โครงร่างวิทยานิพนธ์

THESIS PROPOSAL

ชื่อเรื่อง (ภาษาไทย) โปรแกรมค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

ชื่อเรื่อง (ภาษาอังกฤษ) A Cross-lingual Search Engine for Retrieval of Green House Gas Emission Factor

เสนอโดย นายณฐพจน์ หนูวงษ์

รหัสนิสิต 6770233221

หลักสูตร วิทยาศาสตรมหาบัณฑิต

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ (ภาคนอกเวลาราชการ)

ภาควิชา วิศวกรรมคอมพิวเตอร์

คณะ วิศวกรรมศาสตร์

สถานที่ติดต่อ 90/159 ต.ลาดวาย อ.ลำลูกกา ปทุมธานี. 12150

โทรศัพท์ 094-0768695

อีเมล 6770233221@student.chula.ac.th

อาจารย์ที่ปรึกษา รศ.ดร.ญาใจ ลิ่มปิยะกรณ์

คำสำคัญ (ภาษาไทย) การค้นคืนสารสนเทศข้ามภาษา, ค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก, คำพ้องความ

คำสำคัญ (ภาษาอังกฤษ) Cross-Lingual Information Retrieval, GHGs Emission Factor, Synonym

โครงร่างวิทยานิพนธ์

ภาษาไทย โปรแกรมค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

ภาษาอังกฤษ A Cross-lingual Search Engine For Retrieval of Green House Gas Emission Factor

1. ที่มาและความสำคัญของปัญหา

โลกกำลังเผชิญสภาพอากาศสุดขั้ว (Extreme weather) ซึ่งมีสาเหตุหลักมาจากการปล่อยก๊าซเรือนกระจก (Greenhouse Gas: GHG) สู่ชั้นบรรยากาศในปริมาณมากและต่อเนื่อง ส่งผลกระทบเป็นวงกว้างต่อระบบนิเวศ สภาพภูมิอากาศ ความหลากหลายทางชีวภาพ รวมถึงคุณภาพชีวิตของผู้คนทั้งในระดับท้องถิ่นและระดับโลก ดังนั้น การติดตามและบริหารจัดการข้อมูลการปล่อยก๊าซเรือนกระจกอย่างมีประสิทธิภาพจึงเป็นเรื่องจำเป็นอย่างยิ่ง สำหรับประเทศไทย หน่วยงานหลักที่ทำหน้าที่ประเมิน เก็บรวบรวม จัดทำข้อมูล และเผยแพร่ข้อมูลที่เกี่ยวข้องกับก๊าซเรือนกระจก คือ องค์การบริหารจัดการก๊าซเรือนกระจก หรือ อบก. (Thailand Greenhouse Gas Management Organization–TGO) ซึ่งเป็นหน่วยงานสำคัญภายใต้การกำกับดูแลของกระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม TGO ได้จัดทำฐานข้อมูลและรายงานที่เกี่ยวข้องกับการปล่อยก๊าซเรือนกระจกครอบคลุมหลากหลายภาคส่วน อาทิ ภาคอุตสาหกรรม ภาคพลังงาน เกษตรกรรม และชุมชน รวมถึงข้อมูลการประเมินและคำนวณค่าคาร์บอนฟุตพรินต์ (carbon footprint) การกำหนดค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก (GHGs Emission Factor) โดยอ้างอิงจากมาตรฐานสากลและแนวทางของ IPCC (Intergovernmental Panel on Climate Change)

อย่างไรก็ตาม แม้ TGO จะเผยแพร่ข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกในหลายรูปแบบบนเว็บไซต์ อาทิ ฐานข้อมูลออนไลน์ และไฟล์อิเล็กทรอนิกส์ต่าง ๆ แต่ยังขาดระบบสืบค้นที่สามารถตอบสนองต่อคำค้น EF (Emission Factor) ได้อย่างครอบคลุมและ “ข้ามภาษา” (Cross-lingual) กล่าวคือ ในเนื้อหาบนเว็บไซต์ TGO ชื่อคำค้น EF บางรายการเป็นภาษาอังกฤษ บางรายการเป็นภาษาไทย บางรายการทั้งไทย-อังกฤษ ในด้านผู้ใช้งาน บางส่วนถนัดใช้คำค้นภาษาไทย ขณะที่บางส่วนจะถนัดใช้คำค้นภาษาอังกฤษ สาเหตุปัจจัยเหล่านี้ทำให้การค้นหาแบบปกติที่จำกัดเฉพาะภาษาใดภาษาหนึ่ง อาจค้นหาไม่พบ หรือได้ผลลัพธ์ไม่ถูกต้อง งานวิจัยนี้จึงได้นำเสนอการพัฒนาโปรแกรมค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกครอบคลุมการค้นหาด้วยภาษาไทยและภาษาอังกฤษ รวมทั้งสามารถค้นหาจากการเทียบคู่คำพ้องความ (synonym) เพื่ออำนวยความสะดวกแก่ผู้ใช้งาน ไม่ถูกจำกัดด้วยภาษา สืบค้นได้อย่างรวดเร็วและครอบคลุม ผลลัพธ์การค้นคืนจะแสดงรายการ EF ที่พบในเวอร์ชันปีต่างๆ เพื่อให้ผู้ใช้เลือกค่าได้อย่างถูกต้อง เป็นประโยชน์ต่อผู้ประกอบการในการเลือกใช้ค่า EF ที่ถูกต้องเหมาะสมในการประเมินรายงานคาร์บอนฟุตพรินต์ขององค์กร ส่งผลทำให้ลดข้อผิดพลาด ลดการทำงานซ้ำ และใช้ทรัพยากรอย่างคุ้มค่า

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การค้นคืนสารสนเทศ (Information Retrieval: IR) [1]

2.1.1 แนวคิดพื้นฐานของ IR

• Boolean Model: ใช้ตัวดำเนินการ (Operators) เช่น AND, OR, NOT ในการกำหนดเงื่อนไขค้นหา เอกสารที่ตรงตามเงื่อนไขทั้งหมดจะถูกดึงขึ้นมาแบบ “ตรง-ไม่ตรง” (exact match) แต่ขาดการจัดอันดับตามความเกี่ยวข้อง

• Vector Space Model: แทนเอกสารและคำค้นเป็นเวกเตอร์ในมิติคำ (Term Dimension) แล้วคำนวณความคล้ายโคไซน์ (Cosine Similarity) เพื่อจัดอันดับเอกสารตามความเกี่ยวข้องกับคำค้น

• Probabilistic Models (เช่น BM25): ประเมินความน่าจะเป็นที่เอกสารจะเกี่ยวข้องกับคำค้น โดยนำปัจจัยต่างๆ เช่น ความถี่ของคำ (Term Frequency–TF), ความถี่ในคอร์ปัส (Inverse Document Frequency–IDF), และการปรับสเกลตามความยาวเอกสารมาประกอบ

2.1.2 การประเมินสมรรถนะ (Performance Evaluation)

ตัววัดที่นิยมใช้ในการประเมินความถูกต้องของระบบการค้นคืนสารสนเทศ ประกอบด้วย

• *Precision* ประเมินความถูกต้องของผลลัพธ์ที่ได้จากระบบค้นคืนสารสนเทศ โดยวัดจากสัดส่วนของผลลัพธ์ที่ถูกต้อง (True Positives) เทียบกับผลลัพธ์ทั้งหมดที่ระบบดึงออกมา (ทั้งที่ถูกต้องและผิดพลาด)

$$Precision = \frac{\text{จำนวนคำที่เกี่ยวข้องและถูกดึงมา } (TP)}{\text{จำนวนคำทั้งหมดที่ถูกดึงมา } (TP\ +\ FP)}$$

• *Recall* ประเมินความสามารถของระบบในการค้นคืนผลลัพธ์ที่เกี่ยวข้องทั้งหมด โดยวัดจากสัดส่วนของคำที่เกี่ยวข้องทั้งหมดที่ระบบค้นคืนมาได้ เทียบกับคำที่เกี่ยวข้องจริงทั้งหมด

$$Recall = \frac{\text{จำนวนคำที่เกี่ยวข้องและถูกดึงมา } (TP)}{\text{จำนวนคำที่เกี่ยวข้องทั้งหมด } (TP\ +\ FN)}$$

• Mean Average Precision (MAP) ใช้เมื่อต้องการวัดประสิทธิภาพของการจัดอันดับผลลัพธ์ เป็นตัววัดที่คำนวณค่าเฉลี่ยของค่า Precision ในทุกระดับที่คำที่เกี่ยวข้องปรากฏในลำดับผลลัพธ์ เหมาะสำหรับระบบการค้นคืน

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP(q)$$

2.2 การค้นคืนสารสนเทศข้ามภาษา (Cross-lingual Information Retrieval– CLIR) [2]

โดยทั่วไป ระบบสืบค้นจะสมมติให้คำค้น (Query) และเนื้อหาเอกสารเป็นภาษาเดียวกัน แต่ในกรณี Cross-lingual IR จะเกิดสถานการณ์ที่ผู้ใช้พิมพ์คำค้นเป็นภาษาไทยแต่ต้องการค้นหาเอกสารที่อาจเป็นภาษาอังกฤษหรือทั้งสองภาษาไทย/ อังกฤษ เพื่อให้ CLIR ทำงานได้อย่างมีประสิทธิภาพ ต้องมีกลไกเชื่อมโยงระหว่างภาษาไทยกับภาษาอังกฤษซึ่งมี 2 แนวทางหลักที่สำคัญ ได้แก่

2.2.1 Synonym-based (Dictionary-based) [3]

ใช้คลังคำศัพท์คู่ (Bilingual Dictionary) บรรจุรายการคำพ้องความหมายที่จับคู่คำหรือวลีสำคัญในภาษาไทยและภาษาอังกฤษไว้ล่วงหน้า โดยในขั้นตอน Tokenization และ Indexing, โปรแกรมค้นหาจะขยายคำค้น (Query Expansion) ให้ครอบคลุมคำพ้องความหมาย ในอีกภาษา ยกตัวอย่างเช่น หากผู้ใช้พิมพ์ “ก๊าซเรือนกระจก” ระบบจะสืบค้น “gas greenhouse” หรือ “greenhouse gas” ด้วย ซึ่งคลังคำศัพท์คู่จะอยู่ภายใต้ส่วนของ Domain-specific knowledge ภาพรวมดังแสดงในภาพที่ 1 ประกอบด้วย

• ป้ายกำกับทางเลือก (Alternative Labels) คำศัพท์ที่ใช้แทนกันได้และมีความหมายเหมือนกัน เช่น ตัวย่อ, อักษรย่อ, การสะกดคำผิด หรือการสะกดที่แตกต่างกัน เช่น

CTO => Chief Technology Officer

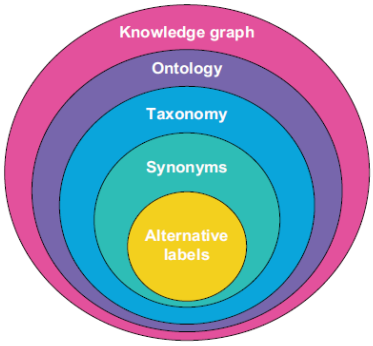
• คำพ้องความ (Synonyms) คำหรือวลีที่สามารถแทนกันได้ โดยแสดงสิ่งที่เหมือนกันหรือใกล้เคียงกัน แม้ว่าจะมีความแตกต่างเล็กน้อยในบริบท เช่น

human => homo sapiens, mankind

• อนุกรมวิธาน (Taxonomy) การจัดหมวดหมู่หรือจัดโครงสร้างข้อมูลให้เป็นลำดับชั้น โดยระบุความสัมพันธ์ระหว่างหมวดหมู่

• ออนโทโลยี (Ontology) การกำหนดความสัมพันธ์ที่ซับซ้อนระหว่างสิ่งต่างๆ ในโดเมน ในความสัมพันธ์เชิงนามธรรม หรือการอ้างอิงลำดับขั้น เช่น พนักงานรายงานต่อหัวหน้า

• กราฟความรู้ (Knowledge Graph) การนำออนโทโลยีมาปฏิบัติจริง โดยระบุเอนทิตีเฉพาะและความสัมพันธ์ระหว่างกัน เช่น ไมเคิลเป็นพนักงาน,ไมเคิลรายงานการทำงานต่อจิม, ดังนั้น จิมเป็นหัวหน้าไมเคิล เป็นต้น



ภาพที่ 1 ภาพรวม Domain-specific knowledge [3]

2.2.2 Embedding-based (Neural / Vector-based)

ใช้โมเดลประมวลภาษาธรรมชาติแบบหลายภาษา (Multilingual NLP) เช่น Multilingual BERT, XLM-R, LaBSE ฯลฯ เพื่อเข้ารหัส (encode) ประโยคหรือข้อความทั้งภาษาไทยและภาษาอังกฤษให้อยู่ในเวกเตอร์ใน latent space เดียวกัน เมื่อผู้ใช้พิมพ์คำค้นภาษาไทย ระบบจะแปลงคำค้นนั้นเป็นเวกเตอร์ และเทียบความคล้ายกับเวกเตอร์ของเอกสารที่อาจจะเป็นภาษาอังกฤษหรือภาษาไทยก็ได้ หากความหมายใกล้เคียงกัน เวกเตอร์ก็จะอยู่ใกล้กัน

2.3 การประมวลภาษาธรรมชาติ (Natural Language Processing – NLP) [4]

การประมวลภาษาธรรมชาติเป็นแขนงหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence– AI) ที่มุ่งเน้นการทำให้เครื่องคอมพิวเตอร์สามารถเข้าใจ ตีความ และจัดการกับภาษามนุษย์ได้อย่างมีประสิทธิภาพ ในบริบทของการค้นคืนสารสนเทศข้ามภาษา NLPเป็นขั้นตอนสำคัญในการเตรียมข้อมูลและสร้างความเข้าใจในภาษาที่ใช้สำหรับการสืบค้น ในงานวิจัยนี้จะเน้นไปที่ ภาษาไทยและภาษาอังกฤษ ซึ่งมีความแตกต่างกันอย่างชัดเจนทั้งในด้านโครงสร้างทางภาษา การตัดคำ และการประมวลผลคำศัพท์เฉพาะทาง

2.3.1 Tokenization / Word Segmentation

ภาษาไทยไม่มีการเว้นวรรคระหว่างคำเหมือนภาษาอังกฤษ ทำให้ต้องใช้เครื่องมือเฉพาะ เช่น Thai tokenizer, ICU tokenizer ใน Elasticsearch เพื่อช่วยตัดคำให้เหมาะสม ภาษาอังกฤษมักใช้ standard tokenizer และอาจเพิ่มขั้นตอน stemming หรือ lemmatization ได้

2.3.2 Stop Words & Synonym

การกำหนด Stop Word เช่น และ, คือ, the, a สามารถช่วยลด noise และเพิ่มประสิทธิภาพในการค้นหา การกำหนด Synonym ทั้งภาษาไทยและอังกฤษ หากใช้ Dictionary-based

2.3.3 Named Entity Recognition (NER)

ในบางกรณีอาจต้องจับชื่อเฉพาะหรือศัพท์เทคนิค เช่น ชื่อสารเคมี ประเภทวัสดุ หรือหน่วยงาน สามารถใช้ NER หรือวิธีการตรวจจับเชิง Lexicon เพื่อเพิ่มประสิทธิภาพการจับคู่

2.4 ค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก (GHGs Emission Factor)

ค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกคำนวณได้จากปริมาณการปล่อยและดูดกลับก๊าซเรือนกระจกต่อหนึ่งหน่วยกิจกรรม ใช้สำหรับประเมินปริมาณการปล่อยก๊าซเรือนกระจกที่เกิดจากกิจกรรมต่าง ๆ เช่น การใช้พลังงาน การเผาไหม้เชื้อเพลิง การขนส่ง หรือกระบวนการผลิตสินค้า ค่าดังกล่าวเป็นตัวแปรสำคัญที่ช่วยให้องค์กร หน่วยงานภาครัฐ และนักวิจัยสามารถคำนวณและวิเคราะห์ปริมาณการปล่อยก๊าซเรือนกระจกเพื่อใช้ในการรายงาน ติดตามผลกระทบ และวางแผนเพื่อลดการปล่อยก๊าซเรือนกระจก โดยใช้ข้อมูลทุติยภูมิจากแหล่งข้อมูลที่เรียงลำดับตามความน่าเชื่อถือจากมากไปน้อย [5-7] ดังนี้

1. ฐานข้อมูลสิ่งแวดล้อมของวัสดุพื้นฐานและพลังงานของประเทศไทย

2. ข้อมูลจากวิทยานิพนธ์และงานวิจัยที่เกี่ยวข้องที่ทำในประเทศไทย ซึ่งผ่านการกรองแล้ว (peer-reviewed publications)
3. ฐานข้อมูลที่เผยแพร่ทั่วไป ได้แก่ LCA Software, ฐานข้อมูลเฉพาะของกลุ่มอุตสาหกรรม, ฐานข้อมูลเฉพาะของแต่ละประเทศ
4. ข้อมูลที่ตีพิมพ์โดยองค์กรระหว่างประเทศ เช่น IPCC สหประชาชาติ

2.5 เทคโนโลยีและแพลตฟอร์มที่ใช้ในงานวิจัย

2.5.1 Elasticsearch [8]

เป็นโปรแกรมค้นหาแบบกระจาย (Distributed search engine) ที่รองรับ Full-text search, Structured search รวมถึง Vector Search มี Plugin หรือ Analyzer สำหรับภาษาไทย (Thai Tokenizer) และสามารถกำหนด Synonym Filter สำหรับ Cross-lingual

• Full-text Search รองรับการค้นหาข้อความทั้งภาษาไทยและอังกฤษด้วยการตั้งค่า Custom Analyzer และ Synonym Filter

• Synonym Matching ใช้ Synonym Filter เพื่อจับคู่คำพ้องความ ระหว่างภาษา เช่น "LPG" ⟷ "Liquified Petroleum Gas" ⟷ "ก๊าซหุงต้ม"

• Vector Search รองรับการค้นหาเชิงความหมายโดยใช้ฟิลด์แบบ Dense Vector และโมเดล NLP เช่น Multilingual BERT

• การจัดอันดับเอกสารหรือคำ คำนวณคะแนนตามความถี่คำ (TF-IDF) และความยาวเอกสาร เพื่อเพิ่มความถูกต้องในการค้นหา

2.5.2 FastAPI [9]

เว็บเฟรมเวิร์กภาษา Python ที่มีประสิทธิภาพสูง ใช้งานง่าย ทำให้สามารถสร้าง REST API เพื่อเชื่อมต่อระหว่าง Frontend กับ Elasticsearch ได้อย่างสะดวก

2.5.3 React.js [10]

ไลบรารี JavaScript สำหรับพัฒนา Frontend มีจุดเด่นด้านการสร้าง UI ที่โต้ตอบผู้ใช้ได้ง่าย ช่วยให้ผู้ใช้สามารถพิมพ์คำค้น แล้วเรียก API ได้ทันที และแสดงผลลัพธ์แบบเรียลไทม์

2.5.4 Apache Airflow [11]

แพลตฟอร์มสำหรับการสร้าง จัดการ และติดตาม Workflow หรือ DAGs (Directed Acyclic Graphs) ซึ่งใช้สำหรับการประมวลผลและจัดการงานต่าง ๆ (Task) โดยเฉพาะอย่างยิ่งในงานด้าน Data Pipeline และ ETL (Extract, Transform, Load)

3. งานวิจัยที่เกี่ยวข้อง

3.1 English-Malayalam Cross-Lingual Information Retrieval – an Experience [12]

งานวิจัยนี้นำเสนอโปรแกรมค้นคืนสารสนเทศข้ามภาษาอังกฤษ-มาลายาลัม ที่รองรับการสืบค้นทั้งภาษาเดียวและข้ามภาษา โดยใช้พจนานุกรมอังกฤษ-มาลายาลัมที่พัฒนาขึ้นเอง พร้อมด้วยเทคนิคการประมวลผลคำ เช่น การตัดคำ, การกำจัดคำหยุด, และการแปลงรากศัพท์ ระบบใช้ Vector Space Model (VSM) ในการจัดอันดับเอกสาร โดยคำนวณน้ำหนักคำผ่าน Local Weighting, Global Weighting (pidf) และ Normalization Factor อินเทอร์เฟซผู้ใช้ถูกพัฒนาด้วย NetBeans 6 และ JDK 1.6 ระบบได้รับการประเมินด้วยคำถาม 25 คำถาม และแสดงผลลัพธ์ที่มีประสิทธิภาพใกล้เคียงกันระหว่างการสืบค้นภาษาเดียวและข้ามภาษา งานวิจัยนี้ยืนยันถึงความเป็นไปได้ในการพัฒนาโปรแกรมค้นคืนสารสนเทศข้ามภาษาสำหรับภาษาอังกฤษและมาลายาลัมภายในระยะเวลาอันสั้นด้วยทรัพยากรภาษาที่เหมาะสม

3.2 Cross-Lingual Information Retrieval Model for Vietnamese-English Web Sites [13]

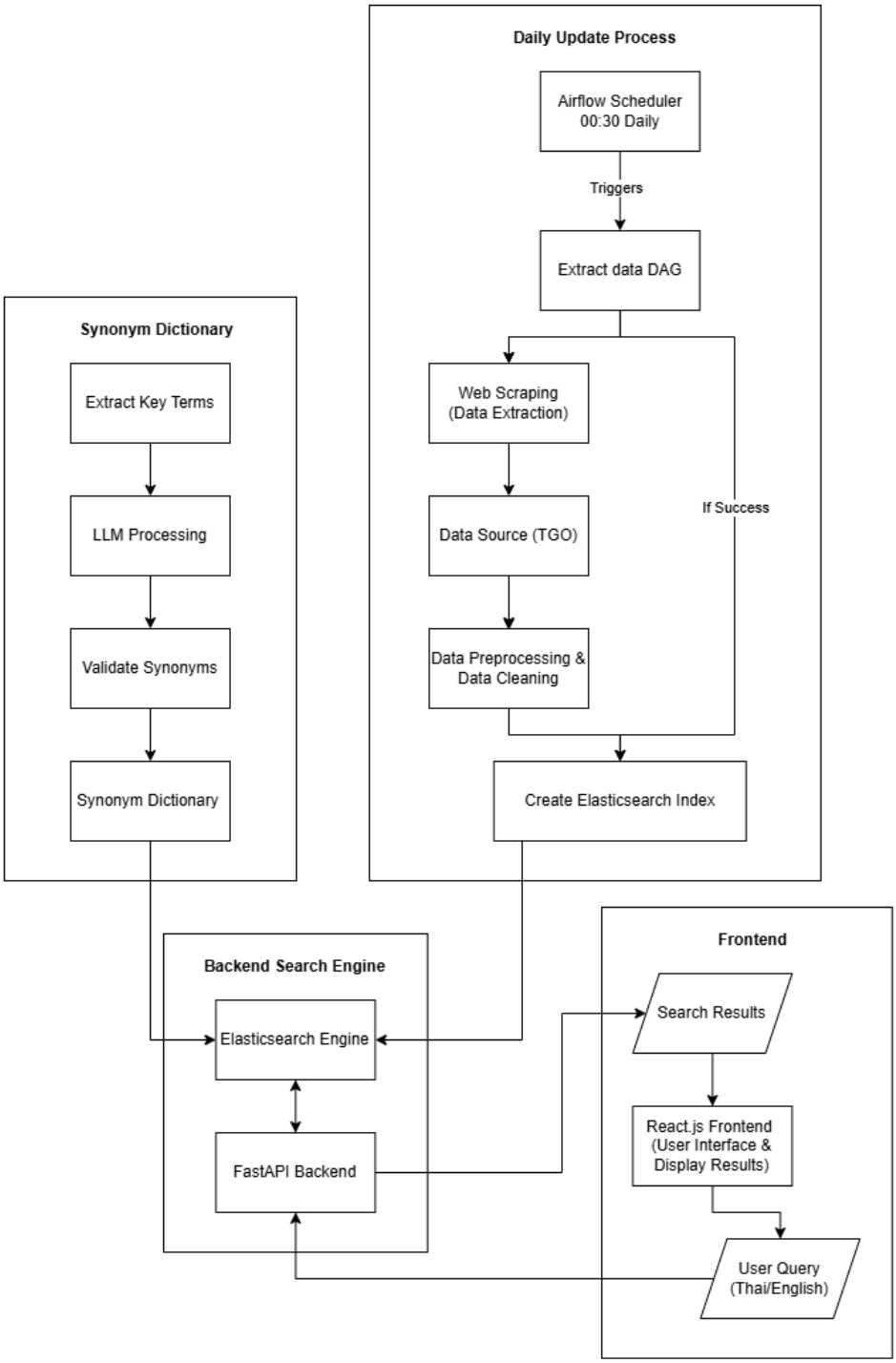
งานวิจัยนี้นำเสนอโมเดล CLIR สำหรับเว็บไซต์สองภาษาที่รองรับภาษาเวียดนามและอังกฤษ ระบบนี้ประกอบด้วย 4 ส่วนหลัก: Web Crawler สำหรับรวบรวมข้อมูล, Translated Document Identifying เพื่อระบุหน้าเว็บคู่แปล, Indexing จัดทำดัชนีแยกตามภาษา และ Searching รองรับการสืบค้นข้อมูลอย่างมีประสิทธิภาพ โมเดลนี้ช่วยลดการประมวลผลซ้ำ และเพิ่มความแม่นยำในการค้นหาโดยใช้ผลการระบุหน้าเว็บคู่แปล ข้อดีของระบบคือ การจัดเก็บข้อมูลในตัวเอง ลดความจำเป็นในการประมวลผลซ้ำ และค้นหาได้แม่นยำมากขึ้น แต่มีข้อจำกัดคือ จำนวนเว็บไซต์สองภาษาในปัจจุบันยังน้อย และไม่รองรับการระบุเอกสารคู่แปลที่อยู่คนละเว็บไซต์ ในอนาคต ระบบอาจถูกพัฒนาให้รองรับการค้นหาเชิงความหมายเพื่อเพิ่มประสิทธิภาพและความครอบคลุม

4. แนวคิดและวิธีการวิจัย

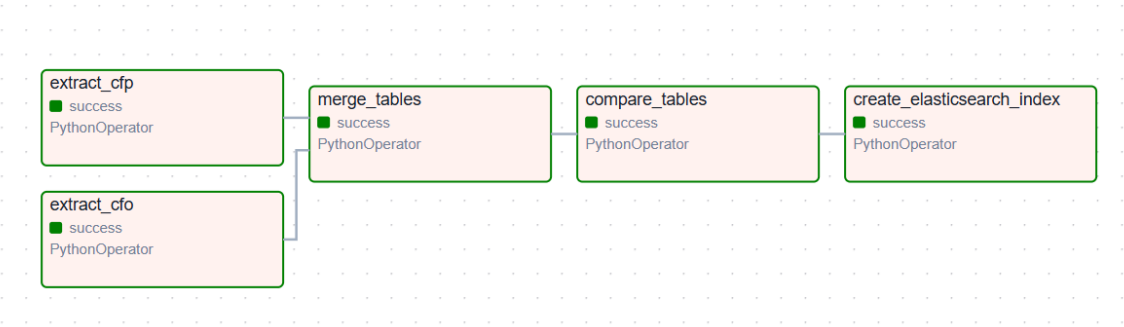
ภาพที่ 2 แสดงกระบวนการวิจัย โดยเริ่มจากการเตรียมข้อมูล EF จาก องค์การบริหารจัดการก๊าซเรือนกระจก (องค์การมหาชน) จากนั้น ทำการตรวจสอบและจัดการความผิดปกติของข้อมูลให้อยู่ในรูปแบบที่เหมาะสม ทำการตั้งค่าการค้นหาของข้อมูลและนำเข้าข้อมูลสู่ Elasticsearch เพื่อแสดงผลต่อไป

4.1 การเก็บรวบรวมข้อมูล (Data Collection)

Airflow Scheduler มีบทบาทสำคัญในกระบวนการเก็บรวบรวมข้อมูล EF จากเว็บ TGO โดยทำหน้าที่จัดการและกำหนดเวลาการทำงานของของระบบแบบอัตโนมัติ โดยเมื่อทำการรันเส้นทางการประมวลผลเพื่อทำในระบบอัปเดตข้อมูลของทุกวันเวลาเที่ยงคืนสามสิบนาที



ภาพที่ 2 ขั้นตอนระเบียบวิจัย



ภาพที่ 3 การทำงานของ data pipeline

4.2 การประมวลผลข้อมูลก่อน (Data Preprocessing)

4.2.1 การดึงข้อมูล (Data Extraction)

ภาพที่ 4 แสดงตัวอย่างตารางค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก (Green House Gas Emission Factor) ที่ได้มาจากองค์การบริหารจัดการก๊าซเรือนกระจก (องค์การมหาชน) ซึ่งจะมีหลายตารางและค่าปล่อยก๊าซเรือนกระจกอีกหลายประเภทที่จะต้องจัดการโดยจะจัดเก็บจะอยู่ในรูปแบบ CSV เพื่อนำไปประมวลผลต่อ

กลุ่ม	ลำดับ	ชื่อ	รายละเอียด	หน่วย	ค่าแฟคเตอร์ (kgCO2e)	ข้อมูลอ้างอิง	วันที่อัพเดท
กลุ่มปิโตรเคมี	1.0	Acrylonitrile Butadiene Styrene (ABS)	ผลิตจากกระบวนการอัลคิลเลชันของเบนซีนและเอทีลีน...	kg	4.1597	Thai National LCI Database, TIIS-MTEC-NSTDA (w...	Update_Dec2019
กลุ่มปิโตรเคมี	2.0	General Purposed Polystyrene (GPPS)	ผลิตจาก Styrene และ Ethylbenzene; LCIA method ...	kg	3.2281	Thai National LCI Database, TIIS-MTEC-NSTDA (w...	Update_Dec2019
กลุ่มปิโตรเคมี	3.0	High Density Polyethylene (HDPE)	ผลิตจาก Ethylene โดยมี 1-Butene และ Propylene ...	kg	6.7071	Thai National LCI Database, TIIS-MTEC-NSTDA (w...	Update_Dec2019
กลุ่มปิโตรเคมี	4.0	High Impact Polystyrene (HIPS)	ผลิตจาก Styrene และ Polybutadiene rubber; LCIA...	kg	3.6843	Thai National LCI Database, TIIS-MTEC-NSTDA (w...	Update_Dec2019
กลุ่มปิโตรเคมี	5.0	Linear Low Density Polyethylene (LLDPE)	ผลิตจากกระบวนการที่เป็น Solution phase และ Gas...	kg	2.1356	Thai National LCI Database, TIIS-MTEC-NSTDA (w...	Update_July2022

ภาพที่ 4 ตัวอย่างตารางค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

(https://thaicarbonlabel.tgo.or.th)

4.2.2 ตรวจหาความผิดปกติของข้อมูล

ทำการกรองข้อมูลที่ไม่จำเป็นออก เช่น ลบแถวที่ข้อมูลขาดหาย

4.2.3 การจัดการรูปแบบของข้อมูล

เนื่องจากลักษณะของข้อมูลที่ได้มาจะมีลักษณะรูปแบบที่ไม่เหมือนกัน จำเป็นต้องจัดการให้ข้อมูลอยู่ในรูปแบบที่ต้องการแบบเดียวกัน เพื่อเพิ่มความสะดวกในการทำงาน

4.2.4 การสร้างดัชนีใน Elasticsearch

ข้อมูลที่ผ่านการเตรียมและประมวลผลก่อนแล้วจะถูกจัดเก็บไว้ใน Elasticsearch Index เพื่อให้การค้นหาสามารถรองรับคำค้นที่หลากหลาย ทั้งภาษาไทยและภาษาอังกฤษและค้นหาได้อย่างมีประสิทธิภาพ ประกอบด้วย

- โครงสร้างดัชนี (Index Structure) จัดระเบียบข้อมูลตามหมวดหมู่ เช่น ชื่อสารเคมี,หน่วยวัด
- การกำหนดตัววิเคราะห์ภาษา (Analyzer) ระบุวิธีการตัดคำและวิเคราะห์คำสำหรับภาษาไทยและอังกฤษ

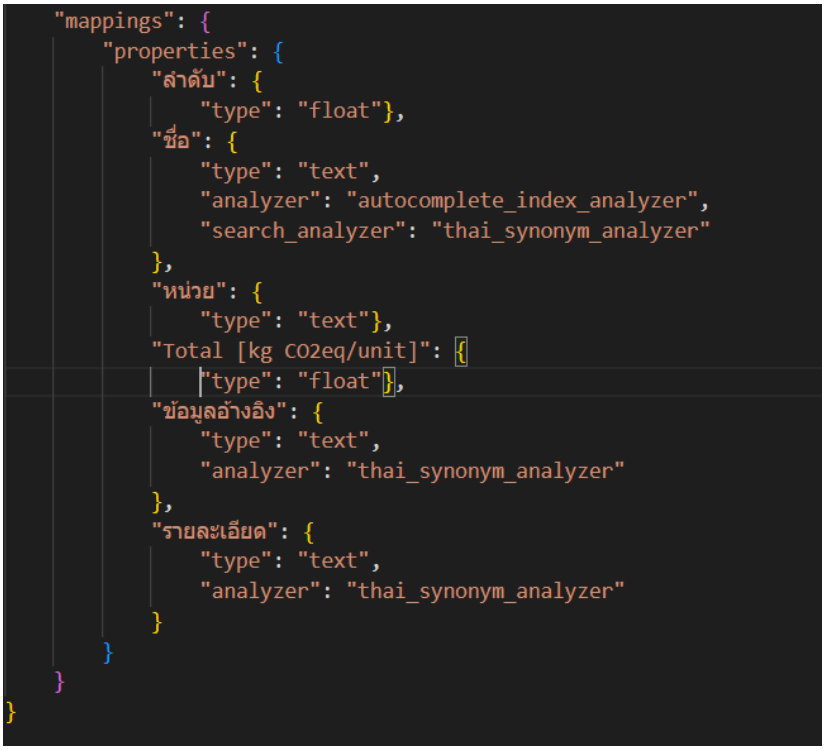
- การใช้การกรองโทเคนคำพ้องความ (Synonym Token Filter) ระหว่างการค้นหา

```
index_settings = {
"settings": {
        "analysis": {
            "filter": {
                "thai_english_synonym_filter": {
                    "type": "synonym",
                    "synonyms_path": "analysis/synonyms.txt"
                },
                "edge_ngram_filter": {
                    "type": "edge_ngram",
                    "min_gram": 1,
                    "max_gram": 20,
                    "token_chars": ["letter", "digit", "whitespace"]
                }
            },
            "analyzer": {
                "autocomplete_index_analyzer": {
                    "type": "custom",
                    "tokenizer": "standard",
                    "filter": ["lowercase","icu_folding","edge_ngram_filter"]
                },
                "autocomplete_search_analyzer": {
                    "type": "custom",
                    "tokenizer": "standard",
                    "filter": ["lowercase","icu_folding"]
                },
                "thai_synonym_analyzer": {
                    "type": "custom",
                    "tokenizer": "standard",
                    "filter": ["lowercase","icu_folding","thai_english_synonym_filter"]
                }
            }
        }
    },
```

ภาพที่ 5 การตั้งค่าต่างๆเพื่อสร้าง index สำหรับการค้นหา

การตั้งค่า search index ประกอบด้วย filter และ analyzer ดังภาพที่ 5

1. Filters เป็นกระบวนการปรับแต่ง Token ที่ได้รับจาก Tokenizer เพื่อการค้นหาที่เหมาะสมยิ่งขึ้น ประกอบด้วย thai_english_synonym_filter สำหรับแปลงคำพ้องความ ในที่นี้จะเป็นคำที่พ้องความทั้งภาษาไทยและอังกฤษ เช่น Anthracite ↔ แอนทราไซต์ ↔ ถ่านหินแข็ง และ edge_ngram_filter ใช้สำหรับการแสดงหรือแนะนำคำที่มีความเป็นไปได้ต่อการค้นหานั้นๆ
2. Analyzers คือชุดการประมวลผลข้อความที่ประกอบด้วย Tokenizer และ Filters เช่น thai_synonym_analyzer สำหรับวิเคราะห์ข้อความภาษาไทยและคำพ้องความ



ภาพที่ 6 การตั้งค่าต่างๆเพื่อสร้าง index สำหรับการค้นหา

ภาพที่ 6 แสดงการกำหนดโครงสร้างของเอกสารที่จัดเก็บใน Index ประกอบด้วย ชื่อคอลัมน์ ลักษณะข้อมูล เพื่อให้ analyzer หรือ search_analyzer ทราบว่าจะมีลักษณะการค้นหาอย่างไร เช่น คอลัมน์ “ชื่อ” จะมีการใช้ทั้ง autocomplete_index_analyzer ซึ่งจะทำการวิเคราะห์ตอนค้นหา และใช้ thai_synonym_analyzer วิเคราะห์เพื่อรองรับคำพ้องความหมายในเวลาเดียวกัน

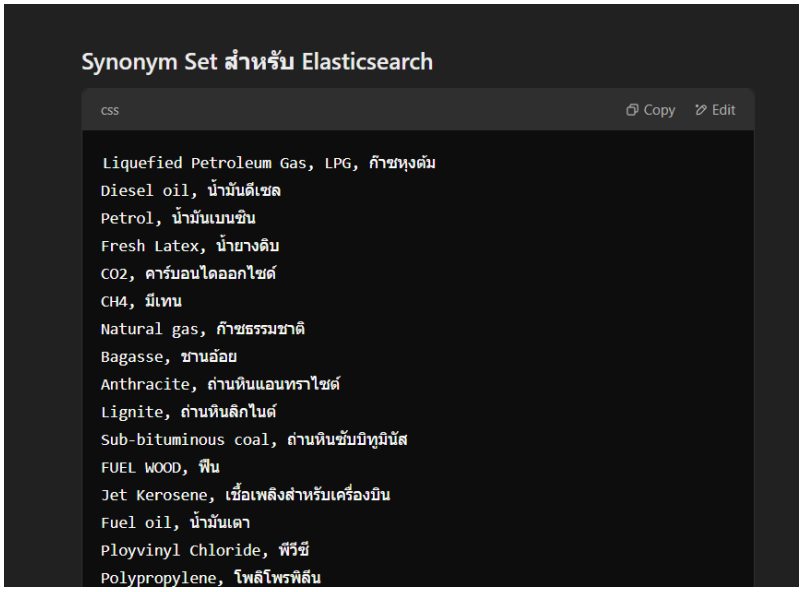
4.3 การสร้างคลังคำพ้องความ (Synonym Dictionary)

สร้างชุดคำพ้องความระหว่างคำศัพท์ภาษาไทยและอังกฤษ โดยคลังคำศัพท์นี้จะถูกนำไปใช้ในขั้นตอนการสืบค้นเพื่อให้ได้ผลลัพธ์ที่ครอบคลุมทั้งสองภาษาซึ่งเป็นกระบวนการรวบรวมคำศัพท์หรือวลีที่มีความหมายเหมือนกันหรือสามารถใช้แทนกันได้ ทั้งในภาษาไทย ภาษาอังกฤษ รวมทั้วตัวย่อหรือสัญลักษณ์ทางเคมีอยู่ด้วยโดยทั้งหมดล้วนแล้วแต่ให้ความหมายเหมือนกันทั้งสิ้น คลังคำพ้องความหมายนี้มีบทบาทสำคัญในการเพิ่มความถูกต้องและครอบคลุมในการสืบค้นข้อมูล เช่น

- Liquified Petroleum Gas, LPG, ก๊าซปิโตรเลียมเหลว, ก๊าซหุงต้ม
- Anthracite, แอนทราไซต์, ถ่านหินแข็ง

กระบวนการที่ได้มาซึ่งชุดคำพ้องความหมายนี้มาจากการใช้ Large Language Model (LLM) ในการค้นหาและประมวลผลคำศัพท์ที่เกี่ยวข้อง โดยเริ่มต้นจากการป้อนข้อมูลคำศัพท์จากตารางแสดงค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกเข้าสู่ LLM เพื่อแปลและเทียบความหมายคำศัพท์ในภาษาไทยและภาษาอังกฤษ ภาพที่ 7 แสดงผลลัพธ์จากคำสั่งหรือ Prompt ที่ใช้คือ

“ช่วยสร้าง Synonym Dictionary แบบ Synonym Set สำหรับคำศัพท์ในหมวดเดียวกันที่มีความหมายเหมือนกันทั้งภาษาไทยและอังกฤษรวมทั้งตัวย่อถ้ามี และให้จัดผลลัพธ์ในรูปแบบที่ Elasticsearch รองรับ เช่น ก๊าซหุงต้ม, LPG, Liquified Petroleum Gas”



ภาพที่ 7 ผลลัพธ์ที่ได้จากการป้อนคำสั่ง

รูปแบบการตั้งค่าที่ Elasticsearch รองรับจะมีอยู่ 3 แบบ

- แบบพ้องความหมายทั่วไป (Synonym Set) คำในบรรทัดเดียวกันจะถือว่ามีความหมายเหมือนกันทั้งหมด สามารถใช้แทนกันได้ในทุกกรณี เช่น word1, word2, word3
- แบบทิศทางเดียว (One-Way Synonym) กำหนดให้คำหนึ่งถูกแทนที่ด้วยอีกคำหนึ่งเสมอ เช่น word1 => word2
- แบบสองทิศทาง (Bi-Directional Synonym) กำหนดให้คำสองคำสามารถแทนที่กันได้ทั้งสองทาง เช่น word1 <=> word2

4.4 Backend

พัฒนาด้วย FastAPI เป็นตัวกลางเชื่อมระหว่าง Frontend และ Elasticsearch เพื่อทำหน้าที่ user query processing รับคำค้นจากผู้ใช้ (ภาษาไทยหรืออังกฤษ) ผ่าน API ส่งคำค้นไปยัง Elasticsearch และรับผลลัพธ์กลับมา จากนั้นประมวลผลผลลัพธ์และจัดเรียงตามความเกี่ยวข้องก่อนส่งไปยัง Frontend

4.5 Frontend

พัฒนาด้วย React.js เพื่อให้ผู้ใช้สามารถสืบค้นข้อมูลได้สะดวก โดยสร้างช่องกรอกคำค้น (Search Box) ที่รองรับภาษาไทยและอังกฤษ แสดงผลลัพธ์การค้นหาในรูปแบบที่เข้าใจง่าย เช่น ตาราง รองรับการกรองข้อมูล (Filter) ตามหมวดหมู่ เช่น ปี, ประเภทสาร

5. วัตถุประสงค์

พัฒนาระบบสืบค้นข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกที่รองรับการค้นหาได้ทั้งภาษาไทยและภาษาอังกฤษ เพื่ออำนวยความสะดวกการค้นหาข้อมูลได้อย่างรวดเร็วและถูกต้องตามปีการประเมิน

6. ขอบเขตการดำเนินงาน

6.1. ใช้ข้อมูล EF ที่เผยแพร่โดย องค์การบริหารจัดการก๊าซเรือนกระจก (องค์การมหาชน)

6.2. ออกแบบให้สืบค้นได้ทั้งภาษาไทยและอังกฤษด้วย Synonym-based Approach

6.3. ระบบสืบค้นที่ยืดหยุ่นสามารถอัปเดตข้อมูลได้อัตโนอัติเมื่อองค์การบริหารจัดการก๊าซเรือนกระจกมีการเปลี่ยนแปลงข้อมูล

6.4. ประเมินระบบด้วยตัววัดมาตรฐานการค้นคืนสารสนเทศ

7. ขั้นตอนการดำเนินงาน

7.1. ศึกษาค้นคว้าทฤษฎีและงานวิจัยที่เกี่ยวข้อง

7.2. จัดเตรียมข้อมูล

7.3. สร้างและปรับแต่งประสิทธิภาพโปรแกรม

7.4. ทดสอบและประเมินผล

7.5. วิเคราะห์ สรุปผลการดำเนินงาน

7.6. เรียบเรียงและจัดทำบทความวิจัย

7.7. จัดทำวิทยานิพนธ์

8. ประโยชน์ที่คาดว่าจะได้รับ

8.1. โปรแกรมค้นหาข้ามภาษาที่ใช้งานจริง ช่วยให้ผู้ใช้สามารถสืบค้นค่า EF ได้สะดวก และเลือกใช้ค่าที่ถูกต้อง ณ เวลาประเมินรายงานคาร์บอนฟุตพรินต์

8.2. ส่งเสริมการรายงานคาร์บอนฟุตพรินต์ โดยเฉพาะธุรกิจ SME

9. รายการอ้างอิง

[1] ว. ศรเลิศล้ำวาณิช, "การค้นคืนสารสนเทศ," สำนักพิมพ์มหาวิทยาลัยธรรมศาสตร์, 2018.

[2] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," *Information processing & management*, vol. 41, no. 3, pp. 523-547, 2005.

[3] T. Grainger, D. Turnbull, and M. Irwin, *AI-Powered Search*. Simon and Schuster, 2025.

[4] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information fusion*, vol. 36, pp. 10-25, 2017.

[5] T. G. G. M. O. P. Organization). "Thai Carbon Label." https://www.tgo.or.th/2023/index.php/th/ (accessed January 15, 2025.

[6] O. Nexus. "OpenLCA Nexus Databases." https://nexus.openlca.org/databases (accessed January 15, 2025.

[7] I. P. o. C. C. (IPCC). "IPCC Emission Factor Database (EFDB)." https://www.ipcc-nggip.iges.or.jp/EFDB/main.php (accessed 15 January 2025).
[8] M. Konda, *Elasticsearch in Action*. Simon and Schuster, 2024.
[9] S. Ramírez. "FastAPI Documentation." Tiangolo. https://fastapi.tiangolo.com/ (accessed January 14, 2025, 2025).
[10] I. Meta. "React: A JavaScript library for building user interfaces." Meta https://reactjs.org/ (accessed January 14, 2025, 2025).
[11] A. S. Foundation. "Apache Airflow: Platform to programmatically author workflows." https://airflow.apache.org/ (accessed January 20, 2025, 2025).
[12] P. Nikesh, S. M. Idicula, and S. D. Peter, "English-Malayalam cross-lingual information retrieval—an experience," in *2008 IEEE International Conference on Electro/Information Technology*, 2008: IEEE, pp. 271-275.
[13] A. F. Abka, M. Pratama, and W. Jatmiko, "Cross-Lingual Summarization: English-Bahasa Indonesia," in *2021 6th International Workshop on Big Data and Information Security (IWBIS)*, 2021: IEEE, pp. 53-58.