การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้อัลกอริทึม FPTC

นางสาวณิชาพร สุระ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ
บัณฑิตวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2549
ลิขสิทธิ์ของสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ

ชื่อ : นางสาวณิชาพร สุระ
ชื่อวิทยานิพนธ์ : การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้อัลกอริทึม FPTC
สาขาวิชา : วิทยาการคอมพิวเตอร์
สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ที่ปรึกษาวิทยานิพนธ์ : ผู้ช่วยศาสตราจารย์ ดร.เยาวดี เต็มธนาภัทร์
ปีการศึกษา : 2549

บทคัดย่อ

ปัจจุบันเอกสารในรูปแบบอิเล็กทรอนิกส์มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น การสืบค้นและการจัดการเอกสารจะง่าย และเป็นไปตามความต้องการ ต้องอาศัยการจัดแบ่งเอกสารเป็นกลุ่มหรือหมวดหมู่ ให้สอดคล้องและตรงกับดัชนี เพื่อให้จัดเก็บและค้นคืนเอกสารได้อย่างรวดเร็วและมีประสิทธิภาพ งานวิจัยนี้ประสงค์เพื่อพัฒนาเครื่องมือในการจำแนกหมวดหมู่เอกสารข้อความภาษาไทยด้วยอัลกอริทึม Feature Projection Text Categorization (FPTC) ซึ่งเป็นอัลกอริทึมที่ปรับมาจาก k-Nearest Neighbor ลักษณะเด่นของ FPTC คือ การแทนคุณลักษณะในแบบภาพฉายของแต่ละคุณลักษณะ การจำแนกหมวดหมู่จะใช้วิธีการเปรียบเทียบความคล้ายของคำที่ปรากฏในเอกสารที่ใช้ทดสอบกับคำที่ปรากฏในเอกสารที่ใช้ในกระบวนการเรียนรู้ เพื่อหาเอกสารที่คล้ายกับเอกสารทดสอบมากที่สุด และกำหนดหมวดหมู่ของเอกสารนั้นให้กับเอกสารทดสอบ โดยจะใช้เอกสารข่าวภาษาไทยจากหนังสือพิมพ์ออนไลน์เป็นกรณีศึกษา

จากผลการทดสอบพบว่า การจำแนกหมวดหมู่ด้วยอัลกอริทึม FPTC สามารถจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิผลดี สำหรับข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน

(วิทยานิพนธ์มีจำนวนทั้งสิ้น 61 หน้า)

คำสำคัญ : การจำแนกหมวดหมู่เอกสาร, โปรเจกชันของคุณลักษณะ

_____อาจารย์ที่ปรึกษาวิทยานิพนธ์

Name : Miss Nichaporn Sura

Thesis Title : Automatic Thai Text Categorization Using FPTC Algorithm

Major Field : Computer Science

King Mongkut's Institute of Technology North Bangkok

Thesis Advisor : Assistant Professor Dr.Yaowadee Temtanapat

Academic Year : 2006

Abstract

Amounts of Electronic documents have more increased and their contexts have more various. It needs competent management and retrieval systems for fast and most satisfying retrieval that required efficient indexing include document categorization. This research experiment Thai text document categorization according to prearranged categories by using feature projection text categorization (FPTC) in training and classification. Classification perform on contents of documents by comparing similarities of terms presented in test documents with similarities of terms presented in training documents

(Total 61 pages)

Keywords:Text Categorization, Text Classification, Feature Projection

_____ Advisor

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือสนับสนุนอย่างดียิ่งของผู้ช่วยศาสตราจารย์ ดร.เยาวดี เต็มธนาภัทร์ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร. กฤษมันต์ วัฒนาณรงค์ และอาจารย์ ดร.เนตรนภา สีหารี คณะกรรมการสอบวิทยานิพนธ์ รวมทั้งอาจารย์ประจำภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศทุกท่านที่ให้คำแนะนำและข้อคิดเห็นต่างๆ ตลอดจนให้ความสนับสนุนทั้งทางด้านเอกสาร ตำรา อุปกรณ์คอมพิวเตอร์ และกำลังใจในการทำวิจัย ผู้จัดทำขอกราบขอบพระคุณในความกรุณาของท่านเป็นอย่างยิ่งไว้ ณ โอกาสนี้ด้วย

ความสำเร็จนี้จะเกิดขึ้นไม่ได้ หากไม่ได้รับการช่วยเหลือสนับสนุนทั้งด้านกำลังกายและกำลังใจจากเพื่อนร่วมงานที่สำนักคอมพิวเตอร์และเทคโนโลยีสารสนเทศทุกท่าน โดยเฉพาะคุณโสภิดา แพรดำ คุณธัญนันท์ กระดาษ คุณทักษพล พึ่งยนต์ และคุณสุทธินนท์ ชื่อสุธรรม ซึ่งเป็นกำลังสำคัญอย่างยิ่ง

ท้ายนี้ผู้จัดทำขอกราบขอบพระคุณบิดา มารดา ท่านผู้มีพระคุณ ญาติพี่น้อง เพื่อนร่วมรุ่น MCS01 เพื่อนร่วมรุ่น CS07 ตลอดจนผู้ร่วมงานทุกท่านในสจพ. ที่สละเวลาให้คำปรึกษาและความช่วยเหลือสนับสนุนจนทำให้การจัดทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ณิชาพร สุระ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

คำอธิบายสัญลักษณ์และคำย่อ

$dist_m$	:	ระยะห่างระหว่างสมาชิกสองค่าบนคุณลักษณะที่ m
df	:	ความถี่ของเอกสาร
idf	:	ส่วนกลับของความถี่เอกสาร
f	:	จำนวนคุณลักษณะ
k	:	จำนวนลำดับของเอกสารที่คล้ายกับต้นแบบมากที่สุด
log	:	ค่า Logarithm ฐานสิบ
P	:	ค่าความแม่นยำ (Precision)
R	:	ค่าความระลึก (Recall)
tf	:	ความถี่ของคำ
tfidf	:	ค่าผลคูณของความถี่ของคำกับส่วนกลับของความถี่เอกสาร
$\sum$	:	การหาผลรวม

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเอกสารในรูปแบบอิเล็กทรอนิกส์มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น การสืบค้นและการจัดการเอกสารจะง่าย และเป็นไปตามความต้องการ ต้องอาศัยการจัดแบ่งเอกสารเป็นกลุ่มหรือหมวดหมู่ ให้สอดคล้องและตรงกับดัชนี เพื่อให้จัดเก็บและค้นคืนเอกสารได้อย่างรวดเร็วและมีประสิทธิภาพ

งานวิจัยทางด้านการจัดแบ่งเอกสารเป็นกลุ่มตามเนื้อหานั้น ได้รับความสนใจ และมีการนำเสนอเทคนิควิธีการจัดกลุ่มหรือหมวดหมู่โดยการเรียนรู้ด้วยคอมพิวเตอร์หลายวิธีการ เทคนิควิธีการเหล่านั้นสามารถนำมาประยุกต์ใช้กับข้อมูลหลายประเภท ทั้งข้อความ รูปภาพ และเสียง Sebastiani [1] ได้สรุปรวบรวมแนวคิดของวิธีการเหล่านั้นไว้อย่างน่าสนใจ และ Marquez [2] ได้ศึกษาเกี่ยวกับการนำวิธีการเรียนรู้ด้วยคอมพิวเตอร์มาประยุกต์ใช้กับการประมวลผลภาษาธรรมชาติ

การจัดแบ่งเอกสารนั้นสามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization)

การจัดกลุ่มเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน งานวิจัยในกลุ่มนี้ ได้แก่ [3], [4]

การจำแนกหมวดหมู่เอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสาร โดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่ เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด สำหรับงานวิจัยนี้จะทำการจำแนกหมวดหมู่เอกสารประเภทข้อความ งานวิจัยในกลุ่มนี้ อาทิ [5], [6], [7], [8], [9] เป็นต้น

ตั้งแต่ทศวรรษที่ 60 เป็นต้นมาจนถึงปัจจุบัน มีการนำเสนอวิธีการในการจำแนกหมวดหมู่หลายวิธีการ เช่น Rocchio [3] Naïve Bayes [6] Decision Tree [9] Neural Network [10] k-NN [11, 12, 14] Centroid-based [7, 8, 15] SVM [16] Hidden Markov Model (HMM) [17] Boosting [8] และ k-NN on Feature Projection [9, 13, 19] นอกจากนี้ยังมีงานวิจัยที่ศึกษาเปรียบเทียบ

ประสิทธิภาพของแต่ละวิธีการ ได้แก่ [7], [9], [12], [13], [19], [20], [21], [22] เป็นต้น

ในปัจจุบันงานวิจัยทางด้านนี้ได้มุ่งเน้นการพัฒนาวิธีการในการจำแนกหมวดหมู่ของเอกสารข้อความให้มีประสิทธิภาพมากขึ้น ทั้งในแง่ของผลลัพธ์ที่มีความถูกต้อง ระยะเวลาในการประมวลผลที่น้อยลง และความสามารถในการทำงานได้กับหลายภาษา

สำหรับการจำแนกหมวดหมู่เอกสารภาษาไทย มีงานวิจัยหลายงานที่ได้ศึกษาทดลองจำแนกหมวดหมู่เอกสารภาษาไทยโดยใช้อัลกอริทึมในการเรียนรู้ที่แตกต่างกัน โดยเท่าที่ศึกษาค้นคว้ามาพบว่า อัลกอริทึมหรือวิธีการที่มีการนำมาใช้กับเอกสารภาษาไทยได้แก่ Rocchio [3] SVM [5] และ Naïve bayes [6]

งานวิจัยฉบับนี้จะนำวิธีการ Feature Projection Text Categorization (FPTC) มาประยุกต์ใช้ในการจำแนกหมวดหมู่เอกสารข้อความภาษาไทย เนื่องจากมีงานวิจัยที่ทำกับภาษาต่างประเทศ [9, 13, 19, 20] ที่แสดงให้เห็นว่า วิธีการนี้มีประสิทธิภาพดีทั้งในแง่ของความถูกต้องในการจัดหมวดหมู่ และระยะเวลาที่ใช้ในการประมวลผล

1.2 วัตถุประสงค์

1.2.1 เพื่อศึกษาเทคนิควิธีการเรียนรู้ FPTC สำหรับการจำแนกหมวดหมู่เอกสารข้อความ

1.2.2 เพื่อพัฒนาระบบที่ใช้สำหรับการจำแนกหมวดหมู่ของเอกสารภาษาไทยอัตโนมัติ

1.2.3 เพื่อทดสอบประสิทธิภาพของอัลกอริทึม FPTC เมื่อนำมาประยุกต์ใช้กับภาษาไทย

1.3 ขอบเขตของการวิจัย

1.3.1 นำอัลกอริทึม FPTC มาใช้สำหรับการจำแนกหมวดหมู่เอกสารข้อความภาษาไทย โดยใช้ข้อมูลข่าวจากเว็บไซต์หนังสือพิมพ์ออนไลน์ที่ถูกจัดหมวดหมู่ไว้แล้วเป็นกรณีศึกษา

1.3.2 ทดสอบประสิทธิภาพของอัลกอริทึม FPTC เมื่อนำมาประยุกต์ใช้กับภาษาไทย โดยศึกษาเพื่อหาลักษณะข้อมูล ค่าความถี่ของเอกสารสำหรับการลดจำนวนคุณลักษณะ และจำนวนเอกสารที่ใกล้เคียงกับเอกสารทดสอบที่เหมาะสมกับตัวจำแนกเอกสารที่พัฒนางานขึ้นในงานวิจัยนี้

1.4 วิธีการวิจัย

1.4.1 ศึกษาอัลกอริทึม FPTC เพื่อนำมาใช้ในการจำแนกหมวดหมู่

1.4.2 ศึกษางานวิจัยที่เกี่ยวข้องกับการจำแนกหมวดหมู่ของเอกสารเพื่อใช้เป็นพื้นฐานในการดำเนินงานวิจัย

1.4.3 ศึกษาและจัดหาเครื่องมือสำหรับการประมวลผลส่วนหน้า เพื่อเตรียมข้อมูลเข้าก่อนการจำแนกเอกสาร

1.4.4 ออกแบบและพัฒนาระบบการจำแนกหมวดหมู่เอกสารข้อความโดยอัตโนมัติด้วยอัลกอริทึม FPTC

1.4.5 จัดเตรียมข้อมูลเข้าสำหรับการทดสอบ โดยเก็บรวมรวมข้อมูลข่าวจากเว็บไซต์

1.4.6 ทดสอบเพื่อประเมินผลการวิจัย

1.4.7 สรุปผลการวิจัย

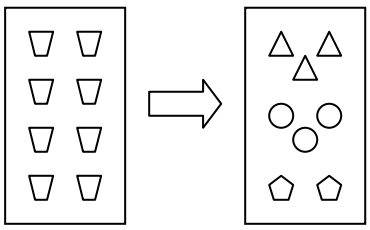
1.5 ประโยชน์ของการวิจัย

งานวิจัยนี้สามารถใช้ประโยชน์ในการจัดเอกสารอัตโนมัติ และเป็นพื้นฐานในการจำแนกหมวดหมู่ สามารถนำมาประยุกต์ใช้กับงานหลายด้าน เช่น การจัดระบบเอกสาร (Document Organization) การคัดกรองเอกสาร (Document Filtering) การจัดทำดัชนีอัตโนมัติเพื่อใช้ในการค้นคืนเอกสาร (Automatic Indexing for IR System) การแก้ปัญหาความหมายกำกวมของคำ (Word-sense Disambiguation) การจัดหมวดหมู่ของเว็บเพจ เป็นต้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

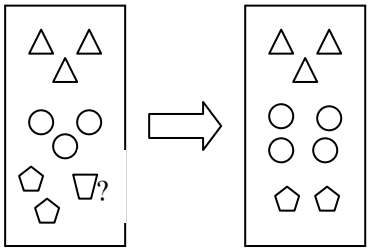
ในช่วงทศวรรษที่ผ่านมา งานทางด้านการจัดการกับเอกสารในแง่ของเนื้อหา หรือที่เรียกกันว่า การค้นคืนสารสนเทศ (Information Retrieval) นั้น ได้รับความสนใจอย่างมากใน เนื่องมาจากการเพิ่มปริมาณของข้อมูลดิจิตอลที่สามารถนำมาใช้งานได้และความต้องการในการเข้าถึงข้อมูลเหล่านั้นด้วยวิธีการที่หลากหลาย การแบ่งเอกสารตามเนื้อหาเพื่อการค้นคืนสารสนเทศนั้นสามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization)

การจัดกลุ่มของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ดังแสดงในภาพที่ 2-1



ภาพที่ 2-1 การจัดกลุ่ม

การจำแนกหมวดหมู่ของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่ เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด ดังแสดงในภาพที่ 2-2



ภาพที่ 2-2 การจำแนกหมวดหมู่

2.1 ความหมายของการจำแนกหมวดหมู่เอกสาร

การจำแนกหมวดหมู่เอกสาร (Text Categorization) คือ กิจกรรมในการแยกเอกสารซึ่งประกอบด้วยภาษาธรรมชาติให้อยู่ภายใต้หมวดหมู่ที่กำหนดไว้ก่อน โดยใช้ใจความสำคัญของเอกสาร การจำแนกหมวดหมู่ของเอกสารมีมาตั้งแต่ช่วงต้นของทศวรรษที่ 60 แต่กลายมาเป็นหัวข้อหนึ่งที่สำคัญในงานวิจัยทางด้านการค้นคืนสารสนเทศเมื่อช่วงต้นทศวรรษที่ 90 โดยในช่วงปลายของทศวรรษที่ 80 วิธีการที่นิยมใช้ในการจำแนกหมวดหมู่เอกสารคือ วิธีการทางด้านวิศวกรรมความรู้ (Knowledge Engineering) ซึ่งเป็นการกำหนดชุดของกฎที่สร้างจากความรู้ของผู้เชี่ยวชาญเพื่อใช้แยกประเภทของเอกสารภายใต้หมวดหมู่ที่กำหนด โดยการกำหนดชุดของกฎดังกล่าวต้องกระทำโดยมนุษย์ ต่อมาในทศวรรษที่ 90 วิธีการนี้ได้รับความนิยมลดน้อยลงเมื่อเทียบกับวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ (Machine Learning) ซึ่งเป็นการสร้างตัวจำแนกหมวดหมู่เอกสารอัตโนมัติ (Automatic Text Classifier) ด้วยเครื่องคอมพิวเตอร์โดยการเรียนรู้ด้วยวิธีการเชิงอุปนัยจากชุดของเอกสารที่ได้จำแนกประเภทไว้ก่อน และลักษณะเฉพาะของหมวดหมู่ที่เกี่ยวข้อง ข้อดีของวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ คือ ความถูกต้องของผลลัพธ์ที่ใกล้เคียงกับผลการจำแนกหมวดหมู่ของเอกสารที่ทำโดยมนุษย์ และการประหยัดแรงงานมนุษย์เป็นอย่างมาก เพราะไม่ต้องใช้ผู้เชี่ยวชาญในการสร้างตัวจำแนกประเภทเอกสาร หรือในการปรับเปลี่ยนหมวดหมู่ของเอกสาร

การจำแนกหมวดหมู่ของเอกสาร เป็นการกำหนดค่าความจริง (Boolean Value) ให้กับคู่ลำดับ $\langle d_j, c_i \rangle \in D \times C$ โดยที่ D เป็นโดเมนของเอกสาร และ $C = \{c_1, ..., c_{|C|}\}$ เป็นเซตของหมวดหมู่เอกสารที่กำหนดไว้ การกำหนดค่า T ให้กับคู่ลำดับ $\langle d_j, c_i \rangle$ จะบ่งบอกว่าเอกสาร d_j อยู่ภายใต้หมวดหมู่ c_i และการกำหนดค่า F ให้กับคู่ลำดับ $\langle d_j, c_i \rangle$ จะบ่งบอกว่าเอกสาร d_j ไม่อยู่ภายใต้หมวดหมู่ c_i หรืออาจกล่าวได้ว่า เป็นการประมาณค่าของฟังก์ชั่นเป้าหมายที่ไม่ทราบค่า $\breve{\phi} : D \times C \rightarrow \{T, F\}$ ด้วยฟังก์ชั่น $\phi : D \times C \rightarrow \{T, F\}$ ซึ่งจะเรียกว่า ตัวจำแนกหมวดหมู่เอกสาร (Classifier) จนกระทั่ง $\breve{\phi}$ และ ϕ มีค่าใกล้เคียงกันมากที่สุดเท่าที่จะเป็นไปได้

การจำแนกหมวดหมู่เอกสารสามารถแบ่งประเภทได้หลายลักษณะขึ้นอยู่กับแง่มุมในการพิจารณา Sebastiani [1] แบ่งการจำแนกหมวดหมู่เอกสารไว้ ดังนี้

2.1.1 การจำแนกแบบหมวดหมู่เดียว (Single-Label Text Categorization) และการจำแนกแบบหลายหมวดหมู่ (Multi-label Text Categorization)

2.1.1.1 การจำแนกแบบหมวดหมู่เดียว เป็นการกำหนดหมวดหมู่ให้กับเอกสาร $d_j \in D$ เพียงเอกสารละหนึ่งหมวดหมู่เท่านั้น

2.1.1.2 การจำแนกแบบหลายหมวดหมู่ เป็นการกำหนดจำนวนหมวดหมู่ให้กับเอกสาร $d_j \in D$ ตั้งแต่ 0 ไปจนถึง $|C|$ หมวดหมู่

2.1.2 การจำแนกโดยใช้หมวดหมู่เป็นตัวหลัก (Category-Pivoted Categorization, CPC) และการจำแนกโดยใช้เอกสารเป็นตัวหลัก (Document-Pivoted Categorization, DPC)

2.1.2.1 การจำแนกโดยใช้หมวดหมู่เป็นตัวหลัก เป็นการพิจารณาหมวดหมู่ $c_i \in C$ แต่ละหมวดหมู่เพื่อหาเอกสาร $d_j \in D$ ทั้งหมดที่ควรอยู่ภายใต้หมวดหมู่ c_i

2.1.2.2 การจำแนกโดยใช้เอกสารเป็นตัวหลัก เป็นการพิจารณาเอกสาร $d_j \in D$ เพื่อหาหมวดหมู่ $c_i \in C$ ทั้งหมดให้กับเอกสาร d_j

2.1.3 การจำแนกหมวดหมู่แบบบ่งชัด (Hard Categorization) และการจำแนกหมวดหมู่แบบจัดลำดับ (Ranking Categorization)

2.1.3.1 การจำแนกหมวดหมู่แบบบ่งชัด เป็นการจำแนกหมวดหมู่โดยตัดสินใจเลือกค่าผลลัพธ์ของฟังก์ชั่นเป้าหมาย $\breve{\phi} : D \times C \rightarrow \{T, F\}$ ว่าจะให้มีค่าเป็น T หรือ F ซึ่งเป็นการจำแนกหมวดหมู่แบบอัตโนมัติ

2.1.3.2 การจำแนกหมวดหมู่แบบจัดลำดับ เป็นการจำแนกหมวดหมู่โดยจัดลำดับค่าผลลัพธ์ของฟังก์ชั่นเป้าหมาย $\breve{\phi} : D \times C \rightarrow [0,1]$ ซึ่งเป็นการจำแนกหมวดหมู่แบบกึ่งอัตโนมัติ

2.2 การเรียนรู้ด้วยคอมพิวเตอร์

การเรียนรู้ด้วยคอมพิวเตอร์เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligent) ที่เกี่ยวข้องกับการออกแบบและพัฒนาอัลกอริทึมและวิธีการที่จะทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ โดยทั่วไปวิธีการเรียนรู้มีอยู่ 2 ประเภท ได้แก่ การเรียนรู้เชิงอุปนัย (Inductive Learning) และการเรียนรู้เชิงอนุมาน (Deductive Learning)

การเรียนรู้ด้วยคอมพิวเตอร์เชิงอุปนัย เป็นการค้นหากฎ ลักษณะแบบแผน หรือข้อสรุปต่างๆ จากการสังเกตกลุ่มข้อมูลขนาดใหญ่ ส่วนการเรียนรู้เชิงอนุมาน เป็นการหาข้อสรุปจากหลักฐานหรือข้อเท็จจริงที่มีอยู่ หลักสำคัญของการวิจัยทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ คือ การสกัดเอาความรู้หรือสารสนเทศจากข้อมูลโดยอัตโนมัติด้วยวิธีการคำนวณหรือวิธีการทางสถิติ ดังนั้นการเรียนรู้ด้วยคอมพิวเตอร์นั้นจึงมีความสัมพันธ์อย่างใกล้ชิดกับการทำเหมืองข้อมูล (Data Mining) และสถิติ

อัลกอริทึมสำหรับการเรียนรู้ด้วยคอมพิวเตอร์ถูกจัดให้อยู่ในกลุ่มของวิทยาศาสตร์หรือวิธีการที่เกี่ยวข้องกับการแบ่งแยกประเภท (Taxonomy) ซึ่งมีที่มาจากผลลัพธ์ที่ได้จากอัลกอริทึม ประเภทของอัลกอริทึมอาจแบ่งได้ ดังต่อไปนี้

2.2.1 การเรียนรู้โดยอาศัยตัวอย่าง (Supervised Learning)

เป็นการเรียนรู้โดยใช้อัลกอริทึมเพื่อสร้างฟังก์ชั่นที่จะทำข้อมูลเข้าให้เป็นผลลัพธ์ที่ต้องการ การจำแนกประเภทเป็นรูปแบบหนึ่งของการเรียนรู้โดยอาศัยตัวอย่าง ตัวเรียนรู้จะต้องศึกษาพฤติกรรมของฟังก์ชั่นที่จะทำการกำหนดเวกเตอร์ $[X_1, X_2, \ldots, X_N]$ ให้อยู่ภายใต้ประเภทใดประเภทหนึ่งจากหลายประเภทโดยสังเกตจากตัวอย่างข้อมูลเข้าและตัวอย่างผลลัพธ์ของฟังก์ชั่น

2.2.2 การเรียนรู้โดยไม่อาศัยตัวอย่าง (Unsupervised Learning)

เป็นการเรียนรู้โดยการจำลองแบบของชุดข้อมูลเข้าจากลักษณะเฉพาะของข้อมูล โดยไม่ต้องใช้ตัวอย่างผลลัพธ์ที่ถูกจัดประเภทไว้แล้ว

2.2.3 การเรียนรู้กึ่งอาศัยตัวอย่าง (Semi-supervised Learning)

เป็นการเรียนรู้ที่อาศัยทั้งตัวอย่างที่ยังไม่ได้จัดประเภทและตัวอย่างที่ถูกจัดประเภทไว้แล้ว เพื่อสร้างฟังก์ชั่นหรือตัวจำแนกประเภทที่เหมาะสม

การวิเคราะห์การคำนวณและประสิทธิภาพของอัลกอริทึมที่ใช้สำหรับการเรียนรู้ด้วยคอมพิวเตอร์เป็นแขนงหนึ่งของทฤษฎีทางวิทยาการคอมพิวเตอร์ที่รู้จักกันในชื่อ ทฤษฎีการเรียนรู้เกี่ยวกับการคำนวณ (Computational Learning Theory)

2.3 การจำแนกหมวดหมู่เอกสารด้วยวิธีการเรียนรู้ด้วยคอมพิวเตอร์

การจำแนกหมวดหมู่เอกสารจะใช้วิธีการเรียนรู้โดยอาศัยตัวอย่าง ส่วนการจัดกลุ่มเอกสารจะใช้วิธีการเรียนรู้โดยไม่อาศัยตัวอย่าง

เอกสารที่ใช้ในกระบวนการเรียนรู้นั้นจะถูกเรียกว่า คลังเอกสารเริ่มต้น (Initial Corpus) $\Omega = \{d_1, \ldots, d_{|\Omega|}\} \subset D$ ซึ่งเป็นเอกสารที่ถูกจัดหมวดหมู่ไว้แล้วภายใต้ $C = \{c_1, ..., c_{|C|}\}$ โดยจะอยู่ในรูปแบบของเมตริกซ์การตัดสินใจ (Decision Matrix) ซึ่งมีลักษณะดังในภาพที่ 2-3

	Training set			**Test set**		
	d_1	...	$d_{\|TV\|}$	$d_{\|TR\|+1}$	...	$d_{\|\Omega\|}$
c_1	$a_{1,1}$	...	$a_{1,\|TV\|}$	$a_{1,\|TR\|+1}$	...	$a_{1,\|\Omega\|}$
...	...	...	...	...	...	...
$c_{\|C\|}$	$a_{\|C\|,1}$	...	$a_{\|C\|,\|TV\|}$	$a_{\|C\|,\|TR\|+1}$	...	$a_{\|C\|,\|\Omega\|}$

ภาพที่ 2-3 เมตริกซ์การตัดสินใจ

ค่าของ $a_{i,j}$ จะมีค่าเป็น 1 เมื่อเอกสาร d_j อยู่ในหมวดหมู่ c_i และจะมีค่าเป็น 0 เมื่อเอกสาร d_j ไม่อยู่ในหมวดหมู่ c_i

เอกสาร d_j ที่มีค่า $a_{i,j}$ เท่ากับ 1 จะเรียกว่าเป็นตัวอย่างด้านบวก (Positive Example) ของหมวดหมู่ c_i และเอกสาร d_j ที่มีค่า $a_{i,j}$ เท่ากับ 0 จะเรียกว่าเป็นตัวอย่างด้านลบ (Negative Example) ของหมวดหมู่ c_i

ในขั้นตอนการสร้างตัวจำแนกหมวดหมู่ คลังเอกสารเริ่มต้นจะถูกแบ่งออกเป็น 2 ชุด โดยไม่จำเป็นต้องมีขนาดเท่ากัน ได้แก่ ชุดฝึกฝน (Training Set) และชุดทดสอบ (Test Set)

ชุดฝึกฝน $TV = \langle d_1,...,d_{|TV|} \rangle$ จะถูกใช้เป็นตัวอย่างเอกสารเพื่อทำการเรียนรู้โดยการสังเกตลักษณะเฉพาะของเอกสารในแต่ละหมวดหมู่ และนำข้อสรุปจากการสังเกตมาสร้างตัวจำแนกหมวดหมู่

ชุดทดสอบ $Te = \langle d_{|TV|+1},...,d_{|\Omega|} \rangle$ จะถูกใช้เพื่อทดสอบประสิทธิภาพของตัวจำแนกหมวดหมู่ โดยการป้อนเอกสารในชุดนี้ให้กับตัวจำแนกหมวดหมู่ที่สร้างขึ้น และนำผลลัพธ์มาเปรียบเทียบกับหมวดหมู่ของเอกสารที่จัดไว้ก่อนแล้ว การวัดประสิทธิภาพจะขึ้นอยู่กับว่าผลลัพธ์ที่ได้จากตัวจำแนกหมวดหมู่จะตรงกับการจัดหมวดหมู่ที่ทำไว้ก่อนแล้วมากเพียงไร

ชุดทดสอบ Te จะไม่ถูกใช้ในการกระบวนการสร้างตัวจำแนกหมวดหมู่ เพื่อความถูกต้องในการวัดผลการทดลองเชิงวิทยาศาสตร์ แต่ในกรณีที่ต้องการปรับค่าพารามิเตอร์ภายในตัวจำแนกหมวดหมู่เพื่อทำให้ตัวจำแนกหมวดหมู่ทำงานได้อย่างมีประสิทธิภาพมากที่สุด โดยต้องทำการทดสอบตัวจำแนกหมวดหมู่เพื่อหาค่าของพารามิเตอร์ที่เหมาะสม ดังนั้นเอกสารในชุดฝึกฝน $TV = \langle d_1,...,d_{|TV|} \rangle$ จะถูกแบ่งย่อยเป็น 2 ส่วน คือ ชุดฝึกฝน $Tr = \langle d_1,...,d_{|Tr|} \rangle$ และชุดทดสอบความเหมาะสม (Validation Set) $Va = \langle d_{|Tr|+1},...,d_{|TV|} \rangle$ โดยชุดทดสอบความเหมาะสม Va จะไม่ถูกใช้ในกระบวนการสร้างตัวจำแนกหมวดหมู่ด้วยเช่นกัน

2.4 การกำจัดคำหยุด

การกำจัดคำหยุด (Stop-Word List Removal) เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึง คำที่ใช้กันโดยทั่วไป ไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ประเภทของคำที่จัดว่าเป็นคำหยุดในภาษาไทย [23] มีดังต่อไปนี้

2.4.1 คำบุพบท

เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน มักใช้นำหน้าคำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อแสดงความสัมพันธ์ของคำหรือกลุ่มคำที่อยู่หลังคำบุพบทว่ามี

ความสัมพันธ์กับคำหรือกลุ่มคำที่อยู่หน้าคำบุพบทอย่างไร ตัวอย่างของคำบุพบท ได้แก่ ของ ใน แก่ แด่ ต่อ ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ช้าแต่ ทาง สู่ แค่ ฯลฯ

2.4.2 คำสันธาน

เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยค ข้อความกับข้อความ เพื่อทำให้ประโยคนั้นกระชับ และสละสลวย โดยมักจะใช้เชื่อมใจความที่คล้อยตามกัน ใจความที่ขัดแย้งกัน ใจความที่เป็นเหตุเป็นผลกัน หรือใจความที่ให้เลือกอย่างใดอย่างหนึ่ง ตัวอย่างของคำสันธาน ได้แก่ เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น หรือไม่ก็ ฯลฯ

2.4.3 คำสรรพนาม

เป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เพื่อไม่ต้องกล่าวคำนามนั้นซ้ำอีก ตัวอย่างของคำสรรพนาม ได้แก่ ฉัน เรา เขา ดิฉัน กระผม กู คุณ ท่าน เธอ ใต้เท้า มัน สิ่ง ใคร บ้าง ต่างก็ ตัวนั้น อันโน้น อะไร ไหน ใด อย่างนี้ ฯลฯ

2.4.4 คำวิเศษณ์

เป็นคำที่ใช้ขยายคำอื่น เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อให้มีความหมายชัดเจนและมีรายละเอียดมากยิ่งขึ้น ตัวอย่างของคำวิเศษณ์ ได้แก่ มาก น้อย ใหญ่ เล็ก มหึมา มโหฬาร อ้วน โต สูง ไพเราะ เยอะ หลาย สวย หอม นุ่ม เผ็ด เช้า ค่ำ นี่ นั่น เอง ทั้งนี้ ค่ะ ครับ ขอรับ จ้า จ๊ะ วะ ไม่ หามิได้ บ่ ทำไม อย่างไร หมด อดีต ปัจจุบัน โบราณ หน้า ฯลฯ

2.4.5 คำอุทาน

เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด รวมทั้งเป็นคำที่ใช้เสริมคำพูด ตัวอย่างของคำอุทานได้แก่ เอ๊ะ อ๊ะ อ๋อ เฮ้อ ว้าย โธ่ โถ อนิจจา สิ หนอย ซะ นะ น่า แหม ตายล่ะ คุณพระช่วย ชิชะ อุวะ ไม่น่าเลย โอ้โห ฯลฯ

คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล

2.5 การหารากศัพท์

รากศัพท์ (Stem) คือ รูปเดิมของคำที่ยังไม่ได้เติมคำอุปสรรค (Prefixes) คำปัจจัย (Suffixes) หรือยังไม่ได้เปลี่ยนรูปไป ตัวอย่างรากศัพท์ของคำแสดงในภาพที่ 2-4 และภาพที่ 2-5

คำศัพท์	รากศัพท์
organization	organize
organisational	organize
organizational	organize
organized	organize
organize	organize
organizer	organize
organizing	organize

ภาพที่ 2-4 ตัวอย่างรากศัพท์คำภาษาอังกฤษ

คำศัพท์	รากศัพท์
พิจารณา	พิจาร
พิจารณ์	พิจาร
การพิจารณา	พิจาร
พิจาร	พิจาร

ภาพที่ 2-5 ตัวอย่างรากศัพท์คำภาษาไทย

การหารากศัพท์ (Stemming) จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่

การหารากศัพท์ของคำภาษาอังกฤษมีขั้นตอนวิธีที่แน่นอนซึ่งสามารถเขียนเป็นอัลกอริทึมในการสกัดคำอุปสรรคและคำปัจจัยได้โดยไม่ต้องเทียบกับรายการคำศัพท์หรือคลังคำ เนื่องมาจากไวยากรณ์ของภาษาอังกฤษมีกฎเกณฑ์ที่แน่นอน ไม่ค่อยมีข้อยกเว้นมากนัก จึงมีความซับซ้อนน้อย ดังนั้นจึงมีผู้สร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้หลายแบบ เช่น Porter Algorithm Lancaster Algorithm เป็นต้น สำหรับงานวิจัยนี้ใช้วิธีการหารากศัพท์ด้วย Porter Algorithm

การหารากศัพท์ของคำภาษาไทยนั้น เท่าที่ศึกษามายังไม่พบว่ามีผู้คิดค้นสร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้ เนื่องมาจากไวยากรณ์ภาษาไทยมีความซับซ้อนมาก และมีข้อยกเว้นหลายประการ การหารากศัพท์ของคำภาษาไทยจึงอาจใช้วิธีการรวบรวมคำศัพท์ที่มีความหมาย

คล้ายกัน หรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์

2.6 การทำดัชนี

คอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ ลักษณะของตัวแทนเอกสารขึ้นอยู่กับสิ่งที่ต้องการพิจารณาว่า ต้องการพิจารณาเฉพาะความหมายของคำ หรือต้องการพิจารณาความหมายตามกฎของภาษา ซึ่งจะสนใจตำแหน่งของคำที่อยู่ในประโยค การจำแนกหมวดหมู่ด้วยวิธีการทางด้านการเรียนรู้ด้วยคอมพิวเตอร์นิยมใช้ลักษณะของตัวแทนเอกสารที่สนใจความหมายของคำ โดยไม่สนใจตำแหน่งของคำ ซึ่งตัวแทนเอกสารมักจะอยู่ในรูปของเวกเตอร์ของน้ำหนักคำ $\vec{d}_j = \langle w_{1j},..., w_{|T|j} \rangle$ โดยที่ $|T|$ คือ เซตของคุณลักษณะ (Feature) ซึ่งคุณลักษณะอาจหมายถึง คำเดี่ยว (Single Word) รากศัพท์ (Stems) วลี (Phrase) ชุดลำดับคำ (N-Gram) หรือสิ่งอื่น แต่โดยมากจะใช้ในรูปแบบของคำ ค่าน้ำหนักของคำจะมีค่าอยู่ระหว่าง 0 และ 1 ซึ่งจะใช้ค่าที่เป็นไบนารีหรือไม่เป็นไบนารีก็ได้ ขึ้นอยู่กับวิธีการคำนวณค่าน้ำหนัก วัลลภ [5] ได้รวบรวมวิธีการคำนวณค่าน้ำหนักไว้หลายวิธี ดังนี้

2.6.1 Boolean Weighting

เป็นการคำนวณค่าน้ำหนักจากการปรากฏของคำในเอกสาร ถ้ามีคำ t_k ปรากฏอยู่ในเอกสารตั้งแต่ 1 ครั้งขึ้นไป จะให้ค่าน้ำหนักเป็น 1 ถ้าคำ t_k ไม่ปรากฏอยู่ในเอกสาร จะให้ค่าน้ำหนักเป็น 0 ค่าน้ำหนักนี้เรียกอีกอย่างว่า ค่าคุณลักษณะความจริง (Boolean Feature) ซึ่งมีค่าเป็นไบนารี

$$w_{kj} = \begin{cases} 1 & if\ \#(t_k, d_j) > 0 \\ 0 & if\ \#(t_k, d_j) \le 0 \end{cases} \qquad (2\text{-}1)$$

โดยที่ w_{kj} คือ น้ำหนักของคำ t_k ในเอกสาร d_j และ $\#(t_k, d_j)$ คือ ความถี่ของคำ t_k ที่ปรากฏในเอกสาร d_j

2.6.2 Word Frequency Weighting

เป็นการคำนวณค่าน้ำหนักจากความถี่ของการปรากฏของคำ t_k ในเอกสาร d_j โดยตรง

$$w_{kj} = \#(t_k, d_j) \tag{2-2}$$

2.6.3 TFIDF Weighting

เป็นการคำนวณค่าน้ำหนักจากความถี่ของการปรากฏของคำ t_k ในเอกสาร d_j และพิจารณาความถี่ของคำ t_k ที่ปรากฏในเอกสารอื่นใน D ร่วมด้วย โดยมีแนวคิดว่า คำที่ปรากฏในเอกสารน้อยฉบับ จะมีค่าน้ำหนักสูง ส่วนคำที่ปรากฏในเอกสารหลายฉบับ จะมีค่าน้ำหนักต่ำ เนื่องจากเป็นคำที่ไม่แสดงถึงคุณลักษณะเฉพาะของเอกสารนั้น

$$w_{kj} = \#(t_k, d_j) \times \log\left(\frac{|Tr|}{\#Tr(t_k)}\right) \tag{2-3}$$

โดยที่ $|Tr|$ คือ จำนวนเอกสารทั้งหมดใน Tr และ $\#Tr(t_k)$ คือ จำนวนเอกสารใน Tr ที่มีคำ t_k ปรากฏอยู่

2.6.4 TFC Weighting

เป็นการคำนวณค่าน้ำหนักคล้ายกับ TFIDF แต่จะมีการพิจารณาจำนวนคำที่แตกต่างกันในแต่ละเอกสารร่วมด้วย ดังนั้นเพื่อปรับบรรทัดฐานให้ทุกเอกสารมีจำนวนคำเท่ากัน จึงปรับค่าน้ำหนักด้วย Cosine Normalization

$$w_{kj} = \frac{\#(t_k, d_j) \times \log\left(\frac{|Tr|}{\#Tr(t_k)}\right)}{\sqrt{\sum_{s=1}^{|T|}\left(\left(\#(t_s, d_j) \times \log\left(\frac{|Tr|}{\#Tr(t_s)}\right)\right)\right)^2}} \tag{2-4}$$

2.6.5 LTC Weighting

เป็นการคำนวณค่าน้ำหนักคล้ายกับ TFC แต่จะมีการพิจารณาความถี่ที่มีค่าแตกต่างกันมาก โดยการเพิ่มฟังก์ชั่น Log เข้ามาเพื่อปรับลดความแตกต่าง

$$w_{kj} = \frac{(1 + \log(\#(t_k, d_j)) \times \log\left(\frac{|Tr|}{\#Tr(t_k)}\right)}{\sqrt{\sum_{s=1}^{|T|}\left(\left((1 + \log(\#(t_s, d_j)) \times \log\left(\frac{|Tr|}{\#Tr(t_s)}\right)\right)\right)^2}} \tag{2-5}$$

2.6.6 Entropy Weighting

เป็นการคำนวณค่าความไม่แน่นอนของคำ โดยค่าความไม่แน่นอนของคำที่ปรากฏอยู่ในเอกสารทุกฉบับจะมีค่าเท่ากับ -1 และค่าความไม่แน่นอนของคำที่ปรากฏอยู่ในเอกสารเพียงฉบับเดียวจะมีค่าเท่ากับ 0

$$\text{ค่าความไม่แน่นอน (Entropy)} = \frac{1}{\log(|Tr|)} \times \sum_{s=1}^{|T|} \frac{\#(t_k, d_s)}{\#Tr(t_k)} \times \log\left(\frac{\#(t_k, d_s)}{\#Tr(t_k)}\right)$$

$$w_{kj} = \log(\#(t_k, d_j) + 1.0 \times (1 + Entropy) \quad (2\text{-}6)$$

นอกจากนี้ Sebastiani [1] ยังได้แสดงวิธีการคำนวณค่าความเป็นหลักการทั่วไป (Generality) ของหมวดหมู่ c_i ในคลังเอกสาร Ω โดยคิดเป็นอัตราร้อยละของเอกสาร d_j ที่อยู่ภายใต้หมวดหมู่ c_i ไว้ดังนี้

$$g_\Omega(c_i) = \frac{\left|\{d_j \in \Omega \mid a_{i,j} = 1\}\right|}{|\Omega|} \quad (2\text{-}7)$$

2.7 การลดขนาดของเอกสาร

เอกสารที่มีขนาดใหญ่ หมายถึง เอกสารที่มีจำนวนคุณลักษณะมาก ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี และเอกสารที่มีจำนวนคุณลักษณะมากอาจก่อให้เกิดปัญหา Overfitting ซึ่งเป็นปรากฏการณ์ที่ตัวจำแนกหมวดหมู่ค้นพบลักษณะโดยบังเอิญของเอกสารตัวอย่าง แทนที่จะค้นพบลักษณะพื้นฐานที่จำเป็นของเอกสารตัวอย่าง ทำให้ตัวจำแนกหมวดหมู่ทำงานผิดพลาด การลดขนาดเอกสารจึงเป็นขั้นตอนหนึ่งที่ต้องทำก่อนการสร้างตัวจำแนกเอกสาร แต่การลดขนาดของเอกสารต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกหมวดหมู่ออกไป

ขอบเขตของการลดขนาดเอกสารมี 2 ลักษณะ ได้แก่ การลดขนาดโดยรวม (Global) ซึ่งจะใช้เซตของคุณลักษณะเดียวกันสำหรับทุกหมวดหมู่ และการลดขนาดเฉพาะหมวดหมู่ (Local) ซึ่งจะใช้เซตของคุณลักษณะที่แตกต่างกันในแต่ละหมวดหมู่

วิธีการลดขนาดของเอกสารมี 2 วิธี ได้แก่ การเลือกคุณลักษณะ (Term Selection) และการสกัดคุณลักษณะ (Term Extraction)

2.7.1 การเลือกคุณลักษณะ เป็นการพิจารณาเลือกเซตของคุณลักษณะใหม่ T' จากเซตของคุณลักษณะเดิม T โดยที่จำนวนของคุณลักษณะ $|T'| << |T|$

วิธีการเลือกคุณลักษณะมี 2 วิธี คือ Wrapper และ Filtering

2.7.1.1. Wrapper จะใช้วิธีการสร้างเซต T' โดยการเพิ่มหรือลดคุณลักษณะจากเซต T แล้วสร้างตัวจำแนกหมวดหมู่จากเซต T' นั้น เพื่อทำการทดสอบปรับค่าพารามิเตอร์ภายในตัวจำแนกเอกสาร เซต T' ที่ให้ผลลัพธ์ที่มีประสิทธิภาพมากที่สุดจะถูกเลือกใช้

2.7.1.2. Filtering จะใช้วิธีการเลือกคุณลักษณะที่มีค่าความสำคัญมากที่สุด วิธีการวัดค่าความสำคัญมีหลายวิธีการ เช่น Chi-Square, Information Gain, Mutual Information ฯลฯ วิธีการที่ง่ายและมีประสิทธิภาพวิธีหนึ่งที่นิยมใช้คือ การวัดความถี่ของคุณลักษณะในเอกสาร $\#Tr(t_k)$ โดยคุณลักษณะที่มีความถี่สูงจะถูกเลือก

2.7.2. การสกัดคุณลักษณะ เป็นการสร้างหรือสังเคราะห์เซตของคุณลักษณะใหม่ T' ขึ้นจากเซตของคุณลักษณะเดิม T โดยที่จำนวนของคุณลักษณะ $|T'| << |T|$

วิธีการสกัดคุณลักษณะมี 2 วิธี คือ Term Clustering และ Latent Semantic Indexing

2.7.2.1. Term Clustering จะใช้วิธีสร้างกลุ่มของคำที่มีความหมายเกี่ยวข้องกันแล้วใช้กลุ่มของคำที่สร้างขึ้นแทนคำเหล่านั้น

2.7.2.2. Latent Semantic Indexing จะใช้วิธีการหารูปแบบของการปรากฏร่วมกันของคุณลักษณะโดยการคำนวณทางสถิติ

2.8 วิธีการเรียนรู้สำหรับการสร้างตัวจำแนกหมวดหมู่เอกสาร

ในการสร้างตัวจำแนกหมวดหมู่แบบจัดลำดับ โดยทั่วไปจะกำหนดนิยามของตัวจำแนกหมวดหมู่ให้อยู่ในรูปของฟังก์ชั่น $CSV_i : D \to [0,1]$ สำหรับแต่ละหมวดหมู่ $c_i \in C$ โดยให้ $d_j \in D$ เป็นข้อมูลเข้า และส่งคืนผลลัพธ์เป็นค่าที่แสดงสถานะการจำแนกหมวดหมู่ ซึ่งมีค่าอยู่ระหว่าง 0 และ 1 เอกสาร d_j จะถูกจัดลำดับตามค่าที่ได้จากฟังก์ชั่น $CSV_i(d_j)$

สำหรับตัวจำแนกหมวดหมู่แบบบ่งชัด จะกำหนดนิยามให้อยู่ในรูปของฟังก์ชั่นได้ 2 รูปแบบ คือ $CSV_i : D \to \{T, F\}$ และ $CSV_i : D \to [0,1]$ ที่มาพร้อมกับนิยามของค่าแบ่ง (Threshold) τ_i โดยที่ถ้า $CSV_i(d_j) \geq \tau_i$ หมายถึงค่า T และถ้า $CSV_i(d_j) < \tau_i$ หมายถึงค่า F

2.8.1 การหาค่าแบ่ง

Sebastiani [1] กล่าวถึงวิธีการหาค่าแบ่งซึ่งมีอยู่หลายแนวทาง ได้แก่

2.8.1.1 Probability Thresholding เป็นการหาค่าแบ่งโดยการคำนวณหาความน่าจะเป็นที่เอกสาร d_j จะอยู่ในหมวดหมู่ c_i

2.8.1.2 CSV Thresholding เป็นการหาค่าแบ่งโดยการกำหนดค่าแบ่ง τ_i หลายค่าแตกต่างกัน เพื่อทดสอบกับเอกสารชุดทดสอบความเหมาะสม Va และเลือกค่าที่ให้ประสิทธิภาพสูงสุด ซึ่งโดยทั่วไปจะได้ค่า τ_i ที่แตกต่างกันสำหรับแต่ละหมวดหมู่ c_i

2.8.1.3 Proportional Thresholding เป็นหาค่าแบ่งโดยการทดสอบทำนองเดียวกับ CSV Thresholding โดยเลือกค่าแบ่ง τ_i ที่ทำให้ $g_{Va}(c_i)$ ใกล้เคียงกับ $g_{Tr}(c_i)$ มากที่สุด ร่วมกับใช้หลักการที่ว่า ควรใช้จำนวนเอกสารในอัตราร้อยละที่เท่ากันทั้งชุดฝึกฝนและชุดทดสอบ

2.8.1.4 Fixed Thresholding เป็นการกำหนดค่าคงที่ k ให้เป็นจำนวนหมวดหมู่ที่จะกำหนดให้เอกสาร d_j ทุกตัว

2.8.2 วิธีการเรียนรู้

วิธีการเรียนรู้ที่ใช้ในการสร้างตัวจำแนกหมวดหมู่เอกสาร (เรียกในรูปแบบของฟังก์ชั่น คือ $CSV_i(d_j)$) นั้นมีอยู่หลายวิธีการด้วยกัน ในที่นี้จะกล่าวถึงแต่ละวิธีการโดยสรุปดังนี้

2.8.2.1 Decision Tree

Tree จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่ การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าความจริงที่ใช้จะมาจากการคำนวณหาค่า Entropy และค่า Information Gain คุณลักษณะใดที่มีค่า Information Gain มากที่สุดจะถูกเลือกเป็นโหนดลูก หากมีค่า Entropy เป็น 0 จะได้เป็นโหนดล่างสุด วิธีการนี้มีข้อดี คือ ง่ายต่อการสร้างตัวจำแนกหมวดหมู่ ส่วนข้อเสีย คือ รองรับจำนวนคุณลักษณะได้น้อย

2.8.2.2 Naïve Bayes

วิธีการนี้จะมอง $CSV_i(d_j)$ ในแง่ของความน่าจะเป็น $P(c_i \mid \vec{d}_j)$ ซึ่งเป็นความน่าจะเป็นที่เอกสาร d_j จะอยู่ในหมวดหมู่ c_i ซึ่งคำนวณตามทฤษฎีของ Bayes ดังนี้

$$P(c_i \mid \vec{d}_j) = \frac{P(c_i)P(\vec{d}_j \mid c_i)}{P(\vec{d}_j)} \tag{2-8}$$

โดย $P(\vec{d}_j)$ คือ ความน่าจะเป็นที่จะหยิบเอกสารที่ใช้เวกเตอร์ $\vec{d}_j$ เป็นตัวแทนได้โดยการสุ่ม และ $P(c_i)$ คือ ความน่าจะเป็นที่จะหยิบเอกสารที่อยู่ในหมวดหมู่ c_i ได้โดยการสุ่ม สำหรับ $P(\vec{d}_j \mid c_i)$ คำนวณได้จาก

$$P(\vec{d}_j \mid c_i) = \prod_{k=1}^{|T|} P(w_{kj} \mid c_i) \tag{2-9}$$

ตัวจำแนกหมวดหมู่ Naïve Bayes ให้ฟังก์ชั่นเป้าหมาย $f : X \to v$ โดยมีคำตอบที่เป็นไปได้ทั้งหมด v_k ค่า และ $X = \{a_1, \ldots, a_T\}$ หลักการของวิธีนี้ คือ คำนวณหาค่าความน่าจะเป็นที่มากที่สุดของ f จากสมมติฐานของ Naïve Bayes จาก $P(a_1, \ldots, a_T \mid v_k) = \prod_t P(a_t \mid v_k)$ หมวดหมู่ที่เป็นผลลัพธ์จะมาจากการคำนวณ $\arg\max P(v_k)\prod_t P(a_t \mid v_k)$ ข้อดีของวิธีการนี้ คือ

ง่ายต่อการสร้างตัวจำแนกหมวดหมู่ รวมทั้งเรียนรู้และจำแนกได้เร็ว แต่มีข้อด้อย คือ ความถูกต้องของผลลัพธ์น้อยกว่าวิธีการอื่น

2.8.2.3 Neural Network

โครงข่ายประสาทเทียมจะมีโหนดนำเข้าสำหรับรับเอกสารที่จะนำเข้ามาเรียนรู้ และมีโหนดส่งออกสำหรับทำนายว่าเอกสารนั้นจะอยู่ในหมวดหมู่ใด น้ำหนักจะอยู่บนขอบที่เชื่อมของหน่วย

การจัดหมวดหมู่เอกสารด้วยโครงข่ายประสาทเทียมนั้น ในขั้นตอนการฝึกฝนโหนดนำเข้า จะเก็บน้ำหนักของคำ และการะจายไปตามโครงข่าย โหนดส่งออกจะตัดสินว่าค่าที่ได้รับอยู่ในกลุ่มใด ถ้ามีการแบ่งกลุ่มผิด จะใช้วิธีการที่เรียกว่า แบคพรอพโพเกชั่น (Back Propagation) เพื่อปรับปรุงพารามิเตอร์ในโครงข่าย พร้อมปรับค่าน้ำหนักคำให้ถูกต้อง โดยวนซ้ำจนได้ค่าความผิดพลาดต่ำที่สุดที่ยอมรับได้

2.8.2.4 k-Nearest Neighbor

หลักการของวิธีการนี้ คือ จะทำการวัดความคล้ายกันของเอกสารทดสอบกับเอกสารที่ใช้ในการเรียนรู้ แล้วจัดลำดับเอกสารตามความคล้าย จำนวน k อันดับ แล้วกำหนดหมวดหมู่ของเอกสารทดสอบ ให้มีค่าเท่ากับหมวดหมู่ที่ปรากฏมากที่สุดของเอกสารจำนวน k ฉบับนั้น

วิธีการนี้มีข้อดี คือ ให้ผลลัพธ์ที่มีความถูกต้องสูง และเรียนรู้เร็ว รวมทั้งรองรับการทำงานกับข้อมูลจำนวนมาก ส่วนข้อเสีย คือ หากมีคุณลักษณะที่ไม่เกี่ยวข้องต่อการจำแนกหมวดหมู่ของเอกสารทดสอบมาก จะทำให้ความถูกต้องในการจำแนกหมวดหมู่ลดลง

2.8.2.5 k-Nearest Neighbor Feature Projection

หลักกการของวิธีการนี้ คล้ายกับวิธีการ k-Nearest Neighbor คือ จะทำการวัดความคล้ายกันของเอกสารทดสอบกับเอกสารที่ใช้ในการเรียนรู้ แล้วจัดลำดับเอกสารตามความคล้าย จำนวน k อันดับ แต่จะมองเอกสารในลักษณะของการฉาย (Projection) บนคุณลักษณะหรือคำของเอกสาร โดยพิจารณาแต่ละคุณลักษณะแยกจากกัน คุณลักษณะใดที่ไม่เกี่ยวข้องจะไม่นำมาพิจารณา การกำหนดหมวดหมู่ของเอกสารทดสอบ จะให้มีค่าเท่ากับหมวดหมู่ที่ปรากฏมากที่สุดของแต่ละคุณลักษณะของเอกสารจำนวน k ฉบับนั้น

ข้อดีของวิธีการนี้ คือ ให้ผลลัพธ์ที่มีความถูกต้องสูงกว่า และทำงานได้เร็วกว่า k-Nearest Neighbor และคุณลักษณะที่ไม่เกี่ยวข้องต่อการจำแนกหมวดหมู่ไม่ค่อยมีผลต่อความถูกต้องในการจำแนกหมวดหมู่ ส่วนข้อด้อย คือ หากกำหนดค่า k มากเกินไป จะทำให้ผลการจำแนกหมวดหมู่มีความถูกต้องน้อยลง

2.8.2.6 Rocchio

วิธีของร็อคชิโอจะสร้างตัวแยกเอกสาร โดยคิดน้ำหนักของเทอม $< w_1,\ldots,w_n >$ กับกลุ่ม c_i ด้วยสูตรการคำนวณ

$$w_{ki} = \left[\frac{B}{\left| \left\{ \vec{d}_j \mid ca_{ij} = 1 \right\} \right|} \times \sum_{\{\vec{d}_j | ca_{ij}=1\}} w_{kj} \right] - \left[\frac{Y}{\left| \left\{ \vec{d}_j \mid ca_{ij} = 0 \right\} \right|} \times \sum_{\{\vec{d}_j | ca_{ij}=0\}} w_{kj} \right] \quad (2\text{-}9)$$

โดยที่ w_{kj} เป็นน้ำหนักของเทอม t_k ซึ่งอยู่ในเอกสารฝึกฝน $\vec{d}_j$ ส่วน B และ Y เป็นค่าพารามิเตอร์ควบคุมความสัมพันธ์ของตัวอย่างที่อยู่ในกลุ่ม และไม่อยู่ในกลุ่ม

ในขั้นตอนการคัดแยกเอกสารจะนำเอาตัวแทนเอกสารไปเปรียบเทียบกับตัวแทนของกลุ่มฝึกฝน ถ้ามากกว่าค่าผ่านอ้างอิงก็สามารถจัดหมวดหมู่เอกสารได้

ข้อดีของวิธีนี้คือ ง่ายต่อการสร้างตัวจำแนกหมวดหมู่ สำหรับข้อเสียคือ ถ้าเอกสารในกลุ่มมีแนวโน้มไม่อยู่ในกลุ่ม จะทำให้การจำแนกหมวดหมู่ผิดพลาด

2.8.2.7 Support Vector Machine

หลักการของ SVM คือการสร้างสมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดย SVM จะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่า (Kernel Function) บน Feature Space

ข้อดีของวิธีการนี้คือ รองรับจำนวนคุณลักษณะได้มาก และมีความถูกต้องสูง ข้อเสียคือ ต้องเลือก Kernel Function ที่เหมาะสม

บทที่ 3
วิธีการดำเนินงานวิจัย

บทนี้อธิบายถึงลักษณะของอัลกอริทึมที่ใช้ในงานวิจัย และรายละเอียดของวิธีการดำเนินงานวิจัย พร้อมทั้งข้อมูลและการเตรียมข้อมูลเพื่อใช้ในการทดสอบ

3.1 ลักษณะของ Feature Projection Text Categorization

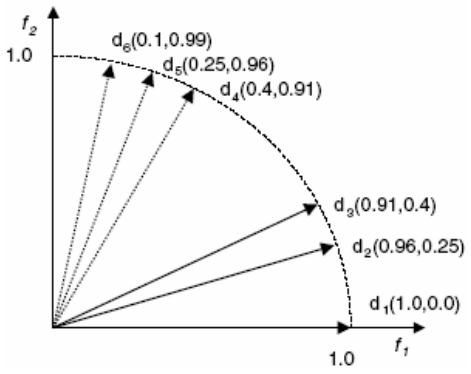
อัลกอริทึม Feature Projection Text Categorization (FPTC) นั้นเป็นอัลกอริทึมที่ถูกดัดแปลงมาจากอัลกอริทึม k-Nearest Neighbor (k-NN) และเป็นอัลกอริทึมแบบ non-incremental นั่นคือ จะทำกระบวนการเรียนรู้เอกสารทั้งหมดพร้อมกันเพียงครั้งเดียว ลักษณะสำคัญที่แตกต่างจาก k-Nearest Neighbor คือ เอกสารทั้งหมดจะถูกเก็บในลักษณะของภาพฉาย (Projections) บนแต่ละคุณลักษณะของเอกสาร เอกสารที่ไม่มีคุณลักษณะใดจะไม่ถูกเก็บบนคุณลักษณะนั้น ระยะห่างระหว่างเอกสารสองเอกสารก็จะถูกพิจารณาบนคุณลักษณะเดียว

ระยะห่างระหว่างเอกสาร d_i และเอกสาร d_j บนคุณลักษณะ t_m จะถูกคำนวณ ดังนี้

$$dist_m(t_m(i), t_m(j)) = \left| w(t_m, \vec{d}_i) - w(t_{(m}, \vec{d}_j) \right| \quad (3\text{-}1)$$

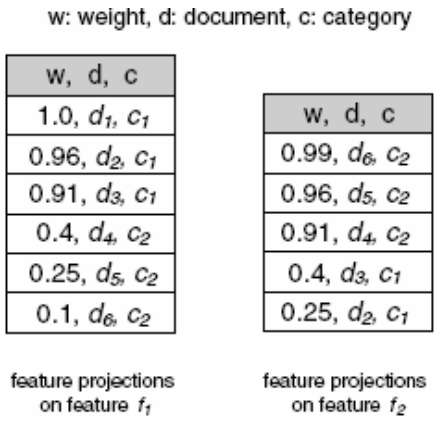
โดยที่ $t_m(i)$ คือ คุณลักษณะ t ลำดับที่ m ในเอกสาร d_i และ $w(t_m, \vec{d}_i)$ คือ ค่าน้ำหนักของคุณลักษณะ t ในเอกสาร d_i

การกำหนดหมวดหมู่ให้กับเอกสารจะเป็นไปตามคะแนนเสียง (Vote) ของเอกสารที่มีค่าน้ำหนักใกล้เคียงกันมากที่สุดจำนวน k ลำดับ บนแต่ละคุณลักษณะ ถ้ามีจำนวนคุณลักษณะ f ชุด วิธีการนี้จะให้ค่าโหวตจำนวน $f \times k$ ชุด ในขณะที่ k-Nearest Neighbor จะให้ค่า k ชุด



ภาพที่ 3-1 การแทนเอกสารในรูปแบบของเว็กเตอร์ (อ้างอิงมาจาก [9])

การแทนเอกสาร (Document Representation) จะอยู่ในรูปแบบของเว็กเตอร์ และการแทนคุณลักษณะ (Feature Representation) จะอยู่ในรูปแบบของภาพฉายของคุณลักษณะ ดังภาพที่ 3-1



ภาพที่ 3-2 การแทนคุณลักษณะในรูปแบบภาพฉายของคุณลักษณะ (อ้างอิงมาจาก [9])

การแทนคุณลักษณะในรูปแบบภาพฉายของคุณลักษณะจะเป็นการแยกเก็บเอกสารบนแต่ละคุณลักษณะ เอกสารที่ไม่มีคุณลักษณะใดจะไม่ถูกจัดเก็บบนคุณลักษณะนั้น ดังภาพที่ 3-1 และภาพที่ 3-2 จะเห็นว่าเอกสาร d_1 ไม่มีคุณลักษณะ f_2 จึงไม่ถูกจัดเก็บบนคุณลักษณะ f_2 ดังนั้นคุณลักษณะที่ไม่เกี่ยวข้อง (irrelevant) จะไม่ถูกพิจารณาในการจำแนกหมวดหมู่บนแต่ละคุณลักษณะ

การจำแนกหมวดหมู่จะแบ่งการทำงานโดยรวมออกเป็นสองขั้นตอน คือ ขั้นตอนการเรียนรู้และขั้นตอนการจำแนกหมวดหมู่ ในขั้นตอนการเรียนรู้ เอกสารจะถูกแบ่งเก็บตามคุณลักษณะ และจะคำนวณค่าน้ำหนักของแต่ละคุณลักษณะในเอกสารเก็บไว้ โดยค่าน้ำหนักของเอกสารจะคำนวณด้วยสมการที่ (2-5)

ในขั้นตอนการจำแนกหมวดหมู่ จะคำนวณค่าน้ำหนักของคุณลักษณะของเอกสารทดสอบแล้วนำมาคำนวณระยะห่างเทียบกับเอกสารที่ใช้เรียนรู้บนแต่ละคุณลักษณะ แล้วจึงจัดลำดับของเอกสารที่มีระยะห่างกับเอกสารทดสอบน้อยที่สุดจำนวน k ลำดับ ซึ่งแต่ละเอกสารจะให้คะแนนเสียงแก่หมวดหมู่ที่เอกสารนั้นอยู่ เอกสารละ 1 คะแนนเสียง จากนั้นจึงนับคะแนนเสียงของแต่ละหมวดหมู่โดยนับรวมกันทุกคุณลักษณะเพื่อหาว่าหมวดหมู่ใดมีความถี่หรือมีคะแนนเสียงมากที่สุดและทำนายหมวดหมู่นั้นให้กับเอกสารทดสอบ โดยรายละเอียดของอัลกอริทึมที่ใช้แสดงดังภาพที่ 3-3

อัลกอริทึมเริ่มจากการกำหนดค่าเริ่มต้นให้แก่ตัวแปรที่เก็บคะแนนเสียงของแต่ละหมวดหมู่ หลังจากนั้นทำการหาเอกสารที่มีค่าน้ำหนักใกล้เคียงกับเอกสารทดสอบมากที่สุดบนแต่ละคุณลักษณะ หลังจากนั้นคำนวณตามขั้นตอนที่อธิบายไว้ข้างต้น

```
Classify(t,k) :
/* t: test instance, k: number of neighbors */
Begin
        For each class c
                vote[c] = 0
        For each feature f
        /* put k nearest neighbors of test instance t on feature f into bag */
        bag = kbag(f,t,k)
        For each class c
        /* count(c,bag) will count c in bag */
        vote[c] = vote[c]+count(c,bag)
        prediction = UNDETERMINED
        For each class c
                If vote[c] > vote[prediction] then
                        prediction = c
        Return (prediction)
End.
```

ภาพที่ 3-3 อัลกอริทึม FPTC (อ้างอิงมาจากงานวิจัย [19])

จากการทดลองเบื้องต้นพบว่า การจำแนกหมวดหมู่เอกสารภาษาไทยที่ได้จากอัลกอริทึมดังภาพที่ 3-3 ยังให้ผลลัพธ์ที่มีความถูกต้องไม่มากนัก สาเหตุส่วนหนึ่งมาจากการที่สมาชิกทุกตัวในคุณลักษณะนั้นมีสิทธิ์ในการให้คะแนนเสียงเท่ากัน ไม่ว่าสมาชิกนั้นจะมีค่าความสำคัญมากหรือน้อย ผลการจัดหมวดหมู่ที่ได้จึงขึ้นอยู่กับคะแนนเสียงทั้งที่เกี่ยวข้องและไม่เกี่ยวข้องกับหมวดหมู่ที่ควรได้ เมื่อคะแนนเสียงจากสมาชิกที่อยู่ภายใต้หมวดหมู่ที่ไม่เกี่ยวข้องมากกว่า ผลการจัดหมวดหมู่จึงไม่ถูกต้อง

ดังนั้นเพื่อให้คะแนนเสียงของแต่ละเอกสารมีค่าความสำคัญที่แตกต่างกัน จึงนำหลักการตามทฤษฎีทางด้านการค้นคืนสารสนเทศ (Information Retrieval) ที่ว่า เอกสารที่มีค่า tfidf สูงกว่านั้นจะมีเป็นประโยชน์ต่อการจำแนกหมวดหมู่มากกว่าเอกสารที่มีค่า tfidf ต่ำกว่า มาประยุกต์เข้ากับอัลกอริทึมดังกล่าว โดยจะกำหนดให้เอกสารที่มีสิทธิ์ให้คะแนนเสียงนั้นต้องมีค่า tfidf สูงกว่าค่าเฉลี่ยของค่า tfidf ทั้งหมดบนคุณลักษณะนั้น การกำหนดความสำคัญในการให้คะแนนเสียง (Voting Score) มีวิธีการ ดังต่อไปนี้

$$vs(c_j, t_m) = w(t_m, \vec{d}) \times r(c_j, t_m) \tag{3-2}$$

โดยที่ $r(c_j, t_m)$ คือ อัตราส่วนการให้คะแนนเสียง (Voting Ratio) ซึ่งมีวิธีการคำนวณ ดังนี้

$$r(c_j, t_m) = \sum_{t_m(l) \in I_m} w(t_m, \vec{d}_l) \times y(c_j, t_m(l)) / \sum_{t_m(l) \in I_m} w(t_m, \vec{d}_l) \quad (3\text{-}3)$$

โดยที่ I_m คือ เอกสารที่มีสิทธิ์ในการให้คะแนนเสียง และ $y(c_j, t_m(l))$ เป็นฟังก์ชันที่จะให้ค่า 1 เมื่อหมวดหมู่ของเอกสาร l นั้นเท่ากับ c_j หรือให้ค่า 0 เมื่อหมวดหมู่ของเอกสาร l นั้นไม่เท่ากับ c_j สมการที่ (3-2) และ (3-3) อ้างอิงมาจาก [9]

โดยจะนำการกำหนดความสำคัญในการให้คะแนนเสียงมาประยุกต์ใช้ในอัลกอริทึม ดังมีรายละเอียดตามภาพที่ 3-4

```
Classify(t,k) :
/* t: test instance, k: number of neighbors */
Begin
        For each class c
                vote[c] = 0
        For each feature f
        /* put k nearest neighbors of test instance t on feature f into bag */
        Bag = kbag(f,t,k)
        For each class c
                vote[c] = vote[c] + vs(cj, tm)
        prediction = UNDETERMINED
        For each class c
                If vote[c] > vote[prediction] then
                        prediction = c
        Return (prediction)
End.
```

ภาพที่ 3-4 อัลกอริทึม FPTC ที่ประยุกต์ใช้คุณสมบัติของค่า tfidf

3.2 ข้อมูลที่ใช้ในการวิจัย

ข้อมูลที่นำมาใช้ในการทำวิจัยเป็นข้อมูลข่าวภาษาไทยจากเว็บไซต์หนังสือพิมพ์ออนไลน์ ได้แก่ หนังสือพิมพ์คมชัดลึก (www.komchadluek.net) หนังสือพิมพ์ผู้จัดการ (www.manager.co.th) หนังสือพิมพ์เดลินิวส์ (www.dailynews.co.th) หนังสือพิมพ์ไทยรัฐ (www.thairath.co.th) อยู่ในช่วงตั้งแต่เดือนมกราคม–เมษายน พ.ศ. 2550 โดยคัดเลือกเฉพาะข่าวที่มีความยาวไม่น้อยกว่า 3 บรรทัด และกำหนดหมวดหมู่ให้กับข่าวตามที่ทางเว็บไซต์หนังสือพิมพ์จัดแบ่งไว้ โดยนำมาประมาณหมวดหมู่ละ 150-200 เอกสาร รวมเป็นจำนวน 1,913 เอกสาร โดยจัดเก็บให้อยู่ในรูปแบบไฟล์ข้อความ แยกเก็บข่าวละหนึ่งไฟล์

3.3 หมวดหมู่ที่ใช้ในการวิจัย

หมวดหมู่ที่ใช้ในการวิจัย นำมาจากหมวดหมู่ที่เว็บไซต์หนังสือพิมพ์ที่กล่าวถึงในหัวข้อ 3.2 ที่ได้จัดแบ่งกลุ่มของข่าวไว้แล้ว โดยจัดเก็บข้อมูลหมวดหมู่ในรูปแบบไฟล์ข้อความ หมวดหมู่ที่คัดเลือกมามี 12 หมวดหมู่ ดังในตารางที่ 3-1

3.4 ระบบจำแนกหมวดหมู่ที่ใช้ในงานวิจัย

ระบบที่พัฒนาขึ้นเพื่อใช้ในงานวิจัยนี้พัฒนาขึ้นโดยใช้ภาษาจาวา ข้อมูลนำเข้าจะเป็นข่าวในรูปแบบไฟล์ข้อความ ผลลัพธ์ที่ได้จากโปรแกรมจะอยู่ในรูปแบบไฟล์ข้อความ ซึ่งมีรายละเอียดประกอบด้วยหมายเลขไฟล์ และหมวดหมู่ของไฟล์ที่ได้ ระบบที่พัฒนาขึ้นในงานวิจัยฉบับนี้มี เป็นตัวจำแนกที่ระบุหมวดหมู่เอกสารแบบบ่งชัด (Hard Categorization) และทำนายหมวดหมู่ให้แก่เอกสารฉบับละหนึ่งหมวดหมู่ (Single-Label Categorization) โดยจะใช้เอกสารเป็นหลักเพื่อหาหมวดหมู่ให้กับเอกสาร (Document-Pivoted Categorization) การเรียนรู้ของตัวจำแนกต้องอาศัยตัวอย่าง (Supervised Learning) นอกจากนี้การลดจำนวนคุณลักษณะของตัวจำแนกหมวดหมู่ในงานวิจัยนี้ ใช้วิธีการกรอง (Filtering) ด้วยค่าความถี่ของเอกสาร ลักษณะโดยรวมของโปรแกรมสรุปได้ดังในภาพที่ 3-5

ตารางที่ 3-1 หมวดหมู่ที่ใช้ในการวิจัย

หมายเลข	ชื่อหมวดหมู่
0	เกษตร
1	เศรษฐกิจ
2	ศาสนา
3	การเมือง
4	การศึกษา วัฒนธรรม
5	กีฬา
6	ตำรวจ อาชญากรรม
7	ท่องเที่ยว
8	บันเทิง
9	ผู้หญิง
10	วิทยาศาสตร์ เทคโนโลยี
11	ไลฟ์สไตล์

Feature	Type	System
Categorization	Single-Label	✓
	Multi-Label	
	Document-Pivoted	✓
	Category-Pivoted	
	Hard	✓
	Raking	
Learning	Supervised	✓
	Unsupervised	
Term Selection	Filtering	✓
	Wrapper	

ภาพที่ 3-5 ลักษณะโดยรวมของระบบ

3.5 อุปกรณ์ที่ใช้ในการวิจัย

อุปกรณ์ที่ใช้ในการวิจัย ประกอบด้วย

ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์ 1 ชุด ประกอบด้วย ซีพียู Pentium M 1.7 GHz หน่วยความจำหลัก 512 MB ฮาร์ดดิสก์ขนาด 60 GB

ซอฟต์แวร์

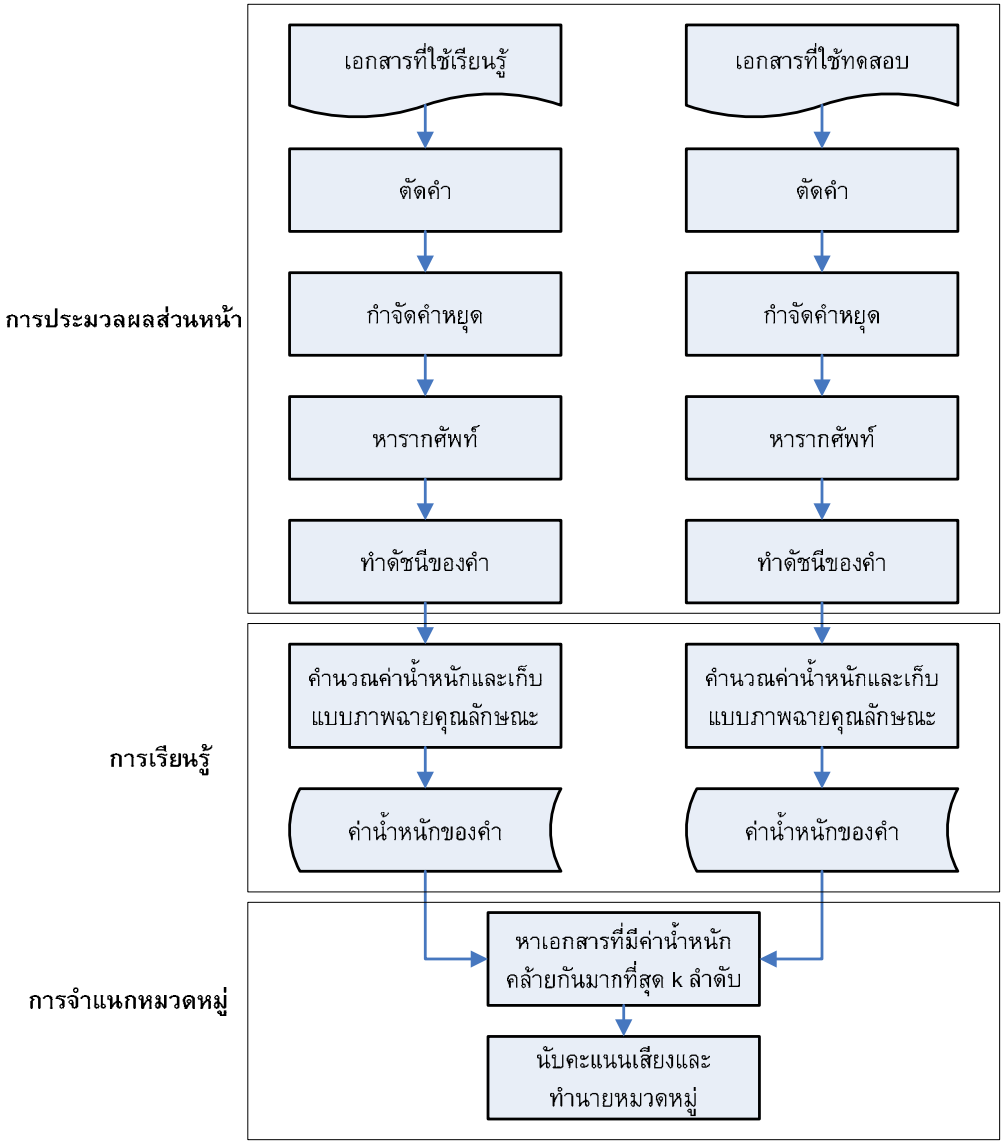
1. ระบบปฏิบัติการ Windows
2. เครื่องมือพัฒนาโปรแกรม j2sdk1.4.2.8
3. โปรแกรมตัดคำ Swath
4. อัลกอริทึมในการหารากศัพท์ของ Porter

3.6 ขั้นตอนการจำแนกหมวดหมู่

ขั้นตอนของการจำแนกหมวดหมู่โดยภาพรวมทั้งระบบ แสดงได้ดังภาพที่ 3-6 โดยสามารถแบ่งเป็น 3 กระบวนการ ดังต่อไปนี้

1. กระบวนการประมวลผลส่วนหน้า ขั้นตอนนี้เป็นขั้นตอนการเตรียมข้อมูล

2. การเรียนรู้ ขั้นตอนนี้เป็นการประยุกต์ใช้อัลกอริทึมเพื่อการเรียนรู้การจำแนกเอกสาร
3. การจำแนกหมวดหมู่ ขั้นตอนนี้เป็นการจัดหมวดหมู่เอกสาร



ภาพที่ 3-6 ขั้นตอนของการจำแนกหมวดหมู่

โดยแต่ละขั้นตอนมีรายละเอียด ดังต่อไปนี้

3.6.1 การประมวลผลส่วนหน้า (Preprocessing)

ขั้นตอนในการประมวลผลส่วนหน้าจะประกอบด้วยการตัดคำ การหารากศัพท์ การกำจัดคำหยุด และการทำดัชนี ซึ่งจะทำทั้งส่วนของเอกสารที่ใช้เรียนรู้และเอกสารทดสอบ

3.6.1.1 การตัดคำ

ในงานวิจัยนี้จะใช้โปรแกรม SWATH (Smart Word Analysis for Thai) ซึ่งนำมาจาก [24] เป็นเครื่องมือที่ใช้ในการตัดคำ งานวิจัยนี้เลือกใช้อัลกอริทึมในการตัดคำแบบ Longest Matching ซึ่งในการทดลองตัดคำเบื้องต้น พบว่าให้ผลในการตัดคำดีที่สุดเมื่อเปรียบเทียบกับผลที่ได้จากอัลกอริทึมอื่นในโปรแกรม swath

3.6.1.2 การหารากศัพท์

การหารากศัพท์ในงานวิจัยนี้ใช้อัลกอริทึมของ Porter ซึ่งเป็นอัลกอริทึมที่นิยมใช้ในการหารากศัพท์ของคำในภาษาอังกฤษ โดยนำมาจาก [25] สำหรับการหารากศัพท์ของคำภาษาไทยนั้นทำเป็นบางส่วนโดยการตัดคำว่า “การ” และ “ความ” ที่อยู่หน้าคำศัพท์ โดยเครื่องมือที่ใช้ในการตัดคำ

3.6.1.3 การกำจัดคำหยุด

หลังจากผ่านการตัดคำและหารากศัพท์แล้ว คำที่ได้จะถูกนำมาเทียบกับรายการคำหยุดทั้งหมด เมื่อไม่ตรงกับคำใดที่อยู่ในรายการคำหยุดจึงจะจัดเก็บคำนั้นลงในดัชนี รายการคำหยุดภาษาอังกฤษนำมาจาก [26] ส่วนรายการคำหยุดภาษาไทยที่ใช้ในงานวิจัยนี้ ส่วนหนึ่งนำมาจากงานวิจัย [23] อีกส่วนหนึ่งมาจากคำศัพท์ที่พบในคลังข้อมูล 300 คำแรกซึ่งนำมาจาก [27] โดยจะนำมาเฉพาะคำที่เป็นคำสันธาน คำบุพบท คำวิเศษณ์ คำอุทาน และคำสรรพนาม นอกจากนี้ยังมีการสร้างคำหยุดเพิ่มขึ้นจากคำหยุดที่ได้มา ดังตัวอย่างในภาพที่ 3-7

ตอน	รวมกับ	ต่อไป	=	ตอนต่อไป
เนื่องจาก	รวมกับ	มา	=	เนื่องมาจาก
ตั้งแต่	รวมกับ	แรก	=	ตั้งแต่แรก
แต่	รวมกับ	อย่างใด	=	แต่อย่างใด
…				

ภาพที่ 3-7 ตัวอย่างของคำหยุดที่เกิดจากการประสมของคำหยุดพื้นฐาน

3.6.1.4 การทำดัชนี

งานวิจัยนี้จัดทำดัชนีแบบ Inverted Index โดยจะนำคำที่ผ่านการหารากศัพท์ และกำจัดคำหยุดแล้วมาเรียงลำดับตามตัวอักษรจากน้อยไปมาก และนับความถี่ของการปรากฏของคำในแต่ละเอกสาร รวมทั้งนับจำนวนเอกสารที่มีคำนั้นปรากฏจัดเก็บลงในดัชนีด้วย ตัวอย่างของดัชนีที่ใช้ในงานวิจัยนี้ เป็นดังภาพที่ 3-8

คำ	จำนวนเอกสารที่พบ	หมายเลขเอกสาร,ความถี่ที่พบในเอกสาร
…		
กก	47	21,3 94,1 95,1 157,3 161,1 …
กกุธภัณฑ์	1	262,1
กฎ	23	172,2 189,3 208,1 210,1 226,1 …
กฎกระทรวง	3	459,1 461,2 466,2
กฎข้อบังคับ	1	274,1
กฎธรรมชาติ	2	227,1 238,1
…		

ภาพที่ 3-8 ตัวอย่างของดัชนี Inverted Index

3.6.1.5 การลดขนาดของคุณลักษณะ

การลดขนาดของคุณลักษณะที่ใช้ในกระบวนการเรียนรู้ในงานวิจัยนี้ จะใช้ค่าความถี่ของเอกสารที่มีคุณลักษณะนั้นปรากฏ (Document Frequency, df) ในการพิจารณาลดขนาด โดยจะเลือกเฉพาะคุณลักษณะที่มีค่าความถี่ของเอกสารมากกว่าค่าที่กำหนด และน้อยกว่าจำนวนเอกสารทั้งหมดที่ใช้เรียนรู้ลบด้วยค่าที่กำหนด การทดลองในงานวิจัยนี้จะใช้ค่าความถี่เอกสารตั้งแต่ 1-3

3.6.2 การเรียนรู้

การเรียนรู้ของอัลกอริทึม FPTC ซึ่งเป็นอัลกอริทึมแบบ Lazy Learning จึงทำเพียงการคำนวณหาค่าน้ำหนักของคุณลักษณะตามสมการที่ (2-5) แล้วทำการเก็บข้อมูลในรูปแบบภาพฉายของคุณลักษณะ โดยในงานวิจัยนี้จะใช้คำ (Word) เป็นคุณลักษณะของเอกสาร

สำหรับการเก็บข้อมูลในรูปของภาพฉายของแต่ละคุณลักษณะ จะเก็บเฉพาะเอกสารที่มีคุณลักษณะนั้นปรากฏ และเลือกเฉพาะเอกสารที่มีค่าน้ำหนักเกินค่าเฉลี่ยของค่าน้ำหนักทั้งหมดของทุกเอกสารในคุณลักษณะนั้น เนื่องจากเอกสารที่มีค่าน้ำหนักน้อยกว่าหรือเท่ากับค่าเฉลี่ยของน้ำหนักทั้งหมดจะไม่มีสิทธิ์ในการให้คะแนนเสียง

3.6.3 การจำแนกหมวดหมู่

ในการจำแนกหมวดหมู่ จะทำการคำนวณค่าน้ำหนักของคำในเอกสารทดสอบ แล้วนำมาเปรียบเทียบความคล้ายกันกับค่าน้ำหนักของคำเดียวกันในเอกสารเรียนรู้โดยการวัดระยะระหว่างค่าน้ำหนักของเอกสารทีละคู่ตามสมการ (3-1) แล้วจึงทำการจัดลำดับของค่าน้ำหนักของเอกสารที่ใช้เรียนรู้ที่มีค่าใกล้เคียงกับค่าน้ำหนักของคุณลักษณะเดียวกันในเอกสารทดสอบมากที่สุดจำนวน k

ลำดับในแต่ละคุณลักษณะ แล้วทำการนับจำนวนหมวดหมู่ของเอกสาร k เอกสารของทุกหมวดหมู่ เพื่อหาว่าหมวดหมู่ใดมีความถี่มากที่สุด แล้วจึงทำนายหมวดหมู่นั้นให้กับเอกสารทดสอบ

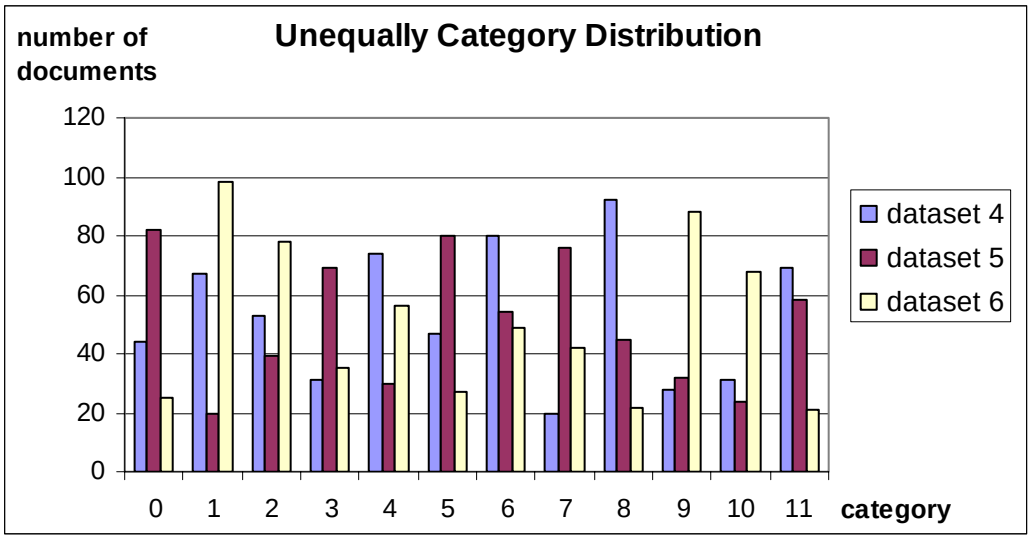
3.7 การออกแบบการทดสอบ

เป้าหมายในการดำเนินการทดสอบในงานวิจัยนี้ คือ การหาค่า k และค่าความถี่เอกสาร (df) ที่ทำให้ได้ผลลัพธ์ในการจำแนกหมวดหมู่ดีที่สุด รวมทั้งการหาลักษณะข้อมูลที่เหมาะสมกับตัวจำแนกหมวดหมู่เอกสารด้วยอัลกอริทึม FPTC โดยจะทำการทดสอบกับข้อมูล 2 ลักษณะคือ ข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน และข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน

ข้อมูลที่ใช้จะถูกแบ่งออกเป็นชุดย่อยจำนวน 6 ชุด โดยข้อมูลชุดที่ 1 ชุดที่ 2 และชุดที่ 3 เป็นข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน ส่วนข้อมูลชุดที่ 4 ชุดที่ 5 และชุดที่ 6 เป็นข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน

ชุดข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากันในแต่ละชุดย่อย จะแบ่งเป็นเอกสารที่ใช้เรียนรู้จำนวนหมวดหมู่ละ 50 เอกสาร และเอกสารที่ใช้ทดสอบจำนวนหมวดหมู่ละ 20 เอกสาร

ชุดข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากันนั้น จะแบ่งเป็นเอกสารที่ใช้ในการเรียนรู้ในแต่ละชุดย่อยและในแต่ละหมวดหมู่จะมีจำนวนเอกสารไม่เท่ากันซึ่งได้มาโดยการสุ่มหยิบ การกระจายตัวของข้อมูลในแต่ละชุดย่อยเป็นดังในภาพที่ 3-9 สำหรับเอกสารที่ใช้ทดสอบจะเป็นชุดเดียวกันกับในชุดข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน



ภาพที่ 3-9 การกระจายตัวของหมวดหมู่ของข้อมูลชุดย่อยที่ 4 ชุดที่ 5 และชุดที่ 6

บทที่ 4
ผลของการวิจัย

ในบทนี้จะกล่าวถึงวิธีการวัดประสิทธิผลของการจำแนกหมวดหมู่ และผลของการจำแนกหมวดหมู่ของข้อมูลข่าวที่ใช้ในการทดสอบ

4.1 วิธีการวัดผลการวิจัย

การประเมินความสามารถของระบบการจำแนกหมวดหมู่นั้น โดยทั่วๆ ไปมักจะวัดกันที่ประสิทธิผล (Effectiveness) มากกว่าที่จะวัดประสิทธิภาพ (Efficiency) นั่นคือ เน้นความสามารถในการตัดสินใจหรือการทำนายหมวดหมู่ที่ถูกต้อง การวัดประสิทธิผลของการจำแนกหมวดหมู่นิยมใช้วิธีการตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ลักษณะของผลการจำแนกหมวดหมู่ที่สามารถเกิดขึ้นได้อาจนำมาเขียนเป็นตารางการณ์ (Contingency Table) ซึ่งอ้างอิงมาจากงานวิจัย [1] ได้ดังนี้

ตารางที่ 4-1 ตารางการณ์ของการจำแนกหมวดหมู่

การจำแนกหมวดหมู่	อยู่ในหมวดหมู่ c_j	ตัดสินโดยผู้เชี่ยวชาญ	
		ใช่	ไม่ใช่
ตัดสินโดยตัวจำแนกอัตโนมัติ	ใช่	TP_j	FP_j
	ไม่ใช่	FN_j	TN_j

TP_j ย่อมาจาก True Positive คือ จำนวนเอกสารที่อยู่ในหมวดหมู่ c_j และตัวจำแนกอัตโนมัติทำนายว่าอยู่ในหมวดหมู่ c_j

FP_j ย่อมาจาก False Positive คือ จำนวนเอกสารที่ไม่อยู่ในหมวดหมู่ c_j แต่ตัวจำแนกอัตโนมัติทำนายว่าอยู่ในหมวดหมู่ c_j

FN_j ย่อมาจาก False Negative คือ จำนวนเอกสารที่อยู่ในหมวดหมู่ c_j แต่ตัวจำแนกอัตโนมัติทำนายว่าไม่อยู่ในหมวดหมู่ c_j

TN_j ย่อมาจาก True Negative คือ จำนวนเอกสารที่ไม่อยู่ในหมวดหมู่ c_j และตัวจำแนกอัตโนมัติทำนายว่าไม่อยู่ในหมวดหมู่ c_j

จากตารางการณ์ สามารถกำหนดวิธีการคำนวณค่าความแม่นยำ (P) และค่าความระลึก (R) ได้ดังนี้

$$P_j = \frac{TP_j}{TP_j + FP_j} \tag{4-1}$$

$$R_j = \frac{TP_j}{TP_j + FN_j} \tag{4-2}$$

P_j คือ ค่าความแม่นยำของการจำแนกหมวดหมู่ c_j และ R_j คือ ค่าความระลึกของการจำแนกหมวดหมู่ c_j

การพิจารณาประสิทธิผลของการจำแนกหมวดหมู่โดยดูค่าความแม่นยำ หรือค่าความระลึกเพียงอย่างเดียวนั้น อาจทำให้ประเมินประสิทธิผลได้ไม่ถูกต้องนัก ทั้งนี้การจำแนกหมวดหมู่ที่ให้ค่าความแม่นยำสูง อาจให้ค่าความระลึกต่ำได้ หากมีจำนวน FN_j มาก หรือให้ค่าความแม่นยำต่ำแต่ให้ค่าความระลึกสูง หากมีจำนวน FP_j มาก ดังนั้นจึงมีวิธีการวัดประสิทธิผลที่นำค่าทั้งสองมาคำนวณร่วมกันเพื่อให้การประเมินค่ามีความถูกต้องมากยิ่งขึ้น นั่นคือ การวัดค่า $F_1 - measure$ โดยมีวิธีการคำนวณ ดังต่อไปนี้

$$F_1 = \frac{2PR}{P + R} \tag{4-3}$$

ค่า F_1 จะมีค่าอยู่ในช่วงระหว่าง 0-1 โดยค่า F_1 ที่มีค่าสูง หมายถึง มีประสิทธิผลในการจำแนกหมวดหมู่สูง ค่า F_1 ที่มีค่าต่ำ หมายถึง มีประสิทธิผลในการจำแนกหมวดหมู่ต่ำ

การวัดค่า F_1 ของการจำแนกหมวดหมู่โดยรวมทั้งระบบ สามารถคำนวณในลักษณะของค่าเฉลี่ยได้ 2 แบบ คือ Micro-average และ Macro-average

ในการคำนวณแบบ Micro-average ค่า F_1 จะถูกคำนวณโดยรวมทุกหมวดหมู่ ค่าความแม่นยำ และค่าความระลึกจะคำนวณจากผลรวมของแต่ละหมวดหมู่ ดังต่อไปนี้

$$P(micro) = \frac{TP}{TP + FP} = \frac{\sum_{j=1}^{|C|} TP_j}{\sum_{j=1}^{|C|} \left(TP_j + FP_j\right)} \tag{4-4}$$

$$R(micro) = \frac{TP}{TP + FN} = \frac{\sum_{j=1}^{|C|} TP_j}{\sum_{j=1}^{|C|} \left(TP_j + FN_j\right)} \tag{4-5}$$

โดย $|C|$ คือ จำนวนหมวดหมู่ทั้งหมด จากสมการ (4-3) ค่า F_1 ในลักษณะของ Micro-average จะคำนวณจากค่าความแม่นยำและค่าความระลึกโดยรวม ดังในสมการที่ (4-6)

$$F_1(micro) = \frac{2P(micro)R(micro)}{P(micro) + R(micro)} \quad (4\text{-}6)$$

ในการคำนวณแบบ Macro-average ค่า F_1 จะถูกคำนวณในแต่ละหมวดหมู่ก่อน แล้วจึงนำค่าที่ได้มาหาค่าเฉลี่ยรวมของทุกหมวดหมู่ ค่าความแม่นยำและค่าความระลึกจะคำนวณจากสมการ (4-7) และ (4-8) ดังต่อไปนี้

$$P(macro) = \frac{\sum_{j=1}^{|C|} P_j}{|C|} \quad (4\text{-}7)$$

$$R(macro) = \frac{\sum_{j=1}^{|C|} R_j}{|C|} \quad (4\text{-}8)$$

จากสมการ (4-3) ค่า F_1 ในลักษณะของ Macro-average จะคำนวณจากค่า F_1 ของแต่ละหมวดหมู่ ดังในสมการที่ (4-9) และ (4-10)

$$F_{1j} = \frac{2P_jR_j}{P_j + R_j} \quad (4\text{-}9)$$

$$F_1(macro) = \frac{\sum_{j=1}^{|C|} F_{1j}}{|C|} \quad (4\text{-}10)$$

ค่าเฉลี่ยแบบ Macro จะเป็นการพิจารณาประสิทธิผลโดยมองว่าแต่ละหมวดหมู่มีน้ำหนักเท่ากัน นั่นคือ ไม่สนใจจำนวนเอกสารในแต่ละหมวดหมู่ ส่วนค่าเฉลี่ยแบบ Micro จะเป็นการพิจารณาโดยมองว่าทุกเอกสารมีน้ำหนักเท่ากัน นั่นคือ พิจารณาจำนวนเอกสารในแต่ละหมวดหมู่ด้วย

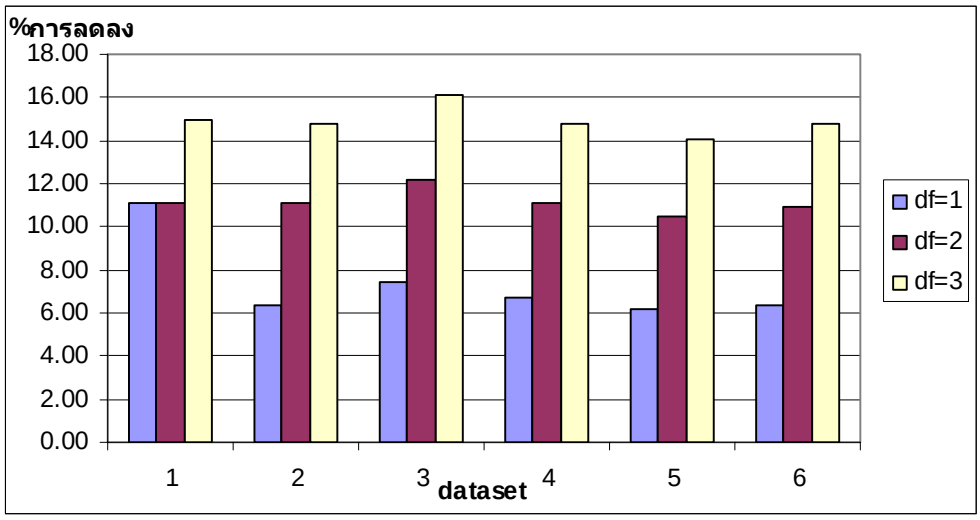
ดังนั้นค่าเฉลี่ยแบบ Macro และค่าเฉลี่ยแบบ Micro อาจให้มีค่าแตกต่างกันได้ โดยเฉพาะอย่างยิ่งเมื่อแต่ละหมวดหมู่มีขอบเขตกว้างมากน้อยต่างกัน หากตัวจำแนกหมวดหมู่ใดให้ค่าประสิทธิผลแบบ Macro-average สูงกว่า Micro-average หมายความว่า ตัวจำแนกหมวดหมู่นั้นทำงานได้ดีกับหมวดหมู่ที่มีขอบเขตแคบหรือมีลักษณะเฉพาะตัวค่อนข้างมาก ทำนองเดียวกันหากตัวจำแนกหมวดหมู่ใดให้ค่าประสิทธิผลแบบ Micro-average สูงกว่า Macro-average หมายความว่าตัวจำแนกหมวดหมู่นั้นทำงานได้ดีกับหมวดหมู่ที่มีขอบเขตกว้างหรือไม่ค่อยมีลักษณะเฉพาะตัวมากนัก ประสิทธิผลของตัวจำแนกหมวดหมู่จึงขึ้นอยู่กับความต้องการหรือวัตถุประสงค์ในการใช้งานตัวจำแนกหมวดหมู่นั้นเป็นสำคัญ

4.2 ผลการวิจัย

ในหัวข้อนี้นำเสนอผลของการทดสอบการจำแนกหมวดหมู่ของข่าวตามเป้าหมายของการทดสอบที่กล่าวไว้ในหัวข้อการออกแบบการทดสอบในบทที่ 3 นั่นคือ การหาค่า k ที่เหมาะสมและค่าความถี่เอกสาร (df) ที่ทำให้ได้ผลลัพธ์ในการจำแนกหมวดหมู่ดีที่สุด

ในการทดสอบใช้ค่า k จำนวน 10 ค่า ได้แก่ 5 10 15 20 25 30 35 40 45 50 สำหรับค่าความถี่ของเอกสารที่ใช้มีจำนวน 4 ค่า ได้แก่ 0 1 2 3 โดย 0 คือ ไม่ลดจำนวนคุณลักษณะเลย

อัตราการลดลงของจำนวนคุณลักษณะเป็นสัดส่วนแปรผันตามค่าความถี่ของเอกสาร โดยมีรายละเอียดดังแสดงในภาพที่ 4-1

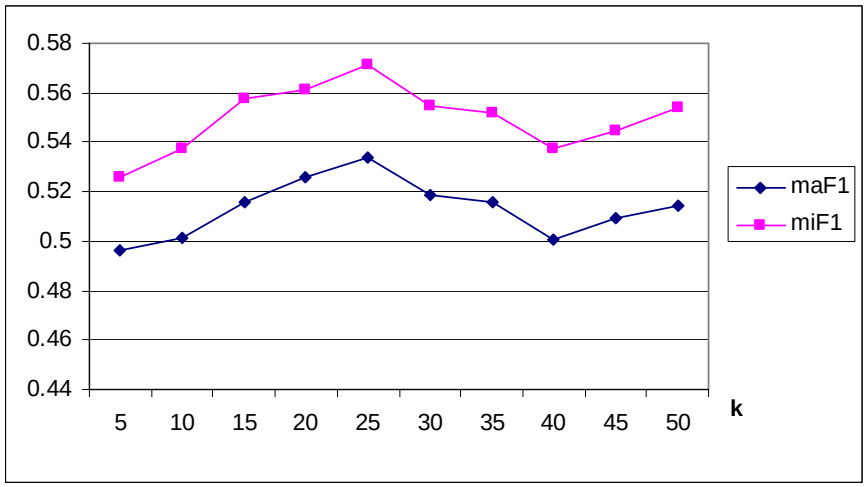


ภาพที่ 4-1 อัตราการลดลงของจำนวนคุณลักษณะ

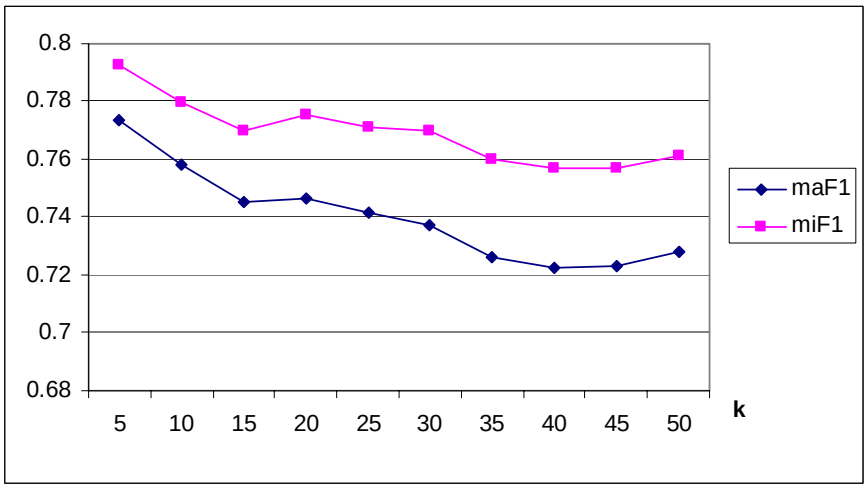
ในการทดสอบจะใช้ข้อมูล 2 ชุด คือ ข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน และข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน เพื่อให้การสรุปผลมีความน่าเชื่อถือ การทดสอบจึงทำกับข้อมูลที่ใช้เรียนรู้ที่แตกต่างกันจำนวน 3 ชุด เมื่อคำนวณหาค่าความแม่นยำ ค่าความระลึก และค่า F_1 ของข้อมูลแต่ละชุดแล้วจึงนำมาหาค่าเฉลี่ยรวมของข้อมูลทั้ง 3 ชุด

4.2.1 การทดสอบกับชุดข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน

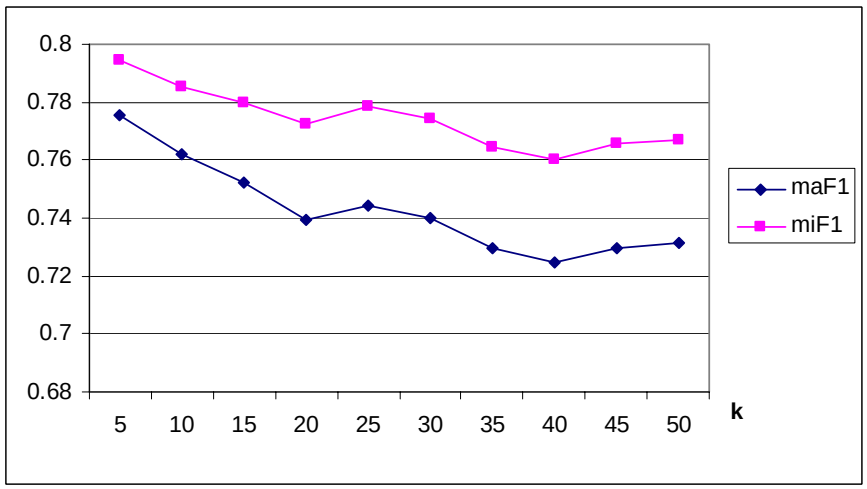
การทดสอบในหัวข้อนี้จะใช้ข้อมูลชุดที่ 1 2 และ 3 ซึ่งประกอบไปด้วยเอกสารที่ใช้เรียนรู้ทั้งหมด 12 หมวดหมู่ หมวดหมู่ละ 50 เอกสารเท่ากัน เมื่อทำการทดสอบด้วยเอกสารทดสอบจำนวนหมวดหมู่ละ 20 เอกสาร และคำนวณหาค่า Micro F_1 และ Macro F_1 แล้ว ได้กราฟแสดงผลการทดสอบสำหรับแต่ละค่า df ดังแสดงในภาพที่ 4-2, 4-3, 4-4 และ 4-5 ตามลำดับ



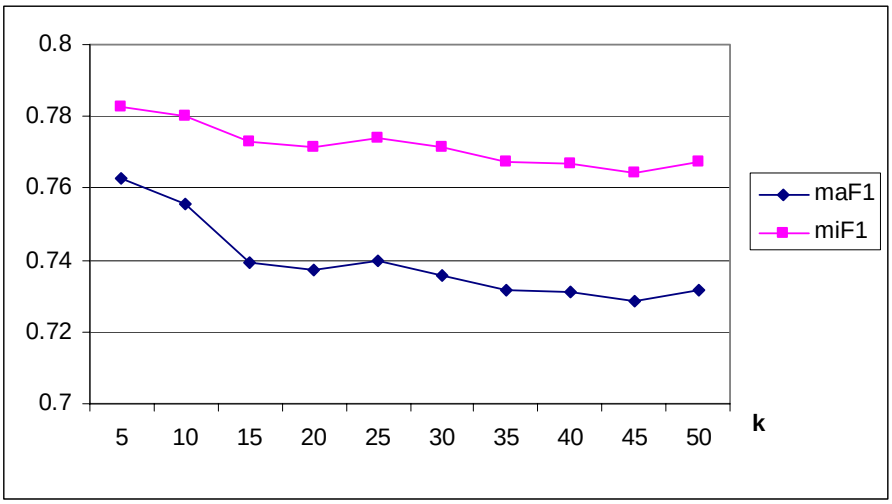
ภาพที่ 4-2 ผลการจำแนกเอกสารที่มีการกระจายตัวเท่ากันและไม่มีการลดจำนวนคุณลักษณะ



ภาพที่ 4-3 ผลการจำแนกเอกสารที่มีการกระจายตัวเท่ากันและลดจำนวนคุณลักษณะด้วย df=1



ภาพที่ 4-4 ผลการจำแนกเอกสารที่มีการกระจายตัวเท่ากันและลดจำนวนคุณลักษณะด้วย df=2



ภาพที่ 4-5 ผลการจำแนกเอกสารที่มีการกระจายตัวเท่ากันและลดจำนวนคุณลักษณะด้วย df=3

ผลการทดลองแสดงให้เห็นว่า การจำแนกเอกสารที่มีการกระจายตัวของหมวดหมู่เท่ากันมีประสิทธิภาพอยู่ในระดับที่ดีพอสมควร เนื่องจากค่า F_1 มีค่าอยู่ในช่วง 0.7 – 0.8

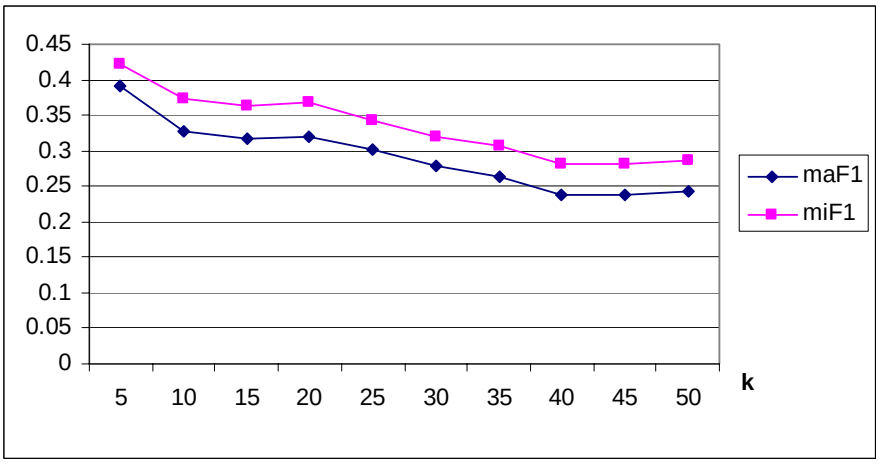
ค่า k ที่ให้ประสิทธิผลในการจำแนกเอกสารในระดับที่ดีสำหรับการทดสอบในงานวิจัยนี้ คือ ค่า k เท่ากับ 5 และมีแนวโน้มว่า ยิ่งค่า k มีค่ามากขึ้นเท่าไหร่ ผลลัพธ์ที่ได้ยิ่งมีคุณภาพน้อยลงเท่านั้น ทั้งนี้น่าจะมีสาเหตุจากการที่ค่า k มีค่ามาก หมายถึง การหยิบเอาเอกสารที่ไม่เกี่ยวข้องกับหมวดหมู่นั้นมาร่วมในการให้คะแนนเสียงมาก ผลการจำแนกหมวดหมู่ที่ได้จึงมีความผิดพลาดมากขึ้น

ส่วนค่าความถี่ของเอกสารที่ใช้ในการลดขนาดของคุณลักษณะนั้น ค่าที่ให้ประสิทธิผลในการจำแนกเอกสารในระดับที่ดี คือ ความถี่มีค่าเป็น 1 และ 2 แสดงให้เห็นว่าการลดขนาดคุณลักษณะของเอกสารลงยิ่งมากยิ่งทำให้จำแนกประเภทได้แย่ลง เนื่องจากจำนวนคุณลักษณะที่ใช้เปรียบเทียบกันได้ระหว่างเอกสารเรียนรู้และเอกสารทดสอบลดลง ส่วนการจำแนกเอกสารโดยไม่ลดขนาดของคุณลักษณะเลย ทำให้จำนวนคุณลักษณะที่ไม่เป็นประโยชน์ต่อการจำแนกหมวดหมู่นั้นมีอยู่มาก

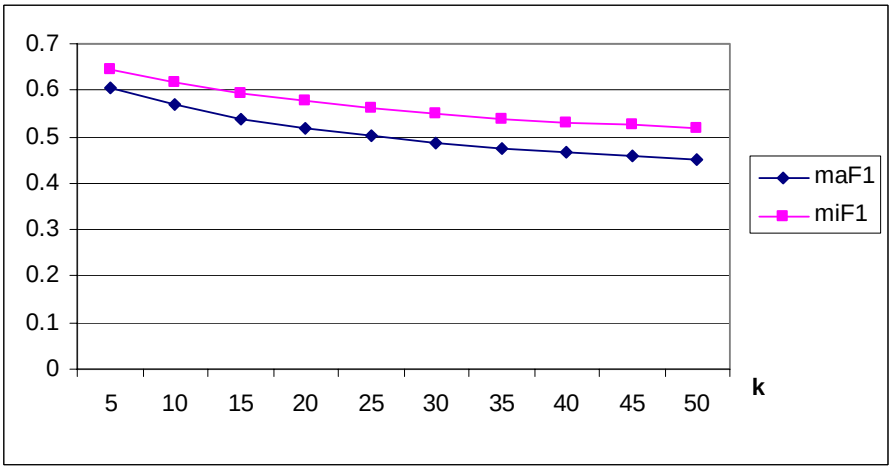
4.2.2 การทดสอบกับชุดข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน

การทดสอบในหัวข้อนี้จะใช้ข้อมูลชุดที่ 4 5 และ 6 ซึ่งมีจำนวนเอกสารที่ใช้ในเรียนรู้ในแต่ละหมวดหมู่ไม่เท่ากัน จำนวนเอกสารที่ใช้ในแต่ละหมวดหมู่แสดงไว้ในภาพที่ 3-8

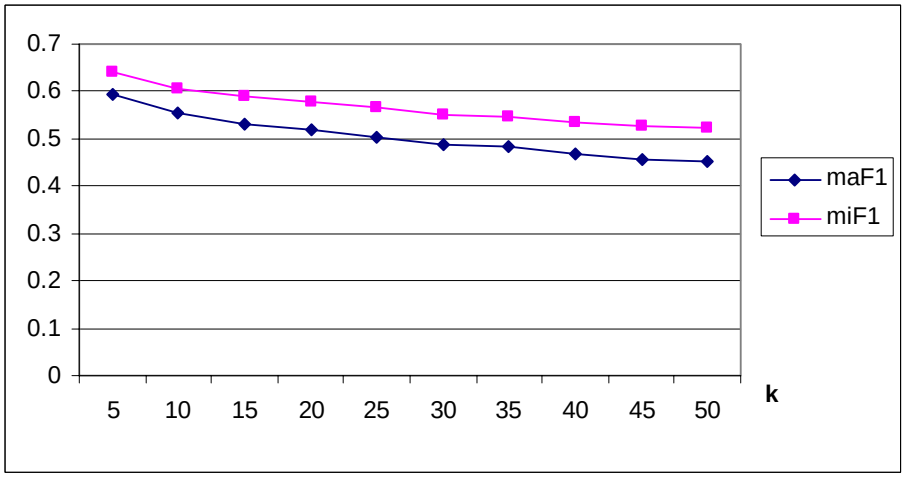
เมื่อทำการทดสอบด้วยเอกสารทดสอบจำนวนหมวดหมู่ละ 20 เอกสาร ซึ่งเป็นข้อมูลชุดเดียวกันกับที่ใช้ในหัวข้อที่ 4.2.1 และคำนวณหาค่า Micro F_1 และ Macro F_1 แล้ว นำมาวาดกราฟแสดงผลการทดสอบสำหรับแต่ละค่า df ได้ดังในภาพที่ 4-6 ภาพที่ 4-7 ภาพที่ 4-8 และภาพที่ 4-9 ตามลำดับ



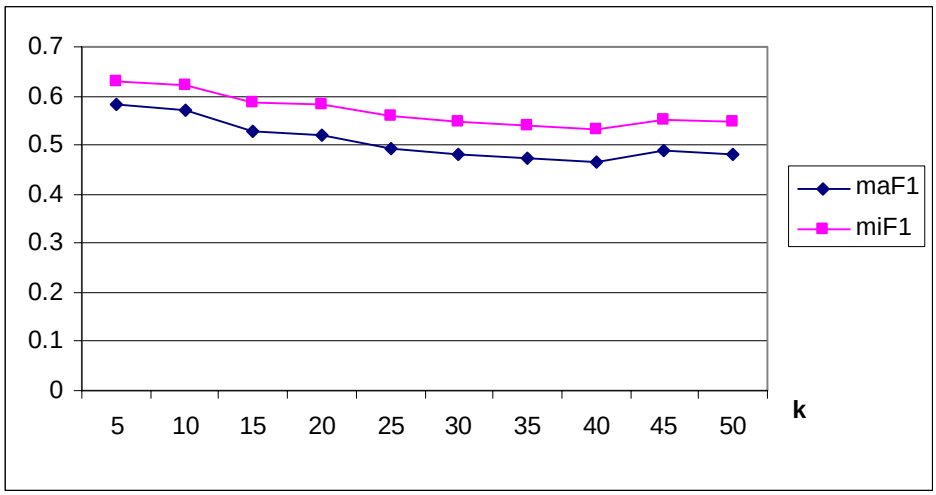
ภาพที่ 4-6 ผลการจำแนกเอกสารที่มีการกระจายตัวไม่เท่ากันและไม่มีการลดจำนวนคุณลักษณะ



ภาพที่ 4-7 ผลการจำแนกเอกสารที่มีการกระจายตัวไม่เท่ากันและลดจำนวนคุณลักษณะด้วย df=1



ภาพที่ 4-8 ผลการจำแนกเอกสารที่มีการกระจายตัวไม่เท่ากันและลดจำนวนคุณลักษณะด้วย df=2



ภาพที่ 4-9 ผลการจำแนกเอกสารที่มีการกระจายตัวไม่เท่ากันและลดจำนวนคุณลักษณะด้วย df=3

ผลการทดลองแสดงให้เห็นว่า การจำแนกเอกสารที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน มีประสิทธิภาพอยู่ในระดับที่ไม่ดี เนื่องจากค่า F_1 ส่วนใหญ่มีค่าอยู่ในช่วง 0.4 – 0.6

ค่า k ที่ให้ประสิทธิผลดีที่สุดในการจำแนกเอกสารข่าวในชุดนี้ คือ ค่า k เท่ากับ 5 และแนวโน้มของค่า k เป็นไปในทำนองเดียวกันกับผลของการทดสอบในหัวข้อที่ 4.2.1 นั่นคือ ค่า k มีค่ามากขึ้นเท่าไหร่ ย่อมหมายถึงการหยิบเอาเอกสารที่ไม่เกี่ยวข้องกับหมวดหมู่นั้นมาร่วมในการให้คะแนนเสียงมากขึ้นท่านั้น ทำให้ผลการจำแนกหมวดหมู่ที่ได้มีความผิดพลาดมากขึ้น

ส่วนค่าความถี่ของเอกสารที่ใช้ในการลดขนาดของคุณลักษณะนั้น ให้ผลลัพธ์ไม่แตกต่างกันมากนัก ยกเว้นการจำแนกเอกสารโดยไม่ลดขนาดของคุณลักษณะเลย ที่มีประสิทธิภาพแย่อย่างเห็นได้ชัด ทั้งนี้เนื่องจากจำนวนคุณลักษณะที่ไม่เป็นประโยชน์ต่อการจำแนกหมวดหมู่นั้นมีอยู่มาก จึงสรุปได้ว่า สำหรับเอกสารที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน ค่าความถี่ที่ใช้ในการลดขนาดของเอกสารทั้งสามค่า มีผลต่อประสิทธิภาพในการจำแนกเอกสารกลุ่มนี้น้อยมาก

จากการทดสอบกับข้อมูลทั้งสองลักษณะ สรุปผลการทดสอบได้ว่า ตัวจำแนกหมวดหมู่อัตโนมัติที่พัฒนาขึ้นใช้ในงานวิจัยนี้ สามารถจำแนกหมวดหมู่เอกสารที่มีการกระจายตัวของหมวดหมู่เท่ากันได้ดีกว่าเอกสารที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน และค่า k ที่ให้ประสิทธิผลดีที่สุด คือ ค่า k เท่ากับ 5 สำหรับค่าความถี่เอกสารที่ใช้ในการลดจำนวนคุณลักษณะที่เหมาะสมมีค่าประมาณ 1-2

บทที่ 5
บทสรุปและแนวทางในอนาคต

ในบทนี้จะกล่าวถึงบทสรุปของการวิจัย ปัญหาและอุปสรรคที่พบ และแนวทางการพัฒนางานวิจัยทางด้านการจำแนกหมวดหมู่เอกสารภาษาไทยในอนาคต

5.1 สรุปผลการวิจัย

งานวิจัยนี้ทำการจำแนกหมวดหมู่เอกสารภาษาไทยด้วยอัลกอริทึม FPTC โดยใช้ข้อมูลข่าวภาษาไทยเป็นกรณีศึกษา วัตุประสงค์หนึ่งในการวิจัย คือ การทดสอบประสิทธิภาพของอัลกอริทึม FPTC เมื่อนำมาประยุกต์ใช้กับภาษาไทย โดยมีการกำหนดเป้าหมายในการทดสอบทั้งหมด 3 ประการ ประการแรก คือ การหาค่า k ที่เหมาะสมกับตัวจำแนกหมวดหมู่เอกสารภาษาไทยในงานวิจัยนี้ ประการที่สองคือ การหาค่าความถี่ของเอกสารที่เหมาะสมในการลดจำนวนคุณลักษณะ ประการที่สามคือ ลักษณะข้อมูลที่เหมาะสมกับตัวจำแนกหมวดหมู่เอกสารภาษาไทยนี้ การทดสอบเพื่อให้ได้ข้อสรุปตามเป้าหมายดังที่กล่าวมา คือ การทดสอบโดยใช้ค่า k ที่แตกต่างกันทั้งหมด 10 ค่า และค่าความถี่เอกสารที่ใช้ลดจำนวนคุณลักษณะที่แตกต่างกันจำนวน 4 คือ รวมทั้งการทดสอบกับข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน และข้อมูลที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน

ผลสรุปที่ได้จากการทดสอบ สามารถสรุปได้ว่า ค่า k ที่เหมาะสมกับตัวจำแนกหมวดหมู่เอกสารภาษาไทยด้วยอัลกอริทึม FPTC ที่ใช้ในงานวิจัยนี้มีค่าเท่ากับ 5 ซึ่งเป็นค่าที่น้อยที่สุด และในการทดสอบพบว่า แนวโน้มของค่า k ยิ่งมีค่ามาก ยิ่งทำให้การจำแนกหมวดหมู่มีประสิทธิผลลดลง เนื่องจากค่า k ที่เพิ่มขึ้น มีส่วนทำให้จำนวนเอกสารที่ไม่เกี่ยวข้องหรือไม่เป็นประโยชน์ต่อการจำแนกเอกสารเพิ่มมากขึ้น

สำหรับค่าความถี่เอกสารที่ใช้ในการลดจำนวนคุณลักษณะลงนั้น ค่าที่เหมาะสม คือ ค่าความถี่เท่ากับ 1 และ 2 ซึ่งสามารถลดจำนวนคุณลักษณะลงได้ในอัตราร้อยละ 6 ถึงร้อยละ 11 ซึ่งจากการทดสอบแสดงให้เห็นว่า การลดจำนวนคุณลักษณะที่น้อยหรือมากเกินไปมีผลต่อประสิทธิผลในการจำแนกเอกสาร

ลักษณะข้อมูลที่เหมาะสมกับตัวจำแนกหมวดหมู่เอกสารในงานวิจัยนี้ คือ ข้อมูลที่มีการกระจายตัวของหมวดหมู่เท่ากัน เนื่องจากให้ผลการจำแนกหมวดหมู่ที่มีประสิทธิผลดีกว่าข้อมูลที่มี

การกระจายตัวของหมวดหมู่ไม่เท่ากัน และมีแนวโน้มว่าตัวจำแนกเอกสารในงานวิจัยนี้ สามารถจำแนกหมวดหมู่ที่มีลักษณะทั่วไป ไม่เฉพาะเจาะจง ได้ดีกว่าหมวดหมู่ที่มีลักษณะเฉพาะตัวสูง เนื่องจากค่า Micro-average มีค่าสูงกว่า Macro-average

5.2 ปัญหาและอุปสรรคในการทำวิจัย

ในการวิจัยพบว่า การตัดคำภาษาไทยไม่ถูกต้องมีผลอย่างมากต่อความถูกต้องในการจำแนกเอกสาร นอกจากนี้การเขียนข่าวภาษาไทยมีการใช้ศัพท์แสลงหรือสำนวนเป็นจำนวนมาก คำที่ใช้ในหมวดหมู่หนึ่งอาจมีความหมายที่แตกต่างไปสำหรับอีกหมวดหมู่หนึ่ง ซึ่งมีผลทำให้จำแนกเอกสารได้ไม่ถูกต้อง ตัวอย่างเช่น ข่าวในหมวดหมู่อาชญากรรมพบคำว่า บุก ยิง ทะลุ จากประโยคตัวอย่างที่ว่า “ผู้ร้ายบุกยิงเหยื่ออย่างอุกอาจ กระสุนทะลุท้ายทอยหนึ่งนัด แล้วยังตามแทงซ้ำนับสิบแผล” ข่าวในหมวดกีฬา ก็พบคำเดียวกันในความหมายของคำแสลงในประโยคที่ว่า “ผีแดงบุกทะลุทะลวงกองหน้าสาริกาดง ยิงกระหน่ำ 3 ต่อ 0” เป็นต้น

5.3 ข้อเสนอแนะและแนวทางการวิจัยในอนาคต

สำหรับแนวทางการวิจัยในอนาคตอาจจะนำการวิเคราะห์โครงสร้างประโยคภาษาไทย และการพิจารณาการปรากฏร่วมกันของคำในเอกสารมาใช้ก่อนที่จะทำการจำแนกหมวดหมู่เพื่อให้การจำแนกหมวดหมู่เอกสารภาษาไทยมีความถูกต้องมากขึ้น

เอกสารอ้างอิง

1. Sebastiani, Fabrizio. "Machine Learning in Automated Text Categorization." **ACM Computing Surveys (CSUR)**. 34 (March 2002) : 1-47.
2. Marquez, Llus. "Machine learning and natural language processing". **Technical Report LSI00-45-R**. Departament de Llenguatges i Sistemes Informatics (LSI), Universitat Politecnica de Catalunya (UPC), Barcelona, Spain, 2000.
3. ณัฐวิทย์ บูรณประภานนท์. **การจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติ**. วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, 2543.
4. ชุลีรัตน์ จรัสกุลชัย, เจษฎา กันทะเสนา และสถาพร คิ้วสุวรรณสุข. "การจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย." **The 5th National Computer Science and Engineering Conference**. 313-324. Chiangmai, Thailand, 2001.
5. วัลลภ อินทร์ฉ่ำ. **ระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ SVM ร่วมกับการประมวลผลภาษา**. วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์, 2548.
6. Kruengkrai, Canasai and Jaruskulchai, Chuleerat. **"**Thai Text Classification based on Naïve Bayes." **Technical Report**. Department of Computer Science, Kasetsart University, 2001.
7. Theeramunkong, Thanaruk and Lertnattee, Verayuth. "Multi-Dimensional Text Classification." **Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)**. 1002-1008. Taiwan, Aug 2002.
8. Han, Eui-Hong and Karypis, George. "Centroid-Based Document Classification: Analysis and Experimental Results". **Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery**. 424-431. London, UK, 2000.
9. Ko, Youngjoong and Seo, Jungyun. "Using the Feature Projection Technique Based on a Normalized Voting Method for Text Classification." **Information Processing & Management**. 40 (March 2002) : 191-208.

10. Holmes, Geoffrey, et all. "Multiclass alternating decision trees". **Proceedings of the 13th European Conference on Machine Learning (ECML)**. 161-172. Helsinki, Finland, 2002.

11. Calvo, Rafael A. "Classifying Financial News with Neural Networks". **Proceedings of the 6th Australasian Document Computing Symposium**. Coffs Harbour, Australia, December 2001.

12. Han , Eui-Hong, et all. "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification". **Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining**. Kowloon, Hong Kong, April 2001.

13. Ilhan, Ufuk. **Application of k-NN and FPTC Based Text Categorization Algorithms to Turkish News Reports**. Master Thesis, Computer Engineering, Institute of Engineering and Science, Bilkent University, 2001.

14. Baoli, Li, Qin, Lu and Shiwen, Yu. "An adaptive *k*-nearest neighbor text categorization strategy." **ACM Transactions on Asian Language Information Processing (TALIP)**. 3 (December 2004) : 215-226.

15. Theeramunkong,Thanaruk and Lertnattee, Verayuth. "Improving Centroid-Based Text Classification Using Term-Distribution Weighting System and Clustering." **Proceedings of International Symposium on Communication and Information Technology (ISCIT-2001)**. Chiangmai, Thailand, Nov 2001.

16. Jaochims, Thorsten. "Text Categorization with Suport Vector Machines: Learning with Many Relevant Features". **Proceedings of the 10th European Conference on Machine Learning**. 137-142. London, UK, 1998.

17. Frasconi, Paolo, Soda, Giovanni and Vullo, Alessandro. "Hidden Markov Models for Text Categorization in Multi-Page Documents." **Journal of Intelligent Information Systems**. 18 (2002) : 195-217.

18. Schapire, Robert E. and Singer, Yoram. "BoosTexter: A Boosting-based System for Text Categorization." **Machine Learning**. 39 (May 2000) : 135-168.

19. Yavuz, Tuba and Guvenir, H. Altay. "Application of k-Nearest Neighbor on Feature Projections Classifier to Text Categorization." **Proceedings of 13th International**

Symposium on Computer and Information Sciences (ISCIS-98). Turkey, October 1998.

20. Ko, Youngjoong and Seo, Jungyun. "Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique." **Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)**. 255-262. Barcelona, Spain, July 2004.
21. Lewis, David D. and Ringuette, Marc. "A Comparison of Two Learning Algorithms for Text Categorization". **Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94)**. 81-93. Las Vegas, USA, 1994.
22. Kehagias, Athanasios, et all. "A Comparison of Word- and Sense-Based Text Categorization Using Several Classification Algorithms." **Journal of Intelligent Information Systems**. 21 (November 2003) : 227-247.
23. Jaruskulchai, Chuleerat. **An Automatic Indexing for Thai Text Retrieval**. PhD Thesis, George Washington University, USA, Aug 31, 1998.
24. Charoenpornsawat, Paisarn. **Software: SWATH - Thai Word Segmentation**. [ออนไลน์] 7 ตุลาคม 2546. [สืบค้นวันที่ 20 พฤษภาคม 2549]. จาก www.cs.cmu.edu/~paisarn/software.html
25. Deng, Xiaotie. **Course Description**. [ออนไลน์] [สืบค้นวันที่ 8 มิถุนายน 2549]. จาก www.cs.cityu.edu.hk/~deng/5286.html
26. Department of Computing Science, University of Glasglow. **Stop Words**. [ออนไลน์] [สืบค้นวันที่ 12 กรกฎาคม 2549]. จาก http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words
27. ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC). **คำศัพท์ที่พบในคลังข้อมูล**. [ออนไลน์] [สืบค้นวันที่ 12 กรกฎาคม 2549]. จาก http://203.185.132.59/thailang/thaichar/word_thai.php

ภาคผนวก ก

ค่าประสิทธิผลของแต่ละชุดข้อมูลโดยละเอียด

ตารางที่ ก-1 ค่าประสิทธิผล F_1- measure ของข้อมูลชุดที่มีการกระจายตัวของหมวดหมู่เท่ากัน

k	dataset	df							
		0		1		2		3	
		maF_1	miF_1	maF_1	miF_1	maF_1	miF_1	maF_1	miF_1
5	1	0.5605	0.5808	0.7614	0.7729	0.7665	0.7817	0.7522	0.7686
	2	0.5485	0.6	0.8095	0.8208	0.8044	0.8167	0.7973	0.8125
	3	0.3804	0.3958	0.7499	0.7833	0.7559	0.7846	0.7381	0.7667
avg		**0.4965**	**0.5255**	**0.7736**	**0.7924**	**0.7756**	**0.7943**	**0.7625**	**0.7826**
10	1	0.5248	0.5502	0.7478	0.7729	0.7593	0.7817	0.7542	0.7773
	2	0.5691	0.6208	0.8056	0.8167	0.8005	0.8167	0.7915	0.8083
	3	0.4106	0.4417	0.7207	0.75	0.7273	0.7583	0.7216	0.7542
avg		**0.5015**	**0.5376**	**0.7581**	**0.7799**	**0.7623**	**0.7856**	**0.7558**	**0.7799**
15	1	0.543	0.5808	0.7352	0.7642	0.7416	0.7729	0.7208	0.7598
	2	0.5458	0.6	0.781	0.7958	0.788	0.8083	0.7793	0.8042
	3	0.4581	0.4917	0.7187	0.75	0.7271	0.7583	0.7174	0.7542
avg		**0.5156**	**0.5575**	**0.745**	**0.77**	**0.7522**	**0.7799**	**0.7392**	**0.7727**
20	1	0.5459	0.559	0.7221	0.7598	0.7196	0.7598	0.7156	0.7555
	2	0.5675	0.6208	0.7804	0.7958	0.7676	0.7917	0.7753	0.8
	3	0.4643	0.5042	0.7366	0.7708	0.7318	0.7667	0.7211	0.7583
avg		**0.5259**	**0.5613**	**0.7464**	**0.7755**	**0.7397**	**0.7727**	**0.7374**	**0.7713**
25	1	0.5454	0.559	0.7137	0.7511	0.7241	0.7642	0.7157	0.7555
	2	0.5796	0.6333	0.7764	0.7917	0.7705	0.7958	0.7745	0.8
	3	0.4759	0.5208	0.7353	0.7708	0.7381	0.775	0.7295	0.7667
avg		**0.5336**	**0.571**	**0.7418**	**0.7712**	**0.7442**	**0.7783**	**0.7399**	**0.774**
30	1	0.5488	0.5633	0.7044	0.7467	0.724	0.7642	0.7158	0.7555
	2	0.5674	0.6208	0.7743	0.7917	0.7661	0.7917	0.7683	0.7958
	3	0.4408	0.4792	0.7333	0.7708	0.7306	0.7667	0.7229	0.7625
avg		**0.519**	**0.5544**	**0.7374**	**0.7697**	**0.7402**	**0.7742**	**0.7357**	**0.7713**
35	1	0.5401	0.5546	0.6996	0.7424	0.712	0.7511	0.7201	0.7598
	2	0.5678	0.6208	0.7624	0.7833	0.7625	0.7875	0.7611	0.7875
	3	0.4387	0.4792	0.716	0.7542	0.7146	0.7542	0.7129	0.7542
avg		**0.5156**	**0.5515**	**0.726**	**0.76**	**0.7297**	**0.7643**	**0.7314**	**0.7672**
40	1	0.5336	0.5459	0.7003	0.7424	0.7115	0.7511	0.7121	0.7511
	2	0.5688	0.6208	0.7549	0.775	0.7621	0.7875	0.7661	0.7917
	3	0.3996	0.4458	0.7128	0.7542	0.6999	0.7417	0.7156	0.7583
avg		**0.5007**	**0.5375**	**0.7227**	**0.7572**	**0.7245**	**0.7601**	**0.7312**	**0.767**
45	1	0.535	0.5459	0.7003	0.7424	0.7194	0.7598	0.7113	0.7511
	2	0.5681	0.6208	0.7553	0.775	0.7621	0.7875	0.7657	0.7917
	3	0.4248	0.4667	0.7134	0.7542	0.7078	0.75	0.7083	0.75
avg		**0.5093**	**0.5445**	**0.723**	**0.7572**	**0.7298**	**0.7658**	**0.7284**	**0.7643**
50	1	0.5387	0.5546	0.7076	0.7467	0.7246	0.7642	0.7176	0.7555
	2	0.5676	0.6208	0.7553	0.775	0.7574	0.7833	0.7657	0.7917
	3	0.4359	0.4875	0.7218	0.7625	0.7117	0.7542	0.712	0.7542
avg		**0.5141**	**0.5543**	**0.7282**	**0.7614**	**0.7312**	**0.7672**	**0.7318**	**0.7671**

ตารางที่ ก-2 ค่าประสิทธิผล F_1 - measure ของข้อมูลชุดที่มีการกระจายตัวของหมวดหมู่ไม่เท่ากัน

k	dataset	df							
		0		1		2		3	
		maF_1	miF_1	maF_1	miF_1	maF_1	miF_1	maF_1	miF_1
5	4	0.3495	0.4000	0.6299	0.6750	0.627	0.6708	0.6098	0.6542
	5	0.4200	0.4500	0.6428	0.6708	0.6361	0.6816	0.6145	0.6542
	6	0.4020	0.4125	0.5388	0.5833	0.5133	0.575	0.5215	0.5833
avg		**0.3905**	**0.4208**	**0.6038**	**0.6431**	**0.5921**	**0.6425**	**0.5819**	**0.6306**
10	4	0.3207	0.375	0.6014	0.6500	0.5971	0.6458	0.5873	0.6375
	5	0.3378	0.375	0.6021	0.6417	0.5907	0.6333	0.568	0.6167
	6	0.3267	0.3708	0.5069	0.5625	0.4727	0.5417	0.5616	0.6167
avg		**0.3284**	**0.3736**	**0.5702**	**0.6181**	**0.5535**	**0.6069**	**0.5723**	**0.6236**
15	4	0.3421	0.3875	0.5912	0.6417	0.5956	0.6458	0.5803	0.6292
	5	0.313	0.3625	0.5369	0.5958	0.5272	0.5917	0.5314	0.5917
	6	0.2933	0.3375	0.481	0.5375	0.4695	0.5333	0.4736	0.5375
avg		**0.3161**	**0.3625**	**0.5364**	**0.5917**	**0.5308**	**0.5903**	**0.5284**	**0.5861**
20	4	0.3443	0.4017	0.5642	0.6208	0.565	0.6167	0.5833	0.6416
	5	0.3165	0.3625	0.5187	0.5792	0.5195	0.5833	0.5183	0.5833
	6	0.2948	0.3375	0.4726	0.5292	0.4736	0.5375	0.4641	0.525
avg		**0.3185**	**0.3672**	**0.5185**	**0.5764**	**0.5194**	**0.5792**	**0.5219**	**0.5833**
25	4	0.2959	0.3333	0.5343	0.5917	0.5449	0.6000	0.527	0.5833
	5	0.3159	0.3625	0.5104	0.575	0.5015	0.5708	0.4944	0.5708
	6	0.2915	0.3333	0.4628	0.5167	0.4638	0.525	0.4605	0.5208
avg		**0.3011**	**0.3431**	**0.5025**	**0.5611**	**0.5034**	**0.5653**	**0.494**	**0.5583**
30	4	0.2933	0.3167	0.5221	0.5833	0.5214	0.5792	0.5116	0.5708
	5	0.2531	0.3083	0.4798	0.5542	0.4894	0.5625	0.4779	0.5583
	6	0.2915	0.3333	0.4619	0.5125	0.4539	0.5125	0.4572	0.5167
avg		**0.2793**	**0.3194**	**0.4879**	**0.55**	**0.4882**	**0.5514**	**0.4822**	**0.5486**
35	4	0.2486	0.2792	0.501	0.5667	0.5122	0.575	0.5029	0.5667
	5	0.2525	0.3083	0.4621	0.5417	0.4813	0.5542	0.4644	0.5417
	6	0.2915	0.3333	0.4557	0.5083	0.4551	0.5125	0.4489	0.5083
avg		**0.2642**	**0.3069**	**0.4729**	**0.5389**	**0.4829**	**0.5472**	**0.4721**	**0.5389**
40	4	0.235	0.2583	0.4919	0.5583	0.4926	0.5583	0.4841	0.5500
	5	0.2481	0.3	0.4588	0.5375	0.4749	0.55	0.4644	0.5417
	6	0.2303	0.2833	0.444	0.4958	0.4391	0.5	0.4465	0.5042
avg		**0.2378**	**0.2806**	**0.4649**	**0.5306**	**0.4689**	**0.5361**	**0.465**	**0.5319**
45	4	0.2367	0.2583	0.4839	0.55	0.4712	0.5417	0.4606	0.5333
	5	0.2481	0.3	0.4499	0.5292	0.4594	0.5375	0.4644	0.5417
	6	0.2303	0.2833	0.4463	0.5	0.4391	0.5	0.5407	0.5833
avg		**0.2384**	**0.2806**	**0.46**	**0.5264**	**0.4566**	**0.5264**	**0.4886**	**0.5528**
50	4	0.2490	0.2792	0.4729	0.5417	0.4654	0.5375	0.4539	0.5292
	5	0.2486	0.3	0.4472	0.525	0.4568	0.5333	0.4562	0.5333
	6	0.2304	0.2833	0.4348	0.4875	0.434	0.4958	0.5352	0.5792
avg		**0.2427**	**0.2875**	**0.4517**	**0.5181**	**0.4521**	**0.5222**	**0.4818**	**0.5472**

ภาคผนวก ข

รายการคำหยุดที่ใช้ในงานวิจัย

รายการคำหยุดที่ใช้ในงานวิจัย

about
above
across
after
afterwards
again
against
all
almost
alone
along
already
also
although
always
am
among
amongst
amount
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anywhere
are
area
areas
around
as
ask
asked
asking
asks
at
away
back
backed
backing
backs
be
became
because
become
becomes
becoming
been
before
beforehand
began
behind
being
beings
below
beside
besides
best
better
between
beyond
big
bill
both
bottom
but
buy
by
call
came
can
cannot
cant
case
cases
certain
certainly
clear
clearly
co
come
computer
con
could
couldnt
cry
d
de
describe
detail
did
differ
different
differently
do
does
done
down
downed
downing
downs
due
during
each
early
eg
eight
either
eleven

else
elsewhere
empty
end
ended
ending
ends
enough
etc
even
evenly
ever
every
everybody
everyone
everything
everywhere
except
face
faces
fact
facts
far
felt
few
fifteen
fify
fill
find
finds
fire
first
five
for
former
formerly
forty
found
four
from
front
full
fully
further
furthered
furthering
furthers
g
gave
general
generally
get
gets
give
given
gives
go
going
good
goods
got
great
greater
greatest
group
grouped
grouping
groups
h
had
has
hasnt
have
having
he
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
high
higher
highest
him
himself
his
how
however
hundred
i
ie
if
important
in
inc
indeed
interest
interested
interesting
interests
into
is
it
its
itself
j
just
k
keep
keeps
kind
knew
know
known
knows
l

large
largely
last
later
latest
latter
latterly
least
less
let
lets
like
likely
long
longer
longest
ltd
m
made
make
making
man
many
may
me
meanwhile
member
members
men
might
mill
mine
more
moreover
most
mostly
move
mr
mrs
much
must
my
myself
n
name
namely
necessary
need
needed
needing
needs
neither
never
nevertheless
new
newer
newest
next
next
nine
no
nobody
non
none
noone
nor
not
nothing
now
nowhere
number
numbered
numbering
numbers
o
of
off
often
old
older
oldest
on
once
one
only
onto
open
opened
opening
opens
or
order
ordered
ordering
orders
other
others
otherwise
our
ours
ourselves
out
over
own
p
part
parted
parting
parts
per
perhaps
place
places
please
point
pointed
pointing
points
possible
present

presented
presenting
presents
problem
problems
put
puts
q
quite
r
rather
re
really
right
room
rooms
s
said
same
saw
say
says
second
seconds
see
seem
seem
seemed
seemed
seeming
seeming
seems
seems
sees
serious
several
shall
she
should
show
showed
showing
shows
side
sides
since
sincere
six
sixty
small
smaller
smallest
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhere
state
states
still
such
sure
system
t
take
taken
ten
than
that
the
their
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereupon
these
they
thick
thin
thing
things
think
thinks
third
this
those
though
thought
thoughts
three
through
throughout
thru
thus
to
today
together
too
took
top
toward
towards
turn
turned
turning
turns
twelve
twenty
two
u

un	where	yet	กล่าว
under	whereafter	you	กล่าวคือ
until	whereas	young	กลุ่ม
up	whereby	younger	กลุ่มก้อน
upon	wherein	youngest	กลุ่มๆ
us	whereupon	your	ก็แล้วแต่
use	wherever	yours	กว่า
used	whether	yours	กว้าง
uses	which	yourself	กว้างขวาง
v	while	yourselves	กว้างๆ
very	whither	z	ก่อน
via	who	เ	ก่อนหน้า
w	whoever	แ	ก่อนหน้านี้
want	whole	ใ	ก่อนๆ
wanted	whom	ไ	กัน
wanting	whose	ก็	กันดีกว่า
wants	why	ก็คือ	กันดีไหม
was	will	ก็แค่	กันเถอะ
way	with	ก็จะ	กันนะ
ways	within	ก็ดี	กันและกัน
we	without	ก็ได้	กันไหม
well	work	ก็ต่อเมื่อ	กันเอง
wells	worked	ก็ตาม	กับ
went	working	ก็ตามแต่	การ
were	works	ก็ตามที	กำลัง
what	would	กระทั่ง	กำลังจะ
whatever	x	กระทำ	กำหนด
when	y	กระนั้น	กู
whence	year	กระผม	เก็บ
whenever	years	กลับ	เกิด

เกิน
เกินๆ
เกี่ยวกัน
เกี่ยวกับ
เกี่ยวข้อง
เกี่ยวเนื่อง
เกี่ยวๆ
เกือบ
เกือบจะ
เกือบๆ
แก
แก่
แก้ไข
ใกล้
ใกล้ๆ
ไกล
ไกลๆ
ขณะ
ขณะเดียวกัน
ขณะใด
ขณะใดๆ
ขณะที่
ขณะนั้น
ขณะนี้
ขณะหนึ่ง
ขวาง
ขวางๆ
ขอ
ของ
ขั้น
ข้า
ข้าง
ข้างเคียง
ข้างต้น
ข้างบน
ข้างล่าง
ข้างๆ
ขาด
ข้าพเจ้า
ข้าฯ
ขึ้น
เขา
เข้า
เข้าใจ
เขียน
คง
คงจะ
คงอยู่
ครบ
ครบครัน
ครบถ้วน
ครั้ง
ครั้งกระนั้น
ครั้งก่อน
ครั้งครา
ครั้งคราว
ครั้งใด
ครั้งที่
ครั้งนั้น
ครั้งนี้
ครั้งละ
ครั้งหนึ่ง
ครั้งหลัง
ครั้งหลังสุด
ครั้งไหน
ครั้งๆ
ครัน
ครับ
คราใด
คราที่
คราวนั้น
คราวนี้
คราวโน้น
คราวละ
คราวหน้า
คราวหนึ่ง
คราวหลัง
คราวไหน
คราวๆ
คราหนึ่ง
คราไหน
คล้าย
คล้ายกัน
คล้ายกันกับ
คล้ายกับ
คล้ายกับว่า
คล้ายว่า
ควร
ความ
ค่อน
ค่อนข้าง
ค่อนข้างจะ
ค่อนไปทาง
ค่อนมาทาง
ค่อย
ค่อยๆ
คะ
ค่ะ
คำ
คิด
คิดว่า
คือ
คุณ
คุณๆ
เคย
เคยๆ
แค่
แค่จะ
แค่นั้น
แค่นี้
แค่เพียง
แค่ว่า
แค่ไหน

ใคร
ใคร่
ใคร่จะ
ใครๆ
ง่าย
ง่ายๆ
ไง
จง
จด
จน
จนกระทั่ง
จนกว่า
จนขณะนี้
จนตลอด
จนถึง
จนทั่ว
จนบัดนี้
จนเมื่อ
จนแม้
จนแม้น
จรด
จรดกับ
จริง
จริงจัง
จริงๆ
จริงๆจังๆ
จวน
จวนจะ
จวนเจียน
จวบ
จวบกับ
จวบจน
จะ
จ้ะ
จ๊ะ
จะได้
จัง
จังๆ
จัด
จัดการ
จัดงาน
จัดแจง
จัดตั้ง
จัดทำ
จัดหา
จัดให้
จับ
จ้า
จ๋า
จาก
จากนั้น
จากนี้
จากนี้ไป
จำ
จำเป็น
จำพวก
จึง
จึงจะ
จึงเป็น
จู่ๆ
ฉะนั้น
ฉะนี้
ฉัน
เฉกเช่น
เฉพาะ
เฉย
เฉยๆ
ไฉน
ช่วง
ช่วงก่อน
ช่วงต่อไป
ช่วงถัดไป
ช่วงท้าย
ช่วงที่
ช่วงนั้น
ช่วงนี้
ช่วงระหว่าง
ช่วงแรก
ช่วงหน้า
ช่วงหลัง
ช่วงๆ
ช่วย
ช่วยๆ
ช้า
ช้านาน
ชาว
ช้าๆ
เช่น
เช่นก่อน
เช่นกัน
เช่นเคย
เช่นดัง
เช่นดังก่อน
เช่นดังเก่า
เช่นดังที่
เช่นดังว่า
เช่นเดียวกัน
เช่นเดียวกับ
เช่นใด
เช่นที่
เช่นที่เคย
เช่นที่ว่า
เช่นนั้น
เช่นนั้นเอง
เช่นนี้
เช่นเมื่อ
เช่นไร
เชื่อ
เชื่อถือ
เชื่อมั่น
เชื่อว่า
ใช่
ใช้
ใช่ไหม
ซะ
ซะก่อน
ซะจน
ซะจนกระทั่ง
ซะจนถึง
ซึ่ง

ซึ่งก็
ซึ่งก็คือ
ซึ่งกัน
ซึ่งกันและกัน
ซึ่งได้แก่
ซึ่งๆ
ณ
ด้วย
ด้วยกัน
ด้วยเช่นกัน
ด้วยที่
ด้วยประการฉะนี้
ด้วยเพราะ
ด้วยว่า
ด้วยเหตุที่
ด้วยเหตุนั้น
ด้วยเหตุนี้
ด้วยเหตุเพราะ
ด้วยเหตุว่า
ด้วยเหมือนกัน
ดัง
ดั่ง
ดังกล่าว
ดังกับ
ดั่งกับ
ดังกับว่า
ดั่งกับว่า
ดังเก่า
ดั่งเก่า
ดังเคย
ดั่งเคย
ดังจะ
ดั่งจะ
ดังเช่น
ดังเช่น
ดั่งเช่น
ดั่งเช่น
ดังเช่นที่
ดั่งเช่นที่
ดังเดิม
ดั่งเดิม
ดังต่อไปนี้
ดังแต่ก่อน
ดั่งแต่ก่อน
ดังที่
ดั่งที่
ดังที่กล่าว
ดังที่เคย
ดังที่จะ
ดังที่เป็น
ดังนั้น
ดังนี้
ดังนี้เช่น
ดังนี้เพราะ
ดังแม้
ดั่งแม้
ดังแม้น
ดั่งแม้น
ดังว่า
ดั่งว่า
ดังเหมือน
ดั่งเหมือน
ด้าน
ด้านๆ
ดำเนิน
ดำเนินการ
ดำเนินงาน
ดำเนินไป
ดิฉัน
ดี
ดีๆ
ดู
ดูจะ
ดูแล
ดูแล้ว
ดูว่า
ดูเหมือน
ดูเหมือนว่า
ดูๆ
เดิม
เดิมที
เดิมๆ
เดียว
เดี๋ยว
เดี๋ยวก่อน
เดียวกัน
เดียวกับ
เดี๋ยวนั้น
เดี๋ยวนี้
แด่
โดย
โดยง่าย
โดยเฉพาะ
โดยเฉพาะอย่างยิ่ง
โดยดี
โดยดุษฎี
โดยตลอด
โดยทั่ว
โดยทั่วกัน
โดยทั่วถึง
โดยทั่วไป
โดยทั่วๆไป
โดยที่
โดยแท้
โดยแท้จริง
โดยนัย
โดยปกติ
โดยมัก
โดยมักจะ
โดยมาก
โดยเมื่อ
โดยรวม
โดยรวมๆ
โดยเร็ว
โดยละม่อม
โดยลำดับ
โดยส่วนมาก
โดยส่วนรวม
โดยส่วนใหญ่
ใด

ใดๆ
ได้
ได้แก่
ได้แต่
ได้ที่
ได้มา
ได้รับ
ตน
ตนเอง
ตนฯ
ตรง
ตรงๆ
ตลอด
ตลอดกาล
ตลอดกาลนาน
ตลอดจน
ตลอดถึง
ตลอดทั้ง
ตลอดทั่ว
ตลอดทั่วถึง
ตลอดทั่วทั้ง
ตลอดปี
ตลอดไป
ตลอดมา
ตลอดระยะเวลา
ตลอดวัน
ตลอดเวลา
ตลอดศก
ต่อ
ต่อกัน
ต่อกับ
ต้อง
ต้องการ
ต่อจากนั้น
ต่อจากนี้
ต่อด้วย
ต่อแต่นี้
ตอน
ตอนก่อน
ตอนใด
ตอนต่อ
ตอนต่อไป
ตอนต่อมา
ตอนถัดไป
ตอนถัดมา
ตอนที่
ตอนที่แล้ว
ตอนนั้น
ตอนนี้
ตอนแรก
ตอนสุดท้าย
ตอนหน้า
ตอนหลัง
ตอนไหน
ตอนๆ
ต่อเนื่อง
ต่อไป
ต่อไปนี้
ต่อมา
ต่อเมื่อ
ต่อว่า
ต่อให้
ต่อๆ
ตะหาก
ตั้ง
ตั้งต้น
ตั้งแต่
ตั้งแต่นั้น
ตั้งแต่นี้
ตั้งแต่แรก
ตั้งที่
ตั้งอยู่
ตั้งๆ
ตัว
ตัวใด
ตัวที่
ตัวนั้น
ตัวนี้
ตัวโน้น
ตัวละ
ตัวไหน
ตัวอย่างเช่น
ตัวเอง
ตัวๆ
ต่าง
ต่างก็
ต่างหาก
ต่างๆ
ตาม
ตามด้วย
ตามแต่
ตามที
ตามที่
ตามๆ
เต็มไปด้วย
เต็มไปหมด
เต็มๆ
แต่
แต่ก็
แต่ก่อน
แต่จะ
แต่เดิม
แต่ต้อง
แต่ถ้า
แต่ทว่า
แต่ที่
แต่นั้น
แต่เพียง
แต่เมื่อ
แต่ไร
แต่ละ
แต่ว่า
แต่ไหน
แต่อย่างใด
โต
โตๆ
ใต้
ถ้า
ถ้าจะ
ถ้าหาก

ถึง
ถึงแก่
ถึงจะ
ถึงบัดนั้น
ถึงบัดนี้
ถึงเมื่อ
ถึงเมื่อใด
ถึงเมื่อไร
ถึงแม้
ถึงแม้จะ
ถึงแม้ว่า
ถึงอย่างไร
ถือ
ถือว่า
ถูก
ถูกต้อง
ถูกๆ
เถอะ
เถิด
ทรง
ทว่า
ทั้ง
ทั้งคน
ทั้งตัว
ทั้งที
ทั้งที่
ทั้งนั้น
ทั้งนั้นด้วย
ทั้งนั้นเพราะ
ทั้งนี้
ทั้งนี้เช่น
ทั้งนี้ด้วย
ทั้งนี้เพราะ
ทั้งปวง
ทั้งเป็น
ทั้งมวล
ทั้งสิ้น
ทั้งหมด
ทั้งหลาย
ทั้งๆ
ทั้งๆที่
ทัน
ทันใด
ทันใดนั้น
ทันที
ทันทีทันใด
ทั่ว
ทั่วกัน
ทั่วถึง
ทั่วถึงกัน
ทั่วทั้ง
ทั่วไป
ทั่วๆ
ทั่วๆไป
ทาง
ทางๆ
ท่าน
ท่ามกลาง
ทำ
ทำงาน
ทำไม
ทำไร
ทำให้
ทำๆ
ที
ที่
ที่จริง
ที่ซึ่ง
ทีเดียว
ทีใด
ที่ใด
ที่ได้
ทีเถอะ
ที่แท้
ที่แท้จริง
ที่นั่น
ทีนี้
ที่นี่
ที่นี้
ทีไร
ทีละ
ที่ละ
ที่แล้ว
ที่ว่า
ที่สุด
ที่แห่งนั้น
ที่ไหน
ทีๆ
ที่ๆ
ทุก
ทุกคน
ทุกครั้ง
ทุกคราว
ทุกคราว
ทุกชิ้น
ทุกตัว
ทุกทาง
ทุกที
ทุกที่
ทุกเมื่อ
ทุกวัน
ทุกวันนี้
ทุกสิ่ง
ทุกหน
ทุกแห่ง
ทุกอย่าง
ทุกอัน
ทุกๆ
เท่า
เท่ากัน
เท่ากับ
เท่าใด
เท่าที่
เท่านั้น
เท่านี้
เท่าไร
เท่าไหร่
แท้
แท้จริง
เธอ

นอก
นอกจาก
นอกจากที่
นอกจากนั้น
นอกจากนี้
นอกจากว่า
นอกนั้น
นอกเหนือ
นอกเหนือจาก
น้อย
น้อยกว่า
น้อยๆ
นะ
น่ะ
นัก
นักๆ
นั่น
นั้น
นั่นไง
นั่นเป็น
นั่นแหละ
นั่นเอง
นั้นๆ
นับ
นับจากนั้น
นับจากนี้
นับตั้งแต่
นับแต่
นับแต่ที่
นับแต่นั้น
นับแต่นี้
น่า
นาง
นางสาว
น่าจะ
นาน
นานๆ
นาย
นำ
นำพา
นำมา
นำมาซึ่ง
นิ
นิด
นิดหน่อย
นิดๆ
นี่
นี้
นี่ไง
นี่นา
นี่แน่ะ
นี่แหละ
นี้แหละ
นี่เอง
นี้เอง
นู่น
นู้น
เน้น
เนี่ย
เนี่ยเอง
เนื่อง
เนื่องจาก
เนื่องด้วย
เนื่องถึง
เนื่องมาจาก
แน่
แน่ะ
โน่น
โน้น
โน่นไง
โน่นแน่ะ
ใน
ในช่วง
ในที่
ในเมื่อ
ในระหว่าง
ในอันที่
ในอันที่จะ
บน
บอก
บอกแล้ว
บอกว่า
บ่อย
บ่อยกว่า
บ่อยครั้ง
บ่อยๆ
บัดดล
บัดเดี๋ยว
บัดเดี๋ยวนี้
บัดนั้น
บัดนี้
บาง
บ้าง
บางกว่า
บางขณะ
บางครั้ง
บางครา
บางคราว
บางที
บางที่
บางหน
บางแห่ง
บางๆ
แบบ
ปฏิบัติ
ประกอบ
ประการ
ประการฉะนี้
ประการใด
ประการหนึ่ง
ประมาณ
ประสบ
ปรับ
ปรากฏ
ปรากฏว่า
ปัจจุบัน
ปิด
เป็น
เป็นด้วย
เป็นดัง

เป็นต้น
เป็นต้นไป
เป็นต้นมา
เป็นแต่
เป็นแต่เพียง
เป็นที
เป็นที่
เป็นที่สุด
เป็นเพราะ
เป็นเพราะว่า
เป็นเพียง
เป็นเพียงว่า
เป็นเพื่อ
เป็นอัน
เป็นอันมาก
เป็นอันว่า
เป็นอันๆ
เป็นอาทิ
เป็นๆ
เปลี่ยน
เปลี่ยนแปลง
เปิด
เปิดเผย
ไป
ไป่
ผ่าน
ผ่านๆ
ผิด
ผิดๆ
ผู้
ผู้ใด
เผิน
เผินๆ
เผื่อ
เผื่อจะ
เผื่อที่
เผื่อว่า
เผื่อๆ
ฝักฝ่าย
ฝักใฝ่
ฝ่าย
ฝ่ายใด
ฝ่ายๆ
พฉฯ
พบ
พบว่า
พยายาม
พร้อม
พร้อมกัน
พร้อมกันกับ
พร้อมกันนั้น
พร้อมกันนี้
พร้อมกับ
พร้อมด้วย
พร้อมทั้ง
พร้อมที่
พร้อมที่จะ
พร้อมเพียง
พร้อมๆ
พร้อมๆกัน
พร้อมๆกับ
พร้อมๆด้วย
พร้อมๆทั้ง
พวก
พวกกัน
พวกกู
พวกแก
พวกเขา
พวกคุณ
พวกฉัน
พวกท่าน
พวกที่
พวกเธอ
พวกนั้น
พวกนี้
พวกมู้น
พวกโน้น
พวกมัน
พวกมึง
พวกเรา
พวกไหน
พวกๆ
พอ
พอกัน
พอควร
พอจะ
พอดี
พอตัว
พอทำเนา
พอที
พอที่
พอที่จะ
พอเพียง
พอแล้ว
พอสม
พอสมควร
พอเหมาะ
พอๆ
พอๆกัน
พัฒนา
พา
พิจารณา
พึง
พึ่ง
พื้นๆ
พูด
เพราะ
เพราะฉะนั้น
เพราะว่า
เพราะๆ
เพิ่ง
เพิ่งจะ
เพิ่ม
เพิ่มเติม
เพียง
เพียงแค่
เพียงใด
เพียงแต่
เพียงพอ
เพียงเพราะ

เพียงเพื่อ
เพียงไร
เพียงไหน
เพื่อ
เพื่อที่
เพื่อที่จะ
เพื่อว่า
เพื่อให้
ภาค
ภาคฯ
ภาย
ภายใต้
ภายนอก
ภายใน
ภายภาค
ภายภาคหน้า
ภายหน้า
ภายหลัง
มอง
มองว่า
มัก
มักจะ
มัน
มันๆ
มั้ย
มั้ยนะ
มั้ยนั่น
มั้ยเนี่ย
มั้ยล่ะ
มา
มาก
มากกว่า
มากมาย
มากๆ
มิ
มิฉะนั้น
มิใช่
มิได้
มี
มีแต่
มึง
มุ่ง
มุ่งเน้น
มุ่งหมาย
เมื่อ
เมื่อก่อน
เมื่อครั้ง
เมื่อครั้งก่อน
เมื่อคราว
เมื่อคราวก่อน
เมื่อคราวที่
เมื่อคืน
เมื่อเช้า
เมื่อใด
เมื่อนั้น
เมื่อนี้
เมื่อเย็น
เมื่อไร
เมื่อวันวาน
เมื่อวาน
เมื่อไหร่
แม้
แม้กระทั่ง
แม้แต่
แม้น
แม้นกระทั่ง
แม้นว่า
แม้นหาก
แม้นเหมือน
แม้ว่า
แม้หาก
ไม่
ไม่ค่อย
ไม่ค่อยจะ
ไม่ค่อยเป็น
ไม่ใช่
ไม่แต่
ไม่เป็นไร
ไม่ว่า
ไม่เสียที
ไม่ไหว
ยก
ยกให้
ยอม
ย่อม
ยอมรับ
ยอมรับว่า
ย่อมๆ
ย่อย
ย่อยๆ
ยัง
ยังคง
ยังงั้น
ยังงี้
ยังโง้น
ยังไง
ยังไงกัน
ยังไงซะ
ยังไงเสีย
ยังจะ
ยังแต่
ยาก
ยาว
ยาวนาน
ยาวๆ
ย้ำ
ย้ำเตือน
ยิ่ง
ยิ่งกว่า
ยิ่งขึ้น
ยิ่งขึ้นไป
ยิ่งจน
ยิ่งจะ
ยิ่งนัก
ยิ่งไปกว่า
ยิ่งไปกว่านั้น
ยิ่งเมื่อ
ยิ่งแล้ว
ยิ่งใหญ่
ยิ่งๆ

ยืนนาน
ยืนยง
ยืนยัน
ยืนยาว
เยอะ
เยอะแยะ
เยอะๆ
แยะ
แยะๆ
รวด
รวดเร็ว
รวม
ร่วม
รวมกัน
ร่วมกัน
รวมด้วย
ร่วมด้วย
รวมถึง
รวมทั้ง
ร่วมมือ
รวมๆ
ระยะ
ระยะๆ
ระหว่าง
รับ
รับรอง
รึ
รึว่า
รือ
รือว่า
เร็ว
เร็วๆ
เรา
เราๆ
เริ่ม
เรียก
เรียกร้อง
เรียบ
เรียบร้อย
เรียบๆ
เรื่อง
เรื่องๆ
เรื่อย
เรื่อยๆ
แรก
แรกๆ
ไร
ลง
ล้วน
ล้วนจน
ล้วนแต่
ล้วนแล้ว
ล้วนแล้วแต่
ล้วนๆ
ละ
ล่ะ
ล่าง
ล่าสุด
เล็ก
เล็กน้อย
เล็กๆ
เลย
เล่า
เล่าว่า
เลือก
แล้ว
แล้วกัน
แล้วแต่
แล้วเสร็จ
และ
วันใด
วันนั้น
วันนี้
วันไหน
วันๆ
ว่า
วาง
วางไว้
ว่าด้วย
ไว้
ส่ง
ส่งๆ
สนใจ
สบาย
สบายๆ
สมัย
สมัยก่อน
สมัยที่
สมัยนั้น
สมัยนี้
สมัยโน้น
สมัยเมื่อ
สร้าง
ส่วน
ส่วนเกิน
ส่วนด้อย
ส่วนดี
ส่วนใด
ส่วนที่
ส่วนน้อย
ส่วนนั้น
ส่วนนี้
ส่วนมาก
ส่วนหนึ่ง
ส่วนใหญ่
ส่วนไหน
สอง
สะดวก
สั่ง
สั้น
สั้นๆ
สามารถ
สำคัญ
สำหรับ
สิ่ง
สิ่งใด
สิ่งนั้น
สิ่งนี้
สิ่งไหน
สิ้น

สิ้นกาลนาน
สืบเนื่อง
สุด
สุดๆ
สู่
สูง
สูงกว่า
สูงส่ง
สูงสุด
สูงๆ
เสมือนกับ
เสมือนว่า
เสร็จ
เสร็จกัน
เสร็จแล้ว
เสร็จสมบูรณ์
เสร็จสิ้น
เสีย
เสียก่อน
เสียจน
เสียจนกระทั่ง
เสียจนถึง
เสียด้วย
เสียนั่น
เสียนั่นเอง
เสียนี่
เสียนี่กระไร
เสียยิ่ง
เสียยิ่งนัก
เสียแล้ว
เสียสิ้น
แสดง
แสดงว่า
หน
หนอ
หนอย
หน่อย
หนอยแน่
หนอยแน่ะ
หน่อยๆ
หนึ่ง
หมด
หมดกัน
หมดสิ้น
หมายความ
หมายความถึง
หมายความว่า
หมายใจ
หมายถึง
หรือ
หรือไง
หรือเปล่า
หรือไม่
หรือยัง
หรือไร
หรือว่า
หรืออย่างไร
หลัก
หลักๆ
หลังจาก
หลังๆ
หลาก
หลากหลาย
หลาย
หลายๆ
หา
หาก
หากกระนั้น
หากแต่
หากทว่า
หากแม้
หากแม้น
หากแม้นว่า
หากแม้ว่า
หากว่า
หาความ
หาใช่
หาใช่ไม่
หาไม่
หารือ
เหตุใด
เหตุนั้น
เหตุนี้
เหตุไร
เห็น
เห็นแก่
เห็นควร
เห็นจะ
เห็นดี
เห็นว่า
เห็นสมควร
เห็นเหมาะ
เห็นๆ
เหมาะ
เหมาะควร
เหมาะสม
เหมาะๆ
เหมือน
เหมือนกัน
เหมือนกันกับ
เหมือนกันว่า
เหมือนกับ
เหมือนแม้น
เหมือนว่า
เหล่า
เหล่านั้น
เหล่านี้
เหลือ
เหลือเกิน
แห่ง
แห่งใด
แห่งนั้น
แห่งนี้
แห่งโน้น
แห่งไหน
แหละ
ให้
ให้แก่
ใหญ่
ใหญ่โต

ใหญ่ๆ
ให้ดี
ให้แค่
ให้ไป
ใหม่
ให้มา
ใหม่ๆ
ไหน
ไหนๆ
อดีต
อนึ่ง
อยาก
อย่าง
อย่างเช่น
อย่างดี
อย่างเดียว
อย่างใด
อย่างที่
อย่างน้อย
อย่างนั้น
อย่างนี้
อย่างโน้น
อย่างมาก
อย่างยิ่ง
อย่างไร
อย่างไรก็
อย่างไรก็ดี
อย่างไรก็ตาม
อย่างไรกัน
อย่างไรซะ
อย่างไรเล่า
อย่างไรเสีย
อย่างละ
อย่างหนึ่ง
อย่างไหน
อย่างๆ
อยู่
อยู่ๆ
ออก
อะ
อ่ะ
อ้ะ
อ๊ะ
อะไร
อ๋ะๆ
อัน
อันจะ
อันใด
อันได้แก่
อันที่
อันที่จริง
อันที่จะ
อันเนื่องจาก
อันเนื่องมาจาก
อันละ
อันไหน
อันๆ
อาจ
อาจจะ
อาจเป็น
อาจเป็นด้วย
อาจเป็นเพราะ
อาจเพราะ
อาทิ
อาทิเช่น
อ่าน
อีก
อื่น
อื่นๆ
เอง
เอ็ง
เอา
เอาแต่
ฯ
ฯล
ฯลฯ
ะ
า
ำ
ิ
ี
ึ
ื
ุ
ู
ๆ

ประวัติผู้วิจัย

ชื่อ : นางสาวณิชาพร สุระ
ชื่อวิทยานิพนธ์ : การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้อัลกอริทึม FPTC
สาขาวิชา : วิทยาการคอมพิวเตอร์

ประวัติ

ประวัติการศึกษา จบการศึกษาระดับปริญญาตรี จากสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ ในสาขาวิชาวิทยาการคอมพิวเตอร์ประยุกต์ คณะวิทยาศาสตร์ประยุกต์ ปีการศึกษา 2539

ประวัติการทำงาน เริ่มรับราชการ เข้าทำงานในตำแหน่งนักวิชาการคอมพิวเตอร์ ที่สำนักคอมพิวเตอร์และเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ เมื่อปี 2540

สถานที่ติดต่อ สำนักคอมพิวเตอร์และเทคโนโลยีสารสนเทศ อาคารเอนกประสงค์ ชั้น 5 สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ