Pattern Analysis and Clustering of Generative AI Prompts

Manuscript Submission Data: 2023, October 10
Article Editing Date: 2023, October 21
Article Accepted Date: 2023, November 2

Tip-aroon Kiawkaew*
Paralee Maneerat*
Suparoek Chootongchai**

ABSTRACT

This research focuses on the analysis and clustering of prompts for a Generative AI system. In this study, prompts are generated from the ChatGPT. Then, techniques in Natural Language Processing (NLP) were applied in the data preparation process. For pattern analysis, the researcher employed three different Word Embedding techniques that are Word2Vec, BERT, and RoBERTa and employed dimensionality reduction methods by using PCA, t-SNE and UMAP. By comparing the clustering results using K-Means and Agglomerative methods, it was observed that the best clustering performance was achieved when combining K-Means clustering with Word2Vec Word Embedding and applying UMAP for dimensionality reduction. This approach resulted in the highest Silhouette Score of 0.5197482, indicating the effectiveness of this method in clustering prompts into six distinct groups.

Keywords: Pattern Analysis, Clustering, Generative AI, Prompts, Artificial Intelligent.

* School of Information Technology, Sripatum University
** Faculty of Education, Chulalongkorn University
Corresponding author e-Mail: tiparoon.ki@spu.ac.th

การวิเคราะห์รูปแบบและจัดกลุ่มข้อความพรอมป์สำหรับเจเนอเรทีฟเอไอ

วันที่ได้รับต้นฉบับบทความ: 10 ตุลาคม 2566
วันที่แก้ไขปรับปรุงบทความ: 21 ตุลาคม 2566
วันที่ตอบรับตีพิมพ์บทความ: 2 พฤศจิกายน 2566

ทิพย์อรุณ เขี้ยวแก้ว*
ปราลี มณีรัตน์*
ศุภฤกษ์ ชูธงชัย**

บทคัดย่อ

งานวิจัยนี้เน้นการวิเคราะห์และจัดกลุ่มข้อความพรอมป์ (Prompt) สำหรับระบบปัญญาประดิษฐ์แบบเจเนอเรทีฟ โดยการวิจัยนี้ ผู้วิจัยใช้ ChatGPT สร้างข้อความพรอมป์ที่มีความหลากหลาย ซึ่งในขั้นตอนการจัดเตรียมข้อมูลได้ประยุกต์วิธีการต่าง ๆ ในกระบวนการประมวลผลภาษาธรรมชาติ โดยกระบวนการวิเคราะห์รูปแบบ (Pattern analysis) ผู้วิจัยใช้เทคนิคการฝังคำ (Word embedding) จำนวน 3 เทคนิค คือ Word2Vec, Bert และ RoBERTa และได้ลดมิติข้อมูลด้วยวิธี PCA, t-SNE และ UMAP จากการเปรียบเทียบการจัดกลุ่มด้วยวิธี K-Means และ Agglomerative แสดงให้เห็นว่าสามารถจัดกลุ่มข้อความพรอมป์ได้ 6 กลุ่ม โดยการจัดกลุ่มแบบ K-Means ร่วมกับการฝังคำแบบ Word2Vec และลดมิติข้อมูลด้วยวิธี UMAP ให้ประสิทธิภาพในการจัดกลุ่มสูงที่สุด มีคะแนน Silhouette อยู่ที่ 0.5197482 คะแนน

คำสำคัญ: การวิเคราะห์รูปแบบ, การจัดกลุ่ม, เจเนอเรทีฟเอไอ, พรอมป์, ปัญญาประดิษฐ์

* คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม
** คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
Corresponding author e-Mail: tiparoon.ki@spu.ac.th

บทนำ

ปัจจุบันถือเป็นยุคของการพัฒนาปัญญาประดิษฐ์ (Artificial intelligent) โดยเฉพาะปัญญาประดิษฐ์แบบเจเนอเรทีฟ (Generative AI) ที่กลายเป็นเครื่องมือที่ได้รับความสนใจในการสร้างเนื้อหาต่าง ๆ เช่น การสร้างงานเขียนที่มีรูปประโยคที่ถูกต้องและสามารถร้อยเรียงคำได้อย่างลื่นไหล (Text generator) การสร้างวิดีโอที่มีความสมจริงมากขึ้น (Video generator) การสร้างสรรค์ผลงานทางศิลปะใหม่ ๆ (AI art generator) อาทิ การสร้างรูปภาพและการสร้างเสียงดนตรี เป็นต้น การประยุกต์ใช้ปัญญาประดิษฐ์เหล่านี้ส่งผลกับหลาย ๆ ด้านของชีวิตมนุษย์ การปรับตัวให้ทันต่อนวัตกรรมใหม่ ๆ จึงเป็นสิ่งที่มีความท้าทายและเป็นเรื่องที่สำคัญอย่างยิ่ง

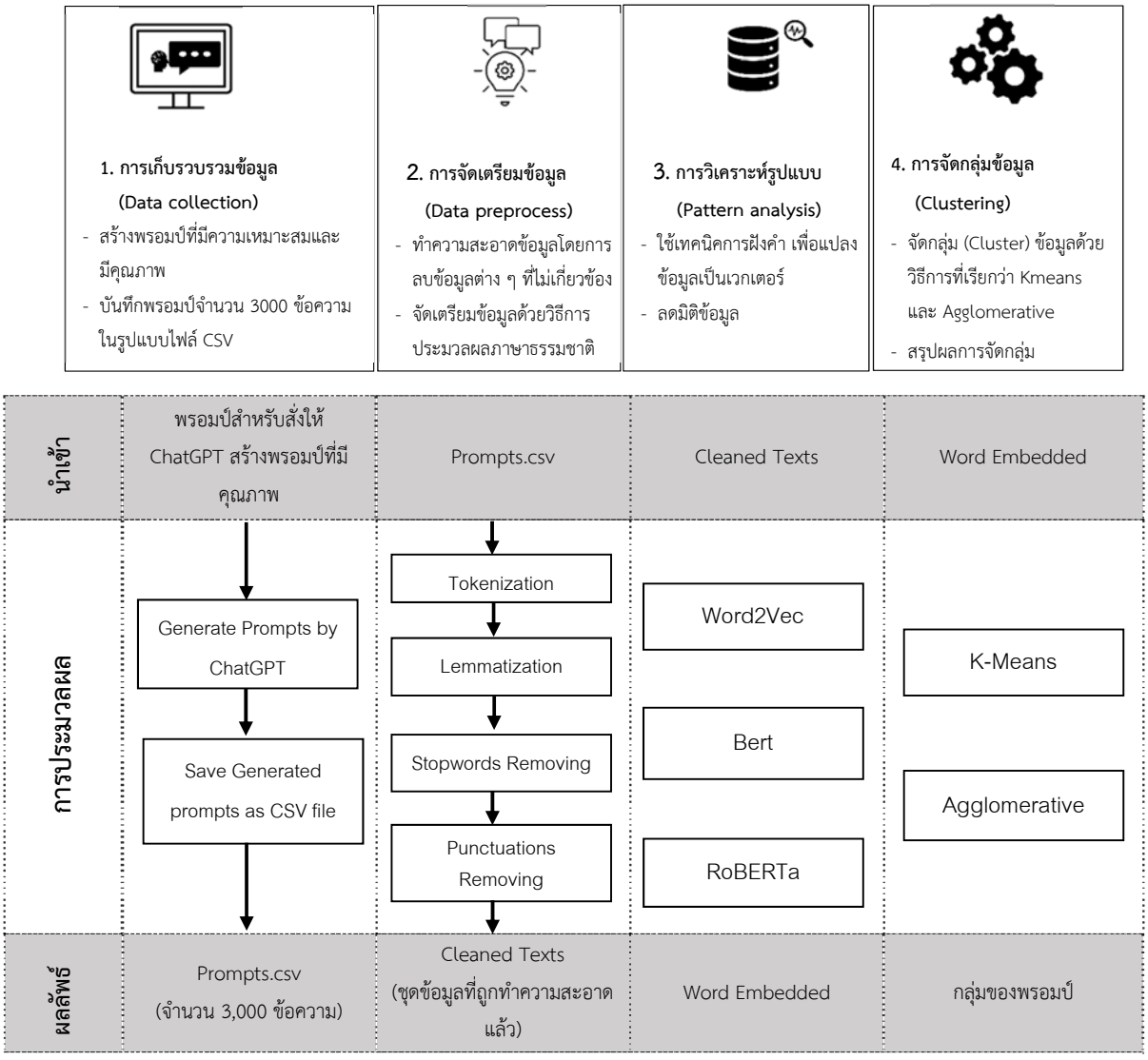
เจเนอเรทีฟเอไอ (Generative AI) คือ ประเภทของปัญญาประดิษฐ์ที่สามารถสร้างข้อมูลและสื่อต่าง ๆ โดยอัตโนมัติ โดยไม่ต้องมีมนุษย์มีส่วนร่วมในกระบวนการสร้าง เจเนอเรทีฟเอไอมีความสามารถในการสร้างข้อมูลที่มีความคล้ายคลึงกับข้อมูลที่มีอยู่ตามคำสั่งหรือคำขอที่ระบุ โดยใช้โมเดลการเรียนรู้เชิงลึก (Deep learning) และเทคนิคการสร้างข้อมูลที่หลากหลาย โดยคำสั่งที่ใช้เพื่อสั่งให้เจเนอเรทีฟประมวลผล เรียกว่า พรอมป์ (Prompt) กล่าวคือ พรอมป์เป็นข้อความหรือคำสั่งที่ผู้ใช้ส่งเข้าสู่ระบบเจเนอเรทีฟเพื่อให้ระบบสร้างเนื้อหาหรือข้อมูลตามคำขอหรือตามรายละเอียดที่ได้ระบุไว้ในข้อความนั้น ๆ พรอมป์จึงเป็นส่วนสำคัญที่มีบทบาทในการกำหนดรูปแบบและเนื้อหาที่ระบบจะสร้าง โดยประสิทธิภาพการทำงานของเจเนอเรทีฟเอไอนั้นขึ้นอยู่กับปัจจัย 2 ข้อ คือ 1) ความสามารถในการประมวลผลของโมเดลที่ใช้ และ 2) การเลือกใช้พรอมป์ที่เหมาะสมสำหรับควบคุมผลลัพธ์ที่ได้จากระบบการประมวลผลพรอมป์เหล่านี้อาศัยกระบวนการประมวลผลภาษาธรรมชาติ (Natural language processing) แต่ทั้งนี้พบว่า ปัญหาหลักที่ส่งผลต่อการใช้งานและประสิทธิภาพของเจเนอเรทีฟเอไอนั้นมาจากผู้ใช้งานสามารถจำแนกได้ 2 สาเหตุ คือ 1) ผู้ใช้ขาดความเข้าใจเกี่ยวกับแนวความคิดและการทำงานของเจเนอเรทีฟเอไอ ส่งผลให้ไม่สามารถสร้างพรอมป์ที่เหมาะสมและมีคุณภาพให้กับระบบ 2) การสร้างพรอมป์ที่ไม่เหมาะสม เช่น ผู้ใช้อาจใช้คำศัพท์หรือรูปประโยคที่ไม่ถูกต้อง ไม่ระบุความต้องการให้ชัดเจน และไม่มีระบุรายละเอียดที่ต้องการ เป็นต้น ส่งผลให้ระบบไม่สามารถสร้างผลลัพธ์ให้สอดคล้องกับความต้องการได้ จากปัญหาที่กล่าวมาข้างต้น ผู้วิจัยจึงมีแนวคิดในการสร้างระบบพรอมป์แบบอัตโนมัติแก่ผู้ใช้งาน ทั้งนี้ก่อนการสร้างพรอมป์ที่เหมาะสม ควรศึกษารูปแบบพรอมป์ที่ดีที่ถูกนำไปใช้ในงานประเภทต่าง ๆ ดังนั้น จึงเกิดเป็นงานวิจัยชิ้นนี้

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและวิเคราะห์รูปแบบข้อความ (Pattern analysis) ที่เหมาะสมสำหรับการสร้างพรอมป์ที่ใช้ในระบบเจเนอเรทีฟเอไอ

2. เพื่อจัดกลุ่มพรอมป์ตามโครงสร้างของข้อความ

กรอบแนวคิดและวรรณกรรมที่เกี่ยวข้อง



ภาพที่ 1 กรอบแนวคิดการทำวิจัย

กรอบแนวคิดการวิจัยนี้ประกอบด้วย 4 กระบวนการ ดังภาพที่ 1 ดังนี้

1. การรวบรวมพรอมป์ (Prompt collection) โดยพรอมป์ที่ถูกรวบรวมมานี้จะเป็นข้อความภาษาอังกฤษที่มีคุณภาพและถูกต้องตามหลักการวิศวกรรมพรอมป์ (Prompt engineering) จำนวน 3,000 ข้อความ ถูกสร้างมาจากระบบ ChatGPT

2. กระบวนการจัดเตรียมข้อมูล (Data preprocessing) เป็นการนำเอาวิธีการต่าง ๆ ในการประมวลผลภาษาธรรมชาติมาใช้ในการจัดเตรียมข้อมูลสำหรับการประมวลผล เช่น การลบเครื่องหมายและอักขระพิเศษ การตัดคำ (Tokenization) และการลบคำฟุ่มเฟือย (Stop words) เป็นต้น

3. กระบวนการวิเคราะห์รูปแบบพรอมป์ (Pattern analysis) โดยข้อความจะผ่านวิธีการฝังคำ (Word embedding) ในงานวิจัยนี้ ผู้วิจัยเลือกทดลองโดยใช้เทคนิคการฝังคำ 3 แบบ คือ Word2Vec, Bert และ RoBERTa

4. การจัดกลุ่ม (Clustering) เป็นการจัดกลุ่มพรอมป์ ด้วย K-Means และ Agglomerative โดย

วรรณกรรมที่เกี่ยวข้องกับงานวิจัยนี้เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ และการสนทนาระหว่างมนุษย์กับระบบปัญญาประดิษฐ์ เป็นเรื่องที่มีความสำคัญอย่างมากในการออกแบบคำสั่งงาน หรือพรอมป์ ที่เหมาะสมและมีคุณภาพให้กับระบบปัญญาประดิษฐ์แบบเจเนอเรทีฟเอไอ เช่น โปรแกรม ChatGPT ซึ่งการเพิ่มประสิทธิภาพของการสนทนาให้มีความชัดเจน และมีความเข้าใจข้อความที่ถูกต้องถือเป็นสิ่งสำคัญ เพราะจะช่วยให้ระบบเข้าใจและตอบสนองตามความคาดหวังของผู้ใช้ได้ดีขึ้น (Dwivedi et al., 2023) การอธิบายอย่างชัดเจน และการเล่าเรื่องในลักษณะที่ตรงไปตรงมาเป็นสิ่งสำคัญเพราะจะช่วยให้ระบบเข้าใจและตอบสนองอย่างเหมาะสม การหลีกเลี่ยงศัพท์เฉพาะหรือคำศัพท์ทางเทคนิคที่ซับซ้อน และการจัดเตรียมตัวอย่างหรือสถานการณ์ตามบริบทก็เป็นอีกปัจจัยสำคัญที่ช่วยให้ระบบสามารถตอบสนองอย่างถูกต้องมากขึ้น

ปัจจุบันในวงการวิศวกรพรอมป์ (Prompt engineer) และเทคโนโลยีเจเนอเรทีฟเอไอ การประมวลผลภาษาธรรมชาติเป็นปัจจัยสำคัญที่ทำให้คอมพิวเตอร์สามารถเข้าใจ และประมวลผลข้อมูล และเนื้อหาที่เกี่ยวข้องกับภาษามนุษย์ได้อย่างรวดเร็วและแม่นยำ (Hosseini, Rasmussen & Resnik, 2023; Sohail et al., 2023) ด้วยความสามารถในการสร้างข้อมูลหรือเนื้อหาใหม่โดยอาศัยการเรียนรู้เชิงลึกและโมเดลปัญญาประดิษฐ์ที่มีความสามารถในการสร้างข้อความที่คล้ายกับข้อความที่มนุษย์เขียน การออกแบบพรอมป์เป็นองค์ประกอบสำคัญที่จะช่วยให้มนุษย์สามารถสร้าง แสดงคำถาม และออกคำสั่งที่เหมาะสม เพื่อให้ระบบเจเนอเรทีฟเอไอเข้าใจและตอบสนองตามความต้องการของมนุษย์ ซึ่งเป็นกระบวนการที่ต้องการความเข้าใจลึกซึ้งในทั้งภาษาธรรมชาติและเทคโนโลยีปัญญาประดิษฐ์ ดังนั้น การแสดงผลลัพธ์ที่แม่นยำและเหมาะสมอีกวิธีหนึ่ง คือ การวิเคราะห์รูปแบบและการจัดกลุ่มของพรอมป์ ซึ่งเป็นเครื่องมือที่มีความสำคัญในการจัดการข้อมูลที่ซับซ้อน การวิเคราะห์รูปแบบช่วยให้ค้นพบความสัมพันธ์และรูปแบบที่อาจไม่เป็นที่รู้จักในข้อมูล โดยใช้เทคนิคและโมเดลทางคณิตศาสตร์และสถิติ นอกจากนี้การจัดกลุ่มช่วยให้สามารถแบ่งข้อมูลออกเป็นกลุ่ม ๆ ตามลักษณะที่คล้ายคลึงกัน เพื่อวิเคราะห์และทำความเข้าใจข้อมูลในมุมมองที่เป็นรูปแบบช่วยในการเข้าใจความสัมพันธ์ระหว่างข้อมูลในมุมมองที่ลึกซึ้งมากขึ้น

ด้วยกลไกของ เจเนอเรทีฟเอไอ ที่เป็นการประมวลผลภาษาธรรมชาติ ซึ่งมีศักยภาพในการวิเคราะห์และคาดการณ์บทสนทนาที่ถูกป้อนเข้าไปแล้วทำการโต้ตอบกลับมา ทำให้การเขียนพรอมป์นั้นเป็นจุดสำคัญที่สุดอย่างหนึ่งในการใช้งานเจเนอเรทีฟเอไอให้มีประสิทธิภาพ ซึ่งในแง่มุมหนึ่งนั้น อาจจะเปรียบได้ว่า พรอมป์ คือการ “สั่งงาน” ให้กับเจเนอเรทีฟเอไอทำงานตามสิ่งที่ผู้ใช้ต้องการ ดังนั้น ทำให้วิธีการเขียนพรอมป์เพื่อสั่งงาน

เจเนอเรทีฟเอไอ ต้องใช้โครงสร้างและรูปแบบการเขียนพรอมป์ที่เหมาะสม ซึ่งมีงานวิจัยต่าง ๆ ที่เกี่ยวข้องกับโครงสร้าง และองค์ประกอบของการเขียนพรอมป์ที่เหมาะสมสามารถสรุปได้ตารางที่ 1

ตารางที่ 1 องค์ประกอบของการเขียนพรอมป์

องค์ประกอบของพรอมป์	Bozkurt and Sharma (2023)	Korzynski, Mazurek, Krzypkowska & Kurasinski (2023)	Alexandra (2023)	Starita (2023)	Leexi (2023)	White et al. (2023)	Kumar (2023)	Tamsin (2023)	Khan (2023)	cloudHQ (2023)
กำหนดวัตถุประสงค์/ เป้าหมาย	•			•	•					
มีความชัดเจน ตรงประเด็น และกระชับ	•	•	•					•		•
กำหนดบทบาท					•	•	•		•	
กำหนดงาน/ คำสั่ง					•	•	•	•	•	•
ให้บริบท	•	•	•		•	•	•	•	•	•
กำหนดขั้นตอน					•					
ให้ตัวอย่าง	•	•		•		•			•	
ระบุรูปแบบที่ต้องการ	•	•		•	•				•	
ระบุรายละเอียดที่สำคัญ	•			•					•	
กำหนดข้อจำกัด					•					

กล่าวโดยสรุป การวิเคราะห์รูปแบบและจัดกลุ่มข้อความของชุดคำสั่งพรอมป์สำหรับเจเนอเรทีฟเอไอเป็นวิธีการที่มีความสำคัญในการจัดการข้อมูลที่ซับซ้อน อาจกล่าวได้ว่าการวิเคราะห์รูปแบบและจัดกลุ่มนี้สามารถช่วยให้ผู้วิจัย และผู้พัฒนาสามารถเข้าใจโครงสร้างของข้อความและสามารถจัดเตรียมข้อความพรอมป์ที่ดี เพื่อนำไปใช้ในงานประเภทต่าง ๆ ได้อย่างมีประสิทธิภาพ ทั้งนี้ยังช่วยให้เกิดความเข้าใจเกี่ยวกับความสัมพันธ์ระหว่างข้อมูลในมุมมองที่ลึกซึ้งมากขึ้น

วิธีดำเนินการวิจัย

หลังจากรวบรวมพรอมป์ จำนวน 3,000 ข้อความ แล้ว ผู้วิจัยจะจัดเตรียมข้อมูลโดยใช้วิธีการเกี่ยวกับการประมวลผลภาษาธรรมชาติ เช่น การลบเครื่องหมายและอักขระพิเศษ การตัดคำ (Tokenization) และการลบคำฟุ่มเฟือย (Stop words) เป็นต้น เพื่อวิเคราะห์รูปแบบพรอมป์ผู้วิจัยได้ทำการแปลงข้อมูลให้อยู่ในรูปของ

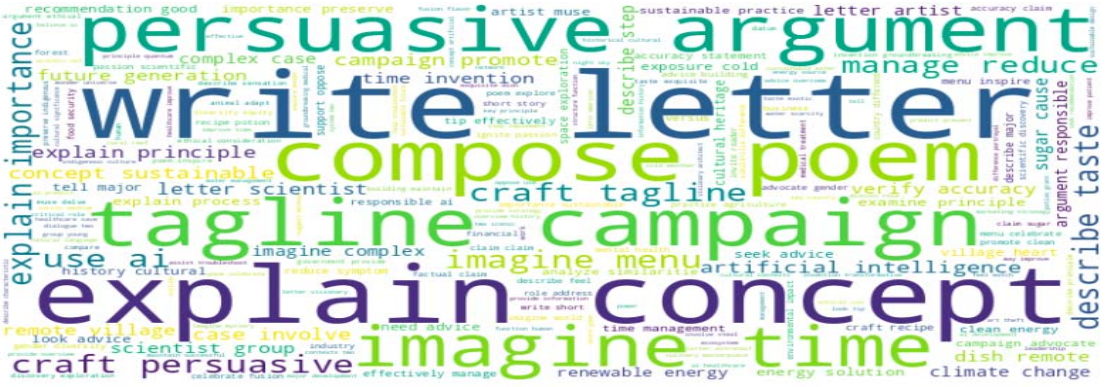
เวกเตอร์ หรือที่เรียกว่า การฝังคำ (Word embedding) และทำการจัดกลุ่ม ข้อความพรอมป์ด้วยวิธี K-Means และ Agglomerative

การเก็บรวบรวมข้อมูล

ในงานวิจัยนี้ ได้ทำการสำรวจข้อมูลจากชุดข้อมูลที่ประกอบด้วยพรอมป์ จำนวน 3,000 ข้อความ ซึ่งเป็นข้อมูลที่ใช้นำเข้าสำหรับการสร้างระบบประมวลผลภาษาธรรมชาติแบบเจเนอทีฟเอไอ ขั้นตอนนี้ผู้วิจัยได้รวบรวมพรอมป์ที่ถูกสร้างขึ้นจาก ChatGPT โดยมีขอบเขตคำสั่งที่ครอบคลุมหัวข้อต่าง ๆ หลากหลายสถานการณ์ รวมถึงมีความแตกต่างทางรูปแบบ และโครงสร้างทางภาษา หลังจากนั้นจะทำความสะอาดข้อมูล ซึ่งรวมถึงการลบคำสั่งที่ไม่เกี่ยวข้อง (Preprocessing) โดยใช้เทคนิคการตัดคำ การลบคำทั่วไปที่ไม่มีความหมาย และการแปลงคำให้เป็นรูปแบบเต็ม (Lemmatization) เพื่อแสดงคำที่พบบ่อยในข้อมูล ทั้งนี้ผลการวิเคราะห์การใช้คำ ความยาวของข้อความ ถูกแสดงดังตารางที่ 2 และผู้วิจัยได้ทำการสร้าง Word Cloud แสดงคำที่พบในชุดข้อมูลดังภาพที่ 2

ตารางที่ 2 ผลการวิเคราะห์ชุดข้อมูลที่ใช้ในงานวิจัย

หัวข้อการวิเคราะห์ข้อมูล	จำนวน	หน่วย
จำนวนคำทั้งหมดที่พบในชุดข้อมูล		
- รวมคำซ้ำ	28,357	คำ
- เฉพาะคำที่ไม่ซ้ำ	2,832	คำ
พรอมป์ที่ยาวที่สุด	28	คำ
พรอมป์ที่สั้นที่สุด	3	คำ



ภาพที่ 2 Word Cloud แสดงคำที่พบในชุดข้อมูล

กระบวนการการทำความสะอาดข้อมูล

ในกระบวนการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสม ประกอบด้วยขั้นตอนการประมวลผล ดังนี้

1. การตัดคำ (Word segmentation) เป็นการแบ่งข้อความออกให้อยู่ในหน่วยของคำ (Token) ซึ่งในงานวิจัยนี้ได้นำโมดูลการตัดคำของไลบรารี่ที่ชื่อว่า “nltk” ในภาษาไพทอนมาใช้งาน

2. การแปลงคำให้อยู่ในรูปพื้นฐาน (Lemmatization) เป็นกระบวนการด้านการประมวลผลภาษาธรรมชาติ โดยระบบจะแปลงคำภาษาอังกฤษให้อยู่ในรูปของรากคำศัพท์นั้นตามพจนานุกรม เพื่อให้คำถูกปรับอยู่ในรูปเดียวกัน ง่ายต่อการนำไปประมวลผล เช่น “running” ปรับเป็น “run”, “better” ปรับเป็น “good” และ “wrote” ปรับเป็น “write” เป็นต้น

3. การกรองคำที่ไม่มีความหมายออกจากข้อมูล (Stopwords removing) เป็นกระบวนการกรองคำที่ไม่มีความหมายออกจากข้อมูล โดย stop_words คือ คำทั่วไปที่ไม่มีความหมาย เช่น “a”, “an” และ “the”

4. การกรองอักขระพิเศษ (Punctuations removing) เป็นกระบวนการลบอักขระพิเศษ รวมถึงสัญลักษณ์ต่าง ๆ ออกจากข้อความ

กระบวนการวิเคราะห์รูปแบบของข้อความ

หลังจากทำความสะอาดข้อมูลเรียบร้อยแล้ว พรอมป์จะถูกนำเข้ากระบวนการฝังคำ ซึ่งเป็นเทคนิคที่ใช้ในการประมวลผลภาษาธรรมชาติ ซึ่งมีด้วยกันหลายเทคนิค เช่น Word2Vec, GloVe (Global Vectors for Word Representation), FastText, BERT (Bidirectional Encoder Representations from Transformers) และ RoBERTa (A Robustly Optimized BERT Pretraining Approach) เป็นต้น ซึ่งทำหน้าที่สร้างเวกเตอร์ที่แสดงความหมายของคำโดยพิจารณาความสัมพันธ์ระหว่างคำที่อยู่ใกล้กันในข้อความ และทำให้คำที่มีความหมายคล้ายคลึงอยู่ใกล้กันในเวกเตอร์ การฝังคำช่วยให้ระบบคอมพิวเตอร์สามารถเรียนรู้และเข้าใจความหมายของคำและข้อความ นอกจากนี้ เทคนิคนี้ยังช่วยลดมิติข้อมูลข้อความขนาดใหญ่และเพิ่มประสิทธิภาพในการประมวลผลข้อความ ในงานวิจัยนี้ผู้วิจัยได้เปรียบเทียบการฝังคำ 3 เทคนิค ได้แก่ 1) Word2Vec 2) BERT และ 3) RoBERTa หลังจากนั้นผู้วิจัยได้ใช้เทคนิคที่ชื่อว่า PCA (Principal Component Analysis) และ t-SNE (t-Distributed Stochastic Neighbor Embedding) เพื่อลดมิติของข้อมูล

การจัดกลุ่มข้อมูล

เป็นกระบวนการในการแบ่งข้อมูลเป็นกลุ่ม โดยคำนึงถึงความคล้ายคลึง หรือความเหมือนกันระหว่างข้อมูลในกลุ่มเดียวกัน และความแตกต่างระหว่างกลุ่มต่าง ๆ ในข้อมูล วัตถุประสงค์หลักของการจัดกลุ่มข้อมูล คือ การหาโครงสร้าง ความหมาย หรือรูปแบบซ่อนที่อาจมีในข้อมูล และช่วยในการวิเคราะห์และเข้าใจข้อมูลอย่างมีประสิทธิภาพ ในงานนี้ผู้วิจัยเลือกเปรียบเทียบการจัดกลุ่มโดยใช้ 2 อัลกอริทึม ดังนี้

1. K-Means โดยหลักการของ K-Means คือ การค้นหาจุดศูนย์กลาง (Centroid) ของกลุ่มแต่ละกลุ่ม โดย k คือ จำนวนกลุ่มที่ต้องการแบ่ง และจากนั้นแบ่งข้อมูลเข้ากลุ่มโดยใช้จุดศูนย์กลางที่ใกล้ที่สุด ซึ่งมีสมการคำนวนระยะทาง (Distance) ที่บ่งบอกความคล้ายคลึงระหว่างข้อมูลและจุดศูนย์กลางที่ใช้มากที่สุด คือ Euclidean Distance ดังสมการที่ 1 : $d(x, y)$

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

โดย x และ y คือ จุดที่ต้องการวัดระยะทางระหว่างกัน
x_i และ y_i คือ ค่าของจุด x และ y ในมิติที่ i
n คือ จำนวนมิติทั้งหมด

2. Agglomerative เป็นการจัดกลุ่มแบบขั้นตอน (Hierarchical clustering) คือ เริ่มจากการสร้างกลุ่มย่อยเล็ก ๆ โดยไม่จำเป็นต้องกำหนดจำนวนกลุ่ม หลังจากนั้น คำนวณระยะทางของกลุ่มข้อมูลแล้วรวมกลุ่มข้อมูลที่มีระยะทางใกล้กันที่สุดเข้าด้วยกัน

ผลการวิจัย

ในงานวิจัยนี้ได้ประเมินคุณภาพของการจัดกลุ่ม เพื่อตรวจสอบความเหมาะสมของการจัดกลุ่มในข้อมูลด้วยคะแนน Silhouette ซึ่งมีคะแนนอยู่ในช่วงระหว่าง -1 ถึง 1 โดยคะแนนที่เข้าใกล้ 1 มากที่สุดจะแสดงถึงการจัดกลุ่มที่ดี และมีคุณภาพ สูตรการคำนวณคะแนน Silhouette มีขั้นตอน ดังนี้

1. คำนวณค่า $a(i)$ ซึ่งเป็นค่าเฉลี่ยของระยะทางระหว่างจุดข้อมูลที่ i กับจุดข้อมูลในกลุ่มเดียวกัน

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j) \quad (2)$$

โดยที่ $a(i)$ คือ คะแนน Silhouette Score สำหรับจุดข้อมูลที่ i, n คือ จำนวนจุดข้อมูลในกลุ่ม (รวมตัวเอง) และ $d(i,j)$ คือ ระยะทางระหว่างจุดข้อมูลที่ i กับจุดข้อมูลที่ j

2. คำนวณค่า $b(i)$ ซึ่งเป็นค่าเฉลี่ยของระยะทางระหว่างจุดข้อมูลที่ i กับจุดข้อมูลในกลุ่มที่ไม่ใช่กลุ่มเดียวกันที่มีค่าเฉลี่ยระยะทางต่ำที่สุด

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \quad (3)$$

โดยที่ $b(i)$ คือ คะแนน Silhouette Score สำหรับจุดข้อมูลที่ i, m คือ จำนวนจุดข้อมูลในกลุ่มที่ไม่ใช่กลุ่มเดียวกัน และ $d(i,j)$ คือ ระยะทางระหว่างจุดข้อมูลที่ i กับจุดข้อมูลที่ j

1. คำนวณคะแนน $s(i)$ สำหรับแต่ละจุดข้อมูล:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

2. คะแนน $Silhouette\ Score$ สำหรับทั้งกลุ่ม คำนวณเฉลี่ยของคะแนน $s(i)$ สำหรับแต่ละจุดข้อมูล

$$Silhouette\ Score\ = (1 / n) * \sum s(i). \tag{5}$$

คะแนน Silhouette Score อยู่ในช่วง [-1, 1] โดยคะแนนที่ใกล้ 1 หมายถึง การจัดกลุ่มที่ดีมาก คะแนนที่ใกล้ 0 หมายถึง การจัดกลุ่มที่ไม่แน่นอน และคะแนนที่ใกล้ -1 หมายถึง การจัดกลุ่มที่ไม่ถูกต้อง โดยผลที่ได้จากการวิจัยนี้แสดงดังตารางที่ 3 และ 4 จากการวิจัยพบว่า สามารถจัดกลุ่มข้อความพรอมป์จากทั้งหมด 3,000 ข้อความ ออกเป็น 6 กลุ่ม โดยวิธีการฝังคำแบบ Word2Vec ร่วมกับการลดมิติข้อมูลด้วย PCA แล้วจัดกลุ่มแบบ K-means ให้ค่าคะแนน Silhouette สูงสุดอยู่ที่ 0.47634807 คะแนน และวิธีการฝังคำแบบ RoBERTa ร่วมกับการลดมิติข้อมูลด้วย PCA แล้วจัดกลุ่มแบบ Agglomerative มีคะแนน Silhouette ต่ำสุด อยู่ที่ 0.31291047 คะแนน

ตารางที่ 3 คะแนน Silhouette แสดงคุณภาพการจัดกลุ่มข้อความพรอมป์โดยลดมิติข้อมูลด้วยวิธี PCA

	Word2Vec	**Bert**	**RoBERTa**
K-Means	0.47634807	0.35758436	0.36506227
Agglomerative	0.44496936	0.31907082	0.31291047

ตารางที่ 4 คะแนน Silhouette แสดงคุณภาพการจัดกลุ่มข้อความพรอมป์โดยลดมิติข้อมูลด้วยวิธี t-SNE

	Word2Vec	**Bert**	**RoBERTa**
K-Means	0.38337007	0.3806843	0.3855886
Agglomerative	0.3160876	0.32726178	0.3597865

อภิปรายผล

จากผลการวิจัยแสดงให้เห็นว่าคะแนน Silhouette อยู่ในช่วง 0.32-0.48 เพื่อปรับปรุงคุณภาพการจัดกลุ่มผู้วิจัยได้นำเทคนิคการลดมิติของข้อมูลที่ชื่อว่า UMAP (Uniform Manifold Approximation and Projection) ซึ่งเป็นเทคนิคทางคณิตศาสตร์ จากการทดลองเพิ่มเติมคะแนน Silhouette ถูกแสดงดังตารางที่ 5 โดยคะแนนเพิ่มขึ้นอยู่ในช่วง 0.46-0.52 ทั้งนี้การจัดกลุ่มด้วย K-Means มีประสิทธิภาพในการจัดกลุ่มสูงที่สุด มีคะแนน Silhouette อยู่ที่ 0.5197482 คะแนน

ตารางที่ 5 คะแนน Silhouette แสดงคุณภาพการจัดกลุ่มพรอมป์โดยลดมิติข้อมูลด้วยวิธี UMAP

	Word2Vec	**Bert**	**RoBERTa**
K-Means	0.5197482	0.51312786	0.48059615
Agglomerative	0.50099605	0.51298964	0.45805013

สามารถสรุปผลการวิจัยตามวัตถุประสงค์ของโครงการวิจัยได้ว่า จากการศึกษาและวิเคราะห์รูปแบบข้อความพรอมป์ที่มีความหลากหลายทางรูปแบบ และโครงสร้างทางภาษา สามารถแบ่งพรอมป์ ออกเป็น 6 กลุ่ม โดยพิจารณาจากความคล้ายคลึงกันของโครงสร้างทางภาษาและคำที่ใช้ ซึ่งวิธีการวิเคราะห์และจัดกลุ่มพรอมป์ที่ให้ประสิทธิภาพสูงสุด คือ ประยุกต์วิธีการต่าง ๆ ในกระบวนการประมวลผลภาษาธรรมชาติในกระบวนการการทำความสะอาดข้อมูล ใช้เทคนิคการฝังคำแบบ Word2Vec แล้วลดมิติข้อมูลด้วยวิธี UMAP ก่อนจัดกลุ่มพรอมป์ด้วยวิธี K-Means

ข้อเสนอแนะและงานวิจัยในอนาคต

วิจัยชิ้นนี้ได้ทำการเก็บรวบรวมพรอมป์จำนวนทั้งสิ้น 3,000 ข้อความ เปรียบเทียบการฝังคำด้วย 3 วิธี คือ Word2Vec, Bert และ RoBERTa หลังจากนั้นทำการลดมิติของข้อมูลเพื่อเพิ่มประสิทธิภาพในการจัดกลุ่มด้วยเทคนิค PCA, t-SNE และ UMAP แล้วได้ทำการจัดกลุ่มพรอมป์ด้วย K-Means และ Agglomerative จากการวิจัยพบว่า สามารถแบ่งกลุ่มพรอมป์ได้เป็น 6 กลุ่ม โดยวิธีการฝังคำแบบ Word2Vec ผ่านการลดมิติข้อมูลแบบ UMAP และจัดกลุ่มด้วย K-Means ให้คะแนน Silhouette สูงสุด ดังนั้น ในงานวิจัยถัดไป ผู้วิจัยมีความสนใจที่จะนำเอาพรอมป์แต่ละกลุ่มไปศึกษาโครงสร้างเพื่อพัฒนาระบบแนะนำพรอมป์สำหรับผู้ใช้ในอนาคต อีกทั้งผลลัพธ์จากวิจัยนี้ช่วยให้ผู้ที่มีความสนใจเกี่ยวกับวิศวกรพรอมป์และเทคโนโลยีเจเนอเรทีฟเอไอสามารถเข้าใจโครงสร้างของข้อความแต่ละกลุ่ม และสามารถจัดเตรียมข้อความพรอมป์ที่ดี เพื่อนำไปใช้ในงานประเภทต่าง ๆ ได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

บรรณานุกรม

Alexandra. (2023). ***Top AI & ChatGPT Prompts for Social Media Post Generation*** (Online). Available: https://socialbee.com/blog/ai-social-media-prompts/ [2023, October 1].

Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. ***Asian Journal of Distance Education***, ***18***, pp. 1-6.

CloudHQ. (2023). ***How to Write ChatGPT Prompts for Email*** (Online). Available: https://blog.cloudhq.net/how-to-write-chatgpt-prompts-for-email/ [2023, October 1].

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M., A., Al-Busaidi, A., S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., & Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. ***International Journal of Information Management***, ***71***, 102642.

Hosseini, M., Rasmussen, L. M., & Resnik, D. B. (2023). Using AI to write scholarly publications. ***Accountability in Research Policies and Quality Assurance***, pp. 1-9.

Korzynski, P., Mazurek, G., Krzypkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. ***Entrepreneurial Business and Economics Review***, ***11***(3), pp. 25-37.

Khan, H. (2023). ***A formula for composing the best prompts*** (Online). Available: https://twitter.com/slow_developer/status/1695027815958405471 [2023, October 20].

Kumar, A. (2023). ***ChatGPT Prompts Design Tips & Examples*** (Online). Available: https://vitalflux.com/chatgpt-prompts-design-tips-examples/ [2023, October 1].

Leexi. (2023). ***Master the art of generating prompts with ChatGPT*** (Online). Available: https://www.leexi.ai/en/business-intelligence/the-7-golden-rules-for-generating-efficient-prompts-with-chatgpt/ [2023, October 1].

Sohail, S. S., Farhat, F., Himeur, Y., Nadeem, M., Madsen, D. Ø., Singh, Y., Atalla, S., & Mansoor, W. (2023). Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions. ***Journal of King Saud University-Computer and Information Sciences***, ***35***(8), 101675.

Starita, L. (2023). ***How to Write AI Prompts: The Key to Better Outputs from Generative AI*** (Online). Available: https://contently.com/2023/04/13/how-to-write-ai-prompts-for-generative-ai/ [2023, October 1].

Tamsin, S. (2023). ***THE ART OF WRITING CHATGPT PROMPTS FOR ANY USE CASE*** (Online). Available: https://sarahtamsin.com/the-art-of-writing-chatgpt-prompts/ [2023, October 1].

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). ***A prompt pattern catalog to enhance prompt engineering with chatGPT*** (Online). Available: https://arxiv.org/pdf/2302.11382.pdf [2023, October 1].