

**โครงงานวิทยาการคอมพิวเตอร์ระดับปริญญาตรี**

**เรื่อง**

**ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ  
ด้วยเทคโนโลยีปัญญาประดิษฐ์  
(Recommended System of Internship's Company)**

**โดย**

**นายทินกฤต สิงห์แก้ว**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**ปีการศึกษา 2565**



**ใบรับรองโครงงานทางวิทยาการคอมพิวเตอร์**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)**

เรื่อง ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ ด้วยเทคโนโลยีปัญญาประดิษฐ์ (Recommended System of Internship's Company)

โดย นายทินกฤต สิงห์แก้ว

**คณะกรรมการพิจารณาเห็นชอบโดย**

อาจารย์ที่ปรึกษาฯ…………………………………………………………...……………..วันที่…………/…………/………..

( ผู้ช่วยศาสตราจารย์ ดร.นงนุช เกตุ้ย )

อาจารย์ที่ปรึกษาร่วมฯ…………………………………………………………............วันที่…………/…………/………..

( อาจารย์วรวิทย์ ฝั้นคำอ้าย )

อาจารย์ผู้รับผิดชอบวิชา…………………………………………………………………..วันที่…………/…………/………..

( อาจารย์ศิริลักษณ์ แก้วศิริรุ่ง )

ประธานหลักสูตร……………………………………………………………..………………วันที่…………/…………/………..

( อาจารย์วรวิทย์ ฝั้นคำอ้าย )

**โครงงานทางวิทยาการคอมพิวเตอร์ระดับปริญญาตรี**

**เรื่อง**

**ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ**

**ด้วยเทคโนโลยีปัญญาประดิษฐ์**

**(Recommended System of Internship's Company)**

**โดย**

**นายทินกฤต สิงห์แก้ว**

**คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาศาสตร์**

**พ.ศ. 2565**

**บทคัดย่อ**

**ชื่อโครงงาน** : ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจด้วยเทคโนโลยี  
 ปัญญาประดิษฐ์ (Recommended System of Internship's  
 Company)

**คณะผู้ศึกษา** : นายทินกฤต สิงห์แก้ว

**อาจารย์ที่ปรึกษาฯ** : ผู้ช่วยศาสตราจารย์ ดร.นงนุช เกตุ้ย

**อาจารย์ที่ปรึกษาร่วมฯ** : อาจารย์วรวิทย์ ฝั้นคำอ้าย

**สาขา** : วิทยาศาสตร์

**หลักสูตร** : วิทยาการคอมพิวเตอร์

**ปีการศึกษา**  : 2565

ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจด้วยเทคโนโลยีปัญญาประดิษฐ์(Recommended System of Internship's Company) เป็นระบบที่ช่วยแนะนำสถานประกอบการสำหรับการฝึกงานของนักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน ในรูปแบบของเว็บแอปพลิเคชัน โดยเป็นการให้นักศึกษาระบุรายละเอียดความสนใจของรูปแบบธุรกิจ หรือรูปแบบของงานที่อยากฝึกงาน เพื่อนำมาวิเคราะห์หาความคล้ายคลึงกับข้อมูลสถานประกอบการที่มีอยู่ในฐานข้อมูลที่ได้รับการจัดกลุ่มไว้แล้วด้วยเทคโนโลยีปัญญาประดิษฐ์ เพื่อเป็นการเพิ่มประสิทธิภาพของนักศึกษาในการฝึกงานในสถานประกอบการที่มีรูปแบบของธุรกิจหรือรูปแบบของงานที่ตรงตามความต้องการ ทำให้การฝึกงานของนักศึกษาได้รับประสบการณ์การทำงานที่ดีและได้รับประโยชน์สูงสุด อีกทั้งยังเป็นช่องทางที่สถานประกอบการจะนำข้อมูลของสถานประกอบการมาเป็นตัวเลือกให้แก่นักศึกษาในต่อไป

**กิตติกรรมประกาศ**

โครงงานทางวิทยาการคอมพิวเตอร์ “ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจด้วยเทคโนโลยีปัญญาประดิษฐ์ (Recommended System of Internship's Company)” เพื่อการสำเร็จการศึกษาระดับปริญญาตรี สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปด้วยดี เนื่องจากได้รับความกรุณาและคำแนะนำจากคณาจารย์หลาย ๆ ท่านในหลักสูตรวิทยาการคอมพิวเตอร์ ที่ได้กรุณาให้ความรู้ แนะแนวทาง ข้อคิด ข้อแนะนำสู่ความสำเร็จและช่วยแก้ไขปัญหาอุปสรรคต่าง ๆ รวมทั้งรูปเล่มให้สำเร็จไปได้ด้วยดี

ขอขอบพระคุณ ผู้ช่วยศาตราจารย์ ดร.นงนุช เกตุ้ย และอาจารย์วรวิทย์ ฟั่นคำอ้าย  
อาจารย์ที่ปรึกษา และอาจารย์ประจำวิชาทุกท่าน ผู้ซึ่งกรุณาให้ความรู้ คำแนะนำแนวทางการสร้างผลงานสู่ความสำเร็จ และช่วยแก้ไขปัญหาอุปสรรคต่าง ๆ รวมทั้งตรวจทานแก้ไขรูปเล่มจนเสร็จสมบูรณ์ ขอขอบคุณบิดา มารดา ผู้มีพระคุณทุกท่าน เพื่อนนักศึกษา และบุคคลที่เกี่ยวข้องที่ยังไม่ได้กล่าวถึง ที่ได้ช่วยออกความคิดเห็น ได้ให้ข้อแนะนำ และอำนวยความสะดวกในด้านต่าง ๆในการทำโครงงานครั้งนี้ไว้ ณ ที่นี้

สุดท้ายนี้ผู้ศึกษาหวังว่าโครงงานฉบับนี้จะเป็นประโยชน์สำหรับมหาวิทยาลัย  
และนักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน และผู้ที่สนใจที่จะศึกษาต่อไป

**สารบัญ**

**หน้า**

**บทคัดย่อ ก**

**กิตติกรรมประกาศ ข**

**สารบัญ ค**

**สารบัญตาราง ง**

**สารบัญภาพ จ**

**บทที่**

1. **บทนำ**
   1. ความเป็นมาและความสำคัญของปัญหา 1

**บทที่ 1**

**บทนำ**

* 1. **ความเป็นมาและความสำคัญของปัญหา**

ในระบบการศึกษาระดับปริญญาตรีนั้นรายวิชาที่มีในการศึกษาปีสุดท้ายของหลักสูตรคือ รายวิชาที่จะต้องให้นักศึกษาแต่ละคนนั้นออกไปฝึกทำงานที่สถานประกอบการต่างๆ ในช่วงระยะเวลาหนึ่ง ซึ่งเป็นสิ่งที่มีความสำคัญและความท้าทายเนื่องจากเป็นการที่นักศึกษาจะได้ทดลองทำงานจริง สถานการณ์จริง สถานที่จริง ในสถานประกอบการที่นักศึกษาได้เลือก

ดังนั้นการเลือกสถานประกอบการสำหรับฝึกงานจึงเป็นเรื่องที่ต้องให้ความสำคัญเป็นอย่างมากเนื่องจากหากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำนั้น ตรงกันกับความสามารถของนักศึกษาก็จะเป็นผลดี เนื่องจากความรู้และทักษะที่ได้จากการทำงานนั้นสามารถนำไปต่อยอดและใช้งานจริงเมื่อจบการศึกษาและเข้าทำงาน แต่หากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำไม่ตรงกับความต้องการหรือทักษะของนักศึกษาอาจทำให้การฝึกงานนั้นล้มเหลว หรืออาจไม่ได้ความรู้และทักษะที่ต้องการได้ และด้วยเทคโนโลยีปัญญาประดิษฐ์ในด้านของการประมวลผลภาษาธรรมชาตินั้นพัฒนาก้าวหน้าขึ้นมาก ทั้งในแง่ของเทคนิค เครื่องมือ และองค์ความรู้ ทำให้เกิดตัวอย่างการนำข้อมูลมาประมวลผลที่มีประสิทธิภาพมากมายในปัจจุบัน

ด้วยเหตุนึ้จึงได้มีการเริ่มโครงงานการพัฒนาเว็บแอปพลิเคชันเพื่อแนะนำบริษัทสำหรับหาสถานประกอบการที่ฝึกงานที่เหมาะสมและเป็นสถานประกอบการที่นักศึกษามีความสนใจในรูปแบบธุรกิจของสถานประกอบการนั้น เพื่อช่วยอำนวยความสะดวกแก่นักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน ในการค้นหาสถานประกอบการสำหรับออกฝึกงาน โดยการใช้วิธีประมวลผลภาษาธรรมชาติเข้ามาช่วยจัดกลุ่มสถานประกอบการและเสนอรายชื่อสถานประกอบการที่เหมาะสมแก่นักศึกษาผ่านทางเว็บแอปพลิเคชัน

* 1. **วัตถุประสงค์**

1.2.1 เพื่อศึกษาและพัฒนาเว็บแอปพลิเคชันเพื่อแนะนำสถานประกอบการตามความสนใจของนักศึกษา

1.2.2 เพื่อศึกษาและวิเคราะห์ความสนใจของนักศึกษาในการหาสถานประกอบการสำหรับฝึกงาน

1.2.3 เพื่อศึกษาประสิทธิภาพการแบ่งกลุ่มของข้อมูลสถานประกอบการด้วยวิธีประมวลผลภาษาธรรมชาติ

### ขอบเขตของโครงงาน

1.3.1 ผู้ใช้สามารถค้นหาสถานประกอบการได้ด้วยรายละเอียดของงานหรือรูปแบบธุรกิจที่สนใจ

1.3.2 ทุกครั้งที่มีการอัพเดทข้อมูลรายละเอียดธุรกิจของสถานประกอบการชุดใหม่เข้ามานั้นข้อมูลทั้งหมดจะถูกจัดกลุ่มใหม่ และต้องกำหนดชื่อกลุ่มใหม่จากป้ายเลขกลุ่ม

1.3.3 ในการจัดกลุ่มข้อมูลนั้นอาจมีข้อมูลสถานประกอบการอื่นเข้ามารวมได้บ้างเนื่องจากข้อมูลของสถานประกอบการนั้นอาจมีความหมายที่ไม่ชัดเจน 100%

* 1. **ประโยชน์ที่คาดว่าจะได้รับ**
     1. ได้พัฒนาเว็บแอปพลิเคชันสำหรับการค้นหาสถานประกอบการสำหรับฝึกงาน
     2. ได้วิเคราะห์รายละเอียดความสนใจของนักศึกษาที่ต้องการฝึกงานในสถานประกอบการต่าง ๆ
     3. เป็นช่องทางสำหรับการเลือกและหาข้อมูลของสถานประกอบการสำหรับออกฝึกงานของนักศึกษา

**1.5 อุปกรณ์เครื่องมือที่ใช้ในโครงงาน**

1.5.1 อุปกรณ์ ประกอบด้วย

1. โทรศัพท์มือถือ
2. คอมพิวเตอร์

1.5.2 ซอฟต์แวร์ ประกอบด้วย

1. Programing language
   1. Python
   2. Javascript
   3. HTML
   4. CSS
2. Web framework
   1. Next.js
3. Database
   1. MongoDB
4. Program
   1. Microsoft excel
   2. Visual studio code
   3. Postman
   4. Firefox
   5. Figma
   6. Notion
5. Version control
   1. Git
   2. GIthub
6. Python library
   1. Pythainlp
   2. Matplotlibs
   3. Pandas
   4. Numpy
   5. Scikit-learn
   6. nltk
   7. fastAPI
   8. python-dotenv
7. Javascript library
   1. Tailwind CSS
   2. cors
   3. dotenv
   4. sweetalert2
   5. headlessui
   6. heroicons
8. Global Network
   1. Cloudflare
9. Cloud computing
   1. Amazon Web Services
   2. Vercel

**1.6 นิยามศัพท์ที่ใช้ในโครงงาน**

1.6.1 ประมวลผลภาษาธรรมชาติ หรือ Natural language processing (NLP) เป็นเทคนิคแขนงหนึ่งในศาสตร์ของเทคโนโลยีปัญญาประดิษฐ์ ซึ่งเป็นการทำให้คอมพิวเตอร์เข้าใจ ตีความ และสื่อสารภาษาของมนุษย์ได้

1.6.2 การจัดกลุ่มข้อมูล หรือ Clustering หมายถึง เป็น Machine learning model ชนิดหนึ่งที่อยู่ในประเภท Unsupervised คือเป็นการที่นำข้อมูลเข้าไปให้ Model ประมวลผลโดยที่ไม่ได้จำกัดคำตอบไว้แต่ให้คอมพิวเตอร์ประมวลผลและกำหนดเองว่าคำตอบควรจะเป็นลักษณะใดบ้าง

1.6.3 Word segmentation หมายถึง ด้วยที่การเขียนภาษาไทยนั้นไม่มีการแยกคำด้วยการเว้นวรรคเหมือนภาษาอังกฤษ หรือ ภาษาอื่นๆ ดังนั้นจึงจำเป็นต้องทำการตัดคำจากประโยคออกมาเป็นคำ ๆ เพื่อให้นำไปประมวลหรือใช้งานต่อได้ด้วยอัลกอริทึมต่าง ๆ

1.6.4 Cosine similarity หมายถึง เป็นการหาค่าความคล้ายคลึงกันระหว่าง Vector A และ Vector B ว่าไปทิศทางเดียวกันหรือไม่โดยการใช้สูตรของกฎสามเหลี่ยมเพื่อหาผลลัพธ์แล้วนำมาเปรียบเทียบกัน

1.6.5 API หรือ Application programing interface หมายถึง ช่องทางสำหรับการสื่อสารกันที่ทาง Server และ Client สร้างขึ้นมาเพื่อเป็นตัวกลางให้โปรแกรม หรือผู้ใช้อื่นๆใช้ติดต่อสื่อสาร เชื่อต่อแลกเปลี่ยนข้อมูลกัน

1.6.6 Cloud computing หมายถึง การเช่าใช้ระบบคอมพิวเตอร์ และทรัพยากรแบบครบวงจรจากผู้ให้บริการต่าง ๆ เช่น Amazon, Google, Microsoft, Huawei โดยสามารถกำหนดรูปแบบของฮาร์ดแวร์และซอฟต์แวร์ที่ต้องการได้ มีให้บริการทั้งเครื่องแม่ข่าย(Server) ฐานข้อมูล(Database) การทดสอบระบบ(Testing) หรือแอปพลิชันสำเร็จรูปในหลายระบบปฏิบัติการ(Platform)

**บทที่ 2**

**ทฤษฎีและงานวิจัยที่เกี่ยวข้อง**

การศึกษาค้นคว้าเพื่อจัดทำโครงงานทางวิทยาการคอมพิวเตอร์ ระบบแนะนำบริษัทสำหรับฝึกงานด้วยเทคโนโลยีปัญญาประดิษฐ์ ผู้ศึกษาได้ ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้องตามหัวข้อลำดับต่อไปนี้

2.1 ทฤษฎีและความรู้ที่เกี่ยวข้อง

2.1.1 ทฤษฎี การประมวลผลภาษาธรรมชาติ หรือ Natural language processing

2.1.2 ทฤษฎี การตัดคำในภาษาไทยหรือ Word segmentation

2.1.3 ทฤษฎี การสกัดใจความของข้อความด้วยเทคนิค TF-IDF

2.1.4 ทฤษฎี การจัดกลุ่มข้อความด้วยอัลกอริทึม K-Means

2.1.5 ทฤษฎี การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow method

2.1.6 ทฤษฎี การคำนวณค่าความคล้ายคลึงด้วยเทคนิค Cosine similarity

2.1.7 ทฤษฎี Amazon web service

2.1.8 ทฤษฎี API

2.1.9 ทฤษฎี Cloudflare

2.1.10 ทฤษฎี Cors

2.1.11 ทฤษฎี Fastapi

2.1.12 ทฤษฎี Git

2.1.13 ทฤษฎี Node.js

2.1.14 ทฤษฎี Matplotlibs

2.1.15 ทฤษฎี Mongodb

2.1.16 ทฤษฎี Next.js

2.1.17 ทฤษฎี Numpy

2.1.18 ทฤษฎี Pandas

2.1.19 ทฤษฎี Pythainlp

2.1.20 ทฤษฎี Scikit-learn

2.1.21 ทฤษฎี Vercel

2.2 งานวิจัยที่เกี่ยวข้อง

**2.1 ทฤษฎีและความรู้ที่เกี่ยวข้อง**

2.1.1 ทฤษฎี การประมวลผลภาษาธรรมชาติ หรือ Natural language processing

NLP ย่อมาจาก Natural Language Processing คือการประมวลผลภาษาธรรมชาติหรือภาษาของมนุษย์ที่ใช้สื่อสารกัน เป็นเทคนิคหนึ่งในเทคโนโลยีปัญญาประดิษฐ์ ที่จะทำให้คอมพิวเตอร์สามารถเข้าใจและเรียนรู้ ประมวลผลภาษาของมนุษย์ได้ ในด้านของการวิเคราะห์ภาษาศาสตร์ การตีความจากบทความ หรือกระทั่งการแปลภาษา NLP นั้นจำเป็นต้องใช้ความรู้จากหลาย ๆ ศาสตร์เข้ามาเช่น Mathematics, Linguistics, Psychology เพื่อพัฒนาประสิทธิภาพการทำงานและความฉลาดของคอมพิวเตอร์ (ตาเยะ, 2022)

จุดเริ่มต้นของ NLP นั้นมีมาตั้งแต่ประมาณปี 1950-1980 ในยุคนั้นวิธีการที่จะให้คอมพิวเตอร์เข้าใจภาษาของมนุษย์นั้นใช้ “Rule-based” เป็นการใช้ if-else ในโปรแกรมที่ตั้งไว้ตามคำที่กำหนด และในต่อมาประมาณปี 1981-2001 เริ่มมีการใช้ ML หรือ Machine leaning ที่ใช้อัลกอริทึมในการประมวลผลเช่น Decision Tree เข้ามาช่วยในการประมวลผล และฝึกสอนคอมพิวเตอร์โดยข้อมูลที่เป็น Dataset ทำให้ความแม่นยำเพิ่มขึ้น และในยุคปัจจุบันยุคที่มี Deep Neural Network เนื่องจากปัจจุบันคอมพิวเตอร์มีความเร็วที่เพิ่มขึ้นและปริมาณข้อมูลนั้นมีมากขึ้นตาม ทำให้การใช้ Deep Neural Network มาสร้างโมเดลสำหรับการทำ NLP เป็นที่นิยมมากยกตัวอย่างเช่น word embeddings คือการหา semantic กับข้อความนั้นๆ (L, 2019)

กระบวนการทำงานของ NLP นั้น มีประกอบไปด้วยหลายส่วนของการประมวลผลและใช้แปลความหมาย ประกอบด้วยดังนี้

1. Tokenization เป็นการตัดคำออกเป็นคำ ๆ เพื่อที่จะนำไปประมวลผลต่อตามรูปแบบของแต่ละภาษา
2. Parsing เป็นการระบุโครงสร้างของข้อความ
3. Lemmatization/stemming คือ การแปลงคําให้อยู่ในรูปแบบดังเดิม
4. Part-of-speech tagging คือ การอธิบายหรือการกำกับว่าในแต่ละคำนั้นมีความหมาย หรือประเภทของคำเป็นอย่างไร
5. Language detection การตรวจสอบภาษาว่าเป็นภาษาอะไร
6. Identification of semantic relationships คือการระบุความสัมพันธ์ของคําต่าง ๆ ในประโยค

ปัจจุบัน NLP นั้นอยู่ในหลายรูปแบบรอบตัวถูกนำไปใช้ในหลาย ๆ ด้านทั้ง Digital marketing, ทางการแพทย์ การแปลภาษา Chatbot และอื่น ๆ

2.1.2 ทฤษฎี การตัดคำในภาษาไทยหรือ Word segmentation

Word segmentation คือการที่นำประโยคมาตัดออกเป็นคำ ๆ เนื่องจากในบางภาษาเช่นภาษาไทยรูปแบบการเขียนนั้นไม่มีการเว้นวรรคของคำต่างจากภาษาอังกฤษที่ใช้การเว้นวรรคในแต่ละคำ ดังนั้นถ้าจะทำ NLP ที่เป็นภาษาไทยนั้นจำเป็นต้องทำ Word segmentation เพื่อให้ได้ชุดคำที่จะนำไปใช้งานต่อ ในปัจจุบันการทำ Word segmentation นั้นมีเครื่องมือให้ใช้อยู่จำนวนมากยกตัวอย่างเช่น Python library pythainlp, nltk หรือสามารถใช้บริการ web API ของ aiforthai (L, 2019)



**ภาพที่ 1** ตัวอย่างการทำ Word segmentation

จากภาพที่ 1 เป็นการทำ Word segmentation ด้วย Python library pythainlp จากคำว่า “นอนตากลมดูดาว” ได้ผลลัพธ์ออกมาเป็น นอน,ตากลม,ดู,ดาว

2.1.3 ทฤษฎี การสกัดใจความของข้อความด้วยเทคนิค TF-IDF

การทำ Term Frequency – Inverse Document Frequency (TF-IDF) เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภายในประโยค เทคนิคนี้มากจาก 2 องค์ประกอบต่อกันคือ Term Frequency (TF) และ Inverse Document Frequency (IDF) องค์ประกอบแรก Term Frequency (TF) นั้นหมายถึงการที่หาคำที่มีการใช้ซ้ำบ่อยที่สุดในเอกสารนั้น ๆ ซึ่งแสดงไปถึงว่าคำนั้นเป็นคำที่มีความสำคัญมากเอกสารนั้น วิธีคำนวณค่าความถี่ของคำใช้การนำจำนวนครั้งของคำที่ปรากฏในเอกสารมาหารด้วยจำนวนคำทั้งหมดในเอกสาร เช่น ต้องการหาค่าความถี่ของคำว่าเว็บไซต์ในเอกสาร

**ตารางที่ 1** ตัวอย่างการคำนวณค่า Term Frequency ที่จำนวนคำทั้งหมดเท่ากับ 7

|  |  |  |  |
| --- | --- | --- | --- |
| คำ | จำนวนคำ | Term Frequency | ผลลัพธ์ |
| เว็บไซต์ | 5 | 5 ÷ 7 | 0.71 |
| หนังสือ | 1 | 1 ÷ 7 | 0.14 |
| ออนไลน์ | 2 | 2 ÷ 7 | 0.29 |
| ขาย | 2 | 2 ÷ 7 | 0.29 |
| เข้าชม | 1 | 1 ÷ 7 | 0.14 |
| มือถือ | 4 | 4 ÷ 7 | 0.57 |
| และ | 3 | 3 ÷ 7 | 0.43 |

จากตัวอย่างจะเห็นได้ว่าคำว่าเว็บไซต์ปรากฏบ่อยในเอกสารทำให้มีค่า Term Frequency สูงจึงเรียกได้ว่าเป็นคำสำคัญของเอกสาร แต่การใช้ค่า Term Frequency เพื่อหาใจความสำคัญเพียงอย่างเดียวนั้นยังไม่ดีพอ จึงต้องใช้องค์ประกอบ Inverse Document Frequency (IDF) เข้ามาช่วยเพิ่มเติม Inverse Document Frequency (IDF) หมายถึง การคำนวณหาน้ำหนักของคำโดยการนำคำสำคัญค้นหาจากหลาย ๆ เอกสารหากคำนั้นมีค่า Inverse Document Frequency (IDF) ต่ำแสดงว่าคำนั้นไม่ได้เป็นคำสำคัญของเอกสารทั้งหมด สมการที่ใช้คำนวณหาค่า Inverse Document Frequency (IDF)

**ตารางที่ 2** ตัวอย่างการคำนวณค่า Inverse Document Frequency ที่จำนวนเอกสารเท่ากับ 10

|  |  |  |  |
| --- | --- | --- | --- |
| คำ | จำนวนเอกสารที่ปรากฏ | Inverse Document Frequency | ผลลัพธ์ |
| เว็บไซต์ | 5 |  | 0.31 |
| หนังสือ | 2 |  | 0.70 |
| ออนไลน์ | 2 |  | 0.70 |
| ขาย | 2 |  | 0.70 |
| เข้าชม | 1 |  | 1.00 |
| มือถือ | 3 |  | 0.52 |
| และ | 2 |  | 0.70 |

จากนั้นคำนวณค่า Term Frequency – Inverse Document Frequency (TF-IDF) ได้โดยสมการ

***ตารางที่ 3*** *ตัวอย่างการคำนวณค่า* TF-IDF

|  |  |  |  |
| --- | --- | --- | --- |
| คำ | TF | IDF | TF-IDF |
| เว็บไซต์ | 0.71 | 0.31 | 0.22 |
| หนังสือ | 0.14 | 0.70 | 0.10 |
| ออนไลน์ | 0.29 | 0.70 | 0.20 |
| ขาย | 0.29 | 0.70 | 0.20 |
| เข้าชม | 0.14 | 1.00 | 0.14 |
| มือถือ | 0.57 | 0.52 | 0.30 |
| และ | 0.43 | 0.70 | 0.30 |

*ดังตัวอย่างจะเห็นได้ว่าบางคำที่มีค่า* TF-IDF *สูงแต่ไม่ได้บ่งบอกถึงลักษณะของข้อความในเอกสารเช่นคำว่า และ ซึ่งถือว่าเป็น* Stop word *ซึ่งเป็นคำที่ไม่สื่อความหมายโดยปกติแล้ว คำเหล่านี้มักถูกกรองออกก่อนที่จะมีการนำข้อความมาทำการประมวลผลทางภาษาจากตัวอย่างจะเห็นว่าเมื่อคำนวณหาค่า* Term Frequency – Inverse Document Frequency (TF-IDF) โดยที่ตัดคำที่ไม่มีความหมายหรือ Stop word ออกแล้วจะเหลือคำว่า มือถือ เว็บไซต์ ออนไลน์ เรียงลำดับความสำคัญจากมากไปน้อยตามลำดับ (Patipan, 2020)

2.1.4 ทฤษฎี การจัดกลุ่มข้อความด้วยอัลกอริทึม K-Means

K-Means เป็นวิธีการหนึ่งใน Data mining อยู่ในกลุ่มของ Unsupervised Learning คือการให้คอมพิวเตอร์เรียนรู้โดยไม่ต้องมีผู้สอน (Chakrit, 2018) เป็นอัลกอริทึมสำหรับการทำ Clustering Model เป็นการจัดกลุ่มข้อมูลด้วยการกำหนดจำนวนกลุ่มก่อนการทำ Clustering ซึ่งแทนด้วยค่า K จากนั้นคำนวณหาจุดกึ่งกลางของแต่ละกลุ่มเรียกว่าจุด Centroid ตามจำนวนกลุ่มที่กำหนดไว้เว้นระยะห่างด้วยการคำนวณระยะห่างด้วยสมการ (DIGI, 2022)

ทำการคำนวณและย้ายจุด Centroid และหาค่าเฉลี่ยจนค่าเฉลี่ยไม่มีการเปลี่ยนแปลงจะได้จุดกึ่งกลางของข้อมูลในแต่ละกลุ่ม

Chart, scatter chart

Description automatically generated

**ภาพที่ 2** ตัวอย่างการจัดกลุ่มข้อมูลด้วย K-Means โดยที่เครื่องหมายกากบาทสีแดงคือจุด Centroid ของแต่ละกลุ่มข้อมูล

2.1.5 ทฤษฎี การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow method

Elbow method เป็นวิธีหนึ่งที่ใช้หาจำนวนของกลุ่มที่เหมาะสมด้วยการวัดข้อผิดพลาด (Error measurement) ผลรวมระยะห่างระหว่างข้อมูลกับจุด Centroid เมื่อข้อผิดพลาดน้อยลงความชันของเส้นโค้งก็จะแบนราบไปตามแกน X จนทำให้เกิดมุมลักษณะเหมือนกับข้อศอกก็จะถือจุดที่อยู่ตรงมุมข้อศอกเป็นจำนวนของกลุ่มข้อมูลที่เหมาะสมดังในภาพตัวอย่างภาพที่ 3 จะเห็นได้ว่าจำนวนกลุ่มที่เหมาะสมคือ 4-5 กลุ่ม (Weerasak, 2017)

**ภาพที่ 3** กราฟที่แสดงจำนวนข้อผิดพลาดเพื่อหาจพนวนกลุ่มที่เหมาะสมที่สุด

2.1.6 ทฤษฎี การคำนวณค่าความคล้ายคลึงด้วยเทคนิค Cosine similarity

Cosine Similarity คือการวัดความเหมือนของ Vector 2 ว่าไปในทิศทางเดียวกันหรือไม่ โดยที่เป็นการตัดขนาด หรือ Magnitude ของ Vector ออกไป (Supalerk, 2020)

Text

Description automatically generated

**ภาพที่ 4** สมการการหาค่า Cosine similarity

อธิบายโดยง่ายคือเป็นการวัดระยะห่างระหว่าง Object A และ Object B ว่ามีความคล้ายกันแค่ไหนยกตัวอย่างเปรียบเทียบระหว่างคำว่า “ยินดีที่ได้รู้จักครับ” และ “ยินดีที่ได้รู้จักค่ะ” ทำการตัดเพื่อหาคำทั้งหมดก่อนครับ คือ [ “ยินดี” , “ที่” , “ได้” , “รู้จัก” , “ครับ” , “ค่ะ”] เราจะได้ Object A และ B เขียนเป็นชุดข้อมูล ได้ดังนี้ (Chakrit, similarity-ความเหมือนที่แตกต่าง, 2019)

1. “ยินดีที่ได้รู้จักครับ” = [ 1 , 1 , 1 , 1 , 1 , 0]
2. “ยินดีที่ได้รู้จักค่ะ” = [ 1 , 1 , 1 , 1 , 0 , 1]

Graphical user interface, text, website

Description automatically generated

**ภาพที่ 5** คำนวณค่า Cosine similarity ด้วย scikit-learn

2.1.7 ทฤษฎี Amazon web service

AWS เป็นตัวย่อของ Amazon Web Services ซึ่งเป็นบริการบนระบบคลาวด์ ที่มีบริการหลากหลายมากกว่า 200 โซลูชัน ถูกใช้งานในธุรกิจและองค์กรทุกประเภทไม่ว่าจะเป็นบริษัทสตาร์ตอัป องค์กรขนาดใหญ่ ไปจนถึงหน่วยงานของรัฐ AWS ให้บริการโครงสร้างพื้นฐานด้านไอที การใช้บริการ Server และ Storage การสร้างและดูแลเว็บไซต์ ไปจนถึง ระบบอี-คอมเมิร์ซ การสร้างแอปพลิเคชัน การส่งเสริมการทำงานแบบ Remote Working การใช้ระบบ IoT เพื่อการสร้างนวัตกรรมใหม่ ๆ รวมถึงโซลูชันอื่น ๆ ในปัจจุบันนี้ AWS เป็นระบบประมวลผลบนคลาวด์ที่มีผู้ใช้บริการมากที่สุดในโลก เพราะได้รับความไว้วางใจจากผู้คนทั่วโลก เนื่องจาก AWS เป็นบริษัทในเครือของ Amazon เว็บไซต์ซื้อขายสินค้าออนไลน์ชื่อดังจากประเทศสหรัฐอเมริกา (CloudHM, 2022)

2.1.8 ทฤษฎี API

API ย่อมาจาก “Application Program Interface” (ส่วนต่อประสานโปรแกรมประยุกต์) ในบริบทของ API คำว่า “Application” หมายถึงทุกซอฟต์แวร์ที่มีฟังก์ชันชัดเจน ส่วน “Interface” อาจถือเป็นสัญญาบริการระหว่างสองแอปพลิเคชัน ใช้สื่อสารกันโดยใช้คำขอ (Request) และการตอบกลับ (Response) ระหว่างเครื่องแม่ข่ายและแอปพลิเคชันอื่น ๆ API คือกลไกที่ช่วยให้ส่วนประกอบซอฟต์แวร์สองส่วนสามารถสื่อสารกันได้โดยใช้ชุดคำจำกัดความและโปรโตคอล ตัวอย่างเช่น ระบบซอฟต์แวร์ของสำนักพยากรณ์อากาศประกอบด้วยข้อมูลสภาพอากาศรายวัน (API คืออะไร)

Graphical user interface, text

Description automatically generated

**ภาพที่ 6** ตัวอย่างการส่งคำขอไปยัง API

Graphical user interface, text, application

Description automatically generated

**ภาพที่ 7** การตอบกลับจาก API

2.1.9 ทฤษฎี Cloudflare

Cloudflare คือ Global Network ที่ถูกออกแบบมาเพื่อให้ทุกสิ่งของคุณที่เชื่อมอยู่บนอินเทอร์เน็ต มีความปลอดภัย (Security) มีประสิทธิภาพ (Performance) และพร้อมใช้งาน (Availability) ซึ่ง Cloudflare จะทำหน้าที่เป็นตัวกลางระหว่างผู้เข้าใช้งานและ Server ที่เก็บข้อมูล โดยผู้เข้าใช้งานจะมาทั้งในรูปแบบของ Visitor, Crawlers & Bots และ Attackers แต่เมื่อใช้งาน Cloudflare การเข้าถึงทุกรูปแบบจะต้องผ่านระบบของ Cloudflare แทน โดย Cloudflare จะเข้ามาช่วยใน 3 เรื่องหลัก ๆ คือ (Cloudflare คืออะไร จะเข้ามาช่วยองค์กรของคุณได้อย่างไร?, 2021)

1. Web Application Firewall (WAF) ป้องกันการโจมตีเว็บไซต์ในรูปแบบ Cloud Security โดย WAF จะช่วยกัน HTTP/HTTPS Traffic ที่เป็นอันตรายออกโดยอัตโนมัติ เช่น Code Injection, Cross-Site-Scripting และ Sensitive Data Exposure

2. Distributed Denial-of-Service (DDoS) คือการโจมตีโดยการส่ง Traffic ปริมาณมากไปยังเว็บไซต์ เพื่อขัดขวางความสามารถในการให้บริการ หรือทำให้ไม่สามารถใช้งานได้ โดย Cloudflare จะเข้ามารับการโจมตีดังกล่าวแทนเว็บไซต์

3. Content Delivery Network (CDN) คือ การกระจายเนื้อหาออกไปตาม Server จุดต่าง ๆ หากมี Traffic ระบบก็จะส่งข้อมูลโดยใช้ Server ที่อยู่ใกล้ที่สุด โดย Cloudflare มี POPs ในไทยมากถึง 6 POPs และมากกว่า 200 POPs ทั่วโลก ช่วยให้เว็บไซต์สามารถใช้งานได้อย่างรวดเร็ว และสเถียร

2.1.10 ทฤษฎี Cors

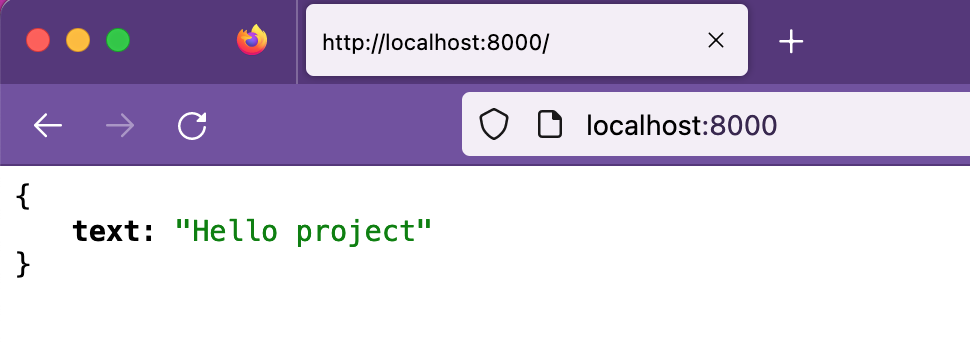
Cross-Origin Resource Sharing (CORS) คือ เป็นกลไกที่ใช้ HTTP headers เพิ่มเติมเพื่อให้บราวเซอร์ได้รับสิทธิ์ในการเข้าถึงทรัพยากรที่เลือกจากเซิร์ฟเวอร์บนโดเมนอื่นมาแสดงบนหน้าเว็บบราวเซอร์ได้ คอมพิวเตอร์แต่ละเครื่องต้องมี Protocol ที่เหมือนกัน ถึงจะสื่อสารกันรู้เรื่อง เว็บบราวเซอร์จะส่ง HTTP request เมื่อต้องการขอข้อมูลข้ามโดเมนหรือ port ที่ต่างกัน และต้องทำตามข้อตกลงการสื่อสาร (Protocol) เพราะปัจจุบันเรามักจะแยกฝั่ง Front-end และ Back-end ออกจากกันเป็นคนละโดเมน ด้วยเหตุผลเรื่องความปลอดภัยของ Browsers HTTP การอนุญาตให้เข้าถึงแหล่งข้อมูลจะต้องอยู่โดเมนเดียวกันเท่านั้น เว้นแต่ว่าแหล่งข้อมูลนั้นจะอนุญาตให้โดเมนของ Browsers สามารถเข้าถึงข้อมูลเหล่านั้นได้ (TAeng Trirong, 2017)

2.1.11 ทฤษฏี Fastapi

fastAPI คือ Framework สำหรับสร้าง Web API ด้วยภาษาที่ได้รับความนิยมคือ Python โดยถูกออกแบบมาให้ง่ายต่อการพัฒนา และสามารถที่จะสร้าง API ขึ้นมาได้อย่างรวดเร็ว โดยประสิทธิภาพการทำงานนั้นเร็วFastAPI นั้นรองรับการทำงานแบบ Asynchronous และมีเจ้าตัว Uvicorn เป็นตัว run server ข้อดีของการใช้งาน fastAPI คือ 1.มีความเร็วของการทำงาเทียบเท่า Node.js และ Go 2.รูปแบบการเขียนฟังก์ชันต่าง ๆ เข้าใจง่ายต่อการศึกษา   
3.ง่ายต่อการใช้งานและพัฒนาต่อ (Natakorn, 2021)



**ภาพที่ 8** ตัวอย่างโค้ดสำหรับการสร้าง Web API ด้วย fastAPI



**ภาพที่ 9** ผลลัพธ์แสดงคำว่า Hello project จาก fastAPI

2.1.12 ทฤษฎี Git

Git คือ Version Control ที่ถูกพัฒนาขึ้นมาเพื่อใช้ในกระบวนการพัฒนาซอฟต์แวร์ แปลให้เข้าใจแบบง่าย ๆ คือ ระบบที่ถูกพัฒนาขึ้นมาเพื่อใช้สำหรับการติดตาม ตรวจสอบ การพัฒนา แก้ไข ซอร์ซโค้ด ซอร์ซไฟล์ ต่าง ๆ ในขั้นตอนการพัฒนา ที่สามารถตรวจสอบได้ทุกตัวอักษร ทุกบรรทัด ทุกไฟล์ ที่มีการแก้ไข ใครเป็นคนแก้ไข และแก้ไข ณ วันที่เท่าไหร่

ระบบการทำงานของ Git ไม่ได้อยู่แค่การตรวจสอบการแก้ไขเท่านั้น ยังสามารถรวมการแก้ไขทั้งหมดเข้าด้วยกันได้อย่างชาญฉลาด เราเรียกขั้นตอนนี้ว่า CI ( Continuous Integration ) และในปัจจุบัน Git VCS ( Version Control System ) มีการควบรวมฟีเจอร์ที่ทำให้นักพัฒนาทำงานได้สะดวกมากขึ้น สามารถทำงานได้ตั้งแต่ขั้นตอนการพัฒนา ไปจนถึงการ Deploy งานขึ้นใช้งานบน Server เราเรียกขั้นตอนนี้ว่า CD ( Continuous Deployment ) รูปแบบการใช้งานของ Git มีด้วยกัน 2 รูปแบบ คือ

* ใช้งานผ่าน Git Command Line ( ใช้งานผ่านการพิมพ์คำสั่งด้วยตัวหนังสือ )
* ใช้งานผ่านโปรแกรม Git GUI ( ใช้งานผ่านโปรแกรมสำเร็จรูป )

2.1.13 ทฤษฎี Node.js

Node.js คือสภาพแวดล้อมการทำงานของภาษา JavaScript นอกเว็บเบราว์เซอร์ที่ทำงานด้วย V8 engine นั่นหมายความว่าเราสามารถใช้ Node.js ในการพัตนาแอพพลิเคชันแบบ Command line แอพพลิเคชัน Desktop หรือแม้แต่เว็บเซิร์ฟเวอร์ได้ โดยที่ Node.js จะมี APIs ที่เราสามารถใช้สำหรับทำงานกับระบบปฏิบัติการ เช่น การรับค่าและการแสดงผล การอ่านเขียนไฟล์ และการทำงานกับเน็ตเวิร์ก เป็นต้น

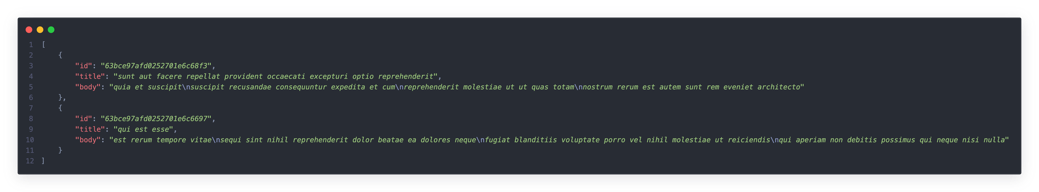
Node.js ถูกพัฒนาและทำงานด้วยใช้ Chrome V8 engine สำหรับคอมไพล์ภาษา JavaScript ให้เป็นภาษาเครื่องด้วยการคอมไพล์แบบ Just-in-time (JIT) เพื่อเพิ่มประสิทธิภาพการทำงานของภาษา JavaScript จากที่แต่เดิมมันเป็นภาษาที่มีการทำงานแบบ Interpreted Node.js เป็นโปรแกรมที่สามารถใช้ได้ทั้งบน Windows, Linux และ Mac OS X นั่นหมายความว่าคุณสามารถเขียนโปรแกรมในภาษา JavaScript และนำไปรันได้ทุกระบบปฏิบัติการที่สนับสนุนโดย Node.js นี่เป็นแนวคิดของการเขียนครั้งเดียวแต่ทำงานได้ทุกที่ (Write once, run anywhere) ข้อดีอีกอย่างหนึ่งในการใช้ภาษา JavaScript ของ Node.js คือทำให้การพัฒนาเว็บไซต์ทำได้อย่างขึ้นสำหรับนักพัฒนา เนื่องจากพวกเขาสามารถใช้ภาษา JavaScript สำหรับทั้ง Front-end และ Back-end ได้โดยไม่ต้องศึกษาภาษาเฉพาะในแต่ละด้าน ตัวอย่างของการพัฒนาเว็บโซต์ในรูปแบบนี้ เช่น React.js ซึ่งเป็นไลบรารีโดย Facebook (ทำความรู้จักกับ Node.js, 2021)

2.1.14 ทฤษฎี Matplotlibs

Matplotlib เป็นโมดูลที่เป็นพื้นฐานของ Python สำหรับการวาดกราฟจากข้อมูลซึ่งจำเป็นมากสำหรับงานทางด้าน Data Analysis, Science, Engineering เป็นตัวช่วยในการวิเคราะห์ข้อมูลโดยใช้รูปแบบของกราฟตัวอย่างประเภทกราฟที่มีให้ใช้ 1.Scatter 2.Bar 3.Stem 4.Step และอื่น ๆ

2.1.15 ทฤษฎี Mongodb

MongoDB เป็น open-source document database โดยเป็นฐานข้อมูลแบบ NoSQL คือไม่มี relation (ความสัมพันธ์) ของตารางแบบ SQL ทั่วๆไป แต่จะเก็บข้อมูลเป็นแบบ JSON (JavaScript Object Notation) แทน การบันทึกข้อมูลทุกๆ record ใน MongoDB เราจะเรียกมันว่า Document ซึ่งจะเก็บค่าเป็น key และ value จะเห็นว่ามันก็คือ JSON นั่นแหละ ตัวอย่างเช่น (Chai, 2015)



**ภาพที่ 10** ตัวอย่างข้อมูลแบบ JSON

ข้อดีของตัวนี้เราจะพูดในหัวข้อถัดไป โดยหลักๆ ก็คือเหมาะกับองค์กรที่อยากจะเข้าถึงข้อมูลได้อย่างรวดเร็ว อีกทั้งยังเหมาะกับการทำ Big Data (PLC, 2022)

**ข้อดีของ MongoDB**

1. ตัว MongoDB สามารถที่จะสร้างเป็น Cluster เพื่อที่จะตอบสนองของคำว่า High Availability (HA) ได้ ซึ่งเราอาจจะเลือก Region ที่เราอยากจะ Deploy บน Cloud Provider นั้นๆ ได้
2. ความรวดเร็วในการเข้าถึงข้อมูล เพราะว่าตัว Database ของมันไม่มี Schema ซึ่งจะต่างกับ SQL โดยพวกนั้นจะอิงจากฐานข้อมูลที่มาจาก Table
3. สามารถทำ Auto Scale ได้ไม่ว่าจะมีการใช้งานมากน้อยแค่ไหน ตัวมันก็สามารถ Adapt กับ Environment นั้นๆ ได้
4. รองรับ Multiple Cloud Provider ซึ่งข้อดีข้อนี้มันจะทำให้ Database ของเรามี High Availability มากขึ้นโดยเราไม่จำเป็นที่จะต้องยึดติดกับ Cloud Provider เจ้าใดเจ้าหนึ่ง

2.1.16 ทฤษฎี Next.js

Next.js เป็น React Web Framework คล้าย ๆ กับ Create React App ที่ช่วยให้เราเขียนเว็บได้สะดวกขึ้น เพราะ Setup และ Config ให้เรียบร้อยครบถ้วน ยกตัวอย่างข้อดีของ Next.js เช่น (Pallop, 2017)

* SSR (server-side rendering)
* Hot rendering
* Static HTML file exportable
* Project Structure
* Routing
* Easy setting up & installation

สามารถทำเว็บไซต์ได้ทั้งแบบ static และ dynamic ซึ่งข้อดีของการเป็น server side rendering คือ ช่วยในเรื่อง SEO หรือ search engine optimization เพราะถ้าทำการ inspect เว็บไซต์ที่สร้างโดย Next.js จะเห็นว่า source จะเป็น html ส่วนใหญ่ ซึ่งทำให้ SEO ค้นผ่าน source เพื่อให้ได้ข้อมูลและจัดหมวดหมู่ได้ง่ายกว่า React ที่เป็น JavaScript มากกว่า ทำให้ Next.js เป็นที่นิยมในหลายๆบริษัท นอกจากนี้ ข้อดีก็คือ render ได้เร็วกว่า React เพราะ Next.js มีสิ่งที่เรียกว่า get static path ซึ่งการสร้าง path แบบ static แบบเว็ปไซต์ html โดยไม่ต้องทำการเชื่อมต่อกับ back end เพื่อให้ได้ data ยิ่งไปกว่านั้น Next.js สามารถรวมเข้ากับ backend ได้ง่ายๆ เพราะ Next.js มีสิ่งที่เรียกว่า API routes ในการรับส่ง request ใน folder ของ page จะมีอีก folder ที่เรียกว่า API ที่ถูกปฏิบัติเป็น endpoint แทนที่จะเป็น page ซึ่ง folder API นี้จะเป็นในส่วนหนึ่งของ server-side เท่านั้น ทำให้ไม่ไปเพิ่ม size ของ client side (frevation, 2021)

2.1.17 ทฤษฎี Numpy

NumPy เป็น Library(ไลบรารี่) ที่รู้จัก และเป็นที่นิยมใช้ในการคำนวณเช่น ใช้คำนวณ Matrix หรือ คำนวณกับ Array ในงาน Data Science, Data analytics และในการทำ Machine Learning (การเรียนรู้ของเครื่องจักร) หรือ Deep Learning (ดีพ เลินนิ่ง) ก็ยังคงต้องใช้ NumPy อยู่ดี NumPy เป็น Library พื้นฐานที่ใช้คำนวณทางคณิตศาสตร์ด้วยภาษา Python สามารถคำนวณ หรือ ดำเนินการทางตรรกะใน Array หลายมิติ หรือ Matrix ได้อย่างรวดเร็ว เพราะ Library เขียนด้วยภาษา C ที่ Compile ไว้แล้ว (mindphp)

Numpy นั้นได้แรงบันดาลใจมาจาก MATLAB ดังนั้นผู้ที่มีประสบการณ์ด้าน MATLAB อยู่แล้วจะทำความเข้าใจ Numpy ได้ไม่ยาก โดยหลักการของ คือการนิยามตัวแปร array หลายมิติ ที่เราคุ้นเคยในคณิตศาสตร์ อาทิเช่น เวกเตอร์ (1 มิติ) เมตริกซ์ (2 มิติ) เทนเซอร์ ( 3 มิติขึ้นไป) เป็นต้นและ operations ของมัน ในการทำความเข้าใจ Numpy นั้นเพื่อนๆ ควรมีความรู้พื้นฐาน Linear algebra พวก vector / matrix ในระดับหนึ่งครับ (JUNG, 2019)

2.1.18 ทฤษฎี Pandas

pandas คือ หนึ่งใน Library สำคัญของภาษา Python เริ่มพัฒนาโดย Wes McKinney นักพัฒนาซอฟต์แวร์ชาวอเมริกัน ปัจจุบัน pandas เป็น open source ให้ทุกคนสามารถใช้ได้แบบฟรี pandas มาจากคำว่า Panel Data (ชุดข้อมูลหลายมิติ) มีจุดเด่นด้านการวิเคราะห์ข้อมูล (Data Analysis) และการทำความสะอาด (Data Cleaning) ซึ่งเป็น Process ที่สำคัญมากในการทำงานกับข้อมูล pandas มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพตั้งแต่ข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่ ทำให้ pandas ตอบโจทย์งานในยุคที่ข้อมูลมีขนาดใหญ่มากขึ้นเรื่อย ๆ ได้ ไม่มีปัญหาติดขัดเหมือนกับ Spreadsheets อื่นๆ (เช่น Excel หรือ Google Sheets ซึ่งจะทำงานได้ช้าลงหากข้อมูลมีขนาดใหญ่ขึ้น) ขั้นตอนการเตรียมข้อมูลนั้นมีความสำคัญมาก และ Data Scientist อาจจะใช้เวลาส่วนใหญ่หมดไปกับขั้นตอนนี้ เพราะหากข้อมูลที่เตรียมได้ไม่มีประสิทธิภาพ การนำ Insights ไปใช้งาน หรือนำข้อมูลไปสร้างโมเดล ย่อมทำให้ได้ข้อมูลที่ไม่น่าเชื่อถือนั่นเอง (Garbage in, Garbage out)

นอกจากนี้ เมื่อเปรียบเทียบกับ Tools วิเคราะห์ข้อมูลอื่นๆ อย่าง Excel หรือ Google Sheets อาจไม่ตอบโจทย์เต็มที่หากต้องการเชื่อมต่อกับแหล่งข้อมูลบางประเภท หรือทำ Automation (ระบบจัดการอัตโนมัติ) ในขณะที่ pandas ซึ่งเป็นส่วนหนึ่งของ Python นั้นสามารถใช้การเขียนโค้ด เพื่อปรับแต่ง หรือเชื่อมต่อกับโปรแกรมอื่นๆ ได้สะดวก (Panchart, 2021)

2.1.19 ทฤษฎี Pythainlp

ในส่วนของ Module pythainlp ก็เป็นเหมือนกันตัว library ที่รวมคำสั่งเกี่ยวกับที่เกี่ยวกับภาษาไทยใน Python ซึ่งก็เป็นตัวช่วยให้การทำงานเกี่ยวกับตัวของภาษาไทย มีประสิทธิภาพและสะดวกมากขึ้น ในการทำงานของ pythainlp ก็จะมีการทำงานเช่น การตัดคำ การแปลไทยเป็นอังกฤษ และการเข้้าถึงรหัส Soundex และยังมีการทำงานๆที่เกี่ยวกับตัวของภาษาไทย มีการแสดง เกี่ยวกับเซ็ตของภาษาไทยทั้งหมด ยังมีในส่วนของการแยกด้วย เช่นเป็นส่วนของพยัญชนะ สระ วรรณยุกต์ เป็นต้น ยังมีในส่วนของเลขไทย มีการเช็คว่าเป็นคำภาษาไทยรึไม่ มีการรนับตัวอักษร ว่าเป็นภาษาไทยกี่เปอร์เซ็นต์ และยังมีส่วนของการแสดงคำอ่านที่เป็นพวกเวลาและยังมีการจัดเรียงคำใน List ให้เรียงกันเป็นลำดับได้ เป็นต้น (mindphp, การใช้งานต่างๆใน PyThaiNLP, 2022) นอกจากนี้ยังใช้สำหรับประมวลผลข้อความ และการวิเคราะห์ทางภาษา คล้ายกับ NLTK แต่ใช้กับภาษาไทยโดยเฉพาะ มีฟังก์ชันการทำงานที่หลากหลาย เช่น Character Set อักษรไทย คำไทย, เรียงคำภาษาไทย, Stop Words ภาษาไทย, ตัดคำภาษาไทย, วิเคราะห์ชนิดของคำทางไวยากรณ์, ตรวจตัวสะกด แก้คำผิด และอีกมากมาย (Surapong, 2020)

2.1.20 ทฤษฎี Scikit-learn

Scikit-learn (เรียกอีกอย่างว่า sklearn) นำเสนอแบบจำลองทางสถิติและการเรียนรู้ของเครื่องที่หลากหลาย แตกต่างจากโมดูลส่วนใหญ่ sklearn ได้รับการพัฒนาใน Python มากกว่า C แม้จะได้รับการพัฒนาใน Python ก็ตาม ประสิทธิภาพของ sklearn นั้นถูกกำหนดให้ใช้ NumPy สำหรับการดำเนินการพีชคณิตเชิงเส้นและอาร์เรย์ที่มีประสิทธิภาพสูง

Scikit-Learn ถูกสร้างขึ้นโดยเป็นส่วนหนึ่งของโครงการ Summer of Code ของ Google และทำให้ชีวิตของนักวิทยาศาสตร์ข้อมูลที่มี Python เป็นศูนย์กลางนับล้านทั่วโลกง่ายขึ้น ส่วนนี้ของซีรีส์มุ่งเน้นไปที่การนำเสนอไลบรารีและมุ่งเน้นไปที่องค์ประกอบเดียว นั่นคือการแปลงชุดข้อมูล ซึ่งเป็นขั้นตอนสำคัญและสำคัญที่ต้องทำก่อนพัฒนาแบบจำลองการทำนาย Scikit-learn เป็นแพ็คเกจ Python โอเพ่นซอร์สพร้อมการวิเคราะห์ข้อมูลที่ซับซ้อนและคุณสมบัติการขุด มันมาพร้อมกับอัลกอริทึมในตัวมากมายที่จะช่วยให้คุณได้รับประโยชน์สูงสุดจากโครงการวิทยาศาสตร์ข้อมูลของคุณ ห้องสมุด Scikit-learn ใช้วิธีดังต่อไปนี้ 1.Classification 2.Regression 3.Clustering 4.Dimensionality reduction 5.Model selection 6. Preprocessing (เจย์, 2021)

2.1.21 ทฤษฎี Vercel

Vercel คือ Cloud Platform ที่ให้บริการทำ Static Hosting Website ต่างๆ และสามารถทำ Serverless Functions บน Cloud รวมทั้งยังสามารถ Integrate และสร้าง Workflow ผ่าน GitHub เพื่อทำ Automated Deployment โค้ดของคุณได้อย่างง่าย Vercel Inc. เดิมชื่อ Zeit เป็นแพลตฟอร์มคลาวด์ของอเมริกาในฐานะบริษัทผู้ให้บริการ บริษัทรักษากรอบการพัฒนาเว็บไซต์ Next.js สถาปัตยกรรมของ Vercel สร้างขึ้นจาก Jamstack และการจัดการการปรับใช้ผ่านที่เก็บ Git Vercel เป็นสมาชิกของ MACH Alliance

**2.2 งานวิจัยที่เกี่ยวข้อง**

วุฒิชัย วิเชียรไชย (2013) ศึกษาเรื่อง การเปรียบเทียบวิธีการแบงแยกคําภาษาไทยดวยโครงสรางการเขียนกับโครงสรางพยางค์ งานวิจัยนี้นำเสนอการแบงแยกคําภาษาไทยโดยเทียบกับโครงสรางการเขียนของภาษาไทยและอัลกอริทึมการแบงแยกคําภาษาไทยโดยโครงสรางพยางค์ เพื่อศึกษาและเปรียบเทียบวิธีการประมวลผลของการแบงแยกคําภาษาไทยและประสิทธิภาพความถูกตองของอัลกอริทึม โดยสามารถแบงงานวิจัยในการแบงแยกคําภาษาไทยไดเปนดังนี้คือวิธีการใชกฎ (Rule base approach) วิธีการใชอัลกอริทึม (Algorithm ap-proach) วิธีการใชพจนานุกรม (Dictionary base approach) และวิธีการใชคลังขอความ (Corpus based approach) ผูวิจัยจึงไดเสนอวิธีการแบงแยกคําภาษาไทย โดยใชโครงสรางการเขียนภาษาไทยเพื่อแกไขลดพื้นที่ในการจัดเก็บคําศัพทในพจนานุกรมและวิธีการแบงแยกคําภาษาไทยดวยโครงสรางพยางคเพื่อลดการสิ้นเปลืองพื้นที่ในการจัดเก็บพจนานุกรม ยกตัวอยางการแบงแยกคําและพยางคของคําวา“ประเทศไทย” จะสามารถแบงแยกคําไดเปน “ประเทศไทย” และแบงพยางคไดเปน “ประ|เทศ|ไทย” จากผลลัพธในการแบงแยกคํานั้นยังขาดความถูกตองในการแบงแยกคําซึ่งสามารถพัฒนาแนวคิดในการศึกษาและสรางกฎเพื่อแบงแยกคําใหถูกตองมากยิ่งขึ้น

จักรินทร์ สันติรัตนภักดี และศุภกฤษฏิ์ นิวัฒนากูล (2021) ศึกษาเรื่อง การออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะเพื่อติดแท็กปัญหาการให้บริการ องค์การขนส่งมวลชนกรุงเทพ (ขสมก.) มีช่องทางในการร้องเรียนรถโดยสารสาธารณะผ่านเว็บบอร์ด ที่ผู้ใช้งานสามารถแสดงความคิดเห็นได้อย่างอิสระ ผู้วิจัยจึงออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ จากข้อร้องเรียนผ่านเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพด้วยกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม แล้วคัดเลือกคำศัพท์ด้วยการวิเคราะห์น้ำหนักของคำ มาสร้างเป็นคลังคำศัพท์ แบ่งเป็น 4 คลาส ได้แก่ คลาสการขับขี่ คลาสผู้ขับขี่และพนักงานผู้ให้บริการ คลาสยานพาหนะและอุปกรณ์ให้บริการ และคลาสเวลาและการเดินรถโดยใช้มอดูลการตัดคำภาษาไทย (Thai Word Segmentation) ดยน าข้อความทั่วไปซึ่งอยู่ในรูปแบบประโยคมาแบ่งออกเป็นค าหรือคุณลักษณะ (Term/Feature) เพื่อแยกส่วนของข้อความออกจากกันก่อนนำไปประมวลผลในขั้นต่อไป แบ่งตามกระบวนการทำงานออกเป็น 3 กลุ่ม ได้แก่ 1) การตัดคำโดยใช้กฎ (Rule-Based Approach) 2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) 3) การตัดคำโดยใช้คลังคำศัพท์ (Corpus-Based Approach) จากการทดลองพบว่าอัลกอริทึมโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวมสูงที่สุด

ปารณีย์ พึ่งวิชา อานนท์ ทับเที่ยง และธัญญา สัตยาอภิธาน (2019) ศึกษาการแบ่งกลุ่มพฤติกรรมของผู้บริโภคที่ซื้อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์ ผู้วิจัยเก็บรวบรวมข้อมูลจากกลุ่มตัวอย่างจำนวน 400 คน ทำการวิเคราะห์แบ่งกลุ่ม ผู้บริโภคด้วยวิธี   
K-Means Clustering เป็น 2 กลุ่มซึ่งมีลักษณะเฉพาะในแต่ละกลุ่ม จากการวิเคราะห์ความแตกต่างระหว่างกลุ่มของทัศนคติด้านพฤติกรรมการซื้อและด้านส่วนประสมทางการตลาดที่มีผลต่อการตัดสินใจซื้อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์เมื่ออยู่ต่างกลุ่มกัน โดยการวิเคราะห์ความแปรปรวนทางเดียว(One-way ANOVA) พบว่า ด้านพฤติกรรมการซื้อทั้ง 2 กลุ่ม มีความถี่ในการซื้อต่างหูแฟชั่นไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนตัวแปรอื่น ๆ นั้นมีความแตกต่างกัน จากการวิเคราะห์แบ่งกลุ่มผู้บริโภคที่ซื้อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์ด้วยวิธี K-mean clustering สามารถจำแนกเป็น 2 กลุ่มโดยแต่ละกลุ่มมีลักษณะเฉพาะดังนี้ กลุ่มที่1 : กลุ่มกระเป๋าหนักจ่ายได้ถ้าชอบ ไม่ค่อยชอบออกสื่อ ลักษณะด้านประชากรศาสตร์ โดยส่วนใหญ่เป็นคนGeneration X เพศหญิงมากกว่าเพศชายมีระดับการศึกษาสูงกว่าปริญญาตรี กลุ่มที่2 : กลุ่มวัยสะอ้อน ชอบออกสื่อ ซื้อน้อยแต่บ่อยครั้ง ลักษณะด้านประชากรศาสตร์ โดยส่วนใหญ่เป็นคนGeneration Yเพศหญิงมากกว่าเพศชาย ส่วนใหญ่มีระดับการศึกษาสูงแต่จะน้อยกว่ากลุ่ม 1โดยมีระดับปริญญาตรีมากที่สุดโดยส่วนใหญ่เป็นพนักงานบริษัทเอกชน (ปารณีย์ พึ่งวิชา, 2019)

ธงชัย คลายคลึง วุฒิชัย สงางาม กิตติวงศ สุธรรมโน และพันธพงศอภิชาตกุลศ ศึกษาเรื่อง เทคนิคการคัดเลือกกลุมโหลดรายอาคารสําหรับรองรับแผนการติดตั้งระบบผลิตไฟฟาพลังงานแสงอาทิตยบนหลังคาเพื่อเพิ่มคาดรรชนีประสิทธิภาพการใชพลังงานไฟฟ้า บทความนี้ตองการนําเสนอเทคนิควิธีการคัดเลือกกลุมโหลดในแตละอาคารที่มีความเหมาะสมสําหรับติดตั้งระบบผลิตไฟฟาจากพลังงานแสงอาทิตยกรณีที่ติดตั้งบนหลังคาของอาคารในศูนยกลางมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานนครราชสีมา ในวิธีการของ K-Means Clustering   
เริ่มตนดวยการจัดแบงขอมูลออกเปน K กลุม กําหนดจุดศูนยกลางเริ่มตนจํานวน K จุดขั้นตอนตอไปคือการสรางกลุมขอมูลและความสัมพันธกับจุดศูนยกลางที่ใกลมากที่สุด จากผลการวิเคราะหการใชพลังงานไฟฟาทั้งหมด 34 อาคารดวยวิธีการ K-Mean Clustering ทําใหสามารถแยกแยะจัดกลุมโหลดรายอาคารไดอยางมีประสิทธิภาพซึ่งแบงได 3 กลุมโดยกลุมที่ 3 จํานวน 19 อาคารนั้นเปนกลุมอาคารที่มีความเหมาะสมทั้งดานพฤติกรรมการใชพลังงานไฟฟาและมีพื้นที่รองรับการติดตั้งระบบผลิตไฟฟาจากพลังงานแสงอาทิตยบนหลังคาได้ (Thongchai Klayklueng, 2019)

วรากร ประดิษฐ์กุล ปราลี มณีรัตน์ และ นิเวศ จิระวิชิตชัย ศึกษาเรื่อง ระบบแนะนํารถยนต์ให้กับลูกค้าโดยกํารวิเคราะห์จากการอ้างอิงถึงพฤติกรรมของผู้ใช้ (Collaborative Filtering) กรณีศึกษาบริษัท โตโยต้า บัสส์ จำกัด ผู้วิจัยได้พัฒนาระบบแนะนำรถยนต์ให้กับลูกค้าโดยการวิเคราะห์จากการอ้างอิงถึงพฤติกรรมของผู้ใช้ เพื่อช่วยให้ลูกค้าได้รับการแนะนำรุ่นรถยนต์ที่เหมาะสม ตรงตามความต้องการของลูกค้า ผู้วิจัยใช้อัลกอริทึมการหาความคล้ายคลึงกันของผู้ใช้ โดยวิเคราะห์จากลูกค้าที่มีพฤติกรรมใกล้เคียงกันด้วยสมการการหาความคล้ายโคไซน์ ซึ่งเป็นฟังก์ชั่นในภาษา Python ในการพัฒนาระบบแนะนำรถยนต์ให้กับลูกค้า ด้วยสมการความคล้ายโคไซน์ (cosine similarity) จากการทดลองเมื่อนำข้อมูลมาจัดลำดับคะแนนความชอบของผู้ใช้แต่ละคน เพื่อเป็นการเพิ่มความเร็วให้อัลกอริทึมของวิธีการกรองแบบร่วมมือ อีกทั้งระบบจะนำค่าความคล้ายคลึงโคไซน์ของผู้ใช้ในระบบกับผู้ใช้เป้าหมายมาทดสอบความแม่นยำของระบบด้วยค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ พบว่ามีค่าเท่ากับ 0.97 เมื่อกำาหนดค่า k ไว้เท่ากับ 5 สรุปได้ว่าระบบมีประสิทธิภาพในการแนะนำรุ่นรถยนต์ที่รวดเร็วและความแม่นยำอยู่ในระดับที่ดี (Warakorn Pradiskul, 2021)

**บทที่ 3**

**ขั้นตอนการดำเนินงาน**

การจัดทำโครงงานทางวิทยาการคอมพิวเตอร์ ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจด้วยเทคโนโลยีปัญญาประดิษฐ์ ผู้จัดทำจึงได้กำหนดขั้นตอนการดำเนินงานดังนี้

3.1 การเตรียมและวิเคราะห์ข้อมูล

3.2 การทำงานของระบบ

3.3 การวิเคราะห์และออกแบบระบบ

3.4 การออกแบบฐานข้อมูล

3.5 การออกแบบหน้าจอ

**3.1 การเตรียมและวิเคราะห์ข้อมูล**

Diagram

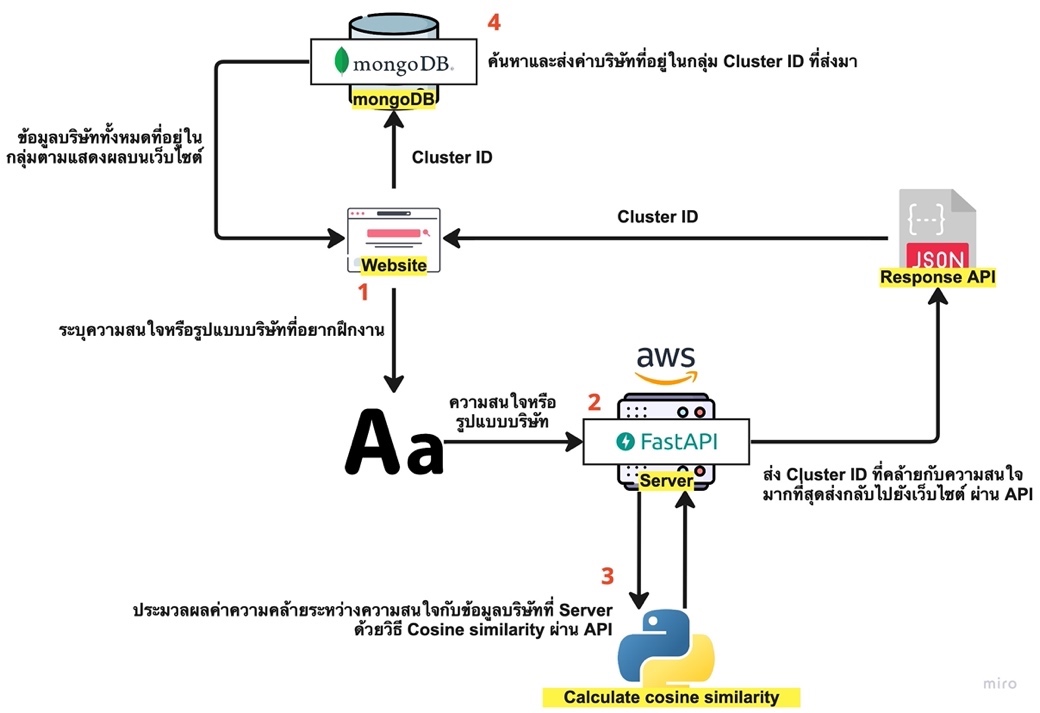
Description automatically generated

**ภาพที่ 11** การเตรียมและวิเคราะห์ข้อมูล

จากภาพที่ 11 แสดงการเตรียมและวิเคราะห์ข้อมูลได้ดังนี้

1. ข้อมูลบริษัทจากสมาคมปัญญาประดิษฐ์แห่งประเทศไทยโดยเป็นไฟล์ข้อมูลแบบ CSV (Comma-Separated Value)
2. ทำความสะอาดข้อมูลลบข้อมูลที่ไม่มีความหมายในตัว ลบตัวเลขที่ไม่จำเป็น แก้คำพิมพ์ผิดและอักษรพิเศษต่าง ๆ
3. เมื่อได้ข้อมูลที่พร้อมแล้วทำการจัดกลุ่มข้อมูลด้วยวิธี K-Means clustering
4. ตั้งชื่อของกลุ่มตามข้อมูลในกลุ่มนั้น ๆ
5. mongoDB เป็นฐานข้อมูลเพื่อเก็บข้อมูลต่าง ๆ ของบริษัทที่ทำการจัดกลุ่มแล้ว

**3.2 การทำงานของระบบ**



**ภาพที่ 12** การทำงานของระบบ

จากภาพที่ 12 แสดงการทำงานของระบบได้ดังนี้

1. Website ใช้ระบุความสนใจเพื่อส่งค่าไปประมวลผลความคล้ายคลึงกับข้อมูลบริษัท
2. Server ใช้ประมวลผลความคล้ายคลึงกันระหว่างความสนใจที่ได้รับมาและข้อมูลบริษัทที่อยู่ในฐานข้อมูลด้วยเทคนิค Cosine similarity โดยภาษา Python และส่งค่ากลับไปเป็น Cluster ID ผ่าน API
3. Response API เป็นข้อมูลในรูปแบบ JSON ที่มีค่า Cosine similarity ของแต่ละกลุ่มและ Cluster ID ที่มีความคล้ายคลึงกับความสนใจมากที่สุด
4. mongoDB เป็นฐานข้อมูลที่เก็บข้อมูลบริษัทไว้ และรอเว็บไซต์เรียกใช้ข้อมูลเพื่อนำไปแสดงผล

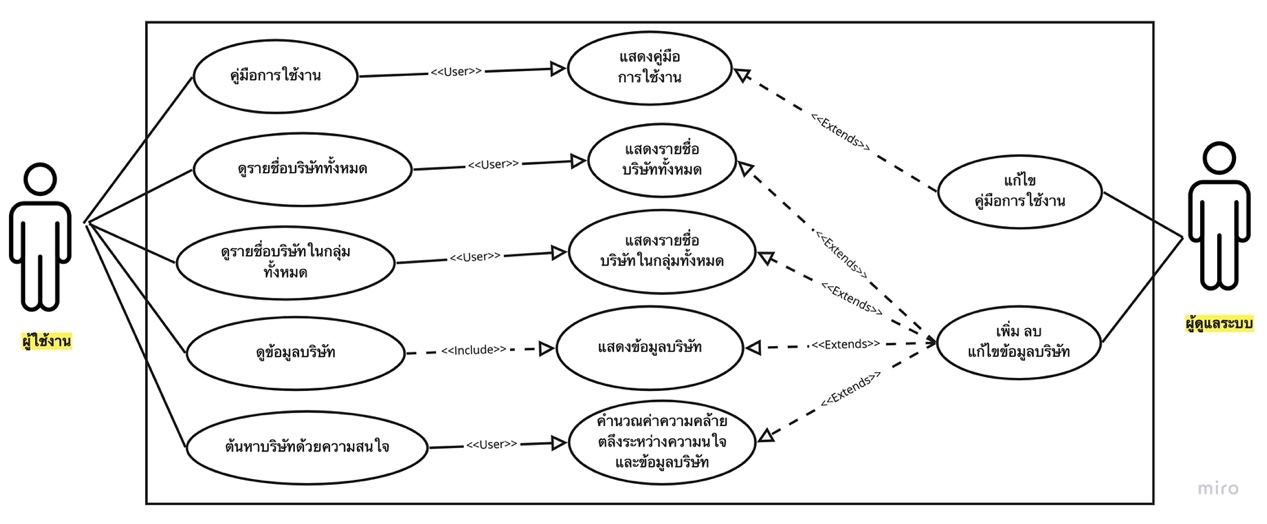
**3.3 การวิเคราะห์และออกแบบระบบ**

3.2.1 การวิเคราะห์ระบบ

การวิเคราะห์ระบบและการออกแบบ (System Analysis and Design) คือ วิธีการที่ใช้ในการสร้างระบบสารสนเทศขึ้นมาใหม่ในธุรกิจใดธุรกิจหนึ่งหรือระบบย่อยของธุรกิจ นอกจากการสร้างระบบสารสนเทศใหม่แล้ว การวิเคราะห์ระบบ ช่วยในการแก้ไขระบบสารสนเทศเดิมที่มีอยู่แล้วให้ดีขึ้น การวิเคราะห์ระบบ คือ การหาความต้องการ (Requirements) ของระบบสารสนเทศว่าคืออะไร หรือต้องการเพิ่มเติมอะไรเข้ามาในระบบ การออกแบบ คือ การนำเอาความต้องการของระบบมาเป็นแบบแผน หรือเรียกว่าพิมพ์เขียวในการสร้างระบบสารสนเทศนั้นให้ใช้งานได้จริง

3.2.2 ยูสเคสไดอะแกรม (Use Case Diagram)

ยูสเคสไดอะแกรม (Use Case Diagram) คือเเผนภาพที่ใช้เเสดงปฎิสัมพันธ์ระหว่างระบบงานเเละสิ่งที่อยู่นอกระบบงาน และเเสดงให้เห้นถึงส่วนประกอบทั้งหมด หรือ ภาพรวมของระบบ เป็นรากฐานในการเริ่มต้นการวิเคราะห์ระบบ โดยค้นหาคำว่าระบบทำอะไร โดยไม่สนใจกลไกการทำงานหรือเทคนิคการทำงานเปรียบเสมือน "กล่องดำ" โดย Use Case Diagram จะช่วยให้ผู้พัฒนาระบบมสามารถแยกแยะกิจกรรมที่อาจจะเกิดขึ้นในระบบ เป็น Diagram พื้นฐาน ที่สามารถอธิบายสิ่งต่าง ๆ ได้โดยใช้รูปภาพที่ไม่ซับซ้อน



**ภาพที่ 13** Use Case Diagram ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ

ด้วยเทคโนโลยีปัญญาประดิษฐ์

**ตารางที่ 1** คำอธิบาย Use case คู่มือการใช้งาน

|  |  |
| --- | --- |
| Use case id: | 1 |
| Use case name: | คู่มือการใช้งาน |
| Actor: | ผู้ใช้งาน |
| Scenario: | การดูวิธีการใช้งานเว็บไซต์ |
| Trigger event: | None |
| Brief Description: | อ่านวิธีการใช้งานเว็บไซต์ |
| Purpose: | เพื่อใช้งานเว็บไซต์ |
| Pre-condition: | เมื่อต้องการใช้งานเว็บไซต์ |
| Main flow: | 1. ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้าเกี่ยวกับ  2. อ่านวิธีการใช้งานเว็บไซต์ |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 2** คำอธิบาย Use case ดูรายชื่อบริษัททั้งหมด

|  |  |
| --- | --- |
| Use case id: | 2 |
| Use case name: | ดูรายชื่อบริษัททั้งหมด |
| Actor: | ผู้ใช้งาน |

**ตารางที่ 2 (ต่อ)**

|  |  |
| --- | --- |
| Scenario: | การดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล |
| Trigger event: | None |
| Brief Description: | ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล |
| Purpose: | ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล |
| Pre-condition: | เมื่อต้องการดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล |
| Main flow: | 1. ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้ารายชื่อบริษัททั้งหมด  2. ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 3** คำอธิบาย Use case ดูรายชื่อบริษัทในกลุ่มทั้งหมด

|  |  |
| --- | --- |
| Use case id: | 3 |
| Use case name: | ดูรายชื่อบริษัทในกลุ่มทั้งหมด |
| Actor: | ผู้ใช้งาน |
| Scenario: | การดูรายชื่อบริษัทในกลุ่มทั้งหมด |
| Trigger event: | กรณีที่แสดงผลจากการค้นหาด้วยความสนใจ หรือกรณีที่คลิกเมนูกลุ่มของบริษัท |
| Brief Description: | ดูรายชื่อบริษัทในกลุ่มที่ต้องการทั้งหมด |
| Purpose: | เพื่อดูรายชื่อบริษัทในกลุ่มทั้งหมด |
| Pre-condition: | เมื่อต้องการดูรายชื่อบริษัทในกลุ่มทั้งหมด |
| Main flow: | 1. ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้ากลุ่มบริษัทที่ต้องการ  2. ดูรายชื่อบริษัทในกลุ่มทั้งหมด |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 4** คำอธิบาย Use case ดูข้อมูลบริษัท

|  |  |
| --- | --- |
| Use case id: | 4 |
| Use case name: | ดูข้อมูลบริษัท |
| Actor: | ผู้ใช้งาน |
| Scenario: | การดูข้อมูลบริษัท |
| Trigger event: | None |
| Brief Description: | ดูข้อมูลบริษัท เช่น ข้อมูลติดต่อ จังหวัด และรูปแบบธุรกิจ |
| Purpose: | เพื่อดูข้อมูลบริษัท |
| Pre-condition: | เมื่อต้องการดูข้อมูลบริษัท |
| Main flow: | 1. ผู้ใช้งานคลิกที่เมนูชื่อของบริษัทที่ต้องการดูข้อมูล  2. ดูข้อมูลบริษัท |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 5** คำอธิบาย Use case ค้นหาบริษัทด้วยความสนใจ

|  |  |
| --- | --- |
| Use case id: | 5 |
| Use case name: | ค้นหาบริษัทด้วยความสนใจ |
| Actor: | ผู้ใช้งาน |
| Scenario: | การค้นหาบริษัทด้วยความสนใจ |
| Trigger event: | None |
| Brief Description: | ค้นหาบริษัทด้วยความสนใจหรือรูปแบบธุรกิจ |
| Purpose: | เพื่อค้นหาบริษัทด้วยความสนใจ |
| Pre-condition: | เมื่อต้องการค้นหาบริษัทด้วยความสนใจ |
| Main flow: | 1. ผู้ใช้งานพิมพ์ความสนใจหรือรูปบบธุรกิจที่ช่องค้นหาและ  2. แสดงรายชื่อบริษัทที่อยู่ในกลุ่มที่ระบบแนะนำ  3. ผู้ใช้คลิกเลือกบริษัทเพื่อดูข้อมูลบริษัท |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 6** คำอธิบาย Use case แก้ไขคู่มือการใช้งาน

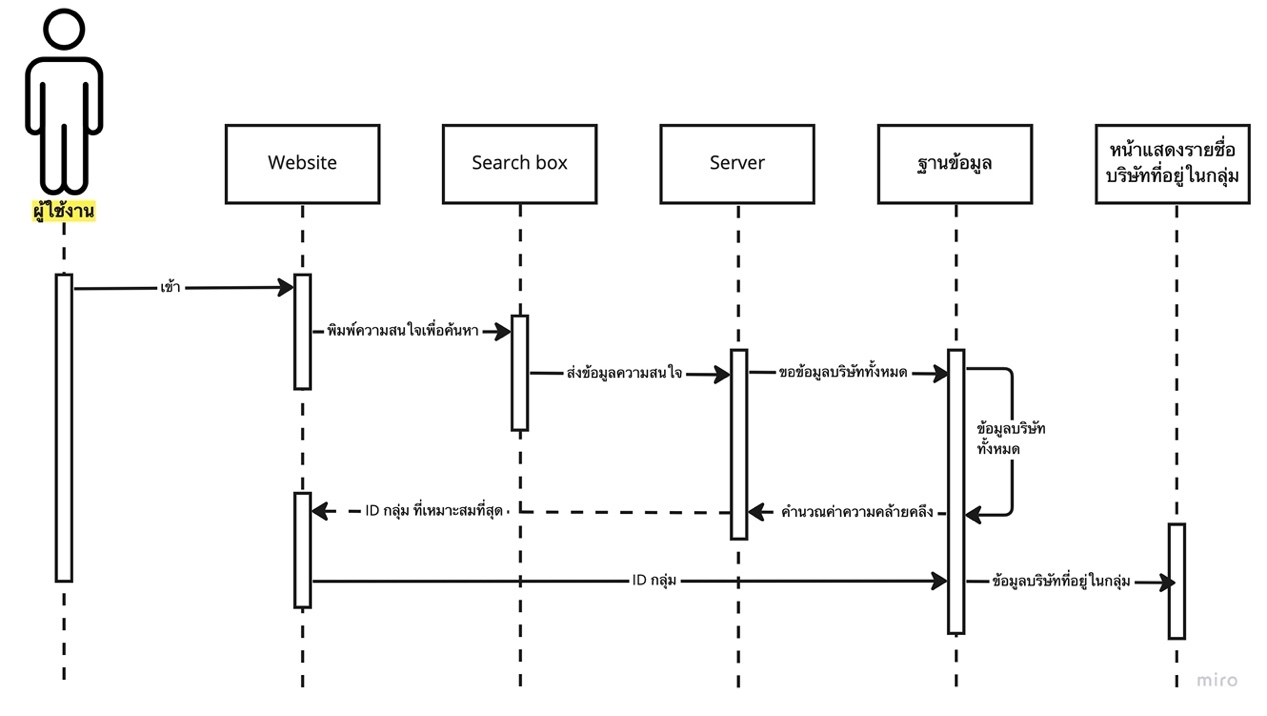
|  |  |
| --- | --- |
| Use case id: | 6 |
| Use case name: | แก้ไขคู่มือการใช้งาน |
| Actor: | ผู้ดูแลระบบ |
| Scenario: | การแก้ไขคู่มือการใช้งาน |
| Trigger event: | None |
| Brief Description: | แก้ไขคู่มือการใช้งาน |
| Purpose: | เพื่อแก้ไข คู่มือการใช้งาน |
| Pre-condition: | เมื่อต้องการแก้ไขคู่มือการใช้งาน |
| Main flow: | 1. ผู้ดูแลระบบแก้ไขข้อมูลคู่มือการใช้งาน  2. Deploy เพื่ออัพเดทระบบ |
| Alternate/Exceptional Flow: | None |

**ตารางที่ 7** คำอธิบาย Use case เพิ่ม ลบ แก้ไขข้อมูลบริษัท

|  |  |
| --- | --- |
| Use case id: | 7 |
| Use case name: | เพิ่ม ลบ แก้ไขข้อมูลบริษัท |
| Actor: | ผู้ดูแลระบบ |
| Scenario: | การเพิ่ม ลบ แก้ไขข้อมูลบริษัท |
| Trigger event: | None |
| Brief Description: | เพิ่ม ลบ แก้ไขข้อมูลบริษัท |
| Purpose: | เพื่อเพิ่ม ลบ แก้ไขข้อมูลบริษัท |
| Pre-condition: | เมื่อต้องการเพิ่ม ลบ แก้ไขข้อมูลบริษัท |
| Main flow: | 1. ผู้ดูแลระบบเพิ่ม ลบ แก้ไขข้อมูลบริษัท ในฐานข้อมูล |
| Alternate/Exceptional Flow: | None |

3.2.3 ซีเควนซ์ไดอะแกรม (Sequence Diagram)

Sequence Diagram เป็นหนึ่งในแผนผังการทำงานแบบ Unified Modeling Language (UML) ใช้สำหรับการสร้างแบบจำลองเชิงวัตถุ โดยข้อแตกต่างจากแผนผังรูปแบบ UML อื่น คือเป็นแผนผังการทำงานที่แสดงลำดับการปฏิสัมพันธ์ (Sequence of interactions)ระหว่างวัตถุที่แสดงภายในระบบต่างๆ อาทิเช่น การส่งข้อความ (messaging) ที่มีการรับส่งข้อมูลระหว่างผู้ใช้ Sequence Diagram เป็นแผนผังการทำงานที่ประกอบไปด้วยคลาส (Class) หรือวัตถุ (Object) เส้นประที่ใช้เพื่อแสดงลำดับเวลา และเส้นที่ใช้เพื่อแสดงกิจกรรมที่เกิดขึ้นจากคลาสหรือวัตถุในแผนผังการทำงานภายใน Sequence Diagram จะใช้สี่เหลี่ยมแทนเสมือนคลาสและวัตถุโดยภายในจะมีชื่อของคลาสหรือวัตถุประกอบอยู่ในรูปแบบ {Object}: Class



**ภาพที่ 14** Sequence Diagram การค้นหาบริษัทด้วยความสนใจของผู้ใช้

**ตารางที่ 8** อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การค้นหาบริษัทด้วยความสนใจของผู้ใช้

|  |  |
| --- | --- |
| **เหตุการณ์ที่เกิดขึ้น** | **คำอธิบาย** |
| เข้า Website | เข้า Website ด้วย Browser |
| พิมพ์ความสนใจเพื่อค้นหา | ระบุความสนใจรูปแบบธุรกิจของบริษัทหรือความสนใจที่อยากฝึกงานของผู้ใช้ |
| ส่งข้อมูลความสนใจ | ส่งข้อมูลความสนใจที่ผู้ใช้ระบุไปประมวลผลที่ Server |
| ขอข้อมูลบริษัททั้งหมด | Server ขอข้อมูลบริษัททั้งหมดจากฐานข้อมูลเพื่อนำมาเก็บไว้รอคำนวณค่าความคล้ายคลึง |
| ข้อมูลบริษัททั้งหมด | ข้อมูลบริษัททั้งหมดในฐานข้อมูล ส่งให้ Server |

**ตารางที่ 8 (ต่อ)**

|  |  |
| --- | --- |
| คำนวณค่าความคล้ายคลึง | คำนวณค่าความคล้ายคลึงระหว่างความสนใจของผู้ใช้และข้อมูลบริษัท |
| ID กลุ่มที่เหมาะสมที่สุด | คืนค่า ID ของกลุ่มบริษัทที่คล้ายกับความสนใจของผู้ใช้ |
| ID กลุ่ม | ส่งค่า ID ของกลุ่มไปยังฐานข้อมูลเพื่อขอข้อมูลบริษัทที่อยู่ในกลุ่มนั้น ๆ |
| ข้อมูลบริษัทที่อยู่ในกลุ่ม | แสดงรายชื่อบริษัทที่อยู่ในกลุ่มในหน้าเว็บ |

Diagram

Description automatically generated

**ภาพที่ 15** Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มบริษัทใหม่

**ตารางที่ 9** อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มบริษัทใหม่

|  |  |
| --- | --- |
| **เหตุการณ์ที่เกิดขึ้น** | **คำอธิบาย** |
| ได้รับข้อมูล | ได้ข้อมูลดิบที่จะนำมาใช้งาน |
| ทำความสะอาดข้อมูล | นำข้อมูลดิบมาทำการลบตัวเลข คำที่ไม่มีความหมายในตัว คำสะกดผิด |
| จัดกลุ่มข้อมูล | ทำการหาคำสำคัญและทำการจัดกลุ่มข้อมูล |
| กำหนดชื่อแต่ละกลุ่มและบันทึกลงฐานข้อมูล | กำหนดชื่อของกลุ่มและบันทึกข้อมูลลงฐานข้อมูลเพื่อใช้ในเว็บไซต์ |

3.2.4 แอคทิวิตี้ไดอะแกรม (Activity Diagram)

Activity Diagram หรือแผนภาพกิจกรรม ใช้อธิบายกิจกรรมที่เกิดขึ้นในลักษณะกระแสการไหลของการทำงาน (Workflow) จะมีลักษณะเดียวกับ Flowchart โดย ขั้นตอนในการทำงานแต่ละขั้นจะเรียกว่า Activity ดังภาพที่ 16 และภาพที่ 17 มีรายละเอียดดังนี้

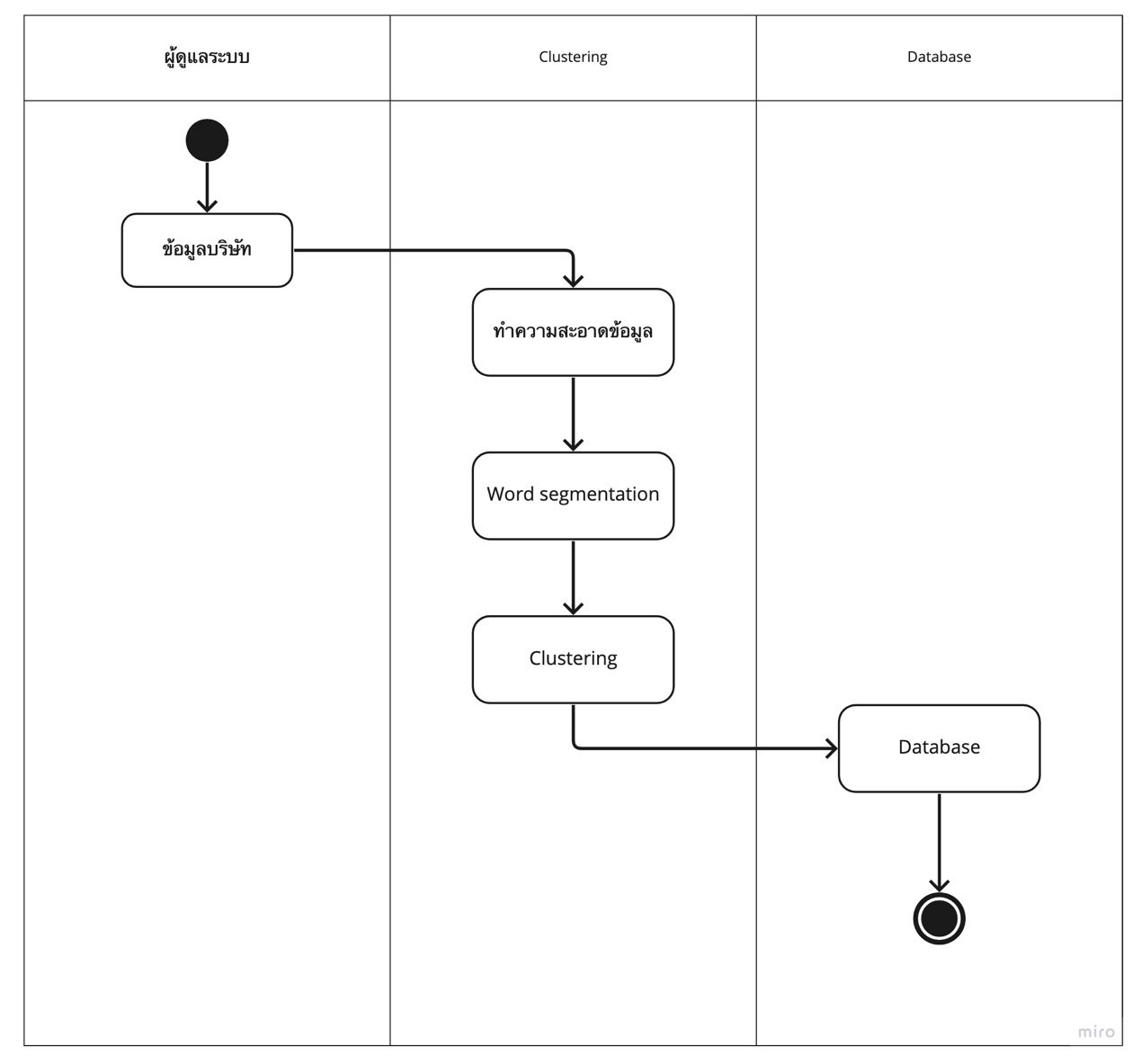
การใช้งาน Activity Diagram

1. อธิบายกระแสการไหลของการทำงาน (Workflow)
2. แสดงขั้นตอนการทำงานของระบบ

Diagram

Description automatically generated

**ภาพที่ 16** Activity Diagram ของผู้ใช้งาน



**ภาพที่ 17** Activity Diagram ของผู้ดูแลระบบ

**3.4 การออกแบบฐานข้อมูล**

3.4.1 E-R Diagram

ER Diagram คือ แบบจำลองที่ใช้อธิบายโครงสร้างของฐานข้อมูลซึ่งเขียนออกมาในลักษณะของรูปภาพ การอธิบายโครงสร้างและความสัมพันธ์ของข้อมูล (Relationship) ประกอบด้วย 1.เอนทิตี้ (Entity) เป็นวัตถุ หรือสิ่งของที่เราสนใจในระบบงานนั้น 2.แอททริบิว (Attribute) เป็นคุณสมบัติของวัตถุที่เราสนใจ 3.ความสัมพันธ์ (Relationship) คือ ความสัมพันธ์ระหว่างเอนทิตี้ ER Diagram มีความสำคัญต่อการพัฒนาระบบงานฐานข้อมูล Application ต่างๆ ที่ต้องการการเก็บข้อมูลอย่างมีระบบ มีโครงสร้าง ดังนั้น ER Diagram จึงใช้เพื่อเป็นเอกสารในการสื่อสารระหว่าง นักออกแบบระบบ และนักพัฒนาระบบ เพื่อให้สื่อสารอย่างตรงกัน

Graphical user interface, application

Description automatically generated with medium confidence

**ภาพที่ 18** ER Diagram ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ

ด้วยเทคโนโลยีปัญญาประดิษฐ์

3.4.2 พจนานุกรมข้อมูล (Data Dictionary)

หลังจากที่วิเคราะห์ระบบแล้ว ผู้ศึกษาได้ออกแบบฐานข้อมูล โดยออกแบบ

โครงสร้างของระบบ ซึ่งประกอบไปด้วยตาราง จำนวน 1 ตาราง และได้อธิบาย ชื่อ ตาราง(File Name), คําอธิบาย(Description), ชื่อข้อมูล(Field Name), ชนิดของข้อมูล(Type), ขนาดที่เก็บ(Length), ลักษณะที่เก็บค่า(Format), ชนิดของคีย์(Key) ดังต่อไปนี้

**ตารางที่ 10** พจนานุกรมข้อมูลบริษัท

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| File name: companies | | | | | |
| Description: ตารางเก็บข้อมูลบริษัททั้งหมด | | | | | |
| Field name | Type | Length | Format | Description | Key |
| \_id | String | 50 | ตัวอักษร | รหัสบริษัท | Primary key |
| short\_company | String | 50 | ตัวอักษร | ชื่อย่อบริษัท | Null |
| th\_company\_name | String | 50 | ตัวอักษร | ชื่อบริษัทภาษาไทย | Null |
| eng\_company\_name | String | 50 | ตัวอักษร | ชื่อบริษัทภาษาอังกฤษ | Null |
| type\_business | String | 50 | ตัวอักษร | ประเภทธุรกิจ | Null |
| product | String | 255 | ตัวอักษร | ประเภทสินค้า | Null |
| type\_innovation | String | 50 | ตัวอักษร | ประเภทเทคโนโลยี | Null |

**ตารางที่ 10 (ต่อ)**

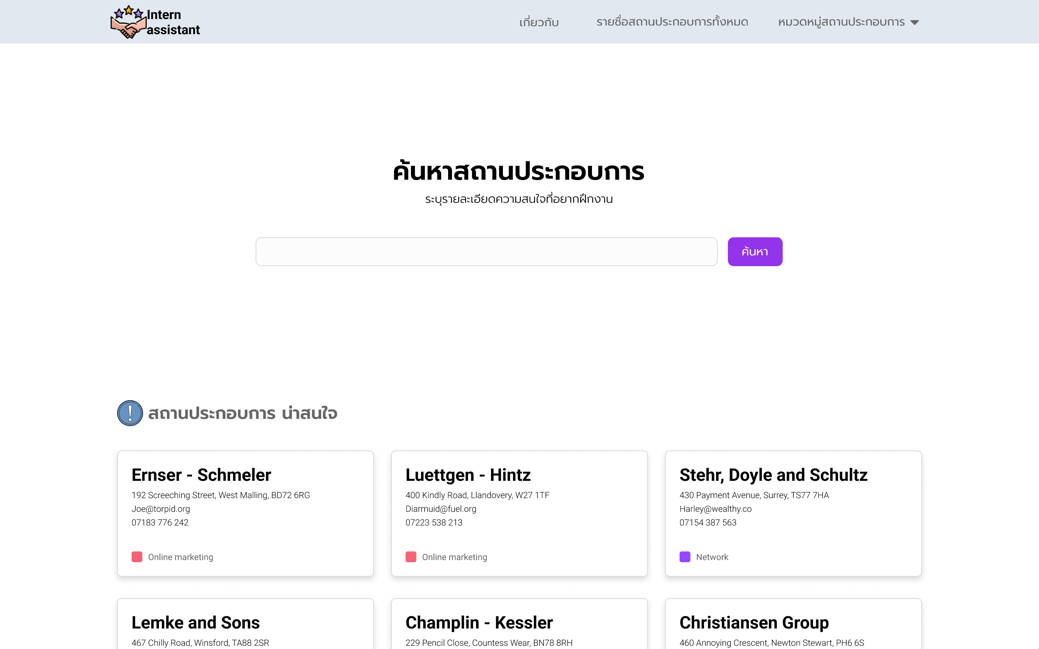
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| detail | String | 255 | ตัวอักษร | รายละเอียดธุรกิจ | Null |
| owner | String | 255 | ตัวอักษร | เจ้าของ | Null |
| province\_base | String | 50 | ตัวอักษร | จังหวัดที่ตั้ง | Null |
| address | String | 255 | ตัวอักษร | ที่อยู่โดยละเอียด | Null |
| phone\_number | String | 20 | ตัวอักษร | เบอร์โทร | Null |
| email | String | 50 | ตัวอักษร | อีเมล | Null |
| website | String | 50 | ตัวอักษร | เว็บไซต์ | Null |
| source | String | 255 | ตัวอักษร | ที่มาข้อมูล | Null |
| cluster | String | 1 | ตัวอักษร | กลุ่ม | Not Null |

**3.5 การออกแบบหน้าจอ**

การออกแบบหน้าจอหรือ UI design นั้นเป็นส่วนที่ผู้พัฒนาโปรแกรมต้องทำเนื่องจากหน้าจอนั้นเป็นส่วนสำคัญที่จะเชื่อมต่อกับผู้ใช้งานโปรแกรมหรือก็คือส่วนที่ผู้ใช้งานจะเห็น สั่งการ และโต้ตอบได้ ซึ่งจะมั่งเน้นการออกแบบในทางด้านของหน้าตา ปุ่ม ช่องที่ใช้พิมพ์สำหรับค้นหา ขนาดตัวอักษร สี และรูปภาพเป็นต้น

การออกแบบหน้าจอสำหรับการค้นหาสถานประกอบการสำหรับฝึกงานด้วยเทคโนโลยีปัญญาประดิษฐ์นั้น จะเน้นในแพลตฟอร์มที่เป็นหน้าจอคอมพิวเตอร์หรือ Desktop เป็นหลัก โดยมีรายละเอียดดังนี้

3.4.1 หน้าแรก



1

2

3

**ภาพที่ 19** หน้าแรก

หมายเลข 1 เมนูหลัก

หมายเลข 2 ช่องค้นหา

หมายเลข 3 รายชื่อบริษัทแนะนำ

3.4.2 หน้าเกี่ยวกับ

Graphical user interface, text, application

Description automatically generated

1

2

**ภาพที่ 20** หน้าเกี่ยวกับ

หมายเลข 1 เมนูหลัก

หมายเลข 2 รายละเอียดเกี่ยวกับเว็บไซต์

3.4.3 หน้าแสดงรายชื่อบริษัทในกลุ่มทั้งหมด

Graphical user interface, application

Description automatically generated

2

1

**ภาพที่ 21** หน้าแสดงรายชื่อบริษัทในกลุ่มทั้งหมด

หมายเลข 1 เมนูหลัก

หมายเลข 2 รายชื่อบริษัททั้งหมดที่อยู่ในกลุ่ม

3.4.4 หน้าแสดงรายชื่อบริษัททั้งหมด

Graphical user interface, text, application, email

Description automatically generated

1

2

**ภาพที่ 22** หน้าแสดงรายชื่อบริษัททั้งหมด

หมายเลข 1 เมนูหลัก

หมายเลข 2 รายชื่อบริษัททั้งหมด

3.4.5 หน้าแสดงผลลัพธ์รายชื่อบริษัท

Graphical user interface, application, Teams

Description automatically generated

3

2

1

**ภาพที่ 23** หน้าแสดงผลลัพธ์รายชื่อบริษัท

หมายเลข 1 เมนูหลัก

หมายเลข 2 ช่องค้นหา

หมายเลข 3 ผลลัพธ์รายชื่อบริษัทที่อยู่ในกลุ่มที่คล้ายคลึงกับความสนใจ

3.4.6 หน้าแสดงข้อมูลบริษัท

Graphical user interface, text, application, email

Description automatically generated

2

4

3

1

**ภาพที่ 24** หน้าแสดงข้อมูลบริษัท

หมายเลข 1 เมนูหลัก

หมายเลข 2 ชื่อบริษัท

หมายเลข 3 ที่อยู่และข้อมูลติดต่อบริษัท

หมายเลข 4 รูปแบบรายละเอียดธุรกิจของบริษัท

**บทที่ 4**

**ผลการดำเนินงาน**

การทำงานของระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจด้วยเทคโนโลยีปัญญาประดิษฐ์ มีขั้นตอนการทำงานดังนี้

**4.1 การวิเคราะห์การทำ Word segmentation**

4.1.1 การวัดค่าความแม่นยำในการตัดคำ

4.1.2 การจัดกลุ่มข้อมูล

4.1.3 การกำหนดชื่อกลุ่ม

4.1.1 การวัดค่าความแม่นยำในการตัดคำ

ในขั้นตอนการทำ Word segmentation นั้นมีขั้นตอนย่อยในการทำร่วมด้วยหลายขั้นตอนหลังจากทำความสะอาดข้อมูล คือการหาคำสำคัญของแต่ละประโยคในที่นี้คือรายละเอียดธุรกิจของแต่ละบริษัท การกำจัดคำที่ไม่สื่อความหมายหรือ Stop word เช่น และ ที่ นี้ อื่น เป็นต้น ออกไปจากประโยคเพื่อให้ได้ประโยคที่มีเนื้อหาใจความดีที่สุด และอีกขั้นตอนสำคัญคือการหาคีย์เวิร์ดของแต่ละประโยคเพื่อที่จะได้ทราบว่าประโยคนั้น ๆ กำลังสื่อถึงเรื่องไหนเป็นสำคัญด้วยเทคนิค TF-IDF ที่เป็นการหาน้ำหนักของคำนั้น ๆ ในประโยค

ในการทำงานประมวลผลเกี่ยวกับการประมวลผลภาษาธรรมชาติจำเป็นต้องมีการตัดคำออกเป็นคำ ๆ เพื่อจะได้ง่ายและนำไปเข้าสู่กระบวนการต่าง ๆ ได้อย่างง่ายโดยปกติแล้วการตัดคำในภาษาอังกฤษนั้นสามารถตัดได้โดยใช้การเว้นวรรคเป็นเงื่อนไขในการตัด แต่ในภาษาไทยนั้นการเขียนนั้นไม่ได้มีการเว้นวรรคคำเหมือนภาษาอังกฤษทำให้การตัดคำนั้นจะใช้เว้นวรรคมาตัดคำตลอดไม่ได้ จำเป็นต้องใช้อัลกอริทึมอื่น ๆ เข้ามาช่วยเช่นการใช้ Dictionary-based, Maximum Matching เป็นต้น

และในภาษา Python ก็มีเครื่องมือที่สามาถช่วยอำนวยความสะดวกในการตัดคำภาษาไทยอย่าง Pythainlp ที่ผู้วิจัยได้เลือกใช้ในโครงงานนี้ในไลบรารี่นั้นสามารถตัดคำได้หลาย Engine ด้วยกันและในแต่ละตัวเลือกก็ใช้อัลกอริทึมต่างกันยกตัวอย่างเช่น newmm, longest, newmm-safe, mm, icu, deepcut, attacut เป็นต้น ดังนั้นก่อนที่จะเลือก Engine มาใช้ตัดคำจำเป็นต้องมีการวัดค่าความถูกต้องเพื่อที่จะได้ผลลัพธ์ที่ตรงกับความต้องการมากที่สุดและในโครงงานนี้ผู้วิจัยได้เลือก Engine มาทดสอบด้วยกันจำนวน 3 ตัวเลือกดังนี้

1. newmm - dictionary-based, Maximum Matching + Thai Character Cluster
2. deepcut - wrapper for DeepCut, learning-based approach
3. longest - dictionary-based, Longest Matching

วิธีทำสอบคือทำการสุ่มเลือกประโยคมาจำนวน 100 ประโยค ทำการตัดคำในแต่ละประโยคเองโดยไม่ใช้ตัวช่วย ใช้ไลบรารี่ Pythainlp ในแต่ละ Engine ตัดคำและนำมาเทียบกับประโยคที่ผู้วิจัยตัดไว้โดย Engine ไหนที่มีความเหมือนกับที่ผู้วิจัยตัดไว้มากที่สุดก็จะถือว่ามีความใกล้เคียงกับความต้องการของผู้วิจัยมากที่สุด

**ผลการทดสอบการตัดคำ**

|  |  |
| --- | --- |
| **Engine** | **Accuracy (%)** |
| newmm | 90.99% |
| longest | 83.04% |
| deepcut | 76.65% |

จากการทดลองพบว่า Engine newmm ในไลบรารี่ Pythainlp นั้นมีความแม่นยำกับที่ผู้วิจัยตัดมากที่สุดที่ 90.99% รองลงมาที่ longest ค่าความแม่นยำอยู่ที่ 83.04% และ  
ความแม่นยำน้อยที่สุดคือ deepcut ที่ความแม่นยำ 76.65% ดังภาพที่ 25 ดังนั้นโครงงานนี้จึงใช้ Engine newmm ในการตัดคำเพื่อนำไปประมวลผลต่อในขั้นตอนอื่น ๆ

Graphical user interface, text, website

Description automatically generated

**ภาพที่ 25** ผลการวัดค่าความแม่นยำในการตัดคำของ Engine ในไลบรารี่ Pythainlp