รายงานการวิจัย
เรื่อง

การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นใน
แบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน

Text Classification Efficiency of Open Ended Questionnaire by
Naïve-Bayes and Support Vector Machine

โดย
อังศุมาลี สุทธภักติ

การวิจัยครั้งนี้ได้รับทุนอุดหนุนการวิจัยจากวิทยาลัยราชพฤกษ์
ปีการศึกษา 2553

รายงานการวิจัย
เรื่อง

การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นใน
แบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน

Text Classification Efficiency of Open Ended Questionnaire by Naïve-Bayes and Support Vector Machine

โดย
อังศุมาลี สุทธภักติ

การวิจัยครั้งนี้ได้รับทุนอุดหนุนการวิจัยจากวิทยาลัยราชพฤกษ์
ปีการศึกษา 2553
ปีที่ทำการวิจัยแล้วเสร็จ 2554

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด และเพื่อทดสอบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด ด้วยวิธีเนอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลที่ได้จากแบบสอบถามเรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร จำนวน 400 ชุด โดยข้อมูลจะถูกแบ่งออกเป็น 2 ส่วน คือ ข้อมูลสำหรับเรียนรู้และข้อมูลทดสอบ (อัตราส่วน 8:2) ซึ่งหมวดหมู่ข้อความแสดงความคิดเห็นสามารถแบ่งออกเป็นเชิงบวก กลาง และลบ

ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่พบว่า การจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชนทั้ง 4 ด้าน (ด้านราคา ด้านสถานที่ ด้านความรู้และทักษะที่ได้รับ และด้านความยอมรับในสังคม) ด้วยวิธีเนอีฟเบย์ มีค่าเอฟเมเชอร์ (F-Measure) อยู่ในช่วงระหว่าง 0.719-0.835 ส่วนการจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชนทั้ง 4 ด้าน ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนมีค่าเอฟเมเชอร์อยู่ในช่วงระหว่าง 0.903-0.963 ดังนั้น การจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิดโดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพดีกว่าวิธีเนอีฟเบย์

คำสำคัญ: การจำแนกข้อความ, แบบสอบถามปลายเปิด, เนอีฟเบย์, ซัพพอร์ตเวกเตอร์แมชชีน

Abstract

The objectives of this research were to classify the text of opinion in Thai language containing in open-ended questionnaire and to determine the efficiency of classification of the text of opinion in Thai language containing in open-ended questionnaire, employing Naïve-Bayes and Support Vector Machine. The data of this research were collected from 400 questionnaires regarding opinion towards the government education and private education. These data were divided into a training set and a testing set in the proportion of 8:2. In addition, the groups of statement were categorized into positive, normal and negative.

The experimental results were stated that the classification of the opinion including price, location, knowledge and skills, and social acceptance affected on both the government education and private education. The F-Measure values for Naïve-Bayes and Support Vector Machine were in the range of 0.719-0.835 and 0.903-0.963, respectively. Hence, the Support Vector Machine had more efficiency to classify the open-ended questionnaire than Naïve-Bayes.

Keywords: Text Classification, Open-Ended Questionnaire, Naïve-Bayes, Support Vector Machine

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณวิทยาลัยราชพฤกษ์ ที่ได้จัดสรรงบประมาณเพื่อสนับสนุนการวิจัยในครั้งนี้ ปี พ.ศ. 2553

ขอขอบคุณ ดร.ชูชาติ หฤไชยะศักดิ์ นักวิจัย ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กระทรวงวิทยาศาสตร์และเทคโนโลยี ที่ได้ให้คำปรึกษาต่างๆ และเป็นผู้สอนเรื่องทำเหมืองข้อความ (Text mining) ทำให้เกิดแนวความคิดในการทำงานวิจัยเล่มนี้ และขอขอบคุณท่านอาจารย์ รองศาสตราจารย์ ดร.สมชาย ปราการเจริญ อาจารย์ประจำคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ที่ให้คำปรึกษาต่างๆ ทำให้รายงานวิจัยเล่มนี้เสร็จสมบูรณ์ได้

ขอขอบคุณผู้ตอบแบบสอบถามทุกท่าน ทำให้ได้ข้อมูลข้อคิดเห็นมาใช้ในการทดลองในงานวิจัยในครั้งนี้ และขอขอบคุณ คุณชัยวุฒิ ชุนช่วยเจริญ ที่ให้ความช่วยเหลือในการเก็บข้อมูลและจัดเตรียมข้อมูลข้อความคิดเห็นเพื่อใช้สำหรับการวิเคราะห์ผล ร่วมทั้งค่อยสนับสนุนในส่วนอื่นๆ ด้วยดีมาตลอด สุดท้ายขอขอบคุณ คุณรพีพงษ์ แย้มสุวรรณ อาจารย์ประจำสาขาคอมพิวเตอร์ธุรกิจ คณะบริหารธุรกิจ วิทยาลัยราชพฤกษ์ที่ให้ความช่วยเหลือทางด้านโปรแกรม และคำปรึกษาต่างๆ ทำให้งานวิจัยเล่มนี้สำเร็จลุล่วงได้

อังศุมาลี สุทธภักติ

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

แบบสอบถาม (Questionnaire) เป็นเครื่องมือวิจัยชนิดหนึ่งที่นิยมใช้กันมากในหมู่นักวิจัย ทั้งนี้เพราะการเก็บรวบรวมข้อมูลด้วยแบบสอบถามเป็นวิธีที่สะดวกและสามารถใช้วัดได้อย่างกว้างขวาง แบบสอบถามส่วนใหญ่จะอยู่ในรูปของคำถามเป็นชุด ๆ ที่ได้ถูกรวบรวมไว้อย่างมีหลักเกณฑ์และเป็นระบบ เพื่อใช้วัดสิ่งที่ผู้วิจัยต้องการจะวัดจากกลุ่มตัวอย่างหรือประชากรเป้าหมายให้ได้มาซึ่งข้อเท็จจริงทั้งในอดีต ปัจจุบันและการคาดคะเนเหตุการณ์ในอนาคต การเก็บข้อมูลด้วยแบบสอบถามสามารถทำได้ด้วยการสัมภาษณ์หรือให้ผู้ตอบตอบด้วยตนเอง (มารยาท โยทองยศ, 2554) ข้อคำถามในแบบสอบถามอาจแบ่งได้เป็น 2 ประเภท คือ คำถามปลายเปิด (Open Ended Question) และคำถามปลายปิด (Closed Question) ซึ่งคำถามปลายเปิดเป็นคำถามที่เปิดโอกาสให้ผู้ตอบสามารถตอบได้อย่างไม่จำกัดวงคำตอบ ซึ่งคาดว่าน่าจะได้คำตอบที่เป็นอิสระ ไม่แน่นอน ผู้ตอบสามารถตอบได้ตามสภาพความเป็นจริงมากกว่าคำตอบที่จำกัดวงให้ตอบ คำถามปลายเปิดจะนิยมใช้กันมากในกรณีที่ผู้วิจัยไม่สามารถคาดเดาได้ล่วงหน้าว่าคำตอบจะเป็นอย่างไร หรือใช้คำถามปลายเปิดในกรณีที่ต้องการได้คำตอบเพื่อนำมาเป็นแนวทางในการสร้างคำถามปลายปิด แบบสอบถามแบบนี้มีข้อเสีย คือ การรวบรวมความคิดเห็น และการแปลผลมีความยุ่งยาก สำหรับคำถามปลายปิดเป็นคำถามที่ผู้วิจัยมีแนวคำตอบไว้ให้ผู้ตอบเลือกตอบจากคำตอบที่กำหนดไว้เท่านั้น คำตอบที่ผู้วิจัยกำหนดไว้ล่วงหน้ามักได้มาจากการทดลองใช้คำถามในลักษณะที่เป็นคำถามปลายเปิด หรือการศึกษากรอบแนวความคิด สมมติฐานการวิจัย และนิยามเชิงปฏิบัติการ คำถามปลายปิดมีวิธีการเขียนได้หลาย ๆ แบบ เช่น แบบให้เลือกตอบอย่างใดอย่างหนึ่ง แบบให้เลือกคำตอบที่ถูกต้องเพียงคำตอบเดียว แบบผู้ตอบจัดลำดับความสำคัญหรือแบบให้เลือกคำตอบหลายคำตอบ (พิชิต ฤทธิ์จรูญ, 2544) การวิเคราะห์ผลทางสถิติในส่วนข้อคิดเห็นหรือข้อเสนอแนะของผู้ตอบแบบสอบถามแบบปลายเปิด เป็นเรื่องยากและต้องใช้เวลามาก โดยเฉพาะในกรณีที่มีการเก็บข้อมูลแบบสำรวจเป็นจำนวนมาก ทำให้ผู้วิเคราะห์ผลไม่นำข้อมูลดังกล่าวมาประมวลผล ซึ่งข้อมูลดังกล่าวมีความสำคัญและเป็นประโยชน์ต่อการวิจัย จึงได้ประยุกต์ใช้การจำแนกหมวดหมู่เอกสารแบบอัตโนมัติ (Automatic Text Classification) กับข้อความภาษาไทยที่เป็นข้อคิดเห็นในแบบสอบถามปลายเปิด

การจำแนกหมวดหมู่เอกสารแบบอัตโนมัติเป็นการนำวิธีการเรียนรู้ด้วยคอมพิวเตอร์ (Machine Learning) ประยุกต์ร่วมกับการประมวลผลภาษาธรรมชาติ การจัดแบ่งกลุ่มเอกสารแบบอัตโนมัติ สามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization) การจัดกลุ่มเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสาร โดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ส่วนการจำแนกหมวดหมู่เอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสาร โดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด โดยผลลัพธ์ที่ได้จากวิธีการเรียนรู้ด้วยคอมพิวเตอร์นั้นความถูกต้องใกล้เคียงกับผลการจำแนกหมวดหมู่ของเอกสารที่ทำโดยมนุษย์ ทำให้ประหยัดแรงงานมนุษย์เป็นอย่างมากเพราะไม่ต้องอาศัยผู้เชี่ยวชาญในการจำแนกประเภทเอกสารหรือปรับเปลี่ยนหมวดหมู่ของเอกสาร (Sebastiani, 2002: 1-47) สำหรับการจำแนกหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ มีงานวิจัยหลายงานที่ได้ศึกษาทดลองจำแนกหมวดหมู่เอกสารภาษาไทยโดยใช้อัลกอริทึมในการเรียนรู้ที่แตกต่างกัน โดยอัลกอริทึมหรือวิธีการที่มีการนำมาใช้กับเอกสารภาษาไทยที่มีประสิทธิภาพดี ได้แก่ เนอีฟเบย์ (Naïve-Bayes) (Kruengkrai and Jaruskulchai, 2001) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) (วัลลภ อินทร์ฉ่ำ, 2548)

จากปัญหาดังกล่าว จึงได้ทำการจำแนกหมวดหมู่ข้อความแสดงข้อคิดเห็นในแบบสอบถามปลายเปิดโดยอัตโนมัติ เพื่อใช้สำหรับการวิเคราะห์ผลทางสถิติ โดยใช้วิธีเนอีฟเบย์เปรียบเทียบประสิทธิภาพกับซัพพอร์ทเวกเตอร์แมชชีน

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด

1.2.2 เพื่อทดสอบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด ด้วยวิธีเนอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน

1.3 สมมติฐานการวิจัย

การจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพดีกว่าเนอีฟเบย์

1.4 ขอบเขตการวิจัย

1.4.1 วิธีที่ใช้ในการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด คือ เนอีฟเบย์และซัพพอร์ทเวคเตอร์แมชชีน

1.4.2 ข้อมูลที่ใช้ในงานวิจัยเป็นข้อมูลที่ได้จากแบบสอบถาม เรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร

1.4.3 การเรียนรู้แบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลสำหรับการเรียนรู้ 80% และข้อมูลสำหรับการทดสอบ 20%

1.4.4 การวิจัยได้แบ่งกลุ่มของการจำแนกหมวดหมู่ข้อความแสดงความคิดเห็น โดยจะแบ่ง 3 หมวดหมู่ ได้แก่ เชิงบวก (Positive) กลาง (Medium) และลบ (Negative)

1.5 คำจำกัดความที่ใช้ในการวิจัย

1.5.1 การจำแนกหมวดหมู่เอกสาร (Text Classification) หมายถึง เป็นกระบวนการในจัดกลุ่มเอกสารข้อความออกเป็นหมวดหมู่ ซึ่งกระบวนการจัดกลุ่มเอกสารนี้จะมีการกำหนดเป้าหมายชุดของเอกสารเอาไว้ล่วงหน้า

1.5.2 เนอีฟเบย์ (Naïve-Bayes) หมายถึง เทคนิคการจำแนกหมวดหมู่ที่ใช้หลักการคำนวณความน่าจะเป็น (กมลาสน์ อุดมล้ำเลิศ และอานนท์ รุ่งสว่าง, 2553: 312)

1.5.3 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) หมายถึง เทคนิคการจำแนกหมวดหมู่ที่ใช้หลักการสร้างสมการเส้นตรงเพื่อแบ่งกลุ่มข้อมูล 2 กลุ่มออกจากกัน โดย SVM พยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มให้มากที่สุด (พรพล ธรรมรงค์รัตน์, ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์, 2008)

1.5.4 แบบสอบถามปลายเปิด (Open Ended Questionnaire) หมายถึง เป็นคำถามที่ไม่ได้กำหนดคำตอบไว้ และเปิดโอกาสให้ผู้ตอบมีอิสระในการตอบด้วยความคิดของตัวเอง ข้อเสียของแบบสอบถามนี้คือ ตอบยากและเสียเวลาในการตอบมาก และต้องใช้การคิดวิเคราะห์อย่างกว้างขวาง (เยาวเรศ จันทะแสน, 2554)

1.5.5 ข้อมูลสำหรับการเรียนรู้ หมายถึง ข้อมูลที่ต้องใช้สำหรับนำข้อมูลเข้าสู่อัลกอริทึมหรือโมเดล เพื่อให้อัลกอริทึมหรือโมเดลเกิดการเรียนรู้และจดจำข้อมูลก่อนการนำข้อมูลมาทดสอบ

1.5.6 ข้อมูลสำหรับการทดสอบ หมายถึง ข้อมูลที่ใช้ในการทดสอบอัลกอริทึมหรือโมเดลเพื่อทดสอบความถูกต้องของการจำแนกข้อมูล

1.5.7 ข้อคิดเห็น หรือความคิดเห็น หมายถึง อารมณ์ ความรู้สึก ความคิด และข้อสันนิษฐานที่ผู้พูด หรือผู้เขียนมีต่อสิ่งใดสิ่งหนึ่ง

1.6 ประโยชน์ของงานวิจัย

1.6.1 เพื่อประยุกต์การจำแนกเอกสารแบบอัตโนมัติกับข้อความของการแสดงความคิดเห็นของผู้ตอบแบบสอบถาม ด้วยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน

1.6.2 ได้วิธีที่มีประสิทธิภาพในการวิเคราะห์ข้อความแสดงความคิดเห็นของผู้ตอบแบบสอบถามปลายเปิด

1.6.3 แก้ปัญหาความยุ่งยากในการวิเคราะห์ข้อมูลของการแสดงข้อคิดเห็นในแบบสอบถามปลายเปิด

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน มีวัตถุประสงค์เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นในแบบสอบถามปลายเปิด และทดสอบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด ด้วยวิธีเนอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน มีเอกสารและงานวิจัยที่เกี่ยวข้องประกอบด้วยหัวข้อดังต่อไปนี้

2.1 แบบสอบถาม (Questionnaire)

2.2 กระบวนการเตรียมข้อมูล (Preprocessing)

2.3 การเรียนรู้และจำแนกหมวดหมู่

2.4 ซัพพอร์ตเวคเตอร์แมชชีน

2.5 เนอีฟเบย์

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 แบบสอบถาม (Questionnaire)

แบบสอบถาม หมายถึง รูปแบบของคำถามเป็นชุด ๆ ที่ได้ถูกรวบรวมไว้อย่างมีหลักเกณฑ์และเป็นระบบ เพื่อใช้วัดสิ่งที่ผู้วิจัยต้องการจะวัดจากกลุ่มตัวอย่างหรือประชากรเป้าหมายให้ได้มาซึ่งข้อเท็จจริงทั้งในอดีต ปัจจุบันและการคาดคะเนเหตุการณ์ในอนาคต แบบสอบถามประกอบด้วยรายการคำถามที่สร้างอย่างประณีต เพื่อรวบรวมข้อมูลเกี่ยวกับความคิดเห็นหรือข้อเท็จจริง โดยส่งให้กลุ่มตัวอย่างตามความสมัครใจ การใช้แบบสอบถามเป็นเครื่องมือในการเก็บรวบรวมข้อมูลนั้น การสร้างคำถามเป็นงานที่สำคัญสำหรับผู้วิจัย เพราะว่าผู้วิจัยอาจไม่มีโอกาสได้พบปะกับผู้ตอบแบบสอบถามเพื่ออธิบายความหมายต่าง ๆ ของข้อคำถามที่ต้องการเก็บรวบรวม

แบบสอบถาม เป็นเครื่องมือวิจัยชนิดหนึ่งที่นิยมใช้กันมาก เพราะการเก็บรวบรวมข้อมูลสะดวกและสามารถใช้วัดได้อย่างกว้างขวาง การเก็บข้อมูลด้วยแบบสอบถามสามารถทำได้ด้วยการสัมภาษณ์หรือให้ผู้ตอบด้วยตนเอง

2.1.1 โครงสร้างของแบบสอบถาม (มารยาท โยทองยศ, 2554)

โครงสร้างของแบบสอบถาม ประกอบไปด้วย 3 ส่วนสำคัญ ดังนี้

2.1.1.1 หนังสือนำหรือคำชี้แจง โดยมากมักจะอยู่ส่วนแรกของแบบสอบถาม อาจมีจดหมายนำอยู่ด้านหน้าพร้อมคำขอบคุณ โดยคำชี้แจงมักจะระบุถึงจุดประสงค์ที่ให้ตอบแบบสอบถาม การนำคำตอบที่ได้ไปใช้ประโยชน์ คำอธิบายลักษณะของแบบสอบถาม วิธีการตอบแบบสอบถามพร้อมตัวอย่าง ชื่อ และที่อยู่ของผู้วิจัย ประเด็นที่สำคัญคือการแสดงข้อความที่ทำให้ผู้ตอบมีความมั่นใจว่า ข้อมูลที่จะตอบไปจะไม่ถูกเปิดเผยเป็นรายบุคคล จะไม่มีผลกระทบต่อผู้ตอบ และมีการพิทักษ์สิทธิของผู้ตอบด้วย

2.1.1.2 คำถามเกี่ยวกับข้อมูลส่วนตัว เช่น เพศ อายุ ระดับการศึกษา อาชีพ เป็นต้น การที่จะถามข้อมูลส่วนตัวอะไรบ้างนั้นขึ้นอยู่กับกรอบแนวความคิดในการวิจัย โดยดูว่าตัวแปรที่สนใจจะศึกษานั้นมีอะไรบ้างที่เกี่ยวกับข้อมูลส่วนตัว และควรถามเฉพาะข้อมูลที่จำเป็นในการวิจัยเท่านั้น

2.1.1.3 คำถามเกี่ยวกับคุณลักษณะหรือตัวแปรที่จะวัด เป็นความคิดเห็นของผู้ตอบในเรื่องของคุณลักษณะ หรือตัวแปรนั้น

2.1.2 ขั้นตอนการสร้างแบบสอบถาม (พิชิต ฤทธิ์จรูญ, 2544)

การสร้างแบบสอบถามประกอบไปด้วยขั้นตอนสำคัญ ดังนี้

ขั้นที่ 1 ศึกษาคุณลักษณะที่จะวัด

การศึกษาคุณลักษณะอาจดูได้จาก วัตถุประสงค์ของการวิจัย กรอบแนวความคิดหรือสมมติฐานการวิจัย จากนั้นจึงศึกษาคุณลักษณะ หรือตัวแปรที่จะวัดให้เข้าใจอย่างละเอียดทั้งเชิงทฤษฎีและนิยามเชิงปฏิบัติการ

ขั้นที่ 2 กำหนดประเภทของข้อคำถาม

ข้อคำถามในแบบสอบถามอาจแบ่งได้เป็น 2 ประเภท คือ

1. คำถามปลายเปิด (Open Ended Question) เป็นคำถามที่เปิดโอกาสให้ผู้ตอบสามารถตอบได้อย่างไม่กำจัดวงคำตอบ ซึ่งคาดว่าน่าจะได้คำตอบที่แน่นอน สมบูรณ์ ตรงกับสภาพความเป็นจริงได้มากกว่าคำตอบที่จำกัดวงให้ตอบ คำถามปลายเปิดจะนิยมใช้กันมากในกรณีที่ผู้วิจัยไม่สามารถคาดเดาได้ล่วงหน้าว่าคำตอบจะเป็นอย่างไร หรือใช้คำถามปลายเปิดในกรณีที่ต้องการได้คำตอบเพื่อนำมาเป็นแนวทางในการสร้างคำถามปลายปิด แบบสอบถามแบบนี้มีข้อเสียคือ มักจะถามได้ไม่มากนัก การรวบรวมความคิดเห็นและการแปลผลมักจะมีความยุ่งยาก

2. คำถามปลายปิด (Closed Question) เป็นคำถามที่ผู้วิจัยมีแนวคำตอบไว้ให้

ผู้ตอบเลือกตอบจากคำตอบที่กำหนดไว้เท่านั้น คำตอบที่ผู้วิจัยกำหนดไว้ล่วงหน้ามักได้มาจากการทดลองใช้คำถามในลักษณะที่เป็นคำถามปลายเปิด หรือการศึกษากรอบแนวความคิดสมมติฐานการวิจัย และนิยามเชิงปฏิบัติการ คำถามปลายเปิดมีวิธีการเขียนได้หลาย ๆ แบบ เช่น

แบบให้เลือกตอบอย่างใดอย่างหนึ่ง แบบให้เลือกคำตอบที่ถูกต้องเพียงคำตอบเดียว แบบผู้ตอบจัดลำดับความสำคัญหรือแบบให้เลือกคำตอบหายคำตอบ

ขั้นที่ 3 การร่างแบบสอบถาม

เมื่อผู้วิจัยทราบถึงคุณลักษณะหรือประเด็นที่จะวัด และกำหนดประเภทของข้อคำถามที่จะมีอยู่ในแบบสอบถามเรียบร้อยแล้ว ผู้วิจัยจึงลงมือเขียนข้อคำถามให้ครอบคลุมทุกคุณลักษณะหรือประเด็นที่จะวัด โดยเขียนตามโครงสร้างของแบบสอบถามที่ได้กล่าวไว้แล้ว และหลักการในการสร้างแบบสอบถาม ดังนี้

1. ต้องมีจุดมุ่งหมายที่แน่นอนว่าต้องการจะถามอะไรบ้าง โดยจุดมุ่งหมายนั้นจะต้องสอดคล้องกับวัตถุประสงค์ของงานวิจัยที่จะทำ

2. ต้องสร้างคำถามให้ตรงตามจุดมุ่งหมายที่ตั้งไว้ เพื่อป้องกันการมีข้อคำถามนอกประเด็นและมีข้อคำถามจำนวนมาก

3. ต้องถามให้ครอบคลุมเรื่องที่จะวัด โดยมีจำนวนข้อคำถามที่พอเหมาะ ไม่มากหรือน้อยเกินไป แต่จะมากหรือน้อยเท่าใดนั้นขึ้นอยู่กับพฤติกรรมที่จะวัด ซึ่งตามปกติพฤติกรรมหรือเรื่องที่จะวัดเรื่องหนึ่งๆ นั้นควรมีข้อคำถาม 25-60 ข้อ

4. การเรียงลำดับข้อคำถาม ควรเรียงลำดับให้ต่อเนื่องสัมพันธ์กัน และแบ่งตามพฤติกรรมย่อยๆ ไว้เพื่อให้ผู้ตอบเห็นชัดเจนและง่ายต่อการตอบ นอกจากนั้นต้องเรียงคำถามง่ายๆ ไว้เป็นข้อแรกๆ เพื่อชักจูงให้ผู้ตอบอยากตอบคำถามต่อ ส่วนคำถามสำคัญๆ ไม่ควรเรียงไว้ตอนท้ายของแบบสอบถาม เพราะความสนใจในการตอบของผู้ตอบอาจจะน้อยลง ทำให้ตอบอย่างไม่ตั้งใจ ซึ่งจะส่งผลเสียต่อการวิจัยมาก

5. ลักษณะของข้อความที่ดี ข้อคำถามที่ดีของแบบสอบถามนั้น ควรมีลักษณะดังนี้

1) ข้อคำถามไม่ควรยาวจนเกินไป ควรใช้ข้อความสั้น กะทัดรัด ตรงกับวัตถุประสงค์และสองคล้องกับเรื่อง

2) ข้อความ หรือภาษาที่ใช้ในข้อความต้องชัดเจน เข้าใจง่าย

3) ค่าเฉลี่ยในการตอบแบบสอบถามไม่ควรเกินหนึ่งชั่วโมง ข้อคำถามไม่ควรมากเกินไปจนทำให้ผู้ตอบเบื่อหน่ายหรือเหนื่อยล้า

4) ไม่ถามเรื่องที่เป็นความลับเพราะจะทำให้ได้คำตอบที่ไม่ตรงกับข้อเท็จจริง

5) ไม่ควรใช้ข้อความที่มีความหมายกำกวมหรือข้อความที่ทำให้ผู้ตอบแต่ละคนเข้าใจความหมายของข้อความไม่เหมือนกัน

6) ไม่ถามในเรื่องที่รู้แล้ว หรือถามในสิ่งที่วัดได้ด้วยวิธีอื่น

7) ข้อคำถามต้องเหมาะสมกับกลุ่มตัวอย่าง คือ ต้องคำนึงถึงระดับการศึกษา ความ

สนใจ สภาพเศรษฐกิจ ฯลฯ

8) ข้อคำถามหนึ่งๆ ควรถามเพียงประเด็นเดียว เพื่อให้ได้คำตอบที่ชัดเจนและตรงจุดซึ่งจะง่ายต่อการนำมาวิเคราะห์ข้อมูล

9) คำตอบหรือตัวเลือกในข้อคำถามควรมีมากพอ หรือให้เหมาะสมกับข้อคำถามนั้น แต่ถ้าไม่สามารถระบุได้หมดก็ให้ใช้ว่า อื่นๆ โปรดระบุ .....

10) ควรหลีกเลี่ยงคำถามที่เกี่ยวกับค่านิยมที่จะทำให้ผู้ตอบไม่ตอบตามความเป็นจริง

11) คำตอบที่ได้จากแบบสอบถาม ต้องสามารถนำมาแปลงออกมาในรูปของปริมาณและใช้สถิติอธิบายข้อเท็จจริงได้ เพราะปัจจุบันนี้นิยมใช้คอมพิวเตอร์ในการวิเคราะห์ข้อมูล ดังนั้นแบบสอบถามควรคำนึงถึงวิธีการประมวลข้อมูลและวิเคราะห์ข้อมูลด้วยโปรแกรมคอมพิวเตอร์ด้วย

ขั้นที่ 4 การปรับปรุงแบบสอบถาม

หลังจากที่สร้างแบบสอบถามเสร็จแล้ว ผู้วิจัยควรนำแบบสอบถามนั้นมาพิจารณาทบทวนอีกครั้งเพื่อหาข้อบกพร่องที่ควรปรับปรุงแก้ไข และควรให้ผู้เชี่ยวชาญได้ตรวจสอบแบบสอบถามนั้นด้วยเพื่อที่จะได้นำข้อเสนอแนะและข้อวิพากษ์วิจารณ์ของผู้เชี่ยวชาญมาปรับปรุงแก้ไขให้ดียิ่งขึ้น

ขั้นที่ 5 วิเคราะห์คุณภาพแบบสอบถาม

เป็นการนำแบบสอบถามที่ได้ปรับปรุงแล้วไปทดลองใช้กับกลุ่มตัวอย่างเล็กๆ เพื่อนำผลมาตรวจสอบคุณภาพของแบบสอบถาม ซึ่งการวิเคราะห์หรือตรวจสอบคุณภาพของแบบสอบถามทำได้หลายวิธี แต่ที่สำคัญมี 2 วิธี ได้แก่

1. ความตรง (Validity) หมายถึง เครื่องมือที่สามารถวัดได้ในสิ่งที่ต้องการวัด โดยแบ่งออกได้เป็น 3 ประเภท คือ

1) ความตรงตามเนื้อหา (Content Validity) คือ การที่แบบสอบถามมีความครอบคลุมวัตถุประสงค์หรือพฤติกรรมที่ต้องการวัดหรือไม่ ค่าสถิติที่ใช้ในการหาคุณภาพ คือ ค่าความสอดคล้องระหว่างข้อคำถามกับวัตถุประสงค์ หรือเนื้อหา (IOC: Index of item Objective Congruence) หรือดัชนีความเหมาะสม โดยให้ผู้เชี่ยวชาญ ประเมินเนื้อหาของข้อถามเป็นรายข้อ

2) ความตรงตามเกณฑ์ (Criterion-related Validity) หมายถึง ความสามารถของแบบวัดที่สามารถวัดได้ตรงตามสภาพความเป็นจริง แบ่งออกได้เป็นความเที่ยงตรงเชิงพยากรณ์และความเที่ยงตรงตามสภาพ สถิติที่ใช้วัดความเที่ยงตรงตามเกณฑ์ เช่น ค่าสัมประสิทธิ์สหสัมพันธ์

(Correlation Coefficient) ทั้งของเพียร์สัน (Pearson) และสเพียร์เมน (Spearman) และการทดสอบค่าที (t-test) เป็นต้น

3) ความตรงตามโครงสร้าง (Construct Validity) หมายถึงความสามารถของแบบสอบถามที่สามารถวัดได้ตรงตามโครงสร้างหรือทฤษฎี ซึ่งมักจะมีในแบบวัดทางจิตวิทยาและแบบวัดสติปัญญา สถิติที่ใช้วัดความเที่ยงตรงตามโครงสร้างมีหลายวิธี เช่น การวิเคราะห์องค์ประกอบ (Factor Analysis) การตรวจสอบในเชิงเหตุผล เป็นต้น

2. ความเที่ยง (Reliability) หมายถึง เครื่องมือที่มีความคงเส้นคงวา นั่นคือ เครื่องมือที่สร้างขึ้นให้ผลการวัดที่แน่นอนคงที่จะวัดกี่ครั้งผลจะได้เหมือนเดิม สถิติที่ใช้ในการหาค่าความเที่ยงมีหลายวิธีแต่นิยมใช้กันคือ ค่าสัมประสิทธิ์แอลฟ่าของคอนบาร์ช (Conbach's Alpha Coefficient: α coefficient) ซึ่งจะใช้สำหรับข้อมูลที่มีการแบ่งระดับการวัดแบบประมาณค่า (Rating Scale)

ขั้นที่ 6 ปรับปรุงแบบสอบถามให้สมบูรณ์

ผู้วิจัยจะต้องทำการแก้ไขข้อบกพร่องที่ได้จากผลการวิเคราะห์คุณภาพของแบบสอบถามและตรวจสอบความถูกต้องของถ้อยคำหรือสำนวน เพื่อให้แบบสอบถามมีความสมบูรณ์และมีคุณภาพผู้ตอบอ่านเข้าใจได้ตรงประเด็นที่ผู้วิจัยต้องการ ซึ่งจะทำให้ผลงานวิจัยเป็นที่น่าเชื่อถือยิ่งขึ้น

ขั้นที่ 7 จัดพิมพ์แบบสอบถาม

จัดพิมพ์แบบสอบถามที่ได้ปรับปรุงเรียบร้อยแล้วเพื่อนำไปใช้จริงในการเก็บรวบรวมข้อมูลกับกลุ่มเป้าหมาย โดยจำนวนที่จัดพิมพ์ควรไม่น้อยกว่าจำนวนเป้าหมายที่ต้องการเก็บรวบรวมข้อมูล และควรมีการพิมพ์สำรองไว้ในกรณีที่แบบสอบถามเสียหรือสูญหายหรือผู้ตอบไม่ตอบกลับ แนวทางในการจัดพิมพ์แบบสอบถามมีดังนี้

1) การพิมพ์แบ่งหน้าให้สะดวกต่อการเปิดอ่านและตอบ
2) เว้นที่ว่างสำหรับคำถามปลายเปิดไว้เพียงพอ
3) พิมพ์อักษรขนาดใหญ่ชัดเจน
4) ใช้สีและลักษณะกระดาษที่เอื้อต่อการอ่าน

2.1.3 หลักการสร้างแบบสอบถาม

1) สอดคล้องกับวัตถุประสงค์การวิจัย
2) ใช้ภาษาที่เข้าใจง่าย เหมาะสมกับผู้ตอบ
3) ใช้ข้อความที่สั้น กระทัดรัด ได้ใจความ
4) แต่ละคำถามควรมีนัย เพียงประเด็นเดียว
5) หลีกเลี่ยงการใช้ประโยคปฏิเสธซ้อน

6) ไม่ควรใช้คำย่อ

7) หลีกเลี่ยงการใช้คำที่เป็นนามธรรมมาก

8) ไม่ชี้นำการตอบให้เป็นไปแนวทางใดแนวทางหนึ่ง

9) หลีดเลี่ยงคำถามที่ทำให้ผู้ตอบเกิดความลำบากใจในการตอบ

10) คำตอบที่มีให้เลือกต้องชัดเจนและครอบคลุมคำตอบที่เป็นไปได้

11) หลีกเลี่ยงคำที่สื่อความหมายหลายอย่าง

12) ไม่ควรเป็นแบบสอบถามที่มีจำนวนมากเกินไป ไม่ควรให้ผู้ตอบใช้เวลาในการตอบแบบสอบถามนานเกินไป

13) ข้อคำถามควรถามประเด็นที่เฉพาะเจาะจงตามเป้าหมายของการวิจัย

14) คำถามต้องน่าสนใจสามารถกระตุ้นให้เกิดความอยากตอบ

2.1.4 เทคนิคการใช้แบบสอบถาม (เกียรติสุดา ศรีสุข, 2552; จินตนา ธนวิบูลย์ชัย, 2545; อุทุมพร จามรมาน, 2544; Neil J. Salkind, 2006)

วิธีใช้แบบสอบถามมี 2 วิธี คือ การส่งทางไปรษณีย์กับการเก็บข้อมูลด้วยตนเอง ซึ่งไม่ว่ากรณีใดต้องมีจดหมายระบุวัตถุประสงค์ของการเก็บข้อมูล ตลอดจนความสำคัญของข้อมูลและผลที่คาดว่าจะได้รับ เพื่อให้ผู้ตอบตระหนักถึงความสำคัญและสละเวลาในการตอบแบบสอบถาม

การทำให้อัตราตอบแบบสอบถามสูงเป็นเป้าหมายสำคัญของผู้วิจัย ข้อมูลจากแบบสอบถามจะเป็นตัวแทนของประชากรได้เมื่อมีจำนวนแบบสอบถามคืนมากว่าร้อยละ 90 ของจำนวนแบบสอบถามที่ส่งไป แนวทางที่จะทำให้ได้รับแบบสอบถามกลับคืนในอัตราที่สูง มีวิธีการดังนี้

1. มีการติดตามแบบสอบถามเมื่อให้เวลาผู้ตอบไประยะหนึ่ง ระยะเวลาที่เหมาะสมในการติดตามคือ 2 สัปดาห์ หลังครบกำหนดส่ง อาจจะติดตามมากกว่าหนึ่งครั้ง

2. วิธีการติดตามแบบสอบถาม อาจใช้จดหมาย ไปรษณีย์ โทรศัพท์ เป็นต้น

3. ในกรณีที่ขอคำถามอาจจะถามในเรื่องของส่วนตัว ผู้วิจัยต้องให้ความมั่นใจว่าข้อมูลที่ได้จะเป็นความลับ

2.1.5 ข้อเด่นและข้อด้อยของการเก็บข้อมูลโดยใช้แบบสอบถาม

การใช้แบบสอบถามในการเก็บรวบรวมข้อมูลมีข้อเด่นและข้อด้อยที่ต้องพิจารณาประกอบในการเลือกใช้แบบสอบถามในการเก็บรวบรวมข้อมูล ดังนี้

ข้อเด่นของการเก็บข้อมูลโดยใช้แบบสอบถามมีดังนี้ คือ

1. ถ้ากลุ่มตัวอย่างมีขนาดใหญ่ วิธีการเก็บข้อมูลโดยใช้แบบสอบถาม จะเป็นวิธีการที่สะดวกและประหยัดกว่าวิธีอื่น

2. ผู้ตอบมีเวลาตอบมากกว่าวิธีการอื่น

3. ไม่จำเป็นต้องฝึกอบรมพนักงานเก็บข้อมูลมากเหมือนกับวิธีการสัมภาษณ์หรือวิธีการสังเกต

4. ไม่เกิดความลำเอียงอันเนื่องมาจากการสัมภาษณ์หรือการสังเกต เพราะผู้ตอบเป็นผู้ตอบข้อมูลเอง

5. สามารถส่งแบบสอบถามให้ผู้ตอบทางไปรษณีย์ได้

6. ประหยัดค่าใช้จ่ายในการเก็บข้อมูล

ข้อด้อยของการเก็บข้อมูลโดยใช้แบบสอบถาม มีดังนี้คือ

1. ในกรณีที่ส่งแบบสอบถามให้ผู้ตอบทางไปรษณีย์ มักจะได้แบบสอบถามกลับคืนมาน้อย และต้องเสียเวลาในการติดตาม อาจทำให้ระยะเวลาการเก็บข้อมูลล่าช้ากว่าที่กำหนดไว้

2. การเก็บข้อมูลโดยวิธีการใช้แบบสอบถามจะใช้ได้เฉพาะกับกลุ่มประชากรเป้าหมายที่อ่านและเขียนหนังสือได้เท่านั้น

3. จะได้ข้อมูลจำกัดเฉพาะที่จำเป็นจริงๆ เท่านั้น เพราะการเก็บข้อมูลโดยวิธีการใช้แบบสอบถามจะต้องมีคำถามจำนวนน้อยข้อที่สุดเท่าที่จะเป็นไปได้

4. การส่งแบบสอบถามไปทางไปรษณีย์ หน่วยตัวอย่างอาจไม่ได้เป็นผู้ตอบแบบสอบถามเองก็ได้ ทำให้คำตอบที่ได้มีความคลาดเคลื่อนไม่ตรงกับความจริง

5. ถ้าผู้ตอบไม่เข้าใจคำถามหรือเข้าใจคำถามผิด หรือไม่ตอบคำถามบางข้อ หรือไม่ไตร่ตรองให้รอบคอบก่อนที่จะตอบคำถาม ก็จะทำให้ข้อมูลมีความคลาดเคลื่อนได้ โดยที่ผู้วิจัยไม่สามารถย้อนกลับไปสอบถามหน่วยตัวอย่างนั้นได้อีก

6. ผู้ที่ตอบแบบสอบถามกลับคืนมาทางไปรษณีย์ อาจเป็นกลุ่มที่มีลักษณะแตกต่างจากกลุ่มผู้ที่ไม่ตอบแบบสอบถามกลับคืนมา ดังนั้นข้อมูลที่นำมาวิเคราะห์จะมีความลำเอียงอันเนื่องมาจากกลุ่มตัวอย่างได้

2.2 กระบวนการเตรียมข้อมูล (Preprocessing)

กระบวนการเตรียมข้อมูล เป็นการเตรียมข้อมูลก่อนทำการเรียนรู้และจำแนกหมวดหมู่ประกอบด้วย

2.2.1 การตัดคำ (Word Segmentation)

การประมวลผลจำแนกหมวดหมู่เอกสารภาษาไทยได้อย่างมีประสิทธิภาพนั้น มีปัญหาเบื้องต้นคือ การตัดคำในภาษาไทย ซึ่งลักษณะการเขียนภาษาไทยจะมีการเขียนติดต่อกันเป็นสายอักขระโดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ดังเช่นภาษาอังกฤษซึ่งใช้ช่องว่าง (Space) คั่นระหว่างคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อแบ่งสายอักขระไทยออกเป็นคำ ๆ จึงได้มีผู้คิดค้น

พัฒนาวิธีการตัดคำแบ่งได้เป็น หลักการตัดคำโดยใช้กฎ (Rule Base Approach) หลักการตัดคำโดยใช้อัลกอริทึม (Algorithm Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus Base Approach)แต่ละวิธีการต่างๆ ก็ให้ผลในด้านความถูกต้อง ความรวดเร็วของการทำงานและปริมาณการใช้ทรัพยากรต่าง ๆ ที่แตกต่างกัน จากการศึกษาเรื่องตัดคำสำหรับการจัดหมวดหมู่เอกสารภาษาไทยพบปัญหาด้านการหาขอบเขตของคำ เนื่องจากไม่มีการเขียนแบ่งพยางค์ คำ หรือประโยค ไม่มีหลักเกณฑ์ตายตัวในการใช้ช่องว่างในภาษาเขียน การสะกดคำมีรูปแบบซับซ้อน มีคำยืม คำทับศัพท์ คำเฉพาะจำนวนมาก และคำมีความกำกวมสูง จากการศึกษาเปรียบเทียบประสิทธิภาพวิธีดังกล่าว พบว่าวิธีตัดคำที่เหมาะสมกับการจัดหมวดหมู่เอกสารคือวิธีการตัดคำแบบยาวที่สุด (Longest Matching) ซึ่งมีวิธีการตรวจสอบสายอักขระ (String) ที่เข้ามาจากซ้ายไปขวากับพยางค์ที่เก็บไว้ในพจนานกรม ในกรณีที่ตรวจสอบแล้วปรากฏว่าพบพยางค์มากกว่า 1 พยางค์ในพจนานุกรม ก็ให้เลือกแบ่งพยางค์โดยเลือกพยางค์ที่ยาวที่สุด แล้วทำต่อไปเรื่อย ๆ จนจบสายอักขระ แต่ถ้ากรณีที่เลือกพยางค์ที่ยาวที่สุดแล้วทำให้เกิดพยางค์ที่ไม่ปรากฏในพจนานุกรมก็ยอมให้มีการย้อนรอย (Back Tracking) กลับไปเลือกพยางค์ที่ยาวรองมาแทนทำต่อไปเรื่อยๆจนสิ้นสุดสายอักขระ (Charoenpornsawat, 1999)

2.2.2 การกำจัดคำหยุด (Stop-Word List Removal)

การกำจัดคำหยุด เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึง คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ประเภทของคำที่จัดว่าเป็นคำหยุดในภาษาไทย (Jaruskulchai, 1998) มีดังต่อไปนี้

1) คำบุพบท

เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน มักใช้นำหน้าคำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อแสดงความสัมพันธ์ของคำหรือกลุ่มคำที่อยู่หลังคำบุพบทว่ามีความสัมพันธ์กับคำหรือกลุ่มคำที่อยู่หน้าคำบุพบทอย่างไร ตัวอย่างของคำบุพบท ได้แก่ ของ ใน แก่ แด่ ต่อ ตั้งแต่ โดย เมื่อ กว่า กับ เป็น ดูก่อน ช้าแต่ ทาง สู่ แค่ ฯลฯ

2) คำสันธาน

เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยค ข้อความกับข้อความ เพื่อทำให้ประโยคนั้นกระชับ และสละสลวย โดยมักจะใช้เชื่อมใจความที่คล้อยตามกัน ใจความที่ขัดแย้งกันใจความที่เป็นเหตุเป็นผลกัน หรือใจความที่ให้เลือกอย่างใดอย่างหนึ่ง ตัวอย่างของคำสันธานได้แก่ เพราะ เพราะว่า และ หรือ จึง ดังนั้น มิฉะนั้น ทั้ง แต่ แต่ว่า ครั้น หรือไม่ก็ ฯลฯ

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

4) คำวิเศษณ์

เป็นคำที่ใช้ขยายคำอื่น เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อให้มีความหมายชัดเจนและมีรายละเอียดมากยิ่งขึ้น ตัวอย่างของคำวิเศษณ์ ได้แก่ มาก น้อย ใหญ่ เล็กมหึมา 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

5) คำอุทาน

เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด รวมทั้งเป็นคำที่ใช้เสริมคำพูดตัวอย่างของคำอุทานได้แก่ เอ๊ะ อ๊ะ อ๋อ เฮ้อ ว้าย โธ่ โถ อนิจจา สิ หนอย ซะ นะ น่า แหม ตายล่ะ คุณพระช่วย ชิชะ อุวะ ไม่น่าเลย โอ้โห ฯลฯ คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเอกสาร และปรากฏในเอกสารเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนก หมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล

2.2.3 การหารากศัพท์

รากศัพท์ (Stem) คือ รูปเดิมของคำที่ยังไม่ได้เติมคำอุปสรรค (Prefixes) คำปัจจัย (Suffixes) หรือยังไม่ได้เปลี่ยนรูปไป ตัวอย่างรากศัพท์ของคำแสดงในภาพที่ 2-1 และภาพที่ 2-2

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

ภาพที่ 2-1 ตัวอย่างรากศัพท์คำภาษาอังกฤษ

คำศัพท์	รากศัพท์
พิจารณา	พิจาร
พิจารณ์	พิจาร
การพิจารณา	พิจาร
พิจาร	พิจาร

ภาพที่ 2-2 ตัวอย่างรากศัพท์คำภาษาไทย

การหารากศัพท์ จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกันเพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่

การหารากศัพท์ของคำภาษาอังกฤษมีขั้นตอนวิธีที่แน่นอนซึ่งสามารถเขียนเป็นอัลกอริทึมในการสกัดคำอุปสรรคและคำปัจจัยได้โดยไม่ต้องเทียบกับรายการคำศัพท์หรือคลังคำ เนื่องมาจากไวยากรณ์ของภาษาอังกฤษมีกฎเกณฑ์ที่แน่นอน ไม่ค่อยมีข้อยกเว้นมากนัก จึงมีความซับซ้อนน้อย ดังนั้นจึงมีผู้สร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้หลายแบบ เช่น อัลกอริทึมปีเตอร์ (Porter Algorithm), อัลกอริทึมแลนแคสเตอร์ (Lancaster Algorithm) เป็นต้น

การหารากศัพท์ของคำภาษาไทยนั้น เท่าที่ศึกษามายังไม่พบว่ามีผู้คิดค้นสร้างอัลกอริทึมสำหรับการหารากศัพท์ไว้ เนื่องมาจากไวยากรณ์ภาษาไทยมีความซับซ้อนมาก และมีข้อยกเว้นหลายประการ การหารากศัพท์ของคำภาษาไทยจึงอาจใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกันหรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์

2.2.4 การทำดัชนี

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบ ที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือการคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (Term Weighting) การสร้างดัชนี โดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม (Sebastiani, 2002; Marquez. 2000; Aas & Eikvil, 1999)

2.3 การเรียนรู้และจำแนกหมวดหมู่

2.3.1 การเรียนรู้ด้วยคอมพิวเตอร์ เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligent) ที่เกี่ยวข้องกับการออกแบบและพัฒนาอัลกอริทึมและวิธีการที่จะทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ การเรียนรู้ด้วยคอมพิวเตอร์เชิงอุปนัย เป็นการค้นหากฎ ลักษณะแบบแผน หรือข้อสรุปต่าง ๆจากการสังเกตกลุ่มข้อมูลขนาดใหญ่ ส่วนการเรียนรู้เชิงอนุมาน เป็นการหาข้อสรุปจากหลักฐานหรือข้อเท็จจริงที่มีอยู่ หลักสำคัญของการวิจัยทางด้านการเรียนรู้ด้วยคอมพิวเตอร์ คือ การสกัดเอาความรู้หรือสารสนเทศจากข้อมูลโดยอัตโนมัติด้วยวิธีการคำนวณหรือวิธีการทางสถิติ ดังนั้นการเรียนรู้ด้วยคอมพิวเตอร์นั้นจึงมีความสัมพันธ์อย่างใกล้ชิดกับการทำเหมืองข้อมูล (Data Mining) และสถิติอัลกอริทึมสำหรับการเรียนรู้ด้วยคอมพิวเตอร์ถูกจัดให้อยู่ในกลุ่มของวิทยาศาสตร์หรือวิธีการที่เกี่ยวข้องกับการแบ่งแยกประเภท (Taxonomy) ซึ่งมีที่มาจากผลลัพธ์ที่ได้จากอัลกอริทึมประเภทของอัลกอริทึมอาจแบ่งได้ ดังต่อไปนี้

1) การเรียนรู้โดยอาศัยตัวอย่าง (Supervised Learning)

เป็นการเรียนรู้โดยใช้อัลกอริทึมเพื่อสร้างฟังก์ชั่นที่จะทำข้อมูลเข้าให้เป็นผลลัพธ์ที่ต้องการ การจำแนกประเภทเป็นรูปแบบหนึ่งของการเรียนรู้โดยอาศัยตัวอย่าง ตัวเรียนรู้จะต้องศึกษาพฤติกรรมของฟังก์ชั่นที่จะทำการกำหนดเวกเตอร์ให้อยู่ $[X_1,X_2,\ldots,X_N]$ ภายใต้ประเภทใดประเภทหนึ่งจากหลายประเภทโดยสังเกตจากตัวอย่างข้อมูลเข้าและตัวอย่างผลลัพธ์ของฟังก์ชั่น

2) การเรียนรู้โดยไม่อาศัยตัวอย่าง (Unsupervised Learning)

เป็นการเรียนรู้โดยการจำลองแบบของชุดข้อมูลเข้าจากลักษณะเฉพาะของข้อมูล โดยไม่ต้องใช้ตัวอย่างผลลัพธ์ที่ถูกจัดประเภทไว้แล้ว

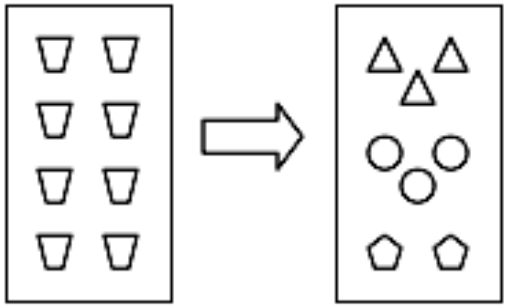
3) การเรียนรู้กึ่งอาศัยตัวอย่าง (Semi-supervised Learning)

เป็นการเรียนรู้ที่อาศัยทั้งตัวอย่างที่ยังไม่ได้จัดประเภทและตัวอย่างที่ถูกจัดประเภทไว้แล้ว เพื่อสร้างฟังก์ชั่นหรือตัวจำแนกประเภทที่เหมาะสมการวิเคราะห์การคำนวณและประสิทธิภาพของอัลกอริทึมที่ใช้สำหรับการเรียนรู้ด้วยคอมพิวเตอร์เป็นแขนงหนึ่งของทฤษฎีทางวิทยาการคอมพิวเตอร์ที่รู้จักกันในชื่อ ทฤษฎีการเรียนรู้เกี่ยวกับการคำนวณ (Computational Learning Theory)

2.3.2 การจำแนกหมวดหมู่

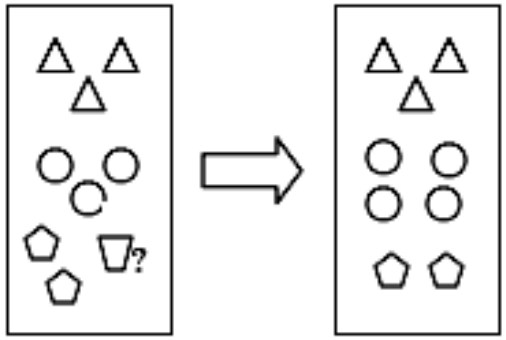
การจำแนกหมวดหมู่สามารถแบ่งได้ 2 ลักษณะ คือ การจัดกลุ่ม (Clustering) และการจำแนกหมวดหมู่ (Classification หรือ Categorization)

การจัดกลุ่มของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยไม่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน ซึ่งจะเป็นการแบ่งกลุ่มตามลักษณะของเอกสาร โดยเอกสารที่มีลักษณะเหมือนกันจะอยู่ด้วยกัน ดังแสดงในภาพที่ 2-3



ภาพที่ 2-3 การจัดกลุ่ม

การจำแนกหมวดหมู่ของเอกสาร คือ การแบ่งกลุ่มตามเนื้อหาของเอกสารโดยที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อน โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่ เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่ต้นแบบมีลักษณะคล้ายกับตัวมันเองมากที่สุด ดังแสดงภาพที่ 2-4



ภาพที่ 2-4 การจำแนกหมวดหมู่

การจำแนกหมวดหมู่เอกสารในภาษาต่างประเทศไม่สามารถนำมาใช้ได้โดยตรงกับการจัดหมวดหมู่ในภาษาไทย เนื่องจากมีปัญหาในการประมวลผลทางด้านภาษาไทยโดยเฉพาะอย่างยิ่งปัญหาในระดับคำ ได้แก่

1. ปัญหาการตัดแบ่งคำ ซึ่งเกิดจากภาษาไทยไม่มีช่องว่างระหว่างคำ ทำให้ไม่สามารถระบุขอบเขตของคำได้อย่างถูกต้อง

2. ปัญหาคำไม่รู้จัก ตัวอย่างเช่น คำศัพท์ด้านเทคนิค คำทับศัพท์ ซึ่งไม่มีการใช้อักษรพิเศษ เพื่อบอกความแต่งแตกระหว่างคำยืมกับคำไทย เช่นเดียวกับภาษาญี่ปุ่นหรือภาษาเกาหลี

3. ไม่มีการใช้เครื่องมือทางภาษา เช่น การใช้อักษรตัวใหญ่ เพื่อบ่งบอกชื่อเฉพาะ เช่นเดียวกับภาษาอังกฤษ

ในการจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติ หากไมมีการแก้ปัญหาในระดับคำจะไม่สามารถกำหนดดัชนีคำหรือวลี เพื่อสร้างกลุ่มต้นแบบในระบบต้นแบบสำหรับจำแนกหมวดหมู่เอกสารอัตโนมัติได้

การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติประกอบด้วย 2 ขั้นตอน ได้แก่ การฝึกฝนระบบสำหรับจัดกลุ่มเอกสารต้นแบบ และการคัดแยกเอกสารด้วยการคำนวณความคล้ายระหว่างเอกสารอินพุตกับกลุ่มเอกสารต้นแบบ สำหรับการฝึกฝนระบบมี 2 วิธี ได้แก่ การเรียนรู้โดยอาศัยตัวอย่าง และการเรียนรู้โดยไม่อาศัยตัวอย่าง

เทคนิคการเรียนรู้โดยอาศัยตัวอย่างมีหลายวิธี เช่น นิวรอนเน็ตเวิร์ก (Neural Network), ริปเปอร์ (Ripper), ร็อกชิโอ (Rocchio) โดยมีการประยุกต์เทคนิคการประมวลผลภาษาธรรมชาติ และปรับโครงสร้างการแทนกลุ่มเอกสาร เทคนิคการประมวลผลภาษาธรรมชาติใช้สำหรับประมวลผลคำและวลี เพื่อใช้เป็นตัวแทนเอกสารที่ใช้ดัชนีคำหรือวลีเพียงอย่างเดียว ส่วนการปรับโครงสร้างการแทนกลุ่มเอกสาร ได้ใช้โครงสร้างแบบมีลำดับชั้นแทนการใช้ระดับเดียว เพราะการค้นหาเอกสารในโครงสร้างแบบมีลำดับชั้นจะตัดกลุ่มเอกสารที่ไม่เกี่ยวข้องออกไป ทำให้สามารถเพิ่มประสิทธิภาพในการค้นหาเอกสารได้เร็วขึ้น

อัลกอริทึมการจัดหมวดหมู่ (Classifier Algorithm) อัลกอริทึมในการจัดหมวดหมู่แบบการเรียนรู้โดยอาศัยตัวอย่าง สามารถแบ่งขั้นตอนวิธีการจัดหมวดหมู่เอกสารแบ่งได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มเอกสารต้นแบบและแยกหมวดหมู่ของเอกสารที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มเอกสารต้นแบบ

2.4 ซัพพอร์ตเวคเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนหรือเอสวีเอ็ม (Support Vector Machine : SVM) จัดเป็นการเรียนรู้ของเครื่องประเภทแบบการเรียนรู้โดยอาศัยตัวอย่างประเภทหนึ่ง ซึ่งมีความสามารถในการจัดหมวดหมู่ และการทำนาย (Regression) โดยเอสวีเอ็มมีพื้นฐานจะมีการคำนวณแบบเชิงเส้น (Linear) ซึ่งจัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ (Decision Hyperplane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน

วิธีการนี้มีชื่อเรียกอีกชื่อหนึ่งว่า จัดหมวดหมู่โดยค่าระยะขอบที่มากที่สุด (Maximum Margin Classifier)

2.4.1 หลักการทำงานของเอสวีเอ็ม ข้อมูลจะถูกเขียนในรูปสมาชิกคู่อันดับดังนี้

$$\{(x_1,c_1),(x_2,c_2),(x_3,c_3),\ldots,(x_n,c_n)\} \quad (2\text{-}1)$$

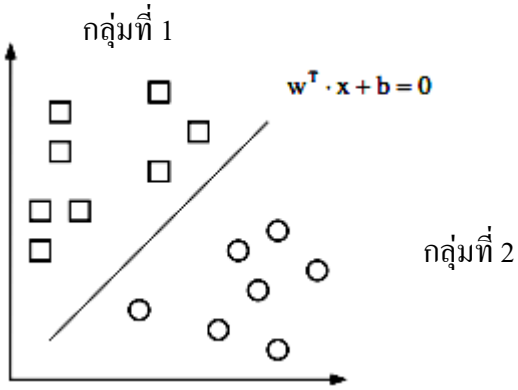
เมื่อ c_i มีค่าเป็น 1 หรือ -1 ซึ่งกำหนดให้เป็นข้อมูลแบ่งกลุ่มของข้อมูล x_i ที่มีค่า c_i เป็น 1 และ x_i ที่มีค่า c_i เป็น -1 โดยที่แต่ละ x_i เป็นค่าข้อมูลมิติของเวกเตอร์จริง เมื่อกำหนดให้ข้อมูลนี้เป็นข้อมูล สำหรับฝึกฝน ซึ่งหมายความว่าการแบ่งกลุ่มของข้อมูลมีความถูกต้อง ดังนั้นเส้นแบ่งกลุ่มข้อมูลที่ ถูกสร้างขึ้นซึ่งเป็นสมการเส้นตรงทั่วไปคือ $y = mx+b$ โดยในที่นี้จะแทน m ด้วย w^T เพื่อกำหนด เป็นสมการ (2-2)

$$w^T.x+b = 0 \quad (2\text{-}2)$$

เมื่อ w^T คือ เวกเตอร์ตั้งฉากของค่าความชัน m ของเส้นแบ่ง

b คือ ค่าคงที่ที่ได้จากค่าของแกน y ของแต่ละข้อมูล x

การแบ่งข้อมูลดังภาพที่ 2-5



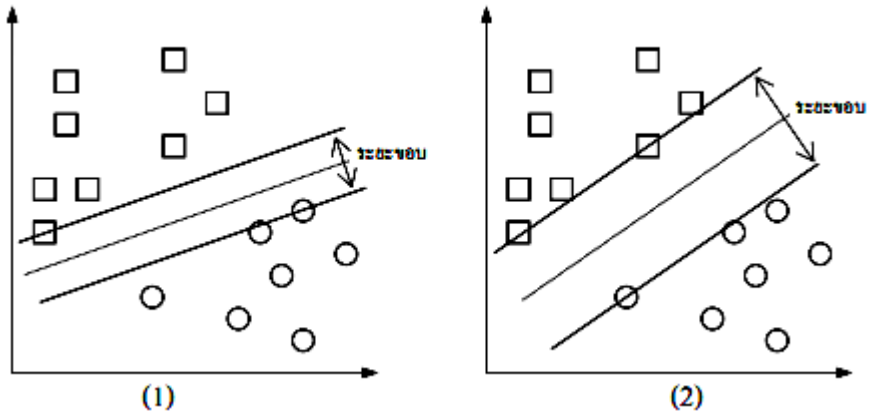
ภาพที่ 2-5 ข้อมูลสองกลุ่มที่ถูกแบ่งด้วยเส้นตรง

ในทางอุดมคติเส้นแบ่งกลุ่มที่ดีที่สุดคือเส้นแบ่งกลุ่มที่ทำให้มีระยะขอบ (Margin) มากที่สุดของเส้นคู่ขนานที่ขยายออกจากเส้นแบ่งไปสัมผัสจุดข้อมูลของทั้งสองกลุ่มที่ใกล้ที่สุด หรือกล่าวได้

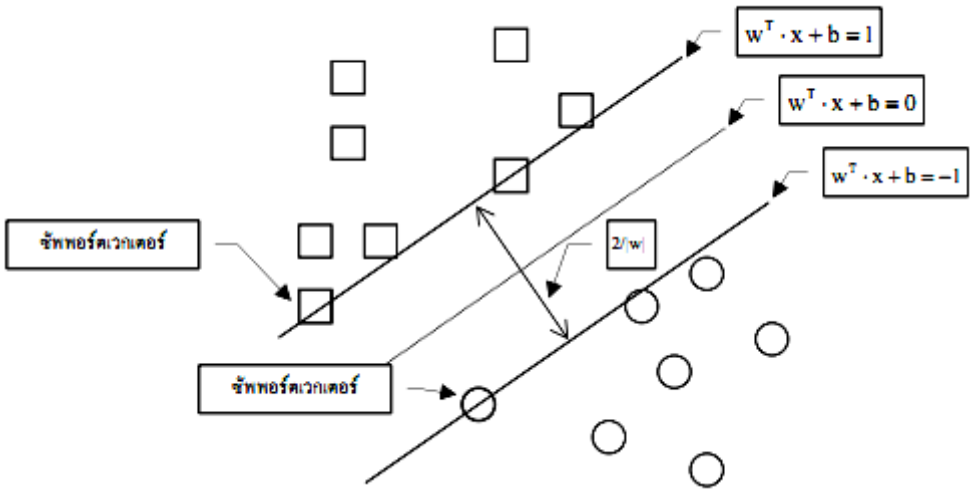
ว่าเป็นเส้นแบ่งที่มีระยะขอบกว้างที่สุด จากภาพที่ 2-6 และภาพที่ 2-7 เรียกจุดข้อมูลอย่างน้อยหนึ่งจุดจากทั้งสองกลุ่มที่สัมผัสกับเส้นขนานที่ขยายออกได้มากที่สุดว่า ซัพพอร์ตเวกเตอร์ โดยค่าของเส้นขอบทั้งสองที่ใช้แบ่งกลุ่มข้อมูลเป็นสองกลุ่มคำนวณได้จากสมการ (2-3) และสมการ (2-4)

$$w^T.x+b = 1 \quad (2\text{-}3)$$

$$w^T.x+b = -1 \quad (2\text{-}4)$$



ภาพที่ 2-6 การปรับความชันของเส้นแบ่งแล้วทำให้ได้ระยะขอบที่มากที่สุด



ภาพที่ 2-7 ระยะขอบที่กว้างที่สุดเมื่อสัมผัสกับซัพพอร์ตเวกเตอร์

เมื่อข้อมูลที่ใช้ฝึกฝนเป็นข้อมูลที่สามารถแบ่งกลุ่มได้ด้วยเส้นตรงใด ๆ ระยะที่ที่กว้างที่สุดของไฮเปอร์แพลน (Hyper plane) ทั้งสองที่ขยายออกไปจนกว่าจะพบจุดข้อมูลของทั้งสองกลุ่มคือ $2/|w|$ โดย ค่า $|w|$ มีค่าน้อยที่สุด ค่าข้อมูลแต่ละ x_i จุดจำแนกอยู่ในกลุ่มใดสามารถพิจารณาได้จากเงื่อนไขดังนี้ ถ้า $w^T.x_i+b \geq 1$ แสดงว่า x_i เป็นกลุ่มที่ 1 และถ้า $w^T.x_i+b \leq -1$ แสดงว่า x_i เป็นกลุ่มที่ 2

ในการคัดแยกจุดข้อมูลใด ๆ ที่นำมาฝึกฝนเพื่อกรองว่าเป็นกลุ่ม 1 สามารถตรวจสอบ ได้จากสมการที่ (2-5)

$$c_i(w \cdot x_i-b) \geq 1 \quad \text{เมื่อ} \quad 1 \leq i \leq n \tag{2-5}$$

ดังนั้นสามารถเขียนเป็นรูปสั้นเพื่อหาระยะขอบที่น้อยที่สุด โดยที่สามารถคำนวณการแบ่งกลุ่มได้ดังสมการที่ (2-6)

$$\text{Minimize}_{w,b} |w| \text{ โดยตรวจสอบ } c_i(w \cdot x_i-b) \geq 1 \quad \text{เมื่อ} \quad 1 \leq i \leq n \tag{2-6}$$

2.4.2 ปัญหาในการเรียนรู้ของเอสวีเอ็ม (Vapnik, 1995) ได้นำเสนอวิธีการปรับปรุงการเรียนรู้แบบ Maximum Margin โดยมีวิธีการดังนี้

2.4.2.1 การแก้ไขปัญหาความผิดพลาดในการจัดกลุ่ม (Misclassification) โดยให้สามารถผ่อนปรนหรืออนุโลมให้มีการละจุดข้อมูลที่ไม่สามารถสร้างเส้นระยะขอบที่มากที่สุดที่ดีที่สุดเพื่อการแบ่งกลุ่ม เรียกว่า ซอฟต์มาร์จิน (Soft Margin) ดังตัวอย่างในภาพที่ 2-8 เรียกจุดข้อมูลนี้ว่า เวกเตอร์อนุโลม โดยการกำหนดให้มีตัวแปรค่าความหยอน (Slack Variable) ξ_i ซึ่งใช้กำหนดค่าการอนุโลมของการจัดกลุ่มไม่ได้จากกลุ่มข้อมูล x_i และตัวแปร C โดยปรับปรุงสูตรการคำนวณใหม่ดังสมการที่ (2-7)

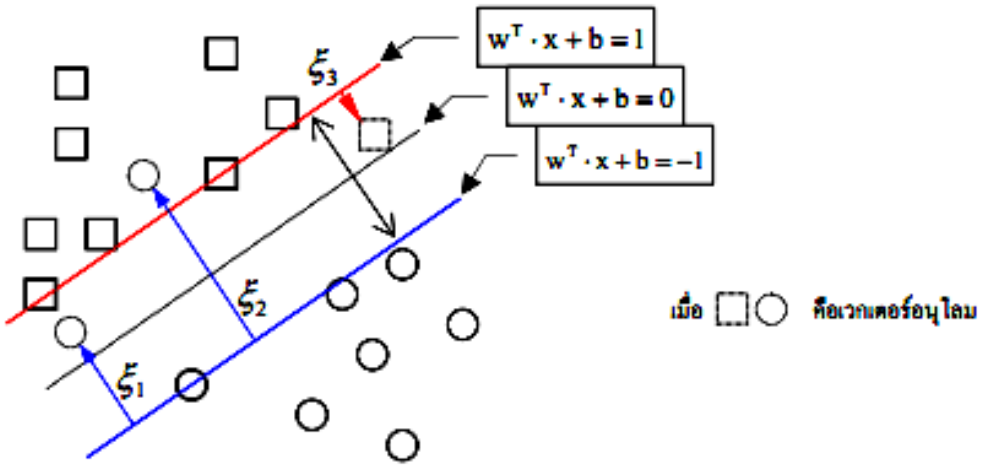
$$\text{minimize}_{w,b} |w| + C\sum_{i=1}^{n} \xi_i \tag{2-7}$$

โดยตรวจสอบ $c_i (w \cdot x_i -b) \geq 1 - \xi_i$

$\xi_i \geq 0$

$C > 0$

เมื่อ $1 \leq i \leq n$

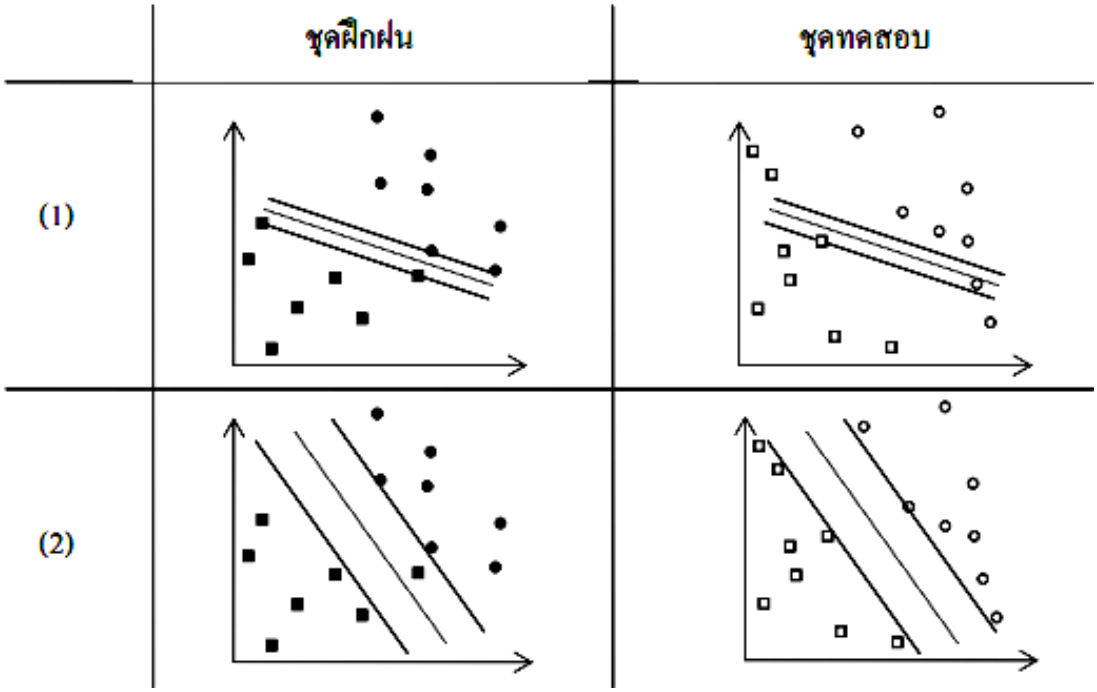


ภาพที่ 2-8 การเพิ่มตัวแปรค่าความหย่อนสำหรับเวกเตอร์อนุโลม

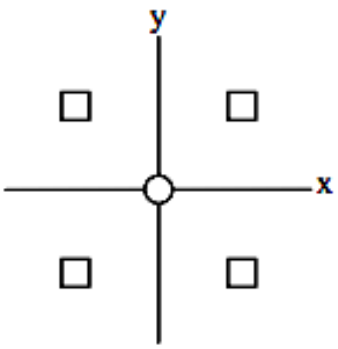
2.4.2.2 การแก้ไขปัญหาเกินพอดี (Over fitting) ซึ่งเกิดจากการปรับให้เอสวีเอ็มเรียนรู้จนได้ค่าที่ดีที่สุด แต่เมื่อนำไปทดสอบกับข้อมูลจริงปรากฏว่ามีความผิดพลาดสูง ดังภาพที่ 2-9 (1) เป็นการเรียนรู้ให้ได้เส้นแบ่งในการแบ่งกลุ่มที่ดีที่สุด แต่เมื่อนำไปทดสอบกับข้อมูลชุดทดสอบปรากฏว่าผิดพลาดมาก เมื่อเปรียบเทียบกับภาพที่ 2-9 (2) ที่มีการผ่อนปรนใหม่ค่าผิดพลาดในการเรียนรู้บ้าง เมื่อนำไปทดสอบกับข้อมูลชุดทดสอบปรากฏว่ามีความผิดพลาดน้อยลงโดยที่ $C\sum^{n}_{i=1}\xi_i$ ทำหน้าที่ในการควบคุมจำนวนจุดข้อมูลที่จัดกลุ่มไม่ได้ ซึ่งค่าของ C จะถูกกำหนดโดยผู้ใช้ ทั้งนี้ C ที่มีค่ามากหมายถึงการยอมให้มีความผิดพลาดได้มาก เมื่อนำไปประยุกต์ใช้ในการคัดแยกข้อมูลจริงมักจะพบกับข้อมูลที่ไม่สมดลซึ่งอาจเกิดจากจำนวนตัวอย่างในแต่ละกลุ่มมีจำนวนไม่เท่ากัน หรือความหนาแน่นของแต่ละคุณลักษณะ (Feature) ที่แตกต่างกันมากในแต่ละกลุ่ม โดยปกติแล้วควรมีการปรับปรุงข้อมูลตัวอย่างให้สมดุลก่อนนำเข้าสู่กระบวนการเรียนรู้จึงทำให้ระบบ สามารถเรียนรู้ได้ดีที่สุด

2.4.3 เอสวีเอ็มกับข้อมูลแบบไม่เชิงเส้น พื้นฐานของเอสวีเอ็มเริ่มใช้ในการคัดแยกกลุ่มข้อมูลในลักษณะเชิงเส้น (Linear Classifier) แต่ในบางครั้งข้อมูลจริงอาจจะมีคุณลักษณะที่ไม่เป็นเชิงเส้นดังตัวอย่างในภาพที่ 2-10 เพื่อแก้ปัญหาดังกล่าวจึงมีวิธีจำลองข้อมูลโดยเพิ่มมิติของข้อมูล (Higher Dimensional) มากขึ้น เช่น ข้อมูลเดิมมีขนาด 2 มิติ คือ แกน x และแกน y จะจำลองมิติที่ 3 ขึ้นมาดังภาพที่ 2-11 ซึ่งเป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2$ และภาพที่ 2-12 เป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2$ และภาพที่ 2-12 เป็นการจำลองมิติที่ 3 ด้วยการสร้างโครงผิวจากสมการ $f(x,y)=x^2+y^2$ เพื่อเป็นการยกข้อมูลให้ไปอยู่บนโครงผิวที่สร้างขึ้น หลังจากนั้นจึงคำนวณหาเส้นระนาบใดๆ ที่สามารถใช้แบ่งข้อมูลจากโครงผิวเป็นสองกลุ่ม

หลังจากนั้นจึงทำการลดมิติกลับไปเป็น 2 มิติ สิ่งที่ได้คือขอบเขตของการแบ่งข้อมูลออกเป็นสองกลุ่ม หลังจากนั้นจึงทำการลดมิติกลับไปเป็น 2 มิติ สิ่งที่ได้คือขอบเขตของการแบ่งข้อมูลออกเป็นสองกลุ่ม ฟังก์ชันที่ใช้ในการคำนวณเพื่อเพิ่มมิติของข้อมูลเรียกว่า ฟังก์ชันแกน (Kernel Function)

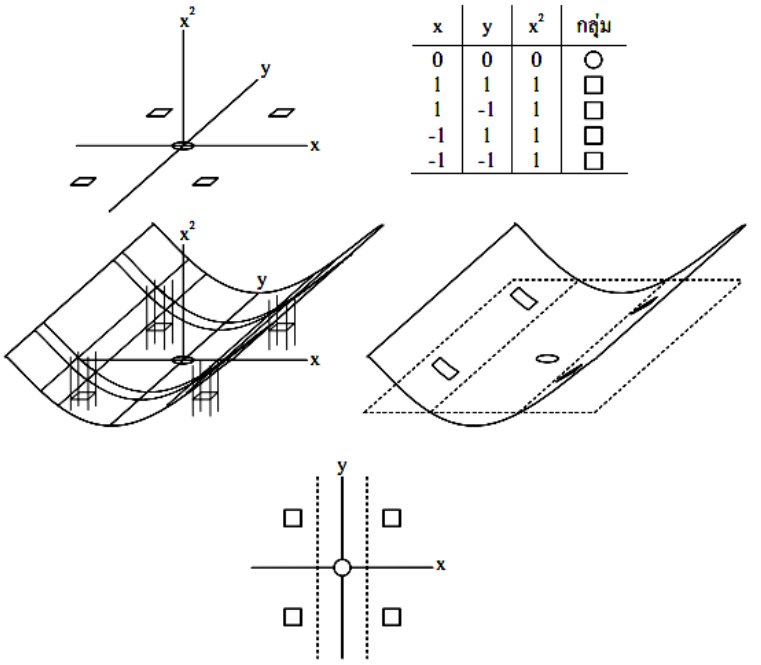


ภาพที่ 2-9 การเปรียบเทียบตัวอย่างการเรียนรู้ของเอสวีเอ็มแบบเกินพอดี



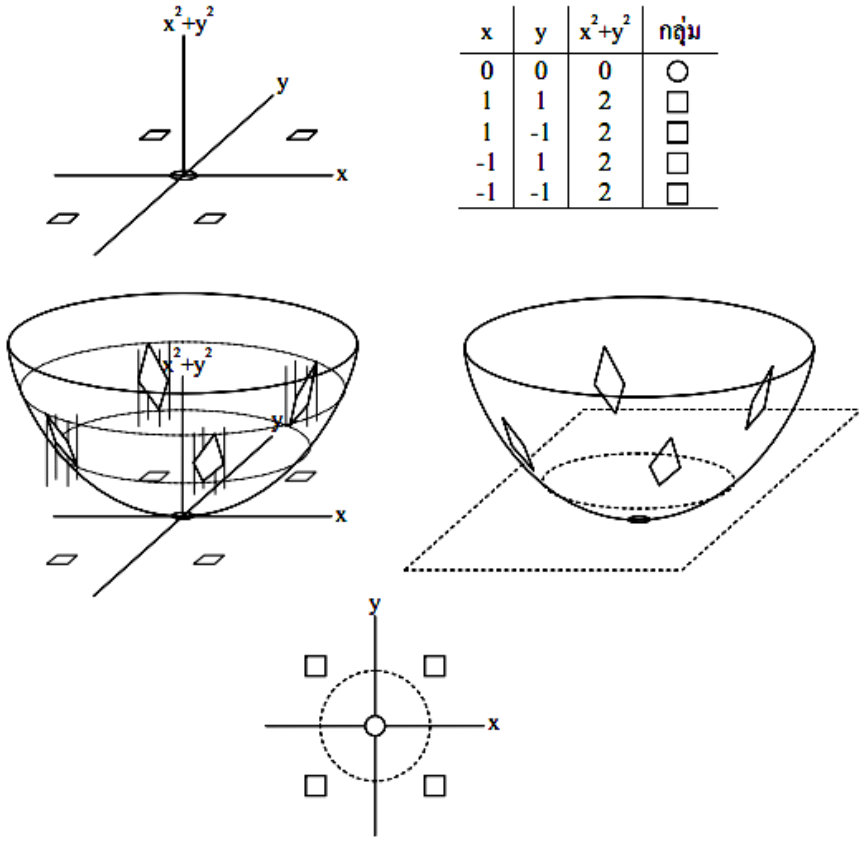
x	y	กลุ่ม
0	0	○
1	1	□
1	-1	□
-1	1	□
-1	-1	□

ภาพที่ 2-10 ตัวอย่างข้อมูลสองมิติที่ไม่สามารถใช้การคัดแยกแบบเชิงเส้นได้



x	y	x^2	กลุ่ม
0	0	0	○
1	1	1	□
1	-1	1	□
-1	1	1	□
-1	-1	1	□

ภาพที่ 2-11 การเพิ่มมิติของข้อมูลด้วย $f(x,y)=x^2$ เพื่อให้สามารถแบ่งกลุ่มข้อมูล



x	y	x^2+y^2	กลุ่ม
0	0	0	○
1	1	2	□
1	-1	2	□
-1	1	2	□
-1	-1	2	□

ภาพที่ 2-12 การเพิ่มมิติข้อมูลด้วย $f(x,y)=x^2+y^2$ เพื่อสามารถแบ่งกลุ่มข้อมูลด้วยระนาบเชิงเส้น

เมื่อข้อมูลแต่ละ x_i ถูกคำนวณใหม่มิติสูงขึ้นด้วยฟังก์ชันแกน Φ จากสมการที่ (2-6) สามารถเขียนใหม่เป็นสมการที่ (2-8)

$$\text{Minimize}_{w,b} |w| \text{ โดยตรวจสอบ } c_i(w \cdot \Phi x - b) \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n \tag{2-8}$$

กำหนดผลของฟังก์ชันแกน $k(x,x')$ ให้เป็นบวกเพื่อใช้ยกข้อมูลให้อยู่ในมิติที่สูงกว่าซึ่งคำนวณได้จากอินเนอร์โปรดักท์ของพื้นที่คุณลักษณะ (Feature Space) ดังสมการ (2-9)

$$k(x, x') = \Phi x \cdot \Phi x' \tag{2-9}$$

สมการที่ (2-9) สามารถเขียนใหม่ในรูปของการรวมฟังก์ชันแกน ได้ดังสมการที่ (2-10)

$$\text{Minimize}_{w,b} |w| \quad \text{โดยตรวจสอบ } c_i(w \cdot k(x, x') - b) \quad \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n \tag{2-10}$$

โดยที่ฟังก์ชันแกนพื้นฐาน ที่ใช้ในการเพิ่มมิติข้อมูลสำหรับประมวลผลข้อมูลแสดงไว้ในตารางที่ 2-1

ตารางที่ 2-1 ฟังก์ชันแกนพื้นฐานสำหรับเอสวีเอ็ม

ชนิดของฟังก์ชัน	รูปสมการ
Linear	$k(x,x') = (x^T x')$
Polynomial	$k(x,x') = (\gamma x^T x' + r)^d, \gamma > 0$
Radius Basic Function	$k(x,x') = \exp(-\gamma \|\| x - x' \|\|^2), \gamma > 0$
Sigmoid	$k(x,x') = \tanh(\gamma x^T x' + r)$
โดยที่ γ , r และ d เป็นพารามิเตอร์ของฟังก์ชันแกน	

ดังนั้นสมการพื้นฐานของเอสวีเอ็มเมื่อรวมฟังก์ชันการคำนวณเพิ่มมิติและการกำหนดเวกเตอร์อนุโลมเข้าด้วยกันสามารถเขียนใหม่ได้ดังสมการ (2-11)

$$\text{minimize}_{w,b} |w| + C\sum_{i=1}^{n} \xi_i \tag{2-11}$$

โดยตรวจสอบ $c_i (w \cdot k(x,x')-b) \geq 1$

$\xi_i \geq 0$

$C > 0$

เมื่อ $1 \leq i \leq n$

2.4.4 การใช้เอสวีเอ็มคัดแยกข้อมูลมากกว่าสองกลุ่ม (Multiclass Classification) เอสวีเอ็มมีพื้นฐานการคัดแยกแบบสองกลุ่ม (Binary Classification) ด้วยการสร้างระนาบตัดสินใจ สำหรับการประยุกต์ใช้เอสวีเอ็มในการคัดแยกกลุ่มข้อมูลมากกว่าสองกลุ่มทำได้โดยการเปรียบเทียบคัดแยกทีละคู่ ๆ หลังจากนั้นจึงหาข้อสรุปเพื่อคัดแยกว่าแต่ละข้อมูลอยู่กลุ่มใด วิธีที่นิยมประยุกต์ใช้มี 2 วิธี (Burges, 1998) คือ การคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด (One-against-The Rest) และ การคัดแยกทีละหนึ่งต่อหนึ่ง (One-against-One) โดยรายละเอียดแต่ละวิธีมีดังนี้

2.4.4.1 วิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด หรือบางครั้งเรียกว่า One-against-All เป็นการเลือกกลุ่มใด ๆ มาเปรียบเทียบกับกลุ่มอื่น ๆ ทีละคู่จนครบทุกกลุ่ม เมื่อมีจำนวนกลุ่มเท่ากับ k กลุ่ม วิธีการนี้จะทำการเรียนรู้การคัดแยกแบบสองกลุ่มจำนวน k รอบ โดยการเปรียบเทียบวนกลุ่มที่เลือกกับกลุ่มที่เหลือ แสดงตัวอย่างดังตารางที่ 2-2

ตารางที่ 2-2 การเปรียบเทียบแบ่งกลุ่มแบบการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมด

กลุ่มที่เลือก		กลุ่มที่เหลือ
1	เปรียบเทียบกับ	2 จนถึงกลุ่มที่ k
2	เปรียบเทียบกับ	1, 3 จนถึงกลุ่มที่ k
3	เปรียบเทียบกับ	1, 2, 4 จนถึงกลุ่มที่ k
⋮	⋮	⋮
k	เปรียบเทียบกับ	1 จนถึงกลุ่มที่ (k-1)

การเปรียบเทียบวนทีละกลุ่มแบบนี้ทำให้ได้ฟังก์ชันตัดสินใจในการแบ่งกลุ่มจำนวน k ฟังก์ชัน คือ

$$(W^1)^T x + b_1$$
$$(W^2)^T x + b_2$$
$$\vdots$$
$$(W^k)^T x + b_k$$

ในการตัดสินใจแบ่งข้อมูลออกเป็นกลุ่มใด ใช้วิธีการคัดกรองกลุ่มที่เหลือออก ยกตัวอย่างเช่นในการตัดสินใจแบ่งข้อมูลเป็นกลุ่มที่ 1 จากจำนวน k กลุ่ม มีการคัดกรองกลุ่มที่เหลือออกคือกลุ่มที่ 2 จนถึงกลุ่มที่ k ด้วยฟังก์ชันตัดสินใจดังนี้

$$(W^{1})^{T} x + b_{1} \geq +1$$
$$(W^{2})^{T} x + b_{2} \leq -1$$
$$\vdots$$
$$(W^{k})^{T} x + b_{k} \leq -1$$

ผลที่ได้คือการคัดแยกข้อมูลออกเป็นกลุ่มต่าง ๆ ตามลักษณะฟังก์ชันตัดสินใจที่สร้างขึ้นในแต่ละรอบ

2.4.4.2 วิธีการคัดแยกทีละหนึ่งต่อหนึ่ง เป็นการเปรียบเทียบกลุ่มข้อมูลจำนวน k กลุ่ม ด้วยการ เปรียบเทียบเพื่อคัดแยกแบบสองกลุ่มทีละคู่แบบไม่ซ้ำกันจนครบทุกกลุ่ม ตัวอย่างเช่น

(1,2),(1,3),(1,4),…,(1,k)

(2,3),(2,4),…,(2,k)

(3,4),…,(2,k)

(k-1,k)

โดยจำนวนครั้งในการเปรียบเทียบของการเรียนรู้เท่ากับ k(k-1)/2 ครั้ง ดังนั้น ฟังก์ชันตัดสินใจมีจำนวน k(k-1)/2 ฟังก์ชัน เช่นกัน ตัวอย่างเช่น หากมีจำนวนกลุ่ม 4 กลุ่ม จำนวนฟังก์ชันตัดสินใจเท่ากับ 4(4-1)/2 = 6 ฟังก์ชัน แสดงตัวอย่างดังตารางที่ 2-3

ตารางที่ 2-3 ตัวอย่างฟังก์ชันตัดสินใจของวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง

$y_i = 1$	$y_i = -1$	ฟังก์ชันตัดสินใจ
กลุ่ม 1	กลุ่ม 2	$f^{12}(x) = (w^{12})^{T} x + b^{12}$
กลุ่ม 1	กลุ่ม 3	$f^{13}(x) = (w^{13})^{T} x + b^{13}$
กลุ่ม 1	กลุ่ม 4	$F^{14}(x) = (w^{14})^{T} x + b^{14}$
กลุ่ม 2	กลุ่ม 3	$F^{23}(x) = (w^{23})^{T} x + b^{23}$
กลุ่ม 2	กลุ่ม 4	$F^{24}(x) = (w^{24})^{T} x + b^{24}$
กลุ่ม 3	กลุ่ม 4	$F^{34}(x) = (w^{34})^{T} x + b^{34}$

ในการตัดสินว่าข้อมูลที่คัดแยกอยู่กลุ่มใด ใช้วิธีพิจารณาผลการโหวตสูงสุดจากการเปรียบเทียบทีละคู่ โดยแต่ละกลุ่มมีโอกาสเปรียบเทียบ จำนวน k-1 ครั้งเท่า ๆ กัน ยกตัวอย่างเช่น ตารางที่ 2-4

ตารางที่ 2-4 ตัวอย่างผลการเปรียบเทียบทีละคู่แบบไม่ซ้ำกัน

คู่เปรียบเทียบ		ผลการเปรียบเทียบ
กลุ่ม 1	กลุ่ม 2	กลุ่ม 1
กลุ่ม 1	กลุ่ม 3	กลุ่ม 1
กลุ่ม 1	กลุ่ม 4	กลุ่ม 1
กลุ่ม 2	กลุ่ม 3	กลุ่ม 2
กลุ่ม 2	กลุ่ม 4	กลุ่ม 4
กลุ่ม 3	กลุ่ม 4	กลุ่ม 3

รวมผลเปรียบเทียบแล้วนำมาพิจารณาตัดสินใจเลือกกลุ่ม ตัวอย่างดังตารางที่ 2-5 ดังนั้นการแบ่งกลุ่ม ครั้งนี้จึงถูกจัดให้เป็นกลุ่มที่ 1 ด้วยผลรวมการเปรียบเทียบทีละคู่ สูงสุดคือ 3 ครั้ง

ตารางที่ 2-5 ผลการตัดสินใจเลือกกลุ่มของการคัดแยก

	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	กลุ่ม 4
ผลรวมการเปรียบเทียบ	3	1	1	1

จากการเปรียบเทียบการทำงานของทั้งสองวิธีในการคัดแยกข้อมูลมากกว่าสองกลุ่มโดยใช้เอสวีเอ็ม พบว่าผลที่ได้มีความถูกต้องใกล้เคียงกัน แต่วิธีการคัดแยกทีละหนึ่งต่อหนึ่งใช้เวลาในการประมวลผลน้อยกว่า เนื่องจากเมื่อพิจารณามิติของข้อมูลจำนวน n ข้อมูล ที่ k กลุ่ม วิธีการคัดแยกทีละหนึ่งต่อหนึ่งใช้การเปรียบเทียบคัดแยกจำนวน k(k-1)/2 ครั้งบนข้อมูลจำนวน 2n/k ข้อมูล เมื่อพิจารณาค่าความซับซ้อนตามขนาดข้อมูลเมื่อประมวลผลด้วยตัวอย่างฟังก์ชันแกนแบบโพลิโนเมียล (Polynomial) ที่มีดีกรีเท่ากับ d คือ $\frac{k(k-1)}{2}O\left(\left(\frac{2n}{k}\right)^d\right)$

ในขณะที่วิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมดใช้การเปรียบเทียบคัดแยกจำนวน k ครั้งบนจำนวนข้อมูล n ข้อมูล ซึ่งมีค่าความซับซ้อนตามขนาดข้อมูลเมื่อประมวลผลด้วยฟังก์ชันแกนแบบโพลิโนเมียลที่มีดีกรีเท่ากับ d คือ $kO(n^d)$

ดังนั้นจะเห็นได้ว่าเมื่อเปรียบเทียบวิธีการประมวลผลด้วยฟังก์ชันแกนเดียวกันบนมิติของจำนวนข้อมูลที่เท่ากันและจำนวนกลุ่มเท่ากัน วิธีการคัดแยกทีละหนึ่งต่อหนึ่งมีค่าความซับซ้อนตามขนาด ข้อมูลน้อยกว่า เพราะฉะนั้นวิธีการคัดแยกทีละหนึ่งต่อหนึ่งจะใช้เวลาในการประมวลผลน้อยกว่าวิธีการคัดแยกทีละหนึ่งเปรียบเทียบกับส่วนที่เหลือทั้งหมดดังนั้นในงานวิจัยนี้จึงเลือกเอสวีเอ็มที่มีความสามารถคัดแยกข้อมูลมากกว่าสองกลุ่ม โดยใช้วิธีการเปรียบเทียบแบบวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง ซึ่งมีการประมวลผลที่เร็วกว่าในขณะที่ ประสิทธิภาพของทั้งสองวิธีไม่แตกต่างกัน

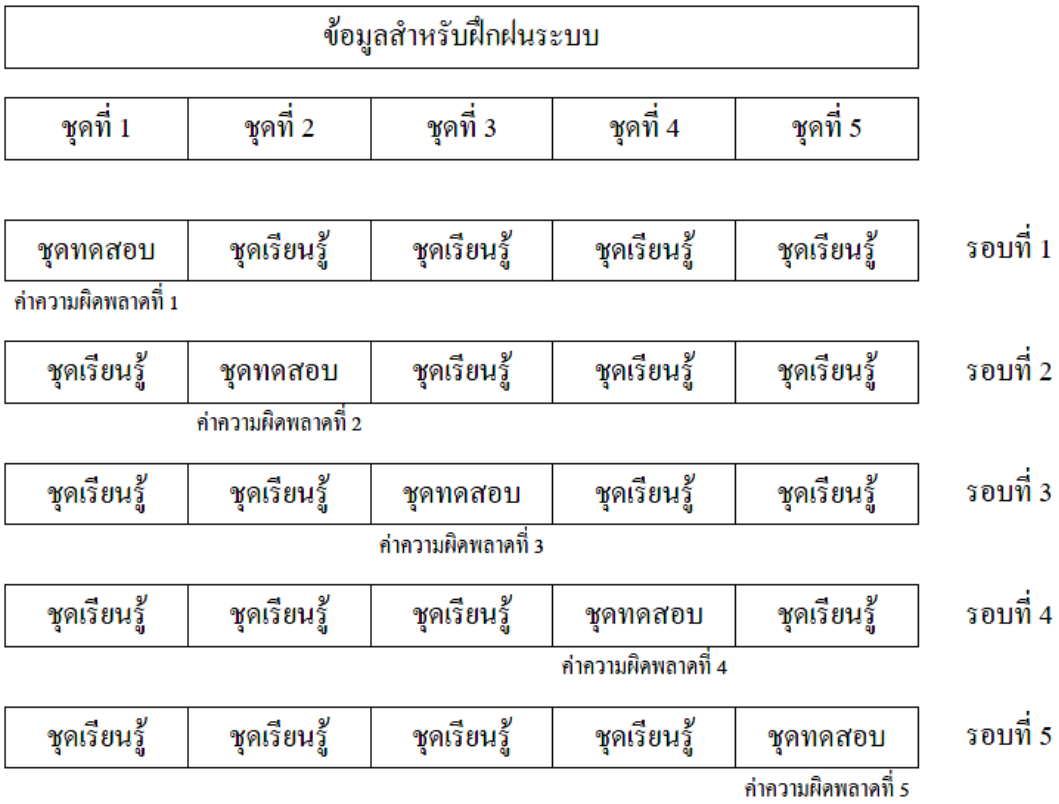
2.4.5 การตรวจสอบความถูกต้องของการเรียนรู้ สามารถสังเกตได้จากค่าความแม่นยำหรือค่าความผิดพลาดที่ได้จากการทำครอสวาลิเดชั่น (Cross Validation) ระหว่างการแบ่งชุดทดสอบในการเรียนรู้ ในที่นี้ผู้วิจัยเลือกพิจารณาความแม่นยำ ซึ่งเป็นสัดส่วนร้อยละของการคัดแยกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยใช้การครอสวาลิเดชั่นแบบเคโฟล์ด (k-fold) ซึ่งพื้นฐานของวิธีการครอสวาลิเดชั่นนี้คือการสุ่มตัวอย่าง (Sampling) ข้อมูลไปเป็นชุดฝึกฝนและชุดทดสอบแบบสลับกัน โดยเริ่ม จากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดทดสอบเพื่อจำลอง ผลลัพธ์จากการ ทำครอสวาลิเดชั่น การแบ่งชุดทดสอบแบบนี้มักถูกใช้เป็นตัวเลือกในการกำหนด โมเดล โดยสังเกตจากผลของการทดสอบในแต่ละครั้งของการปรับค่าพารามิเตอร์แต่ละโมเดล การทำครอสวาลิเดชั่นมีรูปแบบการทดลองเพื่อประเมินหลายวิธีดังนี้

2.4.5.1 วิธีแฮนด์เอาท์วาลิเดชั่น (Handout Validation) เป็นการแบ่งข้อมูลที่จะใช้ตรวจสอบโมเดลโดยการ เลือกสุ่มข้อมูลบางส่วนนำมาเป็นชุดทดสอบจากข้อมูลทั้งหมดที่ใช้เรียนรู้ด้วยการพิจารณาเลือกด้วยมือ โดยในการสุ่มจะต้องมีความสมดุลของการเป็นตัวอย่างที่ดี ส่วนข้อมูลที่เหลือจะใช้เป็นข้อมูล สำหรับฝึกฝนระบบ เมื่อฝึกฝนระบบจนได้โมเดลแล้วจึงนำโมเดลมาทดสอบหาความถูกต้องหรือ ค่าความผิดพลาดด้วยการทดสอบกับชุดข้อมูลทดสอบที่ได้สุ่มเอาไว ซึ่งวิธีการนี้เหมาะสำหรับ ข้อมูลที่นำมาทดสอบมีจำนวนไม่มาก

2.4.5.2 วิธีเคโฟล์ดครอสวาลิเดชั่น เป็นการแบ่งข้อมูลออกเป็น k ชุด เท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด k รอบ โดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล k ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก k-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลยเพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกฝน ดังตัวอย่างในภาพที่ 2-13 เป็นการกำหนดค่า k=5 หมายถึงการนำข้อมูล ทั้งหมดมาแบ่งเป็น 5 ชุด และจะดำเนินการเรียนรู้และทดสอบจำนวน 5 รอบ โดยในแต่ละรอบจะเลือกข้อมูลมา 1 ชุดไม่ซ้ำกันมาเป็นชุดทดสอบ ด้วยการเรียนรู้ของชุดเรียนรู้ที่เหลือ เมื่อทำครบ 5 รอบ หาผลรวมของค่าความผิดพลาดหรือความแม่นยำจากชุดทดสอบในแต่ละรอบแล้วหารด้วยจำนวนรอบ ผลที่ได้เป็นค่าเฉลยของความผิดพลาดในการเรียนรู้ของโมเดล ข้อดีของวิธีการนี้

คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด k-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการคำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีมิติขนาดจำนวนมาก

2.4.5.3 วิธีลีฟวันเอาท์ครอสวาลิเดชั่น (Leave-one-out cross-validation) เป็นการทำครอสวาลิเดชั่นแบบเคโฟล์ดแบบหนึ่ง เมื่อกำหนดให้ k มีค่าเท่ากับจำนวนข้อมูลทั้งหมด ดังนั้นในกรณีที่มีข้อมูล 10 ชุด จะต้องทำการวัดผลความผิดพลาดโดยวิธีการ 10-fold เป็นต้น ข้อดีของวิธีนี้คือ ความเที่ยงตรงสูงใกล้เคียงกับการฝึกฝนในจริง เนื่องจากมีความแตกต่างของข้อมูลที่ใช้ฝึกฝนจริงเพียง 1 ชุดเท่านั้น แต่วิธีนี้ต้องใช้เวลาในการประมวลผลมากกว่าเนื่องจากต้องทำการประมวลผลเป็นจำนวน k รอบ ดังนั้น เมื่อนำไปใช้กับข้อมูลจำนวนมาก ยอมต้องใช้เวลาในการประมวลผลมากขึ้นด้วย โดยมีอัตราการเติบโตของเวลาการประมวลผลเป็น $O = n^2$ ดังนั้นวิธีการนี้จึงเหมาะกับข้อมูลตัวอย่างที่มีน้อย



ภาพที่ 2-13 ตัวอย่างการแบ่งชุดข้อมูลของเคโฟล์ดครอสวาลิเดชั่นเมื่อ k = 5

2.4.6 การคัดเลือกโมเดลจากการเรียนรู้ พิจารณาจากค่าความแม่นยำของแต่ละโมเดลเมื่อนำไปเรียนรู้กับกลุ่มตัวอย่างที่ใช้ฝึกฝน ทั้งนี้ความความแม่นยำขึ้นอยู่กับการปรับพารามิเตอร์ของฟังก์ชันแกนที่เลือกใช้ โดยทั่วไปของการปรับพารามิเตอร์ระหว่างการฝึกฝนมี 2 วิธี คือวิธีการค้นหาแบบตาราง (Grid Search) และวิธีการค้นหาแบบมีรูปแบบ (Pattern Search) โดยรายละเอียดแต่ละวิธีมีดังนี้

2.4.6.1 การค้นหาแบบตารางเป็นการทดสอบค่าพารามิเตอร์ทั้งหมดในช่วงที่กำหนดทีละค่าเพื่อหาค่าพารามิเตอร์ที่ทำให้โมเดลมีค่าความแม่นยำสูงที่สุด ตัวอย่างเช่น ในกรณีที่มีพารามิเตอร์ที่ต้องปรับ 2 พารามิเตอร์ และกำหนดช่วงให้แต่ละพารามิเตอร์มีค่าช่วงที่จะต้องทดสอบตั้งแต่ 0 ถึง 10 ซึ่งมีค่าทดสอบ 11 ค่า ดังนั้นการค้นหาแบบตารางจะมีจำนวนรอบการทดสอบเท่ากับ 11*11 หรือ 121 รอบการทดสอบ เป็นต้น

2.4.6.2 วิธีการค้นหาแบบมีรูปแบบมีหลักการทำงานคือ เริ่มต้นประมวลผลด้วยค่ากึ่งกลางของช่วงค่าพารามิเตอร์ที่กำหนดและทำการกำหนดขั้นของการสุ่มค่าขึ้นและลงของแต่ละค่าพารามิเตอร์ หากค่าใดทำให้ความแม่นยำสูงขึ้น จุดกึ่งกลางในการทดสอบจะย้ายไปยังค่านั้นและทำซ้ำเช่นเดิมอีก หากพบว่าไม่มีค่าใดทำให้โมเดลมีความแม่นยำดีขึ้น ในรอบถัดไปจะมีการกำหนดค่าขั้นของการสุ่มให้ละเอียดขึ้น และทำซ้ำเช่นเดิมไปเรื่อย ๆ จนกว่าค่าความละเอียดของขั้นการสุ่มจะลดลงถึงค่าที่กำหนดให้หยุด ซึ่งวิธีนี้มีจุดดีตรงที่มีจำนวนรอบการทำงานเพื่อทดสอบน้อยกว่าการค้นหาแบบตาราง

2.4.7 ข้อจำกัดของเอสวีเอ็ม Burges ได้กล่าวถึงข้อจำกัดของการเรียนรู้ของเครื่องแบบเอสวีเอ็มไว้ดังนี้

2.4.7.1 การเลือกฟังก์ชันแกนที่เหมาะสม ซึ่งงานวิจัยส่วนใหญ่มักกำหนดฟังก์ชันแกนไว้ก่อน แล้วจึงปรับพารามิเตอร์ของฟังก์ชันแกนเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด ดังนั้นหากต้องการเลือกฟังก์ชันแกนที่เหมาะสมกับข้อมูลนั้นก็ยังต้องมีการวิจัยเพิ่มเติม

2.4.7.2 ความเร็วและขนาดของข้อมูลในการเรียนรู้และทดสอบ โดยเฉพาะเมื่อใช้กับข้อมูลที่มีจำนวนซัพพอร์ตเวกเตอร์จำนวนมาก เช่น มากกว่าล้านซัพพอร์ตเวกเตอร์ เป็นต้น จะทำให้ระบบไม่สามารถทำงานได้

2.4.7.3 ข้อมูลส่วนใหญ่ไม่ได้เป็นข้อมูลแบบสองกลุ่มเสมอไป ดังนั้นการออกแบบวิธีการในการคัดแยกแบบหลายกลุ่ม เป็นประเด็นที่ต้องทำวิจัยเพื่อพัฒนาให้ตอบสนองกับข้อมูลหลากหลายแบบมากขึ้น ในงานวิจัยนี้ผู้วิจัยได้เลือกเอสวีเอ็มเป็นระบบเรียนรู้ในระบบการจัดหมวดหมู่เอกสารแบบอัตโนมัติ โดยทำการศึกษาข้อมูลที่นำมาใช้กับฟังก์ชันแกนแบบต่าง ๆ และทำการปรับพารามิเตอร์ โดยใช้วิธีการแบบการค้นหาแบบตาราง แม้ว่าจะต้องใช้จำนวนรอบใน

การทำงานมากและประมวลผลช้า เนื่องจากผู้วิจัยต้องการศึกษาคุณลักษณะของพารามิเตอร์ที่ส่งผลต่อความถูกต้องของระบบ ส่วนการตรวจสอบความถูกต้องของการปรับพารามิเตอร์ใช้การครอสวาลิเดชั่นแบบเคโฟล์ดบนการคัดแยกข้อมูลแบบหลายกลุ่มด้วยวิธีการคัดแยกทีละหนึ่งต่อหนึ่ง ซึ่งเป็นวิธีที่มีความรวดเร็วในการประมวลผลและมีประสิทธิภาพดี

2.5 เนอีฟเบย์

วิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันมีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นทำให้ผู้วิจัยเลือกวิธีการนี้มาใช้ในงานวิจัย เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่นๆ

กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X=\{ a_1, a_2, ..., a_n\}$ หรือใช้สัญลักษณ์ว่า $P(a_1, a_2, ..., a_n \mid v_j)$ ดังสมการที่ 2-12

$$P(a_1,a_2,...a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j) \tag{2-12}$$

โดยที่ $\prod$ หมายถึง ผลคูณของค่า $P(a_i \mid v_j)$ ทั้งหมด

$i = 1,2,3,\ldots,n$ และ $j = 1,2,3,\ldots,n$

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้ มีวิธีการดังต่อไปนี้ คือ

1. หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, ..., a_n \mid v_j)$ จากสมการที่ 1 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ $P(v_j)$ ได้เท่ากับ v_{NB}

2. นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการที่ 2-13

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j) \tag{2-13}$$

จากสมการที่ 2-13 สามารถเขียนเป็นอัลกอริทึมการเรียนรู้เบย์อย่างง่าย ได้ดังต่อไปนี้

- **Naive_Bayes_Learn(examples)**
  FOR EACH target value v **DO**
  $\overline{P}(v_j) \leftarrow$ estimate $P(v_j)$
  FOR EACH attribute value a of each attribute **DO**
  $\overline{P}(a_i \mid v_j) \leftarrow$ estimate $P(a_i \mid v_j)$
- **Classify_New_Example(x)**

$$v_{NB} = \arg\max_{v_j \in V} \overline{P}(v_j) \times \prod_{i=1}^{n} \overline{P}(a_i \mid v_j)$$

ภาพที่ 2-14 อัลกอริทึมการเรียนรู้เบย์อย่างง่าย

การจัดกลุ่มเอกสารเป็นเทคนิคสำคัญของการจำแนกประเภทข้อความ จุดประสงค์เพื่อจำแนกเอกสารโดยอัตโนมัติ มีอัลกอริทึมมากมายที่ถูกพัฒนาขึ้นสำหรับการจัดกลุ่มเอกสารแบบอัตโนมัติอีกหนึ่งวิธีที่นิยมใช้ทั่วกัน ก็คือ เบย์อย่างง่าย ซึ่งมีงานวิจัยที่ศึกษาอัลกอริทึมแบบเบย์อย่างง่ายมากมาย ในการนำเอกสารมาเรียนรู้เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้อง

ในการเรียนรู้เพื่อจำแนกประเภทเอกสารโดยใช้เบย์อย่างง่าย สมมติว่ามีข้อความที่สนใจกับไม่สนใจ เมื่อทำการเรียนรู้แล้วต้องการทำนายว่าเอกสารหนึ่งๆ จะเป็นเอกสารที่สนใจหรือไม่ สามารถนำประยุกต์ใช้งาน เช่น การกรองข่าวสารเลือกเฉพาะข่าวที่สนใจ เป็นต้น

สมมติให้เอกสารหนึ่ง ๆ คือ ตัวอย่างหนึ่งตัว และแทนเอกสารแต่ละฉบับด้วยเวกเตอร์ของคำโดยใช้คำที่ปรากฏในเอกสารเป็นคุณสมบัติของเอกสาร กล่าวคือ คำที่หนึ่งในเอกสารเป็นคุณสมบัติตัวที่หนึ่ง คำที่สองในเอกสารเป็นคุณสมบัติตัวที่สอง ตามลำดับ ดังนั้นจะได้ว่า a_1 คือคุณสมบัติของคำที่หนึ่ง a_2 คือคุณสมบัติของคำที่สองตามลำดับ จากนั้นสร้างการเรียนรู้ข้อมูลโดยใช้ตัวอย่างสอนเพื่อประมาณค่าความน่าจะเป็นจากสมมติฐานเรื่องความไม่ขึ้นต่อกันของคุณสมบัติของเบย์อย่างง่ายทำให้ได้ สมการหาค่าความน่าจะเป็นของเอกสารตามสมการที่ (2-14)

$$P(doc \mid v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k \mid v_j) \qquad (2\text{-}14)$$

เมื่อ a_i คือ คุณสมบัติตัวที่ i

w_k คือ คำที่ k ในรายการของคำที่เรามีอยู่

$P(a_i = w_k \mid v_j)$ คือ ความน่าจะเป็นที่คำในตำแหน่งที่ i ของคำที่ k ในรายการของคำที่พบในเอกสาร v_j

$P(doc \mid v_j)$ คือ ความน่าจะเป็นของแต่ละเอกสาร v_j

อัลกอริทึมสำหรับการจำแนกประเภทข้อความโดยใช้การเรียนรู้เบย์อย่างง่ายเป็นดังต่อไปนี้

Algorithm: Learn_naive_Bayes_text(*Examples*,V)

1. Collect all words and other tokens that occur in *Examples*.
   - *Vocabulary* ← all distinct words and other tokens in *Examples*.
2. Calculate the required $P(v_j)$ and $P(w_k|v_j)$:
   - **FOR EACH** target value v_j in V **DO**
     - $docs_j$ ← subset of *Examples* for which the target value is v_j
     - $P(v_j) = \frac{|docs_j|}{Examples}$
     - $Text_j$ ← a single document created by concatenating all members of $docs_j$
     - n ← total number of words in $Text_j$ (counting duplicate words multiple times)
     - FOR EACH word w_k in *Vocabulary* **DO**
       - n ← number of times word w_k occurs in $Text_j$
       - $P(w_k|v_j) = \frac{n_k + 1}{n + |Vocabulary|}$

Algorithm: Classify_naive_Bayes_text(*Doc*)

- positions ← all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB}

$$v_{NB} = \arg\max_{v_i \in V} P(v_j) \times \prod_{i \in positions} P(a_i | v_j)$$

ภาพที่ 2-15 อัลกอริทึมการเรียนรู้เบย์อย่างง่ายสำหรับจำแนกประเภทเอกสาร

จากอัลกอริทึมด้านบน สามารถอธิบายรายละเอียดได้ดังนี้

1. นำคำที่ผ่านการตัดคำจากเอกสารมาสร้างเป็นข้อมูลตัวอย่าง

2. คำนวณหาค่าความน่าจะเป็นของกลุ่มเอกสาร $P(v_j)$ และคำนวณหาค่าความน่าจะเป็นของคำที่พบทั้งหมดในเอกสาร $P(a_i | v_j)$ โดยคำนวณได้จาก

2.1 คำนวณหาค่า $P(v_j)$ ได้จากสมการที่ (2-15)

$$P(v_j) = \frac{\text{จำนวนเอกสารในกลุ่มเอกสารนั้น ๆ}}{\text{จำนวนเอกสารทั้งหมด}} \quad (2\text{-}15)$$

2.2 คำนวณหาค่า $P(a_i | v_j)$ ได้จาก

$$P(a_i | v_j) = \frac{\text{ผลรวมความถี่ของคำที่พบในกลุ่มเอกสารนั้น ๆ}}{\text{จำนวนเอกสารในกลุ่มเอกสารนั้น ๆ}} \quad (2\text{-}16)$$

3. คำนวณหาค่า $P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j)$ หมายถึง ค่าความน่าจะเป็นของเอกสารในแต่ละกลุ่มเอกสาร คือ ผลคูณระหว่างค่าความน่าจะเป็นของกลุ่มเอกสารกับค่าความน่าจะเป็นของคำที่พบทั้งหมดในเอกสาร

4. เลือกกลุ่มเอกสารที่มีค่าความน่าจะเป็นมากที่สุด เป็นคำตอบ

(Kruengkrai, Canasai and Jaruskulchai, Chuleerat, 2001; บุญเสริม กิจศิริกุล, 2548)

2.6 งานวิจัยที่เกี่ยวข้อง

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ (2554) ได้นำเสนอแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย โดยลดคุณลักษณะของเอกสารก่อนประมวลผลด้วยเครื่องจักรการเรียนรู้ เพื่อประโยชน์ในการลดมิติของข้อมูล ลดระยะเวลาประมวลผล ประหยัดทรัพยากรของระบบ และเพิ่มประสิทธิภาพในการจัดหมวดหมู่เอกสารภาษาไทย จากการทดลองพบว่าการลดคุณลักษณะด้วยวิธีอินฟอร์เมชันเกรน (Information Gain) และประมวลผลด้วยเครื่องจักรการเรียนรู้แบบต่างๆ บนแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทยโดยวัดประสิทธิภาพการจัดหมวดหมู่เอกสารที่ระดับคุณลักษณะที่ดีสุด ได้ทดสอบกับเอกสารประเภทข่าวอิเล็กทรอนิกส์จากหนังสือพิมพ์ไทยรัฐ จำนวน 10 กลุ่ม ได้แก่ กลุ่มการศึกษา กลุ่มบันเทิง กลุ่มสังคม กลุ่มการเมือง กลุ่มเทคโนโลยี กลุ่มกีฬา กลุ่มข่าวต่างประเทศ กลุ่มเกษตร กลุ่มเศรษฐกิจ กลุ่มวัฒนธรรม โดยมีจำนวนกลุ่มตัวอย่างทั้งหมด 12,000 เอกสาร โดยเป็นกลุ่มตัวอย่างเรียนรู้ (Training) จำนวน 10,000 เอกสาร และกลุ่มทดลอง (Testing) จำนวน 2,000 เอกสาร พบว่าอัลกอริทึมเอสวีเอ็มให้ประสิทธิภาพสูงสุด คือ 94.3% รองลงมาเป็นอัลกอริทึมเบย์ 86.2% อัลกอริทึมอาร์บีเอฟ (RBF) 86.1% อัลกอริทึม J48 79.7% อัลกอริทึมริปเปอร์ 78.9% อัลกอริทึมเคเอ็นเอ็น (KNN) 70.3% ตามลำดับ เมื่อพิจารณาด้านการลดขนาดคุณลักษณะจากกลุ่มตัวอย่างของอัลกอริทึมเอสวีเอ็ม พบว่าสามารถลดลงได้มากถึง 90% โดยการลดลงของคุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด

พิลาพรรณ โพธิ์นรินทร์, สุพจน์ นิตย์สุวัฒน์ และชูชาติ หฤไชยะศักดิ์ (2554) ได้ทำวิจัยโดยนำข้อมูลเชิงบรรณานุกรมซึ่งเป็นข้อมูลรายงานวิจัยด้านการเกษตร จากศูนย์สารสนเทศทางการเกษตรแห่งชาติ สำนักหอสมุด มหาวิทยาลัยเกษตรศาสตร์ จำนวน 2,850 ระเบียน ในการจำแนกหมู่หมวดข้อมูล แบ่งออกเป็น 5 หมวดหมู่ คือ การปรับปรุงพันธุ์ การใช้ปุ๋ยและการปรับปรุงดิน ศัตรูพืช โรคพืช และสภาพแวดล้อม และเลือกใช้คำสำคัญมาทำการสกัดคุณลักษณะ (Feature Extraction) เพื่อดึงคุณลักษณะของคำมา มาใช้ในการทดลองจำแนกหมวดหมู่ข้อมูลแบบมีการเรียนรู้แบบมีผู้สอน (Supervised Learning) โดยเลือกใช้อัลกอริธึมต้นไม้การตัดสินใจ นาอีฟเบย์

และซัพพอร์ทเวกเตอร์แมชชีน จากผลการทดลองอัลกอริธึมที่ดีที่สุด คือ ซัพพอร์ทเวกเตอร์แมชชีน ฟังก์ชันเคอร์เนลแบบเรเดียลเบสิกฟังก์ชัน (Radial Basis Function) มีค่าความถ่วงดุลเท่ากับ 87.4%

กมลาสน์ อุดมล้ำเลิศ และอานนท์ รุ่งสว่าง (2553) กล่าวว่า การนำเสนอเอกสารเว็บเพจที่ถูกจัดแบ่งเป็นหมวดหมู่ตามเนื้อหาที่ปรากฏหรือที่เรียกว่า “เว็บไดเร็กทอรี่” เป็นการให้บริการในอีกรูปแบบหนึ่งที่ได้รับนิยมอย่างแพร่หลายของระบบสืบค้นข้อมูลเว็บยุคใหม่ แต่ด้วยปริมาณเว็บเพจที่มีจำนวนมากขึ้นเรื่อย ๆ จึงได้ทำการงานวิจัยเกี่ยวการจำแนกหมวดหมู่อัตโนมัติโดยใช้เครื่องจักรเรียนรู้ โดยศึกษาตัวจำแนกเว็บเพจภาษาไทยโดยเปรียบเทียบการใช้ตัวคัดแยกที่แตกต่างกัน การลดจำนวนคุณลักษณะโดยการคัดเลือกคุณลักษณะ และการให้น้ำหนัก จากผลการทดลองกับข้อมูลทดสอบ 2 ชุด ซึ่งคัดเลือกมาจากเว็บไดเร็กทอรี่ภาษาไทย 5 หมวดหมู่หลัก ทำให้เห็นว่าการลดคุณลักษณะโดยวิธีการคัดเลือกคุณลักษณะที่เหมาะสมจะไม่ส่งผลต่อความถูกต้องในการจำแนกหมวดหมู่ และการลดคุณลักษณะประกอบกับการใช้ตัวคัดแยกแบบนาอีฟเบย์ร่วมกับเทคนิคการให้น้ำหนักคุณลักษณะจะให้ค่าประสิทธิผลดีที่สุด โดยมีค่าความถูกต้อง (F-measure) 83.4%

จันทิมา พลพินิจ และอุมาภรณ์ สายแสงจันทร์ (2548) กล่าวว่า เนื่องจากระบบเวิร์ลไวด์เว็บได้รับความนิยมในการใช้งานเป็นอย่างมาก เพราะความสามารถในการให้บริการเกี่ยวกับสารสนเทศที่มีความหลากหลาย ทำให้เกิดปัญหาที่ต้องได้รับการแก้ไขเพราะปริมาณของสารสนเทศที่มีจำนวนมหาศาล ซึ่งปัญหาดังกล่าวก็คือ การค้นหาและการเลือกใช้สารสนเทศจะไม่สามารถกระทำได้ในเวลาอันรวดเร็ว การจัดกลุ่มเอกสารข้อความอัตโนมัติจึงกลายเป็นอีกทางเลือกที่สามารถช่วยให้การเข้าถึงสารสนเทศได้เร็วขึ้น จึงได้นำเสนอการจัดกลุ่มเอกสารข้อความภาษาไทยแบบอัตโนมัติบนพื้นฐานของพจนานุกรม และอัลกอริทึมซัพพอร์ตเวคเตอร์แมชชีน ชุดข้อมูลทดสอบคือเอกสารข่าวที่เก็บข้อมูลมาจากหนังสือพิมพ์ไทยรัฐและเดลินิวส์ระหว่างวันที่ 1 มกราคม-31 มีนาคม 2546 โดยจะใช้ข้อมูลจำนวน 1,000 เอกสารสำหรับการเรียนรู้และ 500 เอกสารสำหรับการทดสอบ ภายหลังจากที่ระบบพัฒนาเสร็จสิ้นได้มีการทดสอบประสิทธิภาพของระบบด้วยการวัดค่าเอฟแล้วพบว่าระบบให้ผลของความถูกต้องในการจัดกลุ่มเอกสารอยู่ระหว่างร้อยละ 77.46%-84.70%

บทที่ 3
วิธีการดำเนินงานวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด และเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน โดยใช้ข้อมูลจากแบบสอบถาม เรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร มีวิธีดำเนินการวิจัยดังต่อไปนี้

3.1 ศึกษาวิธีการจำแนกหมวดหมู่

3.2 เครื่องมือในการวิจัย

3.3 ประชากรและกลุ่มตัวอย่างที่ใช้ในการวิจัย

3.4 การเก็บรวบรวมข้อมูล

3.5 ขั้นตอนการจำแนกหมวดหมู่

3.6 การวัดประสิทธิภาพของการจำแนกหมวดหมู่

3.1 ศึกษาวิธีการจำแนกหมวดหมู่

ในงานวิจัยนี้เป็นการจำแนกหมวดหมู่แบบที่ต้องมีตัวอย่างในการเรียนรู้ (Supervised Learning) และข้อมูลที่ใช้เป็นลักษณะข้อความ (Text) แสดงความคิดเห็น ซึ่งวิธีที่เหมาะสมกับการจำแนกหมวดหมู่ ในงานวิจัยนี้ใช้ 2 วิธีเปรียบเทียบกัน คือ วิธีเนอีฟเบย์ (Naïve-Bayes) และซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine : SVM) ซึ่งหลักการของวิธีการเนอีฟเบย์ นี้ใช้การคำนวณความน่าจะเป็น ซึ่งใช้ในการทำนายผลด้วยวิธีเนอีฟเบย์เป็นเทคนิคในการแก้ปัญหาแบบการจำแนกหมวดหมู่ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์การเรียนรู้เบย์อย่างง่าย ส่วนหลักการของวิธีซัพพอร์ตเวคเตอร์แมชชีนนี้เป็นการหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยการใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดยพยายามสร้างเส้นแบ่งกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตทั้งสองกลุ่มมากที่สุด

3.2 เครื่องมือในการวิจัย

เครื่องมือที่ใช้ในการทดลองวิจัยในครั้งนี้ มีดังนี้

3.2.1 แบบสอบถาม เป็นเครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูล ได้สำรวจเรื่อง ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร แบบสอบถามแบ่งออกเป็น 3 ส่วน ดังนี้

ส่วนที่ 1 ข้อมูลพื้นฐานทางประชากรศาสตร์ของกลุ่มตัวอย่าง ได้แก่ เพศ ช่วงอายุ ระดับการศึกษา รายรับต่อเดือน อาชีพปัจจุบันของท่าน

ส่วนที่ 2 ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยภาครัฐและเอกชน แบ่งเป็น 4 ด้าน ได้แก่ *ด้านราคา* เป็นค่าใช้จ่ายทั้งหมดในการศึกษาเล่าเรียนตลอดการศึกษา รวมทั้งค่าดำรงชีพระหว่างการศึกษา ค่าใช้จ่ายส่วนตัว

ด้านสถานที่ เป็นบริเวณโดยรอบ ความสะอาด ความสวยงามของสถานที่ อาคารเรียน สิ่งแวดล้อมต่างๆในสถานการศึกษา

ด้านความรู้และทักษะที่ได้รับ เป็นความรู้และทักษะต่างๆทางด้านวิชาการที่ได้รับในระหว่างการศึกษา

ด้านความยอมรับในสังคม เป็นความยอมรับทั้งหลังจบการศึกษาแล้ว การได้รับการตอบรับจากสังคมและสถานประกอบ

ส่วนที่ 3 ข้อคิดเห็นเพิ่มเติม

3.2.2 Microsoft Office Excel 2007 สำหรับเก็บข้อมูลที่ได้สำรวจจากแบบสอบถามแสดงความคิดเห็นแบบปลายเปิด

3.2.3 โปรแกรม Java EE 6 SDK ใช้ในการเตรียมเอกสารหรือข้อความสำหรับใช้ในการรันโปรแกรม WEKA 3.6.2

3.2.4 โปรแกรม WEKA 3.6.2 ใช้สำหรับการจำแนกหมวดหมู่และทำนายผล โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน

3.3 ประชากรและกลุ่มตัวอย่างที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการเก็บรวบรวมข้อมูล คือ แบบสอบถามความคิดเห็น โดยได้สำรวจเรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร ซึ่งมีประชากรและกลุ่มตัวอย่างที่ใช้ในการวิจัย ดังนี้

3.3.1 ประชากรที่ใช้ในการศึกษาครั้งนี้ คือ ประชากรทั้งเพศชายและเพศหญิงในเขตกรุงเทพมหานคร ผู้ศึกษาจะทำการศึกษาเฉพาะประชากรในเขตกรุงเทพมหานคร ซึ่งมีจำนวนทั้งสิ้น 5,710,883 คน (สำนักทะเบียนกลาง กรมการปกครอง กระทรวงมหาดไทย, 2552)

3.3.2 กลุ่มตัวอย่างของการวิจัยครั้งนี้ คือ ประชาชนที่อาศัยอยู่ในเขตกรุงเทพมหานคร การคำนวณหาขนาดกลุ่มตัวอย่างเพื่อการวิจัยโดยใช้ตารางสุ่มตัวอย่างของทาโร่ยามาเน (Taro Yamane, 1973) ที่ระดับความเชื่อมั่นร้อยละ 95 ที่ระดับความคลาดเคลื่อนร้อยละ 5

สูตร $n = \frac{N}{1+N(e)^2}$

n คือ ขนาดของกลุ่มตัวอย่าง
N คือ ขนาดของประชากรที่ใช้ในการวิจัย
e คือ ค่าเปอร์เซ็นต์ความคลาดเคลื่อนจากการสุ่มตัวอย่าง

$n = \frac{5,710,883}{1+5,710,883(0.05)^2}$

$n \approx 400$

ดังนั้น ขนาดของประชากรในเขตกรุงเทพมหานครจำนวนทั้งสิ้น 5,710,883 คน เก็บข้อมูลโดยการสุ่มขนาดของกลุ่มตัวอย่างที่สามารถเป็นตัวแทนของประชากรจำนวน 400 คน โดยให้เกิดความคลาดเคลื่อนได้ร้อยละ 5 ที่ระดับความเชื่อมั่นร้อยละ 95

3.3.3 การสุ่มกลุ่มตัวอย่าง ผู้วิจัยใช้วิธีการสุ่มตัวอย่างโดยใช้ทฤษฎีความน่าจะเป็น (Probability Sampling) โดยใช้การสุ่มตัวอย่างแบบหลายขั้นตอน (Multi-Stage Sampling) แบ่งได้ดังนี้

3.3.3.1 การสุ่มตัวอย่างแบบแบ่งพื้นที่ (Cluster Sampling) ตามโซนพื้นที่ของจังหวัดกรุงเทพมหานคร ซึ่งมีทั้งหมด 12 พื้นที่ (สำนักงานปลัดกรุงเทพมหานคร ศูนย์ข้อมูลกรุงเทพมหานคร) ได้แก่

โซนพื้นที่ กท.1 หรือ เขตอนุรักษ์เมืองเก่ากรุงรัตนโกสินทร์
โซนพื้นที่ กท.2 หรือ เขตศูนย์กลางธุรกิจ
โซนพื้นที่ กท.3 หรือ เขตเศรษฐกิจใหม่
โซนพื้นที่ กท.4 หรือ เขตเศรษฐกิจใหม่ริมแม่น้ำเจ้าพระยา

โซนพื้นที่ กท.5 หรือ เขตอนุรักษ์เมืองเก่ากรุงธนบุรี

โซนพื้นที่ กท.6 หรือ เขตเศรษฐกิจการจ้างงานใหม่

โซนพื้นที่ กท.7 หรือ เขตที่อยู่อาศัยรองรับการขยายตัวของเมือง ด้านตะวันออกตอนเหนือ

โซนพื้นที่ กท.8 หรือ เขตที่อยู่อาศัยรองรับการขยายตัวของเมือง ด้านตะวันออกตอนใต้

โซนพื้นที่ กท.9 หรือ เขตเกษตรกรรมและที่อยู่อาศัยสภาพแวดล้อมดี

โซนพื้นที่ กท.10 หรือ เขตศูนย์ชุมชนชานเมือง รองรับสนามบิน

โซนพื้นที่ กท.11 หรือ เขตเกษตรกรรมและที่อยู่อาศัยผสมพื้นที่เกษตรกรรม

โซนพื้นที่ กท.12 หรือ เขตเกษตรกรรม อุตสาหกรรม ที่อยู่อาศัยและแหล่งท่องเที่ยว

3.3.3.2 การสุ่มแบบเจาะจง เพื่อเลือกโซนพื้นที่ จากทั้งหมด 12 โซน มา 4 โซน คือ

โซนพื้นที่ กท.1 หรือ เขตอนุรักษ์เมืองเก่ากรุงรัตนโกสินทร์ ประกอบด้วยเขตดุสิต เขตพระนคร เขตป้อมปราบศรัตรูพ่าย และเขตสัมพันธวงศ์

โซนพื้นที่ กท.3 หรือ เขตเศรษฐกิจใหม่ ประกอบด้วย เขตจตุจักร เขตบางซื่อเขตพญาไท เขตดินแดง เขตห้วยขวาง เขตราชเทวี

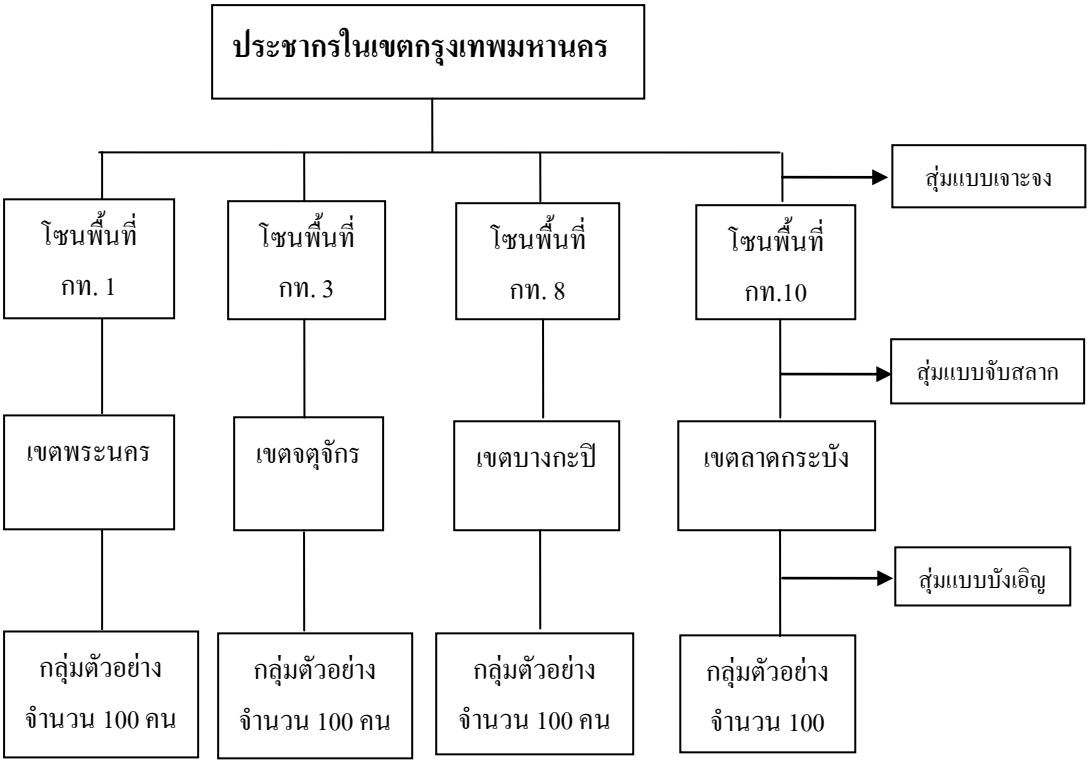
โซนพื้นที่ กท.8 หรือ เขตที่อยู่อาศัยรองรับการขยายตัวของเมือง ด้านตะวันออกตอนใต้ ประกอบด้วย เขตบางกะปิ เขตคันนายาว เขตวังทองหลาง เขตบึงกุ่ม เขตสะพานสูง และเขตสวนหลวง

โซนพื้นที่ กท.10 หรือ เขตศูนย์ชุมชนชานเมือง รองรับสนามบิน ประกอบด้วย เขตลาดกระบัง เขตมีนบุรี และเขตประเวศ

3.1.3.3 การสุ่มแบบจับฉลาก เลือกเขตที่อยู่ในแต่ละโซนมา 1 เขต และทำการสุ่มตัวอย่างที่เป็นตัวแทนในแต่ละเขตโดยการสุ่มตัวอย่างแบบบังเอิญ โดยเก็บตัวอย่างในเขตชุมชนแต่ละเขต แล้วแจกแบบสอบถามเขตละ 100 ชุด จำนวนทั้งหมด 400 ชุด แสดงดังภาพที่ 3-1

ดังนั้น การเก็บกลุ่มตัวอย่าง 400 คน

โดยเก็บกลุ่มตัวอย่าง	เขตพระนคร	100 คน
	เขตจตุจักร	100 คน
	เขตบางกะปิ	100 คน
	เขตลาดกระบัง	100 คน



ภาพที่ 3-1 การสุ่มตัวอย่างเป็นตัวแทนในแต่ละเขต

3.4 การเก็บรวบรวมข้อมูล

ในการวิจัยครั้งนี้ผู้วิจัยได้ดำเนินการเก็บรวบรวมข้อมูลตามขั้นตอนดังนี้

3.4.1 จัดเตรียมเครื่องมือตามจำนวนกลุ่มตัวอย่างจากแบบสอบถาม จำนวน 400 ชุด เพื่อเก็บข้อมูล โดยทำการแจกแบบสอบถามให้กลุ่มตัวอย่างได้แสดงความคิดเห็น

3.4.2 เก็บรวบรวมแบบสอบถามคืนจากกลุ่มตัวอย่าง ตรวจสอบความสมบูรณ์ของข้อมูลให้กลุ่มตัวอย่างกรอกข้อมูลทั้งหมดให้ครบถ้วน

3.4.3 ทำการกรอกข้อความแสดงความคิดเห็น จากแบบสอบถามจำนวน 400 ชุด เก็บลงใน Microsoft Office Excel 2007 แล้วนำมาใช้ในขั้นตอนการจำแนกหมวดหมู่ต่อไป

3.5 ขั้นตอนการจำแนกหมวดหมู่

ขั้นตอนของการจำแนกหมวดหมู่โดยภาพรวมทั้งระบบ แบ่งเป็น 5 ขั้นตอน ดังต่อไปนี้

3.5.1 จัดเตรียมและการคัดเลือกข้อมูลเบื้องต้น (Preprocessing) นำข้อมูลที่ได้จากแบบสอบถามและการคัดเลือกข้อมูล เนื่องจากบางข้อความไม่สามารถระบุชั้วได้ต้องมีการคัดออกร่วมทั้งการแก้ไขคำที่เขียนผิด

3.5.2 เตรียมข้อมูลในรูปแบบไฟล์ XML ข้อมูลที่ได้จากการสำรวจจากแบบสอบถามทั้ง 4 ด้าน ได้แก่ ด้านราคา ด้านสถานที่ ด้านความรู้และทักษะที่ได้รับ และด้านความยอมรับในสังคม นำข้อมูลแต่ละด้านมาแบ่งเป็น 2 ส่วน (อัตราส่วน 8:2) คือ ข้อมูลสำหรับเรียนรู้ (Training) จำนวน 80% และข้อมูลทดสอบ (Test) จำนวน 20%

การเตรียมข้อมูลสำหรับเรียนรู้ เป็นข้อมูลที่ใช้ในการเรียนรู้จากอัลกอริทึมเพื่อใช้ในการสร้างโมเดล โดยนำข้อมูล Training มาระบุขั้ว (Polar Words) เป็นคำที่บ่งบอกถึงระดับความคิดเห็น แบ่งเป็น 3 หมวดหมู่ คือ เชิงบวก (Positive) กลาง (Medium) และลบ (Negative) โดยข้อมูลจะต้องจัดเตรียมให้อยู่ใน XML tag ดังตัวอย่างภาพที่ 3-2

```
<doc>
<category> Positive</category>
<content>สถาบันการศึกษาทางภาครัฐจะมีชื่อเสียงและการยอมรับจากสังคมมาก</content>
</doc>
```

ภาพที่ 3-2 การเตรียมข้อมูลสำหรับเรียนรู้ของข้อคิดเห็นในรูปแบบไฟล์ XML

จากภาพที่ 3-2 เป็นตัวอย่างแสดงผลการเตรียมข้อมูลสำหรับเรียนรู้ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านการยอมรับในสังคมที่มีต่อภาครัฐในรูปแบบไฟล์ XML เห็นได้ว่าจะมีการใส่เนื้อหาลงใน tap<content> และกำกับคำระบุขั้วใน tap <category>

การเตรียมข้อมูลสำหรับทดสอบ เป็นข้อมูลที่ไม่ได้มีการระบุขั้ว ใช้สำหรับการทำนายและทดสอบประสิทธิภาพของการจำแนกหมวดหมู่ โดยข้อมูลจะต้องจัดเตรียมให้อยู่ใน XML tag ดังตัวอย่างภาพที่ 3-3

```
<doc>
<category>?</category>
<content>ได้รับความยอมรับด้านสถาบันน้อยกว่าเอกชน</content>
</doc>
```

ภาพที่ 3-3 การเตรียมข้อมูลทดสอบของข้อคิดเห็นในรูปแบบไฟล์ XML

จากภาพที่ 3-3 เป็นตัวอย่างแสดงผลการเตรียมข้อมูลทดสอบของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านการยอมรับในสังคมที่มีต่อภาครัฐในรูปแบบไฟล์ XML เห็นได้ว่าจะมีการใส่เนื้อหาข้อความลงใน tap <content> และใส่เครื่องหมาย ? ระบุขั้วใน tap <category>

กำหนดให้การสร้างชื่อไฟล์ XML สำหรับการวิเคราะห์ในรูปแบบ XML ดังตารางที่ 3-1

ตารางที่ 3-1 การตั้งชื่อไฟล์สำหรับการวิเคราะห์ในรูปแบบ XML

ปัจจัย	ภาค	ข้อมูลสำหรับเรียนรู้ (Train)	ข้อมูลทดสอบ (Test)
ด้านราคา	ภาครัฐ	price_gov.train.xml	price_gov.test.xml
	เอกชน	price_private.train.xml	price_ private.test.xml
ด้านสถานที่	ภาครัฐ	location_gov.train.xml	location _gov.test.xml
	เอกชน	location _private.train.xml	location _ private.test.xml
ด้านความรู้และทักษะที่ได้รับ	ภาครัฐ	skill_gov.train.xml	skill_gov.test.xml
	เอกชน	skill_private.train.xml	skill_ private.test.xml
ด้านความยอมรับในสังคม	ภาครัฐ	accept_gov.train.xml	accept_gov.test.xml
	เอกชน	accept _private.train.xml	accept_ private.test.xml

3.5.3 การเลือกคำ (Feature Selection) เป็นการเลือกคำที่มีผลต่อการระบุขั้วและเป็นการตัดคำที่ไม่มีผลการระบุขั้วออกโดยอัตโนมัติ ทำได้โดยนำไฟล์ข้อมูลสำหรับเรียนรู้ (Train) ที่อยู่ในรูป XML คือ *.train.xml มาผ่านขั้นตอนต่อไปนี้

3.5.3.1 การตัดคำ (Word Segmentation) ในงานวิจัยนี้จะใช้โปรแกรม LexToPlus.java เป็นโปรแกรมตัดคำแบบอิงพจนานุกรมเล็กซิตรอน (Lexitron) จำนวน 35,936 คำ โดยเลือกแบบยาวที่สุด (Longest Matching Dictionary-Based) และมีอัลกอริทึมที่ช่วยรวบคำไม่รู้จักเข้ามาเป็นคำ (Unknown Merging) โปรแกรมตัดคำนี้สามารถรองรับได้ทั้งข้อความที่เป็นภาษาไทยและอังกฤษ

3.5.3.2 การกำจัดคำหยุด (Stopword Removal) เป็นการจำกัดคำหยุดออกโดยใช้รายการคำหยุดภาษาไทยและอังกฤษ โดยเป็นคำหยุดภาษาไทย จำนวน 115 คำ คำหยุดอังกฤษ จำนวน 241 คำ โดยนำมาเฉพาะคำที่เป็นคำสันธาน คำบุพบท คำวิเศษณ์ คำอุทาน และคำสรรพนาม

3.5.3.3 การหารากศัพท์ (Stemmer) โดยใช้ Stemmer.java เป็นโปรแกรมที่ทำการ stem หรือแปลงคำในภาษาอังกฤษให้อยู่ในรูปแบบที่เป็นรากศัพท์

ในขั้นตอนการเลือกคุณลักษณะคำ (Feature Selection) จะได้ไฟล์ 2 ไฟล์ คือ *.train.category เป็นรายการของหมวดหมู่ความคิดเห็น และ *.train.keyword เป็นรายการของคำที่ผ่านการตัดคำ กำจัดคำหยุด การหารากศัพท์แล้ว ถ้ารายการคำใน *.train.keyword ยังไม่ถูกต้อง ต้องสร้างรายการคำไม่รู้จัก (unknown) เพิ่มในรายการ ซึ่งไม่มีในพจนานุกรมเล็กซิตรอน เพื่อให้โปรแกรม LexToPlus ตัดคำได้ถูกต้องมากยิ่งขึ้น

กำหนดให้ *.train.xml คือ ข้อมูลสำหรับเรียนรู้ที่อยู่ในรูปแบบไฟล์ XML

*.test.xml คือ ข้อมูลทดสอบที่อยู่ในรูปแบบไฟล์ XML

* คือ ให้ใส่ชื่อปัจจัยแต่ละด้านทั้ง 4 ด้าน

*.train.keyword คือ คุณลักษณะที่สกัดได้จากข้อมูลสำหรับเรียนรู้

*.train.category คือ หมวดหมู่ของ Polar Word สกัดได้จากข้อมูลสำหรับเรียนรู้

การเลือกคุณลักษณะคำ (Feature Selection) ในการทดลองมีวิธีการดังนี้

1) เริ่มต้นใส่ไฟล์ทั้งหมดในโฟล์เดอร์เดียวกัน คือ โฟล์เดอร์ ThaiText จากนั้น compile โดยใช้คำสั่ง

```
> javac *.java
```

2) ทำการเลือกคำโดยใช้คำสั่งต่อไปนี้

```
> java FeatureSelection 0 price_gov.train.xml
```

ค่า 0 เป็นการระบุค่าความถี่ขั้นต่ำของคำที่ต้องการ ค่า 0 หมายถึงเอาทุกคำที่เกิดขึ้น ในกรณีใช้งานจริง

ผลลัพธ์ที่ได้จากการเลือก Feature ดังภาพที่ 3-4

```
C:\ThaiText>java FeatureSelection 0 price_gov.train.xml
Status: Reading English stopwords ...
Status: Reading Thai stopwords ...
Status: Analyzing keywords ......

Status: Total number of documents = 314
Status: Total number of categories = 3
Status: Total number of initial words = 302
Status: Total number of selected words = 302
```

ภาพที่ 3-4 ผลลัพธ์ที่ได้จากการเลือก Feature

จากภาพที่ 3-4 บรรทัดแรกเป็นรูปแบบคำสั่งของการเลือกคำ (Feature Section) โดยใช้ไฟล์ข้อมูล price_gov.train.xml ผลการรัน พบว่ามีเอกสารจำนวนทั้งหมด 314 เอกสาร คำที่มีขั้วมี 3 หมวดหมู่ จำนวนคำที่เลือก (Feature Section) จากค่าความถี่ขั้นต่ำ 0 มี 302 คำ ในขั้นตอนนี้โปรแกรมจะสร้างไฟล์ 2 ไฟล์ ได้แก่ price_gov.train.category ซึ่งเป็นรายการของหมวดหมู่ที่โปรแกรมอ่านได้จาก price_gov.train.txt และ price_gov.train.keyword ซึ่งเป็นรายการของคำที่ผ่านค่าความถี่ขั้นต่ำ (หากคำใน price_gov.train.keyword ไม่ถูกต้อง แสดงว่าคำนั้นเป็นคำไม่รู้จัก ผู้ใช้สามารถแก้ไขและเพิ่มคำไม่รู้จักในไฟล์ unknown.txt ได้ จากนั้น run โปรแกรมใหม่อีกครั้ง)

3.5.4 การแปลงไฟล์ให้อยู่ในรูปแบบไฟล์ ARFF เป็นการแปลงไฟล์ทั้งข้อมูลสำหรับเรียนรู้และข้อมูลสำหรับทดสอบให้อยู่ในรูปไฟล์ ARFF ซึ่งสามารถรันด้วยโปรแกรม WEKA ได้ โดยทำการแปลงทั้งไฟล์ trian.xml และ test.xml มีกระบวนการผ่านการการตัดคำ (Word Segmentation) การตัดคำ (Word Segmentation) การหารากศัพท์ (Stemmer) และการแปลงไฟล์ (Formatting) ในรูปแบบไฟล์ ARFF

การแปลงไฟล์เป็น ARFF (FormatArff.java) ใช้รูปแบบคำสั่งต่อไปนี้

> java FormatArff price_gov.train.xml price_gov.train.category price_gov.train.keyword

รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของไฟล์ข้อมูล price_gov train.xml แสดงดังภาพที่ 3-5

```
C:\ThaiText>java FormatArff price_gov.train.xml price_gov.train.category price_g
ov.train.keyword
Status: Reading English stopwords ...
Status: Reading Thai stopwords ...
Status: Number of categories = 3
Status: Number of keywords = 302
Status: Formatting file, price_gov.train.xml...
```

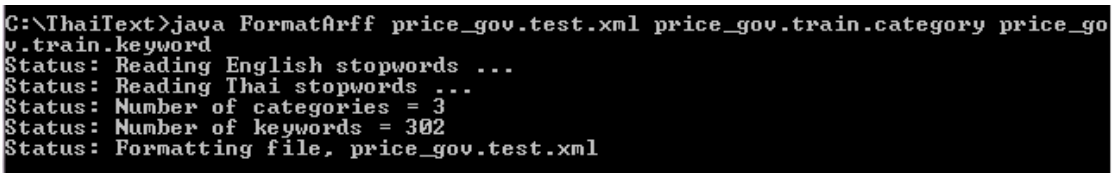
ภาพที่ 3-5 รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของข้อมูลสำหรับเรียนรู้

จากภาพที่ 3-5 บรรทัดแรกเป็นรูปแบบคำสั่งการแปลงไฟล์เป็น ARFF ของข้อมูลสำหรับเรียนรู้ ซึ่งตัวอย่างคือข้อมูลข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ และผลการรัน พบว่าในขั้นตอนนี้โปรแกรมจะสร้างไฟล์ ได้แก่ price_gov.train.arff ซึ่งพร้อมที่จะใช้ในโปรแกรม WEKA ได้ต่อไป

หลังจากนั้นสามารถสร้างไฟล์ ARFF สำหรับ price_gov.test.xml ได้โดยใช้คำสั่งต่อไปนี้

> java FormatArff price_gov.test.xml price_gov.train.category price_gov.train.keyword

รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของไฟล์ข้อมูล price_gov test.xml แสดงดังภาพที่ 3-6

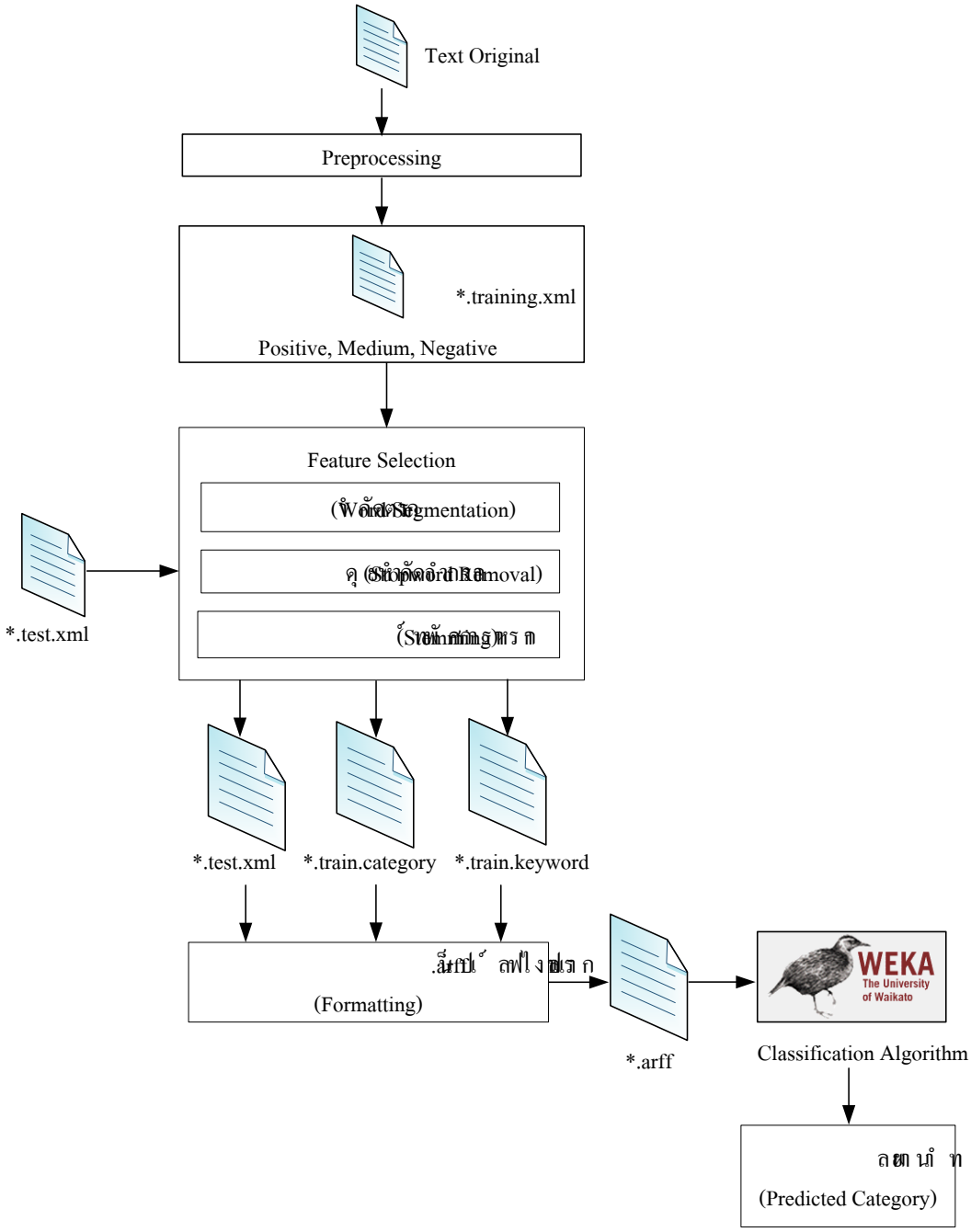


ภาพที่ 3-6 รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของข้อมูลทดสอบ

จากภาพที่ 3-6 บรรทัดแรกเป็นรูปแบบคำสั่งการแปลงไฟล์เป็น ARFF ของข้อมูลทดสอบ ซึ่งตัวอย่างคือข้อมูลข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ และผลการรัน พบว่าในขั้นตอนนี้โปรแกรมจะสร้างไฟล์ ได้แก่ price_gov.test.arff ซึ่งพร้อมที่จะใช้ในโปรแกรม WEKA ได้ต่อไป

3.5.5 การจำแนกหมวดหมู่ด้วยอัลกอริทึม (Classification Algorithm) ขั้นตอนนี้เป็นการประยุกต์ใช้อัลกอริทึมเพื่อเรียนรู้และจำแนกหมวดหมู่ข้อมูล ทำการเรียนรู้ข้อมูลโดยใช้ข้อมูลสำหรับเรียนรู้ที่อยู่ในรูปแบบไฟล์ .arff โดยกำกับระบุขั้ว (Polar Words) ทั้ง 3 หมวดหมู่ คือ เชิงบวก (Positive) กลาง (Medium) และลบ (Negative) จากนั้นนำข้อมูลที่ได้เข้าเรียนรู้และสร้างโมเดล ซึ่งวิธีที่ใช้ในการเรียนรู้และสร้างโมเดล มี 2 วิธี คือ วิธีเนอีฟเบย์และซัพพอร์ตเวกเตอร์แมชชีน แล้วทำการทำนายผล (Predicted Category) โดยใช้ข้อมูลทดสอบที่อยู่ในรูปแบบไฟล์ .arff เข้าโปรแกรม Weka เพื่อวัดประสิทธิภาพของการจำแนกหมวดหมู่ในขั้นตอนต่อไป

ขั้นตอนการจำแนกหมวดหมู่ของของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน ดังภาพที่ 3-7



ภาพที่ 3-7 ขั้นตอนการจำแนกหมวดหมู่ข้อคิดเห็น

3.6 การวัดประสิทธิภาพของการจำแนกหมวดหมู่

การประเมินความสามารถของระบบการจำแนกหมวดหมู่นั้น โดยทั่วๆ วัดประสิทธิผลของการจำแนกหมวดหมู่นิยมใช้วิธีการตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่า *F -measure* กรณีที่มีหมวดหมู่มากกว่า 2 หมวดหมู่ สามารถคำนวณประสิทธิภาพได้ดังต่อไปนี้ (Ozgur, Ozgur and Gungor, 2005)

วิธีการคำนวณค่าความแม่นยำ (*P*) ดังสมการที่ (3-1)

$$P = \frac{\sum_{i=1}^{M} TP_j}{\sum_{i=1}^{M} (TP_i + FP_i)} \quad (3\text{-}1)$$

วิธีการคำนวณค่าความระลึก (*R*) ดังสมการที่ (3-2)

$$R = \frac{\sum_{i=1}^{M} TP_j}{\sum_{i=1}^{M} (TP_i + FN_i)} \quad (3\text{-}2)$$

เมื่อ *P* คือ ค่าความแม่นยำของการจำแนกหมวดหมู่

R คือ ค่าความระลึกของการจำแนกหมวดหมู่

M คือ จำนวนของหมวดหมู่

การพิจารณาประสิทธิผลของการจำแนกหมวดหมู่โดยดูค่าความแม่นยำ หรือค่าความระลึกเพียงอย่างเดียวนั้น อาจทำให้ประเมินประสิทธิผลได้ไม่ถูกต้องนัก ทั้งนี้การจำแนกหมวดหมู่ที่ให้ค่าความแม่นยำสูง อาจให้ค่าความระลึกต่ำได้ หากมีจำนวน *FN* มาก หรือให้ค่าความแม่นยำต่ำ แต่ให้ค่าความระลึกสูง หากมีจำนวน *FP* มาก ดังนั้น จึงมีวิธีการวัดประสิทธิผลที่นำค่าทั้งสองมาคำนวณร่วมกัน เพื่อให้การประเมินค่ามีความถูกต้องมาก มากยิ่งขึ้น นั่นคือ การวัดค่า *F -measure* โดยมีวิธีการคำนวณ ดังต่อไปนี้

$$F = \frac{2PR}{P + R} \quad (3\text{-}3)$$

ค่า *F* จะมีค่าอยู่ในช่วงระหว่าง 0-1

โดย ค่า *F* ที่มีค่าสูง หมายถึง มีประสิทธิผลในการจำแนกหมวดหมู่สูง

ค่า *F* ที่มีค่าต่ำ หมายถึง มีประสิทธิผลในการจำแนกหมวดหมู่ต่ำ

บทที่ 4
ผลการวิจัย

การเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน เรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร ได้ผลการวิจัยดังต่อไปนี้

4.1 ผลการเก็บข้อมูลจากแบบสอบถามและการคัดเลือกข้อมูล

4.2 ผลการจำแนกหมวดหมู่ของข้อคิดเห็น

4.3 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่

4.1 ผลการเก็บข้อมูลจากแบบสอบถามและการคัดเลือกข้อมูล

ประชากรที่ใช้ในการศึกษาครั้งนี้ คือ ประชากรทั้งเพศชายและเพศหญิงในเขตกรุงเทพมหานคร 5,710,883 คน ทำการสุ่มจำนวนกลุ่มตัวอย่าง 400 คน นั้นคือ ในการเก็บข้อมูลจากแบบสอบถามเป็นสอบถามความคิดเห็นแบบปลายเปิด เรื่อง ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร จำนวน 400 ชุด แบบสอบถามแบ่งออกเป็น 3 ส่วน ดังนี้

ส่วนที่ 1 ข้อมูลพื้นฐานทางประชากรศาสตร์ของกลุ่มตัวอย่าง

ส่วนที่ 2 ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยภาครัฐและเอกชน แบ่งเป็น 4 ด้าน ได้แก่ ด้านราคา ด้านสถานที่ ด้านความรู้และทักษะที่ได้รับ และด้านความยอมรับในสังคม

ส่วนที่ 3 ข้อคิดเห็นเพิ่มเติม

ข้อมูลจากแบบสอบถามที่นำมาใช้ในการวิเคราะห์ คือ ส่วนที่ 2 ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยภาครัฐและเอกชนแบ่งเป็น 4 ด้าน ได้แก่ ด้านราคา ด้านสถานที่ ด้านความรู้และทักษะที่ได้รับ และด้านความยอมรับในสังคม

ผลการสำรวจเก็บข้อมูล พบว่ามีผู้ตอบแบบสอบถามไม่ครบถ้วนบางด้าน และข้อคิดเห็นบางข้อความไม่สามารถระบุขั้ว (Polar Words) ของข้อคิดเห็นได้ว่าเป็นความคิดเห็นเชิงบวก กลาง หรือ ลบ จึงจำเป็นต้องมีการคัดเลือกข้อความที่ไม่สามารถระบุขั้วได้ออก ตัวอย่างของลักษณะข้อความที่ไม่สามารถระบุขั้วได้ทางด้านความยอมรับในสังคม (ภาครัฐ) ได้แก่ “ต้องมีความสามารถทางการศึกษา” “ด้านความยอมรับขึ้นอยู่กับคณะที่ศึกษา” “ยอมรับเป็นกรณี” เป็นต้น

หลังการคัดกรองข้อมูลจากแบบสอบถาม 400 ชุด ทำให้ได้ข้อความสำหรับการวิเคราะห์แต่ละด้านไม่เท่ากัน ซึ่งสามารถแบ่งจำนวนข้อมูลสำหรับเรียนรู้ (Train) 80% และข้อมูลทดสอบ (Test) 20% เพื่อใช้ในการวิเคราะห์ได้ ดังตารางที่ 4-1

ตารางที่ 4-1 จำนวนข้อความจากแบบสอบถาม 400 ชุด จำนวนข้อมูลสำหรับเรียนรู้และจำนวนข้อมูลทดสอบ

ข้อคิดเห็น	จำนวนข้อความจากแบบสอบถาม 400 ชุด		จำนวนข้อมูลสำหรับเรียนรู้ (Train)		จำนวนข้อมูลทดสอบ (Test)	
	ภาครัฐ	เอกชน	ภาครัฐ	เอกชน	ภาครัฐ	เอกชน
ด้านราคา	392	389	314	311	78	78
ด้านสถานที่	382	378	306	302	76	76
ด้านความรู้และทักษะที่ได้รับ	387	385	310	308	77	77
ด้านความยอมรับในสังคม	374	388	299	310	75	78

เมื่อเก็บข้อมูลแสดงความคิดเห็นจากแบบทดสอบแบบปลายเปิด เรื่อง ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานครแล้ว จากนั้นกรอกข้อมูลลงใน Microsoft Office Excel 2007 และทำการระบุขั้ว (Polar Words) เป็นคำที่บ่งบอกถึงระดับความคิดเห็น แบ่งเป็น 3 หมวดหมู่ คือ เชิงบวก (Positive) กลาง (Medium) และลบ (Negative) ซึ่งได้แสดงตัวอย่างของข้อความแสดงความคิดเห็นและการระบุขั้ว ดังตารางที่ 4-2

ตารางที่ 4-2 ตัวอย่างข้อความแสดงความคิดเห็นและการระบุขั้ว

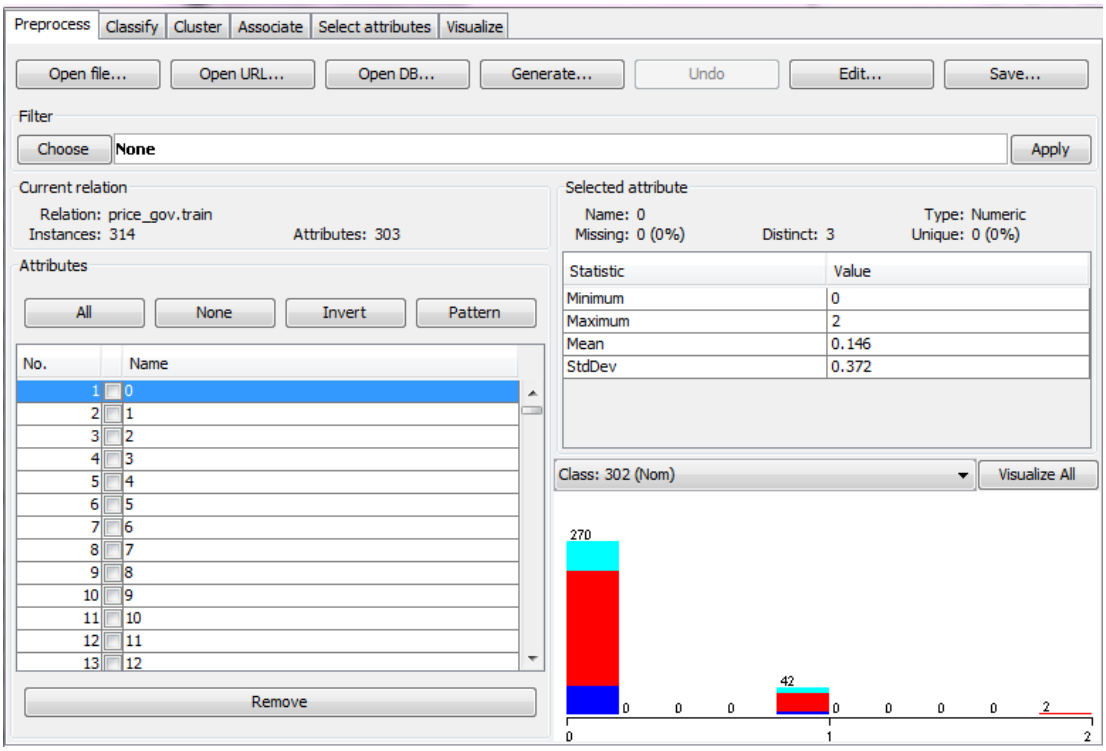
ข้อคิดเห็น	ตัวอย่างข้อความแสดงความคิดเห็นและการระบุขั้ว	
	ภาครัฐ	เอกชน
ด้านราคา	มีราคาเล่าเรียนที่ถูกกว่า<Positive>	ราคาถือว่าไม่แพงเกินไป<Positive>
	ราคาพอใช้ได้<Medium>	พอๆกับรัฐบาล<Medium>
	บางที่แพงมาก ๆ แพงกว่าเอกชนซะอีกๆ บางที่ราคาเหมาะสม และบางครั้งถูกกว่าภาครัฐด้วยซ้ำ<Negative>	ค่าเกี่ยวกับการศึกษาก็มีมากพอสมควร ค่าใช้จ่ายส่วนตัวค่อนข้างสูง<Negative>

ตารางที่ 4-2 ต่อ

ข้อคิดเห็น	ตัวอย่างข้อความแสดงความคิดเห็นและการระบุขั้ว	
	ภาครัฐ	เอกชน
ด้านสถานที่	แน่นอนมหาวิทยาลัยของรัฐจะมีสถานที่ที่ใหญ่เพียงพอต่อจำนวน<Positive>	สถานที่ดูหรูหรา ดูดีกว่าภาครัฐ <Positive>
	มีขนาดกลาง<Medium>	ก็จะคล้ายๆกันกับมหาวิทยาลัยของรัฐ แต่ละที่จะมีขนาดของสถานที่ที่แตกต่างกันไป<Medium>
	สถานที่จะเก่าและไม่ค่อยได้รับการดูแล <Negative>	บางทีก็มีพื้นที่มากอาจจะอยู่ห่างไกล สถานที่สำคัญมาก เดินทางลำบาก <Negative>
ด้านความรู้และทักษะที่ได้รับ	ความรู้มากมาย ทักษะที่ดี<Positive>	บุคลากรเยอะได้ความรู้มากมาย <Positive>
	มีความรู้เท่าเทียมกัน ที่อยู่นักศึกษาเองว่ามีความสนใจมากน้อยเพียงใด<Medium>	ให้ความรู้ไม่แตกต่างกัน<Medium>
	ได้รับความรู้เท่าเอกชนแต่เทคโนโลยีจะด้อยกว่า<Negative>	ความรู้และทักษะน้อยกว่า <Negative>
ด้านความยอมรับในสังคม	สังคมจะยอมรับสถาบันของรัฐบาลมากกว่าเอกชน รวมทั้งสถานประกอบการด้วย<Positive>	ยอมรับในสังคมได้ดี ขึ้นกับผู้เรียน <Positive>
	พอๆ กัน<Medium>	เท่าเทียมกัน<Medium>
	ได้รับการยอมรับน้อยถ้าสถาบันไม่มีชื่อ <Negative>	สังคมและสถานประกอบการจะยอมรับสถาบันเอกชนน้อยว่า <Negative>

4.2 ผลการจำแนกหมวดหมู่ของข้อคิดเห็น

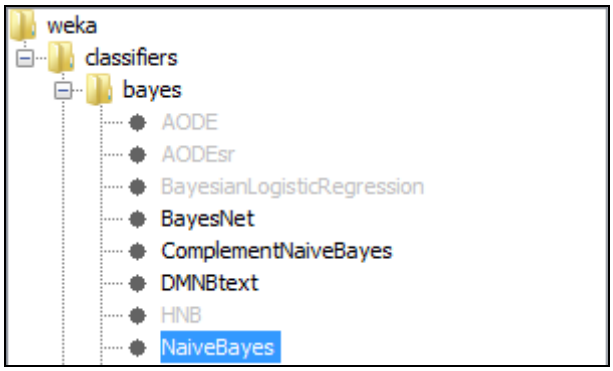
ผลการรันข้อมูลเพื่อการจำแนกหมวดหมู่ด้วยอัลกอริทึม การรันข้อมูลเพื่อการจำแนกหมวดหมู่ด้วยอัลกอริทึมด้วยโปรแกรม WEKA 3.6.2 โดยเปิดไฟล์เพื่อเลือกไฟล์ที่ต้องการวิเคราะห์ตัวอย่างที่นำมาแสดงเป็นข้อมูลที่ได้จากข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ ดังภาพที่ 4-1



ภาพที่ 4-1 ผลการเปิดไฟล์ที่ต้องการวิเคราะห์ด้วยโปรแกรม WEKA

จากภาพที่ 4-1 Current relation ระบุว่า Relation คือ price_gov.test ซึ่งเป็นการเปิดไฟล์ price_gov.test.arff ซึ่งมี จำนวนเอกสาร (Instance) 314 แถว และมี Attributes จำนวน 303 คอลัมน์

หลังจากนำข้อมูลสำหรับการเรียนรู้เข้าสู่โปรแกรม WEKA แล้วเลือกวิธีการเรียนรู้ข้อมูลและสร้างโมเดลด้วยวิธีเนอีฟเบย์ สำหรับการจำแนกหมวดหมู่ข้อคิดเห็นด้วยโปรแกรม WEKA โดยเลือกดังภาพที่ 4-2



ภาพที่ 4-2 การเลือกใช้อัลกอริทึมเนอีฟเบย์ในการเรียนรู้และสร้างโมเดล

ผลการนำข้อมูลสำหรับเรียนรู้และสร้างโมเดลด้วยอัลกอริทึมเนอีฟเบย์ ซึ่งรันในโปรแกรม WEKA ดังภาพที่ 4-3

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     price_gov.train
Instances:    314
Attributes:   303              [list of attributes omitted]
Test mode:    evaluate on training data
=== Classifier model (full training set) ===
Attribute        Medium Positive Negative
                (0.16)   (0.67)   (0.17)
Time taken to build model: 0.08 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances         265               84.3949 %
Incorrectly Classified Instances        49               15.6051 %
Kappa statistic                          0.6726
Mean absolute error                      0.1403
Root mean squared error                  0.2794
Relative absolute error                 42.4274 %
Root relative squared error             68.8188 %
Total Number of Instances              314
```

ภาพที่ 4-3 ผลการนำรันข้อมูลสำหรับเรียนรู้และสร้างโมเดลด้วยวิธีเนอีฟเบย์

จากภาพที่ 4-3 เป็นผลการรันข้อมูลสำหรับเรียนรู้ซึ่งเป็นข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ มีข้อมูลจำนวน 314 ตัวอย่าง ด้วยอัลกอริทึมเนอีฟเบย์ พบว่า มี Attribute จำนวน 3 Attribute ได้แก่ Medium, Positive และ Negative ได้ค่า Correctly Classified Instances เท่ากับ 84.39%

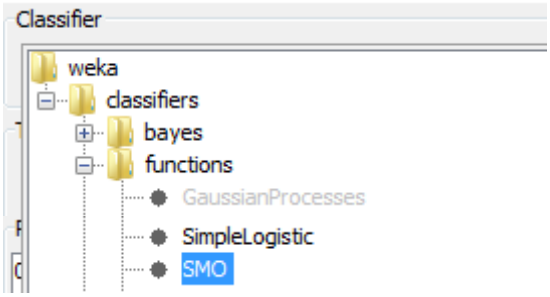
ผลการทำนายด้วยข้อมูลทดสอบ โดยตัวอย่างนี้เป็นข้อมูลข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ โดยใช้ข้อมูลทดสอบ จำนวน 78 ตัวอย่าง ได้ผลการทำนายด้วยวิธีเนอีฟเบย์ด้วยโปรแกรม WEKA ดังภาพที่ 4-4

```
=== Predictions on test set ===

 inst#,     actual, predicted, error, probability distribution
     1        ? 2:Positive     +   0.235 *0.694  0.071
     2        ? 2:Positive     +   0.047 *0.948  0.006
     3        ? 3:Negative     +    0.04   0.178 *0.783
     4        ? 2:Positive     +   0.224 *0.536  0.241
     .          .                    .
     .          .                    .
     .          .                    .
    78        ? 2:Positive     +   0.235 *0.694  0.071
```

ภาพที่ 4-4 ผลการทำนายด้วยวิธีเนอีฟเบย์

หลังจากนำข้อมูลสำหรับการเรียนรู้เข้าสู่โปรแกรม WEKA แล้วเลือกวิธีการเรียนรู้ข้อมูลและสร้างโมเดลด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน สำหรับการจำแนกหมวดหมู่ข้อคิดเห็น ด้วยโปรแกรม WEKA โดยเลือกดังภาพที่ 4-5



ภาพที่ 4-5 การเลือกใช้วิธีซัพพอร์ตเวคเตอร์แมชชีนในการเรียนรู้และสร้างโมเดล

ผลการนำข้อมูลสำหรับเรียนรู้และสร้างโมเดลด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน ซึ่งรันในโปรแกรม WEKA ดังภาพที่ 4-6

=== Evaluation on training set ===		
=== Summary ===		
Correctly Classified Instances	301	95.8599 %
Incorrectly Classified Instances	13	4.1401 %
Kappa statistic	0.914	
Mean absolute error	0.2321	
Root mean squared error	0.2898	
Relative absolute error	70.1791 %	
Root relative squared error	71.3763 %	
Total Number of Instances	314	

ภาพที่ 4-6 ผลการรันข้อมูลสำหรับเรียนรู้และสร้างโมเดลด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน

จากภาพที่ 4-6 เป็นผลการรันข้อมูลสำหรับเรียนรู้ซึ่งเป็นข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ มีข้อมูลจำนวน 314 ตัวอย่าง ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนพบว่า มี Attribute จำนวน 3 Attribute ได้แก่ Medium, Positive และ Negative ได้ค่า Correctly Classified Instances เท่ากับ 95.85%

ผลการทำนายด้วยข้อมูลทดสอบ โดยตัวอย่างนี้เป็นข้อมูลข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยด้านราคาของภาครัฐ โดยใช้ข้อมูลทดสอบ จำนวน 78 ตัวอย่าง ได้ผลการทำนายด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนด้วยโปรแกรม WEKA ดังภาพที่ 4-7

=== Predictions on test set ===				
inst#,	actual, predicted, error, probability distribution			
1	? 2:Positive	+	0.333	*0.667 0
2	? 2:Positive	+	0.333	*0.667 0
3	? 3:Negative	+	0	0.333 *0.66
.	.		.	
.	.		.	
78	? 2:Positive	+	0.333	*0.667 0

ภาพที่ 4-7 ผลการทำนายด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน

4.3 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่

การวัดประสิทธิภาพของการจำแนกหมวดหมู่ โดยนำผลการทำนายที่ได้มาคำนวณประสิทธิภาพของการจำแนกหมวดหมู่ข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน เรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร คำนวณตามสูตรในบทที่ 3 ข้อที่ 3.6 ได้ผลดังตารางที่ 4-3 และ 4-4

ตารางที่ 4-3 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์

ข้อคิดเห็น	ภาค	วิธีเนอีฟเบย์		
		Precision	Recall	F-Measure
ด้านราคา	รัฐบาล	0.849	0.852	0.850
	เอกชน	0.841	0.844	0.841
ด้านสถานที่	รัฐบาล	0.833	0.836	0.834
	เอกชน	0.834	0.838	0.835
ด้านความรู้และทักษะที่ได้รับ	รัฐบาล	0.800	0.741	0.769
	เอกชน	0.838	0.841	0.838
ด้านความยอมรับในสังคม	รัฐบาล	0.744	0.697	0.719
	เอกชน	0.767	0.717	0.742

จากตารางที่ 4-3 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์ ทั้ง 4 ด้าน พบว่า การจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ มีค่า F-Measure อยู่ในช่วงระหว่าง 0.719-0.850 สำหรับการจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน มีค่า F-Measure อยู่ในช่วงระหว่าง 0.742-0.835

ตารางที่ 4-4 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน

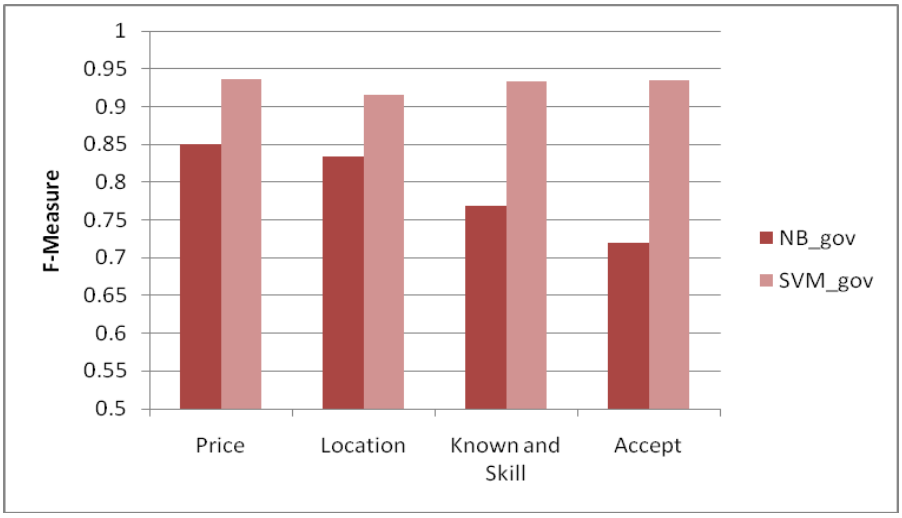
ข้อคิดเห็น	ภาค	วิธีซัพพอร์ตเวคเตอร์แมชชีน		
		Precision	Recall	F-Measure
ด้านราคา	รัฐบาล	0.939	0.938	0.937
	เอกชน	0.959	0.958	0.958
ด้านสถานที่	รัฐบาล	0.927	0.906	0.916
	เอกชน	0.964	0.964	0.963

ตารางที่ 4-4 (ต่อ)

ข้อคิดเห็น	ภาค	วิธีซัพพอร์ตเวคเตอร์แมชชีน		
		Precision	Recall	F-Measure
ด้านความรู้และทักษะที่ได้รับ	รัฐบาล	0.942	0.925	0.933
	เอกชน	0.959	0.958	0.957
ด้านความยอมรับในสังคม	รัฐบาล	0.943	0.926	0.935
	เอกชน	0.913	0.892	0.903

จากตารางที่ 4-4 ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนทั้ง 4 ด้าน พบว่า การจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ มีค่า F-Measure อยู่ในช่วงระหว่าง 0.916-0.937 สำหรับการจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน มีค่า F-Measure อยู่ในช่วงระหว่าง 0.903-0.963

เมื่อนำค่า F-Measure ของการจำแนกหมวดหมู่ข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร ด้วยวิธีเนอีฟเบย์ และวิธีซัพพอร์ตเวคเตอร์แมชชีนทั้ง 4 ด้าน มาสร้างเป็นกราฟเปรียบเทียบประสิทธิภาพระหว่างการจำแนกการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์ และวิธีซัพพอร์ตเวคเตอร์แมชชีนของข้อคิดเห็นที่มีสถาบันการศึกษาไทยภาครัฐ แสดงดังภาพที่ 4-8

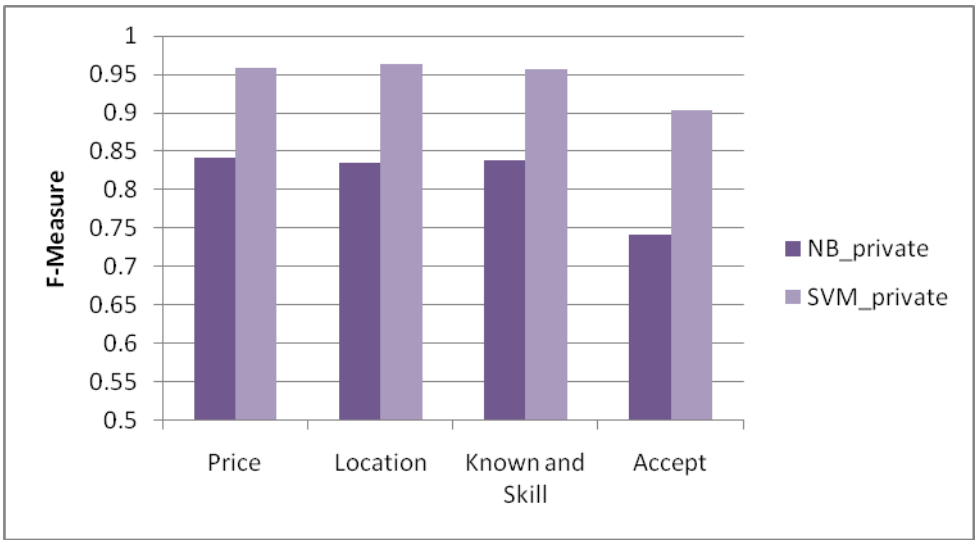


ภาพที่ 4-8 กราฟแสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยในภาครัฐทั้ง 4 ด้าน

จากภาพที่ 4-8 เป็นกราฟแสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยในภาครัฐทั้ง 4 ด้านซึ่งเป็นการวัดประสิทธิภาพด้วยค่า F-Measure โดยพิจารณาเปรียบเทียบระหว่าง 2 วิธีการจำแนกหมวดหมู่ พบว่าประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์มีค่า F-Measure น้อยกว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน แสดงว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนสามารถระบุขั้วความรู้สึกในเชิงบวก กลางและลบได้ดีกว่าการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์

เมื่อพิจารณาค่า F-Measure โดยเปรียบเทียบข้อคิดเห็นแต่ละด้านทั้ง 4 ด้าน ได้แก่ ด้านราคา (Price) ด้านสถานที่ (Location) ด้านความรู้และทักษะที่ได้รับ (Known and Skill) และด้านความยอมรับในสังคม(Accept) พบว่า การจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์ทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยในภาครัฐ ด้านความยอมรับมีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่น ๆ ส่วนการจำแนกหมวดหมู่ด้วยวิธีการซัพพอร์ตเวคเตอร์แมชชีนทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยในภาครัฐ ด้านสถานที่ มีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่น ๆ

ผลการสร้างกราฟเปรียบเทียบประสิทธิภาพระหว่างจำแนกหมวดหมู่ข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยในภาคเอกชนด้วยวิธีเนอีฟเบย์ และวิธีซัพพอร์ตเวคเตอร์แมชชีน แสดงดังภาพที่ 4-9



ภาพที่ 4-9 กราฟแสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยในภาคเอกชนทั้ง 4 ด้าน

จากภาพที่ 4-9 เป็นกราฟแสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยในภาคเอกชนทั้ง 4 ด้านซึ่งเป็นการวัดประสิทธิภาพด้วยค่า F-Measure เมื่อพิจารณาเปรียบเทียบระหว่าง 2 วิธีการจำแนกหมวดหมู่ พบว่าประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์มีค่า F-Measure น้อยกว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน แสดงว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนสามารถระบุขั้วความรู้สึกในเชิงบวก กลางและลบได้ดีกว่าการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์ เมื่อพิจารณาค่า F-Measure โดยเปรียบเทียบข้อคิดเห็นแต่ละด้านทั้ง 4 ด้าน ได้แก่ ด้านราคา (Price) ด้านสถานที่ (Location) ด้านความรู้และทักษะที่ได้รับ (Known and Skill) และด้านความยอมรับในสังคม (Accept) พบว่า ส่วนการจำแนกหมวดหมู่ด้วยวิธีทั้งเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีนทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยในภาคเอกชน ด้านความยอมรับในสังคมมีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่น ๆ

บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผล

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อจำแนกหมวดหมู่ข้อความของข้อคิดเห็นภาษาไทยในแบบสอบถามปลายเปิด และเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน โดยใช้ข้อมูลจากการสำรวจด้วยแบบสอบถาม เรื่องความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนของประชาชนในเขตกรุงเทพมหานคร ซึ่งความคิดเห็นที่มีต่อสถาบันการศึกษาไทยภาครัฐและเอกชนแบ่งเป็น 4 ด้าน ได้แก่ ด้านราคา ด้านสถานที่ ด้านความรู้และทักษะที่ได้รับ และด้านความยอมรับในสังคม การจำแนกหมวดหมู่ระดับความคิดเห็นหรือคำระบุขั้ว แบ่งเป็น 3 หมวดหมู่ คือ เชิงบวก (Positive) กลาง (Medium) และลบ (Negative) สามารถสรุปผลเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน ได้ดังนี้

ผลการวัดประสิทธิภาพของการจำแนกหมวดหมู่ พบว่า การจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชนทั้ง 4 ด้าน ด้วยวิธีเนอีฟเบย์ มีค่า F-Measure อยู่ในช่วงระหว่าง 0.719-0.835 ส่วนการจำแนกหมวดหมู่ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชนทั้ง 4 ด้าน ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน มีค่า F-Measure อยู่ในช่วงระหว่าง 0.903-0.963

การเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่ของข้อคิดเห็นที่มีต่อสถาบันการศึกษาไทยภาครัฐและเอกชน ด้วยด้วยค่า F-Measure พบว่าประสิทธิภาพของการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์มีค่า F-Measure น้อยกว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีน แสดงว่าการจำแนกหมวดหมู่ด้วยวิธีซัพพอร์ตเวคเตอร์แมชชีนสามารถจำแนกหมวดหมู่หรือระบุขั้วความรู้สึกในเชิงบวก กลางและลบได้ดีกว่าการจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์

เมื่อพิจารณาค่า F-Measure โดยเปรียบเทียบข้อคิดเห็นแต่ละด้านทั้ง 4 ด้าน ได้แก่ ด้านราคา (Price) ด้านสถานที่ (Location) ด้านความรู้และทักษะที่ได้รับ (Known and Skill) และด้านความยอมรับในสังคม (Accept) พบว่า การจำแนกหมวดหมู่ด้วยวิธีเนอีฟเบย์ทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยภาครัฐ ด้านความยอมรับ มีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่นๆ

ส่วนการจำแนกหมวดหมู่ด้วยวิธีการซัพพอร์ตเวคเตอร์แมชชีนทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยในภาครัฐ ด้านสถานที่ มีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่นๆ

ส่วนการจำแนกหมวดหมู่ด้วยวิธีทั้งเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีนทำให้ข้อคิดเห็นที่มีต่อสถาบันของการศึกษาไทยภาคเอกชน ด้านความยอมรับในสังคม มีค่า F-Measure ต่ำที่สุดเมื่อเทียบกับด้านอื่นๆ

5.2 อภิปรายผล

ประสิทธิภาพของการจำแนกหมวดหมู่ข้อความแสดงความคิดเห็นยังมีข้อผิดพลาด อยู่เนื่องจากเหตุผลดังต่อไปนี้

5.2.1 การแสดงความคิดเห็น หรือข้อคิดเห็น คือ การแสดงอารมณ์ ความรู้สึก ความคิดเห็นส่วนตัวของผู้พูดหรือผู้เขียนมีต่อสิ่งใดสิ่งหนึ่งหรือเรื่องใดเรื่องหนึ่ง มีข้อกำจัดไม่สามารถระบุขั้วหรือจำแนกหมวดหมู่ได้ เนื่องจากข้อความแสดงความคิดเห็นของไทยมีรูปที่ไม่แน่นอน โดยลักษณะข้อความคิดเห็นบางข้อความไม่สามารถระบุขั่วได้ในข้อมูลจากการสำรวจ 400 ชุด มีตัวอย่างดังนี้

ตัวอย่างด้านความยอมรับในสังคมที่มีต่อสถาบันการศึกษาภาครัฐ เช่น ข้อความว่า

“ต้องมีความสามารถทางการศึกษา” เป็นข้อความไม่สามารถระบุขั้วได้เพราะไม่มีส่วนใดในประโยคบอกว่ายอมรับหรือไม่ระดับใด

“ได้รับการยอมรับน้อยถ้าสถาบันไม่มีชื่อ” เป็นข้อความไม่สามารถระบุขั้วได้เนื่องจากเป็นลักษณะของข้อความที่ต้องมีเงื่อนไขก่อน

“ยอมรับในสังคมได้พอใช้หรือแย่” “บางทีก็ได้รับการยอมรับบางทีก็ไม่ได้รับการยอมรับ” ลักษณะของข้อความที่มีความขัดแย้งกันเอง จึงไม่สามารถระบุขั้วได้ ว่าความคิดเห็นปานกลางหรือเชิงลบ

คำถามในส่วนของด้านการยอมรับ แต่ตอบด้านราคาให้เหตุผลทางด้านราคา เช่น “ค่าใช้จ่ายสูงเลย ทำให้ไม่ค่อยสนใจ” ก็ไม่สามารถระบุขั้วในหัวข้อด้านความยอมรับในสังคมที่มีต่อสถาบันการศึกษาภาครัฐได้

นอกจากนี้ลักษณะของข้อความแสดงความคิดเห็นเชิงเปรียบเทียบไม่สามารถวิเคราะห์ได้ เช่น “อาหารร้าน A แพงกว่าร้าน B” จะไม่สามารถบ่งบอกถึงทัศนะคติได้ว่าข้อความนี้มีความคิดเห็นเชิงบวกหรือกลางหรือลบ แต่ในงานวิจัยนี้ได้แก้ปัญหาในการวิเคราะห์ความคิดเห็นเชิงเปรียบเทียบ โดยกรองข้อความแสดงความคิดเห็นที่มีต่อสถาบันไทยแยกกันระหว่างภาครัฐและเอกชน ตัวอย่างเช่น

ภาครัฐ ข้อความที่ได้ คือ มีราคาถูกมากกว่า อยู่ในหมวดหมู่ ความคิดเห็นเชิงบวก
ภาคเอกชน ข้อความที่ได้ คือ มีราคาแพง อยู่ในหมวดหมู่ ความคิดเห็นเชิงลบ

จากปัญหาที่ไม่สามารถระบุขั้วได้จึงต้องทำการคัดเลือกข้อความบางข้อความออกก่อน ซึ่งมีผลต่อประสิทธิภาพการจำแนกหมวดหมู่ที่ดีขึ้น

5.2.2 ประสิทธิภาพของการจำแนกหมวดหมู่ข้อความแสดงความคิดเห็นยังมีข้อผิดพลาดอยู่ เนื่องจากขั้นตอนการตัดคำ กำจัดคำหยุด และการสกัดคำ ต้องใช้ข้อมูลจากพจนานุกรมคำศัพท์บางคำไม่มีในฐานข้อมูลพจนานุกรมคำศัพท์ รวมทั้งฐานข้อมูลคำหยุดด้วย

จากผลเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน พบว่า การจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีซัพพอร์ตเวคเตอร์แมชชีนมีประสิทธิภาพดีกว่าวิธีเนอีฟเบย์ เป็นไปตามสมมติฐานการวิจัยและสอดคล้องกับผลการวิจัยของนิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ (2554)

5.3 ข้อเสนอแนะ

5.3.1 การเก็บข้อมูลสามารถเก็บข้อมูลด้วยวิธีอื่นร่วมด้วย เช่น เก็บข้อมูลจากแบบสอบถามออนไลด์ เพื่อความรวดเร็วและไม่ต้องพิมพ์ข้อความเก็บลงในไฟล์

5.3.2 สามารถพัฒนาและประยุกต์ใช้ในงานอื่น ๆ ได้ ตัวอย่างเช่น วิเคราะห์ข้อความแสดงความคิดเห็นเกี่ยวกับหัวข้ออื่น ๆ ที่สนใจ ตัวอย่างเช่น การวิเคราะห์ข้อความแสดงความคิดเห็นด้านการบริการหลังการขาย การวิเคราะห์ความคิดเห็นบนเว็บในด้านต่าง ๆ การวิเคราะห์ความคิดเห็นเกี่ยวกับคุณภาพสินค้า เป็นต้น

บรรณานุกรม

กมลาสน์ อุดมล้ำเลิศ และอานนท์ รุ่งสว่าง. (2553). **การคัดแยกหมวดหมู่เว็บเพจภาษาไทยอัตโนมัติด้วยนาอีฟเบย์**. การประชุมทางวิชาการของมหาวิทยาลัยเกษตรศาสตร์ ครั้งที่ 48: 310-317.

เกียรติสุดา ศรีสุข. (2552). **ระเบียบวิธีวิจัย.** เชียงใหม่ : โรงพิมพ์ครองช่าง

จินตนา ธนวิบูลย์ชัย. (2545). **การพัฒนาเครื่องมือสำหรับการประเมินการศึกษา.** หน่วยที่ 8-15 นนทบุรี: โรงพิมพ์มหาวิทยาลัยสุโขทัยธรรมาธิราช.

จันทิมา พลพินิจ และอุมาภรณ์ สายแสงจันทร์. (2548). **ระบบการจัดกลุ่มเอกสารข้อความอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน**. สาขาวิชา วิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ. มหาวิทยาลัยมหาสารคาม.

ณิชาพร สุระ. (2549). **การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้อัลกอริทึม FPTC**. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต. สาขาวิชาวิทยาการคอมพิวเตอร์. ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. (2554). *การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ*. [Online]. วันที่สืบค้น 21 มีนาคม 2553 จาก http://conference.nida.ac.th/rc2010/images/stories/document/group7/Developing%20and%20Effective%20Automatic%20Thai%20Document%20Categorization.pdf

บุญเสริม กิจศิริกุล. (2548). **ปัญญาประดิษฐ์.** เอกสารคำสอนวิชา 2110654. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย. เวอร์ชัน 1.0.2 (มีนาคม 2548) : หน้า 192.

พรพล ธรรมรงค์รัตน์, ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์. (2008). *การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน*. The 12th National Computer Science and Engineering Conference 2008.

พิชิต ฤทธิ์จรูญ. (2544). **ระเบียบวิธีการวิจัยทางสังคมศาสตร์**. กรุงเทพมหานคร : ศูนย์หนังสือราชภัฏพระนคร.

พิลาพรรณ โพธิ์นรินทร์, สุพจน์ นิตย์สุวัฒน์ และชูชาติ หฤไชยะศักดิ์. (2554). *การเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่สำหรับข้อมูลเชิงบรรณานุกรมด้านการเกษตร*. การประชุมทางวิชาการด้านเทคโนโลยีคอมพิวเตอร์และระบบสารสนเทศประยุกต์ระดับชาติ ครั้งที่ 2. มหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ: S3-18.

เพ็ญแข แสงแก้ว. (2541). **การวิจัยทางสังคมศาสตร์**. พิมพ์ครั้งที่ 3. กรุงเทพมหานคร: โรงพิมพ์มหาวิทยาลัยธรรมศาสตร์.

มารยาท โยทองยศ. (2554). *การสร้างแบบสอบถามเพื่อการวิจัย*. [Online]. วันที่สืบค้น 12 มกราคม 2554 จาก http://research.bu.ac.th/knowledge/kn27/Questionnaire_For_Research.pdf

เยาวเรศ จันทะแสน. (2554). *เครื่องมือที่ใช้ในการวัดผล*. วันที่สืบค้น 23 มีนาคม 2554 จาก http://reg.ksu.ac.th/teacher/yahvaret/lession2.html

วัลลภ อินทร์ฉ่ำ. (2548). **ระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ SVM ร่วมกับการประมวลผลภาษา**. วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

สำนักทะเบียนกลาง กรมการปกครอง กระทรวงมหาดไทย. (2552)**.** [Online]. วันที่สืบค้น 4 พฤษภาคม 2554 จาก http://www.dopa.go.th/stat/y_stat.html

อุทุมพร จามรมาน. (2544). **แบบสอบถาม: การสร้างและการใช้**. พิมพ์ครั้งที่ 6. กรุงเทพมหานคร: ฟันนี่พลับบิชชิ่งจำกัด.

Aas, Kjersti, and Eikvil, Line. (1999). **Text Categorization: a Survey**. Report Norwegian Computing Center.

Burges C. (1998). **A tutorial on support vector machines for pattern recognition**. Data Mining and Knowledge Discovery. Kluwer Academic Publishers : Boston. (955-974).

Charoenpornsawat, Paisarn. (1999) . **Feature-based Thai Word Segmentation**. Master's Thesis. Computer Engineering. Chulalongkorn University, Bangkok, Thailand.

Cortes, Corrinna and Vapnik, Vladimir. (1995). **Support-Vector Networks**. Machine Learning.Volume 20, Number 3, September.

Jaruskulchai, Chuleerat. (1998). **An Automatic Indexing for Thai Text Retrieval.** PhD Thesis, George Washington University, USA, Aug 31.

Kruengkrai, Canasai and Jaruskulchai, Chuleerat. (2001)**.** *Thai Text Classification based on Naïve Bayes*. **Technical Report**. Department of Computer Science, Kasetsart University.

Marquez, Llus. (2000). **Machine learning and natural language processing**. Technical Report Departament de Lenguatges Sistemes Informatics (LSI), Barcelona, Spain.

Neil J.Salkind (2006). **Exploring Research** . 6th New jersey ; Pearson Prentice Hall.

Ozgur, Arzucan, Ozgur, Levent and Gungor, Tunga. (2005). Text Categorization with Class-Based and Corpus-Based Keyword Selection. **Department of Computer Engineering**, Bogazici Unversity, Bebek, Istanbul Turkey. pp. 606-615.

Salkind, Neil J. (2006). **Exploring Research**. 6th New jersey ; Pearson Prentice Hall.

Sebastiani, Fabrizio. (2002). *Machine Learning in Automated Text Categorization*. **ACM Computing Surveys (CSUR)**. March.

Vapnik, V. Support-Vector Networks. (1995). **Machine Learning** . Kluwer Academic Publishers. pp. 273-297.

ภาคผนวก ก
ตัวอย่างแบบสอบถาม

แบบสอบถาม

เรื่อง ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชน ของประชาชนในเขตกรุงเทพมหานคร

ผู้ศึกษาขอความร่วมมือผู้ตอบแบบสอบถามทุกท่านตอบแบบสอบถามตามความเป็นจริง แบบสอบถามออกเป็น 3 ส่วน ดังนี้

ส่วนที่ 1 ข้อมูลพื้นฐานทางประชากรศาสตร์ของกลุ่มตัวอย่าง

ส่วนที่ 2 ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชน

ส่วนที่ 3 ข้อคิดเห็นเพิ่มเติม

<u>ส่วนที่ 1</u> ข้อมูลพื้นฐานทางประชากรศาสตร์ของกลุ่มตัวอย่าง

คำชี้แจง : โปรดทำเครื่องหมาย ／ ลงหน้าข้อความที่ต้องการ

1. เพศ

☐ 1) หญิง ☐ 2) ชาย

2. ช่วงอายุ

☐ 1) 15 – 25 ปี ☐ 3) 36 – 50 ปี

☐ 2) 26 – 35 ปี ☐ 4) 51 – 60 ปี

3. ระดับการศึกษา

☐ 1) มัธยมศึกษาตอนต้น ☐ 3) ปริญญาตรี หรือเทียบเท่า

☐ 2) มัธยมศึกษาตอนปลาย ☐ 4) สูงกว่าปริญญาตรี

4. รายรับต่อเดือน

☐ 1) ต่ำกว่า 5,000 บาท ☐ 4) 20,001 บาท – 30,000 บาท

☐ 2) 5,001 บาท – 10,000 บาท ☐ 5) มากกว่า 30,000 บาท

☐ 3) 10,001 บาท – 20,000 บาท

5. อาชีพปัจจุบันของท่าน

☐ 1) นักเรียน / นักศึกษา ☐ 5) พ่อบ้าน / แม่บ้าน

☐ 2) รับจ้าง ☐ 6) พนักงานบริษัท

☐ 3) ข้าราชการ / พนักงานของรัฐ ☐ 7) อื่นๆ ระบุ.....

☐ 4) ผู้ประกอบการ

<u>ส่วนที่ 2</u> ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐและเอกชน

คำชี้แจ้ง : กรุณากรอกข้อความตามความคิดเห็นของท่านที่มีต่อสถาบันการศึกษาไทยทั้งภาครัฐและเอกชนให้ครบทุกข้อ

ปัจจัย	**ความคิดเห็นที่มีต่อสถาบันการศึกษา**	
	ภาครัฐ	**เอกชน**
1. ด้านราคา เป็นค่าใช้จ่ายทั้งหมดในการศึกษาเล่าเรียนตลอดการศึกษา รวมทั้งค่าดำรงชีพระหว่างการศึกษา ค่าใช้จ่ายส่วนตัว	....	....
2. **ด้านสถานที่** เป็นบริเวณโดยรอบความสะอาด ความสวยงามของสถานที่ อาคารเรียน สิ่งแวดล้อมต่างๆในสถานการศึกษา	....	....
3. ด้านความรู้และทักษะที่ได้รับ เป็นความรู้และทักษะต่างๆทางด้านวิชาการที่ได้รับในระหว่างการศึกษา	....	....
4. ด้านความยอมรับในสังคม เป็นความยอมรับทั้งหลังจบการศึกษาแล้ว การได้รับการตอบรับจากสังคมและสถานประกอบ	....	....

<u>ส่วนที่ 3</u> ข้อคิดเห็นอื่นๆ เพิ่มเติม

.....

.....

.....

.....

.....

.....

.....

.....

.

ขอขอบคุณทุกท่านที่ให้ความร่วมมือ

ภาคผนวก ข
ข้อมูลที่ได้จากแบบสอบสอบ
รายการคำหยุดที่ใช้ในงานวิจัย

ข้อมูลที่ได้จากแบบสอบสอบ

ข้อมูลที่ได้จากแบบสอบสอบ เป็นข้อมูลที่ได้การสำรวจด้วยแบบสอบสอบ เรื่อง ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชน ของประชาชนในเขตกรุงเทพมหานคร จำนวน 400 ชุด โดยได้ยกตัวอย่างข้อมูลทาง 4 ด้าน ด้านละ 40 ตัวอย่าง

ตารางที่ ข-1 ข้อมูลจากแบบสอบถามความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนด้านราคา

ลำดับ	**ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ**	**ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน**
1	พอสมควร บางทีก็แพงกว่าเอกชน	มีเงินเข้าได้อย่างเดียว ค่าใช้จ่ายแพงกว่า
2	ถูกกว่า	แพงกว่า ใช้แต่เงิน
3	ราคาพอสมควร บางทีก็แพงกว่าเอกชน	มีเงินเข้าได้อย่างเดียวก็เข้าเรียนได้
4	ค่าเกี่ยวข้องกับด้านการศึกษาก็พอสมควร ค่าใช้จ่ายส่วนตัวไม่มากเพราะไม่มีการแข่งขันระหว่างนักศึกษามากเท่ากับเอกชน	ค่าเกี่ยวกับการศึกษาก็มีมากพอสมควร ค่าใช้จ่ายส่วนตัวค่อนข้างสูง
5	ค่าใช้จ่ายต่ำ	ค่าใช้จ่ายสูง
6	ค่าใช้จ่ายในสถาบันของรัฐนั้นจะมีค่าใช้จ่ายที่ถูกกว่าสถาบันของเอกชน แต่สถาบันของรัฐส่วนใหญ่จะได้รับความเชื่อถือจากสถานบันประกอบการมากกว่าเอกชน	ค่าใช้จ่ายของเอกชนนั้นจะแพงกว่าสถาบันของรัฐเยอะ แต่สถาบันเอกชนบางที่มีศักยภาพจริงที่จะได้รับความเชื่อถือในด้านศักยภาพน้อยกว่าภาครัฐ
7	พอๆกัน	พอๆกัน
8	มีค่าใช้จ่ายที่เหมาะสม น่าสนใจ ไม่ต้องคิดมาก	มีค่าใช้จ่ายสูงมากเพราะถ้าเข้าเรียนต้องคิดหนักเรื่องค่าเทอม
9	ค่าใช้จ่ายไม่มาก	ค่าใช้จ่ายสูง
10	ค่าใช้จ่ายถูกกว่า	ค่าใช้จ่ายแพงกว่า
11	ราคาพอใช้ได้	ค่าใช้จ่ายเยอะมาก
12	ค่าเทอมถูกกว่าเอกชน	ค่าเทอมราคาสูง
13	ค่าใช้จ่ายไม่สูงจนเกินไป	ค่าใช้จ่ายสูงเกินไป
14	ค่าใช้จ่ายที่เหมาะสม	ค่าใช้จ่ายมีราคาแพงเกินไป
15	ค่าใช้จ่ายต่างๆต่ำ	ค่าใช้จ่ายต่างๆค่อนข้างสูง
16	ราคาใช้จ่ายต่ำ	ราคาใช้จ่ายสูง
17	ราคาถูกกว่า	ราคาแพงกว่า
18	ค่าเทอมถูกกว่าเอกชน	ค่าเทอมแพง
19	มีค่าใช้จ่ายที่พอเหมาะไม่สูงเกินไป	มีค่าใช้จ่ายสูงแต่ก็บริหารดี
20	จะเสียค่าใช้จ่ายน้อยกว่าเอกชน	จะมีการเก็บเงินด้านอื่นอีกมากมาย

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
21	ค่าใช้จ่ายต่างๆในการศึกษาถูก	ค่าใช้จ่ายแพง
22	ราคาถูก	ราคาแพง
23	ราคาถูก	ราคาแพง
24	ประหยัดมากกว่า	ค่าใช้จ่ายมากกว่ารัฐ
25	มีราคาถูก	มีราคาแพง
26	ราคาถูก	ราคาแพง
27	ราคาถูก	ราคาค่อนข้างแรง
28	มีค่าใช้จ่ายถูก	มีค่าใช้จ่ายค่อนข้างแพง ค่าใช้จ่ายจุจิกมากมาย
29	มีราคาเล่าเรียนที่ถูกกว่า	มีราคาเล่าเรียนที่สูงกว่ารัฐบาล
30	ราคาก็ปกติ	แพงมาก
31	สูงกว่า	แพง
32	ถูกกว่า	แพงกว่าถึงแพงมาก
33	ดี	ดี
34	ค่าเทอมมีราคาถูกว่าเอกชน	มีราคาค่าเทอมแพงกว่ารัฐบาล
35	ค่อนข้างน้อย	สูงมาก
36	พอสมควร	แพง
37	ราคาอยู่ในเกณฑ์ปานกลาง ไม่สูงหรือต่ำเกินไป	ราคาค่าใช้จ่ายยังสูงอยู่
38	1.ค่าเล่าเรียนถุกกว่า 2.สามารถเบิกได้เต็มจำนวน	1.ค่าเล่าเรียนแพง 2. สามารถเบิกได้แค่ครึ่งเดียว 3. เรียนง่าย, อิสระ(การแต่งกาย) 4. ความเป็นกันเองกับอาจารย์มากกว่า
39	ต่ำ	สูง
40	ถูก	แพง

ตารางที่ 2 ข้อมูลจากแบบสอบถามความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนด้านสถานที่

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
1	ใหญ่เป็นบางที่ แต่โรงเรียนภาครัฐชอบอยู่แต่ในเมือง หรือกรุงเทพ	สถาบันเอกชนส่วนมากจะเล็กกว่าภาครัฐ และสถานที่อยู่แต่ปริมณฑณ
2	ส่วนมากจะอยู่ในเมือง	จะอยู่ค่อนข้างชานเมือง
3	ใหญ่เป็นบางมหาวิทยาลัย	จะดีกว่าหรูกว่ารัฐบาล พื้นที่ใหญ่กว้าง
4	ภาครัฐส่วนมากอยู่ในตัวเมือง เดินทางสะดวกแต่พื้นที่น้อย สถานที่จึงเล็ก	ภาคเอกชนส่วนมากจะอยู่แถบชานเมืองซะส่วนใหญ่ พื้นที่มาก จึงใหญ่กว้างขวาง
5	รัฐบาลจะมีพื้นที่มากกว่า	จะมีพื้นที่ที่กว้างไกล
6	แน่นอนมหาวิทยาลัยของรัฐจะมีสถานที่ที่ใหญ่เพียงพอต่อจำนวนนักศึกษา เช่น จุฬาลงกรณ์มหาวิทยาลัย,ม.พระจอมเกล้าฯ,ม.ธรรมศาสตร์	ก็จะคล้ายๆกันกับมหาวิทยาลัยของรัฐ แต่ละที่จะมีขนาดของสถานที่ที่แตกต่างกันไป
7	ดี	พอใช้
8	บางทีอาจจะเล็กไป	กว้างไกลน่าสนใจ ร่มรื่นหน้าอยู่
9	สถานที่ตั้งจะมีพื้นที่เล็กกว่าเอกชน	มีบริเวณที่ตั้งที่เหมาะสม ตั้งอยู่ในที่ที่คนผ่านไปมารู้จักเยอะ
10	รัฐบาลจะมีพื้นที่มากกว่า	เอกชนมีพื้นที่น้อยกว่า
11	บางที่อาจจะเล็กไปหน่อย แต่ความรู้ไม่ได้ต่างกัน	ดีพอสมควร เพราะโปรโมทเยอะชื่อเสียงเลยมากทำให้คนเข้ามาศึกษา
12	รัฐบาลจะมีพื้นที่มากกว่า	เทคโนโลยีกว้างไกล
13	สถานที่จะเล็กกว่า	สถานที่จะใหญ่กว่า
14	จะมีพื้นที่ที่พอดีกับสถานที่	จะมีพื้นที่ที่กว้างไกล
15	สถานที่พร้อมสำหรับการเรียนและกว้างขวาง	สถานที่พร้อมสำหรับการเรียน
16	สถานที่กว้างขวาง	สถานที่เล็กมาก
17	จะมีขนาดใหญกว่า	พื้นที่จะเล็กน้อยกว่าภาครัฐ
18	ไม่ค่อยกว้างเท่าไร จำกัดในการรับนักศึกษา	กว้างสามารถรองรับนักศึกษาได้มาก
19	มหาวิทยาลัยมีความทันสมัยและมีมาตรฐานที่ดี	บางสถานศึกษายังมีมาตรฐานที่ยังไม่ค่อยดีนัก
20	สถานที่จะเก่าและไม่ค่อยได้รับการดูแล	ได้รับการปรับปรุงอย่างเสมอๆ
21	ใหญ่และมาตรฐาน	เล็กเกินไปบางที่
22	เล็ก	ใหญ่
23	สถานที่เล็ก	สถานที่ใหญ่
24	ดี	ดีมากกว่า

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
25	มีขนาดกลาง	มีขนาดใหญ่
26	ขนาดกลาง	ขนาดใหญ่
27	เล็ก	ใหญ่
28	สถานที่มีขนาดเล็ก (เป็นบางที่)	มีขนาดใหญ่
29	รัฐบาลมีพื้นที่การศึกษาที่เล็กกว่า	มีพื้นที่ขนาดใหญ่กว่ารัฐบาล
30	สถานที่ก็ขนาดกลาง	สถานที่ก็เหมือน ๆ กัน แต่ก็ทันสมัยกว่า
31	ไม่ค่อยรมรื่น บรรยากาศไม่ค่อยดี	ร่มรื่น มีอุปกรณ์ในการศึกษามาก
32	ไม่ค่อยร่มรื่น ขาดแคลนอุปกรณ์หรือสื่อการสอน	ร่มรื่น น่าอยู่ มีอุปกรณ์การศึกษาครบ
33	ดี	ดีมาก
34	ไม่ค่อยมีความพร้อม	มีความพร้อมเพียง
35	มีพื้นที่ใหญ่และสะดวกสะบาย	เล็ก แออัด
36	ไม่ค่อยดี	อาจดี
37	มีสถานที่ที่เป็นสถาบัน การศึกษาเพียงพอในการรองรับ	สถานที่ยังไม่เพียงพอ
38	แล้วแต่ละมหาวิทยาลัย	จะมีพื้นที่ที่กว้างไกล
39	ค่อนข้างใหญ่	เล็ก
40	ไกล ชนบท เดินทางไม่สะดวก	ใกล้ รถติด
41	พื้นที่มาก	พื้นที่น้อย
42	ดูหรูหราเป็นที่น่าเรียนเป็นอย่างมาก และมีคนมาเรียนเยอะ	อาจดูไม่หรูหราเหมือนภาครัฐ แต่ก็มีคนเรียนเหมือนกัน
43	ก็จะมีอายุมากกว่า เพราะที่ของรัฐส่วนใหญ่เป็นของเก่า ส่วนใหญ่จะไม่กว้างมาก	ก็จะใหญ่เพราะเอกชนซื้อที่ทำตึกเรียน
44	มีจำนวนพื้นที่มากกว่า	พื้นที่น้อยกว่า เพราะข้อจำกัดด้านเงินทุน
45	สถานที่กว้างแต่มีคนจำนวนมาก	สถานที่กว้างขวาง
46	ส่วนใหญ่แล้วจะมีพื้นที่มากกว่า	บางทีก็มีพื้นที่มากอาจจะอยู่ห่างไกลสถานที่สำคัญมาก เดินทางลำบาก
47	สะดวก	สะอาดมาก
48	ร่มรื่นดี	สะอาดดีน่าอยู่กว่าสถาบันอื่น
49	มีพื้นที่มากกว่าเอกชน	มีพื้นที่น้อยกว่ารัฐ
50	กว้างขวางพื้นที่ค่อนใช้สอย	กว้างขวางมากพอสมควร

ตารางที่ 3 ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนด้านความรู้และความทักษะที่ได้รับ

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
1	บางทีก็ให้ความรู้มากกว่าเอกชน	ให้ความรู้เรื่องทางภาษามากกว่ารัฐบาล แต่แค่บางสถาบัน
2	แล้วแต่ความคิดของแต่ละคน	เหมือนกัน
3	รัฐบาลต้องใช้ความรู้ในการเข้าสอนเข้าเรียน มีความรู้แน่นมากกว่า	ความรู้ไม่ค่อยแน่นเรียนให้ผ่านๆไป
4	แล้วแต่สถานศึกษาแต่ละสถาบันบางที่ก็เด่นบางวิชา วิชาอื่นๆก็ระดับปานกลาง	เอกชนเด่นทางด้านเทคโนโลยีซะส่วนใหญ่
5	อาจจะได้ทักษะและความรู้มากหรือน้อยขึ้นอยู่กับผู้เรียนและอาจารย์ผู้สอน	อาจจะมีเทคโนโลยีที่เหนือกว่า
6	มีความรู้เท่าเทียมกัน ที่อยู่นักศึกษาเองว่ามีความสนใจมากน้อยเพียงใด	มีความรู้เท่าเทียมกัน
7	ดี	ดี
8	บุคลากรน้อยทำให้ได้รับความรู้ไม่เพียงพอ	บุคลากรเยอะได้ความรู้มากมาย
9	ได้รับความรู้เท่าเอกชนแต่เทคโนโลยีจะด้วยกว่า	ได้รับความรู้ทัดเทียมกับรัฐบาล
10	ได้รับความรู้ดี	ได้รับความรู้ดี
11	ได้รับความรู้เท่ากับที่เอกชน	ทักษะก็ไม่ต่างกันมากแตกต่างแค่ที่ชื่อเสียง
12	ได้รับความรู้ดี	ได้รับความรู้ดี
13	ได้รับความรู้เท่าๆกับที่เอกชน	ทักษะก็ไม่ต่างกันมาก
14	บุคลากรน้อย ความรู้ไม่ครบ	บุคลากรมากพอ
15	ได้รับความรู้อย่างครบถ้วน	ได้รับความรู้อย่างครบถ้วน
16	ไม่แตกต่างกัน	ไม่แตกต่างกัน
17	ได้ความรู้ดี	ได้ความรู้ดี
18	มีความรู้แน่นกว่า	ความรู้พอประมาณ
19	มีความเหมาะสมสามารถนำไปใช้ในชีวิตประจำวันได้ดี	มีความเหมาะสมและสามารถนำไปใช้ได้ในชีวิตประจำวัน
20	เนื้อหาจะแน่นและมากกว่าเอกชน	จะเจาะจงเฉพาะบางด้านและบางวิชาที่สถานศึกษาเอกชนถนัด
21	วิชาการเยอะ	วิชาการน้อยกิจกรรมเยอะ
22	รัฐเน้นวิชาการ	เอกชนเน้นกีฬากิจกรรมต่างๆ
23	เน้นวิชาการต่างๆ	เน้นการทำกิจกรรมร่วมกันต่างๆ

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
24	ผมไม่ค่อยได้รับรู้ข่าวสาร	น้อย
25	ด้านวิชาการดี	ด้านวิชาการดี
26	ดีมาก ในด้านวิชาการ	ดีมากในด้านวิชาการ
27	รัฐบาลจะเน้นวิชาการมากกว่า	เอกชนจะเน้นกิจกรรมเป็นส่วนใหญ่
28	การเรียนการสอนดี	การเรียนการสอนดี มีการเรียนรู้เพิ่มเติมในสิ่งที่ใหม่มาก
29	รัฐบาลมีทักษะที่ดีกว่า	ทีทักษะน้อยกว่ารัฐบาล
30	ภาครัฐจะเน้นด้านวิชาการ	เอกชนเน้นด้านกิจกรรม ค่าใช้จ่ายเยอะ
31	มีบุคลากร ครูน้อย	มีบุคลากร ครูมาก
32	มีบุคลากรน้อยไม่เพียงพอการศึกษา	บุคลากรมีมากพอในการศึกษา
33	ธรรมดา	ดี
34	ค่อนข้างดี	เน้นทักษะและความรู้ต่าง ๆ แบบเจาะลึก
35	ได้รับความรู้ค่อนข้างน้อย	ความรู้ที่ได้รับมาตรงตามความต้องการ และมีความรู้สูงมาก
36	ค่อนข้างดี	เหมือนกัน
37	ความรู้ที่ได้รับดี อยู่ในเกณฑ์ที่เหมาะสม	มีการเจาะลึกในการให้ความรู้ที่ลึกมากกว่า
38	ได้ทักษะที่แน่นกว่า	เหมือนกัน
39	น้อย	มาก
40	เอาใจใส่นักศึกษา	ไม่ค่อยเอาใจใส่นักศึกษา
41	มีความเท่าเทียมกันกับภาคเอกชน	เท่าเทียมกับภาครัฐ
42	อาจได้เท่าเทียมกันขึ้นอยู่กับทักษะของนักศึกษา	เท่าเทียมหรือมากกว่านั้นขึ้นอยู่กับไหวพริบหรือความรู้
43	ก็ดี เพราะภาครัฐก็จะสอนเข้มงวดกว่า	ไม่ค่อยดูเท่าที่ควรด้วยความที่เป็นเอกชนจะทำให้ไม่โหด
44	ความรู้เท่าเทียมกัน	เหมือนกัน
45	พอสมควร	พอสมควร
46	จะได้รับมากน้อยขึ้นอยู่กับคนเรียนไม่ว่าที่ไหน ก็จะให้ความรู้มากเท่ากัน	เหมือนกัน
47	ได้รู้รอบด้าน	รู้สิ่งใหม่ ๆ
48	ได้ความรู้อะไรหลายๆอย่างจากภาครัฐ	ได้นำความรู้ไปใช้งานทางด้านเอกชน
49	มีความเท่าเทียมกัน	มีความเท่าเทียมกัน
50	อาจให้ความรู้ไม่ทั่วถึงเพราะผู้เรียนเยอะ	อาจใช้ความรู้ดีกว่า เพราะต่อหนึ่งห้องมีคนเรียนน้อย

ตารางที่ 4 ความคิดเห็นที่มีต่อสถาบันการศึกษาไทยระหว่างภาครัฐและเอกชนด้านความยอมรับในสังคม

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
1	ด้านการยอมรับภาครัฐส่วนมากยอมรับ	ยิ่งมีชื่อเสียงมาก ก็ยิ่งแพงมาก
2	บางที่ก็ได้รับความยอมรับ	บางทีก็ได้ความมีชื่อเสียง
3	รัฐบาลสังคมยอมรับมากกว่าเพราะเป็นการสอบเข้าโดยใช้ความรู้และความสามารถ	มีเงินยัดเงินก็เรียนได้ ถึงโง่แค่ไหนก็จบได้
4	ภาครัฐได้รับความน่าเชื่อถือมากกว่าเพราะต้องใช้ความสามารถสอบเข้า	บางที่ก็ได้รับการยอมรับบางที่ก็ยังไม่มีการยอมรับ
5	อาจจะยอมรับมากกว่าเอกชน	ขึ้นอยู่กับผู้เรียน
6	สังคมจะยอมรับสถาบันของรัฐบาลมากกว่าเอกชน รวมทั้งสถานประกอบการด้วย	สังคมและสถานประกอบการจะยอมรับสถาบันเอกชนน้อยว่า
7	พอๆกัน	พอๆกัน
8	ผู้สนใจเยอะ เพราะถูก	ค่าใช้จ่ายสูงเลย ทำให้ไม่ค่อยสนใจ
9	มีการยอมรับน้อย	มีการยอมรับเป็นส่วนใหญ่
10	มีความยอมรับมากกว่าเอกชน	มีความยอมรับน้อยกว่ารัฐ
11	เท่าเทียมกัน	เท่าเทียมกัน
12	มีความยอมรับมากกว่าเอกชน	สู้รัฐบาลไม่ได้
13	เท่าเทียมกัน	เท่าเทียมกัน
14	ยอมรับน้อย	ยอมรับเยอะ เพราะบุคลากรครบ
15	ได้รับความยอมรับในสังคมมาก	ไม่ค่อยได้รับความยอมรับในสังคมเท่าที่ควร
16	มีการตอบรับดีมาก	ไม่ค่อยได้รับทางสังคม
17	ใช้ได้อยู่ในเกณฑ์ดี	อยู่ในเกณฑ์ดี
18	รัฐบาลเป็นที่ยอมรับกว่า	ยอมรับน้อย
19	มีการพิจารณาที่ดีกว่าเมื่อเทียบกับเอกชน	ค่อนข้างน้อยที่จะได้รับการพิจารณาก่อน
20	ถ้าจบสถาบันรัฐจะได้รับความยอมรับมากกว่า	ความยอมรับน้อยเป็นเฉพาะบางแห่ง
21	ได้รับการยอมรับเพราะมีคุณภาพ	ไม่ค่อยยอมรับ เพราะอิสระเกินไป
22	นักศึกษานิยมเพราะราคาถูก	น้อย
23	จะไม่ค่อยน่าเชื่อถือ	จะยอมรับเป็นส่วนใหญ่
24	น่าจะยอมรับมากกว่าเอกชน	มีบ้างแต่ไม่มากพอในสังคม
25	มีความยอมรับมากกว่าเอกชน	น้อย
26	นักศึกษาที่เลือกภาครัฐเพราะราคาถูก	นักศึกษาที่เลือกเอกชนเพราะชอบความสะดวกสบาย

ลำดับ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาครัฐ	ความคิดเห็นที่มีต่อสถาบันการศึกษาภาคเอกชน
27	ด้านรัฐจะมีการยอมรับน้อยกว่า	ด้านเอกชนจะเป็นที่ยอมรับมากกว่า
28	เนื่องจากมีค่าใช้จ่ายน้อยจึงเป็นที่นิยม	เป็นที่ยอมรับของคนมีตังค์
29	รัฐบาลดูมีมาตรฐานการศึกษามากกว่าเอกชน	เอกชนดูมีประสิทธิภาพการศึกษาน้อยกว่า
30	คนที่เรียนสถาบันรัฐส่วนใหญ่เพราะราคาถูก	คนที่เรียนสถาบันเอกชน ส่วนใหญ่จะมีฐานะ
31	ยอมรับน้อย เพราะไม่ค่อยได้รับความยอดนิยม	ยอมรับมากเพราะมีครบพร้อมทุกอย่าง
32	ยอมรับน้อย เพราะไม่ค่อยได้รับความยอดนิยม	ยอมรับมากทั้งทางด้านการศึกษา กีฬา เป็นต้น
33	ธรรมดา	ดีมาก
34	เป็นที่ยอมรับในบางส่วน	เป็นที่ยอมรับค่อนข้างมาก
35	ไม่ค่อยได้รับความนิยม	นิยมมาก
36	ยอมรับสูง	น่าจะสูง
37	ได้รับการยอมรับ มีสถานประกอบการยอมรับ	เมื่อจบแล้ว สถานประกอบการยังไม่ยอมรับเท่าที่ควร
38	จะสูงกว่า	ต่ำกว่า
39	ไม่ค่อยยอมรับ	ได้รับความนิยม
40	น้อย	มาก ยิ่งค่าเทอมมากยิ่งดัง
41	สังคมยอมรับมากกว่า	ยอมรับน้อยกว่า
42	เป็นที่ยอมรับมากกว่าเพราะเป็นที่ได้ยินชื่อเสียงมากกว่า	อาจได้ยอมรับน้อยกว่า เพราะอาจมาใหม่ อาจไม่ได้รับการยอมรับ
43	ยอมรับดี	ยอมรับประมาณหนึ่ง
44	ในรัฐจะได้ความยอมรับกว่าเอกชน	เหมือนกัน
45	เป็นที่ยอมรับมาก ผู้คนให้ความสนใจ	ไม่เป็นที่ยอมรับเท่าที่ควร ส่วนใหญ่เป็นพวกไม่มีที่เรียน เข้ารัฐไม่ได้
46	สังคมได้รับการยอมรับมากกว่าทั้งที่จะเก่ง ไม่เก่งขึ้นอยู่กับคนมากกว่า	สังคมให้การยอมรับน้อย เพราะเข้าง่ายจะโดนดูถูกกว่า ไม่ได้สอบเข้า ไม่มีความสามารถ แต่จริงไม่สำคัญเลยเรียนที่ไหนถ้าไม่จบก็ไม่มีงานทำก็เท่านั้น
47	ปานกลาง	ยอมรับเป็นกรณี
48	ดีมากได้คุณภาพ	ไม่ค่อยยอมรับ
49	มีความยอมรับสูงกว่าเอกชน	บางที่ไม่ค่อยมีความยอมรับเท่าที่ควร
50	จะหางานยาก	จะหางานง่ายกว่า

รายการคำหยุดที่ใช้ในงานวิจัย

ตัวอย่างรายการคำหยุดที่ใช้ในงานวิจัยภาษาอังกฤษมีดังต่อไปนี้

anyway	url	during	could	he	info	north	real
never	become	about	did	held	inside	not	really
while	but	three	do	help	interest	of	related
before	don't	above	does	hence	into	offer	reserved
since	i'm	all	each	her	issue	offering	right
must	know	also	etc	here	it	old	said
itself	looking	am	every	hereby	its	on	sample
con	myself	an	everything	herein	join	one	san
led	need	and	fact	hereof	just	only	see
high	now	anything	faq	hereon	known	onto	she
via	want	are	favorite	hereto	large	open	short
try	useful	around	few	herewith	let	or	should
then	along	available	find	high	like	other	simple
next	any	be	first	him	little	our	small
url	across	because	for	yet	man	out	so
var	afb	been	form	his	many	over	some
them	after	best	from	hold	me	own	start
can't	against	between	full	hot	medium	page	still
long	alt	big	get	hour	men	part	such
several	done	both	great	how	mixed	past	than
off	next	buy	had	however	more	pic	that
take	taken	can	hardly	if	most	plus	the
throughout	last	center	has	include	much	present	their
down	week	com	have	including	new	provide	there
without	as	come	having	index	nor	providing	
when	whereby	whether	who	whose	will	within	
where	wherein	which	whom	why	with	would	
they	through	to	unlisted	used	was	what	
thing	thus	too	unto	using	we	your	
this	time	two	us	various	well	these	
those	title	under	use	very	were	you	

ตัวอย่างรายการคำหยุดที่ใช้ในงานวิจัยภาษาไทยมีดังต่อไปนี้

ที่	อย่าง	ทุก	น่า	แต่	ลง	โดย	ผ่าน
ใน	เมื่อ	จึง	ถ้า	แล้ว	แห่ง	จาก	ตาม
ว่า	ทำ	คือ	ตั้งแต่	ยัง	เห็น	ความ	ทำให้
ได้	รับ	มีค	เฉพาะ	ต้อง	สำหรับ	ซึ่ง	ขณะ
เป็น	เพื่อ	ถูก	เลย	วัน	เอง	แรก	
มี	กล่าว	หาก	เริ่ม	ก็	อาจ	เช่น	
จะ	มาก	ไว้	เคย	ถึง	เปิด	เปิดเผย	
และ	เพราะ	ดัง	ตั้ง	กัน	ช่วง	คง	
การ	ทาง	จัด	เรา	ขึ้น	ต่าง	ผล	
ให้	ต่อ	นัก	หลัง	อยู่	พบ	ร่วม	
ของ	นั้น	ราย	ที่สุด	ด้วย	หนึ่ง	รวม	
นี้	อีก	ส่ง	ทั้งนี้	ส่วน	เป็นการ	ด้าน	
มา	เข้า	แบบ	ต่างๆ	กว่า	หลาย	อยาก	
ไม่	ทั้ง	พร้อม	เดียว	ก่อน	สุด	อะไร	
ไป	หรือ	ระหว่าง	เดียวกัน	นำ	นอกจาก	เขา	
กับ	ออก	ขอ	บาง	ครั้ง	เนื่องจาก	หลังจาก	