



โครงการทางวิทยาการคอมพิวเตอร์ระดับปริญญาตรี

เรื่อง

ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัด
กลุ่มเมล็ด (K-Means)

COMPUTER INTERNSHIP RECOMMENDATION SYSTEM WITH
K-MEANS CLUSTERING

โดย

นายทินกฤต สิงห์แก้ว

มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน
ปีการศึกษา 2565



ใบรับรองโครงการวิทยาการคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน^๔
วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)

เรื่อง ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน
(K-Means)

Computer Internship Recommendation System With K-Means Clustering
โดย นายทินกฤต สิงห์แก้ว

คณะกรรมการพิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษา..... วันที่...../...../.....
(ผู้ช่วยศาสตราจารย์ ดร.นงนุช เกตุย)

อาจารย์ที่ปรึกษาร่วม..... วันที่...../...../.....
(อาจารย์วรวิทย์ พันคำอ้าย)

อาจารย์ผู้รับผิดชอบวิชา..... วันที่...../...../.....
(อาจารย์ปกรณ์ สุนทรเมธ)

ประธานหลักสูตร..... วันที่...../...../.....
(อาจารย์วรวิทย์ พันคำอ้าย)

โครงการทางวิทยาการคอมพิวเตอร์ระดับปริญญาตรี

เรื่อง

ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่ม
เคลื่อน (K-Means)

Computer Internship Recommendation System With K-Means
Clustering

โดย

นายทินกฤต สิงห์แก้ว

คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร
มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาศาสตร์
พ.ศ. 2565

บทคัดย่อ

ชื่อโครงงาน	: ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means)
	Computer Internship Recommendation System With K-Means Clustering
ผู้ศึกษา	: นายพินกฤต สิงห์แก้ว
อาจารย์ที่ปรึกษา	: ผู้ช่วยศาสตราจารย์ ดร. นงนุช เกตุชัย
อาจารย์ที่ปรึกษาร่วม	: อาจารย์วรวิทย์ ผันคำอ้าย
สาขาวิชา	: วิทยาศาสตร์
หลักสูตร	: วิทยาการคอมพิวเตอร์
ปีการศึกษา	: 2565

ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) (Computer Internship Recommendation System With K-Means Clustering) เป็นระบบที่ช่วยแนะนำสถานประกอบการสำหรับการฝึกงานของนักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน ในรูปแบบของเว็บแอปพลิเคชันที่พัฒนาโดยใช้ Next.js เว็บเฟรมเวิร์ค (Web Framework) สำหรับพัฒนาเว็บแอปพลิเคชัน โดยให้นักศึกษาระบุรายละเอียดความสนใจตามรูปแบบธุรกิจ หรือรูปแบบของงาน เพื่อนำมาวิเคราะห์หาความคล้ายคลึงกันข้อมูลสถานประกอบการที่มีอยู่ในฐานข้อมูล ที่ผ่านกระบวนการจัดกลุ่มข้อมูลเคลื่อน (K-Means) ซึ่งเป็นส่วนหนึ่งของเทคโนโลยีปัญญาประดิษฐ์ โดยการใช้เทคโนโลยีประมวลผลภาษาธรรมชาติ (Natural Language Processing) เพื่อจัดกลุ่มข้อมูลสถานประกอบการด้านคอมพิวเตอร์จากสมาคมปัญญาประดิษฐ์แห่งประเทศไทย ซึ่งการจัดกลุ่มข้อมูลด้วยเคลื่อน (K-Means) อยู่ในกลุ่มของการให้คอมพิวเตอร์เรียนรู้โดยไม่มีผู้สอน (Unsupervised Learning) โดยผลการทดลองใช้พบว่า นักศึกษาได้ใช้ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) (Computer Internship Recommendation System With K-Means Clustering) และมีค่าเฉลี่ยค่าความพึงพอใจอยู่ที่ 4.01 ซึ่งอยู่ในระดับดี

กิจกรรมประการ

โครงการทางวิทยาการคอมพิวเตอร์ “ระบบแนะนำสถานที่ทำงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคเม่น (K-Means) (Computer Internship Recommendation System With K-Means Clustering)” เพื่อการสำเร็จการศึกษาระดับปริญญาตรี สามารถดำเนินการจนประสบความสำเร็จลุล่วงไปด้วยดี เนื่องจากได้รับความกรุณาและคำแนะนำจากคณาจารย์หลาย ๆ ท่านในหลักสูตรวิทยาการคอมพิวเตอร์ ที่ได้กรุณาให้ความรู้และแนวทาง ข้อคิด ข้อแนะนำสู่ความสำเร็จและช่วยแก้ไขปัญหาอุปสรรคต่าง ๆ รวมทั้งรูปเล่มให้สำเร็จไปได้ด้วยดี

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. นงนุช เกตุชัย และอาจารย์วรวิทย์ พันคำ อ้ายอาจารย์ที่ปรึกษา และอาจารย์ประจำวิชาทุกท่าน ผู้ซึ่งกรุณาให้ความรู้ คำแนะนำแนวทาง การสร้างผลงานสู่ความสำเร็จ และช่วยแก้ไขปัญหาอุปสรรคต่าง ๆ รวมทั้งตรวจทานแก้ไข รูปเล่มจนเสร็จสมบูรณ์ ขอขอบคุณบิดา มารดา ผู้มีพระคุณทุกท่าน เพื่อนนักศึกษา และบุคคล ที่เกี่ยวข้องที่ยังไม่ได้กล่าวถึง ที่ได้ช่วยออกแบบคิดเห็น ได้ให้ข้อแนะนำ และคำนวณความ ละเอียดในด้านต่าง ๆ ในการทำโครงการครั้งนี้ไว้ ณ ที่นี่

สุดท้ายนี้ผู้ศึกษาหวังว่าโครงการฉบับนี้จะเป็นประโยชน์สำหรับมหาวิทยาลัย และ นักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน และผู้ที่สนใจที่จะศึกษาต่อไป

สารบัญ

	หน้า
บทคัดย่อ	๑
กิตติกรรมประกาศ	๒
สารบัญ	๓
สารบัญตาราง	๔
สารบัญภาพ	๕
บทที่ 1 บทนำ	๖
1.1 ความเป็นมาและความสำคัญของปัญหา	๑
1.2 วัตถุประสงค์	๑
1.3 ขอบเขตของโครงการ	๑
1.4 ประโยชน์ที่คาดว่าจะได้รับ	๒
1.5 อุปกรณ์เครื่องมือที่ใช้ในโครงการ	๒
1.6 นิยามศัพท์ที่ใช้ในโครงการ	๓
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	๕
2.1 ทฤษฎีและความรู้ที่เกี่ยวข้อง	๕
2.2 งานวิจัยที่เกี่ยวข้อง	๑๘
บทที่ 3 ขั้นตอนการดำเนินงาน	๒๑
3.1 การเตรียมและวิเคราะห์ข้อมูล	๒๑
3.2 การทำงานของระบบ	๒๘
3.3 การวิเคราะห์และออกแบบระบบ	๒๘
3.4 การออกแบบฐานข้อมูล	๓๕
3.5 การออกแบบหน้าจอ	๓๗
3.6 การใช้งานระบบ	๔๑
บทที่ 4 ผลการดำเนินงาน	๔๒
4.1 การวิเคราะห์และการตัดคำ (Word segmentation)	๔๒
4.2 ขั้นตอนการใช้งานสำหรับผู้ใช้งาน	๕๒
4.3 การวัดค่าความคล้ายคลึง	๕๕
บทที่ 5 สรุปผลการดำเนินงาน	๕๘
5.1 สรุปผลการดำเนินงาน	๕๘

สารบัญ (ต่อ)

	หน้า
5.2 สรุปปัญหาที่เกิดระหว่างการดำเนินงาน	59
5.3 แนวทางพัฒนาระบบในอนาคต	59
5.4 แบบประเมินความพึงพอใจของผู้ใช้ ภาคผนวก ก คู่มือการติดตั้งระบบ ภาคผนวก ข คู่มือการใช้งาน	59 64 80
ประวัติผู้ศึกษา	88

สารบัญตาราง

ตารางที่	หน้า
1 ตัวอย่างการคำนวณค่า Term Frequency ที่จำนวนคำทั้งหมดเท่ากับ 7	7
2 ตัวอย่างการคำนวณค่า Inverse Document Frequency ที่จำนวนเอกสารเท่ากับ 10	8
3 ตัวอย่างการคำนวณค่า TF-IDF	8
4 การวิเคราะห์ข้อมูล	21
5 คำอธิบาย Use case คุณมีการใช้งาน	29
6 คำอธิบาย Use case ดูรายชื่อบริษัททั้งหมด	30
7 คำอธิบาย Use case ดูรายชื่อบริษัทในกลุ่มทั้งหมด	30
8 คำอธิบาย Use case ดูข้อมูลบริษัท	30
9 คำอธิบาย Use case คนหาบริษัทด้วยความสนใจ	31
10 คำอธิบาย Use case แก้ไขคุณมีการใช้งาน	31
11 คำอธิบาย Use case เพิ่ม ลบ แก้ไขข้อมูลบริษัท	32
12 อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การค้นหาบริษัทด้วยความสนใจ	33
13 อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มใหม่	34
14 พจนานุกรมข้อมูลบริษัท	36
15 ตารางตัวอย่างการวัดค่าความแม่นยำในการตัดคำ	44
16 ผลการทดสอบความแม่นยำการตัดคำ	45
17 แสดงการนับจำนวนบริษัทแต่ละประเทศในการจัดกลุ่มทั้งหมด 1,643 รายการ	49
18 แสดงจำนวนค่าเฉลี่ยความพึงพอใจต่อระบบ	60

สารบัญภาพ

ภาพที่	หน้า
1 ตัวอย่างการทำ Word segmentation	7
2 การกำหนดสุ่มกำหนดจุด Centroid	9
3 จุด Centroid ที่อยู่ตรงกลางและจุดข้อมูลทุกจุดไม่เปลี่ยนแปลง	10
4 ตัวอย่างการกลุ่มข้อมูลที่มีจุด Centroid เป็นกลางบทสีแดง	10
5 กราฟที่แสดงจำนวนของผิดพลาดเพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด	11
6 ตัวอย่างโค้ดสำหรับการสร้าง Web API ด้วย fastAPI	13
7 ผลลัพธ์แสดงคำว่า Hello project จาก fastAPI	14
8 ตัวอย่างข้อมูลแบบ JSON	15
9 การเตรียมและวิเคราะห์ข้อมูล	22
10 การเรียกใช้ไลบรารี (Library) สำหรับคำนวนค่า TF-IDF	23
11 การอ่านข้อมูลจากไฟล์และกำหนดตัวกรองการตัดคำ	23
12 พังค์ชันสำหรับใช้ลบตัวเลข และอักษรพิเศษ	24
13 พังค์ชันสำหรับใช้ลบคำที่ไม่สื่อความหมายและตัวเลขโดย	24
14 การวนซ้ำข้อมูลเพื่อตัดคำและทำความสะอาดข้อมูล	24
15 การเหรอและการทดสอบโมเดลการคำนวนค่า TF-IDF	25
16 ผลลัพธ์การคำนวนค่า TF-IDF	25
17 การเรียกใช้ไลบรารีสำหรับการ จัดกลุ่มข้อมูลด้วยเคมีน (K-Means)	26
18 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k และการจัดกลุ่มข้อมูล	26
19 การเรียกใช้งานไฟล์ clustering.py เพื่อจัดกลุ่มข้อมูลและบันทึกผลลัพธ์	27
20 การนำเข้าข้อมูลลงสู่ฐานข้อมูล MongoDB	27
21 การทำงานของระบบ	28
22 Use Case Diagram ของระบบ	29
23 Sequence Diagram การคุนหาบริษัทด้วยความสนใจของผู้ใช้	32
24 Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มบริษัทใหม่	33
25 Activity Diagram ของผู้ใช้งาน	34
26 Activity Diagram ของผู้ดูแลระบบ	35
27 ER Diagram ระบบแนะนำบริษัทสำหรับผู้ใช้งานตามความสนใจ	36
28 หน้าแรก	38
29 หน้าเกี่ยวกับ	38
30 หน้าแสดงรายชื่อบริษัทในกลุ่มทั้งหมด	39
31 หน้าแสดงรายชื่อบริษัททั้งหมด	39
32 หน้าแสดงผลลัพธ์รายชื่อบริษัท	40
33 หน้าแสดงข้อมูลบริษัท	40
34 ตัวอย่างผลลัพธ์จากการค้นหาด้วยความสนใจของผู้ใช้	41
35 ตัวอย่างข้อมูลต้นฉบับ	43

สารบัญภาพ (ต่อ)

ภาคที่	หน้า
36 ตัวอย่างการตัดคำโดยใช้ Engine newmm	43
37 ตัวอย่างการตัดคำโดยใช้ Engine longest	43
38 ตัวอย่างการตัดคำโดยใช้ Engine deepcut	44
39 ผลการวัดค่าความแม่นยำในการตัดคำของ Engine ในไลบรารี Pythainlp	45
40 ตัวอย่างตาราง TF-IDF และน้ำหนักของคำ	46
41 ตัวอย่างการตัดคำและลบ Stop word	46
42 การทำ Elbow method	47
43 จัดกลุ่มข้อมูลจำนวน 9 กลุ่ม	47
44 จัดกลุ่มข้อมูลจำนวน 8 กลุ่ม	48
45 จัดกลุ่มข้อมูลจำนวน 7 กลุ่ม	48
46 จัดกลุ่มข้อมูลจำนวน 6 กลุ่ม	48
47 เปรียบเทียบอัตราการเติบโตการจัดกลุ่มข้อมูล	50
48 หน้าแรกเว็บไซต์ Intern-assistant	52
49 คนหาบริษัท	53
50 หน้าแสดงผลลัพธ์การค้นหา	53
51 หน้าเกี่ยวกับ	54
52 หน้าแสดงรายชื่อบริษัททั้งหมด	54
53 หน้าแสดงรายชื่อบริษัทในกลุ่มทั้งหมด	55
54 หน้ารายละเอียดบริษัท	55
55 ตัวอย่างการคำนวณค่า Cosine similarity	56
56 ตัวอย่างการคำนวณค่า Cosine similarity ผ่าน API และคืนค่าความคล้ายคลึง	56
57 ตัวอย่างการคำนวณค่า Cosine similarity ผ่าน API และคืนค่าเป็นข้อมูลบริษัทที่อยู่ในกลุ่มที่คล้ายที่สุด	57
58 โค้ดคำสั่งในไฟล์ clustering.py ใช้ในการจัดกลุ่มข้อมูล	65
59 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k และการจัดกลุ่มข้อมูล	66
60 แสดงการใช้งานคำสั่งจัดกลุ่มข้อมูลใน Terminal	66
61 แสดงไฟล์ clustered_company.csv	67
62 แสดงหน้าการจัดการ Cluster MongoDB	67
63 แสดงหน้าตั้งค่าและสร้าง Cluster MongoDB	68
64 แสดงหน้าสร้างบัญชีสำหรับจัดการฐานข้อมูล	68
65 แสดงหน้าเพิ่ม IP address ที่สามารถเชื่อมต่อฐานข้อมูลได้	69
66 แสดงหน้าจัดการ Cluster MongoDB	69
67 แสดงการสร้างฐานข้อมูล MongoDB	70
68 แสดงหน้าต่างการสร้างฐานข้อมูลและ Collection	70
69 ตัวอย่างการเลือกตั้งค่าการดาวน์โหลดโปรแกรม MongoDB compass	71
70 หน้าต่างเลือกเชื่อมต่อกับ Cluster	71

สารบัญภาพ (ต่อ)

ภาพที่	หน้า
71 หน้าต่างข้อมูลการเชื่อมต่อ Cluster กับ MongoDB compass	72
72 หน้าต่างโปรแกรม MongoDB compass สำหรับเชื่อมต่อ Cluster	72
73 หน้าต่างโปรแกรมแสดงข้อมูลใน Collection	73
74 หน้าต่าง Import ข้อมูลนามสกุลไฟล์ csv	73
75 หน้าต่างแสดงข้อมูลใน Collection ในโปรแกรม MongoDB compass	74
76 หน้าเว็บไซต์ Amazon Web Services	74
77 หน้าแสดงการเลือกสร้าง Instance ใหม่	75
78 หน้าแสดงการตั้งค่า Instance	75
79 ตัวอย่างการเชื่อมต่อเข้าไปยัง Instance	76
80 การดาวน์โหลดโปรเจคจาก Github ด้วยคำสั่ง git clone	76
81 สร้างไฟล์ใหม่ชื่อ .env และสร้างตัวแปรเพื่อกีบค่าเชื่อมต่อฐานข้อมูล	77
82 ตัวอย่างการรีมตัน Server Web API เพื่อคำนวณค่า Cosine similarity บน AWS	77
83 สร้างโปรเจคใหม่ใน Vercel	78
84 หน้าแสดงรายชื่อ Repository	78
85 หน้าการตั้งค่าโปรเจคตอน Deploy	79
86 ตัวอย่างหน้าเว็บไซต์	79
87 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k	81
88 แสดงการใช้คำสั่งจัดกลุ่มข้อมูลใน Terminal	81
89 แสดงไฟล์ clustered_company.csv	82
90 หน้าต่าง Import ข้อมูลนามสกุลไฟล์ csv	82
91 ข้อมูลใน Collection ในโปรแกรม MongoDB compass	83
92 หน้าเว็บไซต์ intern-assistant.vercel.app	83
93 หน้าแสดงผลลัพธ์เมื่อคนหาบริษัท	84
94 หน้าแสดงข้อมูลบริษัท	84
95 ตัวอย่างการส่งคำขอไปยัง https://iamonze.tech/allcompanies	85
96 ตัวอย่างการส่งคำขอไปยัง https://iamonze.tech/company/1	85
97 ตัวอย่างการส่งคำขอไปยัง https://iamonze.tech/cluster	86
98 ตัวอย่างการส่งคำขอไปยัง https://iamonze.tech/search	86
99 ตัวอย่างการส่งคำขอไปยัง https://iamonze.tech/searchcompany	87

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในระบบการศึกษาระดับปริญญาตรีนั้นรายวิชาที่มีในการศึกษาปีสุดท้ายของหลักสูตรคือรายวิชาที่จะต้องให้นักศึกษาแต่ละคนนั้นออกใบฝึกทำงานที่สถานประกอบการต่าง ๆ ในช่วงระยะเวลาหนึ่ง ซึ่งเป็นสิ่งที่มีความสำคัญและความท้าทายเนื่องจากเป็นการที่นักศึกษาจะได้ทดลองทำงานจริง สถานการณ์จริง สถานที่จริง ในสถานประกอบการที่นักศึกษาได้เลือก

ดังนั้นการเลือกสถานประกอบการสำหรับฝึกงานจึงเป็นเรื่องที่ต้องให้ความสำคัญเป็นอย่างมากเนื่องจากหากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำนั้น ตรงกันกับความสามารถของนักศึกษาก็จะเป็นผลดี เนื่องจากความรู้และทักษะที่ได้จากการทำงานนั้นสามารถนำไปต่อยอดและใช้งานจริงเมื่อจบการศึกษาและเข้าทำงาน แต่หากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำไม่ตรงกับความต้องการหรือทักษะของนักศึกษาอาจทำให้การฝึกงานนั้นล้มเหลว หรืออาจไม่ได้ความรู้และทักษะที่ต้องการได้ และด้วยเทคโนโลยีปัญญาประดิษฐ์ในด้านของการประมวลผลภาษาธรรมชาตินั้นพัฒนาอย่างรวดเร็วมาก ทั้งในแง่ของเทคนิค เครื่องมือ และองค์ความรู้ ทำให้เกิดตัวอย่างการนำข้อมูลมาประมวลผลที่มีประสิทธิภาพมากมายในปัจจุบัน

ด้วยเหตุนี้จึงได้มีการเริ่มโครงการพัฒนาเว็บแอปพลิเคชันที่นักศึกษามีความสนใจในรูปแบบธุรกิจของสถานประกอบการนั้น เพื่อช่วยอำนวยความสะดวกแก่นักศึกษาให้สามารถเข้าถึงข้อมูลของสถานประกอบการสำหรับฝึกงาน โดยการใช้วิธีประมวลผลภาษาธรรมชาติเข้ามาช่วยจัดกลุ่มสถานประกอบการและเสนอรายชื่อสถานประกอบการที่เหมาะสมแก่นักศึกษาผ่านทางเว็บแอปพลิเคชัน

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาและพัฒนาเว็บแอปพลิเคชันเพื่อแนะนำสถานประกอบการตามความสนใจของนักศึกษา

1.2.2 เพื่อศึกษาและวิเคราะห์ความสนใจของนักศึกษาในการหาสถานประกอบการสำหรับฝึกงาน

1.2.3 เพื่อศึกษาประสิทธิภาพการแบ่งกลุ่มของข้อมูลสถานประกอบการด้วยวิธีประมวลผลภาษาธรรมชาติ

1.3 ขอบเขตของโครงการ

1.3.1 ผู้ใช้สามารถค้นหาสถานประกอบการได้ด้วยรายละเอียดของงานหรือรูปแบบธุรกิจที่สนใจ

1.3.2 เว็บแอปพลิเคชันสามารถให้ข้อมูลบริษัทเพื่อการตัดสินใจในการเลือกบริษัท ผู้งานได้

1.3.3 ใช้เทคโนโลยีการประมวลผลภาษาธรรมชาติ เพื่อจัดการคำและแบ่งกลุ่มข้อมูล โดยใช้เทคโนโลยีการจัดกลุ่มเดமีน (K-Means) และหาความคล้ายของข้อมูลด้วยการคำนวณ ค่าความคล้ายคลึง (Cosine similarity)

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้พัฒนาเว็บแอปพลิเคชันสำหรับการค้นหาสถานประกอบการสำหรับผู้งาน

1.4.2 สามารถนำระบบประมวลผลภาษาธรรมชาติมาใช้ในการจัดกลุ่มข้อมูลได้อย่าง แม่นยำ

1.4.3 เป็นช่องทางสำหรับการเลือกและหาข้อมูลของสถานประกอบการสำหรับออก ผู้งานของนักศึกษา

1.5 อุปกรณ์เครื่องมือที่ใช้ในโครงการ

1.5.1 Programming language

1. Python
2. Javascript
3. HTML
4. CSS

1.5.2 Framework

1. Next.js
2. fastAPI

1.5.3 Database

1. MongoDB

1.5.4 Program

1. Microsoft Excel
2. Visual studio code
3. Postman
4. Firefox
5. Figma
6. Notion

1.5.5 Version control

1. Git
2. Github

1.5.6 Python library

1. Pythainlp
2. Matplotlibs
3. Pandas
4. Numpy
5. Sci-kit learn
6. nltk
7. python-dotenv

1.5.7 Javascript library

1. Tailwind CSS
2. cors
3. dotenv
4. sweetalert2
5. headlessui
6. heroicons

1.5.8 Global network

1. Cloudflare

1.5.9 Cloud computing

1. Amazon Web Services
2. Vercel

1.6 นิยามศัพท์ที่ใช้ในโครงงาน

1.6.1 การประมวลผลภาษาธรรมชาติ (Natural language processing:NLP) เป็นเทคนิคแขนงหนึ่งในศาสตร์ของเทคโนโลยีปัญญาประดิษฐ์ ซึ่งเป็นการทำให้คอมพิวเตอร์เข้าใจ ตีความ และสื่อสารภาษาของมนุษย์ได้

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

1.6.3 การตัดคำ (Word segmentation) หมายถึง ตัวย่อที่การเขียนภาษาไทยนั้นไม่มีการแยกคำด้วยการเว้นวรรคหรืออ่านภาษาอังกฤษ หรือ ภาษาอื่นๆ ดังนั้นจึงจำเป็นต้องทำการตัดคำจากประโยคออกมานเป็นคำ ๆ เพื่อให้นำไปประมวลหรือใช้งานต่อได้ด้วยอัลกอริทึมต่าง ๆ

1.6.4 การหาความคล้ายคลึง (Cosine similarity) ระหว่างเวกเตอร์เอ (Vector A) และ เวกเตอร์บี (Vector B) ว่าไปพิศทางเดียวกันหรือไม่โดยการใช้สูตรของกฎสามเหลี่ยมเพื่อหาผลลัพธ์แล้วนำมาเปรียบเทียบกัน

1.6.5 ช่องทางสำหรับการสื่อสารกัน (Application programming interface:API) ระหว่าง เครื่องแม่ข่าย (Server) และ เครื่องลูกข่าย (Client) สร้างขึ้นมาเพื่อเป็นตัวกลางให้โปรแกรม หรือผู้ใช้อินเทอร์เฟซติดต่อสื่อสารเชื่อมต่อแลกเปลี่ยนข้อมูลกัน

1.6.6 การเข้าใช้ระบบคอมพิวเตอร์ (Cloud computing) และทรัพยากรแบบครบวงจร

ผู้ให้บริการต่าง ๆ เช่น Amazon, Google, Microsoft, Huawei โดยสามารถกำหนดรูปแบบของ ชาร์ดแวร์และซอฟต์แวร์ที่ต้องการได้ มีให้บริการทั้งเครื่องแม่ข่าย (Server) ฐานข้อมูล (Database) การทดสอบระบบ (Testing) หรือแอปพลิชันสำเร็จรูปในหลายระบบปฏิบัติการ (Platform)

1.6.7 สถานประกอบการ หมายถึง บริษัทที่ประกอบอาชีพทางด้านศาสตร์ของ คอมพิวเตอร์และเทคโนโลยีต่าง ๆ

1.6.8 ตรงกับความต้องการ หมายถึง การนำความต้องการของผู้ใช้มาเปรียบเทียบกับ ข้อมูลและคำนวณคาดคะมานโดยคลึง

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การศึกษาค้นคว้าเพื่อจัดทำโครงการทางวิทยาการคอมพิวเตอร์ ระบบแนะนำสถานที่ ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลมีน (K-Means) ผู้ศึกษาได้ศึกษาค้นคว้า เอกสารและงานวิจัยที่เกี่ยวข้องตามหัวข้อลำดับต่อไปนี้

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 การจัดการระบบคลาวด์ (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

การประมวลผลภาษาธรรมชาติ (Natural Language Processing:NLP) หรือภาษาของมนุษย์ที่ใช้สื่อสารกัน เป็นเทคนิคหนึ่งในเทคโนโลยีปัญญาประดิษฐ์ ที่จะทำให้คอมพิวเตอร์

สามารถเข้าใจและเรียนรู้ ประมวลผลภาษาของมนุษย์ได้ ในด้านของการวิเคราะห์ภาษาศาสตร์ การตีความจากบุคคล หรือการทั้งการแปลภาษา NLP นั้นจำเป็นต้องใช้ความรู้จากหลาย ๆ ศาสตร์เข้ามา เช่น Mathematics, Linguistics, Psychology เพื่อพัฒนาประสิทธิภาพการทำงาน และความฉลาดของคอมพิวเตอร์

จุดเริ่มต้นของ NLP นั้นมีมาตั้งแต่ประมาณปี 1950–1980 ในยุคนั้นวิธีการที่จะให้คอมพิวเตอร์เข้าใจภาษาของมนุษย์นั้นใช้ “Rule-based” เป็นการใช้ if–else ในโปรแกรมที่ตั้งไว้ตามคำที่กำหนด และในต่อมาประมาณปี 1981–2001 เริ่มมีการใช้ ML หรือ Machine learning ที่ใช้อัลกอริทึมในการประมวลผล เช่น “Decision Tree” เข้ามาช่วยในการประมวลผล และฝึกสอนคอมพิวเตอร์โดยข้อมูลที่เป็น Dataset ทำให้ความแม่นยำเพิ่มขึ้น และในยุคปัจจุบันยุคที่มี

Deep Neural Network เนื่องจากปัจจุบันคอมพิวเตอร์มีความสามารถที่เพิ่มขึ้นและปริมาณข้อมูลนั้น มีมากขึ้นตามทำให้การใช้ Deep Neural Network มาสร้างโมเดลสำหรับการทำ NLP เป็นที่นิยมมากยกตัวอย่างเช่น word embeddings คือการหา semantic กับข้อความนั้น ๆ

กระบวนการทำงานของ NLP นั้น มีประกอบไปด้วยหลายส่วนของการประมวลผล และใช้เปลี่ยนความหมาย ประกอบด้วยดังนี้

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

ปัจจุบัน NLP นั้นอยู่ในหลายรูปแบบรอบตัวถูกนำมาใช้ในหลาย ๆ ด้านทั้งในด้านการทำ Digital marketing, ทางการแพทย์ การแปลงภาษา Chatbot และอื่น ๆ (ตาเยะ, 2022)

2.1.2 ทฤษฎี การตัดคำในภาษาไทย (Word segmentation)

การที่นำประโยชน์มาตัดคำออกเป็นคำ ๆ (Word segmentation) เนื่องจากในบางภาษา เช่นภาษาไทยรูปแบบการเขียนนั้นไม่มีการเว้นวรรคของคำ ต่างจากภาษาอังกฤษที่ใช้การเว้นวรรคในแต่ละคำดังนั้นถ้าจะทำ NLP ที่เป็นภาษาไทยนั้นจำเป็นต้องทำ Word segmentation เพื่อให้ได้ชุดคำที่จะนำไปใช้งานต่อ ในปัจจุบันการทำ Word segmentation นั้นมีเครื่องมือให้ใช้อยู่จำนวนมากยกตัวอย่างเช่น Python library pythainlp, nltk หรือสามารถใช้บริการ web API ของ aiforthai

```
onze@Tinngrits-MacBook-Pro:~/desktop/final_project
..final_project (-zsh)          #1 ..project/report (-zsh)      #2
python report.py
['นอน', 'ตากลม', 'ดู', 'ดาว']
~/d/final_project main ! 1 ?3  4s backup_finalproject * 14:06:09
```

ภาพที่ 1 ตัวอย่างการทำ Word segmentation

จากภาพที่ 1 เป็นการทำ Word segmentation ด้วย Python library pythainlp จากคำว่า “นอนตากลมดูดาว” ได้ผลลัพธ์ออกมาเป็น นอน, ตากลม, ดู, ดาว (L, 2019)

2.1.3 ทฤษฎี การสกัดใจความของข้อความด้วยเทคนิค TF-IDF

การสกัดใจความของข้อความ (Term Frequency – Inverse Document Frequency: TF-IDF) เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภาษาในประโยชน์ เทคนิคนี้มากจาก 2 องค์ประกอบต่อ กันคือ Term Frequency (TF) และ Inverse Document Frequency (IDF) องค์ประกอบแรก Term Frequency (TF) นั้นหมายถึงการที่หาคำที่มีการใช้ซ้ำบ่อยที่สุดในเอกสารนั้น ๆ ซึ่งแสดงไปถึงว่าคำนั้นเป็นคำที่มีความสำคัญมากเอกสารนั้น วิธีคำนวนค่าความถี่ของคำใช้การนำจำนวนครั้งของคำที่ปรากฏในเอกสารมาหารด้วยจำนวนคำทั้งหมด ในเอกสาร เช่น ต้องการหาค่าความถี่ของคำว่า “เว็บไซต์” ในเอกสาร (CHAKRIT, 2019)

$$TF(\text{ของคำคำหนึ่ง}) = \frac{\text{จำนวนของคำนั้นที่มีในเอกสาร}}{\text{จำนวนคำทั้งหมดที่มีในเอกสาร}}$$

ตารางที่ 1 ตัวอย่างการคำนวนค่า Term Frequency ที่จำนวนคำทั้งหมดเท่ากับ 7

คำ	จำนวนคำ	Term Frequency	ผลลัพธ์
เว็บไซต์	5	$5 \div 7$	0.71
หนังสือ	1	$1 \div 7$	0.14
ออนไลน์	2	$2 \div 7$	0.29
ขาย	2	$2 \div 7$	0.29
เข้าชม	1	$1 \div 7$	0.14
มือถือ	4	$4 \div 7$	0.57
และ	3	$3 \div 7$	0.43

จากตัวอย่างจะเห็นได้ว่าคำว่า “เว็บไซต์” ปรากฏบ่อยในเอกสารทำให้มีค่า Term Frequency สูงจึงเรียกได้ว่าเป็นคำสำคัญของเอกสาร แต่การใช้ค่า Term Frequency เพื่อหาใจความสำคัญเพียงอย่างเดียวนั้นยังไม่ดีพอ จึงต้องใช้องค์ประกอบ Inverse Document Frequency (IDF) เพิ่มเติม Inverse Document Frequency (IDF) หมายถึง การคำนวณหน้าหนักของคำโดยการนำคำสำคัญดันหากลาย ๆ เอกสารหากคำนั้นมีค่า Inverse Document Frequency (IDF) ต่ำ แสดงว่าคำนั้นไม่ได้เป็นคำสำคัญของเอกสารทั้งหมด สมการที่ใช้คำนวณหาค่า Inverse Document Frequency (IDF)

$$IDF(\text{ของคำคำหนึ่ง}) = \log\left(\frac{\text{จำนวนเอกสารทั้งหมด}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏ}}\right)$$

ตารางที่ 2 ตัวอย่างการคำนวณค่า Inverse Document Frequency ที่จำนวนเอกสารเท่ากับ 10

คำ	จำนวนเอกสารที่ปรากฏ	Inverse Document Frequency	ผลลัพธ์
เว็บไซต์	5	$\log(10 \div 5)$	0.31
หนังสือ	2	$\log(10 \div 2)$	0.70
ออนไลน์	2	$\log(10 \div 2)$	0.70
ขาย	2	$\log(10 \div 2)$	0.70
เข้าชม	1	$\log(10 \div 1)$	1.00
มือถือ	3	$\log(10 \div 3)$	0.52
และ	2	$\log(10 \div 2)$	0.70

จำนวนคำคำนวณค่า Term Frequency – Inverse Document Frequency (TF-IDF) โดยสมการ

$$TF - IDF = TF \times IDF$$

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

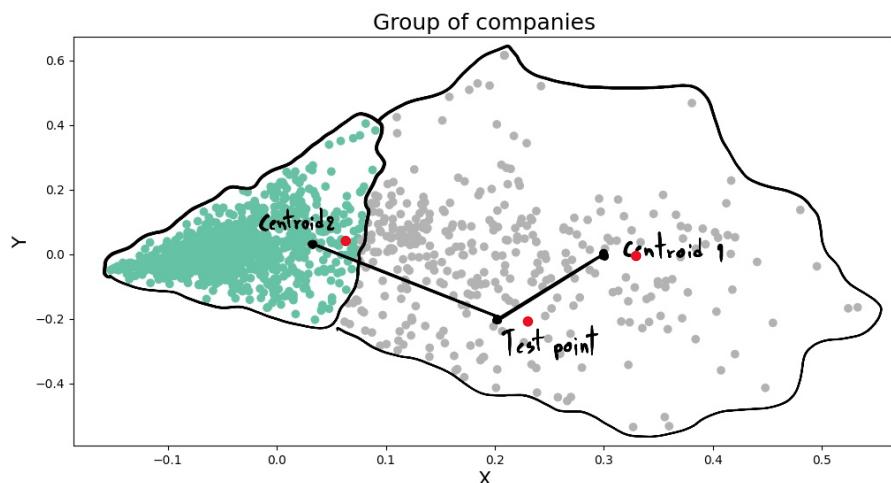
ดังตัวอย่างจะเห็นได้ว่าบางคำที่มีค่า TF-IDF สูงแต่ไม่ได้บ่งบอกถึงลักษณะของข้อความ ในเอกสาร เช่นคำว่า “และ” ซึ่งถือว่าเป็น Stop word ซึ่งเป็นคำที่ไม่สื่อความหมาย โดยปกติแล้ว คำเหล่านี้มักถูกกรองออกก่อนที่จะมีการนำข้อความมาทำการประมวลผลทางภาษา จากตัวอย่างจะเห็นว่าเมื่อคำนวณหาค่า Term Frequency – Inverse Document Frequency (TF-IDF) โดยที่ตัดคำที่ไม่มีความหมายหรือ Stop word ออกแล้วจะเหลือคำว่า “มือถือ” “เว็บไซต์” “ออนไลน์” เรียงลำดับความสำคัญจากมากไปน้อยตามลำดับ (Patipan, 2020)

2.1.4 ทฤษฎี การจัดกลุ่มข้อมูลด้วยอัลกอริทึม K-Means

K-Means เป็นวิธีการหนึ่งใน Data mining อยู่ในกลุ่มของ Unsupervised Learning คือการให้คอมพิวเตอร์เรียนรู้โดยไม่ต้องมีผู้สอน (Chakrit, ว่าด้วย-k-means-และการประยุกต์, 2018) เป็นอัลกอริทึมสำหรับการทำ Clustering Model เป็นการจัดกลุ่มข้อมูลด้วยการกำหนดจำนวนกลุ่มก่อนการทำ Clustering ซึ่งแทนด้วยค่า K จากนั้นคำนวณหาจุดกึ่งกลางของแต่ละกลุ่มเรียกว่าจุด Centroid ตามจำนวนกลุ่มที่กำหนด ไว้เงินระยะห่างด้วยการคำนวณระยะห่าง ด้วยสมการ (Chakrit, ว่าด้วย-k-means-และการประยุกต์, 2018)

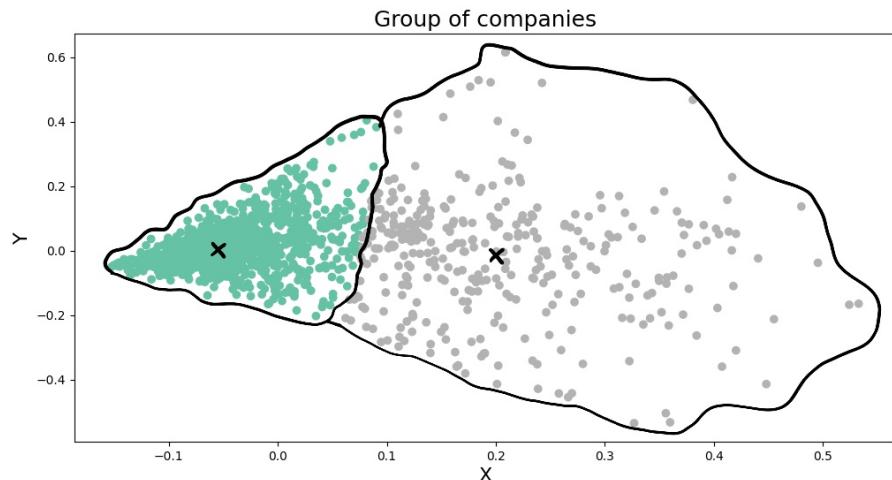
$$Distance = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

ตัวอย่างการคำนวณค่าการกำหนดจุด Centroid เมื่อกำหนด k เท่ากับ 2 ทำการสุมจุดข้อมูลจากข้อมูลทั้งหมด



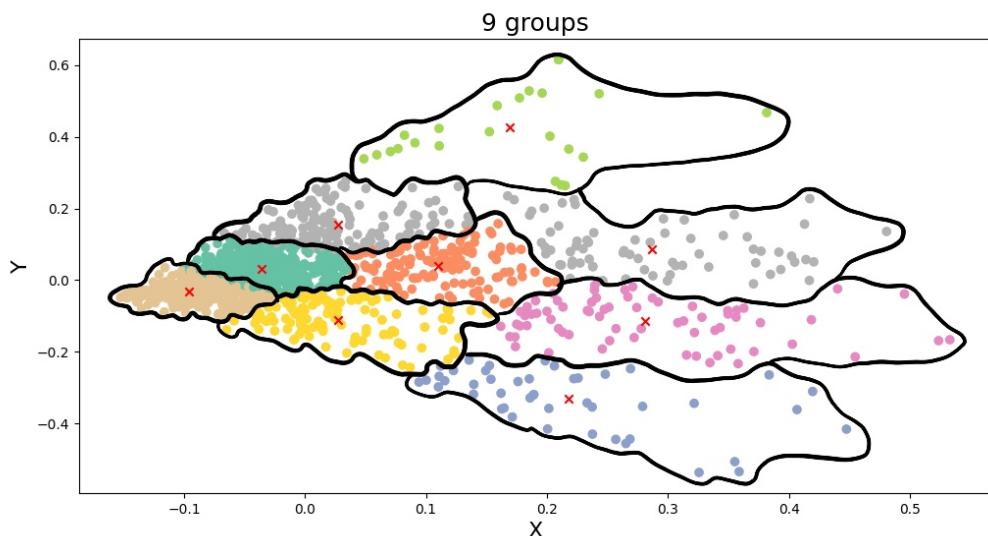
ภาพที่ 2 การกำหนดสุมกำหนดจุด Centroid

ทำการทำซ้ำการกำหนดจุด centroid จนกว่าตำแหน่งของข้อมูลทุกด้วยจะไม่เปลี่ยนแปลง



ภาพที่ 3 จุด Centroid ที่อยู่ตรงกลางและจุดข้อมูลทุกจุดไม่เปลี่ยนแปลง

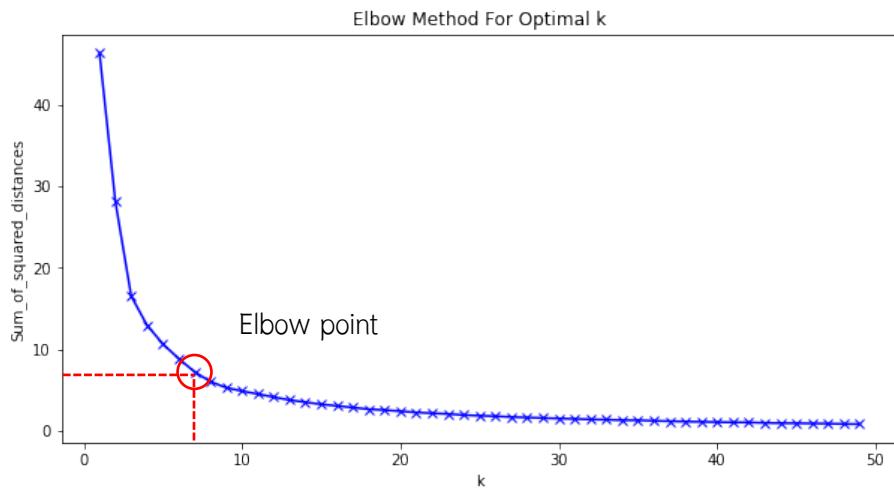
ทำการคำนวณและย้ายจุด Centroid และหาค่าเฉลี่ยของค่าเฉลี่ยไม่มีการเปลี่ยนแปลงจะได้จุดกึ่งกลางของข้อมูลในแต่ละกลุ่ม



ภาพที่ 4 ตัวอย่างการกลุ่มข้อมูลที่มีจุด Centroid เป็นกากบาทสีแดง

2.1.5 ทฤษฎี การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow method

Elbow method เป็นวิธีหนึ่งที่ใช้หาจำนวนของกลุ่มที่เหมาะสมสมด้วยการวัดข้อผิดพลาด (Error measurement) ผลรวมระยะห่างระหว่างข้อมูลกับจุด Centroid เมื่อข้อผิดพลาดน้อยลงความชันของเส้นโค้งจะแบนราบไปตามแกน X จนทำให้เกิดมุมลักษณะเหมือนกับข้อศอกก็จะถือว่าที่อยู่ตรงมุมข้อศอกเป็นจำนวนของกลุ่มข้อมูลที่เหมาะสมสมดังในภาพตัวอย่างภาพที่ 5 จะเห็นได้ว่าจำนวนกลุ่มที่เหมาะสมคือ 6-7 กลุ่ม (Paul, 2021)



ภาพที่ 5 กราฟที่แสดงจำนวนข้อผิดพลาดเพื่อหาจำนวนกลุ่มที่เหมาะสมที่สุด

2.1.6 ทฤษฎี การคำนวณค่าความคล้ายคลึงด้วยเทคนิค Cosine similarity

การวัดความเหมือนของ Vector 2 (Cosine Similarity) ว่าไปในทิศทางเดียวกันหรือไม่ โดยที่เป็นการตัดขนาด หรือ Magnitude ของ Vector ออกไปหากค่าได้จากการนี้

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

อธิบายโดยง่ายคือเป็นการวัดระยะห่างระหว่าง Object A และ Object B ว่ามีความคล้ายกันแค่ไหน ยกตัวอย่างเปรียบเทียบระหว่างคำว่า “ยินดีที่ได้รู้จักรับ” และ “ยินดีที่ได้รู้จัคคะ” ทำการตัดเพื่อหาคำทั้งหมดก่อนคือ “ยินดี” , “ที่” , “ได้” , “รู้จัก” , “ครับ” , “ค่ะ” จากนั้นจะได้ Object A และ B เช่นเป็นชุดข้อมูล ดังนี้ (Supalerk, 2020)

- A. “ยินดีที่ได้รู้จักรับ” = [1 , 1 , 1 , 1 , 1 , 0]
- B. “ยินดีที่ได้รู้จัคคะ” = [1 , 1 , 1 , 1 , 0 , 1]

$$\text{similarity} = \frac{4}{\sqrt{(1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2)} \times \sqrt{(1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 1^2)}}$$

$$\text{similarity} = \frac{4}{\sqrt{6} \times \sqrt{6}} = \frac{4}{6} = 0.67$$

$$\text{similarity} = 0.8$$

2.1.7 ทฤษฎี การจัดการระบบคลาวด์ (Amazon web service)

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

2.1.8 ทฤษฎี API

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

2.1.9 ทฤษฎี Cloudflare

คลาวด์เฟร์ (Cloudflare) คือ Global Network ที่ถูกออกแบบมาเพื่อให้ทุกสิ่งที่เชื่อมอยู่บนอินเทอร์เน็ต มีความปลอดภัย (Security) มีประสิทธิภาพ (Performance) และพร้อมใช้งาน (Availability) ซึ่ง Cloudflare จะทำหน้าที่เป็นตัวกลางระหว่างผู้เข้าใช้งานและ Server ที่เก็บข้อมูล โดยผู้เข้าใช้งานจะมาทั้งในรูปแบบของ Visitor, Crawlers & Bots และ Attackers แต่เมื่อใช้งาน Cloudflare การเข้าถึงทุกรูปแบบจะต้องผ่านระบบของ Cloudflare แทนโดย 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 หรือทำให้ไม่สามารถใช้งานได้โดย Cloudflare จะเข้ามารับการโจมตีดังกล่าวแทนเว็บไซต์

3. Content Delivery Network (CDN) คือ การกระจายเนื้อหาออกไปตาม Server จุดต่าง ๆ หากมี Traffic ระบบจะส่งข้อมูลโดยใช้ Server ที่อยู่ใกล้ที่สุดโดย Cloudflare มี POPs

ในไทยมากถึง 6 POPs และมากกว่า 200 POPs ทั่วโลก ช่วยให้เว็บไซต์สามารถใช้งานได้อย่างรวดเร็ว และ สลับ

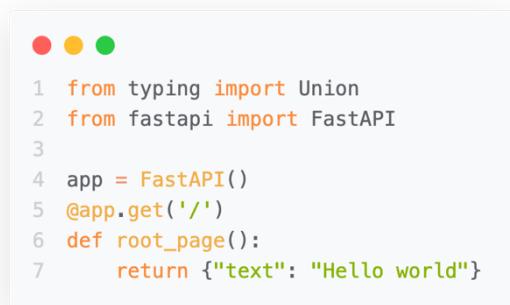
2.1.10 ทฤษฎี Cors

การอนุญาตการแบ่งปันข้อมูลกัน (Cross-Origin Resource Sharing:CORS) เป็นกลไกที่ใช้เพิ่มเติมเพื่อให้เบราว์เซอร์ได้รับสิทธิ์ในการเข้าถึงทรัพยากรที่เลือกจากเซิร์ฟเวอร์บนโดเมนอื่นมาแสดงบนหน้าเว็บเบราว์เซอร์ได้ คอมพิวเตอร์แต่ละเครื่องต้องมี Protocol ที่เหมือนกัน ถึงจะสื่อสารกันได้ เช่น HTTP request เมื่อต้องการขอข้อมูล ข้ามโดเมนหรือ port ที่ต่างกัน และต้องทำตามข้อตกลงการสื่อสาร (Protocol) เพราะปัจจุบันเรา มักจะแยกผัง Front-end และ Back-end ออกจากกันเป็นคนละโดเมน ด้วยเหตุผลเรื่องความปลอดภัยของ Browsers HTTP การอนุญาตให้เข้าถึงแหล่งข้อมูลจะต้องอยู่ในโดเมนเดียวกันเท่านั้น เว้นแต่ว่าแหล่งข้อมูลนั้นจะอนุญาตให้โดเมนของ Browsers สามารถเข้าถึงข้อมูลเหล่านั้นได้ (TAeng Trirong, 2017)

2.1.11 ทฤษฎี Fastapi

เฟรมเวิร์คสำหรับพัฒนาส่วนต่อประสานเครื่องแม่ข่ายกับเครื่องลูกข่ายด้วยภาษาไพธอน (Python) เฟรมเวิร์ค fastAPI ถูกออกแบบมาให้ง่ายต่อการพัฒนา และสามารถที่จะสร้าง API ขึ้นมาได้อย่างรวดเร็ว โดยประสิทธิภาพการทำงานนั้นเร็ว fastAPI นั้นรองรับการทำงานแบบ Asynchronous และมี Uvicorn เป็นตัว run server ข้อดีของการใช้งาน fastAPI คือ (Natakorn, 2021)

1. มีความเร็วของการทำงานเทียบเท่า Node.js และ Go
2. รูปแบบการเขียนฟังก์ชันต่าง ๆ เข้าใจง่ายต่อการศึกษา
3. ง่ายต่อการใช้งานและพัฒนาต่อ

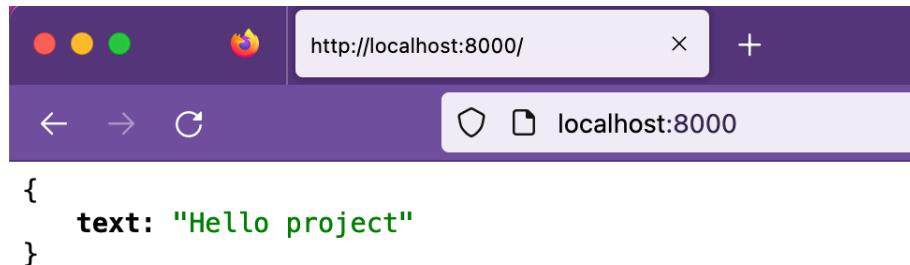


```

● ● ●
1 from typing import Union
2 from fastapi import FastAPI
3
4 app = FastAPI()
5 @app.get('/')
6 def root_page():
7     return {"text": "Hello world"}

```

ภาพที่ 6 ตัวอย่างโค้ดสำหรับการสร้าง Web API ด้วย fastAPI



ภาพที่ 7 ผลลัพธ์แสดงคำว่า Hello project จาก fastAPI

2.1.12 ทฤษฎี Git

Git คือ Version Control ที่ถูกพัฒนาขึ้นมาเพื่อใช้ในกระบวนการพัฒนาซอฟต์แวร์ อย่างเป็นระบบ ให้เข้าใจโดยง่าย คือ ระบบที่ถูกพัฒนาขึ้นมาเพื่อใช้สำหรับการติดตาม ตรวจสอบ การพัฒนา แก้ไข ซอฟต์แวร์ โค้ด ซอฟต์แวร์ไฟล์ต่าง ๆ ในขั้นตอนการพัฒนา ที่สามารถตรวจสอบได้ทุก ตัวอักษร ทุกบรรทัด ทุกไฟล์ ที่มีการแก้ไข ใครเป็นคนแก้ไข และแก้ไข ณ วันที่เท่าไหร่

ระบบการทำงานของ Git ไม่ได้อยู่แค่การตรวจสอบการแก้ไขเท่านั้น ยังสามารถ รวมการแก้ไขทั้งหมดเข้าด้วยกันได้อย่างชัมภูณ์แล้วเรียกว่า CI (Continuous Integration) และในปัจจุบัน Git VCS (Version Control System) มีการควบรวมฟีเจอร์ที่ทำให้ นักพัฒนาทำงานได้สะดวกมากขึ้น สามารถทำงานได้ตั้งแต่ขั้นตอนการพัฒนาไปจนถึงการ Deploy งานขึ้นสู่งานบน Server เรียกว่า CD (Continuous Deployment) รูปแบบการใช้งานของ Git มีด้วยกัน 2 รูปแบบ คือ (codebee, 2020)

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

2.1.13 ทฤษฎี Node.js

Node.js คือสภาพแวดล้อมการทำงานของภาษา JavaScript นอกเว็บเบราว์เซอร์ที่ ทำงานด้วย V8 engine นั่นหมายถึงสามารถใช้ Node.js ในการพัฒนาแอปพลิเคชันแบบ Command line และแพลตฟอร์ม Desktop หรือแม้แต่เว็บเซิร์ฟเวอร์ได้โดยที่ Node.js จะมี APIs ที่ สามารถใช้สำหรับทำงานกับระบบปฏิบัติการ เช่น การรับค่าและการแสดงผล การอ่านเขียน ไฟล์ และการทำงานกับเน็ตเวิร์ก เป็นต้น

Node.js ถูกพัฒนาและทำงานด้วยใช้ Chrome V8 engine สำหรับคอมโพล์ภาษา JavaScript ให้เป็นภาษาเครื่องด้วยการคอมไพล์แบบ Just-in-time (JIT) เพื่อเพิ่มประสิทธิภาพ การทำงานของภาษา JavaScript จากที่เต็มมันเป็นภาษาที่มีการทำงานแบบ Interpreted Node.js เป็นโปรแกรมที่สามารถใช้ได้ทั้งบน Windows, Linux และ Mac OS X นั่นหมายความว่า สามารถเขียนโปรแกรมในภาษา JavaScript และนำไปรันได้ทุกระบบปฏิบัติการที่สนับสนุนโดย Node.js นี้ เป็นแนวคิดของการเขียนครั้งเดียวแต่ทำงานได้ทุกที่ (Write once, run anywhere) ข้อดีอีกอย่างหนึ่งในการใช้ภาษา JavaScript ของ Node.js คือทำให้การพัฒนาเว็บไซต์ทำได้ง่าย ขึ้น สำหรับนักพัฒนา เนื่องจากสามารถใช้ภาษา JavaScript สำหรับทั้ง Front-end และ

Back-end ได้โดยไม่ต้องศึกษาภาษาเฉพาะในแต่ละด้าน ตัวอย่างของการพัฒนาเว็บไซต์ในรูปแบบนี้ 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 (หัด Python สำหรับคนเป็น Excel : ตอนที่ 8 – การสร้างกราฟด้วย Matplotlib)

2.1.15 ทฤษฎี Mongodb

MongoDB เป็น open-source document database โดยเป็นฐานข้อมูลแบบ NoSQL คือไม่มีความสัมพันธ์ (No relationship) ของตารางแบบ SQL ทั่วไป แต่จะเก็บข้อมูลเป็นแบบ JSON

(JavaScript Object Notation) แทนการบันทึกข้อมูลทุก ๆ Record ใน MongoDB และเรียกว่า Document ซึ่งจะเก็บค่าเป็น key และ value จะเห็นว่าคือ JSON (Chai, 2015) ตัวอย่างเช่น

```

1  {
2    "userId": 1,
3    "id": 1,
4    "title": "delectus aut autem",
5    "completed": false
6  },
7  {
8    "userId": 1,
9    "id": 2,
10   "title": "quis ut nam facilis et officia qui",
11   "completed": false
12 },

```

ภาพที่ 8 ตัวอย่างข้อมูลแบบ JSON

โดยหลัก ๆ เหามากกับองค์กรที่อยากระบบฐานข้อมูลโดยย่างรวดเร็ว ยืดหยุ่นมากกับการทำ Big Data และอื่น ๆ ดังนี้ (PLC, 2022)

1. ตัว MongoDB สามารถที่จะสร้างเป็น Cluster เพื่อที่จะตอบสนองของคำว่า High Availability (HA) ได้ ซึ่งสามารถเลือก Region ที่อยากระบบ Cloud Provider นั้น ๆ ได้
2. ความรวดเร็วในการเข้าถึงข้อมูล เพราะว่า Database ไม่มี Schema ซึ่งจะต่างกับ SQL โดยแบบนั้นจะอิงจากฐานข้อมูลที่มาจากการ Table
3. สามารถทำ Auto Scale ได้ไม่ว่าจะมีการใช้งานมากน้อยแค่ไหนก็สามารถปรับใช้กับ Environment นั้น ๆ ได้

4. รองรับการทำ Multiple Cloud Provider ซึ่งข้อดีข้อนี้สามารถทำให้ Database ที่ใช้นั้น มี High Availability มากรีนโดยเราไม่จำเป็นที่จะต้องยึดติดกับ Cloud Provider เจ้าใดเจ้าหนึ่ง

2.1.16 ทฤษฎี Next.js

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

1. SSR (server-side rendering)
2. Hot rendering
3. Static HTML file exportable
4. Project Structure
5. Routing
6. 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 ส่วนใหญ่ซึ่งทำให้ SEO ค้นพบ source เพื่อให้ได้ข้อมูลและจัดหมวดหมู่ได้ง่ายกว่า React ที่เป็น JavaScript มากกว่า ทำให้ Next.js เป็นที่นิยมในหลาย ๆ บริษัท นอกจากนี้ ข้อดีก็คือ render ได้เร็วกว่า React เพราะ Next.js มีลิ้งที่เรียกว่า get static path ซึ่งการสร้าง path แบบ static แบบเว็บไซต์ html โดยไม่ต้องทำการเชื่อมต่อกับ backend เพื่อให้ได้ data ยิ่งไปกว่านั้น Next.js สามารถรวมเข้ากับ backend ได้ง่าย ๆ เพราะ Next.js มีลิ้งที่เรียกว่า 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 ไลบรารี (Library) ที่รู้จัก และเป็นที่นิยมใช้ในการคำนวณ เช่น ใช้คำนวณ Matrix หรือ คำนวณกับ Array ในงาน Data Science, Data analytics และในการเรียนรู้ของเครื่องจักร (Machine Learning) หรือ ดีพ เลิร์นนิ่ง (Deep Learning) ก็ต้องใช้ Numpy ด้วยที่ Numpy เป็น Library พื้นฐานที่ใช้คำนวณทางคณิตศาสตร์ด้วยภาษา Python สามารถคำนวณหรือดำเนินการทางตรรกيةใน Array หลายมิติหรือ Matrix ได้อย่างรวดเร็ว เพราะ Library เชียนด้วยภาษา C ที่ Compile ไว้แล้ว (mindphp, Numpy คืออะไร)

Numpy นั้นได้แรงบันดาลใจมาจาก MATLAB ดังนั้นผู้ที่มีประสบการณ์ด้าน MATLAB อุปถัมภ์จะทำความเข้าใจ Numpy ได้ไม่ยาก โดยหลักการคือการนิยามตัวแปร Array หลายมิติที่คุณเคยในคณิตศาสตร์ อาทิ เช่น เวกเตอร์ (1 มิติ) เมตริก (2 มิติ) เтенเซอร์ (3 มิติขึ้นไป) เป็นต้น และ operations ของมัน ในการทำความเข้าใจ Numpy นั้นควรมีความรู้พื้นฐาน Linear algebra ในเรื่องของ vector / matrix ในระดับหนึ่ง (JUNG, 2019)

2.1.18 ทฤษฎี Pandas

Pandas คือ หนึ่งใน Library สำคัญของภาษา Python เวิร์กพัฒนาโดย Wes McKinney นักพัฒนาซอฟต์แวร์ชาวอเมริกัน ปัจจุบัน Pandas เป็น open source ให้ทุกคนสามารถใช้ได้แบบฟรี Pandas มาจากคำว่าชุดข้อมูลหลายมิติ (Panel Data) มีจุดเด่นด้านการวิเคราะห์ข้อมูล (Data Analysis) และการทำความสะอาด (Data Cleaning) ซึ่งเป็น Process ที่สำคัญมากในการทำงานกับข้อมูล Pandas มีความสามารถในการจัดการ และวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพตั้งแต่ข้อมูลขนาดเล็กไปจนถึงข้อมูลขนาดใหญ่ทำให้ Pandas ตอบโจทย์งานในยุคที่ข้อมูลมีขนาดใหญ่มากขึ้นเรื่อยๆ ได้ไม่เป็นหน้าติดขัดเมื่อกับ Spreadsheets อื่นๆ เช่น Excel หรือ Google Sheets ซึ่งจะทำงานได้ช้าลงหากข้อมูลมีขนาดใหญ่ขึ้น ขั้นตอนการเตรียมข้อมูลนั้นมีความสำคัญมาก และ Data Scientist อาจจะใช้เวลาส่วนใหญ่หมดไปกับขั้นตอนนี้ เพราะหากข้อมูลที่เตรียมได้ไม่มีประสิทธิภาพการนำ Insights ไปใช้งานหรือนำข้อมูลไปสร้างโมเดลย่อมทำให้ได้ข้อมูลที่ไม่精准เสียอีก

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

2.1.19 ทฤษฎี Pythainlp

โมดูล pythainlp เป็นเหมือนกับ library ที่รวมคำสlangที่เกี่ยวกับภาษาไทยใน Python ซึ่งก็เป็นตัวช่วยให้การทำงานเกี่ยวกับตัวของภาษาไทย มีประสิทธิภาพและสะดวกมากขึ้น ในการทำงานของ pythainlp ก็จะมีการทำงาน เช่น การตัดคำ การแปลไทยเป็นอังกฤษ และการเข้าถึงรหัส Soundex และยังมีการทำงานที่เกี่ยวกับตัวของภาษาไทย มีการแสดงผล เกี่ยวกับเช็ตของภาษาไทยทั้งหมด ยังมีในส่วนของการแยกตัวอักษร เป็นส่วนของพยัญชนะ สระ วรรณยุกต์ เป็นต้น ยังมีในส่วนของเลขไทย มีการเช็คว่าเป็นคำภาษาไทยหรือไม่ มีการรับตัวอักษร ว่าเป็นภาษาไทยกี่เปอร์เซ็นต์ และยังมีส่วนของการแสดงคำอ่านที่เป็นพากเกราและยังมีการจัดเรียงคำใน List ให้เรียงกันเป็นลำดับได้ เป็นต้น (mindphp, 2022) นอกจากนี้ยังใช้สำหรับประมวลผลข้อความ และการวิเคราะห์ทางภาษาคล้ายกับ NLTK แต่ใช้กับภาษาไทยโดยเฉพาะ มีฟังก์ชันการทำงานที่หลากหลาย เช่น Character Set อักษรไทย คำไทย, เรียงคำภาษาไทย, Stop Words ภาษาไทย, ตัดคำภาษาไทย, วิเคราะห์ชนิดของคำทางไวยากรณ์, ตรวจตัวสะกด แก้คำผิด และอีกมากmany (Surapong, 2020)

2.1.20 ทฤษฎี Scikit-learn

Scikit-learn หรือ sklearn นำเสนอแบบจำลองทางสถิติและการเรียนรู้ของเครื่องที่หลากหลายแตกต่างจากโมดูลส่วนใหญ่ sklearn ได้รับการพัฒนาใน Python มากกว่า C แม้จะได้รับการพัฒนาใน Python ก็ตาม ประสิทธิภาพของ sklearn นั้นสูงกำหนดให้ใช้ NumPy สำหรับการดำเนินการพื้นฐานเชิงเส้นและ多元 regression ที่มีประสิทธิภาพสูง (เจร์, 2021)

Scikit-Learn ถูกสร้างขึ้นโดยเป็นส่วนหนึ่งของโครงการ Summer of Code ของ Google และทำให้ชีวิตของนักวิทยาศาสตร์ข้อมูลที่มี Python เป็นคุณย์กลางนับล้านทั่วโลกง่ายขึ้น ส่วนนี้ของชีรีส์มุ่งเน้นไปที่การนำเสนอไลบรารีและมุ่งเน้นไปที่องค์ประกอบเดียว นั่นคือการแปลงชุดข้อมูล ซึ่งเป็นขั้นตอนสำคัญและสำคัญที่ต้องทำก่อนพัฒนาแบบจำลองการทำงานอย่าง Scikit-learn เป็นแพ็คเกจ Python โอเพ่นซอร์สพร้อมการวิเคราะห์ข้อมูลที่ซับซ้อน และมาพร้อมกับอัลกอริทึมในตัวมากมายที่จะช่วยให้คุณได้รับประโยชน์สูงสุดจากการวิทยาศาสตร์ข้อมูลของคุณไลบรารี Scikit-learn มีให้เลือกใช้ได้รายการดังนี้ 1. Classification 2.Regression 3.Clustering 4.Dimensionality reduction 5. Model selection 6. Preprocessing

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 โดยได้อย่างง่าย Vercel Inc. เติมชื่อ Zeit เป็นแพลตฟอร์มคลาวด์ของเมริกาในฐานะบริษัทผู้ให้บริการบริษัทรักษาการของการพัฒนาเว็บไซต์ Next.js สถาปัตยกรรมของ Vercel สร้างขึ้นจาก Jamstack และการจัดการการปรับใช้ผ่านที่เก็บ Git Vercel เป็นสมาชิกของ MACH Alliance (Huangsri, 2021)

2.2 งานวิจัยที่เกี่ยวข้อง

จักรินทร์ สันติรัตนภักดี และศุภกฤษฐ์ นิวัฒนาภูล ศึกษาเรื่อง การออกแบบและพัฒนากระบวนการจำแนกชั้นของเรียนรถโดยสารสาธารณะเพื่อติดแท็กป้ายหากการให้บริการ องค์การขนส่งมวลชนกรุงเทพ (ขสมก.) มีช่องทางในการรับเรียนรถโดยสารสาธารณะผ่านเว็บบอร์ด ที่ผู้ใช้งานสามารถแสดงความคิดเห็นได้อย่างอิสระ ผู้วิจัยจึงออกแบบและพัฒนากระบวนการจำแนกชั้นของเรียนรถโดยสารสาธารณะ จากชั้นของเรียนผ่านเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพด้วยกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม และวัดเลือกคำศัพท์ด้วยการวิเคราะห์หนังสือของคำ มาสร้างเป็นคลังคำศัพท์ แบ่งเป็น 4 คลาส โดยแก้ คลาสการขับซึ่คลาสผู้ขับซึ่และพนักงานผู้ให้บริการ คลาสขายนพาหนะและอุปกรณ์ให้บริการ และคลาสวีลา และการเดินรถโดยใช้มอดูลการตัดคำภาษาไทย (Thai Word Segmentation) ด้วยชั้นความทั่วไปซึ่งอยู่ในรูปแบบประโยชน์ค่าแบ่งออกเป็นคำหรือคุณลักษณะ (Term/Feature) เพื่อแยกส่วนของข้อความออกจากกันก่อนนำไปประมวลผลในขั้นต่อไป แบ่งตามกระบวนการทำงานออกเป็น 3 กลุ่ม โดยแก้ 1) การตัดคำโดยใช้กฎ (Rule-Based Approach) 2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) 3) ภาตต์ดอยไซซ์คลังคำศัพท์ (Corpus-Based Approach) จากการทดลองพบว่าอัลกอริทึมโครงข่ายประสาทเทียบแบบเพอร์เซ็ปตรอนหลายชั้น มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวมสูงที่สุด (จักรินทร์ สันติรัตนภักดี, 2021)

วุฒิชัย วิเชียรไชย ศึกษาเรื่อง การเปรียบเทียบวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างการเขียนกับโครงสร้างพยานค์ งานวิจัยนี้นำเสนอการแบ่งแยกคำภาษาไทยโดยเทียบกับโครง

สร้างการเขียนของภาษาไทยและอักษรที่มีการแบ่งแยกคำภาษาไทยโดยโครงสร้างพยานค์เพื่อศึกษาและเปรียบเทียบวิธีการประมวลผลของการแบ่งแยกคำภาษาไทย และประสิทธิภาพความถูกต้องของอักษรที่มี โดยสามารถแบ่งงานวิจัยในการแบ่งแยกคำภาษาไทยได้เป็นดังนี้ คือวิธีการใช้กฎ (Rule base approach) วิธีการใช้อักษรที่มี (Algorithm ap-proach) วิธีการใช้ พจนานุกรม (Dictionary base approach) และวิธีการใช้คลังข้อความ (Corpus based approach) ผู้วิจัยจึงได้เสนอวิธีการแบ่งแยกคำภาษาไทย โดยใช้โครงสร้างการเขียนภาษาไทยเพื่อแก้ไขลดพื้นที่ในการจัดเก็บคำศัพท์ในพจนานุกรม และวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างพยานค์เพื่อลดการลับเลื่อนพื้นที่ในการจัดเก็บพจนานุกรม ยกตัวอย่างการแบ่งแยกคำและพยานค์ของคำว่า “ประเทศไทย” จะสามารถแบ่งแยกคำได้เป็น “ประเทศไทย” และแบ่งพยานค์ได้เป็น “ประเทศไทย” จากผลลัพธ์ในการแบ่งแยกคำนั้นยังขาดความถูกต้องในการแบ่งแยกคำซึ่งสามารถพัฒนาแนวคิดในการศึกษาและสร้างกฎเพื่อแบ่งแยกคำให้ถูกต้องมากยิ่งขึ้น (วุฒิชัย, 2013)

ปราณี พึงวิชา 璇ันท์ ทับเที่ยง และธัญญา สัตยาภิชาณ (2019) ศึกษาการแบ่งกลุ่มพุทธิกรรมของผู้บริโภคที่ชื่อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์ ผู้วิจัยเก็บรวบรวมข้อมูลจากกลุ่มตัวอย่างจำนวน 400 คน ทำการวิเคราะห์แบ่งกลุ่ม ผู้บริโภคด้วยวิธีการจัดกลุ่มด้วยเคลื่อน (K-Means Clustering) เป็น 2 กลุ่ม ซึ่งมีลักษณะเฉพาะในแต่ละกลุ่ม จากการวิเคราะห์ความแตกต่างระหว่างกลุ่มของทัศนคติต้านพุทธิกรรมการซึ่งและด้านส่วนประสบทางการตลาดที่มีผลต่อการตัดสินใจชื่อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์เมื่อออยู่ต่างกลุ่มกัน โดยการวิเคราะห์ความแปรปรวนทางเดียว (One-way ANOVA) พบว่า ด้านพุทธิกรรมการซึ่งหั้ง 2 กลุ่ม มีความถี่ในการซื้อต่างๆ เช่น ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับ .05 ส่วนตัวแปรอื่น ๆ นั้นมีความแตกต่างกัน จากการวิเคราะห์แบ่งกลุ่มผู้บริโภคที่ชื่อเครื่องประดับผ่านเครือข่ายสังคมออนไลน์ด้วยวิธี K-mean clustering สามารถจำแนกเป็น 2 กลุ่มโดยแต่ละกลุ่มมีลักษณะเฉพาะดังนี้ กลุ่มที่1 : กลุ่มกระเปาหนักจ่ายได้ถ้าชอบ ไม่ค่อยชอบ ออกสื่อ ลักษณะด้านประชากรศาสตร์ โดยส่วนใหญ่เป็นคน Generation X เพศหญิงมากกว่าเพศชาย มีระดับการศึกษาสูงกว่าปริญญาตรี กลุ่มที่2 : กลุ่มวัยสะอ่อน ชอบออกสื่อ ชื่อน้อย แต่บ่อยครั้ง ลักษณะด้านประชากรศาสตร์ โดยส่วนใหญ่เป็นคน Generation Y เพศหญิงมากกว่าเพศชาย ส่วนใหญ่มีระดับการศึกษาสูงแต่จะน้อยกว่ากลุ่ม 1 โดยมีระดับปริญญาตรีมากที่สุดโดยส่วนใหญ่เป็นพนักงานบริษัทเอกชน (ปราณี พึงวิชา, 2019)

ธงชัย คล้ายคลึง วุฒิชัย สง่างาม กิตติวงศ์ สุธรรมโน และพันธ์พงศ์อภิชาตกุลศ (2019) ศึกษาเรื่อง เทคนิคการคัดเลือกกลุ่ม ให้ลดรายอาควรสำหรับองรับแผนการติดตั้งระบบผลิตไฟฟ้าพลังงานแสงอาทิตย์ บนหลังคาเพื่อเพิ่มค่าครองน้ำประลิทธิภาพการใช้ไฟฟ้าพลังงานไฟฟ้า บทความนี้ต้องการนำเสนอเทคนิควิธีการคัดเลือกกลุ่มโดยลดในแต่ละอาคารที่มีความเหมาะสมสำหรับติดตั้งระบบผลิตไฟฟ้าจากพลังงานแสงอาทิตย์กรณีที่ติดตั้งบนหลังคาของอาคารในศูนย์กลางมหาวิทยาลัยเทคโนโลยีราชมงคลธัญญาชลี สำนักครรราชสีมา ในวิธีการของ K-Means Clustering

เริ่มต้นด้วยการจัดแบ่งข้อมูลออกเป็น K กลุ่ม กำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด ขั้นตอนต่อไปคือการสร้างกลุ่มข้อมูลและความสัมพันธ์กับจุดศูนย์กลางที่ใกล้มากที่สุด จากผล การวิเคราะห์การใช้พลังงานไฟฟ้าทั้งหมด 34 อาคารด้วยวิธีการ K-Mean Clustering ทำให้สามารถแยกแยะจัดกลุ่มให้ครายอาคารได้อย่างมีประสิทธิภาพซึ่งแบ่งได้ 3 กลุ่มโดยกลุ่มที่ 3 จำนวน 19 อาคารนั้นเป็นกลุ่มอาคารที่มีความเหมาะสมทั้งด้านพฤติกรรมการใช้พลังงานไฟฟ้า และมีพื้นที่รองรับการติดตั้งระบบผลิตไฟฟ้าจากพลังงานแสงอาทิตย์บนหลังคาได้ (Thongchai Klayklueng, 2019)

รายงาน ประดิษฐ์กุล ปราลี มณีรัตน์ และ นิเวศ จิระวิชิตชัย (2021) ศึกษาเรื่อง ระบบแนะนำรายนต์ให้กับลูกค้าโดยการวิเคราะห์จากการอ้างอิงถึงพฤติกรรมของผู้ใช้ (Collaborative Filtering) กรณีศึกษาบริษัท โตโยต้า บัสส จำกัด ผู้วิจัยได้พัฒนาระบบแนะนำรายนต์ให้กับลูกค้าโดยการวิเคราะห์จากการอ้างอิงถึงพฤติกรรมของผู้ใช้ เพื่อช่วยให้ลูกค้าได้รับการแนะนำรุ่นรถยนต์ที่เหมาะสม ตรงตามความต้องการของลูกค้า ผู้วิจัยใช้อัลกอริทึมการหาความคล้ายคลึงกันของผู้ใช้ โดยวิเคราะห์จากลูกค้าที่มีพฤติกรรมใกล้เคียงกันด้วยสมการการหาความคล้ายๆ กันของโคไซน์ ซึ่งเป็นฟังก์ชันในภาษา Python ในการพัฒนาระบบแนะนำรายนต์ให้กับลูกค้า ด้วยสมการความคล้ายๆ กันของโคไซน์ (cosine similarity) จากการทดลองเมื่อนำข้อมูลมาจัดลำดับคะแนนความชอบของผู้ใช้แต่ละคน เพื่อเป็นการเพิ่มความเร็วให้อัลกอริทึมของวิธีการกรองแบบร่วมมือ อีกทั้งระบบจะนำความคล้ายคลึงๆ กันของผู้ใช้ในระบบกับผู้ใช้เป้าหมาย มาทดสอบความแม่นยำของระบบด้วยค่าเฉลี่ยความคลาดเคลื่อนสมบูรณ์ พบว่ามีค่าเท่ากับ 0.97 เมื่อกำหนดค่า k ให้เท่ากับ 5 สรุปได้ว่าระบบมีประสิทธิภาพในการแนะนำรุ่นรถยนต์ที่รวดเร็วและแม่นยำอยู่ในระดับที่ดี (Warakorn Pradiskul, 2021)

บทที่ 3

ขั้นตอนการดำเนินงาน

การจัดทำโครงการทางวิทยาการคอมพิวเตอร์ ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) ผู้จัดทำได้กำหนดขั้นตอนการดำเนินงานดังนี้

- 3.1 การเตรียมและวิเคราะห์ข้อมูล
- 3.2 การทำงานของระบบ
- 3.3 การวิเคราะห์และออกแบบระบบ
- 3.4 การออกแบบฐานข้อมูล
- 3.5 การออกแบบหน้าจอ
- 3.6 การใช้งานระบบ

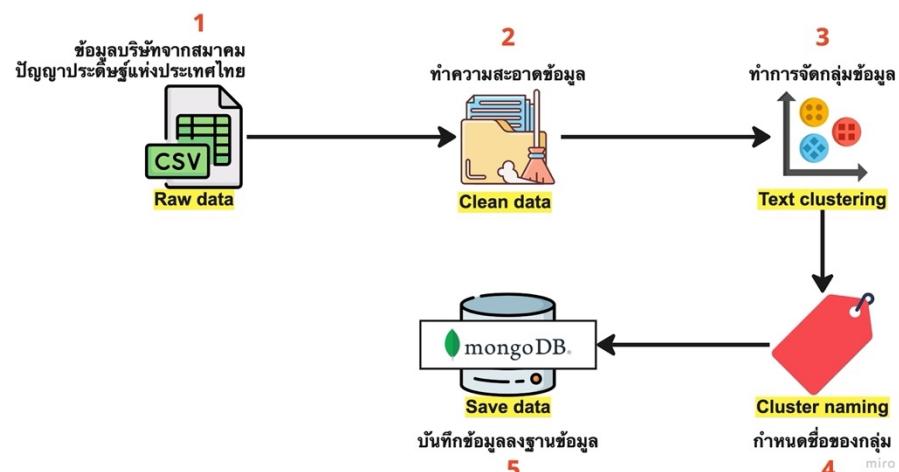
3.1 การเตรียมและวิเคราะห์ข้อมูล

ตารางที่ 4 การวิเคราะห์ข้อมูลประเภทของสถานประกอบการ

ข้อมูล	จำนวน	หน่วย
จำนวนข้อมูลสถานประกอบการทั้งหมด	1,643	รายการ
artificial Intelligence	31	รายการ
internet of things	105	รายการ
chatbot	46	รายการ
big data	84	รายการ
machine learning	61	รายการ
data science	102	รายการ
face recognition	20	รายการ
face detection	5	รายการ
optical character recognition	3	รายการ
data mining	9	รายการ
natural language processing	11	รายการ
data visualization	1	รายการ
image processing	20	รายการ
robotics	45	รายการ
computer vision	14	รายการ
speech recognition	2	รายการ
automatic license plate recognition	1	รายการ

ตารางที่ 4 (ต่อ)

ข้อมูล	จำนวน	หน่วย
e-kyc	1	รายการ
biometrics	9	รายการ
biometric authentication	3	รายการ
sentiment analysis	3	รายการ
text mining	2	รายการ
embedded system	1	รายการ
machine translation	1	รายการ
ไม่มีประเภท	1,318	รายการ
จำนวนคำทั้งหมด	9.856	คำ



ภาพที่ 9 การเตรียมและวิเคราะห์ข้อมูล

จากการที่ 9 แสดงการเตรียมและวิเคราะห์ข้อมูลได้ดังนี้

- ข้อมูลบริษัทจากสมาคมปัญญาประดิษฐ์แห่งประเทศไทยโดยเป็นไฟล์ข้อมูลแบบ CSV (Comma-Separated Value)
- ทำความสะอาดข้อมูลบข้อมูลที่ไม่มีความหมายในตัว ลบตัวเลขที่ไม่จำเป็น แก้คำพิมพ์ผิดและอักษรพิเศษต่าง ๆ
- จัดกลุ่มข้อมูลด้วยวิธี K-Means clustering
- ตั้งชื่อของกลุ่มข้อมูลโดยอ้างอิงจากงานด้านไอทีจากเว็บไซต์ th.jobsdb.com
- นำผลการจัดกลุ่มจัดเก็บลงฐานข้อมูล MongoDB

3.1.1 การนำเข้าข้อมูลไฟล์ .csv เพื่อทำการจัดกลุ่มข้อมูล

1. ดาวน์โหลดโปรเจคจาก https://github.com/slapexs/final_project
 2. นำเข้าไฟล์ข้อมูลลงในโฟลเดอร์ data_csv

3.1.2 การทำ Word segmentation

```
1 import pandas as pd
2 from pythainlp.corpus import thai_stopwords
3 from nltk.corpus import stopwords
4 from pythainlp.tokenize import word_tokenize
5 from sklearn.feature_extraction.text import TfidfVectorizer
6 import string
7 import numpy as np
```

ภาพที่ 10 การเรียกใช้ไลบรารี (Library) สำหรับคำนวณค่า TF-IDF

จากภาพที่ 10 แสดงการเรียกใช้ฟังก์ชันจากไลบรารี (Library) ที่ใช้ในการคำนวณค่า TF-IDF ประกอบไปด้วย

1. Pandas ใช้ในการอ่านข้อมูลในไฟล์
 2. thai_stopword เป็นรายการคำที่ไม่สื่อความหมายในภาษาไทย
 3. stopword เป็นรายการคำที่ไม่สื่อความหมายในภาษาอังกฤษ
 4. word_tokenize เป็นฟังก์ชันที่ใช้ในการตัดคำแยกเป็นคำ ๆ จากประโยค
 5. TfidfVectorizer เป็นฟังก์ชันสำหรับคำนวณหาค่า TF-IDF จากประโยคที่ตัดคำแล้ว
 6. String เป็นคลาสของภาษา Python ที่ใช้แสดงข้อมูลตัวอักษรต่าง ๆ
 7. Numpy ใช้ในการสร้างอาเรย์สำหรับการใช้งานในการอ่านข้อมูล

ภาพที่ 11 การอ่านข้อมูลจากไฟล์และกำหนดตัวกรองการตัดคำ

จากภาพที่ 11 แสดงการอ่านข้อมูลจากไฟล์ และการกำหนดตัวกรอกในการตัดคำทั้งคำที่ไม่สื่อความหมายในภาษาไทย ภาษาอังกฤษ ตัวเลขไทย และอักษรพิเศษ พร้อมทั้งประกาศตัวแปรเพื่อแก้ไขการจัดการตัดคำ

```

● ● ●
1 def clean_string(detail:list) -> list:
2     temp_clean = []
3     for i in detail:
4         if i not in string.punctuation and i not in string.digits and i not in spx_char and i not in th_number:
5             temp_clean.append(i.lower())
6     return ''.join(temp_clean)
7

```

ภาพที่ 12 พังก์ชันสำหรับใช้ลบตัวเลข และอักขระพิเศษ

จากภาพที่ 12 แสดงการลบตัวเลข และอักขระพิเศษออกจากประโภคที่รับเข้ามาและทำการเชื่อมประโยคและคืนค่ากลับออกໄປ

```

● ● ●
1 def clean_stopword(token:list) -> list:
2     temp = []
3     for i in token:
4         if i not in th_stopword and i not in eng_stopword and i not in th_number:
5             temp.append(i)
6     return temp

```

ภาพที่ 13 พังก์ชันสำหรับใช้ลบคำที่ไม่สื่อความหมายและตัวเลขไทย

จากภาพที่ 13 แสดงการลบคำที่ไม่สื่อความหมายในภาษาไทย ภาษาอังกฤษ และตัวเลขไทยออกໄປจากข้อมูลที่รับเข้ามาและคืนค่ากลับออกໄປ

```

● ● ●
1 for i in range(len(df)):
2     sample = clean_string(str(df.iloc[i]['detail']).lower())
3     text_cleaned = clean_stopword(word_tokenize(sample, None, 'newmm', False))
4     list_company_detail.append(text_cleaned)

```

ภาพที่ 14 การวนซ้ำข้อมูลเพื่อตัดคำและทำความสะอาดข้อมูล

จากภาพที่ 14 แสดงการวนซ้ำการส่งข้อมูลที่อ่านจากไฟล์เพื่อนำไปลบคำที่ไม่สื่อความหมาย ตัวเลข และอักขระพิเศษออกจากข้อมูลตัวแรกถึงตัวสุดท้าย และนำผลลัพธ์ไปเก็บไว้ในตัวแปร

```

1 def fake_tokenize(word):
2     return word
3
4 vectorizer = TfidfVectorizer(
5     analyzer='word',
6     tokenizer=fake_tokenize,
7     preprocessor=fake_tokenize,
8     token_pattern=None,
9     lowercase=True,
10 )
11 tfidf_vector = vectorizer.fit_transform(list_company_detail)
12 tfidf_array = np.array(tfidf_vector.todense())
13 df_tfidf = pd.DataFrame(tfidf_array, columns=vectorizer.get_feature_names_out())
14 df_tfidf = df_tfidf.drop(df_tfidf.columns[[k for k in range(-15, 0, 1)]], axis = 1)
15 print(df_tfidf)

```

ภาพที่ 15 การเทรนและการทดสอบโมเดลการคำนวณค่า TF-IDF

จากภาพที่ 15 แสดงการสร้างเวกเตอร์ของการคำนวณค่า TF-IDF การเทรนข้อมูล และการทดสอบการประมวลผลจากโมเดลที่เทรนประกอบไปด้วย

1. vectorizer เป็นการเทรนโมเดลสำหรับการคำนวณค่า TF-IDF
2. tfidf_vector เป็นการทดสอบและสร้างเวกเตอร์ของคำในแต่ละประโยค
3. tfidf_array เป็นการนำเวกเตอร์มาเปลี่ยนให้อยู่ในรูปแบบของอาเรย์
4. df_tfidf เป็นการนำข้อมูลในอาเรย์มาสร้างเป็นตารางข้อมูล

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	628	629	630	631	632	633	634	635	636	637	638	639	640	641	642	643	644	645	646	647	648	649	650	651	652	653	654	655	656	657	658	659	660	661	662	663	664	665	666	667	668	669	670	671	672	673	674	675	676	677	678	679	680	681	682	683	684	685	686	687	688	689	690	691	692	693	694	695	696	697	698	699	700	701	702	703	704	705	706	707	708	709	710	711	712	713	714	715	716	717	718	719	720	721	722	723	724	725	726	727	728	729	730	731	732	733	734	735	736	737	738	739	740	741	742	743	744	745	746	747	748	749	750	751	752	753	754	755	756	757	758	759	760	761	762	763	764	765	766	767	768	769	770	771	772	773	774	775	776	777	778	779	780	781	782	783	784	785	786	787	788	789	790	791	792	793	794	795	796	797	798	799	800	801	802	803	804	805	806	807	808	809	810	811	812	813	814	815	816	817	818	819	820	821	822	823	824	825	826	827	828	829	830	831	832	833	834	835	836	837	838	839	840	841	842	843	844	845	846	847	848	849	850	851	852	853	854	855	856	857	858	859	860	861	862	863	864	865	866	867	868	869	870	871	872	873	874	875	876	877	878	879	880	881	882	883	884	885	886	887	888	889	890	891	892	893	894	895	896	897	898	899	900	901	902	903	904	905	906	907	908	909	910	911	912	913	914	915	916	917	918	919	920	921	922	923	924	925	926	927	928	929	930	931	932	933	934	935	936	937	938	939	940	941	942	943	944	945	946	947	948	949	950	951	952	953	954	955	956	957	958	959	960	961	962	963	964	965	966	967	968	969	970	971	972	973	974	975	976	977	978	979	980	981	982	983	984	985	986	987	988	989	990	991	992	993	994	995	996	997	998	999	1000
--	---	---	---	---	---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	------

ภาพที่ 16 ผลลัพธ์การคำนวณค่า TF-IDF

จากภาพที่ 16 แสดงผลลัพธ์คำนวณค่า TF-IDF สำหรับแต่ละคำจากที่ผ่านการคำนวณค่า TF-IDF และแสดงออกมากเป็นตารางข้อมูลเรียงลำดับข้อมูลตั้งแต่ประโภคแรกถึงสุดท้าย

3.1.3 การจัดกลุ่มข้อมูล (Clustering)



```
1 from sklearn.cluster import KMeans
2 from sklearn.decomposition import PCA
3 from sklearn.preprocessing import StandardScaler
```

ภาพที่ 17 การเรียกใช้ไลบรารีสำหรับการจัดกลุ่มข้อมูลด้วยเมล์ (K-Means)

จากภาพที่ 17 แสดงการนำเข้าไลบรารีจาก Scikit-learn เพื่อทำการจัดกลุ่มข้อมูลโดยมีรายละเอียดดังนี้

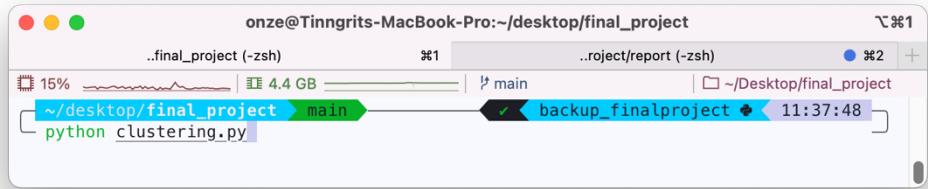
1. KMeans เป็นการเรียกใช้อัลกอริทึมสำหรับการจัดกลุ่ม
2. PCA (Principle Components Analysis) ซึ่งเป็นวิธีการลด dimension ของ Feature ลงช่วยลดTHONความซ้ำซ้อนของข้อมูลทำให้ train model ได้ง่ายขึ้น
3. StandardScaler เป็นตัวแปลงค่าตัวเลขให้อยู่ในปริมาณที่ใกล้เคียงกัน



```
1 k = 7
2 kmeans = KMeans(n_clusters=k, random_state=1)
3 # Fit model
4 kmeans.fit(df_tfidf[['x_value', 'y_value']])
5 clusters = kmeans.labels_
```

ภาพที่ 18 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k และการจัดกลุ่มข้อมูล

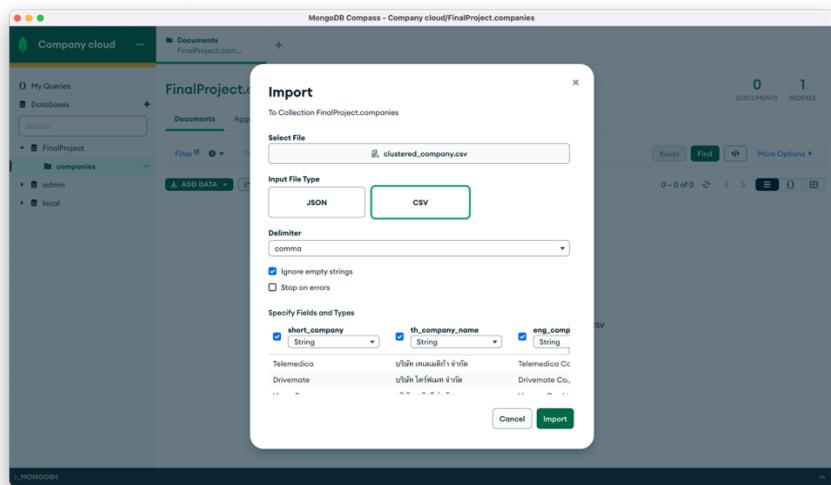
4. เมื่อกำหนดค่าการจัดกลุ่มเรียบร้อยเรียกใช้ไฟล์ clustering.py ใน Terminal เพื่อทำการจัดกลุ่มและบันทึกผลลัพธ์



ภาพที่ 19 การเรียกใช้งานไฟล์ clustering.py เพื่อจัดกลุ่มข้อมูลและบันทึกผลลัพธ์

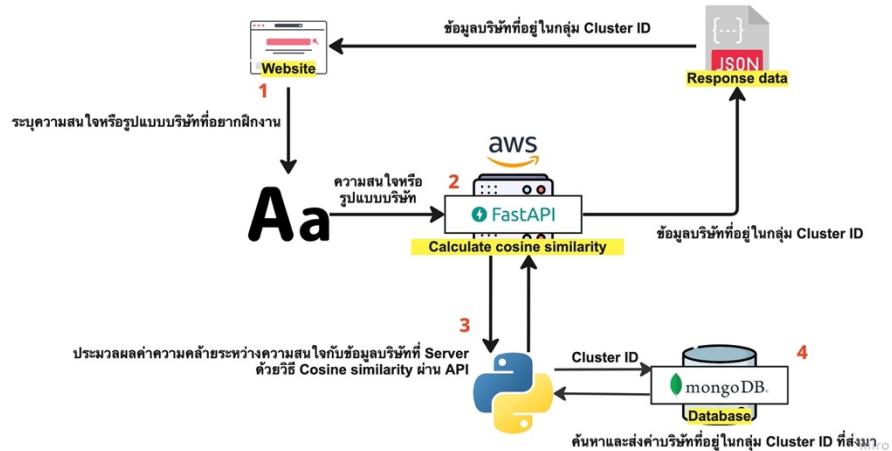
3.1.4 การนำเข้าข้อมูล Clustering เข้าสู่ฐานข้อมูล

1. นำเข้าข้อมูลลงฐานข้อมูล MongoDB โดยโปรแกรม MongoDB Compass
2. เชื่อมต่อ MongoDB Compass กับ Mongodb Atlas
3. นำเข้าข้อมูลด้วยไฟล์ .CSV ที่เป็นผลลัพธ์จากการจัดกลุ่มข้อมูล



ภาพที่ 20 การนำเข้าข้อมูลลงสู่ฐานข้อมูล MongoDB

3.2 การทำงานของระบบ



ภาพที่ 21 การทำงานของระบบ

จากภาพที่ 21 แสดงการทำงานของระบบได้ดังนี้

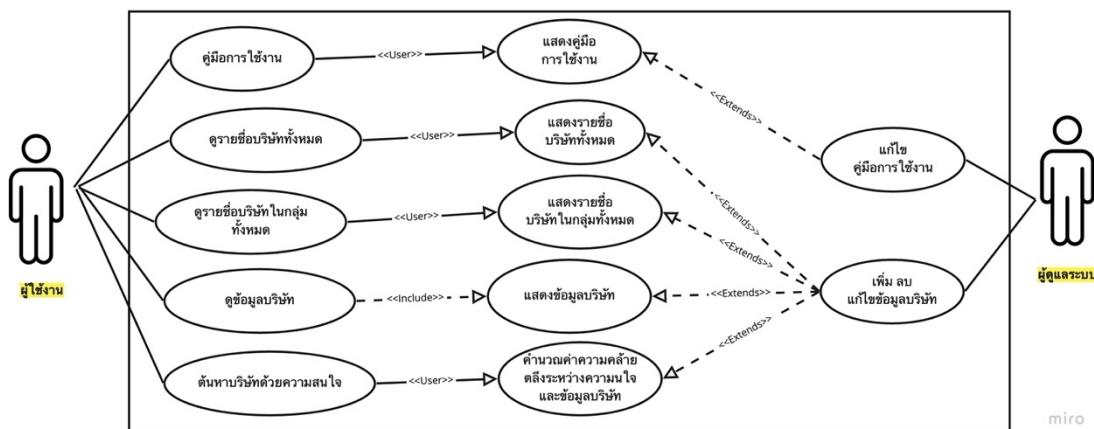
1. Website ใช้ระบุความสนใจเพื่อส่งค่าไปประมวลผลความคล้ายคลึงกับข้อมูลบริษัท
2. Server ใช้ประมวลผลความคล้ายคลึงกันระหว่างความสนใจที่ได้รับมา และข้อมูลบริษัทที่อยู่ในฐานข้อมูลด้วยเทคนิค Cosine similarity โดยภาษา Python และส่งค่ากลับไปเป็น Cluster ID
3. เมื่อได้ Cluster ID แล้วนำไปค้นหาบริษัทที่ Cluster ID ตรงกันในฐานข้อมูลและคืนค่า Response API เป็นข้อมูลในรูปแบบ JSON ที่มีข้อมูลบริษัทที่อยู่ใน Cluster ID นั้น
4. MongoDB เป็นฐานข้อมูลที่เก็บข้อมูลบริษัทไว้ และรอให้เซิร์ฟเวอร์เรียกใช้ข้อมูลเพื่อนำไปแสดงผล

3.3 การวิเคราะห์และออกแบบระบบ

3.3.1 การวิเคราะห์ระบบ

การวิเคราะห์ระบบและการออกแบบ (System Analysis and Design) คือวิธีการที่ใช้ในการสร้างระบบสารสนเทศซึ่งมาใหม่ในธุรกิจ ได้แก่ วิธีการที่ใช้ในการสร้างระบบสารสนเทศใหม่ แล้ว การวิเคราะห์ระบบช่วยในการแก้ไขระบบสารสนเทศเดิมที่มีอยู่แล้วให้ดีขึ้น การวิเคราะห์ระบบ คือ การหาความต้องการ (Requirements) ของระบบสารสนเทศว่าคืออะไร หรือต้องการเพิ่มเติมอะไรเข้ามาในระบบ การออกแบบ คือ การนำเอาความต้องการของระบบมาเป็นแบบแผน หรือเรียกว่าพิมพ์เขียวในการสร้างระบบสารสนเทศนี้ให้ใช้งานได้จริง

3.3.2 ຢູ່ລາຍລະອຽດຂອງແກຣມ (Use Case Diagram)



ภาพที่ 22 Use Case Diagram ของระบบ

ตารางที่ 5 คำอธิบาย Use case คู่มือการใช้งาน

Use case id:	1
Use case name:	คุ้มครองการใช้งาน
Actor:	ผู้ใช้งาน
Scenario:	การตัดสินใจการใช้งานเว็บไซต์
Trigger event:	None
Brief Description:	อ่านวิธีการใช้งานเว็บไซต์
Purpose:	เพื่อใช้งานเว็บไซต์
Pre-condition:	เมื่อต้องการใช้งานเว็บไซต์
Main flow:	<ol style="list-style-type: none"> ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้าเกี่ยวกับ อ่านวิธีการใช้งานเว็บไซต์
Alternate/Exceptional Flow:	None

ตารางที่ 6 คำอธิบาย Use case ดูรายชื่อบริษัททั้งหมด

Use case id:	2
Use case name:	ดูรายชื่อบริษัททั้งหมด
Actor:	ผู้ใช้งาน
Scenario:	การดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล
Trigger event:	None
Brief Description:	ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล
Purpose:	ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล
Pre-condition:	เมื่อต้องการดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล
Main flow:	<ol style="list-style-type: none"> 1. ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้ารายชื่อบริษัททั้งหมด 2. ดูรายชื่อบริษัททั้งหมดที่มีในฐานข้อมูล
Alternate/Exceptional Flow:	None

ตารางที่ 7 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:	<ol style="list-style-type: none"> 1. ผู้ใช้งานเปิดเว็บไซต์เข้าไปยังหน้ากลุ่มบริษัทที่ต้องการ 2. ดูรายชื่อบริษัทในกลุ่มทั้งหมด
Alternate/Exceptional Flow:	None

ตารางที่ 8 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

ตารางที่ 8 (ต่อ)

Purpose:	เพื่อคูช้อมูลบริษัท
Pre-condition:	เมื่อต้องการคูช้อมูลบริษัท
Main flow:	<ol style="list-style-type: none"> ผู้ใช้งานคลิกที่เมนูชื่อของบริษัทที่ต้องการคูช้อมูล คูช้อมูลบริษัท
Alternate/Exceptional Flow:	None

ตารางที่ 9 คำอธิบาย Use case คนหาบริษัทด้วยความสนใจ

Use case id:	5
Use case name:	คนหาบริษัทด้วยความสนใจ
Actor:	ผู้ใช้งาน
Scenario:	การคนหาบริษัทด้วยความสนใจ
Trigger event:	None
Brief Description:	คนหาบริษัทด้วยความสนใจหรือรูปแบบธุรกิจ
Purpose:	เพื่อคนหาบริษัทด้วยความสนใจ
Pre-condition:	เมื่อต้องการคนหาบริษัทด้วยความสนใจ
Main flow:	<ol style="list-style-type: none"> ผู้ใช้งานพิมพ์ความสนใจหรือรูปแบบธุรกิจที่ชองค์หนาและ แสดงรายชื่อบริษัทที่อยู่ในกลุ่มที่ระบบแนะนำ ผู้ใช้คลิกเลือกบริษัทเพื่อคูช้อมูลบริษัท
Alternate/Exceptional Flow:	None

ตารางที่ 10 คำอธิบาย Use case แก้ไขคูมีการใช้งาน

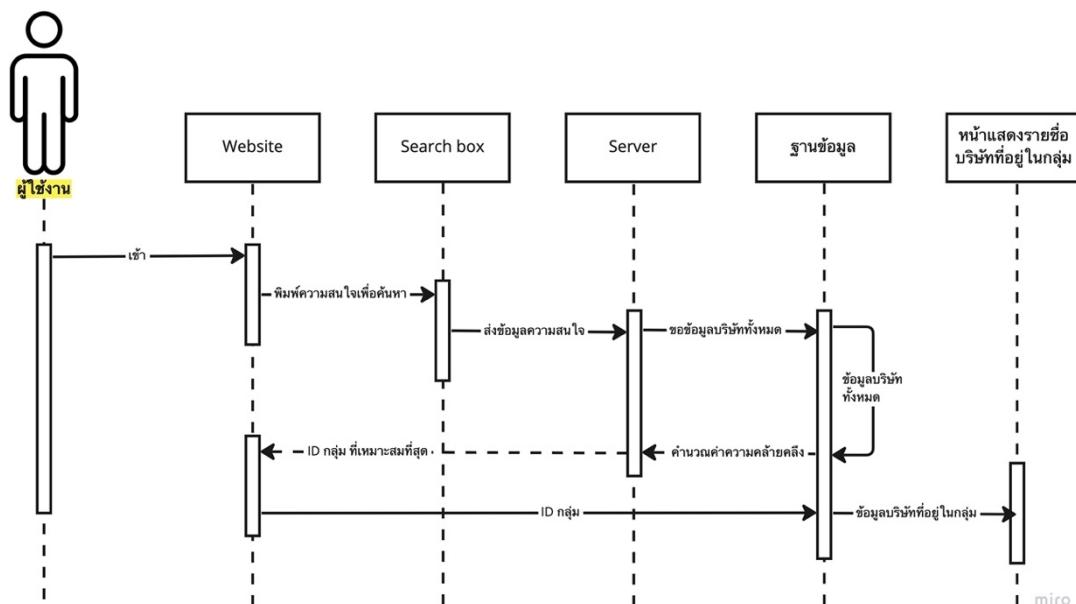
Use case id:	6
Use case name:	แก้ไขคูมีการใช้งาน
Actor:	ผู้ดูแลระบบ
Scenario:	การแก้ไขคูมีการใช้งาน
Trigger event:	None
Brief Description:	แก้ไขคูมีการใช้งาน
Purpose:	เพื่อแก้ไข คูมีการใช้งาน
Pre-condition:	เมื่อต้องการแก้ไขคูมีการใช้งาน
Main flow:	<ol style="list-style-type: none"> ผู้ดูแลระบบแก้ไขคูช้อมูลคูมีการใช้งาน Deploy เพื่ออัปเดตระบบ
Alternate/Exceptional Flow:	None

ตารางที่ 11 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)

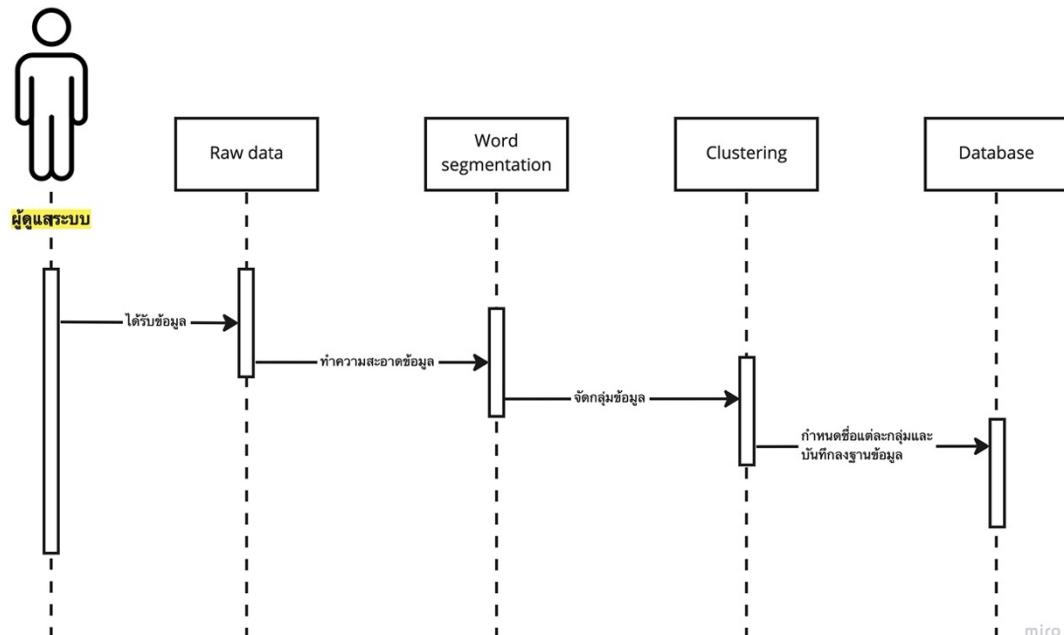
Sequence Diagram เป็นหนึ่งในแผนผังการทำงานแบบ Unified Modeling Language (UML) ใช้สำหรับการสร้างแบบจำลองเชิงวัตถุ โดยข้อแตกต่างจากแผนผังรูปแบบ UML อื่นคือเป็นแผนผังการทำงานที่แสดงลำดับการปฏิสัมพันธ์ (Sequence of interactions) ระหว่างวัตถุที่แสดงภายใต้ระบบต่าง ๆ เช่น การส่งข้อความ (messaging) ที่มีการรับส่งข้อมูลระหว่างผู้ใช้ Sequence Diagram เป็นแผนผังการทำงานที่ประกอบไปด้วยคลาส (Class) หรือวัตถุ (Object) เส้นประที่ใช้เพื่อแสดงลำดับเวลา และเส้นที่ใช้เพื่อแสดงริบิกรรมที่เกิดขึ้นจากคลาสหรือวัตถุในแผนผังการทำงานภายใต้ Sequence Diagram จะใช้สีเหลี่ยมแทนสมือนคลาส และวัตถุโดยภายนอกจะมีชื่อของคลาสหรือวัตถุประกอบอยู่ในรูปแบบ [Object]: Class



ภาพที่ 23 Sequence Diagram การค้นหาบริษัทด้วยความสนใจของผู้ใช้

ตารางที่ 12 อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การค้นหาบริษัทด้วยความสนใจ

เหตุการณ์ที่เกิดขึ้น	คำอธิบาย
เข้า Website	เข้า Website ด้วย Browser
พิมพ์ความสนใจเพื่อค้นหา	ระบุความสนใจรูปแบบอุปกรณ์ของบริษัทหรือความสนใจที่อยากฝึกงานของผู้ใช้
ส่งข้อมูลความสนใจ	ส่งข้อมูลความสนใจที่ผู้ใช้ระบุไปประมวลผลที่ Server
ขอข้อมูลบริษัททั้งหมด	Server ขอข้อมูลบริษัททั้งหมดจากฐานข้อมูลเพื่อนำมาเก็บไว้รอคำนวณคาดคะเนความคล้ายคลึง
ข้อมูลบริษัททั้งหมด	ข้อมูลบริษัททั้งหมดในฐานข้อมูล ส่งให้ Server
คำนวณคาดคะเนความคล้ายคลึงระหว่างความสนใจของผู้ใช้และข้อมูลบริษัท	คำนวณคาดคะเนความคล้ายคลึงระหว่างความสนใจของผู้ใช้และข้อมูลบริษัท
ID กลุ่มที่เหมาะสมที่สุด	คืนค่า ID ของกลุ่มบริษัทที่คล้ายกับความสนใจของผู้ใช้
ID กลุ่ม	ส่งค่า ID ของกลุ่มไปยังฐานข้อมูลเพื่อขอข้อมูลบริษัทที่อยู่ในกลุ่มนั้น ๆ
ข้อมูลบริษัทที่อยู่ในกลุ่ม	แสดงรายชื่อบริษัทที่อยู่ในกลุ่มในหน้าเว็บ



ภาพที่ 24 Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มบริษัทใหม่

ตารางที่ 13 อธิบายเหตุการณ์ที่เกิดขึ้นใน Sequence Diagram การเพิ่มข้อมูลและจัดกลุ่มใหม่

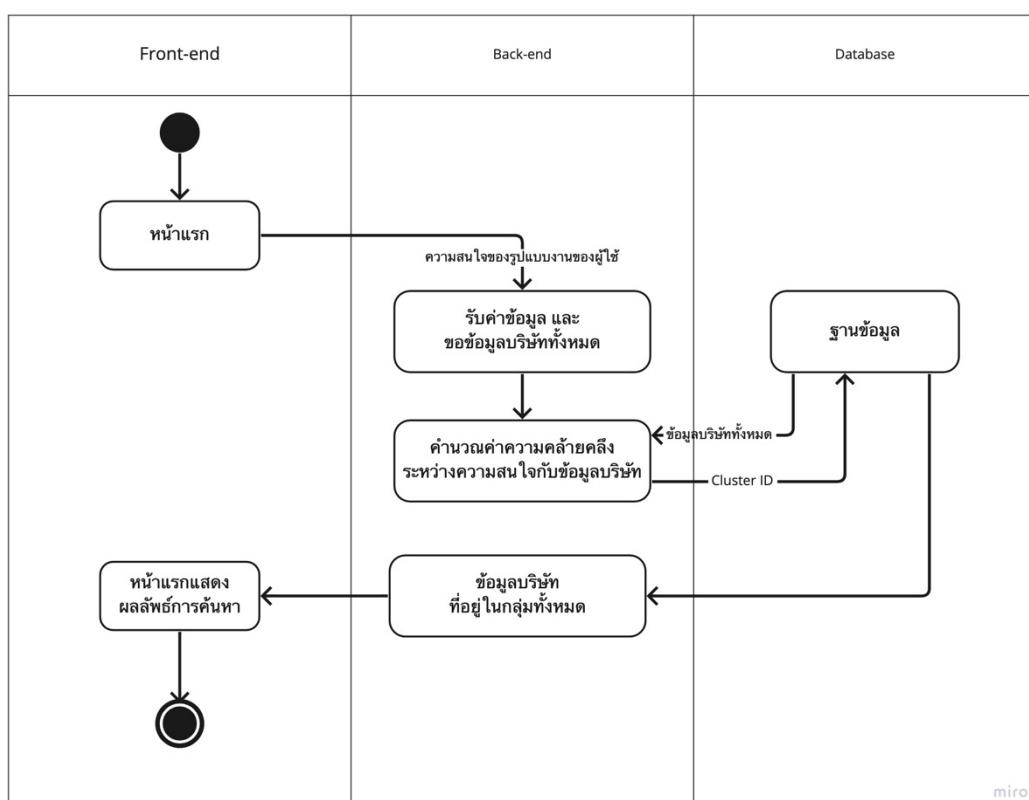
เหตุการณ์ที่เกิดขึ้น	คำอธิบาย
ได้รับข้อมูล	ได้ข้อมูลดิบที่จะนำมาใช้งาน
ทำความสะอาดข้อมูล	นำข้อมูลดิบมาทำการลบตัวเลข คำที่ไม่มีความหมายในตัว คำສະกัดผิด
จัดกลุ่มข้อมูล	ทำการหาคำสำคัญและทำการจัดกลุ่มข้อมูล
กำหนดชื่อแต่ละกลุ่มและบันทึกลงฐานข้อมูล	กำหนดชื่อของกลุ่มและบันทึกข้อมูลลงฐานข้อมูลเพื่อใช้ในเว็บไซต์

3.3.4 แอคทิวิตี้โดยแกรม (Activity Diagram)

Activity Diagram หรือแผนภาพกิจกรรม ใช้อธิบายกิจกรรมที่เกิดขึ้นในลักษณะกระแสการไหลของการทำงาน (Workflow) จะมีลักษณะเดียวกับ Flowchart โดย ขั้นตอนในการทำงานแต่ละขั้นจะเรียกว่า Activity

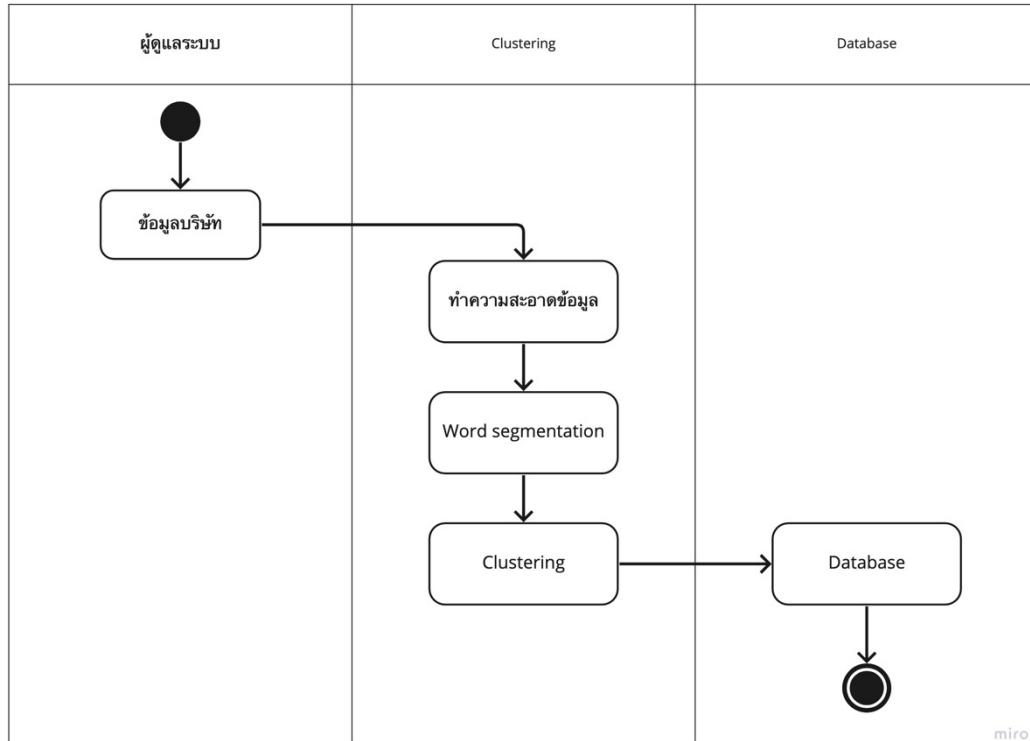
การใช้งาน Activity Diagram

- อธิบายกระแสการไหลของการทำงาน (Workflow)
- แสดงขั้นตอนการทำงานของระบบ



ภาพที่ 25 Activity Diagram ของผู้ใช้งาน

จากภาพที่ 25 แสดงขั้นตอนการทำงานของระบบเมื่อผู้ใช้คนหนึ่งห้ามบริษัทด้วยความสนใจของผู้ใช้การทำงานจะเริ่มต้นจากการที่ผู้ใช้ระบุความสนใจ จากนั้นทำการส่งข้อมูลไปคำนวณค่าความคล้ายคลึงผ่าน API และเรียกขอມูลบริษัทที่มีความคล้ายมากที่สุดมาแสดงผลหน้าเก็บใช้



ภาพที่ 26 Activity Diagram ของผู้ดูแลระบบ

จากภาพที่ 26 แสดงการทำงานของการจัดกลุ่มข้อมูลเพื่อนำไปใช้ในระบบโดยการทำงานเริ่มต้นที่นำข้อมูลที่ได้เก็บรวบรวมลงในไฟล์จากนั้นทำการทำ Word segmentation และทำการจัดกลุ่มข้อมูล ลูกท้ายบันทึกข้อมูลเพื่อนำไปนำเข้าลงฐานข้อมูล MongoDB

3.4 การออกแบบฐานข้อมูล

3.4.1 แบบจำลองโครงสร้างของฐานข้อมูล (Entity–Relationship Diagrams: ER Diagram)

แบบจำลองโครงสร้างของฐานข้อมูล (Entity–Relationship Diagrams: ER Diagram) เป็นเครื่องมือในการออกแบบ การอธิบายโครงสร้างและความสัมพันธ์ของข้อมูล (Relationship) ประกอบด้วย 1. เอ็นทิตี้ (Entity) เป็นวัตถุ หรือสิ่งของที่เราสนใจในระบบงานนั้น 2. 属性 (Attribute) เป็นคุณสมบัติของวัตถุที่เราสนใจ 3. ความสัมพันธ์ (Relationship) คือ ความสัมพันธ์ระหว่างเอนทิตี้ ER Diagram มีความสำคัญต่อการพัฒนาระบบงานฐานข้อมูล Application ต่างๆ ที่ต้องการการเก็บข้อมูลอย่างมีระบบ มีโครงสร้าง

ตั้งนั้น ER Diagram จึงใช้เพื่อเป็นเอกสารในการสื่อสารระหว่าง นักออกแบบระบบ และ นักพัฒนาระบบ เพื่อให้สื่อสารอย่างตรงกัน

companies	
_id	PK
short_company	
th_company_name	
eng_company_name	
type_business	
product	
type_innovation	
detail	
owner	
province_base	
address	
phone_number	
email	
website	
source	
cluster	

ภาพที่ 27 ER Diagram ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยี การจัดกลุ่มเค้มีน (K-Means)

3.4.2 พจนานุกรมข้อมูล (Data Dictionary)

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

ตารางที่ 14 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	ชื่อบริษัท ภาษาไทย	Null
eng_company_name	String	50	ตัวอักษร	ชื่อบริษัท ภาษาอังกฤษ	Null
type_business	String	50	ตัวอักษร	ประเภท ธุรกิจ	Null

ตาราง 14 (ต่อ)

product	String	255	ตัวอักษร	ประเภทสินค้า	Null
type_innovation	String	50	ตัวอักษร	ประเภทเทคโนโลยี	Null
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 นั้นเป็นส่วนที่ผู้พัฒนาโปรแกรมต้องทำเนื่องจากหน้าจอจะเป็นส่วนสำคัญที่จะเชื่อมต่อกับผู้ใช้งานโปรแกรมหรือก็คือส่วนที่ผู้ใช้งานจะเห็น สำหรับการติดต่อบอกได้ ซึ่งจะมุ่งเน้นการออกแบบในทางด้านของรูปแบบหน้าตา บุ่มกัด ช่องที่ใช้พิมพ์สำหรับคนหา ขนาดตัวอักษร สี และรูปภาพ เป็นต้น

การออกแบบหน้าจอสำหรับระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) จะเน้นในแพลตฟอร์มเว็บแอพลิเคชันที่เป็นหน้าจอคอมพิวเตอร์หรือ Desktop เป็นหลัก โดยมีรายละเอียดดังนี้

3.5.1 หน้าแรก

The screenshot shows the application's header with the logo 'Intern assistant' and a red circular badge with the number '1'. Below the header, there are two tabs: 'เกี่ยวกับ' (About) and 'รายชื่อสถานประกอบการทั้งหมด' (List of all business operators). A purple button labeled 'สมัคร!' (Sign up!) is located on the right.

ค้นหาสถานประกอบการ
ระบุชื่อองค์กรค้นหาข้อมูลของคุณ

สถานประกอบการ บล๊าสบิว

3 Ernsner - Schmeler
192 Screeching Street, West Malling, BD72 6RG
Joe@topdog.org
07183 776 242
Online marketing

Luettgen - Hintz
400 Kindy Road, Llandover, W27 1TF
Dermalogical.org
07223 938 213
Online marketing

Stehr, Doyle and Schultz
430 Payment Avenue, Surrey, TS77 7HA
Hansel@wealthy.co
07154 387 563
Network

Lemke and Sons
467 Chilly Road, Winsford, TA88 2SR

Champlin - Kessler
229 Pencil Close, Countess Wear, BN78 8RH

Christiansen Group
466 Annoying Crescent, Newton Stewart, PH6 6S

ภาพที่ 28 หน้าแรก

จากภาพที่ 28 แสดงการออกแบบหน้าแรกของเว็บไซต์ซึ่งประกอบไปด้วยส่วนประกอบดังนี้

หมายเลข 1 เมนูหลักสามารถเชื่อมโยงไปยังหน้าแสดงรายละเอียดของเว็บไซต์ หน้าสถานประกอบการทั้งหมด และหน้าสถานประกอบการแต่ละประเภท

หมายเลข 2 เป็นช่องสำหรับระบุความสนใจของผู้ใช้เพื่อนำไปค้นหากลุ่มบริษัทที่มีความคล้ายคลึงกับความสนใจของผู้ใช้

หมายเลข 3 เป็นการแนะนำบริษัทที่น่าสนใจจากที่อยู่ในฐานข้อมูล

3.5.2 หน้าเกี่ยวกับ

The screenshot shows the application's header with the logo 'Intern assistant' and a red circular badge with the number '1'. Below the header, there are three tabs: 'เกี่ยวกับ' (About), 'รายชื่อสถานประกอบการทั้งหมด' (List of all business operators), and 'หน้าห้องลับสถานประกอบการ' (Hidden page of business operators). A purple button labeled 'สมัคร!' (Sign up!) is located on the right.

เกี่ยวกับ

Repellendus laudantium dignissimos deleniti. Officiis et maiores quod veritatis dignissimos voluptatem possimus. Magni tempore sed. Animis eum voluptas forum esse amet quisquam. Tempore suscipit animi harum voluptatem impedit. Temporibus sed aperiam impedit cum modi. Autem architecto est eveniet Et dolores assumenda numquam qui qui. Debitis et perspicillatis ad iste. Minima repudiandae dolor rerum et aut. Atque magni ullam assumenda a consecutetur molestias tenetur dolorem eveniet. Et qui ea quam ea quia. Expedita libero enim ut. Vitae iusto sed molestiae ut optio dolor perferendis perferendis. Praesentium ipsum provident qui ut error beatae quibusdam. Et qui animi qui in voluptas sed.

รูปแบบการทำงาน

นี่เป็นรูปแบบการทำงานที่สำคัญ

```

graph LR
    A[CSV] --> B[ ]
    B --> C[Database]
    
```

SERVICE

API Github Document Article

ภาพที่ 29 หน้าเกี่ยวกับ

หมายเลข 1 คือการอธิบายรายละเอียดเกี่ยวกับระบบและรูปแบบการทำงานที่ดำเนินการในส่วนต่าง ๆ

3.5.3 หน้าแสดงรายชื่อปริษัทในกลุ่มทั้งหมด

1 Ernsler - Schmeler
Tramp LLC
Gusikowski - Considine
Dicki, Welch And Rippin
Boyer - Daugherty
Gutkowski - Kautzer
Ledner - Pfeffer
Mohr, Herzog And Terry
Harris, Veum And Kertzmann
Rolfson - Wiza
Baumbach - Raynor
Dach, Cummings And Lindgren
Halvorson - Nikolaus
Mueller - Hodkiewicz

Von, Kunde And Stracke
Bode And Sons
Grady, Huels And Runte
Balistrieti Inc
Cartwright, Goyette And Watsica
Schulist LLC
Olson LLC
Davis, Dibbert And Schuster
Ferry - Howe
Konopelski Group
Rutherford - Denesik
Grant And Sons
Wilkinson - Hegmann
Conn - Bergstrom

ภาพที่ 30 หน้าแสดงรายชื่อปริษัทในกลุ่มทั้งหมด

หมายเหตุ 1 แสดงรายชื่อปริษัทที่อยู่ในกลุ่มที่เลือกดูทั้งหมดและสามารถคลิกเพื่อรายละเอียด

3.5.4 หน้าแสดงรายชื่อปริษัททั้งหมด

1 Ernsler - Schmeler
Tramp LLC
Gusikowski - Considine
Dicki, Welch And Rippin
Boyer - Daugherty
Gutkowski - Kautzer
Ledner - Pfeffer
Mohr, Herzog And Terry
Harris, Veum And Kertzmann
Rolfson - Wiza
Baumbach - Raynor
Dach, Cummings And Lindgren
Halvorson - Nikolaus
Mueller - Hodkiewicz

Von, Kunde And Stracke
Bode And Sons
Grady, Huels And Runte
Balistrieti Inc
Cartwright, Goyette And Watsica
Schulist LLC
Olson LLC
Davis, Dibbert And Schuster
Ferry - Howe
Konopelski Group
Rutherford - Denesik
Grant And Sons
Wilkinson - Hegmann
Conn - Bergstrom

ภาพที่ 31 หน้าแสดงรายชื่อปริษัททั้งหมด

หมายเหตุ 1 รายชื่อปริษัททั้งหมดที่อยู่ในฐานข้อมูลและสามารถคลิกดูข้อมูลปริษัทได้

3.5.5 หน้าแสดงผลลัพธ์รายชื่อบริษัท

ค้นหาสถานประกอบการ
ระบุชื่อของธุรกิจค้าขายที่คุณต้องการ

1 SEO
X Home

2 Online marketing
3 Top LLC
4 Gusikowski - Considine
5 Dicki, Welch And Rippin
6 Boyer - Daugherty
7 Von, Kunde And Stracke
8 Bode And Sons
9 Grady, Huels And Runte
10 Ballistri Inc
11 Cartwright, Goyette And Watsica

ภาพที่ 32 หน้าแสดงผลลัพธ์รายชื่อบริษัท

หมายเหตุ 1 ช่องสำหรับระบุความสนใจของผู้ใช้เพื่อค้นหาบริษัทที่คล้ายคลึง
หมายเหตุ 2 ผลลัพธ์รายชื่อบริษัทที่มีความคล้ายคลึงกับความสนใจที่ผู้ใช้ระบุ

3.5.6 หน้าแสดงข้อมูลบริษัท

1 Ernsler - Schmeler
Online marketing

2 467 Chilly Road, Winsford, TA88 2SR
pe@torpid.org
07183 776 242

3 Requiescam sub libato

ภาพที่ 33 หน้าแสดงข้อมูลบริษัท

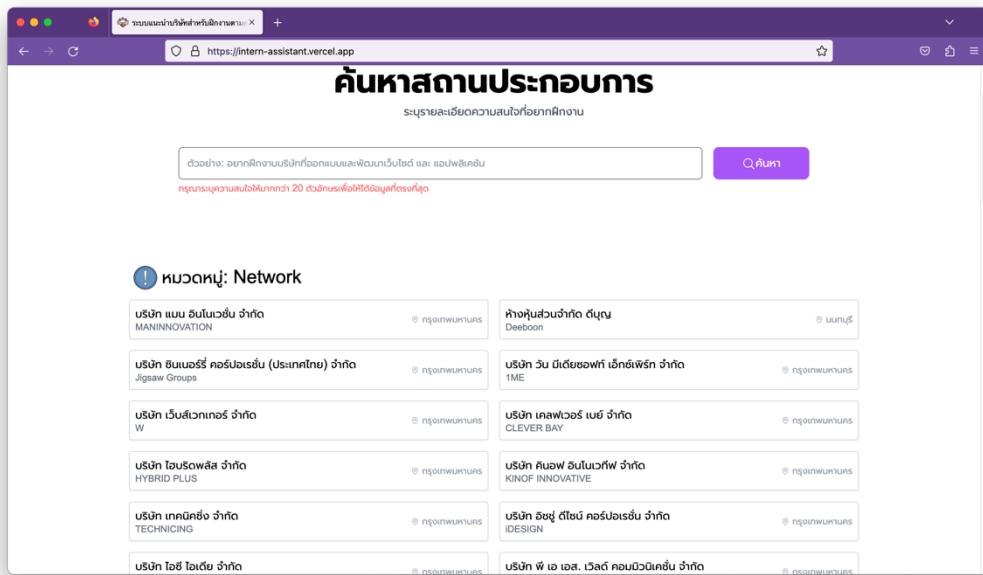
หมายเหตุ 1 แสดงชื่อของบริษัท

หมายเหตุ 2 แสดงรายละเอียดข้อมูลที่อยู่ ข้อมูลการติดต่อของบริษัท

หมายเหตุ 3 แสดงข้อมูลรายละเอียดรูปแบบธุรกิจที่บริษัทดำเนินกิจการอยู่

3.6 การใช้งานระบบ

1. ระบุความสนใจในรูปแบบธุรกิจ หรือสิ่งที่อยากร่วมในการฝึกงานลงในช่องคนหา
2. เมื่อได้ผลลัพธ์บริษัทที่มีความคล้ายคลึงกับสิ่งที่ค้นหาแล้วผู้ใช้สามารถทำการพิจารณาบริษัทเพื่อเลือกตัดสินใจในการฝึกงานได้



រាយទី 34 តារាងផលិតផលិតផ្ទាល់របស់ក្រសួងការងារក្នុងការគាំទ្រិតរាយការងារ

บทที่ 4

ผลการดำเนินงาน

การทำงานของระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเดjmén (K-Means) มีขั้นตอนการทำงานดังนี้

4.1 การวิเคราะห์และการตัดคำ (Word segmentation)

4.1.1 การวัดค่าความแม่นยำในการตัดคำ

ในขั้นตอนการทำ Word segmentation นั้นมีขั้นตอนอยู่อย่างในการทำร่วมด้วยหลายขั้นตอนหลังจากทำความสะอาดข้อมูล คือการหาคำสำคัญของแต่ละประโยคในที่นี่คือรายละเอียดธุรกิจของแต่ละบริษัท การทำการตัดคำที่ไม่สื่อความหมายหรือ Stop word ออกไปจากประโยคเพื่อให้ได้ประโยคที่มีเนื้อหาใจความดีที่สุด และอีกขั้นตอนสำคัญคือการหาคีย์เวิร์ดของแต่ละประโยคเพื่อที่จะได้ทราบว่าประโยคนั้น ๆ กำลังสื่อถึงเรื่องไหนเป็นลำดับด้วยเทคนิค TF-IDF ที่เป็นการหาหนังสือของคำนั้น ๆ ในประโยค

ในการทำงานประมวลผลเกี่ยวกับการประมวลผลภาษาธรรมชาติจำเป็นต้องมีการตัดคำออกเป็นคำ ๆ เพื่อจะได้ง่ายและนำไปเข้าสู่กระบวนการต่าง ๆ ได้อย่างง่ายโดยปกติแล้วการตัดคำในภาษาอังกฤษสามารถตัดได้โดยใช้การเก็บรวมเป็นเงื่อนไขในการตัด แต่ในภาษาไทยนั้นการเขียนนั้นไม่ได้มีการเว้นวรรคคำเหมือนภาษาอังกฤษทำให้การตัดคำนั้นจะใช้เก็บรวมมาตัดคำตลอดไม่ได้ จำเป็นต้องใช้อัลกอริทึมอื่น ๆ เข้ามาช่วย เช่นการใช้ Dictionary-based, Maximum Matching เป็นต้น

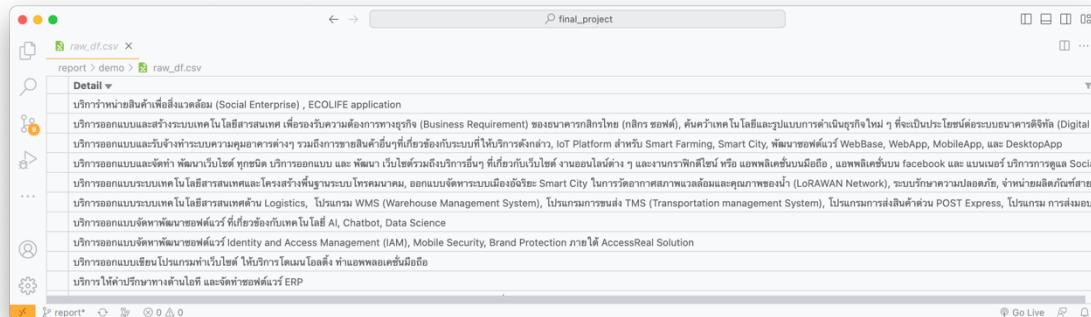
และในภาษา Python มีเครื่องมือที่สามารถช่วยอำนวยความสะดวกในการตัดคำภาษาไทยอย่าง Pythainlp ที่ผู้วิจัยได้เลือกใช้ในโครงงานนี้ในไลบรารีนั้นสามารถตัดคำได้หลาย Engine ด้วยกันและในแต่ละตัวเลือกก็ใช้อัลกอริทึมต่างกันยกตัวอย่างเช่น newmm, longest, newmm-safe, mm, icu, deepcut, attacut เป็นต้น

1. การวัดค่าความแม่นยำในการตัดคำ

การเลือก 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

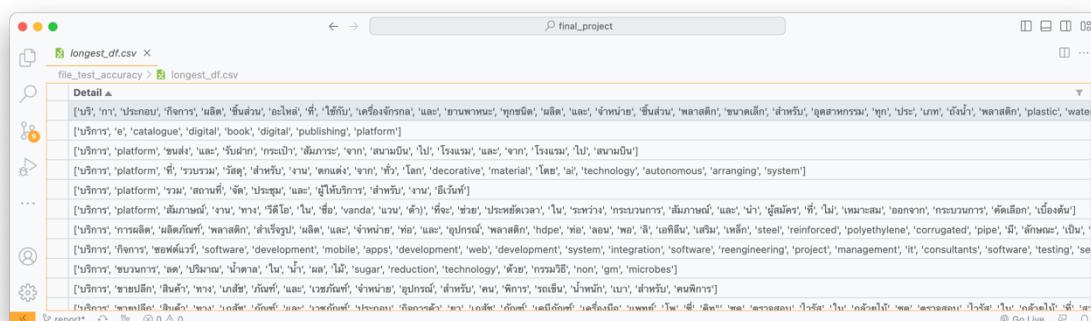
วิธีทดสอบคือทำการสุ่มเลือกประโยคมาจำนวน 100 ประโยค ทำการตัดคำในแต่ละประโยคเองโดยไม่ใช้ตัวช่วย จากนั้นใช้ Engine ในไลบรารี Pythainlp เพื่อตัดคำและนำมาเทียบกับประโยคที่ผู้วิจัยตัดได้โดย Engine ให้หนที่มีความเหมือนกับที่ผู้วิจัยตัดไว้มากที่สุดก็จะถือว่ามีความใกล้เคียงกับความต้องการของผู้วิจัยมากที่สุด



ภาพที่ 35 ตัวอย่างข้อมูลต้นฉบับ



ภาพที่ 36 ตัวอย่างการตัดคำโดยใช้ Engine newmm



ภาพที่ 37 ตัวอย่างการตัดคำโดยใช้ Engine longest

ภาพที่ 38 ตัวอย่างการตัดคำโดยใช้ Engine deepcut

ยกตัวอย่างการทำการทำทดสอบการตัดคำโดยใช้ Engine newmm, longest และ deepcut โดยที่ใช้วิธีเทียบการตัดคำโดยที่ผู้วิจัยทำการตัดคำเองและนำไปเทียบกับการใช้โลบรารีตัดคำโดยนับจำนวนเฉพาะคำที่โลบรารีตัดตรงกับผู้วิจัยด้วยเหตุนี้จึงถือว่า Engine ได้มีการตัดคำที่ตรงกับผู้วิจัยมากที่สุดจะถือเป็นตรงกับมนุษย์มากที่สุด

โจทย์: บริการ Software Business Solutions ระบบปรับสมัครบุคลากรออนไลน์ ระบบลงทะเบียนเรียนออนไลน์

ตารางที่ 15 ตารางตัวอย่างการวัดค่าความแม่นยำในการตัดคำ

ลำดับ	Human	Newmm	Longest	Deepcut
1	บริการ	บริการ	บริการ	บริการ
2	software	software	software	software
3	Business	business	business	business
4	Solutions	solutions	solutions	solutions
5	ระบบ	ระบบ	ระบบ	ระบบ
6	รับสมัคร	รับสมัคร	รับสมัคร	รับ
7	บุคลากร	บุคลากร	บุคลากร	สมัคร
8	ออนไลน์	ออนไลน์	ออนไลน์	บุคลากร
9	ระบบ	ระบบ	ไลน์	ออนไลน์
10	ลงทะเบียนเรียน	ลงทะเบียนเรียน	ระบบ	ระบบ
11	ออนไลน์	ออนไลน์	ลงทะเบียนเรียน	ลง
12			ออนไลน์	ลงทะเบียน
13				เรียน
14				ออนไลน์
รวม		11	10	9

จากตารางที่ 15 จะเห็นว่าเมื่อทำการใช้ไลบรารีในการตัดคำแล้วสามารถสรุปได้ว่า Engine ที่มีความคล้ายกับผู้วิจัยมากที่สุดคือ newmm รองลงมาคือ longest และ deepcut
ตารางที่ 16 ผลการทดสอบความแม่นยำการตัดคำ

Engine	Accuracy (%)
newmm	90.99%
longest	83.04%
deepcut	76.65%

จากการทดลองจากข้อมูลปริมาณ 100 รายการพบว่า Engine newmm ในไลบรารี Pythainlp นั้นมีความแม่นยำกับผู้วิจัยตัดมากที่สุดที่ 90.99% รองลงมาที่ longest ค่าความแม่นยำอยู่ที่ 83.04% และความแม่นยำน้อยที่สุดคือ deepcut ที่มีความแม่นยำ 76.65% ดังภาพที่ 39 ดังนั้นโครงงานนี้จึงใช้ Engine newmm ในการตัดคำเพื่อนำไปประมวลผลต่อในขั้นตอนอื่น ๆ

```
onze@Tinngrits-MacBook-Pro:~/Documents/final_project
..final_project (-zsh) ❶ ..final_project (-zsh) ❷
└── python calculate_accuracy.py
newmm: 90.99%
deepcut: 76.65%
longest: 83.04%
[~/Documents/final_project] report +1 !2 ?2 5s 14:41:13
```

ภาพที่ 39 ผลการวัดค่าความแม่นยำในการตัดคำของ Engine ในไลบรารี Pythainlp

4.1.2 การคำนวณค่า TF-IDF

เมื่อได้เครื่องมือที่จะช่วยตัดคำแล้วขั้นตอนต่อไปคือการนำมาทำการจัดกลุ่มข้อมูลโดยเทคนิคที่เลือกใช้คือการหาค่า TF-IDF (Term Frequency–Inverse Document Frequency) เพื่อหาว่าคำไหนในประโยคนั้นเป็นคำสำคัญของประโยคนั้น ๆ โดยการวัดจากน้ำหนักของคำด้วยวิธีดังกล่าว

	aaa	ab	abap	abeam	ablerex	abroad	academic	acceptance	access	accessories	...	ໄອທີ	ໄອສ ຕຽມ	ໄອ ເຕືອ	ໄອ ແພດ	ໄອ ໂນມາຍ	ໄອ ໂສ	ໄອ ໂຄຣ	ໄອ ໂຫວງເຈນ
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	
1640	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1641	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1642	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1643	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1644	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1645	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

ภาพที่ 40 ตัวอย่างตาราง TF-IDF แสดงน้ำหนักของคำ

เมื่อได้ชุดของประโยชน์ที่ทำการตัดคำเรียบร้อยแล้วจึงนำเข้าสู่กระบวนการกรองคำที่เป็น Stop word หรือคำที่ไม่สื่อความหมายออกและตัวอักษรพิเศษต่าง ๆ ด้วยฟังก์ชันในไลบรารี Pythainlp และ nltk ยกตัวอย่างเช่น

ประโยชน์เริ่มต้น: บริการ Platform รวมสถานที่จัดประชุมและผู้ให้บริการสำหรับงานอีเว้นท์ที่ดีที่สุด

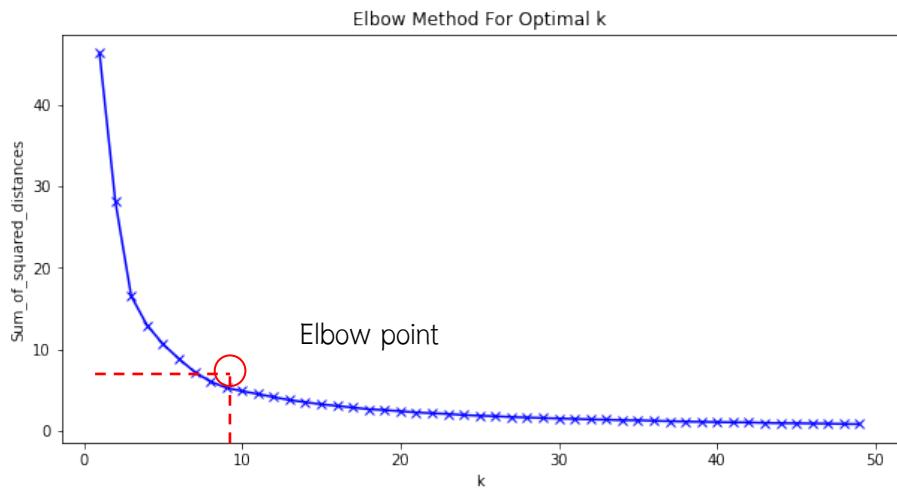
คำที่ไม่สื่อความหมาย (Stop word): รวมจัดและที่ที่สุด

ประโยชน์ใหม่: บริการPlatformสถานที่ประชุมผู้ให้บริการสำหรับงานอีเว้นท์ที่ดี

```
onze@Tinngrits-MacBook-Pro:~/Documents/final_project
..final_project (-zsh)  xt1 ..final_project (-zsh)  xt2 +
23% python report/demo/main.py
Default: นิรภัย Platform รวม สถานที่ จัด ประชุม และ ผู้ให้บริการ สำหรับ งาน อีเว้นท์ ที่ ดี ที่สุด
Keep stop word: ['นิรภัย', 'Platform', 'รวม', 'สถานที่', 'จัด', 'ประชุม', 'และ', 'ผู้ให้บริการ', 'สำหรับ', 'งาน', 'อีเว้นท์', 'ที่', 'ดี', 'ที่สุด']
Removed stop word: ['นิรภัย', 'platform', 'สถานที่', 'ประชุม', 'ผู้ให้บริการ', 'สำหรับ', 'งาน', 'อีเว้นท์', 'ที่']
```

ภาพที่ 41 ตัวอย่างการตัดคำและลบ Stop word

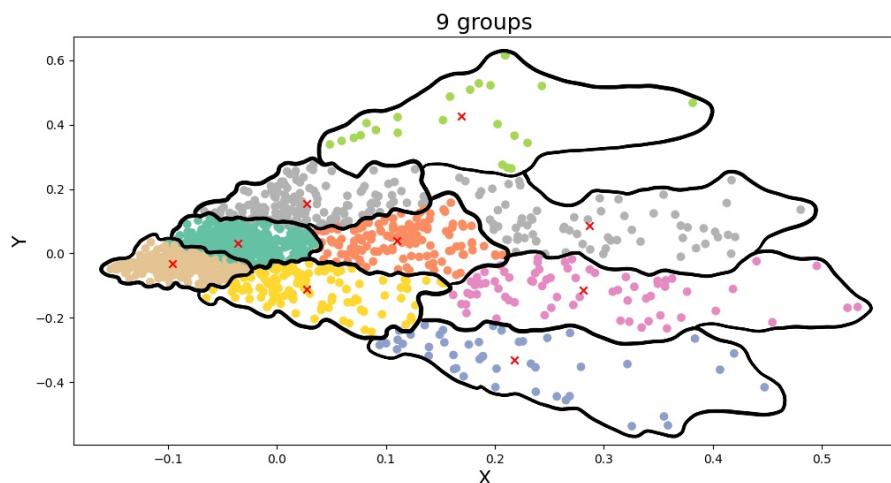
การจัดกลุ่มของข้อมูลหรือการทำ Text clustering นั้นโครงงานนี้จะใช้เทคนิค K-Means มาใช้ในการจัดกลุ่มโดยเลือกจำนวนกลุ่มจากการทำ Elbow method เพื่อหาจำนวนกลุ่มที่ดีที่สุดได้ดังภาพที่ 42 โดยจำนวนกลุ่มที่เหมาะสมจะอยู่ที่บริเวณส่วนโคงคล้ายข้อศอกในที่นี่จะประมาณกลุ่มได้ 6-9 กลุ่ม



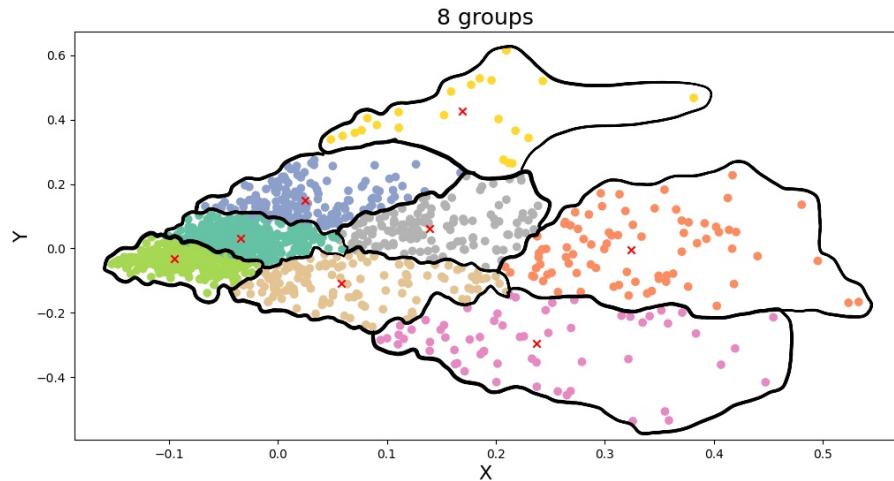
ภาพที่ 42 การทำ Elbow method

จากนั้นทำการทดลองจัดกลุ่มข้อมูลด้วยจำนวนกลุ่มที่แตกต่างกันแต่ข้อมูลเดียวกัน ทำการทดลองจัดกลุ่มด้วย 4 กรณีได้ดังนี้

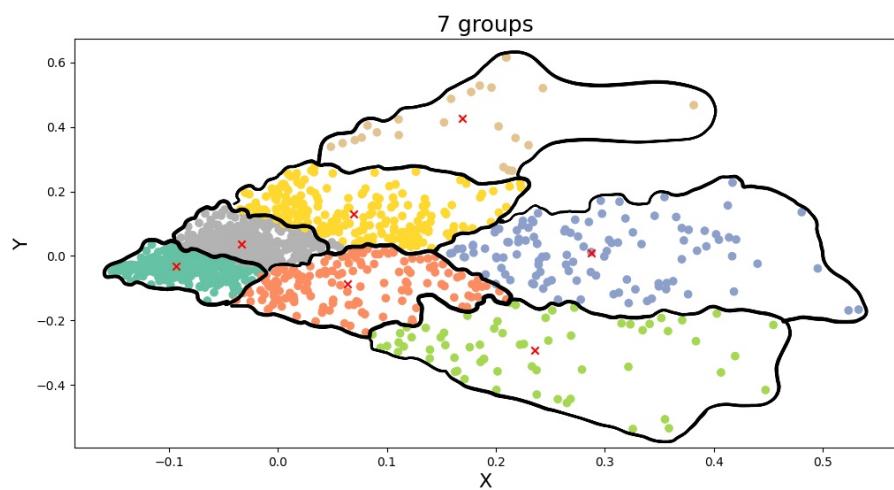
- จำนวน 9 กลุ่ม พบร่วงขบวนเขตของข้อมูลค่อนข้างแคบมากและมีกลุ่มที่มีเนื้อหาซ้ำกันมากกว่า 1 กลุ่ม
- จำนวน 8 กลุ่ม พบร่วงขบวนเขตของข้อมูลค่อนข้างแคบและเนื้อหาบางส่วนจะปนอยู่ในกลุ่มอื่น ๆ ที่ไม่ใกล้กัน
- จำนวน 7 กลุ่ม พบร่วงขบวนเขตของข้อมูลค่อนข้างดีและเนื้อหาในกลุ่มนั้นมีปะปนกันน้อยมากและไม่มีกลุ่มที่มีเนื้อหาซ้ำกัน
- จำนวน 6 กลุ่ม พบร่วงขบวนเขตของข้อมูลกว้างมากและทำให้ใน 1 กลุ่มนั้นมีเนื้อหาที่มากกว่า 1 อย่างทำให้ไม่สามารถระบุแน่ชัดได้ว่าเกี่ยวกับเรื่องใดเป็นหลัก



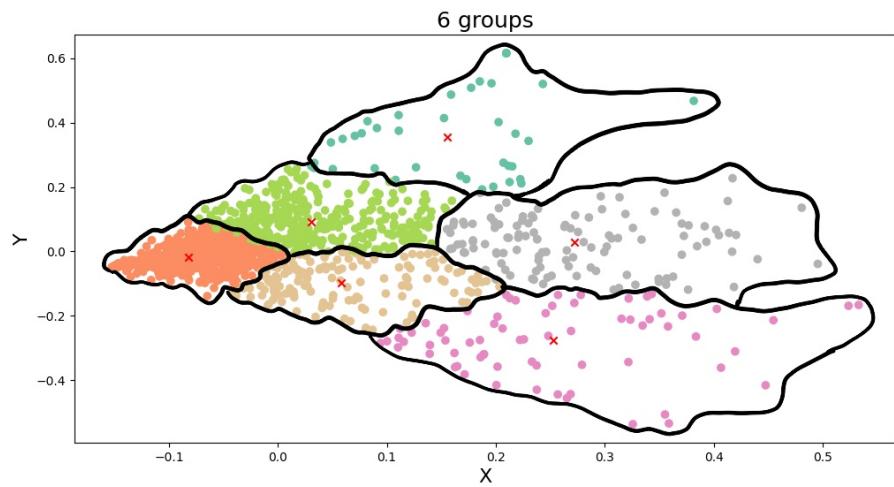
ภาพที่ 43 จัดกลุ่มข้อมูลจำนวน 9 กลุ่ม



ภาพที่ 44 จัดกลุ่มข้อมูลจำนวน 8 กลุ่ม



ภาพที่ 45 จัดกลุ่มข้อมูลจำนวน 7 กลุ่ม



ภาพที่ 46 จัดกลุ่มข้อมูลจำนวน 6 กลุ่ม

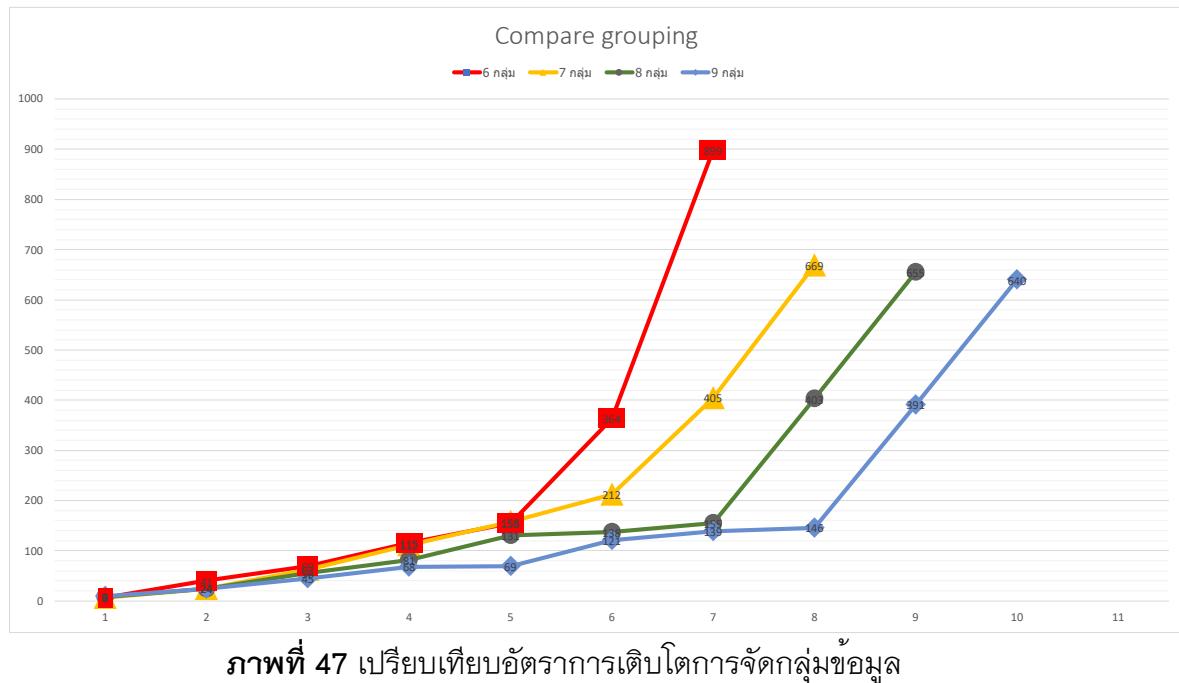
ทำการทดลองนับจำนวนรายการบริษัทแต่ละประเภทในแต่ละกลุ่มเพื่อคิดค่าเฉลี่ยและ
การกระจายตัวของข้อมูลบริษัท

ตารางที่ 17 แสดงการนับจำนวนบริษัทแต่ละประเภทในการจัดกลุ่มทั้งหมด 1,643 รายการ

จำนวนกลุ่ม	ประเภทของบริษัท	ค่าเฉลี่ย
6	<ul style="list-style-type: none"> - กลุ่ม 1 มี 115 รายการ - กลุ่ม 2 มี 899 รายการ - กลุ่ม 3 มี 69 รายการ - กลุ่ม 4 มี 155 รายการ - กลุ่ม 5 มี 41 รายการ - กลุ่ม 6 มี 364 รายการ 	469.43 รายการ
7	<ul style="list-style-type: none"> - กลุ่ม 1 มี 405 รายการ - กลุ่ม 2 มี 24 รายการ - กลุ่ม 3 มี 112 รายการ - กลุ่ม 4 มี 158 รายการ - กลุ่ม 5 มี 669 รายการ - กลุ่ม 6 มี 63 รายการ - กลุ่ม 7 มี 212 รายการ 	234.71 รายการ
8	<ul style="list-style-type: none"> - กลุ่ม 1 มี 403 รายการ - กลุ่ม 2 มี 81 รายการ - กลุ่ม 3 มี 655 รายการ - กลุ่ม 4 มี 155 รายการ - กลุ่ม 5 มี 138 รายการ - กลุ่ม 6 มี 24 รายการ - กลุ่ม 7 มี 131 รายการ - กลุ่ม 8 มี 56 รายการ 	205.38 รายการ
9	<ul style="list-style-type: none"> - กลุ่ม 1 มี 391 รายการ - กลุ่ม 2 มี 69 รายการ - กลุ่ม 3 มี 640 รายการ - กลุ่ม 4 มี 68 รายการ - กลุ่ม 5 มี 45 รายการ - กลุ่ม 6 มี 121 รายการ - กลุ่ม 7 มี 139 รายการ - กลุ่ม 8 มี 146 รายการ 	182.56 รายการ

ตารางที่ 17 (ต่อ)

จำนวนกลุ่ม	ประเภทของบริษัท	ค่าเฉลี่ย
9	- กลุ่ม 1 มี 115 รายการ	469.43 รายการ



จากการที่ 43–46 จะเห็นได้ว่าการกระจายตัวของข้อมูลนั้นมีความแตกต่างกันสังเกตได้จากสีที่ระบุตำแหน่งของกลุ่มในแต่ละภาพโดยเฉพาะเมื่อเปรียบเทียบการแบ่งกลุ่มที่ 6 และ 8 จะเห็นว่าขอบเขตของข้อมูลของการแบ่ง 6 กลุ่มนั้นมีความกว้างมากและห่างไกลจากจุดกึ่งกลางข้อมูล (Centroids point) จึงประมาณได้ว่าการแบ่งกลุ่มที่จำนวน 6 กลุ่มนั้นอาจไม่ได้ประสิทธิภาพความแม่นยำมากพอ และในตรงกันข้ามการแบ่งกลุ่มข้อมูลที่ 9 กลุ่มนั้นจะเห็นได้ว่าขอบเขตของข้อมูลนั้นแคบมากกันถึงทับซ้อนกันในแต่ละกลุ่มถึงแม้ขอบเขตของข้อมูลจะอยู่ใกล้ๆ กันก็ตาม ดังนั้นผู้วิจัยจึงตัดตัวเลือกการแบ่งกลุ่มข้อมูลที่ 6 กลุ่มออกเหลือเพียง 7 และ 9 กลุ่ม

และจากการ 47 จะเห็นได้ว่าการเติบโตของข้อมูลเมื่อนำมาวิเคราะห์ในแต่ละประเภทมาเรียงจากน้อยไปมากนั้น อัตราการเติบโตของการจัดกลุ่มข้อมูลที่ 7 กลุ่มนั้นมีอัตราการเติบโตที่คงที่มากที่สุด

จากนั้นทดลองทำการสูมเรียกข้อมูลในแต่ละกรณีออกมาเพื่อประกอบการตัดสินใจที่จะเลือกจำนวนกลุ่มของข้อมูล

จากการทดลองสูมเรียกข้อมูลหลาย ๆ ครั้งพบว่าจำนวนของการจัดกลุ่มที่มีค่าข้อมูลทับซ้อนกันน้อยที่สุดอยู่ที่ 7 กลุ่มทำให้ผู้วิจัยเลือกที่จะแบ่งกลุ่มข้อมูลที่ 7 กลุ่มแต่ทั้งนี้ก็ยังมีข้อมูลที่ทับซ้อนกันอยู่บ้างเล็กน้อยซึ่งอยู่ในระดับที่รับได้ และเมื่อทำการแบ่งกลุ่มข้อมูล

เรียบเรียงแล้วจึงสามารถบันทึกข้อมูลพร้อมกับ ID ของกลุ่มเพื่อนำไปนำเข้าฐานข้อมูลและใช้งานต่อไป

4.1.3 การกำหนดชื่อกลุ่ม

เมื่อได้ข้อมูลที่สมบูรณ์อยู่ในฐานข้อมูลแล้วนั้นการแสดงผลข้อมูลของกลุ่มจากหน้าเว็บไซต์จะเป็นตัวองมีการตั้งชื่อกลุ่มนี้องจากข้อมูลที่ได้จากการจัดกลุ่มคือ ID ซึ่งคือตัวเลขตั้งแต่ 0-6 เนื่องจากทำการกำหนดจำนวนกลุ่มไว้ที่ 7 กลุ่มดังนั้นเพื่อให้การแสดงผลในหน้าเว็บไซต์และให้การเรียกกลุ่มง่ายขึ้น จึงทำการตั้งชื่อกลุ่มโดยชื่อจะอยู่ในประเภทของเทคโนโลยีเนื่องจากข้อมูลที่มีนั้นเป็นบริษัทที่ทำเกี่ยวกับอุตสาหกรรมเทคโนโลยีหรือบริษัททางด้านโอดีการตั้งชื่อของกลุ่มข้อมูลนั้นได้ทำการอ้างอิงชื่อจากประเภทงานโอดีจากเว็บไซต์ th.jobsdb.com เป็นหลัก ซึ่งมีอยู่ 18 ประเภท รายชื่อประเภทงานโอดีที่มีในเว็บไซต์ th.jobsdb.com มีดังนี้

1. งาน Application Network
2. งาน Software
3. งาน Database
4. นักวิทยาศาสตร์ข้อมูล
5. งาน Hardware
6. งาน IT Audit
7. งานปรึกษาโอดี
8. งาน IT Project
9. งานดูแลระบบ SEO
10. งาน MIS
11. งาน Mobile งาน Wireless communications
12. งานดูแลระบบ Network
13. งานโปรแกรมเมอร์
14. งาน IT Security
15. งาน IT Support
16. งาน Software Tester
17. นักออกแบบ UI/UX
18. งานโอดีอื่น ๆ

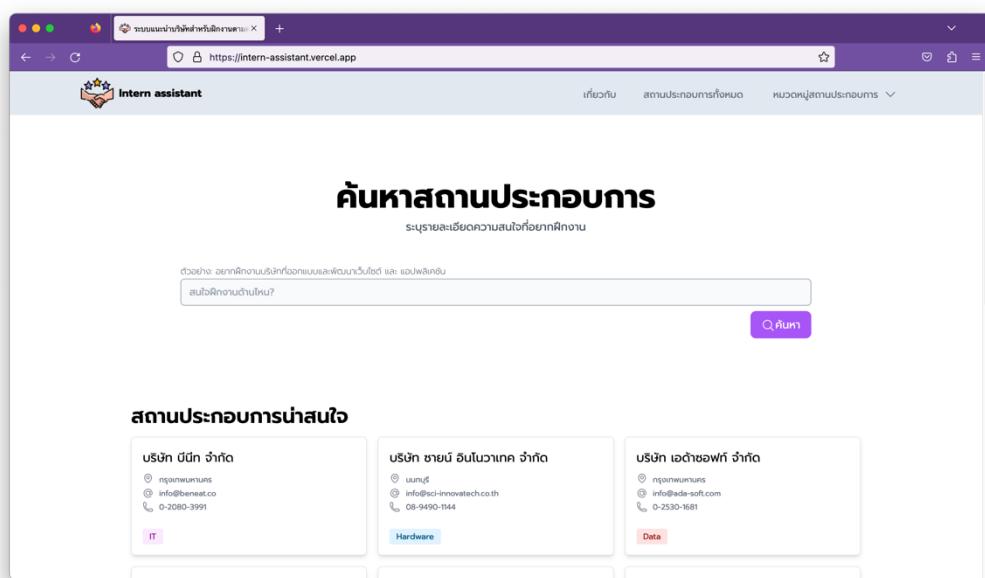
แต่ในโครงงานนี้มีกลุ่มข้อมูลเพียง 7 กลุ่มดังนั้นจึงต้องเลือกประเภทงานที่ตรงกับข้อมูลในกลุ่มมากที่สุดเท่านั้น โดยวิธีที่ใช้เลือกคือการสุ่มข้อมูลในแต่ละกลุ่มตั้งแต่ 0-6 และตรวจดูว่าควรจะได้ชื่อกลุ่มเป็นประเภทงานไหน

จากการทำการสุ่มเรียกข้อมูลดูทั้ง 7 กลุ่ม รายชื่อประเภทงานที่สามารถใช้ตั้งชื่อกลุ่มข้อมูลโดยอ้างอิงของเว็บไซต์ th.jobsdb.com มีดังนี้

1. Data analysis
2. Online marketing
3. Software
4. Hardware
5. Network
6. IT
7. Other

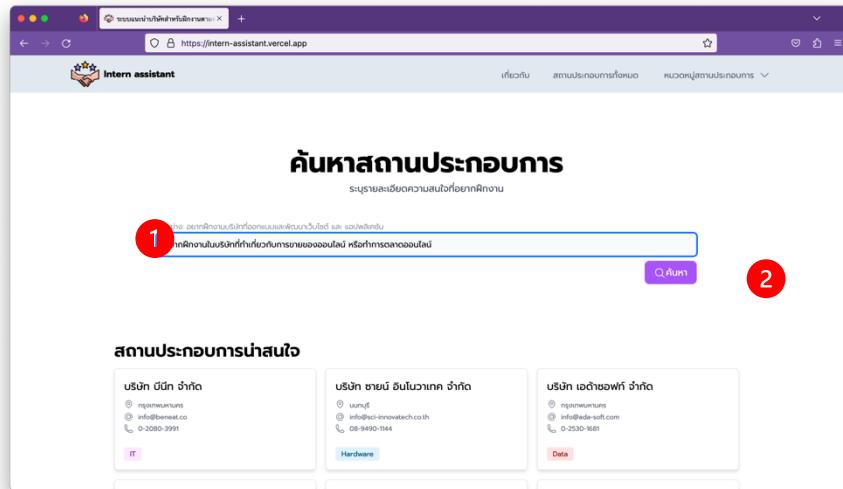
4.2 ขั้นตอนการใช้งานสำหรับผู้ใช้งาน

4.2.1 หน้าแรกเว็บไซต์ Intern-assistant



ภาพที่ 48 หน้าแรกเว็บไซต์ Intern-assistant

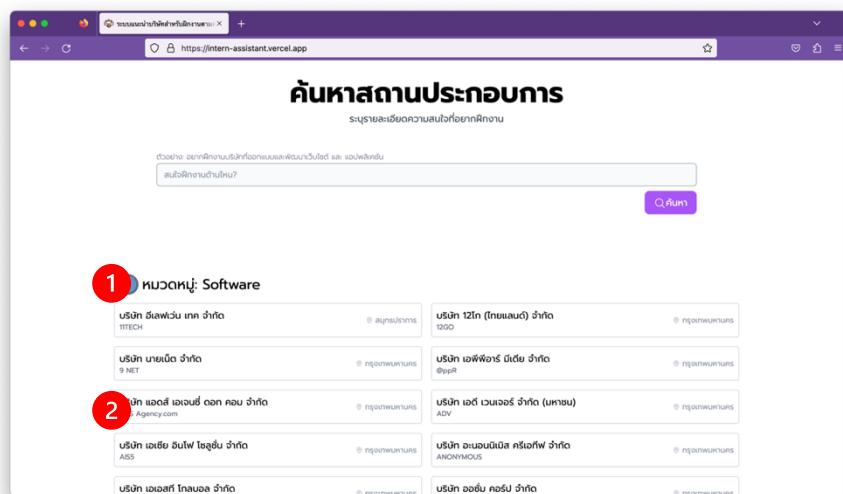
4.2.2 គណនោបរិម្ភ័



រាជធានី 49 គណនោបរិម្ភ័

ឈ្មោះលេខ 1 ពិម័យគណនោបរិម្ភ័ទាំងអស់ ដែលត្រូវបានបង្កើតឡើង និងបានបង្កើតឡើង ឈ្មោះលេខ 2 ក្នុងបញ្ជីការងារ

4.2.3 អនុការណ៍ផលលំដាប់ការគណនោ

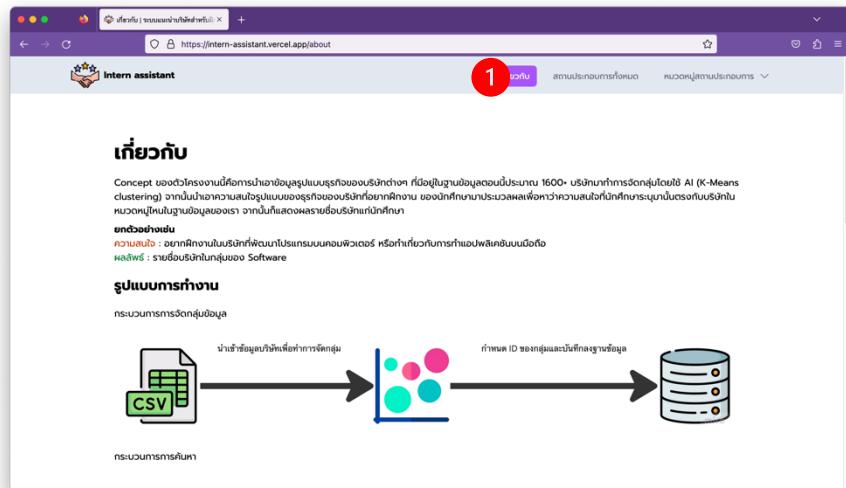


រាជធានី 50 អនុការណ៍ផលលំដាប់ការគណនោ

ឈ្មោះលេខ 1 ឱ្យចុះឈ្មោះក្រុមហ៊ុន

ឈ្មោះលេខ 2 រាយក្រឹងបរិម្ភ័ទាំងអស់ ដែលត្រូវបានបង្កើតឡើង

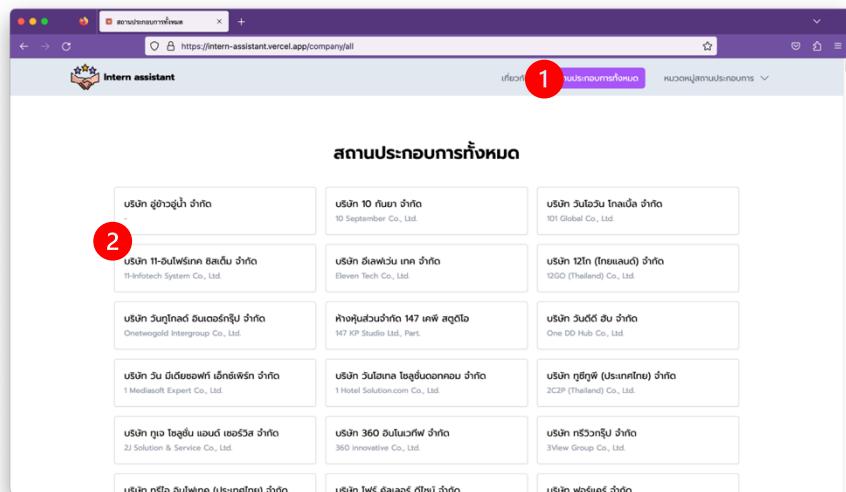
4.2.4 អនុការណ៍វក្សា



ภาพที่ 51 หน้าเกี่ยวกับ

หมายเหตุ 1 เมนูหน้าเกี่ยวกับแสดงข้อมูลเกี่ยวกับระบบและการทำงาน

4.2.5 หน้าแสดงรายชื่อบริษัททั้งหมด

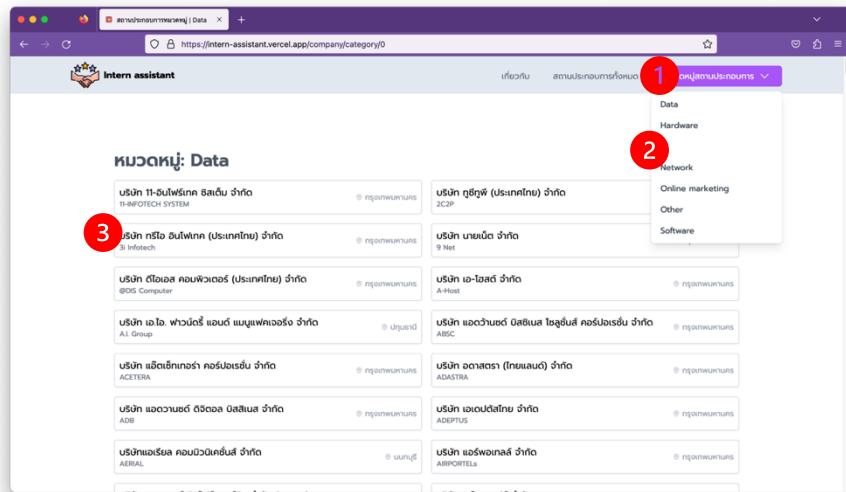


ภาพที่ 52 หน้าแสดงรายชื่อบริษัททั้งหมด

หมายเหตุ 1 เมนูหน้ารายชื่อบริษัททั้งหมด

หมายเหตุ 2 รายชื่อบริษัททั้งหมด

4.2.6 หน้าแสดงรายชื่อบริษัทในกลุ่มทั้งหมด



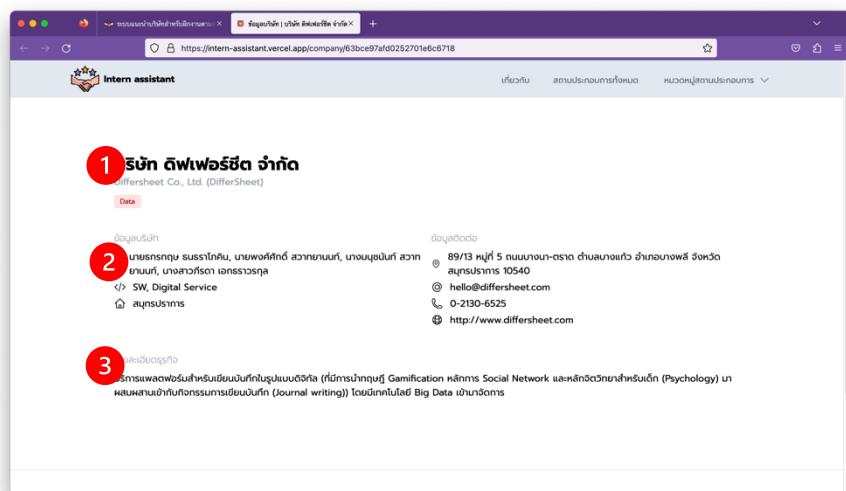
ภาพที่ 53 หน้าแสดงรายชื่อปริญญาในกลุ่มทั้งหมด

หมายเหตุ 1 เมนูรายชื่อคลิมบริษัท

หมายเลขอีก 2 รายซึ่งออกกล่าวบริษัท

หมายเลขอ ๓ รายชื่อบริษัทในกลุ่ม

4.2.7 หน้ารายละเอียดบริษัท



ภาพที่ 54 หน้าร้ายลักษณะอิมดบริษัท

4.3 การวัดค่าความคงถาวรของ

Cosine similarity เป็นเทคนิคที่นำมาใช้หาความคล้ายคลึงระหว่างความสนใจของผู้ใช้ และข้อมูลบริษัทที่อยู่ในฐานข้อมูลยิ่งค่า Cosine similarity เข้าใกล้ 1 แสดงว่าประโยชน์นั้นมีความคล้ายคลึงกับข้อมูลบริษัทในกลุ่มนั้นมากดังภาพที่ 55

```
onze@Tinngrits-MacBook-Pro:~/Desktop/final_project
..final_project (~zsh)  #1 ..final_project (~zsh)  #2
20% 4.4 GB develop + ~ /Desktop/final_project
python cosine similarity.py
Keyword: ออกแนวเรื่องไซต์ด้วย react js ท่าเกี่ยวกับการเชื่อมเว็บ การตลาดออนไลน์ด้วย และ SEO
cluster: 6
cosine similarity: 0.07871941760441518
~/Desktop/final_project develop !3 ??
7s 12:56:32
```

ภาพที่ 55 ตัวอย่างการคำนวณค่า Cosine similarity

4.3.1 การหากลุ่มที่มีความคล้ายคลึงมากที่สุด

1. คำนวณค่า Cosine similarity ทุกลุ่มจากความสนใจของผู้ใช้ผ่าน API ที่สร้างไว้เพื่อคำนวณโดยเฉพาะ

My Workspace

main testing / test cluster api

POST https://lamonze.tech/search

Body (JSON)

```
{
  "keyword": "ออกแบบเว็บไซต์ด้วย react และ angular"
}
```

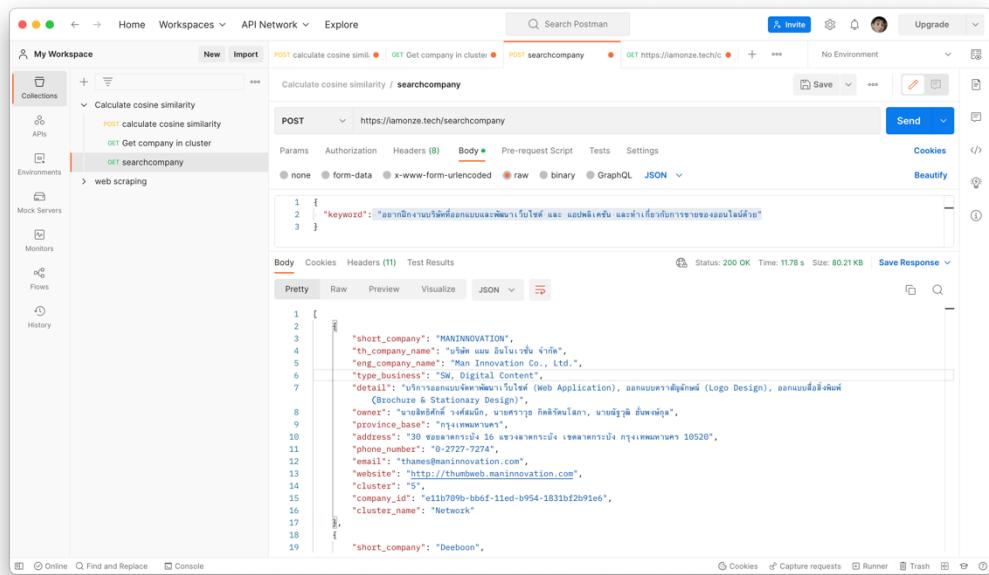
Status: 200 OK Time: 11.85 s Size: 814 B

```

1   "cosine_similarity": [
2     0.03592689333934359,
3     0.0,
4     0.16852858937859051,
5     0.0388431775312984,
6     0.027678186645230848,
7     0.3397083961645713,
8     0.3397083961645713,
9     0.0400188240297087
10    1,
11    "max_cosine_similarity": 0.3397083961645713,
12    "cluster": 5
13  ]

```

ภาพที่ 56 ตัวอย่างการคำนวณค่า Cosine similarity ผ่าน API และคืนค่าความคล้ายคลึง



ภาพที่ 57 ตัวอย่างการคำนวณค่า Cosine similarity ผ่าน API และศึกษาเป็นข้อมูลบริษัทที่อยู่ในกลุ่มที่คล้ายที่สุด

บทที่ 5

สรุปผลการดำเนินงาน

การพัฒนาระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) ในครั้งนี้สามารถสรุปการดำเนินงาน ปัญหาที่เกิดขึ้นระหว่างการดำเนินงาน ข้อเสนอและแนวทางพัฒนาต่อไปดังนี้

5.1 สรุปผลการดำเนินงาน

จากการดำเนินงานได้นำข้อมูลบริษัทจากสมาคมปัญญาประดิษฐ์แห่งประเทศไทยมาทำการจัดกลุ่มโดยใช้เทคนิคการจัดกลุ่มข้อมูล (K-Means) และใช้เทคนิคการคำนวณความคล้ายคลึง (Cosine similarity) ในการหาความคล้ายคลึงของความสนใจรูปแบบงานของผู้ใช้กับข้อมูลบริษัทได้ผลลัพธ์ดังนี้

การจัดกลุ่มข้อมูลทั้งหมดจำนวน 4 กลุ่มได้แก่ การจัดกลุ่มข้อมูลที่ 6 7 8 และ 9 โดยผลลัพธ์ดังนี้ กรณีแบ่งกลุ่มข้อมูลที่ 6 กลุ่มนั้นมีอัตราเรียกข้อมูลต่ำและพบว่าข้อบ阙ข้อมูลนั้นกร้างเกินไปและมีข้อมูลทับซ้อนกันจำนวนมาก กรณีแบ่งข้อมูลที่ 7 กลุ่มพบว่ามีอัตราเรียกดูข้อมูล ข้อมูลมีความทับซ้อนกันน้อยมากและข้อบ阙ของข้อมูลก็อยู่ในระดับที่เหมาะสมยอมรับได้ กรณีแบ่งกลุ่มที่ 8 กลุ่มพบว่ามีความคล้ายเดียวกับการแบ่งกลุ่มที่ 7 กลุ่มแต่ข้อบ阙ของข้อมูลบางกลุ่มนั้นแคบเกินไปทำให้มีเนื้อหาที่ซ้ำกันกับกลุ่มอื่น และกรณีแบ่งกลุ่มที่ 9 กลุ่มพบว่าข้อบ阙ของข้อมูลนั้นแคบที่สุดและแต่ละกลุ่มนั้นมีความทับซ้อนกันค่อนข้างมากจึงเกิดกลุ่มที่มีเนื้อหาแบบเดียวกันแต่อยู่คนละกลุ่ม

ซึ่งจะเห็นได้ว่ากรณีการแบ่งกลุ่มที่ 6 และ 9 กลุ่มนั้นข้อบ阙ของเนื้อหานั้นอยู่ในระดับที่ไม่ค่อยดีนักเมื่อเทียบกับการแบ่งกลุ่มที่ 7 และ 8 กลุ่ม และในการทดลองวิธี Elbow method จุดที่อยู่ตรงกันอยู่ระหว่างจุดที่ 7-8 ผู้วิจัยจึงเลือกเปรียบเทียบกันและหาข้อสรุปได้ว่าเลือกแบ่งกลุ่มที่ 7 กลุ่มเป็นการแบ่งกลุ่มข้อมูลที่มีประสิทธิภาพมากที่สุดจากการสูงเรียกดูข้อมูลของแต่ละกรณีเพื่อมาใช้ในการพัฒนาระบบ

ในการหาความคล้ายคลึงของข้อมูลด้วยวิธี Cosine similarity นั้นผลลัพธ์การคำนวณและการวิเคราะห์เพื่อหากลุ่มที่เหมาะสมกับความสนใจของผู้ใช้นั้นพบว่าเมื่อได้กลุ่มจากที่การคำนวณแล้วนั้นรูปแบบธุรกิจค่อนข้างตรงกับความสนใจที่ผู้ใช้ส่วนมาก

5.1.1 จุดเด่นของระบบ

1. ขั้นตอนการใช้งานของผู้ใช้นั้นถูกออกแบบมาให้ใช้งานง่ายและสะดวก รวดเร็ว
2. ออกแบบหน้าจอแสดงผล (User interface) เข้าใจง่ายและใช้งานได้สะดวก ไม่ซับซ้อน สามารถรองรับได้ทุกอุปกรณ์
3. ระบบสามารถเข้าถึงง่ายเนื่องจากพัฒนาอยู่ในรูปแบบของ Web application ทำให้ไม่ต้องติดตั้งก่อนใช้งานสามารถใช้งานผ่าน Browser ได้ในทุกอุปกรณ์

5.2 สรุปปัญหาที่เกิดระหว่างการดำเนินงาน

5.2.1 ในกรณีที่จะเพิ่มข้อมูลบริษัทลงในฐานข้อมูลเพิ่มเติมจำเป็นต้องทำการ Word segmentation ข้อมูลใหม่ที่จะเข้ามาร่วมกับข้อมูลเดิมที่มีอยู่แล้วดังนั้นข้อมูลของแต่ละบริษัทก็จะถูกเปลี่ยนกลุ่มไปทุกครั้งทั้งมีการเพิ่มข้อมูลใหม่

5.2.2 เมื่อทำการ Clustering ข้อมูลใหม่แล้วต้องทำการตั้งชื่อให้กับกลุ่มข้อมูลใหม่ เพราะเมื่อมีข้อมูลที่เปลี่ยนไปเนื้อหาบริษัทในกลุ่มเดิมก็อาจเปลี่ยนไปยกตัวอย่างเช่น กลุ่มที่ 0 เดิมเป็นกลุ่มของ Network แต่เมื่อมีการเพิ่มข้อมูลใหม่และทำการ Clustering ใหม่กลุ่ม 0 ก็อาจจะกลายเป็น Data เพราะเนื้อหาในกลุ่มนั้นเปลี่ยนไป หรืออาจมีกลุ่มเพิ่มเติมขึ้นมา นอกจากเหนือจากปัจจุบัน

5.2.3 ข้อมูลที่ได้รับมาเมื่อมีคำที่สะกดผิดจนไม่สามารถแก้ไขได้ด้วยผู้พัฒนาเอง การทำ Clustering นั้นอาจจะไม่ได้แยกแยะข้อมูลได้ดียกตัวอย่างเช่น ประยุคที่มีคำว่า “แอปพลิเคชัน” ที่สะกดถูกต้องอาจจะอยู่คนละกลุ่มกับประยุคที่มีคำว่า “แอปพลิเคชั่น” ที่มีการสะกดผิด

5.3 แนวทางพัฒนาระบบในอนาคต

5.3.1 พัฒนาความแม่นยำในการจัดกลุ่มข้อมูล

5.3.2 พัฒนาให้สามารถแนะนำตำแหน่งงานในบริษัทได้

5.3.3 เพิ่มชุดข้อมูลให้มากขึ้นเพื่อเพิ่มความแม่นยำของเว็บไซต์

5.3.4 พัฒนาความเร็วของอัลกอริทึมในการคำนวณความคล้ายคลึง

5.3.5 พัฒนาชั้นตอนการเรียกดูข้อมูลให้ง่ายขึ้น

5.3.6 พัฒนาให้รองรับการกรองตัวเลือกที่จะค้นหา

5.3.7 พัฒนาเว็บไซต์ให้มีความปลอดภัยมากขึ้น

5.4 แบบประเมินความพึงพอใจของผู้ใช้

จากการประเมินความพึงพอใจของผู้ใช้ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) (Computer Internship Recommendation System With K-Means Clustering) สามารถสรุปผลได้ดังนี้

การกำหนดเกณฑ์การพิจารณา

เกณฑ์การพิจารณาระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) มีดังนี้

1. เกณฑ์การให้คะแนน ได้กำหนดเกณฑ์การให้คะแนนไว้ 5 ระดับดังนี้

5 คะแนน หมายถึง ระดับความพึงพอใจ/ความเข้าใจดีมาก

4 คะแนน หมายถึง ระดับความพึงพอใจ/ความเข้าใจดี

3 คะแนน หมายถึง ระดับความพึงพอใจ/ความเข้าใจปานกลาง

2 คะแนน หมายถึง ระดับความพึงพอใจ/ความเข้าใจน้อย

1 คะแนน หมายถึง ระดับความพึงพอใจ/ความเข้าใจน้อยมาก

2. เกณฑ์การแบ่งช่วงคะแนนค่าเฉลี่ย เกณฑ์การแบ่งช่วงคะแนนค่าเฉลี่ยได้กำหนดเกณฑ์การประเมินไว้ดังนี้

ค่าเฉลี่ย 4.51 – 5.00 หมายถึง ระดับความพึงพอใจ/ความเข้าใจดีมาก
 ค่าเฉลี่ย 3.51 – 4.50 หมายถึง ระดับความพึงพอใจ/ความเข้าใจดี
 ค่าเฉลี่ย 2.51 – 3.50 หมายถึง ระดับความพึงพอใจ/ความเข้าใจปานกลาง
 ค่าเฉลี่ย 1.51 – 2.50 หมายถึง ระดับความพึงพอใจ/ความเข้าใจน้อย
 ค่าเฉลี่ย 1.00 – 1.50 หมายถึง ระดับความพึงพอใจ/ความเข้าใจน้อยมาก

ตารางที่ 18 แสดงจำนวนค่าเฉลี่ยความพึงพอใจต่อระบบ

รายละเอียด	ระดับความพึงพอใจ/ความเข้าใจ							เกณฑ์การประเมิน
	5	4	3	2	1	ค่าเฉลี่ย		
มีการออกแบบหน้าจอสำหรับผู้ใช้งานอย่างเหมาะสม	7	6	3	1	0	4.12		ดี
วิธีการใช้งานง่ายต่อการทำความเข้าใจ	6	7	3	1	0	4.06		ดี
ประสิทธิภาพความเสถียรในการทำงานของเว็บแอปพลิเคชัน	2	9	6	0	0	3.76		ดี
ความเหมาะสมของรูปแบบของหน้าจอแจ้งเตือนต่างๆ	5	7	5	0	0	4		ดี
ความพึงพอใจในภาพรวม	5	9	3	0	0	4.12		ดี
ค่าเฉลี่ยรวม						4.01		ดี

จากตารางที่ 18 พบร้าผู้ใช้มีความพึงพอใจต่อระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) โดยภาพรวมอยู่ในเกณฑ์ประเมินที่มีค่าเฉลี่ยเท่ากับ 4.01

ขอเสนอแนะ

สามารถสรุปความพึงพอใจและข้อเสนอแนะของผู้ใช้ที่มีต่อ ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means) ได้ดังนี้

- ควรจะมีฟังก์ชันที่ให้เลือกจังหวัดว่าเราอยู่ภาคกลางของจังหวัดไหน เพราะบางคนอยากทำงานใกล้ๆ บ้าน
- อย่างให้มีการเลือกชนิดของภาคอย่างภาคเหนือหรือภาคอีสานแต่โดยรวมทำได้ดีแล้ว
- ต้องทำให้ผู้ใช้งาน ใช้งานได้ง่ายกว่าเดิม และการค้นหาบางอย่างก็ไม่ตรงกับความต้องการที่ค้นหาเท่าไหร่

เอกสารอ้างอิง

- ตาเยะ, ช. (2565). NLP คืออะไร. สีบคน 8 กุมภาพันธ์ 2566.
 แหล่งที่มา : <https://www.mindphp.com/คุณมี/73-คืออะไร/8859-nlp.html>
- D'Agostino, A. (2564). Text Clustering with TF-IDF in Python. สีบคน 11 กุมภาพันธ์ 2566
 แหล่งที่มา : <https://medium.com/mlearning-ai/text-clustering-with-tf-idf-in-python-c94cd26a31e7>
- Amazon. API คืออะไร. (2566). สีบคน 11 กุมภาพันธ์ 2565.
 แหล่งที่มา : <https://aws.amazon.com/th/what-is/api/>
- บทความ E-R Diagram คืออะไร. (2557). สีบคน 22 กุมภาพันธ์ 2566
 แหล่งที่มา : <https://www.9experttraining.com/articles/บทความ-e-r-diagram-คืออะไร>
- Mindphp. (2566). สีบคน 21 กุมภาพันธ์ 2566
 แหล่งที่มา : <https://www.mindphp.com/บทความ/31-ความรู้ทั่วไป/6870-use-case-diagram.html>
- mindphp. การวิเคราะห์ระบบและการออกแบบ System Analysis and Design (ซีสเต็ม อนาคต ดีไซน์). (2022). สีบคน 21 กุมภาพันธ์ 2566
 แหล่งที่มา: <https://www.mindphp.com/ บทความ/31-ความรู้ทั่วไป/4084-system-analysis-and-design.html>
- mindphp. (2566). NumPy คืออะไร. สีบคน 18 กุมภาพันธ์ 2566.
 แหล่งที่มา: <https://www.mindphp.com/บทเรียนออนไลน์/83-python/8492-what-is-the-numpy.html>
- marcuscode. (2564). ทำความรู้จักกับ Node.js. สีบคน 17 กุมภาพันธ์ 2566
 แหล่งที่มา: <http://marcuscode.com/tutorials/nodejs/introducing-nodejs>
- nipa. (2564). Cloudflare คืออะไร จะเข้ามาช่วยองค์กรของคุณได้อย่างไร?.
 สีบคน 11 กุมภาพันธ์ 2566 แหล่งที่มา : <https://web.nipa.cloud/how-cloudflare-protect-your-corporate>
- Warakorn Pradiskul, (2564). Recommender System Using Collaborative Filtering A Case Study of Toyota Buzz Company Limited, หน้า 11-21.
- Thongchai Klayklueng, (2562). เทคนิคการตัดเลือกกลุ่มให้ตรงรายอาชารสำหรับ รองรับ แผนการติดตั้งระบบผลิตไฟฟ้าพลังงานแสงอาทิตย์บนหลังคาเพื่อเพิ่มค่าครรชนี ประสิทธิภาพการใช้พลังงานไฟฟ้า Load Clustering Technique Application to PV Solar Rooftop Installation Planning for Improving Energy Efficiency, หน้า 134-148.
- ปราณีย์ พึงวิชา, (2562). ศึกษาการแบ่งกลุ่มพฤติกรรมของผู้บริโภคที่ซื้อเครื่องประดับผ่าน เครื่อขายสัมมออนไลน์ Clustering of Jewellery Purchasing Behaviour through Social Network, หน้า 213-224.

เอกสารอ้างอิง (ต่อ)

- จักรินทร์ สันติรัตนภักดี, ศ. น. (2564). การออกแบบและพัฒนาระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะเพื่อติดแท็กปัญหาการให้บริการ, หน้า 77–90.
- วุฒิชัย, (2556). การเปรียบเทียบวิธีการแบ่งแยกคำภาษาไทยด้วยโครงสร้างการเขียนกับโครงสร้างพยางค์ The Comparison of Thai Word Segmentation with Thai Writing Structures and Syllable Structures, หน้า 504–509.
- เจย์. (2564). A Beginner's Guide to Scikit-learn. สืบค้น 18 กุมภาพันธ์ 2566.
แหล่งที่มา: <https://hashdork.com/th/scikit-learn/>
- mindphp. (2565). การใช้งานต่างๆ ใน PyThaiNLP. สืบค้น 18 กุมภาพันธ์ 2566
แหล่งที่มา: <https://www.mindphp.com/บทความ/it-news/8778-การใช้งานต่างๆ ใน-pythainlp.html>
- JUNG. (2564). พื้นฐาน Python และ Numpy สำหรับ Deep Learning. สืบค้น 18 กุมภาพันธ์ 2566 แหล่งที่มา: <https://www.kaggle.com/code/ratthachat/python-numpy-deep-learning#Numpy>
- frevation. (2564). Next js. สืบค้น 17 กุมภาพันธ์ 2566.
แหล่งที่มา: <https://www.frevation.com/blog/web-development/next-js/>
- CloudHM. (2564). บริการของ AWS มีจุดเด่นและนำไปใช้ประโยชน์ในด้านใดได้บ้าง.
สืบค้น 11 กุมภาพันธ์ 2566. แหล่งที่มา: <https://blog.cloudhm.co.th/what-is-and-what-business-need-aws/>
- Chakrit. (2562). similarity – ความเหมือนที่แตกต่าง. สืบค้น 11 กุมภาพันธ์ 2566.
แหล่งที่มา: <https://www.softnix.co.th/2019/05/29/similarity–ความเหมือนที่แตกต่าง/>
- DIGI. (2564). รู้จัก Clustering Model คืออะไร. สืบค้น 11 กุมภาพันธ์ 2566.
แหล่งที่มา : <https://digi.data.go.th/blog/what-is-clustering-model-and-example/>
- Chakrit. (2561). ว่าด้วย k-means – และการประยุกตร. สืบค้น 11 กุมภาพันธ์ 2566.
แหล่งที่มา : <https://www.softnix.co.th/2018/09/06/ว่าด้วย-k-means–และการประยุกตร/>
- CHAKRIT. (2562). TF-IDF ทำงานยังไง. สืบค้น 11 กุมภาพันธ์ 2566.
แหล่งที่มา : <https://www.softnix.co.th/2019/05/28/tf-idf–ทำงานยังไง/>
- Paul. (2564). K-Means Clustering with Elbow Method. สืบค้น 11 กุมภาพันธ์ 2566.
แหล่งที่มา : <https://medium.com/kbtg-life/k-means-clustering-with-elbow-method-8d02b35aaa2e>
- Rungnapha, K. (2561). Sequence Diagram แผนผังการทำงานแบบลำดับปฏิสัมพันธ์.
สืบค้น 22 กุมภาพันธ์ 2566.
แหล่งที่มา : <https://www.gurugeek.com/education/sequencediagram/>
- Surapong, K. (2563). PyThaiNLP คืออะไร Tutorial สอนใช้งาน PyThaiNLP Library NLP ภาษาไทย สำหรับ Python เป็นต้น – PyThaiNLP ep.1. สืบค้น 18 กุมภาพันธ์ 2566.
แหล่งที่มา: <https://www.bualabs.com/archives/3234/what-is-pythainlp-tutorial-teach-basic-how-to-use-pythainlp-library-nlp-in-python-pythainlp-ep-1/>

เอกสารอ้างอิง (ต่อ)

- Panchart, M. (2564). DATA รู้จัก pandas – Library อันดับ 1 สำหรับการทำ Data Analysis. สืบค้น 18 กุมภาพันธ์ 2566. แหล่งที่มา : <https://blog.skooldio.com/what-is-pandas>
- Pallop, C. (2560). Next.js คืออะไร?. สืบค้น 17 กุมภาพันธ์ 2566. แหล่งที่มา: <https://medium.com/hamcompe/next-js-คืออะไร-8fbb36e68b0>
- PLC, V. M. (2565). เขียนและจัดการข้อมูลได้ง่ายๆ ด้วย MongoDB. สืบค้น 17 กุมภาพันธ์ 2566. แหล่งที่มา: <https://www.proen.cloud/en/blogs/mongodb/>
- Chai, P. (2558). MongoDB คืออะไร? + สอนวิธีใช้งานเบื้องต้น. สืบค้น 17 กุมภาพันธ์ 2566. แหล่งที่มา : <https://devahoy.com/blog/2015/08/getting-started-with-mongodb>
- Natakorn, C. (2564). FastAPI คืออะไร และการใช้งานเบื้องต้น. สืบค้น 17 กุมภาพันธ์ 2566. แหล่งที่มา : <https://natakornch.medium.com/fastapi-คืออะไร-และการใช้งานเบื้องต้น-4f2d0fd91bcd>
- TAeng Trirong, P. (2560). Cross-Origin Resource Sharing (CORS) เป็นสิ่งที่ Web Developer ต้องควรรู้. สืบค้น 11 กุมภาพันธ์ 2566. แหล่งที่มา : <https://medium.com/nellika/cors-เป็นสิ่ง-ที่-web-developer-ต้องควรรู้-c906b1b47958>
- Supalerk, P. (2563). เมื่อสาย DATA อยากจะกิน Pizza (โดยใช้ Jaccard Similarity และ Cosine Similarity). สืบค้น 11 กุมภาพันธ์ 2566. แหล่งที่มา: <https://medium.com/data-cafe-thailand/เมื่อสาย-data-อยากจะกิน-pizza-โดยใช-jaccard-similarity-และ-cosine-similarity-f921fa4ab043>
- Weerasak, T. (2560). การหาจำนวน k ที่เหมาะสมที่สุดด้วยวิธี Elbow Method. สืบค้น 11 กุมภาพันธ์ 2566. แหล่งที่มา: <https://medium.com/espressofx-notebook/การหา-จำนวน-k-ที่เหมาะสมที่สุดด้วยวิธี-elbow-method-79b9a75f934>
- Patipan, P. (2563). สรุปใจความสำคัญของข้อความด้วยเทคนิคการประมวลผลทางภาษา เป้าหมาย: TF-IDF, Part 1. สืบค้น 11 กุมภาพันธ์ 2566. แหล่งที่มา : <https://bigdata.go.th/big-data-101/tf-idf-1/>
- L, M. (2562). NLP(Natural Language Processing) ศาสตร์(ไม่)ใหม่ ศาสตร์แห่งเจ้า: แยกประเภทอีเมลล์ด้วยพลังฟอร์ซ. สืบค้น 8 กุมภาพันธ์ 2566. แหล่งที่มา : <https://medium.com/mmp-li/nlp-natural-language-processing-ศาสตร์-ไม่-ใหม่-ศาสตร์แห่งเจ้า-แยกประเภทอีเมลล์ด้วยพลังฟอร์ซ-66b8bdff2e42>

ການພັນວົງ ດ
ຄູມື່ອກາຮຕິດຕັ້ງຮະບບ

ການພັນວາງ ກ ຄູ່ມືອກາຮຕິດຕັ້ງ

การเติร์ยมข้อมูลและติดตั้งโปรแกรมสำหรับการพัฒนาระบบแนะนำสถานที่ฝึกงาน
ด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเคลื่อน (K-Means)

1. การจัดกลุ่มข้อมูลด้วยเทคนิคทางปัญญาประดิษฐ์เคลื่อน (K-Means)

1.1 ทำการดาวน์โหลด Repository จาก https://github.com/slapexs/final_project ด้วยการใช้คำสั่ง `git clone https://github.com/slapexs/final_project` เพื่อดownloadไฟล์มาไว้ในเครื่อง

1.2 บันทึกไฟล์ข้อมูลบริษัทนามสกุล .csv ในโฟลเดอร์ data_csv

1.3 เปิดไฟล์ clustering.py ในโปรแกรม Text editor

ภาพที่ 58 โค้ดคำสั่งในไฟล์ clustering.py ใช้ในการจัดกลุ่มข้อมูล

1.4 ทำการกำหนดจำนวนของกลุ่มข้อมูลที่ต้องการลงในตัวแปร k

```

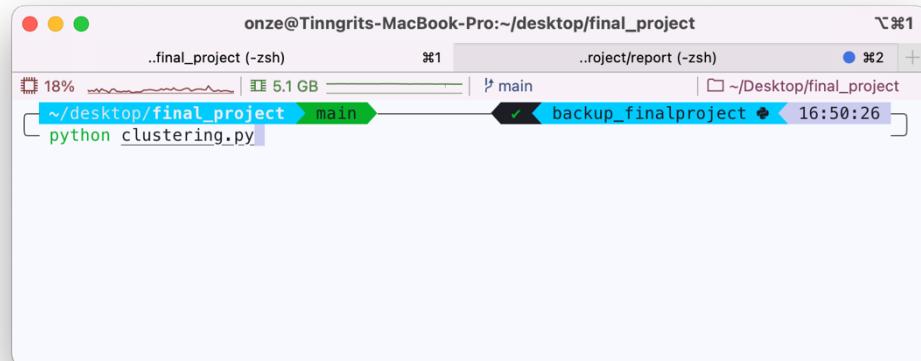
● ○ ●
1 k = 7
2 kmeans = KMeans(n_clusters=k, random_state=1)
3 # Fit model
4 kmeans.fit(df_tfidf[['x_value', 'y_value']])
5 clusters = kmeans.labels_

```

ภาพที่ 59 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k และการจัดกลุ่มข้อมูล

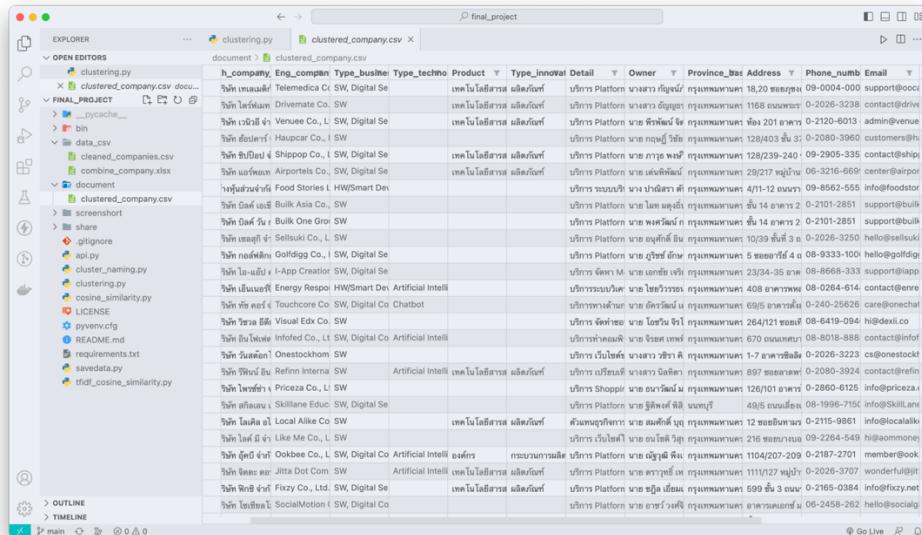
จากภาพที่ 59 แสดงการจัดกลุ่มข้อมูลด้วยเคมีน (K-Means) โดยการที่กำหนดจำนวนกลุ่มข้อมูลที่ต้องการลงในตัวแปร k จากนั้นทำการจัดกลุ่มข้อมูลด้วยฟังก์ชัน KMeans และคืนค่ากลับมา�ังตัวแปร kmeans และเก็บป้ายชื่อของกลุ่มลงในตัวแปร clusters

1.5 เรียกใช้ไฟล์เพื่อทำการจัดกลุ่มข้อมูล



ภาพที่ 60 แสดงการใช้งานคำสั่งจัดกลุ่มข้อมูลใน Terminal

1.6 ได้ไฟล์ clustered_company.csv ในโฟลเดอร์ document ที่เป็นผลลัพธ์การจัดกลุ่ม



The terminal window shows the command:

```
mongorestore --db cluster --collection company -f clustered_company.csv
```

Output from the command:

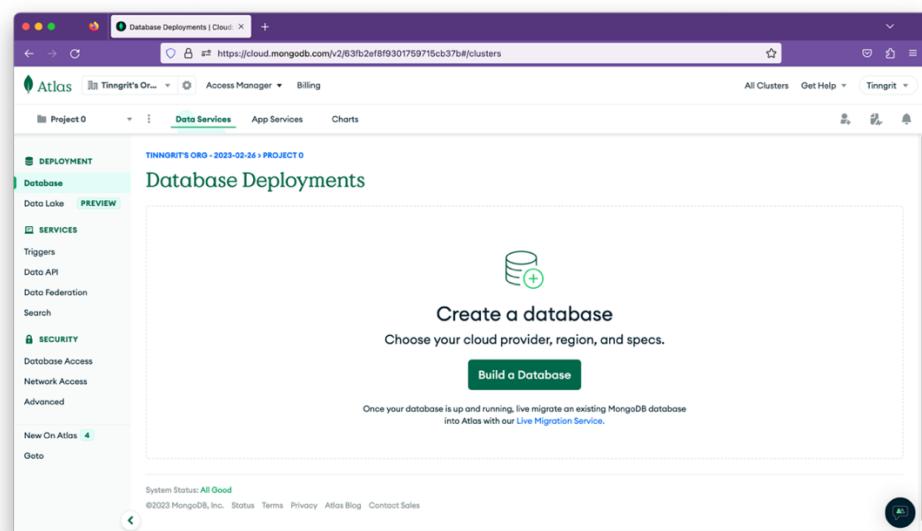
```
clustered_company.csv: 100 documents restored.
```

ภาพที่ 61 แสดงไฟล์ clustered_company.csv

2. การสร้างคลัสเตอร์ (Cluster) ของฐานข้อมูลมองโกลดีบี (MongoDB) บนเว็บไซต์

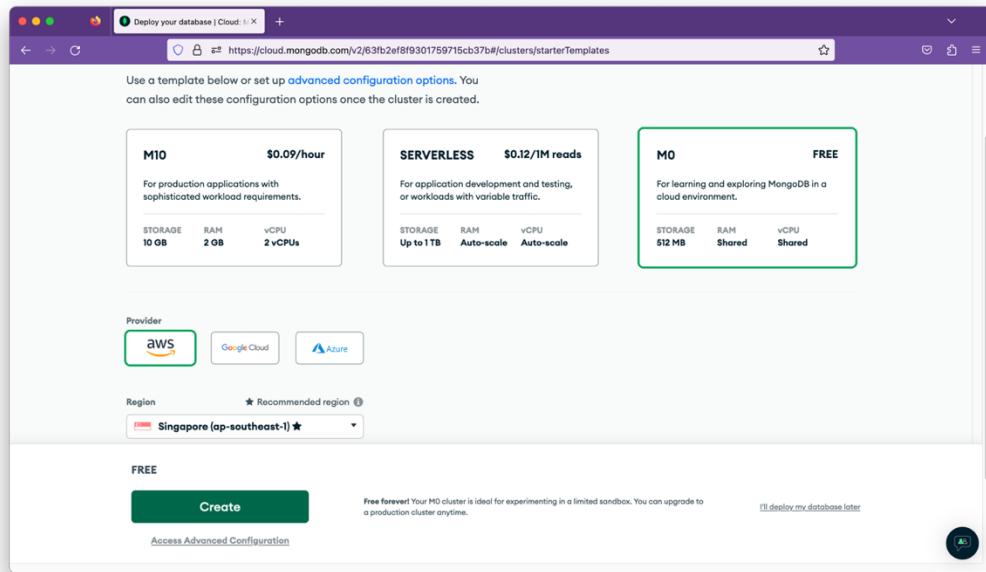
2.1 การใช้งาน Cloud MongoDB เข้าเว็บไซต์ <https://www.mongodb.com> จากนั้นเข้าสู่ระบบ

2.2 สร้าง Cluster โดยการกดปุ่ม Build a Database



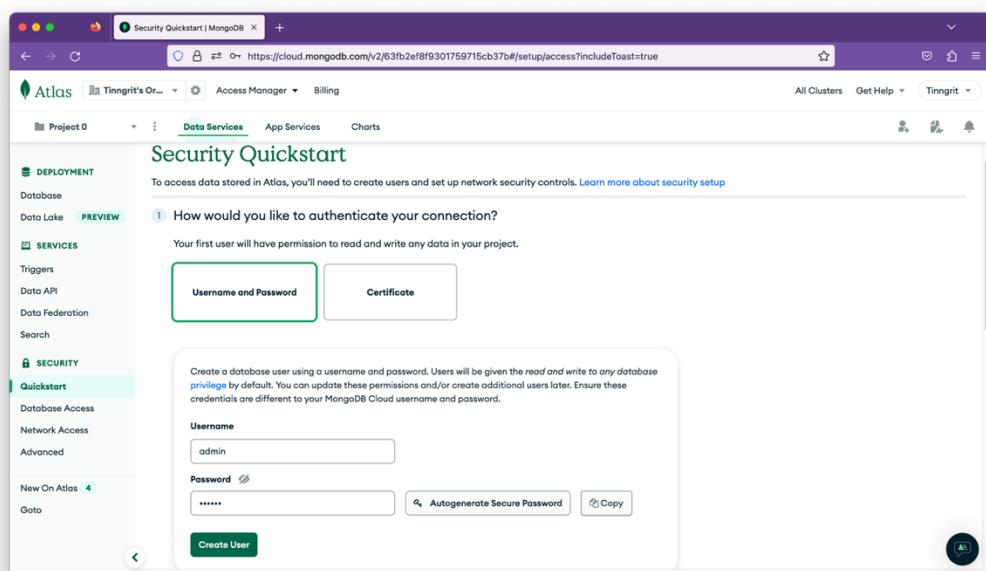
ภาพที่ 62 แสดงหน้าการจัดการ Cluster MongoDB

2.3 เลือกการตั้งค่าของ Cluster และตั้งชื่อ จากนั้นกดปุ่ม Create เพื่อสร้างฐานข้อมูล โดยที่เลือกผู้ให้บริการเป็น AWS และเลือกภูมิภาคเป็นประเทศไทย



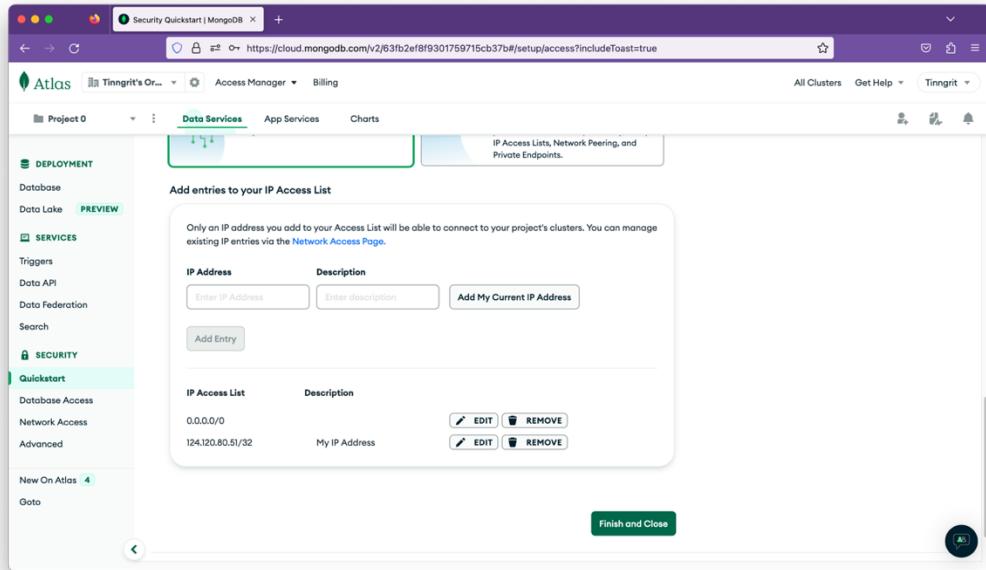
ภาพที่ 63 แสดงหน้าตั้งค่าและสร้าง Cluster MongoDB

2.4 สร้างบัญชีสำหรับใช้งานฐานข้อมูลตั้งค่า username และ password จากนั้นกดปุ่ม Create User



ภาพที่ 64 แสดงหน้าสร้างบัญชีสำหรับจัดการฐานข้อมูล

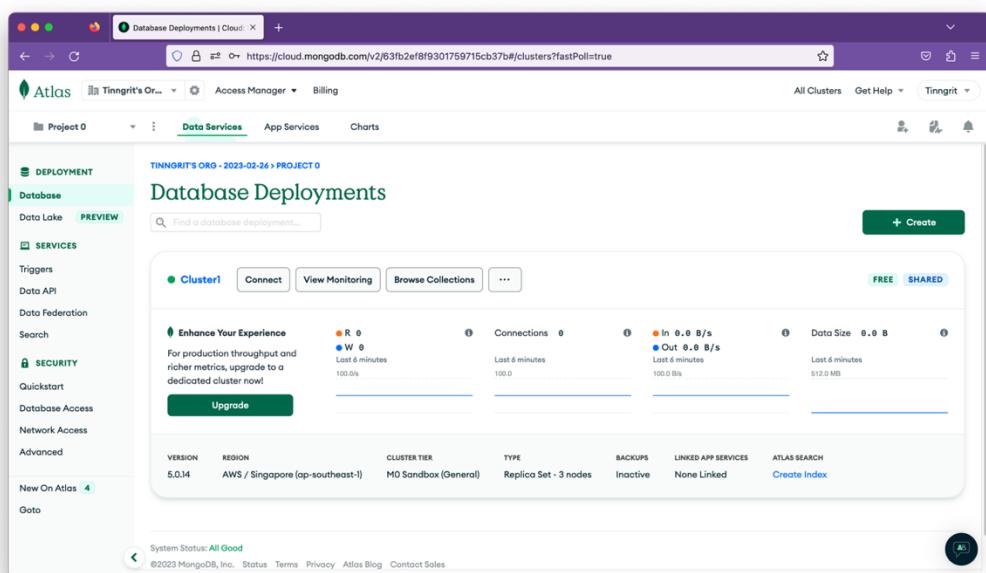
2.4 เพิ่มรายชื่อ IP address ที่สามารถเชื่อมต่อฐานข้อมูลได้ (กรณีเป็น 0.0.0.0 หมายถึงทุก IP address สามารถเชื่อมต่อเข้ามาได้) จากนั้นกดปุ่ม Add Entry และปุ่ม Finish and Close



ภาพที่ 65 แสดงหน้าเพิ่ม IP address ที่สามารถเชื่อมต่อฐานข้อมูลได้

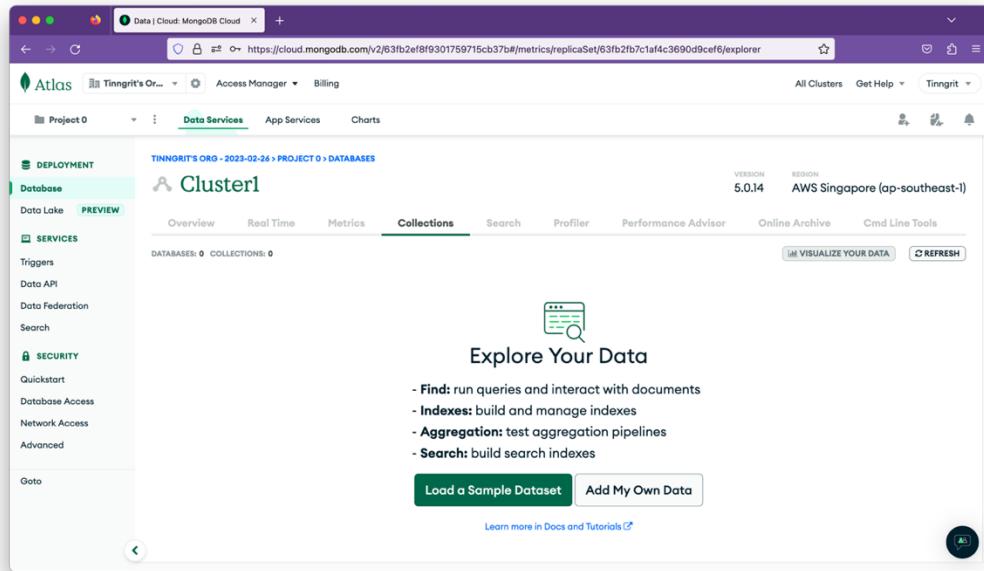
3. การสร้างฐานข้อมูลในโปรแกรม mongod (MongoDB)

3.1 สร้างฐานข้อมูลและ Collection กดที่ชื่อ Cluster



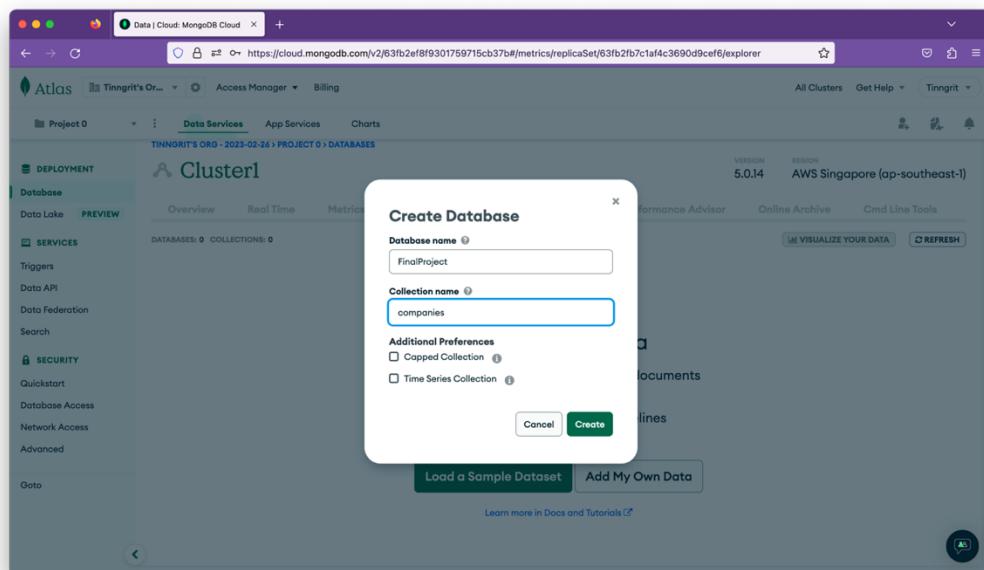
ภาพที่ 66 แสดงหน้าจัดการ Cluster MongoDB

3.2 กดที่เแคบเมนู Collections และกดปุ่ม Add My Own Data เพื่อสร้างฐานข้อมูล



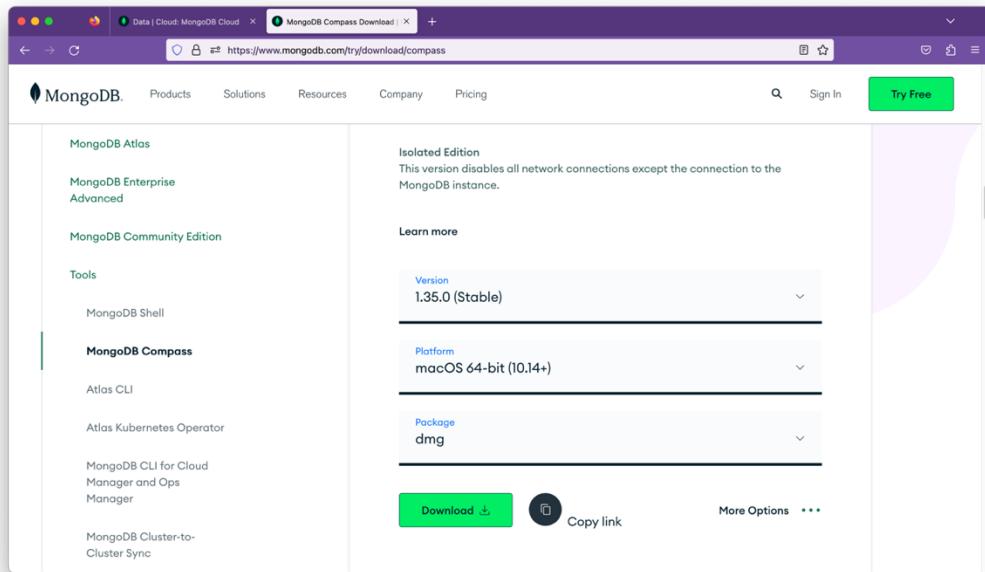
ภาพที่ 67 แสดงการสร้างฐานข้อมูล MongoDB

3.3 กำหนดชื่อฐานข้อมูลและชื่อ Collection ที่อยู่ในฐานข้อมูล



ภาพที่ 68 แสดงหน้าต่างการสร้างฐานข้อมูลและ Collection

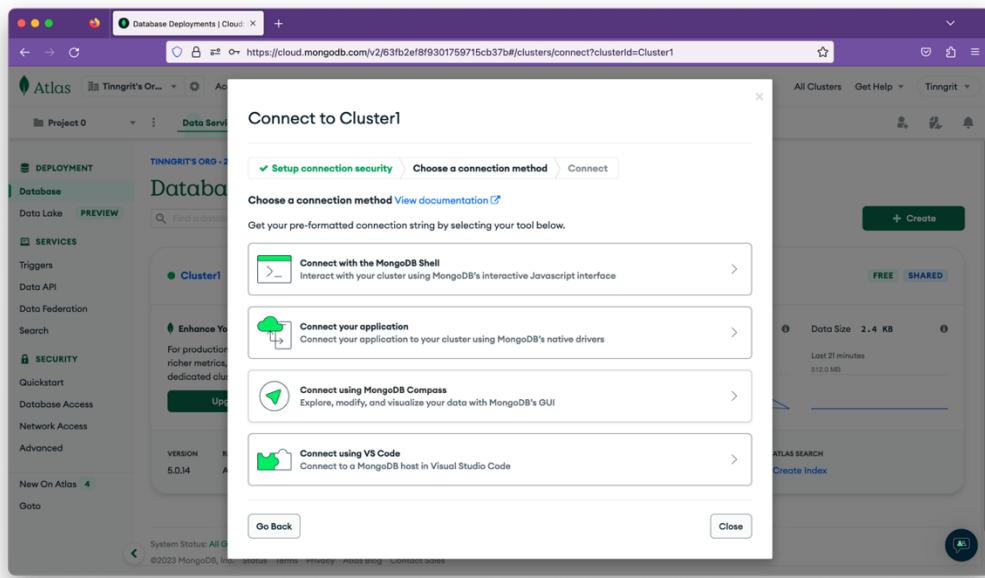
3.3 ดาวน์โหลดโปรแกรม MongoDB compass ที่เว็บไซต์ของ MongoDB และเลือก Version Platform และ Package ตามระบบปฏิบัติการที่ใช้และกดปุ่ม Download



ภาพที่ 69 ตัวอย่างการเลือกตั้งค่าการดาวน์โหลดโปรแกรม MongoDB compass

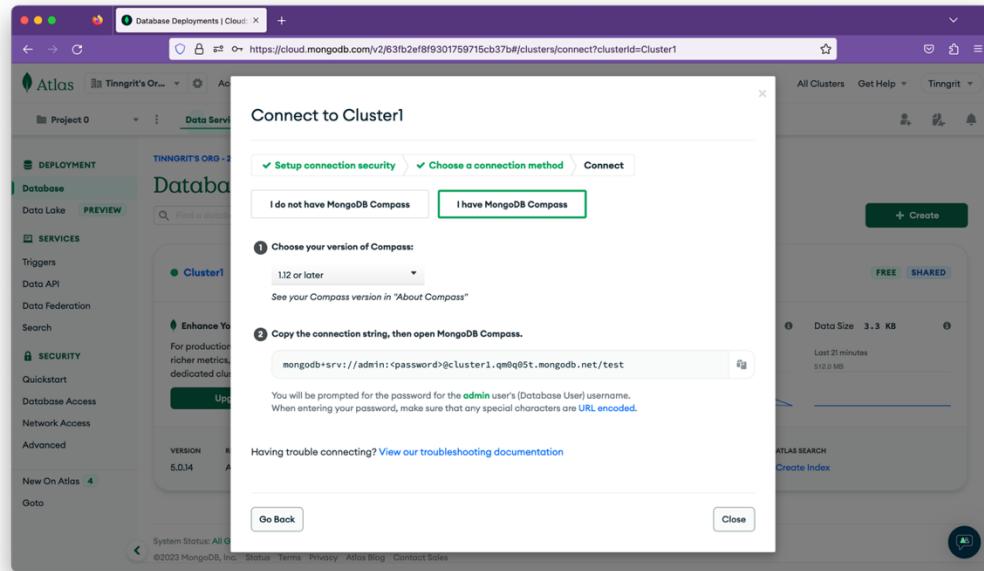
4. การนำข้อมูลเข้าสู่ฐานข้อมูลของโกลดีบี (MongoDB)

4.1 เปิดเว็บไซต์หน้าจัดการ Cluster กดปุ่ม Connect และกดปุ่ม Connect using MongoDB Compass



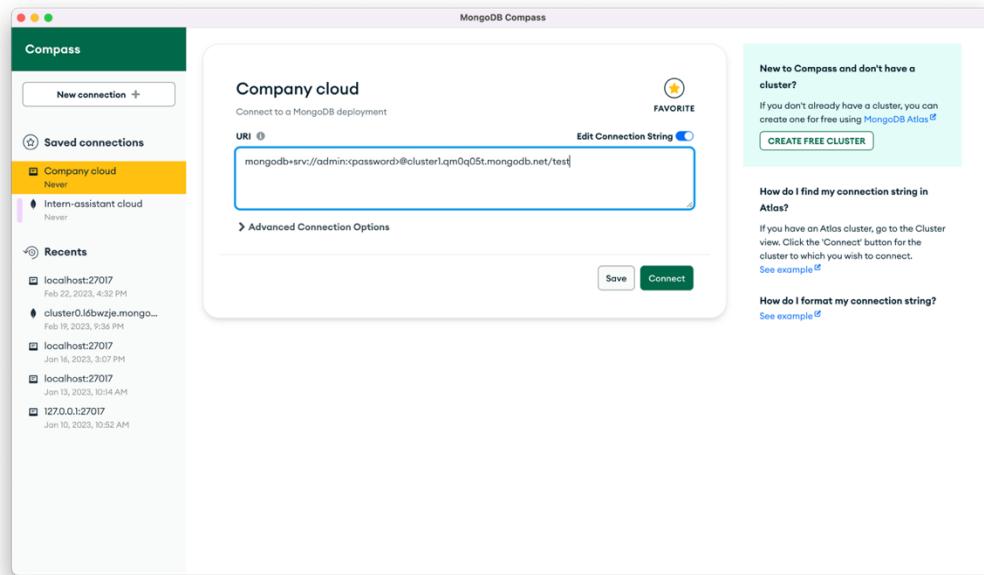
ภาพที่ 70 หน้าต่างเลือกเชื่อมต่อ กับ Cluster

4.2 ກົດປຸ່ມ I have MongoDB Compass ແລະ ຄົດລອກ Connection string ໃນຂອງ 2



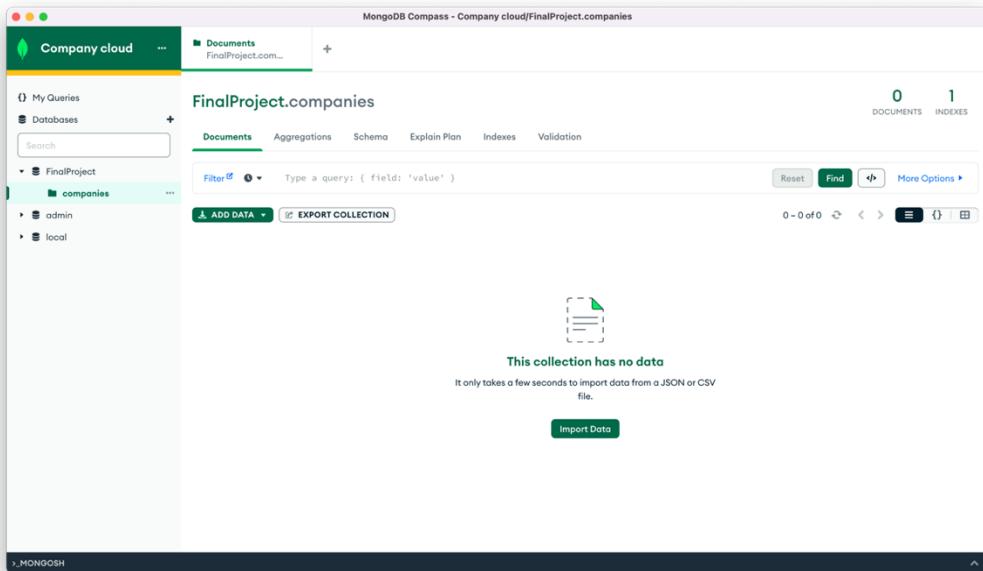
ກາພທີ 71 ໜ້າຕ່າງຂອງລາຍການເຊື່ອມຕ່ອງ Cluster ລັບ MongoDB compass

4.3 ເປີດໂປຣແກຣມ MongoDB compass ແລະ ວາງລິງคາກເຊື່ອມຕ່ອງ URI ແກ້ໄຂ username ແລະ password ໄທຕຽງກັບທີ່ສ້າງບັງວິຈາກນັ້ນກົດປຸ່ມ Connect



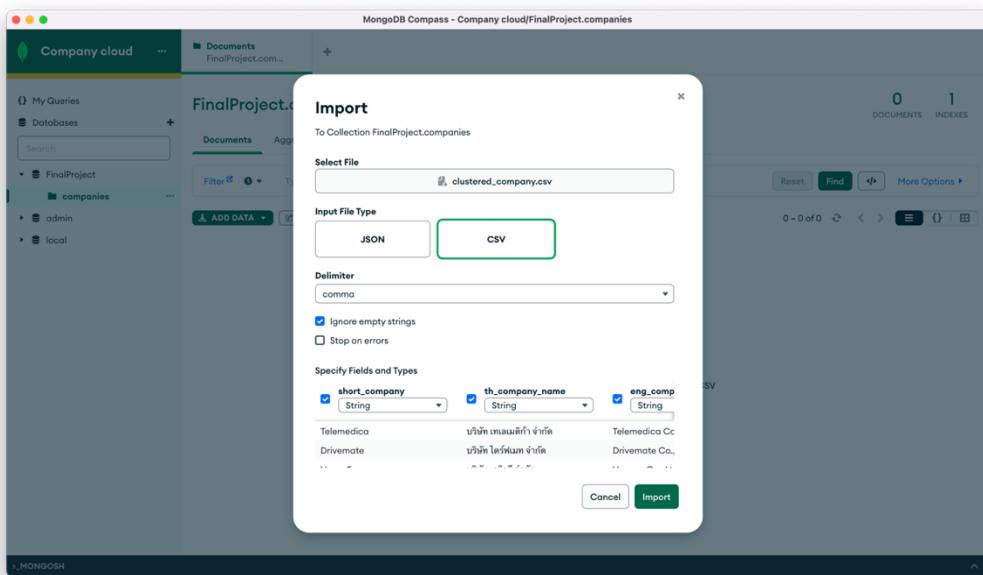
ກາພທີ 72 ໜ້າຕ່າງໂປຣແກຣມ MongoDB compass ສໍາຮັບເຊື່ອມຕ່ອງ Cluster

4.4 เลือกฐานข้อมูลและ Collection ที่เมนูด้านซ้ายและกดปุ่ม Import Data



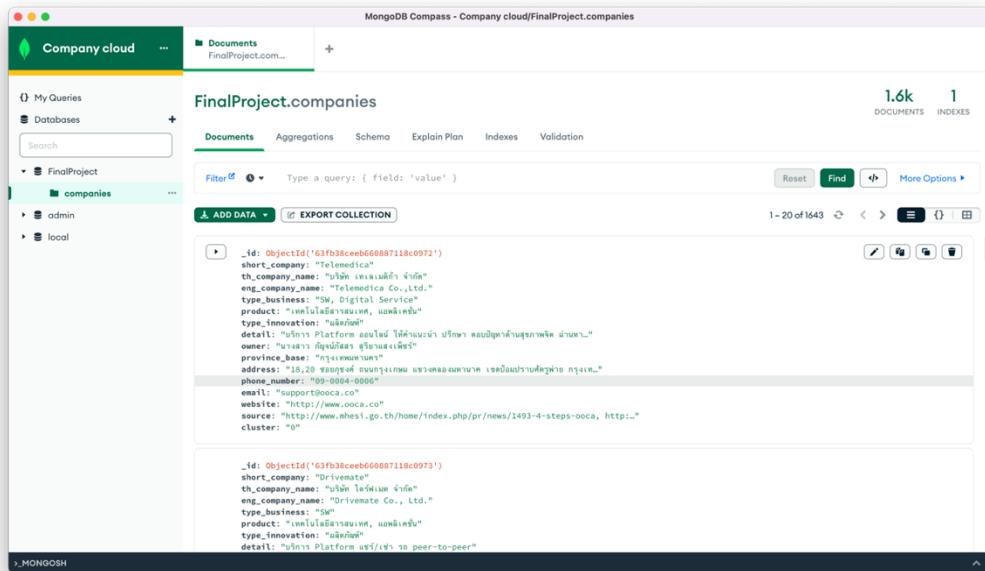
ภาพที่ 73 หน้าต่างโปรแกรมแสดงข้อมูลใน Collection

4.5 เลือกไฟล์ข้อมูลบริษัทที่จัดกู้มแล้วกดปุ่ม CSV เพื่อ Import ข้อมูลแบบไฟล์นามสกุล csv และกดปุ่ม Import และกดปุ่ม Done เพื่อเสร็จสิ้นกระบวนการ



ภาพที่ 74 หน้าต่าง Import ข้อมูลนามสกุลไฟล์ csv

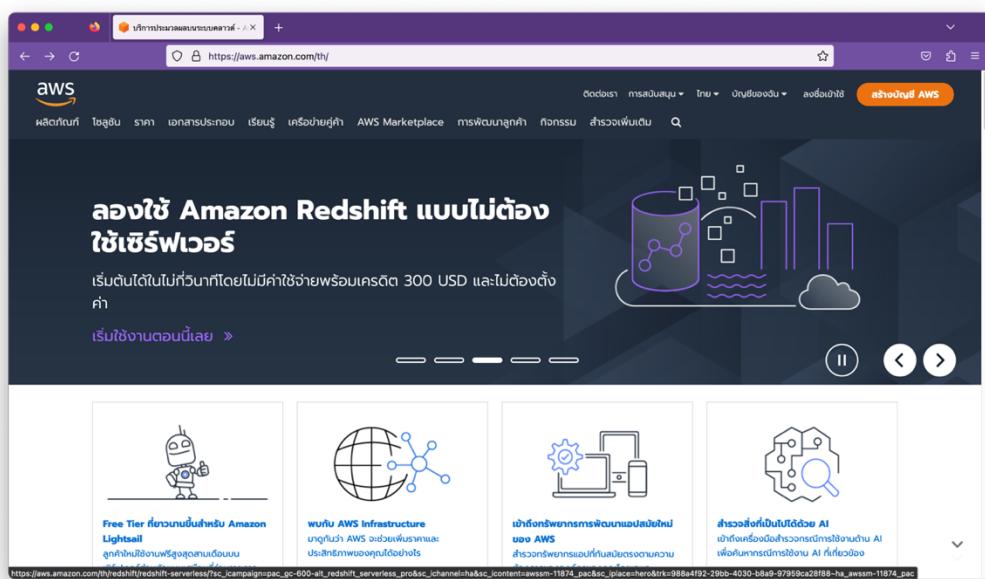
4.6 เมื่อ Import ข้อมูลสำเร็จจะได้ข้อมูลอยู่ใน Collection



ภาพที่ 75 หน้าต่างแสดงข้อมูลใน Collection ในโปรแกรม MongoDB compass

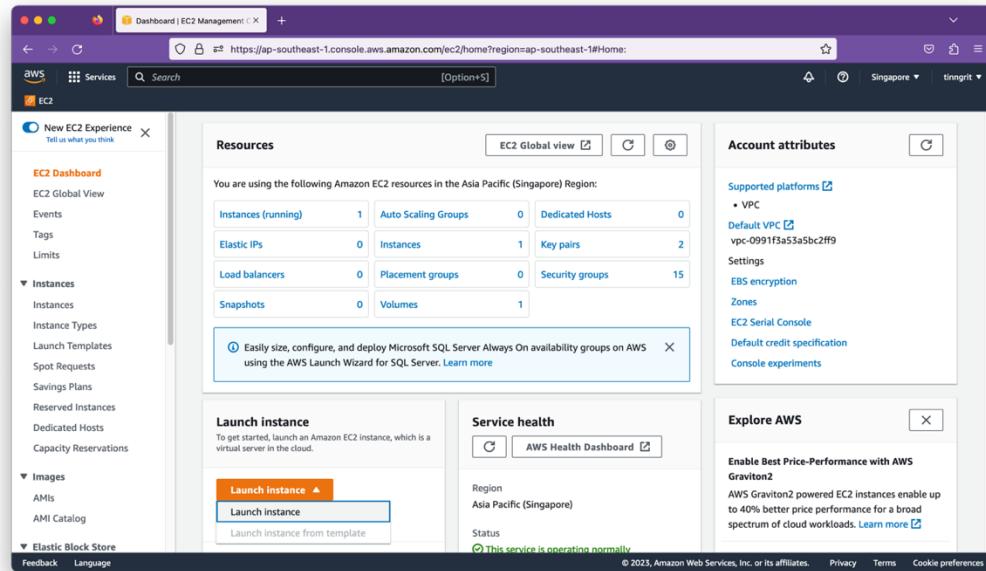
5. การใช้งาน Cloud computing ของ Amazon Web Services

5.1 เข้าไปยังเว็บไซต์ <https://aws.amazon.com/th/> เข้าสู่ระบบที่เมนู ลงชื่อเข้าใช้



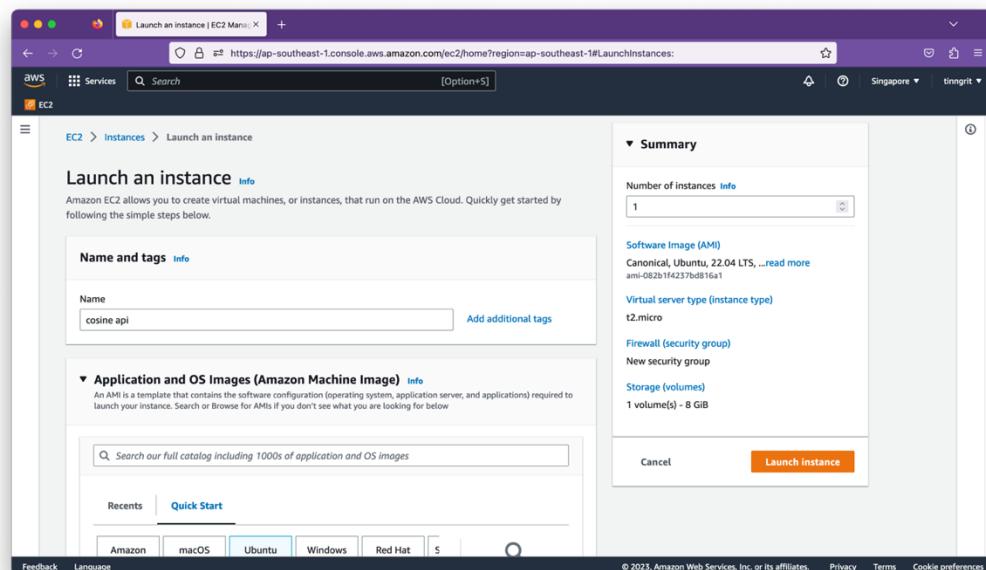
ภาพที่ 76 หน้าเว็บไซต์ Amazon Web Services

5.2 กดที่เมนู EC2 และกดปุ่ม Launch instance เพื่อสร้าง Instance ใหม่



ภาพที่ 77 หน้าแสดงการเลือกสร้าง Instance ใหม่

5.3 ตั้งค่าเครื่อง Instance ตามต้องการและกดปุ่ม Launch instance



ภาพที่ 78 หน้าแสดงการตั้งค่า Instance

5.4 ทำการเชื่อมต่อไปยัง Instance ด้วยวิธี SSH โดยใช้ Terminal

```

ubuntu@ip-172-31-35-185: ~
cd desktop/aws
~/desktop/aws
ssh -i "finalproject.pem" ubuntu@ec2-52-221-246-234.ap-southeast-1.compute.amazonaws.com
Welcome to Ubuntu 20.04.5 LTS (GNU/Linux 5.15.0-1028-aws x86_64)

 * Documentation: https://help.ubuntu.com
 * Management: https://landscape.canonical.com
 * Support: https://ubuntu.com/advantage

System information as of Sun Feb 26 12:32:02 UTC 2023

System load: 0.24 Processes: 102
Usage of /: 22.8% of 7.57GB Users logged in: 0
Memory usage: 25% IPv4 address for eth0: 172.31.35.185
Swap usage: 0%

Expanded Security Maintenance for Applications is not enabled.

24 updates can be applied immediately.
18 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

New release '22.04.2 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Last login: Sun Feb 26 12:31:43 2023 from 124.120.80.51
ubuntu@ip-172-31-35-185:~$
```

ภาพที่ 79 ตัวอย่างการเชื่อมต่อเข้าไปยัง Instance

6. การติดตั้งและใช้งาน Web API สำหรับคำนวณค่าความคล้ายคลึง (Cosine similarity)

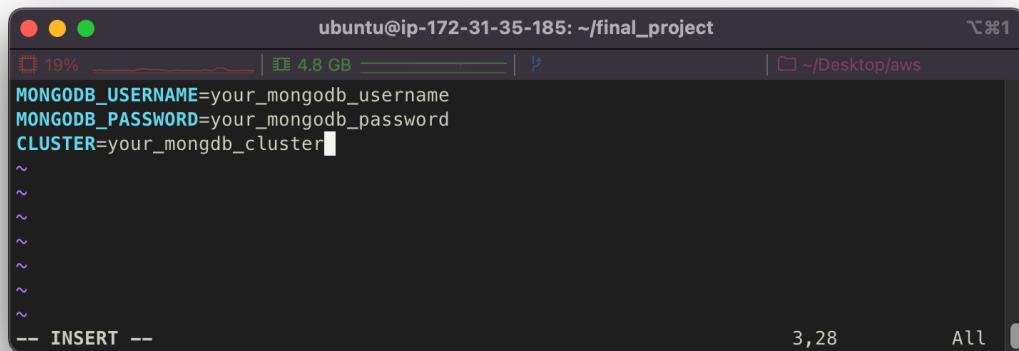
6.1 ทำการ Clone project จาก Github ที่ลิงค์
https://github.com/slapexs/final_project.git

```

ubuntu@ip-172-31-35-185: ~
git clone https://github.com/slapexs/final_project.git
Cloning into 'final_project'...
remote: Enumerating objects: 769, done.
remote: Counting objects: 100% (65/65), done.
remote: Compressing objects: 100% (51/51), done.
remote: Total 769 (delta 27), reused 47 (delta 14), pack-reused 704
Receiving objects: 100% (769/769), 17.56 MiB | 14.54 MiB/s, done.
Resolving deltas: 100% (369/369), done.
ubuntu@ip-172-31-35-185:~$
```

ภาพที่ 80 การดาวน์โหลดโปรเจกจาก Github ด้วยคำสั่ง git clone

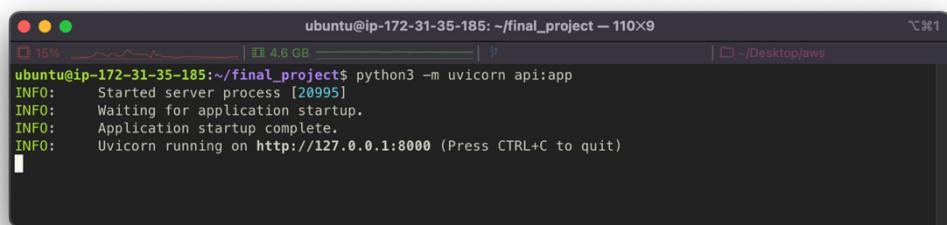
6.2 สร้างไฟล์ใหม่ในโฟลเดอร์ตั้งชื่อว่า .env และสร้างตัวแปรชื่อว่า MONGODB_USERNAME MONGODB_PASSWORD และ CLUSTER เพื่อใช้เก็บข้อมูลเชื่อมต่อฐานข้อมูล



```
ubuntu@ip-172-31-35-185: ~/final_project
MONGODB_USERNAME=your_mongodb_username
MONGODB_PASSWORD=your_mongodb_password
CLUSTER=your_mongodb_cluster
~
```

ภาพที่ 81 สร้างไฟล์ใหม่ชื่อ .env และสร้างตัวแปรเพื่อเก็บค่าเชื่อมต่อฐานข้อมูล

6.3 ใช้คำสั่ง pip3 install -r requirements.txt เพื่อทำการติดตั้ง library ที่จำเป็นและใช้คำสั่ง python3 -m uvicorn api:app เพื่อใช้งาน Server Web API



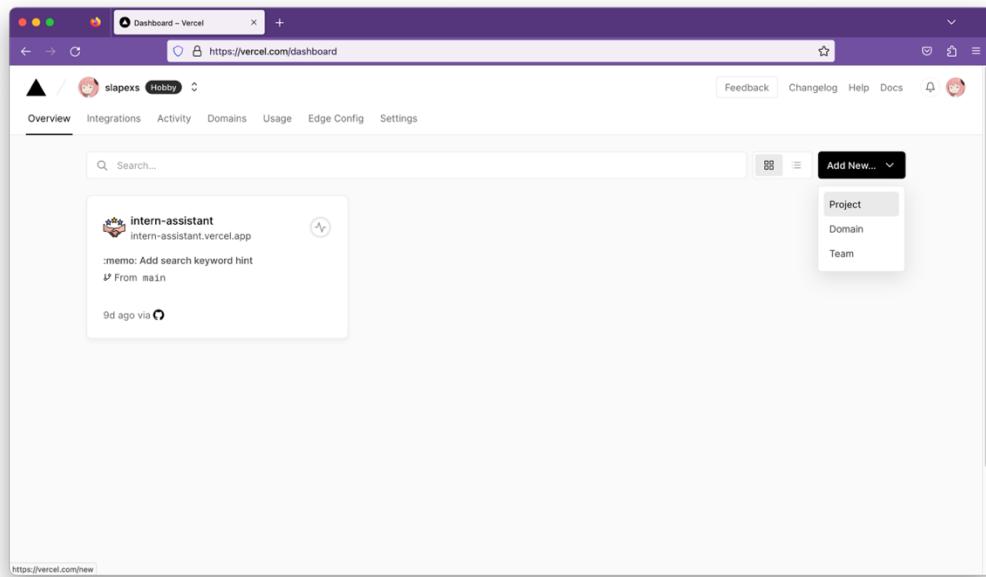
```
ubuntu@ip-172-31-35-185:~/final_project$ python3 -m uvicorn api:app
INFO:     Started server process [20995]
INFO:     Waiting for application startup.
INFO:     Application startup complete.
INFO:     Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)
```

ภาพที่ 82 ตัวอย่างการเริ่มต้น Server Web API เพื่อคำนวณค่า Cosine similarity บน AWS

7. การติดตั้งและใช้งานเว็บแอปพลิเคชัน (Web application)

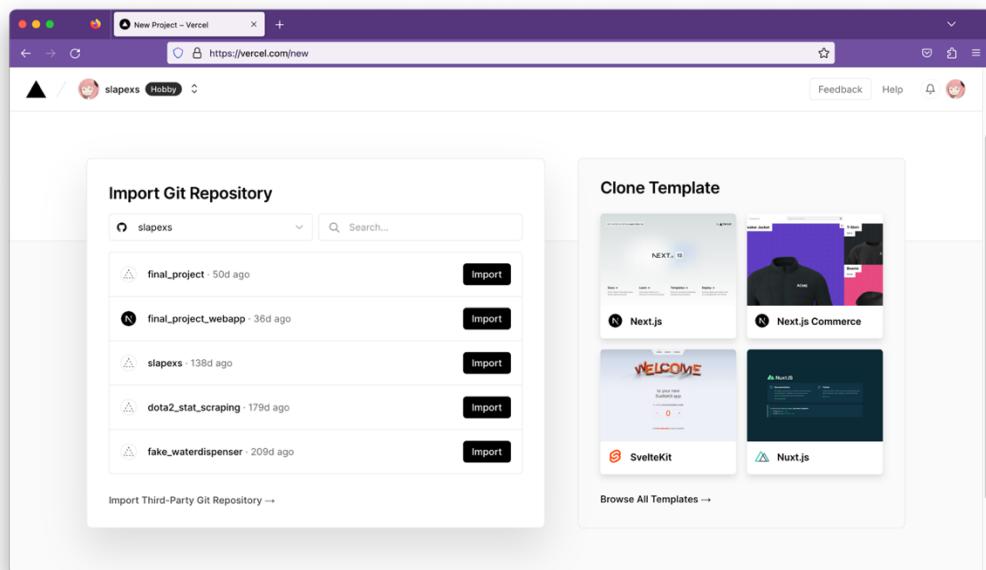
7.1 เข้าเว็บไซต์ <https://vercel.com> และเข้าสู่ระบบด้วยบัญชี Github

7.2 ทำการสร้างโปรเจคใหม่กดปุ่ม Add New และเลือก Project



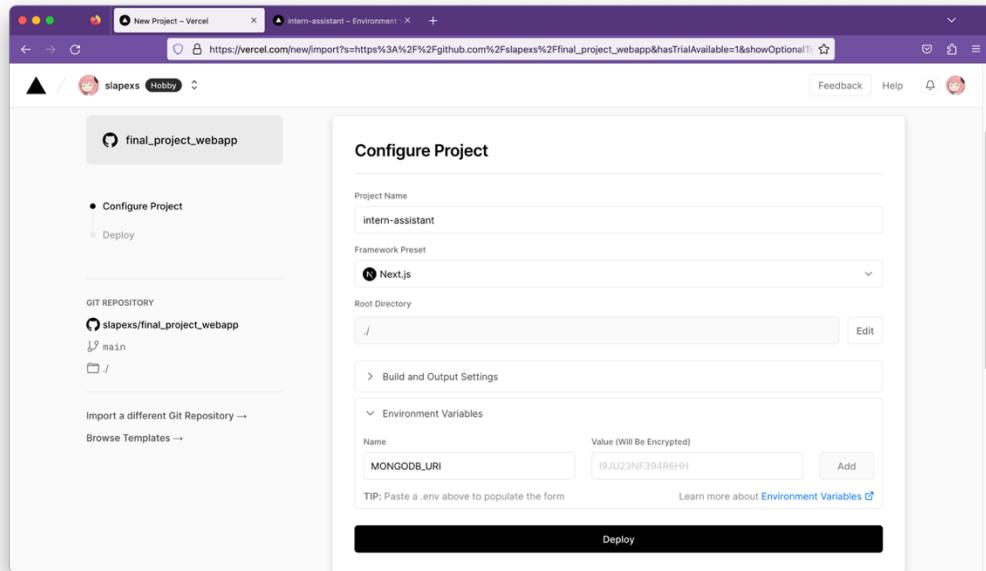
ภาพที่ 83 สร้างโปรเจคใหม่ใน Vercel

7.3 เลือก Repository ที่ต้องการจะ deploy และกดปุ่ม Import



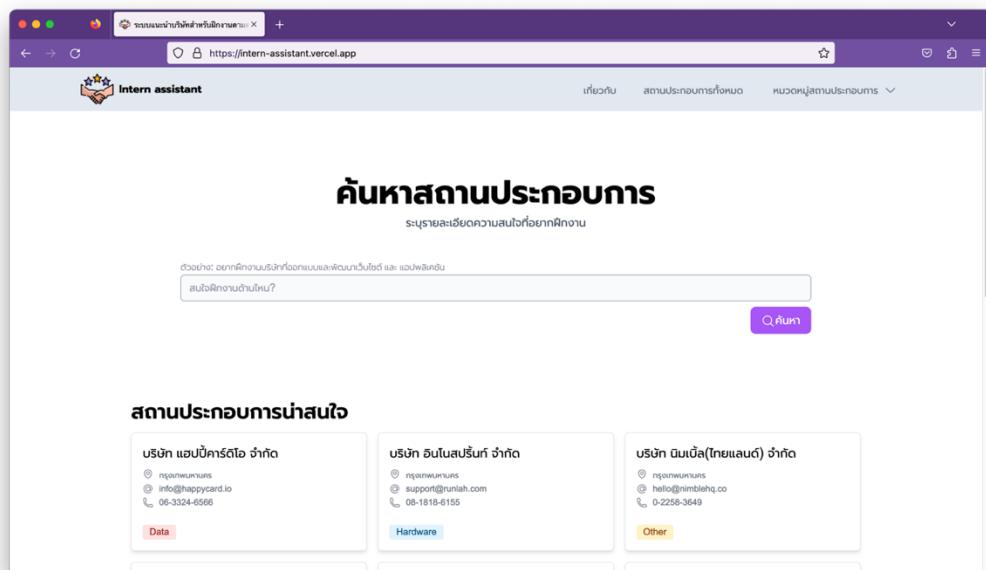
ภาพที่ 84 หน้าแสดงรายชื่อ Repository

7.4 ตั้งค่าโปรเจก และเพิ่มตัวแปร MONGODB_URI ในส่วนของ Environment Variables และใส่ค่าเป็น Connection string ของ MongoDB Atlas จากนั้นกดปุ่ม Deploy เพื่อทำการเผยแพร่สู่สาธารณะ



ภาพที่ 85 หน้าการตั้งค่าโปรเจกตอน Deploy

7.5 ตัวอย่างหน้าเว็บไซต์เมื่อ Deploy เรียบร้อย



ภาพที่ 86 ตัวอย่างหน้าเว็บไซต์

រាជធានីភ្នំពេញ

ក្រសួងពេទ្យ

ជាតិ

ภาคผนวก ข คู่มือการใช้งาน

ระบบแนะนำสถานที่ฝึกงานด้านคอมพิวเตอร์ ด้วยเทคโนโลยีการจัดกลุ่มเครื่อง (K-Means) สามารถแบ่งได้ 3 ส่วนดังนี้

1. ผู้ดูแลระบบ

- 1.1 การจัดกลุ่มข้อมูล
1. เปิดไฟล์ clustering.py ในโปรแกรม Text editor
2. ทำการกำหนดจำนวนของกลุ่มข้อมูลที่ตัวแปร k



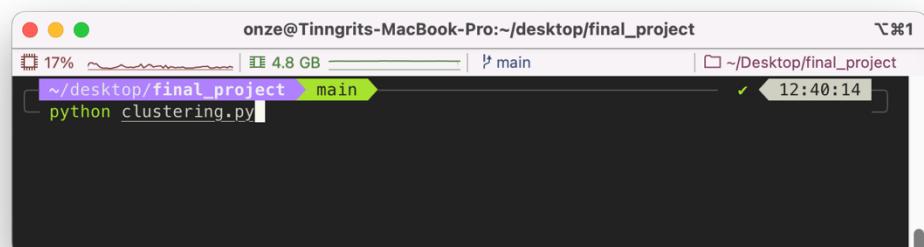
```

● ○ ●
1 k = 7
2 kmeans = KMeans(n_clusters=k, random_state=1)
3 # Fit model
4 kmeans.fit(df_tfidf[['x_value', 'y_value']])
5 clusters = kmeans.labels_

```

ภาพที่ 87 แสดงกำหนดจำนวนกลุ่มที่ตัวแปร k

3. ใช้คำสั่งใช้งานไฟล์เพื่อจัดกลุ่มข้อมูล



Terminal window showing the command:

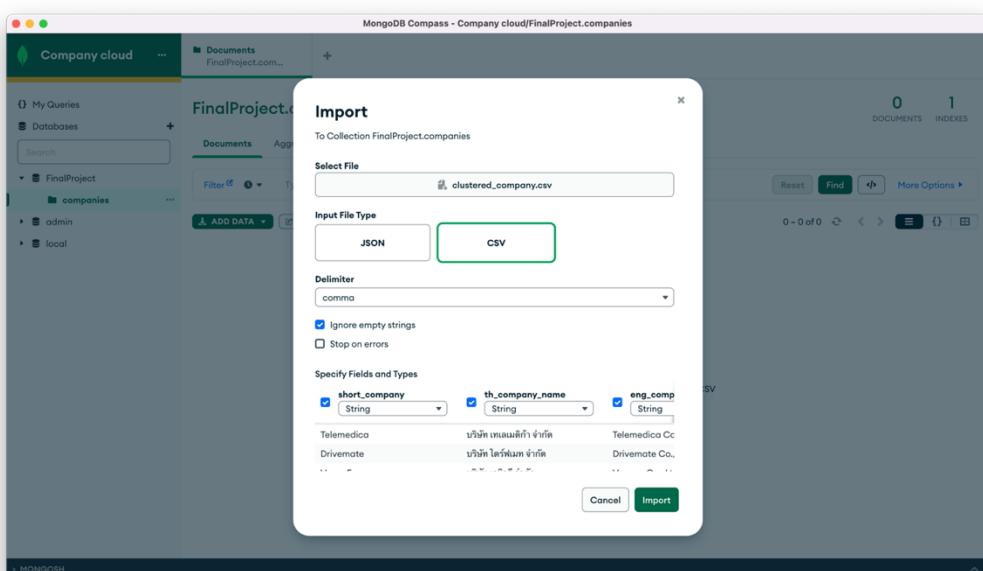
```
onze@Tinngrits-MacBook-Pro:~/desktop/final_project
~/desktop/final_project/main$ python clustering.py
```

ภาพที่ 88 แสดงการใช้คำสั่งจัดกลุ่มข้อมูลใน Terminal

4. ได้ไฟล์ clustered_company.csv ในโฟลเดอร์ document ที่เป็นผลลัพธ์การจัดกลุ่มข้อมูล

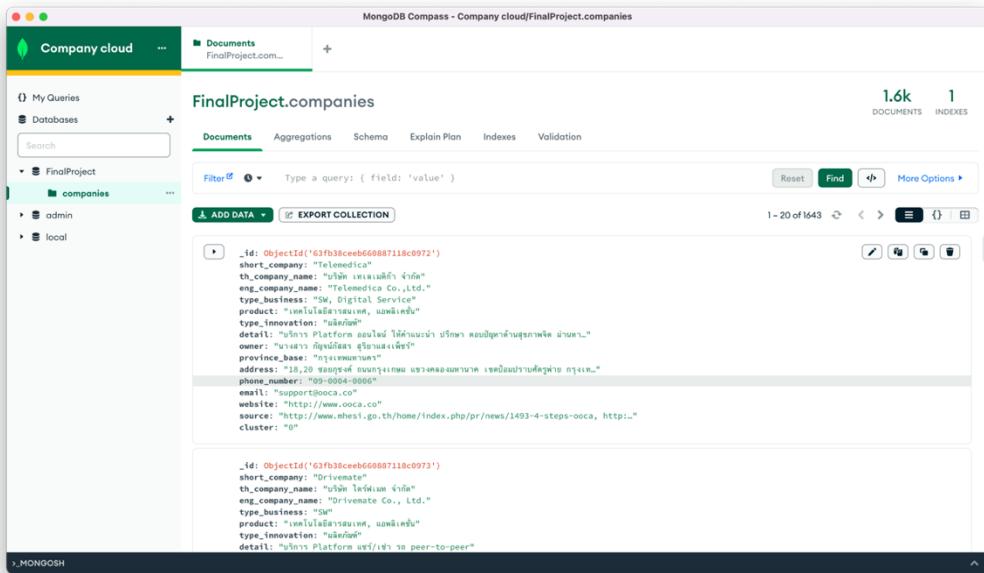
ภาพที่ 89 แสดงไฟล์ clustered_company.csv

5. เลือกไฟล์ข้อมูลบริษัทที่จัดกลุ่มแล้ว กดปุ่ม CSV เพื่อ Import ข้อมูลแบบไฟล์นามสกุล CSV และกดปุ่ม Import และกดปุ่ม Done เพื่อเสร็จสิ้นกระบวนการ



ภาพที่ 90 หน้าต่าง Import ข้อมูลนามสกุลไฟล์ CSV

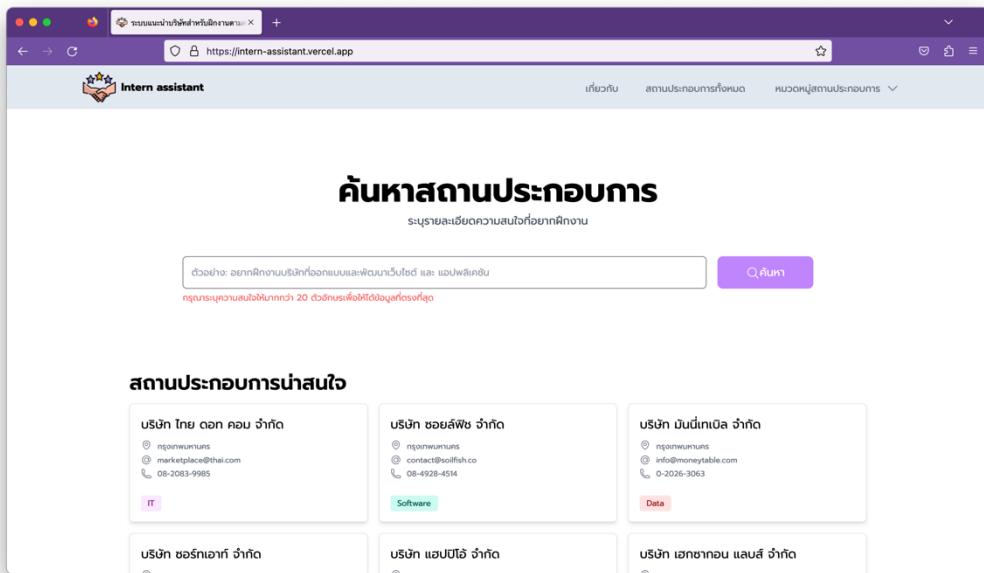
6. เมื่อ Import ข้อมูลสำเร็จจะได้ข้อมูลอยู่ใน Collection



ภาพที่ 91 ข้อมูลใน Collection ในโปรแกรม MongoDB compass

2. ผู้ใช้งาน

2.1 เข้าเว็บไซต์ <https://intern-assistant.vercel.app> จากนั้นทำการค้นหาบริษัทตัวய ความสนใจในรูปแบบงานของผู้ใช้



ภาพที่ 92 หน้าเว็บไซต์ intern-assistant.vercel.app

2.2 เมื่อเจอรายชื่อบริษัทผู้ใช้สามารถกดดูข้อมูลบริษัทได้ตามต้องการ

หมวดหมู่: Data	
บริษัท เทเลเมดี้ จำกัด Telemedia	◎ กรุงเทพมหานคร
บริษัท อินไฟฟ์ด์ จำกัด Infofed	◎ กรุงเทพมหานคร
บริษัท แม็ก เทค ทัวร์เตอร์ จำกัด TourTourist	◎ กรุงเทพมหานคร
บริษัท ออโต้เพียร์ จำกัด AUTOPAIR	◎ กรุงเทพมหานคร
บริษัท อายา จำกัด AIYA	◎ กรุงเทพมหานคร
บริษัท ไอ-แอป คิดเรชั่น จำกัด IAppCreation	◎ กรุงเทพมหานคร
บริษัท สวีฟท์ ดิโนมิกส์ จำกัด Swift Dynamics	◎ กรุงเทพมหานคร
บริษัท นิสิส ไนซ์โซลูชันส์ จำกัด NYSIS Solutions	◎ กรุงเทพมหานคร
บริษัท ไลเก็ล อะลایก์ จำกัด Local Alike	◎ กรุงเทพมหานคร
บริษัท เบลลักก์ กลุ๊ป จำกัด Bellugg	◎ กรุงเทพมหานคร

ภาพที่ 93 หน้าแสดงผลลัพธ์เมื่อค้นหาบริษัท

2.3 ผู้ใช้สามารถดูข้อมูลบริษัทเพื่อประกอบการตัดสินใจในการเลือกสถานประกอบการเพื่อฝึกงานได้

บริษัท วัน เมเดียซอฟต์ เอ็กซ์เพร็ค จำกัด
1 Mediasoft Expert Co., Ltd. (1ME)

Network

บริษัท วัน เมเดียซอฟต์ เอ็กซ์เพร็ค จำกัด
32/1 ซอยอุดมสุข 30 แยก 2 แขวงบางนาเหนือ เขตบางนา กรุงเทพมหานคร 10260
customer_care@1mediasoft.com
0-2105-4602
http://www.1mediasoft.com

ภาพที่ 94 หน้าแสดงข้อมูลบริษัท

3. Web API

เป็นเว็บ API ที่ทำหน้าที่ให้ข้อมูลที่ต้องการและมีหน้าที่ในการคำนวณความคล้ายคลึง (Cosine similarity) ข้อมูลที่สามารถคืนไปยังคำขอได้มีดังนี้

3.1 ข้อมูลบริษัททั้งหมดในฐานข้อมูล

```

{
  "short_company": "Telemedica",
  "th_company_name": "เทลเมเดีย เทคโนโลยี จำกัด",
  "eng_company_name": "Telemedica Co.,Ltd.",
  "type": "บริษัท",
  "product": "Software Digital Service",
  "product": "medical device",
  "type_innovation": "นวัตกรรม",
  "type": "บริษัท",
  "detail": "บริษัท Platform ออนไลน์ ให้ค่าตอบแทน บริษัท คอบลูฟ์ ค่าจ้างช่างภาพอาชีวะ ผู้เชี่ยวชาญการทดสอบความปลอดภัยและเชื่อมต่อสิ่งแวดล้อม ให้ค่าตอบแทน บริษัท ค่าจ้างอาชีวะ ผู้เชี่ยวชาญการทดสอบความปลอดภัยและเชื่อมต่อสิ่งแวดล้อม ให้ค่าตอบแทน บริษัท ค่าจ้างอาชีวะ"
}

```

ภาพที่ 95 ตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/allcompanies>

จากภาพที่ 95 แสดงตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/allcompanies> เมื่อ click GET เพื่อขอข้อมูลบริษัททั้งหมดในฐานข้อมูลด้วยโปรแกรม Postman

3.2 ข้อมูลบริษัทที่อยู่ในกลุ่มที่กำหนด

```

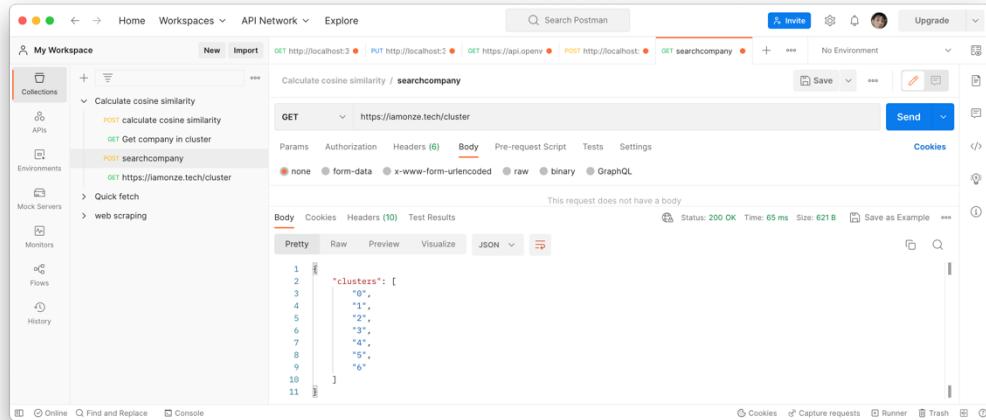
{
  "short_company": "SOFTLAVU",
  "th_company_name": "ซอฟต์ ลาวา จำกัด",
  "eng_company_name": "Softlau Co., Ltd.",
  "type": "บริษัท",
  "detail": "บริษัทผู้ผลิตซอฟต์แวร์ (Online Platform - Web/Mobile Application) สำหรับพิมพ์ธุรกิจ (Business Management Software) เช่น LavaPOS (Lava's Point of Sale System)",
  "owner": "วุฒิชัย ศรีนิลกุล นางสาวนราฯ ลีลารัตน์",
  "province_base": "กรุงเทพมหานคร",
  "address": "บ้านปัน ใจเพื่อน บ้านนาฯ ที่ 2, แขวง 59/13 หมู่ 2 แขวงราษฎร์ ตำบลราษฎร์ เขตฯ ล่างกา"
}

```

ภาพที่ 96 ตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/company/1>

จากภาพที่ 96 แสดงตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/company/1> ด้วย เมื่อ click GET เพื่อขอข้อมูลบริษัทที่อยู่ในกลุ่มที่ 1 ทั้งหมดในฐานข้อมูลด้วยโปรแกรม Postman

3.3 ຂໍ້ມູນລາຍກາຮັບສົດຂອງກຸລຸມ (Cluster ID) ທີ່ໜໍາມາໃນຈຸນຂອ່ມູນ

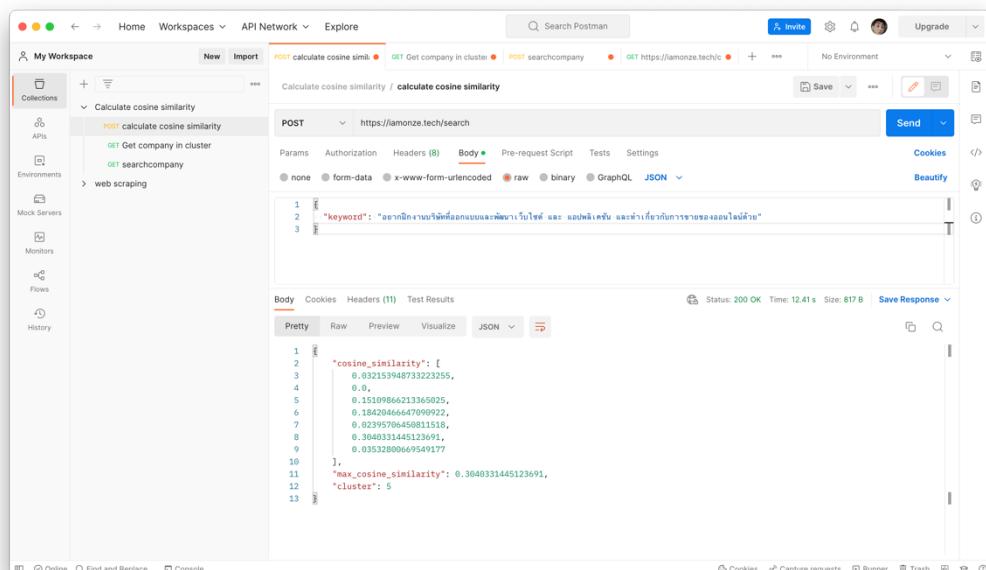


ກາພທີ 97 ຕ້ວອຍ່າງກາຮັບສົດຂອ່ມູນ https://iamonze.tech/cluster

ຈາກກາພທີ 97 ແສດງຕ້ວອຍ່າງກາຮັບສົດຂອ່ມູນ https://iamonze.tech/cluster ຕ້ວຍເມືອນ
ອດ GET ເພື່ອຂໍ້ມູນລາຍກາຮັບສົດຂອງກຸລຸມຂອ່ມູນ (Cluster ID) ທີ່ໜໍາມາໃນຈຸນຂອ່ມູນດ້ວຍ
ໂປຣແກຣມ Postman

3.4 ກາຮັບສົດຄາເພື່ອຄຳນວນຄາຄວາມຄລ້າຍຄລື່ງ (Cosine similarity)

1. ສົ່ງກລັບຂໍ້ມູນມາເປັນຜລລັບພົດຄວາມຄລ້າຍຄລື່ງແລະຮັບກຸລຸມ (Cluster ID)



ກາພທີ 98 ຕ້ວອຍ່າງກາຮັບສົດຂອ່ມູນ https://iamonze.tech/search

จากภาพที่ 98 แสดงตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/search> และแนบข้อมูลรายละเอียดความสนใจของผู้ใช้ไปด้วยเพื่อคำนวณค่าความคล้ายคลึง (Cosine similarity) ด้วยเมธอด POST และผลลัพธ์ที่คืนค่ากลับมาจะเป็นค่าความคล้ายคลึงของแต่ละกลุ่ม และรหัสกลุ่มข้อมูลที่มีความคล้ายคลึงมากที่สุด

2. ส่งกลับข้อมูลมาเป็นข้อมูลบริษัทที่อยู่ในกลุ่มที่มีความคล้ายมากที่สุด

```

POST https://iamonze.tech/searchcompany
{
  "keyword": "ออกแบบงานเว็บไซต์"
}

```

```

{
  "short_company": "MANINNOVATION",
  "th_company_name": "มน อินโน เดไซน์ จำกัด",
  "eng_company_name": "Man Innovation Co., Ltd.",
  "type_business": "SM, Digital Content",
  "detail": "บริษัทออกแบบและพัฒนาเว็บไซต์ (Web Application), ออกแบบตราสิทธิ์ (Logo Design), ออกแบบเบอร์ชิป (Brochure & Stationary Design)",
  "owner": "นายสมชาย วงศ์วนิช, นายศรีวุฒิ งามเจริญ",
  "province_base": "กรุงเทพมหานคร",
  "address": "39 ถนนเพชรบุรี 16 แขวงคลองเตยเหนือ เขตคลองเตย กรุงเทพมหานคร 10520",
  "phone": "+66 92 242 7222",
  "email": "thanes@maninnovation.com",
  "website": "http://thumweb.maninnovation.com",
  "cluster": "5",
  "company_id": "e11b799b-bb6f-11ed-b954-1831bf2b91e6",
  "cluster_name": "Network"
},
{
  "short_company": "Deeboon",
}

```

ภาพที่ 99 ตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/searchcompany>

จากภาพที่ 99 แสดงตัวอย่างการส่งคำขอไปยัง <https://iamonze.tech/searchcompany> และแนบข้อมูลรายละเอียดความสนใจของผู้ใช้ไปด้วยเพื่อคำนวณค่าความคล้ายคลึง (Cosine similarity) มีลักษณะคล้ายกับการทำงานในภาพที่ 98 แต่ผลลัพธ์ที่คืนค่ากลับมาจะเป็นข้อมูลบริษัทที่อยู่ในกลุ่มที่มีความคล้ายมากที่สุด

ประวัติผู้ศึกษา



ชื่อ-นามสกุล : นายพินกฤต สิงห์แก้ว
รหัสนักศึกษา : 64342205007-7
วันเดือนปีเกิด : 22 สิงหาคม พ.ศ. 2541
ที่อยู่ปัจจุบัน : 209 ม.5 ต.ร้องกวาง อ.ร้องกวาง จ.แพร่ 54140
E-mail : tinngrit@outlook.com

ประวัติการศึกษา

พ.ศ. 2554 – พ.ศ. 2557 : สำเร็จการศึกษาระดับมัธยมศึกษาตอนต้น
 โรงเรียนร้องกวางอนุสรณ์ จ.แพร่

พ.ศ. 2557 – พ.ศ. 2560 : สำเร็จการศึกษามัธยมศึกษาตอนปลาย
 โรงเรียนร้องกวางอนุสรณ์ จ.แพร่

พ.ศ. 2560 – พ.ศ. 2563 : สำเร็จการศึกษาระดับประกาศนียบัตรวิชาชีพชั้นสูง
 สาขาวิชาเทคโนโลยีคอมพิวเตอร์
 สาขางานคอมพิวเตอร์ซอฟต์แวร์
 วิทยาลัยเทคนิคแพร่

พ.ศ. 2564 – ปัจจุบัน : กำลังศึกษาระดับปริญญาตรี สาขาวิชาศาสตร์
 หลักสูตรวิทยาการคอมพิวเตอร์
 คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร
 มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน