แบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

นางสาวประภัสสร ข่ายกระโทก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลย มหาวิทยาลัยสุโขทัยธรรมาธิราช
พ.ศ. 2564

An Optimal Model for Text Categorization for Knowledge Dissemination Book

Miss Prapatsorn Khaikratoke

A Thesis Submitted in Partial Fulfillment of the Requirements for

the Degree of Master of Science in Information and Communication Technology

School of Science and Technology

Sukhothai Thammathirat Open University

2021

หัวข้อวิทยานิพนธ์ แบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

ชื่อและนามสกุล นางสาวประภัสสร ข่ายกระโทก

แขนงวิชา เทคโนโลยีสารสนเทศและการสื่อสาร

สาขาวิชา วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช

อาจารย์ที่ปรึกษา 1. อาจารย์ ดร. ศรันย์ นาคฉนอม

2. อาจารย์ ดร. เตชค์ฐสิณป์ เพียซ้าย

วิทยานิพนธ์นี้ ได้รับความเห็นชอบให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรระดับปริญญาโท เมื่อวันที่ 27 กันยายน 2565

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.พงศ์ศรัณย์ บุญโญปกรณ์)

..... กรรมการ

(อาจารย์ ดร. ศรันย์ นาคฉนอม)

..... กรรมการ

(อาจารย์ ดร. เตชค์ฐสิณป์ เพียซ้าย)

..... ประธานกรรมการบัณฑิตศึกษา

(รองศาสตราจารย์ ดร.นราธิป ศรีราม)

ชื่อวิทยานิพนธ์ แบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้
ผู้วิจัย นางสาวประภัสสร ข่ายกระโทก **รหัสนักศึกษา** 2629600616
ปริญญา วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร)
อาจารย์ที่ปรึกษา (1) อาจารย์ ดร. ศรันย์ นาคถนอม (2) อาจารย์ ดร. เตชค์ฐสิณป์ เพียซ้าย
ปีการศึกษา 2564

บทคัดย่อ

การวิจัยนี้มีวัตถุประสงค์เพื่อ 1) ศึกษาและวิเคราะห์ระบบหนังสือเผยแพร่ความรู้ 2) พัฒนาแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ 3) ประเมินแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

การวิจัยใช้ชุดข้อมูลหนังสือเผยแพร่ความรู้ ตั้งแต่ปี 2553 – 2563 ในรูปแบบระบบวารสารอิเล็กทรอนิกส์ออนไลน์ และทำการวิเคราะห์คำสำคัญจากหนังสือเผยแพร่ความรู้จำนวน 948 คำเพื่อจัดกลุ่มจำแนกหมวดหมู่ด้วยการพัฒนาสร้างแบบจำลองจาก 4 อัลกอริทึม ได้แก่ อัลกอริทึมต้นไม้การตัดสินใจ อัลกอริทึมนาอีฟเบย์ อัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน และอัลกอริทึมการถดถอยโลจีสติก โดยใช้การเลือกคุณลักษณะแบบค่าเกณฑ์ความรู้เพื่อประเมินประสิทธิภาพของแต่ละอัลกอริทึม การทดสอบแบ่งชุดข้อมูลเป็น 2 ส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกสอน และชุดข้อมูลสำหรับการทดสอบ

จากผลการประเมินพบว่า 1) ระบบหนังสือเผยแพร่ความรู้สามารถจำแนกหมวดหมู่ข้อความได้ 4 หมวดหมู่ ได้แก่ หมวดหมู่ยุทธศาสตร์ หมวดหมู่ยุทธการ หมวดหมู่ยุทธวิธี และหมวดหมู่อื่นๆ และ 2) การพัฒนาสร้างแบบจำลองจากทั้ง 4 อัลกอริทึมด้วยการประเมินประสิทธิภาพเพื่อจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้พบว่า 3) อัลกอริทึมการถดถอยโลจีสติกและอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน มีค่าความถูกต้องเท่ากับ 0.87 อัลกอริทึมนาอีฟเบย์มีค่าความถูกต้องเท่ากับ 0.85 และอัลกอริทึมต้นไม้การตัดสินใจมีค่าความถูกต้องเท่ากับ 0.82 และแบบจำลองที่เหมาะสมได้แก่ อัลกอริทึมการถดถอยโลจีสติกเนื่องจากอัลกอริทึมการถดถอยโลจีสติกใช้เวลาในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ใช้เวลาน้อยกว่าอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีนเป็นร้อยละ 11.8

คำสำคัญ การวิเคราะห์ข้อมูล, การจัดหมวดหมู่, การจำแนกหมวดหมู่ข้อความ

Thesis title: An Optimal Model for Text Categorization for Knowledge Dissemination Book
Researcher: Miss Prapatsorn Khaikratoke; **ID:**2629600616; **Degree:** Master of Science (Science and Technology); **Thesis advisors:** (1) Dr. Sarun Nakthanom; (2) Dr. Tejtasin Phiasai; **Academic year:** 2021

Abstract

The objectives of this research were 1) to study and analyze the knowledge dissemination book system, 2) to develop a model for categorizing the text of knowledge dissemination books, and 3) to evaluate a text classification model of the knowledge dissemination books.

In this research, the knowledge book dataset that was an online e-journal system from 2010 to 2020 was analyzed to classify a category of the knowledge books with 948 keywords by using the development of modeling from 4 algorithms that included decision tree algorithm,Naïve Bayes algorithm, support vector machine algorithm, and logistic regression algorithm. All algorithms used the knowledge threshold feature selection to assess the performance of the algorithms. The test data was divided into two parts that were training dataset and testing dataset.

The evaluation results revealed that 1) all algorithms could categorize books into 4 categories that included strategic category, operational category, tactical category, and other categories and 2) the developed model was developed by the performance assessment from the four algorithms to categorize books and 3) each algorithm found that the logistic regression algorithm and the support vector machine algorithm had an accuracy of 0.87, the Naïve Bayes algorithm had an accuracy of 0.85 and the decision tree algorithm had an accuracy of 0.82. For an optimal model, it was the logistic regression algorithm that it took a time for processing less than the support vector machine algorithm to categorize knowledge dissemination books approximately 11.8 percent.

Keywords: Data Analysis, Categorization, Text Categorization

กิตติกรรมประกาศ

ขอกราบขอบพระคุณ อาจารย์ ดร. ศรันย์ นาคฉนอม อาจารย์ที่ปรึกษาวิทยานิพนธ์ และอาจารย์ ดร. เตชค์ฐสิณป์ เพียซ้าย อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม เป็นอย่างยิ่งที่ได้เสียสละเวลาให้ คำปรึกษา คำแนะนำ ชี้แนวทางในการดำเนินงาน และให้ความช่วยเหลือเมื่อเกิดปัญหาในขณะ ดำเนินงาน ทำให้การจัดทำวิทยานิพนธ์นี้สมบูรณ์ด้วยดี

ขอกราบขอบพระคุณ ผศ.ดร.พงศ์ศรัณย์ บุญโญปกรณ์ ที่ได้สละเวลาเป็นกรรมการสอบปกป้องวิทยานิพนธ์ และให้คำแนะนำสำหรับนำไปปรับปรุงการดำเนินงานให้เป็นไปอย่างราบรื่น

ขอขอบพระคุณ คณาจารย์ทุกท่านในสาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช ที่อบรมและให้ความรู้ตลอดหลักสูตร

ประภัสสร ข่ายกระโทก

พฤศจิกายน 2565

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

สารบัญภาพ (ต่อ)

หน้า

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันเราอยู่ในยุคดิจิทัลที่มีการแสวงหาเนื้อหาใหม่ๆ เพื่อเติมเต็มความพึงพอใจแบบทันทีทันใดแนวโน้มจำนวนผู้ใช้อุปกรณ์พกพามีเพิ่มขึ้นมีการใช้เวลาบนช่องทางนี้ทุกวันมากกว่า 5 ชั่วโมงต่อวัน ทำให้พฤติกรรมของผู้บริโภค สื่อที่เปลี่ยนไป ทำให้รูปแบบการสื่อสารเป็นดิจิทัลมากขึ้น อย่างหลีกเลี่ยงไม่ได้ Digital Publishing จึงเป็นทางเลือกที่ดีสำหรับธุรกิจ หรือองค์กรที่ต้องการ พัฒนาการสื่อสารดิจิทัลให้พร้อมเชื่อมต่อและโต้ตอบกับได้สะดวก รวดเร็ว ทุกที่ ทุกเวลาที่ผู้บริโภค เป้าหมายต้องการ จะเห็นได้ว่ามีการนำเทคโนโลยี Digital Publishing หรือ การจัดพิมพ์ดิจิทัล มาใช้ ในการแปลงสิ่งพิมพ์จากเดิมที่มีลักษณะรูปแบบเป็นเพียงแผ่นกระดาษในเล่มหนังสือที่พิมพ์ซึ่งมีแค่ ข้อความและภาพนิ่งเท่านั้น แปลงให้เป็นดิจิทัลสำหรับ แจกจ่าย เผยแพร่หรือแบ่งปัน เพื่อใช้งานผ่าน อุปกรณ์อิเล็กทรอนิกส์ (สมาร์ทโฟน แท็บเล็ต โน๊ตบุ๊ค คอมพิวเตอร์ตั้งโต๊ะ) รวมทั้งการนำไปใช้เพื่อจัด จำหน่ายผ่านเว็บ แอปพลิเคชันบนมือถือ หรือผู้ให้บริการทางอิเล็กทรอนิกส์ เทคโนโลยี Digital Publishing ที่เราสามารถมองเห็นภาพได้ชัดเจนในปัจจุบัน คือ Blog หรือ eBook ยังคงครอบคลุม สื่อและเนื้อหาทุกชนิดที่เผยแพร่ผ่านระบบอินเทอร์เน็ต อาทิ หนังสือ, เพลง, วิดีโอ, ข่าว, เกม, เว็บ, แอปพลิเคชันบนมือถือและอื่นๆ ซึ่งจากผลการสำรวจพฤติกรรมผู้ใช้อินเทอร์เน็ตในประเทศไทยของ ETDA พบว่าคนไทยมีการใช้อินเทอร์เน็ตเพิ่มขึ้นอย่างก้าวกระโดดราว 150% มีจำนวนผู้ใช้ อินเทอร์เน็ต 47.5 ล้านคน หรือราว 70% ของประชาชนทั้งประเทศ และผลจากการเปลี่ยนแปลง นวัตกรรมด้านเทคโนโลยีดิจิทัลทำให้ส่งผลดีต่อการปรับปรุงสิ่งพิมพ์ให้ออนไลน์ ทุกสำนักพิมพ์มีความคิดสร้างสรรค์ มีกลยุทธ์การสร้างสรรค์เนื้อหาที่เหมาะสมสามารถเผยแพร่เนื้อหาได้เองทาง ออนไลน์ สร้างฐานผู้ชมหรือผู้อ่าน และพัฒนาวิธีหารายได้ใหม่ๆ จากเนื้อหาดิจิทัลได้อย่างง่ายดายจึง เป็นหนึ่งเหตุผลที่ทำให้การสื่อสารดิจิทัลได้รับความนิยมและมีแนวโน้มที่จะเติบโตกลายเป็นอนาคตของสื่อสิ่งพิมพ์ ผู้วิจัยทำการศึกษาค้นคว้างานวิจัย จากแหล่งค้นคว้าออนไลน์เพื่อนำองค์ความรู้ด้านสื่อสิ่งพิมพ์ออนไลน์ และการจำแนกหมวดหมู่ข้อความมาจัดเก็บและพัฒนาเผยแพร่ทางดิจิทัลจากการศึกษาได้พบทฤษฎีที่มีลักษณะเกี่ยวข้อง ตัวอย่างเช่น การสกัดข้อมูลเทคนิคการเรียนรู้ของเครื่อง โดยการใช้เทคนิค

การทำเหมืองข้อความ (กาญจนา สุดาทิพย์, 2559) การวิเคราะห์ข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความ (ทิชากร เนตรสุวรรณ์, 2559) พัฒนาระบบจำแนกหมวดหมู่การแจ้งซ่อมบ้านออนไลน์โดยใช้เหมืองข้อความ (ประเดิม วงค์กระโซ่, 2560) เป็นต้นจากงานวิจัยที่ได้ศึกษาไปข้างต้น สามารถนำมาจัดหมวดหมู่ข้อความ และสืบค้นคำได้ตามความต้องการ ผู้วิจัยจึงมีแนวคิดนำเทคนิคการประมวลผลภาษาธรรมชาติ และองค์ความรู้ดังกล่าวข้างต้น มาปรับปรุงพัฒนาประยุกต์ใช้กับงานวิจัยของตน ซึ่งเดิมหน่วยงานมีการตีพิมพ์เผยแพร่หนังสือเผยแพร่ความรู้ทางทหารให้แก่กำลังพล รวมถึงประชาชนบุคคลทั่วไปที่มีความสนใจ โดยสามารถศึกษาค้นคว้าในรูปแบบของเล่มหนังสือ และสามารถดาวน์โหลดข้อมูลเอกสารในรูปแบบอิเล็กทรอนิกส์ได้ที่เว็บไซต์ของหน่วยงานได้บางส่วน ด้วยเนื้อหาในหนังสือเผยแพร่ความรู้ส่วนใหญ่พบว่าเป็นลักษณะของข้อความเป็นส่วนใหญ่ และจำนวนหน้าของเอกสารอิเล็กทรอนิกส์ที่มีจำนวนมาก จึงพบปัญหาทำให้การอ่าน การค้นหาข้อมูลที่ต้องการเป็นไปได้ยาก ค้นหาได้เฉพาะชื่อเรื่อง ปีที่จัดพิมพ์ ไม่มีการจัดหมวดหมู่ของคำหรือข้อความทำให้เสียเวลาในการค้นหา หากมีการรวบรวมและจัดกลุ่มของข้อความเป็นหมวดหมู่ จะทำให้การค้นหาข้อมูลสะดวกรวดเร็ว ตรงตามความต้องการของผู้ใช้งาน และพบว่าในการพัฒนาหนังสือเผยแพร่ความรู้ของหน่วยงาน ยังไม่สามารถวิเคราะห์องค์ความรู้ที่ได้จัดทำหนังสือฯ ขึ้นมา ยังไม่สามารถบอกแนวโน้ม ทิศทาง ความต้องการองค์ความรู้ของการพัฒนาหนังสือเผยแพร่ความรู้ ต่อไปว่าควรเป็นไปในทิศทางใด

ผู้วิจัยจึงมีแนวคิดพัฒนาจากเดิมมีการจัดพิมพ์บนกระดาษไปสู่การจัดพิมพ์ดิจิทัล และพัฒนาระบบการเผยแพร่หนังสือแบบออนไลน์ เพื่อให้ผู้ใช้งานสามารถเข้าถึงได้สะดวกจากทุกที่ทุกเวลาโดยใช้ข้อมูลเดิมที่มีอยู่เพื่อให้การค้นหาข้อมูลได้สะดวกรวดเร็ว ตรงตามความต้องการ และได้ประยุกต์ใช้การจำแนกข้อความ (Text Classification) สำหรับการจำแนกประเภท แยกแยะหรือจัดข้อความเป็นหมวดหมู่หรือเป็นกลุ่ม เพื่อระบุและเพื่อการวิเคราะห์ว่าข้อความควรจัดจำแนกอยู่ใน หมวดหมู่ใดโดยใช้คำสำคัญจากเนื้อหาภายในหนังสือฯ แต่ละเล่มและได้พัฒนาการจำแนกหมวดหมู่ หนังสือเผยแพร่ความรู้ โดยใช้เทคนิคเหมืองข้อความจัดกลุ่มของข้อมูลให้เป็นหมวดหมู่และการตัดคำ ด้วย Python และใช้อัลกอริทึม Decision Trees, Support Vector Machine, Naive Bayes และ Logistic regression มาใช้สำหรับเพื่อสร้างกระบวนการวิเคราะห์ความรู้สึกสร้างแบบจำลองการจำแนกหมวดหมู่ไปใช้วิเคราะห์กับข้อมูลคำสำคัญที่ได้มีการรวบรวมเอาไว้ทั้งสิ้น 948 คำ และนำมาทดลองใช้กับกลุ่มของคำที่ถูกแบ่งไว้สำหรับทดสอบเพื่อเปรียบเทียบประสิทธิภาพในการจำแนกจัดประเภทหมวดหมู่หนังสือเผยแพร่ความรู้ และนำมาใช้ในกำหนดแนวทางในการพัฒนาเอกสารเผยแพร่ความรู้ของหน่วยต่อไป

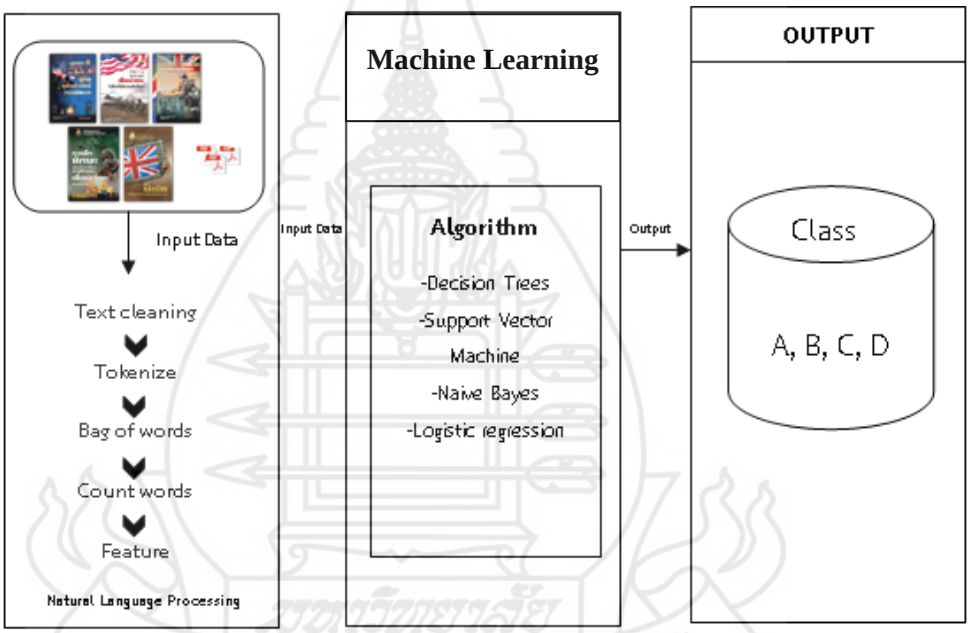
2. วัตถุประสงค์การวิจัย

2.1 เพื่อศึกษาและวิเคราะห์ระบบหนังสือเผยแพร่ความรู้

2.2 เพื่อพัฒนาแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

2.3 เพื่อประเมินแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

3. กรอบแนวคิดการวิจัย



ภาพที่ 1.1 กรอบแนวคิดการวิจัย

4. สมมติฐานการวิจัย

ได้แบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้จากอัลกอริทึมที่เหมาะสม โดยมีการเปรียบเทียบและประเมินประสิทธิภาพค่าความถูกต้อง ค่าความแม่นยำของแบบจำลอง

5. ขอบเขตของการวิจัย

5.1 ด้านเนื้อหาการวิจัย การวิจัยมีรูปแบบเนื้อหาเป็นลักษณะการพัฒนาแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ โดยมีการเปรียบเทียบประสิทธิภาพแบบจำลอง และเปรียบเทียบประสิทธิภาพของอัลกอริทึม Logistic Regression, Decision Tree, Naïve Bayes และ Support Vector Machine หาค่าความถูกต้อง (Accuracy) และค่าความแม่นยำ (Precision) เพื่อหาแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความฯ

5.2 ด้านประชากรและกลุ่มตัวอย่าง การออกแบบและพัฒนาแบบจำลองในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ผู้วิจัยรวบรวมข้อมูลคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้ จำนวน 948 คำ ซึ่งได้มีการรวบรวมไว้ตั้งแต่ปี พ.ศ. 2553 ถึง 2563 แปลงข้อมูลให้อยู่ในรูปแบบไฟล์ .csv เพื่อนำแบบจำลองที่สร้างขึ้นมาใช้จำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ และวิเคราะห์แนวโน้มทิศทางความต้องการองค์ความรู้ตามระดับดังกล่าวเพื่อการพัฒนาต่อไปด้วยการใช้การจำแนกหมวดหมู่ต่อไป

5.3 ด้านระยะเวลา ระยะเวลาในการทำวิจัยตั้งแต่ 2562 ถึง ตุลาคม 2565 ดังแสดงในตารางระยะเวลา และขั้นตอนการดำเนินงานวิจัย

6. นิยามศัพท์เฉพาะ

6.1 การวิเคราะห์ข้อมูล (Data Analysis) คือ กระบวนในการนำข้อมูลต่างๆ ที่ได้จากการการเก็บรวบรวมไว้มาทำการเรียบเรียง จัดกลุ่ม/แยกประเภทชุดข้อมูล หาความสัมพันธ์ของชุดข้อมูลแต่ละชุดในรูปแบบต่างๆ เพื่อหาความหมาย หรือคำตอบตามเป้าหมาย หรือวัตถุประสงค์ที่วางไว้

6.2 แบบจำลอง (Model) คือ สิ่งที่ได้รับการพัฒนาขึ้นเพื่ออธิบายหรือแสดงให้เห็นถึงองค์ประกอบสำคัญๆของเรื่องใดเรื่องหนึ่งให้เข้าใจได้ง่ายขึ้น สามารถทำความเข้าใจการทำงานของระบบจริงได้ง่ายกว่าการศึกษาจากระบบจริงโดยตรงเพื่อเป็นแนวทางในการดำเนินการอย่างใดอย่างหนึ่งต่อไป

6.3 การจำแนกหมวดหมู่ เป็นการจัดและแยกหมวดหมู่ของสิ่งที่เราต้องการที่มีลักษณะคล้าย หรือใกล้เคียงกันและสัมพันธ์กันให้อยู่ในประเภท หรือหมวดหมู่เดียวกันเพื่อความเป็นระเบียบเรียบร้อยในการจัด เพื่อความสะดวกและรวดเร็วในการค้นหาสิ่งที่สัมพันธ์กันในส่วน

ของการจัดหมวดหมู่หนังสือนั้นมีลักษณะการจัดที่คล้ายกัน โดยจะพิจารณาจากเนื้อหา เนื้อเรื่อง คำสำคัญในเล่ม หรือลักษณะการแต่งของผู้เขียนที่มีความคล้ายคลึงกันไว้ในหมวดหมู่เดียวกัน

6.4 ประสิทธิภาพ หมายถึง การหาหรือการพิจารณาในการวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก ค่าความถ่วงดุล และระยะเวลาที่ใช้ในการประมวลผล

7. ประโยชน์ที่คาดว่าจะได้รับ

7.1 ได้กระบวนการวิเคราะห์การจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ที่เกิดขึ้นใหม่

7.2 ได้อัลกอริทึมที่เหมาะสมสำหรับการวิเคราะห์จำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

7.3 ได้ข้อมูลที่ผ่านการวิเคราะห์ การจำแนกหมวดหมู่ข้อความ และแนวโน้มทิศทางความต้องการองค์ความรู้ตามระดับ เพื่อนำมาใช้ในการกำหนดแนวทางในการพัฒนาเอกสารเผยแพร่ความรู้ของหน่วยได้

8. วิธีการดำเนินการวิจัย

ผู้วิจัยแบ่งขั้นตอนระยะเวลาและขั้นตอนในการดำเนินงานวิจัย ดังนี้

8.1 ศึกษาค้นคว้า และสืบค้นงานวิจัยที่เกี่ยวข้อง

8.2 การจัดเตรียมรวบรวมข้อมูล นำข้อมูลมาวิเคราะห์

8.3 การพัฒนา Open Journal System (OJS)

8.4 การสร้างแบบจำลอง และการประเมินประสิทธิภาพ

8.5 ปรับปรุง และแก้ไขชิ้นงาน

8.6 สรุปผล และจัดทำรายงานและเตรียมการนำเสนอ

	งาน	เดือน 1	เดือน 2	เดือน 3	เดือน 4	เดือน 5	เดือน 6	เดือน 7	เดือน 8
1	ศึกษาค้นคว้า และสืบค้นงานวิจัยที่เกี่ยวข้อง								
2	การจัดเตรียมรวบรวมข้อมูลสำหรับนำข้อมูลมาวิเคราะห์								
3	การพัฒนา Open Journal System (OJS)								
4	การสร้างแบบจำลอง และการประเมินประสิทธิภาพ								
5	ปรับปรุง และแก้ไขชิ้นงาน								
6	สรุปผลและจัดทำรายงานและเตรียมการนำเสนอ								

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

งานวิจัยนี้ได้รวบรวมข้อมูลคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้ มาทำการประยุกต์ใช้การจำแนกข้อความ การจัดกลุ่ม สำหรับการวิเคราะห์ว่าข้อความควรจัดจำแนกอยู่ในหมวดหมู่ใด และสร้างแบบจำลองเพื่อใช้ในการจำแนกหมวดหมู่ไปใช้วิเคราะห์กับข้อมูลเพื่อให้การค้นหาข้อมูลได้สะดวกรวดเร็ว ตรงตามความต้องการผู้วิจัยได้ศึกษาค้นคว้าทฤษฎีและงานวิจัยที่เกี่ยวข้องกับงานวิจัยของตน สำหรับนำมาประยุกต์ใช้ในการทำงานวิจัย มีรายละเอียดดังต่อไปนี้

1. หนังสือเผยแพร่ความรู้

หนังสือเผยแพร่ความรู้ คือ หนังสือหรือเอกสารตำราที่หน่วยด้านการศึกษาของกองทัพบกได้จัดพิมพ์ขึ้นมาเพื่อเผยแพร่และแจกจ่ายให้กับหน่วยในกองทัพ เพื่อเผยแพร่วิทยาการทางทหาร เป็นเรื่องที่เหมาะในการพัฒนาทางด้านความคิด การเรียนรู้ให้กับกำลังพลของ ทบ. โดยความมุ่งหมายในการจัดพิมพ์ คือ เพื่อเผยแพร่วิทยาการทางทหาร สำหรับให้หน่วยไว้ทำการศึกษาหาความรู้เพิ่มเติม และสำหรับให้กำลังพลไว้ทำการศึกษาหาความรู้เพิ่มเติม เนื้อหาภายในเล่มจะเป็นลักษณะของคู่มือการฝึก การศึกษา การรบ ความคิดเห็น แนะแนวทางการปฏิบัติ และประสบการณ์ที่ได้เผชิญของผู้เขียน เป็นต้น

2. แบบจำลอง (Model)

2.1 แบบจำลอง (Model) คือสิ่งที่มนุษย์ได้สร้างขึ้นมาใช้แทนสิ่งนั้นๆ เพื่อให้การศึกษาและการ ทำความเข้าใจการทำงานเป็นไปได้ง่ายมากกว่าการทำความเข้าใจระบบจริงโดยตรงรูปแบบลักษณะ ของแบบจำลองอาจมีรูปแบบที่ต่างกันออกไปทั้งลักษณะ รูปแบบ และขนาดแต่สิ่งที่เหมือนกันคือ ลักษณะในการพัฒนาแบบจำลอง เช่น วัตถุประสงค์ ข้อสมมติในการพัฒนาไม่ซับซ้อน มีการกำหนด ขอบเขตชัดเจน (Dr.WeerakasetSuanpaga, (ม.ป.ป)) ประเภทแบบจำลอง สามารถแบ่งได้เป็น 5 ลักษณะ ดังนี้

2.1.1 แบบจำลองเชิงแนวคิด แสดงแนวคิด กระบวนการ โดยใช้ภาพเขียน หรือภาพวาดโครงสร้างระบบ

2.1.2 แบบจำลองเชิงกายภาพ สร้างขึ้นเพื่อให้เห็นรูปทรง หรือรูปร่างทางกายภาพ อาจใช้มาตราส่วนย่อ หรือขยายส่วน จากของจริงเพื่อให้เห็นภาพชัดมากขึ้น

2.1.3 แบบจำลองเชิงคณิตศาสตร์และสถิติ แบบจำลองเชิงคณิตศาสตร์ ใช้รูปแบบสมการทางคณิตศาสตร์ อธิบายความสัมพันธ์ระหว่างองค์ประกอบภายในระบบ ส่วนแบบจำลองทางสถิติต่างจากแบบจำลอง เชิงคณิตศาสตร์คือจะเกี่ยวข้องกับการวิเคราะห์ความผันแปร

2.1.4 แบบจำลองเชิงภาพเคลื่อนไหว มีการนำเทคนิคทางเทคโนโลยีด้านการสร้างภาพ แปลงข้อมูลดิบ เป็นภาพที่แสดงรูปทรง อาจเป็นลักษณะภาพ 2 มิติ หรือ 3 มิติ เคลื่อนไหวได้

2.1.5 แบบจำลองเชิงซิตเต็มไดนามิคส์ สามารถคำนวณการทำงานระบบคอมพิวเตอร์ได้

2.2 แบบจำลองข้อมูล หรือแบบจำลองฐานข้อมูล คือขั้นตอนแรกของการออกแบบฐานข้อมูลที่ กำหนดโครงสร้างฐานข้อมูลจะถูกนำไปใช้ในการจัดเก็บรวมถึงการจัดการข้อมูลผู้ใช้งานฐานข้อมูล นอกจากนี้อาจหมายถึงการระบบถึงแบบจำลองข้อมูลเพื่อใช้ในการกำหนดขอบเขตของปัญหา แบบจำลองข้อมูลมีลักษณะการแสดงผลในรูปแบบของแผนภาพที่แสดงถึงโครงสร้างซับซ้อนของ ฐานข้อมูลคุณลักษณะ ข้อจำกัด การเปลี่ยนแปลง เป็นต้น (อ.ดร.โกเมศ อัมพวัน, ม.ป.ป))

2.3 แบบจำลอง (Model) คือ สัญลักษณ์ใช้จำลองข้อเท็จจริงต่างๆ ที่เกิดขึ้นในระบบ แบบจำลอง ประกอบไปด้วย แผนภาพชนิดต่างๆ ที่แสดงให้เห็นถึงมุมมองของระบบ นอกจากนี้แบบจำลองยังเป็น เครื่องมือที่ช่วยในการสื่อสารระหว่างบุคคลมีความถูกต้องตรงกัน แบบจำลองเป็นสิ่งที่ได้มาจากการ วิเคราะห์ความต้องการของผู้ใช้งานทั้งในด้านระบบและซอฟต์แวร์ ทำให้มองเห็นถึงความต้องการของ ระบบได้ชัดเจนว่ามีหน้าที่ทำอะไรและอย่างไร (Softengthai, 2012)

3. วารสารอิเล็กทรอนิกส์

3.1 วารสารอิเล็กทรอนิกส์ (E-journal) เป็นวารสารรูปแบบใหม่ที่มีการจัดเก็บ บันทึก เผยแพร่ สารนิเทศทางวิชาการในรูปแบบของแฟ้มคอมพิวเตอร์และทางสื่ออิเล็กทรอนิกส์มี มีกำหนดการ เผยแพร่วารสารที่แน่นอน สม่ำเสมอ สามารถสืบค้นและสั่งซื้อ รวมถึงการเป็นสมาชิก โดยสามารถแบ่งวารสารอิเล็กทรอนิกส์ได้เป็น 3 รูปแบบ คือวารสารอิเล็กทรอนิกส์รูปแบบ

ฐานข้อมูลระบบ ออนไลน์ วารสารอิเล็กทรอนิกส์ฐานข้อมูลซีดี-รอม ฉบับเต็มและวารสารอิเล็กทรอนิกส์รูปแบบ เครือข่ายคอมพิวเตอร์ (อ.ปราณี วงศ์จำรัส และคณะ, (ม.ป.ป))

3.2 วารสารอิเล็กทรอนิกส์ เป็นวารสารรูปแบบดิจิทัลที่ผลิตและเผยแพร่ผ่านทางอินเทอร์เน็ต โดย สำนักพิมพ์วารสาร สถาบันการศึกษา ตัวแทนจัดทำฐานข้อมูลได้มีการจัดพิมพ์ร่วมกับวารสารใน รูปแบบสิ่งพิมพ์ หรือจัดพิมพ์ในรูปแบบอิเล็กทรอนิกส์รูปแบบเดียว และมีการนำเสนอเนื้อหา เหมือนกับวารสารรูปแบบสิ่งพิมพ์ วารสารอิเล็กทรอนิกส์สามารถเข้าถึงได้โดยไม่ต้องเสียค่าใช้จ่าย (วราพรรณ อภิศุภะโชค, 2550)

3.3 วารสารอิเล็กทรอนิกส์ หมายถึง วารสารรูปแบบใหม่ที่มีการจัดเก็บ บันทึก และพิมพ์เผยแพร่ สารนิเทศทางวิชาการไว้ในรูปแฟ้มคอมพิวเตอร์ และสื่ออิเล็กทรอนิกส์มีกำหนดออกแน่นอน สม่ำเสมอ โดยสามารถทำการเข้าถึง สืบค้นข้อมูล และสั่งซื้อหรือบอกรับเป็นสมาชิกได้จากฐานข้อมูล ซีดี-รอม ฐานข้อมูลออนไลน์ และเครือข่ายคอมพิวเตอร์ (วัลลภา อุทยาน , 2547) รูปแบบของวารสารอิเล็กทรอนิกส์แบ่งเป็น 3 รูปแบบ ได้แก่

3.3.1 วารสารอิเล็กทรอนิกส์ในรูปของฐานข้อมูลระบบออนไลน์ (Online Based Electronic Journal) เป็นวารสารที่มีเนื้อหาฉบับเต็มที่สามารถสืบค้นข้อมูลด้วยระบบออนไลน์จากฐานข้อมูลพาณิชย์ โดย ผู้ใช้สามารถเข้าถึงข้อมูลของผู้ผลิตหรือแหล่งผลิตได้ด้วยการเชื่อมตรง (On-line)

3.3.2 วารสารอิเล็กทรอนิกส์ในรูปฐานข้อมูลซีดี-รอมฉบับเต็ม (CD-ROM Electronic) เป็นเทคโนโลยี การจัดเก็บ การบันทึกข้อมูลในรูปดิจิตอล จัดเป็นสื่อประเภทออปติคอล (Optical media) ที่ใช้แสง เลเซอร์ในการอ่านและบันทึกข้อมูล ซีดี-รอม เป็นสื่อบันทึกข้อมูลชนิดสื่อผสมหรือมัลติมีเดีย (Multimedia) ที่ใช้บันทึกข้อมูล เช่น ตัวอักษร ตัวเลข ข้อความ ภาพ สัญลักษณ์ และเสียง

3.3.3 วารสารอิเล็กทรอนิกส์ในรูปแบบเครือข่าย (Network Electronic Journals) เป็นวารสารใน รูปแบบสื่ออิเล็กทรอนิกส์ฉบับเต็มที่เผยแพร่ และให้บริการในระบบเครือข่ายอินเทอร์เน็ต ปัจจุบัน วารสารอิเล็กทรอนิกส์ที่พบในระบบเครือข่าย สามารถแบ่งได้เป็น 2 ประเภท คือ

1) วารสารที่มีการเสนอเนื้อหาในลักษณะบทความ ข้อมูลในแต่ละฉบับจะประกอบด้วย บทความจากวารสารต่างๆ ซึ่งอาจจะมีการคัดเลือกบทความที่ตีพิมพ์เผยแพร่โดยมีคณะกรรมการ พิจารณา และสามารถบอกรับเป็นสมาชิกวารสารได้เช่นเดียวกับวารสารทางวิชาการที่พิมพ์เผยแพร่ใน รูปสิ่งพิมพ์ เช่น วารสาร Interpersonal Computing and Technology

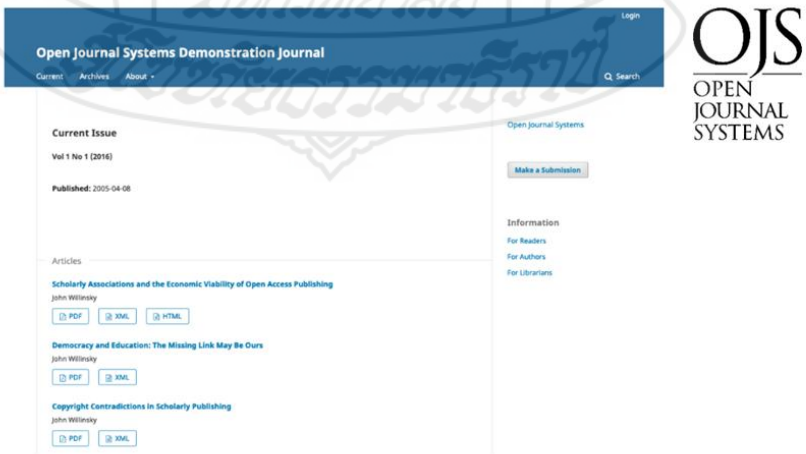
(IPCT) ซึ่งเป็นวารสารที่เกี่ยวกับเทคโนโลยีสารสนเทศ จัดพิมพ์เผยแพร่โดย Center for Teaching and Technology ประเทศสหรัฐอเมริกา

2) วารสารที่มีการเสนอเนื้อหาในลักษณะจดหมายข่าว เป็นวารสารอิเล็กทรอนิกส์ที่ สามารถบอกรับเป็นสมาชิกโดยกลุ่มผู้ใช้บริการข่าวสาร (Listsery) ซึ่งจะให้ข่าวสาร ข้อมูล และมีการแลกเปลี่ยนความรู้ ความคิดเห็นซึ่งกันและกันในลักษณะไปรษณีย์อิเล็กทรอนิกส์ และการประชุม ทางไกลด้วยคอมพิวเตอร์

จากความหมายของวารสารอิเล็กทรอนิกส์ที่กล่าวข้างต้น พอจะสรุปได้ว่า วารสารอิเล็กทรอนิกส์ หมายถึง วารสารที่มีการจัดเก็บ บันทึก และเผยแพร่ผ่านทางเครือข่ายอินเทอร์เน็ตรูปแบบออนไลน์มีการนำเสนอเนื้อหา เหมือนกับวารสารรูปแบบสิ่งพิมพ์ทุกประการ โดยวารสารอิเล็กทรอนิกส์สามารถเข้าถึง สืบค้น ได้ทุกที่ทุกเวลาที่ต้องการไม่เสียค่าใช้จ่าย

4. Open journal system (OJS)

4.1 Open Journal System (OIS) เป็นซอฟต์แวร์ในกลุ่มรหัสเปิด (Open Source) เพื่อใช้ใน การบริหารจัดการวารสาร ระบบการตีพิมพ์และเผยแพร่วารสารวิชาการแบบออนไลน์ พัฒนาโดย Public Knowledge Project (PKP) ได้รับการออกแบบและพัฒนาขึ้นเพื่อให้เป็นระบบที่ช่วย สนับสนุนและเพิ่มประสิทธิภาพในการจัดทำวารสารวิชาการ และจัดเป็นระบบที่มีความยืดหยุ่นสูง มากทั้งในส่วนกระบวนการจัดทำต้นฉบับและการเผยแพร่ (Editorial and Publishing Process) โดยได้นำนวัตกรรมด้านเทคโนโลยีสารสนเทศและการสื่อสารมาช่วยอำนวยความสะดวกสำหรับ ผู้จัดทำวารสาร ในทุกขั้นตอนของการทำงานอย่างครบวงจร (สารานุกรมเสรี วิกิพีเดีย, ม.ป.ป))



ภาพที่ 2.1 ตัวอย่างหน้าต่างการทำงานของ Open journal system (OJS)

4.2 ระบบการทำงานของโปรแกรม Open Journal System หรือ (OJS) สามารถติดตั้งและควบคุมได้ เองโดยอิสระ กองบรรณาธิการสามารถกำหนดความต้องการเกี่ยวกับกรอบการทำงานได้ สำหรับ ขั้นตอนการส่งบทความและการจัดการเนื้อหาสามารถดำเนินการได้แบบออนไลน์ และโมดูลการ บอกรับการเป็นสมาชิกพร้อมกับตัวเลือกสำหรับการเข้าถึงวารสารและบทความ การจัดทำดัชนีเพื่อ การค้นหาเนื้อหา นอกจากนั้นยังมีเครื่องมือช่วยในการอ่านและค้นหาเนื้อหา พร้อมการแจ้งข้อความ ผ่านทางอีเมล์และการส่งข้อคิดเห็นจากผู้อ่านได้ด้วย ส่วนประกอบที่สำคัญ ของ OJS ประกอบด้วย (คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร, ม.ป.ป))

5. Google Colab

Google colab คือโฮสต์โปรแกรม Jupyter notebook บน Cloud ของ Google ชื่อเต็ม คือ Google Colaboratory มีการใช้ภาษา python3 เป็นภาษาหลักที่ใช้ในการเขียนและรันงานบน colab นี้ ทำไมถึงสร้าง Machine Learning บน Google Colab การสร้าง Model จะต้องใช้เครื่องคอมพิวเตอร์ที่มีความเร็วหรือ ประสิทธิภาพสูง เพื่อลดระยะเวลาในการประมวลผลของเครื่อง Google Colab นั้นมีความเร็วหลายเท่าถ้าเปรียบเทียบกับคอมพิวเตอร์ที่ใช้ (yingphan.ch, 2021)

ข้อดี คือ ความเร็วของ CPU และ GPU ใช้ได้ฟรี สามารถเชื่อมต่อกับ Google Drive ได้ รองรับ Tensorflow

ข้อเสีย คือ อาจจะหยุดการทำงานได้ เมื่อมีเวลาการรันต่อครั้งที่ 12 ชั่วโมง ขึ้นไป

6. การจับสำนวนคำ (Word Cloud)

Word cloud คือ กลุ่มคำ เกิดจากการรวมกลุ่มของคำที่มีลักษณะแตกต่างกัน และชัดเจนใน การแสดงภาพที่สวยงาม ซึ่งภายในจะแสดงข้อความต่างๆ ที่มีขนาดเล็ก ใหญ่ขึ้นอยู่กับพารามิเตอร์ที่ ใช้ ที่เรียกว่า text cloud, tag cloud, weighted List หรือ Wordle Word cloud คือ วิธีการแสดง ข้อมูลที่เป็นข้อความสำคัญและจำเป็น คำสำคัญอาจประกอบด้วยคีย์เวิร์ดที่สำคัญและไม่ซ้ำกัน (แมตต์มิลส์, 2564)

ภาพที่ 2.2 แสดงตัวอย่างการจับสำนวนคำ (Word Cloud)

จากภาพด้านบนลักษณะการแสดงคำใน Word cloud พบว่าหากข้อความเด่นชัดและมีขนาดใหญ่มากเท่าใด ระดับดับความเกี่ยวข้องจะยิ่งสูงขึ้น และสำคัญมากขึ้นเท่านั้น จึงเป็นวิธีที่เหมาะสมในการแบ่งและดึงส่วนที่เกี่ยวข้อง โดนเด่น ออกจากข้อมูลหรือข้อความ โดยปกติ Word cloud จะทำหน้าที่แสดงภาพหรือแทนค่าที่ทำได้ในหนึ่งคำ เน้นความเกี่ยวข้องและความสำคัญของทุกข้อความที่เลือกแสดง ถูกกำหนดโดยการใช้สีเฉพาะพร้อมกับรูปแบบตัวอักษรและขนาด ซึ่งระบบ และวิธีการนี้ช่วยให้บุคคลสามารถเปรียบเทียบ และเปรียบเทียบพื้นที่เฉพาะและชัดเจนสองส่วนของ ข้อความที่เลือกได้ ต่อไปทำให้สามารถตรวจจับและค้นหาความคล้ายคลึงกันที่มีอยู่ในถ้อยคำได้

อาจกล่าวได้ว่า Word cloud เป็นวิธีการกรองคำที่จำเป็นหรือเกี่ยวข้องออกเพื่อแยกและดึง เอาคำที่มีความสำคัญหรือความสำคัญสูงสุดและน้อยที่สุดออกสำหรับการเลือกหัวข้อเฉพาะในหมู่คน อื่น ๆ ด้วยวิธีการออกเสียงลงคะแนนเพื่อสร้างวิธีการสื่อสารที่สนุก น่าตื่นเต้น และโต้ตอบได้

Tag cloud / Word cloud คือ การจับสำนวนหรือคำของเนื้อหาบนเว็บไซต์ จากนั้นแยก ลำดับสถิติจำนวนนับเพื่อนำมาจับกลุ่มกันลักษณะคล้ายก้อนเมฆ มีการกำหนดขนาดตัวอักษรให้มี ขนาดใหญ่ ขนาดเล็กแตกต่างกันไป เพื่อแสดงให้ผู้เข้าชมเว็บไซต์ได้เห็นความสำคัญของ คำ (tag) นั้นๆ ตามลำดับ จุดประสงค์หลักๆ เพื่อสร้างความแปลกใหม่ ในการเข้าถึงเนื้อหา และความทันสมัยบนเว็บไซต์ที่ใช้เทคโนโลยีเว็บ 2.0 และกระตุ้นการคลิกเข้าไปดูในเนื้อหาตามคำที่สนใจ

ได้ง่ายขึ้นในปัจจุบันส่วนใหญ่ มักจะเป็นเว็บประเภทบล็อก เครือข่ายสังคม และวิดีโอออนไลน์ โดยเว็บกลุ่มนี้จะใช้แท็กคลาวด์จาก กลุ่มคำค้น (keyword) ที่ผู้เข้าชมพิมพ์ค้นหาภายในเว็บไซต์ (สารานุกรมเสรี วิกิพีเดีย, (ม.ป.ป))

Word cloud คือ กลุ่มคำที่จับตัวกันมีลักษณะเหมือนก้อนเมฆ เป็นเทคนิคที่ใช้สำหรับการจับกลุ่มคำในภาษาใดๆ โดยการนับจำนวนคำจากมากไปหาน้อย จากนั้นทำการแสดงผลโดยคำที่พบ เจอมากก็จะกลุ่มกันเป็นก้อนเมฆที่ใหญ่ และไล่ลำดับลงมาเป็นก้อนเมฆเล็กๆ ตามลำดับ Word Cloud ถือ เป็น Visual Design ที่สวยงาม และดีต่อสายตามากๆ จากการแสดงผลทำให้เรามองเห็นสถิติการใช้คำที่มีจำนวนมากที่สุดไปจนถึงจำนวนน้อยที่สุด (My Little Learn, 2563)

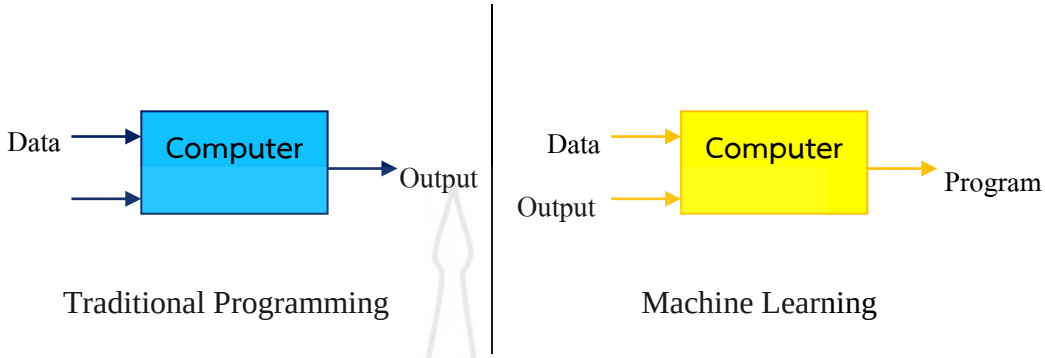
Word cloud มีประโยชน์ในด้านการนับสถิติการแสดงผลกลุ่มคำ เช่น ระบบการรีวิว ของ ร้านอาหาร โรงแรม มีการนำข้อเสนอแนะของผู้เข้าใช้บริการมาวิเคราะห์ว่ามีการกล่าวถึงสิ่งใดมาก ที่สุด หรือการใช้ Word Cloud วิเคราะห์การพูดถึงไวรัส COVID-19 ใน Social Media ว่าบุคคล ทั่วไปมีความสนใจในประเด็นอะไรเกี่ยวกับ COVID-19 มากที่สุด เป็นต้น

Word Cloud หรือกลุ่มคำ คือ การจับกลุ่มคำโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุด เป็น ประโยชน์ในการทำรายงานข้อความ เพื่อให้มองเห็นคำที่ถูกใช้มากที่สุดได้ง่ายขึ้น จะอธิบายให้เข้าใจ ง่ายๆก็คือ Module word cloud จะทำงานโดยการนับ คำที่ซ้ำกันแล้วมาแสดง เช่น กากา มามา หาหา ลาลา มันก็จะแสดงคำว่า กา มา หา ลา ออกมาเป็นตัวหนังสือที่ใหญ่สุด (tatiya, 2562)

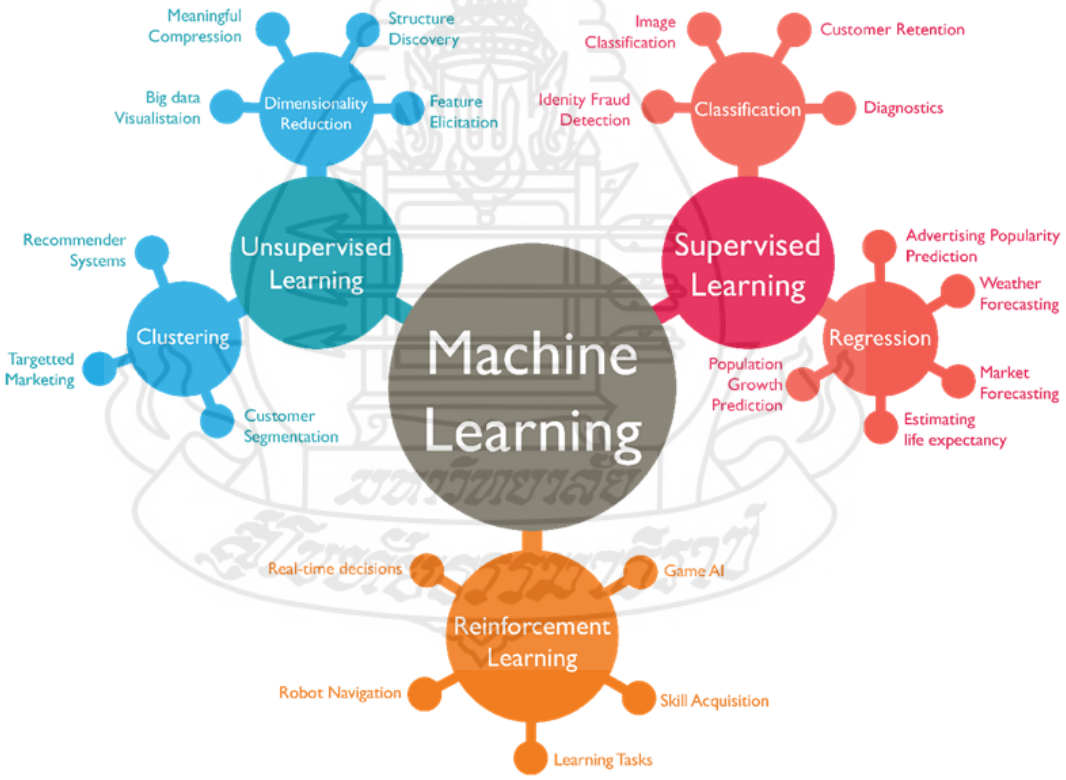
จากความหมายของการจับสำนวนคำที่กล่าวข้างต้น พอจะสรุปได้ว่า การจับสำนวนคำ หมายถึง การรวมกลุ่มคำที่จับตัวกันมีลักษณะเหมือนก้อนเมฆ แสดงเป็นภาพเพื่อสร้างความแปลกใหม่ ในการเข้าถึงเนื้อหา โดยภายในจะแสดงข้อความหรือคำต่างๆ ที่มีขนาดเล็ก ขนาดใหญ่ขึ้นแตกต่างกันอยู่กับพารามิเตอร์ เป็นคำที่ไม่ซ้ำกัน เพื่อแสดงให้ผู้เข้าชมเว็บไซต์ได้เห็นความสำคัญของคำนั้นๆ และประโยชน์ในด้านการทำรายงานข้อความ

7. การเรียนรู้ของเครื่อง (Machine Learning)

7.1 Machine Learning คือ การทำให้ระบบคอมพิวเตอร์เรียนรู้ได้ด้วยตนเอง โดยใช้ “ข้อมูล” ซึ่งจะแตกต่างจากการเขียนโปรแกรมโดยทั่วไป เพราะ Programming จะใส่ ข้อมูล (Data) และ Program เข้าไป เพื่อให้ได้ Output แต่ Machine Learning ไม่ได้ใส่ข้อมูล (Data) และ Output (ผลลัพธ์) เข้าไป เพื่อให้หา Program ที่จะนำไปตอบในอนาคตได้ว่า input แบบนี้ Output จะเป็นอย่างไร (Vithan Minaphinant, 2561)



ภาพที่ 2.3 เปรียบเทียบการทำงาน Traditional Programming และ Machine Learning



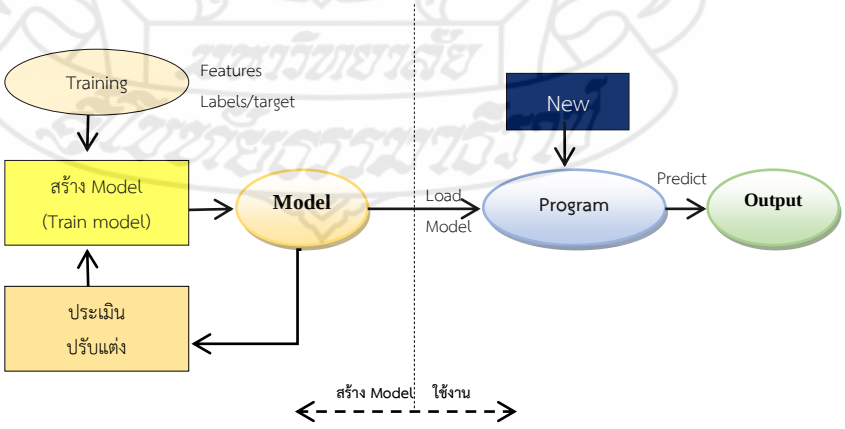
ภาพที่ 2.4 แสดงการแบ่งประเภทของ Machine Learning

Machine Learning สามารถแบ่งออกได้เป็น 3 ประเภท ดังนี้

- Supervised Learning เรียนรู้โดยมี data มาสอน
- Unsupervised Learning เรียนรู้โดยไม่มี data สอน
- Reinforcement Learning เรียนรู้ตามสภาพแวดล้อม

7.2 การเรียนรู้ของเครื่อง เป็นส่วนหนึ่งของปัญญาประดิษฐ์ อาจกล่าวได้ ว่าเป็นการพัฒนาคอมพิวเตอร์ให้สามารถเรียนรู้ได้โดยการสร้างโปรแกรมคอมพิวเตอร์จากการ วิเคราะห์ชุดข้อมูล เพื่อเพิ่มประสิทธิภาพการแก้ไขปัญหาที่อาจเกิดขึ้นในด้านต่างๆ ลักษณะการ เรียนรู้ของเครื่องเป็นการสร้างอัลกอริทึม หรือ โปรแกรมคอมพิวเตอร์ จากการให้ข้อมูลฝึก (Training Data) เพื่อสอนคอมพิวเตอร์ให้เรียนรู้นำมาสู่สมมติฐานเพื่อนำมาใช้ในการแยกแยะวัตถุ สามารถแบ่งเทคนิคการเรียนรู้ของเครื่องได้ 3 ลักษณะ คือ 1)การเรียนรู้แบบมีผู้สอน (Supervised Learning) 2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และ 3)การเรียนรู้แบบกึ่งมีผู้สอน (Semi Supervised Learning) (พัชราภรณ์ สิทธิคำฟู, 2557)

การเรียนรู้ของเครื่อง คือ กระบวนการที่คล้ายกับการสร้างโปรแกรมอย่างหนึ่ง โดยนำเข้า ข้อมูล (Data set) มาสอน (Train) ให้เครื่องคอมพิวเตอร์ เมื่อทำการสอนแล้ว คอมพิวเตอร์จะสร้าง แบบจำลอง (Model) ขึ้นมา ซึ่งเป็นเสมือนกลไก กฎเกณฑ์ หลักการคิดคำนวณประมวลผลของ โปรแกรม ซึ่งบางครั้งเรียกว่า “ส่วนสมองของเครื่องจักร” (ผศ.ดร.กอบเกียรติ สระอุบล, 2563)

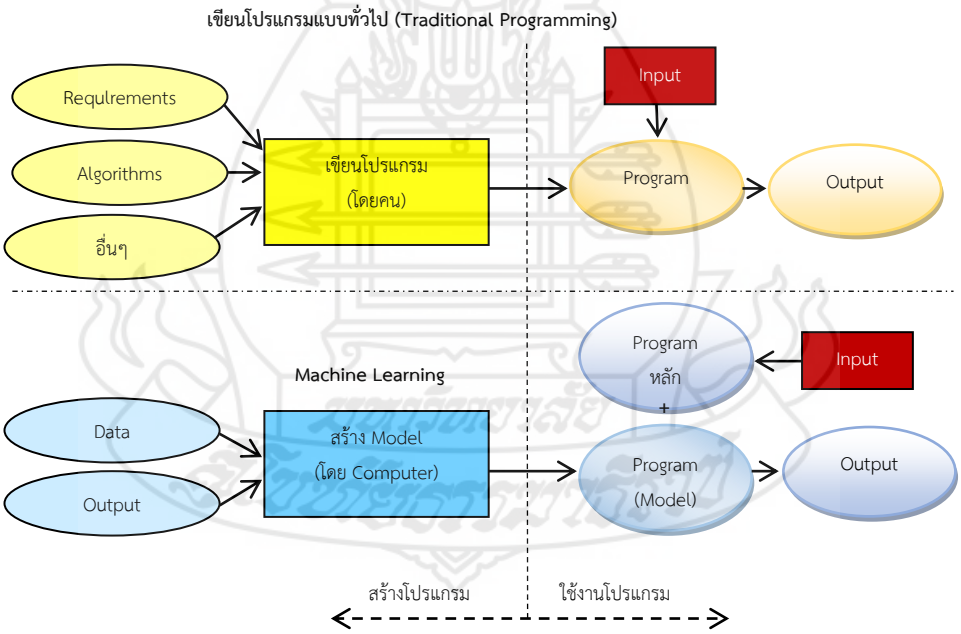


ภาพที่ 2.5 แสดงขั้นตอน Machine Learning

การสร้าง Model เป็นขั้นตอนการนำชุดข้อมูลที่ประกอบด้วยค่า หรือ ตัวแปร ที่เป็นคุณลักษณะเด่น (Features) และ Output (Labels/targets) มาสอนให้กับคอมพิวเตอร์ ขั้นตอนนี้จะทำให้ได้ Model โดยข้อมูลที่นำมาสอนอาจเรียกว่า Training set และขั้นตอนการสอนเรียกว่า Train

การใช้งาน เมื่อสร้างและได้ Model แล้ว ต้องพัฒนาโปรแกรมหลักเพิ่มเติม ทำหน้าที่ Interface รับข้อมูล Input นำเข้าไปประมวลผล หรือ ทำนายผลจนได้ Output ออกมา

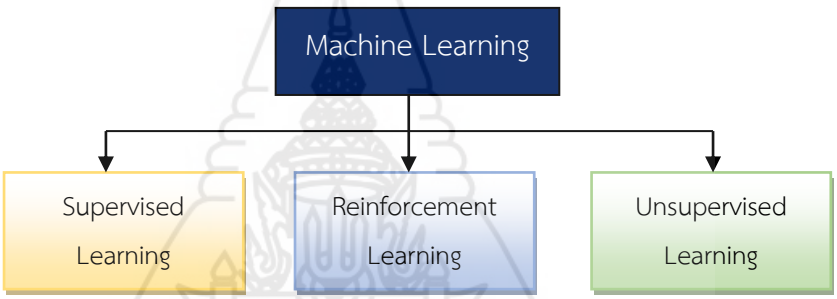
การสร้างหรือเขียนโปรแกรมโดยปกติทั่วไป (Traditional Programming) มีวิธีการคือ ผู้พัฒนาทำการรวบรวม ศึกษาวิเคราะห์ความต้องการของลักษณะงาน รายละเอียดต่างๆ พร้อมทั้งยังเลือกหรือกำหนดวิธีการคำนวณประมวลผลหรืออัลกอริทึม (Algorithm) จากนั้นทำการเขียนโปรแกรมให้มีการทำงาน (Logic) ตามความต้องการ อาจสรุปได้ว่าผู้พัฒนาโปรแกรมเป็นคนสร้างโปรแกรมในส่วนของการคำนวณ การกำหนดเงื่อนไข การตัดสินใจต่างๆ ตามลักษณะของงาน ส่วนในขั้นตอนของการใช้งานโปรแกรมนั้นก็รันโปรแกรมแล้วใส่ Input นำมาสู่ผลลัพธ์ Output



ภาพที่ 2.6 แสดงกระบวนการพัฒนาโปรแกรม

Machine Learning แตกต่างตรงที่ ส่วนโปรแกรมที่ทำการคำนวณประมวลผลการตั้งเงื่อนไข การตัดสินใจต่างๆ (เรียกว่า Model) ถูกสร้างโดยคอมพิวเตอร์ ผู้พัฒนาจะนำชุดข้อมูล (Dataset) ทั้งที่เป็น Data และ Output เข้าไปให้คอมพิวเตอร์เรียนรู้ เพื่อสร้าง Model ซึ่งเป็นเสมือนสมองของโปรแกรม จากนั้นผู้พัฒนาจะเขียนโปรแกรมหลักเพิ่มเติม นำมาสู่โปรแกรมที่สมบูรณ์ สามารถรับ Input ประมวลผล และได้ Output

จะเห็นได้ว่า กระบวนการสร้างโปรแกรมมีส่วนแตกต่างกัน โดย Machine Learning จะให้เครื่องจักร (เครื่องคอมพิวเตอร์) ทำการเรียนรู้จากข้อมูลที่นำเข้าไปสอน แต่จะรู้เพียงเฉพาะส่วนที่เกี่ยวข้องกับสิ่งที่สอน (Train) เท่านั้น Machine Learning สามารถแบ่งเป็น 3 ประเภท ดังนี้



ภาพที่ 2.7 แสดงประเภทของ Machine Learning

Supervised Learning คือ การเรียนรู้แบบมีการสอน (Train) คือทำให้คอมพิวเตอร์เรียนรู้จากข้อมูลที่ได้นำเข้าไปสอน (ข้อมูลที่นำเข้าไปสอน เรียกว่า Training data หรือ Training set) สามารถแบ่งย่อยได้ 2 กลุ่ม คือ Classification เป็นการจำแนก คัดแยก แยกแยะ และ Regression เป็นการคำนวณทำนายค่าเป็นตัวเลข

Unsupervised Learning คือ การเรียนรู้แบบไม่มีการสอน ไม่จำเป็นต้องใช้ข้อมูลที่มี Output (ไม่ต้องมีข้อมูล Label/Target) การทำงานคือป้อนข้อมูลที่ต้องการทำนาย จากนั้นระบบจะทำการประมวลผลข้อมูลให้เอง

Reinforcement Learning คือ การเรียนรู้ที่ต้องอาศัยการป้อนกลับจากนั้นให้ระบบเรียนรู้แล้วปรับปรุงตัวเอง ระบบจะทำการป้อนกลับแล้วนำค่าข้อมูลต่างๆ มาปรับปรุงตัวเอง จนสามารถทำงานได้เสถียรภาพ

8. การจัดกลุ่มข้อมูล

การจัดกลุ่มสามารถแบ่งออกได้หลากหลายวิธี ในส่วนนี้จะกล่าวถึงเทคนิคสำหรับการจัดกลุ่ม ซึ่งแบ่งได้ 2 ประเภท ได้แก่

8.1 การจัดกลุ่มตัวแปร คือ เทคนิคที่ใช้สำหรับการจัดกลุ่มตัวแปรที่มีลักษณะหลายตัว ซึ่งมี ความสัมพันธ์แบ่งออกเป็นกลุ่มย่อย โดยตัวแปรที่อยู่ในกลุ่มหรือปัจจัยเดียวกันนั้นจะต้องมีส่วนร่วม คล้ายหรือสัมพันธ์กันอยู่ สามารถแบ่งได้เป็น 2 เทคนิค ได้แก่ การวิเคราะห์ด้วยประกอบหลัก (Principal Component Analysis) และ การวิเคราะห์ปัจจัย (Factor Analysis)

8.2 การจัดกลุ่มข้อมูล คือ เทคนิคที่ใช้สำหรับจัดกลุ่มคน สัตว์ สิ่งของ องค์กร หรืออื่นๆ ที่มี ลักษณะคล้ายกันหรือเหมือนกันจัดไว้ในกลุ่มเดียวกัน แต่หากต่างกันจะถูกจัดไว้อีกกลุ่ม การศึกษาความ คล้ายกันนั้นพิจารณาจากตัวแปรที่ใช้ในการจัดกลุ่ม มีด้วยกันหลายเทคนิค ในที่นี้จะยกตัวอย่าง 3 เทคนิค ได้แก่ การวิเคราะห์กลุ่ม (Cluster Analysis) การวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) และการวิเคราะห์ด้วยถดถอยโลจิสติก (Logistic Regression Analysis) ซึ่งทั้ง 3 เทคนิคที่ กล่าวไปข้างต้นมีรายละเอียดข้อมูลไม่ว่าจะเป็นวัตถุประสงค์ หลักการที่ใช้ในการวิเคราะห์ที่แตกต่าง กันสำหรับเทคนิคในการวิเคราะห์กลุ่มนั้นสามารถที่จะใช้ในการจัดกลุ่มข้อมูล คน สัตว์ สิ่งของ หรือ อื่นๆ และยังใช้ในการจัดกลุ่มตัวแปรได้อีกด้วย หากแต่ว่าส่วนใหญ่ที่นิยมใช้ในการจัดกลุ่มข้อมูลนั้นมี มากกว่าการใช้จัดกลุ่มตัวแปร โดยในการจัดกลุ่มตัวแปรนั้นนิยมใช้เทคนิคการวิเคราะห์ปัจจัย

8.2.1 ความหมายการวิเคราะห์กลุ่มข้อมูลด้วยเทคนิค (Cluster Analysis) คือ เทคนิคที่ใช้ในการแบ่งกลุ่มหน่วยข้อมูล หรือใช้ในการแบ่งคน สัตว์ สิ่งของ องค์กร อื่นๆ ซึ่งแบ่งออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม โดยจะใช้ หลักเกณฑ์ในการแบ่งกล่าวคือ “ให้หน่วยที่อยู่ในกลุ่มเดียวกันมีลักษณะที่สนใจเหมือนกันหรือ คล้ายกัน แต่หน่วยที่อยู่ต่างกลุ่มกันนั้นจะมีลักษณะสนใจที่ต่างกัน”

คำว่าลักษณะที่สนใจ อาจมีหลายๆ ตัวแปร เช่น สนใจความคิดทางด้านการเมือง จะมีคำถาม หลายคำถามเป็นไปในด้านทางการเมือง จากนั้นจะนำคำตอบเหล่านั้นมาแบ่งกลุ่ม (กัลยา วานิชย์ บัญชา, 2552)

การจัด Case (Case หมายถึง คน สัตว์ สิ่งของ องค์กร หรือ อื่นๆ) เป็นการจัดตัวแปร ออกเป็นกลุ่มย่อยๆ จำนวนตั้งแต่ 2 กลุ่มขึ้นไป Case ที่อยู่ในกลุ่มเดียวกันจะมีลักษณะที่คล้ายหรือ เหมือนกัน ส่วน Case ที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกันออกไป

ตัวแปรที่อยู่ในกลุ่มเดียวกันนั้น จะมีความสัมพันธ์กันที่มากกว่าตัวแปรที่อยู่ต่างกลุ่มกัน ส่วนตัวแปรที่อยู่ต่างกลุ่มกันจะมีความสัมพันธ์กันน้อย หรือ ไม่มีความสัมพันธ์กันเลยก็ได้ เช่น มีวัตถุประสงค์ที่จะจัดสิ่งของ (Object) n. สิ่งให้อยู่เป็นกลุ่มๆ โดยที่สิ่งของที่อยู่ในกลุ่มเดียวกันจะมี ลักษณะคล้ายกัน (Similarity) หรือใกล้ชิดกัน (Closeness) ตัวอย่างเช่น

	ตัวแปรที่ 1	ตัวแปรที่ 2	ตัวแปรที่ K…	ตัวแปรที่ P
หน่วยที่ 1:	X_{11}	X_{12}	X_{1k}	X1p
หน่วยที่ 2:	X_{21}	X_{22}	X_{2k}	X2p
หน่วยที่ j:	X_{j1}	X_{j2}	Xjk	Xjp
หน่วยที่ n:	X_{n1}	X_{n2}	Xnk	Xnp

โดย X_{jk} หมายถึง การวัดหรือค่าครั้งที่ j ของตัวแปร k

จากข้อมูลข้างบนจะเห็นว่าแต่ละหน่วยถูกวัดด้วยตัวแปร p ตัว ซึ่งหน่วยที่ก็คือสิ่งของที่ นั้นเอง ในการวิเคราะห์จัดกลุ่มอาจเรียกทับศัพท์ เช่น item case หรือ object แล้วแต่ว่าคำไหนสื่อความหมายได้มากน้อยกว่ากัน แต่อย่างไรก็ตามในการวิเคราะห์จัดกลุ่มนอกจากจะใช้จัดสิ่งของให้อยู่เป็นกลุ่มๆ แล้วยังสามารถใช้แนวคิดนี้จัดตัวแปร p ตัวให้อยู่เป็นกลุ่ม ๆ ได้อีกด้วย

8.2.2 วัตถุประสงค์การแบ่งกลุ่มข้อมูลด้วยเทคนิคการวิเคราะห์กลุ่ม

การวิเคราะห์กลุ่ม คือเทคนิคที่ใช้ในการจัดกลุ่มโดยไม่ทราบมาก่อนว่าควรมีกี่กลุ่ม จะแบ่ง ตามค่าของตัวแปรที่ได้นำมาใช้ในการแบ่งกลุ่ม โดยให้หน่วยที่อยู่ในกลุ่มเดียวกัน มีความคล้ายกันใน ตัวแปรที่ศึกษา และวัตถุประสงค์ของการจัดกลุ่มหรือจัดกลุ่มจะขึ้นกับสาขาที่จะนำไปประยุกต์ใช้ ดังนี้

1) ด้านการแพทย์ ใช้จัดกลุ่มคนไข้ตามอาการหรือความรุนแรงของโรค เพื่อใช้วิธีการรักษาที่ แตกต่างกันตามความรุนแรงของโรค จัดกลุ่มโรงพยาบาลที่มีประสิทธิภาพคล้ายกันไว้ด้วยกัน และจัด กลุ่มประเทศต่างๆ ตามความเจริญด้านสาธารณสุข โดยใช้ตัวแปรหรือดัชนีด้านสาธารณสุข เช่น อัตรา คนป่วยโรคต่างๆ อายุเฉลี่ย ค่ารักษาพยาบาลเฉลี่ยต่อประชากร 1 คน เป็นต้น

2) ด้านการตลาด ใช้แบ่งผู้บริโภคหรือลูกค้าตามพฤติกรรมการบริโภคสินค้าต่าง ๆ โดยให้ลูกค้า ที่พฤติกรรมการบริโภคหรือการซื้อสินค้าที่คล้ายกันอยู่ในกลุ่มเดียวกัน ส่วนลูกค้าที่มีพฤติกรรมการ บริโภคต่างกันจะอยู่ต่างกลุ่มกัน เมื่อจัดกลุ่มแล้วจะทำให้สามารถ

วางแผนกลยุทธ์ทางการตลาด สำหรับลูกค้าแต่ละกลุ่มได้อย่างมีประสิทธิภาพ ตัวแปรที่จะนำมาใช้ในการจัดกลุ่มอาจใช้ตัวแปรด้าน พฤติกรรมต่าง ๆ ของลูกค้า นอกจากนี้ยังสามารถใช้วางแผนด้านการตลาดในพื้นที่ที่แตกต่างกันได้อีก ด้วย

3) ด้านการศึกษา ใช้จัดกลุ่มนักเรียนตามผลการเรียน (GPAX) ระดับสติปัญญา (IQ) ระดับ การศึกษาของผู้ปกครอง เพื่อให้ได้นักเรียนในกลุ่มเดียวกัน ผลการเรียนระดับสติปัญญา และระดับ การศึกษาของผู้ปกครองใกล้เคียงกัน ส่วนนักเรียนที่อยู่ต่างกลุ่มกันจะมีผลการเรียนระดับ สติปัญญา และการศึกษาของผู้ปกครองต่างกัน เพื่อให้ครูผู้สอนสามารถวางแผนหรือเลือกเนื้อหา และหาวิธีการ สอนตามความเหมาะสมของแต่ละกลุ่ม โดยต่างกลุ่มกันต้องใช้วิธีการสอนที่แตกต่างกันเพื่อให้เกิดผล สัมฤทธิ์มากที่สุด

วัตถุประสงค์ของเทคนิควิธี Cluster Analysis ที่สำคัญมีอยู่ 2 ประการ คือ การจัดกลุ่ม หน่วยวิเคราะห์ การจัดกลุ่มตัวแปร ซึ่งมีความสอดคล้องกับ (กัลยา วานิชย์บัญชา, 2548) และ สามารถกล่าวได้ว่าคือ มีวัตถุประสงค์เพื่อจัดกลุ่ม Case ซึ่งจะเป็นประโยชน์ในงานด้านต่างๆ เช่น ด้านการตลาด ด้านการแพทย์ ด้านการปกครอง เป็นต้น (สุชาติ ประสิทธิรัฐสินธ์, 2540)

8.2.3 คุณสมบัติของเทคนิควิธี Cluster Analysis

คุณสมบัติของเทคนิควิธี Cluster Analysis มีด้วยกันหลายประการ ซึ่งมีรายละเอียด ดังนี้ (สุชาติ ประสิทธิรัฐสินธ์, 2540)

1) ความต้องการทางด้านข้อมูล คือ ส่วนที่ใช้สำหรับการวิเคราะห์จัดกลุ่มหน่วยวิเคราะห์ ผู้วิจัย จะใช้ข้อมูลที่ระบุหน่วยวิเคราะห์และตัวแปรตามที่จัดเก็บมาได้เลย ส่วนในการวิเคราะห์จัดกลุ่มตัว แปร ผู้วิจัยไม่อาจจะใช้แฟ้มข้อมูลได้ แต่สามารถใช้เมตริกแสดงความสัมพันธ์ระหว่างตัวแปรแทนได้

2) แนวคิดพื้นฐาน คือสิ่งสำคัญที่สุดของการวิเคราะห์การจัดกลุ่มคือ ตัวแปรที่ใช้ หากผู้วิจัย ไม่ได้เก็บข้อมูลเกี่ยวกับตัวแปรที่สำคัญๆ ผลที่ได้ก็จะไม่ดีหรือทำให้ผิดพลาดได้ ทั้งนี้อาจเพราะตัวแปร ที่ได้เลือกไว้ตั้งแต่ต้นจะกำหนดคุณสมบัติของสิ่งที่ระบุความเป็นกลุ่มย่อย เช่น การจัดกลุ่มโรงเรียนใน เมือง หากผู้วิจัยไม่เก็บข้อมูลเกี่ยวกับจำนวนนักเรียนและครูขนาดของโรงเรียนก็ไม่อาจเป็นเกณฑ์ที่ใช้ ในการจัดกลุ่มได้

3) ความคล้ายกันของหน่วย คือความคิดเกี่ยวกับความคล้ายของหน่วยที่ศึกษา เป็นเทคนิคของ การวิเคราะห์ทางสถิติซึ่งมีหลายวิธี โดยทั่วไปการวัดความคล้ายจะพิจารณาจากความห่างระหว่างวัตถุ หรือพิจารณาจากความคล้ายกัน

4) การวัดความห่าง คือวิธีการวัดความห่างสามารถวัดได้หลายวิธี และที่นิยมใช้วัดกันมากก็คือ วิธีที่เรียกว่า ระยะห่างเชิงยูคลิดยกกำลังสอง (Squared Euclidean distance) คือ ผลรวมของ ผลต่างยกกำลังสองของทุกตัวแปร

8.2.4 ประเภทของเทคนิค Cluster Analysis

เทคนิคของ Cluster Analysis สามารถแบ่งได้เป็นหลายประเภทหรือเทคนิคย่อย โดย เทคนิคที่นิยมใช้กันมากนั้นมี 2 เทคนิค คือ การวิเคราะห์กลุ่มแบบขั้นตอน (Hierarchical Cluster Analysis) และการวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis หรือ บางครั้งเรียกว่า K - Means Cluster Analysis) นอกจากนี้ ยังมีเทคนิค 2 Step Cluster Analysis ซึ่งเทคนิคดังกล่าวจะมีวัตถุประสงค์ และวิธีการที่แตกต่างกันออกไป

1) เทคนิค Hierarchical Cluster Analysis เป็นเทคนิคที่นิยมใช้กันมากในการจัดกลุ่ม Case หรือจัดกลุ่มตัวแปร โดยมีเงื่อนไขในกรณีที่ใช้ในการแบ่ง Case นั้น จำนวน Case ต้องไม่มากนัก (จำนวน Case ควรต่ำกว่า 200 ถ้าตั้งแต่ 200 ขึ้นไปใช้ K-Means Cluster) และจำนวนตัวแปรต้อง ไม่มากเช่นกัน ไม่จำเป็นต้องทราบจำนวนกลุ่มมาก่อน และไม่จำเป็นต้องทราบว่าตัวแปรใด หรือ Case ใดอยู่กลุ่มใดมาก่อน เทคนิค Hierarchical Cluster สามารถแบ่งเป็น 2 เทคนิคย่อย คือ Agglomerative Hierarchical Cluster Analysis และDivisive Hierarchical Cluster Analysisสำหรับโปรแกรมสำเร็จรูปทั่วไปจะใช้เทคนิค Agglomerative Hierarchical Cluster Analysis

Agglomerative Hierarchical Cluster Analysis ในการเริ่มต้นจะสมมติว่ามี กลุ่มย่อย สิ่งของ หรือ item ที่มีระยะสั้นที่สุด หรือคล้ายกันมากที่สุดจะรวมเข้าด้วยกันเป็นกลุ่มก่อน จึง จะเหลือ n-1 กลุ่มย่อย จากนั้นหาระยะทางหรือความคล้ายจาก n-1 กลุ่มย่อยใหม่แล้วดูว่ากลุ่มย่อย ใดมีระยะทางสั้นที่สุดหรือคล้ายกันมากที่สุดแล้วทำการรวมกลุ่มย่อยนั้นเข้าด้วยกันทำเช่นนี้ต่อไป เรื่อยๆ จนท้ายที่สุดแล้วจะเหลือเพียง 1 กลุ่มซึ่งประกอบด้วยสิ่งของ ก สิ่ง

Divisive Hierarchical Cluster Analysis ในการเริ่มต้นจะสมมติว่ามีกลุ่มที่ประกอบด้วยสิ่งของ หรือ item จำนวน n สิ่ง จากนั้นก็จะแบ่งออกเป็น 2 กลุ่มชนิดที่สิ่งของในกลุ่มมี ระยะทางไกลที่สุด ขั้นต่อไปก็จะมี 3 กลุ่มย่อย ทำเช่นนี้ต่อไปเรื่อยๆ จนที่สุดแล้วจะมี n กลุ่มย่อยซึ่ง แต่ละกลุ่มย่อยประกอบด้วยสิ่งของ 1 สิ่ง หนังสือส่วนมากจะไม่มีรายละเอียดเกี่ยวกับวิธีนี

2) เทคนิค K-Means Clustering เป็นเทคนิคการจำแนก Case ออกเป็นกลุ่มย่อยจะถูกใช้ เมื่อมีจำนวน Case ที่มากโดยจะต้องกำหนดจำนวนกลุ่มหรือจำนวน Cluster ที่ต้องการก่อน เช่น กำหนดให้มี k กลุ่ม เทคนิค K-Means จะมีการทำงานหลายๆ รอบ (iteration)

โดยในแต่ละรอบจะมี การรวม Cases ให้ไปอยู่ในกลุ่มใดกลุ่มหนึ่ง โดยเลือกกลุ่มที่ Case นั้นมีระยะห่างจากค่ากลางของ กลุ่มน้อยที่สุด แล้วจะทำการคำนวณค่ากลางของกลุ่มใหม่ ทำเช่นนี้ไปเรื่อยๆ จนกระทั่งค่ากลางของ กลุ่มไม่มีการเปลี่ยนแปลง หรือครบจำนวนรอบที่กำหนดไว้ ตัวแปรที่ใช้ในเทคนิค K-Means Clustering จะต้องเป็นตัวแปรเชิงปริมาณ คือ เป็นสเกลอันตรภาค (Interval Scale) หรือสเกล อัตราส่วน (Ration Scale) โดยไม่สามารถใช้กับข้อมูลที่อยู่ในรูปความถี่ หรือ Binary เหมือนเทคนิค Hierarchical ได้

ข้อแตกต่างระหว่างเทคนิค Hierarchical กับวิธี K-Means ได้จำแนกข้อแตกต่างระหว่าง เทคนิค Hierarchical กับวิธี K-Means ไว้ดังนี้ (กัลยา วานิชย์บัญชา, 2548) เทคนิค K-Means ใช้เมื่อมีจำนวน Case หรือ จำนวนข้อมูลมาก โดยทั่วไปนิยมใช้เมื่อ $n \geq 200$ เพราะเมื่อ n มาก เทคนิค K-Means จะง่ายกว่า และใช้ระยะเวลาในการคำนวณที่น้อยกว่า การใช้เทคนิค Hierarchical หรือกล่าวได้ว่าเมื่อมีจำนวน Case ไม่มากควรใช้เทคนิค Hierarchical เทคนิค K-Means ผู้ใช้จะต้องกำหนดจำนวนกลุ่มที่แน่นอนไว้ก่อนล่วงหน้า กรณีที่ผู้ วิเคราะห์ยังไม่แน่ใจว่าควรมีกี่กลุ่มจึงจะเหมาะสม ผู้วิเคราะห์อาจจะใช้วิธีใดวิธีหนึ่ง ดังต่อไปนี้ ทำการวิเคราะห์ด้วยวิธี K-Means หลายๆ ครั้ง แต่ละครั้งกำหนดจำนวนที่กลุ่มแตกต่างกัน เช่น เป็น 3, 4 หรือ 5 กลุ่ม แล้วพิจารณาหาจำนวนกลุ่มที่เหมาะสม แต่เมื่อมีข้อมูลมาก วิธีนี้จะทำให้ เสียเวลามาก ควรใช้ข้อมูลบางส่วนทำการวิเคราะห์โดยวิธี Hierarchical เพื่อหาจำนวนกลุ่มที่ควรจะ เป็น จากนั้นจึงใช้เทคนิค K-Means กับข้อมูลทั้งหมดที่มีเทคนิค Hierarchical ผู้วิเคราะห์จะ Standardized ข้อมูลหรือไม่ก็ได้แต่โดยวิธี K-Means จะต้องทำการ Standardized ข้อมูลก่อนเสมอ วิธี K-Means จะหาระยะห่างโดยวิธี Euclidean Distance โดยอัตโนมัติ ในขณะที่วิธี Hierarchical ผู้วิเคราะห์มีสิทธิ์ที่จะเลือกวิธีการคำนวณระยะห่าง หรือความคล้ายได้

9. การจำแนกประเภทข้อมูล (Classification)

9.1 การจำแนกประเภท (Classification) เป็นลักษณะของการสร้างแบบจำลอง หรือโมเดล เพื่อ ใช้ในการจำแนกประเภทข้อมูลจากคุณสมบัติ (Attribute) ของข้อมูล (Class) ให้อยู่ในกลุ่มหรือ หมวดหมู่ที่ได้กำหนดไว้ เช่น การจัดกลุ่มนักเรียนเป็น 3 กลุ่ม ได้แก่กลุ่มดี ปานกลาง และอ่อน โดยทำ การพิจารณาจากผลการเรียนของนักเรียน (ศศิมา มณฑาสุวรรณ, 2557)

การจำแนกประเภท (Classification) คือการจำแนกข้อมูลหรือการแบ่งประเภทของข้อมูล ซึ่งทำการหาต้นแบบหรือสำรวจสิ่งที่เป็นจุดเด่นจุดด้อยในข้อมูลชุดนั้นที่ปรากฏ โดยใช้ข้อมูลจำนวน หนึ่งเพื่อนำมาสร้างต้นแบบ และตัวแบบที่ได้นั้นสามารถนำไปใช้กำหนดข้อมูล และ

จัดสรรข้อมูลทั้ง เก่าและใหม่ของข้อมูลในชุดนั้นได้เหมาะสมว่าควรอยู่หมวดหมู่ใด มีกี่ประเภท อย่างไร (อาจารย์ อนุพงศ์ สุขประเสริฐ, (ม.ป.ป))

9.2 วัตถุประสงค์ของการจำแนกประเภทข้อมูล คือสร้างโมเดลการแยกแอทริบิวท์หนึ่งโดยขึ้นกับ แอทริบิวท์อื่น โมเดลที่ได้มาจากการจำแนกประเภทนั้นจะทำให้เราสามารถพิจารณาคลาสข้อมูลที่ยัง ไม่ได้จัดกลุ่มในอนาคตได้

9.3 การจำแนกประเภทข้อมูล (Data Classification) หมายถึง การเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยการสร้างแบบจำลอง หรือโมเดลในการจำแนกประเภทสำหรับการพยากรณ์ หรือ ทำนายกลุ่มข้อมูลใหม่ ในการสร้างแบบจำลองข้อมูลเกิดมาจากการหาความสัมพันธ์ โดยข้อมูลจะถูกแบ่งไว้เป็น 2 ส่วน คือ ส่วนเรียนรู้ข้อมูล (Training Data) ใช้สำหรับสร้างแบบจำลองจำแนกประเภทข้อมูลใหม่ขึ้น เพื่อให้แบบจำลองที่สร้างได้เกิดการเรียนรู้ข้อมูล และส่วนที่สองใช้ในการทดสอบแบบจำลองที่ได้สร้างขึ้นมา (Testing Data) เป็นชุดข้อมูลประเมินความถูกต้องของแบบจำลองจำแนกประเภทข้อมูล (อนันตต์ชัย ชุติภาสเจริญ และดร.จรัญ แสนราช, 2561)

จากความหมายของการจำแนกประเภทที่กล่าวข้างต้น พอจะสรุปได้ว่า จำแนกประเภทหมายถึง การจำแนกข้อมูลหรือแบ่งประเภทของข้อมูล โดยใช้ข้อมูลจำนวนหนึ่งเพื่อสร้างตัวต้นแบบ และตัวแบบที่ได้นั้นสามารถนำไปใช้กำหนดข้อมูล และจัดสรรข้อมูลทั้งเก่าและใหม่ของข้อมูลในชุดนั้นได้เหมาะสมว่าควรอยู่หมวดหมู่ใด มีกี่ประเภท และสามารถใช้ประกอบการพิจารณาข้อมูลที่ยังไม่ได้จัดกลุ่มในอนาคตได้

10. เทคนิคการจำแนกประเภทข้อมูล (Classification Techniques)

เทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆ ได้มีการถูกกำหนดไว้แล้วสร้างแบบจำลองสำหรับการพยากรณ์ค่าข้อมูล (Predictive Model) ในอนาคต เรียกว่า เทคนิคการเรียนรู้แบบมีผู้สอน (Supervised learning) แบบการจำแนกประเภทข้อมูลโดยจะใช้อัลกอริทึมสำหรับการพยากรณ์หรือทำนายแนวโน้ม ตัวอย่างเช่น (อาจารย์อนุพงศ์ สุขประเสริฐ, (ม.ป.ป.))

10.1 ต้นไม้ตัดสินใจ (Decision Tree) วิธีต้นไม้ตัดสินใจ (Decision Tree) เป็นเทคนิคการเรียนรู้โดยการจำแนกประเภท (Classification) ข้อมูลออกเป็นกลุ่ม (class) ต่างๆ โดยใช้คุณลักษณะ (attribute) ข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบว่า คุณลักษณะใดเป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญ

มากน้อยต่างกันอย่างไร จึงมีประโยชน์ช่วยให้สามารถวิเคราะห์ข้อมูล และตัดสินใจได้ถูกต้องยิ่งขึ้น

ส่วนประกอบของผลลัพธ์ของวิธีต้นไม้ตัดสินใจ

1) โหนดภายใน (internal node) คือ คุณลักษณะต่างๆ ของข้อมูล ซึ่งเมื่อข้อมูลใดๆ ตกลงมาที่โหนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปในทิศทางใด โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก

2) กิ่ง (branch, link) เป็นค่าของคุณลักษณะในโหนดภายในที่แตกกิ่งนี้ออกมาซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าของคุณลักษณะในโหนดภายในนั้น

3) โหนดใบ (leaf node) คือกลุ่มต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล

ต้นไม้ตัดสินใจ (Decision tree learning) เป็นวิธีการเรียนรู้ซึ่งใช้ในสถิติ, การเรียนรู้ของเครื่อง และการทำเหมืองข้อมูล โดยพิจารณาการสังเกตการแบ่งแยกของข้อมูลโดยการพิจารณาข้อมูล ในการเรียนรู้ของเครื่อง (machine learning) ต้นไม้ตัดสินใจ เป็นโมเดลทางคณิตศาสตร์ที่ใช้ทำนายประเภทของวัตถุโดยพิจารณาจากลักษณะของวัตถุ บัพภายใน (inner node) ของต้นไม้จะแสดงตัวแปร ส่วนกิ่งจะแสดงค่าที่เป็นไปได้ของตัวแปร ส่วนบัพใบ (leaf node) จะแสดงประเภทของวัตถุ ต้นไม้ตัดสินใจที่บัพใบแสดงถึงข้อมูลที่เป็นข้อมูลไม่ต่อเนื่อง (discrete values) จะเรียกว่าต้นไม้ตัดสินใจแบบจำแนก (classification trees) และต้นไม้ตัดสินใจที่บัพใบเป็นข้อมูลต่อเนื่อง (continuous values) จะเรียกว่าต้นไม้ตัดสินใจแบบถดถอย (regression trees) (สารานุกรมเสรีวิกิพีเดีย. (ม.ป.ป))

การพัฒนาขั้นตอนวิธี (algorithm) ในการสอน (training) ต้นไม้การตัดสินใจในปัจจุบันนั้นมีมากมาย ซึ่งมาจากวิธีพื้นฐานวิธีหนึ่งซึ่งเป็นการค้นหาแบบละโมภ (greedy search) จากบนลงล่าง (top-down) ชื่อว่า ID3 ซึ่งถูกพัฒนาโดย John Ross Quinlan ในปี 1986

เอนโทรปี (Entropy) ID3 สร้างต้นไม้การตัดสินใจจากบนลงล่างด้วยการถามว่าลักษณะใด (“ลักษณะ” แทน “ตัวแปรต้น”) เป็นรากของต้นไม้การตัดสินใจ และถามซ้ำๆ ไปเรื่อยๆ เพื่อหาต้นไม้ทั้งต้นด้วยการเขียนโปรแกรมด้วยความสัมพันธ์แบบเวียนเกิด (recursion) โดยในการเลือกว่าลักษณะใดดีที่สุดนั้นดูจากค่าของลักษณะเรียกว่าเกนความรู้ (Information gain) ก่อนที่จะรู้จักเกนความรู้จะต้องนิยามค่าหนึ่งที่ใช้บอกความไม่บริสุทธิ์ของข้อมูลก่อน เรียกว่าเอนโทรปี (Entropy) โดยนิยามเอนโทรปีของต้นไม้การตัดสินใจในตัวในเซตของตัวอย่าง S คือ E (S) ดังนี้

$$E(S) = \sum_{j=1}^{n} ps(j)\ log_2\ ps(j) \qquad (2\text{-}1)$$

โดยที่

S คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลายๆ กรณี

$ps(j)$ คือ อัตราส่วนของกรณี S ที่ตัวแปรตาม หรือ ผลลัพธ์มีค่า j

โดยสำหรับต้นไม้การตัดสินใจที่มีผลลัพธ์เป็นแค่เพียงค่าตรรกะ (Boolean) ใช่ กับ ไม่ใช่ จะมีเอนโทรปี คือ

$$E(S) = -p_{yes}\ log_2(p_{yes}) - p_{no}\ log_2(p_{no}) \qquad (2\text{-}2)$$

จะเห็นว่าเอนโทรปีจะมีค่าอยู่ระหว่าง 0 กับ 1 โดยจะมีค่าเป็นศูนย์เมื่อทุกๆ กรณีมีผลลัพธ์เพียงแบบเดียว เช่น ใช่ทั้งหมด หรือ ไม่ใช่ทั้งหมด และจะมีค่ามากขึ้นเมื่อเริ่มมีค่าที่แตกต่างกันมากขึ้น หรือเอนโทรปีจะมีค่ามากขึ้นหากข้อมูลไม่บริสุทธิ์ และจะตัดสินใจได้ว่าผลลัพธ์จะเป็นอะไรเมื่อเอนโทรปีเป็น 0 เท่านั้น

10.2 นาอีฟเบย์ (Naïve Bayes) นาอีฟเบย์ เป็นเครื่องจักรเรียนรู้ที่อาศัยหลักการความน่าจะเป็น ตามทฤษฎีของเบย์ (Bayes Theorem) ซึ่งมีอัลกอริทึมที่ไม่ซับซ้อน เป็นขั้นตอนวิธีในการจำแนกข้อมูล โดยเรียนรู้จากการเกิดเหตุการณ์หรือปัญหาต่างๆที่เกิดขึ้น เพื่อนำมาวิเคราะห์สร้างเงื่อนไขการจำแนกข้อมูลใหม่ โดยหลักการของนาอีฟเบย์จะการคำนวณหาความน่าจะเป็นในการทำนายผล เป็นเทคนิคในการแก้ปัญหาแบบการจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้ จะทำการวิเคราะห์ความสัมพันธ์ ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน (อนัตต์ชัย ชุติภาสเจริญ และ ดร.จรัญ แสนราช, 2561)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2\text{-}3)$$

โดยที่

$P(A|B)$ คือ ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้น ถ้าเหตุการณ์ B เกิดขึ้นแล้ว

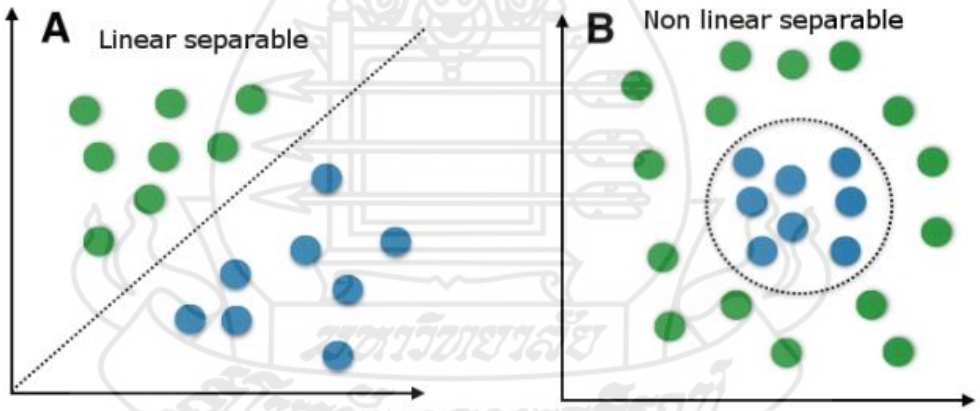
$P(B|A)$ คือ ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น ถ้าเหตุการณ์ A เกิดขึ้นแล้ว

$P(A)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A

$P(B)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B

10.3 Support Vector Machines (SVM) คือ อัลกอริทึมที่สามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด (Jaruwit Pratancheewin, 2562)

Support Vector Machine คือ Algorithm แบบ Supervised Learning ที่ใช้สำหรับแก้ปัญหาการจัดกลุ่มข้อมูล Classification และการวิเคราะห์การถดถอย Regression ซึ่งจะมีความคล้ายคลึงกับ Logistic Regression (LR) (natthasath, 2561)

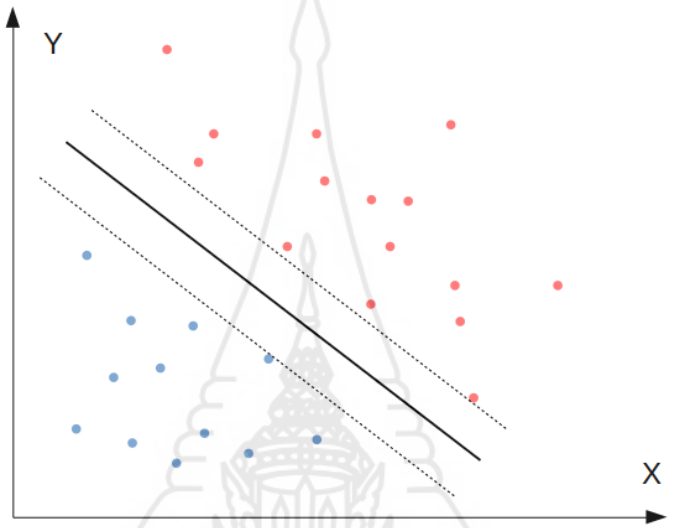


ภาพที่ 2.8 แสดงการแบ่ง Support Vector Machine ด้วยสมการเชิงเส้น

Support Vector Machine จะเป็นการจัดกลุ่มข้อมูล Classification โดยการแบ่ง Class ของข้อมูลออกจากกัน ซึ่งสามารถใช้การแบ่งด้วยสมการเชิงเส้นได้ทั้ง Linear และ Non Linear ดังภาพด้านบน

Support Vector Machines หรือ SVM ยืดหยุ่นและทำงานได้ดี โดยเฉพาะอย่างยิ่งเมื่อข้อมูลมีความซับซ้อน (หลาย Feature) แต่จำนวนตัวอย่างไม่มาก (ต่ำกว่าแสนรายการ) แม้จะถูก

ออกแบบมาสำหรับ Binary classification แต่สามารถที่จะนำไปประยุกต์ใช้กับ Multiclass classification และ Linear regression ได้โดยใช้หลักการเดิมแต่เปลี่ยนรายละเอียดเล็กน้อย ซึ่งในระดับการใช้งาน สามารถใส่ข้อมูลแบบ Multiclass ลงไปได้เลย ส่วน SVM สำหรับ Linear regression ให้เรียกใช้ Class LinearSVR จากโมดูล sklearn.svm (ชิตพงษ์ กิตตินราดร, 2563)



ภาพที่ 2.9 Support Vector Machines (SVM)

จากภาพเป็นปัญหา Binary classification ต้องการจำแนกข้อมูลออกเป็นสองกลุ่ม คือสีน้ำเงินและสีแดง สิ่งที่ SVM ทำ คือการหาเส้นแบ่งการตัดสินใจที่เป็นเส้นทึบ ซึ่งเส้นนี้จะเกิดขึ้นระหว่างกลางของเส้นประด้านซ้ายและขวา โดยมีเงื่อนไขว่าจะต้องหาคู่ของเส้นประที่กว้างที่สุดเท่าที่จะเป็นไปได้ จะมีสองแบบ คือ 1) Hard margin classification คือคู่เส้นประที่ห้ามไม่ให้มีจุดข้อมูลอยู่ในพื้นที่ระหว่างเส้นประ และ 2) Soft margin classification คืออนุญาตให้มีข้อมูลอยู่ในพื้นที่ระหว่างเส้นประได้บ้าง

Support Vector Machines หรือ SVM ใช้ Hypothesis function แบบเส้นตรงเหมือนกับ Linear regression

$$h_\theta(x) = w_1 x_1 + w_2 x_2 + \cdots w_n x_n + b \\ = w^T x + b \qquad (2\text{-}4)$$

โดยถ้าผลลัพธ์เป็นบวก จะทำนาย Class $\hat{y}$ ว่าเป็น 1 ถ้าเป็นลบจะทำนายว่าเป็น 0 สามารถเขียนวิธีการตัดสินใจตามเงื่อนไขได้ดังนี้

$$\hat{y} = \begin{cases} 0 \text{ if } w^T x + b < 0, \\ 1 \text{ if } w^T x + b \geq 0 \end{cases} \quad (2\text{-}5)$$

เมื่อนิยามเส้นแบ่งการตัดสินใจแล้ว (เส้นทึบ) กำหนดเส้นประทั้งสองด้านของเส้นทึบ โดยเส้นประแต่ละด้านคือตำแหน่งที่ $h_\theta(x)$ เท่ากับ -1 และ 1

10.4 การถดถอยโลจิสติก (Logistic regression)

การถดถอยโลจิสติก (logistic regression) เป็นเทคนิคการวิเคราะห์ทางสถิติใช้สำหรับการพยากรณ์ความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจจากชุดตัวแปรอิสระ ซึ่งเป็นที่รู้จักกันดีในหมู่นักสถิติ โดยเริ่มใช้ในการวิจัยทางการแพทย์และสาธารณสุขก่อน และแพร่หลายเข้าสู่การวิจัยในสาขาสังคมศาสตร์ และพฤติกรรมศาสตร์ในภายหลังเมื่อมีโปรแกรมคอมพิวเตอร์ที่สามารถวิเคราะห์โมเดลนี้ได้ เช่น GLIM BMDP SAS และ SPSS (ศิริเดช สุชีวะ, 2558)

การวิเคราะห์การถดถอยโลจิสติค มีลักษณะคล้ายการวิเคราะห์ถดถอยโดยทั่วไป เป็นการหาความสัมพันธ์ระหว่างตัวแปรตาม และตัวแปรอิสระ ซึ่งเมื่อได้แบบแผนความสัมพันธ์ (สมการพยากรณ์) แล้ว สามารถนำแบบแผนดังกล่าวไปใช้ประมาณค่าตัวแปรตามได้โดยข้อแตกต่างระหว่างการวิเคราะห์การถดถอยโลจิสติค และ การวิเคราะห์ถดถอยโดยทั่วไป ก็คือ ตัวแปรตาม (Y) เป็นตัวแปรเชิงกลุ่ม แต่ในการวิเคราะห์ถดถอยโดยทั่วไปตัวแปรตาม (Y) จะต้องเป็นตัวแปรเชิงปริมาณ (สืบค้นจาก https://dspace.bru.ac.th/xmlui/bitstream/handle/123456)

10.4.1 ประเภทของการวิเคราะห์การถดถอยโลจิสติค แบ่งเป็น 2 ประเภท คือ

1) Binary Logistic จะใช้เมื่อตัวแปรตาม เป็นตัวแปรเชิงกลุ่มที่มีค่าได้เพียง 2 ค่า เช่น Y= 1 ถ้านักศึกษาสอบผ่าน หรือ = 0 ถ้านักศึกษาสอบไม่ผ่าน

2) Multinomial Logistic จะใช้เมื่อตัวแปรตาม เป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ค่า เช่น Y = 1 หมายถึงไม่เป็นโรคมะเร็ง Y = 2 หมายถึงกรเป็นมะเร็งขั้นต้น ...Y = 5 หมายถึงการ เป็นมะเร็งขั้นสุดท้ายสำหรับในบทนี้จะพิจารณาเฉพาะกรณีที่ตัวแปรตาม Y มีค่าเพียงแค่สองค่า (dichotomous) ซึ่งเรียกว่า Binary Logistic Regression

วิธีการวิเคราะห์ข้อมูลด้วยเทคนิค Logistic Regression เป็นการประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ในกรณีที่มีตัวแปรอิสระเพียงตัวเดียว Logistic Regression Model

$$\text{Prob (event)} = \frac{e^{(\beta_0+\beta_1\times)}}{1+e^{(\beta_0+\beta_1\times)}} \quad (2\text{-}6)$$

หรือ $$\text{Prob (event)} = \frac{1}{1+e^{-(\beta_0+\beta_1\times)}} \quad (2\text{-}7)$$

โดยที่ β_0 และ β_1 คือ ค่าพารามิเตอร์ เมื่อทำการประมาณค่าได้

X คือ ตัวแปรอิสระ

e คือ ค่า natural logarithm ในทางคณิตศาสตร์มีค่าประมาณ 2.71828 หากกรณีมีตัวแปรอิสระมากกว่า 1 ตัวแปร สามารถเขียนได้ดังนี้

$$\text{Prob (event)} = \frac{e^{(\beta_0+\beta_1\times_1+\beta_2\times_2+\beta_3\times_3+\cdots+\beta_p\times p)}}{1+e^{(\beta_0+\beta_1\times_1+\beta_2\times_2+\beta_3\times_3+\cdots+\beta_p\times p)}} \quad (2\text{-}8)$$

หรือ $$\text{Prob (event)} = \frac{1}{1+e^{(\beta_0+\beta_1\times_1+\beta_2\times_2+\beta_3\times_3+\cdots+\beta_p\times p)}} = \pi \quad (2\text{-}9)$$

เมื่อได้ค่าความน่าจะเป็นของการเกิดเหตุการณ์ก็จะสามารถคำนวณค่าความน่าจะเป็นของการไม่เกิดเหตุการณ์นั้นได้คือ

$$\text{Prob(no event)} = 1 - \frac{1}{1+e^{(\beta_0+\beta_1\times_1+\beta_2\times_2+\beta_3\times_3+\cdots+\beta_p\times p)}} = 1 - \pi \quad (2\text{-}10)$$

11. การวัดประสิทธิภาพแบบจำลอง

การประเมินค่าความแม่นยำของแบบจำลองสำหรับการสร้างความน่าเชื่อถือให้แบบจำลองที่ได้พัฒนาขึ้น เป็นการนำวิธีการตรวจสอบไขว้กัน (K-Fold Cross-Validation) เป็นวิธีการตรวจสอบค่าความผิดพลาด ในการคาดการของแบบจำลองโดยแบ่งข้อมูลออกเป็น K กลุ่ม (K-Fold) เท่าๆ กัน ในขั้นตอนแรกเลือกข้อมูลกลุ่มที่เป็นข้อมูลสำหรับทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลสำหรับเรียนรู้ นำข้อมูลไปจัดหมวดหมู่ จากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบ และข้อมูลกลุ่มอื่น ๆ ที่เหลือเป็นชุดสำหรับเรียนรู้ทำเช่นนี้ไปเรื่อยๆ จนครบ K กลุ่ม

ขั้นตอนสุดท้ายจะหาค่าเฉลี่ยของค่าความถูกต้องในแต่ละกลุ่มวิธีการนี้ข้อมูลทุกตัวอย่างจะเป็นทั้งสำหรับเรียนรู้ และสำหรับการทดสอบ

การวัดประสิทธิภาพของแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ สามารถพิจารณาได้จากค่าความถูกต้อง โดยวัดที่ประสิทธิผลของการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศ คือ Confusion Matrix ประกอบด้วย ค่าความถูกต้อง (Accuracy) ค่า Sensitivity ของคลาสหลัก (True Positive Rate) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure) (พัชราภรณ์ สิทธิคำฟู, 2557)

ทำนาย (Prediction)

ของจริง (Actual)	Positive (1)	Negative (0)
Positive (1)	True Positive (TP)	False Positive (FP)
Negative (0)	False Negative (FN)	True Negative (TN)

ภาพที่ 2.10 Confusion Matrix

โดยที่

TP คือ สิ่งที่ทำนาย ตรง กับสิ่งที่เกิดขึ้นจริง ในกรณีทำนายว่า จริง และสิ่งที่เกิดขึ้ คือ จริง
TN คือ สิ่งที่ทำนาย ตรง กับสิ่งที่เกิดขึ้น ในกรณีทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น คือ ไม่จริง
FP คือ สิ่งที่ทำนาย ไม่ตรง กับสิ่งที่เกิดขึ้น ในกรณีทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง
FN คือ สิ่งที่ทำนาย ไม่ตรง กับที่เกิดขึ้นจริง ในกรณีทำนายว่า ไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

สามารถใช้ Confusion Matrix คำนวนการประเมินประสิทธิภาพของการทำนายด้วยแบบจำลองโดย Accuracy คือ การหาค่าความถูกต้องที่เราทายได้ตรงกับสิ่งที่เกิดขึ้นจริง

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (2\text{-}11)$$

Precision คือ การคำนวณค่าความแม่นยำเป็นการเปรียบเทียบ การทำนายที่ถูกต้องว่า จริง และก็เกิดขึ้นจริง (TP) กับ การทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง (FP)

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2\text{-}12)$$

Recall คือ การคำนวณหาค่าความระลึกของการทำนายว่าเป็น “จริง” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งทำนาย และ เกิดขึ้น ว่า “เป็นจริง”

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2\text{-}13)$$

F1 score คือค่าเฉลี่ยระหว่าง precision และ recall จุดประสงค์ของการสร้าง F1 ขึ้นมา คือ เพื่อเป็น single metric ที่วัดความสามารถของโมเดล

$$\textbf{F1 = 2 x (Precision x Recall) / (Precision + Recall)} \quad (2\text{-}14)$$

F-Measure ค่าความถ่วงดุล คือ การวัดประสิทธิภาพโดยรวมของทั้งสองค่าระหว่างค่าความแม่นยำและค่าความระลึก พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

$$F - \text{Measure} = \frac{2*\text{Precision}+\text{Recall}}{\text{Precision}+\text{Recall}} \quad (2\text{-}15)$$

12. งานวิจัยที่เกี่ยวข้อง

ที่	ชื่องานวิจัย/ผู้วิจัย	ระบบ/ชิ้นงาน	วิธีการวิเคราะห์ ออกแบบ พัฒนา	อัลกอริทึม/เทคนิค
1	การสกัดข้อมูลเทคนิคการเรียนรู้ของเครื่องโดยใช้เทคนิคการทำเหมืองข้อความ : กาญจนา สุดาทิพย์ (2559)	Web Application	-Text Analysis -Natural Language Processing -Named-entity recognition -Processing Module -Extraction Module -Generator Module -Use Case -Sequence Diagram	-Decision Trees -Naive Bayes -Artificial Neural Network -Text Mining -Ontology Based Knowledge
2	การวิเคราะห์ข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความ : ทิชากร เนตรสุวรรณ์ (2559)	Web Application	-Crime Data Extraction -Text Mining -Natural Language Processing -Online Analytical Processing -Data warehouse	-Decision Trees -Artificial Neural Network -Naive Bayes -Text Mining -GATE Framework -Neural Network
3	การประยุกต์ใช้เหมืองข้อความร่วมกับความสัมพันธ์ของคำโดยกราฟในการให้คำแนะนำ : วัชรีวรรณ จิตต์สกุล (2559)	Web Application	- Graph Theory -Text Mining Process	-Decision Trees -Naive Bayes -Support Vector Machine - k-Nearest Neighbor -Random Forest -Bayesian Logistic Regression
4	การตัดคำภาษาไทยโดยใช้เหมืองข้อมูลของข้อความ: เกรียงไกร เชาว์นิธิ (2558)	Web Application	-Flowchart -Natural Language Processing -Text Mining	-Hybrid Method

ที่	ชื่องานวิจัย/ ผู้วิจัย	ระบบ/ ชิ้นงาน	วิธีการวิเคราะห์ ออกแบบ พัฒนา	อัลกอริทึม/เทคนิค
5	แบบจำลองการจำแนกไอซีดี-เทน ทีเอ็ม ข้ามภาษาโดยใช้เหมืองข้อความ: พรรัศมิ์ จตุนราพิทย์ (2558)	Web Application	-Use Case -Sequence Diagram -Text Mining	-Decision Trees -Naive Bayes -Support Vector Machine
6	ระบบจำแนกหมวดหมู่การแจ้งซ่อมบ้านออนไลน์โดยใช้เหมืองข้อความ : ประเดิม วงค์กระโซ่ (2560)	Web Application	-Flowchart -Use case -Text Mining	-Decision Trees -Naive Bayes - K-Nearest Neighbor
7	การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ : พรรณนาภรณ์ เกตุภู่พงษ์ (2561)	Web Application	-Use case -Text Mining	-Decision Trees -Naive Bayes -Support Vector Machine -Artificial Neural Network -Hold Method

ที่	ชื่องานวิจัย/ผู้วิจัย	ระบบ/ชิ้นงาน	วิธีการวิเคราะห์ ออกแบบ พัฒนา	อัลกอริทึม/เทคนิค
8	การสกัดปัจจัยที่ส่งผลต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุในระบบรถไฟด้วยเหมืองข้อความ : ธันยพร กรานสุข (2560)	Web Application	- CRISP-DM -Text Mining	-Naive Bayes -Support Vector Machine
9	ระบบวิเคราะห์ความคิดเห็นต่อธุรกิจด้วยการทำเหมืองข้อความบนทวิตเตอร์ : ชนิตา ลิสิริกุล (2560)	Web Application	-ER Diagram -Text Mining	-Naive Bayes -Support Vector Machine
10	เว็บท่าเชิงความหมายสำหรับการจัดกลุ่มข้อมูลโดยใช้เทคนิคเหมืองข้อความ กรณีศึกษามะเร็ง: สุภาพร วีระพันธ์ยานนท์ (2561)	Web Application	-Graph Visualization -Text Extraction -Text Mining	-Hierarchical
11	แบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้: ประภัสสร ข่ายกระโทก (.....)	Open Journal System (OJS)	-Text classification - Word cloud	Decision Trees -Support Vector Machine -Naive Bayes -Logistic regression

12.1 การสกัดข้อมูลเทคนิคการเรียนรู้ของเครื่องโดยใช้เทคนิคการทำเหมืองข้อความ : กาญจนา สุดาทิพย์ (2559) ได้ทำวิจัยการสกัดข้อมูลเทคนิคการเรียนรู้ของเครื่องโดยใช้เทคนิคการทำเหมืองข้อความ โดย ประยุกต์ใช้การทำเหมืองข้อความกับการใช้ฐานความรู้ออนโทโลยีในการสกัดข้อมูลและจัดเก็บคำสำคัญภายใต้ขอบเขต แนวคิดการเรียนรู้ของเครื่อง เพื่อระบุขอบเขตของข้อมูลที่นำมาใช้ เช่น งานวิจัยส่วนใหญ่จะอยู่ในรูปแบบไฟล์ PDF พบว่ามีปัญหาการ สะกดคำผิด มีปัญหาเกี่ยวกับสระและวรรณยุกต์ และลักษณะคำในภาษาไทยที่ไม่มีการเว้นช่องว่างเหมือนภาษาอังกฤษอาจจะมีปัญหาในการตัดคำ จึงต้องใช้ฐานความรู้ออนโทโลยีมาช่วยในการสกัดข้อมูล หลังจากทราบปัญหาแล้วจึงทำการรวบรวมข้อมูลที่ใช้ในงานวิจัยทั้งภาษาไทยและภาษาอังกฤษ จากเว็บไซต์ Thai Journal Online (ThaiJo) เว็บไซต์ Thailand Library Integrated System (Thailis) เว็บไซต์ ACM Digital Library (ACM) และเว็บไซต์ IEEE Xplore Digital Library (IEEE) รวมทั้งหมด 607 ฉบับซึ่งในการทำงานประกอบไปด้วย3 ส่วนสำคัญ ได้แก่ การเตรียมข้อมูล (Processing Module) การสกัด ข้อมูล (Extraction Module) และการแปลผลลัพธ์การสกัดข้อมูล (Generator Module) ซึ่งในแต่ละส่วนจะประกอบไป ด้วยขั้นตอนย่อย ๆ ในส่วนการเตรียมข้อมูล เช่น การสกัดข้อความจากไฟล์ PDF การตัดแบ่งคำโดยใช้เล็กซิตรอน (Lexitron) การกำจัดคำหยุด การแปลผลลัพธ์การสกัดข้อมูลโดยนำข้อมูลมาเรียงตามลำดับความถี่ เป็นต้น หลังจากการออกแบบและพัฒนาระบบสกัดข้อมูล ผู้วิจัยได้ทำการประเมินประสิทธิภาพโดยใช้ผู้เชี่ยวชาญ 3 ท่าน และเปรียบเทียบผลลัพธ์ที่ได้จากบทความชุดเดียวกัน ซึ่งการประเมินแบ่งได้เป็น 3 ส่วน ได้แก่ 1) ประเมินขั้นตอนวิธีการสกัดข้อมูล พบว่ามีผลการประเมินโดยภาพรวมอยู่ในระดับดี (ค่าเฉลี่ยเท่ากับ 4.22 คะแนนและค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.9) 2) ประเมินการใช้งานระบบสกัดข้อมูล พบว่ามีผลการประเมินโดยภาพรวมอยู่ในระดับดี (ค่าเฉลี่ยเท่ากับ 4.11 คะแนนและค่าส่วนเบี่ยงเบนมาตรฐาน 0.29) และ 3) ผลลัพธ์การสกัดข้อมูลจากระบบเปรียบเทียบกับผลลัพธ์ที่ได้จากผู้เชี่ยวชาญโดยใช้บทความวิชาการชุดเดียวกัน พบว่ามีค่าความความถูกต้องเฉลี่ยของบทความวิชาการภาษาไทยอยู่ที่ร้อยละ 71.34 และบทความวิชาการภาษาอังกฤษอยู่ที่ร้อยละ 87 จึงแสดงให้เห็นว่าการออกแบบขั้นตอนวิธีและการพัฒนาระบบสกัดข้อมูลพัฒนาขึ้น สามารถที่จะนำไปประยุกต์ใช้ให้เกิดประโยชน์ด้านการสกัดข้อมูลได้อย่างมีประสิทธิภาพ

12.2 การวิเคราะห์ข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความ : ทิชากร เนตรสุวรรณ์ (2559) ได้เสนอการวิเคราะห์ข่าวภาษาอังกฤษด้านอาชญากรรมออนไลน์ด้วยเทคนิคการทำเหมืองข้อความและเทคนิคการประมวลผลภาษาธรรมชาติ

(GATE Framework) และนำเสนอกรอบการทำงานสำหรับการจำแนกประเภทข้อมูลข่าวฯ เป็น 5 หมวดหมู่ มาช่วยแก้ปัญหาในการสกัดคำสำคัญ (Keyword) ที่ซ่อนอยู่ในเนื้อหาข่าว และจำแนกประเภทของข่าวอาชญากรรม นอกจากนี้ช่วยแก้ในด้านข้อมูลที่มีเพิ่มมากขึ้นทำให้เกิดการซับซ้อนของข้อมูลส่งผลให้ประสิทธิภาพการนำเสนอข้อมูลอาจผิดพลาดตามไปด้วย และช่วยในการรายงานข่าวอาชญากรรมของสำนักงานตำรวจที่มีเพียงการสรุปผลรายงานสถิติไม่มีการนำเสนอในรูปแบบอื่นที่หลากหลาย เช่น นำเสนอในรูปแบบตาราง อาทิ ช่วงเวลา ประเภทของการเกิดเหตุ ซึ่งการนำเสนอแบบหลายมิติต้องใช้ระยะเวลาในการรวบรวมและเก็บข้อมูล โดยผู้วิจัยได้พัฒนาต้นแบบคลังข้อมูล (Data warehouse) สำหรับเก็บข้อมูลอาชญากรรม ช่วยจัดเก็บข้อมูลที่มีปริมาณมาก ลดปัญหาเรื่องของประสิทธิภาพในการประมวลผลข้อมูล และสนับสนุนการประมวลผลในเชิงวิเคราะห์แบบออนไลน์ (Online Analysis Processing - OLAP) ซึ่งจะสามารถนำไปใช้ในระบบสนับสนุนการตัดสินใจได้มีประสิทธิภาพช่วยวิเคราะห์ข้อมูลในด้านการเปรียบเทียบ การนำเสนอในมุมมองเฉพาะ สามารถสรุปผลการรายงาน และรายงานข้อมูลที่มีความสัมพันธ์กันของข้อมูลอาชญากรรมโดยการสร้างการวิเคราะห์การประมวลผลแบบออนไลน์ (Online Analytical) ในรูปแบบของตารางหลายมิติ (Multidimensional table) เพื่ออำนวยความสะดวกรวดเร็ว ใช้ในการระบุความเสี่ยงของการเกิดอาชญากรรม รวมถึงพัฒนาต้นแบบคลังข้อมูลสำหรับเก็บข้อมูล เพื่อนำไปประยุกต์ใช้ในระบบต่อไปและใช้เป็นระบบสนับสนุนการตัดสินใจให้กับสำนักงานตำรวจ) ซึ่งมีการแสดงสถิติของการเกิดอาชญากรรม เพื่อสำนักงานตำรวจสามารถใช้ในการกำหนดนโยบายสำหรับการสร้างความปลอดภัยให้แก่นักท่องเที่ยวและประชาชนในพื้นที่เมืองพัทยา โดยจากการผลการทดลอง พบว่าโดยการใช้แบบจำลองโครงข่ายประสาทเทียม (Natural Network) ให้ประสิทธิภาพการจำแนกได้ดี มีค่าความแม่นยำ (Precision) เท่ากับ 84.16% ค่าความระลึก (Recall) เท่ากับ 83.40% และ F-measure เท่ากับ 83.49%

12.3 การประยุกต์ใช้เหมืองข้อความร่วมกับความสัมพันธ์ของคำโดยกราฟในการให้คำแนะนำ : วัชรีวรรณ จิตต์สกุล (2559) นำเสนอการประยุกต์ใช้เหมืองข้อความร่วมกับความสัมพันธ์ของคำโดยกราฟในการให้คำแนะนำ โดยแบ่งการดำเนินงานออกเป็น 3 ส่วน ดังนี้ ส่วนที่ 1 ทำการคัดเลือกอัลกอริทึมที่มีความเสถียรในการจำแนก โดยคัดเลือกจาก 4 พื้นฐานได้แก่ ฐานกฎ ฐานต้นไม้ตัดสินใจ ฐานความน่าจะเป็น และฐานเรียนรู้ ส่วนที่ 2 สร้างความสัมพันธ์ของคำด้วยวิธี Co-occurrence และส่วนที่ 3 การให้คำแนะนำ โดยพิจารณาคำที่ได้จากอัลกอริทึมที่มีความเสถียรในการจำแนกร่วมกับคำสำคัญที่ได้จากความสัมพันธ์ของคำโดยกราฟ ผลการวิจัยพบว่าอัลกอริทึมที่มีความเสถียรในการจำแนก ได้แก่ อัลกอริทึม Random Forest โดยให้ประสิทธิภาพ

ในการจำแนกดีที่สุด ซึ่งสามารถแบ่งกลุ่มคำเป็น 3 กลุ่ม คือ คำที่ตรงกัน คำที่คล้ายกัน และคำที่ไม่มี และเมื่อนำอัลกอริทึม Random Forest มาพิจารณาร่วมกับความสัมพันธ์ของคำที่สร้างขึ้น พบว่าให้ค่าความถูกต้องเฉลี่ยในการทำนายสูงถึง 81.34% และแสดงการให้คำแนะนำคำได้อย่างเหมาะสม

12.4 แบบจำลองการจำแนกไอซีดี-เทน ทีเอ็ม ข้ามภาษาโดยใช้เหมืองข้อความ : พรรัศมิ์ จตุนราพิทย์ (2558) ปัจจุบันเทคโนโลยีมีส่วนเข้ามาช่วยงานทางการแพทย์มากมาย ซึ่งส่วนหนึ่งของงานเวชระเบียนในการจำแนกโรคเพื่อบันทึก ไอซีดี-เทน ทีเอ็ม ก็เช่นกัน แต่เนื่องด้วยระบบปัจจุบันโดยหลักมนุษย์ยังเป็นผู้ดำเนินการและต้องอาศัยความรู้เฉพาะทาง ความชำนาญของผู้ทำงาน ผู้วิจัยได้นำเสนอแบบจำลองในการจำแนก ไอซีดี-เทน ทีเอ็มอัตโนมัติโดยใช้เหมืองข้อความซึ่งได้นำการค้นคืนมาช่วยในการวิเคราะห์คำ เพิ่มความแม่นยำในการจำแนกไอซีดี-เทน ทีเอ็ม และอัลกอริทึมที่ใช้ในการจำแนก ไอซีดี-เทน ทำการเลือกโดยเปรียบเทียบผลคำวินิจฉัย 3,000 คำ วินิจฉัยจากผู้เชี่ยวชาญกับอัลกอริทึมนาอีฟเบย์ การตัดสินใจต้นไม้ ซัพพอร์ตเวกเตอร์แมชชีน เพื่อหาค่าความถูกต้อง ค่าความแม่นยำ และค่าระลึก ผลลัพธ์ที่ได้อัลกอริทึมนาอีฟเบย์ให้ค่าความถูกต้อง และค่าความแม่นยำสูงสุด ผู้วิจัยจึงเลือกอัลกอริทึมนาอีฟเบย์เพื่อทำการสร้างแบบจำลองการจำแนก ไอซีดี-เทน ทีเอ็ม

12.5 ระบบจำแนกหมวดหมู่การแจ้งซ่อมบ้านออนไลน์โดยใช้เหมืองข้อความ : ประเดิม วงค์กระโซ่ (2560) นี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลหรือแบบจำลองระบบจำแนกหมวดหมู่การแจ้งซ่อมบ้านออนไลน์โดย โดยใช้โปรแกรม Rapid Miner Studio 8.0 ในการพัฒนาแบบจำลองเพื่อเปรียบเทียบประสิทธิภาพค่าความถูกต้องที่ดีที่สุดในการจำแนกหมวดหมู่และเทคนิคเหมืองข้อความเพื่อมาช่วยในการจำแนกข้อความรายการแจ้งซ่อมของระบบบริหารจัดการแจ้งซ่อมหลังการขาย โดยใช้ข้อมูลที่มีอยู่แล้วมาทำการเรียนรู้ด้วยเครื่องและมาใช้ในการจัดหมวดหมู่ข้อความ เพื่ออำนวยความสะดวกให้กับเจ้าหน้าที่บริการ หลังการขาย ช่วยให้การทำงานรวดเร็วขึ้น และลดข้อผิดพลาดในการเลือกหมวดหมู่แจ้งซ่อมบ้านเดิมระบบมีลักษณะการทำงานในรูปแบบของเว็บแอปพลิเคชันอินเตอร์เฟสดำเนินการเกี่ยวกับรายการแจ้งซ่อมของลูกบ้าน มีการจัดหมวดหมู่รายการแจ้งซ่อมเพื่อเปิดใบแจ้งซ่อม โดยเจ้าหน้าที่บริการหลังการขายจะทำหน้าที่เปิดใบแจ้งซ่อม เลือกหมวดหมู่แจ้งซ่อมบ้าน บันทึกรายการแจ้งซ่อม ซึ่งเกิดข้อผิดพลาดในการเปิดใบแจ้งซ่อมแยกตามหมวดหมู่ หรือใส่หมวดหมู่รายการหมวดหลัก หมวดย่อย ผิดประเภททำให้เกิดการยกเลิกหรือแก้ไขรายการแจ่งซ่อมอยู่บ่อยครั้ง ทำให้เกิดล่าช้าในการปฏิบัติงาน ทั้งนี้ได้มีการเลือกใช้

เทคนิคการจำแนกข้อความด้วยกัน 3 เทคนิค ได้แก่ เทคนิคนาอีฟเบย์ เทคนิคต้นไม้การตัดสินใจ และเทคนิคการหาเพื่อนบ้านใกล้ที่สุดเพื่อเปรียบเทียบประสิทธิภาพค่าความ ถูกต้องที่ดีที่สุด โดยวัดประสิทธิภาพค่าความถูกต้องในการจำแนกหมวดหมู่ด้วยวิธีการวัดประสิทธิภาพโดยการแบ่งข้อมูลออกเป็นหลายส่วนเท่าๆ กัน และจากการทดสอบเทคนิคดังกล่าว พบว่าเทคนิคนาอีฟเบย์มีประสิทธิภาพความถูกต้องดีที่สุดร้อยละ 84.22 หลังจากนั้นได้นำแบบจำลอง นาอีฟเบย์มาเป็นโมดูลในการพัฒนาระบบจำแนกหมวดหมู่การแจ้งซ่อมบ้านออนไลน์โดยใช้เหมืองข้อความของระบบบริหารจัดการแจ้งซ่อมหลังการขาย และทำการประเมินความพึงพอใจของผู้ใช้งาน พบว่าผู้ใช้งานมีความพึงพอใจด้านการใช้งานโดยรวมมีค่าเฉลี่ยเท่ากับ 4.39 ค่าเบี่ยงเบนมาตรฐาน เท่ากับ 0.06 ด้านการรับรู้ความง่ายของระบบ มีค่าเฉลี่ยเท่ากับ 4.43 คะแนน ค่าเบี่ยงเบนมาตรฐาน เท่ากับ 0.02 ด้านความครบถ้วนและความถูกต้องของฟังก์ชันการทำงานของระบบ มีค่าเฉลี่ยเท่ากับ 4.44 คะแนน ค่าเบี่ยงเบนมาตรฐานเท่ากับ 0.005 ด้านการรับรู้ประโยชน์ของระบบ มีค่าเฉลี่ยเท่ากับ 4.30 ค่าเบี่ยงเบนมาตรฐานเท่ากับ 0.03 และด้านประสิทธิภาพของระบบ มีค่าเฉลี่ยเท่ากับ 4.49 ค่าเบี่ยงเบนมาตรฐานเท่ากับ 0.03

12.6 การประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ : พรรณนาภรณ์ เกตุภู่พงษ์ (2561) ได้พัฒนาระบบสำหรับการจำแนกประเภทของโรคจากอาการที่เป็นข้อความภาษาไทยและอังกฤษ โดยการสร้างแบบจำลองเชิงทำนายเพื่อประมวลผลหาชื่อโรคและจำแนกโรคที่มีความน่าจะเป็นจากข้อมูลบนเวชระเบียนผู้ป่วย พบข้อจำกัด 2 ประการ ประการแรกการวินิจฉัยโรคของแพทย์ไม่แม่นยำเสมอไปเกิดการคลาดเคลื่อน มีข้อผิดพลาด ประการที่สองมีการบันทึกคำวินิจฉัยของแพทย์ลงในเวชระเบียน มีการบันทึกรหัสไอซีดีเทนซีเอ็มกำกับทุกครั้งใช้ในการจำแนกซึ่งใช้ระยะเวลานาน (รหัสไอซีดีเทนซีเอ็ม คือ รหัสโรคสากลที่ใช้ในทุกโรงพยาบาลสามารถจำแนกได้จากการระบุคำอาทิ ชื่อโรค ชนิดโรค ตำแหน่งของโรค) และพบปัญหาการทำงานล่าช้า เนื่องจากจำนวนบุคลากรไม่เพียงพอต่อการทำงาน เจ้าหน้าที่เวชสถิติ 1 คน ทำหน้าที่จำแนกรหัสฯจากแพทย์มากกว่า 1 คน ผู้วิจัยจึงแนวคิดที่จะนำเสนอแบบจำลองสำหรับจำแนกประเภทโรคจากอาการ โดยการประยุกต์ใช้การทำเหมืองข้อความ เพื่อช่วยแพทย์ในการวินิจฉัยโรคและจำแนกรหัสไอซีดีเทนซีเอ็มได้ด้วยข้อมูลอาการของผู้ป่วย ซึ่งการสร้างแบบจำลองในงานวิจัยนี้จะเลือกใช้ตัวจำแนกประเภทที่นิยมใช้ในการทำเหมืองข้อความ ได้แก่ ต้นไม้ตัดสินใจ การเรียนรู้เบส์อย่างง่าย ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม มาเปรียบเทียบกันโดยใช้ระยะเวลาที่ใช้ในการสร้างแบบจำลอง ระยะเวลาที่แบบจำลองใช้ในการทำนาย กราฟเส้นโค้งอาร์โอซี อัตราผลบวกจริง อัตราผลบวกเท็จ ค่าความเที่ยง และค่าความแม่นมาเป็นตัวชี้วัด ซึ่งผลลัพธ์ที่

ได้พบว่าการใช้โครงข่ายประสาทเทียมเป็นตัวจำแนกประเภทในการสร้างแบบจำลองมีความเหมาะสมที่สุดสำหรับงานวิจัยนี้ เนื่องจากให้อัตราผลบวกจริงสูงสุดที่ร้อยละ 89.03 และมีพื้นที่ใต้เส้นโค้งของกราฟเส้นโค้งอาร์โอซีมากที่สุด นอกจากนี้ผลลัพธ์ที่ได้จะช่วยให้แพทย์วินิจฉัยโรคได้เร็ว และการจำแนกประเภทชนิดโรคจากอาการ การเข้าถึงรหัสฯ สะดวกรวดเร็วมีประสิทธิภาพมากขึ้น

12.7 การสกัดปัจจัยที่ส่งผลต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุในระบบรถไฟด้วยเหมืองข้อความ : ธันยพร กรานสุข (2560) การขนส่งสาธารณะระบบราง รถไฟ เป็นอีกรูปแบบการเดินทางที่ใช้ขนส่งผู้โดยสารที่มีความสำคัญต่อประเทศ ดังนั้นเมื่อเกิดอุบัติเหตุ จึงทำให้เกิดมูลค่าความเสียหายที่สูง หากสามารถวิเคราะห์สืบหาสาเหตุปัจจัยที่ส่งผลต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุได้อย่างรวดเร็ว จะสามารถช่วยลดความรุนแรงของอุบัติเหตุ และเพิ่มประสิทธิภาพในการบริหารจัดการ จนนำไปสู่มาตรการป้องกัน หรือแผนการปรับปรุงแก้ไขได้อย่างมีประสิทธิภาพ ดังนั้น ผู้วิจัยได้ประยุกต์ใช้เทคนิคเหมืองข้อความในการนำมาสกัดเอาคุณลักษณะเฉพาะ และสร้างแบบจำลองในการจำแนกปัจจัยที่ส่งผลต่อการผิดพลาดที่ก่อให้เกิดอุบัติเหตุโดยทำการเปรียบเทียบประสิทธิภาพระหว่าง 2 อัลกอริทึม คือ อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน และอัลกอริทึมนาอีฟเบย์ ผลจากการทดสอบพบว่าอัลกอริทึมที่มีความแม่นยำ และมีประสิทธิภาพในการจำแนกมากที่สุดคือ อัลกอริทึมซัพพอร์ตเวกเตอร์ แมชชีน โดยมีค่าผลการวัดประสิทธิภาพโดยรวม (F-Measure) เท่ากับ 97.87 เปอร์เซ็นต์

12.8 ระบบวิเคราะห์ความคิดเห็นต่อธุรกิจด้วยการทำเหมืองข้อความบนทวิตเตอร์ : ชนิตา ลิสิริกุล (2560) ได้พัฒนาเว็บแอปพลิเคชันสำหรับวิเคราะห์ความคิดเห็นจากข้อความบนทวิตเตอร์ในการรับฟังเสียงของลูกค้าสถาบันการเงินแห่งหนึ่งอย่างอัตโนมัติโดยใช้ทวิตเตอร์เอพีไอรวบรวมข้อมูลความคิดเห็นต่อธุรกิจจากคำที่มีความหมายเกี่ยวข้องกับธนาคารกรุงเทพบนทวิตเตอร์มาวิเคราะห์ด้วยการทำมือ (Manual) ทีละข้อความคิดเห็นและกำหนดคุณสมบัติย่อยแต่ละข้อความและนำมาตัดคำกำจัดคำหยุดและพิจารณาค่าน้ำหนักของคำสำคัญเพื่อนับความถี่ของคำที่ปรากฏในเอกสารและนำมาวิเคราะห์ทัศนคติเชิงบวกและเชิงลบและจากการรวบรวมข้อมูลความคิดเห็นจากทวิตเตอร์ตั้งแต่เดือนมกราคม 2559 ถึงตุลาคม 2560 ได้นำมาพัฒนาเป็นแบบจำลองที่ใช้เทคนิคการทำเหมืองข้อความโดยใช้โปรแกรม Rapidminer สำหรับการจำแนกข้อความออกเป็น 4 ด้าน คือ ด้านการบริการ ด้านผลิตภัณฑ์ ด้านสถานที่ และด้านเทคโนโลยีสารสนเทศ และจำแนกทัศนคติจากข้อความออกเป็นด้านบวกและด้านลบ อีกทั้งได้เปรียบเทียบ

ประสิทธิภาพของแบบจำลองโดยใช้อัลกอริทึมนาอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน เพื่อหาวิธีการจำแนกข้อความและจำแนกทัศนคติที่ดีที่สุด ผลการเปรียบเทียบประสิทธิภาพอัลกอริทึมสำหรับการจำแนกข้อความและจำแนกทัศนคติพบว่า ประสิทธิภาพการจำแนกข้อความทวิตเตอร์โดยใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ดีกว่าอัลกอริทึมนาอีฟเบย์โดยมีค่าความถูกต้องที่ 76.47% และประสิทธิภาพการจำแนกทัศนคติโดยใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ดีกว่าอัลกอริทึมนาอีฟเบย์โดยมีค่าความถูกต้องที่ 84.15% จากนั้นจึงเลือกแบบจำลองโดยใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเพื่อนำไปพัฒนาเป็นเว็บแอปพลิเคชันสำหรับวิเคราะห์ความคิดเห็นซึ่งจะช่วยให้องค์กรได้นำความคิดเห็นนั้นไปพัฒนาและปรับปรุงเพื่อให้ตรงกับความต้องการของลูกค้าได้อย่างรวดเร็ว

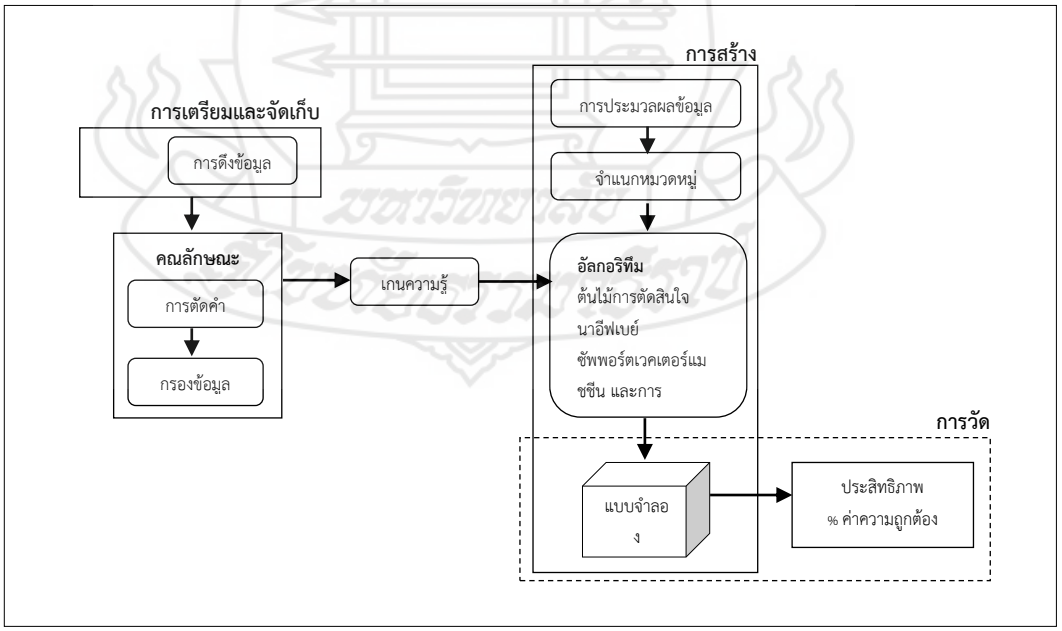
12.9 เว็บท่าเชิงความหมายสำหรับการจัดกลุ่มข้อมูลโดยใช้เทคนิคเหมืองข้อความกรณีศึกษามะเร็ง : สุภาพร วีระพันธ์ยานนท์ (2561) นำเสนอโมเดลการจัดกลุ่มข้อมูลกรณีศึกษามะเร็งนำข้อมูลที่สืบค้นได้จากเสิรชเอนจินมารวบรวมการจัดหมวดหมู่โดยใช้เทคนิคเหมืองข้อความ (Text mining) เดิมพบปัญหาการค้นหาที่ยากเนื่องจากไม่มีการจัดหมวดหมู่ ผู้วิจัยเสนอการจัดกลุ่มเอกสารข้อความภาษาไทยจากข่าวหนังสือพิมพ์ จัดกลุ่มจากเอกสารเว็บและจัดกลุ่มผลลัพธ์การสืบค้นเว็บภาษาไทย และจัดกลุ่มเสิรชเอนจิน แต่ยังขาดการจัดกลุ่มผลลัพธ์จากการสืบค้นข้อมูลภาษาไทยและพัฒนาเป็นเว็บท่าเชิงความหมายสำหรับการจัดกลุ่มมะเร็งและแสดงผลในลักษณะกราฟวิชวลไลซ์เซชัน โดยการพัฒนาอัลกอริทึมเพื่อจัดกลุ่มข้อมูลจากเว็บไซต์ไทยที่มีความน่าเชื่อถือ เช่นเว็บไซต์โรงพยาบาล แพทย์ ผู้เชี่ยวชาญ สถาบันทางการแพทย์ โดยชุดข้อมูลทดสอบจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวข้องกับมะเร็ง เช่น มะเร็ง การรักษามะเร็งอาการของมะเร็งอาหารสำหรับผู้ป่วยมะเร็งอาหารเสริมต้านมะเร็ง สมุนไพรรักษามะเร็ง เป็นตัวทดสอบโมเดลที่สร้างกรณีศึกษามะเร็งการดำเนินการวิจัยมีลำดับขั้นตอนดังนี้ เริ่มต้นด้วยการพัฒนาส่วนการประมวลผลข้อความภาษาไทยเพื่อสร้างออนโทโลยีมะเร็งโดยเสนออัลกอริทึมการตัดคำโดยการนำออนโทโลจีร่วมกับอัลกอริทึมการตัดคำที่มี และสร้างพจนานุกรมเฉพาะด้านมะเร็งเพื่อเป็นตัวเพิ่มประสิทธิภาพในการตัดคำ จากนั้นทำการสกัดข้อความและสืบค้นเชิงความหมายในฐานข้อมูลโดยใช้ออนโทโลยีมะเร็งและพจนานุกรมมะเร็งที่สร้างขึ้น ต่อไปทำการการเปรียบเทียบสมการการทำดัชนีเอกสารได้แก่ TFIDF, WTFIDF และ FTFIDF ร่วมกับการทดลองการจัดกลุ่มข้อความแบบ Hierarchical เปรียบเทียบวิธีการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบ ด้วยวิธี Single link วิธี Average link และวิธี Complete link เพื่อเปรียบเทียบว่าสมการใดให้ประสิทธิภาพการจัดกลุ่มข้อความวิธีใดดีที่สุด ผลการทดลองแสดงให้เห็นว่าการใช้สมการ WTFIDF ร่วมกับ

อัลกอริทึมประสิทธิภาพการจัดกลุ่มด้วยอัลกอริทึมการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบการจัดกลุ่มแบบ Complete link ให้ค่าความถูกต้องในการจำแนกกลุ่มได้ประสิทธิภาพสูงสุด และ SSE ต่ำที่สุดจากการแบ่งข้อมูลมะเร็งได้ 6 กลุ่มในการจัดกลุ่มข้อความดีกว่าเมื่อเปรียบเทียบกับอัลกอริทึมอื่น ๆและท้ายสุดทำการพัฒนาเว็บท่าเชิงความหมาย

ผู้วิจัยจึงมีแนวคิดพัฒนาจากเดิมมีการจัดพิมพ์บนกระดาษไปสู่การจัดพิมพ์ดิจิทัล และพัฒนาระบบการเผยแพร่หนังสือแบบออนไลน์ เพื่อให้ผู้ใช้งานสามารถเข้าถึงได้สะดวกจากทุกที่ทุกเวลา ทุกอุปกรณ์ โดยใช้ข้อมูลเดิมที่มีอยู่ เพื่อให้การค้นหาข้อมูลได้สะดวกรวดเร็ว ตรงตามความต้องการ และได้ประยุกต์ใช้ การจำแนกข้อความ (Text Classification) สำหรับการจำแนกประเภท แยกแยะ หรือจัดข้อความเป็นหมวดหมู่หรือเป็นกลุ่ม เพื่อระบุและเพื่อการวิเคราะห์ว่าข้อความควรจัดจำแนกอยู่ในหมวดหมู่ใด โดยใช้คำสำคัญจากเนื้อหาภายในหนังสือฯ แต่ละเล่มและได้พัฒนาการจำแนกหมวดหมู่หนังสือเผยแพร่ความรู้ โดยใช้เทคนิคเหมืองข้อความจัดกลุ่มของข้อมูลให้เป็นหมวดหมู่ และการตัดคำด้วย Python ใช้อัลกอริทึม ได้แก่ Decision Trees, Support Vector Machine, Naive Bayes และ Logistic regression มาใช้เพื่อสร้างกระบวนการวิเคราะห์ความรู้สึก และสร้างแบบจำลองการจำแนกหมวดหมู่ไปใช้วิเคราะห์กับข้อมูลคำสำคัญที่ได้มีการรวบรวมเอาไว้ทั้งสิ้น 948 คำ และนำมาทดลองใช้กับกลุ่มของคำที่ถูกแบ่งไว้สำหรับทดสอบเพื่อเปรียบเทียบประสิทธิภาพในการจำแนก จัดประเภทหมวดหมู่หนังสือเผยแพร่ความรู้และนำมาใช้ในการกำหนดแนวทางในการพัฒนาเอกสารเผยแพร่ความรู้ของหน่วยต่อไป

บทที่ 3
วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัยของงานวิจัยนี้ ผู้วิจัยมีแนวคิดเพื่อพัฒนาแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ ซึ่งจากเดิมมีการจัดพิมพ์บนกระดาษ พัฒนาในรูปแบบวารสารออนไลน์ เพื่อให้ผู้ใช้งานสามารถเข้าถึงข้อมูล และค้นหาได้สะดวกรวดเร็ว โดยใช้ข้อมูลเดิมที่มีอยู่ และได้ประยุกต์ใช้การจำแนกข้อความ (Text Classification) สำหรับการจำแนกประเภท แยกแยะ หรือจัดข้อความเป็นหมวดหมู่หรือเป็นกลุ่ม เพื่อระบุและเพื่อการวิเคราะห์ว่าข้อความควรจัดจำแนกอยู่ในหมวดหมู่ใด โดยใช้อัลกอริทึม ได้แก่ Decision Trees, Support Vector Machine, Naive Bayes และ Logistic regression มาใช้เพื่อสร้างกระบวนการวิเคราะห์ความรู้สึก และสร้างแบบจำลองการจำแนกหมวดหมู่ไปใช้วิเคราะห์กับข้อมูลคำสำคัญที่ได้มีการรวบรวมเอาไว้ทั้งสิ้น 948 คำ และนำมาทดลองใช้กับกลุ่มของคำที่ถูกแบ่งไว้สำหรับทดสอบเพื่อเปรียบเทียบประสิทธิภาพในการจำแนก จัดประเภทหมวดหมู่หนังสือเผยแพร่ความรู้ และนำมาใช้ในการกำหนดแนวทางในการพัฒนาเอกสารเผยแพร่ความรู้ของหน่วยต่อไป



ภาพที่ 3.1 การวิเคราะห์ขั้นตอนการทำงาน

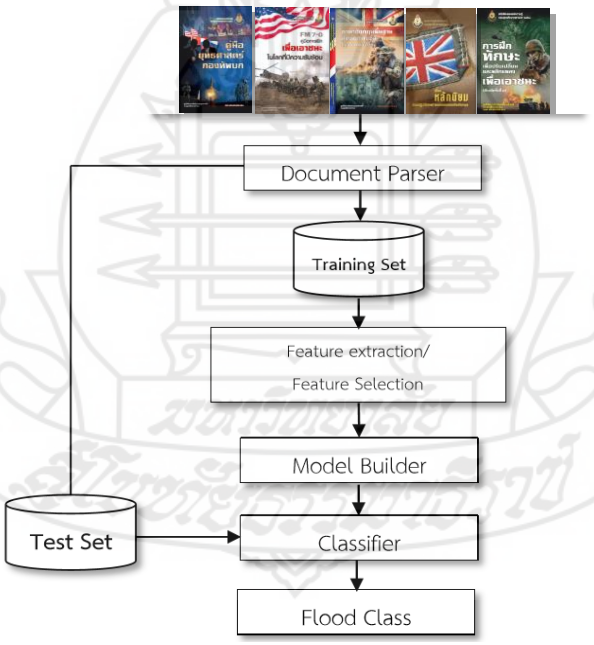
1. การศึกษาแหล่งข้อมูล

งานวิจัยนี้ผู้วิจัยมุ่งเน้นไปที่ข้อมูลในหนังสือเผยแพร่ความรู้ โดยเป็นข้อความ คำสำคัญ ซึ่งจากเดิมมีการจัดพิมพ์ในรูปเล่มหนังสือ มาพัฒนาให้เป็นในรูปแบบวารสารออนไลน์ เพื่อให้ผู้ใช้งานสามารถเข้าถึงข้อมูล และค้นหาได้สะดวกรวดเร็ว โดยใช้ข้อมูลเดิมที่มีอยู่แล้วจากหนังสือ

2. ประชากรและกลุ่มตัวอย่าง

ข้อมูลคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้ จำนวน 948 คำ ที่ได้มาซึ่งมีการรวบรวมไว้จากหนังสือฯ ตั้งแต่ปี พ.ศ. 2553 ถึง 2563

3. กรอบแนวคิดในการพัฒนาแบบจำลอง

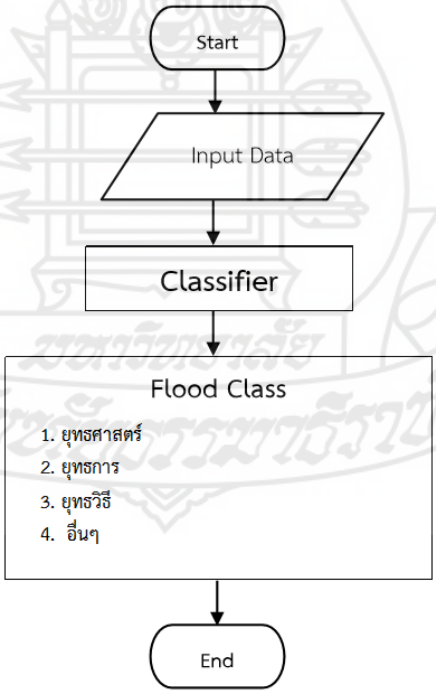


ภาพที่ 3.2 กรอบแนวคิดในการพัฒนาแบบจำลอง

จากภาพที่ 3-1 เป็นกรอบแนวคิดในการพัฒนาแบบจำลอง โดยเมื่อได้ข้อมูลที่ได้จากการเก็บรวบรวมข้อมูลเรียบร้อยแล้ว ขั้นต่อไปเป็นการนำข้อมูลที่ได้ มาทำการ Document Parser

เพื่อทำให้ข้อมูลอยู่ในรูปแบบเดียวกันและเหมาะสมสำหรับการสร้างแบบจำลองการจำแนกหมวดหมู่ โดยแบ่งข้อมูลเป็น 2 ส่วน ได้แก่ ส่วนที่ใช้สำหรับการเรียนรู้ (Training Set) และส่วนที่ใช้เพื่อการทดลองการจำแนกหมวดหมู่ จากนั้นทำการสกัดคุณลักษณะ (Feature Extraction) ของข้อความด้วยการตัดคำ และสร้างเมตริกซ์ของกลุ่มข้อความขึ้นจากเวกเตอร์ข้อความทั้งหมดจนได้เมตริกซ์ของคำ (Term Document Matrix) เนื่องจากข้อความที่ได้มีปริมาณสูงจึงทำการเลือกคุณลักษณะ (Feature Selection) ของข้อความด้วยการลดจำนวนแอตทริบิวต์ของข้อความ โดยลดจำนวนแอตทริบิวต์ของข้อความเพื่อเปรียบเทียบวิธีการลดจำนวนแอตทริบิวต์ที่เหมาะสมทำให้แบบจำลองการจำแนกหมวดหมู่ข้อความให้ค่าความถูกต้องในการจำแนกถูกต้องมากที่สุด จากนั้นเข้าสู่กระบวนการสร้างแบบจำลองการจำแนกหมวดหมู่ด้วยขั้นตอนวิธีการใช้อัลกอริทึม ได้แก่ Decision Trees, Support Vector Machine, Naive Bayes และ Logistic regression จากนั้นนำแบบจำลองที่ได้มาทำการจำแนกหมวดหมู่ข้อความ/คลาส ให้กับหนังสือเผยแพร่ความรู้ประกอบด้วย 4 คลาส คือ ยุทธศาสตร์ ยุทธการ ยุทธวิธี และอื่นๆ

4. การวิเคราะห์และการเตรียมข้อมูล



ภาพที่ 3.3 ขั้นตอนการทำงานการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

ผู้วิจัยรวบรวมข้อมูลจากหนังสือเผยแพร่ความรู้ โดยใช้คำสำคัญในเล่มฯ มาจำแนกหมวดหมู่เป็น 4 หมู่หรือคลาส คือ ยุทธศาสตร์ ยุทธการ ยุทธวิธี และอื่นๆ เพื่อให้ตรงกับความต้องการ เพื่อช่วย ในการจัดการงานให้กับหน่วยงานต่อไป ในการแบ่งนั้นผู้วิจัยได้อ้างอิงคำศัพท์ทางทหารมากำหนด แบ่งเป็นแบ่งคลาส และอ้างอิงความหมายจากเว็บไซต์ พจนานุกรม ฉบับราชบัณฑิตยสถานออนไลน์ พ.ศ. 2554 ได้ความหมายไว้ดังนี้

ตารางที่ 3.1 แสดงรายละเอียดคำค้นจาก พจนานุกรม ฉบับราชบัณฑิตยสถานออนไลน์

คำค้น	ความหมาย
1. ยุทธศาสตร์	(1) น. วิชาว่าด้วยการพัฒนาและการใช้อำนาจทางการเมือง เศรษฐกิจ จิตวิทยา และกำลังรบทางทหารตามความจำเป็นทั้งในยามสงบและยามสงคราม. (2) ว. ที่มีความสำคัญทางการเมือง เศรษฐกิจ จิตวิทยา และกำลังรบทางทหาร ทั้งในยามสงบและยามสงคราม เช่น จุดยุทธศาสตร์
2. ยุทธการ	น. การรบ, การทำสงคราม
3. ยุทธวิธี	น. วิธีและอุบายของการรบ.
4. อื่นๆ	ว. นอกออกไป, ต่างออกไป

ภาพที่ 3.4 แสดงการวิเคราะห์และการจัดเตรียมข้อมูล

ตารางที่ 3.2 แสดงรายละเอียดที่ใช้จำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้

Class	รายละเอียด
ยุทธศาสตร์	การวางแผนคล้ายๆ การกำหนดนโยบายเป็นการวางแผนในมุมกว้างๆ ใช้สำหรับการบริหารคนหมู่มาก
ยุทธการ	แผนปฏิบัติการ เป็นการแปรแผนยุทธศาสตร์ลงไปเป็นแผนยุทธวิธี ซึ่งโดยหลักแล้วเกี่ยวข้องกับการจัดกำลังพลและกำหนดภาระหน้าที่ที่ต้องกระทำรวมทั้งกำหนดเป้าหมายที่ต้องบรรลุ การรบ การทำสงคราม
ยุทธวิธี	กระบวนการปฏิบัติงาน การเตรียมแนวทางปฏิบัติเพื่อใช้สำหรับการตัดสินใจดำเนินการหรือตอบสนองต่ออุปสรรคเฉพาะหน้าในรูปแบบต่างๆ โดยพื้นฐานความเข้าใจทั่วๆไป เป็นไปเพื่อให้บรรลุผลดีที่สุดและส่งประโยชน์เฉพาะ วิธี และอุบายของการรบ
อื่นๆ	ข้อความอื่นๆ ความคิดเห็นทั่วไปของผู้เขียนหนังสือ

การรวบรวมข้อมูลจากหนังสือเผยแพร่ความรู้ จัดเก็บในรูปแบบไฟล์นามสกุล .csv เพื่อใช้ในการนำเข้าสู่กระบวนการจำแนกหมวดหมู่โดยทำการแปลงข้อมูลให้อยู่ในรูปแบบรายการข้อมูล ซึ่งจะประกอบไปด้วย ID, รายการ และ Class ซึ่งรายการข้อมูลประกอบด้วย 4 คลาส คือ A (ยุทธศาสตร์) B (ยุทธการ) C (ยุทธวิธี) D (อื่นๆ) ในแต่ละรายการสร้างมาจากคำสำคัญของเนื้อหาในหนังสือเผยแพร่ความรู้

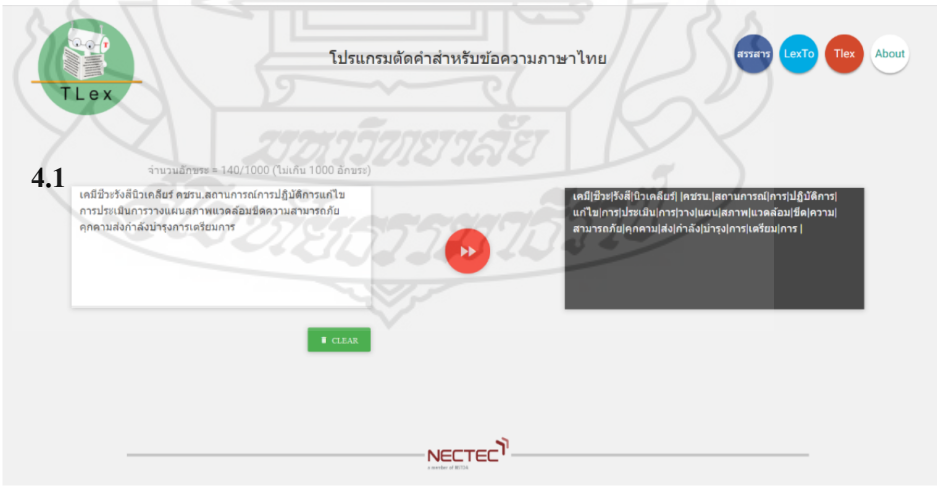
ตารางที่ 3.3 แสดงตัวอย่างข้อมูลคำสำคัญหนังสือเผยแพร่ความรู้ที่นำไปจำแนกคลาส

ID	รายการ	Class
30	การประเมิน, การวางแผน, สภาพแวดล้อม, ขีดความสามารถ	A
31	ส่งกำลังบำรุง, การเตรียมการ, การประสานงาน	B
36	การตั้งรับ, การรบ, การจัดหน่วย, การวางกำลังในที่มั่น, การเปลี่ยนผ่าน	C
37	เสถียรภาพ, ความเชื่อมั่น, ความมั่นคง, การทำลายล้าง	A
43	บทบาท, ยุทธศิลป์, ยุทธการ	B
122	ฐานบิน แอลซี อิงลิช, บันทึก, ประสบการณ์, นายทหาร, เวียดนาม	D

เมื่อได้ข้อมูลข้อความคำสำคัญที่รวบรวมไว้ จากนั้นทำการกำจัดคำข้อความที่ซ้ำกันออก และ ในส่วนข้อความที่รวบรวมไว้นั้นยังไม่เครื่องหมายหรือสิ่งอื่นใดที่จะใช้สำหรับในการแบ่งคำ หรือเว้น วรรค ผู้วิจัยจึงทำการตัดคำ โดยผู้วิจัยได้เลือกใช้เป็นโปรแกรมตัดคำที่มีให้ประชาชนทั่วไปได้ดาวน์ โหลดและใช้งานฟรีที่มีชื่อว่า TLexs เป็นโปรแกรมตัดคำแบบอิงการเรียนรู้ของเครื่อง ถูกพัฒนาขึ้นมา โดยหน่วยปฏิบัติการศูนย์เทคโนโลยีอิเล็กทรอนิกส์ และคอมพิวเตอร์แห่งชาติ (NECTEC)

TLexPlus เป็นระบบแบ่งคำโดยใช้เทคนิคการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการ Conditional Random Fields(CRFs) ร่วมกับคลังข้อมูลของ BEST2009 ขนาด 9 ล้านคำ ในการเรียนรู้ จุดเด่นของ TLexPlus คือสามารถแบ่งคำที่เกิดขึ้นใหม่ คำ ในภาษาต่างประเทศ หรือคำแสลงใหม่ได้อย่างถูกต้อง โดยไม่ต้องอาศัยพจนานุกรม เหมาะสำหรับ นำไปใช้แบ่งคำเพื่อหานิพจน์ระบุนาม (Named Entities) หรือชื่อเฉพาะต่างๆ เป็นต้น

คุณสมบัติของ TLexPlus สามารถทำงานได้กับทุกระบบปฏิบัติการ Windows, Unix based, OSX) รองรับการทำงานในรูปแบบเซอร์วิส (REST Fut Service) ประมวลผลได้อย่างรวดเร็ว รองรับการจัดการคำที่เกิดขึ้นใหม่ คำในภาษาต่างประเทศ หรือคำแสลงใหม่ รองรับการจัดการคำที่ไม่ อยู่ในพจนานุกรมอย่างชาญฉลาด เรียนรู้จากคลังข้อมูลของ BEST2009 ขนาด 9 ล้านคำ มีโมเดลการ เรียนรู้ให้เลือกหลายขนาดตามความเหมาะสมของการใช้งาน และมีระบบจัดการคำศัพท์เพื่อให้ TLexPlus สามารถแบ่งคำได้ตามที่ต้องการได้



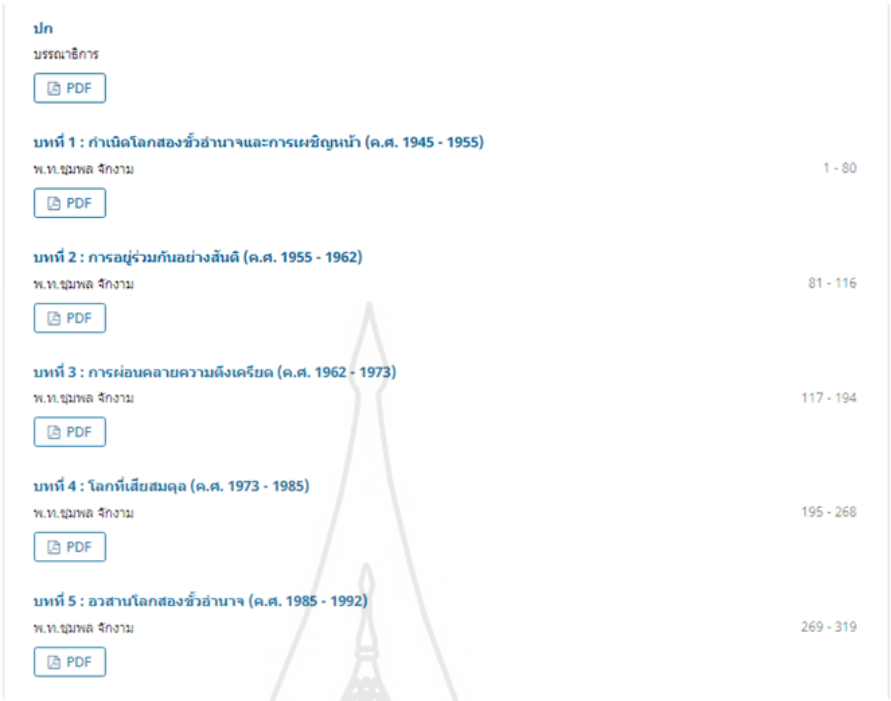
ภาพที่ 3.5 แสดงการตัดคำด้วยโปรแกรม TLexS

5. การพัฒนาระบบ

จากการจัดเตรียมและนำเข้าข้อมูล ผู้วิจัยได้ทำการพัฒนา Open Journal System ซึ่งเป็นการบริหารจัดการการตีพิมพ์และการเผยแพร่หนังสือเผยแพร่ความรู้แบบออนไลน์ ซึ่งรูปแบบคล้ายการเผยแพร่วารสารวิชาการแบบออนไลน์ และพัฒนาประสิทธิภาพในการดำเนินการให้ดียิ่งขึ้นให้สามารถตอบสนองความต้องการของผู้ใช้งาน

ภาพที่ 3.6 แสดงหน้าต่างการใช้งานหน้าแรก

ภาพที่ 3.7 แสดงบทคัดย่อ ปีที่พิมพ์ และชื่อผู้เรียบเรียงหนังสือ



ภาพที่ 3.8 แสดงรายละเอียดเนื้อหาภายใน โดยแบ่งออกเป็นบท

6. การสร้างแบบจำลอง

ผู้วิจัยสร้างแบบจำลองสำหรับการจำแนกหมวดหมู่ข้อความโดยใช้เทคนิคมาจำแนกหมวดหมู่ข้อความ เลือกอัลกอริทึม 4 เทคนิค คือ Decision Tree, Naïve Bayes, Support Vector Machine และ Logistic Regression

นำเข้าข้อมูลคำสำคัญของหนังสือเผยแพร่ความรู้ จากไฟล์นามสกุล .csv จากนั้นอ่านข้อมูลจากไฟล์ดังกล่าวที่ได้มีการนำเข้ามา

	ID	keyword	class
0	27	เคมี, ชีวะ, รังสี, นิวเคลียร์, คชรน., สถานการณ...	C
1	30	การประเมิน, การวางแผน, สภาพแวดล้อม, ขีดความสาม...	A
2	31	ส่งกำลังบำรุง, การเตรียมการ, การประสานงาน	B
3	32	การตอบสนอง, การเตรียมพร้อม, ขั้นการปฏิบัติ, สิ...	B
4	33	การฟื้นฟู, ทำลายล้างพิษ, การเปลี่ยนผ่าน, ถอนกำลัง	C
...	...	...	...
124	222	การประเมินผล, การปฏิบัติ	C
125	224	โครงสร้าง, โครงสร้างอาเซียน, กฎบัตรอาเซียน	A
126	225	เสาหลักอาเซียน, แนวความคิด	B
127	227	บทบาท, การสนับสนุน	B

ภาพที่ 3.9 แสดงรายละเอียดข้อมูลที่นำเข้า

ถัดมาทำการตัดคำ ลบคำ stopword และ punctuation (เครื่องหมายวรรคตอน) โดยเริ่มจากการดึง array ของ stopwords หรือคำที่ไม่ค่อยสื่อความหมาย มาเก็บไว้ที่ตัวแปร thai_stopwords

ภาพที่ 3.10 แสดงตัวอย่างข้อความ คำที่ไม่สื่อความหมาย

การตัดคำ (Word Tokenize) ลบ stopword และ punctuation (เครื่องหมายวรรคตอน) ออกจากข้อความ และเปลี่ยนข้อความให้มีช่องว่างระหว่างคำ) เพื่อนำไปประมวลผลกับ Word Cloud และสร้าง Bag of Word (Bow)

	ID	keyword	class	text_tokens
0	27	เคมี, ชีวะ, รังสี, นิวเคลียร์, คชรน., สถานการณ...	C	เคมี ชีวะ รังสี นิวเคลียร์ คช รน สถานการณ์ การ...
1	30	การประเมิน, การวางแผน, สภาพแวดล้อม, ขีดความสาม...	A	การประเมิน การวางแผน สภาพแวดล้อม ขีดความสามารถ...
2	31	ส่งกำลังบำรุง, การเตรียมการ, การประสานงาน	B	ส่ง กำลัง บำรุง การ เตรียมการ การประสานงาน
3	32	การตอบสนอง, การเตรียมพร้อม, ขั้นการปฏิบัติ, ลิ...	B	การ ตอบสนอง การเตรียมพร้อม ขั้น การปฏิบัติ ลิ่...
4	33	การฟื้นฟู, ทำลายล้างพิษ, การเปลี่ยนผ่าน, ถอนกำลัง	C	การฟื้นฟู ทำลายล้าง พิษ การ เปลี่ยน ผ่าน ถอนกำลัง
...	...	...	...	...
124	222	การประเมินผล, การปฏิบัติ	C	การประเมินผล การ ปฏิบัติ
125	224	โครงสร้าง, โครงสร้างอาเซียน, กฎบัตรอาเซียน	A	โครงสร้าง โครงสร้าง อาเซียน กฎบัตร อาเซียน
126	225	เสาหลักอาเซียน, แนวความคิด	B	เสาหลัก อาเซียน แนวความคิด
127	227	บทบาท, การสนับสนุน	B	บทบาท การ สนับสนุน
128	229	สภาพแวดล้อมทางยุทธศาสตร์, ยุทธศาสตร์, กำหนดยุท...	A	สภาพแวดล้อม ทาง ยุทธศาสตร์ ยุทธศาสตร์ กำหนด ยุ...

129 rows × 4 columns

ภาพที่ 3.11 แสดงตัวอย่างการตัดคำ (Word Tokenize)

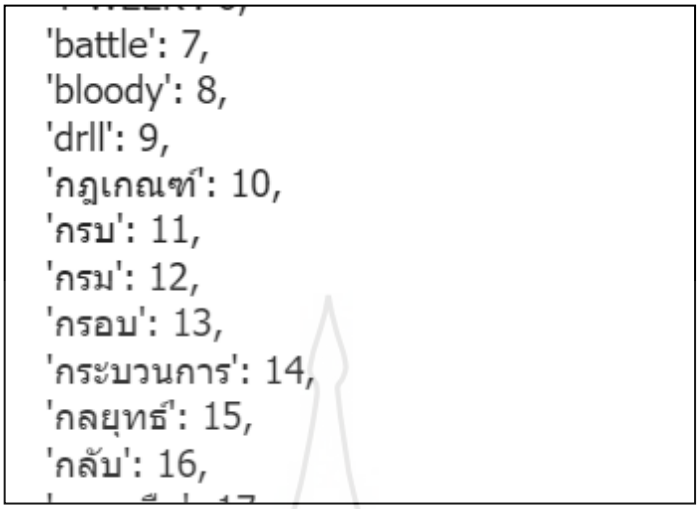
การแสดงผลความถี่ของคำด้วย Word Cloud ทำให้ทราบเบื้องต้นว่าคำที่พบส่วนใหญ่ที่ปรากฏในข้อความมีคำว่าอะไรบ้าง โดยถ้าการแสดงผลของคำใน Word Cloud มีขนาดใหญ่เท่าไหร่จะหมายถึงจำนวนของความถี่ของคำคำนั้นที่ปรากฏอยู่ในข้อความนั้นๆ

ภาพที่ 3.12 แสดง Word Cloud ของ Class (A)

ภาพที่ 3.13 แสดง Word Cloud ของ Class (B)

ภาพที่ 3.14 แสดง Word Cloud ของ Class (C)

ภาพที่ 3.15 แสดง Word Cloud ของ Class (D)



ภาพที่ 3.16 แสดง Word Cloud ของ Class (D)

จากภาพ 3.14 เป็นการดึงคำทั้งหมดออกมาจากข้อความ และจัดเก็บในรูปแบบ Vector (มีลักษณะคล้ายๆ ลักษณะของพจนานุกรมที่มีการระบุตัวเลข index ของแต่ละคำ)

แบ่งข้อมูลเป็น 2 ส่วน คือส่วนของการฝึกฝน (train) แบบจำลอง 70% และส่วนของการทดสอบ (test) แบบจำลอง 30% โดย X คือตัวแปรต้นที่เป็นข้อความ และ y คือตัวแปรตามที่เป็น sentiment (A, B, C, D)

สร้าง Bag-of-Words (BoW) เพื่อใช้ในการฝึกฝนแบบจำลอง ซึ่งมีลักษณะคล้ายกับตารางที่มีแถวเป็นข้อความ คอลัมน์เป็นคำทั้งหมด และค่าคือจำนวนคำที่ปรากฏในข้อความ

text_tokens	-	0	7	FM	JOPP	METL	T-WEEK	I
ภารกิจ	0	0	0	0	0	0	0	
การ ตอบสนอง การเตรียมพร้อม ขั้น การปฏิบัติ สิ่ง บอก เหตุ การ แจ้ง เตือน มาตร การ ควบคุม	0	0	0	0	0	0	0	
การ บูรณาการ การ ยุทธ์	0	0	0	0	0	0	0	
การ เดินทาง ลาดตระเวน กลางคืน	0	0	0	0	0	0	0	
ลาดตระเวน กลับ	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	
ฉาก นำ ปะทะ	0	0	0	0	0	0	0	
ยุทธศาสตร์ ของ จีน ปัจจัยภายใน ยุทธศาสตร์ การปกครอง กลไก การตัดสินใจ ความ มั่นคง โครงสร้าง	0	0	0	0	0	0	0	
ค่านิยม ตัว แบบ ภาวะผู้นำ	0	0	0	0	0	0	0	
เพิ่ม ศักยภาพ ผู้บัญชาการ กรม การ ดำเนิน กลยุทธ์	0	0	0	0	0	0	0	
ความกล้าหาญ กำลัง รบ	0	0	0	0	0	0	0	

90 rows × 224 columns

Warning: Total number of columns (224) exceeds max_columns (20) limiting to first (20) columns.

ภาพที่ 3.17 แสดง Bag-of-Words (BoW)

สร้างแบบจำลองสำหรับการจำแนกประเภท หรือ หมวดหมู่ โดยให้ตัวแปรต้นเป็น BoW ที่สร้างจากข้อความสำหรับการฝึกฝน และตัวแปรตามคือ y_train หรือก็คือคอลัมน์ sentiment ที่แบ่งไว้สำหรับการฝึกฝน

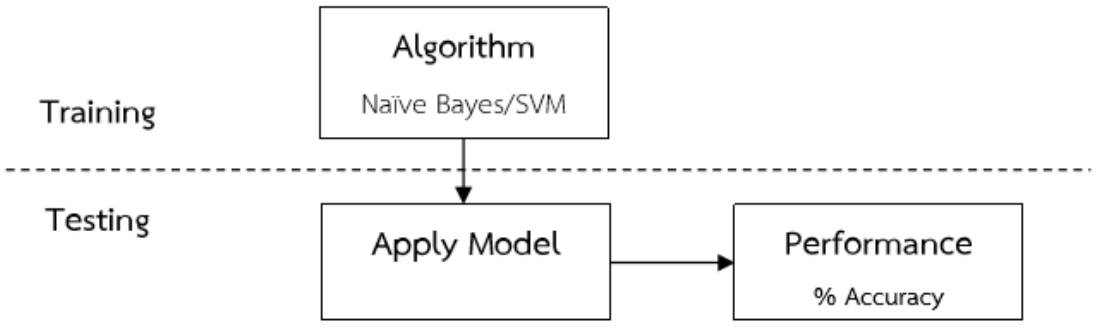
ทดสอบแบบจำลอง โดยใช้ sklearn ในการทดสอบแบบจำลองว่ามีความแม่นยำมากน้อยเพียงใด โดยดูจาก precision และ recall

	precision	recall	f1-score	support
A	0.80	1.00	0.89	4
B	0.78	0.78	0.78	9
C	1.00	0.75	0.86	4
D	0.91	0.91	0.91	22
accuracy			0.87	39
macro avg	0.87	0.86	0.86	39
weighted avg	0.88	0.87	0.87	39

ภาพที่ 3.18 แสดงตัวอย่างการทดสอบแบบจำลอง

7. การประเมินประสิทธิภาพแบบจำลอง

การประเมินประสิทธิภาพของแบบจำลอง เป็นการประเมินประสิทธิภาพของผลลัพธ์จากแบบจำลองที่ได้พัฒนาขึ้น รวมถึงการวิเคราะห์ข้อมูลว่าครอบคลุม และสามารถตอบโจทย์หรือวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรกได้ หรือไม่



การทดสอบใช้วิธี Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับฝึกฝนหรือสร้างแบบจำลอง (Training Set) และ ข้อมูลสำหรับการทดสอบแบบจำลอง (Test Set) แบบจำลอง แบ่งข้อมูล 2 ส่วน โดยกำหนดให้ Train 70% และให้ Test 30% การทดสอบแบบ Split Test นี้ทำการสุ่มข้อมูลเพียงครั้งเดียวซึ่งในบางครั้งถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้สร้างโมเดลทำให้ผลการวัด ประสิทธิภาพได้ออกมาดี ในทางตรงข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สร้างโมเดลมากทำให้ผลการวัดประสิทธิภาพได้ออกมาแย่ ดังนั้นจึงควรใช้วิธี Split Test นี้หรือทำการสุ่ม หลายๆ ครั้ง แต่ข้อดีของวิธีการนี้คือใช้เวลาในการสร้างโมเดลน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่มาก

การประเมินประสิทธิภาพ เพื่อให้ได้ข้อสรุปมีความแม่นยำเพียงไร เหมาะสมที่จะนำไปใช้หรือไม่ ซึ่งจะสามารถดูจากผลการทำนายแบบจำลองหาได้จาก Confusion Matrix เป็นการประเมินผลการทำนายกับผลลัพธ์จริงที่หาได้ และค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure)

ตารางที่ 3.4 แสดง Confusion Matrix

	ทำนาย (Prediction)	
ของจริง (Actual)	Positive (1)	Negative (0)
Positive (1)	True Positive (TP)	False Positive (FP)
Negative (0)	False Negative (FN)	True Negative (TN)

โดยที่

(TP) สิ่งที่ทำนาย ตรงกับสิ่งที่เกิดขึ้นจริง ในกรณี ทำนายว่าจริง และสิ่งที่เกิดขึ้น ก็คือ จริง

(TN) สิ่งที่ทำนายตรงกับสิ่งที่เกิดขึ้น ในกรณี ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น ก็คือ ไม่จริง

(FP) สิ่งที่ทำนายไม่ตรงกับสิ่งที่เกิดขึ้น คือทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง

(FN) สิ่งที่ทำนายไม่ตรงกับที่เกิดขึ้นจริง คือทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น คือ จริง

สามารถใช้ Confusion Matrix คำนวณการประเมินประสิทธิภาพของการทำนายด้วยแบบจำลองโดย Accuracy คือค่าความถูกต้องที่เราทายได้ตรงกับสิ่งที่เกิดขึ้นจริง

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3\text{-}1)$$

Precision คือค่าความแม่นยำเป็นการเปรียบเทียบ การทำนายที่ถูกต้องว่า จริง และก็เกิดขึ้นจริง (TP) กับ การทำนายว่า จริง แต่สิ่งที่เกิดขึ้น คือ ไม่จริง (FP)

$$\text{Precision} = \frac{TP}{TP+FP}\frac{TP+TN}{TP+TN+FP+FN} \qquad (3\text{-}2)$$

Recall คือ การคำนวณหาค่าความระลึกของการทำนายว่าเป็น “จริง” เทียบกับ จำนวนครั้งของเหตุการณ์ทั้งทำนาย และ เกิดขึ้น ว่า “เป็นจริง”

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (3\text{-}3)$$

F1 score คือค่าเฉลี่ยระหว่าง precision และ recall จุดประสงค์ของการสร้าง F1 ขึ้นมา คือ เพื่อเป็น single metric ที่วัดความสามารถของโมเดล

F1 = 2 x (Precision x Recall) / (Precision + Recall) (3-4)

F-Measure ค่าความถ่วงดุล คือ การวัดประสิทธิภาพโดยรวมของทั้งสองค่าระหว่างค่าความแม่นยำและค่าความระลึก พร้อมกันของโมเดล โดยพิจารณาแยกทีละคลาส

$$F-\text{Measure} = \frac{2*\text{Precision}+\text{Recall}}{\text{Precision}+\text{Recall}} \qquad (3\text{-}5)$$

บทที่ 4
ผลการวิเคราะห์ข้อมูล

การพัฒนาแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ โดยดำเนินการรวบรวมข้อมูลคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้ จำนวน 948 คำ ตั้งแต่ปี พ.ศ. 2553 - 2563 โดยใช้ข้อมูลเดิมที่มีอยู่ และได้ประยุกต์ใช้การจำแนกข้อความ (Text Classification) สำหรับการจำแนกประเภท แยกแยะ หรือจัดข้อความเป็นหมวดหมู่หรือเป็น กลุ่ม เพื่อระบุและเพื่อการวิเคราะห์ว่าข้อความควรจัดจำแนกอยู่ในหมวดหมู่ใด โดยใช้อัลกอริทึม ใน การประเมินประสิทธิภาพแบบจำลอง ได้แก่ Decision Trees, Support Vector Machine, Naive Bayes และ Logistic regression มีรายละเอียดผลการวิเคราะห์การดำเนินงานแบ่งออกเป็น 5 ตอน ดังต่อไปนี้

ตอนที่ 1 การประเมินประสิทธิภาพ ด้วย Decision Tree

ตอนที่ 2 การประเมินประสิทธิภาพ ด้วย Naïve Bayes

ตอนที่ 3 การประเมินประสิทธิภาพ ด้วย Support Vector Machine

ตอนที่ 4 การประเมินประสิทธิภาพ ด้วย Logistic Regression

ตอนที่ 5 ค่าความแม่นยำการจำแนกหมวดหมู่ด้วยเทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ต เวกเตอร์แมชชีน และการถดถอยโลจีสติก

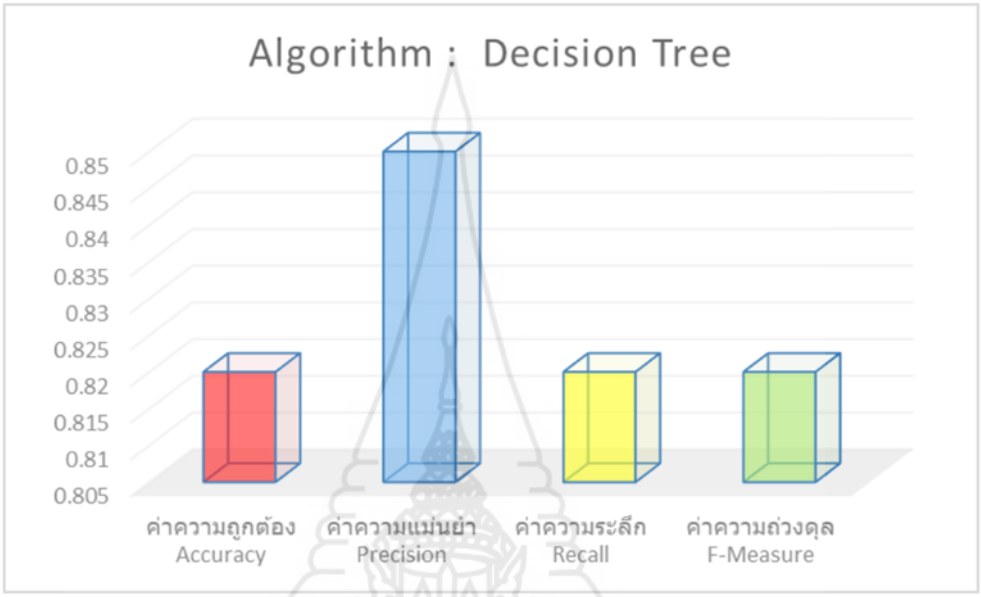
ตอนที่ 6 ตารางการจำแนก Class A, Class B, Class C และ Class D

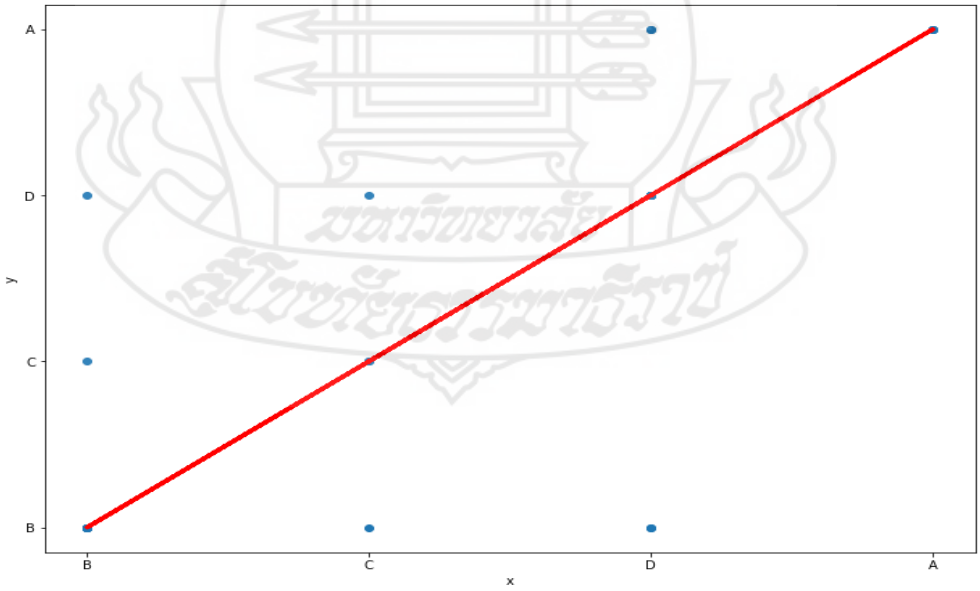
ตอนที่ 1 การประเมินประสิทธิภาพ ด้วย Decision Tree

ตารางที่ 4.1 การประเมินประสิทธิภาพ ด้วย Decision Tree

Algorithm	Accuracy ค่าความถูกต้อง	Precision ค่าความแม่นยำ	Recall ค่าความระลึก	F-Measure ค่าความถ่วงดุล
Decision Tree	0.82	0.85	0.82	0.82

จากตารางที่ 4.1 แสดงให้เห็นถึงการประเมินประสิทธิภาพของแบบจำลอง ด้วยเทคนิค Decision Tree โดยมี ค่าความถูกต้อง (Accuracy) เท่ากับ 0.82 ค่าความแม่นยำ (Precision) เท่ากับ 0.85 ค่าความระลึก (Recall) เท่ากับ 0.82 และค่าความถ่วงดุล (F-Measure) เท่ากับ 0.82





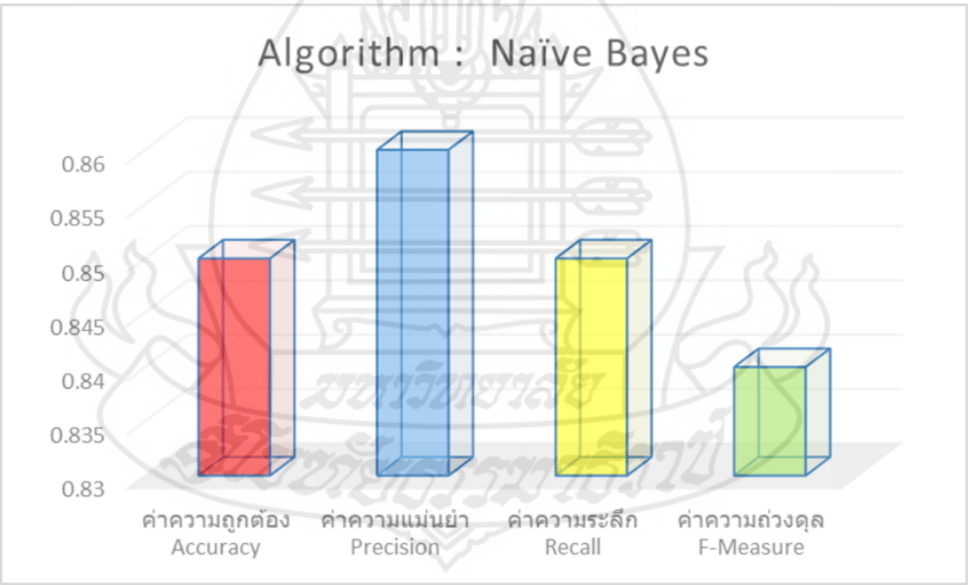
ภาพที่ 4.1 แผนภูมิกราฟเส้นการประเมินประสิทธิภาพ ด้วย Decision Tree

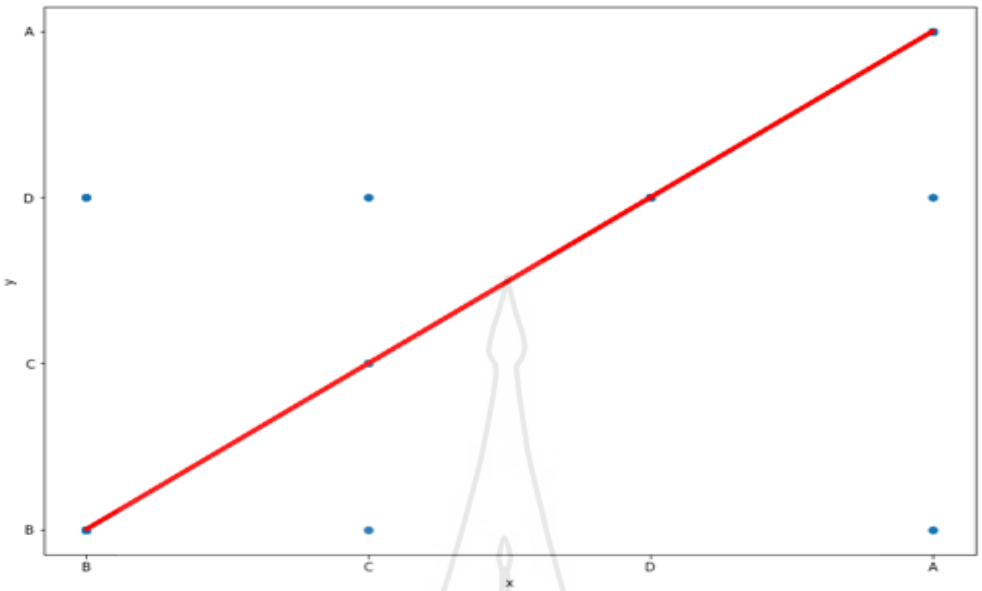
ตอนที่ 2 การประเมินประสิทธิภาพ ด้วย Naïve Bayes

ตารางที่ 4.2 การประเมินประสิทธิภาพ ด้วย Naïve Bayes

Algorithm	Accuracy ค่าความถูกต้อง	Precision ค่าความแม่นยำ	Recall ค่าความระลึก	F-Measure ค่าความถ่วงดุล
Naïve Bayes	0.85	0.86	0.85	0.84

จากตารางที่ 4.2 แสดงให้เห็นถึงการประเมินประสิทธิภาพของแบบจำลอง ด้วยเทคนิค Naïve Bayes โดยมี ค่าความถูกต้อง (Accuracy) เท่ากับ 0.85 ค่าความแม่นยำ (Precision) เท่ากับ 0.86 ค่าความระลึก (Recall) เท่ากับ 0.85 และค่าความถ่วงดุล (F-Measure) เท่ากับ 0.84





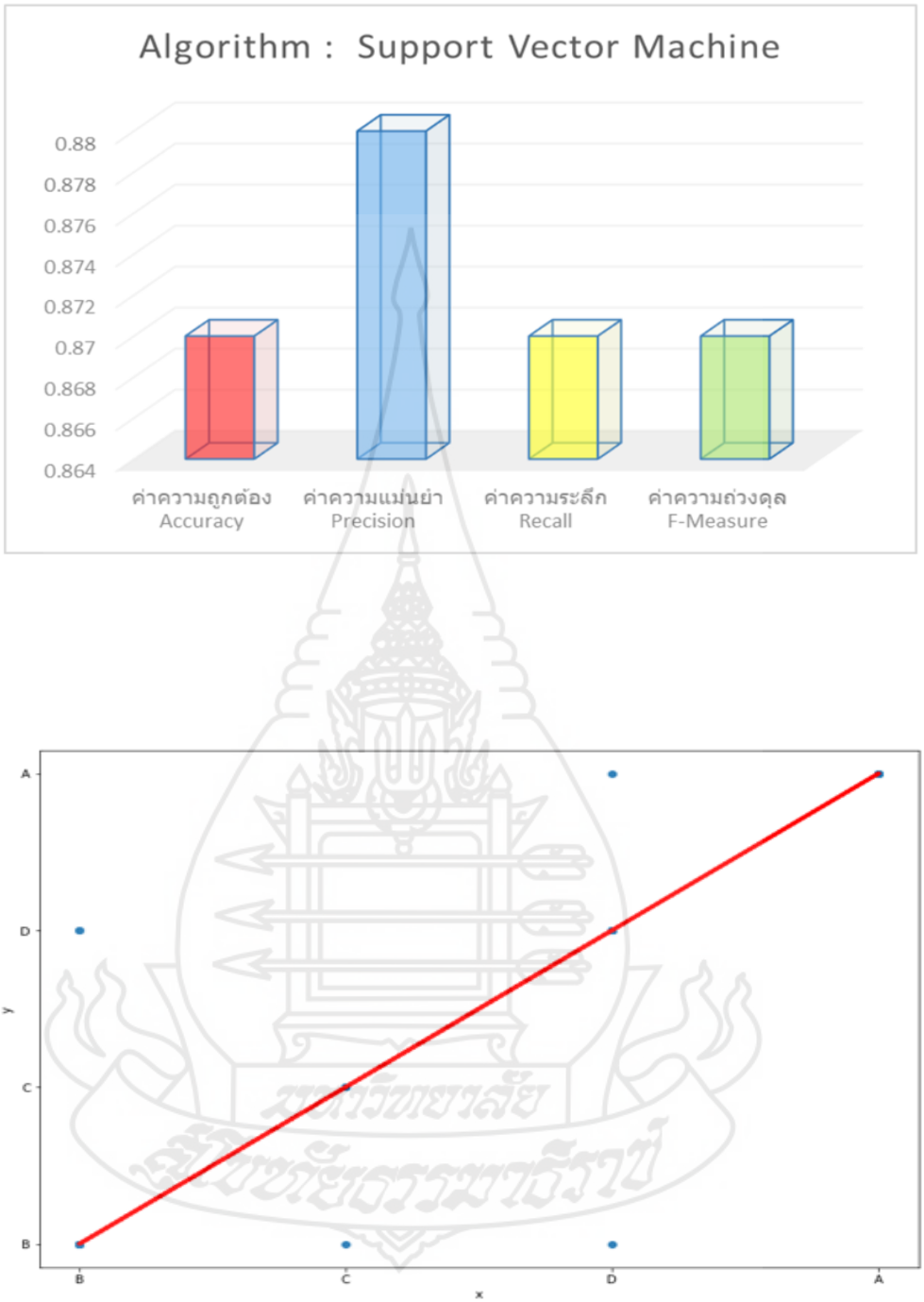
ภาพที่ 4.2 แผนภูมิกราฟเส้นการประเมินประสิทธิภาพ ด้วย Naïve Bayes

ตอนที่ 3 การประเมินประสิทธิภาพ ด้วย Support Vector Machine

ตารางที่ 4.3 การประเมินประสิทธิภาพ ด้วย Support Vector Machine

Algorithm	**Accuracy** ค่าความถูกต้อง	**Precision** ค่าความแม่นยำ	**Recall** ค่าความระลึก	**F-Measure** ค่าความถ่วงดุล
Support Vector Machine	0.87	0.88	0.87	0.87

จากตารางที่ 4-3 แสดงให้เห็นถึงการประเมินประสิทธิภาพของแบบจำลอง ด้วยเทคนิค Support Vector Machine โดยมี ค่าความถูกต้อง (Accuracy) เท่ากับ 0.87 ค่าความแม่นยำ (Precision) เท่ากับ 0.88 ค่าความระลึก (Recall) เท่ากับ 0.87 และค่าความถ่วงดุล (F-Measure) เท่ากับ 0.87



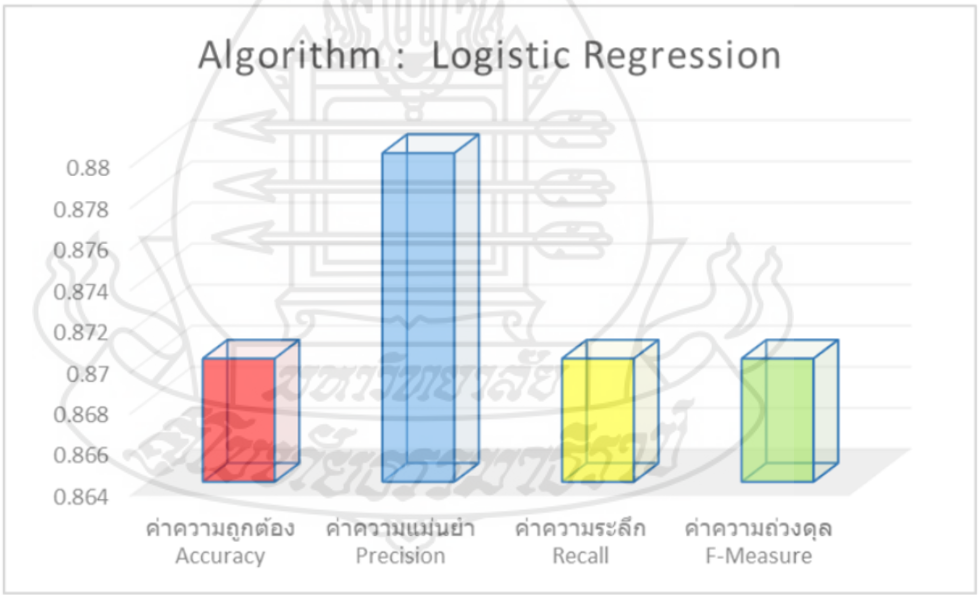
ภาพที่ 4.3 แผนภูมิกราฟเส้นการประเมินประสิทธิภาพ ด้วย Support Vector Machine

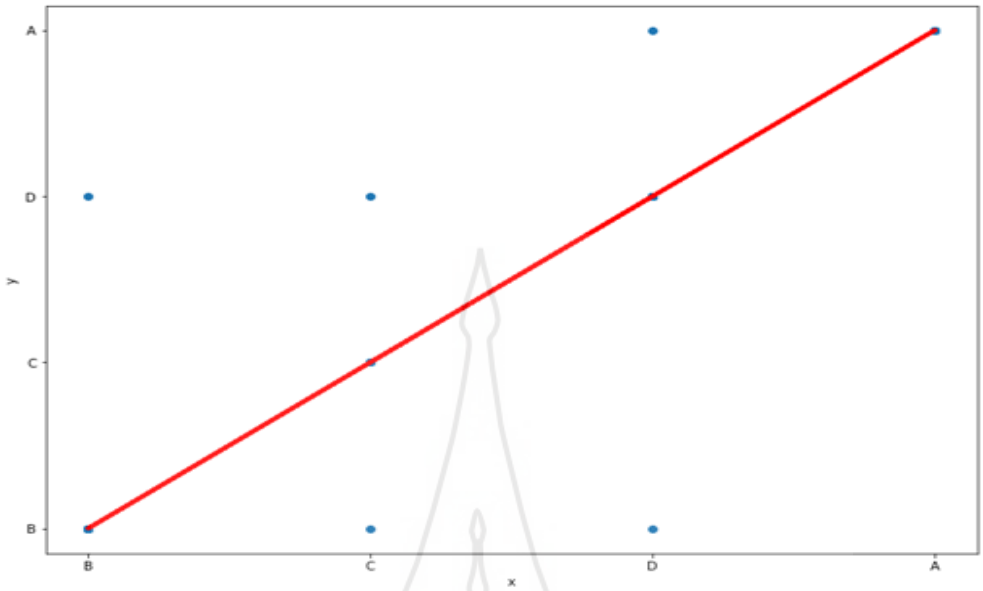
ตอนที่ 4 การประเมินประสิทธิภาพ ด้วย Logistic Regression

ตารางที่ 4.4 การประเมินประสิทธิภาพ ด้วย Logistic Regression

Algorithm	Accuracy ค่าความถูกต้อง	Precision ค่าความแม่นยำ	Recall ค่าความระลึก	F-Measure ค่าความถ่วงดุล
Logistic Regression	0.87	0.88	0.87	0.87

จากตารางที่ 4.4 แสดงให้เห็นถึงการประเมินประสิทธิภาพของแบบจำลอง ด้วยเทคนิค Logistic Regression โดยมี ค่าความถูกต้อง (Accuracy) เท่ากับ 0.87 ค่าความแม่นยำ (Precision) เท่ากับ 0.88 ค่าความระลึก (Recall) เท่ากับ 0.87 และค่าความถ่วงดุล (F-Measure) เท่ากับ 0.87





ภาพที่ 4.4 แผนภูมิกราฟเส้นการประเมินประสิทธิภาพ ด้วย Logistic Regression

ตอนที่ 5 ค่าความแม่นยำการจำแนกหมวดหมู่ด้วยเทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ต เวกเตอร์แมชชีน และการถดถอยโลจีสติก

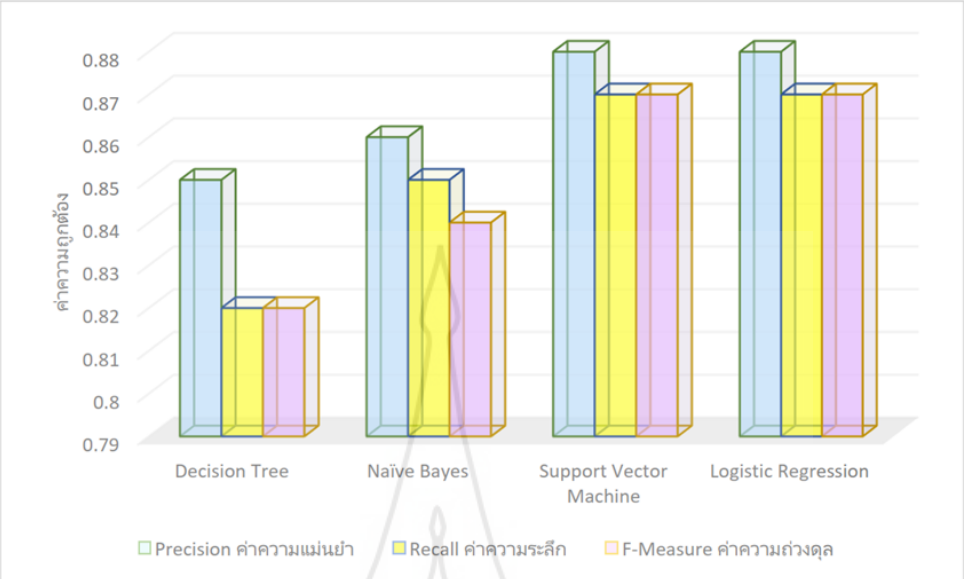
ตารางที่ 4.5 ค่าความแม่นยำการจำแนกหมวดหมู่ด้วยเทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ต เวกเตอร์แมชชีน และการถดถอยโลจีสติก

Algorithm	Precision	Recall	F1-Score	Support
Decision Tree				
A	0.60	1.00	0.75	3
B	0.67	0.86	0.75	7
C	0.67	0.50	0.57	4
D	0.95	0.84	0.89	25
Naïve Bayes				
A	1.00	0.71	0.83	7
B	0.78	0.78	0.78	9
C	1.00	0.60	0.75	5
D	0.82	1.00	0.90	18

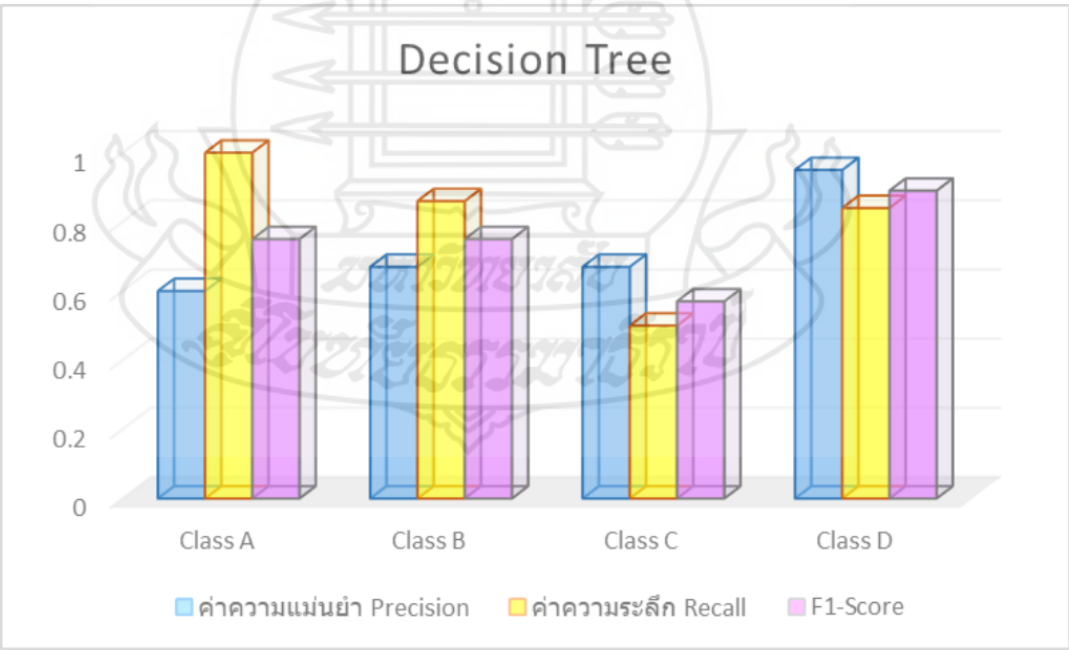
ตารางที่ 4.5 (ต่อ)

Algorithm	Precision	Recall	F1-Score	Support
Support Vector Machine				
A	0.80	1.00	0.89	4
B	0.78	0.78	0.78	9
C	1.00	0.75	0.86	4
D	0.91	0.91	0.91	22
Logistic Regression				
A	0.80	1.00	0.89	4
B	0.78	0.88	0.82	8
C	1.00	0.60	0.75	5
D	0.91	0.91	0.91	22

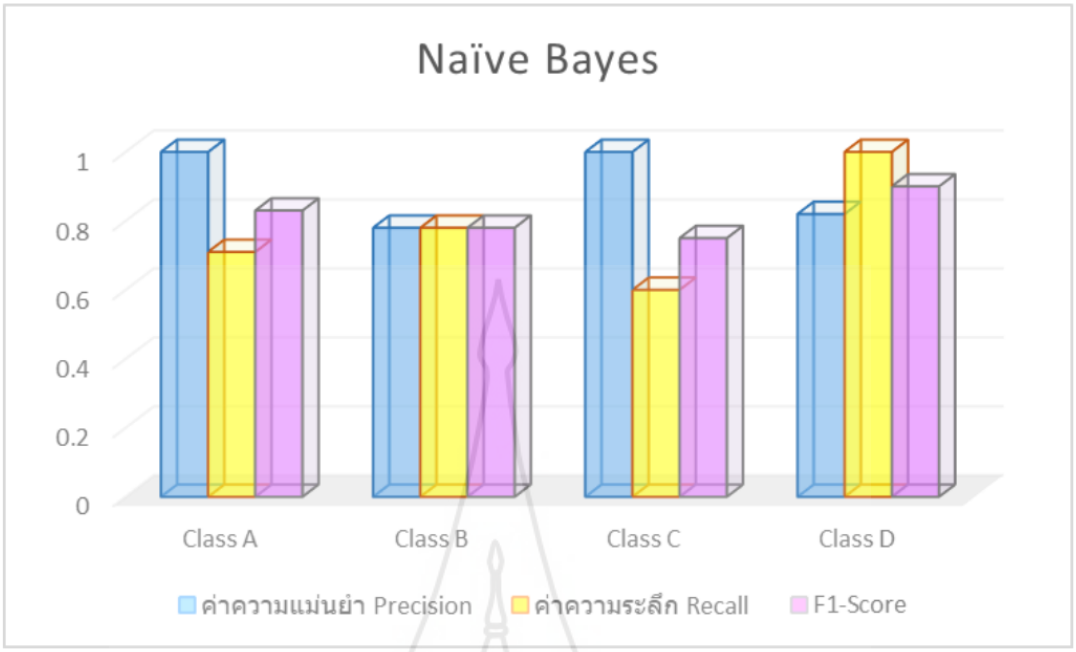
จากตารางที่ 4.5 แสดงเปรียบเทียบค่าความแม่นยำของการจำแนกหมวดหมู่ด้วยเทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ต เวกเตอร์แมชชีน และการถดถอยโลจีสติก โดยแบ่งตามคลาสประกอบด้วย A B C และ D สามารถสรุปผลได้ว่า แบบจำลอง Decision Tree ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class D เท่ากับ 0.95 ถัดมา Class B และ Class C เท่ากับ 0.67 Class A เท่ากับ 0.60 ตามลำดับ แบบจำลอง Naive Bayes ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class A และ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.82 และ Class B เท่ากับ 0.78 ตามลำดับ แบบจำลอง Support Vector Machine ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.91 Class A เท่ากับ 0.80 และ Class B เท่ากับ 0.78 ตามลำดับ แบบจำลอง Logistic Regression ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.91 Class A เท่ากับ 0.80 และ Class B เท่ากับ 0.78 ตามลำดับ



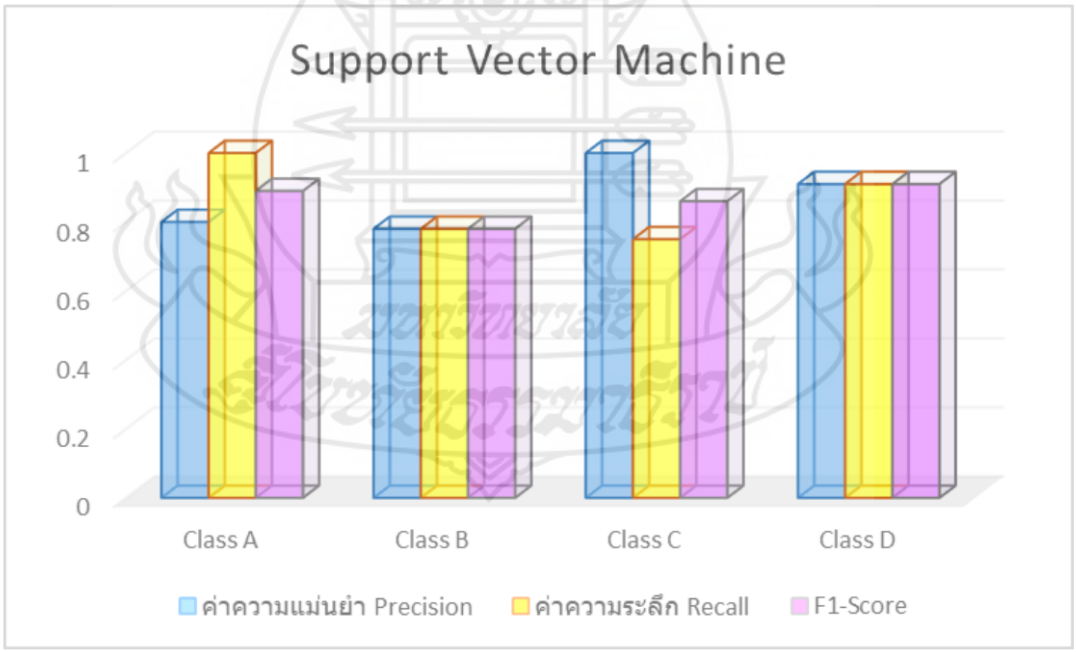
ภาพที่ 4.5 แผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกหมวดหมู่ ทั้ง 4 เทคนิค ประกอบด้วย เทคนิคต้นไม้ตัดสินใจ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน และการถดถอยโลจีสติก



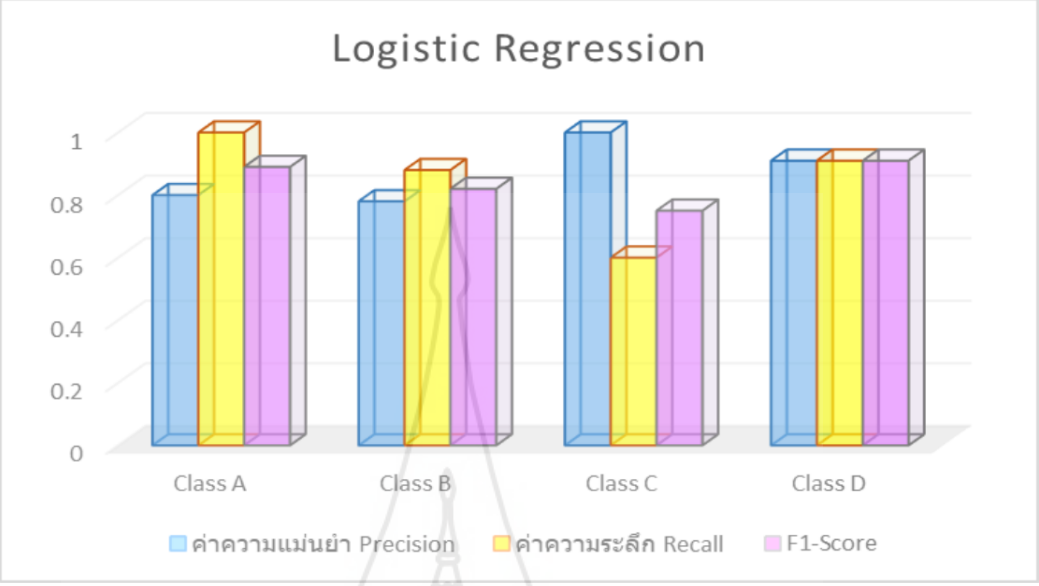
ภาพที่ 4.6 แผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกหมวดหมู่เทคนิคต้นไม้ตัดสินใจ



ภาพที่ 4.7 แผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกหมวดหมู่เทคนิคนาอีฟเบย์



ภาพที่ 4.8 แผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกหมวดหมู่เทคนิคซัพพอร์ตเวกเตอร์แมชชีน



ภาพที่ 4.9 แผนภูมิเปรียบเทียบค่าความถูกต้องการจำแนกหมวดหมู่เทคนิคการถดถอยโลจีสติก

ตอนที่ 6 ตารางการจำแนก Class A, Class B, Class C และ Class D

ตารางที่ 4.6 ตัวอย่างตารางการจำแนก Class A, Class B, Class C และ Class D

ที่	รายการ	A	B	C	D	IOC-A	IOC-B	IOC-C	IOC-D	สรุป
1	เคมี			1	2	0.0000	0.0000	0.3333	0.6667	D
2	ชีวะ			1	2	0.0000	0.0000	0.3333	0.6667	D
3	รังสี			1	2	0.0000	0.0000	0.3333	0.6667	D
4	นิวเคลียร์			1	2	0.0000	0.0000	0.3333	0.6667	D
5	คชรน.			1	2	0.0000	0.0000	0.3333	0.6667	D
6	สถานการณ์			3		0.0000	0.0000	1.0000	0.0000	C
7	การปฏิบัติการ			3		0.0000	0.0000	1.0000	0.0000	C
8	แก้ไข			2	1	0.0000	0.0000	0.6667	0.3333	C
9	การประเมิน	1	2			0.3333	0.6667	0.0000	0.0000	B
10	การวางแผน	1	2			0.3333	0.6667	0.0000	0.0000	B

ตารางที่ 4.6 (ต่อ)

ที่	รายการ	A	B	C	D	IOC-A	IOC-B	IOC-C	IOC-D	สรุป
11	สภาพแวดล้อม	1			2	0.3333	0.0000	0.0000	0.6667	D
12	ขีดความสามารถ	1	2			0.3333	0.6667	0.0000	0.0000	B
13	ภัยคุกคาม	1			2	0.3333	0.0000	0.0000	0.6667	D
14	ส่งกำลังบำรุง		1	2		0.0000	0.3333	0.6667	0.0000	C
15	การเตรียมการ		1	2		0.0000	0.3333	0.6667	0.0000	C
16	การประสานงาน		1	2		0.0000	0.3333	0.6667	0.0000	C
17	การตอบสนอง		1	2		0.0000	0.3333	0.6667	0.0000	C
18	การเตรียมพร้อม		1	2		0.0000	0.3333	0.6667	0.0000	C
19	ขั้นการปฏิบัติ		1	2		0.0000	0.3333	0.6667	0.0000	C
20	สิ่งบอกเหตุ		1		2	0.0000	0.3333	0.0000	0.6667	D

จากตารางที่ 4.6 แสดงตัวอย่างตารางการจำแนก Class A, Class B, Class C และ Class D พบว่าองค์ประกอบทุกด้านมีค่าความสอดคล้อง IOC (Index of Congruence: IOC) มีค่าความสอดคล้องระหว่าง 0.90 – 0.92 ซึ่งมีค่ามากกว่า 0.60 สามารถนำไปใช้งานได้

บทที่ 5
สรุปการวิจัย อภิปรายผล และข้อเสนอแนะ

1. สรุปการวิจัย

ในการวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษาและวิเคราะห์ระบบหนังสือเผยแพร่ความรู้ พัฒนาแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ และประเมินแบบจำลองการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ โดยทำการศึกษาข้อมูลคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้ จำนวน 948 คำ ตั้งแต่ปี พ.ศ. 2553 - 2563 มีขั้นตอนการดำเนินการวิจัย การพัฒนา Open Journal System (OIS) ซึ่งเป็นการจัดการเว็บไซต์และการเผยแพร่หนังสือเผยแพร่ความรู้ออนไลน์โดยมีการจำแนกหมวดหมู่คำสำคัญของหนังสือเผยแพร่ความรู้ สำหรับทำการวิเคราะห์ ข้อความหนังสือเผยแพร่ความรู้ เปรียบเทียบประสิทธิภาพการใช้แบบจำลองที่เหมาะสม และเปรียบเทียบประสิทธิภาพของอัลกอริทึม เพื่อหาแบบจำลองที่ดีและเหมาะสมในการจำแนกหมวดหมู่ของข้อความ เลือกใช้เทคนิคการจำแนกข้อมูล 4 เทคนิคโดยใช้อัลกอริทึม ได้แก่ Logistic Regression, Decision Tree, Naive Bayes และ Support Vector Machine และทำการแสดงผลโดย Word Cloud เพื่อให้ได้ข้อมูลมาใช้สำหรับการจำแนกหมวดหมู่ต่อไป ได้แก่ หมวดหมู่ยุทธศาสตร์ หมวดหมู่ยุทธการ หมวดหมู่ยุทธวิธี และหมวดหมู่อื่นๆ ในรูปแบบขั้นตอนการวิจัยและพัฒนา ประกอบด้วย 1) ศึกษารูปแบบที่เหมาะสมในการจำแนกข้อความหนังสือเผยแพร่ความรู้ 2) วิเคราะห์ข้อความหนังสือเผยแพร่ความรู้ 3) ประเมินประสิทธิภาพข้อมูลที่ได้จากการวิเคราะห์ข้อความหนังสือเผยแพร่ความรู้ จากผลการประเมินพบว่า อัลกอริทึมทั้งหมดสามารถจำแนกหมวดหมู่หนังสือเผยแพร่ความรู้ได้ 4 หมวดหมู่ ได้แก่ หมวดหมู่ยุทธศาสตร์ หมวดหมู่ยุทธการ หมวดหมู่ยุทธวิธี และหมวดหมู่อื่นๆ โดยอัลกอริทึมการถดถอยโลจีสติกและอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน มีค่าความถูกต้องเท่ากับ 0.87 อัลกอริทึมนาอีฟเบย์มีค่าความถูกต้องเท่ากับ 0.85 และอัลกอริทึมต้นไม้การตัดสินใจมีค่าความถูกต้องเท่ากับ 0.82 และพบว่าอัลกอริทึมการถดถอยโลจีสติกใช้เวลาในการจำแนกหมวดหมู่หนังสือเผยแพร่ความรู้ดีกว่าอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีนประมาณร้อยละ 11.8

2. อภิปรายผล

จากการทดสอบโดยใช้ตัวแปรข้อมูลคำสำคัญจากหนังสือฯ จำนวน 948 คำ ซึ่งเป็นข้อมูลข้อความที่เป็นทั้งภาษาไทยและภาษาอังกฤษ นำมาทำการเปรียบเทียบประสิทธิภาพการจำแนกประเภทข้อมูล 4 วิธี คือใช้อัลกอริทึม Decision Tree, Naive Bayes, Support Vector Machine และ Logistic Regression ซึ่งทำการทดสอบโดยการใช้วิธี Split Test เป็นการทดสอบข้อมูลด้วยการแบ่งข้อมูลออกเป็น 2 ส่วน คือส่วนข้อมูลสำหรับฝึกฝนหรือสร้างแบบจำลอง (Training Set) และส่วนข้อมูลสำหรับการทดสอบแบบจำลอง (Test Set) แบบจำลองแบ่งข้อมูล 2 ส่วน โดยกำหนดให้ Train 70% และให้ Test 30% จากผลการทดลองพบว่าในการเปรียบเทียบค่าความแม่นยำของแบบจำลองการจำแนกหมวดหมู่ด้วยอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) อัลกอริทึมนาอีฟเบย์ (Naive Bayes) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และอัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) โดยแบ่งตามคลาสประกอบด้วย คลาส A, คลาส B, คลาส C และ D ผลการประเมินพบว่า

อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class D เท่ากับ 0.95 ถัดมา Class B และ Class C เท่ากับ 0.67 Class A เท่ากับ 0.60 ตามลำดับ อัลกอริทึมนาอีฟเบย์ (Naive Bayes) ให้ค่าความแม่นยำ (Precision) มากสุด ที่ Class A และ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.82 และ Class B เท่ากับ 0.78 ตามลำดับ

อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.91 Class A เท่ากับ 0.80 และ Class B เท่ากับ 0.78 ตามลำดับ

อัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) ให้ค่าความแม่นยำ (Precision) มากสุดที่ Class C เท่ากับ 1.00 ถัดมา Class D เท่ากับ 0.91 Class A เท่ากับ 0.80 และ Class B เท่ากับ 0.78 ตามลำดับ

ส่วนการประเมินประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) นั้น เทคนิคอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเทคนิคอัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) มีค่าความถูกต้อง (Accuracy) สูงสุดเท่ากับ 87% อัลกอริทึมนาอีฟเบย์ (Naive Bayes) มีค่าความถูกต้อง (Accuracy) เท่ากับ 85% และอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) มีค่าความถูกต้อง (Accuracy) เท่ากับ 82% ตามลำดับ ซึ่งจากผลการวิจัยนี้ทั้งอัลกอริทึมซัพพอร์ต เวกเตอร์แมชชีน (Support Vector Machine) และ

อัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) มีผลการวัดประเมินประสิทธิภาพที่เท่ากัน โดยมีค่าความถูกต้อง (Accuracy) เท่ากับ 0.87 หรือ 87% และค่าความแม่นยำ (Precision) เท่ากับ 0.88 หรือ 88%

ผู้วิจัยให้ผลการประเมินประสิทธิภาพของเทคนิคอัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) ดีกว่าเทคนิคอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เนื่องจากมีระยะเวลาในการสร้างแบบจำลองที่ดีกว่า โดยเทคนิคอัลกอริทึมการถดถอยโลจีสติก (Logistic Regression) ใช้เวลา 0.363 วินาที และเทคนิคอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ใช้เวลา 0.412 วินาที ซึ่งประมาณร้อยละ 11.8 และจากการใช้ทั้ง 4 อัลกอริทึม สามารถจำแนกหมวดหมู่หนังสือเผยแพร่ความรู้ออกมาได้เป็น 4 หมวดหมู่ ได้แก่ หมวดหมู่ยุทธศาสตร์ หมวดหมู่ยุทธการ หมวดหมู่ยุทธวิธี และหมวดหมู่อื่นๆ ในการศึกษางานวิจัยที่เกี่ยวข้องทำให้ผู้วิจัยมองเห็นวิธีการ แนวทางในการทำวิจัย การ จำแนกเกี่ยวกับการหมวดหมู่ ซึ่งพบว่าเป็นการใช้เทคนิคการเรียนรู้ของเครื่องในการสกัดข้อมูลในการ จัดหมวดหมู่ และมีการประยุกต์ใช้ในรูปแบบของการพัฒนาระบบวารสารออนไลน์Open Journal System (OIS) มีการใช้เทคนิคการประมวลผลข้อความ กระบวนการจัดกลุ่มการจำแนก การจัดหมวดหมู่ของข้อความ ความคิดเห็น และอื่นๆ และจากงานวิจัยที่ได้ศึกษามีการใช้เทคนิคการใช้งานอย่างแพร่หลายเพราะให้ประสิทธิภาพในการทำงานที่ดีและง่ายในการทำความเข้าใจ เช่น Decision Trees, Support Vector Machine , Naive Bayes เพื่อพัฒนาแบบจำลองในการจำแนกหมวดหมู่ข้อความ นำมาใช้วิเคราะห์ เปรียบเทียบประสิทธิภาพค่าความถูกต้อง ในการทำงานวิจัยครั้งนี้เพื่อเปรียบเทียบหาประสิทธิภาพของแบบจำลองในการจำแนกหมวดหมู่ที่ดี ได้พิจารณาคำสำคัญจากข้อความในหนังสือเผยแพร่ความรู้รูปแบบไฟล์ PDF ผู้วิจัยจึงสนใจศึกษาเทคนิคและเลือกใช้อัลกอริทึม Decision Trees, Support Vector Machine, Naive Bayes, Logistic regression เพื่อหาค่าความถูกต้อง ค่าความแม่นยำ เลือกใช้เทคนิคการจำแนกประเภท (Text classification) เพื่อมาช่วย ในการจำแนกข้อความ และจากการศึกษางานวิจัยสอดคล้องกับงานวิจัยของ (พรรณนาภรณ์ เกตุภู พงษ์, 2561) ที่มีการใช้เทคนิค Decision Trees, Naive Bayes และ Support Vector Machine มา ประยุกต์ใช้การทำเหมืองข้อความเพื่อจำแนกประเภทโรคจากอาการ และจากการศึกษางานวิจัยที่ เกี่ยวข้องยังไม่มีงานวิจัยใดประยุกต์ใช้ Word cloud ผู้วิจัยจึงเลือกนำมาใช้การแสดงผลลัพธ์ในด้าน แสดงความถี่ของคำด้วย Word cloud และจำแนกหมวดหมู่โดยแบ่งเป็น 4 หมวดให้เหมาะกับหนังสือเผยแพร่ความรู้เพื่อเป็นการเพิ่มประสิทธิภาพแบบจำลองสำหรับการจำแนกหมวดหมู่ข้อความ ที่ดีและเหมาะสม และพัฒนาประสิทธิภาพในการดำเนินการให้ดียิ่งขึ้นให้สามารถตอบสนองความ ต้องการของหน่วยงานและผู้สนใจต่อไปในอนาคต

3. ข้อเสนอแนะ

การพัฒนาแบบจำลองที่เหมาะสมในการจำแนกหมวดหมู่ข้อความหนังสือเผยแพร่ความรู้ใน ส่วนการเตรียมข้อมูลสำหรับนำเข้า ควรมีการรวบรวมข้อมูลคำสำคัญจากหนังสือเผยแพร่ความรู้ที่มี ปริมาณที่มากขึ้น เพื่อให้การวัดประประสิทธิภาพแบบจำลองมีความถูกต้อง แม่นยำมากขึ้นและทำ การเปรียบเทียบอัลกอริทึมอื่นๆ ที่อาจเหมาะสม และพร้อมใช้งานมากยิ่งขึ้นไป

บรรณานุกรม

บรรณานุกรม

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI:Machine Learning ด้วย Python*. กรุงเทพมหานคร: ค่านสุทธาการพิมพ์ จำกัด.

โกเมศ อัมพวัน. (ม.ป.ป). *แบบจำลองข้อมูล.* สืบค้นจาก https://staff.informatics.buu.ac.th/~komate/886301/DB-Chapter-2.pdf

คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร. (ม.ป.ป). *คู่มือการใช้งานระบบวารสาร.* สืบค้นจาก https://www.sci.nu.ac.th/sciencejournal/files/Manual%20OJS%20for%20Author.pdf

ชิตพงษ์ กิตตินราดร. (2563). *Support Vector Machines*. สืบค้นจาก https://guopai.github.io/ml-blog08.html

ปราณี วงศ์จำรัส และคณะ. (ม.ป.ป). *วารสารอิเล็กทรอนิกส์(E-journal).* สืบค้นจาก https://lis.human.cmu.ac.th/web2018/e-learning/009230/lesson1/e-journal.htm

พัชราภรณ์ สิทธิคำฟู. (2557). *การจำแนกหมวดหมู่ข้อความข่าวสารภัยพิบัติอุทกภัยจากแหล่งข้อมูลสาธารณะภาษาไทย (สารนิพนธ์ วิทยาศาสตรมหาบัณฑิต).* มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, กรุงเทพมหานคร.

แมตต์มิลส์. (2564). *Word Cloud ดีสำหรับอะไร.* สืบค้นจาก https://itigic.com/th/what-are-word-clouds-good-for/

วราพรรณ อภิศุภะโชค. (2550). *การจัดการวารสารอิเล็กทรอนิกส์.* วารสารมนุษยศาสตร์, 14(2), 34. สืบค้นจาก https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiao43i6df5AhXmT2wGHaIpDVAQFnoECCIQAQ&url=https%3A%2F%2Fso04.tci-thaijo.org%2Findex.php%2Fabc%2Farticle%2Fdownload%2F52802%2F43843%2F122336&usg=AOvVaw3UN9DDv6ETsYQu21eDzRJI

วัลลภา อุทยาน. (2547). *วารสารอิเล็กทรอนิกส์.* สืบค้นจาก https://www.nectec.or.th/schoolnet/library/createweb/10000/generality/10000-12235.html

ศศิมา มณฑาสุวรรณ. (2557). *การพัฒนาระบบค้นหารหัส ICD – 10 สำหรับงานเวชระเบียน (วิทยานิพนธ์ปริญญามหาบัณฑิต).* มหาวิทยาลัยศิลปากร, กรุงเทพมหานคร. สืบค้นจาก http://www.sure.su.ac.th/xmlui/bitstream/id/7d338c62-839e-43f0-8af9-2473ac83d7a0/Sasima_Monthasuwan_fulltext.pdf?attempt=2

ศิริเดช สุชีวะ. (2558). *การวิเคราะห์ถดถอยโลจิสติก: แนวคิด การวิเคราะห์ และการแปรความหมาย*. สืบค้นจาก https://portal.edu.chula.ac.th/pub/jrm/index.php//jrm/article/view/193

สารานุกรมเสรี วิกิพีเดีย. (ม.ป.ป). *Open Journal System*. สืบค้นจาก https://th.wikipedia.org/wiki/Open_journal_system

สารานุกรมเสรี วิกิพีเดีย. (ม.ป.ป). *ต้นไม้ตัดสินใจ*. สืบค้นจาก https://th.wikipedia.org/wiki/%E0%B8%95%E0%B9%89%E0%B8%99%E0%B9%84%E0%B8%A1%E0%B9%89%E0%B8%95%E0%B8%B1%E0%B8%94%E0%B8%AA%E0%B8%B4%E0%B8%99%E0%B9%83%E0%B8%88

สารานุกรมเสรี วิกิพีเดีย. (ม.ป.ป). *แท็กคลาวด์*. สืบค้นจาก https://th.wikipedia.org/wiki/%E0%B9%81%E0%B8%97%E0%B9%87%E0%B8%81%E0%B8%84%E0%B8%A5%E0%B8%B2%E0%B8%A7%E0%B8%94%E0%B9%8C

อนัตต์ชัย ชุติภาสเจริญ และ จรัญ แสนราช. (2561). *การเปรียบเทียบประสิทธิภาพของอัลกอริทึมและการคัดเลือกคุณลักษณะที่เหมาะสมเพื่อการพยากรณ์โอกาสความสำเร็จในการโอนเงินข้ามประเทศของบุคคลทั่วไป*. วารสารวิจัย มหาวิทยาลัยขอนแก่น (ฉบับบัณฑิตศึกษา) สาขามนุษยศาสตร์และสังคมศาสตร์, 6(3), 107. สืบค้นจาก https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwi0l_6R8e75AhVETmwGHTiQABcQFnoECAYQAw&url=https%3A%2F%2Fso04.tci-thaijo.org%2Findex.php%2Fgskkuhs%2Farticle%2Fdownload%2F156370%2F113480%2F426187&usg=AOvVaw2YvMw_Vfszk8gUq0_LvDAx

อนันตต์ชัย ชุติภาสเจริญ และ จรัญ แสนราช. (2561). *การจำแนกประเภทข้อมูล*. วารสารวิจัยมหาวิทยาลัยขอนแก่น (ฉบับบัณฑิตศึกษา), 6(3), 106-107. สืบค้นจาก https://so04.tcithaijo.org/index.php/gskkuhs/article/download/156370/113480/426187

อนุพงศ์ สุขประเสริฐ. (ม.ป.ป). *Chapter 8 Classification*. สืบค้นจาก https://slideplayer.in.th/slide/14904208/

Jaruwit Pratancheewin. (2562). *เรียนรู้และทำความเข้าใจเรื่อง Support Vector Machine (SVM) คืออะไร*. สืบค้นจาก https://www.glurgeek.com/education/support-vector-machine/

My Little Learn. (2563). *สอนทำ Word Cloud ด้วย MS Word 2 นาทีรู้เรื่อง*. สืบค้นจาก https://mylittlelearn.com/knowledge-for-your-kids/%E0%B8%AA%E0%B8%AD%E0%B8%99%E0%B8%97%E0%B8%B3-word-cloud-with-ms-word-2-minutes/

natthasath. (2561). *Support Vector Machine*. สืบค้นจาก https://codeinsane.wordpress.com/2018/12/08/support-vector-machine/

Softengthai. (2012). *แบบจำลองระบบ*. สืบค้นจาก
http://lprusofteng.blogspot.com/2012/03/blog-post_11.html

tatiya. (2562). *มาทำความรู้จักกับ Word Cloud*. สืบค้นจาก
https://www.mindphp.com/forums/viewtopic.php?f=144&t=49427

VithanMinaphinant. (2561). *Machine Learning คืออะไร?*. สืบค้นจาก
https://medium.com/investic/machine-learning%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-fa8bf6663c07

Weerakaset Suanpaga. (ม.ป.ป). *โมเดล หรือแบบจำลอง*. สืบค้นจาก
https://pirun.ku.ac.th/~fengwks/SD/2model.pdf

ภาคผนวก

ภาคผนวก ก

รายการหนังสือเผยแพร่ความรู้ ย้อนหลัง 5 ปี (2559 – 2563)

รายการหนังสือเผยแพร่ความรู้ ย้อนหลัง 5 ปี (2559 – 2563)

ปี	รายการ
2559	1. ความรู้พื้นฐานการพัฒนาหลักนิยมกำลังรบผสมเหล่าระดับกรม (กรมผสม) ทบ. ไทย
	2. สภาพแวดล้อมทางยุทธศาสตร์หลังสงครามโลกครั้งที่ 2
2560	3. ยุทธศาสตร์จีนบนเวทีโลก (ฉบับปรับปรุง)
	4. ยุทธศาสตร์กองทัพบก
	5. บันทึกประสบการณ์ของนายทหารใหม่ในเวียดนาม
2561	6. คู่มือการฝึกเพื่อเอาชนะในโลกที่มีความซับซ้อน FM 7-0
	7. คู่มือภาษาอังกฤษพื้นฐานสำหรับกำลังพลในกองทัพบกไทย (ฉบับปรับปรุงครั้งที่ 2)
	8. การฝึกทักษะเพื่อปรับเปลี่ยนและพลิกแพลงเพื่อเอาชนะ
2562	9. การพิทักษ์สุขภาพกำลังรบ
	10. การวางแผนการทัพ
	11. กองทัพบกในการช่วยเหลือภัยพิบัติ
2563	12. การปฏิบัติการแก้ไขสถานการณ์ อันเนื่องมาจากเคมี ชีวะ รังสีและนิวเคลียร์
	13. กำลังรบผสมเหล่า ระดับกรม (กรมผสม)
	14. การวางแผนปฏิบัติการยุทธ์ร่วม

ภาคผนวก ข

ตัวอย่างโค้ดโปรแกรมการทำงาน

ตัวอย่างโค้ดการนำเข้าไฟล์ข้อมูล

```
df = pd.read_csv('data_key.csv', sep='\t', names=['ID', 'keyword'
, 'class'], header=None)
df
```

ตัวอย่างโค้ดการตัดคำ

```
from pythainlp.corpus.common import thai_stopwords
thai_stopwords = list(thai_stopwords())
thai_stopwords
```

```
from pythainlp import word_tokenize
def text_process(text):
    final = "".join(u for u in text if u not in ("?", ".", ";", "
:", "!", '"', "ๆ", "ฯ","," ))
    final = word_tokenize(final)
    final = " ".join(word for word in final)
    final = " ".join(word for word in final.split()
                     if word.lower not in thai_stopwords)
    return final
df['text_tokens'] = df['keyword'].apply(text_process)
df
```

ตัวอย่างโค้ดการสร้าง WordCloud

```
from wordcloud import WordCloud, STOPWORDS
df_pos = df[df['class'] == 'A']
pos_word_all = " ".join(text for text in df_pos['text_tokens'])
reg = r"[ก-ฮa-zA-Z']+"
fp = 'THSarabunNew.ttf'
wordcloud = WordCloud(stopwords=thai_stopwords, background_color
= 'white', max_words=2000, height = 2000, width=4000, font_path=f
p, regexp=reg).generate(pos_word_all)
plt.figure(figsize = (16,8))
plt.imshow(wordcloud)
plt.axis('off')
plt.show()
```

ตัวอย่างโค้ดการสร้างแบบจำลอง

```
from sklearn.model_selection import train_test_split
X = df[['text_tokens']]
y = df['class']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_si
ze=0.3, random_state=101)

from sklearn.feature_extraction.text import CountVectorizer
cvec = CountVectorizer(analyzer=lambda x:x.split(' '))
cvec.fit_transform(X_train['text_tokens'])
cvec.vocabulary_

train_bow = cvec.transform(X_train['text_tokens'])
pd.DataFrame(train_bow.toarray(), columns=cvec.get_feature_names(
), index=X_train['text_tokens'])
```

ตัวอย่างโค้ดการประเมินประสิทธิภาพแบบจำลอง

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(train_bow, y_train)

from
sklearn.metrics import confusion_matrix,classification_report

test_bow = cvec.transform(X_test['text_tokens'])

test_predictions = lr.predict(test_bow)
print(classification_report(test_predictions, y_test))
```

ประวัติผู้วิจัย

ชื่อ	นางสาวประภัสสร ข่ายกระโทก
วัน เดือน ปีเกิด	5 ตุลาคม 2538
สถานที่เกิด	อำเภอโชคชัย จังหวัดนครราชสีมา
ประวัติการศึกษา	ปริญญาตรีบริหารธุรกิจบัณฑิต มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน พ.ศ. 2560
สถานที่ทำงาน	กรุงเทพมหานคร
ตำแหน่ง	พนักงานคอมพิวเตอร์