

การสืบค้นเชิงความหมายด้วยเวกเตอร์สำหรับฐานข้อมูลโครงการมหาวิทยาลัย โดยใช้
Transformer Embeddings และการค้นหา Approximate Nearest Neighbor
A Semantic Vector-Based Retrieval Framework for University Project Repositories
Using Transformer Embeddings and Approximate Nearest Neighbor Search

สุรพล ชุ่มกลิ่น¹, ภาณุวัฒน์ ชันจา^{2*} และ มานิตย์ พ่วงบางโพ³

Suraphon Chumklin¹, Phanuwat Khanja^{2*} and Manit Puangbangpo³

คณะวิทยาศาสตร์และเทคโนโลยี¹ และ คณะเทคโนโลยีอุตสาหกรรม^{2,3} มหาวิทยาลัยราชภัฏอุดรดิตถ์

Faculty of Science and Technology¹, and Faculty of Industrial Technology^{2,3}, Uttaradit Rajabhat University

E-mail : suraphon@uru.ac.th¹, phanuwat.kha@uru.ac.th², manit.phu@uru.ac.th³

(Received : January 16, 2026; Revised : May 12, 2026; Accepted : May 15, 2026)

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) พัฒนาระบบการสืบค้นเชิงความหมายบนพื้นฐานเวกเตอร์สำหรับคลังข้อมูลโครงการของมหาวิทยาลัย 2) ศึกษาผลการทดลองใช้ระบบสืบค้นเชิงความหมายที่ผสานโมเดล Sentence Transformer ร่วมกับอัลกอริทึม Approximate Nearest Neighbor และ 3) ศึกษาผลการประเมินประสิทธิภาพและคุณภาพการจับคู่เชิงความหมายของระบบในบริบทข้อมูลโครงการของมหาวิทยาลัย การดำเนินการวิจัยนี้ประกอบด้วยรวบรวมข้อมูลกลุ่มตัวอย่าง 937 โครงการ การออกแบบโครงสร้างข้อมูล การพัฒนาแบบจำลองและการทดสอบประสิทธิภาพของระบบ เครื่องมือที่ใช้ในการวิจัย ได้แก่ ฐานข้อมูลซอฟต์แวร์ประมวลผลข้อมูล และกรอบการประเมินคุณภาพการสืบค้นเชิงความหมายเพื่อใช้วิเคราะห์ประสิทธิภาพของระบบ

ผลการวิจัยพบว่า 1) กรอบการสืบค้นเชิงความหมายที่พัฒนาขึ้นสามารถค้นคืนโครงการที่มีความเกี่ยวข้องเชิงบริบทได้อย่างมีประสิทธิภาพ โดยมีเวลาตอบสนองอยู่ในระดับที่เหมาะสมสำหรับระบบสืบค้นเชิงโต้ตอบ และสามารถแยกแยะโครงการที่เกี่ยวข้องสูงออกจากโครงการที่เกี่ยวข้องต่ำได้อย่างชัดเจน 2) การศึกษาผลการทดลองใช้ระบบสืบค้นเชิงความหมาย พบว่าการใช้ Sentence Transformer ร่วมกับดัชนี ANN (Annoy) ช่วยเพิ่มความแม่นยำของผลลัพธ์เมื่อเทียบกับการสืบค้นแบบใช้คำสำคัญ โดยเฉพาะในข้อมูลภาษาไทยที่มีความหลากหลายของถ้อยคำและเมทาเดตาไม่สมบูรณ์ และ 3) ผลการประเมินประสิทธิภาพระบบ พบว่าค่าตัวชี้วัด Response Level อยู่ในระดับดี สะท้อนความสามารถของระบบในการตอบสนองต่อเจตนาของผู้ใช้ และแสดงให้เห็นศักยภาพของการประยุกต์ใช้การสืบค้นเชิงความหมายเพื่อสนับสนุนการจัดการองค์ความรู้และการตัดสินใจเชิงข้อมูลในสถาบันอุดมศึกษา

คำสำคัญ : การสืบค้นเชิงความหมาย, เซนเทนซ์ทรานส์ฟอร์เมอร์, การค้นหาค่าใกล้เคียงแบบประมาณค่า, คลังโครงการของมหาวิทยาลัย, การประมวลผลภาษาธรรมชาติภาษาไทย

ABSTRACT

This research aimed to: 1) develop a vector-based semantic search framework for a university project repository; 2) study the results of implementing a semantic search system integrating Sentence Transformer models with Approximate Nearest Neighbor (ANN) algorithms; and 3) examine the results of evaluating the system's performance and semantic matching quality in the context of university project data. The research process consisted of collecting data from 937 projects, designing the data structure, developing the model, and testing the system performance.

The research instruments included a database system, data processing software, and a semantic search quality evaluation framework used to analyze the system performance.

The research findings revealed that: 1) the developed semantic search framework could effectively retrieve contextually relevant projects, with response times suitable for interactive search systems, and could clearly distinguish highly relevant projects from less relevant ones; 2) the results of implementing the semantic search system showed that the use of Sentence Transformer models together with an ANN index (Annoy) improved the accuracy of retrieval results compared with keyword-based search, particularly for Thai-language data with lexical diversity and incomplete metadata; and 3) the system performance evaluation showed that the Response Level metric was at a good level, reflecting the system's ability to respond to user intent and demonstrating the potential of applying semantic search to support knowledge management and data-driven decision-making in higher education institutions.

Keywords : Semantic Search, Sentence Transformers, Approximate Nearest Neighbor, University Project Repository, Thai NLP

บทนำ

ปัจจุบันสถาบันอุดมศึกษามีการพัฒนาคลังข้อมูลดิจิทัลเพื่อจัดเก็บและเผยแพร่ข้อมูลงานวิจัยและโครงการในการกิจต่าง ๆ ซึ่งมีบทบาทสำคัญต่อการบริหารจัดการองค์ความรู้ การติดตามผลการดำเนินงาน และการสนับสนุนการตัดสินใจเชิงนโยบายในระดับองค์กร มหาวิทยาลัยราชภัฏอุดรดิตถ์ได้บูรณาการข้อมูลโครงการหลายประเภท เช่น โครงการ U2T งานวิจัย โครงการตามเป้าหมาย SDGs และโครงการพัฒนาชุมชน เข้าสู่คลังข้อมูลกลางของสถาบันผ่านระบบสารสนเทศงานวิจัยและพันธกิจสัมพันธ์ [1] อย่างไรก็ตาม ระบบสืบค้นส่วนใหญ่ยังคงใช้การค้นหาแบบอิงคำสำคัญ ซึ่งมีข้อจำกัดในการจัดการข้อมูลภาษาไทยที่มีความหลากหลายด้านคำศัพท์และรูปแบบการเขียน ส่งผลให้ข้อมูลที่มีความหมายใกล้เคียงกันไม่สามารถถูกค้นคืนได้อย่างครบถ้วน [2]-[4]

ความก้าวหน้าด้านการประมวลผลภาษาธรรมชาติและการเรียนรู้เชิงลึก โดยเฉพาะแนวคิดการสืบค้นเชิงความหมายและโมเดล Transformer เช่น BERT และ Sentence-BERT ช่วยให้สามารถแทนข้อความในรูปเวกเตอร์เชิงความหมายและจับคู่ข้อมูลตามบริบทได้อย่างมีประสิทธิภาพ [5], [6] นอกจากนี้ อัลกอริทึม Approximate Nearest Neighbor เช่น FAISS และ Annoy ยังช่วยเพิ่มความรวดเร็วในการค้นหาในข้อมูลขนาดใหญ่ [7], [8] แม้งานวิจัยด้านนี้จะมีการประยุกต์ใช้อย่างแพร่หลายในต่างประเทศ แต่เมื่อพิจารณาในบริบทของคลังข้อมูลโครงการและสารสนเทศงานวิจัยของมหาวิทยาลัยไทย ยังพบช่องว่างเชิงวิชาการที่สำคัญ กล่าวคือ งานวิจัยส่วนใหญ่ยังมุ่งเน้นการศึกษาบนชุดข้อมูลมาตรฐานหรือข้อมูลภาษาอังกฤษ และขาดการพัฒนาในบริบทข้อมูลจริงที่มีลักษณะเฉพาะ เช่น ความหลากหลายของถ้อยคำภาษาไทยและความไม่สมบูรณ์ของเมทาดาทา นอกจากนี้ ยังมีข้อจำกัดในด้านการบูรณาการเทคนิคให้เป็นกรอบระบบสืบค้นเชิงความหมายแบบครบวงจร รวมถึงการประเมินผลที่สะท้อนการใช้งานจริงในระดับองค์กร [9]-[11] ดังนั้น งานวิจัยนี้จึงมุ่งพัฒนากรอบการสืบค้นเชิงความหมายบนฐานเวกเตอร์สำหรับคลังข้อมูลโครงการของมหาวิทยาลัยราชภัฏอุดรดิตถ์ โดยผสาน Sentence Transformer กับอัลกอริทึม ANN และการจัดอันดับด้วย Cosine Similarity เพื่อเพิ่มประสิทธิภาพการค้นคืนโครงการที่มีความหมายใกล้เคียงกัน แม้ใช้ถ้อยคำต่างกัน แนวทางดังกล่าวมีศักยภาพในการสนับสนุนการค้นหาโครงการข้ามคณะและข้ามหมวด SDGs ตลอดจนเสริมสร้างการจัดการองค์ความรู้และการวิเคราะห์ข้อมูลเชิงสถาบันในระยะยาว

1. วัตถุประสงค์การวิจัย

- 1.1 พัฒนารอบการสืบค้นเชิงความหมายบนพื้นฐานเวกเตอร์สำหรับคลังข้อมูลโครงการของมหาวิทยาลัย
- 1.2 ศึกษาผลการทดลองใช้ระบบสืบค้นเชิงความหมายที่ฐานโมเดล Sentence Transformer ร่วมกับอัลกอริทึม Approximate Nearest Neighbor
- 1.3 ศึกษาผลการประเมินประสิทธิภาพและคุณภาพการจับคู่เชิงความหมายของระบบในบริบทข้อมูลโครงการของมหาวิทยาลัย

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 Semantic Search เป็นแนวทางการค้นคืนสารสนเทศที่มุ่งเน้นการทำความเข้าใจความหมายและบริบทของคำค้นและเอกสาร แทนการจับคู่คำแบบตรงตัว ซึ่งช่วยลดข้อจำกัดจากความหลากหลายของภาษา โครงสร้างประโยค และคำพ้องความหมาย โดยเฉพาะในคลังข้อมูลขนาดใหญ่ที่มีคำอธิบายไม่เป็นมาตรฐาน [3] แนวคิดหลักคือการแทนข้อความด้วยเวกเตอร์เชิงความหมาย (Semantic Embeddings) เพื่อสะท้อนความสัมพันธ์เชิงแนวคิดระหว่างข้อความ ทำให้สามารถค้นคืนเอกสารที่มีความหมายใกล้เคียงได้แม้ใช้ถ้อยคำแตกต่างกัน [4] การพัฒนาเทคนิคนี้ได้รับแรงสนับสนุนจากความก้าวหน้าด้าน NLP และโมเดลเชิงลึกกลุ่ม Transformer ซึ่งช่วยเพิ่มความแม่นยำของผลลัพธ์การค้นคืนอย่างมีนัยสำคัญเมื่อเทียบกับวิธี Keyword-based Search ในข้อมูลจริงที่มีความไม่สม่ำเสมอ [5], [6], [12]

2.2 Sentence Transformer Embeddings เป็นสถาปัตยกรรมโครงข่ายประสาทเชิงลึกที่พัฒนาต่อยอดจากโมเดล Transformer โดยเฉพาะ BERT เพื่อสร้างตัวแทนเชิงเวกเตอร์ระดับประโยค ซึ่งสะท้อนความหมายเชิงบริบทของข้อความและสามารถนำไปเปรียบเทียบความคล้ายคลึงกันของข้อความจำนวนมากได้อย่างมีประสิทธิภาพ โมเดลดังกล่าวใช้โครงสร้างแบบ Siamese หรือ Triplet Network ทำให้สามารถคำนวณค่าความคล้ายคลึงด้วย Cosine Similarity ได้โดยตรง [6] ลดภาระการคำนวณเมื่อประยุกต์ใช้กับชุดข้อมูลขนาดใหญ่ ด้วยคุณสมบัติในการแปลงข้อความเป็นเวกเตอร์หลายมิติที่รักษาความสัมพันธ์เชิงความหมาย Sentence Transformers จึงถูกนำไปใช้อย่างแพร่หลายในงานด้านการจัดกลุ่มเอกสาร ระบบถาม-ตอบเชิงความหมาย การวัดความคล้ายคลึงของข้อความ และระบบแนะนำข้อมูล [13], [14] ซึ่งสะท้อนถึงความยืดหยุ่นและประสิทธิภาพของโมเดลในการจัดการข้อมูลข้อความเชิงความหมาย

2.3 Approximate Nearest Neighbor (ANN) เป็นเทคนิคการค้นหาเวกเตอร์ในชุดข้อมูลขนาดใหญ่และมีมิติสูงที่มุ่งเพิ่มประสิทธิภาพด้านเวลาและการใช้หน่วยความจำ เมื่อเทียบกับการค้นหาแบบตรงตัว (Exact Nearest Neighbor) โดยยอมลดความแม่นยำลงเล็กน้อยเพื่อแลกกับความเร็วในการตอบสนอง จึงเหมาะสำหรับระบบขนาดใหญ่และระบบเรียลไทม์ [15] ANN อาศัยโครงสร้างข้อมูลและกลไกการจัดดัชนี เช่น random projection, multiple randomized trees หรือโครงสร้างแบบกราฟ เพื่อลดพื้นที่ค้นหาและจำกัดการสำรวจเฉพาะเวกเตอร์ที่มีแนวโน้มใกล้เคียง ส่งผลให้เวลาในการประมวลผลลดลงอย่างมีนัยสำคัญ ด้วยเหตุนี้ ANN จึงถูกนำไปใช้อย่างแพร่หลายในงานค้นหาเชิงความหมาย ระบบแนะนำ และงานประมวลผลเวกเตอร์จากโมเดลการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกบนข้อมูลขนาดใหญ่

2.4 งานวิจัยที่เกี่ยวข้อง งานวิจัยในช่วงทศวรรษที่ผ่านมาแสดงให้เห็นว่าการผสมผสานเทคนิค Semantic Embeddings ร่วมกับอัลกอริทึม Approximate Nearest Neighbor (ANN) ได้กลายเป็นกรอบหลักของระบบสืบค้นสารสนเทศสมัยใหม่ โดยเฉพาะในบริบทของข้อมูลขนาดใหญ่และข้อมูลที่มีความซับซ้อนเชิงความหมาย [8], [15] งานของ Reimers และ Gurevych [6] ได้นำเสนอ Sentence-BERT ซึ่งพัฒนาจาก BERT เพื่อสร้าง Embeddings ระดับประโยคสำหรับงาน Semantic Similarity และ Semantic Search และถูกนำไปใช้อย่างแพร่หลายในหลายโดเมน งานด้าน Semantic Search และ Document Understanding รายงานว่าการใช้ Representation เชิงเวกเตอร์ช่วยเพิ่มความแม่นยำในการค้นคืนข้อมูลและลดข้อจำกัดของการสืบค้นแบบใช้คำสำคัญ [3], [17]

งานด้านอัลกอริทึม ANN ตั้งแต่ Indyk และ Motwani [15] จนถึงระบบสมัยใหม่ เช่น FAISS และ Annoy [7], [8] ยืนยันว่าเทคนิคเหล่านี้สามารถรักษาสมาดุลระหว่างความเร็วและความแม่นยำได้ดี เหมาะสำหรับระบบสืบค้นที่มีข้อมูลจำนวนมาก สำหรับบริบทภาษาไทย งานวิจัยได้พัฒนาชุดข้อมูลและกรอบการประเมิน Semantic Similarity [2] รวมถึงการประยุกต์ใช้ Semantic Search ร่วมกับ Ontology ในระบบสารสนเทศงานวิจัยและโดเมนเฉพาะหลายประเภท [9], [10] และงานล่าสุดของ Jearanaitanakij และคณะ [11] แสดงให้เห็นว่าการผสาน Sentence-BERT กับ FAISS สามารถเพิ่มความแม่นยำของระบบถาม-ตอบภาษาไทยได้อย่างมีนัยสำคัญ

จากการทบทวนงานวิจัยที่เกี่ยวข้องทั้งหมด สามารถสรุปได้ว่าการใช้ Semantic Embeddings ร่วมกับ ANN เป็นแนวทางที่ได้รับการยอมรับทั้งในเชิงทฤษฎีและเชิงประยุกต์ งานเหล่านี้จึงเป็นฐานความรู้สำคัญที่สนับสนุนความเหมาะสมของกรอบแนวคิดและสถาปัตยกรรมระบบที่นำเสนอในงานวิจัยนี้ โดยเฉพาะในบริบทของระบบสืบค้นข้อมูลเชิงความหมายภายในสถาบันอุดมศึกษาและข้อมูลภาษาไทย

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงพัฒนา (Research and Development) โดยมุ่งพัฒนาระบบสืบค้นเชิงความหมายสำหรับคลังข้อมูลโครงการมหาวิทยาลัย เครื่องมือการวิจัยประกอบด้วยเครื่องมือด้านซอฟต์แวร์ชุดข้อมูล และกรอบการประเมินผล ดังรายละเอียดต่อไปนี้

1.1 ชุดข้อมูลโครงการมหาวิทยาลัยราชภัฏอุดรดิตถ์ ได้แก่ งานวิจัย โครงการ U2T กิจกรรมบริการวิชาการ และโครงการยุทธศาสตร์ [1] ข้อมูลแต่ละระเบียบประกอบด้วยฟิลด์สำคัญ เช่น ชื่อโครงการ คำอธิบายโครงการ ประเภทโครงการ คณะ ปีดำเนินการ และหมวดเป้าหมายการพัฒนาที่ยั่งยืน (SDGs)

1.2 เครื่องมือสร้างตัวแทนเชิงความหมาย (Semantic Embedding Tool) ใช้ไลบรารี Sentence Transformers พร้อมโมเดล paraphrase-multilingual-MiniLM-L12-v2 สำหรับแปลงข้อความโครงการและข้อความค้นของผู้ใช้เป็นเวกเตอร์เชิงความหมายแบบหนาแน่น (Dense Embeddings) ขนาด 384 มิติ [6], [18]

1.3 เครื่องมือจัดทำดัชนีและสืบค้นแบบ Approximate Nearest Neighbor (ANN) ใช้ไลบรารี Annoy (Approximate Nearest Neighbors Oh Yeah) สำหรับจัดทำดัชนีเวกเตอร์เชิงความหมาย [8], [15] โดยตั้งค่าเมตริกระยะทางแบบ Angular Distance ซึ่งสอดคล้องกับ Cosine Similarity [4] และใช้จำนวนต้นไม้ (Number of Trees) เท่ากับ 10 Annoy

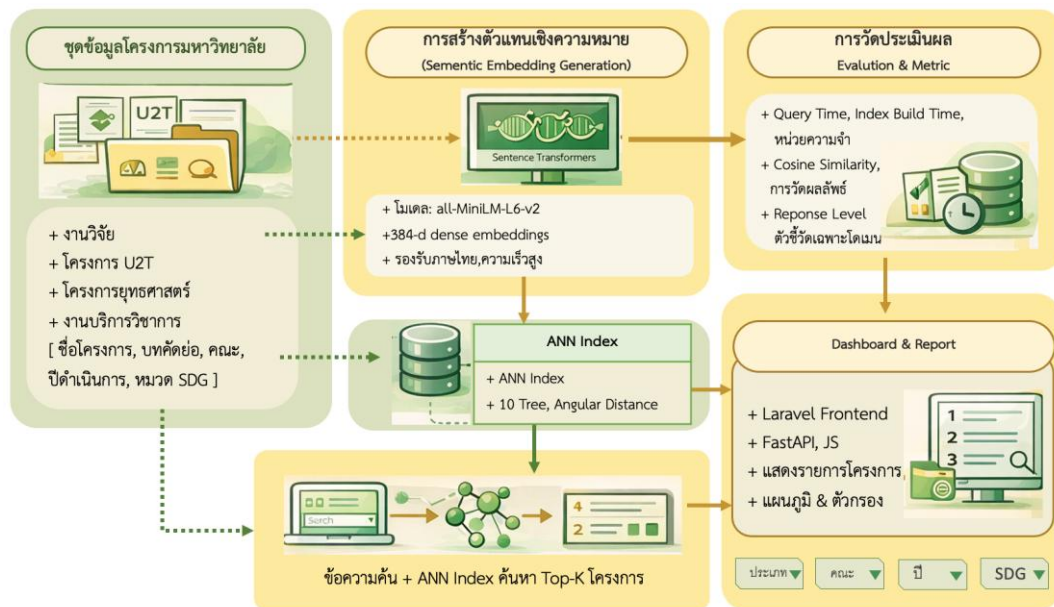
1.4 เครื่องมือประเมินคุณภาพการสืบค้นเชิงความหมาย ได้แก่ 1) ตัวชี้วัดด้านประสิทธิภาพการค้นหา เช่น Query Time, Index Build Time และการใช้หน่วยความจำ [3], [17] 2) ตัวชี้วัดด้านคุณภาพการจับคู่เชิงความหมาย (Cosine Similarity) [4] 3) ตัวชี้วัดเฉพาะโดเมน (Response Level) เพื่อประเมินสัดส่วนของผลลัพธ์ที่มีความเกี่ยวข้องกับความต้องการของผู้ใช้ในบริบทของข้อมูลเฉพาะด้าน [12], [13]

2. ประชากรและกลุ่มตัวอย่าง

2.1 ประชากรที่ใช้ในการวิจัย คือ ข้อมูลโครงการทั้งหมดที่จัดเก็บอยู่ในคลังข้อมูลกลางของมหาวิทยาลัยราชภัฏอุดรดิตถ์ จำนวน 937 รายการ โดยข้อมูลแต่ละระเบียบประกอบด้วยชื่อโครงการ คำอธิบายโครงการ ประเภทโครงการ คณะ/หน่วยงาน ปีดำเนินการ และหมวดเป้าหมายการพัฒนาที่ยั่งยืน (SDGs)

2.2 กลุ่มตัวอย่าง ได้แก่ ข้อมูลโครงการทั้งหมดที่มีอยู่ในคลังข้อมูลกลางของมหาวิทยาลัย ณ ช่วงเวลาที่ทำการศึกษา โดยใช้วิธีการคัดเลือกแบบทั้งประชากร (Total Sampling) นอกจากนี้ ในการประเมินเชิงผู้ใช้ ผู้วิจัยได้กำหนดกลุ่มตัวอย่างจำนวน 30 คน ประกอบด้วยอาจารย์ บุคลากรที่เกี่ยวข้องกับโครงการ และผู้ใช้งานระบบภายในมหาวิทยาลัย โดยใช้วิธีการคัดเลือกแบบเจาะจง (Purposive Sampling) เพื่อให้ได้ผู้ประเมินที่มีประสบการณ์หรือความเกี่ยวข้องกับข้อมูลโครงการและระบบสืบค้นเชิงความหมายโดยตรง

3. ขั้นตอนการดำเนินการวิจัย



ภาพที่ 1 แผนภาพกรอบแนวคิดในการวิจัย

จากภาพที่ 1 กรอบแนวคิดในการวิจัย ประกอบด้วย วิธีการเตรียมชุดข้อมูล กลไกการจัดทำดัชนีแบบ ANN กระบวนการสืบค้น และกรอบการวัดผลประเมินผล กระบวนการทั้งหมดออกแบบให้สามารถทำซ้ำได้ (Reproducible) และรองรับการประเมินทั้งเชิงปริมาณและเชิงคุณภาพ ซึ่งมีรายละเอียดดังนี้

3.1 ชุดข้อมูล (Dataset)

งานวิจัยใช้ข้อมูลของมหาวิทยาลัย ได้แก่ งานวิจัย โครงการ U2T กิจกรรมบริการวิชาการ และโครงการตามยุทธศาสตร์มหาวิทยาลัย โดยมีฟิลด์ข้อมูลทั้งเชิงข้อความและเชิงหมวดหมู่ซึ่งเป็นฐานสำคัญสำหรับการสร้างตัวแทนเชิงความหมาย (semantic embeddings) ตามหลักการในงาน NLP-IR [3], [18]

ตารางที่ 1 ฟิลด์ข้อมูลและคุณลักษณะของชุดข้อมูลโครงการ

ฟิลด์	คำอธิบาย	ตัวอย่างค่า	หมายเหตุ
title	ชื่อโครงการ	“พัฒนาชุมชนด้วยนวัตกรรมดิจิทัล”	มีทุกระเบียน
description	บทคัดย่อ/สรุปโครงการ	“ยกระดับคุณภาพชีวิตชุมชน...”	หลายระเบียนขาดข้อมูล
type	ประเภทโครงการ	“U2T”, “Research”, “Service”	ใช้ในการกรอง
faculty	หน่วยงานที่รับผิดชอบ	“คณะเทคโนโลยีอุตสาหกรรม”	—
year	ปีของโครงการ	2024	ใช้กรองผลลัพธ์
sdg	หมวด SDG	“SDG 4 – Quality Education”	ไม่ครบทุกระเบียน

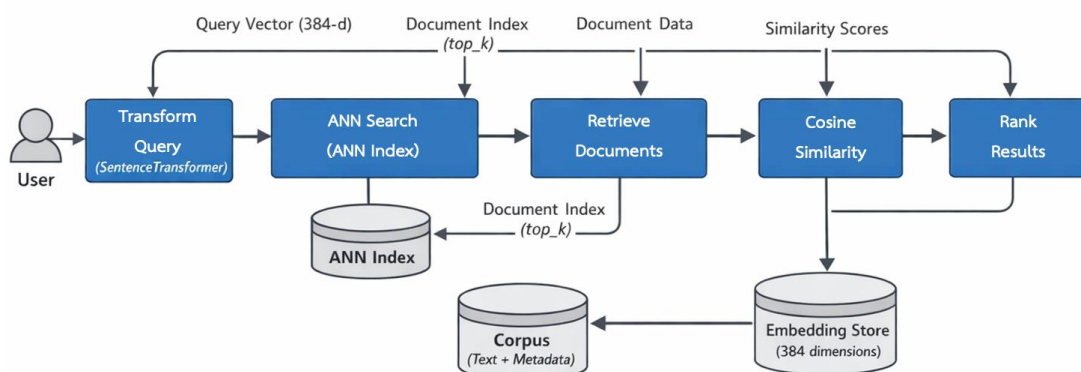
จากตารางที่ 1 แสดงฟิลด์ข้อมูลและคุณลักษณะของชุดข้อมูลโครงการ ตัวอย่างค่า และหมายเหตุเกี่ยวกับความสมบูรณ์ของข้อมูล ซึ่งสะท้อนความท้าทายของข้อมูลจริงที่มีความไม่ครบถ้วนของคำอธิบาย (Missing Descriptions) ลักษณะปัญหาที่พบทั่วไปในงานสืบค้นข้อความภาษาไทย [2]

3.2 ขั้นตอนการเตรียมข้อมูล (Data Preprocessing) การเตรียมข้อมูลถูกออกแบบเพื่อรองรับความหลากหลายของภาษาไทยและความไม่เป็นมาตรฐานของข้อมูล [3] โดยประกอบด้วยการปรับข้อความให้อยู่ในรูปแบบมาตรฐาน (Text Normalization) การรวมฟิลด์ title และ description เพื่อเพิ่มบริบทของข้อมูล การจัดการคำที่หายไปโดยใช้เฉพาะ title ในกรณีที่ไม่มีความอธิบายโครงการ และการตัดคำพร้อมสร้างตัวแทนเชิงความหมายด้วยโมเดล paraphrase-multilingual-MiniLM-L12-v2 ซึ่งรองรับภาษาไทยได้อย่างมีประสิทธิภาพ [2] ผลลัพธ์จะถูกแปลงเป็นเวกเตอร์เชิงความหมายสำหรับกระบวนการสืบค้นในขั้นตอนถัดไป [6]

3.3 การสร้างเวกเตอร์เชิงความหมาย (Semantic Embedding Generation) งานวิจัยนี้ใช้โมเดล SentenceTransformer โดยเลือกใช้รุ่น paraphrase-multilingual-MiniLM-L12-v2 เพื่อแปลงข้อความโครงการเป็นเวกเตอร์เชิงความหมายสำหรับคำนวณค่าความคล้ายเชิงมุมด้วย Cosine Similarity ในการทดลองนี้ใช้ชุดข้อมูลจำนวน 937 รายการ โดยแต่ละรายการถูกแทนด้วยเวกเตอร์ขนาด 384 มิติ ทำให้ได้เมทริกซ์ Embeddings ขนาด 937×384 ซึ่งจัดเก็บในรูปแบบ NumPy array เพื่อรองรับการคำนวณ Cosine Similarity และการสร้างดัชนีแบบ ANN ด้วย Annoy สำหรับการสืบค้นในขั้นตอนถัดไป [8], [15]

3.4 การสร้างดัชนี ANN (Annoy Index Construction) งานวิจัยนี้ใช้ไลบรารี Annoy สำหรับสร้างดัชนีค้นหาแบบ Approximate Nearest Neighbor (ANN) [8] โดยกำหนดมิติของเวกเตอร์เท่ากับ 384 และใช้ระยะเชิงมุมสอดคล้องกับ Cosine Similarity เป็นเกณฑ์วัดความใกล้เคียง [4], [6] จากนั้นนำเวกเตอร์ของเอกสารทั้งหมดเข้าสู่ดัชนีและสร้างโครงสร้างต้นไม้แบบสุ่มจำนวน 10 ต้น เพื่อรักษาสมดุลระหว่างความเร็วและความแม่นยำของผลลัพธ์ และการตั้งค่า Metric แบบ Angular เป็นแนวปฏิบัติมาตรฐาน [15], [18] ที่สามารถให้เวลาตอบสนองในระดับต่ำกว่า 200 มิลลิวินาทีอย่างสม่ำเสมอ จึงเหมาะสมสำหรับระบบสืบค้นเชิงความหมายแบบโต้ตอบในบริบทคลังข้อมูลโครงการมหาวิทยาลัย [1], [9]-[11]

3.5 พังค์ชันสืบค้นเชิงความหมาย (Semantic Search Pipeline) เป็นกลไกหลักของระบบ ทำหน้าที่แปลงคำค้นของผู้ใช้เป็นเวกเตอร์แทนความหมายด้วยโมเดล SentenceTransformer [6] และใช้โครงสร้าง Approximate Nearest Neighbor (ANN) เพื่อค้นหาเอกสารที่มีความใกล้เคียงเชิงความหมายในเบื้องต้น [7] จากนั้นระบบจะคำนวณค่า Cosine Similarity ระหว่างคำค้นและเอกสารที่ค้นพบ เพื่อจัดอันดับผลลัพธ์ตามระดับความเกี่ยวข้องเชิงความหมายก่อนส่งกลับให้ผู้ใช้ [3], [4]



ภาพที่ 2 แผนภาพการไหลของข้อมูลในกระบวนการค้นหาเชิงความหมาย

จากภาพที่ 2 แผนภาพแสดงลำดับการไหลของข้อมูลเริ่มจากการรับคำค้น การสร้างเวกเตอร์เชิงความหมาย การค้นหาเอกสารกลุ่ม Top-k จาก ANN index การดึงข้อมูลจริงจากคลังข้อมูล และการคำนวณค่าความคล้ายเชิงความหมายเพื่อนำมาจัดอันดับผลลัพธ์จากมากไปน้อย นอกจากนี้ ระบบยังรองรับการกรองผลลัพธ์ด้วยข้อมูลกำกับ เช่น ปีดำเนินการ คณะ และเป้าหมายการพัฒนาที่ยั่งยืน (SDGs) เพื่อเพิ่มความแม่นยำเชิงบริบท

สถาปัตยกรรมการสืบค้นนี้สอดคล้องกับแนวทางระบบ Information Retrieval เชิงความหมายสมัยใหม่ที่ผสานการสร้าง Embeddings ด้วยโมเดล Transformer และการค้นหาเวกเตอร์ด้วยอัลกอริทึม ANN [3], [5], [8]

4. สถิติที่ใช้ในการวิจัย ได้แก่

4.1 ค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน เวลาในการตอบสนองของคำค้น (Query Time)

4.2 ค่า Cosine Similarity มีการกำหนดเกณฑ์ระดับความเกี่ยวข้องเป็น 5 ระดับ [3], [4] (Similarity ≥ 0.80 = สูงมาก, $0.70 - 0.79$ = สูง, $0.60 - 0.69$ = ปานกลาง, $0.50 - 0.59$ = ต่ำ, และ < 0.50 = ต่ำมาก)

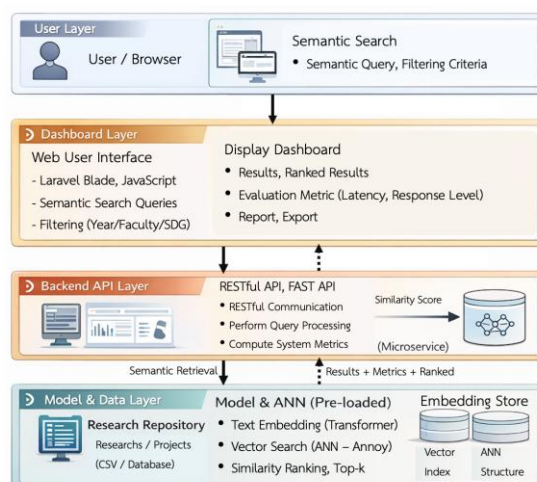
4.3 ตัวชี้วัด Response Level สำหรับการประเมินคุณภาพผลลัพธ์การสืบค้นตามกรอบแนวคิด Information Retrieval มีเกณฑ์การแปลผล 5 ระดับ [3], [17] (Response Level ≥ 0.80 = แม่นยำมาก, $0.70 - 0.79$ = ดี, $0.60 - 0.69$ = ปานกลาง, $0.50 - 0.59$ = ต่ำ, และ < 0.50 = ควรปรับปรุง)

4.4 การประเมินผู้ใช้ Mean Opinion Score (MOS) ร่วมกับมาตรไลเคิร์ต 5 ระดับ (คะแนนเฉลี่ย $4.51-5.00$ = มากที่สุด, $3.51-4.50$ = มาก, $2.51-3.50$ = ปานกลาง, $1.51-2.50$ = น้อย, และ $1.01-1.50$ = น้อยที่สุด)

ผลการวิจัย

1. ผลการพัฒนาระบบสารสนเทศ

ระบบสารสนเทศออกแบบตามสถาปัตยกรรมแบบไมโครเซอร์วิสซึ่งแยกส่วนการทำงานระหว่างส่วนติดต่อผู้ใช้ ตรรกะของระบบ และกระบวนการประมวลผลภาษาธรรมชาติอย่างชัดเจน เพื่อรองรับการขยายระบบ และการบำรุงรักษาในระยะยาว ระบบส่วนติดต่อผู้ใช้พัฒนาด้วย Laravel ทำหน้าที่เป็นแดชบอร์ดสำหรับป้อนคำค้น กำหนดตัวกรองตามปี คณะ และเป้าหมายการพัฒนาที่ยั่งยืน รวมถึงแสดงผลลัพธ์ในรูปแบบตารางและกราฟ และรองรับการส่งออกข้อมูลเป็นไฟล์ Excel และ PDF ขณะที่ระบบประมวลผลด้านหลังพัฒนาในรูปแบบ RESTful API ด้วย FastAPI โดยใช้โมเดล SentenceTransformer แบบหลายภาษาในการแปลงคำค้นและข้อมูลโครงการเป็นเวกเตอร์เชิงความหมาย และประยุกต์ใช้เทคนิค ANN ด้วยไลบรารี Annoy เพื่อจัดทำดัชนีและค้นหาเอกสารที่เกี่ยวข้อง ระบบจัดอันดับผลลัพธ์และกรองข้อมูล พร้อมคำนวณตัวชี้วัดด้านประสิทธิภาพ แสดงการออกแบบดังภาพที่ 3



ภาพที่ 3 การออกแบบสถาปัตยกรรมระบบสารสนเทศเพื่อการสืบค้นเชิงความหมาย

ผลการทดลองใช้ระบบสารสนเทศ โดยเปรียบเทียบกับวิธีการสืบค้นข้อมูลด้วยข้อความ “ความเข้มแข็งชุมชน” พบว่า ระบบสืบค้นแบบใช้คำสำคัญ (Keyword-based Search) สามารถค้นคืนข้อมูลได้เพียง 1 โครงการ ในทางตรงกันข้าม ระบบสืบค้นที่ประยุกต์ใช้เทคนิค Semantic Search สามารถค้นคืนโครงการที่มีความหมายใกล้เคียงกับคำค้นได้จำนวน 14 รายการ ภายใต้ค่า Top-K ที่ครอบคลุมข้อมูลทั้งหมด 937 โครงการ และจัดอันดับผลลัพธ์ตามค่าความคล้ายคลึงเชิงความหมาย โดยกำหนดค่า Threshold ไว้ที่ 0.60 ดังแสดงในภาพที่ 4 ซึ่งสะท้อนความสามารถของระบบในการจับความสัมพันธ์เชิงมิติของข้อมูลได้ดีกว่าวิธีการสืบค้นแบบใช้คำสำคัญ

นอกจากนี้ ระบบสารสนเทศที่พัฒนาขึ้นยังรองรับการกรองผลลัพธ์ตามเงื่อนไขต่าง ๆ เช่น ปีดำเนินการ คณะ/หน่วยงาน และเป้าหมายการพัฒนาที่ยั่งยืน (SDGs) รวมถึงสามารถส่งออกผลลัพธ์ในรูปแบบ Excel และ PDF เพื่อสนับสนุนการวิเคราะห์ข้อมูลและการจัดทำรายงานเชิงวิชาการได้

Semantic Search
ระบบค้นหาข้อมูลเชิงความหมาย

ความเข้มแข็งชุมชน ปีทั้งหมด ทุกคณะ ทุก SDG ค้นหา

Response Level
1.49%
Support Items: 14 / 937

Search Results

Title	Faculty	Year	Similarity
แนวทางการเสริมสร้างความเข้มแข็งชุมชนสู่การกำหนดปัจจัยความสำเร็จในการพัฒนาชุมชนท้องถิ่นจังหวัดอุดรธานี	คณะมนุษยศาสตร์และสังคมศาสตร์	2563	0.8091
การสร้างความเข้มแข็งของครอบครัวผ่านทุนทางสังคม กรณีตำบลป่าเซ่า จังหวัดอุดรธานี	คณะวิทยาศาสตร์และเทคโนโลยี	2564	0.7159
การพัฒนาขีดความสามารถของชุมชนจัดการตนเองเพื่อสร้างความมั่นคงทางอาหารและยกระดับเศรษฐกิจฐานรากภายใต้สถานการณ์วิกฤต โดยความร่วมมือของภาคีเครือข่ายในพื้นที่ของจังหวัดอุดรธานี	คณะมนุษยศาสตร์และสังคมศาสตร์	2565	0.6923
การศึกษาห่วงโซ่อุปทานกองทุนหมู่บ้านเพื่อยกระดับประสิทธิภาพการจัดการสู่สถาบันการเงินชุมชนด้วยหลักปรัชญาเศรษฐกิจพอเพียง ของสำนักงานกองทุนหมู่บ้านและชุมชนเมืองแห่งชาติ สาขา 2	คณะวิทยาการจัดการ	2563	0.6700
โครงการ การพัฒนาทักษะฝีมือแรงงาน "ช่างชุมชน" เพื่อยกระดับรายได้เศรษฐกิจฐานรากและส่งเสริมคุณภาพชีวิตชุมชน	คณะเทคโนโลยีอุตสาหกรรม	2565	0.6699
โครงการพัฒนาศักยภาพผู้นำชุมชนเพื่อพัฒนาผลิตภัณฑ์ชุมชนจังหวัดอุดรธานี	คณะวิทยาการจัดการ	2563	0.6480
การถ่ายทอดเทคโนโลยีเศรษฐกิจเพื่อชุมชน	คณะเทคโนโลยีอุตสาหกรรม	2563	0.6457
โครงการพัฒนาชุมชนต้นแบบตามแนวปรัชญาเศรษฐกิจพอเพียง ในชุมชนท้องถิ่น	คณะเกษตรศาสตร์	2562	0.6349
โครงการบริหารจัดการเพื่อพัฒนาคุณภาพชีวิตและยกระดับรายได้ให้กับคนในชุมชน	วิทยาลัยน่าน	2563	0.6238
การพัฒนาศักยภาพการบริหารจัดการกลุ่มอาชีพในชุมชนที่สอดคล้องกับบริบทและทรัพยากรพื้นที่ บ้านนาคูหา ตำบลสวนเขื่อน อำเภอเมือง จังหวัดแพร่	คณะมนุษยศาสตร์และสังคมศาสตร์	2563	0.6182
โครงการส่งเสริมโครงการพัฒนาคุณภาพชีวิตและยกระดับรายได้ให้กับคนในชุมชนฐานราก จังหวัดอุดรธานี แพร่ และน่านมหาวิทยาลัยราชภัฏอุดรธานี	คณะวิทยาศาสตร์และเทคโนโลยี	2564	0.6133
โครงการยกระดับความรู้ด้านโลจิสติกส์ของกลุ่มวิสาหกิจชุมชน ในเขต ต.วังกะพ้อ อ.เมือง จ.อุดรธานี เพื่อเพิ่มประสิทธิภาพการดำเนินงานและการสร้างความได้เปรียบทางการแข่งขันทางการค้า	คณะเทคโนโลยีอุตสาหกรรม	2563	0.6112
โครงการโครงการพัฒนาคุณภาพชีวิตและยกระดับรายได้ให้กับคนในชุมชนฐานราก จ.แพร่	คณะมนุษยศาสตร์และสังคมศาสตร์	2565	0.6106
โครงการพัฒนาศักยภาพกลุ่มเกษตรกรพัฒนาตำบลน้ำทิพย์สู่การเป็นผู้ประกอบการชุมชนเพื่อการสร้างฐานทางเศรษฐกิจ	วิทยาลัยน่าน	2564	0.6003

Export PDF Export Excel

Copyright © 2025. Uttaradit Rajabhat University. All rights reserved. URU Engagement

ภาพที่ 4 แดชบอร์ดระบบค้นหาข้อมูลเชิงความหมาย

2. ผลการหาประสิทธิภาพของระบบ

การประเมินประสิทธิภาพของระบบครอบคลุมสามมิติหลัก ได้แก่ (1) ประสิทธิภาพการสืบค้น (2) ความถูกต้องของการจับคู่เชิง และ (3) การรับรู้ความเกี่ยวข้องในมุมมองผู้ใช้ มีรายละเอียดดังนี้

2.1 ประสิทธิภาพการสืบค้น (Search Efficiency)

การประเมินระบบสืบค้นเชิงความหมายดำเนินการกับคำค้นตัวแทน 5 คำ โดยพิจารณาตัวชี้วัดด้านเวลาในการตอบสนอง (Query Time) และคุณภาพเชิงความหมาย ได้แก่ ค่าเฉลี่ย ค่าสูงสุด และค่าต่ำสุดของ Cosine Similarity [4] รวมถึงค่า Response Level ซึ่งเป็นตัวชี้วัดมาตรฐานในงานสืบค้นบนเวกเตอร์มิติสูง โดยอาศัยกรอบแนวคิดของการประเมินระบบสืบค้นสารสนเทศ [3], [17] และแนวทางการสืบค้นแบบ Approximate Nearest Neighbor [15], [16] ดังแสดงในตารางที่ 2

ตารางที่ 2 แสดงผลการประเมินประสิทธิภาพระบบค้นหาเชิงความหมาย

Query	Query Time (ms)	Average Cosine Similarity	Max Similarity	Min Similarity	Response Level (%)
ท้องถิ่น	46.8159	0.6317	0.6441	0.6262	60.0
การพัฒนาชุมชน	29.2618	0.7672	0.8011	0.7330	100.0
สิ่งแวดล้อม	26.3030	0.6944	0.6944	0.6944	10.0
การศึกษา	29.6378	0.6373	0.7310	0.6071	100.0
นวัตกรรม	40.7741	0.6023	0.6590	0.6056	80.0

ผลการทดลองกำหนดค่า Top-K = 10 คิดเป็นร้อยละ 1.07 ของชุดข้อมูลทั้งหมด และตั้งค่า threshold ของ Similarity ที่ 0.60 เพื่อจัดอันดับผลลัพธ์ พบว่าระบบสามารถตอบสนองคำค้นภายในเวลาเฉลี่ย 34.56 มิลลิวินาที ซึ่งต่ำกว่า 200 มิลลิวินาที ถือว่าอยู่ในระดับที่เหมาะสมสำหรับระบบสืบค้นเชิงโต้ตอบ ตามข้อเสนอในงานวิจัยด้าน Approximate Vector Search และ ANN-based retrieval [5], [7], [9] และให้ค่าความคล้ายเชิงความหมายเฉลี่ย 0.67 อยู่ในระดับปานกลาง สำหรับค่า Response Level มีการแปรผันตามลักษณะของคำค้น จากการทดสอบมีค่าเฉลี่ยเท่ากับร้อยละ 70.0 ซึ่งสามารถจัดอยู่ในระดับดี ทั้งนี้คำค้นที่มีความเฉพาะเจาะจงให้ผลลัพธ์ที่มีความเกี่ยวข้องสูงกว่า ขณะที่คำค้นเชิงนามธรรมมีความท้าทายมากกว่า

2.2 คุณภาพการจับคู่ความหมาย (Semantic Matching Quality)

คุณภาพการจับคู่เชิงความหมายประเมินจากค่า Cosine Similarity และตัวชี้วัด Response Level ซึ่งเป็นมาตรฐานที่ใช้กันอย่างแพร่หลายในงาน Semantic Similarity และ Text Retrieval [3], [4] ดังนี้

1) ค่าความคล้ายคลึง (Cosine Similarity) โดย Angular Distance เป็นค่าที่ใช้วัดความห่างเชิงมุม ระหว่างเวกเตอร์ของคำค้นและเอกสาร ถ้าค่ายิ่งต่ำหมายถึงเวกเตอร์ยิ่งอยู่ใกล้กัน และ Cosine Similarity เป็นค่าที่แปลงจาก Angular Distance ค่าอยู่ในช่วง $[0,1]$ ค่ายิ่งสูงหมายถึงความหมายยิ่งใกล้เคียง

ผลลัพธ์ในตารางที่ 3 แสดงให้เห็นว่าระบบสามารถแยกโครงการที่มีความเกี่ยวข้องสูงออกจากโครงการที่มีความเกี่ยวข้องต่ำได้อย่างชัดเจน ซึ่งสอดคล้องกับงานวิจัยที่รายงานว่า Semantic Embeddings สามารถจับความคล้ายคลึงเชิงบริบทได้แม่นยำกว่าวิธีการสืบค้นแบบคำสำคัญ [4], [6] ผลการสืบค้นสำหรับคำค้น “ดิจิทัล” พบว่าเอกสารที่ถูกจัดอันดับในลำดับต้นมีค่า Angular Distance ต่ำและมีค่า Cosine Similarity สูง ขณะที่เอกสารในลำดับถัดไปมีค่า Angular Distance เพิ่มขึ้นและค่า Cosine Similarity ลดลงอย่างต่อเนื่อง แนวโน้มดังกล่าวสะท้อนว่าเอกสารที่อยู่ในอันดับต้นมีความใกล้เคียงเชิงความหมายกับคำค้นมากกว่า และยืนยันว่าโครงสร้าง ANN สามารถจัดอันดับเอกสารตามระดับความใกล้เคียงเชิงความหมายได้อย่างเป็นระบบ [15], [18]

ตารางที่ 3 ผลการสืบค้นสำหรับคำค้น “ดิจิทัล”

ลำดับ	Document Index	Angular Distance	Cosine Similarity	การตีความ
1	263	0.8437	0.6441	ความหมายใกล้เคียงระดับค่อนข้างสูง
2	264	0.8647	0.6262	ใกล้เคียงเชิงความหมาย
3	58	0.8662	0.6248	ใกล้เคียงเชิงบริบท
4	278	0.9750	0.5247	ความเกี่ยวข้องปานกลาง
5	271	0.9887	0.5112	ใกล้เคียงขั้นต่ำ

2) การวิเคราะห์ตัวชี้วัด Response Level การประเมินประสิทธิภาพของระบบด้วยตัวชี้วัด Response Level ดำเนินการโดยพิจารณาสัดส่วนของผลลัพธ์ที่มีค่าความคล้ายคลึงเชิงความหมาย (Cosine Similarity) อยู่ในระดับที่กำหนดไว้ในช่วงมากกว่า 0.60 (Threshold = 0.6) ซึ่งเป็นแนวทางที่เหมาะสมสำหรับชุดข้อมูลที่ไม่มีป้ายกำกับ (Unlabeled Datasets) [2], [17] และ ใช้เป็นการกรองข้อมูลออกจากผลลัพธ์ โดยไม่เข้มงวดเกินไปจนทำให้ Recall ลดลง สอดคล้องกับแนวทางในงาน Information Retrieval ในบริบท Semantic Search [3,17]

ผลการทดลองสำหรับคำค้น “ท้องถิ่น” พบว่า ระบบสามารถดึงผลลัพธ์ที่มีค่าความคล้ายคลึงสูงกว่าค่าขีดระดับได้จำนวน 3 รายการจากทั้งหมด 5 รายการ ส่งผลให้มีค่า Response Level เท่ากับ 60% ซึ่งสะท้อนว่ามากกว่ากึ่งหนึ่งของผลลัพธ์ มีความสอดคล้องกับความหมายของคำค้นในระดับที่ยอมรับได้ ซึ่งสอดคล้องกับข้อจำกัดของข้อมูลจริงที่มีความหลากหลายทางภาษา แต่ผลลัพธ์ยังแสดงผลได้ในสัดส่วนที่น่าพอใจ

2.3 การประเมินเชิงผู้ใช้ (User-Centered Evaluation) การประเมินเชิงผู้ใช้ใช้แนวคิด Mean Opinion Score (MOS) ร่วมกับมาตรไลเคิร์ต 5 ระดับ เพื่อสะท้อนการรับรู้คุณภาพของผู้ใช้จริง กลุ่มผู้ประเมินจำนวน 30 คน ซึ่งเป็นบุคลากรสายวิชาการ จำนวน 20 คน และสายสนับสนุนของมหาวิทยาลัย จำนวน 10 คน ได้ทดลองใช้งานระบบสืบค้นเชิงความหมายผ่านแดชบอร์ดด้วยคำค้นตัวอย่าง และประเมินคุณภาพใน 3 มิติ ได้แก่ ความเกี่ยวข้องของผลลัพธ์ ความชัดเจนของข้อมูล และความรวดเร็วในการตอบสนอง [3], [9], [17] แสดงผลการประเมินดังตารางที่ 4

ตารางที่ 4 ผลการประเมินเชิงผู้ใช้ด้วย Mean Opinion Score (MOS)

มิติการประเมิน	ค่าเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน	ระดับการประเมิน
ความเกี่ยวข้องของผลลัพธ์	4.37	0.49	ระดับมาก
ความชัดเจนของข้อมูลที่แสดง	4.20	0.41	ระดับมาก
ความรวดเร็วในการตอบสนอง	4.43	0.50	ระดับมาก
ค่าเฉลี่ย	4.33	0.47	ระดับมาก

ผลการประเมินพบว่า ระบบมีคุณภาพในระดับมากในทุกมิติ โดยมีค่าเฉลี่ยรวม 4.33 และส่วนเบี่ยงเบนมาตรฐาน 0.47 สะท้อนว่าการประยุกต์ใช้การสืบค้นเชิงความหมายช่วยเพิ่มความพึงพอใจของผู้ใช้เมื่อเทียบกับการสืบค้นแบบใช้คำสำคัญดั้งเดิม อย่างไรก็ตาม มิติด้านความชัดเจนของข้อมูลยังมีความแปรผันบ้าง เนื่องจากข้อจำกัดของคำอธิบายโครงการที่ไม่สมบูรณ์ในบางระเบียบ ซึ่งเป็นประเด็นที่ควรปรับปรุงในงานวิจัยในอนาคต

อภิปรายผลการวิจัย

จากผลการวิจัยสามารถสรุปผลตามวัตถุประสงค์ได้ดังนี้ 1) การพัฒนารอบการสืบค้นเชิงความหมายบนพื้นฐานเวกเตอร์ พบว่า กรอบที่พัฒนาขึ้นประกอบด้วยโมดูลหลัก ได้แก่ การสร้างตัวแทนเชิงความหมายด้วย Sentence Transformer การจัดทำดัชนีด้วยอัลกอริทึม Approximate Nearest Neighbor (ANN: Annoy) และการจัดอันดับผลลัพธ์ด้วย Cosine Similarity โดยระบบสามารถค้นคืนข้อมูลที่มีความเกี่ยวข้องเชิงบริบทได้อย่างมีประสิทธิภาพ และมีเวลาตอบสนอง 34.56 มิลลิวินาที ซึ่งเหมาะสมต่อการใช้งานจริงในคลังข้อมูลโครงการมหาวิทยาลัย 2) ผลการทดลองใช้ระบบสืบค้นเชิงความหมาย พบว่า ระบบสามารถค้นคืนโครงการที่เกี่ยวข้องได้ดีกว่าการสืบค้นแบบใช้คำสำคัญ โดยสามารถค้นพบข้อมูลที่มีความหมายใกล้เคียงกันแม้ใช้ถ้อยคำแตกต่างกันอย่างมีนัยสำคัญ และมีค่าความคล้ายคลึงเชิงความหมายอยู่ในระดับปานกลาง สะท้อนประสิทธิภาพของการผสานเทคนิค Sentence Transformer และ ANN ในบริบทภาษาไทย 3) ผลการประเมินประสิทธิภาพและคุณภาพของระบบพบว่า (1) ระบบมีประสิทธิภาพด้านเวลาในการตอบสนองอยู่ในระดับรวดเร็ว (2) ค่า Cosine Similarity สามารถแยกความเกี่ยวข้องของข้อมูลได้อย่างชัดเจน และ (3) ค่า Response Level อยู่ในระดับดี ขึ้นอยู่กับลักษณะของคำค้น นอกจากนี้ ผลการประเมินจากผู้ใช้มีค่าเฉลี่ย 4.33 อยู่ในระดับมาก สะท้อนว่าระบบที่พัฒนาขึ้นสามารถผสานเทคนิคการประมวลผลภาษาธรรมชาติร่วมกับอัลกอริทึมการค้นหาเวกเตอร์ได้อย่างเป็นระบบ ส่งผลให้สามารถสืบค้นข้อมูลโครงการที่มีความหมายใกล้เคียงกันได้อย่างมีประสิทธิภาพ แม้ใช้ถ้อยคำที่แตกต่างกัน

ผลวิจัยดังกล่าวสอดคล้องกับงานของ Reimers และ Gurevych [6] ที่พบว่า การใช้ Sentence-level Embedding สามารถเพิ่มความแม่นยำในการจับคู่ความหมายของข้อความและสนับสนุนการจัดอันดับเอกสารในระบบสืบค้นได้อย่างมีประสิทธิภาพ รวมถึงงานของ Manning [3] ที่ชี้ให้เห็นว่าการใช้ Representation เชิงความหมายช่วยแก้ข้อจำกัดของ Keyword-based Retrieval ในสภาพข้อมูลจริงที่มีความหลากหลายของภาษา ผลการวิจัยแสดงให้เห็นว่า เหตุที่ระบบสามารถตอบสนอง 34.56 มิลลิวินาที เนื่องมาจากการเลือกใช้อัลกอริทึม ANN ซึ่งลดความซับซ้อนของการค้นหาเวกเตอร์ในปริมาณความมิติสูง สอดคล้องกับงานวิจัย [7], [8], [15]

ในด้านคุณภาพการจับคู่เชิงความหมาย ค่าความคล้ายคลึง (Cosine Similarity) ที่สามารถแยกโครงการที่เกี่ยวข้องสูงออกจากโครงการที่เกี่ยวข้องต่ำได้อย่างชัดเจน สะท้อนว่าโมเดล Sentence Transformer สามารถเรียนรู้บริบทเชิงความหมายของข้อความภาษาไทยได้อย่างมีประสิทธิภาพ ผลดังกล่าวสอดคล้องกับงานของ Devlin และคณะ [5] และงานภาษาไทยของ Netisopakul และคณะ [2]

อย่างไรก็ตาม ผู้วิจัยได้ดำเนินการวิเคราะห์ผลการศึกษาทั้งในมิติของประสิทธิภาพและคุณภาพการจับคู่เชิงความหมายของระบบ ภายใต้บริบทข้อมูลโครงการของมหาวิทยาลัย โดยพิจารณาผลลัพธ์การสืบค้นร่วมกับค่า Cosine Similarity และค่า Response Level ทั้งนี้ พบว่าความคลาดเคลื่อนของระบบอาจเกิดขึ้นได้จากหลายปัจจัยที่เกี่ยวข้องกับลักษณะข้อมูล กระบวนการสร้างตัวแทนเชิงความหมาย และกลไกการจัดอันดับผลลัพธ์ ได้แก่ (1) ลักษณะของคำค้นที่มีความกำกวมหรือเป็นนามธรรม เช่น “สิ่งแวดล้อม” ซึ่งทำให้การตีความเจตนาของผู้ใช้คลาดเคลื่อน (2) ความไม่สมบูรณ์ของข้อมูล โดยเฉพาะระเบียบที่ไม่มีคำอธิบายโครงการ ส่งผลให้การสร้างเวกเตอร์เชิงความหมายขาดบริบทที่เพียงพอ (3) ข้อจำกัดของอัลกอริทึม Approximate Nearest Neighbor (ANN) ซึ่งให้ผลลัพธ์แบบประมาณค่า อาจไม่ใช่เพื่อนบ้านที่ใกล้เคียงที่สุดจริง และ (4) การกำหนดค่า Threshold ของ Cosine Similarity ที่ระดับ 0.60 ซึ่งอาจก่อให้เกิดทั้งกรณีผลลัพธ์ที่เกี่ยวข้องถูกตัดออก (False Negative) และผลลัพธ์ที่ไม่เกี่ยวข้องถูกจัดอันดับสูง (False Positive) ผลการวิเคราะห์ข้อผิดพลาดดังกล่าวสะท้อนให้เห็นถึงปัจจัยสำคัญที่มีผลต่อประสิทธิภาพของระบบ และชี้ให้เห็นแนวทางในการพัฒนาต่อยอดซึ่งจะช่วยยกระดับความแม่นยำของระบบสืบค้นเชิงความหมายได้อย่างมีประสิทธิภาพ

โดยสรุปผลแสดงให้เห็นว่ากรอบการสืบค้นเชิงความหมายที่พัฒนาขึ้นสามารถตอบโจทย์วัตถุประสงค์การวิจัยทั้งด้านประสิทธิภาพ และความเหมาะสมต่อการใช้งานจริงในบริบทของคลังข้อมูลโครงการมหาวิทยาลัยไทย ซึ่งสอดคล้องกับแนวโน้มงานวิจัยด้าน Semantic Search และ Information Retrieval สมัยใหม่อย่างชัดเจน

ข้อเสนอแนะ

1. ควรนำระบบสืบค้นเชิงความหมายไปบูรณาการกับคลังข้อมูลกลางของมหาวิทยาลัย เพื่อสนับสนุนการค้นคืนข้อมูลโครงการข้ามคณะและการวิเคราะห์เชิงนโยบาย
2. งานวิจัยในอนาคตควรศึกษาเปรียบเทียบประสิทธิภาพของ ANN หลายอัลกอริทึม เช่น FAISS และ HNSW รวมถึงการปรับจูนโมเดล Embedding สำหรับภาษาไทยโดยเฉพาะ

เอกสารอ้างอิง

- [1] Uttaradit Rajabhat University, "Engagement Information System: Semantic Search System," 2025. [Online]. Available: <https://info.uru.ac.th/engagement/semantic-search>.
- [2] P. Netisopakul, G. Wohlgenannt, A. Pulich, and Z. Z. Hlaing, "Improving the state-of-the-art in Thai semantic similarity using distributional semantics and ontological information," PLOS ONE, vol. 16, no. 2, p. e0246751, 2021, doi: 10.1371/journal.pone.0246751.
- [3] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008.
- [4] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," Journal of Artificial Intelligence Research, vol. 37, pp. 141–188, 2010, doi: 10.1613/jair.2834.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [6] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [7] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019, doi: 10.1109/TBDATA.2019.2921572.
- [8] E. Bernhardsson, "Annoy: Approximate nearest neighbors in C++/Python," 2018. [Online]. Available: <https://github.com/spotify/annoy>.
- [9] T. Junlabuddee and K. Tuamsuk, "Semantic search system for research data in information science," Journal of Information Science (Thailand), vol. 39, no. 3, pp. 43–61, 2021. (in Thai)
- [10] N. Kaewboonma, J. Panawong, E. Pianhanuruk, and M. Buranarach, "Development of intelligent semantic search system for rubber research data in Thailand," in AIP Conference Proceedings, vol. 1891, no. 1, p. 020073, 2017, doi: 10.1063/1.5005333.
- [11] K. Jearanaitanakij et al., "Thai question-answering system using similarity search and LLM," ECTI Transactions on Computer and Information Technology (ECTI-CIT), vol. 18, no. 3, pp. 406–416, 2024, doi: 10.37936/ecti-cit.2024183.255313.
- [12] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 2016, pp. 55–64, doi: 10.1145/2983323.2983769.
- [13] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [14] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," ACM Computing Surveys, vol. 52, no. 1, pp. 1–38, 2019, doi: 10.1145/3285029.

- [15] P. Indyk and R. Motwani, "Approximate nearest neighbors: Toward removing the curse of dimensionality," in Proceedings of the 30th Annual ACM Symposium on Theory of Computing, Dallas, TX, USA, 1998, pp. 604–613, doi: 10.1145/276694.276876.
- [16] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval: The Concepts and Technology Behind Search, 2nd ed. Boston, MA, USA: Addison-Wesley, 2011.
- [17] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.
- [18] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 824–836, 2020, doi: 10.1109/TPAMI.2018.2889473.