โปรแกรมค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

A CROSS-LINGUAL SEARCH ENGINE FOR RETRIEVAL OF GREENHOUSE GAS EMISSION FACTORS

จัดทำโดย นายณฐพจน์ หนูวงษ์
รหัสนิสิต 6770233221

นิสิตปริญญาโท สาขาสาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาคนอกเวลาราชการ
อาจารย์ที่ปรึกษา รศ.ดร.ญาใจ ลิ่มปิยะกรณ์

Outline

1.ที่มาและความสำคัญของปัญหา

2.ทฤษฎีที่เกี่ยวข้อง

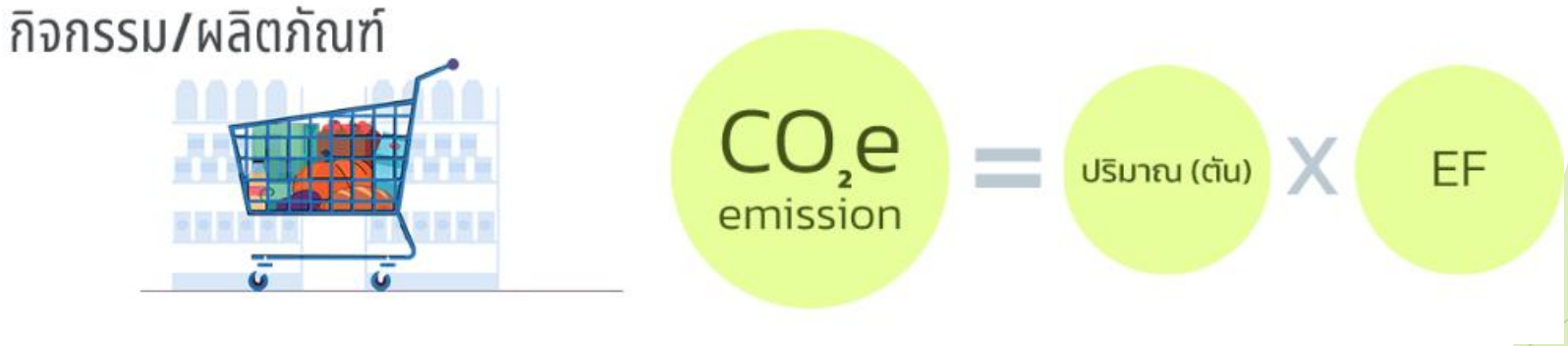
3.แนวคิดและวิธีการวิจัย

4.การทดลองและวิธีการวิจัย

5.ประเมินงานผลวิจัยและสรุป ข้อจำกัด ข้อเสนอแนะ

1.ที่มาและความสำคัญของปัญหา

ค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก (GHGs Emission Factor: EF) เป็นค่าการปล่อยก๊าซเรือนกระจกจากการผลิตหรือการบริการ ที่คิดรวมค่าการปล่อยก๊าซเรือนกระจกที่ก่อให้เกิดภาวะโลกร้อน (Climate Change) อาทิ ก๊าซคาร์บอนไดออกไซด์ (CO2) ก๊าซมีเทน (CH4) ก๊าซไนตรัสออกไซด์ (N2O) เป็นต้น ค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกนี้ มีความสำคัญต่อการประเมินคาร์บอนฟุตพรินต์เป็นอย่างมาก



https://circularmaterialhub.com/Calculate.php

1.ที่มาและความสำคัญของปัญหา

เนื่องจากการค้นหาค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกบนเว็บไซต์ขององค์การบริหารจัดการก๊าซเรือนกระจกไม่ยืดหยุ่น โดยจะต้องค้นหาคำที่เฉพาะเจาะจงเท่านั้น ดังนั้นจึงมีความสนใจที่จะทำให้การค้นหานั้นยืดหยุ่นขึ้นโดยสามารถที่จะค้นหาได้ทั้งภาษาไทยและภาษาอังกฤษที่มีความหมายเหมือนกัน หรือระบบค้นหาแบบ2 ภาษา

Emission Factor (CFP)

Emission Factor (CFP) ทั้งหมด 1 รายการ

sulfur

กลุ่ม	ลำดับ	ชื่อ	รายละเอียด	หน่วย	ค่าแฟคเตอร์ ($kgCO_2e$)	ข้อมูลอ้างอิง	วันที่อัพเดท
กลุ่มเคมีภัณฑ์ (Chemicals)	1.	Sulfur	Sulfur from Refinery	kg	0.4529	Thai National LCI Database/MTEC	Update_24Sep12

Emission Factor (CFP)

Emission Factor (CFP) ทั้งหมด 1 รายการ

ซัลเฟอร์

กลุ่ม	ลำดับ	ชื่อ	รายละเอียด	หน่วย	ค่าแฟคเตอร์ ($kgCO_2e$)	ข้อมูลอ้างอิง	วันที่อัพเดท
กลุ่มพลังงาน: เชื้อเพลิงเหลว และเชื้อเพลิงแข็ง	1.	ซัลเฟอร์	ซัลเฟอร์ที่ได้จากกระบวนการกลั่นน้ำมันดิบ; LCIA method IPCC 2013 GWP 100a V1.03	kg	0.2390	Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)	Update_Dec2019

https://thaicarbonlabel.tgo.or.th/index.php?lang=TH&mod=Y0hKdlpIVmpkSE5mWlcxcGMzTnBiMjQ9

ขอบเขตการวิจัย

1.ใช้ข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกที่เผยแพร่โดย องค์การบริหารจัดการก๊าซเรือนกระจก (องค์การมหาชน)

2.ออกแบบให้สืบค้นได้ทั้งภาษาไทยและอังกฤษด้วย Synonym-based Approach

3.ระบบสืบค้นที่ยืดหยุ่นสามารถอัปเดตข้อมูลได้อัตโนอัติเมื่อองค์การบริหารจัดการก๊าซเรือนกระจกมีการเปลี่ยนแปลงข้อมูล

4.ประเมินระบบด้วยตัววัดมาตรฐานการค้นคืนสารสนเทศ

ประโยชน์ที่คาดว่าจะได้รับ

1. โปรแกรมค้นหาข้ามภาษาที่ช่วยให้ผู้ใช้สามารถค้นคืนค่า EF ได้สะดวก และเลือกใช้ค่าที่ถูกต้อง ณ เวลาประเมินรายงานคาร์บอนฟุตพรินต์

2.ส่งเสริมการรายงานคาร์บอนฟุตพรินต์ โดยเฉพาะธุรกิจ SME

2.ทฤษฎีที่เกี่ยวข้อง

การค้นคืนสารสนเทศข้ามภาษา (Cross-Lingual Information Retrieval - CLIR) เป็นเทคนิคที่ช่วยให้ผู้ใช้สามารถค้นหาข้อมูลโดยใช้ภาษาใดภาษาหนึ่ง แต่ยังสามารถดึงข้อมูลจากเอกสารที่เขียนในภาษาอื่น ๆ ได้ โดยไม่จำเป็นต้องแปลข้อความทั้งหมดด้วยตนเอง โดยจะสร้างคลังคำพ้อง (Synonym Dictionary) ที่จับคู่คำในภาษาไทยและภาษาอังกฤษ

ขั้นตอนการทำงาน

1. Tokenization & Normalization: แยกคำในภาษาไทยและภาษาอังกฤษเพื่อให้ระบบเข้าใจคำที่ต้องการค้นหา

2. Synonym Matching: ขยายคำค้นโดยใช้ Synonym Dictionary เช่น

LPG ⟷ Liquified Petroleum Gas ⟷ ก๊าซหุงต้ม

3. Indexing & Searching: ใช้ Elasticsearch จัดเก็บข้อมูลและกำหนดให้ค้นหาผ่านชุดคำพ้อง

2.ทฤษฎีที่เกี่ยวข้อง

Tokenization คือกระบวนการแบ่งข้อความออกเป็นหน่วยย่อย (Tokens) เช่น คำ วลี หรืออักขระ ซึ่งช่วยให้ระบบสามารถวิเคราะห์และทำการค้นหาได้อย่างถูกต้อง

ข้อความต้นฉบับ	ผลลัพธ์ของ Tokenization
"ก๊าซเรือนกระจกสูงขึ้น"	["ก๊าซเรือนกระจก", "สูง", "ขึ้น"]
"การค้นคืนสารสนเทศข้ามภาษา"	["การ", "ค้นคืน", "สารสนเทศ", "ข้าม", "ภาษา"]

สำหรับภาษาไทยการตัดคำเป็นเรื่องที่ท้าทายเนื่องจาก **ไม่มีการเว้นวรรคระหว่างคำ** เหมือนภาษาอังกฤษ เช่น "ก๊าซเรือนกระจก" ควรจะเป็น 1 คำ แต่ระบบทั่วไปอาจตัดเป็น ["ก๊าซ", "เรือน", "กระจก"] ซึ่งอาจทำให้ผลลัพธ์การค้นหาไม่ถูกต้อง

2.ทฤษฎีที่เกี่ยวข้อง

ICU Tokenizer เป็นตัวตัดคำที่ใช้ **International Components for Unicode (ICU)** ซึ่งรองรับการตัดคำในหลายภาษา รวมถึงภาษาไทยโดย**อาศัยกฎทางภาษาศาสตร์และโมเดลสถิติ** แทนการใช้พจนานุกรมแบบตายตัว

หลักการทำงานของ ICU Tokenizer

1.ใช้กฎทางภาษาศาสตร์ (Rule-based Tokenization)

1. วิเคราะห์โครงสร้างประโยคและบริบทของคำ
2. รองรับภาษาที่ไม่มีการเว้นวรรค เช่น ภาษาไทย ญี่ปุ่น จีน

2.ใช้โมเดลสถิติช่วยในการตัดคำ (Statistical Model)

1. แยกคำโดยดูจากความน่าจะเป็นของการเกิดขึ้นของคำ
2. หากไม่มีคำในพจนานุกรม ระบบจะพิจารณาความเป็นไปได้ของการเป็นคำ

Synonym-based (Dictionary-based)

ใช้คลังคำศัพท์คู่ (Bilingual Dictionary) บรรจุรายการคำพ้องความหมายที่จับคู่คำหรือวลีสำคัญในภาษาไทยและภาษาอังกฤษไว้ ซึ่ง คำพ้องความ (Synonyms) มีความหมายคือคำหรือวลี(กลุ่มคำ) ที่สามารถแทนกันได้ โดยแสดงสิ่งที่เหมือนกันหรือใกล้เคียงกัน แม้ว่าจะมีความแตกต่างเล็กน้อยในบริบท แม้กระทั่งตัวย่อหรือสูตรทางเคมี

English Term	Thai Term
Agriculture	การเกษตร, เกษตรกรรม
Anthracite	ถ่านหินแอนทราไซต์
Bagasse	ชานอ้อย
Biogas	ก๊าซชีวภาพ
Cob	ซังข้าวโพด
CO2	คาร์บอนไดออกไซด์
LPG, Liquified Petroleum Gas	ก๊าซหุงต้ม

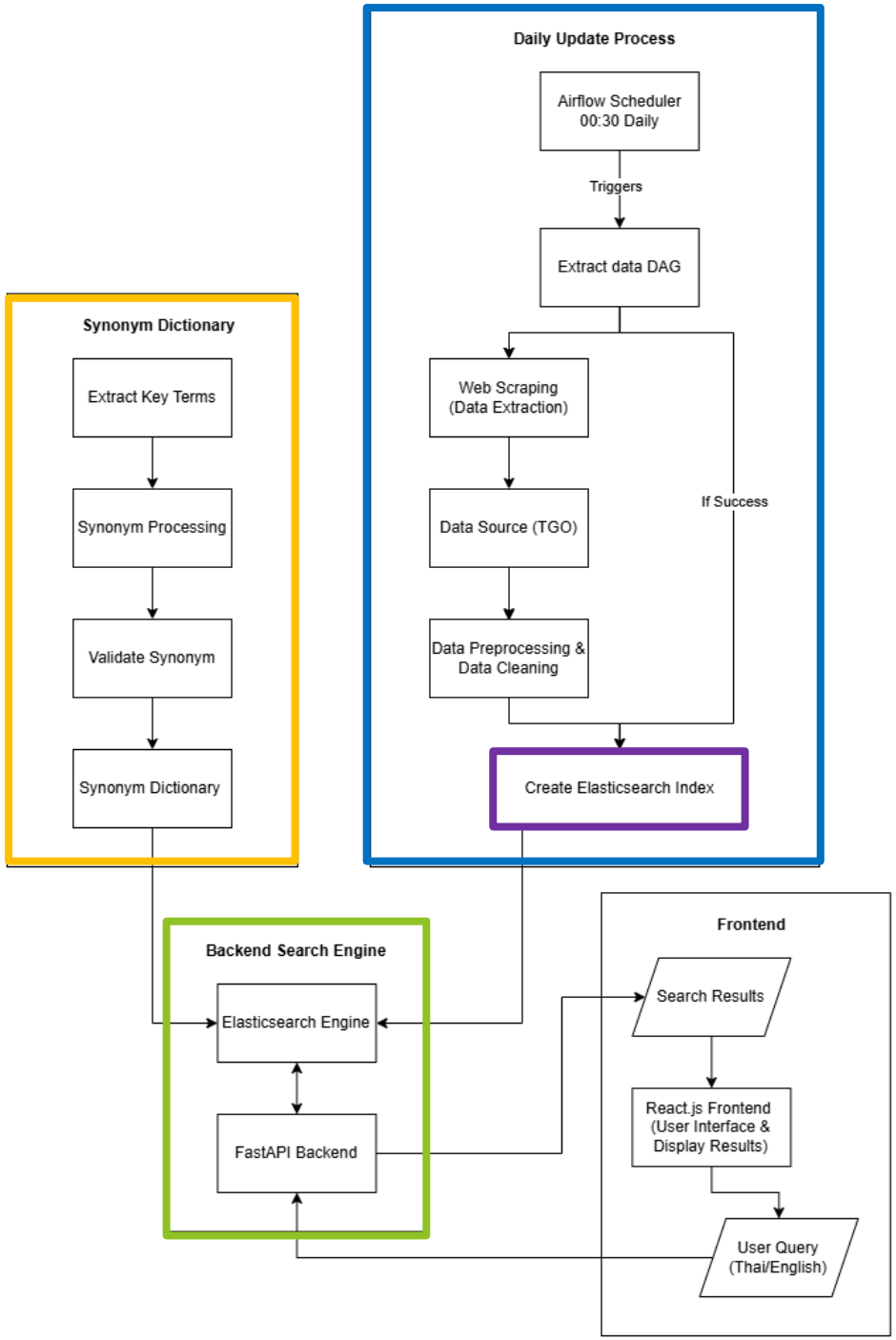
เทคโนโลยีและแพลตฟอร์มที่ใช้

Elasticsearch มีคุณสมบัติเหมาะสมกับการค้นคืนข้อมูลข้ามภาษา (Cross-Lingual Information Retrieval) โดยรองรับการทำงานที่ซับซ้อน เช่น การวิเคราะห์คำพ้องความหมาย (Synonym Matching) และการค้นหาแบบ Full-Text Search ได้อย่างมีประสิทธิภาพ

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

Apache Airflow แพลตฟอร์มสำหรับการสร้าง จัดการ ติดตาม Workflow

3.แนวคิดและวิธีการวิจัย

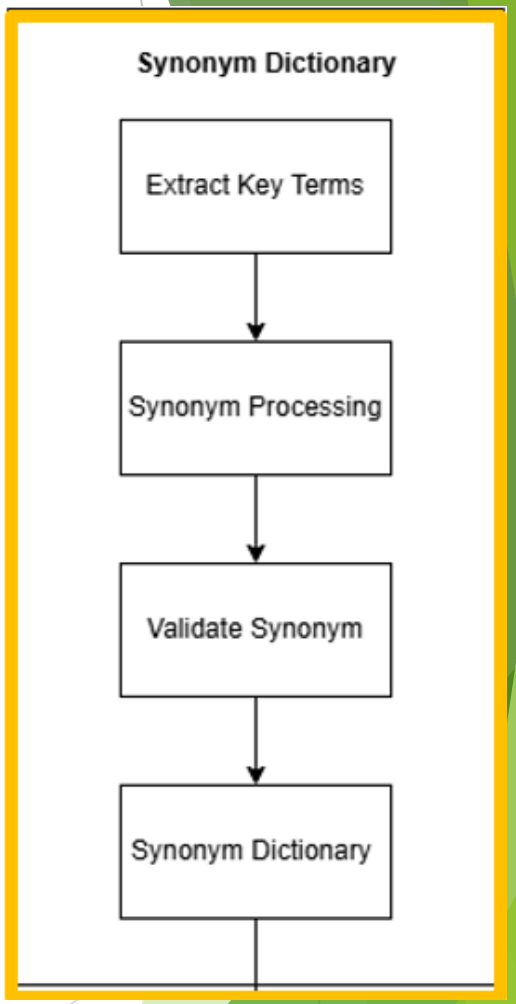


3.แนวคิดและวิธีการวิจัย

Synonym dictionary

มีขั้นตอนดังนี้

1. รวมข้อมูลจากทั้ง3 ตารางคือ CFP,CFO, Carbon Label มาไว้ด้วยกัน
2. Extract ข้อมูลมาจากตารางที่รวมด้วยโดย Extract มาจากชื่อเป็นหลักเพื่อทำการสร้างคำพ้องความหมาย
3. เมื่อได้คำทั้งหมดมาแล้วทำไมจับกลุ่ม เนื่องจากว่ารายงานที่มีทั้งหมดนั้นค่อนข้างเยอะแต่ จะเป็นกลุ่มคำที่คล้ายกัน
4. หลังจากนั้นก็ค้นหาคำค้นจากแหล่งอ้างอิงที่เชื่อถือได้ เช่น พจนานุกรมต่างๆ , ecoinvent หรือเป็นเว็บไซต์เกี่ยวกับด้านสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกด้วยเฉพาะ



3.แนวคิดและวิธีการวิจัย

Synonym dictionary

รูปแบบการจัดเก็บเพื่อจะเรียกใช้คำพ้องความหมายหรือ synonyms จะเก็บอยู่ในรูปแบบนี้ csv ไฟล์โดยแบบลักษณะคำที่เรียกว่าพ้องความหมายทั่วไป (Synonym Set) มีความหมายว่า “**คำในบรรทัดเดียวกันจะถือว่ามีความหมายเหมือนกันทั้งหมด สามารถใช้แทนกันได้ในทุกกรณี**” เช่น

Sulfur, กำมะถัน, ซัลเฟอร์, Sulphur

LPG, Liquified Petroleum Gas, ก๊าซหุงต้ม

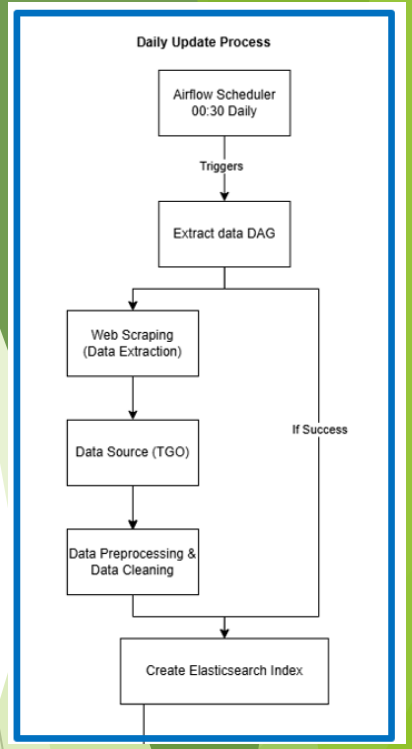
จากด้านตัวอย่างด้านบนมีความหมายว่า 1 บรรทัดจะแทนคำ 1 คำซึ่งแต่ละคำมีความหมายเท่ากันทั้งหมด กล่าวคือถ้าเราค้นหาคำว่า Sulfur แล้วในระบบมีทั้ง Sulfur และกำมะถัน ทั้ง2คำก็จะถูกแสดงขึ้นมาทั้งคู่

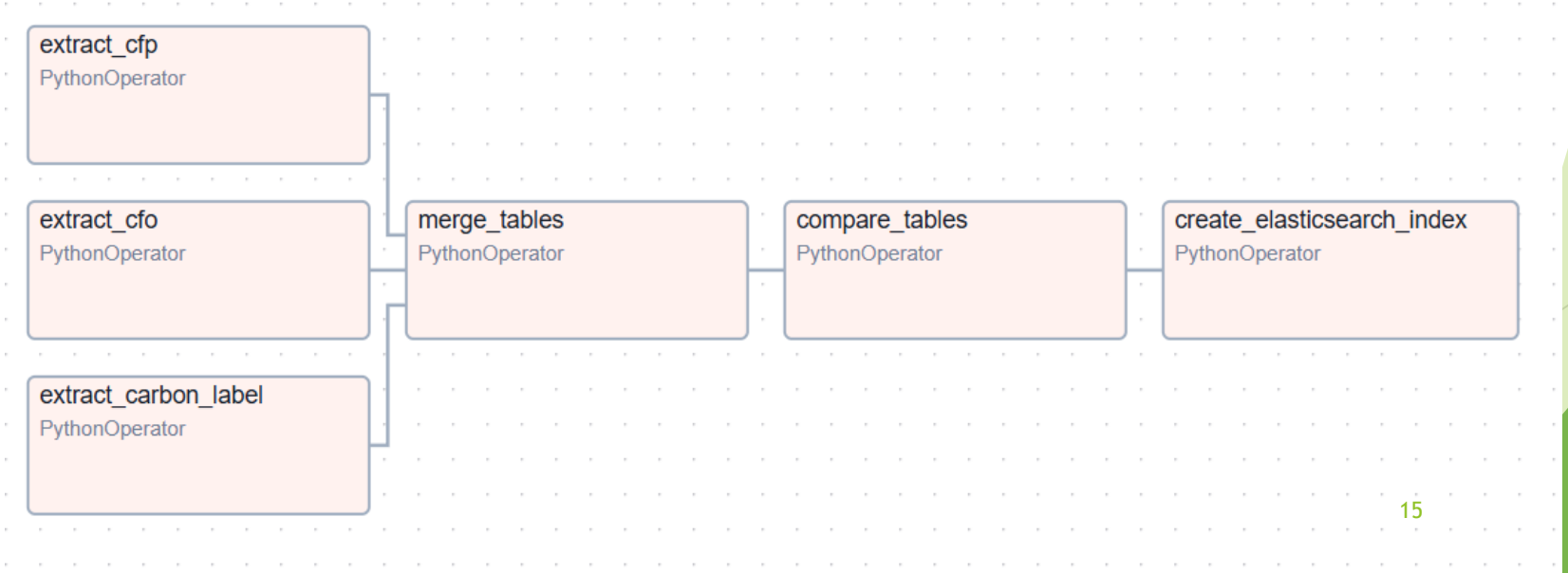
3.แนวคิดและวิธีการวิจัย

Daily Batch

Data pipeline ของการดึงขอข้อมูลจากเว็บไซต์ขององค์การบริหารจัดการก๊าซเรือนกระจก โดยจะแบ่งการทำงานออกเป็น 2 ส่วนใหญ่ๆคือการนำข้อมูลที่ได้มาจากองค์การบริหารจัดการก๊าซเรือนกระจก มาประมวลผล

จากนั้นอัปเดตข้อมูลในกรณีที่ข้อมูลเกิดการเปลี่ยนแปลงค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก





3.แนวคิดและวิธีการวิจัย

โดยจากรูป เป็นข้อมูลตัวอย่างที่แสดงหลังจากรวมข้อมูลทั้ง 3 ตารางเข้าไว้ด้วยกันแล้ว


```
[
    {
        "Category":"กลุ่มปิโตรเคมี",
        "Name":"Acrylonitrile Butadiene Styrene (ABS)",
        "Detail":"ผลิตจากกระบวนการอัลคิลเลชันของเบนซีนและเอทีลีน; LCIA method IPCC 2013 GWP 100a V1.03",
        "Unit":"kg",
        "Factor":4.1597,
        "Reference":"Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)",
        "Last_Updated":"Dec 2019"
    },
    {
        "Category":"กลุ่มปิโตรเคมี",
        "Name":"General Purposed Polystyrene (GPPS)",
        "Detail":"ผลิตจาก Styrene และ Ethylbenzene; LCIA method IPCC 2013 GWP 100a V1.03",
        "Unit":"kg",
        "Factor":3.2281,
        "Reference":"Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)",
        "Last_Updated":"Dec 2019"
    },
    {
        "Category":"กลุ่มปิโตรเคมี",
        "Name":"High Density Polyethylene (HDPE)",
        "Detail":"ผลิตจาก Ethylene โดยมี 1-Butene และ Propylene เป็น Comonomer; LCIA method IPCC 2013 GWP 100a V1.03",
        "Unit":"kg",
        "Factor":6.7071,
        "Reference":"Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)",
        "Last_Updated":"Dec 2019"
    },
```

3.แนวคิดและวิธีการวิจัย

Elasticsearch Index

งานวิจัยนี้จะนำเสนอการค้นหาแบบวิธี hybrid ซึ่งวิธีการค้นหาแบบ hybrid คือการรวมกันระหว่าง keyword search และ vector search

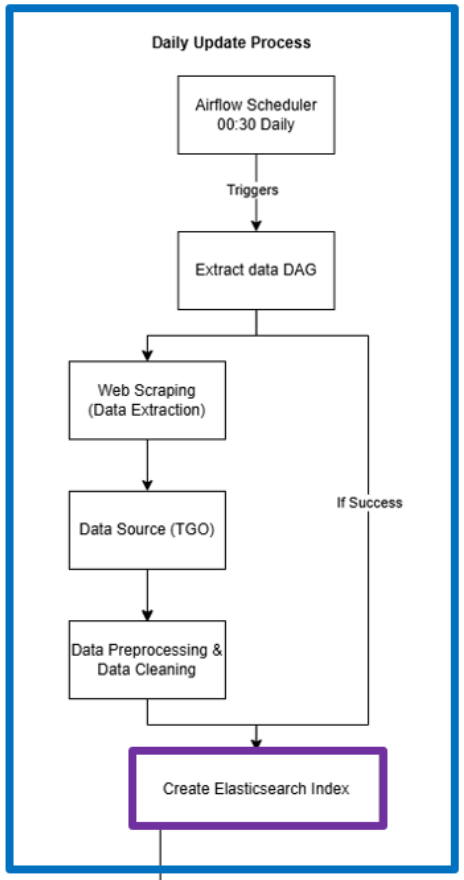
Keyword/Lexical Search คือ การค้นหาตามคำที่ตรงกันแบบตัวอักษร จะมีประสิทธิภาพดีเมื่อมีการควบคุมข้อมูลที่มีโครงสร้างและผู้ใช้มีความชัดเจนในสิ่งที่ต้องการค้นหา

Vector search คือ ค้นหาตามความหมายผ่าน Dense Vector เหมาะสำหรับการค้นหาแนวคิดที่คล้ายกัน และการค้นหาแบบสนทนาที่ผู้ใช้ไม่สามารถระบุคำค้นที่แม่นยำได้

3.แนวคิดและวิธีการวิจัย

Elasticsearch Index

องค์ประกอบหลัก	หน้าที่หลัก
Document	หน่วยข้อมูลพื้นฐานใน Elasticsearch (JSON Format)
Index	กลุ่มของ Documents ที่ใช้เก็บข้อมูล
Shards	แบ่งข้อมูลเป็นส่วนย่อยเพื่อกระจายโหลด
Mappings	กำหนดโครงสร้างข้อมูล เช่น ประเภทของฟิลด์
Analyzers	ช่วยตัดคำและจัดรูปแบบข้อมูล
Inverted Index	โครงสร้างข้อมูลที่ช่วยค้นหาข้อมูลเร็วขึ้น



3.แนวคิดและวิธีการวิจัย

```
index_settings = {
    "settings": {
        "analysis": {
            "filter": {
                "synonym_filter": {
                    "type": "synonym_graph",
                    "synonyms":thai_synonyms,
                    "expand": True,
                    "updateable": True
                }
            },
            "analyzer": {
                "thai_eng_analyzer": {
                    "type":"custom",
                    "tokenizer": "icu_tokenizer",
                    "filter": [
                        "lowercase",
                        "icu_folding"
                    ]
                },
                "thai_eng_search_analyzer":{
                    "type": "custom",
                    "tokenizer": "icu_tokenizer",
                    "filter": [
                        "lowercase",
                        "icu_folding",
                        "synonym_filter"
                    ]
                }
            }
        }
    },
```

ส่วนของการconfig เพื่อที่ใช้ในการสร้าง index ใน elasticsearch โดยจะแบ่งเป็น 3 ส่วนหลักๆ

synonym หรือคำพวกความหมาย

- types:synonym graph filter ที่รองรับ multi-word synonyms
- Synonyms: thai_synonyms เป็นตัว list ของ synonym ที่อ่านมาจาก csv ไฟล์
- Expand, updateable : สามารถอัพเดท synonym list ได้โดยไม่ต้อง reindex ข้อมูลใหม่

Analyzer เพื่อใช้ทั้งปรับแต่งคำ, แยกข้อความ, เตรียมข้อมูล

โดยใช่ icu_tokenizer เพื่อรองรับทั้งคำไทยและอังกฤษ และเรียกใช้ synonym_filter ที่ประกาศไว้ด้านบนมาใช้

3.แนวคิดและวิธีการวิจัย

```
    "mappings": {
        "properties": {
            "Category": {
                "type": "text"
            },
            "Name": {
                "type": "text",
                "analyzer": "thai_eng_analyzer",
                "search_analyzer": "thai_eng_search_analyzer"
            },
            "Unit": {
                "type": "text"
            },
            "Factor": {
                "type": "text"
            },
            "Reference": {
                "type": "text"
            },
            "Last Updated": {
                "type": "text"
            }
            ,
            "text_vector": {
                "type": "dense_vector",
                "dims": 768,
                "index": True,
                "similarity": "cosine"
            }
        }
    }
}
```

Mapping คือส่วนที่ประกาศโครงสร้างของ index ใน elasticsearch โดยจะมี column

Category, Name, Unit, Factor, Reference, Last Update ,text_vector

Name : จะเป็น column หลักที่เรียกใช้ analyzer ที่ประกาศใช้ไปก่อนหน้านี้ใช้รองรับทั้งการค้นหาแบบคำพ้องความหมาย การตัดคำ, และสามารถค้นได้ทั้งภาษาไทยและภาษาอังกฤษ

Text_vector: จะเป็น column ที่สร้างขึ้นเพื่อรองรับการค้นหาแบบ hybrid โดยจะใช้ model “intfloat/multilingual-e5-base” เพื่อ embedding ข้อความที่เราต้องการให้เปลี่ยนรูปแบบให้อยู่ใน vector ที่มีความยาวหลัง embedded 768 และใช้ cosine similarity ในการคำนวณความใกล้เคียง

3.แนวคิดและวิธีการวิจัย

การ embedding คือการแปลงข้อความให้เป็น **ตัวเลข vector** ที่คอมพิวเตอร์เข้าใจและประมวลผลได้ โดยคงความหมายและความสัมพันธ์ของข้อมูลเดิมไว้

เช่น "รถยนต์สีแดง" Vector: [0.2, -0.1, 0.8, 0.3, ..., 0.5] // 768 มิติ

หลังจากเมื่อทำการค้นหาคำเช่นๆ ก็จะทำการเป็น vector เหมือนกันแล้วมาเปรียบเทียบด้วยวิธี **cosine similarity**

ซึ่งค่าที่ออกมาจะมีค่าระหว่าง [-1,1] ซึ่งถ้ามีค่าใกล้ 1 แสดงว่าคำค้นนั้นๆมีค่าเหมือนกันกับผลลัพธ์มากๆ

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

ข้อมูลที่ใช้ในการทดลอง

ฐานข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกที่ใช้ในการทดลองประกอบด้วยข้อมูลทั้งหมด 9990 รายการ รวบรวมจากเว็บไซต์ อบก. ประกอบด้วย

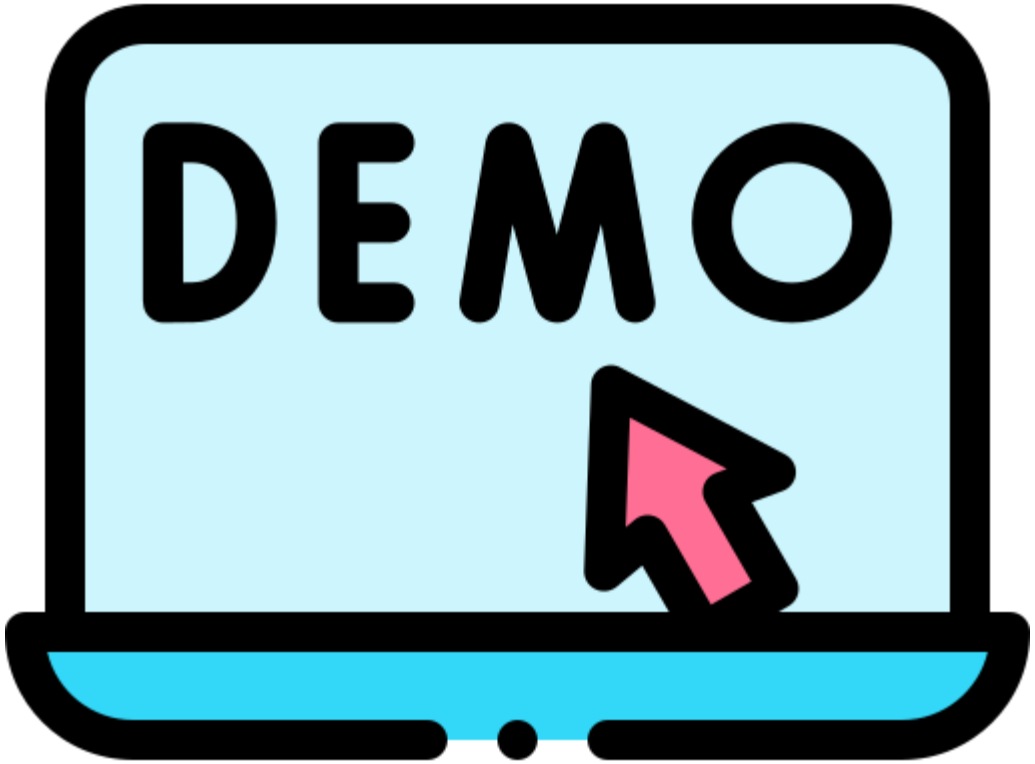
1) Carbon Footprint for Organization (CFO) จำนวน 49 รายการ
2) Carbon Footprint for Product (CFP) จำนวน 762 รายการ
3) Carbon Label Project จำนวน 9179 รายการ

4.การทดลองและวิธีการวิจัย

การประเมินสมรรถนะ (performance evaluation) ของระบบการค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกในงานวิจัยนี้ ใช้วิธีการเปรียบเทียบกับ Ground Truth การทดลองแบ่งออกเป็น 4 แบบ ดังนี้

1) Keyword matching: ค้นหาแบบใช้คำสำคัญอย่างเดียว
2) Keyword and synonyms matching: ค้นหาแบบใช้คำสำคัญร่วมกับคำพ้องความหมาย
3) Vector search: ค้นหาแบบเวกเตอร์
4) Hybrid search and synonyms matching: ค้นหาแบบผสมร่วมกับคำพ้องความหมาย

4.การทดลองและวิธีการวิจัย



https://cdn-icons-png.flaticon.com/512/5110/5110617.png

5.ประเมินงานผลวิจัยและสรุป ข้อจำกัด ข้อเสนอแนะ

ประเมินผลงานวิจัย ใช้ Precision เพื่อพิจารณาผลการค้นคืน 15 รายการแรก การประเมินจะพิจารณาความถูกต้องจากคำพ้องความหมายด้วย ไม่ใช่แค่ exact match

และ Recall ใช้การประเมินความครอบคลุมในของคำที่เกี่ยวข้องทั้งหมด

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

$$Precision@15 = \frac{\text{จำนวนคำที่เกี่ยวข้องและถูกดึงมา } (TP)}{\text{จำนวนคำทั้งหมดที่ถูกดึงมา } 15 \text{ รายการแรก } (TP + FP)}$$

5.ประเมินงานผลวิจัยและสรุป ข้อจำกัด ข้อเสนอแนะ

Recall

Query input	Keyword	Keyword + synonyms	Hybrid	Hybrid + synonyms	Ground truth
ก๊าซหุงต้ม	8	20	10	20	20
กาแฟ Arabica	19	24	19	25	25
Arabica	0	23	1	23	25
ถั่วเหลือง	162	162	162	162	16
แก๊สโซลีน	8	13	8	14	14
Gasoline	7	13	8	13	14
A4 Paper	4	5	4	5	5
E85	0	12	0	13	14
Ethanol	2	3	3	3	3
Sulphur	0	4	1	4	4
Sulfur	1	4	1	4	4
รถบรรทุก 10 ล้อ	56	56	56	56	56
ผ้าถักฝ้าย สีดำ	29	29	29	29	29
กระดาษคราฟท์	32	32	32	32	32

Query input	Keyword	Keyword + synonyms	Hybrid	Hybrid + synonyms
ก๊าซหุงต้ม	0.40	1.00	0.50	1.00
กาแฟ Arabica	0.76	0.96	0.76	1.00
Arabica	0.00	0.92	0.04	0.92
ถั่วเหลือง	1.00	1.00	1.00	1.00
แก๊สโซลีน	0.57	0.93	0.57	1.00
Gasoline	0.50	0.93	0.57	0.93
A4 Paper	0.80	1.00	0.80	1.00
E85	0.00	0.86	0.00	0.93
Ethanol	0.14	0.21	0.21	0.21
Sulphur	0.00	1.00	0.25	1.00
Sulfur	0.25	1.00	0.25	1.00
รถบรรทุก 10 ล้อ	0.98	0.31	0.87	0.98
ผ้าถักฝ้าย สีดำ	1.00	1.00	1.00	1.00
กระดาษคราฟท์	0.97	1.00	0.97	1.00
Macro-average Recall	0.53	0.91	0.57	**0.93**

5.ประเมินงานผลวิจัยและสรุป ข้อจำกัด ข้อเสนอแนะ

Precision@15

Query input	Keyword	Keyword + synonyms	Hybrid	Hybrid + synonyms
ก๊าซหุงต้ม	8	15	5	15
กาแฟ Arabica	4	15	13	15
Arabica	0	15	0	15
ถั่วเหลือง	15	15	10	15
แก๊สโซลีน	8	13	8	14
Gasoline	8	13	5	14
A4 Paper	4	5	4	5
E85	0	13	0	14
Ethanol	2	3	3	3
Sulphur	0	4	1	4
Sulfur	1	4	1	4
รถบรรทุก 10 ล้อ	15	15	15	15
ผ้าถักฝ้าย สีดำ	2	2	2	2
กระดาษคราฟท์	15	15	15	15

Query input	Keyword	Keyword + synonyms	Hybrid	Hybrid + synonyms
ก๊าซหุงต้ม	0.53	1.00	0.33	1.00
กาแฟ Arabica	0.27	1.00	0.87	1.00
Arabica	0.00	1.00	0.00	1.00
ถั่วเหลือง	1.00	1.00	0.67	1.00
แก๊สโซลีน	0.53	0.87	0.53	0.93
Gasoline	0.53	0.87	0.33	0.93
A4 Paper	0.27	0.33	0.27	0.33
E85	0.00	0.87	0.00	0.93
Ethanol	0.13	0.20	0.20	0.20
Sulphur	0.00	0.27	0.07	0.27
Sulfur	0.07	0.27	0.07	0.27
รถบรรทุก 10 ล้อ	1.00	1.00	1.00	1.00
ผ้าถักฝ้าย สีดำ	0.13	0.13	0.13	0.13
กระดาษคราฟท์	1.00	1.00	1.00	1.00
Macro-average Precision@15	0.39	0.70	0.39	**0.71**

ข้อจำกัดในงานวิจัย

ข้อจำกัดคือความครอบคลุมของพจนานุกรมคำพ้องความหมาย แม้ว่าจะได้รับการพัฒนาอย่างละเอียดแต่ยังไม่ครอบคลุมคำศัพท์ใหม่ๆ ที่เกิดขึ้นอย่างต่อเนื่องในสาขาสิ่งแวดล้อมและเทคโนโลยีสะอาด รวมถึงคำศัพท์ท้องถิ่นและการเรียกชื่อแบบไม่เป็นทางการที่ใช้กันอย่างแพร่หลายในหมู่ผู้ประกอบการ

ข้อเสนอแนะ

จากข้อจำกัดและผลการวิจัยที่ได้ การวิจัยเพื่อปรับปรุงประสิทธิภาพการค้นหาเชิงความหมายโดยใช้เทคโนโลยี AI ที่ทันสมัยเป็นสิ่งจำเป็น การนำ Large Language Models (LLMs) ที่รองรับภาษาไทยมาใช้ อาจช่วยปรับปรุงความเข้าใจในบริบทและความหมายของคำค้น รวมถึงการพัฒนาระบบที่สามารถเรียนรู้จากพฤติกรรมการใช้งานของผู้ใช้เพื่อปรับปรุงผลลัพธ์การค้นหาให้ดีขึ้นอย่างต่อเนื่อง