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

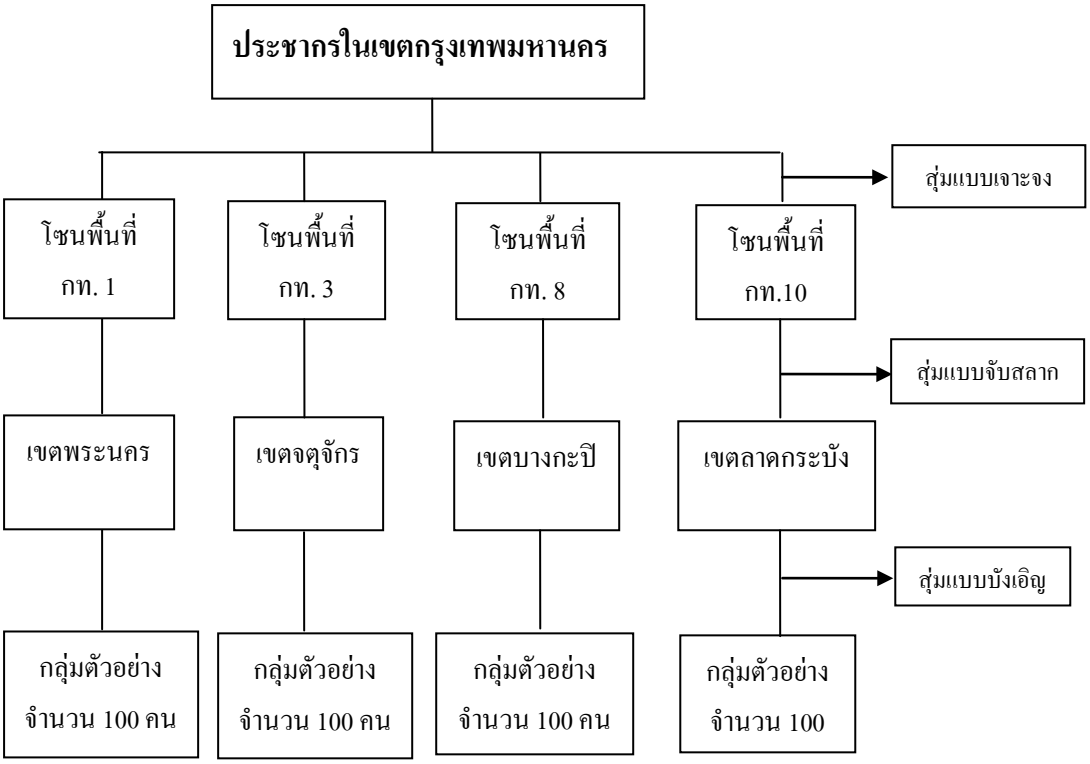
โซนพื้นที่ กท.8 หรือ เขตที่อยู่อาศัยรองรับการขยายตัวของเมือง ด้านตะวันออกตอนใต้ ประกอบด้วย เขตบางกะปิ เขตคันนายาว เขตวังทองหลาง เขตบึงกุ่ม เขตสะพานสูง และเขตสวนหลวง

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...
```

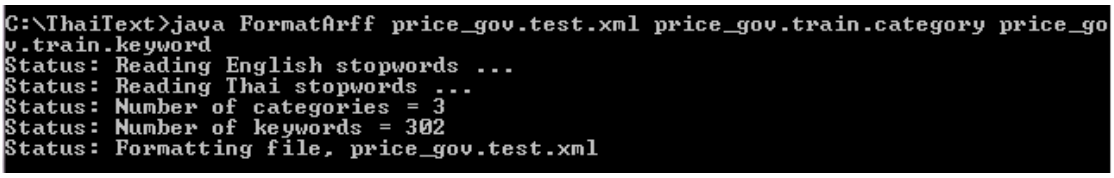
ภาพที่ 3-5 รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของข้อมูลสำหรับเรียนรู้

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

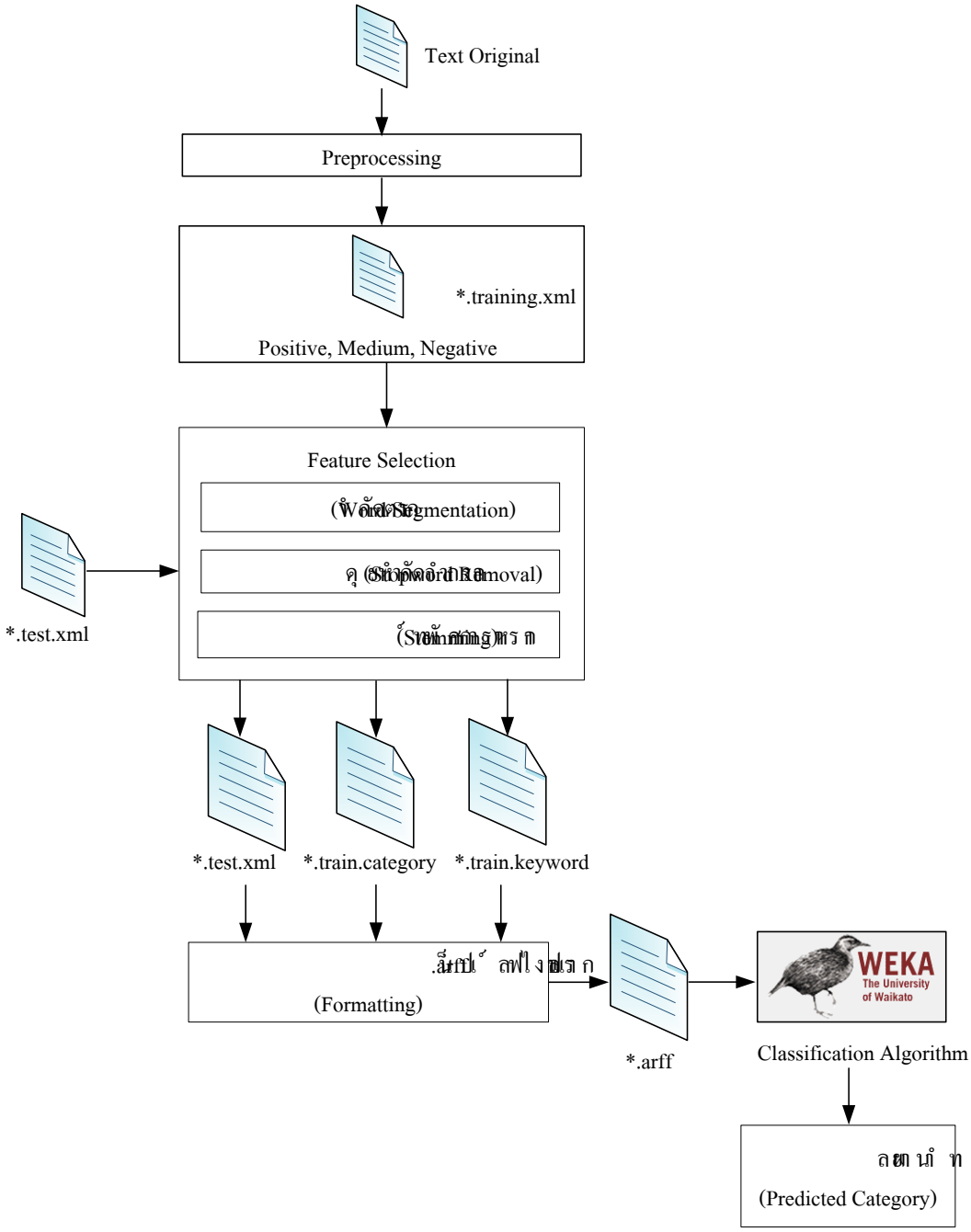


ภาพที่ 3-6 รูปแบบคำสั่งและผลการรันการแปลงไฟล์เป็น ARFF ของข้อมูลทดสอบ

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