การเปรียบเทียบประสิทธิภาพวิธีหาความเหมือนของคำอธิบายรายวิชาเพื่อเทียบโอนหน่วยกิตด้วยการแปลงเชิงปริมาณ

Efficiency Comparison of Methods for Finding Similarity of Course Descriptions to Transfer Credits using Word Embedding

กมลรัตน์ สมใจ[1] และ แสงดาว นพพิทักษ์[2*]

Kamonrat Somjai [1] and Sangdaow Noppitak[2*]

[1,2] สาขาเทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์, มหาวิทยาลัยราชภัฏบุรีรัมย์

[1,2] Department of Information Technology, Faculty of Science, Buriram Rajabhat University

[*]**Corresponding author, E-mail:** sangdaow.np@bru.ac.th

บทคัดย่อ

การเปรียบเทียบประสิทธิภาพวิธีหาความเหมือนของคำอธิบายรายวิชาเพื่อเทียบโอนหน่วยกิตด้วยการแปลงเชิงปริมาณ โดยงานวิจัยนี้ มุ่งเน้นหาวิธีเพิ่มประสิทธิภาพความแม่นยำวิธีหาความเหมือนของคำอธิบายรายวิชา ข้อมูลที่ใช้ในงานวิจัยนี้ รวบรวมจากรายวิชาของหลักสูตรเทคโนโลยีสา รสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ จำนวน 58 รายวิชา ขั้นตอนการเตรียมข้อมูลประกอบด้วย การตัดคำ (Word Tokenize) ให้อยู่ในรูปแบบคำศัพท์ด้วยหลักการประมวลผลภาษาธรรมชาติ (Natural Language Processing) จากนั้นกำจัดคำหยุด (Stop-Word Elimination) เพื่อกำจัดคำที่ไม่เกี่ยวข้องกับเอกสาร และหารากศัพท์ (Stemming) สำหรับงานวิจัยนี้ นำวิธี TF-IDF หรือ Term Frequency Inverse Document Frequency ซึ่งเป็นเทคนิคการคัดแยกคำตามความสำคัญในการสร้างเวกเตอร์คุณลักษณะจากคำ ผลการทดลอง การหาค่าความเหมือนของข้อความด้วยการหาระยะทาง พบว่า วิธีระยะห่างโคไซน์มีค่าต่ำกว่าวิธีระยะห่างยูคลิเดียน และผลการวัดความเหมือนของข้อความด้วยการหาค่าโคไซน์และเคอร์นอล พบว่า วิธีที่มีประสิทธิภาพดีที่สุดโดยเฉลี่ย ได้แก่ rbf kernel 0.97 ลำดับถัดมาเป็นวิธี Laplacian kernel 0.90 ส่วนวิธี Cosine Similarity, Linear Kernel และ Pairwise Kernels ผลการวัดความเหมือนของข้อความโดยเฉลี่ยเท่ากันคือ 0.62

คำสำคัญ : ความเหมือนของคำ, การแปลงเชิงปริมาณ, หลักการประมวลผลภาษาธรรมชาติ

Abstract

Comparing the efficiency of methods for finding similarities in course descriptions for credit transfer using quantitative transformation. This research focuses on improving accuracy and discovering similarities in course descriptions. The information used in this

study is compiled from the subjects of the Information Technology curriculum at the Faculty of Science, Buriram Rajabhat University, comprising 58 courses. The data preparation process involves word tokenization into vocabulary using the principles of Natural Language Processing, followed by the elimination of stop words and stemming. Stemming aims to find the roots of words. For this research, the Term Frequency Inverse Document Frequency (TF-IDF) method is employed. TF-IDF is a technique for ranking words according to their importance in creating feature vectors from words. Experimental results reveal the similarity value of the text and distance. It is found that the cosine distance method yields lower values compared to the Euclidean distance method. Furthermore, the results of measuring text similarity using cosine and kernel values show that the rbf kernel method performs the best on average, achieving a similarity score of 0.97. Following closely is the Laplacian kernel method with a score of 0.90. The Cosine Similarity, Linear Kernel, and Pairwise Kernels yield an average text similarity measurement result of 0.62.

Keyword : Similarity, Word Embedding, Natural Language Processing

บทนำ

การจัดการเรียนในระดับอุดมศึกษานั้นผู้เรียนสามารถนำรายวิชาจากสถาบันที่ตนเองจบเข้ามาเทียบโอนกับรายวิชาในหลักสูตรที่สอบเข้าได้ และสาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ก็เช่นกัน ได้เปิดรับสมัครโดยกำหนดคุณสมบัติของผู้เรียนที่จบการศึกษาในระดับชั้นมัธยมศึกษาตอนปลายและผู้เรียนที่จบการศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูง (ปวส.) สามารถสมัครเรียนและเทียบโอนหน่วยกิตรายวิชาที่เรียนมาแล้วได้ โดยกำหนดเกณฑ์ในการเทียบโอนรายวิชาจากคำอธิบายรายวิชาที่ขอเทียบโอนจะต้องมีข้อความที่มีความหมายเหมือนตรงกัน อย่างน้อยร้อยละ 80 และผลการเรียนในรายวิชานั้นต้องเป็นเกรด C หรือ เกรด 2 ขึ้นไป จึงจะสามารถขอเทียบโอนได้

จากการศึกษางานวิจัยด้านการทำเหมืองข้อความ (Text Mining) ที่มีการนำมาประยุกต์ใช้ในการวิเคราะห์คำเพื่อหาความเหมือน (Similarity) และวัดความเหมือนกันของข้อความหรือเอกสาร พบว่ามีงานวิจัยจำนวนมากที่นำเสนอวิธีที่แม่นยำในการตรวจวัดความเหมือนกันของข้อความ ดังเช่น การประยุกต์ใช้เอ็นแกรม (N-Gram) ในการหาความคล้ายคลึงของรายวิชาที่เขียนด้วยภาษาอารบิค ซึ่งพบว่า 3-Grams ให้ผลการวิจัยที่ดีที่สุด (Al-Ramahi & Mustafa, 2012) การเปรียบเทียบความคล้ายคลึงของข้อวามเพื่อจำแนกประเภทข้อความในสื่อโซเชียลออนไลน์ โดยใช้สมการวัดความคล้ายคลึงทั้งหมด 10 แบบด้วยกัน ดังนี้ Euclidean, Chebyshev Bray Curtis, Correlation, Canberra, Minkowski, Cosine, Jaccard, City Block, Roger-Tanimoto และได้ทำการกำหนดน้ำหนักของคำก่อนที่จะนำมาเข้าสู่กระบวนการวัดความคล้ายด้วยค่า TF และ TF-IDF ซึ่งผลวิจัย พบว่า สมการที่สามารถวัดความคล้ายคลึงได้ดีที่สุดคือ

Bray Curtis ร่วมกับการกำหนดคุณลักษณะแบบ TF (Viriyavisuthisakul et al., 2015) นอกจากนี้ยังมีวิธีวัดความเหมือนของข้อความอีกหลายวิธี เช่นการวัดความเหมือนระหว่างออบเจ็กต์หนึ่ง (Object) กับอีกออบเจ็กต์หนึ่งด้วยวิธีการทำสถิติ ซึ่งสามารถนำมาประยุกต์ใช้เพื่อวัดความเหมือนของคำรายวิชาภาษาไทยได้ดี ดังนั้นผู้วิจัยจึงสนใจที่จะศึกษาการเปรียบเทียบประสิทธิภาพวิธีหาความเหมือนของคำอธิบายรายวิชาในภาษาไทย ทั้งนี้เพื่อหาวิธีเพิ่มประสิทธิภาพความแม่นยำในการเทียบโอนหน่วยกิตของสาขาวิชาต่อไป

วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบประสิทธิภาพวิธีหาความเหมือนของคำอธิบายรายวิชาเพื่อเทียบโอนหน่วยกิตด้วยการแปลงเชิงปริมาณ

แนวคิด ทฤษฎี

1. การทำเหมืองข้อความ

การทำเหมืองข้อความ (Text Mining) เป็นเทคนิคเพื่อค้นหารูปแบบ (Pattern) หรือความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความจำนวนมหาศาล โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง (Machine Learning) การประมวลผลภาษาธรรมชาติ (Natural Language Processing) และการรู้จำแบบ (Pattern Recognition) ซึ่งอาจแบ่งได้ 3 ประเภท ได้แก่ การสรุปเอกสารข้อความ (Document Summarization) การแบ่งประเภทเอกสาร (Document Classification) และการแบ่งกลุ่มเอกสารข้อความ (Document Clustering) ที่รวมไปถึงการวัดความเหมือนและความแตกต่างของเอกสารและข้อความ (สุขุมาลย์ หนกหลัง, 2563)

การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) (พีรพัฒ โชคสุวัฒนสกุล และคณะ, 2566) มีจุดมุ่งหมายให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้ดียิ่งขึ้น เนื่องจากคอมพิวเตอร์ถูกออกแบบมาให้เหมาะสมกับการเข้าใจข้อมูลของตัวเลขหรือรหัส ซึ่งไม่ตรงกับวิธีการสื่อสารของมนุษย์ NLP จึงเกิดขึ้นเพื่อลดช่องว่างในการสื่อสารระหว่างมนุษย์กับคอมพิวเตอร์ ซึ่งกระบวนการประมวลผลข้อความเพื่อการทำเหมืองข้อความ ประกอบด้วย การตัดคำ (Word Segmentation) การกำจัด คำหยุด (Stop-Word Elimination) การหารากศัพท์ (Stemming) การคัดเลือกคุณลักษณะ (Feature Selection) เป็นต้น (จักรินทร์ สันติรัตนภักดี และศศิธร อิ่มวุฒิ, 2564)

2. การแปลงเชิงปริมาณ

การแปลงเชิงปริมาณ (Word Embedding) คือ การสร้างเวกเตอร์คุณลักษณะจากคำ ซึ่งเป็นการแปลงชุดของคำให้เป็นตัวเลขที่ไม่ซ้ำกันโดยการคำนวนความน่าจะเป็นของคำ และบริบทภายในประโยค ซึ่งนำเสนอในรูปของเวกเตอร์ จุดเด่นของเวกเตอร์คุณลักษณะที่ได้จาก Word Embedding นั้น คือ สามารถนำไปคำนวนความคล้ายคลึงกับคำอื่น ๆ ในบริบทของคำที่แตกต่างกันได้ (ชาลิสา จิตบุญญาพินิจ

et al., 2565) ปัจจุบันได้มีการพัฒนาตัวแบบจำลองสำหรับการทำ Word Embedding ออกมาให้ใช้งานอย่างสะดวก เช่น 1) Word2Vec เป็นแบบจำลองที่ใช้สร้างการฝังคำ หรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ของคำต่าง ๆ ถูกคำนวณจากบริบทรอบข้างโดยใช้เทคนิคโครงข่ายประสาทเทียม (Neural Network) (Mikolov et al., 2013) 2) Glove หรือ Global Vectors for Word Representation เป็นแบบจำลองถดถอยล็อก-ไบลิเนียร์ (Log-bilinear Regression) ที่รวม 2 ตัวแบบเข้าด้วยกัน คือ โกลบอลเมตริกซ์แฟคเตอร์ไรเซชัน (Global Matrix Factorization) และโลคอลคอนเท็กซ์วินโดว์ (Local Context Window) และเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) (Pennington et al., 2014) 3) TF-IDF หรือ Term Frequency Inverse Document Frequency เป็นเทคนิคการคัดแยกคำตามความสำคัญที่ถูกใช้ในการสร้างเวกเตอร์ (สราธิป เตชาวิวัฒนบูลย์, 2560) และปัจจุบันวิธีที่ได้รับความนิยม คือ LLM (Large Language Model) เป็นเทคนิคด้านปัญญาประดิษฐ์ที่ใช้เทคโนโลยีการเรียนรู้ของเครื่อง (Machine Learning) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เพื่อสร้างโมเดลภาษาขนาดใหญ่ที่สามารถเข้าใจและสร้างข้อความภาษาธรรมชาติได้โดยอัตโนมัติ และมีหลักการทำงานคือการใช้โครงข่ายประสาทเทียม (Neural Networks) ที่มีโครงสร้างซับซ้อนในการประมวลผลข้อมูลภาษา โดยมีชั้นการเรียนรู้หลายชั้น (Deep Learning) ที่ช่วยให้โมเดลสามารถเรียนรู้และเข้าใจลักษณะเฉพาะของภาษา รวมถึงความสัมพันธ์และรูปแบบของข้อมูลที่มีอยู่ในภาษาต่าง ๆ (Brown et al., 2020; Chen et al., 2021) แต่มีข้อจำกัดด้านความน่าเชื่อถือในกรณีที่มีข้อมูล (Training data) ไม่เพียงพอต่อการเรียนรู้ (Training)

3. Term Frequency - Inverse Document Frequency

Term Frequency - Inverse Document Frequency หรือ TF-IDF เป็นการสกัดคุณลักษณะของเอกสาร เพื่อหาลักษณะของตัวแทนของเอกสารนั้น ๆ โดยพิจารณาจากความถี่ของคำที่ปรากฏในเอกสาร (Term Frequency :TF) และการหาน้ำหนักของคำที่ปรากฏในเอกสารด้วยค่า TF-IDF ซึ่งเป็นการวัดค่าน้ำหนักที่เอาไว้ประเมินค่าความสำคัญของคำนั้น ๆ ในกลุ่มของเอกสาร ที่คำนวณได้จากสมการ (1) และ (2) (Al-Obaydy et al., 2022)

$$Weight = tf \times idf \quad (1)$$

โดย tf แทนด้วยจำนวนที่มีการปรากฏ term t ในเอกสาร d และ idf ได้มาจากการคำนวณด้วยสมการ

$$idf(t, D) = log \frac{N}{|\{d \epsilon D : t \epsilon d\}|} \quad (2)$$

โดย N แทนจำนวนเอกสารทั้งหมด และ $|\{d \epsilon D: t \epsilon d\}|$ คือจำนวนเอกสาร (d) ที่มี term t ปรากฏอยู่

4. การวัดความเหมือนของข้อความ

การวัดค่าความเหมือน (Cosine Similarity) เป็นการเลือกคุณลักษณะแบบกรอง เป็นการหาค่าความเหมือนหรือความคล้ายคลึงคงที่ได้จากค่าความต่างของมุมของวัตถุ 2 อย่างที่เกิดขึ้นบนพื้นที่เว็กเตอร์ โดยความคล้ายคลึงกันในแบบของ Cosine นั้น จะมีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น จึงเป็นวิธีการที่นิยมและมีประสิทธิภาพสูง (Christopher D. Manning et al., 2008) ดังสมการ (3)

$$similarity = \cos(\theta) = \frac{A}{\|A\|\|B\|} \quad (3)$$

โดย A คือ คุณลักษณะแต่ละตัวของข้อมูลทั้งหมด

B คือ ประเภทของกลุ่มในการจำแนกของแต่ละตัวอย่าง

งานวิจัยที่เกี่ยวข้อง

สราธิป เตชาวิวัฒนบูลย์ (2560) ได้ทำวิจัยเรื่องการเปรียบเทียบวิธีวัดคล้ายคลึงของข้อความบนฐานเวกเตอร์เพื่อการเทียบโอนหน่วยกิตของรายวิชาระดับมหาวิทยาลัย โดยมีวัตถุประสงค์เพื่อวิเคราะห์และเปรียบเทียบประสิทธิภาพของวิธีวัดความคล้ายคลึงของข้อความที่แทนด้วยเวกเตอร์ในการเปรียบเทียบคำอธิบายรายวิชาของมหาวิทยาลัย เพื่อการเทียบโอนหน่วยกิต โดยวิธีวัดความคล้ายคลึง 3 วิธี ได้แก่ วิธีวัดระยะทางแบบเบร์-เคอร์ทิส (Bray-Curtis Distance) วิธีวัดระยะทางเชิงมุม (Cosine Distance) วิธีวัดระยะทางแบบยูคลิด (Euclidean Distance) ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลคำอธิบายรายวิชาที่เขียนด้วยภาษาไทย จากคณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี จำนวน 33 รายวิชา ซึ่งผ่านการตัดคำด้วยโปรแกรมทีเล็กส์ (TLex) แล้วแปลงอยู่ในรูปแบบเวกเตอร์ของความถี่ของเทอม (TF) และการให้น้ำหนักของเทอม (IT-IDF) ผลการวิจัยพบว่า วิธีวัดความคล้ายคลึงของข้อมูลที่แทนด้วยเวกเตอร์ของความถี่ของเทอม (TF) ร่วมกับวิธีวัดระยะทางเชิงมุม (Cosine Distance) ได้ค่าความคล้ายคลึงสูงที่สุดที่ระดับร้อยละ 71.71 ซึ่งสามารถนำไปประยุกต์ใช้ในการตรวจสอบความคล้ายคลึงของคำอธิบายรายวิชาเพื่อการเทียบโอนหน่วยกิตได้

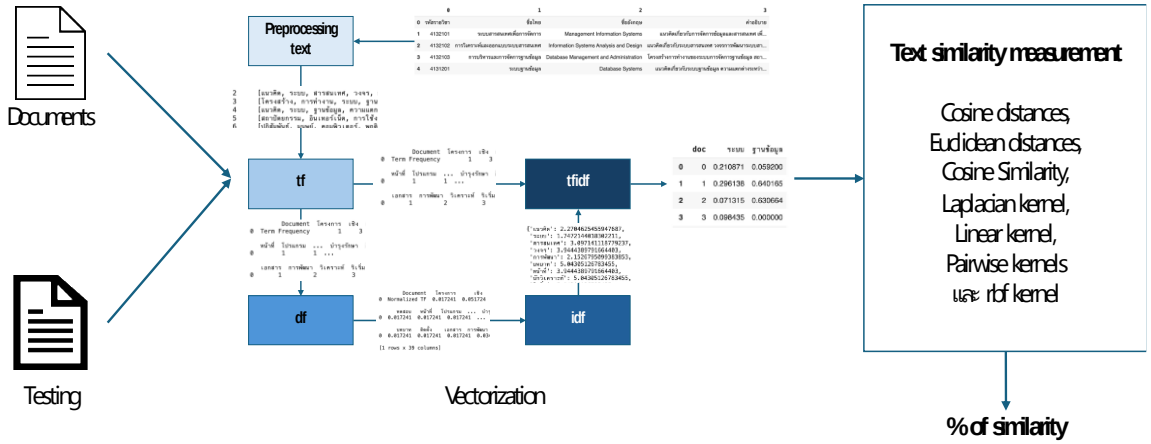
Zoupanos et al. (2022) ทำการเปรียบเทียบความคล้ายคลึงกันของคำอธิบายหลักสูตรระหว่างเทคนิค FAISS และ Elasticsearch กับชุดข้อมูลขนาดใหญ่ โดยเทคนิค FAISS (Facebook AI Similarity Search) เป็นเครื่องมือที่พัฒนาโดย Facebook AI Research เพื่อใช้ในการค้นหาข้อมูลที่คล้ายคลึงกันอย่างรวดเร็วในชุดข้อมูลขนาดใหญ่ โดยเฉพาะสำหรับการค้นหาแบบเวกเตอร์ (Vector search) ส่วนเทคนิค Elasticsearch เป็นเครื่องมือในการค้นหาข้อมูลที่มีความยืดหยุ่นสูง ซึ่งสามารถใช้สำหรับการจัดเก็บข้อมูล

ที่มีโครงสร้างและไม่มีโครงสร้างได้ นอกจากนี้ยังมีความสามารถในการค้นหาข้อมูลโดยอัตโนมัติและการทำความเข้าใจภาษาธรรมชาติ (NLP) มีความสามรถในการค้นหาและการจัดการข้อมูลที่มีประสิทธิภาพในระดับข้อมูลขนาดใหญ่ จากการประเมินประสิทธิภาพของวิธีการเปรียบเทียบเวกเตอร์สองวิธีคือ FAISS และ Elasticsearch พบว่า FAISS มีประสิทธิภาพเหนือกว่า Elasticsearch ในสภาพแวดล้อมแบบรวมศูนย์ด้วยชุดข้อมูลขนาดใหญ่ และ FAISS จัดทำดัชนีสำหรับคำตอบได้อย่างรวดเร็วในการเปรียบเทียบเวกเตอร์

Sitikhu et al. (2019) ได้ศึกษาเรื่องการเปรียบเทียบความคล้ายคลึงกันทางความหมายเพื่อเพิ่มความสามารถในการตีความของมนุษย์ โดยนำเสนอวิธีการที่แตกต่างกันสามวิธี ได้แก่ 1) Cosine Similarity ด้วย TF-IDF 2) Cosine Similarity ด้วย Word Embedding และ 3) Soft Cosine Similarity ด้วย Word Embedding ผลการวิจัย พบว่า วิธีของ Cosine Similarity ด้วย TF-IDF มีประสิทธิภาพดีที่สุดในการค้นหาความคล้ายคลึงระหว่างข้อความ อีกทั้งง่ายต่อการตีความและสามารถนำไปใช้งานในแอปพลิเคชันด้านการค้นคืน

ผลการศึกษาและอภิปรายผล

ในการเปรียบเทียบวิธีวัดความเหมือนของคำอธิบายรายวิชาเพื่อการเทียบโอนหน่วยกิต ผู้วิจัยได้ดำเนินการเป็นไปตามขั้นตอนการวิจัยดังภาพที่ 1 กรอบแนวคิดขั้นตอนที่ใช้ในการดำเนินการวิจัย ซึ่งมีรายละเอียด ดังนี้



ภาพที่ 1 กรอบแนวคิดขั้นตอนที่ใช้ในการดำเนินการวิจัย

1. การรวบรวมข้อมูล (Data Collection)

ผู้วิจัยได้รวบรวมข้อมูลเพื่อการเทียบโอนหน่วยกิตของสาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ จำนวน 58 รายวิชา และมีจำนวนรายวิชาที่นักศึกษายื่นคำร้อง

ขอเทียบโอนหน่วยกิต ประจำปีการศึกษา 2565 จำนวนทั้งสิ้น 14 รายวิชา ทำการบันทึกสร้างเก็บเป็นไฟล์ Microsoft Excel

2. การเตรียมข้อมูล (Pre-Processing)

เนื่องจากภาษาไทยไม่ใช้อักขระช่องว่าง (White Space) เพื่อบ่งชี้ขอบเขตของแต่ละคำเหมือนภาษาอังกฤษ อีกทั้งคอมพิวเตอร์ยังแยกคำภาษาไทยแต่ละคำเป็นประโยคไม่ได้ เพราะไม่รู้จักแต่ละคำเหมือนมนุษย์ ดังนั้นผู้วิจัยจึงได้ดำเนินการดังนี้

2.1 การตัดคำ (Word Tokenize) ตัดคำออกจากข้อความให้อยู่ในรูปแบบคำศัพท์ด้วยหลักการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ซึ่งงานวิจัยนี้ ผู้วิจัยใช้ไลบรารีของ pythainlp ที่ชื่อ tokenze ซึ่งสามารถแสดงตัวอย่างคำอธิบายรายวิชา 4132103 “โครงสร้างการทำงานของระบบการจัดการฐานข้อมูล สถาปัตยกรรมระบบ ฐานข้อมูล ระบบฐานข้อมูลเชิงสัมพันธ์ ระบบฐานข้อมูลแบบกระจาย ระบบฐานข้อมูล แบบอ็อบเจ็กต์ ระบบฐานข้อมูลบนเว็บ การสร้างระบบฐานข้อมูล” และทำการตัดคำได้ผลดังนี้

'โครงสร้าง', 'การทำงาน', 'ของ', 'ระบบ', 'การ', 'จัดการ', 'ฐานข้อมูล', ' ', 'สถาปัตยกรรม', 'ระบบ', ' ', 'ฐานข้อมูล', ' ', 'ระบบ', 'ฐานข้อมูล', 'เชิง', 'สัมพันธ์', ' ', 'ระบบ', 'ฐานข้อมูล', 'แบบ', 'กระจาย', ' ', 'ระบบ', 'ฐานข้อมูล', ' ', 'แบบ', 'อ็อบเจ็กต์', ' ', 'ระบบ', 'ฐานข้อมูล', 'บน', 'เว็บ', ' ', 'การ', 'สร้าง', 'ระบบ', 'ฐานข้อมูล', ' ', 'การ', 'สร้าง', 'พจนานุกรม', ' ', 'ฐานข้อมูล', ' ', 'การ', 'จัดการ', 'ดัชนี', 'ข้อมูล', ' ', 'การ', 'จัดการ', 'บูรณภาพ', 'ของ', 'ข้อมูล', ' ', 'การ', 'จัดการ', 'ความมั่นคง', 'ของ', 'ฐานข้อมูล', ' ', 'การ', 'กำหนด', 'ผู้ใช้', ' ', 'การ', 'จัดการ', 'สิทธิ์', 'ใน', 'การใช้งาน', 'ฐานข้อมูล', 'ระบบ', 'ความปลอดภัย', ' ', 'การ', 'ถ่ายโอน', 'ข้อมูล', 'ระหว่าง', 'ฐานข้อมูล', ' ', 'การ', 'กำหนด', 'กลยุทธ์', 'ใน', 'การสำรอง', 'ข้อมูล', ' ', 'การกู้', 'คืน', 'เมื่อ', 'ระบบ', 'ล้มเหลว', ' ', 'การประมวลผลข้อมูล', 'แบบ', 'รายการ', ' ', 'และ', 'ปัญหา', 'และ', 'วิธี', 'แก้ปัญหา', 'การ', 'เข้าถึง', ' ', 'ข้อมูล', 'พร้อมกัน', ' ', 'รวมถึง', 'การ', 'ปรับ', 'ระบบ', 'ให้', 'มีประสิทธิภาพ', 'การใช้งาน', 'ดี', 'ที่สุด

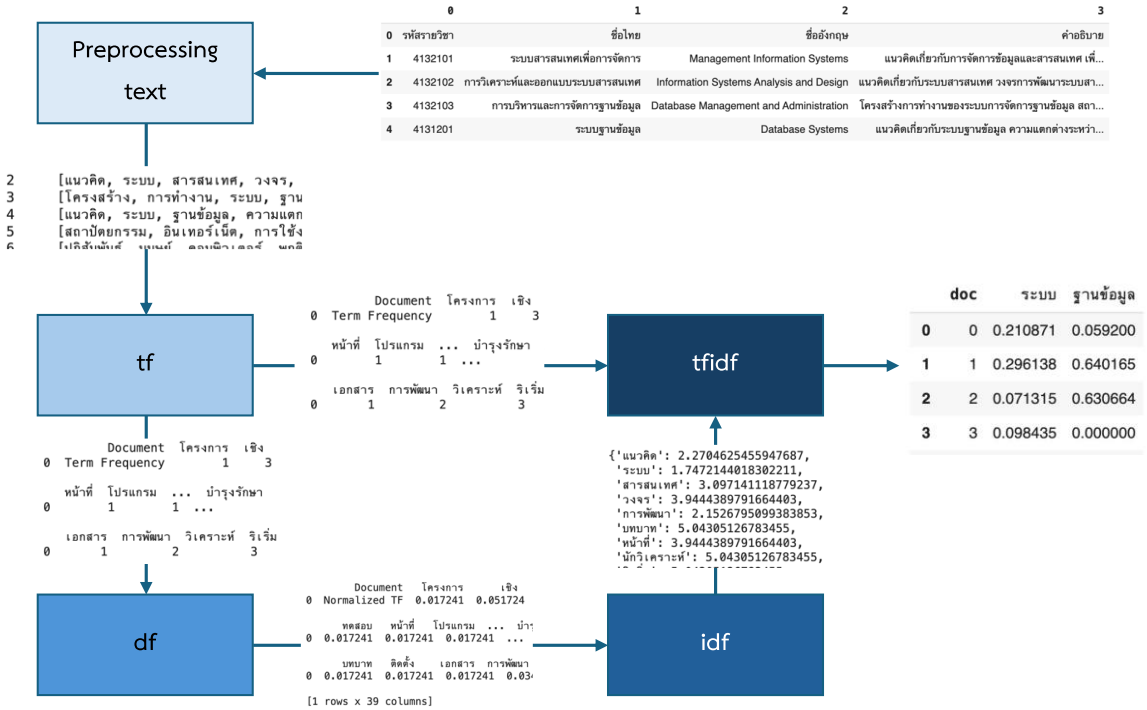
2.2 การกำจัดคำหยุด (Stop-Word Elimination) เป็นการนำคำที่ไม่เกี่ยวข้องกับเอกสารหรือคำที่ปรากฏบ่อยครั้งในเอกสารแต่ไม่ได้บ่งบอกคุณลักษณะของเอกสารนั้น เป็นคำขยายให้แก่คำอื่น เช่น คำเชื่อม คำสันธาน คำบุพบท หรือคำที่มักจุถูกใช้บ่อย ๆ เช่น “การ” “ความ” “ที่” “จะ” “ใช้” เป็นต้น คำเหล่านี้จะถูกตัดทิ้งโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนไป ดังตัวอย่าง

['โครงสร้าง', 'การทำงาน', '~~ของ~~', 'ระบบ', '~~การ~~', '~~จัดการ~~', 'ฐานข้อมูล', ' ', 'สถาปัตยกรรม', 'ระบบ', ' ', 'ฐานข้อมูล', ' ', 'ระบบ', 'ฐานข้อมูล', 'เชิง', 'สัมพันธ์', ' ', 'ระบบ', 'ฐานข้อมูล', '~~แบบ~~', 'กระจาย', ' ', 'ระบบ', 'ฐานข้อมูล', ' ', '~~แบบ~~', 'อ็อบเจ็กต์', ' ', 'ระบบ', 'ฐานข้อมูล', '~~บน~~', 'เว็บ', ' ', '~~การ~~', 'สร้าง', 'ระบบ', 'ฐานข้อมูล', ' ', 'การ', 'สร้าง', 'พจนานุกรม', ' ', 'ฐานข้อมูล', ' ', '~~การ~~', '~~จัดการ~~', 'ดัชนี', 'ข้อมูล', ' ', '~~การ~~', '~~จัดการ~~', 'บูรณภาพ', '~~ของ~~', 'ข้อมูล', ' ', '~~การ~~', '~~จัดการ~~', 'ความมั่นคง', '~~ของ~~', 'ฐานข้อมูล', ' ', 'การกำ', '~~หน~~', 'ด', 'ผู้ใช้', ' ', 'การ', 'จัดการ', 'สิทธิ์', 'ใน', 'การใช้งาน', 'ฐานข้อมูล', 'ระบบ', 'ความปลอดภัย', ' ', '~~การ~~', 'ถ่ายโอน', 'ข้อมูล', '~~ระหว่าง~~', 'ฐานข้อมูล', ' ', 'การกำ', 'หน', 'ดก', 'ล', 'ยุทธ์', 'ใน', 'การสำรอง', 'ข้อมูล', ' ', 'การกู้', 'คืน', '~~เมื่อ~~', 'ระบบ', 'ล้มเหลว', ' ', 'การประมวลผลข้อมูล', '~~แบบ~~', 'รายการ', ' ', '~~และ~~', 'ปัญหา', '~~และ~~', 'วิธี', 'แก้ปัญหา', '~~การ~~', 'เข้าถึง', ' ', 'ข้อมูล', '~~พร้อมกัน~~', ' ', '~~รวมถึง~~', '~~การ~~', '~~ปรับ~~', 'ระบบ', '~~ให้~~', 'มีประสิทธิภาพ', 'การใช้งาน', 'ดี', '~~ที่สุด~~']

2.3 การหารากศัพท์ (Stemming) เป็นการหา Stem หมายถึงรากศัพท์ คำหลัก หรือต้นตอของศัพท์ เพื่อให้การอธิบายเป็นไปในทางเดียวกัน ซึ่งจะได้ผลลัพธ์การเตรียมข้อมูลดังนี้

['โครงสร้าง', 'การทำงาน', 'ระบบ', 'ฐานข้อมูล', ' ', 'สถาปัตยกรรม', 'ระบบ', ' ', 'ฐานข้อมูล', ' ', 'ระบบ', 'ฐานข้อมูล', 'เชิง', 'สัมพันธ์', ' ', 'ระบบ', 'ฐานข้อมูล', 'กระจาย', ' ', 'ระบบ', 'ฐานข้อมูล', ' ', 'อ็อบเจ็กต์', ' ', 'ระบบ', 'ฐานข้อมูล', 'เว็บ', ' ', 'สร้าง', 'ระบบ', 'ฐานข้อมูล', ' ', 'สร้าง', 'พจนานุกรม', ' ', 'ฐานข้อมูล', ' ', 'ดัชนี', 'ข้อมูล', ' ', 'บูรณภาพ',

3. การหาค่าน้ำหนักของคำ (Term Weighting) การหาค่าน้ำหนักของคำหรือเทอม (Term) เพื่อสกัดคุณลักษณะของเอกสาร ด้วยการคำนวณหา TF-IDF แปลงให้อยู่ในรูปแบบของเวกเตอร์ที่แสดงขั้นตอนได้ตามภาพที่ 2 ทั้งนี้ผู้วิจัยได้ทดสอบพร้อมกำหนดค่าขั้นต่ำ Document Frequency ของ Term อยู่ระหว่าง 2-21 จึงจะทำให้ได้ค่าที่เหมาะสม และสามารถแสดงตัวอย่างค่าน้ำหนักของคำอธิบายวิชาที่คำนวณได้ดังตารางที่ 1



ภาพที่ 2 แสดงขั้นตอนการหาค่าน้ำหนักของคำ เพื่อแปลงให้อยู่ในรูปของเวกเตอร์

ตารางที่ 1 ตัวอย่างค่าน้ำหนักของการคำนวณหา TF-IDF ของคำอธิบายรายวิชา

รายวิชา	แนวคิด	ข้อมูล	สารสนเทศ	ระบบ	การพัฒนา
4132101	18	1	10	7	0
4132102	18	2	-	8	0
4132103	18	2	-	8	-
SC312003	-	2	-	9	-

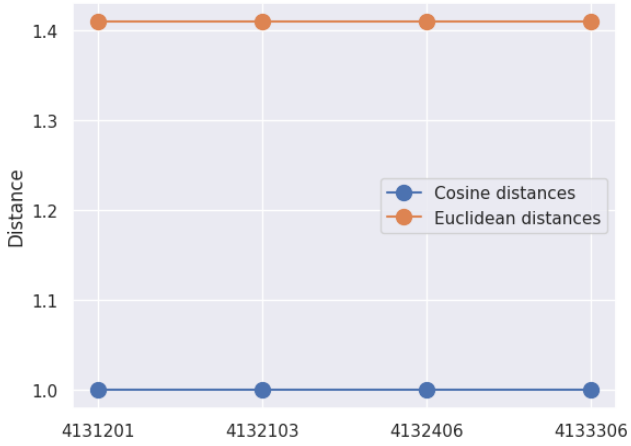
4. การวัดความเหมือนของข้อความ (Text Similarity Measurement)

หลังจากคำนวณค่าน้ำหนักของคำในเอกสารด้วยวิธี TF และTF-IDF แล้ว ผู้วิจัยได้ทำการทดสอบหาประสิทธิภาพการวัดความเหมือนของข้อความด้วยเมตริกแบบคู่ 7 วิธี ได้แก่ Cosine Similarity, Cosine distances, Euclidean distances, Laplacian kernel, Linear kernel, Pairwise kernels และ rbf kernel โดยคำนวณในแต่ละคู่เวกเตอร์ของคำอธิบายรายวิชาที่ขอเทียบโอน การทดลองนี้ใช้รายวิชา sc12003 เปรียบเทียบความเหมือนกับอีก 4 รายวิชาได้แก่ 4131201, 4132103, 4134206, และ 4133306 ประกอบด้วยขั้นตอนในการวัดความเหมือนของข้อความ ดังนี้

4.1 ผลการวัดความเหมือนของข้อความด้วยการหาระยะทาง

ในการทดลองนี้เป็นการวัดความเหมือนของข้อความด้วยการหาระยะทางจาก 2 วิธี ระยะห่างโคไซน์ (Cosine distances) และระยะห่างยูคลิเดียน (Euclidean distances) โดยวิธีระยะห่างโคไซน์

เน้นการหาค่าจากเวกเตอร์สองเวกเตอร์มีทิศทางที่ตรงกันหรือมุมระหว่างทั้งสองเวกเตอร์เป็นมุมมุ่งอยู่ในทิศทางเดียวกัน ส่วนระยะห่างยูคลิเดียนเน้นพิจารณาทั้งทิศทางและความยาวของเวกเตอร์ ดังนั้น เมื่อเทียบรายวิชา sc12003 กับอีก 4 รายวิชา 4131201, 4132103, 4134206, และ 4133306 พบว่า วิธีระยะห่างยูคลิเดียน มีความห่างจากกันและมีความยาวของเวกเตอร์เท่ากัน คือ 1.4 หน่วย และส่วนวิธีระยะห่างโคไซน์ เมื่อเทียบรายวิชา sc12003 กับอีก 4 รายวิชา 4131201, 4132103, 4134206, และ 4133306 พบว่า เวกเตอร์สองเวกเตอร์มีทิศทางที่ตรงกันเท่ากับ 1.0 เท่ากันทั้ง 4 รายวิชา ดังภาพที่ 3



ภาพที่ 3 ค่าความเหมือนของข้อความด้วยการหาระยะทางด้วยวิธีระยะห่างโคไซน์ และระยะห่างยูคลิเดียน

4.2 ผลการวัดความเหมือนของข้อความด้วยการหาค่าโคไซน์และเคอร์นอล

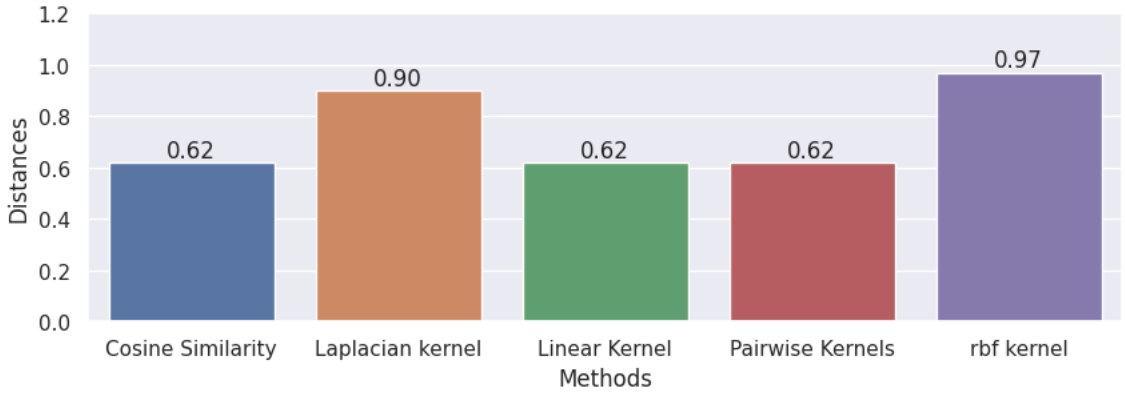
ในการทดลองนี้เป็นการวัดความเหมือนของข้อความด้วยการหาค่าโคไซน์และเคอร์นอล 5 วิธีได้แก่ Cosine Similarity, Laplacian kernel, Linear Kernel, Pairwise Kernels, และ rbf kernel ตารางที่ 2 ค่าความเหมือนของข้อความระหว่าง sc12003 กับ 4 รายวิชา 4131201, 4132103, 4134206, และ 4133306 ในแต่ละวิธี พบว่า 1) ค่าความเหมือนของข้อความด้วย Cosine Similarity ที่ดีที่สุดเท่ากับ 0.82 จากการเทียบความเหมือนระหว่าง sc12003 กับ 4131201 2) ค่าความเหมือนของข้อความด้วย Laplacian kernel ที่ดีที่สุดเท่ากับ 0.91 เท่ากัน 2 รายวิชาจากการเทียบความเหมือนระหว่าง sc12003 กับ 4131201 และ 4132103 3) ค่าความเหมือนของข้อความด้วย Linear Kernel ที่ดีที่สุดเท่ากับ 0.82 จากการเทียบความเหมือนระหว่าง sc12003 กับ 4131201 4) ค่าความเหมือนของข้อความด้วย Pairwise Kernels ที่ดีที่สุดเท่ากับ 0.82 จากการเทียบความเหมือนระหว่าง sc12003 กับ 4131201 และ 5) ค่าความเหมือนของข้อความด้วย rbf kernel ที่ดีที่สุดเท่ากับ 0.98 เท่ากัน 2 รายวิชาจากการเทียบความเหมือนระหว่าง sc12003 กับ 4131201 และ 4132103

ภาพที่ 4 พบว่า ค่าความเหมือนของข้อความระหว่าง sc12003 กับ 4 รายวิชา 4131201, 4132103, 4134206, และ 4133306 วิธีที่มีค่าความเหมือนเฉลี่ยอยู่ในระดับกลางได้แก่วิธี

Cosine Similarity, Linear Kernel และ Pairwise Kernel โดยมีค่าความเหมือนเท่ากับ 0.62 ทั้ง 3 วิธี และวิธีที่มีค่าความเหมือนเฉลี่ยอยู่ในระดับสูงได้แก่วิธี Laplacian Kernel และ RDF Kernel โดยมีค่าความเหมือนเท่ากับ 0.90 และ 0.97 ตามลำดับ

ตารางที่ 2 ตัวอย่างผลการคำนวณค่าความเหมือนของรายวิชา

รายวิชาที่นำมาเทียบ	รายวิชาที่ถูกเทียบ	วิธีการวัดความเหมือนของข้อความ (Text Similarity Measurement)				
		Cosine Similarity	Laplacian kernel	Linear Kernel	Pairwise Kernels	rbf kernel
sc312003	4131201	0.82	0.91	0.82	0.82	**0.98**
	4132103	0.76	0.91	0.76	0.76	**0.98**
	4134206	0.54	0.90	0.54	0.54	0.96
	4133306	0.34	0.89	0.34	0.34	0.94



ภาพที่ 4 ค่าความเหมือนของข้อความเฉลี่ยด้วยการหาค่าโคไซน์และเคอร์นอล

สรุปผลและข้อเสนอแนะ

จากการวิเคราะห์และเปรียบเทียบประสิทธิภาพวิธีหาความเหมือนด้วยเมตริกแบบคู่ (Pairwise Metrics) ในการเทียบโอนคำอธิบายรายวิชาของสาขาวิชาเทคโนโลยีสารสนเทศ ซึ่งมีขั้นตอนการทำงานในการเตรียมข้อมูลโดยการตัดคำภาษาไทย การกำจัดคำหยุด การคำนวณน้ำหนักของคำเพื่อสกัดคุณลักษณะของข้อความแปลงให้อยู่ในรูปเวกเตอร์และคำนวณความเหมือนเพื่อเปรียบเทียบประสิทธิภาพด้วยวิธีเมตริกแบบคู่ 7 วิธี การหาค่าความเหมือนของข้อความด้วยการหาระยะทางด้วยวิธีระยะห่างโคไซน์มีค่าต่ำกว่าวิธีระยะห่างยูคลิเดียน และผลการวัดความเหมือนของข้อความด้วยการหาค่าโคไซน์และเคอร์นอล พบว่า วิธีที่มีประสิทธิภาพดีที่สุดโดยเฉลี่ย ได้แก่ rbf kernel 0.97 ลำดับถัดมาเป็นวิธี Laplacian kernel 0.90 ส่วนวิธี Cosine Similarity, Linear Kernel และ Pairwise Kernels ผลการวัดความเหมือน

ของข้อความโดยเฉลี่ยเท่ากันคือ 0.62 ดังนั้น งานวิจัยนี้สามารถนำมาใช้ในการเพิ่มประสิทธิภาพความแม่นยำในการเทียบโอนหน่วยกิตได้ ซึ่งสอดคล้องกับงานวิจัยของ สราธิป เตชาวิวัฒนบูลย์ (2560) อย่างไรก็ตามวิธีหาความเหมือนของข้อความด้วยอัลกอริทึมแบบต่าง ๆ กันนั้น การเตรียมข้อมูลเป็นสิ่งสำคัญ ควรทำความสะอาดข้อมูลและวิเคราะห์อย่างละเอียดก่อนนำมาใช้งาน

เอกสารอ้างอิง

จักรินทร์ สันติรัตนภักดี และศศิธร อิ่มวุฒิ. (2564). การออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะ ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมเพื่อจำแนกปัญหาการให้บริการ. การประชุมวิชาการเสนอผลงานวิจัยระดับชาติด้านวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏจันทรเกษม ครั้งที่ 4, 1(1), 63.

ชาลิสา จิตบุญญาพินิจ, ปราลี มณีรัตน์, และนิเวศ จิระวิชิตชัย. (2565). การพัฒนาแบบจำลองการวิเคราะห์ความรู้สึกบนสื่อสังคมออนไลน์ไทยโดยใช้เทคนิคการเรียนรู้เชิงลึก. วารสารวิทยาศาสตร์และเทคโนโลยี หัวเฉียวเฉลิมพระเกียรติ, 8(2), 1–11.

พีรพัฒ โชคสุวัฒนสกุล, ศดานันท์ อาศัยบุญ, และอรรถพล ธำรงรัตนฤทธิ์. (2566). การประยุกต์ใช้วิทยาศาสตร์ข้อมูลในการวิเคราะห์ข้อมูลจากการวิเคราะห์ผลกระทบและประเมินผลสัมฤทธิ์ของกฎหมาย. วารสารนิติศาสตร์ มหาวิทยาลัยธรรมศาสตร์, 52(2), 247–273.

สราธิป เตชาวิวัฒนบูลย์. (2560). การเปรียบเทียบวิธีวัดความคล้ายคลึงของข้อความบนฐานเวกเตอร์เพื่อการเทียบโอนหน่วยกิต ของรายวิชาระดับมหาวิทยาลัย. The 12th National Conference and 2017 International Conference on Applied Computer Technology and Information Systems and 2017 National Conference On Business Administration, 118–124.

สุขุมาลย์ หนกหลัง. (2563). Text mining in practice with R. วารสารวิธีวิทยาการวิจัย, 32(3).

Al-Obaydy, W. N. I., Hashim, H. A., Najm, Y. A., & Jalal, A. A. (2022). Document classification using term frequency-inverse document frequency and K-means clustering. Indonesian Journal of Electrical Engineering and Computer Science, 27(3), 1517–1524.

Al-Ramahi, M. A., & Mustafa, S. H. (2012). N-gram-based techniques for arabic text document matching; case study: courses accreditation. Abhath Al-Yarmouk. Basic Sciences and Engineering, 21(1), 85–105.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. de O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., & Brockman, G. (2021). Evaluating large language models trained on code. ArXiv Preprint ArXiv:2107.03374.

Christopher D. Manning, Prabhakar Raghavan, & Hinrich Schütze. (2008). Introduction to Information Retrieval. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). A Comparison of Semantic Similarity Methods for Maximum Human Interpretability. 2019 Artificial Intelligence for Transforming Business and Society (AITB), 1–4. https://doi.org/10.1109/AITB48515.2019.8947433

Viriyavisuthisakul, S., Sanguansat, P., Charnkeitkong, P., & Haruechaiyasak, C. (2015). A comparison of similarity measures for online social media Thai text classification. 2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 1–6.

Zoupanos, S., Kolovos, S., Kanavos, A., Papadimitriou, O., & Maragoudakis, M. (2022, September 7). Efficient comparison of sentence embeddings. ACM International Conference Proceeding Series. https://doi.org/10.1145/3549737.3549752