเป็นเครื่องมือที่ช่วยให้การพัฒนาและการแสดงผลบนเว็บแอปพลิเคชัน มีความสะดวกสบาย สามารถพัฒนาต่อยอดผลงานได้อย่างมีประสิทธิภาพ [28]

- Pandas

เป็น Library Python แบบ Open Source ที่มีเครื่องมือจัดการและวิเคราะห์ข้อมูลประสิทธิภาพสูง โดยใช้โครงสร้างข้อมูลที่ชื่อ Pandas มาจากคำว่า Panel Data (ชุดข้อมูลหลายมิติ) มีจุดเด่นด้านการวิเคราะห์ข้อมูล (Data Analysis) และการทำความสะอาด (Data Cleaning) ซึ่งเป็น Process ที่สำคัญมากในการทำงานกับข้อมูล โดยมี Feature การทำงานตั้งนี้ [29]

- DataFrame ที่รวดเร็วและมีประสิทธิภาพ พร้อมการสร้าง Index เริ่มต้นและ Index ที่กำหนดเองได้
- เป็นเครื่องมือสำหรับโหลดข้อมูลลงใน In-memory Data Objects จากสกุลไฟล์ต่าง ๆ
- การจัดตำแหน่งข้อมูลและจัดการข้อมูลที่ขาดหายไป
- Reshaping และ Pivoting data
- การทำ label สำหรับการ Slicing, การ Indexing และ Subsetting ชุดข้อมูลที่มีขนาดใหญ่

- โครงสร้างข้อมูลสามารถ Delete หรือ Insert ได้
- จัดกลุ่มตาม Engine เพื่อให้สามารถใช้การดำเนินการ Split-Apply-Combine กับ Data Set
- การ Merging และ Joining ของ Data Set ที่มีประสิทธิภาพสูง
- การสร้าง Range ของวันและความถี่ของการเปลี่ยนแปลง การย้าย Window Statistics การย้าย Window Linear Regressions การ Shift วัน และการ Lagging
- NumPy

เป็น Library ที่ใช้ในการคำนวณทางคณิตศาสตร์ในภาษา Python [30] ซึ่งภายในถูกเขียนด้วยภาษา C จึงทำงานได้เร็วและมีประสิทธิภาพ โดย NumPy มีความสามารถในการจัดการกับอาร์rayหลายมิติและข้อมูลแบบแมทริกซ์
- Natural Language Toolkit

หรือเรียกว่า NLTK เป็น Library สำหรับการประมวลผลภาษาธรรมชาติ (NLP) ของภาษาอังกฤษ โดยเขียนในภาษา Python พัฒนาขึ้นโดย Steven Bird และ Edward Loper NLTK มีฟังก์ชันสำหรับการแบ่งประเภท ตัดคำ กำกับไวยากรณ์ประโยค และอื่น ๆ
- SpaCy

เป็น Library Open Source สำหรับการประมวลผลภาษาธรรมชาติขั้นสูง (NLP) ที่เขียนด้วยภาษา Python พร้อมด้วยองค์ประกอบใน Cython (ส่วนต่อขยาย C ของ Python ออกแบบมาเพื่อให้ C มีประสิทธิภาพเข้ากับโปรแกรม python) จึงเป็น Library ที่ค่อนข้างเร็ว ซึ่ง SpaCy ให้ API ใน การเข้าถึงเครื่องและคุณสมบัติที่สามารถควบคุม โดยการสร้างแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) และ Deep Learning นอกจากนี้ SpaCy ยังสามารถทำงานร่วมกับ Library อื่นได้ เช่น NLTK SpaCy ถูกออกแบบสำหรับการใช้งานและช่วยสร้างแอปพลิเคชันที่ประมวลผลและเข้าใจข้อความจำนวนมาก สามารถใช้ในการสร้างการดึงข้อมูลหรือระบบความเข้าใจภาษาธรรมชาติหรือข้อความก่อนการประมวลผลสำหรับ Deep Learning คุณสมบัติบางอย่าง SpaCy มีให้ได้แก่ การแบ่งคำ (Tokenization) การติดแท็กส่วนของคำพูด (Part-of-Speech Tagging) การจำแนกข้อความ (Sentence Boundary Detection) และการจดจำ-entity ที่มีชื่อ (Named Entity Recognition) [31]
- Scikit-learn

เป็น Library ในการทำ Machine Learning ที่ครุศึกษาเอาไว้ เนื่องจากมีความง่ายและมีประสิทธิภาพในการทำ Predictive Data Analysis เช่น Support Vector Machines(SVMs) Random Forests Gradient Boosting K-means และ DBSCAN ซึ่งใช้อัลกอริทึมเหล่านี้เมื่อเราสร้างและฝึก Model สำหรับทำ Machine Learning [32]
- Keras

เป็น Deep Learning Library ที่ได้รับความนิยมอย่างรวดเร็ว เนื่องจากใช้งานง่ายแต่มีประสิทธิภาพสูงในการรัน Model ซึ่ง Backend ของ Keras มีทั้ง Tensorflow และ Theano ซึ่งจะเป็น Deep Learning Library ที่มีสมรรถนะสูงทั้งคู่ Keras สามารถใช้เพื่อทำงานประมวลผล Regression Classification หรือประมวลผลรูปภาพได้ [32]
- Tensorflow

เป็น Library สำหรับสร้าง Machine Learning Models แบบ Open Source จาก Google สามารถใช้งานได้กับภาษา Python แต่ก็ใช้งานร่วมกับภาษาอื่นๆ เช่น C, Java หรือ Go ได้เช่นกัน และยังมี Community ขนาดใหญ่ ทำให้สามารถค้นหาข้อมูลหรือสอบถามเวลาเจอปัญหาได้ง่าย [32]
- BeautifulSoup

เป็น Library สำหรับทำ Web scraping หรือการดึงข้อมูลผ่านทางหน้าเว็บไซต์ ซึ่งหลังจากดึงข้อมูลเสร็จ จะเข้าสู่กระบวนการสกัด (Extract) เอาเฉพาะข้อมูลที่ต้องการ เพื่อนำมาเก็บไว้ในรูปแบบที่ต้องการเป็นแหล่ง Data Source เพื่อใช้งานต่อไป โดยจะใช้ภาษา Python ในการเขียน Script [33]
- Selenium

เป็นเครื่องมือที่ช่วยให้สร้างโปรแกรมสำหรับการทดสอบเว็บไซต์ [34] โดยจะสร้าง Web Browser จำลองขึ้นมาโดยใช้การเขียนโปรแกรมในภาษาต่าง ๆ เช่น Java Python หรือ Javascript ติดต่อกับไลบรารีของ WebDriver เพื่อเข้าถึงคุณโගรุที่แสดงผ่าน Web Browser ได้ แล้วทำการควบคุมให้คลิกแบบอัตโนมัติ ควบคุมปรับเปลี่ยนตัวเลือกหรือใส่ค่าต่าง ๆ ที่จำเป็นแบบอัตโนมัติ รวมถึงสามารถแก้ไข Code HTML บางส่วนให้เป็นไปตามที่ต้องการเพื่อทดสอบการแสดงผลของเบราว์เซอร์ได้ นอกจากนี้การใช้ Selenium เพื่อสร้าง Web Browser จำลองยังสามารถนำมาใช้ร่วมกับการดึงข้อมูลจากหน้าเว็บไซต์ เพื่อใช้ข้อมูลจากเว็บไซต์นั้น ๆ ได้

2.4.5 Infrastructure Tools

- Google Cloud Platform (GCP)

เป็นบริการ Cloud ที่ให้บริการโดยเบริชต์ Google ซึ่งมีบริการที่ส่วนส่วนเสริมการพัฒนาแอปพลิเคชันให้มีความสะดวกสบายมากยิ่งขึ้น ช่วยให้ลดค่าใช้จ่ายในการลงทุน Server หรือ Hardware ต่าง ๆ ที่ใช้ในการรันแอปพลิเคชัน และมีผู้ดูแลระบบให้ตลอด 24 ชม. อีกทั้งในบางบริการยังมีการคิดเงินแบบ Pay as you go หมายถึงเราจะจ่ายเงินเฉพาะสิ่งที่เราใช้จริง ๆ เท่านั้น ไม่ติดค่าใช้จ่ายขณะที่เราไม่ได้ใช้งาน ซึ่ง Google มีจุดเด่นมากต่อ Network อยู่ทั่วโลกมากกว่า 33 ประเทศ และมีบริการให้เลือกใช้งานมากมาย [35]

- Cloud Pub/Sub

บริการรับส่งข้อมูลระหว่าง Service เป็น Managed Service ที่ให้บริการโดย GCP จะมีแบ่งเป็น 3 ส่วนหลัก ๆ คือ Publisher คือผู้ส่งข้อมูลโดยจะมีการกำหนด topic เพื่อให้ Subscriber หรือผู้รับ รอรับข้อมูลจาก topic นั้น ๆ โดยจะมี Cloud Pub/Sub เป็นสื่อการในการสื่อสารระหว่าง Publisher กับ Subscriber ซึ่งจะเป็นการสื่อสารแบบ asynchronously [36] ซึ่งการนำ Cloud Pub/Sub มาใช้งานยังมีข้อดีอื่น ๆ อีกเช่น

- ไม่ต้องติดตั้ง หรือคูແລນែងจากเป็นบริการให้บริการโดย Google สามารถใช้งานได้ทันที
- มีความเสถียร มั่นใจได้ว่าข้อมูลที่ publish ออกไปจะไม่สูญหาย
- มีการป้องกันการได้รับข้อความซ้ำ

- Google Container Registry (GCR)

บริการจัดเก็บ Image ของ Docker ให้บริการบน GCP มีหน้าที่ในการจัดการเก็บ และทำ Version Control ของ Image ซึ่งจะเป็น Private Registry และสามารถทำ Version ของแต่ละ Image ได้ทำให้มีอิเกิดปัญหาสามารถย้อนกลับไปใช้ Version ก่อนหน้าได้ ซึ่งทำให้สามารถนำ Image ที่อยู่บน Container Registry ไปสร้างเป็น Instance ได้ [37]

- Cloud Run

เป็น Compute Platform แบบ Serverless ที่ให้บริการบน GCP มีลักษณะการให้บริการแบบ PaaS (Platform as a Service) มีความสามารถในการสเกลขึ้นลงได้โดยอัตโนมัติตามจำนวนการใช้งาน ซึ่งทำให้ประหยัดค่าใช้จ่ายในส่วนการดูแล Server ซึ่งสามารถใช้ Image จาก Google Container Registry มาจับเป็นแอปพลิเคชันได้ทันที [38]

- Google Translate

เป็นโปรแกรมแปลภาษาของ Google สามารถแปลได้ทั้งคำ ประโยค เนื้อหาเป็นย่อหน้า จนที่สุดแปลงเว็บก็สามารถทำได้ เป็นโปรแกรมที่สร้างขึ้นโดยใช้กลวิธี อ้างอิงกฎ การนิยามศัพท์ สร้างกลุ่มคำ ใช้สถิติ การบรรจุเรื่องไว้การณ์ภาษาเข้าไปกระบวนการประมวลผลการแปลโดยคอมพิวเตอร์หลายเครื่องนี้ เรียกว่า "Statistical Machine Translation" หรือ "ระบบการแปลภาษาเชิงสถิติ" [39] ซึ่งในที่นี้ดำเนิน Google Translate มาใช้เพื่อแปลเนื้อหาจาก Web Link ที่เป็นภาษาไทยให้เป็นภาษาอังกฤษก่อนที่จะนำไปใช้งาน

2.4.6 Application Testing Tools

- K6

เป็นเครื่องมือที่ใช้ในการทดสอบประสิทธิภาพของเว็บไซต์ โดยสามารถทดสอบได้ทั้งการทดสอบความแข็งแกร่งของเว็บไซต์ ได้หลากหลายรูปแบบ เช่น การทดสอบการรับส่งข้อมูลผ่าน API, จำนวน request ต่อวินาที, จำนวนผู้ใช้งานที่เข้ามาใช้งานพร้อมกัน และอื่น ๆ โดยสามารถเขียนโปรแกรมเพื่อทดสอบได้ด้วยภาษา Javascript [40]

- Lighthouse

เป็นเครื่องมือที่ใช้ในการวัดประสิทธิภาพของเว็บไซต์ โดยสามารถวัดได้ทั้งความเร็วในการโหลดหน้าเว็บไซต์ ความสามารถในการทำงานของเว็บไซต์ ความเข้ากันได้ของเว็บไซต์กับอุปกรณ์ที่ใช้ในการเข้าถึงเว็บไซต์และความปลอดภัยของเว็บไซต์ [41]

2.4.7 Project Management

- ClickUp

เป็น Project Management Software ที่สามารถนำมายังจัดการวางแผนงาน ติดตามงาน ไฟล์เอกสาร แฟช และอีกมากมาย ทำให้สามารถจัดการทุกอย่างได้ในที่เดียว ช่วยทำให้การทำงานมีประสิทธิภาพมากขึ้น ซึ่ง ClickUp สามารถสร้าง

Workflow ของทุกโครงการได้ ทำให้สามารถมองเห็นภาพรวมของงานหรือจัดการโครงการได้อย่างเป็นระบบ สามารถวัดผลการทำงานของแต่ละโครงการในรูปแบบของ Dashboard ได้ ช่วยในการติดตาม Progress ของงาน เพื่อทำให้บรรลุเป้าหมายได้อย่างมีประสิทธิภาพที่สุด และเป็นตัวช่วยในการกระจายงานให้แก่สมาชิกในทีมแต่ละคนได้อย่างมีประสิทธิภาพ [42]

- Miro

เป็นแพลตฟอร์มที่ช่วยให้การทำงานออนไลน์เป็นทีมได้ดียิ่งขึ้น สามารถรวมความคิดแบบเรียลไทม์ได้จากทุกที่ผ่านทางเว็บไซต์ เป็นกระดานร่วมความคิดของแต่ละคน สร้างการมีส่วนร่วมและเก็บทุกความคิดไว้ไม่ให้ตกหล่น และไม่จำเป็นต้องจดรายงานการประชุมเพราะ เนื่องจากใน Miro ทุกอย่างจะถูกบันทึกไว้บนกระดานและสมาชิกในทีมแต่ละคนสามารถแสดงความคิดเห็นให้แต่ละข้อเสนอได้ ทำให้การจัดการความคิดทำได้ง่ายขึ้นและรวดเร็วขึ้น [43]

- Discord

เป็นแพลตฟอร์มที่ใช้งานได้ฟรี โดยเป็นการผสมผสานระหว่าง Slack ที่มีจุดเด่นด้านการสื่อสารที่สามารถสนทนาแบบส่วนตัวและสนทนาแบบกลุ่มได้ กับ Skype ที่มีจุดเด่นด้านการแชทในรูปแบบ Voice Chat และ Video Chat จึงเป็นอีกหนึ่งแพลตฟอร์มที่ได้รับความนิยมอย่างมาก

นอกจากนี้ Discord สามารถสร้างห้องสนทนาไว้พูดคุยกันในกลุ่มเฉพาะหรือเป็นแหล่งพบปะสังสรรค์ของกลุ่มคนที่สนใจในเรื่องเดียวกัน รวมถึงค้นหาหรือเพิ่มเพื่อนเข้ามากลุ่มสนทนาได้อย่างง่ายดาย จึงสามารถใช้เป็นสื่อกลางในการทำงานร่วมกับผู้อื่นได้ [44]

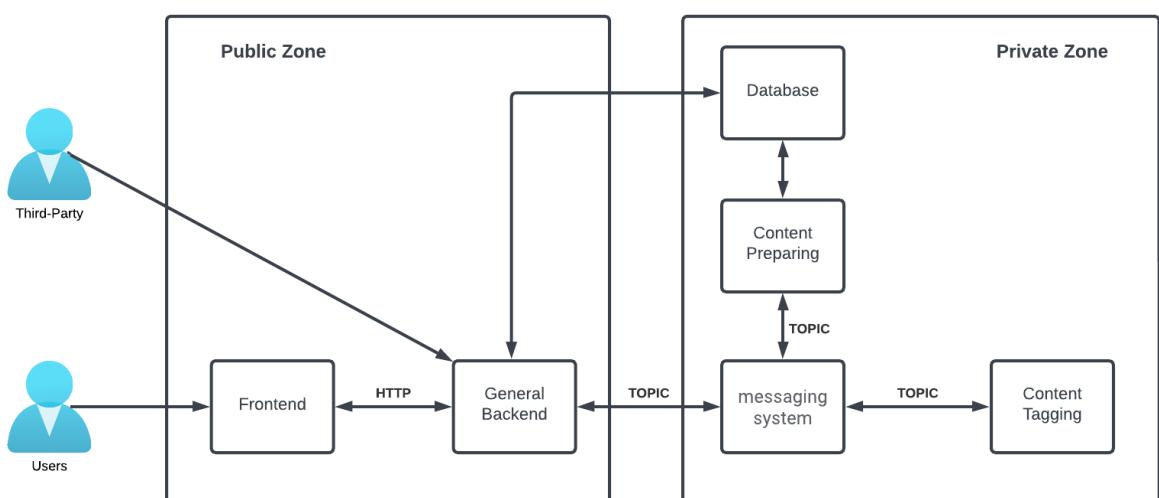
บทที่ 3 วิธีการดำเนินงาน

3.1 ข้อกำหนดและความต้องการของระบบ

- ผู้ใช้งานสามารถสมัครสมาชิกเพื่อเข้ามาใช้งานภายในระบบได้
- ผู้ใช้งานที่สมัครสมาชิกแล้ว สามารถเข้าสู่ระบบได้
- ผู้ใช้งานสามารถออกจากระบบได้
- ผู้ใช้งานสามารถค้นหาบทความตามหมวดหมู่ที่ต้องการค้นหาได้
- ผู้ใช้งานสามารถค้นหาบทความด้วยเลขอ้างอิงเนื้อหาได้
- สมาชิกสามารถตรวจสอบข้อมูลprofileของตนเองได้
- สมาชิกสามารถสร้างคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยการส่ง Web Link ให้กับระบบได้
- สมาชิกสามารถสร้างคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยการส่ง Web Link ผ่าน Api-Key ให้กับระบบได้
- สมาชิกสามารถสร้างคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยการส่ง Web Link ผ่าน Api-Key ให้กับระบบได้
- สมาชิกสามารถดูคำร้องขอการวิเคราะห์หมวดหมู่ของบทความของตนเองได้
- ระบบสามารถจัดหมวดหมู่ให้กับบทความได้ โดยการวิเคราะห์เนื้อหาภายในบทความ
- ผู้จัดการระบบสามารถจัดการผู้ใช้งานและเนื้อหาได้
- ผู้จัดการระบบสามารถเพิ่มและจัดการหมวดหมู่ได้ด้วยตนเองทาง Frontend

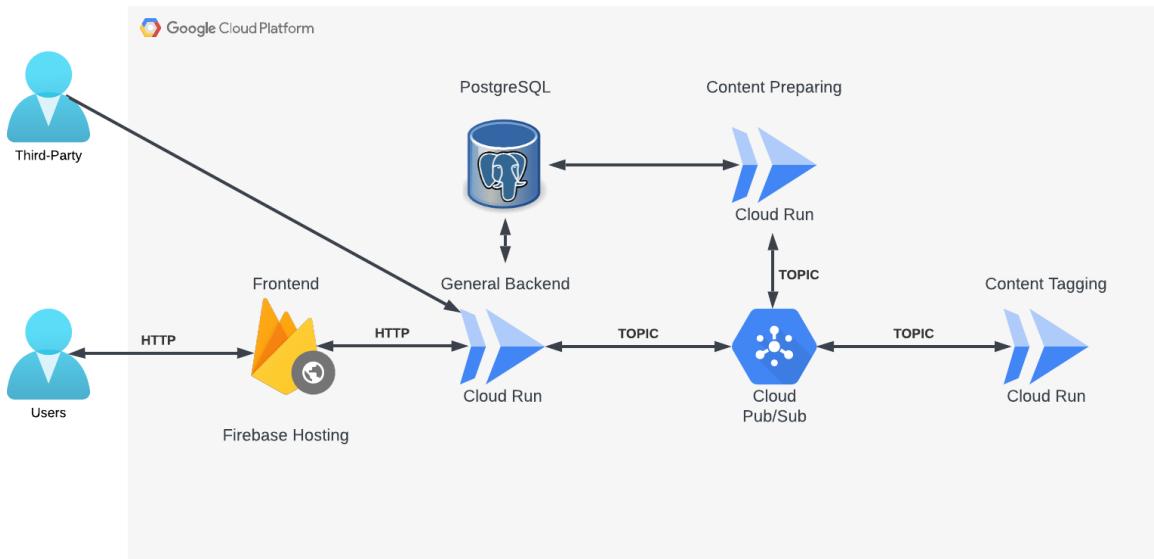
3.2 สถาปัตยกรรมระบบ

3.2.1 System Overview



รูปที่ 3.1: โครงสร้างทางสถาปัตยกรรมของระบบ

ระบบจะถูกแบ่งเป็น 2 ส่วนหลัก ๆ ได้แก่ ส่วนที่ 1 จะเป็นส่วนที่ผู้ใช้งานสามารถเข้าถึงได้ Ftp จะแบ่งเป็น Frontend ของเว็บไซต์ และ Backend ของเว็บไซต์ที่จะประมวลผลเบื้องต้น เช่น การเข้าสู่ระบบหรือการสมัครสมาชิก ซึ่งผู้ใช้งานจะสามารถติดต่อกับระบบได้ 2 วิธีด้วยกัน วิธีที่ 1 คือ การทำงานผ่านทางหน้าเว็บไซต์ โดยใช้ Email และ Password ในการเข้าสู่ระบบก่อนสร้างคำร้องขอต่าง ๆ และวิธีที่ 2 คือ การใช้งานผ่าน Api-Key หรือเรียกว่า Third-Party ซึ่งจะติดต่อกับ Backend โดยตรง โดย Backend จะส่งคำร้องขอไปยังส่วนที่ 2 ที่เป็นส่วนประมวลผลวิเคราะห์ข้อมูลและไม่สามารถเข้าถึงได้จากภายนอก นอกจากนี้ระบบจะมีการสื่อสารระหว่างระบบด้วย Messaging System เพื่อให้สามารถทำงานแบบ Asynchronous ได้ เนื่องจากการประมวลผลในแต่ละขั้นตอนอาจจะใช้เวลานาน การสื่อสารแบบ Asynchronous จะช่วยให้จัดการการทำงาน ไม่จำเป็นต้องรอให้ส่วนวิเคราะห์เสร็จสิ้นตามลำดับ ทำให้สามารถประมวลผลส่วนอื่นได้ เมื่อประมวลผลเสร็จแล้ว ระบบก็จะส่งข้อมูลผ่าน Messaging System กลับมาซึ่งจะเป็นการทำงานแบบ Event Driver Architecture ซึ่งจะทำให้ระบบสามารถทำงานได้เร็วขึ้น และสามารถทำงานได้หลายอย่างพร้อมกันได้



รูปที่ 3.2: เครื่องมือในโครงสร้างทางสถาปัตยกรรมของระบบ

จากการออกแบบแบบโครงสร้างทางสถาปัตยกรรมของระบบข้างต้น (รูปที่: 3.2) ทำให้สามารถเลือกใช้เทคโนโลยีได้ดังภาพ ในส่วนของ Frontend จะใช้ Firebase Hosting ในการนำเว็บไซต์ที่พัฒนาแล้ว ไปติดตั้งให้ผู้ใช้งานสามารถเข้าถึงได้ ในส่วนของ Server ประมวลผลต่างๆ จะมีการนำ Cloud Run มาใช้ในการประมวลผล และในส่วนของ Messaging System จะใช้ Cloud Pub/Sub ซึ่งบริการทั้งหมดจะอยู่บน Google Cloud Platform

3.2.2 Content-Tagging

ในส่วนของการนำโมเดลไป Deploy ใช้งานเป็นเว็บแอปพลิเคชันนั้น ต้องทำการแปลงโมเดลให้อยู่ในรูปแบบที่พ้อ้มนำไปใช้งาน เช่น .joblib, .sav และ .h5 เมื่อนำเข้าโมเดลมาข้างเว็บแอปพลิเคชัน ขั้นตอนถัดไปคือการประมวลผลเกี่ยวกับการ Preprocessing ซึ่งในขั้นตอนนี้ต้องทำท่านเดินเหมือนในช่วงการเตรียมข้อมูลสำหรับการพัฒนาโมเดล จากนั้นส่งข้อมูลที่ผ่านการ Preprocessing ให้โมเดลทำการทำงานยผลลัพธ์ โดยป้อน Input เป็นเนื้อหาต่าง ๆ และส่งผ่าน Cloud Pub/Sub เมื่อระบบประมวลผลเสร็จก็จะทำการส่งผลลัพธ์กลับไปยัง Cloud Pub/Sub เพื่อให้ Service อื่น ๆ รับทราบและนำไปประมวลผลต่อไป

3.2.3 Content-Preparing

ในส่วนของการจัดเตรียมเนื้อหา (Content-Preparing) จะรับการส่งคำร้องขอวิเคราะห์ที่มีความทันสมัย เช่น Cloud Pub/Sub จากนั้นจะทำการตัดจัดหน้าเว็บไซต์และดึงข้อมูลเฉพาะส่วนที่ต้องการ แล้วจึงทำการแปลงภาษาเป็นภาษาอังกฤษเพื่อเก็บข้อมูลไปยังฐานข้อมูลและทำการกระจายข้อมูลไปยัง Cloud Pub/Sub เพื่อให้ Service อื่น ๆ ได้ประมวลผลต่อ

3.2.4 General-Backend

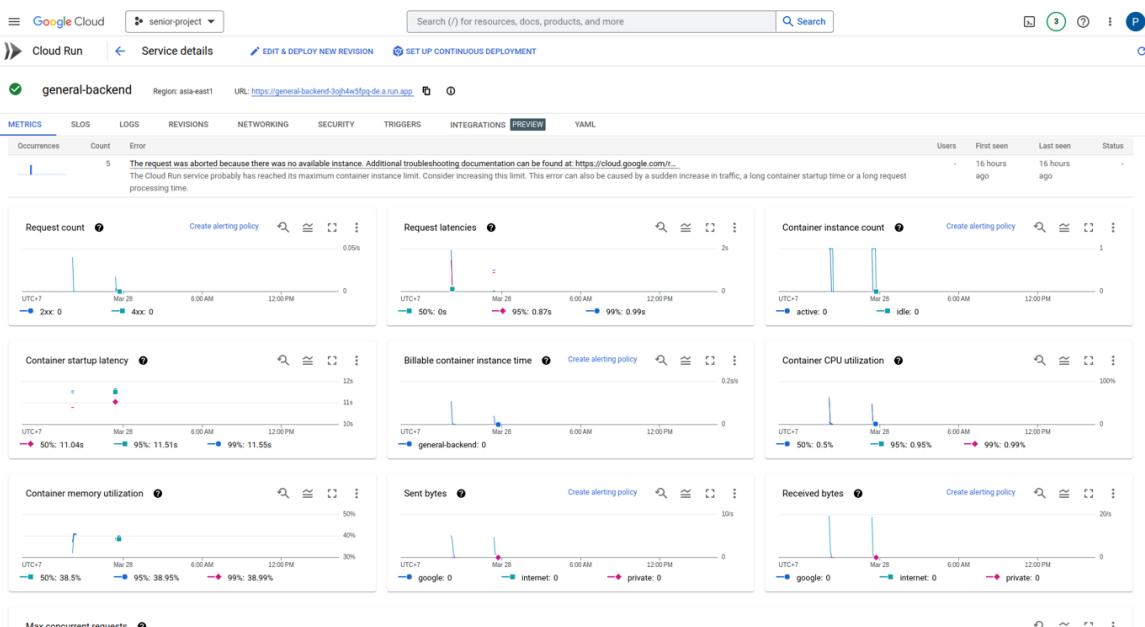
ในส่วนของการติดต่อกับ Frontend จะเป็นหน้าที่ของ General-Backend ซึ่งมีหน้าที่ต่าง ๆ เช่น การสมัครสมาชิก เข้าสู่ระบบ และสำหรับการใช้งาน Cloud Pub/Sub ขั้นแรกจะต้องทำการสร้าง Topic สำหรับใช้เป็นหัวข้อในการกระจายข้อมูลไปยัง Subscriber ต่าง ๆ ที่ทำการ Subscribe กับ Topic นั้น ๆ ซึ่งหนึ่ง Topic สามารถมี Subscriber ได้หลายตัวโดยจะเป็นในลักษณะของ 1-to-Many ดังนั้น เมื่อมีข้อมูลที่ส่งมาอย่าง Topic Cloud Pub/Sub จะทำการกระจายข้อมูลที่ได้รับไปยัง Subscriber ทุก Subscriber ทำให้มี Service อย่างได้ข้อมูล ก็สามารถทำการ Subscriber ไปยัง Topic ที่ต้องการได้ทันที ทำให้ระบบสามารถ Scale ฟังก์ชันการทำงานได้อย่างง่าย ไม่จำเป็นต้องเพิ่ม Code สำหรับการกระจายข้อมูลไปยังที่ใหม่

The screenshot shows the Google Cloud Pub/Sub Topics page. On the left sidebar, under the 'Pub/Sub' section, there are four categories: Topics, Subscriptions, Snapshots, and Schemas. The 'Topics' tab is selected. At the top right, there are buttons for 'CREATE TOPIC' and 'DELETE'. Below these are two tabs: 'LIST' (selected) and 'METRICS'. A 'Filter' dropdown is present. The main area displays a table with three rows:

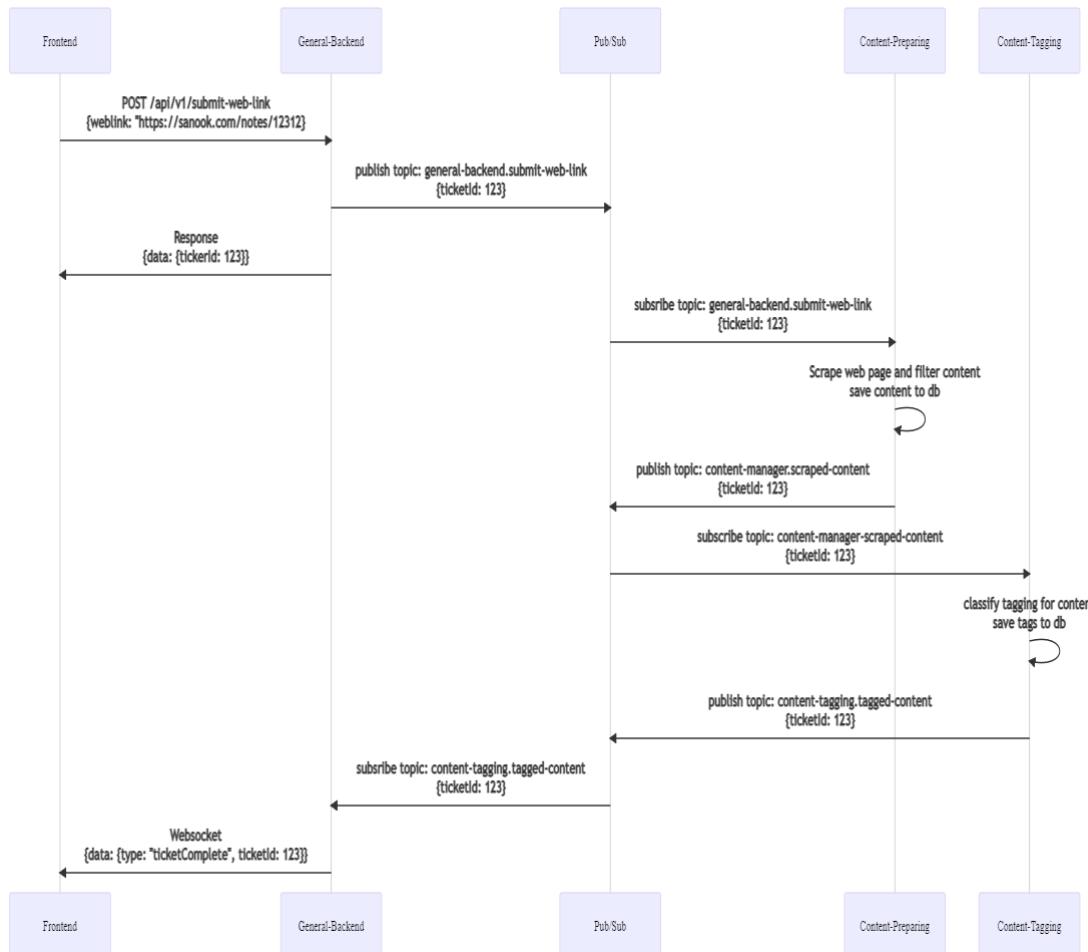
Topic ID	Encryption key	Topic name	Retention
content-preparing-prepared-content	Google-managed	projects/senior-project-364818/topics/content-preparing-prepared-content	-
general-backend.submit-web-link	Google-managed	projects/senior-project-364818/topics/general-backend.submit-web-link	-

รูปที่ 3.3: รายการ Topic ที่ใช้ในการสื่อสารในระบบ

เนื่องจาก Service ทำงานโดยใช้บริการ Cloud Run ทำให้สามารถ Monitor การใช้งานต่าง ๆ เช่น Memory, CPU ได้

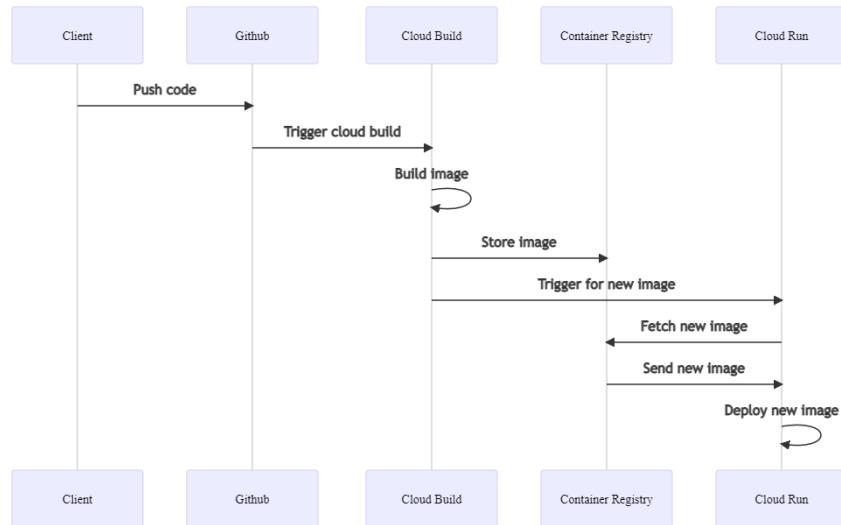


รูปที่ 3.4: การ Monitor การใช้งานทรัพยากรต่าง ๆ บน Cloud Run



รูปที่ 3.5: Sequence Diagram จำลองการส่งประมวลผลการวิเคราะห์หมวดหมู่ของบทความ

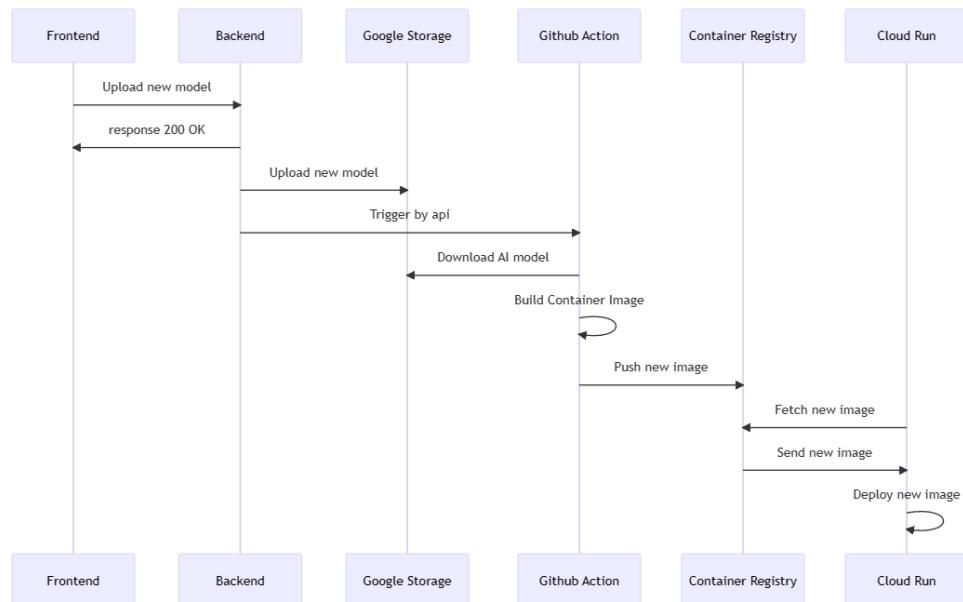
ภาพข้างต้น (รูปที่ 3.5) คือ Time Sequence จำลองการส่งประมวลผลวิเคราะห์หมวดหมู่ของบทความ โดยจะเริ่มจาก Frontend ส่งคำร้องขอมายัง Backend จากนั้น Backend จะ Publish ข้อมูลไปยัง Cloud Pub/Sub และเข้าสู่กระบวนการวิเคราะห์ข้อมูล โดยเริ่มจากการที่ระบบ Content-Preparing จะทำการไปเก็บข้อมูลจากเว็บไซต์ปลายทาง ได้แก่ หัวข้อและเนื้อหา แล้วจึงส่งกลับมายัง Cloud Pub/Sub ซึ่งระบบ Content-Tagging จะทำการนำเนื้อหาที่ได้ไปวิเคราะห์ต่อไป และเมื่อกระบวนการรุ่งหมดเสร็จสิ้น ก็จะทำการส่งกลับมายัง Backend และส่งต่อให้ Frontend ต่อไป



รูปที่ 3.6: Sequence Diagram ของ DevOps Life Cycle

ในส่วนของการบันการพัฒนาและติดตั้งระบบสำหรับใช้งาน (รูปที่:3.6) จะใช้เทคโนโลยีหลัก ๆ ได้แก่

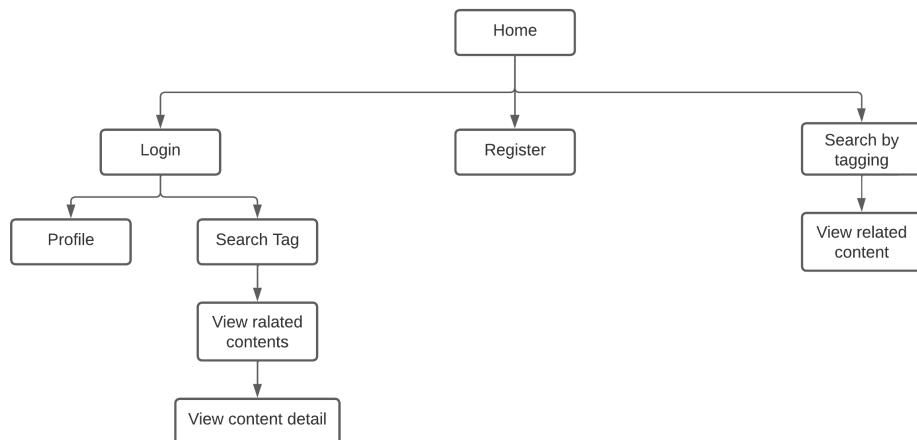
1. Github สำหรับเก็บ Source Code และมีมีการ Push Code ใหม่เข้ามา จะส่งผลให้เกิดการ Trigger ไปยัง Cloud Build
2. Cloud Build จะทำการดึง Source Code จาก Github และทำการ Build เป็น Container Image สำหรับสร้าง Application ในรูปแบบของ Container และจะส่งไปเก็บไว้ใน Container Registry
3. Container Registry จะเป็นส่วนที่เก็บ Image ของ Application ที่ต้องการนำไปทำงาน ซึ่งจะถูก Cloud Run นำไปสร้างเป็น Container สำหรับประมวลผลต่อไป
4. Cloud Run จะเป็น Serverless สำหรับประมวลผลคำร้องขอต่าง ๆ



รูปที่ 3.7: Sequence Diagram ของ Update New AI-Model Process

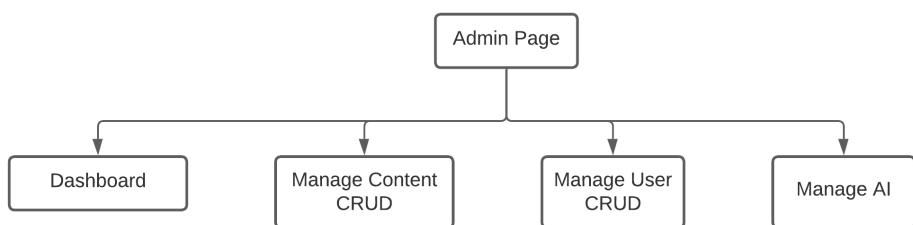
จากภาพข้างต้น (รูปที่ 3.7) ในส่วนของการแก้ไข Model ของ AI จะทำในลักษณะที่ใช้กระบวนการ DevOps โดยที่จะทำการอัพโหลด Model ใหม่ผ่าน Admin UI จากนั้น Backend จะทำการ Clone โครงการของ Content-Tagging และนำไฟล์ที่อัพโหลดผ่าน Admin UI เพิ่มเข้าไป แล้วจึงทำการ Push Code กลับมายัง Github จากนั้น Pipeline ของการ Deploy ก็จะทำการ Trigger และวิ่งทำงานต่อไปจน Deploy ขึ้นไปยัง Cloud Run

3.2.5 Navigation Map



รูปที่ 3.8: Navigation Map ของเว็บไซต์สำหรับผู้ใช้งาน

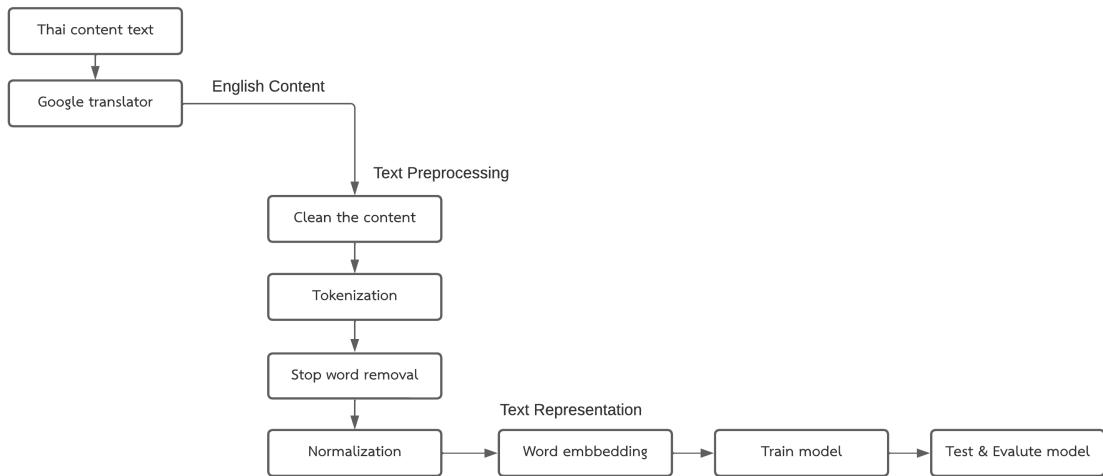
เมื่อผู้ใช้งานเข้าใช้งานเว็บไซต์ AI Platform Thai Content Tagging เริ่มไปที่จะเริ่มต้นที่หน้า Landing ซึ่งหน้า Landing จะประกอบไปด้วยรายละเอียดต่าง ๆ เช่น ข้อมูลภาพรวมของเว็บไซต์ กระบวนการในการประมวลผลหมวดหมู่โดย AI และวิธีการใช้งานการวิเคราะห์หมวดหมู่ของบทความ โดยในหน้า Landing ระบบจะแบ่งออกเป็น 3 ทางเลือก ได้แก่ หน้า Register สำหรับผู้ใช้งานที่ยังไม่มีบัญชีในระบบ หน้า Login สำหรับผู้ใช้งานที่มีบัญชีในระบบแล้ว และหน้า Search By Tagging สำหรับผู้ใช้งานที่ต้องการค้นหาบทความ ผู้ใช้งานที่ต้องการใช้งานจำเป็นที่จะต้องมีบัญชีในการใช้งานจึงจะสามารถเข้าสู่ระบบเพื่อรับบริการได้ เมื่อเข้าสู่ระบบแล้ว ผู้ใช้งานจะสามารถส่งคำร้องขอเพื่อการวิเคราะห์หมวดหมู่ของบทความที่ต้องการได้ 2 วิธี คือ Web Link และ Raw Content เมื่อส่งคำร้องขอวิเคราะห์สำเร็จ ผู้ใช้งานจะสามารถค้นหาบทความด้วยหมวดหมู่และสามารถดูบทความที่เกี่ยวข้องเพิ่มเติมได้ นอกจากนี้ผู้ใช้งานยังสามารถจัดการไฟล์ของตนเองได้และสามารถดูประวัติการร้องขอของตนเองได้เช่นกัน



รูปที่ 3.9: Navigation Map ของเว็บไซต์สำหรับผู้จัดการระบบ

ในส่วนของผู้จัดการระบบ จะเป็นส่วนที่จัดการระบบหลังบ้านของเว็บไซต์ทั้งหมด เช่น จัดการข้อมูลผู้ใช้งาน เนื้อหาที่ผู้ใช้งานได้ส่งคำร้องขอ ซึ่งสามารถเข้าถึงได้เฉพาะผู้จัดการระบบเท่านั้น

3.2.6 Data Science Processing



รูปที่ 3.10: Data Processing

ในกระบวนการทำโมเดลจะมีขั้นตอนหลัก ๆ ดังนี้

1. การเตรียมข้อมูลสำหรับทำโมเดล

ในขั้นตอนของการเตรียมข้อมูลสำหรับทำโมเดล จะมีทั้งหมด 2 ขั้นตอนหลัก ๆ ด้วยกัน ได้แก่

 - (a) การรวบรวมข้อมูลบทความต่าง ๆ จากแพลตฟอร์มออนไลน์ เช่น เว็บ Sanook, Thai PBS หรือไทยรัฐ ด้วยการทำ Web Scraping โดยการใช้ Selenium เป็น Web Driver และใช้ BeautifulSoup ในการดึงข้อมูลผ่านทางหน้าเว็บไซต์และสกัดเอามาเฉพาะส่วนที่ต้องการอ่านมา ได้แก่ ส่วนหัวข้อและส่วนเนื้อหาของบทความ ซึ่งจะได้บันทุกมาจากเว็บไซต์ที่มีเนื้อหาเป็นภาษาไทยอ่อนๆ จากนั้นจึงรวมเป็นไฟล์ข้อความตามหมวดหมู่ โดยรวมประมาณ 1,500 ชุด
 - (b) เมื่อมีจำนวนเครื่องมือสำหรับการพัฒนาโมเดลที่รองรับและจัดการข้อมูลที่เป็นภาษาอังกฤษได้มากกว่าภาษาไทย ดังนั้นคุณผู้อ่านจึงเลือกที่จะใช้ข้อมูลทั้ง 2 ภาษาสำหรับการพัฒนาโมเดล ได้แก่ ภาษาไทยและภาษาอังกฤษ และเปรียบเทียบผลลัพธ์ว่าชุดข้อมูลภาษาใดส่งผลให้โมเดลมีประสิทธิภาพสูงที่สุด ซึ่งชุดข้อมูลที่เป็นภาษาอังกฤษจะได้มาจากการนำบทความภาษาไทยที่ทำการรวบรวมด้วยการทำ Web Scraping มาแปลงเป็นภาษาอังกฤษด้วยบริการของ Google Translator แล้วจึงนำข้อมูลไปใช้ในกระบวนการถัดไป
2. การทำ Text Preprocessing
 - (a) ทำความสะอาดข้อมูล คือ ทำการเปลี่ยนตัวอักษรให้เป็นตัวพิมพ์เล็กทั้งหมด ตัดอักษรพิเศษที่ไม่ใช่ตัวอักษรภาษาอังกฤษออก เช่น อ็โมิ ตัวเลข หรือ Punctuation และลบเว้นวรรคที่ติดกันมากกว่า 1 เว้นวรรคเพื่อให้เหลือเว้นวรรคเดียว ซึ่งการทำในลักษณะนี้จะเป็นการลดสัญญาณรบกวนในข้อมูล
 - (b) Tokenization คือ การแยกคำแต่ละคำในประโยคออกจากกัน โดยที่ยังมีความหมายถูกต้องสมบูรณ์ตามฐานข้อมูลพจนานุกรม ดังนั้นจึงใช้ word_tokenize ใน Library NLTK ในการ拆分แบ่งคำในประโยค
 - (c) การลบ Stop Word คือ การลบคำที่ไม่ใช่คำที่มีความหมายและมีความสำคัญในประโยค เช่น a, also, unless เพื่อให้เหลือเพียงคำที่ค่อนข้างสื่อความหมายและมีความสำคัญในประโยค โดยใช้ Library NLTK ในการเรียกใช้ Stop Words
 - (d) Normalization คือ การลดความซับซ้อนของคำให้อยู่ในรูปของคำพื้นฐาน เพื่อกำจัด Inflection ของคำ เพื่อลดความซับซ้อนของการนำไปประมวลผล Model ด้วยการใช้ Lemmatization จาก Library Spacy เพื่อให้คำที่มีความหมายเหมือนกันอยู่ในรูปเดียวกัน เช่น is am are เป็น be จากนั้นจึงรวมคำทุกคำให้กล้ายเป็นประโยคเพื่อนำไปใช้ในขั้นตอนถัดไป
3. การทำ Text Representation คือ การแปลงคำให้เป็น Vector ที่อยู่ในรูปตัวเลขเพื่อนำมาใช้ในการพัฒนาโมเดล ซึ่งคุณผู้อ่านได้เลือกใช้ลักษณะ TF-IDF ในกระบวนการคำนวณหาค่าความสำคัญของคำแต่ละคำในชุดข้อมูล แล้วจึงแทนข้อความด้วยตัวเลขที่เป็นค่าความสำคัญของคำในแต่ละคำ เพื่อทำ Text Representation ให้เป็น Vector โดยใช้ TfidfVectorizer ใน Library Scikit-Learn

ในการสร้างโมเดล TF-IDF สำหรับแปลงข้อความเป็น Vector และปรับ tune พารามิเตอร์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ซึ่งพารามิเตอร์ที่ต้องปรับ tune มีดังนี้

- min_df คือ พารามิเตอร์ที่ใช้กำหนดค่าความถี่ของคำที่สำคัญที่จะถือว่าสำคัญพอที่จะนำมาใช้ในการสร้าง Text Representation โดยมีค่าเป็นจำนวนเต็มบวกหรือเป็นสัดส่วนที่ระบุจำนวนคำที่ต้องปรากฏในเอกสารอย่างน้อยกี่ครั้งถึงจะถือว่าเป็นคำที่สำคัญพอที่จะนำมาใช้ในการคำนวน TF-IDF ทำให้คำที่ปรากฏในเอกสารถ้าหากว่าคำ min_df จะไม่ถูกนับเข้าไปในการคำนวน TF-IDF ซึ่งจะช่วยลดคำที่ไม่สำคัญและเพิ่มความสมมั่นคงกับคำที่มีความถี่สูงมากขึ้น
- max_df คือ พารามิเตอร์ที่ใช้กำหนดค่าความถี่ของคำที่สำคัญมากพอที่จะนำมาใช้ในการสร้าง Text Representation โดยมีค่าเป็นจำนวนเต็มบวกหรือเป็นสัดส่วนที่ระบุจำนวนคำที่ต้องปรากฏในเอกสารมากที่สุดกี่ครั้งถึงจะถือว่าเป็นคำที่ไม่สำคัญพอที่จะนำมาใช้ในการคำนวน TF-IDF ทำให้คำที่ปรากฏในเอกสารถ้าหากว่าคำ max_df จะไม่ถูกนับเข้าไปในการคำนวน TF-IDF ซึ่งจะช่วยลดคำที่ไม่สำคัญและเพิ่มความสมมั่นคงกับคำที่มีความถี่สูงต่อลง
- max_features คือ พารามิเตอร์ที่ใช้กำหนดจำนวนคำสูงสุดที่จะนำไปใช้ในการสร้าง Text Representation โดยมีค่าเป็นจำนวนเต็มบวกที่ระบุจำนวนคำที่ต้องการให้เป็น feature ในโมเดล เมื่อกำหนดค่า max_features ไว้ใน TfidfVectorizer โมเดลจะเลือกเฉพาะคำที่ถูกเลือกมาจากข้อมูลในจำนวนที่กำหนด ซึ่งจำนวนคำที่ถูกเลือกจะเป็นคำที่มีความถี่สูงตามลำดับจนถึงค่า max_features หากมีข้อมูลมากและมีคำซ้ำกันมาก ๆ คำ max_features ควรเลือกสูง เพื่อให้ feature มีความหลากหลายมากขึ้น
- norm คือ พารามิเตอร์ที่ใช้กำหนดวิธีการปรับค่าขนาดของเวกเตอร์ TF-IDF หลังจากที่คำนวนค่า TF-IDF ของแต่ละคำในเอกสาร
- stop_words คือ รายการคำ stopwords ที่ต้องการให้ตัดออกในการคำนวนค่า TF-IDF
- ngram_range คือ จำนวนคำติดกันที่นำมาใช้ในการคำนวนค่า TF-IDF

4. การพัฒนาโมเดล Model

เป็นการสร้างโมเดล ซึ่งจะทำทั้งหมด 3 โมเดลด้วยกัน ได้แก่ Random Forest, K-Nearest Neighbor และ LSTM โดยแต่ละโมเดลจะมีขั้นตอนในการพัฒนาโมเดล ดังนี้

- Random Forest
 - (a) แบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนแรกสำหรับการพัฒนาโมเดลและส่วนที่สองสำหรับการประเมินโมเดล ซึ่งจะแบ่งชุดข้อมูลแต่ละรูปแบบออกเป็นอัตราส่วน 8:2 (ชุดข้อมูลสำหรับการพัฒนาโมเดล : ชุดข้อมูลสำหรับการประเมินโมเดล)
 - (b) นำชุดข้อมูลสำหรับการพัฒนาโมเดลมาแปลงเป็น Vector ด้วยโมเดล TF-IDF ที่สร้างไว้ เพื่อใช้ในการพัฒนาโมเดล
 - (c) สร้างโมเดลสำหรับทำ Text Classification โดยใช้ RandomForestClassifier ใน Library Scikit-Learn และปรับ tune พารามิเตอร์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ซึ่งพารามิเตอร์ที่ต้องปรับ tune มีดังนี้
 - n_estimators คือ จำนวน Decision Tree ที่จะถูกสร้างแต่ละคำให้ในการทำงานหรือตัดสินใจในการจัดกลุ่ม (Classification) ใน Random Forest ซึ่งการกำหนดค่า n_estimators ควรพิจารณาความซับซ้อนของปัญหา ขนาดข้อมูลที่ใช้ในการฝึกโมเดล และเวลาที่ใช้ในการสร้างและทำงาน โดยค่าที่เหมาะสมสำหรับ n_estimators อาจแตกต่างกันไปในแต่ละงานหรือปัญหาที่ต้องการใช้ Random Forest ดังนั้นจึงต้องมีการทดลองและทดสอบการใช้งานกับค่า n_estimators ต่าง ๆ เพื่อหาค่าที่เหมาะสม
 - max_depth คือ จำนวนชั้นของ Decision Tree ที่ถูกสร้างขึ้นในโมเดล Random Forest ซึ่งการกำหนดค่า max_depth จะส่งผลในการควบคุมความซับซ้อนของโมเดล ดังนั้นจึงต้องมีการทดลองและทดสอบการใช้งานกับค่า max_depth ต่าง ๆ เพื่อหาค่าที่เหมาะสม
 - min_samples_split คือ จำนวนตัวอย่างขั้นต่ำที่จำเป็นต้องมีในโหนดก่อนที่จะทำการแยกสร้างโหนดย่อยใน Decision Tree ที่ถูกสร้างขึ้นในโมเดล Random Forest ซึ่งการกำหนดค่า min_samples_split สามารถช่วยลดความซับซ้อนของต้นไม้และป้องกันการเกิด Overfitting ดังนั้นจึงต้องมีการทดลองและทดสอบการใช้งานกับค่า min_samples_split ต่าง ๆ เพื่อหาค่าที่เหมาะสม
 - min_samples_leaf คือ จำนวนตัวอย่างขั้นต่ำที่จำเป็นต้องมีในโหนดย่อยสุดท้ายของ Decision Tree ซึ่งการกำหนดค่า min_samples_split ควรพิจารณาความซับซ้อนของปัญหา ขนาดข้อมูลที่ใช้ในการฝึกโมเดล และความต้องการในการควบคุมความลึกและการแยกสร้างโหนดย่อยใน Random Forest ดังนั้นจึงต้องมีการทดลองและทดสอบการใช้งานกับค่า min_samples_leaf ต่าง ๆ เพื่อหาค่าที่เหมาะสม
 - (d) นำชุดข้อมูลที่แปลงเป็น Vector มาทำการ fit กับโมเดลที่สร้างไว้

- K-Nearest Neighbor

- แบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนแรกสำหรับการพัฒนาโมเดลและส่วนที่สองสำหรับการประเมินโมเดล ซึ่งจะแบ่งชุดข้อมูลแต่ละรูปแบบออกเป็นอัตราส่วน 8:2 (ชุดข้อมูลสำหรับการพัฒนาโมเดล : ชุดข้อมูลสำหรับการประเมินโมเดล)
- นำชุดข้อมูลสำหรับการพัฒนาโมเดลมาแปลงเป็น Vector ด้วยโมเดล TF-IDF ที่สร้างไว้ เพื่อใช้ในการพัฒนาโมเดล
- สร้างโมเดลสำหรับทำ Text Classification โดยใช้ KNeighborsClassifier ใน Library Scikit-Learn และปรับ tune ารามณ์ต่อร์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ซึ่งพารามณ์ต่อร์ที่ต้องปรับ tune มีดังนี้
 - n_neighbors คือ การกำหนดจำนวนเพื่อนบ้านที่ใช้ในการพิจารณาเพื่อตัดสินใจในการจัดกลุ่มหรือทำนายค่าของตัวอย่างใหม่ จำนวนนี้จะใช้วิธีการหาค่าเฉลี่ยหรือการโหวตเพื่อตัดสินใจในการจัดกลุ่มหรือทำนายค่าของตัวอย่างใหม่ ดังนั้นจึงต้องมีการทดลองและทดสอบการใช้งานกับค่า n_neighbors ต่าง ๆ เพื่อหาค่าที่เหมาะสม
 - weights คือ การกำหนดน้ำหนักให้กับตัวอย่างในกระบวนการคำนวณระยะห่างหรือความคล้ายคลึงใน K-Nearest Neighbor ซึ่งจะมีทั้งหมด 2 ค่า ได้แก่ uniform ที่ทำให้ตัวอย่างทุกตัวจะมีน้ำหนักเท่ากัน ไม่สนใจระยะห่างระหว่างตัวอย่าง และ distance ที่ทำให้ตัวอย่างใกล้ที่สุดจะมีน้ำหนักมากกว่าตัวอย่างที่ห่างออกไป
- (d) นำชุดข้อมูลที่แปลงเป็น Vector มาทำการ fit กับโมเดลที่สร้างไว้

- LSTM

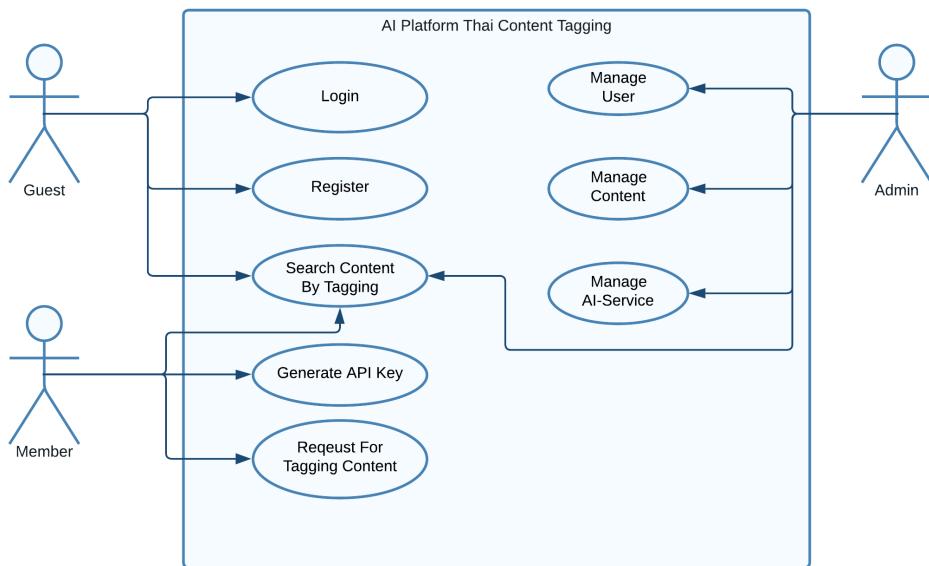
- แบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนแรกสำหรับการพัฒนาโมเดลและส่วนที่สองสำหรับการประเมินโมเดล ซึ่งจะแบ่งชุดข้อมูลแต่ละรูปแบบออกเป็นอัตราส่วน 8:2 (ชุดข้อมูลสำหรับการพัฒนาโมเดล : ชุดข้อมูลสำหรับการประเมินโมเดล)
- นำชุดข้อมูลสำหรับการพัฒนาโมเดลมาแปลงเป็น Vector ด้วยโมเดล TF-IDF ที่สร้างไว้ เพื่อใช้ในการพัฒนาโมเดล
- สร้างโมเดลสำหรับทำ Text Classification ทั้งหมด 3 Layer ได้แก่ Embedding Layer, LSTM Layer และ Dense Layer โดยใช้ Embedding, LSTM และ Dense ใน Library Tensorflow ตามลำดับ จากนั้นจึงทำการ Compile โมเดลด้วย Optimizer โดยในทุก Layer จะต้องมีการปรับ tune พารามณ์ต่อร์เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ซึ่งพารามณ์ต่อร์ที่ต้องปรับ tune มีดังนี้
 - units คือ จำนวนของ Nueral ที่มีใน LSTM Layer
 - dropout คือ อัตราส่วนที่จะทำการสุ่มคลบบางส่วนของ Nueral ใน LSTM Layer ซึ่งช่วยลดการเกิด Overfitting และช่วยให้โมเดลทำงานอย่างมีประสิทธิภาพมากขึ้น
 - activation คือ ฟังก์ชันที่ใช้ในการรับข้อมูลหลังจากผ่านการคำนวณเพื่อสร้างผลลัพธ์ออกมา ซึ่งในโมเดลนี้จะเลือกใช้ Softmax เนื่องจากเป็นการคำนวณผลลัพธ์จากหลายผลลัพธ์ที่ได้ทำการเรียงรูป
 - optimizers คือ อัลกอริทึมที่ใช้ในกระบวนการปรับค่าพารามณ์ต่อร์ของโมเดลในระหว่างการฝึกสอน เพื่อลดค่าของ Loss Function และปรับปรุงประสิทธิภาพของโมเดลให้ดีขึ้น ซึ่งการเลือก optimizers ที่เหมาะสมขึ้นอยู่กับปัญหาที่กำลังจะแก้ไขและลักษณะของข้อมูล ดังนั้นจึงต้องทบทวนการทดลองและประเมินผลลัพธ์ของโมเดลกับ optimizers ต่างๆเพื่อหาวิธีการปรับแต่งที่ให้ผลลัพธ์ที่ดีที่สุด
 - loss คือ ตัวชี้วัดที่ใช้ในการประเมินความคลาดเคลื่อนระหว่างผลลัพธ์ที่โมเดลคำนวณกับผลลัพธ์จริงในขณะฝึกสอน โดยค่า loss จะแสดงถึงความต่างหรือความสูญเสียระหว่างค่าที่โมเดลคำนวณกับค่าจริงของข้อมูล ฝึกสอน ซึ่งในโมเดลนี้จะเลือกใช้ Categorical Cross-Entropy เนื่องจากเป็นฟังก์ชันสำหรับงานที่เกี่ยวกับการจำแนกหลายคลาส ใช้ในการคำนวณค่าความคลาดเคลื่อนระหว่างความน่าจะเป็นที่คำนวณโดยโมเดลกับความน่าจะเป็นจริงของแต่ละคลาส
 - learning_rate คือ ค่าที่ใช้ในการกำหนดอัตราการปรับค่าพารามณ์ต่อร์ในกระบวนการฝึกสอนของโมเดล อัตราการเรียนรู้จะมีผลต่อความเร็วและความน่าเชื่อถือในการปรับปรุงโมเดล ดังนั้นจึงต้องทบทวนการทดลองและประเมินผลลัพธ์ของโมเดลกับ learning_rate ต่างๆเพื่อหาวิธีการปรับแต่งที่ให้ผลลัพธ์ที่ดีที่สุด
- (d) นำชุดข้อมูลที่แปลงเป็น Vector มาทำการ fit กับโมเดลที่สร้างไว้

5. การทดสอบและประเมินผลโมเดล

เป็นการทดสอบและวัดผลประสิทธิภาพของโมเดล โดยในที่นี้จะวัดผลด้วยการใช้ค่า Accuracy, Precision, Recall และ F1 Score

3.3 Use Case Analysis

3.3.1 Use Case Diagram



รูปที่ 3.11: Use Case Diagram

3.3.2 Use Case Narrative

3.3.2.1 ผู้ใช้งานสมัครสมาชิก

Name: ผู้ใช้งานสมัครสมาชิก

Actors: ผู้ใช้งานทั่วไป

Goal: ผู้ใช้งานต้องการสมัครสมาชิกกับระบบ

Precondition: -

Main success scenario:

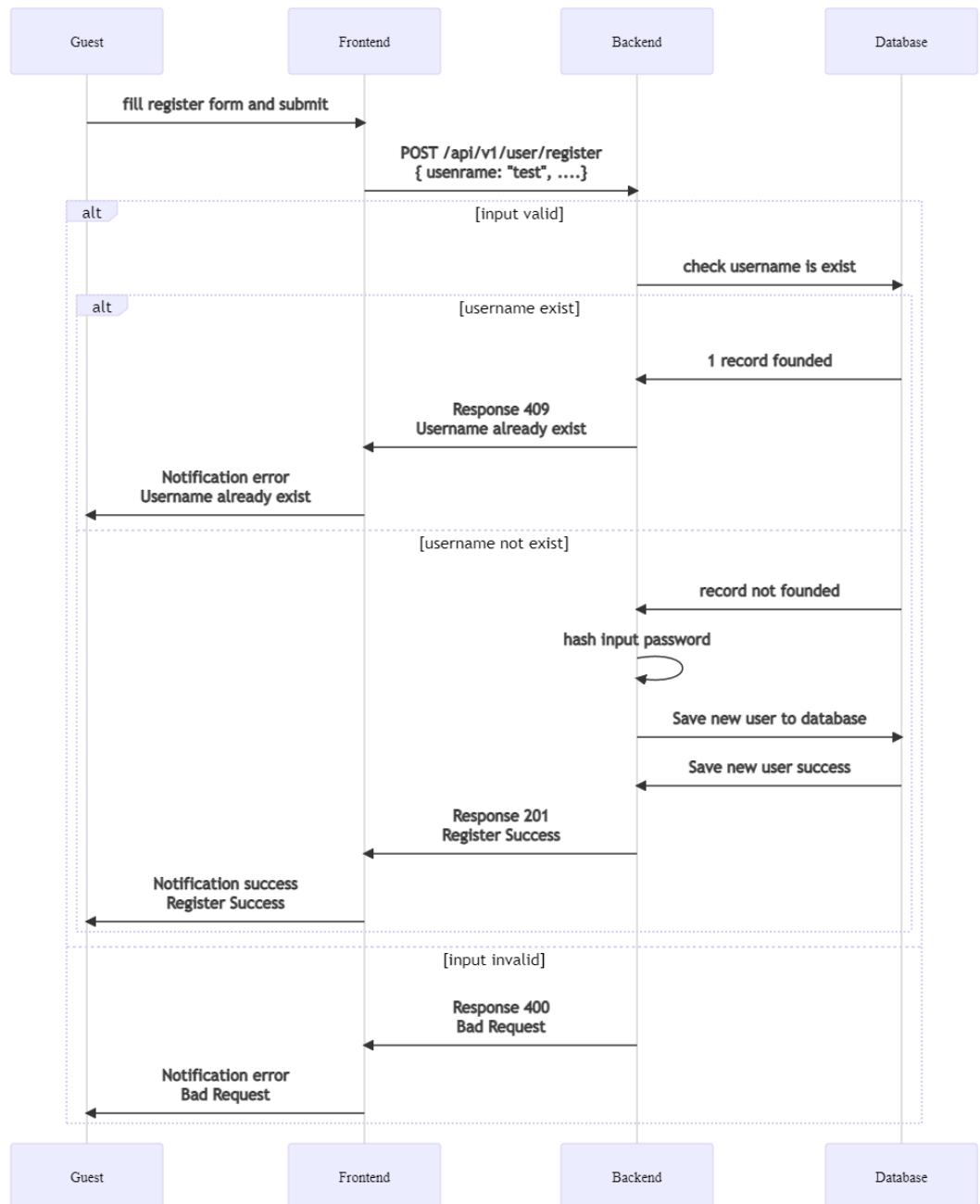
1. ผู้ใช้งานกรอกข้อมูลส่วนตัว รวมไปถึง Email และ Password
2. ผู้ใช้งานส่งแบบฟอร์มข้อมูลมายังระบบ
3. ระบบตรวจสอบข้อมูลและทำการบันทึกลงฐานข้อมูล
4. ผู้ใช้งานลงทะเบียนเสร็จสมบูรณ์

Extension (a):

- 3a. ระบบตรวจสอบข้อมูลพบว่าผู้ใช้งานกรอกข้อมูลไม่ครบ
- 4a. ผู้ใช้งานได้รับแจ้งเตือนว่ากรอกข้อมูลไม่ครบ
- 5a. กลับไปที่ข้อที่ 1

Extension (b):

- 3a. ระบบตรวจสอบข้อมูลพบว่ามี Email อยู่แล้วในระบบ
- 4a. ผู้ใช้งานได้รับแจ้งเตือนว่า Email ซ้ำ
- 5a. กลับไปที่ข้อที่ 1



รูปที่ 3.12: Sequence diagram ของการสมัครสมาชิก

ตารางที่ 3.1: API Document ของ User ในการสร้างบัญชีสำหรับใช้งานภายในระบบ

URL	/api/v1/user/register
Description	สร้างบัญชีสำหรับใช้งานภายในระบบ
METHOD	POST
HEADER	-
PAYLOAD	email: string password: string firstName: string lastname: string image: file
Response Status	Response Data
201	<pre>{ message: "register successfull" }</pre>
409	<pre>{ errors: [{ "message": "Email already exist" }] }</pre>
400	<pre>{ errors: [{ "field": "Email", "message": "This field is required" }, { "field": "password", "message": "This field is required" }] }</pre>

3.3.2.2 ผู้ใช้งานเข้าสู่ระบบ

Name: ผู้ใช้งานเข้าสู่ระบบ

Actors: ผู้ใช้งานทั่วไป

Goal: ผู้ใช้งานสามารถเข้าสู่ระบบได้

Precondition: ผู้ใช้งานสมัครสมาชิกแล้ว

Main success scenario:

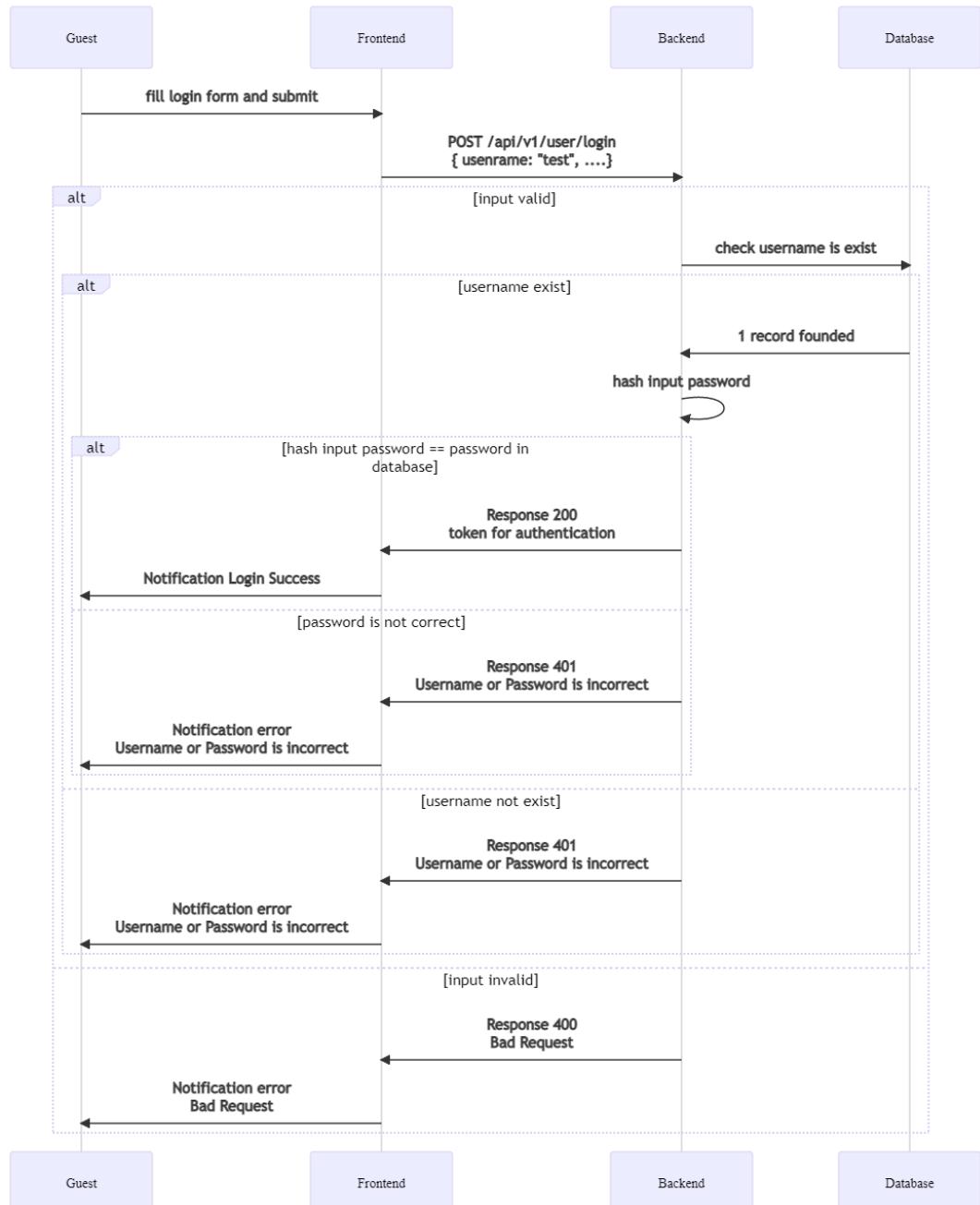
1. ผู้ใช้งานกรอก Email และ Password
2. ผู้ใช้งานส่งคำร้องขอเข้าสู่ระบบ
3. ระบบตรวจสอบ Email และ Password ตรงกับในฐานข้อมูล
4. ผู้ใช้งานสามารถเข้าสู่ระบบได้

Extension (a):

- 3a. ระบบตรวจสอบข้อมูลพบว่าผู้ใช้งานกรอกข้อมูลไม่ครบ
- 4a. ผู้ใช้งานได้รับแจ้งเตือนว่ากรอกข้อมูลไม่ครบ
- 5a. กลับไปที่ข้อที่ 1

Extension (b):

- 3a. ระบบตรวจสอบข้อมูลพบว่าไม่มี Email และ Password ตรงกับในฐานข้อมูล
- 4a. ผู้ใช้งานได้รับแจ้งเตือนว่า Email และ Password ไม่ถูกต้อง
- 5a. กลับไปที่ข้อที่ 1



รูปที่ 3.13: Sequence diagram ของการเข้าสู่ระบบ

ตารางที่ 3.2: API Document ของ User ในการเข้าสู่ระบบบัญชีผู้ใช้งาน

URL	/api/v1/user/login
Description	เข้าสู่ระบบบัญชีผู้ใช้งาน
METHOD	POST
HEADER	-
PAYLOAD	email: string password: string
Response Status	Response Data
200	<pre>{ message: "login successfull", token: "xxxxxxxxxxxxxx" }</pre>
400	<pre>{ errors: [{ "message": "Email or password is invalid" }] }</pre>

3.3.2.3 ตรวจสอบข้อมูลโปรแกรมของตนเอง

Name: สมาชิกตรวจสอบข้อมูลโปรแกรมของตนเอง

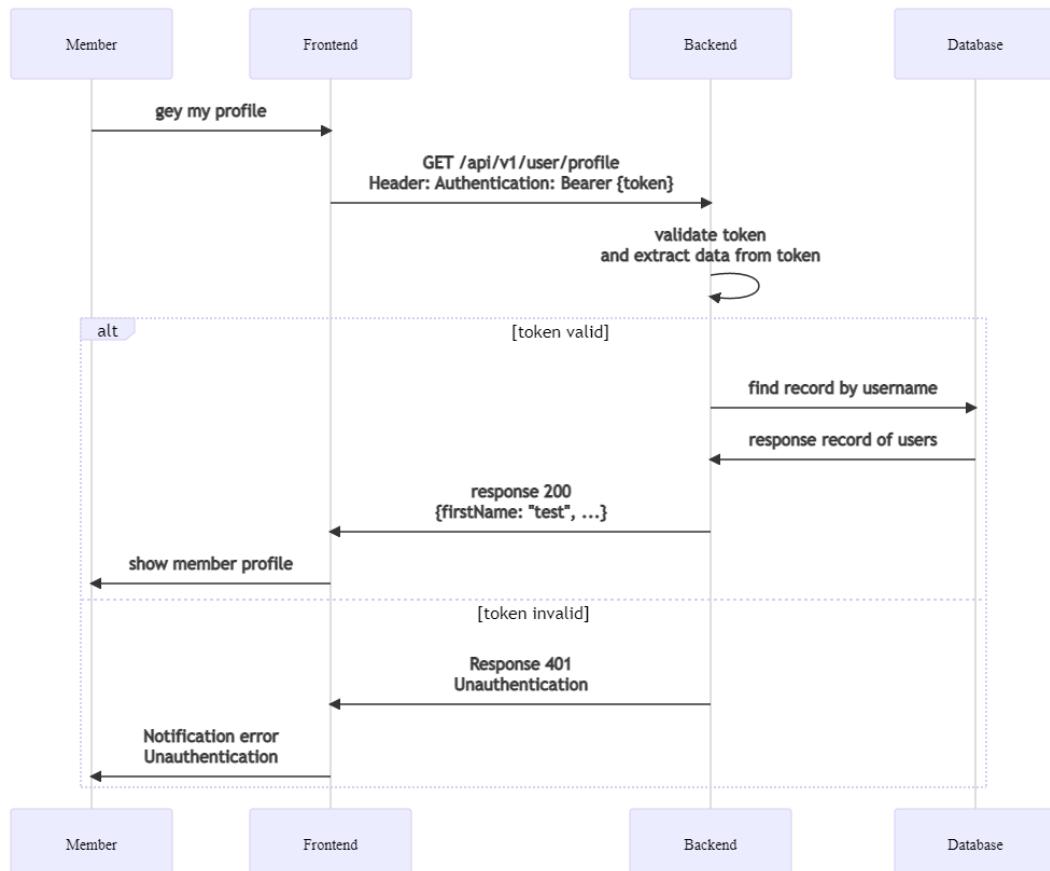
Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกสามารถเรียกดูข้อมูลโปรแกรมของตนเองได้

Precondition: สมาชิกต้องทำการเข้าสู่ระบบແລ້ວ

Main success scenario:

1. สมาชิกร้องขอดูข้อมูลโปรแกรมของตนเอง
2. ระบบดึงข้อมูลโปรแกรมแสดงผลให้สมาชิก



รูปที่ 3.14: Sequence diagram ของการตรวจสอบข้อมูลโปรไฟล์

ตารางที่ 3.3: API Document ของ User ในการดูรายละเอียดโปรไฟล์ของตนเอง

URL	/api/v1/user/profile
Description	ดูรายละเอียดโปรไฟล์ของตนเอง
METHOD	GET
HEADER	Authorization: Bearer {token}
PAYLOAD	-

Response Status	Response Data
200	<pre>{ id: "1", firstName: "Ponlawat", lastName: "Suparat" email: ""test@gmail.com"", role: "Member", api_key: "asdasfwaefdfaxcvadasv324234" }</pre>
401	<pre>{ errors: [{ "message": "Unauthorization" }] }</pre>

3.3.2.4 สร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link

Name: สมาชิกสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link

Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาไปยังระบบสำเร็จ

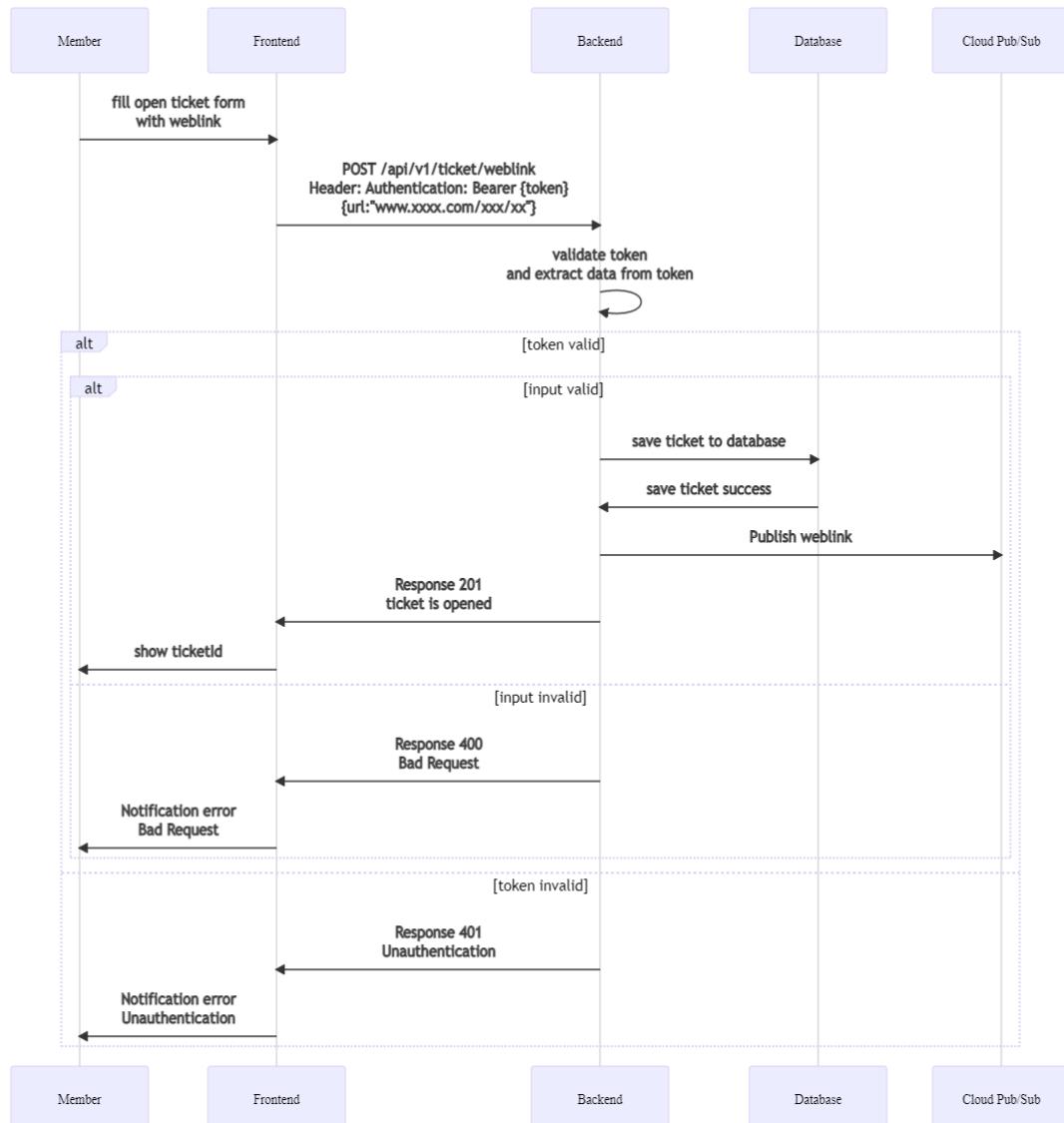
Precondition: สมาชิกต้องทำการเข้าสู่ระบบแล้ว

Main success scenario:

1. สมาชิกกรอกแบบฟอร์ม รวมไปถึงกรอก Web Link ที่ต้องการวิเคราะห์หมวดหมู่ของเนื้อหา
2. สมาชิกส่งแบบฟอร์มมายังระบบ
3. ระบบทำการตอบกลับด้วยเลขคำร้องขอ
4. ระบบประมวลผลเสร็จสิ้นและส่งการแจ้งเตือนกลับไปยังสมาชิก

Extension (a):

- 3a. ระบบตรวจสอบข้อมูลพบว่าสมาชิกกรอกข้อมูลไม่ครบ
- 4a. สมาชิกได้รับแจ้งเตือนว่ากรอกข้อมูลไม่ครบ
- 5a. กลับไปที่ข้อที่ 1



รูปที่ 3.15: Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link

ตารางที่ 3.4: API Document ของ Ticket ในการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link

URL	/api/v1/ticket/weblink/
Description	สร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link
METHOD	POST
HEADER	Authorization: Bearer {token}
PAYLOAD	url: string

Response Status	Response Data
201	{ ticketId: 123, status: "Open" }
400	{ errors: [{ "field": "url", "message": "This field is required" }] }
401	{ errors: [{ "message": "Unauthorization" }] }

3.3.2.5 สร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความ

Name: สมาชิกสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความ

Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาไปยังระบบสำเร็จ

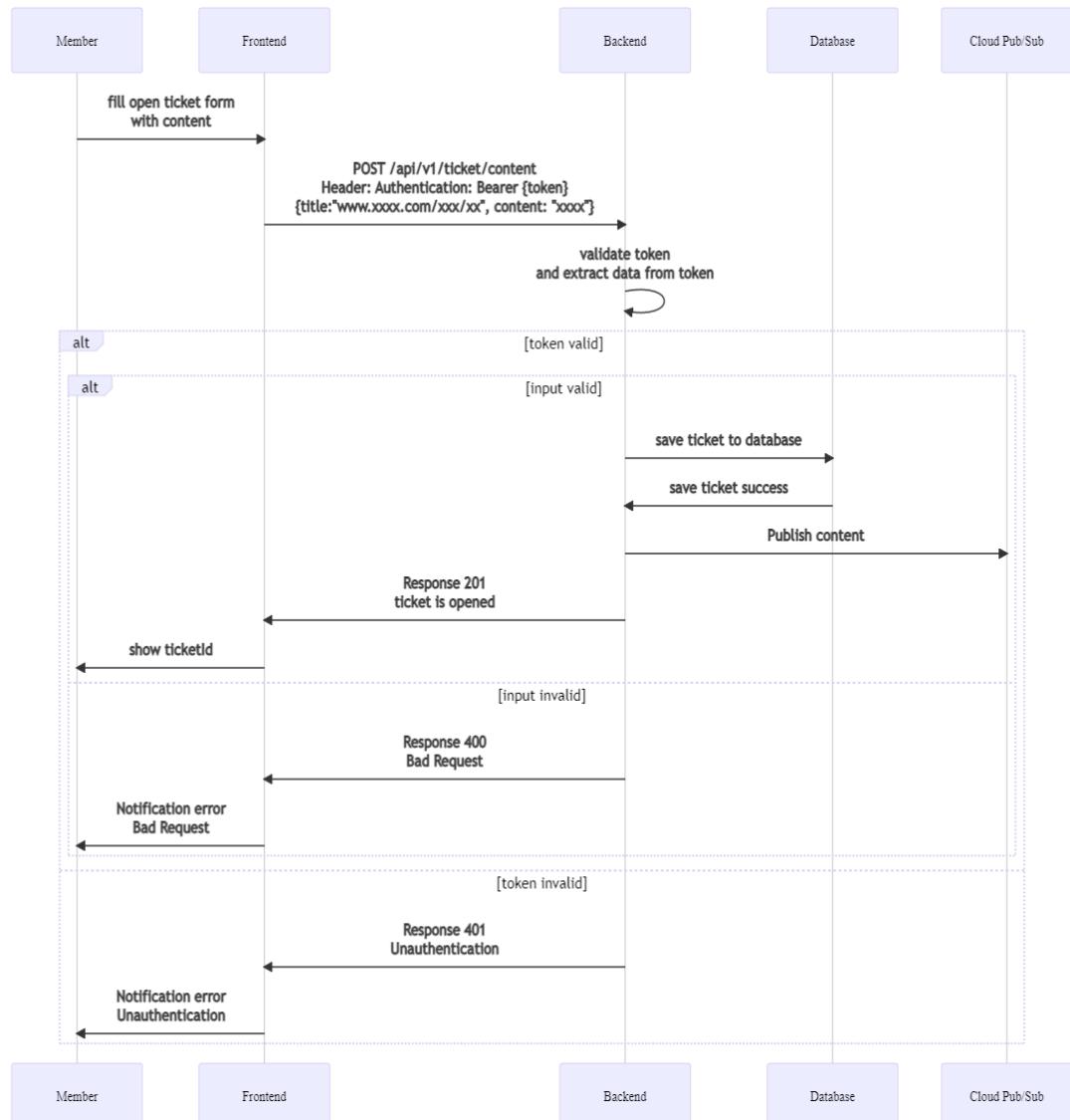
Precondition: สมาชิกต้องทำการเข้าสู่ระบบแล้ว

Main success scenario:

1. สมาชิกกรอกแบบฟอร์ม รวมไปถึงกรอกบุคคลที่ต้องการวิเคราะห์หมวดหมู่ของเนื้อหา
2. สมาชิกส่งแบบฟอร์มมายังระบบ
3. ระบบทำการตอบกลับด้วยเลขคำร้องขอ
4. ระบบประมวลผลเสร็จสิ้นและส่งการแจ้งเตือนกลับไปยังสมาชิก

Extension (a):

- 3a. ระบบตรวจสอบข้อมูลพบว่าสมาชิกกรอกข้อมูลไม่ครบ
- 4a. สมาชิกได้รับแจ้งเตือนว่ากรอกข้อมูลไม่ครบ
- 5a. กลับไปที่ข้อที่ 1



รูปที่ 3.16: Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความ

ตารางที่ 3.5: API Document ของ Ticket ในการสร้างคำร้องขอวิเคราะห์ด้วยเนื้อหาที่เตรียมไว้

URL	/api/v1/ticket/content/				
Description	สร้างคำร้องขอวิเคราะห์ด้วยเนื้อหาที่เตรียมไว้				
METHOD	POST				
HEADER	Authorization: Bearer {token}				
PAYLOAD	<table> <tr> <td>title:</td> <td>string</td> </tr> <tr> <td>content:</td> <td>string</td> </tr> </table>	title:	string	content:	string
title:	string				
content:	string				

Response Status	Response Data
201	<pre>{ ticketId: 123, status: "Open" }</pre>
400	<pre>{ errors: [{ "field": "title", "message": "This field is required" }, { "field": "content", "message": "This field is required" }] }</pre>
401	<pre>{ errors: [{ "message": "Unauthorization" }] }</pre>

3.3.2.6 สร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link ผ่าน Api-Key

Name: สมาชิกสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link ผ่าน Api-Key

Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาไปยังระบบสำเร็จ

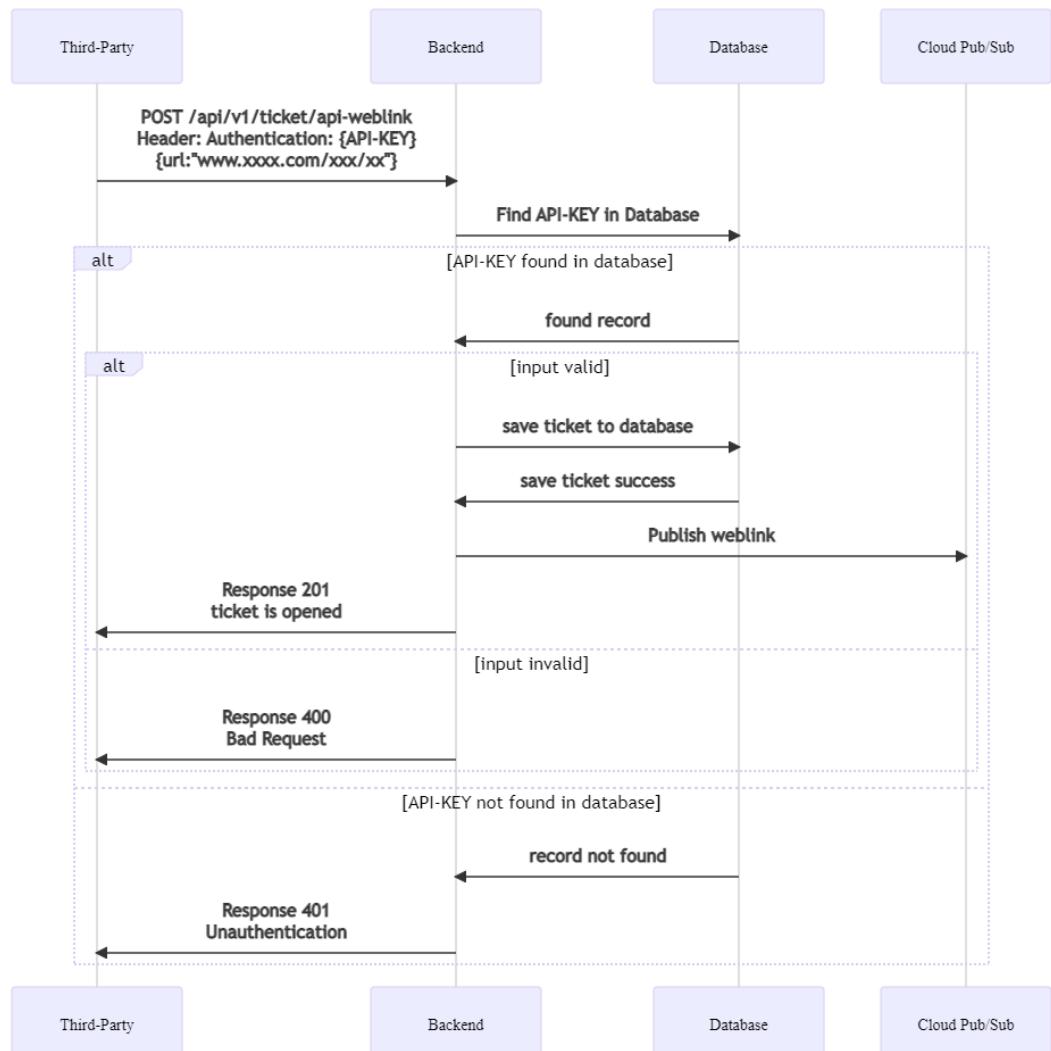
Precondition: สมาชิกต้องทำการสร้าง Api-Key และลงทะเบียน Web Hook สำหรับรับข้อมูลการประมวลผล

Main success scenario:

1. สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link โดยระบบทั้งหมดพร้อมกับ Api-Key
2. ระบบทำการตอบกลับด้วยเลขคำร้องขอ
3. ระบบประมวลผลเสร็จสิ้นและส่งการแจ้งเตือนกลับไปยังสมาชิกผ่าน Web Hook Url

Extension (a):

- 2a. ระบบตรวจสอบข้อมูลพิบảว่าสมาชิกส่งข้อมูลไม่ครบ
- 3a. สมาชิกได้รับแจ้งเตือนว่าส่งข้อมูลไม่ครบ
- 4a. กลับไปที่ข้อที่ 1



รูปที่ 3.17: Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วย Web Link ผ่าน Api-Key

ตารางที่ 3.6: API Document ของ Ticket ในการสร้างคำร้องขอวิเคราะห์ด้วย Web Link (Api-key)

URL	/api/v1/ticket/api-weblink/
Description	สร้างคำร้องขอวิเคราะห์ด้วย Web Link (Api-key)
METHOD	POST
HEADER	Authorization: {api-key}
PAYLOAD	url: string

Response Status	Response Data
201	{ ticketId: 123, status: "Open" }
400	{ errors: [{ "field": "url", "message": "This field is required" }] }
401	{ errors: [{ "message": "Unauthorization" }] }

3.3.2.7 สร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความผ่าน Api-Key

Name: สมาชิกสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความผ่าน Api-Key

Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาไปยังระบบสำเร็จ

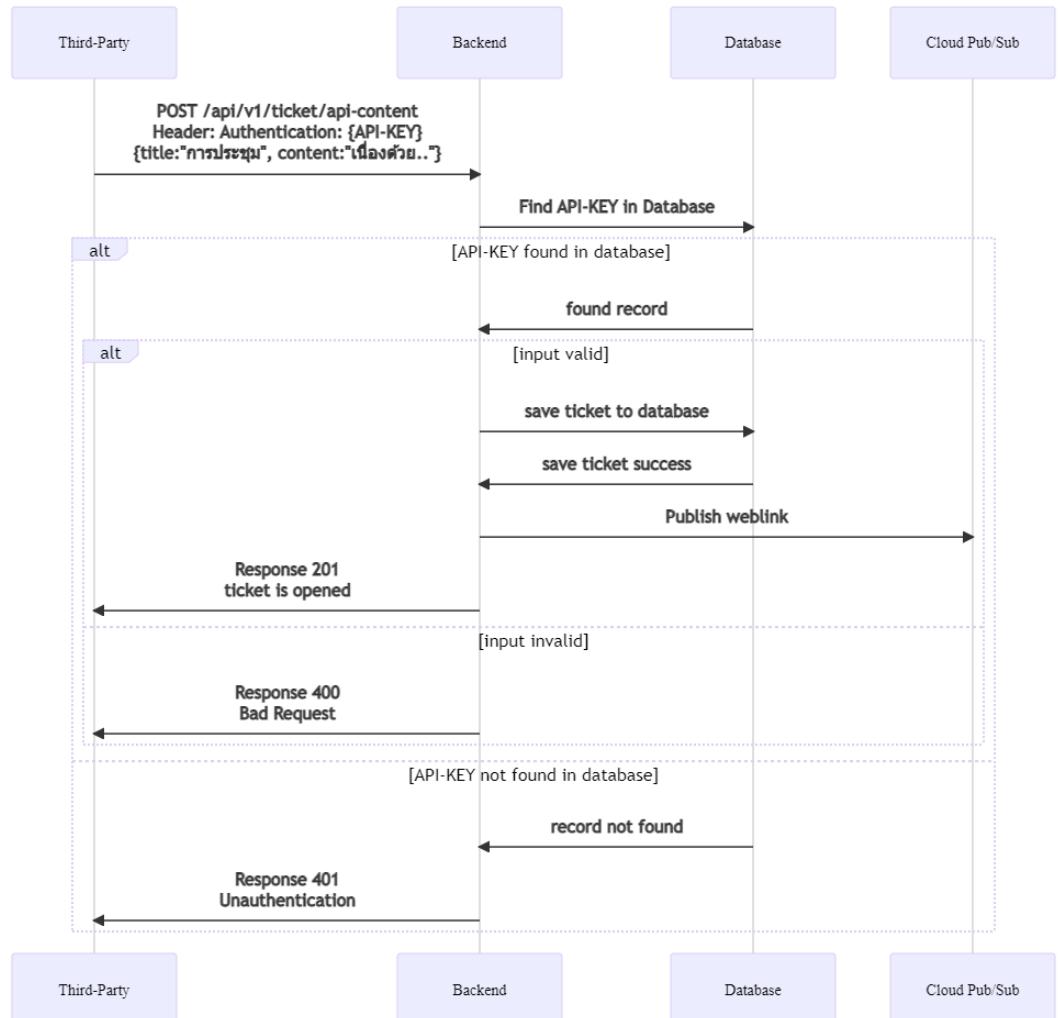
Precondition: สมาชิกต้องทำการสร้าง Api-Key และลงทะเบียน Web Hook สำหรับรับข้อมูลการประมวลผล

Main success scenario:

1. สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบทความมายังระบบ พร้อมกับ Api-Key
2. ระบบทำการตอบกลับด้วยเลขคำร้องขอ
3. ระบบประมวลผลเสร็จสิ้นและส่งการแจ้งเตือนกลับไปยังสมาชิกผ่าน Web Hook Url

Extension (a):

- 2a. ระบบตรวจสอบข้อมูลพบว่าสมาชิกส่งข้อมูลไม่ครบ
- 3a. สมาชิกได้รับแจ้งเตือนว่าส่งข้อมูลไม่ครบ
- 4a. กลับไปที่ข้อที่ 1



รูปที่ 3.18: Sequence diagram ของการสร้างคำร้องขอวิเคราะห์หมวดหมู่ของเนื้อหาด้วยบุคลากรผ่าน Api-Key

ตารางที่ 3.7: API Document ของ Ticket ในการสร้างคำร้องขอวิเคราะห์ที่ด้วยเนื้อหาที่เตรียมไว้ (Api-Key)

URL	/api/v1/ticket/api-content/				
Description	สร้างคำร้องขอวิเคราะห์ที่ด้วยเนื้อหาที่เตรียมไว้ (Api-Key)				
METHOD	POST				
HEADER	Authorization: {Api-Key}				
PAYLOAD	<table> <tr> <td>title:</td> <td>string</td> </tr> <tr> <td>content:</td> <td>string</td> </tr> </table>	title:	string	content:	string
title:	string				
content:	string				

Response Status	Response Data
201	<pre>{ ticketId: 123, status: "Open" }</pre>
400	<pre>{ errors: [{ "field": "title", "message": "This field is required" }, { "field": "content", "message": "This field is required" }] }</pre>
401	<pre>{ errors: [{ "message": "Unauthorization" }] }</pre>

3.3.2.8 ดูคำร้องขอการวิเคราะห์หมู่ของบทความของตนเอง

Name: สมาชิกดูคำร้องขอการวิเคราะห์หมู่ของบทความของตนเอง

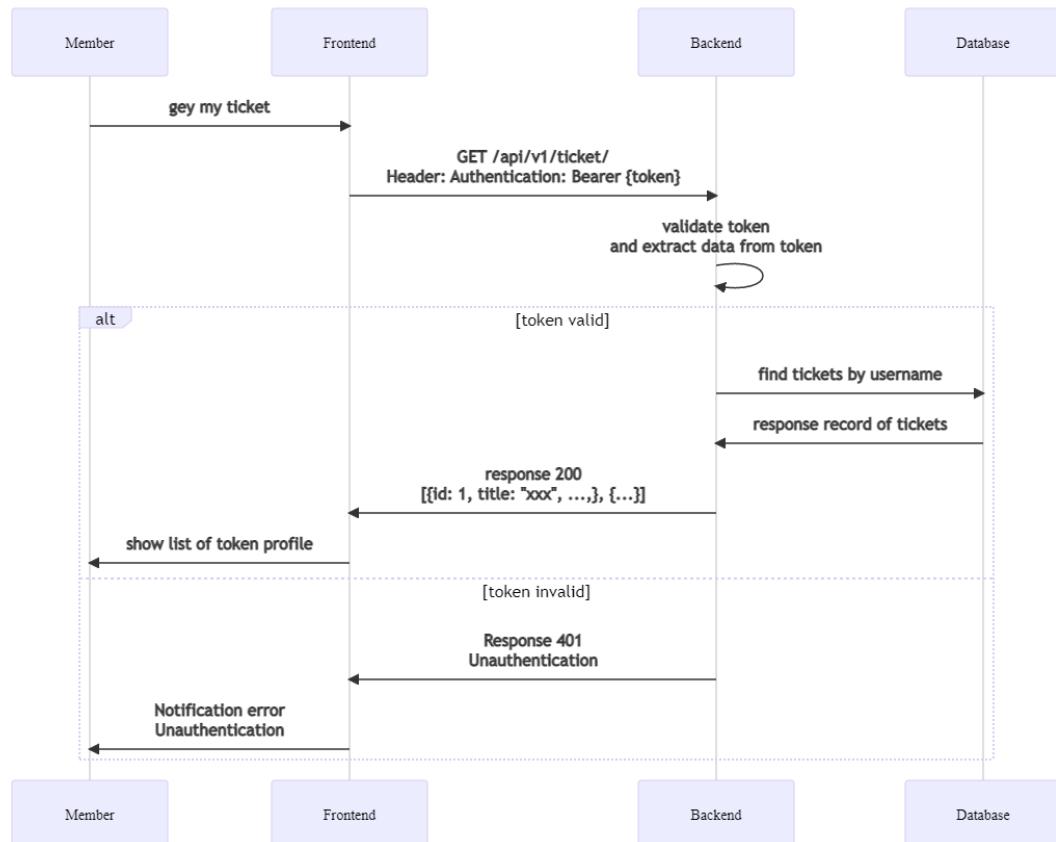
Actors: สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว)

Goal: สมาชิกส่งคำร้องขอการวิเคราะห์หมู่ของบทความไปยังระบบสำเร็จ

Precondition: สมาชิกต้องทำการเข้าสู่ระบบแล้ว

Main success scenario:

1. สมาชิกร้องขอดูข้อมูลคำร้องขอของตนเอง
2. ระบบดึงข้อมูลคำร้องขอของสมาชิกและแสดงผลให้สมาชิก



รูปที่ 3.19: Sequence diagram ของการดูแลรักษาการริบิเคราะห์ที่มุ่งของบทความ

ตารางที่ 3.8: API Document ของ Ticket ในการดูรายการคำร้องขอของตนเอง

URL	/api/v1/ticket/
Description	ดูรายการคำร้องขอของตนเอง
METHOD	GET
HEADER	Authorization: Bearer {token}
PAYLOAD	

Response Status	Response Data
200	<pre>[{ id: 1, title: "การประชุม...", status: "processing", category: [], detail: "www.ai-tagging.com/content/1" }, { id: 2, title: "ตารางนั่งและพื้นที่", status: "success", category: ["บันเทิง", "กีฬา"], detail: "www.ai-tagging.com/content/2" }]</pre>
401	<pre>{ errors: [{ "message": "Unauthorization" }] }</pre>

3.3.2.9 ค้นหาบทความโดยใช้ตัวกรองต่าง ๆ

Name: การค้นหาเนื้อหาโดยใช้ตัวกรองต่าง ๆ

Actors: ผู้ใช้งาน สมาชิก (ผู้ใช้งานที่ลงทะเบียนแล้ว) ผู้จัดการระบบ

Goal: ค้นหาบทความตามตัวกรองต่าง ๆ ได้

Precondition: -

Main success scenario:

1. ผู้ใช้งานเลือกค่าตัวกรองต่าง ๆ เช่น เรื่อง รีบประทายทางของข้อมูล หมวดหมู่ของเนื้อหา
2. ระบบดึงข้อมูลและแสดงผลบทความตามค่าตัวกรองกลับไปยังผู้ใช้งาน



รูปที่ 3.20: Sequence diagram ของการค้นหาบทความโดยใช้ตัวกรองต่าง ๆ

ตารางที่ 3.9: API Document ของการแสดงรายการค้นหาด้วยตัวกรองต่าง ๆ

URL	/api/v1/content/search?filterweb=sanook, true&filtercategory=sport,politic
Description	แสดงรายการค้นหาด้วยตัวกรองต่าง ๆ
METHOD	GET
HEADER	-
QUERY	filterweb: string[] filtercategory: string[]
Response Status	Response Data
200	[{ id: 1, weblink: "www.sanook.com/xxx/yyy", status: "success", category: ["การเมือง"], title: "การประชุมเพื่อการปกครอง" content: "เนื้องด้วย....." }, { id: 2, weblink: "www.sanook.com/xxx/yyy", status: "success", category: ["การเมือง"], title: "การประชุมเพื่อการปกครอง" content: "เนื้องด้วย....." }]

3.3.2.10 ค้นหาบุคลากรโดยใช้เลขอ้างอิงเนื้อหา

Name: การค้นหาเนื้อหาโดยใช้เลขอ้างอิงเนื้อหา

Actors: ผู้ใช้งาน สมมุติ (ผู้ใช้งานที่ลงทะเบียนแล้ว) ผู้จัดการระบบ

Goal: ค้นหาบุคลากรตามเลขอ้างอิงเนื้อหา

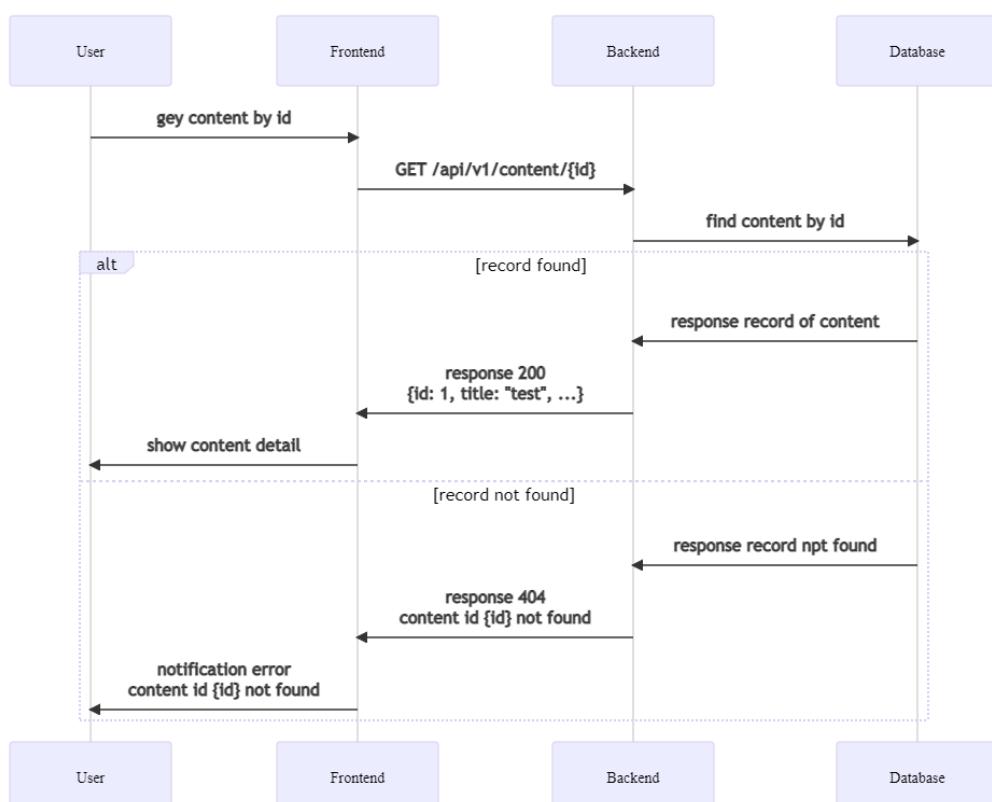
Precondition: -

Main success scenario:

1. ผู้ใช้งานทำการค้นหาบุคลากรโดยใช้เลขอ้างอิงเนื้อหา
2. ระบบค้นหาบุคลากรโดยใช้เลขอ้างอิงเนื้อหา
3. ระบบแสดงผลบุคลากรกลับไปยังผู้ใช้งาน

Extension (a):

- 2a. ระบบไม่พบบุคลากรตามเลขอ้างอิง
- 3a. ระบบแจ้งผู้ใช้งานว่าไม่พบเนื้อหาตามเลขอ้างอิง



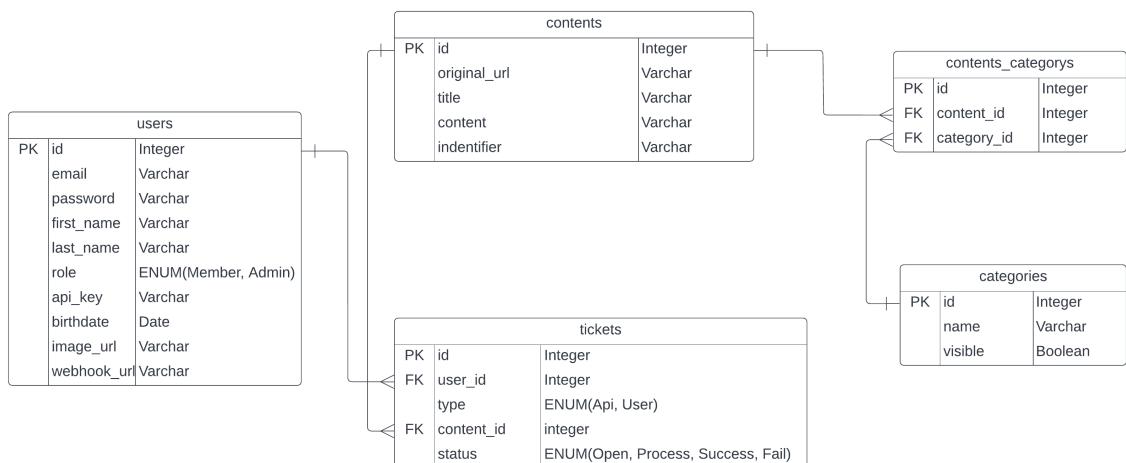
รูปที่ 3.21: Sequence diagram ของการค้นหาบุคลากรโดยใช้เลขอ้างอิงเนื้อหา

ตารางที่ 3.10: API Document ของการแสดงผลข้อมูลเกี่ยวกับเนื้อหาตามเลขอ้างอิงเนื้อหา

URL	/api/v1/content/{id}
Description	แสดงผลข้อมูลเกี่ยวกับเนื้อหาตามเลขอ้างอิงเนื้อหา
METHOD	GET
HEADER	-
PAYLOAD	url: string

Response Status	Response Data
200	{ id: 1, weblink: "'www.sanook.com/xxx/yyy'", status: "success", category: ["การเมือง"], title: "การประชุมเพื่อการปกคล่อง" content: "เนื่องด้วย....." }
404	{ errors: [{ message: "content id 1 doesn't not exist" }] }

3.4 โครงสร้างฐานข้อมูล



รูปที่ 3.22: โครงสร้างฐานข้อมูล

ตารางที่ 3.11: ตารางข้อมูลของสมาชิกในระบบ

Key	Attribute	Domain	Null	Description	Example
PK	id	int	No	เลขสำคัญประจำตัวผู้ใช้งานเป็นตัวเลขต่อเนื่องไม่มีซ้ำ	1
	password	varchar(64)	No	รหัสผ่านของบัญชีผู้ใช้ที่ต้องการเข้าสู่ระบบที่ถูกเข้ารหัสด้วย Argon2 algorithm	\$argon2d\$v=19\$m=12,t=3,p=1\$MGtvM3Rzc2R6bWpjMDAwMA\$09LkH+4Z7FPBQsVfa3O2rw
	first_name	varchar(64)	No	ชื่อของผู้ใช้งาน	พลวัต
	last_name	varchar(64)	No	นามสกุลของผู้ใช้งาน	สุการ์ตัน

ตารางที่ 3.11: ตารางข้อมูลของสมาชิกในระบบ

Key	Attribute	Domain	Null	Description	Example
	email	varchar(64)	No	อีเมล์ของผู้ใช้งาน ซึ่งจะไม่ซ้ำกับผู้ใช้งานคนอื่น ๆ	ponlawat@gmail.com
	role	varchar(64)	ENUM (Member, Admin)	สิทธิ์การเข้าใช้งาน	Admin
	api_key	varchar(255)	Yes	รหัสสำหรับการยืนยันตัวตนการเข้าใช้งานผ่านระบบภายนอก	8d969eef6ecad3c29a3a 629280e686cf03f5d5a86 aff3ca12020c923adc6c92
	birthdate	date	No	วันเกิดในรูปแบบ พ.ศ.	22/01/2544
	image_url	varchar(64)	Yes	ลิงค์สำหรับเก็บรูปภาพ	www.abc.com/ profile.png
	webhook_url	varchar(64)	Yes	ลิงค์ Url สำหรับส่งข้อมูลกลับ	www.efg.com/content/ webhook

ตารางที่ 3.12: ตารางข้อมูลคำร้องขอการวิเคราะห์หมวดหมู่ของบทความ

Key	Attribute	Domain	Null	Description	Example
PK	id	int	No	เลขลำดับเรียงอิงเนื้อหาเป็นตัวเลขต่อเนื่องไม่ซ้ำ	1
FK	user_id	int	No	เลขลำดับเรียงตามผู้ใช้งาน	2
	type	Enum(Api, User)	No	ประเภทของการสร้างคำร้องขอ	User
FK	content_id	int	No	เลขลำดับเรียงอิงเนื้อหา	3
	status	Enum(Open, Processing, Success, Fail)	No	สถานะการดำเนินขอร้องขอ	Processing

ตารางที่ 3.13: ตารางบทความที่ทำการวิเคราะห์

Key	Attribute	Domain	Null	Description	Example
PK	id	int	No	เลขลำดับเรียงอิงเนื้อหาเป็นตัวเลขต่อเนื่องไม่ซ้ำ	1
	original_url	varchar(255)	Yes	ลิงค์ url ที่ต้องการวิเคราะห์เนื้อหา	https://www.sanook.com/money/ 888331/
	title	varchar(255)	Yes	หัวข้อของเนื้อหาที่ต้องการวิเคราะห์	APEC 2022 พาณิชย์ ยกขวนสินค้า GI ขั้นเวทีเอเปค
	content	text	Yes	เนื้อหาที่ต้องการวิเคราะห์	พาณิชย์ ยกขวนสินค้า GI ร่วมต้อนรับ ^{สุดยอดผู้นำในการประชุมเอเปค 2022} Soft Power นำอัตลักษณ์ของถิ่นไทยสู่ สายตาชาวโลก นายรุตติ์ไกร สีรีระพันธุ์ อธิบดีกร.....
	identifier	varchar(64)	Yes	ตัวอ้างอิงเนื้อหาซึ่งจะอยู่ในรูปแบบ {website}:{reference}	sanook:money/888331

ตารางที่ 3.14: ตารางหมวดหมู่ที่ใช้งานภายในระบบ

Key	Attribute	Domain	Null	Description	Example
PK	id	int	No	เลขลำดับระบุอ้างอิงเนื้อหาเป็นตัวเลขต่อเนื่องไม่ซ้ำ	1
	name	varchar(64)	No	ชื่อของประเภทหมวดหมู่	การเมือง
	visible	boolean	No	สถานะการแสดงผลให้ผู้อื่นเห็น	TRUE

ตารางที่ 3.15: ตารางความสัมพันธ์ของเนื้อหา กับ หมวดหมู่ต่าง ๆ

Key	Attribute	Domain	Null	Description	Example
PK	id	int	No	เลขลำดับระบุอ้างอิงเนื้อหาเป็นตัวเลขต่อเนื่องไม่ซ้ำ	1
FK	content_id	int	No	เลขลำดับระบุอ้างอิงเนื้อหา	3
FK	category_id	int	No	เลขลำดับระบุอ้างอิงหมวดหมู่	4

3.5 User Interface Design

3.5.1 User Interface Design Rule

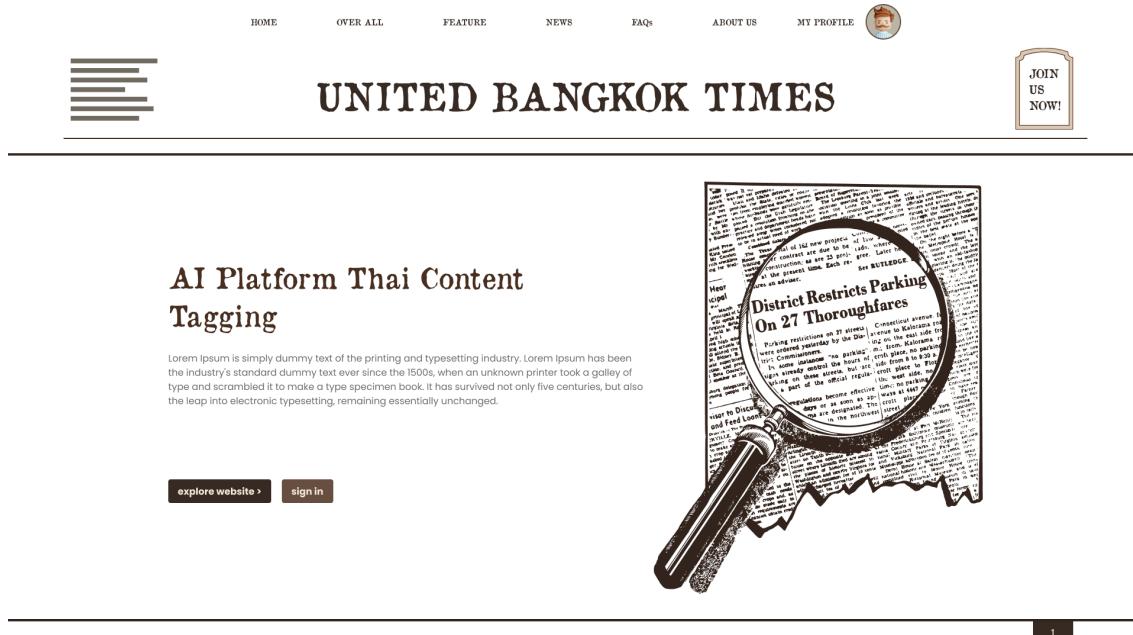
การออกแบบเว็บแอปพลิเคชัน AI Platform Thai Content Tagging มีหลักการออกแบบดังนี้

- มีการป้องกันการผิดพลาดของการกรอกข้อมูลใน Form เพื่อไม่ให้มีปัญหาต่อผู้ใช้และระบบ
- มี Navbar เพื่อสามารถนำผู้ใช้งานหน้าที่ต้องการได้
- มี Response กลับมาอย่างผู้ใช้งานในการกระทำนั้น ๆ ทุกครั้ง
- ไม่เปลี่ยนแปลงพฤติกรรมที่ผู้ใช้งานคุ้นเคย เช่น ตำแหน่งของปุ่มต่าง ๆ
- ใช้ Font OldNewpaperTypes, Poppins และ Lato
- โหนสีในการออกแบบ



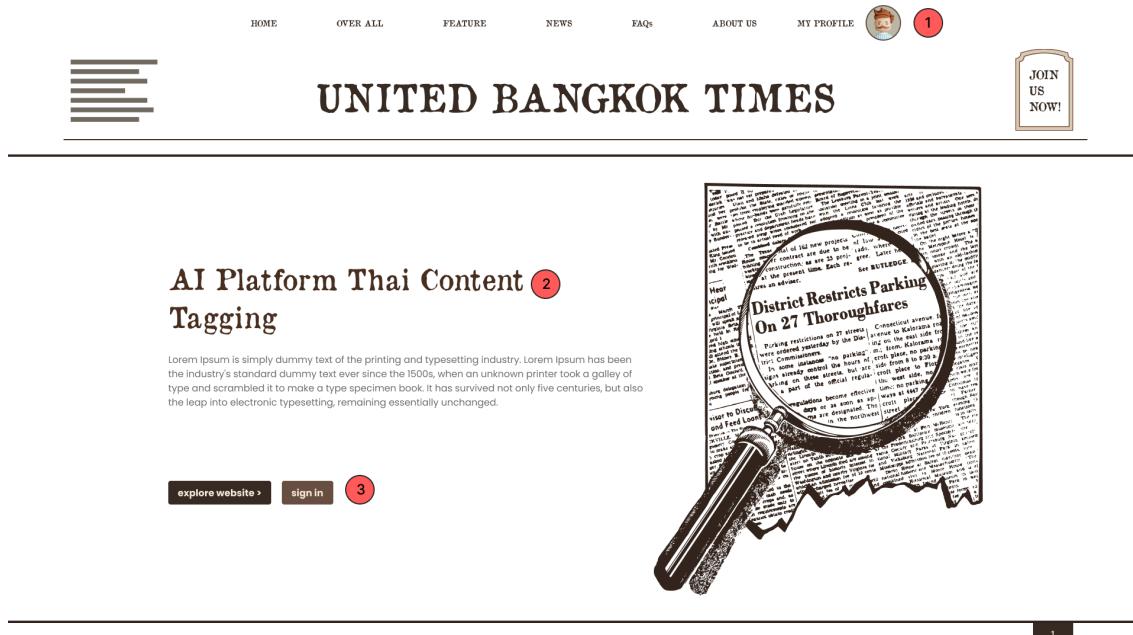
3.5.2 User Interface สำหรับผู้ใช้งาน

3.5.2.1 หน้า Landing



รูปที่ 3.23: หน้า Landing

เป็นหน้าแรกของเว็บแอปพลิเคชันที่ผู้ใช้งานจะได้พบเจอ ซึ่งหน้า Landing เป็นหน้าที่ผู้ใช้งานที่ว่าไปและสามารถเข้าถึงได้โดยที่ผู้ใช้งานที่ว่าไปสามารถเลือกได้ว่าจะรับข่าวเรื่องแอปพลิเคชันหรือจะสมัครสมาชิก โดยหน้า Landing page จะเป็นหน้าที่แสดงให้ผู้ใช้งานที่ว่าไปและสมาชิกทราบได้ว่าเว็บไซต์นี้เป็นเว็บแอปพลิเคชันเกี่ยวกับอะไรและสามารถใช้งานอย่างไร



รูปที่ 3.24: Landing page ที่แสดงแต่ละองค์ประกอบ

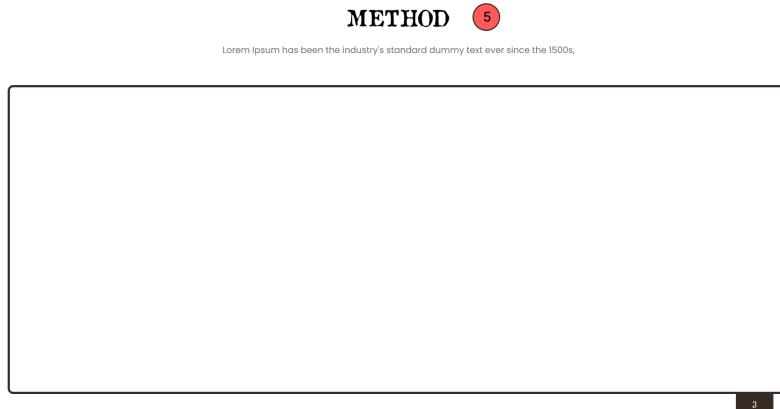
ในภาพนี้ (รูปที่:3.24) จะมีอยู่ 3 ส่วนด้วยกัน 'ได้แก่'

1. Navigation Bar (NavBar) เป็นส่วนที่จะนำพาผู้ใช้งานไปยังหน้าอื่น ๆ ซึ่งผู้ใช้งานสามารถเลือกดูหน้าเว็บแอปพลิเคชันทั้งหมดคร่าวๆ ได้ โดยแต่ละหน้าก็จะบอกรายละเอียดเกี่ยวกับเว็บแอปพลิเคชัน
2. ส่วนรายละเอียดเว็บแอปพลิเคชัน เป็นส่วนที่จะอธิบายรายละเอียดและการทำงานของเว็บแอปพลิเคชันคร่าวๆ เพื่อให้ผู้ใช้งานรู้จัก และเข้าใจแนวคิดของเว็บแอปพลิเคชันมากขึ้น
3. หากผู้ใช้งานต้องการเข้าชมหรือศึกษารายละเอียดของเว็บแอปพลิเคชันมากขึ้น ผู้ใช้งานสามารถเลือกคลิกที่ Explore Website ได้ ซึ่งเว็บแอปพลิเคชันก็จะพาผู้ใช้งานไปยังหน้า Home Page (รูปที่:3.39) หรือถ้าหากผู้ใช้งานต้องการเข้าสู่ระบบ ผู้ใช้งานสามารถเลือกคลิกที่ Sign In ได้ ซึ่งเว็บแอปพลิเคชันก็จะแสดงหน้า Sign In เพื่อให้ผู้ใช้งานได้ทำการเข้าสู่ระบบ (รูปที่:3.28)



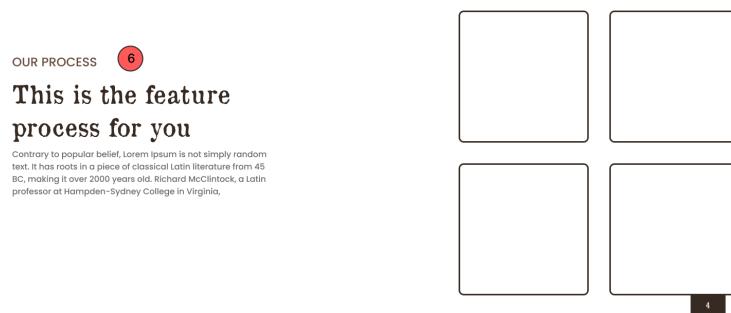
รูปที่ 3.25: หน้า Landing ในส่วนของ Overall

ในส่วนที่ 4 จะเป็นส่วนที่ยังคงอยู่ในหน้า Landing Page โดยเป็นส่วนที่ให้ข้อมูลและรายละเอียดเกี่ยวกับเว็บแอปพลิเคชัน AI Platform Thai Content Tagging ว่าเว็บแอปพลิเคชันนี้ให้ Service ที่มีบริการใดบ้าง



รูปที่ 3.26: หน้า Landing ในส่วนของ Method

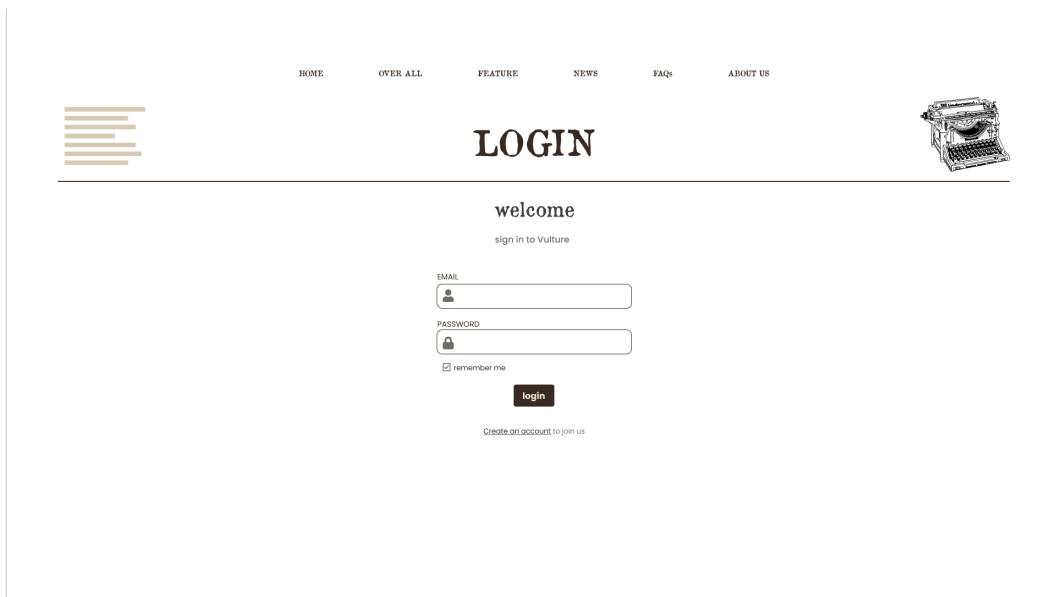
ในส่วนที่ 5 จะเป็นส่วนที่ยังคงอยู่ในหน้า Landing Page โดยเป็นส่วนที่แสดงกระบวนการการทำงานของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging



รูปที่ 3.27: หน้า Landing ในส่วนของ Feature

ในส่วนที่ 6 จะเป็นส่วนที่ยังคงอยู่ในหน้า Landing Page โดยเป็นส่วนที่อธิบายให้ผู้ใช้งานทราบว่าผู้ใช้งานจะสามารถใช้งานบริการของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging ได้อย่างไร

3.5.2.2 หน้า Login



รูปที่ 3.28: หน้า Login

เป็นหน้าสำหรับผู้ใช้งานที่ต้องการเข้าสู่ระบบเพื่อใช้บริการของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging โดยหากผู้ใช้งานทำการกรอก Email และ Password ที่ถูกต้อง เว็บแอปพลิเคชันก็จะพาผู้ใช้งานไปยังหน้า Home Page (รูปที่:3.39) ซึ่งในหน้า Login จะมีองค์ประกอบดัง ๆ ได้แก่

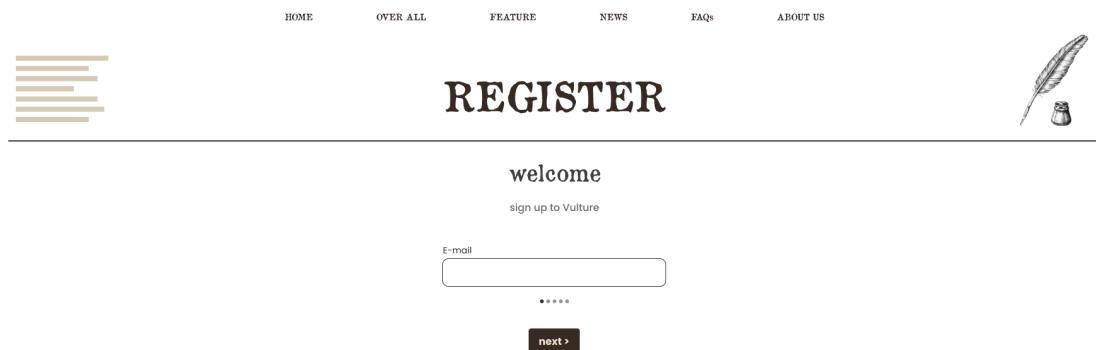
- Email เป็นส่วนสำหรับกรอก Email ที่ใช้ในการเข้าสู่ระบบ
- Password เป็นส่วนสำหรับกรอก Password ที่ใช้ในการเข้าสู่ระบบ
- Remember Me เป็นส่วนที่จดจำ Username และ Password เพื่อให้สามารถเข้าสู่ระบบได้ทันทีในครั้งถัดไป
- Create an account หากกรณีที่ผู้ใช้งานยังไม่มีบัญชี ผู้ใช้งานสามารถสมัครบัญชีได้ด้วยการคลิกที่ Create an account เว็บแอปพลิเคชันก็จะพาผู้ใช้งานไปยังหน้า Register (รูปที่:3.30)

คุณกรอกอีเมล์หรือพาสเวิร์ดผิดพลาด X

รูปที่ 3.29: กรอก Username หรือ Password ผิดพลาด

กรณีที่ผู้ใช้งานกรอก Email และ Password ผิดพลาด เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอก Email และ Password ใหม่อีกครั้ง

3.5.2.3 หน้า Register



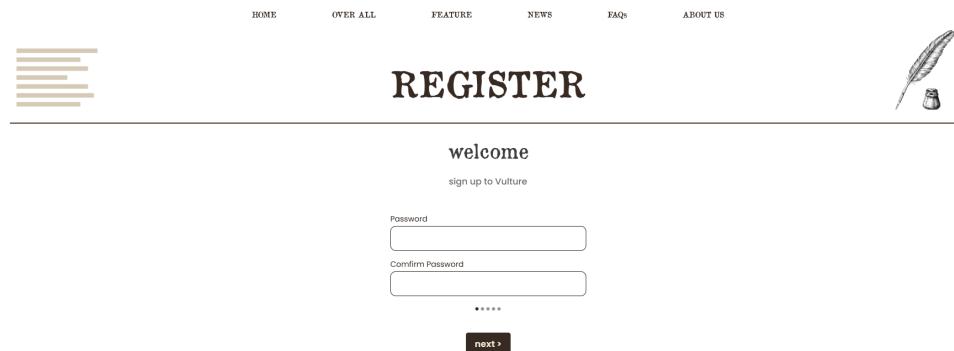
รูปที่ 3.30: หน้า Register ขั้นตอนกรอก Email

เป็นหน้าสำหรับผู้ใช้งานที่ต้องการสมัครสมาชิกเพื่อเป็นสมาชิกของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging โดยหากผู้ใช้งานทำการกรอก Email ถูกต้อง เว็บแอปพลิเคชันจะพาผู้ใช้งานไปยังหน้า Register ในขั้นตอนถัดไป (รูปที่: 3.32)

กรุณากรอกอีเมล์ของคุณให้ถูกต้อง X

รูปที่ 3.31: กรอก Email ผิดพลาด

กรณีที่ผู้ใช้งานกรอก Email ผิดพลาด เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอก Email ใหม่อีกครั้ง



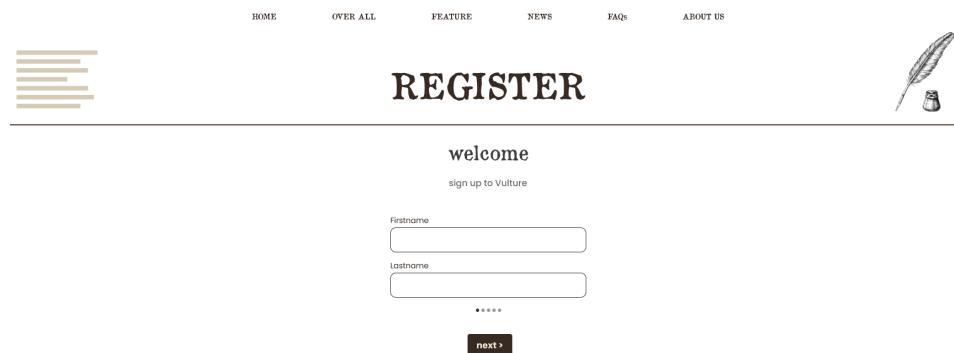
รูปที่ 3.32: หน้า Register ขั้นตอนกรอก Password

เป็นหน้าสำหรับให้ผู้ใช้งานตั้งและยืนยัน Password เพื่อสมัครสมาชิกของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging โดยเป็นขั้นตอนถัดจากการกรอก Email (รูปที่:3.30) หากผู้ใช้งานทำการกรอก Password ถูกต้อง เว็บแอปพลิเคชันจะพานำไปยังหน้า Register ในขั้นตอนถัดไป (รูปที่:3.34)



รูปที่ 3.33: กรอก Password ไม่ตรงกัน

กรณีที่ผู้ใช้งานกรอก Password 'ไม่ตรงกัน' เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอก Password 'ใหม่อีกครั้ง'



รูปที่ 3.34: หน้า Register ขั้นตอนกรอกชื่อและนามสกุล

เป็นหน้าสำหรับให้ผู้ใช้งานกรอกชื่อและนามสกุล เพื่อสมัครสมาชิกของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging โดยเป็นขั้นตอนถัดจากการกรอก Password (รูปที่:3.32) หากผู้ใช้งานทำการกรอกชื่อและนามสกุลถูกต้อง เว็บแอปพลิเคชันก็จะพาผู้ใช้งานไปยังหน้า Register ในขั้นตอนถัดไป (รูปที่:3.36)

✖

กรุณากรอกชื่อเป็นภาษาไทยเท่านั้น

รูปที่ 3.35: กรอกชื่อและนามสกุลเป็นภาษาอื่น

กรณีที่ผู้ใช้งานกรอกชื่อและนามสกุลเป็นภาษาอื่นที่ไม่ใช่ภาษาไทย เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอกชื่อและนามสกุลใหม่อีกครั้ง

The screenshot shows a registration form. At the top, there's a navigation bar with links: HOME, OVER ALL, FEATURE, NEWS, FAQS, and ABOUT US. To the right of the navigation is a feather quill icon. Below the navigation is a large 'REGISTER' button. Underneath the button, the word 'welcome' is displayed, followed by a link 'sign up to Vulture'. A large rectangular input field is present, with a placeholder 'Jennie Kim' and a timestamp '12/03/44'. Below this input field are two smaller input fields: 'Display Name' containing 'Jennie Kim', 'Birthdate' containing '12/03/44', and 'Gender'. A 'next >' button is located at the bottom right of the form area.

รูปที่ 3.36: หน้า Register ขั้นตอนกรอกข้อมูลของผู้ใช้งาน

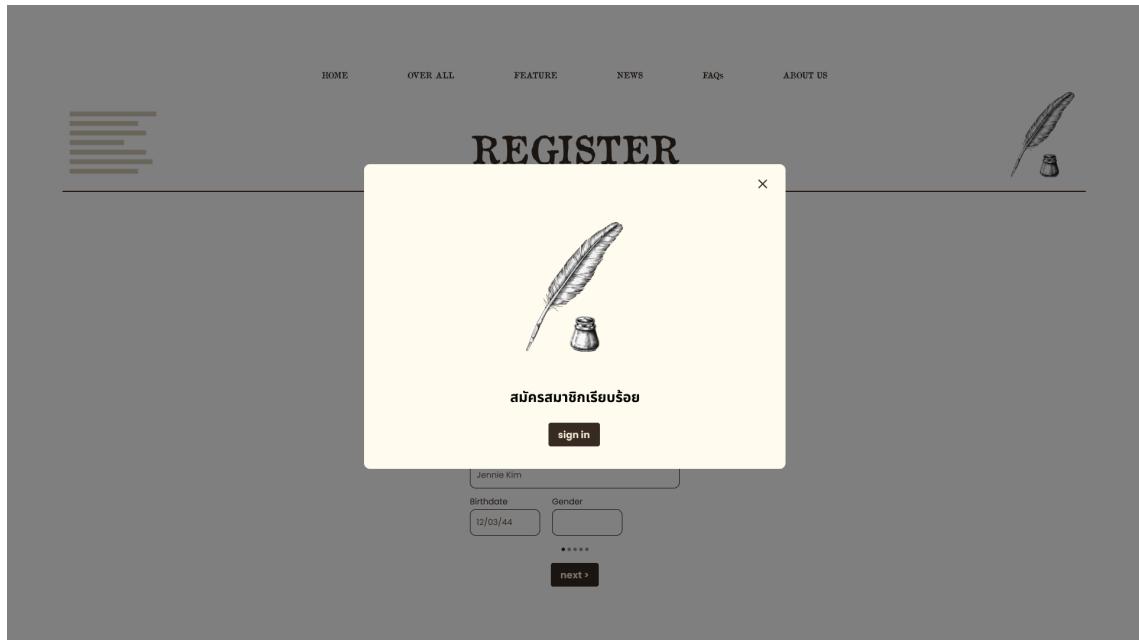
เป็นหน้าสำหรับให้ผู้ใช้งานกรอกข้อมูลชื่อเล่น วันเกิด และเพศ พร้อมแนบรูปถ่าย เพื่อสมัครสมาชิกของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging โดยเป็นขั้นตอนถัดจากการกรอกชื่อและนามสกุล (รูปที่:3.34) หากผู้ใช้งานทำการกรอกข้อมูลครบถ้วน เว็บแอปพลิเคชันก็จะถือว่าผู้ใช้งานทำการสมัครสมาชิกเสร็จสิ้น (รูปที่:3.38)

✖

กรุณากรอกข้อมูลให้ครบถ้วน

รูปที่ 3.37: กรอก Password ไม่ตรงกัน

กรณีที่ผู้ใช้งานกรอกข้อมูลไม่ครบถ้วน เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานกรอกข้อมูลให้ครบถ้วน



รูปที่ 3.38: หน้า Register เมื่อสมัครสมาชิกสำเร็จ

เมื่อผู้ใช้งานสมัครสมาชิกของเว็บแอปพลิเคชัน AI Platform Thai Content Tagging เสร็จสิ้นแล้ว ระบบแอปพลิเคชันจะทำการแจ้งเตือนว่าผู้ใช้งานได้สมัครสมาชิกเรียบร้อยแล้ว

3.5.2.4 หน้า Home

UNITED BANGKOK TIMES

Red Velvet เพย์พาทเชอร์ใหม่ของ 3 สาว 'Irene-Seulgi-Wendy'
สำหรับ 'The ReVe Festival 2022 - Birthday'

วันนี้ (22 พฤษภาคม 2565) กลุ่มศิลปินวง K-pop ร้องเริง如 Red Velvet ของบริษัท SM Entertainment ปล่อยภาพกิ๊ฟเซ็ตในวันเกิดเป็นปีที่ 2 "The ReVe Festival 2022 - Birthday" ...

รูปที่ 3.39: หน้า Home

เป็นหน้าที่ผู้ใช้งานทั่วไปและสามารถเข้าถึงได้ สามารถอ่านบทความต่าง ๆ ได้จากหน้า Home Page โดยบทความในหน้า Home Page จะถูกจำแนกตามหมวดหมู่ที่เว็บแอปพลิเคชัน AI Platform Thai Content Tagging ได้จัดไว้

โดยจากภาพข้างต้น (รูปที่:3.39) ก็จะมีอยู่ 2 ส่วนด้วยกัน ได้แก่

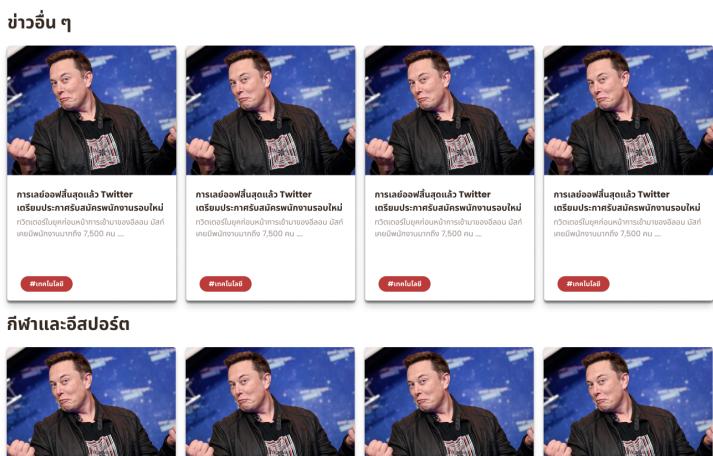
1. Tag แสดงหมวดหมู่ของเนื้อหาในบทความนั้น ๆ ที่ได้มากจากการวิเคราะห์หมวดหมู่ของบทความโดยเว็บแอปพลิเคชัน AI Platform Thai Content Tagging
2. เป็นส่วนหัวข้อของบทความ พร้อมกับมีคำอธิบายที่แสดงเนื้อหาภายในบทความคร่าวๆ



รูปที่ 3.40: หน้า Home ส่วนที่เกี่ยวข้องกับการวิเคราะห์หมวดหมู่ของสมาชิก

จากภาพข้างต้น (รูปที่:3.40) จะเป็นส่วนที่อ้างอิงถึงหน้า Home Page แต่จะเป็นส่วนที่เข้าถึงได้แค่สมาชิกเท่านั้น ผู้ใช้งานที่ไม่สามารถเข้าถึงได้ เนื่องจากเป็นส่วนที่เกี่ยวข้องกับข้อมูลของสมาชิกที่เคยใช้บริการเว็บแอปพลิเคชัน AI Platform Thai Content Tagging ซึ่งจะมีทั้งหมด 3 ส่วน ได้แก่

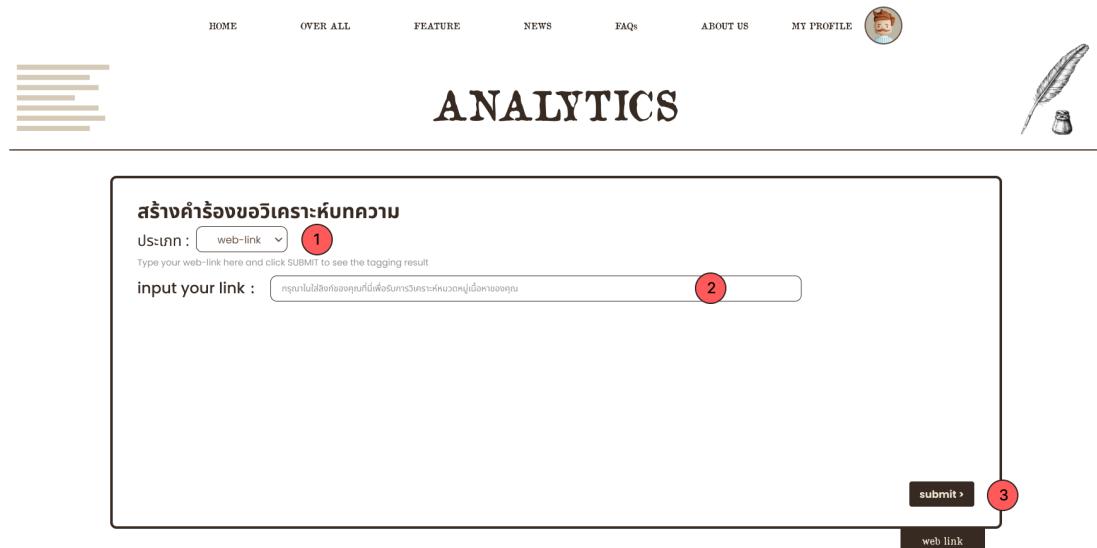
- ส่วนที่ 3 เป็นส่วนที่แสดงด้วยย่อการวิเคราะห์หมวดหมู่ของบทความล่าสุดของสมาชิกที่ได้ทำการร้องขอการวิเคราะห์หมวดหมู่
- ส่วนที่ 4 เป็นส่วนกรอบโปรไฟล์ของสมาชิก โดยสามารถตรวจสอบว่าบทความที่ทำการส่งไปจัดหมวดหมู่ประมวลผลสำเร็จ รอการประมวลผล หรือล้มเหลวหรือไม่ และจะแสดงบทความก่อน ๆ ที่เคยส่งเพื่อใบวิเคราะห์หมวดหมู่ นอกจากนี้ยังสามารถส่งคำร้องขอเพื่อวิเคราะห์การจัดหมวดหมู่ของบทความ และสามารถดูประวัติการส่งคำร้องได้
- ส่วนที่ 5 เป็นส่วนที่แสดงหมวดหมู่ที่รับแอปพลิเคชัน AI Platform Thai Content Tagging เปิดให้บริการ



รูปที่ 3.41: หน้า Home ส่วนแสดงข่าวอื่น ๆ เพิ่มเติม

ในส่วนที่ 6 จะเป็นส่วนที่แสดงเนื้อหาข่าวอื่น ๆ เพิ่มเติม ซึ่งเป็นส่วนที่ผู้ใช้งานที่ไม่ได้และสมาชิกสามารถเข้าถึงได้

3.5.2.5 หน้า Analytic



รูปที่ 3.42: หน้ากรณ์ที่สามารถเลือกตั้งการวิเคราะห์หมวดหมู่ของบทความด้วย Web Link

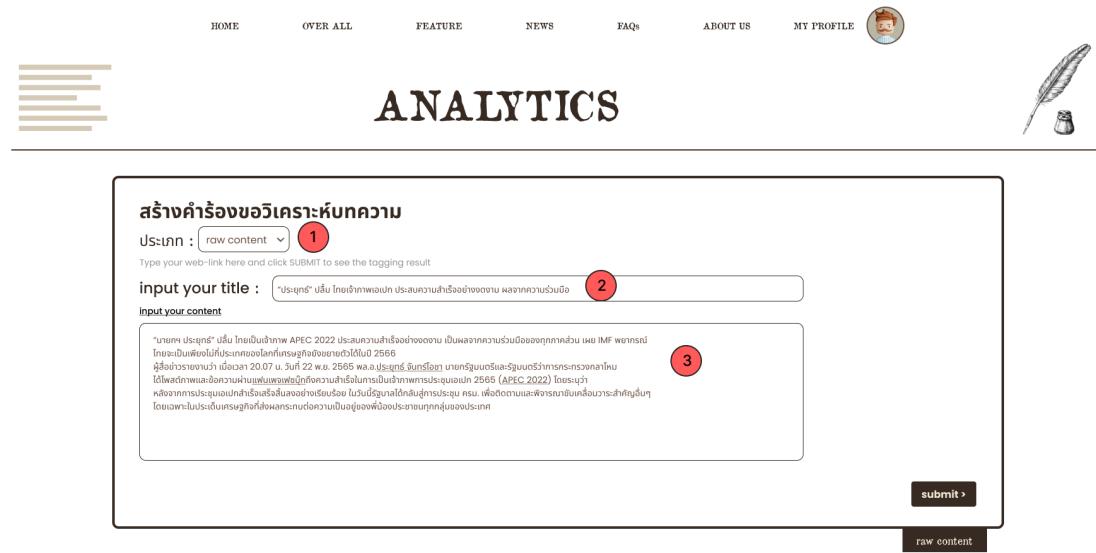
เป็นหน้าสำหรับการส่งคำขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วย Web Link ซึ่งจะมีอยู่ 3 ส่วน ได้แก่

1. เป็นส่วนที่ให้สามารถเลือกว่าจะส่งคำขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยอะไร โดยจะแบ่งเป็น 2 ประเภท คือ การส่งคำขอด้วย Web Link และการส่งคำขอด้วยบทความ โดยสามารถเลือกได้จาก Drop Down
2. เป็นส่วนสำหรับใส่ Web Link ที่ต้องการวิเคราะห์หมวดหมู่ของบทความ
3. ปุ่มยืนยันเพื่อส่งคำขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วย Web Link



รูปที่ 3.43: กรณ์ที่กรอก Web Link ไม่ถูกต้อง

กรณ์ที่ผู้ใช้งานกรอก Web Link ไม่ถูกต้อง เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอก Web Link ใหม่อีกครั้ง



รูปที่ 3.44: หน้ากรณีที่สามารถตั้งการวิเคราะห์หมวดหมู่ของบทความด้วยเนื้อหาในบทความ

เป็นหน้าสำหรับการส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยเนื้อหาในบทความ ซึ่งจะมีอยู่ 3 ส่วนหลัก ๆ ได้แก่

1. เป็นส่วนที่ให้สามารถเลือกว่าจะส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยอะไร โดยจะแบ่งเป็น 2 ประเภท คือ การส่งคำร้องขอด้วย Web Link และการส่งคำร้องขอด้วยบทความ โดยสามารถเลือกได้จาก Drop Down
2. เป็นส่วนสำหรับใส่หัวข้อของเนื้อหาในบทความที่ต้องการวิเคราะห์หมวดหมู่
3. เป็นส่วนสำหรับใส่เนื้อหาในบทความที่ต้องการวิเคราะห์หมวดหมู่ของบทความ และจึงกดปุ่มยืนยันเพื่อส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยเนื้อหาในบทความ

กรุณากรอกข้อมูลให้ครบถ้วน



รูปที่ 3.45: กรณีที่กรอกข้อมูลไม่ครบถ้วน

กรณีที่ผู้ใช้งานกรอกข้อมูลไม่ครบถ้วน เว็บแอปพลิเคชันจะทำการแจ้งเตือนให้ผู้ใช้งานทราบและให้ผู้ใช้งานกรอกข้อมูลให้ครบถ้วน

3.5.2.6 หน้า My Ticket

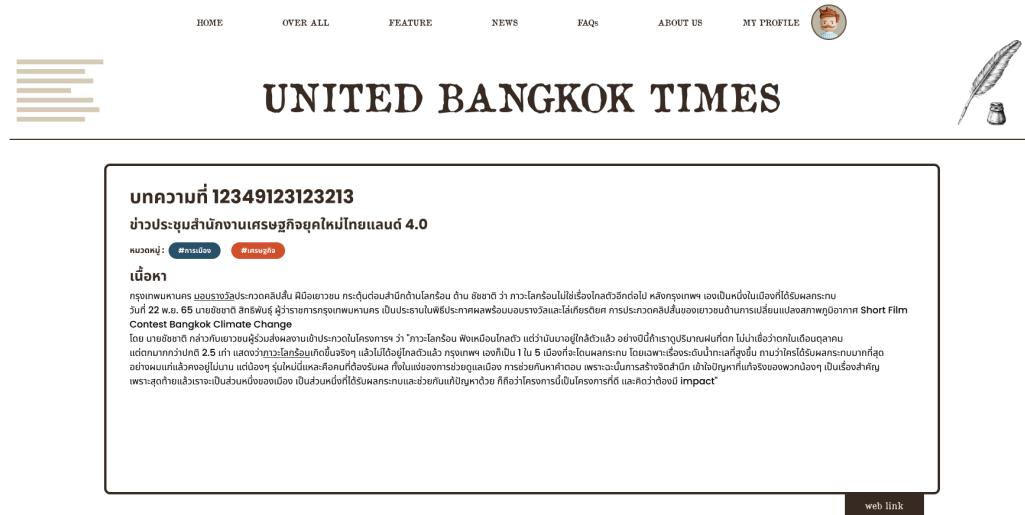
Reference no.	Title	Status	Category	Detail
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคให...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	success	#มีการปิด #การเมือง	ผู้สนับสนุน
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	failed	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-
12349123123213	ประชุมสานักงานนเครื่องก่ออุคใหป...	pending	-	-

รูปที่ 3.46: หน้า My Ticket

เป็นหน้าแสดงผลสถานะคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบุคลากรของสมาคมฯ ซึ่งจะมีอยู่ 6 ส่วน ได้แก่

- ส่วนแสดงเลขที่อ้างอิงของบุคลากรที่สมาคมฯได้ส่งคำร้องเพื่อวิเคราะห์หมวดหมู่ของบุคลากร
- ส่วนแสดงชื่อหัวข้อของบุคลากร
- ส่วนแสดงสถานะดำเนินการของคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบุคลากรนั้น ๆ โดยมี 3 สถานะด้วยกัน ได้แก่ อยู่ในระหว่างรอประเมินผล (Pending) สำเร็จ (Success) และล้มเหลว (Failed)
- ส่วนแสดงหมวดหมู่ที่ได้จากการวิเคราะห์หมวดหมู่ของบุคลากรโดยเว็บแอปพลิเคชัน AI Platform Thai Content Tagging หากล้มเหลวหรือดำเนินการจะไม่แสดงผลในส่วนนี้
- ส่วนรายละเอียดเพิ่มเติม หากดำเนินการวิเคราะห์หมวดหมู่ของบุคลากรสำเร็จ จะสามารถคลิกดูเพิ่มเติมได้
- ส่วนค้นหาคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบุคลากรของสมาคมฯ

3.5.2.7 หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ

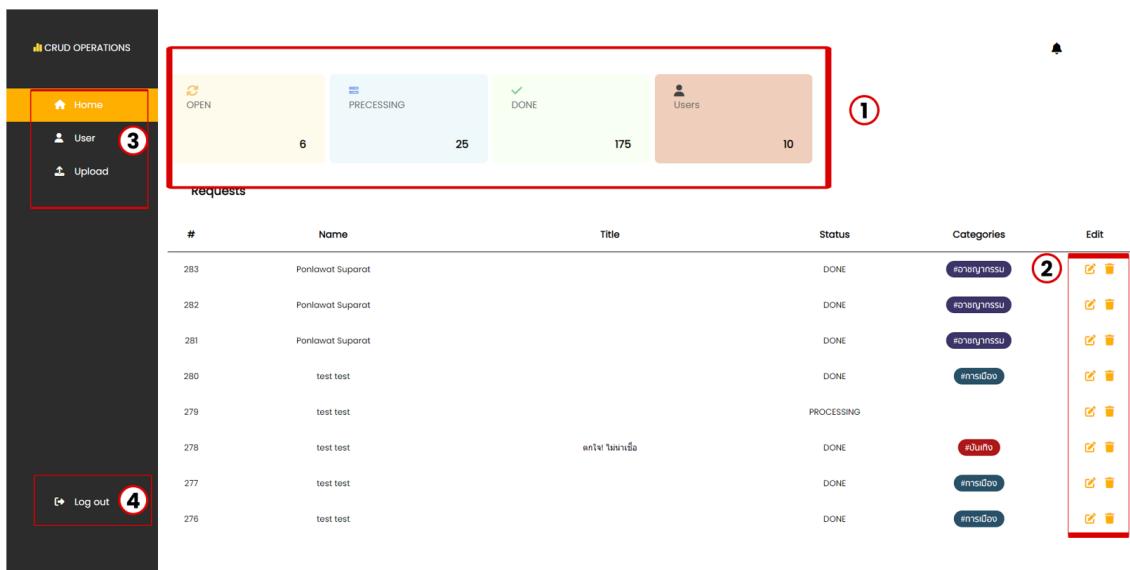


รูปที่ 3.47: หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ

เป็นหน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ โดยจะแสดงเลขจำนวน หัวข้อบทความ เนื้อหาในบทความ และหมวดหมู่ที่ได้จากการวิเคราะห์โดยเว็บแอปพลิเคชัน AI Platform Thai Content Tagging

3.5.3 User Interface สำหรับผู้ดูแลระบบ

3.5.3.1 หน้า Home

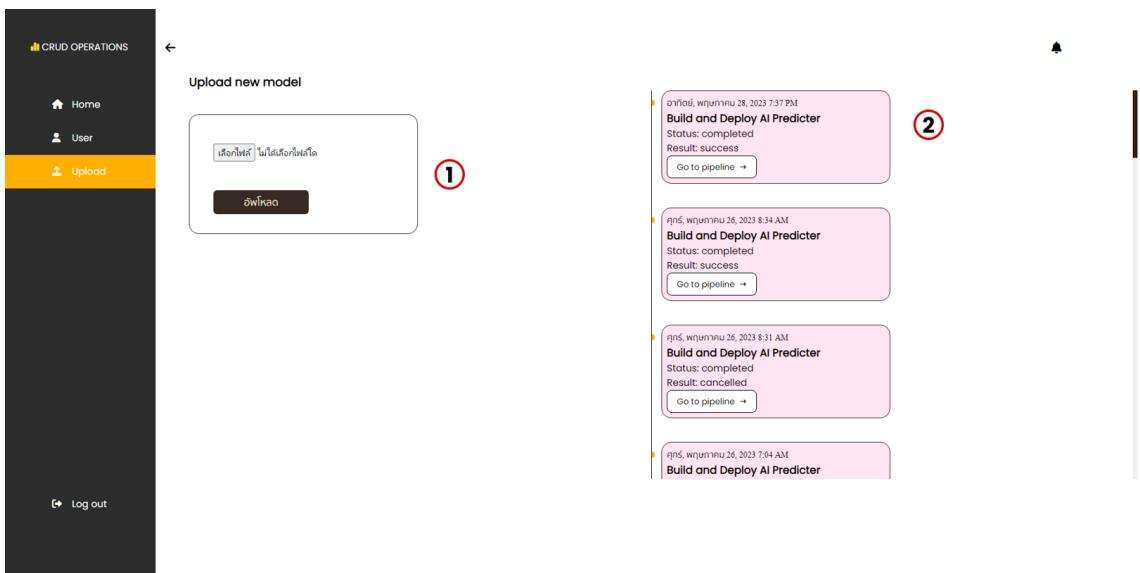


รูปที่ 3.48: หน้า Home ของผู้ดูแลระบบ

เป็นหน้าสำหรับผู้ดูแลระบบ เพื่อใช้ในการจัดการข้อมูลและระบบต่าง ๆ ได้แก่ ข้อมูลของผู้ใช้งานหรือหมวดหมู่ของเนื้อหาในระบบ โดยในภาพนี้ (รูปที่ 3.48) จะมีอยู่ 4 ส่วนด้วยกัน ได้แก่

- เป็นส่วนที่แสดงสถานะต่าง ๆ ได้แก่ จำนวนผู้ใช้งานที่ส่งคำร้องขอในการวิเคราะห์หมวดหมู่ให้กับบทความ จำนวนบทความที่กำลังประมวลผล จำนวนบทความที่ประมวลผลเสร็จเรียบร้อย และจำนวนสมาชิกทั้งหมดภายในระบบ
 - เป็นส่วนที่แสดงรายละเอียดของบทความต่าง ๆ ที่ผู้ใช้งานได้ส่งคำร้องใน การวิเคราะห์เข้ามาในระบบ ซึ่งผู้ดูแลระบบสามารถแก้ไขหรือลบบทความได้ผ่านทางหน้านี้
 - เมนูสำหรับไปยังหน้าต่าง ๆ ประกอบไปด้วย หน้าหลัก หน้าผู้ใช้งาน และหน้านำเข้าโนเดล ตามลำดับ
 - ปุ่มสำหรับออกจากระบบ

3.5.3.2 หน้า Upload New Model



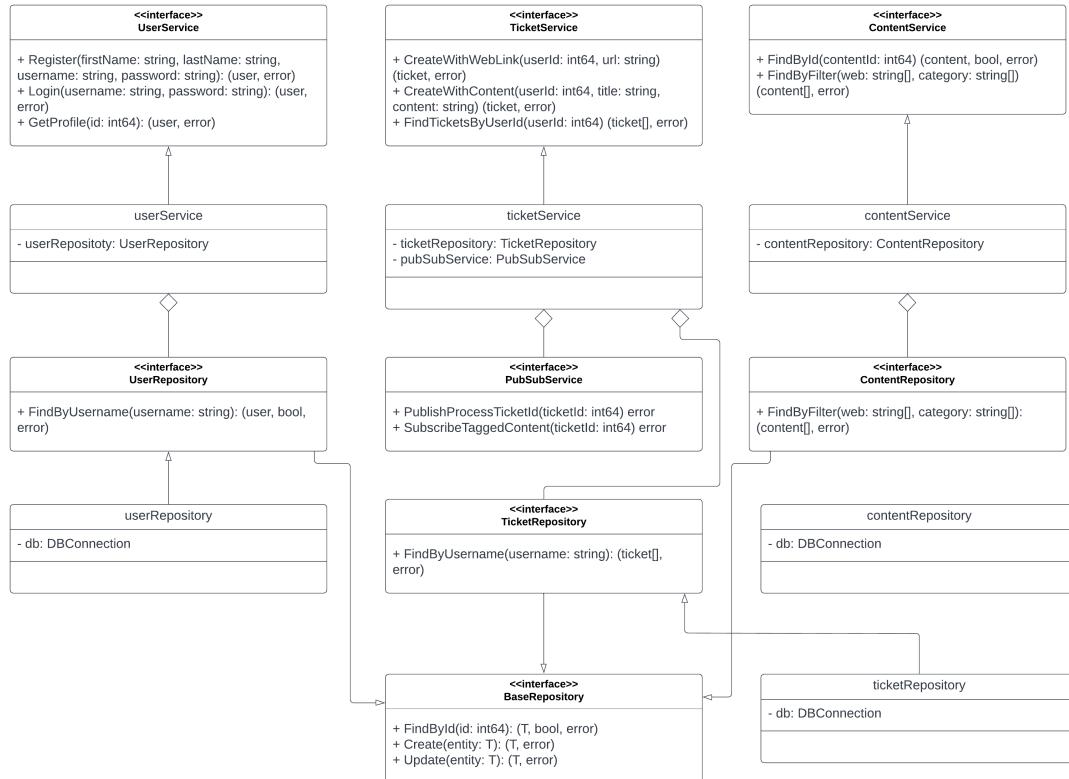
รูปที่ 3.49: หน้า Upload โมเดลใหม่

เป็นหน้าสำหรับผู้ดูแลระบบ เพื่อใช้ในอัพเดตโมเดลเข้าไปใหม่ในระบบ โดยในภาพนี้ (รูปที่ 3.49) จะมีอยู่ 2 ส่วนด้วยกัน ได้แก่

- ผู้ดูแลระบบสามารถอัพเดทโมเดลผ่านทางเว็บแอปพลิเคชันด้วยการเลือกที่ปุ่มอัพโหลดไฟล์ `joblib` เท่านั้น
 - เป็นส่วนที่แสดงให้เห็นสถานะการอัพเดทโมเดล โดยจะสามารถเลือกที่ปุ่ม "Go to pipeline" เพื่อดูรายละเอียดเพิ่มเติมได้

3.6 Class Diagram

การออกแบบ Class Diagram จะออกแบบให้มีความ Loose Coupling โดยจะมีการทำ Interface และ Dependency Injection ในการเชื่อมต่อ กันระหว่างคลาสต่าง ๆ เพื่อให้โค้ดสามารถทดสอบได้ง่าย



รูปที่ 3.50: Class Diagram ของระบบ

3.7 การออกแบบการทดลอง

3.7.1 User Evaluation

ทดสอบเว็บแอปพลิเคชันกับผู้ใช้งานจำนวน 20 คน โดยเก็บข้อมูลทั้งการตอบสนองต่อความต้องการและประสบการณ์การใช้งานของผู้ใช้งาน เพื่อเก็บรวบรวมเป็นข้อมูลสำหรับนำไปพัฒนาปรับปรุงเว็บแอปพลิเคชันให้สามารถตอบสนองความต้องการได้ดียิ่งขึ้น

3.7.2 Application Evaluation

การประเมินผล Application จะประเมินใน 2 รูปแบบ ได้แก่

- การประเมินความสามารถของ Application นั้น ๆ ว่าสามารถทำฟังก์ชันไหนได้บ้าง เช่น สามารถนำข้อมูลที่มีอยู่มาจัดเรียงตามที่ต้องการ สามารถแสดงผลการวิเคราะห์ที่หมวดหมู่ของบทความได้
- การประเมินประสิทธิภาพของ Application ว่าสามารถทำงานได้ดีแค่ไหน เช่น จำนวนข้อมูลที่สามารถรองรับได้ ความเร็วในการทำงาน ความแม่นยำในการทำงาน

3.7.3 Artificial Intelligence Evaluation

การประเมินผลติดตามการใช้งาน Artificial Intelligence ก็มีส่วนสำคัญในการพัฒนา Artificial Intelligence ให้มีความแม่นยำมากยิ่งขึ้น ซึ่งจำเป็นต้องใช้ข้อมูลที่ถูกต้องเป็นจำนวนมาก สำหรับการทดสอบความแม่นยำของ Artificial Intelligence เพื่อให้ Artificial Intelligence สามารถใช้เป็นแหล่งข้อมูลอ้างอิงและวิเคราะห์ผลได้แม่นยำมากยิ่งขึ้น กล่าวคือต้องใช้บทความที่มีการจัดหมวดหมู่อย่างถูกต้องจำนวนมาก เพื่อนำมาทดสอบกับ Artificial Intelligence และเปรียบเทียบผลลัพธ์ว่าตรงตามหมวดหมู่ของบทความที่

เตรียมมาหรือไม่ โดยจะใช้ตัววัดผลคือ ค่า Accuracy, Precision, Recall, F1 Score และ Loss Function ในการวัดประสิทธิภาพของ Artificial Intelligence

บทที่ 4 ผลการดำเนินงาน

4.1 การเก็บข้อมูล

ในส่วนของการเก็บข้อมูลสำหรับการพัฒนาโมเดล จะได้ผลลัพธ์ดังนี้

การเก็บ URL จากหน้าเว็บไซต์ ซึ่งในขั้นตอนนี้สามารถทำได้ 2 วิธี ได้แก่ วิธีแรก คือ การบันทึกหน้าเว็บไซต์ที่มีข้อมูลบทความจำนวนมาก นำมาอ่านข้อมูล HTML และจึงเก็บ URL ของบทความต่าง ๆ และวิธีที่สอง คือ การอ่าน URL ผ่าน Request โดยเป็นไปได้จากการส่ง Request ร้องขอเพื่อดูบทความเพิ่มเติม เมื่อได้ URL มาครบตามที่ต้องการแล้ว จะทำการดึงข้อมูลเป็นไฟล์ HTML จาก URL ที่เก็บมาจากหน้าเว็บไซต์ จากนั้นทำการเข้าลีฟ์ไฟล์ HTML ที่ได้จาก URL ที่เก็บมาจากหน้าเว็บไซต์และทำการดึงเนื้อหาบทความด้วยการใช้ Python BeautifulSoup มาเก็บไว้ เพื่อเป็นชุดข้อมูลสำหรับพัฒนาโมเดล ได้ผลลัพธ์ดังรูปที่ 4.1

รูปที่ 4.1: ตัวอย่างการดึงเนื้อหาบทความจากไฟล์ HTML ด้วยการใช้ Python BeautifulSoup

ทำการแปลงเนื้อหาที่ได้จากรูปภาพที่ 4.1 ซึ่งเป็นเนื้อหาที่เป็นภาษาไทยให้กลายเป็นภาษาอังกฤษด้วยการใช้ Translator Service เพื่อใช้สำหรับการพัฒนาโมเดลที่ใช้ข้อมูลเป็นภาษาอังกฤษ ได้ผลลัพธ์ดังรูปที่ 4.1

1 The young man drove the rich on the highway. Lift them up the police station Punch the police, the nose is broken-the little finger is broken.[The young man lifted up the police station. I asked the reason to set the rubber cone on the expressway. After driving until the car is damaged The police clarified it was punching the nose. The little hand of the right hand (28 Nov 64) at 8:00 pm, Police Lieutenant Aran Chavanon, Deputy Deputy Chief of Staff, Express Traffic Control Center, Kor Kor. Yesterday (27 Nov 64) at approximately 23.00 hrs. While he performed his duties at the Expressway 1 police station, he had a name later. The police station by informing that at 8:00 pm, he drove a white Honda Jazz brand on the Port 2 Expressway heading to Bang Na. Before Chonchon Rich Tang Which is located on the traffic surface Until causing the car to be damaged He therefore tried to clarify the cause of the rubber cone and explain the law to the police. Revealed that after that, Mr. Sittichok has argued. Then became anger and resentment He therefore informed him to contact the traffic officer during office hours. But the man spoke loudly And bring the phone to take pictures So he raised his hand to close the camera and told him to calm down. Mr Sittthok used the fist to fight into his face until the nosebleeds flowed. While the group of male friends who came with the attitude to attack, Pol. After Mr Sittthok caused the incident, he drove out of the police station. He was sent to treatment at the police hospital. The doctor sewed the wounds around the nose and diagnosed that the right hand bone right hand, broken. Revealed that After the incident, he learned that while he was treated Mr Sittichok, the parties, reported to condone themselves to assault with the investigating officers at the Port Police Station, by claiming that the side of the police tried to draw a gun And punching Mr. Sittthok first He therefore reported to the Port Police Station. Initially, in the matter of reporting charges At this time, it is not possible to notify. Since the inquiry official is in the process of collecting evidence Including requesting Images from CCTV on the expressway To be considered

2 Grandma led the grandchildren to report more. The monk uses a rocket.Claiming 4 herds of hernia|Grandmother, Primary 5, reported to the offense. While another grade 6 grandmother came out, revealed that the grandchildren were hit 2 times, including 4 children being done, expected to have more victims (29 Nov 64) officials from social development and human security. , The police suppressing human trafficking And related agencies Traveled to meet Phra Sombat, 40 years old, Abbot of Pradu Temple To investigate the facts After becoming news Forcing grade 3 students and grade 5, spinning the genitals until the orgasm And later a grade 1 student came out and revealed that 5 years ago was hit by the abbot Doing the same. Most recently, there was another grade 6 grandmother to report. After the grandchildren told The abbot has done the same 2 times. In total, there were 4 children being done to report 2 monks and expected to add more victims. Because of this kind of way for many years, Mr Thongthip Phu Si to specialist in specialized social developers The Office of Social Development and Human Security states that there must be a thorough investigation. That the story that the child told Is it true? There will be many agencies together. Which must be fair to all parties Especially the abbot Who still denied that they did not act Confessed only that To massage the genitals only. Later, Grandma, the 5th grade boy who came out to give information to the first teacher. Traveled to meet the inquiry official Confirmed to condone the abbot or His Majesty's abundance While the children of Primary 5 were found, the inquiry official was 58 years old. That was abbot Do the same 2 times but do not dare to tell Grandma. Until the same school students revealed Therefore dare to tell Grandma After this, it will be allowed to follow the legal procedures. To protect your own grandchildren

รูปที่ 4.2: ตัวอย่างการแปลเนื้อหาจากภาษาไทยเป็นภาษาอังกฤษ

4.2 การออกแบบโมเดลสำหรับวิเคราะห์หมวดหมู่ของเนื้อหาในบทความต่าง ๆ

4.2.1 การเตรียมข้อมูลสำหรับการพัฒนาโมเดล

ข้อมูลที่ใช้ในการพัฒนาโมเดลจะมีทั้งหมด 2 รูปแบบด้วยกัน ได้แก่ ชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยและชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ ชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยจะมีทั้งหมด 9,000 ชุด และชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษจะมีทั้งหมด 9,000 ชุด ซึ่งมีลักษณะข้อมูลเป็นไฟล์ Text และมีหมวดหมู่กำกับในแต่ละชุดข้อมูล โดยมีรายละเอียดดังนี้

ตารางที่ 4.1: ตารางชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาไทย

หมวดหมู่	จำนวนชุดข้อมูลข่าว
อาชญากรรม	1,500
กีฬา	1,500
การเมือง	1,500
เศรษฐกิจ	1,500
บันเทิง	1,500
เทคโนโลยี	1,500
รวม	9,000

title	content	tag	
เล่นได้ทุกวัน! เปิดวาร์ป "คิม มี-จู" นางแบบแคน...	พาไปรู้สักกัน คิม มี-จู หนึ่งในคนดังทางโซเชียล...	sport	
อกใหญ่เป็นเหด!	อกใหญ่มีไว้ด้วยกัน! "มินามิ วาจิ" จากอดีตสาววายน์...	สือเป็นอีกหนึ่งสาวสวยแคนปลารักที่มีแฟชั่นคลับหน...	sport
ชีล่า : กีฬาสายแหวกมากความหมายที่นิยมถังแต่อ...	โลกใบนี้ยังมีอะไรอีกมากให้เราค้นหา กีฬากีฬา...	sport	
สรุปผลบอลพรีเมียร์ลีก นัดที่ 2 รันที่ 13-15 ส...	หัวตาราง แมนเชสเตอร์ ยูไนเต็ด เปิดบ้านชนะ 4-0 ในนัดนี้ ชี...	sport	
นางฟ้าชัดๆ "ฟิกเกอร์ เพ็ญฟิกเกอร์" นางแบบสาวสุดเอ...	สือเป็นอีกหนึ่งสาวที่ขึ้นชื่อบนการออกกำลังกายด้วย...	sport	

รูปที่ 4.3: ตัวอย่างชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาไทย

ตารางที่ 4.2: ตารางชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาอังกฤษ

หมวดหมู่	จำนวนชุดข้อมูลข่าว
อาชญากรรม	1,500
กีฬา	1,500
การเมือง	1,500
เศรษฐกิจ	1,500
บันเทิง	1,500
เทคโนโลยี	1,500
รวม	9,000

title	content	tag
Can play every day!Opening the warp "Kim Mi-Ju..."	Take to know Kim Mi-Ju, one of the famous Sout...	sport
Big chest is the reason!"Minami Wiji" from a f...	Regarded as another beautiful girl, raw fish t...	sport
Summary of the Premier League football match 2...	The head of the Man City opened the house, won...	sport
Clear angels "Phen Pheng", the most female mod...	Considered as another girl who likes to exerci...	sport
Come quickly!The most famous guru "Ten Hak" wa...	The famous Richard Key Guru revealed that the ...	sport

รูปที่ 4.4: ตัวอย่างชุดข้อมูลข่าวที่มีเนื้อหาเป็นภาษาอังกฤษ

นอกจากนี้คณะผู้จัดทำได้ทำการแบ่งชุดข้อมูลออกเป็น 2 ส่วน ส่วนแรกสำหรับการพัฒนาโมเดลและส่วนที่สองสำหรับการประเมินโมเดล ซึ่งจะแบ่งชุดข้อมูลแต่ละรูปแบบออกเป็นอัตราส่วน 8:2 (ชุดข้อมูลสำหรับการพัฒนาโมเดล : ชุดข้อมูลสำหรับการประเมินโมเดล) โดยมีรายละเอียดดังนี้

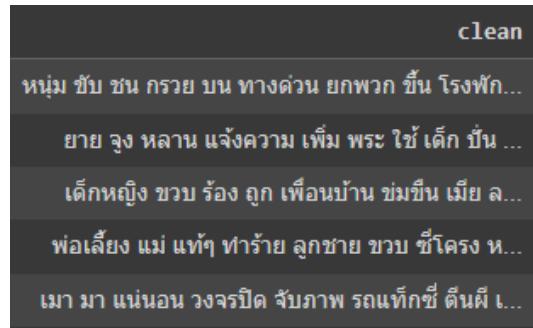
ตารางที่ 4.3: ตารางชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยสำหรับใช้พัฒนาและทดสอบโมเดล

หมวดหมู่	จำนวนชุดข้อมูล	จำนวนชุดข้อมูล
	สำหรับพัฒนาโมเดล	สำหรับประเมินโมเดล
อาชญากรรม	1,200	300
กีฬา	1,200	300
การเมือง	1,200	300
เศรษฐกิจ	1,200	300
บันทิง	1,200	300
เทคโนโลยี	1,200	300
รวม	7,200	1,800

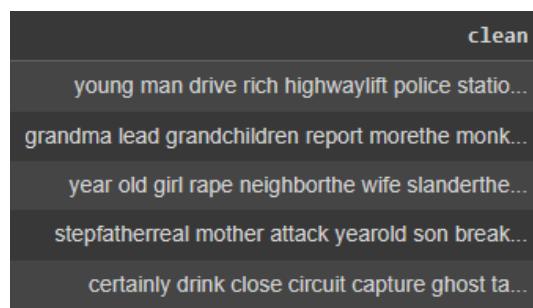
ตารางที่ 4.4: ตารางชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษสำหรับใช้พัฒนาและทดสอบโมเดล

หมวดหมู่	จำนวนชุดข้อมูล	จำนวนชุดข้อมูล
	สำหรับพัฒนาโมเดล	สำหรับประเมินโมเดล
อาชญากรรม	1,200	300
กีฬา	1,200	300
การเมือง	1,200	300
เศรษฐกิจ	1,200	300
บันทิง	1,200	300
เทคโนโลยี	1,200	300
รวม	7,200	1,800

การเตรียมข้อมูลสำหรับการพัฒนาโมเดล จะมีการทำการลบคำที่ไม่จำเป็น ลบตัวเลข หรือช่องว่าง เพื่อเป็นการลบการระบุของข้อมูล จากนั้นจึงทำการแบ่งคำสำหรับลบ Stop word และปรับคำให้อยู่ในรูปแบบที่เป็นบรรทัดฐานเพื่อลดความซ้ำซ้อนของข้อมูล แล้วจึงนำมาเรียบเรียงต่อกันเป็นประโยคใหม่ โดยนำหัวข้อและเนื้อหาร่วมกันเป็นบทความเดียว ทำให้สามารถถึง Feature ออกจากข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น



รูปที่ 4.5: ตัวอย่างชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยที่ผ่านการเตรียมข้อมูลแล้ว

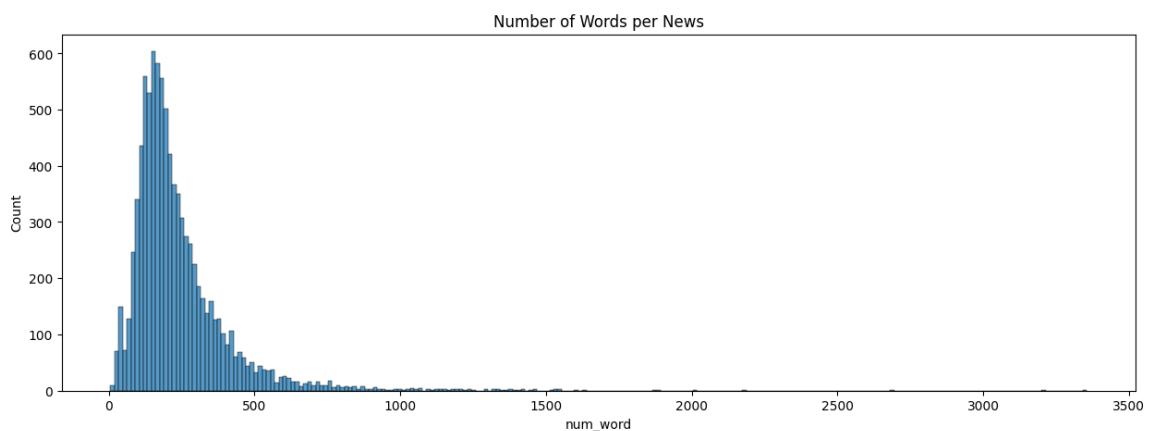


รูปที่ 4.6: ตัวอย่างชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษที่ผ่านการเตรียมข้อมูลแล้ว

4.2.2 Exploratory Data Analysis

4.2.2.1 ชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย

จากการสำรวจชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยทั้งหมด 9,000 ชุด พบร่ว่า แต่ละชุดข้อมูลมีการกระจายตัวของขนาดของข้อมูลดังนี้



รูปที่ 4.7: การกระจายตัวของจำนวนคำในแต่ละชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย

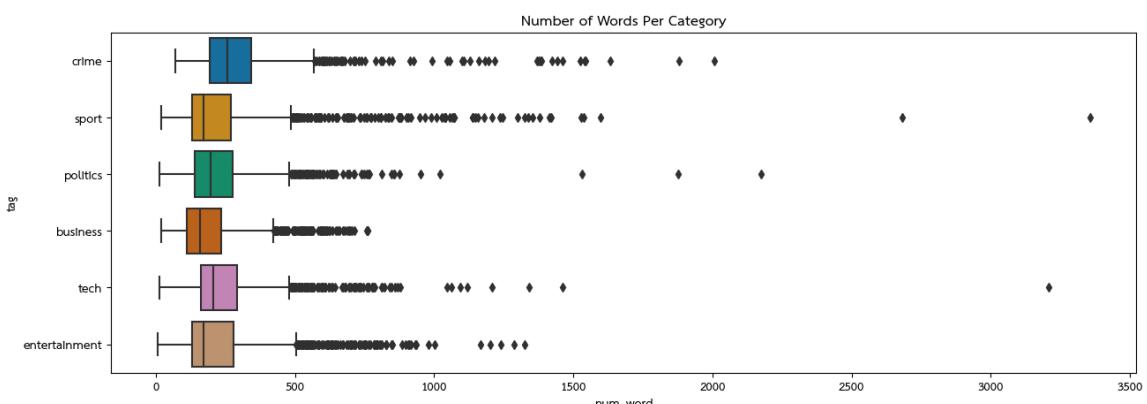
ตารางที่ 4.5: ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาไทยทั้งหมดหลังทำความสะอาดข้อมูล

Mean	STD	Min	Q1	Q2	Q3	Max
239.14	175.10	6.00	139.00	195.00	286.00	3356.00

จะเห็นได้ว่าการกระจายตัวของจำนวนคำในแต่ละชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย มีการกระจายตัวแบบเบื้องขวาและมี Outlier (รูปที่:4.7) ดังนั้นค่าเฉลี่ยของจำนวนคำในชุดข้อมูลจึงมากกว่าค่ากลางของจำนวนคำในชุดข้อมูล และเมื่อถูกการกระจายตัวประกอบกับค่าทางสถิติของจำนวนคำในเนื้อหาทั้งหมด (ตารางที่:4.5) พบว่า จำนวนคำในชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยส่วนใหญ่จะมีค่าอยู่ที่ 140-290 คำ โดยจำนวนคำที่น้อยที่สุดในชุดข้อมูล คือ 6 คำ และจำนวนคำที่มากที่สุดในชุดข้อมูล คือ 3,356 คำ

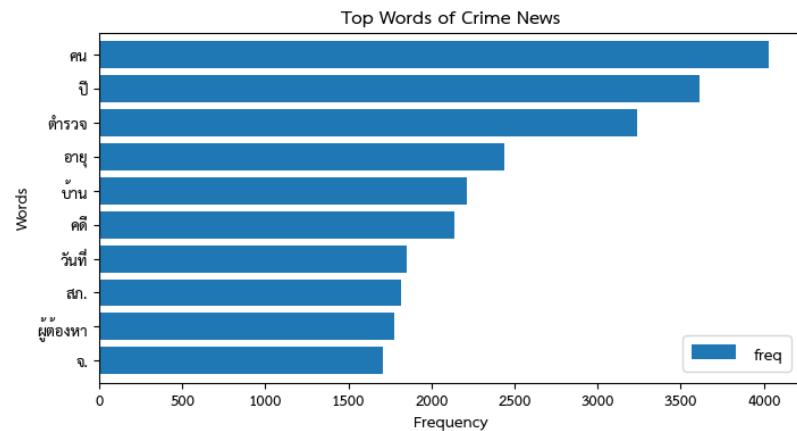
ตารางที่ 4.6: ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาไทยของแต่ละหมวดหมู่หลังทำความสะอาดข้อมูล

หมวดหมู่	Mean	STD	Min	Q1	Q2	Q3	Max
อาชญากรรม	294.51	182.57	71.00	193.00	257.00	343.00	2006.00
กีฬา	239.04	222.40	21.00	129.00	173.50	271.00	3356.00
การเมือง	227.62	147.87	12.00	140.75	196.00	276.00	2174.00
เศรษฐกิจ	190.63	127.83	21.00	110.00	160.00	235.25	762.00
บันเทิง	233.04	167.31	6.00	129.00	173.00	279.25	1327.00
เทคโนโลยี	249.99	171.35	12.00	163.00	206.00	291.00	3209.00

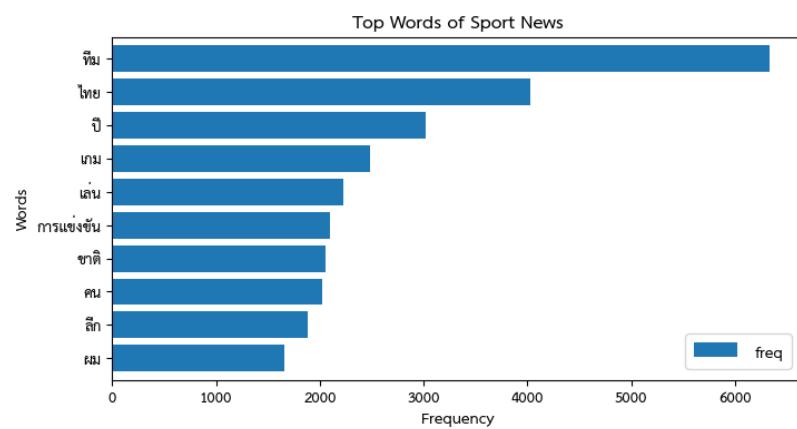


รูปที่ 4.8: ลักษณะของจำนวนคำในเนื้อหาที่เป็นภาษาไทยของแต่ละหมวดหมู่หลังทำความสะอาดข้อมูล

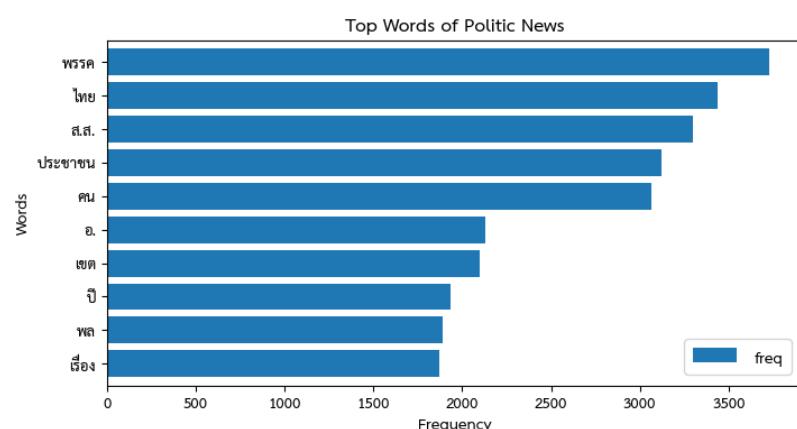
เนื่องจากในแต่หมวดหมู่ของชุดข้อมูลมีขนาดของข้อมูลแตกต่างกัน ซึ่งในชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยจะมีทั้งหมด 6 หมวดหมู่ด้วยกัน ดังนั้นจึงต้องมีการสำรวจการกระจายตัวของขนาดของข้อมูลในแต่ละหมวดหมู่ ซึ่งจากการที่ 4.6 และรูปภาพที่ 4.8 พบว่า จำนวนคำของเนื้อหาในหมวดหมู่อาชญากรรมส่วนใหญ่จะมีค่ามากกว่าหมวดหมู่อื่น ในขณะที่จำนวนคำของเนื้อหาในหมวดหมู่เศรษฐกิจ ส่วนใหญ่จะมีค่าน้อยกว่าหมวดหมู่อื่น นอกจากนี้ยังพบว่า จำนวนคำของชุดข้อมูลในหมวดหมู่กีฬามีการกระจายตัวของกลุ่มข้อมูลสูงสุด และจำนวนคำของชุดข้อมูลในหมวดหมู่เศรษฐกิจมีการกระจายตัวของกลุ่มข้อมูลต่ำสุด ซึ่งสังเกตได้จากค่า STD ในตารางที่ 4.6



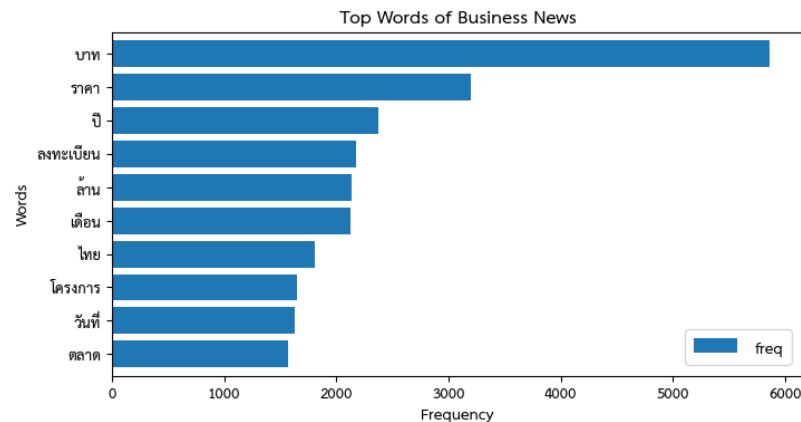
รูปที่ 4.9: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่อาชญากรรม



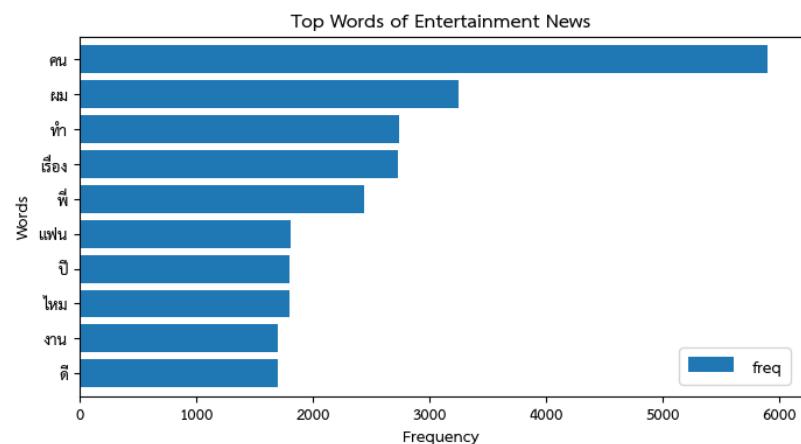
รูปที่ 4.10: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่กีฬา



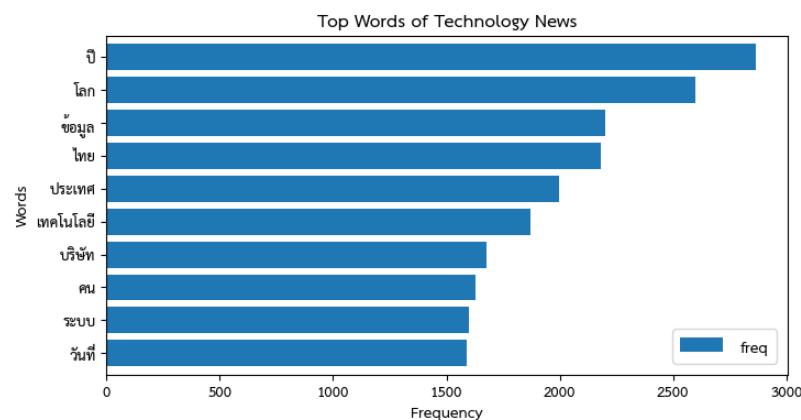
รูปที่ 4.11: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่การเมือง



รูปที่ 4.12: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่เศรษฐกิจ



รูปที่ 4.13: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่บันเทิง



รูปที่ 4.14: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของหมวดหมู่เทคโนโลยี

จากรูปภาพที่ 4.9 ถึงรูปภาพที่ 4.11 เป็นการแสดงผลคำที่พบมากที่สุด 10 อันดับแรกในชุดข้อมูลของแต่ละหมวดหมู่ เพื่อใช้สำหรับคุณวิเคราะห์ในมุมของคำที่พบบ่อยในแต่ละหมวดหมู่ ซึ่งจากข้อมูลข้างต้นพบว่า แต่ละหมวดหมู่ค่อนข้างที่จะมีคำที่พบบ่อยแตกต่างกัน แต่จะมีบางคำที่พบบ่อยในทุกหมวดหมู่ เช่น ปี คน ซึ่งถือว่าเป็นคำที่ว้าวไป ไม่ได้มีความสำคัญมากสำหรับข้อมูล

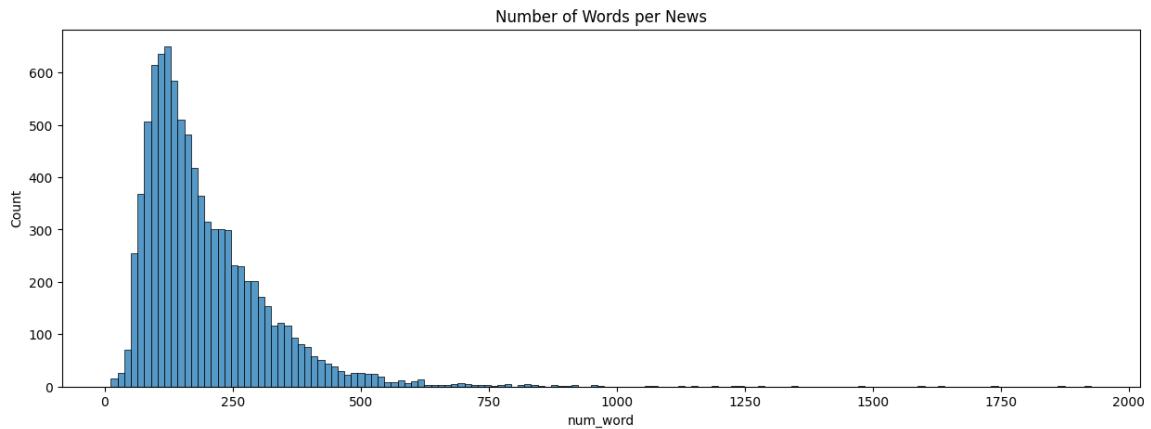


รูปที่ 4.15: 100 อันดับคำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของแต่ละหมวดหมู่

จากรูปภาพที่ 4.15 เป็นการแสดงผลคำที่พบมากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของแต่ละหมวดหมู่ โดยใช้ Word Cloud ในการแสดงผล

4.2.2.2 ชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ

จากการสำรวจชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษทั้งหมด 9,000 ชุด พบร้า แต่ละชุดข้อมูลมีการกระจายตัวของขนาดของข้อมูลดังนี้



รูปที่ 4.16: การกระจายตัวของจำนวนคำในแต่ละชุดข้อมูลที่มีเนื้อหาเป็นอังกฤษ

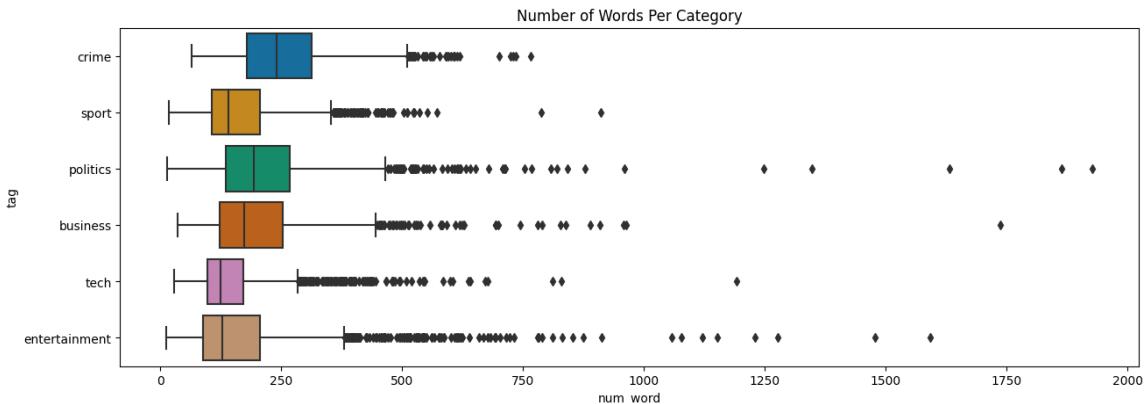
ตารางที่ 4.7: ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษทั้งหมดหลังทำความสะอาดข้อมูล

Mean	STD	Min	Q1	Q2	Q3	Max
195.38	125.35	12.00	112.00	163.00	248.00	1927.00

จะเห็นได้ว่าการกระจายตัวของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษทั้งหมดมีการกระจายตัวแบบเบื้องขวาและมี Outlier (รูปที่:4.16) ดังนั้นค่าเฉลี่ยของจำนวนคำในชุดข้อมูลสูงกว่าค่ากลางของจำนวนคำในชุดข้อมูล และเมื่อดูการกระจายตัวประกอบกับค่าทางสถิติของจำนวนคำในเนื้อหาทั้งหมด (ตารางที่:4.7) พบร้า จำนวนคำในชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษส่วนใหญ่จะมีค่าอยู่ที่ 110-250 คำ โดยจำนวนคำที่น้อยที่สุดในชุดข้อมูล คือ 12 คำ และจำนวนคำที่มากที่สุดในชุดข้อมูล คือ 1,927 คำ

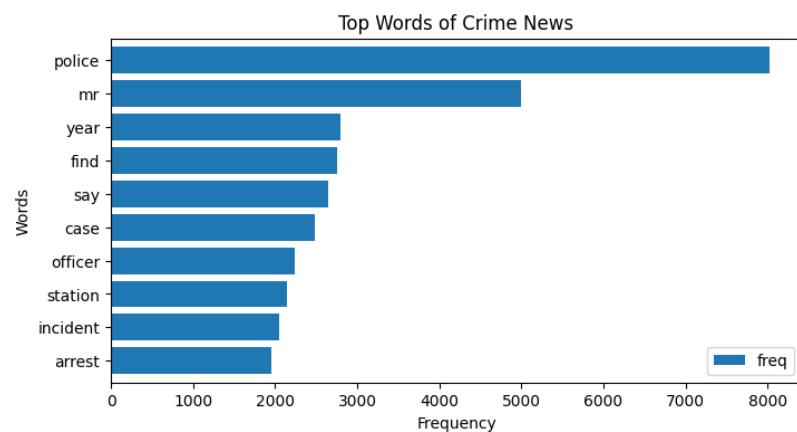
ตารางที่ 4.8: ตารางแสดงค่าทางสถิติของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่หลังทำความสะอาดข้อมูล

หมวดหมู่	Mean	STD	Min	Q1	Q2	Q3	Max
อาชญากรรม	254.96	104.65	64.00	179.00	239.00	312.00	766.00
กีฬา	167.08	91.22	18.00	106.00	141.00	206.00	911.00
การเมือง	218.88	140.71	14.00	135.00	193.00	268.00	1927.00
เศรษฐกิจ	201.04	123.04	35.00	122.75	173.00	252.00	1736.00
บันเทิง	175.07	151.97	12.00	88.00	127.00	205.00	1592.00
เทคโนโลยี	155.22	100.30	28.00	97.00	125.00	172.00	1191.00

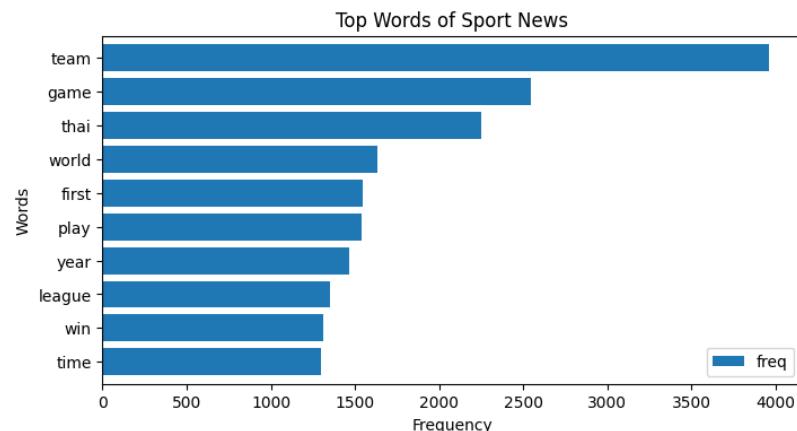


รูปที่ 4.17: ลักษณะของจำนวนคำในเนื้อหาที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่ทั้งทำความสะอาดข้อมูล

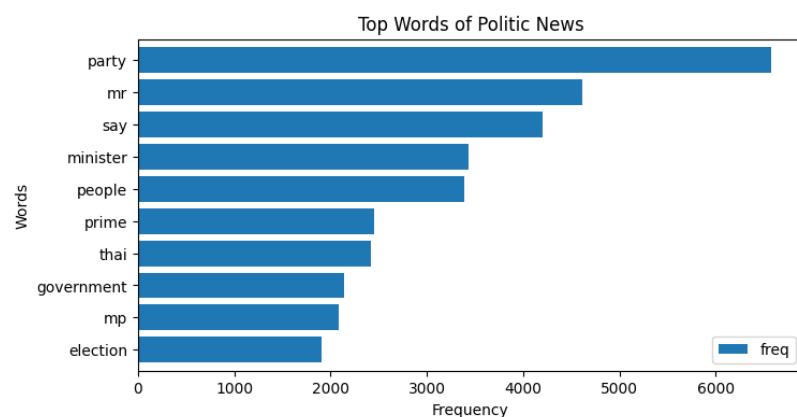
เนื่องจากในแต่ละหมวดหมู่ของชุดข้อมูลมีขนาดของข้อมูลแตกต่างกัน ซึ่งในชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษจะมีทั้งหมด 6 หมวดหมู่ด้วยกัน ดังนั้นจึงต้องมีการสำรวจการกระจายตัวของขนาดของข้อมูลในแต่ละหมวดหมู่ ซึ่งจากการที่ 4.8 และรูปภาพที่ 4.17 พบว่า จำนวนคำของเนื้อหาในหมวดหมู่อื่นๆ ค่อนข้างส่วนใหญ่จะมีค่ามากกว่าหมวดหมู่อื่น ในขณะที่จำนวนคำของเนื้อหาในหมวดหมู่เทคโนโลยี ส่วนใหญ่จะมีค่าน้อยกว่าหมวดหมู่อื่น นอกจากนี้ยังพบว่า จำนวนคำของชุดข้อมูลในหมวดหมู่บันเทิงมีการกระจายตัวของกลุ่มข้อมูลสูงสุด และจำนวนคำของชุดข้อมูลในหมวดหมู่กีฬามีการกระจายตัวของกลุ่มข้อมูลต่ำสุด ซึ่งสังเกตได้จากค่า STD ในตารางที่ 4.8



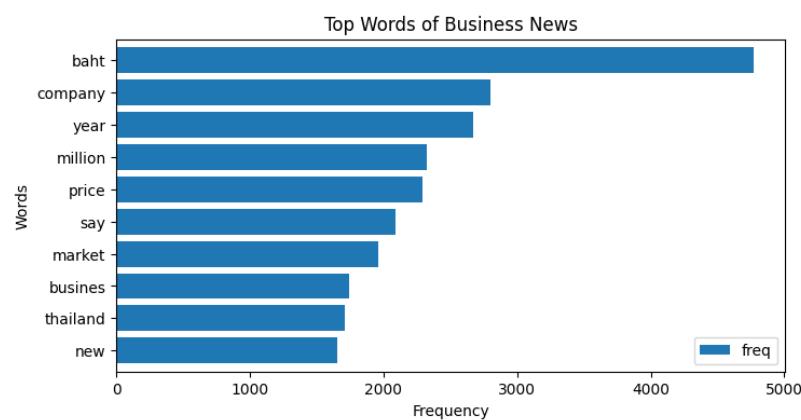
รูปที่ 4.18: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่อื่นๆ



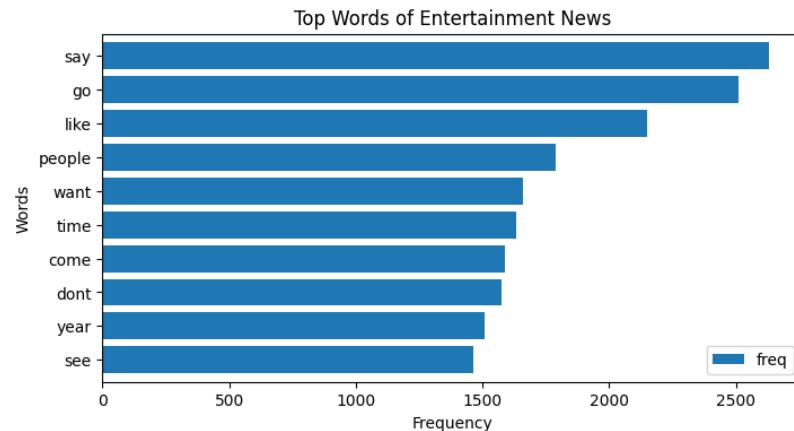
รูปที่ 4.19: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่กีฬา



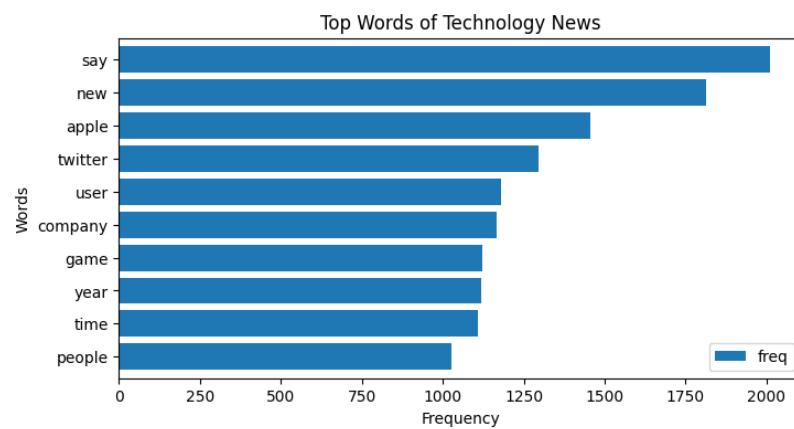
รูปที่ 4.20: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่การเมือง



รูปที่ 4.21: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่เศรษฐกิจ

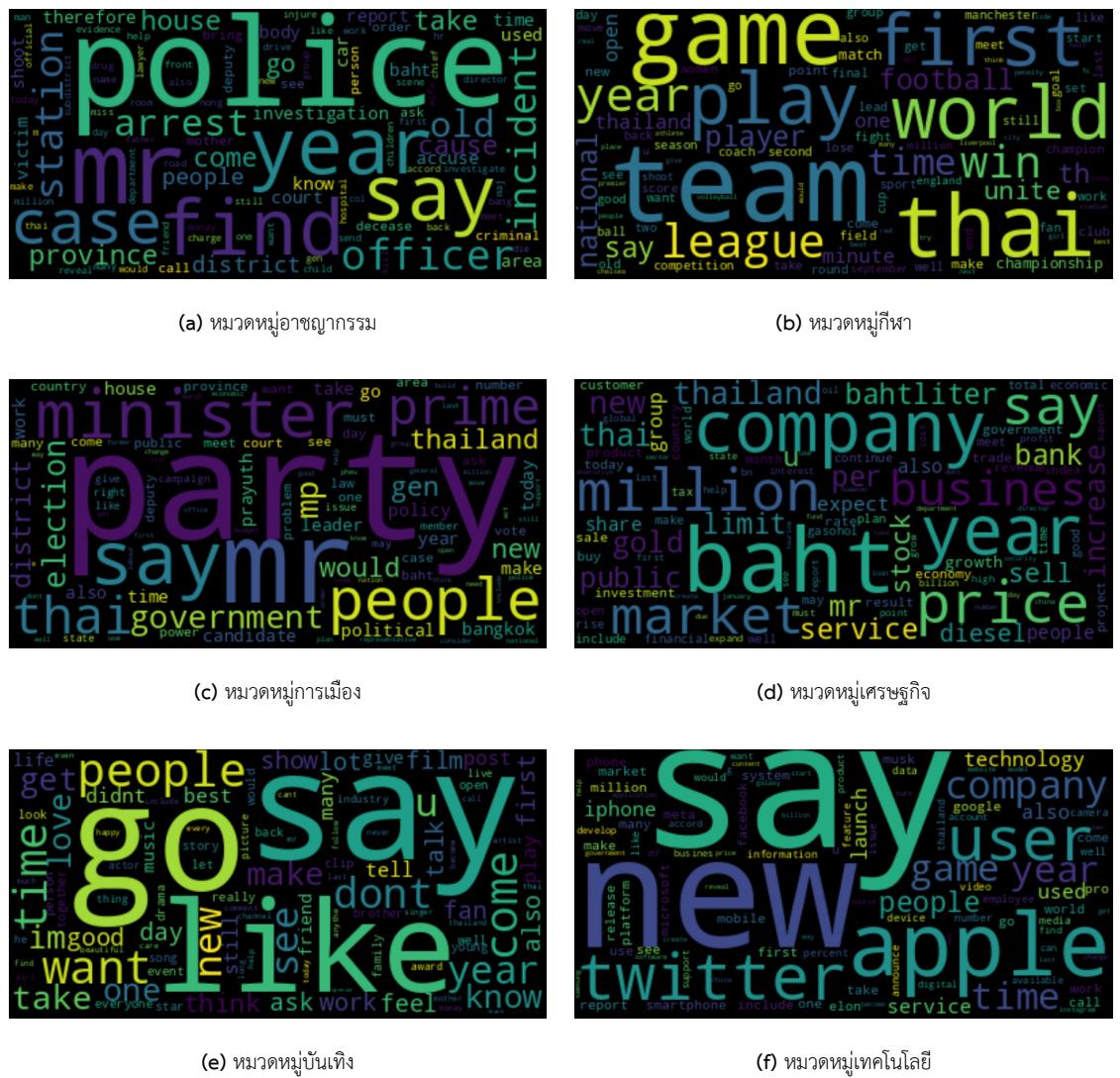


รูปที่ 4.22: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่บันเทิง



รูปที่ 4.23: 10 คำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของหมวดหมู่เทคโนโลยี

จากรูปภาพที่ 4.18 ถึงรูปภาพที่ 4.23 เป็นการแสดงผลคำที่พบมากที่สุด 10 อันดับแรกในชุดข้อมูลของแต่ละหมวดหมู่ เพื่อใช้สำหรับคุณวินัยของคำที่พบบ่อยในแต่ละหมวดหมู่ ซึ่งจากข้อมูลข้างต้นพบว่า แต่ละหมวดหมู่ค่อนข้างที่จะมีคำที่พบบ่อยแตกต่างกัน แต่จะมีบางคำที่พบบ่อยในทุกหมวดหมู่ เช่น say year ซึ่งถือว่าเป็นคำที่ว้าวไป ไม่ได้มีความสำคัญมากสำหรับข้อมูล



รูปที่ 4.24: 100 อันดับคำที่พบได้มากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่ โดยใช้ Word Cloud ในการแสดงผล

จากรูปภาพที่ 4.24 เป็นการแสดงผลคำที่พบมากที่สุด 100 อันดับแรกในชุดข้อมูลที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่ โดยใช้ Word Cloud ในการแสดงผล

4.2.3 การทำ Text Representation ด้วย TF-IDF

เนื่องจากคณะผู้จัดทำได้นำชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยและภาษาอังกฤษมาทำ Text Representation ด้วยการใช้ TF-IDF ซึ่งใช้โครงสร้างของโมเดลแตกต่างกัน ดังนี้

ตารางที่ 4.9: โครงสร้างของโมเดล TF-IDF ในรูปแบบภาษาไทยและภาษาอังกฤษ

โครงสร้าง	TF-IDF สำหรับชุดข้อมูลภาษาไทย	TF-IDF สำหรับชุดข้อมูลภาษาอังกฤษ
tokenizer	split('')	defalut
stop_words	thai_stopwords()	english
ngram_range	(1,1)	(1,2)
min_df	5	5
max_df	0.95	0.95
max_features	50% ของจำนวนคำที่ไม่ซ้ำกันทั้งหมด	50% ของจำนวนคำที่ไม่ซ้ำกันทั้งหมด
norm	l2	l2
encoding	utf-8	utf-8

0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.0	0.391722	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.064794
0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000

รูปที่ 4.25: ตัวอย่าง Text Representation ของชุดข้อมูลที่ได้จากการทำ TF-IDF

จะเห็นได้ว่าผลลัพธ์ของการทำ Text Representation ด้วย TF-IDF (รูปที่: 4.25) เป็น Vector ที่ถูกแสดงออกมาในรูปแบบของตัวเลข ซึ่งตัวเลขแต่ละค่าจะสื่อถึงความสำคัญหรือความเกี่ยวข้องของคำในชุดข้อมูลนั้น ๆ โดยชุดข้อมูลในแต่ละหมวดหมู่จะมีคำที่สำคัญที่แตกต่างกัน ดังนี้

ตารางที่ 4.10: แสดง 3 คำที่มีความสำคัญมากที่สุดในชุดข้อมูลที่เป็นภาษาไทยของแต่ละหมวดหมู่

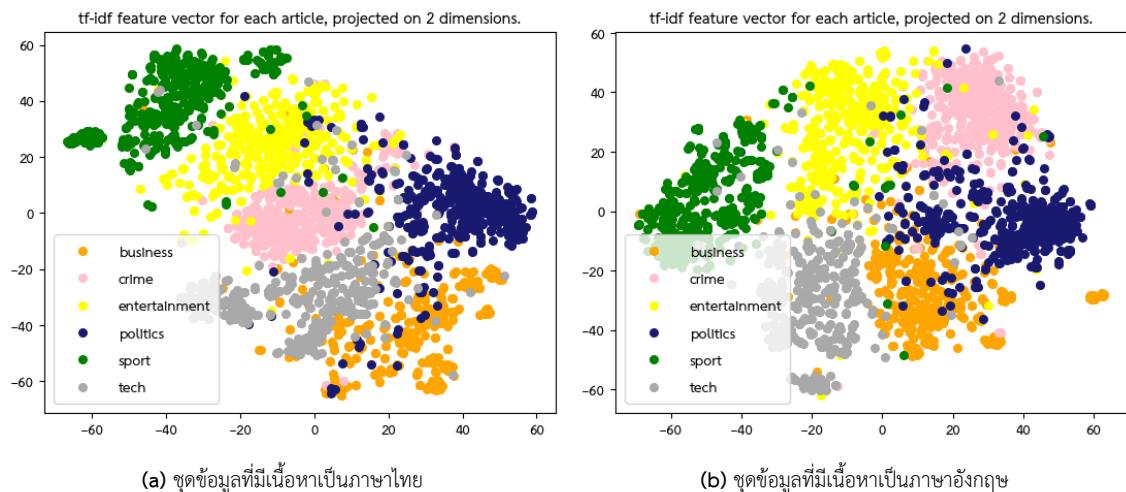
หมวดหมู่	Unigrams	หมวดหมู่	Unigrams
อาชญากรรม	- สภ. - ผู้ต้องหา - ตำรวจ	เศรษฐกิจ	- ลงทะเบียน - บาท - ราคา
กีฬา	- ลีก - เกม - ทีม	บันเทิง	- โพสต์ - ใหม่ - พี่
การเมือง	- เลือกตั้ง - ส.ส. - พรรคร	เทคโนโลยี	- หุ่นยนต์ - ดวงจันทร์ - อวกาศ

ตารางที่ 4.11: แสดง 3 คำที่มีความสำคัญมากที่สุดในชุดข้อมูลที่เป็นภาษาอังกฤษของแต่ละหมวดหมู่

หมวดหมู่	Unigrams	Bigram	หมวดหมู่	Unigrams	Bigram
อาชญากรรม	- station	- year old	เศรษฐกิจ	- index	- million baht
	- incident	- police officer		- stock	- company limit
	- police	- police station		- baht	- public company
กีฬา	- game	- premier league	บันเทิง	- film	- feel like
	- league	- manchester unite		- im	- post picture
	- team	- national team		- love	- entertainment industry
การเมือง	- prime	- gen prayuth	เทคโนโลยี	- iphone	- artificial intelligence
	- minister	- thai party		- twitter	- iphone pro
	- party	- prime minister		- apple	- elon musk

เนื่องจากในภาษาอังกฤษมักมีคำประสมอยู่ เช่น Police officer, Premier League ดังนั้นจึงมีการทำ Bigram เพื่อหาคำประสมในชุดข้อมูลที่เป็นภาษาอังกฤษ แต่สำหรับภาษาไทย คำประสมส่วนใหญ่มักเป็นคำที่ติดกันและไม่มีการเว้นวรรค เช่น สถานีตำรวจน้ำ แม่น้ำ ซึ่ง เมื่อนำมาแบ่งแยกเป็นคำๆ ก็อาจนับเป็นคำเดียว จึงทำเพียง Unigram ในชุดข้อมูลที่เป็นภาษาไทยเท่านั้น

การลั่นเกตเพียงคำสำคัญที่มากที่สุดนั้นยังไม่สามารถเห็นการแบ่งกลุ่มของคำแต่ละคำในเนื้อหาได้อย่างชัดเจน ดังนั้นจึงมีการทำ T-distributed Stochastic Neighbor Embedding (t-SNE) เพื่อลดมิติข้อมูลแบบไม่เชิงเส้น ช่วยให้เห็นรูปแบบและความสัมพันธ์ของคำในชุดข้อมูลมากยิ่งขึ้น



รูปที่ 4.26: ผลลัพธ์จากการทำ t-SNE ด้วยข้อมูลที่ได้จากการทำ TF-IDF

จากรูปภาพที่ 4.26 จะเห็นได้ว่าคำในชุดข้อมูลของแต่ละหมวดหมู่ค่อนข้างมีความสัมพันธ์กันอย่างชัดเจน การกระจายตัวของข้อมูลในแต่ละหมวดหมู่จะอยู่ในบริเวณที่ใกล้เคียงกัน แต่จะมีหมวดหมู่การเมืองที่ความสัมพันธ์ของข้อมูลค่อนข้างกระจายตัว และมีบางคำที่มีความสัมพันธ์ใกล้เคียงกับหมวดหมู่อาชญากรรมและเศรษฐกิจ นอกจากนี้หมวดหมู่เทคโนโลยีในชุดข้อมูลที่เป็นภาษาไทยมีความสัมพันธ์ของข้อมูลค่อนข้างกระจายตัวเช่นกัน แต่ก็ยังมีการกระจายตัวที่อยู่ในบริเวณใกล้เคียงกับคำในหมวดหมู่ตุนเอง

4.2.4 การเลือก Model ระหว่าง Random Forest KNN และ NN

คณะกรรมการผู้จัดทำได้ทำการศึกษาและพัฒนาโมเดลทั้งหมด 3 โมเดล ได้แก่ Random Forest, K-Nearest Neighbor และ LSTM โดยใช้การทำ Text Representation ด้วย TF-IDF เป็นข้อมูลนำเข้าสำหรับการพัฒนาโมเดล ซึ่งได้ประสิทธิภาพดังนี้

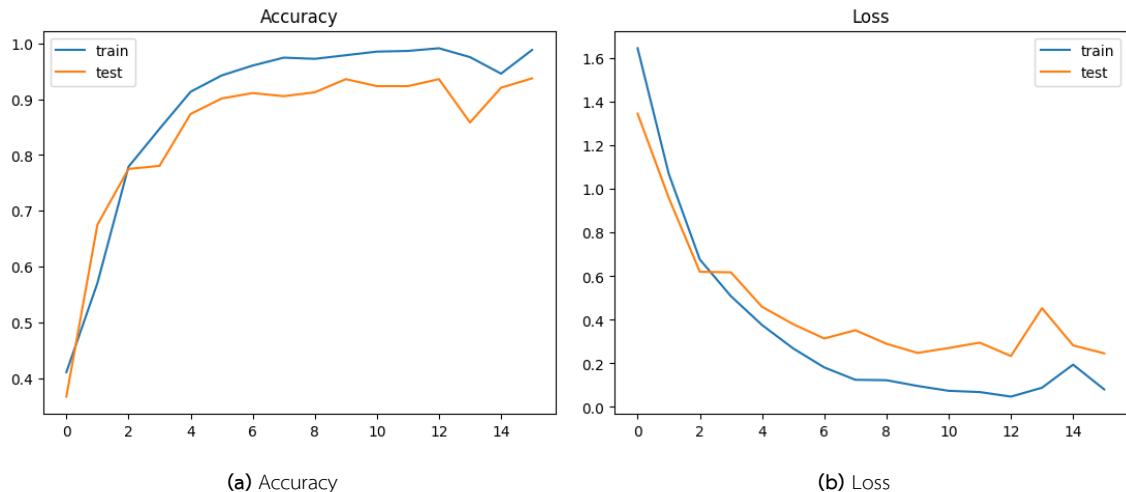
ตารางที่ 4.12: ค่า Accuracy ของแต่ละโมเดลเมื่อใช้ TF-IDF กับชุดข้อมูลที่เป็นภาษาไทย

ชื่อโมเดล	โครงสร้างของโมเดล	ค่า Accuracy
Random Forest	n_estimators: 180, max_depth: 110, min_samples_split: 9, min_samples_leaf: 1	0.9372
K-Nearest Neighbor	n_neighbors: 10, weights: 'distance', metric: 'euclidean', n_jobs = -1	0.9261
LSTM	units: 104, dropout: 0.102, activation: softmax, optimizers: Adam, loss: categorical_crossentropy, learning_rate=0.00018	0.1994

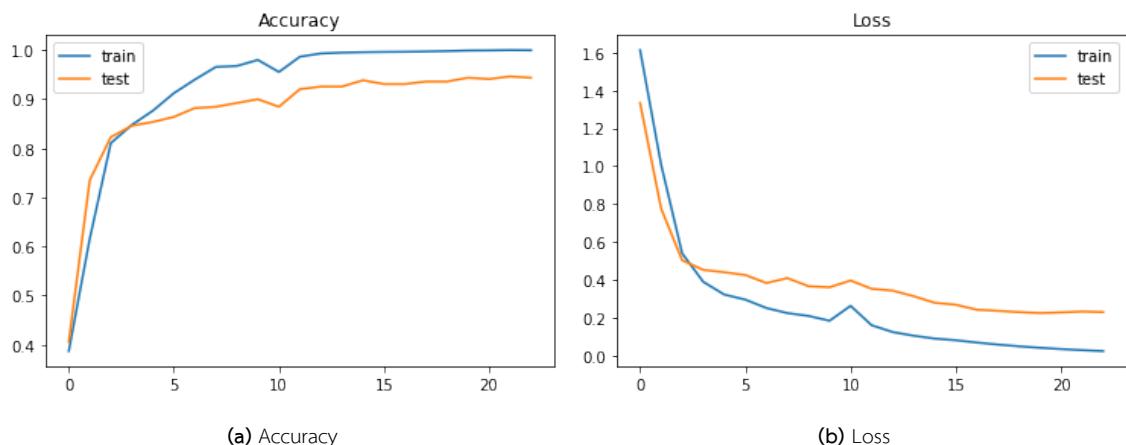
ตารางที่ 4.13: ค่า Accuracy ของแต่ละโมเดลเมื่อใช้ TF-IDF กับชุดข้อมูลที่เป็นภาษาอังกฤษ

ชื่อโมเดล	โครงสร้างของโมเดล	ค่า Accuracy
Random Forest	n_estimators: 450, max_depth: 114, min_samples_split: 2, min_samples_leaf: 1	0.9456
K-Nearest Neighbor	n_neighbors: 15, weights: 'distance', metric: 'euclidean', n_jobs = -1	0.9311
LSTM	units: 108, dropout: 0.174, activation: softmax, optimizers: Adam, loss: categorical_crossentropy, learning_rate=0.00040	0.2094

จากตารางที่ 4.12 และตารางที่ 4.13 พบว่า การทำ Text Representation ด้วย TF-IDF เมื่อนำมาใช้งานกับโมเดล Random Forest และ K-Nearest Neighbor ค่า Accuracy ที่ได้ค่อนข้างสูง ในขณะที่ ค่า Accuracy ที่ได้จากการทำ Text Representation ที่ไม่ใช้ TF-IDF อาจมีขนาดใหญ่เกินไปหรือไม่ เหมาะสมกับโมเดล LSTM ดังนั้นคณะกรรมการผู้จัดทำจึงเลือกที่จะเปลี่ยนวิธีการทำ Text Representation สำหรับโมเดล LSTM ด้วยการทำ Pad Sequence แทน ทำให้ได้ประสิทธิภาพของโมเดลดังนี้



รูปที่ 4.27: กราฟแสดงค่า Accuracy และ Loss ของโมเดล LSTM สำหรับภาษาไทย



รูปที่ 4.28: กราฟแสดงค่า Accuracy และ Loss ของโมเดล LSTM สำหรับภาษาอังกฤษ

ตารางที่ 4.14: แสดงค่า Evaluation Metrics ของโมเดล LSTM

ชื่อโมเดล	Accuracy	Precision	Recall	F1-score
LSTM สำหรับภาษาไทย	0.9233	0.9237	0.9227	0.9226
LSTM สำหรับภาษาอังกฤษ	0.9378	0.9372	0.9370	0.9371

จากตารางที่ 4.14 เมื่อเปลี่ยนวิธีการทำ Text Representation สำหรับโมเดล LSTM ด้วยการทำ Pad Sequence แทน พบร้า โมเดล LSTM มีประสิทธิภาพสูงขึ้น โดยมีค่า Accuracy สูงถึง 0.9233 สำหรับชุดข้อมูลที่เป็นภาษาไทยและมีค่า Accuracy สูงถึง 0.9378 สำหรับชุดข้อมูลที่เป็นภาษาอังกฤษ เมื่อนำโมเดล LSTM ที่ทำการพัฒนาแล้วมาเปรียบเทียบกับประสิทธิภาพของโมเดล Random Forest และ K-Nearest Neighbor ที่ทำไว้ จะได้ผลลัพธ์ดังนี้

ตารางที่ 4.15: แสดงค่า Evaluation Metrics ของโมเดลแต่ละโมเดลสำหรับชุดข้อมูลที่เป็นภาษาไทย

หมวดหมู่	Random Forest			K-Nearest Neighbor			LSTM		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
อาชญากรรม	0.94	0.95	0.95	0.91	0.94	0.92	0.93	0.95	0.94
กีฬา	0.95	0.97	0.96	0.96	0.98	0.97	0.96	0.97	0.97
การเมือง	0.94	0.88	0.91	0.95	0.84	0.89	0.83	0.90	0.86
เศรษฐกิจ	0.94	0.93	0.93	0.88	0.92	0.90	0.90	0.90	0.90

อังกฤษจะทำให้ได้โมเดลที่มีประสิทธิภาพมากกว่าการทำโมเดลด้วยชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย ทั้งนี้อาจเกิดจากการแปลภาษาไทยให้กลายเป็นภาษาอังกฤษส่งผลให้ความซับซ้อนของคำในประโยคลดลง หรืออาจเป็นเพราะการใช้ Library ตัดคำของภาษาไทยไม่ได้มีประสิทธิภาพในการตัดคำที่เป็นคำทับศัพท์ เนื่อง ชื่อของคนที่ทับศัพท์มาจากภาษาอังกฤษ จึงส่งผลต่อประสิทธิภาพของการทำโมเดลด้วยชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ ซึ่งโมเดลที่มีค่า Accuracy สูงสุด คือ โมเดล Random Forest ที่ใช้ชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ โดยมีค่า Accuracy สูงถึง 0.9456

4.2.5 การทดสอบโมเดล

นำโมเดลที่ได้ค่า Accuracy สูงที่สุดมาทดสอบด้วยชุดข้อมูลอื่น โดยจะใช้ชุดข้อมูลที่อยู่ภายใน Domain ที่ระบบรองรับอย่างละ 100 ชุดต่อหมวดหมู่ และใช้ชุดข้อมูลที่ไม่ได้อยู่ภายใน Domain ที่ระบบรองรับอย่างละ 10 ชุดต่อหมวดหมู่ ซึ่งได้ผลลัพธ์ดังนี้

ตารางที่ 4.17: ผลการทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายใน Domain ที่ระบบรองรับ

หมวดหมู่	ผลลัพธ์ที่ทำนายได้						F1-Score (%)
	อาชญากรรม	กีฬา	การเมือง	เศรษฐกิจ	บันเทิง	เทคโนโลยี	
อาชญากรรม	96	0	0	2	2	0	93.50
กีฬา	0	97	0	0	3	0	98.00
การเมือง	6	0	87	1	3	3	93.00
เศรษฐกิจ	1	0	1	93	0	5	94.00
บันเทิง	3	0	0	0	97	0	94.50
เทคโนโลยี	0	1	0	2	1	96	94.00
เฉลี่ย							94.50

จากตาราง 4.17 จะเห็นได้ว่าเมื่อทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายในโมเดล พบร้า หมวดหมู่กีฬาเป็นหมวดหมู่ที่ทำนายได้แม่นยำมากที่สุดและหมวดหมู่การเมืองเป็นหมวดหมู่ที่ทำนายได้แม่นยำน้อยที่สุด

ตารางที่ 4.18: ผลการทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายนอก Domain ที่ระบบรองรับ

หมวดหมู่	ผลลัพธ์ที่ทำนายได้						F1-Score (%)
	อาชญากรรม	กีฬา	การเมือง	เศรษฐกิจ	บันเทิง	เทคโนโลยี	
อาชญากรรม	9	0	1	0	0	0	90.00
กีฬา	0	10	0	0	0	0	100.00
การเมือง	1	0	9	0	0	0	90.00
เศรษฐกิจ	0	0	0	10	0	0	88.46
บันเทิง	0	0	0	0	10	0	100.00
เทคโนโลยี	0	0	0	3	0	7	85.00
เฉลี่ย							92.24

จากตาราง 4.18 จะเห็นได้ว่าเมื่อทดสอบโมเดลด้วยชุดข้อมูลที่อยู่ภายในโมเดล พบร้า หมวดหมู่กีฬาและบันเทิงเป็นหมวดหมู่ที่ทำนายได้แม่นยำมากที่สุดและหมวดหมู่การเทคโนโลยีเป็นหมวดหมู่ที่ทำนายได้แม่นยำน้อยที่สุด ทั้งนี้เป็นเพราะว่าบทความที่คนผู้จัดทำได้นำมาใช้ทดสอบในหมวดหมู่เทคโนโลยี เป็นบทความที่มีความคล้ายคลึงกับหมวดหมู่เศรษฐกิจ เนื่องจากมีการพูดถึงการลงทุนด้านเทคโนโลยีหรือพูดถึงจำนวนเงิน ซึ่งทำให้โมเดลเกิดการผิดพลาดในการทำนายว่าเป็นหมวดหมู่เศรษฐกิจ

4.3 Application

4.3.1 User Interface Design

ในส่วนของ User Interface ได้มีการออกแบบเพื่อรองรับการใช้งานของผู้ใช้งานที่ต้องการใช้บริการของ Thai Tag โดยผู้ใช้งานสามารถเยี่ยมชมเว็บแอปพลิเคชัน เข้าสู่ระบบ การส่งคำร้องวิเคราะห์เนื้อหาต่าง ๆ ดูรายละเอียดบทความ และติดตามคำร้องวิเคราะห์ของคนเองได้อีก ซึ่งจากการเว็บแอปพลิเคชันของคณะผู้จัดทำ จึงได้ User Interface ของเว็บแอปพลิเคชัน ดังนี้

4.3.1.1 หน้าแรก

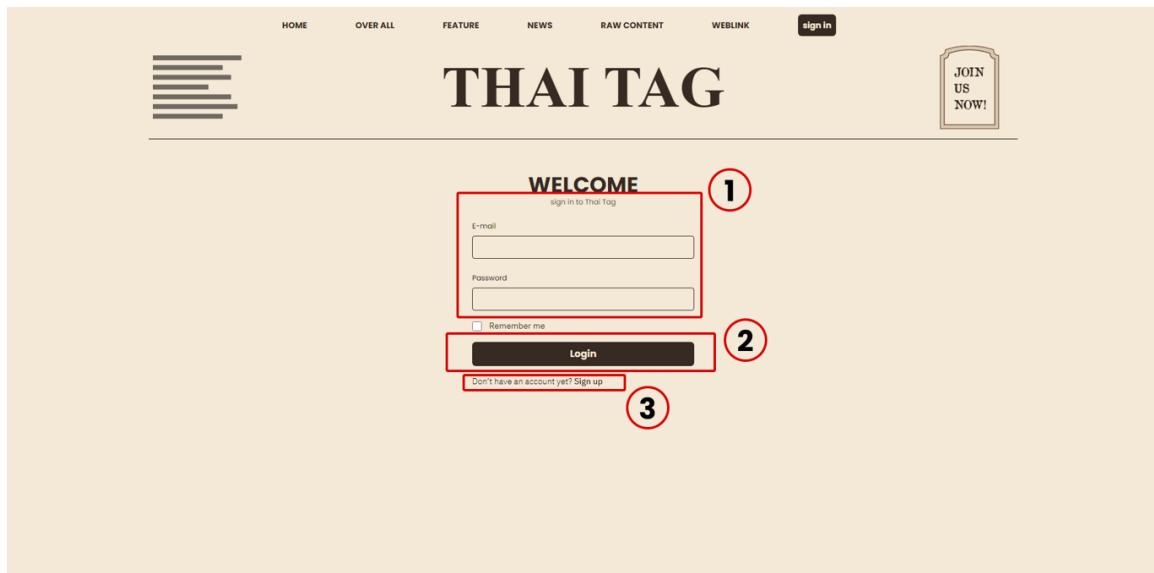


รูปที่ 4.29: หน้าแรก

เป็นหน้าแรกที่ผู้ใช้งานจะได้เห็นเมื่อมาเยี่ยมชมเว็บแอปพลิเคชัน ซึ่งจะให้ข้อมูลเกี่ยวกับเว็บแอปพลิเคชันว่ามีกระบวนการในการประมวลผลอย่างไร รวมไปถึงให้รายละเอียดเกี่ยวกับวิธีการใช้งานเว็บแอปพลิเคชันและรายละเอียดอื่น ๆ โดยหน้านี้จะมีอยู่ 3 ส่วนหลัก ๆ ด้วยกัน ได้แก่

1. แนะนำเกี่ยวกับเว็บแอปพลิเคชันและบริการต่าง ๆ ของเว็บแอปพลิเคชัน
2. ปุ่มเข้าสู่หน้าหลักเว็บไซต์และปุ่มเข้าสู่ระบบ ซึ่งหากผู้ใช้งานต้องการเยี่ยมชมเว็บแอปพลิเคชันเพียงอย่างเดียว ผู้ใช้งานสามารถกดที่ปุ่ม Explore Website เพื่อเข้าชมเว็บแอปพลิเคชันเบื้องต้น แต่ถ้าต้องการเข้าสู่ระบบให้กดที่ปุ่ม Sign In
3. ปุ่มเข้าสู่ระบบ หากผู้ใช้งานไม่เข้าสู่ระบบเพื่อยืนยันตัวตน ผู้ใช้งานจะไม่สามารถเข้าใช้บริการของเว็บแอปพลิเคชันได้ สามารถเยี่ยมชมได้เพียงแค่หน้าแรกและหน้าหลักเท่านั้น

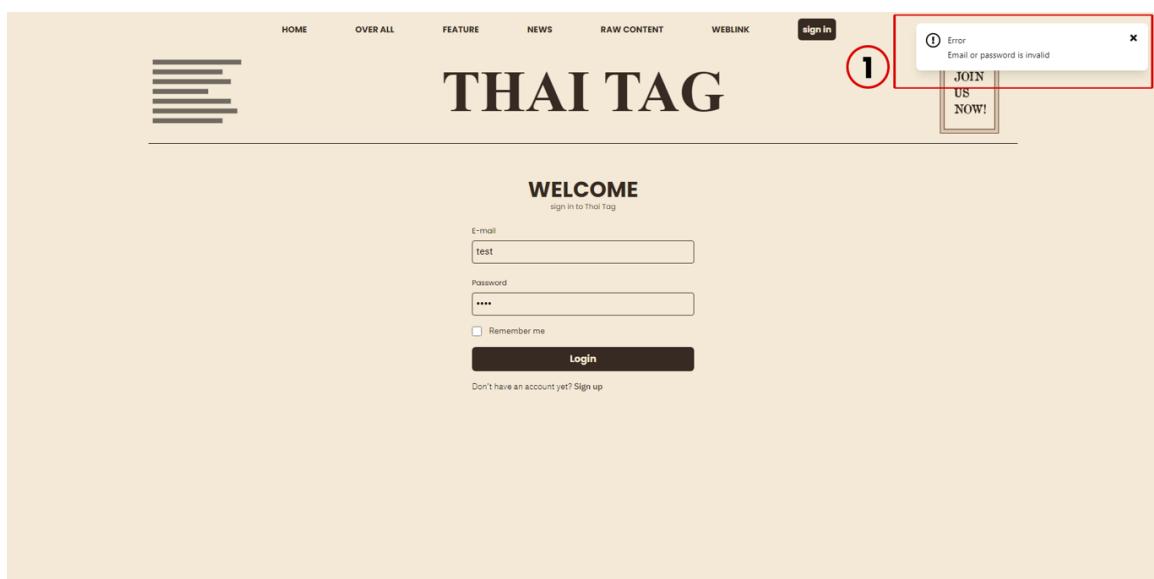
4.3.1.2 หน้าเข้าสู่ระบบ



รูปที่ 4.30: หน้าเข้าสู่ระบบ

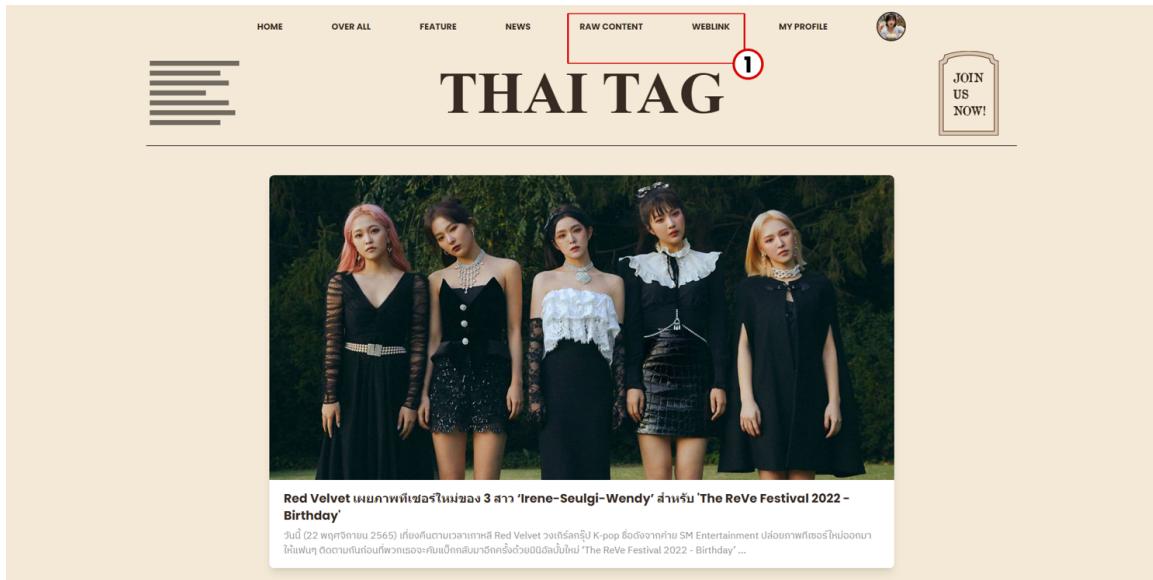
เป็นหน้าที่ผู้ใช้งานต้องกรอก Email และรหัสผ่านให้ถูกต้องเพื่อเข้าสู่ระบบ ถ้าผู้ใช้งานต้องการเข้าสู่ระบบแต่ไม่มีบัญชีผู้ใช้ ก็สามารถสมัครบัญชีได้ที่หน้านี้ โดยหน้านี้จะมีอยู่ 3 ส่วนหลัก ๆ ด้วยกัน ได้แก่

1. ช่องสำหรับใส่ Email และรหัสผ่าน
2. ปุ่มยืนยันเพื่อเข้าสู่ระบบ
3. สำหรับผู้ใช้งานที่ไม่มีบัญชีผู้ใช้ จะต้องทำการสมัครบัญชีผ่านช่องทางนี้



รูปที่ 4.31: หน้าเข้าสู่ระบบ กรณ์เข้าสู่ระบบผิดพลาด

กรณีที่ผู้ใช้งานเข้าสู่ระบบผิดพลาด เริบแอปพลิเคชันจะแจ้งเตือนให้ผู้ใช้งานกรอก Email และรหัสผ่านให้ถูกต้องตามวิธีหมายเลข 1 ในภาพที่ 4.30



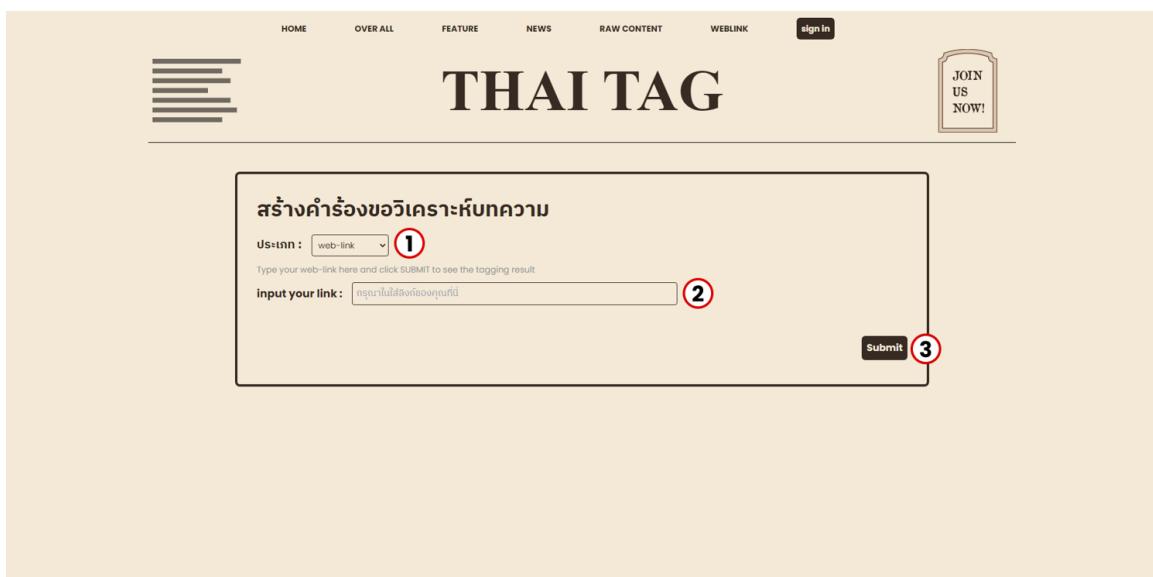
รูปที่ 4.32: หน้าหลัก

4.3.1.3 หน้าหลัก

เป็นหน้าที่ผู้ใช้งานทั่วไปและสมาชิกสามารถเข้าถึงได้ สามารถอ่านบทความต่าง ๆ ได้ ซึ่งผู้ใช้งานสามารถวิเคราะห์หมวดหมู่ของเนื้อหาได้ผ่านทางหน้านี้ โดยการเลือกบันทึกหมายเลข 1 ในภาพที่ 4.32 ดังนี้

1. ถ้าหากต้องการวิเคราะห์หมวดหมู่ของเนื้อหาโดยใช้บทความ ให้เลือกที่ Raw Content
2. ถ้าหากต้องการวิเคราะห์หมวดหมู่ของเนื้อหาโดยใช้ Weblink ให้เลือกที่ Weblink

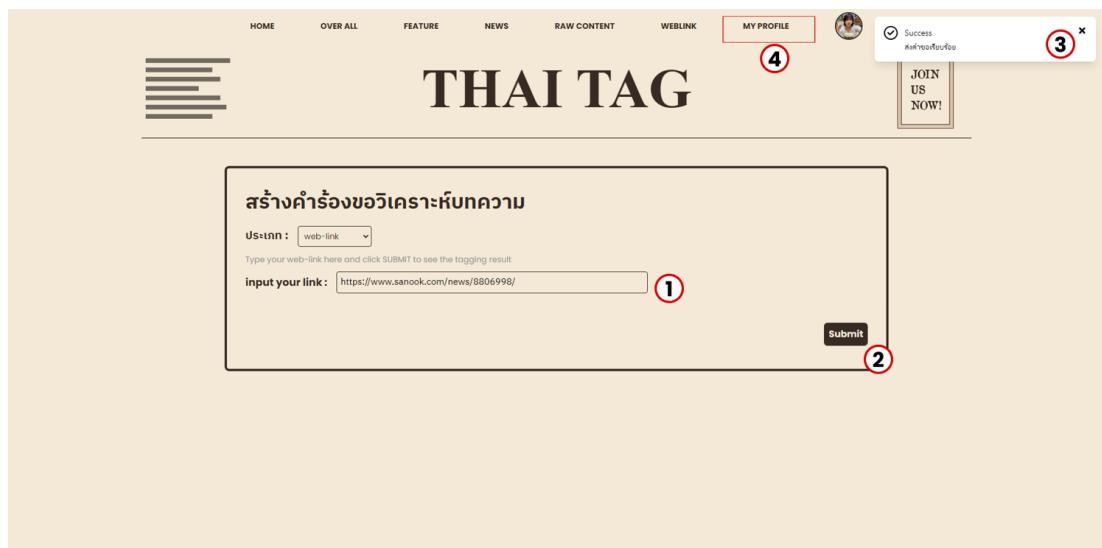
4.3.1.4 หน้าสร้างคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link



รูปที่ 4.33: หน้าสร้างคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link

เป็นหน้าสำหรับการส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วย Web Link ซึ่งจะมีอยู่ 3 ส่วน "ได้แก่"

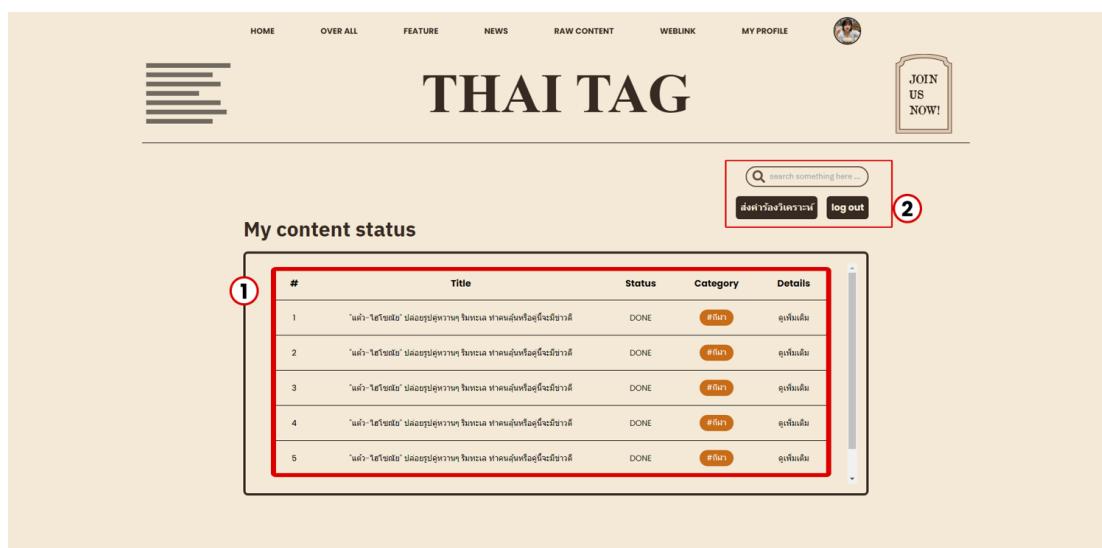
1. เป็นส่วนที่ให้สมาชิกเลือกว่าจะส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วยอะไร โดยจะแบ่งเป็น 2 ประเภท คือ การส่งคำร้องขอด้วย Web Link และการส่งคำร้องขอด้วยบทความ โดยสมาชิกสามารถเลือกได้จาก Drop Down
2. ส่วนสำหรับใส่ Web Link ที่ต้องการวิเคราะห์หมวดหมู่ของบทความ
3. ปุ่มยืนยันเพื่อส่งคำร้องขอเพื่อวิเคราะห์หมวดหมู่ของบทความด้วย Web Link



รูปที่ 4.34: กรณีส่งคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link สำเร็จ

เมื่อผู้ใช้งานส่งคำร้องขอวิเคราะห์หมวดหมู่ของบทความด้วย Web Link สำเร็จ เว็บแอปพลิเคชันจะแสดงกล่องข้อความว่า "ส่งคำร้องขอเรียบร้อย" ตามหมายเลข 3 ในภาพที่ 4.34 และสามารถติดตามผลการร้องขอได้โดยการเลือกที่ "My Profile" ตามหมายเลข 4 ในภาพที่ 4.34

4.3.1.5 หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ



รูปที่ 4.35: หน้าแสดงผลการวิเคราะห์หมวดหมู่ของบทความ

หลังจากที่สมาชิกส่งคำร้องขอวิเคราะห์หมวดหมู่ของบทความ ระบบจะทำการประมวลผลและแสดงรายละเอียดต่าง ๆ ประกอบไปด้วยลำดับบทความ รายละเอียดบทความคร่าว สถานะการประมวลผล หมวดหมู่ที่วิเคราะห์ได้ และปุ่มดูรายละเอียดเพิ่มเติม ซึ่งหน้านี้จะมีอยู่ 2 ส่วน ได้แก่

1. ตารางแสดงรายละเอียดต่าง ๆ เกี่ยวกับการวิเคราะห์หมวดหมู่ของบทความเบื้องต้น
2. ช่องสำหรับค้นหาบทความ ปุ่มสำหรับส่งคำร้องอีคัช และปุ่มออกจากระบบ

ถ้าหากต้องการส่งคำร้องขอวิเคราะห์หมวดหมู่ของบทความอีคัช หรือออกจากระบบ สมาชิกสามารถดำเนินการได้ตามวงกลมหมายเลข 2 ในภาพที่ **4.35**

4.3.2 การทดสอบความสามารถของระบบ

ในการพัฒนาเว็บแอปพลิเคชันต้องทำการทดสอบการใช้งานของเว็บแอปพลิเคชันว่าสามารถทำงานได้ที่คาดหวังหรือไม่ โดยวิธีการทดสอบสามารถแบ่งได้เป็น 2 วิธี ได้แก่

1. การทดสอบการทำงานของเว็บแอปพลิเคชัน โดยใช้วิธีการ Manual Testing

ตารางที่ 4.19: แสดงผลการทดสอบเว็บแอปพลิเคชันผู้ใช้งาน

รายการทดสอบที่	รายละเอียด	ผลที่คาดหวัง	ผ่านการทดสอบ
1	สมัครสมาชิกข้อมูลครบถ้วนสมบูรณ์, Email ไม่เข้าในระบบ	สมัครสมาชิกสำเร็จ	✓
2	สมัครสมาชิกข้อมูลครบถ้วนสมบูรณ์, Email ข้าในระบบ	สมัครสมาชิกไม่สำเร็จ	✓
3	สมัครสมาชิกข้อมูลไม่ครบถ้วนสมบูรณ์	สมัครสมาชิกไม่สำเร็จ	✓
4	เข้าสู่ระบบ Email และ Password ถูกต้อง	เข้าสู่ระบบสำเร็จ	✓
5	เข้าสู่ระบบ Email ถูกต้อง แต่ Password ผิด	เข้าสู่ระบบไม่สำเร็จ	✓
6	เข้าสู่ระบบ และไม่มี Email อยู่ในระบบ	เข้าสู่ระบบไม่สำเร็จ	✓
7	ส่ง Weblink ที่ระบบรองรับ	ประมวลผล วิเคราะห์จัดหมวดหมู่สำเร็จ	✓
8	ส่ง Weblink ที่ระบบไม่รองรับ	ระบบปฏิเสธรับคำร้องขอ	✓
9	ส่ง Weblink ที่ระบบรองรับ แต่หน้าเว็บไม่มีอยู่จริง	ระบบประมวลผลผิดพลาด	✓
10	ส่งเนื้อหา เพื่อวิเคราะห์หมวดหมู่	ประมวลผล วิเคราะห์จัดหมวดหมู่สำเร็จ	✓
11	ดูคำร้องขอ	สามารถดูคำร้องขอของตนเองได้	✓
12	ออกจากระบบ	สามารถออกจากระบบได้	✓

ตารางที่ 4.20: แสดงผลการทดสอบเว็บแอปพลิเคชันผู้ดูแลระบบ

รายการทดสอบที่	รายละเอียด	ผลที่คาดหวัง	ผ่านการทดสอบ
1	จัดการและลบบัญชีผู้ใช้งานได้	สามารถลบบัญชีใช้งานได้	✓
2	จัดการคำร้องขอเพื่อจัดหมวดหมู่ของเนื้อหา	สามารถเพิ่ม/ลบ/แก้ไข คำร้องขอได้	✓
3	จัดการเนื้อหาของบทความออนไลน์บนเว็บแอปพลิเคชัน	สามารถเพิ่ม/ลบ/แก้ไข เนื้อหาได้	✓
4	จัดการหมวดหมู่ที่จัดเตรียมเป็นแบบแอปพลิเคชัน	สามารถเพิ่ม/ลบ/แก้ไข หมวดหมู่ได้	✓
5	แก้ไขและพัฒนาเทคโนโลยีปัญญาประดิษฐ์	สามารถแก้ไขเทคโนโลยีปัญญาประดิษฐ์ได้	✓

2. การทดสอบการทำงานของเว็บแอปพลิเคชัน โดยใช้วิธีการ Automate Testing ในที่นี้จะใช้วิธีการ Unit Testing โดยใช้ JUnit ในการทดสอบการทำงานของ Code ที่เขียนขึ้นมาซึ่งหลังจากที่ทำการทดสอบแล้ว จะสามารถสร้างรายงานการทดสอบได้ เป็นการทดสอบ

ที่ครอบคลุมทั้ง 19 กรณีตามหลักในการใช้งานเว็บแอปพลิเคชันทั่วไป ซึ่งเป็นส่วนที่ช่วยให้การพัฒนาซอฟต์แวร์ให้มีคุณภาพมากยิ่งขึ้นและสามารถพัฒนาไฟเจอร์ใหม่ ๆ ได้อย่างมีความมั่นใจ เนื่องจากสามารถตรวจสอบได้จากการนิทดสอบเดิมที่ทำการเขียนไว้ อีกทั้งยังเป็นส่วนช่วยให้นักพัฒนาฝึกความรับผิดชอบมากขึ้นในการพัฒนาเว็บแอปพลิเคชัน ซึ่งจากการทดสอบพบว่า ไม่มีกรณีไหนที่ล้มเหลว

Test Summary



Packages Classes

Class	Tests	Failures	Ignored	Duration	Success rate
me.nonlawat.domain.content.internal.ContentServiceImplTest\$TestGetContentById	2	0	0	0.606s	100%
me.nonlawat.domain.ticket.internal.TicketServiceImplTest\$TestGetTicketById	2	0	0	0.109s	100%
me.nonlawat.domain.ticket.internal.TicketServiceImplTest\$TestSubmitWeblink	1	0	0	0.028s	100%
me.nonlawat.domain.user.internal.UserServiceImplTest	1	0	0	0.191s	100%
me.nonlawat.domain.user.internal.UserServiceImplTest\$TestLogin	3	0	0	0.018s	100%
me.nonlawat.domain.user.internal.UserServiceImplTest\$TestProfile	2	0	0	0.003s	100%
me.nonlawat.domain.user.internal.UserServiceImplTest\$TestRegister	2	0	0	0.008s	100%
me.nonlawat.infrastructure.auth.AdminRequiredFilterTest	3	0	0	0.023s	100%
me.nonlawat.infrastructure.auth.ApiKeyRequiredFilterTest	3	0	0	0.006s	100%

รูปที่ 4.36: รายการ Unit tests ภายใน General Backend

4.3.3 การทดสอบประสิทธิภาพของระบบ

ในการทดสอบประสิทธิภาพของระบบจะทำการทดสอบด้วยวิธีการ Load Testing โดยจะเป็นการทดสอบประสิทธิภาพการทำงานส่วนที่เกี่ยวข้องกับการวิเคราะห์หมวดหมู่ของบทความ โดยเป็นการจำลองการใช้งานของผู้ใช้งานพร้อมกันจำนวน 10 คน ในระยะเวลา 30 วินาที โดยจะใช้โปรแกรม K6 ที่ทำการเขียน script ให้ทำการส่งคำขอในการประมวลผลไปยังระบบ AI ซึ่งการประเมินประสิทธิภาพจะวัดเฉพาะขนาดการเตรียมข้อมูล และการทำนาย โดยประสิทธิภาพจะประเมินจาก จำนวนคำร้องขอที่สามารถรับได้ต่อวินาที (Request per second) และได้ผลลัพธ์ดังนี้

ตารางที่ 4.21: ผลการทดสอบความเร็วในการจัดหมวดหมู่ที่เก็บเนื้อหาของเว็บแอปพลิเคชัน

จำนวนคำของเนื้อหารีบต้น	จำนวนคำของเนื้อหาหลังจากการทำ Preprocessing	จำนวนคำร้อง/วินาที
100	71	20.87
200	119	20.23
500	305	20.06
1000	502	17.24

จากผลลัพธ์จะเห็นได้ว่า ในข้อตอนการเตรียมข้อมูลจะทำให้จำนวนคำในบทความถูกตัดทอนออกไป ก่อนจะเข้าสู่กระบวนการทำนาย แสดงให้เห็นว่าบทความต้องตัด แม้ในตอนแรกจะมีจำนวนมาก แต่หลังจากการตัดทอนก็สามารถมีจำนวนคำที่ใช้สำหรับการทำนายใกล้เคียงกันได้ ซึ่งจะส่งผลการเวลาที่ใช้ในการทำนาย ซึ่งจะแปรผันตรงกับจำนวนคำหลังจากการจัดเตรียมข้อมูล

4.4 การประเมินผลและความเห็นจากผู้ใช้งาน

คณะผู้จัดทำจะให้ผู้ที่ทำแบบสอบถามใช้งานเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่เนื้อหาออนไลน์ จำนวนจึงทำแบบสอบถามซึ่งในแบบสอบถามของคณะผู้จัดทำจะถูกแบ่งออกเป็น 3 ส่วน ดังนี้

- ข้อมูลทั่วไป เป็นการสอบถามข้อมูลเบื้องต้น รวมถึงประสบการณ์การจัดหรือจำแนกหมวดหมู่เนื้อหาของผู้ที่ทำแบบสอบถาม
- แบบประเมินแบบให้คะแนน เป็นแบบประเมินความพึงพอใจในการใช้งานเว็บแอปพลิเคชัน ทั้งส่วนการออกแบบ ประสิทธิภาพ และการจัดหมวดหมู่ในเนื้อหาของบทความ ซึ่งจะแบ่งเป็น 2 ส่วนย่อยด้วยกัน ได้แก่ ส่วนของเว็บแอปพลิเคชันและส่วนของการจัดหมวดหมู่ในเนื้อหาของบทความแบบอัตโนมัติ

3. แบบสอบถามแบบถาม-ตอบ เป็นการสอบถามความรู้สึกและความคิดเห็นที่มีต่อเว็บแอปพลิเคชัน รวมถึงข้อเสนอแนะและสิ่งที่ต้องปรับปรุงด้วย

ตัวอย่างแบบสอบถามความพึงพอใจหลังจากการใช้งานเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทย โดยใช้เทคโนโลยีปัญญาประดิษฐ์

คำอธิบาย แบบสอบถามนี้มีทั้งหมด 3 ส่วน ขอให้ผู้ตอบแบบสอบถามตอบให้ครบทั้ง 3 ส่วน

ตอนที่ 1 ข้อมูลทั่วไป

1. เพศ ชาย หญิง

2. อายุ

<input type="checkbox"/> ต่ำกว่า 18 ปี	<input type="checkbox"/> 18 - 20 ปี	<input type="checkbox"/> 21 - 25 ปี
<input type="checkbox"/> 26 - 30 ปี	<input type="checkbox"/> 31 - 40 ปี	<input type="checkbox"/> มากกว่า 40 ปี

3. อาชีพ

<input type="checkbox"/> ข้าราชการ/พนักงานรัฐวิสาหกิจ/เจ้าหน้าที่รัฐ	<input type="checkbox"/> พนักงาน/ลูกจ้างเอกชน
<input type="checkbox"/> ค้ายา/ประกอบธุรกิจส่วนตัว	<input type="checkbox"/> เกษตรกร
<input type="checkbox"/> รับจ้างทั่วไป	<input type="checkbox"/> นักเรียน/นักศึกษา
<input type="checkbox"/> ไม่ได้ประกอบอาชีพ	<input type="checkbox"/> อื่นๆ (โปรดระบุ) _____

4. เคยค้นหาเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยหมวดหมู่

เคย ไม่เคย

5. เคยจัดหรือจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยตนเอง

เคย ไม่เคย

ตอนที่ 2 แบบประเมินแบบให้คะแนน

หมวดหมู่	ระดับความพึงพอใจ/ความรู้ความเข้าใจ				
	มากที่สุด	มาก	ปานกลาง	น้อย	น้อยที่สุด
ด้านการออกแบบ					
1. ความสูงของเว็บแอปพลิเคชัน					
2. สีสันในการออกแบบมีความเหมาะสม					
3. การจัดวางองค์ประกอบของเว็บแอปพลิเคชัน					
4. การใช้งานเว็บแอปพลิเคชันง่ายและสะดวก					
5. ความน่าสนใจในการนำเสนอเว็บแอปพลิเคชัน					
6. ความสะดวกในการเชื่อมโยงเนื้อหาภายในเว็บแอปพลิเคชัน					
ด้านประสิทธิภาพ					
1. ความรวดเร็วในการแสดงผลของเว็บแอปพลิเคชัน					
2. ความรวดเร็วของระบบจัดหมวดหมู่ให้กับเนื้อหาออนไลน์					
3. ความแม่นยำของระบบจัดหมวดหมู่ของบทความตัวย่อให้กับเนื้อหาในบทความ					
4. ความแม่นยำของระบบจัดหมวดหมู่ของบทความด้วย Web Link					
5. ความรวดเร็วโดยรวมของเว็บแอปพลิเคชัน					

ตอนที่ 3 แบบสอบถามแบบตาม - ตอบ

1. หลังการทดลองใช้งาน ท่านคิดว่าเว็บแอปพลิเคชันนี้จะทำให้ท่านจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ได้สะดวกกว่าจัดหมวดหมู่ด้วยตนเองหรือไม่ อ่านว่า
-
-
-

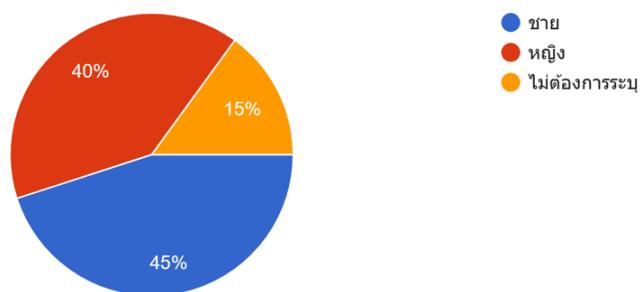
2. ท่านคิดว่าเว็บแอปพลิเคชันนี้สามารถช่วยจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ได้โดยไม่ต้องอาศัยความรู้หรือไม่ อย่างไร

3. ท่านคิดเห็นอย่างไรกับการที่เว็บแอปพลิเคชันสามารถจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ได้โดยการใช้เนื้อหานะบความหรือ Web Link

4. ข้อควรปรับปรุง/ข้อเสนอแนะเพิ่มเติม

4.4.1 ผลสำรวจจากการตอบแบบสอบถามของผู้ใช้งาน

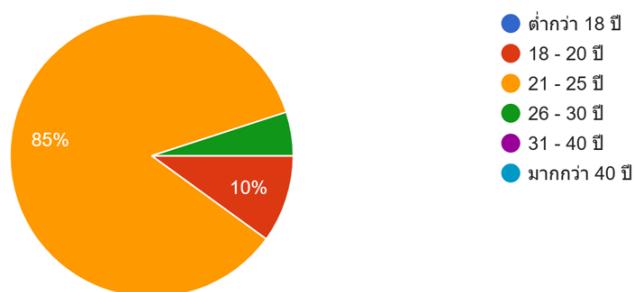
เพศ
20 responses



รูปที่ 4.37: สัดส่วนเพศของผู้ตอบแบบสอบถาม

จากการเก็บผลสำรวจ คณะผู้จัดทำได้เก็บข้อมูลจากเพศชาย 45% เพศหญิง 40% และไม่ต้องการระบุเพศ 15%

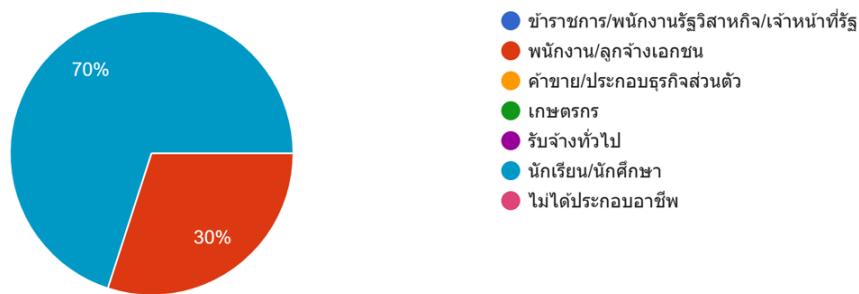
อายุ
20 responses



รูปที่ 4.38: สัดส่วนอายุของผู้ตอบแบบสอบถาม

จากการเก็บผลสำรวจ ส่วนใหญ่จะเป็นกลุ่ม Early Adopter ซึ่งมีอายุ 21 - 25 ปี เนื่องจากการทำแบบทดสอบนี้เป็นการทำแบบทดสอบผ่านทางออนไลน์ จึงเข้าถึงกลุ่มอายุ 21 - 25 ปีมากกว่ากลุ่มอายุอื่น

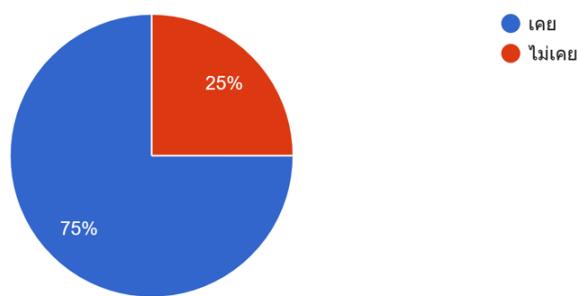
อาชีพ
20 responses



รูปที่ 4.39: สัดส่วนอาชีพของผู้ตอบแบบสอบถาม

จากการเก็บผลสำรวจ ส่วนใหญ่จะเป็นกลุ่มนักเรียน/นักศึกษา โดยคิดเป็น 70% และเป็นกลุ่มพนักงาน/ลูกจ้างเอกชน 30 %

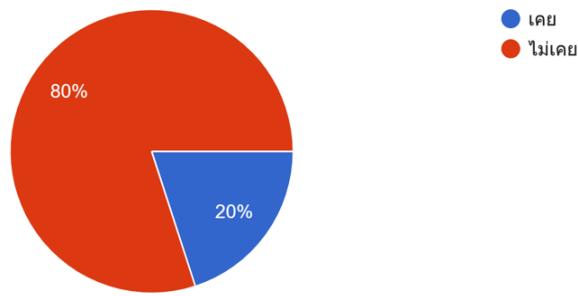
เคยค้นหาเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยหมวดหมู่
20 responses



รูปที่ 4.40: สัดส่วนผู้ตอบแบบสอบถามที่เคยค้นหาเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยหมวดหมู่

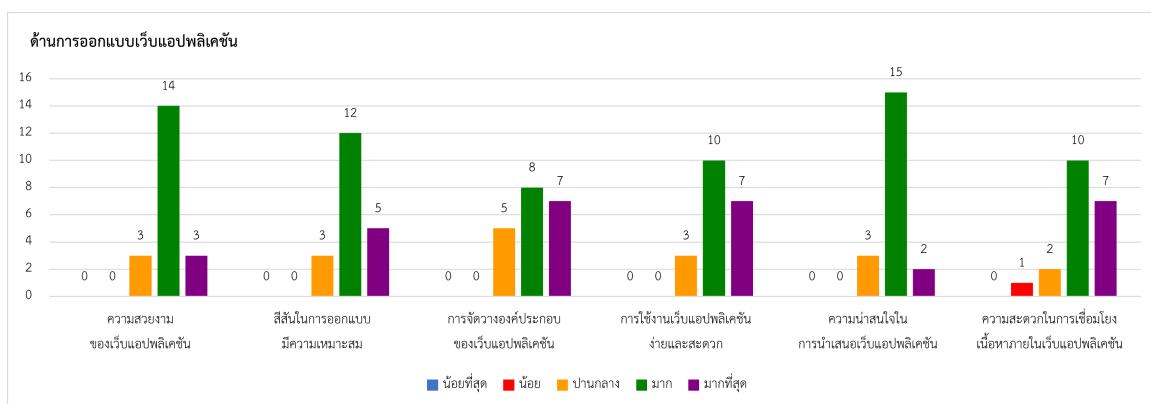
จากการเก็บผลสำรวจ พบร่วมบุคคลที่เคยค้นหาเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยหมวดหมู่ถึง 75% ซึ่งเป็นผลดีต่อการทดสอบเนื่องจากทำให้ได้รับข้อมูลเพิ่มเติมจากผู้ที่เคยมีประสบการณ์จริง

เคยจัดหรือจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยตนเอง
20 responses



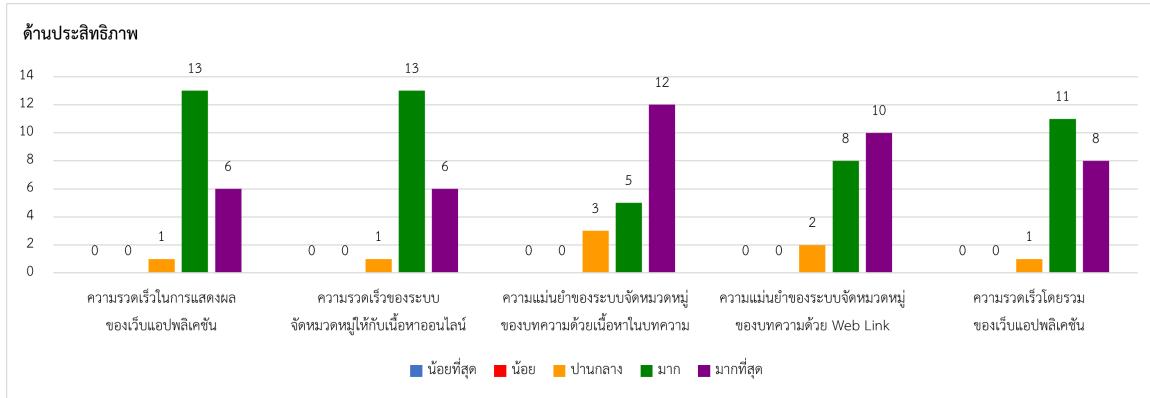
รูปที่ 4.41: สัดส่วนผู้ตอบแบบสอบถามที่เคยจัดหรือจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยตนเอง 20%

จากการเก็บผลสำรวจ พบร่วมบุคคลที่เคยจัดหรือจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทยด้วยตนเอง 20%



รูปที่ 4.42: การให้คะแนนความพึงพอใจต่อเว็บไซต์ในด้านการออกแบบของผู้ตอบแบบสอบถาม

สำหรับผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ทำแบบสอบถามในด้านการออกแบบ ผู้ทำแบบสอบถามส่วนใหญ่ให้คะแนน "มาก" ในเรื่องของความสวยงาม สีสัน และความน่าสนใจ ส่วนในเรื่องของการจัดวางองค์ประกอบ การใช้งานง่าย และความสะดวกในการเชื่อมโยงเนื้อหาภายใน ผู้ทำแบบสอบถามส่วนใหญ่ให้คะแนนอยู่ในช่วง "มาก" ถึง "มากที่สุด" แสดงให้เห็นว่าโดยภาพรวมผู้ทำแบบสอบถาม มีความพึงพอใจในการออกแบบเว็บแอปพลิเคชันของคณะผู้จัดทำ



รูปที่ 4.43: การให้คะแนนความพึงพอใจต่อเว็บไซต์ในด้านประสิทธิภาพของผู้ที่ทำแบบสอบถาม

สำหรับผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ที่ทำแบบสอบถามในด้านประสิทธิภาพ ผู้ที่ทำแบบสอบถามส่วนใหญ่ให้คะแนน "มาก" ถึง "มากที่สุด" ไม่ว่าจะเป็นในเรื่องของความรวดเร็วหรือความแม่นยำของเว็บแอปพลิเคชัน แสดงให้เห็นว่าโดยภาพรวมผู้ที่ทำแบบสอบถาม มีความพึงพอใจในประสิทธิภาพของเว็บแอปพลิเคชัน

4.4.2 สรุปผลการสำรวจจากแบบสอบถาม

ตารางที่ 4.23: ผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ที่ทำแบบสอบถามในด้านการออกแบบ

หัวข้อ	\bar{X}	S.D.
ความสวยงามของเว็บแอปพลิเคชัน	4	0.56
สีสันในการออกแบบมีความเหมาะสม	4.1	0.64
การจัดวางองค์ประกอบของเว็บแอปพลิเคชัน	4.1	0.79
การใช้งานเว็บแอปพลิเคชันง่ายและสะดวก	4.2	0.70
ความน่าสนใจในการนำเสนอเนื้อหาภายในเว็บแอปพลิเคชัน	3.95	0.51
ความสะดวกในการเชื่อมโยงเนื้อหาภายในเว็บแอปพลิเคชัน	4.15	0.81
เฉลี่ย	4.08	0.67

ตารางที่ 4.24: ผลลัพธ์ของการให้คะแนนความพึงพอใจของผู้ที่ทำแบบสอบถามในด้านการออกแบบ

หัวข้อ	\bar{X}	S.D.
ความรวดเร็วในการแสดงผลของเว็บแอปพลิเคชัน	4.25	0.55
ความรวดเร็วของระบบจัดหมวดหมู่ให้กับเนื้อหาออนไลน์	4.25	0.55
ความแม่นยำของระบบจัดหมวดหมู่ของบทความตามหัวข้อที่ระบุไว้	4.45	0.76
ความแม่นยำของระบบจัดหมวดหมู่ของบทความด้วย Web Link	4.44	0.68
ความรวดเร็วโดยรวมของเว็บแอปพลิเคชัน	4.35	0.59
เฉลี่ย	4.35	0.63

จากการสำรวจผู้ที่ทำแบบทดสอบทั้งหมด 20 คน สามารถแบ่งได้เป็น 2 กลุ่ม ได้แก่ กลุ่มที่เคยจัดการเกี่ยวกับหมวดหมู่ของเนื้อหาออนไลน์และกลุ่มที่ไม่เคยจัดการเกี่ยวกับหมวดหมู่ของเนื้อหาออนไลน์ ซึ่งคิดเป็น 20% และ 80% ตามลำดับ พบว่าในภาพรวมผู้ที่ทำแบบทดสอบมีความพึงพอใจในการใช้งานเว็บแอปพลิเคชัน โดยพิจารณาจากค่าเฉลี่ยการให้คะแนนในด้านต่าง ๆ ของผู้ที่ทำแบบทดสอบ ซึ่งมีค่าเฉลี่ยคะแนนด้านการออกแบบอยู่ที่ 4.08 และมีค่าเฉลี่ยคะแนนด้านประสิทธิภาพอยู่ที่ 4.35 ซึ่งเป็นค่าในช่วงเกณฑ์ "มาก" ถึง "มากที่สุด" ประกอบกับมีค่าส่วนเบี่ยงเบนมาตรฐานอยู่ในช่วง 0.60 ถึง 0.70

บทที่ 5 บทสรุปและอภิปราย

5.1 สรุปผลโครงการ

ในการพัฒนาเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ คณะผู้จัดทำได้นำชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทย มาแปลงเป็นภาษาอังกฤษ เพื่อใช้ในการเปรียบเทียบว่าการพัฒนาด้วยชุดข้อมูลภาษาใดทำให้ประสิทธิภาพของโมเดลตีที่สุด ซึ่งชุดข้อมูลทั้ง 2 ภาษาได้ถูกนำมาใช้กับโมเดลทั้ง 3 รูปแบบด้วยกัน ได้แก่ Random Forest, K-Nearest Neighbor และ LSTM โดยใช้ TF-IDF เป็น Text Representation แต่การทำ TF-IDF กับโมเดล LSTM ให้ผลลัพธ์ที่ไม่มีประสิทธิภาพมากพอ โดยมีค่า Accuracy อยู่ที่ 0.1994 สำหรับชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยและมีค่า Accuracy อยู่ที่ 0.2094 สำหรับชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ ดังนั้นคณะผู้จัดทำจึงได้เปลี่ยนวิธีการทำ Text Representation ของโมเดล LSTM ส่งผลให้โมเดล LSTM มีประสิทธิภาพมากขึ้น แต่สุดท้ายโมเดลที่มีประสิทธิภาพมากที่สุดคือโมเดล Random Forest ที่ค่า Accuracy อยู่ที่ 0.9372 สำหรับชุดข้อมูลที่มีเนื้อหาเป็นภาษาไทยและมีค่า Accuracy อยู่ที่ 0.9456 สำหรับชุดข้อมูลที่มีเนื้อหาเป็นภาษาอังกฤษ เพราะฉะนั้นคณะผู้จัดทำจึงเลือกวิธีการแปลงเนื้อหาตามความยาก ภาษาไทยให้เป็นภาษาอังกฤษและใช้โมเดล Random Forest มาพัฒนาเว็บแอปพลิเคชันสำหรับการจัดหมวดหมู่ให้กับเนื้อหาออนไลน์ที่เป็นภาษาไทย ซึ่งทำให้เริ่มแอปพลิเคชันนี้เป็นแพลตฟอร์มสำหรับการจัดหมวดหมู่ เนื้อหาตามหมวดหมู่ที่ต้องการได้ ซึ่งเว็บแอปพลิเคชันสามารถรองรับโดยเมินได้ไม่น้อยกว่า 5 โดเมนและรองรับหมวดหมู่ได้ถึง 6 หมวดหมู่ ได้แก่ การเมือง กีฬา เทคโนโลยี บันเทิง เศรษฐกิจ และอาชญากรรม นอกจากนี้เว็บแอปพลิเคชันยังมีระบบที่รองรับผู้จัดการระบบให้สามารถเพิ่มและจัดการหมวดหมู่ได้ด้วยตนเองผ่านทางระบบหน้าบ้าน ซึ่งจากการทดสอบความแม่นยำในการทำงานผ่านเว็บแอปพลิเคชัน พบว่า เว็บแอปพลิเคชันสามารถจำแนกหมวดหมู่ได้ทั้งชุดข้อมูลที่อยู่ภายใต้ Domain ที่ระบบรองรับและชุดข้อมูลที่อยู่ภายใต้ Domain ที่ระบบรองรับ แต่เว็บแอปพลิเคชันจะจำแนกหมวดหมู่ที่อยู่ภายใต้ Domain ที่ระบบรองรับได้แม่นยำกว่า ทั้งนี้อาจเป็นเพราะเนื้อหาออนไลน์ที่นำมาทดสอบอาจจะมีลักษณะคล้ายกัน เช่น ข้อมูลที่ถูกใช้ในการพัฒนาเทคโนโลยีปัญญาประดิษฐ์

5.2 ปัญหาที่พบและการแก้ไข

1. **ปัญหา:** ในกรณีที่ต้องการนำข้อมูลจากเว็บไซต์มาใช้ในระบบ พบว่า ไม่สามารถนำข้อมูลมาใช้ได้ในบางเว็บไซต์ เนื่องจากในปัจจุบันการสร้างหน้าเว็บไซต์สามารถแบ่งออกเป็น 2 รูปแบบหลัก ๆ ด้วยกัน ได้แก่ SSR (Server Side Rendering) และ CSR (Client Side Rendering) ซึ่งในส่วนของ SSR ไม่พับปัญหาในการ Scrape ข้อมูล เนื่องจากในส่วนของ HTML มีการสร้างมาโดยสมบูรณ์ผ่านในเว็บไซต์ตั้งแต่ต้น แต่ในส่วนของ CSR จะมี Code HTML บางส่วนเท่านั้น แล้วจะทำการดึงข้อมูลเพิ่มเติมในภายหลัง ทำให้ไม่สามารถใช้ HTML ที่ได้มาใน Request แยกโดยตรงได้

แก้ไข: ใช้ Selenium ในการจำลองการทำงานของเว็บไซต์ขึ้นมา โดยให้ประมาณผลดึงข้อมูลเพิ่มเติมให้เสร็จสิ้นก่อน จากนั้นจึงทำการดึงข้อมูล HTML ที่สร้างเสร็จแล้วมาใช้งาน

2. **ปัญหา:** การออกแบบการอัปเดต AI ในตอนแรกคณะผู้จัดทำวางแผนว่าจะใช้การอัปเดตผ่าน Github โดยการอัปโหลดโมเดลใหม่เข้าไป ทำให้เกิดการเปลี่ยนแปลงบน Repository และเกิดการกระตุน Github Action ให้ดำเนินการสร้าง Container Image ใหม่จากนั้นจึงนำไป Deploy บน Cloud Run แต่เนื่องจากโมเดลมีขนาดใหญ่จนไม่สามารถอัปโหลดขึ้นใน Github ได้ ทำให้ไม่สามารถใช้วิธีที่ออกแบบไว้แต่แรกได้

แก้ไข: เปลี่ยนกระบวนการอัปโหลดโมเดลไปยังการจัดเก็บอื่น ๆ แล้วสั่งกระตุน Github Action ผ่าน Rest API แทน และเปลี่ยน Workflow ให้ทำการดาวน์โหลดโมเดลเข้ามาก่อนในขณะที่ทำการสร้าง Container Image

3. **ปัญหา:** ในกระบวนการสร้างโมเดล คณะผู้จัดทำได้วางแผนที่จะนำชุดข้อมูลมาทำ Text Representation ด้วย TF-IDF ก่อนที่จะนำไปใช้ในการพัฒนาโมเดลทั้ง 3 โมเดล ได้แก่ Random Forest, K-Nearest Neighbor และ LSTM แต่เมื่อนำวิธี TF-IDF มาใช้กับโมเดล LSTM ปรากฏว่าไม่สามารถใช้ได้ เนื่องจากขนาด Vector ที่ใหญ่เกินไป สร่งผลให้ไม่สามารถพัฒนาโมเดลได้อย่างเต็มที่ เมื่อทำการลดขนาด Vector ลงเพื่อให้สามารถใช้งานกับโมเดลได้ ผลที่ได้คือประสิทธิภาพของโมเดลที่ได้ไม่เป็นไปตามที่คิดไว้ โดยมีค่า Accuracy ประมาณ 0.20 ซึ่งถือเป็นค่าที่ยอมรับไม่ได้ ดังนั้นจึงไม่สามารถใช้การทำ Text Representation ด้วย TF-IDF กับโมเดล LSTM ได้

แก้ไข: เปลี่ยนวิธีการทำ Text Representation ด้วยการใช้ Tokenizer ใน Keras Library ร่วมกับการทำ pad_sequences เพื่อแปลงคำให้เป็น Vector และนำมาใช้กับโมเดล LSTM ซึ่งวิธีการนี้ทำให้โมเดล LSTM มีประสิทธิภาพมากขึ้น โดยมีค่า Accuracy สูงประมาณ 0.92

5.3 ข้อจำกัดและข้อเสนอแนะ

5.3.1 ข้อจำกัด

- เนื่องจาก Library ที่ใช้สำหรับจัดการชุดข้อมูลที่เป็นภาษาไทยนั้มีอย่างจำกัด ทำให้การจัดการข้อมูลที่เป็นภาษาไทยต้องมีประสิทธิภาพทำได้ยากขึ้น อย่างเช่นการใช้ Library PyThai ในการช่วยตัดคำก็ยังมีข้อจำกัดอยู่ในกรณีที่เจอคำที่ไม่มีในชุด Library เช่น คำทับศัพท์จากภาษาต่างประเทศอย่างเช่นคน จึงเกิดการตัดคำผิดพลาดขึ้น โดยในกรณีผิดพลาดที่ค่อนข้างผู้จัดทำพบคือ ตัดแยกคำที่มีการันต์ออกโดยแยกระหว่างคำนับกับตัวการันต์ ทำให้หลังจากการตัดคำได้การันต์จำนวนมากและคำที่ตัดได้ก็มีปริบบทต่างไปจากเดิม และเมื่อนำมาใช้ในปัจจุบันนี้เป็นผลให้ความแม่นยำของโมเดลลดลง ซึ่งแตกต่างจากภาษาอังกฤษที่สามารถตัดคำได้โดยใช้วิเคราะห์เป็นตัวแบ่ง
- คำในชุดข้อมูลบางคำมีความกำหนดหรือคำบางคำในแต่ละหมวดหมู่ที่มีลักษณะของคำที่คล้ายกัน กรณีที่ค่อนข้างทำพบคือ คำในบางชุดข้อมูลที่อยู่ในหมวดหมู่การเมืองมีความคล้ายคลึงกับคำในบางชุดข้อมูลที่อยู่ในหมวดหมู่เศรษฐกิจและอาชญากรรม ทำให้ในบางครั้งโมเดลไม่สามารถจำแนกได้ถูกต้องกับคำที่คล้ายกันนี้ครั้งได้ในหมวดหมู่ใด ส่งผลให้ประสิทธิภาพของโมเดลลดลง
- เนื่องจากคอมพิวเตอร์ที่ใช้ในการพัฒนาโมเดลไม่ได้มีประสิทธิภาพเพียงพอต่อการพัฒนาโมเดล จึงต้องใช้ Google Colaboratory เป็นเครื่องมือในการพัฒนาโมเดล ซึ่ง Google Colaboratory ก็มีการจำกัดการใช้ RAM และ GPU ทำให้ไม่สามารถพัฒนาโมเดลได้อย่างเต็มที่เท่าที่ควร กระบวนการพัฒนาโมเดลจึงทำได้ช้าลง
- โมเดลที่พัฒนาขึ้นสามารถจัดหมวดหมู่ได้เพียง 1 หมวดหมู่ต่อข้อความ ยังไม่สามารถทำนายข่าวที่มีหลายหมวดหมู่ในข่าวเดียวอย่างแม่นยำได้

5.3.2 ข้อเสนอแนะและแนวทางในการพัฒนา

- อาจใช้ Transformer ในการช่วยแปลงคำให้เป็น Vector เนื่องจากมี Transformer หลายรูปแบบที่สามารถแปลงคำให้คล้ายเป็น Vector ได้อย่างมีประสิทธิภาพ และบาง Transformer ก็รองรับข้อมูลที่เป็นภาษาไทย เช่น WangchanBERT
- อาจมีการสร้างโมเดลเพื่อจำแนกว่าคำแต่ละคำควรอยู่หมวดหมู่ไหน ก่อนที่จะมีการพัฒนาโมเดลเพื่อจำแนกหมวดหมู่ของเนื้อหาออนไลน์
- พัฒนาโมเดลให้เป็น Multi-labels เพื่อที่จะสามารถจำแนกหมวดหมู่ให้กับเนื้อหาออนไลน์ที่มีหลายหมวดหมู่ในบทความเดียวได้
- พัฒนาโมเดลให้สามารถจำแนกหมวดหมู่ได้อย่างหลากหลายมากขึ้นกว่าปัจจุบันและพัฒนาให้เว็บแอปพลิเคชันสามารถรองรับโดยมีมากกว่าปัจจุบัน
- พัฒนาโมเดลให้สามารถจำแนกหมวดหมู่ได้อย่างรวดเร็วและมีความแม่นยำมากกว่าปัจจุบัน

หนังสืออ้างอิง

1. Orn Smith, 2017, “แนะนำวิธีใช้ Category และ Tag คืออะไร อันไหนใช้เมื่อไรย่างไร,” จากเว็บไซต์ <https://contentshifu.com/blog/category-vs-tag-best-practices>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
2. MK, 2016, “Pantip.com จับมือ ม.เกษตร เปิดตัวระบบ Auto Tag ด้วย Machine Learning,” จากเว็บไซต์ <https://www.blognone.com/node/79057>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
3. Michael Andrews, 2022, “AI-based auto-tagging of content: what you need to know,” จากเว็บไซต์ <https://kontent.ai/blog/ai-based-auto-tagging-of-content-what-you-need-to-know/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
4. SAS, 2022, “การประมวลผลภาษาธรรมชาติ (Natural language processing),” จากเว็บไซต์ https://www.sas.com/th_th/insights/analytics/what-is-natural-language-processing-nlp.html, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
5. Patipan Prasertsom, 2021, “การค้นหาตัวแทนเชิงความหมายของข้อความ: Word2Vec Word Embedding,” จากเว็บไซต์ <https://bigdata.go.th/big-data-101/word2vec/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
6. Patipan Prasertsom, 2020, “สกัดใจความสำคัญของข้อความด้วยเทคนิคการประมวลผลทางภาษาเบื้องต้น: TF-IDF,” จากเว็บไซต์ <https://bigdata.go.th/big-data-101/tf-idf-1/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
7. CHAKRIT, 2019, “TF-IDF ทำงานอย่างไร?,” จากเว็บไซต์ <https://www.softnix.co.th/2019/05/28/tf-idf-ทำงานยังไง/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
8. ไกรศักดิ์ เกษร, 2021, **วิทยาศาสตร์ข้อมูล(Data Science)**, ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์มหาวิทยาลัยนเรศวร, p.352.
9. AWS, 2022, “RESTful API คืออะไร,” จากเว็บไซต์ <https://aws.amazon.com/th/what-is/restful-api/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
10. somkiat, 2020, “Vue.js 101 ฉบับมาเร็ว ไปเร็วๆ 4 รูปแบบของ Event-Driven Architecture ที่ใช้บ่อย ๆ,” จากเว็บไซต์ <https://www.somkiat.cc/summary-event-driven-architecture/>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
11. Nutti Saelor, 2017, “Dependency Injection คืออะไร,” จากเว็บไซต์ <https://medium.com/@leelorz6/dependency-injection-คืออะไร-6a1a8a2996be>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
12. Peem Srinikorn, 2022, “ทำความรู้จักกับ CI/CD services บน Google Cloud,” จากเว็บไซต์ <https://cloud-ace.co.th/blogs/w0w7e1-ci-cd-services-google-cloud>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
13. JobsDB, 2022, “Agile คืออะไร แนวคิดการทำงานในองค์กรรุ่นใหม่ ต้องรู้หรือไม่?,” จากเว็บไซต์ <https://th.jobsdb.com/th-th/articles/ความเข้าใจผิดการทำงานแบบagile/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
14. S. Joni, Y. Vignesh, C. Juan, J. Bernard, and J. Soon-Gyo, 2019, “Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type,” **Journal of Business Research**, vol. 101, pp. 203–217, Aug 2019.
15. AWS, 2022, “Amazon Comprehend,” จากเว็บไซต์ <https://aws.amazon.com/th/comprehend/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
16. Marcuscode, 2020, “แนะนำภาษา JavaScript,” จากเว็บไซต์ <http://marcuscode.com/lang/javascript/introducing-to-javascript>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
17. Java, 2022, “Java คืออะไร จาวา คือภาษาคอมพิวเตอร์ สำหรับเขียนโปรแกรมเชิงวัตถุ,” จากเว็บไซต์ <https://www.mindphp.com/คุณมี/73-คืออะไร/2185-java-คืออะไร.html>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
18. Skooldio, 2021, “TECHNOLOGY Golang คืออะไร? ดียังไง? ทำไมคนใช้เยอะ? รวมลิสต์ที่ควรรู้เกี่ยวกับ Golang,” จากเว็บไซต์ <https://blog.skooldio.com/what-is-golang/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
19. AWS, 2022, “Python คืออะไร,” จากเว็บไซต์ <https://aws.amazon.com/th/what-is/python/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
20. Hardwarelibre, 2022, “Google Collab หรือ Google Colaboratory: มันคืออะไร,” จากเว็บไซต์ <https://www.hwlibre.com/th/google-colaboratory/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
21. Patchara Boonmathanaruk, 2022, “Figma คืออะไร? ทำไม่ถึงเป็น Tool มาแรงที่สุดในการ Design!,” จากเว็บไซต์ <https:////blog.skooldio.com/figma-ui-design-tool/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].

22. Linux-Console.net, 2019, ``PostgreSQL คืออะไร? PostgreSQL ทำงานอย่างไร," จากเว็บไซต์ <https://th.linux-console.net/?p=1801>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
23. Codebee, 2020, ``Git คืออะไร การใช้งาน Git จะบันทึกอะไรบ้าง," จากเว็บไซต์ <https://www.codebee.co.th/labs/git-คืออะไร-การใช้งาน-git/>, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
24. Fern, 2021, ``GitHub คืออะไร," จากเว็บไซต์ <https://wrksoftware.co/th/github-คืออะไร>, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
25. AWS, 2021, ``GitHub คืออะไร," จากเว็บไซต์ <https://aws.amazon.com/th/docker/>, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
26. พรหมพัฒน์ ไชยเดช, 2019, ``Vue.JS สร้าง Web Application สุด Wow สู่เส้นทาง Front-End Developer," จากเว็บไซต์ <https://page.futureskill.co/vuejs>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
27. Ltd Developer Team BorntoDev Co., 2020, ``Vue.js 101 ฉบับมาเร็ว ไปเร็วๆ," จากเว็บไซต์ <https://www.borntodev.com/2020/07/14/vue-js-101/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
28. นานพ กองอุ่น, 2019, ``เริ่มต้นทำ Web Application ด้วย NuxtJS," จากเว็บไซต์ <https://www.programmerthailand.com/tutorial/post/view?id=314>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
29. Mindphp.com, 2022, ``Pandas คืออะไร?," จากเว็บไซต์ <https://www.mindphp.com/บทเรียนออนไลน์/83-python/8493-what-is-the-pandas.html>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
30. Ltd Developer Team BorntoDev Co., 2022, ``NumPy คืออะไร?," จากเว็บไซต์ <https://www.borntodev.com/2020/04/16/ทั้งหมดการใช้-numpy-ใน-python-3/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
31. Top'p Kullawattana, 2019, ``Introduction NLP with spaCy in Software Engineering," จากเว็บไซต์ <https://suttipong-kull.medium.com/introduction-nlp-in-software-engineering-88ff3cd5ca86t>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
32. Punna Rirkvaleekul, 2020, ``สิ่งจำเป็นที่สาย Machine Learning ต้องรู้," จากเว็บไซต์ <https://twinsynergy.co.th/สิ่งจำเป็นที่สาย-machine-learning-ต้อง/>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
33. Sonny, 2020, ``Web Scraping (Python) ด้วย beautifulsoup และ requests," จากเว็บไซต์ <https://stackpython.co/tutorial/web-scraping-python-beautifulsoup-requests>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
34. Thichaluk.s, 2015, ``สร้างโปรแกรมทดสอบเว็บแอพพลิเคชันอัตโนมัติด้วย Selenium WebDriver," จากเว็บไซต์ <https://sysadmin.psu.ac.th/2015/08/31/selenium-webdriver-part1>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
35. BERYL8, 2019, ``เหตุผลที่หลาย ๆ ธุรกิจเลือกใช้ GCP (Google Cloud Platform) เพื่อเพิ่มประสิทธิภาพในการทำงาน," จากเว็บไซต์ <https://www.beryl8.com/th/newsroom/insights/102/เหตุผลที่หลาย ๆ ธุรกิจเลือกใช้-gcp-google-cloud-platform-เพื่อเพิ่มประสิทธิภาพในการทำงาน>, [ออนไลน์; วันที่สืบค้น 11 พฤษภาคม 2022].
36. Niw Popkorn, 2020, ``ออกแบบระบบให้คุยกับ module กันได้ ด้วย Pub/Sub," จากเว็บไซต์ <https://blog.niwpopkorn.com/2020/04/module-pubsub.html>, [ออนไลน์; วันที่สืบค้น 12 พฤษภาคม 2022].
37. Sirawit, 2018, ``ส่ง Docker image ขึ้น Google Container Registry (GCR)," จากเว็บไซต์ <https://medium.com/@sirawit/ส่ง-docker-image-ขึ้น-google-container-registry-gcr-92a4686a341/>, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
38. NUUNEOI, 2019, ``Cloud Run คืออะไร," จากเว็บไซต์ https://nuuneozi.com/blog/blog.php?read_id=973/, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
39. Mdsiglobal, 2019, ``Google Translate คืออะไร," จากเว็บไซต์ <https://www.mdsiglobal.com/google-translate/>, [ออนไลน์; วันที่สืบค้น 13 พฤษภาคม 2022].
40. iFew, 2020, ``แนะนำ k6 สำหรับทำ Load Testing และ Automation Testing," จากเว็บไซต์ <https://siamchamnankit.co.th/แนะนำ-k6-สำหรับทำ-load-testing-และ-automation-testing-43774a041c6f>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
41. Panacholn Pongsamritphol, 2018, ``วัดความเร็ว และปรับปรุงประสิทธิภาพของ Progressive Web App ด้วย Lighthouse," จากเว็บไซต์ <https://engineering.thinknet.co.th/วัดความเร็ว-และปรับปรุงประสิทธิภาพของ-progressive-web-app-ด้วย-lighthouse-8a9aeb39b7b3>, [ออนไลน์; วันที่สืบค้น 1 มิถุนายน 2023].
42. Cartoon Tanaporn, 2020, ``ไม่ใช่ไม่ได้แล้ว! ทำไม ClickUp ถึงหมายเป็นตัวช่วยในการ Work From Home ในองค์กรของคุณ," จากเว็บไซต์ <https://thegrowthmaster.com/blog/clickup-work-from-home>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].

43. 1TOALL COMPANY LIMITED, 2021, ``MIRO : THE ONLINE COLLABORATIVE WHITEBOARD PLATFORM," จากเว็บไซต์ <https://www.1-to-all.com/blog/miro-the-online-collaborative-whiteboard-platform>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].
44. I Plan Digital Agency, 2022, ``Discord คืออะไร? ต่างจากแพลตฟอร์มแชットอื่น ๆ อย่างไร?," จากเว็บไซต์ <https://www.iplandigital.co.th/collaborative-apps/what-is-discord/>, [ออนไลน์; วันที่สืบค้น 6 กันยายน 2022].

ภาคผนวก A

ชื่อภาคผนวกที่ 1

ใส่หัวข้อตามความเหมาะสม

This is where you put hardware circuit diagrams, detailed experimental data in tables or source codes, etc..