โปรแกรมค้นหาข้ามภาษาสำหรับค้นคืนค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

A CROSS-LINGUAL SEARCH ENGINE FOR RETRIEVAL OF GREEN HOUSE GAS EMISSION FACTOR

จัดทำโดย นายณฐพจน์ หนูวงษ์
รหัสนิสิต 6770233221

นิสิตปริญญาโท สาขาสาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาคนอกเวลาราชการ
อาจารย์ที่ปรึกษา รศ.ดร.ญาใจ ลิ่มปิยะกรณ์

Outline

1.ที่มาและความสำคัญของปัญหา
2.ทฤษฎีที่เกี่ยวข้อง
3.งานวิจัยที่เกี่ยวข้อง
4.แนวคิดและวิธีการวิจัย
5.วัตถุประสงค์
6.ขอบเขตการวิจัย
7.ประโยชน์ที่คาดว่าจะได้รับ

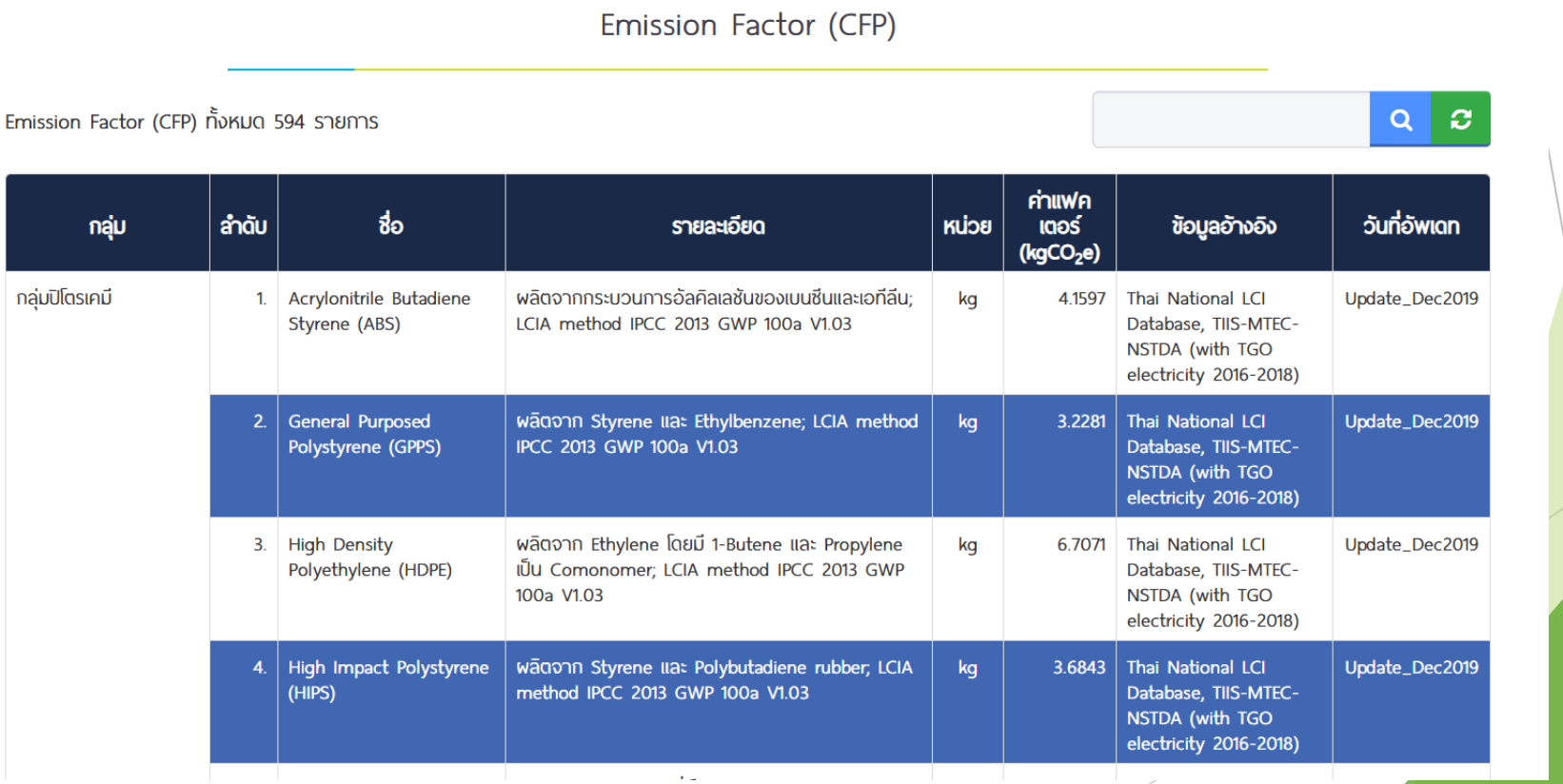
1.ที่มาและความสำคัญของปัญหา

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

https://circularmaterialhub.com/Calculate.php

1.ที่มาและความสำคัญของปัญหา

เนื่องจากการค้นหาค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกบนเว็บไซต์ขององค์การบริหารจัดการก๊าซเรือนกระจกไม่ยืดหยุ่น โดยจะต้องค้นหาคำที่เฉพาะเจาะจงเท่านั้น ดังนั้นจึงมีความสนใจที่จะทำให้การค้นหานั้นยืดหยุ่นขึ้นโดยสามารถที่จะค้นหาได้ทั้งภาษาไทยและภาษาอังกฤษที่มีความหมายเหมือนกัน หรือระบบค้นหาแบบ2 ภาษา



Emission Factor (CFP)

Emission Factor (CFP) ทั้งหมด 594 รายการ

กลุ่ม	ลำดับ	ชื่อ	รายละเอียด	หน่วย	ค่าแฟคเตอร์ ($kgCO_2e$)	ข้อมูลอ้างอิง	วันที่อัพเดท
กลุ่มปิโตรเคมี	1.	Acrylonitrile Butadiene Styrene (ABS)	ผลิตจากกระบวนการอัลคิลเลชันของเบนซีนและเอทีลีน; LCIA method IPCC 2013 GWP 100a V1.03	kg	4.1597	Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)	Update_Dec2019
	2.	General Purposed Polystyrene (GPPS)	ผลิตจาก Styrene และ Ethylbenzene; LCIA method IPCC 2013 GWP 100a V1.03	kg	3.2281	Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)	Update_Dec2019
	3.	High Density Polyethylene (HDPE)	ผลิตจาก Ethylene โดยมี 1-Butene และ Propylene เป็น Comonomer; LCIA method IPCC 2013 GWP 100a V1.03	kg	6.7071	Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)	Update_Dec2019
	4.	High Impact Polystyrene (HIPS)	ผลิตจาก Styrene และ Polybutadiene rubber; LCIA method IPCC 2013 GWP 100a V1.03	kg	3.6843	Thai National LCI Database, TIIS-MTEC-NSTDA (with TGO electricity 2016-2018)	Update_Dec2019

https://thaicarbonlabel.tgo.or.th/index.php?lang=TH&mod=Y0hKdlplVmpkSE5mWlcxcGMzTnBiMjQ9

2.ทฤษฎีที่เกี่ยวข้อง

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

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

2.ทฤษฎีที่เกี่ยวข้อง

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

2.ทฤษฎีที่เกี่ยวข้อง

เทคโนโลยีและแพลตฟอร์มที่ใช้

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

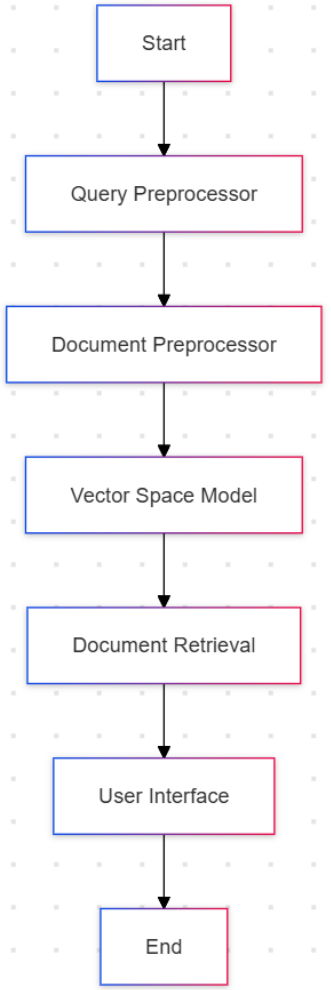
เฟรมเวิร์กภาษา Python ที่มีประสิทธิภาพสูง ใช้งานง่าย ทำให้สามารถสร้าง REST API เพื่อเชื่อมต่อระหว่าง Frontend กับ Elasticsearch ได้อย่างสะดวก

Apache Airflow แพลตฟอร์มสำหรับการสร้าง จัดการ ติดตาม Workflow

3.งานวิจัยที่เกี่ยวข้อง

English-Malayalam Cross-Lingual Information Retrieval – an Experience

งานวิจัยนี้นำเสนอระบบค้นคืนสารสนเทศข้ามภาษาอังกฤษ-มาลายาลัม (CLIR) ที่รองรับการสืบค้นทั้งภาษาเดียวและข้ามภาษา โดยใช้พจนานุกรมอังกฤษ-มาลายาลัมที่พัฒนาขึ้นเอง พร้อมด้วยเทคนิคการประมวลผลคำ เช่น การตัดคำ, การแปลงรากศัพท์
ระบบใช้ Vector Space Model (VSM) ในการจัดอันดับเอกสาร
ประเมินด้วยคำถาม 25 คำถาม และแสดงผลลัพธ์ที่มีประสิทธิภาพใกล้เคียงกันระหว่างการสืบค้นภาษาเดียวและข้ามภาษา งานวิจัยนี้ยืนยันถึงความเป็นไปได้ในการพัฒนาระบบ CLIR สำหรับภาษาอังกฤษและมาลายาลัมภายในระยะเวลาอันสั้นด้วยทรัพยากรภาษาที่เหมาะสม



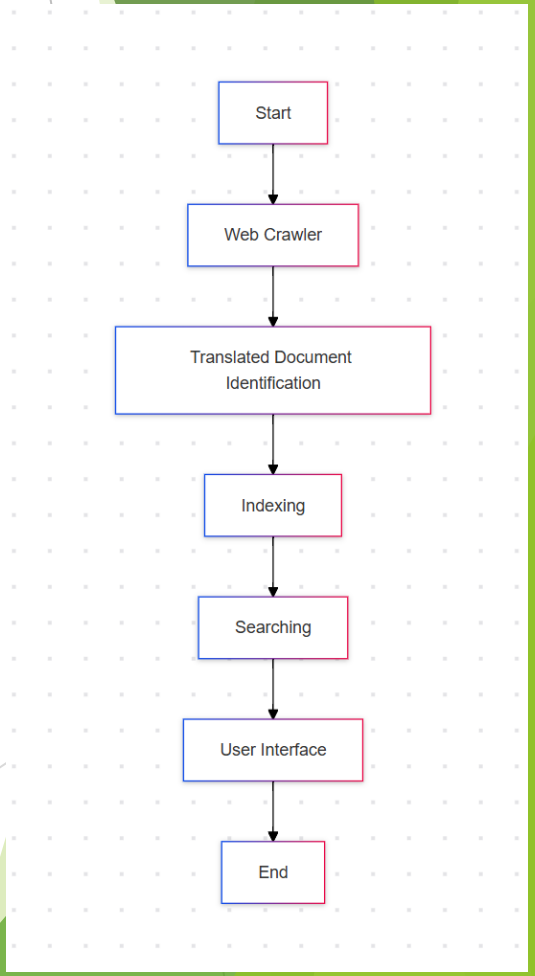
3.งานวิจัยที่เกี่ยวข้อง

Cross-Lingual Information Retrieval Model for Vietnamese-English Web Sites

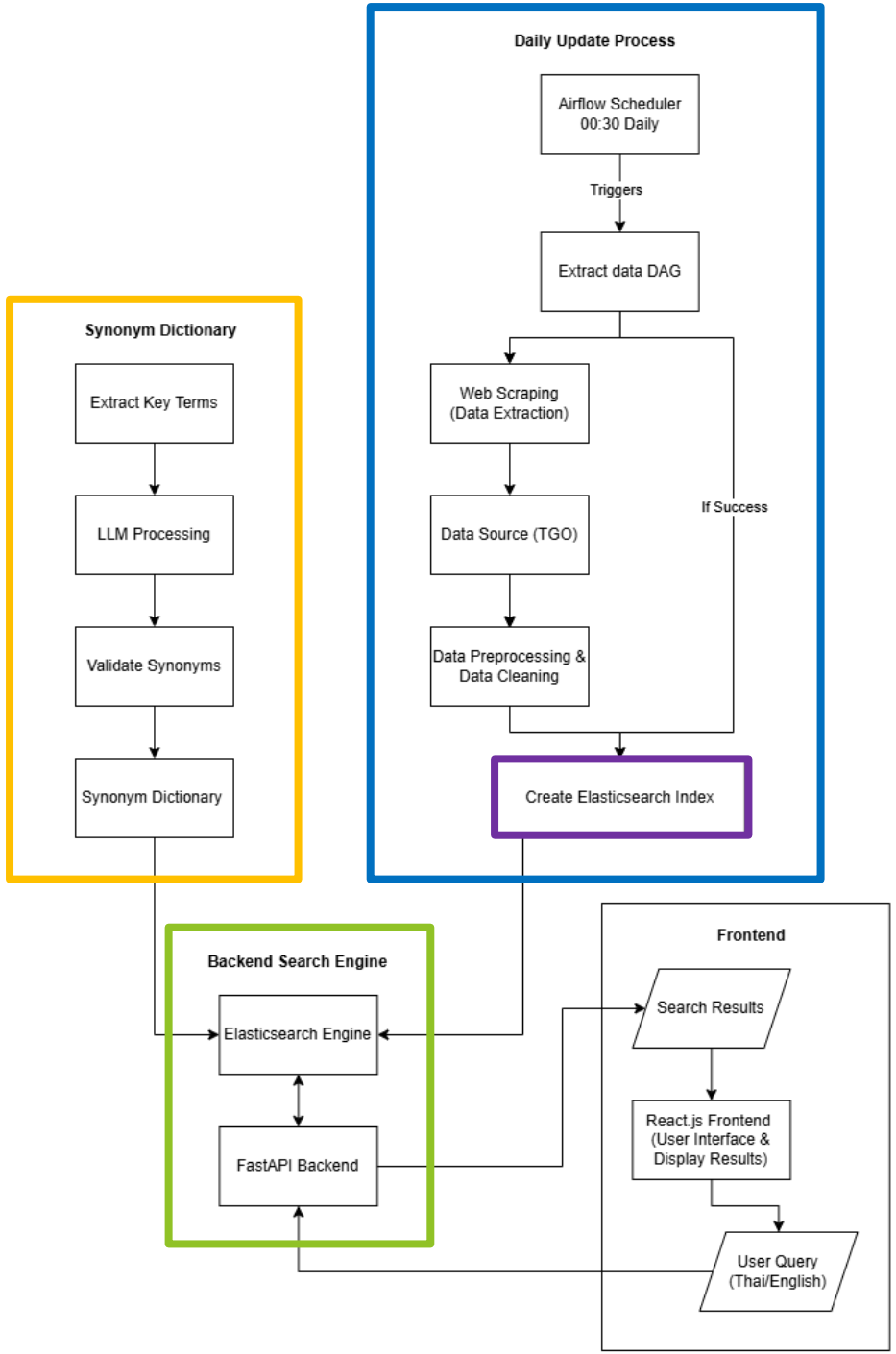
งานวิจัยนี้นำเสนอโมเดลระบบการสืบค้นข้อมูลข้ามภาษา (CLIR) สำหรับเว็บไซต์สองภาษาที่รองรับภาษาเวียดนามและอังกฤษ ระบบนี้ประกอบด้วย 4 ส่วนหลัก

1. Web Crawler สำหรับรวบรวมข้อมูล
2. Translated Document Identifying เพื่อระบุหน้าเว็บคู่แปล
3. Indexing ตามภาษา
4. Searching รองรับการสืบค้นข้อมูล

โมเดลนี้ช่วยลดการประมวลผลซ้ำ และเพิ่มความแม่นยำในการค้นหาโดยใช้ผลการระบุหน้าเว็บคู่แปล ข้อจำกัดคือจำนวนเว็บไซต์สองภาษาในปัจจุบันยังน้อย



4.แนวคิดและวิธีการวิจัย



4.แนวคิดและวิธีการวิจัย

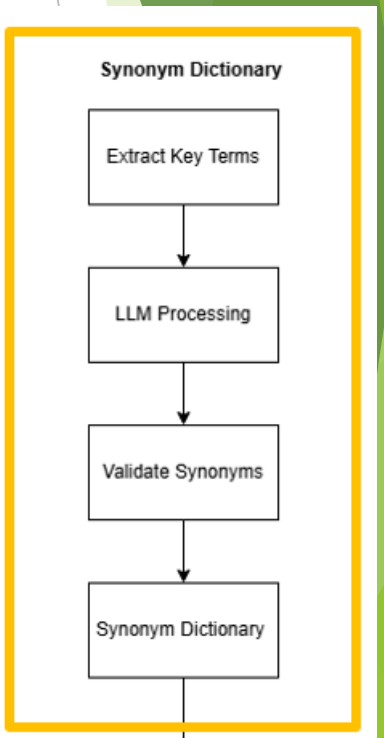
Synonym dictionary

ใช้ LLM เพื่อช่วยสร้าง Synonym Dictionary แบบ Synonym Set สำหรับคำศัพท์ในหมวดเดียวกันที่มีความหมายเหมือนกันทั้งภาษาไทยและอังกฤษรวมทั้งตัวย่อถ้ามี และให้จัดผลลัพธ์ในรูปแบบที่ Elasticsearch รองรับ เช่น ก๊าซหุงต้ม, LPG, Liquified Petroleum Gas

Synonym Set สำหรับ Elasticsearch

css Copy Edit

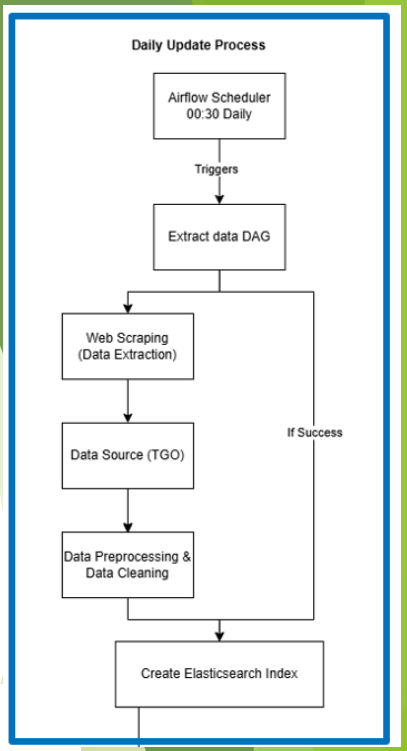
```
Liquefied Petroleum Gas, LPG, ก๊าซหุงต้ม
Diesel oil, น้ำมันดีเซล
Petrol, น้ำมันเบนซิน
Fresh Latex, น้ำยางดิบ
CO2, คาร์บอนไดออกไซด์
CH4, มีเทน
Natural gas, ก๊าซธรรมชาติ
Bagasse, ชานอ้อย
Anthracite, ถ่านหินแอนทราไซด์
Lignite, ถ่านหินลิกไนต์
Sub-bituminous coal, ถ่านหินซับบิทูมินัส
FUEL WOOD, ฟืน
Jet Kerosene, เชื้อเพลิงสำหรับเครื่องบิน
Fuel oil, น้ำมันเตา
Ploy vinyl Chloride, พีวีซี
Polypropylene, โพลิโพรพิลีน
```

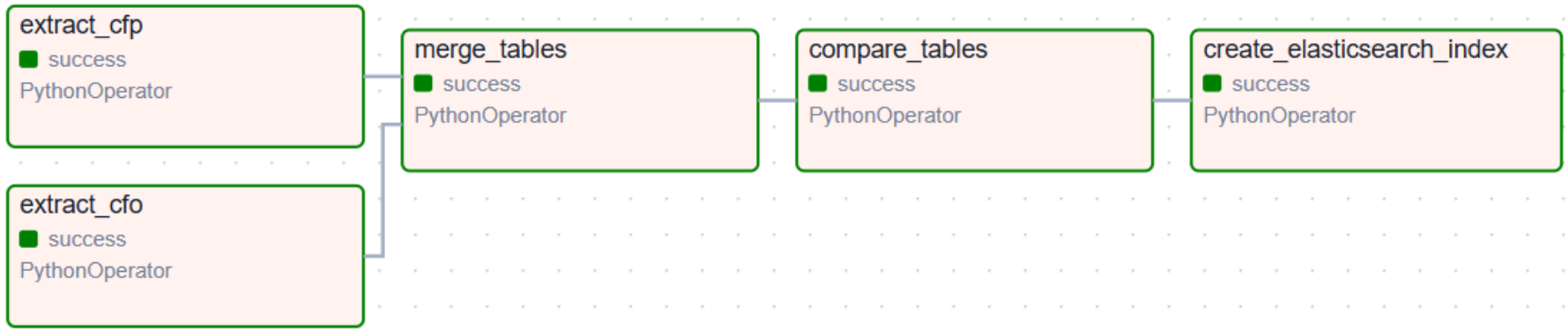


4.แนวคิดและวิธีการวิจัย

Daily Batch

Data pipeline ของการดึงขอข้อมูลจากเว็บไซต์ขององค์การบริหารจัดการก๊าซเรือนกระจก โดยจะแบ่งการทำงานออกเป็น 2 ส่วนใหญ่ๆคือการนำข้อมูลที่ได้มาจากองค์การบริหารจัดการก๊าซเรือนกระจก มาประมวลผล
จากนั้นอัปเดตข้อมูลในกรณีที่ข้อมูลเกิดการเปลี่ยนแปลงค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจก

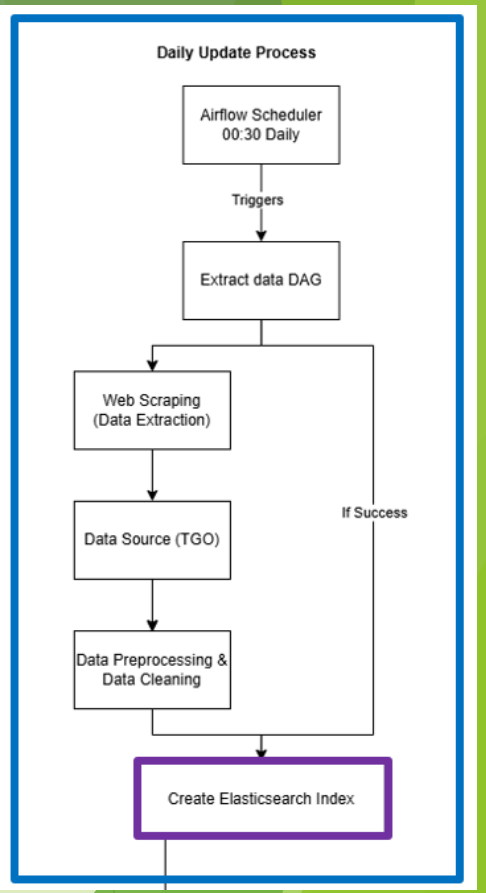




4.แนวคิดและวิธีการวิจัย

Elasticsearch Index

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



4.แนวคิดและวิธีการวิจัย

```
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"]
                }
            }
        }
    },
```

```
    "mappings": {
        "properties": {
            "ลำดับ": {
                "type": "float"},
            "ชื่อ": {
                "type": "text",
                "analyzer": "autocomplete_index_analyzer",
                "search_analyzer": "thai_synonym_analyzer"
            },
            "หน่วย": {
                "type": "text"},
            "Total [kg CO2eq/unit]": {
                "type": "float"},
            "ข้อมูลอ้างอิง": {
                "type": "text",
                "analyzer": "thai_synonym_analyzer"
            },
            "รายละเอียด": {
                "type": "text",
                "analyzer": "thai_synonym_analyzer"
            }
        }
    }
}
```

5.วัตถุประสงค์

พัฒนาระบบสืบค้นข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกที่รองรับการค้นหาได้ทั้งภาษาไทยและภาษาอังกฤษ เพื่ออำนวยความสะดวกการค้นหาข้อมูลได้อย่างรวดเร็วและถูกต้องตามปีการประเมิน

6.ขอบเขตการวิจัย

1.ใช้ข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกที่เผยแพร่โดย องค์การบริหารจัดการก๊าซเรือนกระจก (องค์การมหาชน)

2.ออกแบบให้สืบค้นได้ทั้งภาษาไทยและอังกฤษด้วย Synonym-based Approach

3.ระบบสืบค้นที่ยืดหยุ่นสามารถอัปเดตข้อมูลได้อัตโนอัติเมื่อองค์การบริหารจัดการก๊าซเรือนกระจกมีการเปลี่ยนแปลงข้อมูล

4.ประเมินระบบด้วยตัววัดมาตรฐานการค้นคืนสารสนเทศ

7.ประโยชน์ที่คาดว่าจะได้รับ

ระบบสืบค้นข้ามภาษาที่ใช้งานจริงช่วยให้ผู้ใช้สามารถค้นหาข้อมูลค่าสัมประสิทธิ์การปล่อยก๊าซเรือนกระจกได้อย่างสะดวก โดยไม่ถูกจำกัดด้วยภาษา ซึ่งช่วยให้การเข้าถึงข้อมูลเป็นไปอย่างมีประสิทธิภาพยิ่งขึ้นอีกทั้งช่วยให้ธุรกิจ โดยเฉพาะกลุ่มธุระกิจขนาดเล็ก (SME) สามารถปรับตัวให้เข้ากับมาตรฐานด้านสิ่งแวดล้อมได้อย่างมีประสิทธิภาพ