

**โครงงานวิทยาการคอมพิวเตอร์ระดับปริญญาตรี**

**เรื่อง**

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

**โดย**

**นายทินกฤต สิงห์แก้ว**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**ปีการศึกษา 2565**



**ใบรับรองโครงงานทางวิทยาการคอมพิวเตอร์**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)**

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

โดย นายทินกฤต สิงห์แก้ว

**คณะกรรมการพิจารณาเห็นชอบโดย**

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

( ผู้ช่วยศาสตราจารย์ ดร.นงนุช เกตุ้ย )

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

( อาจารย์วรวิทย์ ฝั้นคำอ้าย )

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

( อาจารย์ศิริลักษณ์ แก้วศิริรุ่ง )

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

( อาจารย์วรวิทย์ ฝั้นคำอ้าย )

**โครงงานทางวิทยาการคอมพิวเตอร์ระดับปริญญาตรี**

**เรื่อง**

**ระบบแนะนำบริษัทสำหรับฝึกงานตามความสนใจ**

**ด้วยเทคโนโลยีปัญญาประดิษฐ์**

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

**โดย**

**นายทินกฤต สิงห์แก้ว**

**คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร**

**มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน**

**เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรบัณฑิต สาขาวิทยาศาสตร์**

**พ.ศ. 2565**

**บทคัดย่อ**

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

**คณะผู้ศึกษา** : นายทินกฤต สิงห์แก้ว

**อาจารย์ที่ปรึกษาฯ** : ผู้ช่วยศาสตราจารย์ ดร.นงนุช เกตุ้ย

**อาจารย์ที่ปรึกษาร่วมฯ** : อาจารย์วรวิทย์ ฝั้นคำอ้าย

**สาขา** : วิทยาศาสตร์

**หลักสูตร** : วิทยาการคอมพิวเตอร์

**ปีการศึกษา**  : 2565

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

**กิตติกรรมประกาศ**

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

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

สุดท้ายนี้ผู้ศึกษาหวังว่าโครงงานฉบับนี้จะเป็นประโยชน์สำหรับมหาวิทยาลัย  
และนักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน และผู้ที่สนใจที่จะศึกษาต่อไป

**สารบัญ**

**หน้า**

**บทคัดย่อ ก**

**กิตติกรรมประกาศ ข**

**สารบัญ ค**

**สารบัญตาราง ง**

**สารบัญภาพ จ**

**บทที่**

1. **บทนำ**
   1. ความเป็นมาและความสำคัญของปัญหา 1

**บทที่ 1**

**บทนำ**

* 1. **ความเป็นมาและความสำคัญของปัญหา**

ในระบบการศึกษาระดับปริญญาตรีนั้นรายวิชาที่มีในการศึกษาปีสุดท้ายของหลักสูตรคือ รายวิชาที่จะต้องให้นักศึกษาแต่ละคนนั้นออกไปฝึกทำงานที่สถานประกอบการต่างๆ ในช่วงระยะเวลาหนึ่ง ซึ่งเป็นสิ่งที่มีความสำคัญและความท้าทายเนื่องจากเป็นการที่นักศึกษาจะได้ทดลองทำงานจริง สถานการณ์จริง สถานที่จริง ในสถานประกอบการที่นักศึกษาได้เลือก

ดังนั้นการเลือกสถานประกอบการสำหรับฝึกงานจึงเป็นเรื่องที่ต้องให้ความสำคัญเป็นอย่างมากเนื่องจากหากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำนั้น ตรงกันกับความสามารถของนักศึกษาก็จะเป็นผลดี เนื่องจากความรู้และทักษะที่ได้จากการทำงานนั้นสามารถนำไปต่อยอดและใช้งานจริงเมื่อจบการศึกษาและเข้าทำงาน แต่หากสถานประกอบการที่เลือกนั้นรูปแบบธุรกิจหรืองานที่ทำไม่ตรงกับความต้องการหรือทักษะของนักศึกษาอาจทำให้การฝึกงานนั้นล้มเหลว หรืออาจไม่ได้ความรู้และทักษะที่ต้องการได้ และด้วยเทคโนโลยีปัญญาประดิษฐ์ในด้านของการประมวลผลภาษาธรรมชาตินั้นพัฒนาก้าวหน้าขึ้นมาก ทั้งในแง่ของเทคนิค เครื่องมือ และองค์ความรู้ ทำให้เกิดตัวอย่างการนำข้อมูลมาประมวลผลที่มีประสิทธิภาพมากมายในปัจจุบัน

ด้วยเหตุนึ้จึงได้มีการเริ่มโครงงานการพัฒนาเว็บแอปพลิเคชันเพื่อแนะนำบริษัทสำหรับหาสถานประกอบการที่ฝึกงานที่เหมาะสมและเป็นสถานประกอบการที่นักศึกษามีความสนใจในรูปแบบธุรกิจของสถานประกอบการนั้น เพื่อช่วยอำนวยความสะดวกแก่นักศึกษามหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา น่าน ในการค้นหาสถานประกอบการสำหรับออกฝึกงาน โดยการใช้วิธีประมวลผลภาษาธรรมชาติเข้ามาช่วยจัดกลุ่มสถานประกอบการและเสนอรายชื่อสถานประกอบการที่เหมาะสมแก่นักศึกษาผ่านทางเว็บแอปพลิเคชัน

* 1. **วัตถุประสงค์**

1.2.1 เพื่อศึกษาและพัฒนาเว็บแอปพลิเคชันเพื่อแนะนำสถานประกอบการตามความสนใจของนักศึกษา

1.2.2 เพื่อศึกษาและวิเคราะห์ความสนใจของนักศึกษาในการหาสถานประกอบการสำหรับฝึกงาน

1.2.3 เพื่อศึกษาประสิทธิภาพการแบ่งกลุ่มของข้อมูลสถานประกอบการด้วยวิธีประมวลผลภาษาธรรมชาติ

### ขอบเขตของโครงงาน

1.3.1 ผู้ใช้สามารถค้นหาสถานประกอบการได้ด้วยรายละเอียดของงานหรือรูปแบบธุรกิจที่สนใจ

1.3.2 ทุกครั้งที่มีการอัพเดทข้อมูลรายละเอียดธุรกิจของสถานประกอบการชุดใหม่เข้ามานั้นข้อมูลทั้งหมดจะถูกจัดกลุ่มใหม่ และต้องกำหนดชื่อกลุ่มใหม่จากป้ายเลขกลุ่ม

1.3.3 ในการจัดกลุ่มข้อมูลนั้นอาจมีข้อมูลสถานประกอบการอื่นเข้ามารวมได้บ้างเนื่องจากข้อมูลของสถานประกอบการนั้นอาจมีความหมายที่ไม่ชัดเจน 100%

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

**1.5 อุปกรณ์เครื่องมือที่ใช้ในโครงงาน**

1.5.1 อุปกรณ์ ประกอบด้วย

1. โทรศัพท์มือถือ
2. คอมพิวเตอร์

1.5.2 ซอฟต์แวร์ ประกอบด้วย

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

**1.6 นิยามศัพท์ที่ใช้ในโครงงาน**

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

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

1.6.3 Word segmentation หมายถึง ด้วยที่การเขียนภาษาไทยนั้นไม่มีการแยกคำด้วยการเว้นวรรคเหมือนภาษาอังกฤษ หรือ ภาษาอื่นๆ ดังนั้นจึงจำเป็นต้องทำการตัดคำจากประโยคออกมาเป็นคำ ๆ เพื่อให้นำไปประมวลหรือใช้งานต่อได้ด้วยอัลกอริทึมต่าง ๆ

1.6.4 Cosine similarity หมายถึง เป็นการหาค่าความคล้ายคลึงกันระหว่าง Vector A และ Vector B ว่าไปทิศทางเดียวกันหรือไม่โดยการใช้สูตรของกฎสามเหลี่ยมเพื่อหาผลลัพธ์แล้วนำมาเปรียบเทียบกัน

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

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

**บทที่ 2**

**ทฤษฎีและงานวิจัยที่เกี่ยวข้อง**

การศึกษาค้นคว้าเพื่อจัดทําโครงงานทางวิทยาการคอมพิวเตอร์ ระบบแนะนำบริษัทสำหรับฝึกงานด้วยเทคโนโลยีปัญญาประดิษฐ์ ผู้ศึกษาได้ ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้องตามหัวข้อลำดับต่อไปนี้

2.1 ทฤษฎีและความรู้ที่เกี่ยวข้อง

2.1.1 ทฤษฎี การประมวลผลภาษาธรรมชาติ หรือ Natural language processing

2.1.2 ทฤษฎี การตัดคำในภาษาไทยหรือ Word segmentation

2.1.3 ทฤษฎี การสกัดใจความของข้อความด้วยเทคนิค TF-IDF

2.1.4 ทฤษฎี การจัดกลุ่มข้อความด้วยอัลกอริทึม K-Means

2.1.5 ทฤษฎี การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow method

2.1.6 ทฤษฎี การคำนวณค่าความคล้ายคลึงด้วยเทคนิค Cosine similarity

2.1.7 ทฤษฎี Amazon web service

2.1.8 ทฤษฎี API

2.1.9 ทฤษฎี Cloudflare

2.1.10 ทฤษฎี Cors

2.1.11 ทฤษฎี CSS

2.1.12 ทฤษฎี Dotenv

2.1.13 ทฤษฎี Fastapi

2.1.14 ทฤษฎี Figma

2.1.15 ทฤษฎี Firefox

2.1.16 ทฤษฎี Git

2.1.17 ทฤษฎี Github

2.1.18 ทฤษฎี Heroicons

2.1.19 ทฤษฎี HTML

2.1.20 ทฤษฎี Javascript

2.1.21 ทฤษฎี Matplotlibs

2.1.22 ทฤษฎี Microsoft excel

2.1.23 ทฤษฎี Mongodb

2.1.24 ทฤษฎี Next.js

2.1.25 ทฤษฎี Nltk

2.1.26 ทฤษฎี Notion

2.1.27 ทฤษฎี Numpy

2.1.28 ทฤษฎี Pandas

2.1.29 ทฤษฎี Postman

2.1.30 ทฤษฎี Pythainlp

2.1.31 ทฤษฎี Python

2.1.32 ทฤษฎี Python-dotenv

2.1.33 ทฤษฎี Scikit-learn

2.1.34 ทฤษฎี Sweetalert2

2.1.35 ทฤษฎี Tailwind css

2.1.36 ทฤษฎี Vercel

2.1.27 ทฤษฎี Visual studio code

2.2 งานวิจัยที่เกี่ยวข้อง

**2.1 ทฤษฎีและความรู้ที่เกี่ยวข้อง**

2.1.1 ทฤษฎี การประมวลผลภาษาธรรมชาติ หรือ Natural language processing

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

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

กระบวนการทำงานของ NLP นั้น มีประกอบไปด้วยหลายส่วนของการประมวลผลและใช้แปลความหมาย ประกอบด้วยดังนี้

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

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

2.1.2 ทฤษฎี การตัดคำในภาษาไทยหรือ Word segmentation

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



**ภาพที่ 1** ตัวอย่างการทำ Word segmentation

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

2.1.3 ทฤษฎี การสกัดใจความของข้อความด้วยเทคนิค TF-IDF

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

**ตารางที่ 1** ตัวอย่างการคำนวณค่า Term Frequency ที่จำนวนคำทั้งหมดเท่ากับ 7

|  |  |  |  |
| --- | --- | --- | --- |
| คำ | จำนวนคำ | Term Frequency | ผลลัพธ์ |
| เว็บไซต์ | 5 | 5 ÷ 7 | 0.71 |
| หนังสือ | 1 | 1 ÷ 7 | 0.14 |
| ออนไลน์ | 2 | 2 ÷ 7 | 0.29 |
| ขาย | 2 | 2 ÷ 7 | 0.29 |

**ตารางที่ 1 (ต่อ)**

|  |  |  |  |
| --- | --- | --- | --- |
| เข้าชม | 1 | 1 ÷ 7 | 0.14 |
| มือถือ | 4 | 4 ÷ 7 | 0.57 |
| และ | 3 | 3 ÷ 7 | 0.43 |

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

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

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

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| คำ | TF | IDF | TF-IDF |
| เว็บไซต์ | 0.71 | 0.31 | 0.22 |
| หนังสือ | 0.14 | 0.70 | 0.10 |
| ออนไลน์ | 0.29 | 0.70 | 0.20 |

***ตารางที่* 3 (*ต่อ*)**

|  |  |  |  |
| --- | --- | --- | --- |
| ขาย | 0.29 | 0.70 | 0.20 |
| เข้าชม | 0.14 | 1.00 | 0.14 |
| มือถือ | 0.57 | 0.52 | 0.30 |
| และ | 0.43 | 0.70 | 0.30 |

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

2.1.4 ทฤษฎี การจัดกลุ่มข้อความด้วยอัลกอริทึม K-Means

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

ทำการคำนวณและย้ายจุด Centroid และหาค่าเฉลี่ยจนค่าเฉลี่ยไม่มีการเปลี่ยนแปลงจะได้จุดกึ่งกลางของข้อมูลในแต่ละกลุ่ม

Chart, scatter chart

Description automatically generated

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

2.1.5 ทฤษฎี การหาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow method

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

**ภาพที่ 3** กราฟที่แสดงจำนวนข้อผิดพลาดเพื่อหาจพนวนกลุ่มที่เหมาะสมที่สุด

2.1.6 ทฤษฎี การคำนวณค่าความคล้ายคลึงด้วยเทคนิค Cosine similarity

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

Text

Description automatically generated

**ภาพที่ 4** สมการการหาค่า Cosine similarity

อธิบายโดยง่ายคือเป็นการวัดระยะห่างระหว่าง Object A และ Object B ว่ามีความคล้ายกันแค่ไหนยกตัวอย่างเปรียบเทียบระหว่างคำว่า “ยินดีที่ได้รู้จักครับ” และ “ยินดีที่ได้รู้จักค่ะ” ทำการตัดเพื่อหาคำทั้งหมดก่อนครับ คือ [ “ยินดี” , “ที่” , “ได้” , “รู้จัก” , “ครับ” , “ค่ะ”] เราจะได้ Object A และ B เขียนเป็นชุดข้อมูล ได้ดังนี้

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

Graphical user interface, text, website

Description automatically generated

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