Science and Technology

RMUTT Journal

Vol.6 No.1 (2016) : 94-101

www.sci.rmutt.ac.th/stj

Online ISSN 2229-1547

การจำแนกกลุ่มข้อความรีวิว โดยใช้เทคนิคเหมืองข้อมูล

Text Review using data mining Classification Technique

ประพัฒน์ พรมน้ำอ่าง[1] วสุวรรธน์ พงศ์ขจร[1] และ นิเวศ จิระวิชิตชัย[1*]

[1]หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาระบบสารสนเทศคอมพิวเตอร์

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม 61 ถนนพหลโยธิน เขตจตุจักร กรุงเทพมหานคร 10900

*E-mail: nivet.ch@spu.ac.th

บทคัดย่อ

ปัจจุบันเทคโนโลยีมีบทบาทในชีวิตประจำวันมากมายและมีข้อความวิจารณ์มากมายบนเว็บที่ใช้ในการบ่งบอกถึงความน่าสนใจในตัวสินค้าที่ขายในเวปไซต์ วิธีดังกล่าวสามารถช่วยให้ผู้ซื้อสามารถอ่านข้อความวิจารณ์ของสินค้าได้ก่อนการตัดสินใจสั่งซื้อ ทำให้มันมีประโยชน์เป็นอย่างมากในการเสนอขายสินค้าบนเวปไซต์ บทความนี้นำเสนอการแบ่งกลุ่มข้อความจากข้อความรีวิว โดยใช้เทคนิคเหมืองข้อมูล ซึ่งประกอบด้วยเทคนิค SVM เทคนิค Decition Tree เทคนิค k-NN และ เทคนิค Naïve Bayes จากการทดลองพบว่าโดยเทคนิค SVMได้ค่าความถูกต้องสูงที่สุดอยู่ที่ 86.26 %

คำสำคัญ: ข้อความรีวิว เหมืองข้อมูล

Abstract

Recently, technology plays a lot of role in daily life and has plenty text review on the web which is used to specify the interesting points in the product selling in the website. This method helps the purchaser able to read the text review of the product before deciding to make the purchase order which makes high advantage in proposing selling products on the website. This article presents the Movie Text Review using data mining classification technique consists of SVM, Decision Tree, K-NN and Naïve Bayes techniques. From the experiment, found that by using the SVM technique makes the highest accurate value of 86.26%

Keywords: Text review, Data Mining

Received: December 11, 2015

Revised: April 22, 2016

Accepted: April 29, 2016

1. บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ปัจจุบันมีการใช้อินเตอร์เน็ตกันอย่างแพร่หลาย มีการแลกเปลี่ยนความคิดเห็นเกี่ยวกับเรื่องต่างๆ ตามหัวข้อที่ผู้ใช้งานมีความสนใจในเรื่องนั่น จึงเกิดเป็นเครือข่ายทางสังคมขึ้นมา อาทิเช่น ข้อมูลสินค้า สถานที่ท่องเที่ยว การเข้าพักโรงแรม แหล่งรับประทานอาหาร เป็นต้น ข้อมูลการรีวิวภาพยนต์ก็ถือเป็นอีกเรื่องหนึ่งที่ผู้ใช้งานอินเตอร์เน็ตเข้ามาแลกเปลี่ยนประสบการณ์ในการรับชมและความคิดเห็นในรูปแบบของข้อความล้วนแต่จะทำให้เกิดข้อมูลจากการรีวิวขึ้นมากมาย ทำให้การค้นหาข้อมูลเกิดความยุ่งยากและใช้เวลาในการค้นหาข้อมูลที่สนใจมีระยะเวลาที่นาน บางครั้งข้อมูลที่ได้รับก็ไม่ได้ตรงตามความต้องการของผู้ใช้งานเพื่อให้เกิดประโยชน์มากที่สุดเราจำเป็นต้องนำข้อมูลมากมายเหล่านี้มา ทำการวิเคราะห์ (analyze)

เทคนิคการจำแนกประเภทข้อมูล (Data Classification) เป็นเทคนิคหนึ่งที่สำคัญของการสืบค้นความรู้บนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Database: KDD) หรือดาต้าไมน์นิง (Data Mining) ข้อมูลที่นำมาจำแนกอาจเป็นข้อมูลที่มีโครงสร้าง (Structured Data) หรือไม่มีโครงสร้าง (Unstructured Data) การจำแนกข้อความเป็นกระบวนการทำเหมืองข้อความ (Text Mining) เพื่อทำการค้นหาองค์ความรู้ที่ซ่อนอยู่ในข้อความ จากการจำแนกข้อมูลทำให้ได้ข้อมูลที่ง่ายต่อการสืบค้นต่อไป [1-2]

1.2 วัตถุประสงค์ของการศึกษา

1.2.1 เพื่อพัฒนารูปแบบในการจำแนกข้อความ

1.2.2 เพื่อหาความแม่นยำของรูปแบบในการจำแนกข้อความรีวิวโดยใช้เทคนิค Data Mining

1.3 ขอบเขตของการศึกษา

1.3.1 สามารถนำหลักการ เทคนิค SVM (Linear) เทคนิค Decition Tree เทคนิค k-NN และเทคนิค Naïve Bayes มาประยุกต์ใช้ในการจำแนกข้อความรีวิว

1.3.2 สามารถวัดประสิทธิภาพของแต่ละเทคนิคที่นำมาเปรียบเทียบว่าเทคนิคใดสามารถทำนายหาค่าความแม่นยำที่ดีที่สุด

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ทราบถึงการะบวนการทางเทคนิคในการประยุกต์ใช้การทำเหมืองข้อความในการจำแนกชุดข้อความที่มีประสิทธิภาพสูงสุด

2. วัสดุอุปกรณ์และวิธีดำเนินการวิจัย

2.1 วัสดุอุปกรณ์

องค์ประกอบหลักที่จำเป็นในการพัฒนารูปแบบในการจำแนกข้อความ มีองค์ประกอบหลักด้วยกัน 2 องค์ประกอบคือ ด้านฮาร์ดแวร์และซอฟต์แวร์

2.1.1 ความต้องการทางด้านฮาร์ดแวร์

ฮาร์ดแวร์ที่รองรับมีองค์ประกอบ ดังนี้

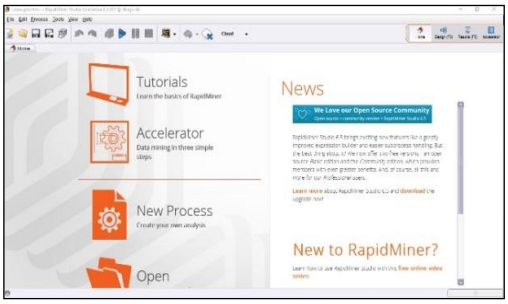
เครื่องคอมพิวเตอร์

- หน่วยประมวลผลกลาง 2.0 GHz
- Hard disk ขนาด 750 GB
- RAM 8 GB

2.1.2 ความต้องการทางด้านซอฟต์แวร์

ซอฟต์แวร์ที่ช่วยสนับสนุนในการค้นหาประสิทธิภาพการจำแนกข้อความที่เป็นสแปม ดังนี้

1. Operation System : Microsoft Windows 10
2. Application : RapidMiner Studio 6.5



รูปที่ 1 แสดงตัวอย่างหน้าเริ่มต้นของโปรแกรมสำเร็จรูป RapidMiner Studio

RapidMiner [4-5] เป็นแพลตฟอร์มซอฟต์แวร์ที่พัฒนา โดยบริษัท RapidMiner มีสำนักงานใหญ่อยู่ที่ประเทศสหรัฐอเมริกา เป็นเครื่องมือสำหรับวิเคราะห์ การเรียนรู้ของเครื่อง (machine learning) การทำเหมืองข้อมูล (data mining) การทำเหมืองข้อความ (text mining) การวิเคราะห์เชิงพยากรณ์ (predictive analytics) และการวิเคราะห์เชิงธุรกิจ (business analytics) สร้างเพื่อใช้สำหรับธุรกิจ การวิจัยการศึกษา RapidMiner เป็นซอฟต์แวร์แบบ open source ได้รับการรับรองใน Sourceforge มีทั้งแบบ Starter Edition Professional Edition และ Enterprise Edition

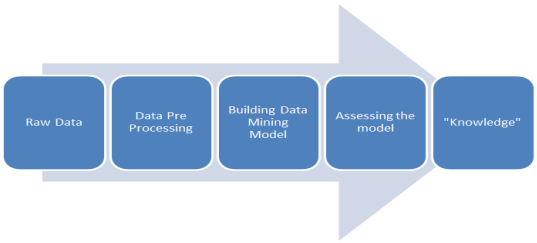
2.2 วิธีดำเนินการวิจัย

การศึกษาทฤษฎีวิธีการของการทำเหมืองข้อมูล (Data mining Algorithms) Data Mining คือ ชุด software วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนการตัดสินใจของผู้ใช้ มันเป็น software ที่สมบรูณ์ทั้งเรื่องการค้นหา การทำรายงาน และโปรแกรมในการจัดการ ซึ่งเราคุ้นเคยดีกับคำว่า Executive Information System (EIS) หรือระบบข้อมูลสำหรับการตัดสินใจในการบริหาร ซึ่งเป็นเครื่องมือชิ้นใหม่ที่สามารถค้นหาข้อมูลในฐานข้อมูลขนาดใหญ่หรือข้อมูลที่เป็นประโยชน์ในการบริหาร ซึ่งเป็นการเพิ่มคุณค่าให้กับฐานข้อมูลที่มีอยู่ [1-3]

ระบบสนับสนุนการตัดสินใจ (Decision Support System) คือทำอย่างไรให้ข้อมูลที่เรามีอยู่กลายเป็นความรู้อันมีค่าได้สร้างคำตอบของอนาคตได้

ฐานข้อมูลขนาดใหญ่จะประกอบไปด้วยข้อมูลเป็นพันๆ ล้านไบต์ ยากแก่การค้นหาได้อย่างทันกาลด้วยวิธี DBMS (Database Management System) โดยทั่วไป ข้อมูลที่เป็นที่สนใจของผู้บริหารธุรกิจวันนี้สามารถจะค้นหาได้ง่ายขึ้นแล้ว ซึ่งจะเป็นประโยชน์อย่างยิ่งในการค้นหาข้อมูลที่ต้องการในมหาสมุทรข้อมูลเพื่อนำมาเทียบเคียงและดูแนวโน้ม และนำข้อมูลที่จำเป็นของบริษัทส่งกลับให้ผู้บริหารตัดสินใจได้อย่างทันกาล

นี่คือจุดประสงค์ของ Data Mining ที่จะมาช่วยในเรื่องของการค้นข้อมูลสำคัญที่ปะปนกับข้อมูลอื่น ๆ ในฐานข้อมูลที่ไม่ใช่แค่การสุ่มหา บางคนเรียกว่า KDD (Knowledge Discovery in Database) หรือ การค้นหาข้อมูลด้วยความรู้ และนั่นก็คือ Data Mining



รูปที่ 2 แสดงตัวอย่างการทำเหมืองข้อมูล

เทคนิค Naive Bayes [7-8] คือโมเดลการคัดแยกประเภทข้อมูลที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem และสมมติฐานที่ให้การเกิดของเหตุการณ์ต่างๆเป็นอิสระต่อกัน (independence) กำหนดให้ P(H) ความน่าจะเป็นที่จะเกิดเหตุการณ์ H และ P(H|E) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E จากตัว

แปรที่กำหนดและแนวคิดของ Bayes' Theorem นั้น เราสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่างๆ ได้ดังสมการ

$$P(H|E) = [P(E|H)xP(H)]/P(E)$$

เทคนิค K-Nearest Neighbor (k-NN) [6-7] เป็นอัลกอลิทึ่มที่ใช้ในการจัดกลุ่มข้อมูล (Classification) ซึ่งเป็นอัลกอลิทึ่มที่อยู่ในกลุ่มของ Supervised learning (ข้อมูลที่นำมาเรียนรู้จะต้องมี Label คอยบอกไว้ว่าคือข้อมูลอะไร) ดังนั้น จึงเป็นการจัดกลุ่มข้อมูลที่อยู่ใกล้กันเข้าไว้ด้วยกัน โดยที่ k หมายถึงจำนวนข้อมูลที่นำมาพิจารณา

- กำหนดขนาดของ K (ควรกำหนดให้เป็นเลขคี่)
- คำนวณระยะห่าง (Distance) ของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลตัวอย่าง
- จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน K ที่กำหนดไว้
- พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่ม (Class) ไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด
- กำหนด class ให้กับจุดที่พิจารณา จากตัวอย่างข้างต้น อาจกำหนดให้ k= 5 ดังนั้นสังเกตว่าระยะทางที่ใกล้กับจุด (3,3) มากที่สุด 5 ลำดับ

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) [6-7] เป็นโครงสร้างข้อมูลชนิดเป็นลำดับชั้น (hierarchy) ใช้สนับสนุนการตัดสินใจ โดยจะมีลักษณะคล้ายต้นไม้จริงกลับหัวที่มีโหนดรากอยู่ด้านบนสุดและโหนดใบอยู่ล่างสุดของต้นไม้ ภายในต้นไม้จะประกอบไปด้วยโหนด (node) ซึ่งแต่ละโหนดจะมีคุณลักษณะ (attribute) เป็นตัวทดสอบ กิ่งของต้นไม้ (branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (leaf) ซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไม้ตัดสินใจแสดงถึงกลุ่มของข้อมูล (class) หรือนั่นก็คือผลลัพธ์ที่ได้จากการทำนาย โหนดที่อยู่บนสุดของต้นไม้เรียกว่าโหนดราก (root node)ดังแสดงโครงสร้างของต้นไม้ตัดสินใจตัดสินใจ

เทคนิค Support Vector Machine [6-7] เป็นอัลกอริทึมในการคัดแยกที่มีการนำมาใช้กันอย่างกว้างขวางในด้านการประมวลผลเป็นภาพดิจิตอล หลักการของ SVM คือการให้อินพุทที่ใช้ฝึกเป็นเวคเตอร์ในสเปซ N มิติ เช่นถ้าในกรณีของ 2 มิติ และ 3 มิติ จะเป็นจุดที่อยู่ในระนาบ xy และสเปซ xyz ตามลำดับ จากนั้นทำการสร้างไฮเปอร์เพลน (Hyperplane) ที่จะแยกกลุ่มของเวคเตอร์อินพุทออกเป็นประเภทต่างๆ ในกรณีที่เป็น 2 มิติ และ 3 มิติ ไฮเปอร์เพลปน คือเส้นตรงและระนาบตามลำดับ ข้อเด่นของ SVM จะทำการเก็บแมพ (Map) เวคเตอร์ในสเปซอินพุทให้เข้าสู่ Feature Space โดยใช้ฟังก์ชั่นหรือเรียกว่าเคอร์นัล (kernel) ชนิดต่างๆ เช่น โพลีโนเมียล (Polynomial) เรเดียล (Radial) เป็นต้น ใน Feature Space ดังกล่าวเวคเตอร์อินพุท สามารถแยกประเภทได้โดยไฮเปอร์เพลน

การออกแบบระบบการจัดกลุ่มของข้อความรีวิว โดยใช้โปรแกรม RapidMiner Studio เพื่อใช้ในการวิเคราะห์ข้อมูลต่างๆ ใช้ โอเปอเรเตอร์ดังต่อไปนี้

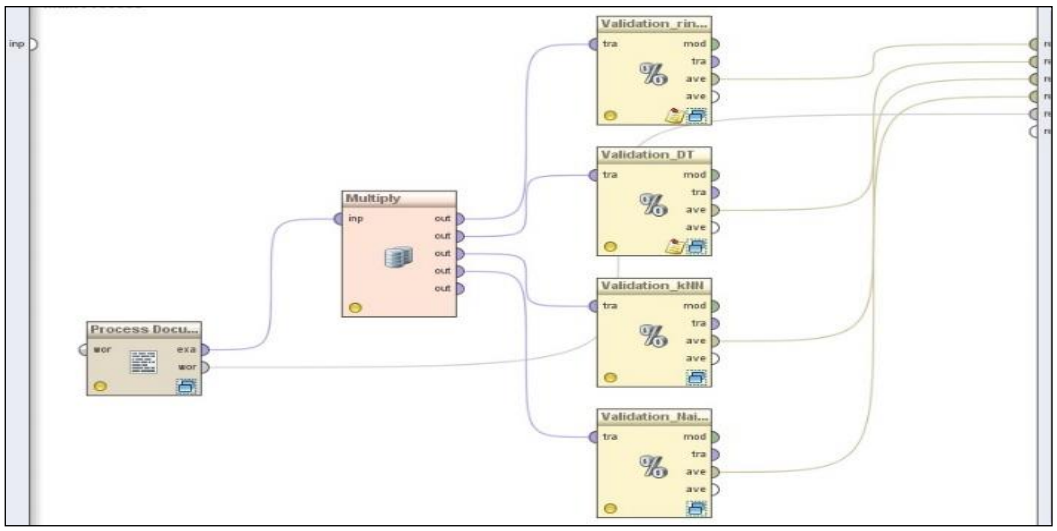
Process Document From Files(Text Processing)

- Tokenize
- Filter Tokens (by Length)
- Stem (poter)
- Filter stopwords (English)

Multiply(RapidMiner Studio Core)

X-Validation(RapidMiner Studio Core)

- Validation Naive Bayes
- Validation k-NN
- Validation Decition Tree
- Validation SVM(Linear)



รูปที่ 3 แสดงการออกแบบระบบการจำแนข้อความรีวิว

3. ผลการวิจัยและวิจารณ์ผลการวิจัย

การรายงานผลประสิทธิภาพการจำแนกข้อความรีวิว รายละเอียด ดังนี้

3.1 เทคนิค Naïve Bayes

PerformanceVector

```
PerformanceVector:
accuracy: 80.77% +/- 0.50% (mikro: 80.77%)
ConfusionMatrix:
True:   positive        negative
positive:       10106   2413
negative:       2394    10087
precision: 80.82% +/- 0.54% (mikro: 80.82%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       10106   2413
negative:       2394    10087
recall: 80.70% +/- 0.76% (mikro: 80.70%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       10106   2413
negative:       2394    10087
AUC (optimistic): 0.911 +/- 0.005 (mikro: 0.911) (positive class: negative)
AUC: 0.812 +/- 0.006 (mikro: 0.812) (positive class: negative)
AUC (pessimistic): 0.812 +/- 0.006 (mikro: 0.812) (positive class: negative)
```

รูปที่ 4 การรายงานผลประสิทธิภาพการจำแนกข้อความรีวิวโดยใช้เทคนิค Naïve Bayes

- ค่า true positive คือ ค่าที่ทำนายว่าถูกจริง คือ 10,106 จาก 12,500 ข้อความ
- ค่า true negative คือ ค่าที่ทำนายถูกว่าผิดจริง คือ 10,087 จาก 12,500 ข้อความ
- ค่าความถูกต้องของเทคนิค Naive Bayes คือ 80.77 %

3.2 เทคนิค k-NN

PerformanceVector

```
PerformanceVector:
accuracy: 68.54% +/- 1.09% (mikro: 68.54%)
ConfusionMatrix:
True:   positive        negative
positive:       9189    4554
negative:       3311    7946
precision: 70.61% +/- 1.49% (mikro: 70.59%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       9189    4554
negative:       3311    7946
recall: 63.57% +/- 1.13% (mikro: 63.57%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       9189    4554
negative:       3311    7946
AUC (optimistic): 0.903 +/- 0.007 (mikro: 0.903) (positive class: negative)
AUC: 0.500 +/- 0.000 (mikro: 0.500) (positive class: negative)
AUC (pessimistic): 0.467 +/- 0.014 (mikro: 0.467) (positive class: negative)
```

รูปที่ 5 การรายงานผลประสิทธิภาพการจำแนกข้อความรีวิวโดยใช้เทคนิค k-NN

- ค่า true positive คือ ค่าที่ทำนายว่าถูกจริง คือ 9,189 จาก 12,500 ข้อความ
- ค่า true negative คือ ค่าที่ทำนายถูกว่าผิดจริง คือ 7,946 จาก 12,500 ข้อความ
- ค่าความถูกต้องของเทคนิค K-NN คือ 68.54 %

3.3 เทคนิค ต้นไม้ตัดสินใจ (Decition Tree)

PerformanceVector

```
PerformanceVector:
accuracy: 62.58% +/- 0.90% (mikro: 62.58%)
ConfusionMatrix:
True:   positive        negative
positive:       12230   9084
negative:       270     3416
precision: 92.69% +/- 1.81% (mikro: 92.67%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       12230   9084
negative:       270     3416
recall: 27.33% +/- 1.72% (mikro: 27.33%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       12230   9084
negative:       270     3416
AUC (optimistic): 0.979 +/- 0.005 (mikro: 0.979) (positive class: negative)
AUC: 0.624 +/- 0.010 (mikro: 0.624) (positive class: negative)
AUC (pessimistic): 0.270 +/- 0.017 (mikro: 0.270) (positive class: negative)
```

รูปที่ 6 การรายงานผลประสิทธิภาพการจำแนกข้อความรีวิวโดยใช้เทคนิค ต้นไม้ตัดสินใจ (Decition Tree)

- ค่า true positive คือ ค่าที่ทำนายว่าถูกจริง คือ 12,230 จาก 12,500 ข้อความ
- ค่า true negative คือ ค่าที่ทำนายถูกว่าผิดจริง คือ 3,416 จาก 12,500 ข้อความ
- ค่าความถูกต้องของเทคนิคต้นไม้ตัดสินใจ (Decition Tree) คือ 62.58 %

3.4 เทคนิค SVM(Linear)

PerformanceVector

```
PerformanceVector:
accuracy: 86.26% +/- 0.50% (mikro: 86.26%)
ConfusionMatrix:
True:   positive        negative
positive:       11141   2077
negative:       1359    10423
precision: 88.47% +/- 0.75% (mikro: 88.47%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       11141   2077
negative:       1359    10423
recall: 83.38% +/- 0.89% (mikro: 83.38%) (positive class: negative)
ConfusionMatrix:
True:   positive        negative
positive:       11141   2077
negative:       1359    10423
AUC (optimistic): 0.938 +/- 0.003 (mikro: 0.938) (positive class: negative)
AUC: 0.938 +/- 0.003 (mikro: 0.938) (positive class: negative)
AUC (pessimistic): 0.938 +/- 0.003 (mikro: 0.938) (positive class: negative)
```

รูปที่ 7 การรายงานผลประสิทธิภาพการจำแนกข้อความรีวิวโดยใช้เทคนิค SVM(Linear)

- ค่า true positive คือ ค่าที่ทำนายว่าถูกจริง คือ 11,141 จาก 12,500 ข้อความ
- ค่า true negative คือ ค่าที่ทำนายถูกว่าผิดจริง คือ 10,423 จาก 12,500 ข้อความ
- ค่าความถูกต้องของเทคนิค SVM(Linear) คือ 86.26 %

3.5 WordList(Process Document from Files)

แสดงข้อมูลจำนวนคำที่ได้จากการประมวณผลคุณลักษณะคำ จากกระบวนการประมวณผลของโปรแกรม RapidMiner ทั้งหมด จะได้คำต่างๆที่ออกมาเป็นจำนวน 1,433 คำ จากข้อความรีวิว 12,500 ข้อความ

Word	Attribute Name	Total Occurences	Document Occurences	positive	negative
abandon	abandon	269	251	130	139
abil	abil	589	543	331	258
abl	abl	1367	1221	745	622
abov	abov	840	795	436	404
absolut	absolut	1840	1666	797	1043
absurd	absurd	321	289	114	207
abus	abus	395	303	203	192
accent	accent	671	549	223	448
accept	accept	813	691	464	349
accid	accid	370	323	230	140
accord	accord	278	264	114	164
account	account	304	273	153	151
accur	accur	401	357	238	163
achiev	achiev	530	478	328	202
act	act	7466	6132	2904	4562
action	action	3525	2477	1708	1817
actor	actor	6719	5080	3064	3655
actress	actress	1456	1235	755	701
actual	actual	4943	3873	1904	3039
ad	ad	638	610	345	293
adapt	adapt	755	603	404	351
add	add	369	355	254	115
addit	addit	379	365	226	153
admir	admir	443	408	264	179
admit	admit	665	631	341	324
adult	adult	738	613	425	313
adventur	adventur	726	598	477	249

รูปที่ 8 การแสดงข้อมูลจำนวนคำที่ได้จากการประมวณผล

4. สรุปผลการวิจัย

ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (Precision) และค่าความอ่อนไหว (recall) ดังแสดงในตาราง

Performance	Naïve Bayes (%)	K-NN (%)	Decision Tree (%)	SVM (Linear) (%)
Accuracy	80.77	68.54	62.58	86.26
Precision	80.82	70.61	92.69	88.47
Recall	80.70	63.57	27.33	83.38

จากการทดสอบโดยใช้โปรแกรม RapidMiner นั้น ทางเทคนิค SVM (Linear)ให้ค่าความถูกต้อง (accuracy) ดีที่สุดคือ 86.26 % จึงเป็นเทคนิคที่น่าสนใจ ที่จะนำไปประยุกต์ใช้ในการจำแนกกลุ่มข้อความวิจารณ์ภาพยนตร์ จากกระบวนการที่นำเสนอข้างต้นยังสามารถนำไปประยุกต์ใช้ในงานประเภทอื่นๆได้อีกด้วย เช่น ข้อมูลรีวิวสินค้า ข้อมูลการรีวิวโรงแรม เป็นต้น

5. เอกสารอ้างอิง

[1] Vishal Gupta. “A Survey of text Mining Techniques and Applications.” Journal of Emerging technologies in web intelligence, 2009

[2] Markus Hofmann, Ralf Klinkenberg. Rapidminer Data Mining Use Cases and Business Analytics Applications: CRC Press, 2013

[3] Mohammed Zaki , Wagner Meira Jr. Data Fundamental Concepts and Algorithms Data Mining and Analysis: Cambridge University Press,2014.

[4] เอกสิทธิ์ พัชรวงศ์ศักดา, Introduction to Business Analytics with RapidMiner Studio 6 (ฉบับภาษาไทย), กรุงเทพฯ: หสม. ดาต้า คิวบ์, 2557.

[5] เอกสิทธิ์ พัชรวงศ์ศักดา, คู่มือการวิเคราะห์ข้อมูลด้วย RapidMiner Studio 6, กรุงเทพฯ: หสม. ดาต้า คิวบ์, 2557.

[6] นิเวศ จิระวิชิตชัย “การจำแนกความคิดเห็นโดยใช้การลดคุณลักษณะร่วมกับการกำหนดค่าน้ำหนักดัชนีของคำ” วารสารวิศวสารลาดกระบัง (Ladkrabang Engineering Journal) คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ปี 2554

[7] นิเวศ จิระวิชิตชัย “แบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ” วารสารวิชาการเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี 2555

[8] ฐิติมา เกษมศรีธนาวัฒน์, ธนัสนี เพียรตระกูล. การจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบเบย์ร่วมกับการเลือกคุณลักษณะด้วยอัลกอริทึมรีลีฟ: CIT&UniNOMS2011, 2011