ระบบแนะนำภาพยนตร์โดยการทำเหมืองข้อความจากคำวิจารณ์ของผู้ชม

MOVIE RECOMMENDATION SYSTEM USING TEXT MINING FROM AUDIENCE'S COMMENTS

ฉัฐเมศร์ พูลทรัพย์
อุษนิษา เถาว์โท

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2560

ระบบแนะนำภาพยนตร์โดยการทำเหมืองข้อความจากคำวิจารณ์ของผู้ชม

MOVIE RECOMMENDATION SYSTEM USING TEXT MINING FROM AUDIENCE'S COMMENTS

ณัฐเมศร์ พูลทรัพย์
อุษนิษา เถาว์โท

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2560

ปริญญานิพนธ์ปีการศึกษา 2560

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง ระบบแนะนำภาพยนตร์โดยการทำเหมืองข้อความจากคำวิจารณ์ของผู้ชม

MOVIE RECOMMENDATION SYSTEM USING TEXT MINING FROM AUDIENCE'S COMMENTS

ผู้จัดทำ

1. นายฉัฐเมศร์ พูลทรัพย์ รหัสนักศึกษา 57010466
2. นางสาวอุษนิษา เถาว์โท รหัสนักศึกษา 57011553

_____ อาจารย์ที่ปรึกษา

(รศ. ดร.เกียรติกูล เจียรนัยธนะกิจ)

ระบบแนะนำภาพยนตร์โดยการทำเหมืองข้อความจากคำวิจารณ์ของผู้ชม

นายฉัฐเมศร์	พูลทรัพย์	57010466
นางสาวอุษนิษา	เถาว์โท	57011553
รศ. ดร.เกียรติกูล	เจียรนัยธนะกิจ	อาจารย์ที่ปรึกษา

ปีการศึกษา 2560

บทคัดย่อ

ปริญญานิพนธ์นี้ได้ถูกจัดทำขึ้นโดยมีวัตถุประสงค์เพื่อศึกษากระบวนการประมวลผลทางภาษาและการเรียนรู้ของเครื่องและทดลองสร้างโมเดลจำแนกข้อความภาษาไทยที่วิจารณ์ภาพยนตร์ออกมาเป็นแง่มุม โดยแบ่งออกเป็น 5 แง่มุม ได้แก่ นักแสดง, เสียง, ภาพหรือกราฟิก บท และภาพรวม และวิเคราะห์ความรู้สึกต่อแง่มุมนั้น ๆ เป็นแง่บวกและแง่ลบ โมเดลที่ใช้จำแนกทั้งสองประเด็นจะใช้อัลกอริทึมที่แตกต่างออกไปโดยโมเดลสำหรับวิเคราะห์แง่มุมของภาพยนตร์ด้วยอัลกอริทึมต้นไม้ตัดสินใจส่วนโมเดลวิเคราะห์ความรู้สึกของข้อความด้วยอัลกอริทึมป่าแบบสุ่ม ในการทดสอบประสิทธิภาพของโมเดลทั้งสองโมเดล จะนำชุดข้อมูลที่ใช้ทดสอบนำเข้าไปเพื่อให้โมเดลทำนายและสรุปผลของประสิทธิภาพการจำแนกแง่มุมและความรู้สึกที่มีต่อภาพยนตร์แต่ละคลาสด้วยคอนฟิวชันเมทริกซ์ โมเดลนี้จะนำมาจะจัดกลุ่มข้อความวิจารณ์ภาพยนตร์ออกมาเป็นแง่มุมและความรู้สึกต่อแง่มุม นำมาเป็นคะแนนของภาพยนตร์แต่ละเรื่อง หลังจากนั้นจะมีระบบแนะนำภาพยนตร์อื่นให้แก่ผู้ใช้โดยดูจากคะแนนที่ใกล้เคียงกันกับภาพยนตร์ที่ผู้ใช้กำลังรับชม นอกจากนี้ยังสามารถคัดกรองภาพยนตร์โดยอาศัยเกณฑ์การตัดสินใจตามที่ผู้ใช้งานกำหนดได้

MOVIE RECOMMENDATION SYSTEM USING TEXT MINING FROM AUDIENCE'S COMMENTS

Mr. Chattamet	Poonsub	57010466
Ms. Usanisa	Taoto	57011553
Assoc.Prof.Dr. Kietikul	Jearanaitanakij	Advisor

Academic Year 2017

ABSTRACT

The purpose of this research is to build two classifiers for categorizing audience's comments about the movie into various aspects and predicting the sentiments of those aspects. While the aspect of the movie can be classified into five classes, i.e., actor, graphic, sound, plot and general, the sentiment of the aspect can be categorized into positive and negative feelings. We collect audience's comments from ten volunteers and feed the training data to decision tree and random forest classifiers. The performance of each classifier model is shown in confusion matrix. The model can be applied to classify the aspect and the sentiment from audience's comment. In addition, our movie recommendation system also suggests related movies based on similar summary scores. We can customize the search filter to fit the need of each user as well.

กิตติกรรมประกาศ

คณะผู้จัดทำขอขอบพระคุณ รศ. ดร.เกียรติกูล เจียรนัยธนะกิจ อาจารย์ที่ปรึกษาโครงงานที่คอยให้คำแนะนำ ให้ความช่วยเหลือทั้งในเรื่องความรู้ แนวทางการพัฒนาโครงงานและสนใจสอบถามความคืบหน้าของโครงงานนี้อย่างสม่ำเสมอ

ขอขอบคุณเพื่อน ๆ นักศึกษาในสาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ ที่เป็นกำลังใจ ให้คำปรึกษา ช่วยเหลือซึ่งกันและกันเสมอมา รวมทั้งขอขอบพระคุณสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่เป็นสถานศึกษาที่ดีและมอบโอกาสต่าง ๆ ให้เสมอมา

ฉัฐเมศร์ พูลทรัพย์
อุษนิษา เถาว์โท

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

เนื่องจากในปัจจุบันระบบวิจารณ์ภาพยนตร์ตามเว็บไซต์ชื่อดังเช่น Rotten Tomato, IMDB หรือเว็บไซต์อื่น ๆ จะให้นักวิจารณ์ภาพยนตร์ได้ทำการเขียนบทวิจารณ์ภาพยนตร์ในหลาย ๆ องค์ประกอบของภาพยนตร์พร้อมกับระบุคะแนนของภาพยนตร์เรื่องนั้น คะแนนและบทความจากผู้วิจารณ์ภาพยนตร์เป็นสิ่งที่ทำให้ผู้ชมมีความสนใจในตัวภาพยนตร์ แต่การที่ผู้ชมหนึ่งคนจะเข้าใจว่าหนังเรื่องนี้มีองค์ประกอบไหนที่ดีหรือไม่ดีบ้าง ผู้ชมจะต้องไปไล่อ่านบทวิจารณ์ของนักวิจารณ์ภาพยนตร์ทำให้ผู้ชมต้องใช้เวลาในการอ่านค่อนข้างนาน อีกทั้งบทวิจารณ์ส่วนใหญ่จะใช้ภาษาอังกฤษทำให้ผู้ที่ไม่ถนัดในการอ่าน และแปลภาษาอังกฤษต้องใช้เวลาเพิ่มขึ้นไปอีก เพื่อที่จะทำให้ผู้ชมภาพยนตร์หรือผู้ที่ต้องการอ่านบทวิจารณ์ลดเวลาในการอ่าน ผู้จัดทำจึงออกแบบระบบที่สามารถแปลงบทวิจารณ์ภาพยนตร์ให้อยู่ในรูปแบบอื่นที่ผู้ชมภาพยนตร์สามารถเข้าใจได้ถึงบทวิจารณ์ได้ภายในเวลาอันสั้น

จากปัญหาที่กล่าวมาข้างต้นทำให้ผู้จัดทำได้คิดระบบที่สามารถเปลี่ยนบทวิจารณ์ให้กลายเป็นคะแนนได้โดยอาศัยหลักการของการเรียนรู้ของเครื่อง (Machine Learning) เพื่อที่จะทำให้ระบบสามารถเรียนรู้ถึงบทวิจารณ์ของภาพยนตร์ และเปลี่ยนบทวิจารณ์เหล่านั้นกลายเป็นคะแนนด้วยตนเอง และระบบนี้จะทำการวิเคราะห์ไปถึงแต่ละองค์ประกอบของภาพยนตร์ทำให้แยกแยะได้ว่าองค์ประกอบไหนของภาพยนตร์ดีหรือไม่ดีอย่างไร

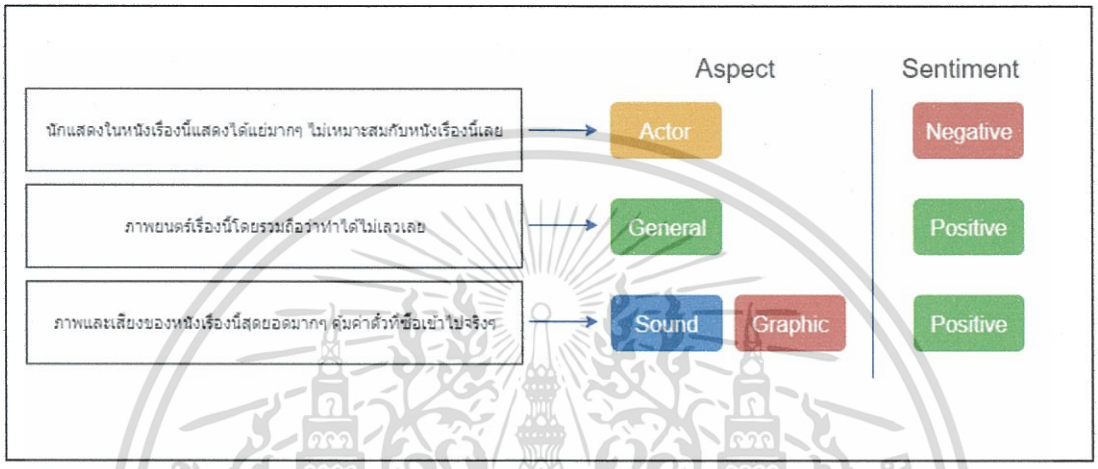
ระบบที่ได้กล่าวไปนั้นจะทำให้ผู้ชมภาพยนตร์ไม่ต้องไปใช้เวลาในการที่จะอ่านบทวิจารณ์และสามารถสรุปได้ในทันทีว่าภาพยนตร์ในเรื่องที่ตนเองต้องการจะไปดูนั้น มีจุดเด่น จุดด้อยอย่างไรบ้าง มีความน่าชมมากน้อยเพียงใด และช่วยในการตัดสินใจเพื่อการรับชมภาพยนตร์

1.2 วัตถุประสงค์ของการศึกษา

1) เพื่อนำความรู้หลาย ๆ ด้าน เช่น การเรียนรู้ของเครื่อง (Machine Learning), การประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาประยุกต์ใช้เพื่อพัฒนาระบบการประเมินบทวิจารณ์ได้ด้วยตัวเอง
2) ลดเวลาการอ่านบทวิจารณ์ของผู้ชมภาพยนตร์

1.3 ขอบเขตของโครงงาน

ระบบการวิเคราะห์บทวิจารณ์ภาพยนตร์เป็นระบบที่สามารถแยกแง่มุมของภาพยนตร์จากบทวิจารณ์ออกเป็น 5 แง่มุม ได้แก่ นักแสดง (Actor & Actress), เสียง (Sound), บทหรือเนื้อหา(plot), ภาพหรือกราฟฟิก (Graphic) และภาพรวม (General) ระบบจะทำการวิเคราะห์อีกด้วยว่าบทวิจารณ์ดังกล่าวเป็นบทวิจารณ์เชิงบวกหรือเชิงลบ



รูป 1.1 ตัวอย่างผลลัพธ์ระบบวิเคราะห์บทวิจารณ์ภาพยนตร์

การพัฒนาระบบวิเคราะห์บทวิจารณ์ภาพยนตร์ผลลัพธ์จะถูกแบ่งออกเป็น 2 ส่วนได้แก่ องค์ประกอบภาพยนตร์ (Aspect) และความรู้สึกที่มีต่อภาพยนตร์ (Sentiment) ทั้ง 2 ส่วนจะมีกระบวนการดังนี้

1) องค์ประกอบของภาพยนตร์ (Aspect) การพัฒนาจะถูกแบ่งเป็น 5 ส่วน ได้แก่ การเตรียมข้อมูลบทวิจารณ์ภาพยนตร์, การคัดกรองข้อมูล (Data Preprocessing), การคัดเลือกคำสำคัญ (Feature Word), การทดลองวัดความแม่นยำความถูกต้องของระบบ, การประยุกต์ผลลัพธ์ไปใช้งาน
2) ความรู้สึกที่มีต่อภาพยนตร์ (Sentiment) การพัฒนาจะถูกแบ่งออกเป็น 7 ส่วนได้แก่ การเตรียมข้อมูลบทวิจารณ์ภาพยนตร์, การคัดกรองข้อมูล (Data Preprocessing), การคัดเลือกคำสำคัญ (Feature Word), การคัดเลือกข้อมูลต้นแบบ, การเลือกใช้การแบ่งประเภทข้อมูล (classifier), การทดลองวัดความแม่นยำความถูกต้องของระบบ, การประยุกต์ผลลัพธ์ไปใช้งาน

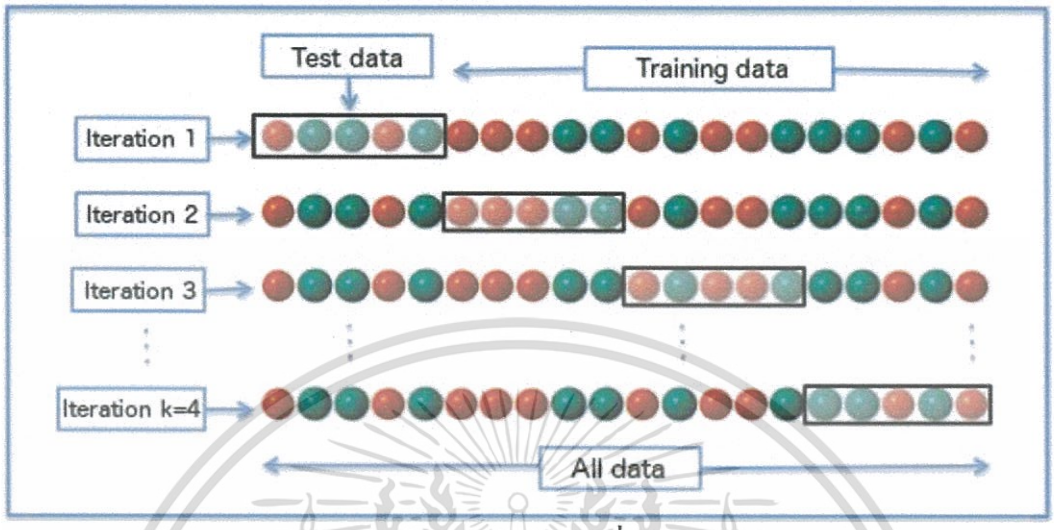
ขอบเขตของโครงงานจะอยู่ที่ส่วนสุดท้ายของการพัฒนาของทั้ง 2 ส่วน คือ การประยุกต์ผลลัพธ์ไปใช้งาน โดยจะพัฒนาในรูปแบบของเว็บแอปพลิเคชัน

1.4 ขั้นตอนการดำเนินงาน

การพัฒนาระบบวิเคราะห์บทวิจารณ์ภาพยนตร์นั้นจะมีขั้นตอนการดำเนินงานดังนี้

1) ศึกษาพื้นฐานการใช้ภาษาไพทอน (Python) และการใช้ไพทอนเพื่อทำการประมวลผลทางภาษาธรรมชาติ (Natural Language Processing)
2) เตรียมข้อมูลบทวิจารณ์ภาพยนตร์จากเว็บไซต์ต่าง ๆ เช่น พันทิป (Pantip), เฟซบุ๊ก (Facebook) และทำการเก็บข้อมูลเหล่านั้นอยู่ในรูปของเจสัน (JavaScript Object Notation: JSON) โดยใช้ไลบรารีที่มีชื่อว่าเซเลเนียม (Selenium) ร่วมกับส่วนต่อประสานโปรแกรมประยุกต์ (application programming interface: API) ของเว็บไซต์นั้น ๆ
3) ทำการคัดกรองข้อมูล โดยให้นำสัญลักษณ์ต่างๆที่ไม่เป็นประโยชน์ต่อการวิเคราะห์ของระบบออก ได้แก่ นำเครื่องหมาย “(”, “)”, “ๆ” เป็นต้น ออก รวมไปถึงการตัดประโยคออกเป็นคำ ๆ โดยในระบบนี้จะใช้ไลบรารี่ที่มีชื่อว่าไพไทยเอ็นแอลพี (Pythainlp) ในการทำงานดังกล่าว
4) การคัดเลือกคำสำคัญ โดยจะถูกแยกออกเป็น 3 ส่วน
   a) คำสำคัญของส่วนแง่มุมของภาพยนตร์ (Aspect) คือคำนามที่ระบุถึงองค์ประกอบของภาพยนตร์โดยตรง ได้แก่ นักแสดง, บท, พล็อต เป็นต้น และอาจรวมไปถึงกริยาที่ระบุถึงแง่มุมของภาพยนตร์ ได้แก่ แสดง, กำกับ เป็นต้น
   b) คำสำคัญของส่วนความรู้สึกที่ต่อภาพยนตร์ (Sentiment) คือคำกริยาวิเศษณ์ต่าง ๆ ทั้งหมด
   c) คำอื่น ๆ ที่ไม่ได้ปรากฏจะถือว่าเป็นคำไม่สำคัญ และตัดทิ้งออกจากการวิเคราะห์ของระบบไป
5) การคัดเลือกข้อมูลต้นแบบ จะทำการคัดเลือกประโยคหรือชุดข้อความที่มีความหมายที่แสดงถึงองค์ประกอบของภาพยนตร์และเป็นความคิดเห็นของการชมภาพยนตร์อย่างชัดเจน และทำการจัดหมวดหมู่ให้สอดคล้องกับข้อมูลต้นแบบ
6) การเลือกใช้การแบ่งประเภทข้อมูล (classifier) ระบบจะใช้การแบ่งประเภทข้อมูล จะไลบรารีที่มีชื่อว่า ไซคิตเลิร์น (SciKit-learn) โดยเริ่มจากการที่แยกข้อมูลต้นแบบออกเป็น 2 ส่วน (ไม่จำเป็นต้องเท่ากัน) ส่วนแรกจะนำไปเป็นข้อมูลที่มีพร้อมผลเฉลยเข้าไปให้กับการแบ่งประเภทข้อมูล และอีกส่วนหนึ่งจะเป็นข้อมูลชุดทดสอบเพื่อตรวจสอบความถูกต้องของการแบ่งประเภทข้อมูล ทั้งนี้ในไลบรารีของไซคิตเลิร์น จะมีการแบ่งประเภทข้อมูล ให้เลือกใช้หลายตัว
7) การทดลองวัดความแม่นยำความถูกต้องของระบบ ขั้นตอนนี้จะทำการตรวจสอบจากข้อมูลชุดทดสอบว่าตัวการแบ่งประเภทข้อมูล (classifier) นั้นตอบถูกแค่ไหนในระบบนี้จะ

คิดจากอัลกอริทึมการเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K-fold cross validation) ในการตรวจสอบ (ในระบบใช้ K=10)



รูป 1.2 การเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม (K=4)

8) การประยุกต์ผลลัพธ์ไปใช้งาน จะทำการนำผลลัพธ์ที่ได้จากขั้นตอนที่กล่าวมาทั้งหมดประยุกต์ไปกลายเป็นฟังก์ชันบนเว็บแอปพลิเคชัน เช่น สามารถรับความคิดเห็นของผู้ชมภาพยนตร์ และนำมาคิดคะแนนให้กับภาพยนตร์เรื่องนั้น ๆ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถพัฒนาระบบที่มีประโยชน์ให้กับผู้ชมภาพยนตร์โดยลดเวลาที่ต้องใช้ในการอ่านบทวิจารณ์ภาพยนตร์
2) ได้รับความรู้ในการทำการประมวลผลภาษาธรรมชาติ (Natural Language Process) และนำความรู้ไปใช้กับอย่างอื่นได้ในอนาคต
3) รู้จักการแก้ปัญหาเพื่อให้โครงงานสำเร็จได้ด้วยดี

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ภาษาไพทอน (Python)

รูป 2.1 สัญลักษณ์ภาษาไพทอน

ภาษาไพทอน (Python) เป็นภาษาที่ถูกสร้างขึ้นมาจากภาษาซี โดยประมวลผลทีละบรรทัด ภาษาไพทอนจะมีจุดเด่นมากมายโดยเฉพาะการเป็น โอเพนซอร์ส (open source) และมีแพ็กเกจ (package) ที่เหมาะกับการทำวิทยาศาสตร์ข้อมูล (Data science) เช่น แพนดาส์ (Pandas) ไซคิตเลิร์น (Scikit-learn), เทนเซอร์โฟล(Tensorflow) อีกทั้งภาษาไพทอน ยังเป็นภาษาที่เรียนรู้ง่ายเหมาะกับคนที่ต้องการศึกษาด้านการเรียนรู้ของเครื่อง

2.1.2 กฎการเขียนภาษา Python เบื้องต้น

1) ไม่จำเป็นต้องมี ";" (semicolon) ปิดท้ายทุกบรรทัด
2) การประกาศตัวแปรสามารถทำได้ทันทีไม่มีคำเพื่อกำหนดชนิดของตัวแปร
3) ทุกคำสั่งของการวนลูปซ้ำ และเงื่อนไข จะต้องเปิดด้วย ":" (colon) และส่วน body ของคำสั่งเหล่านั้นจะต้องถูกย่อหน้า (indent) 1 ครั้งเพื่อเป็นการระบุว่านี่คือส่วน body ทุกๆบรรทัด

โปรแกรม 2.1 การใช้คำสั่งวนซ้ำในภาษาไพทอน

```
for i in range(0,10):
  print("this is my loop")
  print("body part")
```

4) การแสดงข้อความออกทางหน้าจอใช้คำสั่ง print ดังโปรแกรม 2.1

5) การรับข้อมูลขาเข้าสามารถใช้คำสั่ง input ดังโปรแกรม 2.2 โดยค่าที่รับมานั้นจะมีชนิดข้อมูลเป็นสายอักขระ (string)

โปรแกรม 2.2 ตัวอย่างการใช้ input ในภาษา Python

```
var1 = input("Enter your number : ")
print(var1)
```

6) ประโยคเงื่อนไขจะใช้คำว่า elif (condition) แทนคำว่า else if(condition) กรณีอื่นยังคงใช้เหมือนเดิม ดังโปรแกรม 2.3

โปรแกรม 2.3 ตัวอย่างการใช้ If-Else condition ในภาษา Python

```
age = 20
if(age > 18):
  print("you're studying at university")
elif(age > 13 and age <= 18):
  print("you're studying at high school")
else:
  print("you're studying at elemantary school")
```

2.2 การเรียนรู้ของเครื่อง (machine learning)

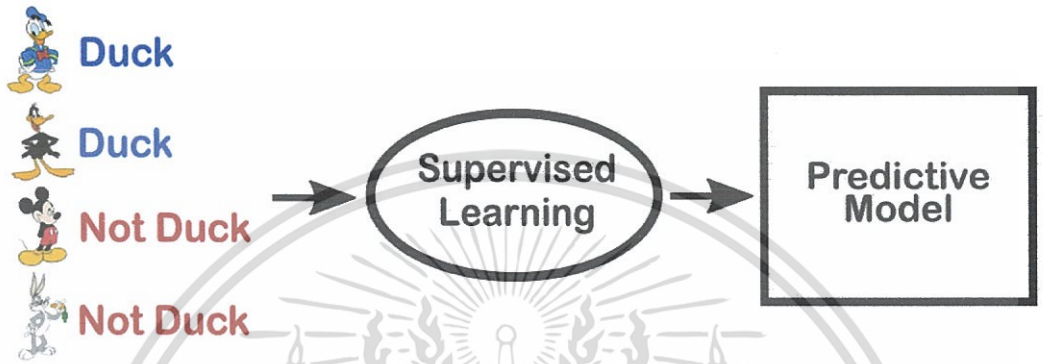
การเรียนรู้ของเครื่องเป็นกระบวนการที่ทำให้คอมพิวเตอร์สามารถที่จะเรียนรู้ด้วยตนเอง หรืออีกนัยหนึ่งคือสามารถคิดหรือทำนายสิ่งที่ต้องทำได้ด้วยตนเอง โดยปราศจากการทำงานตามลำดับคำสั่งโปรแกรม โดยการเรียนรู้ของเครื่องเป็นการรวมศาสตร์หลายแขนงเข้าด้วยกันได้แก่ วิทยาการคอมพิวเตอร์, วิศวกรรม และสถิติ

การเรียนรู้ของเครื่องนั้นถูกแบ่งออกเป็น 3 ประเภท ได้แก่ การเรียนรู้แบบมีผู้สอน (supervised learning), การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) และการเรียนรู้แบบเสริมกำลัง (reinforcement learning) ในที่นี้จะไม่กล่าวถึงการเรียนรู้แบบเสริมกำลัง เนื่องจากไม่เกี่ยวข้องกับการทำระบบวิจารณ์ภาพยนตร์

2.2.1 การเรียนรู้แบบมีผู้สอน (supervised learning)

การเรียนรู้แบบมีผู้สอนเป็นกระบวนการเรียนรู้ด้วยตนเองของคอมพิวเตอร์โดยให้มีมนุษย์เป็นผู้สอน โดยจะเริ่มจากการนำข้อมูลทั้งหมดมาแยกประเภทด้วยตนเองก่อน หรือบอกว่าข้อมูลเหล่านี้คืออะไรก่อน สามารถเอาค่าเหล่านั้นไปสอนให้กับคอมพิวเตอร์ว่าข้อมูลแบบนี้มันจะ

มีความหมายว่าอย่างไร หลังจากนั้นจึงทำการแยกข้อมูลทั้งหมดออกเป็น 2 ส่วน ส่วนแรกเป็นข้อมูลไว้ใช้ฝึกให้กับคอมพิวเตอร์หรือเรียกว่าข้อมูลสอน (training data) และอีกส่วนหนึ่งเป็นข้อมูลที่ใช้สำหรับทดสอบว่าคอมพิวเตอร์นั้นสามารถตอบคำตอบที่ถูกต้องได้เรียกอีกอย่างว่า ข้อมูลทดสอบ (testing data) จากรูป 2.2 จะเห็นได้ว่าต้องระบุก่อนว่าตัวการ์ตูนด้านซ้ายมือเป็นเป็ดใช่หรือไม่ก่อนที่จะนำเข้าสู่กระบวนการการเรียนรู้แบบไม่มีผู้สอน (supervised learning)



รูป 2.2 การใช้การเรียนรู้แบบมีผู้สอน จะถูกเรียกว่าการแบ่งประเภทข้อมูล (classification)

2.2.2 การเรียนรู้แบบไม่มีผู้สอน (unsupervised learning)

การเรียนรู้แบบไม่มีผู้สอนเป็นกระบวนการเรียนรู้ด้วยตนเองโดยที่ไม่มีมนุษย์เป็นผู้สอน ข้อมูลจะไม่ถูกทำการ จำแนกออกเป็นกลุ่มใด ๆ ไว้ โดยจะมีหลักการในการจำแนกจากการจับแบบรูป (pattern) ของข้อมูลว่ารูปแบบนี้ควรอยู่ใน label ไหน สิ่งที่มนุษย์ต้องกำหนดคือข้อมูลจะต้องถูกแบ่งออกเป็นกี่ label เพราะฉะนั้นคำตอบที่ได้จะเป็นเพียงการจัดกลุ่มว่าข้อมูลนี้อยู่ในกลุ่มอะไร จะไม่เหมือนการเรียนรู้แบบมีผู้สอน (supervised learning) ที่สามารถระบุได้เลยว่าข้อมูลแบบนี้คืออะไร การแบ่งกลุ่มของ การเรียนรู้แบบไม่มีผู้สอน เรียกอีกอย่างได้ว่า การแบ่งกลุ่มข้อมูล (clustering)

2.3 หลักไวยากรณ์ภาษาไทย

ภาษาไทยเป็นคำโดด ไม่มีการผันรูปตามเวลาหรือปริมาณ คำภาษาไทยไม่มีการระบุเพศ เมื่อต้องการขยายความต้องนำคำมาเรียกต่อกัน โดยคำขยายจะวางไว้หลังคำที่ต้องการขยาย และหากต้องการสร้างประโยคจะเรียงเป็น “ประธาน-กริยา-กรรม”

2.3.1 คำ

คำคือเสียงที่เปล่งออกมาแล้วมีความหมาย คำเป็นหน่วยที่เล็กที่สุดของภาษา พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554 ได้แบ่งคำไทยออกเป็น 8 ชนิดได้แก่

1) กริยา คือคำที่แสดงอาการของนามหรือสรรพนาม

2) นาม คือคำที่ใช้สำหรับเรียกคน สัตว์ หรือสิ่งของ

3) นิบาต คือคำที่มีชนิดของคำไม่ชัดเจน

4) บุพบท คือคำที่ทำหน้าที่เชื่อมต่อคำ อยู่หน้าคำนาม คำสรรพนามหรือคำกริยา บอกให้รู้ถึงตำแหน่งหรือหน้าที่

5) วิเศษณ์ (คุณศัพท์หรือกริยาวิเศษณ์) คือคำที่ขยายคำนาม คำกริยาหรือคำวิเศษร์ เพื่อบอกลักษณะหรือปริมาณ

6) สรรพนาม คือคำที่ใช้เรียกแทนคำนาม

7) สันธาน คือคำที่ใช้เชื่อมประโยคให้ต่อเนื่องกัน

8) อุทาน คือเสียงหรือคำที่เปล่งออกมาเพื่อแสดงอารมณ์หรือความรู้สึก

2.3.2 ประโยค

ประโยคคือหน่วยของภาษาที่มาจากกระประกอบคำแล้วได้ความหมายสมบูรณ์ ประกอบไปด้วยภาคประธาน และภาคแสดง ประโยคสามารถแบ่งได้ 2 แบบ ได้แก่ การแบ่งชนิดของประโยคตามเจตนารมย์ของผู้เสื่อสาร และการแบ่งประโยคตามลักษณะของโครงสร้างของประโยค การแบ่งประโยคตามลักษณะโครงสร้างของประโยคแบ่งออกเป็น 3 ชนิด ได้แก่

1) ประโยคความเดียว คือประโยคที่มีใจความสำคัญอย่างเดียว ประโยคชนิดนี้จะประกอบไปด้วยภาคประธาน และภาคแสดงอย่างละหนึ่งส่วน

2) ประโยคความรวม คือประโยคที่มีใจความสำคัญตั้งแต่สองใจความขึ้นไป ใจความสำคัญของประโยคย่อยอาจจะมีเนื้อหาคล้อยตามกัน ขัดแย้งกัน ให้เลือกอย่างใดอย่างหนึ่งหรือมีความเป็นเหตุเป็นผลกันก็ได้ แต่ทุกประโยคย่อยมีน้ำหนักของใจความสำคัญเท่ากัน

3) ประโยคความซ้อน คือประโยคที่ประโยคย่อยมีน้ำหนักของใจความสำคัญไม่เท่ากัน โดยจะมีส่วนหนึ่งเป็นประโยคหลัก นอกนั้นเป็นใจความเสริมประโยคหลัก

2.4 การประมวลผลภาษาธรรมชาติ (Natural language processing)

การประมวลผลภาษาทางธรรมชาติ (Natural language processing) เป็นศาสตร์หนึ่งในสาขาปัญญาประดิษฐ์ (Artificial intelligence) และ ภาษาศาสตร์คอมพิวเตอร์ (Computational linguistics) ซึ่งเกี่ยวกับการที่คอมพิวเตอร์ทำความเข้าใจ และมีปฏิสัมพันธ์กับภาษาธรรมชาติของมนุษย์ กระบวนการที่คอมพิวเตอร์ทำการวิเคราะห์ ทำความเข้าใจ หาความหมายจากภาษาธรรมชาติของมนุษย์ (ภาษาที่มนุษย์ใช้ติดต่อสื่อสารกัน เช่น ข้อความ, ภาษาพูด, ภาษามือ เป็นต้น) โดยจะหาวิธีการที่ทำให้คอมพิวเตอร์เข้าใจความหมายของคำ ประโยครวมไปถึงไวยกรณ์ของภาษาธรรมชาติ

ซึ่งกระบวนการเหล่านี้ทำให้คอมพิวเตอร์ทำงานเกี่ยวกับภาษาของมนุษย์ เช่น การแปลภาษา, การสืบค้นข้อมูลสารสนเทศ, การจดจำ และสังเคราะห์เสียงพูด, การทำเหมืองข้อความ เป็นต้น

2.4.1 การประมวลผลภาษาไทย

ในส่วนของการประมวลผลภาษาไทยนั้น เนื่องจากภาษาไทยจัดอยู่ในกลุ่มของภาษาที่ไม่ตัดคำ ไม่มีตัวอักษรใด ๆ บ่งบอกขอบเขตของคำอย่างชัดเจน ทำให้เกิดปัญหาเรื่องคำที่ไม่รู้จัก และคำกำกวม ทำให้ยังไม่มีเทคนิคในการตัดคำที่ให้ความถูกต้องในทุกกรณี

2.4.1.1 การตัดคำภาษาไทย (Thai word segmentation)

การตัดคำเป็นพื้นฐานที่สำคัญในการวิเคราะห์ และพัฒนาระบบที่เกี่ยวข้องกับการประมวลผลภาษา ปัจจุบันเทคนิคที่ใช้ในการตัดคำมี 3 วิธีหลัก ได้แก่

1) การใช้ไวยากรณ์ทางภาษา (Rule-based) เป็นวิธีตัดคำที่รวดเร็ว และไม่ต้องไม่ต้องเก็บคำจากพจนานุกรมไว้ในหน่วยความจำ
2) การอ้างอิงคำจากพจนานุกรม (Dictionary-based) มีความแม่นยำสูงแต่ต้องใช้พื้นที่ในหน่วยความจำขนาดใหญ่เพื่อเก็บคำในพจนานุกรม และในการปรับปรุงเวอร์ชันของพจนานุกรมมีความยุ่งยาก
3) การสร้างโมเดลเรียนรู้จากฐานข้อความขนาดใหญ่ (Corpus-based) เป็นวิธีที่ใช้การเรียนรู้ของเครื่อง (Machine Learning) ทำให้แก้ปัญหาเรื่องคำที่ไม่รู้จักได้จากการเรียนรู้จากข้อความขนาดใหญ่

2.4.1.2 การกำหนดหน้าที่ของคำ (Part-Of-Speech Tagging)

ในการตัดคำด้วยเทคนิคการสร้างโมเดลเรียนรู้จากฐานข้อความขนาดใหญ่จะต้องมีการจัดกลุ่มของคำตามหน้าที่ของคำในไวยากรณ์อย่างไร้ความกำกวม ซึ่งจะต้องคำนึงถึงความสัมพันธ์ของคำรอบข้าง ในการกำหนดหน้าที่ของคำ

ตาราง 2.1 หน้าที่ของคำ (Part-Of-Speech) ทั้งหมดพร้อมคำอธิบายและตัวอย่าง

ที่	POS	คำอธิบาย	ตัวอย่าง
1	NPRP	คำนามชื่อเฉพาะ	กรุงเทพ, ดารารัตน์, ยุโรป,
2	NCNM	ตัวเลขที่ใช้บอกปริมาณ	หนึ่ง, สอง, 1, 2
3	NONM	ตัวเลขที่ใช้บอกลำดับที่	ที่1, ที่สอง
4	NALM	คำกริยาวิเศษณ์ที่ทำหน้าที่เหมือนคำนาม	วันนี้, เมื่อก่อน
5	NAJN	คำคุณศัพท์ที่ทำหน้าที่เหมือนคำนาม	เชิงกายภาพ, ทางสิ่งแวดล้อม
6	NABL	คำนามที่ใช้กำกับป้าย	1, 2, 3, a, b, ก, ข
7	NCMN	คำนามที่ใช้เรียกชื่อทั่วไป	ผู้ใช้, หนังสือ, ม้า
8	NTTL	คำนามที่ใช้นำหน้าชื่อ	คุณ, นาย,พลเอก

9	NCLT	คำนามที่ใช้เรียกคำนามที่อยู่รวมกันเป็นกลุ่ม	กอง, กลุ่ม, คณะ
10	PPRS	คำสรรพนามแทนบุคคล	คุณ, ฉัน, เขา, มัน
11	PDMN	คำสรรพนามที่ใช้ชี้ระยะใกล้ ไกล	นั่น, โน่น, ที่นี่
12	PDEF	คำสรรพนามที่ใช้ชี้เฉพาะเจาะจง	เหล่านี้, ทั้งหมด
13	PIND	คำสรรพนามที่ไม่ชี้เฉพาะเจาะจง	บ้าง, ทั้งหลาย
14	PNTR	คำสรรพนามใช้ถาม	ใคร, อะไร
15	PREL	คำสรรพนามใช้ชี้ซ้ำคำนามที่อยู่ข้างหน้า	ที่, ซึ่ง, ว่า, อัน
16	VACT	คำกริยาที่มีการแสดงอย่างเป็นรูปธรรม	กิน, บรรยาย, ดู, คิด
17	VSTA	คำกริยาที่ไม่ได้มีการแสดงอย่างเป็นรูปธรรม	เป็น, อยู่, คือ, รู้, ราคา
18	VATT	คำวิเศษณ์ที่บอกลักษณะหรือคุณสมบัติ	ใหญ่, อ้วน, ร้อน, ดี
19	XVBM	คำกริยาช่วยที่นำหน้าคำว่า “ไม่”	จะ, กำลัง, อาจ
20	XVAM	คำกริยาช่วยที่ตามหลังคำว่า “ไม่”	ค่อย, น่า, ได้
21	XVMM	คำกริยาช่วยที่อยู่ได้ทั้งนำหน้าหรือหรือตามหลังคำว่า “ไม่”	ควร, เคย, ต้อง
22	XVBB	คำกริยาช่วยที่ขึ้นต้นประโยค	กรุณา, เชิญ, จง
23	XVAE	คำกริยาช่วยที่ตามหลังคำกริยา	แล้ว, อยู่
24	DDAN	คำบ่งชี้ที่อยู่หลังคำนาม	บ้านนี้, โน้น, นั่น
25	DDAC	คำบ่งชี้ที่อยู่หลังลักษณะนาม	สบู่ก้อนนั้น, นี้, นี่
26	DDBQ	คำบ่งชี้ที่อยู่ระหว่างคำนามและลักษณะนาม	ทั้ง, ตั้ง, เพียง
27	DDAQ	คำบ่งชี้ที่อยู่หลังคำวิเศษณ์บอกจำนวนในรูปแบบไม่มีเศษ	พอดี, ถ้วน
28	DIAC	คำบอกจำนวน	ตัวอื่น, ต่างๆ, ไหน
29	DIBQ	คำบอกจำนวนโดยการประมาณ	ประมาณ, ราว
30	DIAQ	คำบอกจำนวนซึ่งอยู่ในรูปแบบมีเศษ	เศษ, กว่า
31	DCNM	คำแสดงจำนวนนับ	สิบ, ร้อย, 1000
32	DONM	คำแสดงลำดับที่	ที่หนึ่ง
33	ADVN	กริยาวิเศษณ์ในรูปปกติ	เก่ง, ง่าย, ตรง
34	ADVI	กริยาวิเศษณ์ที่อยู่ในรูปแสดงความซ้ำ	ช้าๆ, เสมอๆ

35	ADVP	กริยาวิเศษที่ตามหลังคำอุปสรรค	อย่างรวดเร็ว, โดยเร็ว
36	ADVS	กริยาวิเศษณ์ที่ขยายความทั้งประโยค	ทั่วไป, โดยปกติ, ธรรมดา
37	CNIT	ลักษณะนามบอกรูปร่างสัณฐาน	ใบ, ตัว, เล่ม
38	CLTV	ลักษณะนามบอกหมวดหมู่	คู่, ฝูง
39	CMTR	ลักษณะนามที่เกี่ยวกับมาตรวัด	กิโลกรัม, วินาที, องศา
40	CFQC	ลักษณะนามบอกจำนวน	ครั้ง, เที่ยว
41	CVBL	ลักษณะนามบอกอาการ	กำ, ม้วน
42	JCRG	คำสันธานเชื่อมประโยคในระดับเดียวกัน	และ, แต่
43	JCMP	คำสันธานเชื่อมเนื้อความเพื่อเปรียบเทียบ	มากกว่า, เหมือนกับ
44	JSBR	คำสันธานเชื่อมเพื่อแสดงความเป็นเหตุเป็นผล	ดังนั้น, เพราะ
45	RPRE	คำบุพบท	ของ, บน, ใน
46	INT	คำอุทาน	อุ๊ย, แหม
47	FIXP	คำอุปสรรค	ความ, อย่าง
48	EAFF	คำลงท้ายในประโยคบอกเล่าเพื่อตอบรับ	ครับ, นะ, หน่อย
49	EITT	คำลงท้ายในประโยคคำถาม	เหรอ, ไหม, มั้ย
50	NEG	นิเสธ	ไม่, มิ
51	PUNC	เครื่องหมายวรรคตอน	?, !

2.5 เว็บแอปพลิเคชัน

2.5.1 จาวาสคริปต์ (JavaScript)

2.5.1.1 แนะนำจาวาสคริปต์เบื้องต้น

จาวาสคริปต์ (JavaScript) เป็นภาษาคอมพิวเตอร์ที่นิยมใช้ในการพัฒนาเว็บแอปพลิเคชันเนื่องจากภาษาจาวาสคริปต์มีความสามารถในการจัดการได้ทั้งฝั่งไคลเอนต์ (client) และฝั่งเซิร์ฟเวอร์ (server) ภาษาจาวาสคริปต์เป็นภาษาที่มีคุณสมบัติอะซิงโครนัส (asynchronous) ซึ่งจะแก้ไขปัญหาการขัดขวางคำสั่งถัดไปของภาษาที่เป็นซิงโครนัส (synchronous)

โปรแกรม 2.4 ตัวอย่างคุณสมบัติอะซิงโคตรนัสในภาษาจาวาสคริปต์

```
setTimeout(function(){
  console.log('Step 1')
}, 3000);

console.log('Step 2')
```

จากโปรแกรมดังกล่าวในฟังก์ชัน setTimeout จะทำงานฟังก์ชันด้านในเมื่อเวลาผ่านไป 3000 มิลลิวินาที แต่ด้วยคุณสมบัติอะซิงโครนัสของจาวาสคริปต์ ซึ่งจะทำให้โปรแกรมนี้แสดงผลว่า "step 2" ออกมาก่อน "step 1"

ตัวอย่าง 2.1 ผลลัพธ์การรันโปรแกรมที่ 2.4

```
"Step 2"
"Step 1"
```

ปัญหาของภาษาที่เป็นซิงโครนัสคือการที่ทำตามคำสั่งแบบเรียงลำดับแต่ถ้าคำสั่งก่อนหน้ายังประมวลผลไม่เสร็จจะทำให้คำสั่งถัดไปต้องรอเพราะฉะนั้นจากโปรแกรม 2.4 ถ้าภาษาที่ใช้มีคุณสมบัติซิงโครนัสจะทำให้การแสดง "step 2" ออกทางคอนโซลจะต้องรอเวลา 3 วินาทีก่อนถึงจะแสดงผลได้ และออกหลังจากการแสดง "step 1" ด้วย คุณสมบัติอื่นๆของจาวาสคริปต์นอกจากคุณสมบัติอะซิงโครนัสแล้ว จาวาสคริปต์เองสามารถที่จะทำให้ผู้ใช้งานตอบโต้หรือมีปฏิสัมพันธ์กับเว็บไซต์ได้ดี ยกตัวอย่างการคลิกที่ปุ่มใดที่สามารถใช้จาวาสคริปต์ในการตรวจจับการคลิกของผู้ใช้งานและควบคุมให้เว็บไซต์แสดงผลตามที่ต้องการเป็นต้น นอกจากนี้จาวาสคริปต์ยังสามารถเข้าไปเปลี่ยนแปลงค่าของส่วนประกอบเอชทีเอ็มแอล (HTML Element) เพื่อเปลี่ยนรูปแบบการแสดงผลของเว็บไซต์ได้ด้วย

2.5.1.2 การเขียนจาวาสคริปต์เบื้องต้น

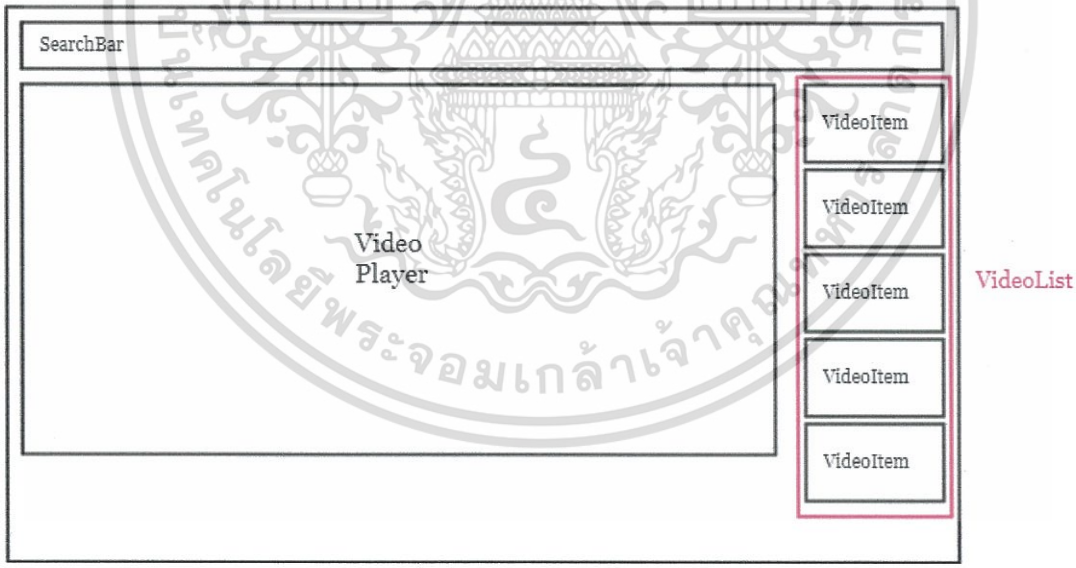
1) ตัวแปรจะไม่มีการกำหนด type ใดๆให้กับตัวแปรสามารถประกาศตัวแปรได้ด้วย keyword ต่างๆ หลักๆจะมี 3 ตัว let, var, const
2) สามารถใช้คำสั่ง console.log("some message") เพื่อให้แสดงออกข้อความทางคอนโซล
3) เครื่องหมายทางคณิตศาสตร์สามารถใช้ บวก (+), ลบ (-), คูณ (*), หาร (/), หารเอาเศษ(%) รวมไปถึงสามารถใช้ตัวดำเนินการกำหนดค่าแบบผสมได้ทุกตัว
4) คำสั่งที่ใช้เปรียบเทียบได้แก่ >=, >, <, <=, ==, !=, ===, !== หรือ greater than or equal, greater than, lower than, lower than or equal, equal, not-

equal, strict equal และ strict not-equal ตามลำดับ โดยเครื่องหมายที่ใช้เปรียบเทียบว่าเท่ากันจะมี 2 แบบ มีความแตกต่างกันที่ == จะวัดว่าค่าเท่ากันหรือไม่โดยไม่สนว่าเป็น type อะไรแต่ === จะตรวจสอบเรื่อง type ด้วย

5) คำสั่งเชื่อมนิพจน์ได้แก่ &&, ||, ! หรือ and, or, not ตามลำดับ
6) คำสั่งวนซ้ำได้แก่ for, while และ do...while
7) ฟังก์ชันในจาวาสคริปต์สามารถประกาศเป็นตัวแปรได้

2.5.2 รีแอกต์เฟรมเวิร์ก (React framework)

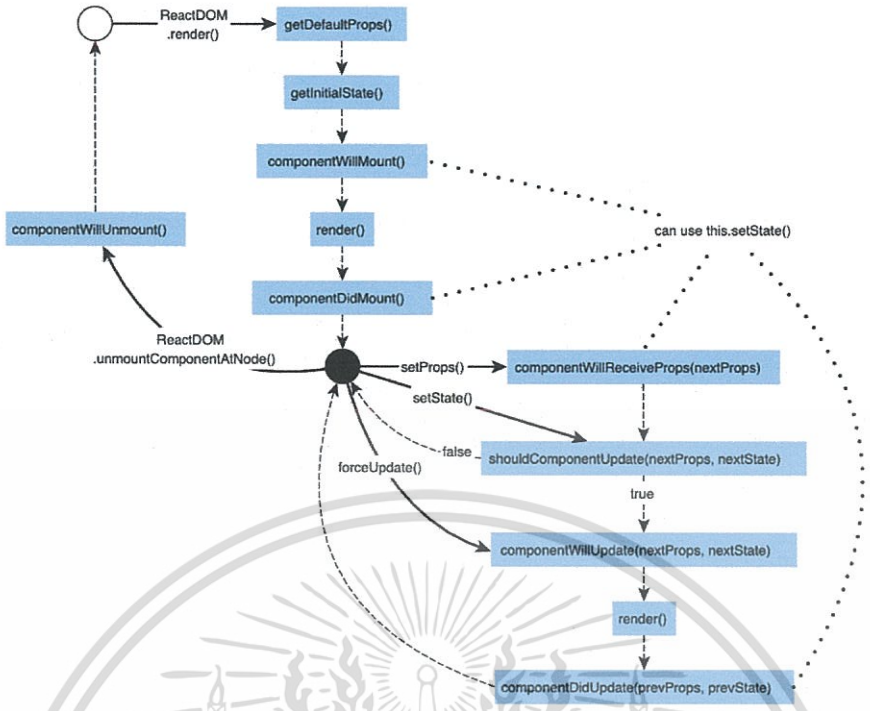
รีแอกต์เฟรมเวิร์กเป็นเฟรมเวิร์กส่วนหน้า (frontend framework) ที่ใช้ภาษาจาวาสคริปต์โดยมีคุณสมบัติในการที่แบ่งเว็บแอปพลิเคชันออกเป็นส่วนโปรแกรม (component) ที่ต่างกันหรือแบ่งเป็นส่วน โดยสำหรับผู้พัฒนาเว็บแอปพลิเคชันจะต้องออกแบบโดยคำนึงว่าในแต่ละหน้ามีส่วนประกอบอะไรบ้าง และแต่ละส่วนจะต้องทำหน้าที่อย่างไร รองรับและโต้ตอบกับผู้ใช้งานอย่างไรบ้าง มีตัวแปรใดบ้างที่ทำงานในแต่ละส่วนของเว็บแอปพลิเคชัน ยกตัวอย่างเช่น ผู้พัฒนาเว็บแอปพลิเคชันต้องการออกแบบหน้าเว็บไซต์สำหรับรับชมวิดีโอโดยที่มีเมนูสำหรับค้นหาและมีวิดีโอแนะนำให้ผู้ใช้งานอาจจะสามารถออกแบบได้ดังนี้



รูป 2.3 ตัวอย่างการแบ่งส่วนโปรแกรม (component) ในเว็บแอปพลิเคชัน

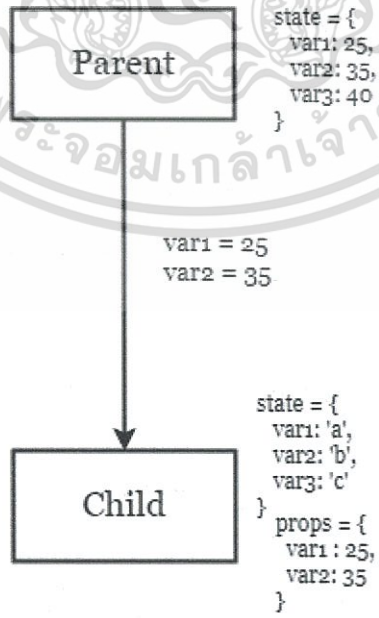
นอกจากนี้การออกแบบส่วนโปรแกรมในแต่ละส่วนของรีแอกต์จะสามารถมีตัวแปร (state) และฟังก์ชันได้ด้วย รวมไปถึงสามารถส่งค่าตัวแปรในส่วนโปรแกรมให้แก่ส่วนโปรแกรมตัวอื่น ๆ ได้ด้วย ทำให้รีแอกต์มีคุณสมบัติที่คล้าย ๆ กับภาษาจาวา คุณสมบัติของรีแอกต์เฟรมเวิร์กสามารถสรุปได้ 3 ข้อได้แก่

1) การนำมาใช้ซ้ำของส่วนโปรแกรม (Component reusability) สามารถออกแบบส่วนโปรแกรมเพียงอันเดียวแต่นำไปใช้หลาย ๆ รอบได้ และแต่ละ ส่วนโปรแกรมที่นำมาใช้ซ้ำ ๆ ไม่จำเป็นจะต้องเหมือนกันทุกประการ สามารถปรับเปลี่ยนค่าในแต่ละส่วนโปรแกรมเพื่อให้แสดงผลลัพธ์ได้อย่างถูกต้อง
2) ไลฟ์ไซเคิลเมท็อด (Lifecycle method) เป็นเมท็อดที่สำหรับการควบคุมลำดับการแสดงผล การเข้าถึงข้อมูลของส่วนโปรแกรมสำหรับภาษาจาวาสคริปต์ที่มีคุณสมบัติอะซิงโครนัสทำให้บางครั้งตัวแปรในส่วนโปรแกรมเกิดการแสดงผลก่อนที่ตัวแปรนั้นจะมีค่าตามที่กำหนด ยกตัวอย่างในกรณีที่ตัวแปรต้องรอค่าจากเอพีไอ เป็นอาทิ ทุก ๆ ส่วนโปรแกรมจะถูกควบคุมด้วยฟังก์ชัน ReactDOM.render() เป็น เมท็อดที่สามารถควบคุมว่าส่วนโปรแกรมใดบ้างจะถูกไปแสดงผล เมื่อฟังก์ชัน ReactDOM.render() ถูกเรียกใช้จะเริ่มกระบวนการนำส่วนโปรแกรมไปแสดงผล ก่อนหน้าที่จะแสดงผลได้เมท็อดที่มีชื่อว่า getDefaultProps, getInitialState และ componentWillMount จะถูกเรียกใช้งานก่อน และสามารถเรียกใช้คำสั่งอื่น ๆ เช่น การส่งรีเควสไปที่ตัวบริการ (server) ใด ๆ ภายในเมท็อดเหล่านี้ได้เพื่อเตรียมค่าตัวแปรให้มีค่าตามที่ต้องการก่อนที่จะถูกแสดงผล เมื่อส่วนโปรแกรมถูกแสดงผลจะมีเมท็อดที่ถูกเรียกใช้หลังจากการแสดงผลคือ componentDidMount ต่อมาเมื่อเกิดการเปลี่ยนแปลงใดๆของตัวแปรในส่วนโปรแกรมรวมไปถึงการเปลี่ยนแปลงในฟังก์ชัน componentDidMount ด้วย จะถูกเรียก render() ใหม่อีกครั้งเพื่อแสดงค่าใหม่ที่ได้รับมา และเมื่อผู้ใช้งานไม่ได้เรียกใช้งานส่วนโปรแกมนี้ต่อไปแล้วเช่นย้ายไปหน้าอื่น ๆ ของเว็บแอปพลิเคชัน เมท็อด componentWillUnmount จะถูกเรียกใช้งาน



รูป 2.4 ไลฟ์ไซเคิลเมท็อดของรีแอกต์เฟรมเวิร์ก

3) กระแสข้อมูลแบบน้ำตก (Waterfall dataflow) ในส่วนโปรแกรมทุกตัวสามารถที่จะส่งค่าไปให้กับส่วนโปรแกรมตัวอื่น ๆ ได้ โดยการส่งค่านั้นจะมีลักษณะที่ส่วนโปรแกรมที่เป็นพ่อ (parent) จะส่งข้อมูลให้กับตัวลูก (child) โดยที่ตัวลูกจะสามารถรับค่าจากคุณสมบัติ (props) ของตัวมันเอง



รูป 2.5 ตัวอย่างการส่งข้อมูลในระบบกระแสข้อมูลแบบน้ำตก

บทที่ 3

การออกแบบและการพัฒนา

3.1 การวิเคราะห์ความต้องการของระบบ

1) ระบบสามารถวิเคราะห์ข้อมูลได้ทั้ง 2 รูปแบบได้แก่ แง่มุมภาพยนตร์ (aspect) และความรู้สึกที่มีต่อภาพยนตร์ (sentiment)
2) ระบบสามารถวิเคราะห์ข้อมูลความคิดเห็นภาพยนตร์ที่เป็นภาษาไทยได้
3) ระบบสามารถใช้อัลกอริทึมที่เรียนรู้แบบมีผู้สอน (supervised learning) ในการวิเคราะห์ความคิดเห็นภาพยนตร์ได้
4) ระบบสามารถจำแนกองค์ประกอบภาพยนตร์และแสดงผลลัพธ์ได้โดยแบ่งเป็น 4 ประเภทได้แก่ ทั่วไป (general), นักแสดง (actor & actress), เสียง (sound), บทหรือเนื้อหา (plot) และ ภาพหรือกราฟิก (graphic)
5) ระบบสามารถจำแนกความรู้สึกที่มีต่อภาพยนตร์และแสดงผลลัพธ์ได้โดยแบ่งเป็น 2 ประเภทได้แก่ ความรู้สึกเชิงบวก (positive) และ ความรู้สึกเชิงลบ (negative)

3.2 การเลือกใช้เครื่องมือที่ใช้พัฒนา

3.2.1 ภาษา

ระบบจะพัฒนาด้วยภาษาไพทอน เนื่องจากเป็นภาษาที่ศึกษาง่าย และ มีแพ็กเกจ (package) สำหรับทำการเรียนรู้ของเครื่อง (machine learning) ได้ดี และเป็นโอเพนซอร์ส (open source)

ระบบจะใช้ package ต่างๆดังต่อไปนี้เพื่อทำการพัฒนาระบบ

3.2.1.1 ไพไทยเอ็นแอลพี (Pythainlp)

ที่มา: https://github.com/PyThaiNLP/pythainlp

เวอร์ชันที่ใช้ในระบบ: 1.4.1

รูป 3.1 คลังข้อมูลชิ้นงาน (repository) ของแพ็กเกจไพไทยเอ็นแอลพี

ไพไทยเอ็นแอลพีเป็นแพ็กเกจที่ใช้ตัดคำจากประโยคต่าง ๆ ได้ ในระบบจะนำคำที่ตัดได้มารวบรวมและนับความถี่ของคำเพื่อนำมาหาคำสำคัญที่เป็นประโยชน์ต่อการวิเคราะห์ความคิดเห็นของภาพยนตร์ ตัวอย่างของการตัดคำและการเรียกใช้จะพบได้ในโปรแกรม 3.1

โปรแกรม 3.1 แสดงตัวอย่างการตัดคำของแพ็คเกจไพไทยเอ็นแอลพี

```
from pythainlp.tokenize import word_tokenize
text='ผมรักคุณนะครับโอเคบ่พวกเราเป็นคนไทยรักภาษาไทยภาษาบ้านเกิด'
a=word_tokenize(text,engine='icu')
# ['ผม', 'รัก', 'คุณ', 'นะ', 'ครับ', 'โอ', 'เค', 'บ่', 'พวก', 'เรา',
'เป็น', 'คน', 'ไทย', 'รัก', 'ภาษา', 'ไทย', 'ภาษา', 'บ้าน', 'เกิด']
b=word_tokenize(text,engine='dict')
# ['ผม', 'รัก', 'คุณ', 'นะ', 'ครับ', 'โอเค', 'บ่', 'พวกเรา', 'เป็น', 'คน
ไทย', 'รัก', 'ภาษาไทย', 'ภาษา', 'บ้านเกิด']
c=word_tokenize(text,engine='mm')
# ['ผม', 'รัก', 'คุณ', 'นะ', 'ครับ', 'โอเค', 'บ่', 'พวกเรา', 'เป็น', 'คน
ไทย', 'รัก', 'ภาษาไทย', 'ภาษา', 'บ้านเกิด']
d=word_tokenize(text,engine='pylexto')
# ['ผม', 'รัก', 'คุณ', 'นะ', 'ครับ', 'โอเค', 'บ่', 'พวกเรา', 'เป็น', 'คน
ไทย', 'รัก', 'ภาษาไทย', 'ภาษา', 'บ้านเกิด']
e=word_tokenize(text,engine='newmm')
# ['ผม', 'รัก', 'คุณ', 'นะ', 'ครับ', 'โอเค', 'บ่', 'พวกเรา', 'เป็น', 'คน
ไทย', 'รัก', 'ภาษาไทย', 'ภาษา', 'บ้านเกิด']
```

3.2.1.2 ไซคิตเลิร์น (Scikit-learn)

ที่มา: http://scikit-learn.org/stable/

เวอร์ชันที่ใช้ในระบบ: 0.19

รูป 3.2 หน้าเว็บไซต์หลักของแพ็คเกจไซคิตเลิร์น (Scikit-learn package)

ไซคิตเลิร์นเป็นแพ็กเกจสำหรับทำการเรียนรู้ของเครื่อง (machine learning) และทำโมเดลทำนาย (predictive model) ด้วยภาษาไพทอน โดยในระบบจะใช้ในส่วนการแบ่งประเภทข้อมูล (classification) ของแพ็คเกจ เพื่อทำการจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment)

รูป 3.3 ตัวอย่างอัลกอริทึมในส่วนของการแบ่งประเภทข้อมูลของไซคิตเลิร์น

3.2.1.3 สปลินเตอร์ (Splinter)

ที่มา: http://splinter.readthedocs.io/

เวอร์ชันที่ใช้ในระบบ : 0.7.6

สปลินเตอร์ (Splinter) เป็น แพ็กเกจ (package) สำหรับดึงข้อมูลจากเว็บไซต์ (web scraping) ในระบบจะใช้แพ็กเกจนี้เพื่อรวบรวมชุดมูลอันได้แก่ข้อความที่เกี่ยวกับการวิจารณ์ภาพยนตร์

ตามเว็บไซต์ต่าง ๆ โดยใช้หลักการเอชทีทีพี รีเควส (HTTP Request) ไปยังเว็บเซิร์ฟเวอร์ (web server) เมื่อได้รับข้อมูลเป็นดีโอเอ็ม เอลิเมนต์ (DOM Element) ก็สามารถเลือกข้อมูลที่ต้องการดึงจากเว็บไซต์นั้นได้

โปรแกรม 3.2 แสดงรายชื่อกระทู้ในเว็บไซต์พันทิปห้องภาพยนตร์

```
from splinter import Browser
url = "https://pantip.com/tag/ภาพยนตร์"
with Browser('chrome', headless=True) as browser:
  browser.visit(url)
  forum_name = browser.find_by_css('.post-item-title')
  for item in forum_name:
      print(item.text)
```

3.2.2 โปรแกรมที่ใช้ในการพัฒนา

3.2.2.1 Anaconda Interpreter

ตัวแปลภาษาไพทอนให้เข้าสู่ภาษาเครื่อง (Machine code) โดย Anaconda จะเป็นตัวที่รวมแพ็กเกจพื้นฐานทั้งหมดของภาษาไพทอนเอาไว้ ทำให้สามารถเขียนโปรแกรมภาษาไพทอนได้ทันที เมื่อติดตั้ง Anaconda เสร็จ

3.2.2.2 jetBrain PyCharm Community Edition 2017.1

สิ่งแวดล้อมสำหรับการพัฒนาแบบเบ็ดเสร็จ (Integrated Development Environment) ที่เหมาะกับการเขียนโปรแกรมด้วยภาษาไพทอน และมีระบบดีบักเกอร์ (debugger) ที่ดี และยังมีระบบออโตคอมพลีต (autocomplete) ในการทำให้ผู้พัฒนาสามารถเขียนโค้ดได้ง่ายขึ้นรวมทั้งมีข้อผิดพลาดน้อยลง

3.2.2.3 Visual Studio Code

โปรแกรมแก้ไขข้อความ (text editor) ที่มีความเร็วมากที่สุด และมีส่วนเสริม (extension) มากมาย ในที่นี้ทางผู้พัฒนาได้ใช้เพื่อทำข้อมูลที่ใช้ฝึก (training data) ให้กับการแบ่งประเภทข้อมูล (classifier)

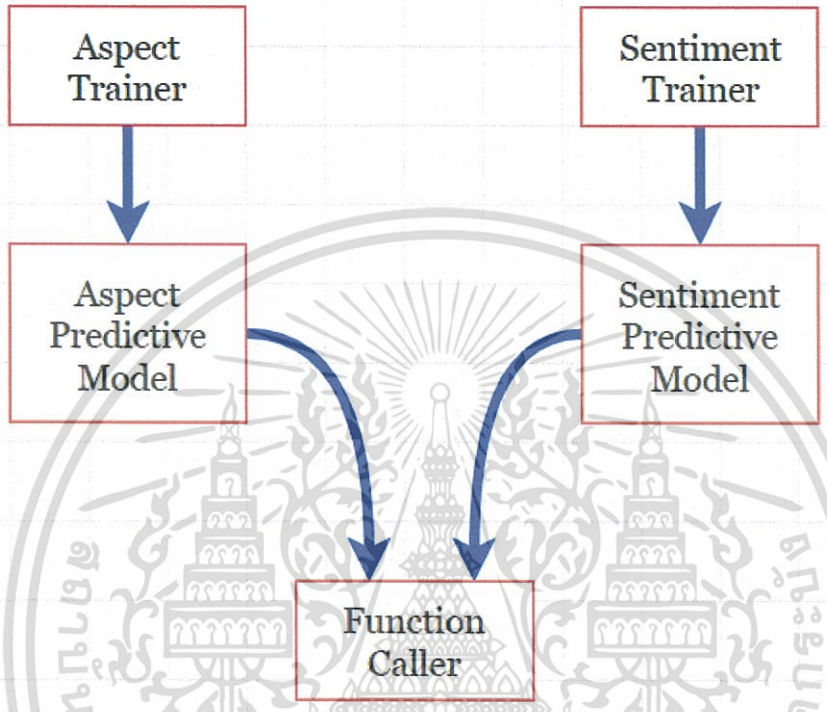
3.2.2.4 Firebase

บริการส่วนหลัง (backend service) ของกูเกิล (google) นำมาใช้เพื่อเป็นฐานข้อมูลสำหรับระบบแนะนำภาพยนตร์ (recommendation system) สำหรับการปรับเปลี่ยนคะแนนของภาพยนตร์ในแต่ละเรื่อง รวมไปถึงใช้รับความคิดเห็นที่มีต่อภาพยนตร์จากผู้ใช้งาน

3.3 การออกแบบ

3.3.1 ระบบวิเคราะห์บทวิจารณ์ภาพยนตร์

ระบบได้ถูกออกแบบให้มีการทำงานดังรูป 3.4 ซึ่งจะมีรายละเอียดดังต่อไปนี้



รูป 3.4 การแบ่งระบบวิจารณ์ภาพยนตร์

3.3.1.1 การฝึกจำแนกแง่มุมภาพยนตร์ (Aspect trainer)

ส่วนที่จะใช้ทำการสร้างการเรียนรู้ให้กับคอมพิวเตอร์ในเรื่องขององค์ประกอบภาพยนตร์ โดยในส่วนนี้จะต้องสามารถนำข้อมูลบทวิจารณ์ภาพยนตร์มาตัดคำและหลังจากนั้นจะต้องทำการเลือกคำที่ส่งผลหรือมีความหายโดยตรงกับองค์ประกอบภาพยนตร์มาสร้างถุงคำ (bag of word) ของแต่ละองค์ประกอบของภาพยนตร์ได้ และนำไปใช้ในส่วนของโมเดลจำแนกแง่มุมภาพยนตร์ (Aspect predictive model) ต่อไป

3.3.1.2 โมเดลจำแนกแง่มุมภาพยนตร์ (Aspect predictive model)

ส่วนที่นำถุงคำ (bag of word) จากส่วนแรกมาทำเป็นโมเดลการทำนาย โดยในส่วนนี้จะต้องทำการสร้างฟังก์ชันที่สามารถใส่ข้อมูลนำเข้าเป็นความคิดเห็นของภาพยนตร์ได้และส่งผลลัพธ์ออกมาเป็นองค์ประกอบภาพยนตร์ทั้งนี้ฟังก์ชันสามารถคืนค่าองค์ประกอบภาพยนตร์ได้ในรูปแบบลิสต์หรือตอบได้หลายคำตอบ

3.3.1.3 การฝึกจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment trainer)

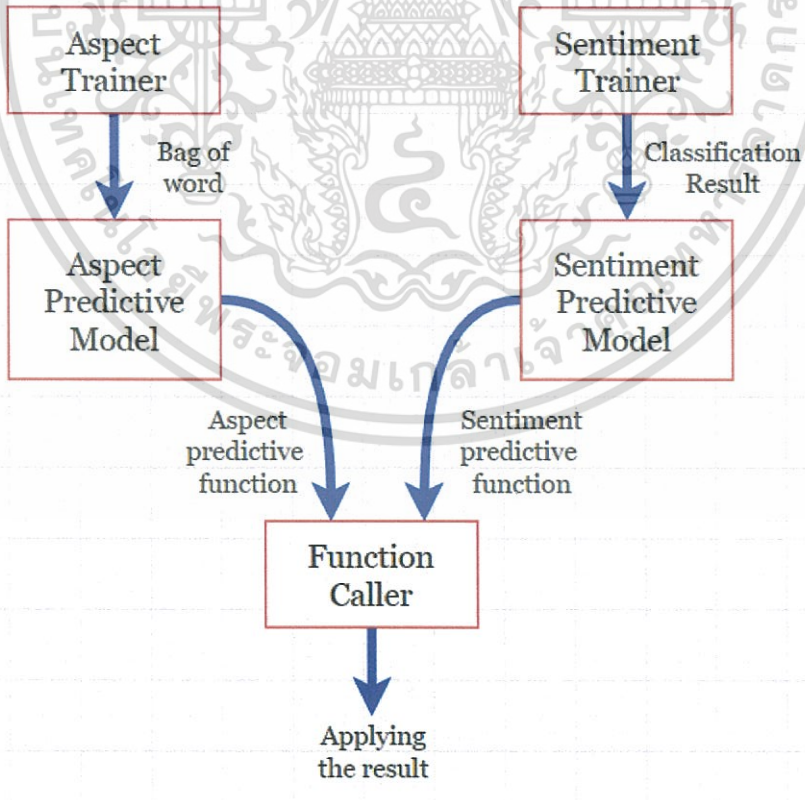
ส่วนที่จะใช้ทำการสร้างการเรียนรู้ให้กับคอมพิวเตอร์ในเรื่องของความรู้สึกที่มีต่อภาพยนตร์โดยในระบบจะต้องนำข้อมูลบทวิจารณ์มาตัดคำและหลังจากนั้นจะทำการคัดเลือกคำที่มีความหมายกับความรู้สึกที่มีต่อภาพยนตร์เป็นคำสำคัญ หลังจากนั้นจะนำข้อมูลบทวิจารณ์ภาพยนตร์ที่ผ่านการคัดกรองและจัดหมวดหมู่มาแล้ว มาเป็นข้อมูลเพื่อใช้ฝึกการเรียนรู้ของคอมพิวเตอร์ โดยใช้คำสำคัญเป็นเกณฑ์ตัดสิน หลังจากนั้นก็จะนำเข้าส่วนของการแบ่งประเภทข้อมูล (classification) เพื่อไปใช้ในส่วนถัดไป

3.3.1.4 โมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model)

ส่วนที่นำผลลัพธ์ของ การแบ่งประเภทข้อมูลเพื่อทำการทำนายความคิดเห็นภาพยนตร์มาใช้โดยในส่วนนี้จะต้องทำสร้างฟังก์ชันที่สามรถใส่ข้อมูลนำเข้าเป็นความคิดเห็นของภาพยนตร์ได้และส่งผลลัพธ์ออกมาเป็นความรู้สึกที่มีต่อภาพยนตร์

3.3.1.5 ฟังก์ชันคอลเลอร์ (Function caller)

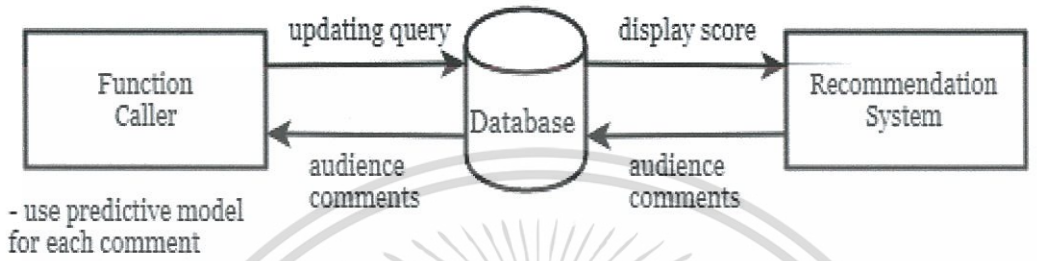
จากหัวข้อ 3.3.1.2 และหัวข้อ 3.3.1.4 จะได้ฟังก์ชันเพื่อหาคำตอบของความคิดเห็นภาพยนตร์ ในส่วนนี้จะทำการเรียกใช้ฟังก์ชันแยกออกมาเพื่อที่จะนำผลลัพธ์ไปต่อยอดหรือนำผลลัพธ์ที่ได้ไปใช้ในรูปแบบอื่น



รูป 3.5 การแบ่งระบบวิจารณ์ภาพยนตร์พร้อมแสดงผลลัพธ์ที่ควรจะได้ในแต่ละส่วนของระบบ

3.3.2 ระบบแนะนำภาพยนตร์

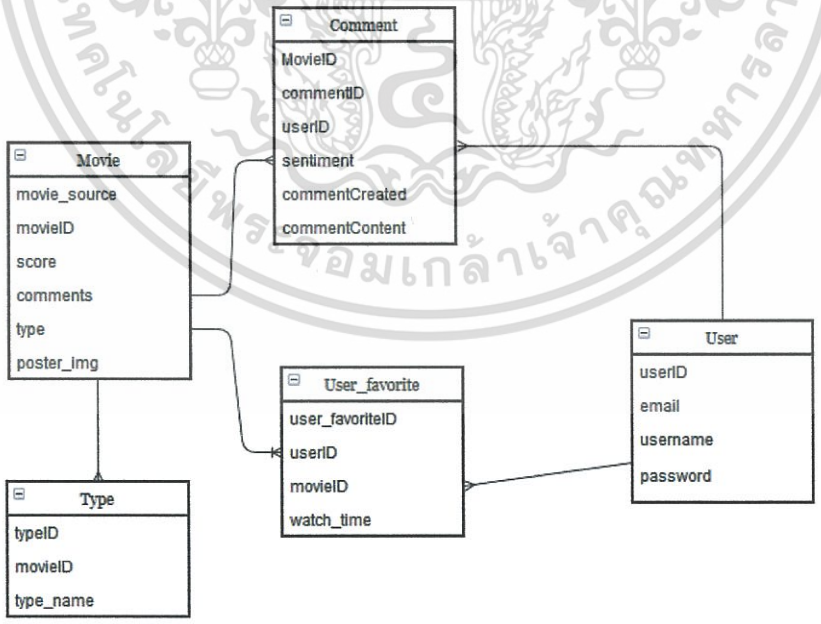
จากข้อ 3.3.1.5 สามารถเรียกใช้โมเดลจำแนกแง่มุมและความรู้สึกทั้ง 2 โมเดล จากฟังก์ชันคอลเลอร์ได้โดยการรับข้อมูลจากฐานข้อมูลซึ่งฐานข้อมูลจะรวบรวมความคิดเห็นของภาพยนตร์แต่ละเรื่องไว้ หลังจากนั้นจึงทำการเรียกใช้ฟังก์ชั่นเพื่อทำนายแง่มุมและความรู้สึกที่มีต่อภาพยนตร์และทำการคำณวนคะแนนและส่งกลับไปที่ฐานข้อมูล



รูป 3.6 การส่งข้อมูลในระบบแนะนำภาพยนตร์

3.3.2.2 ฐานข้อมูล (Database)

ฐานข้อมูลในระบบนี้จะต้องเก็บข้อมูลของภาพยนตร์เอาไว้ เช่น คะแนนของภาพยนตร์แต่ละเรื่อง แหล่งที่มาของวิดีโออื่น ๆ และในภาพยนตร์แต่ละเรื่องจะต้องเก็บความคิดเห็นของภาพยนตร์แต่ละเรื่องทำให้สรุปออกมาเป็นแผนภาพได้ดังนี้



รูป 3.7 แบบจำลองความสัมพันธ์เอนทิตีของระบบแนะนำภาพยนตร์

3.3.2.3 ระบบแนะนำ (recommendation system)

ในระบบแนะนำมีหน้าที่ในการติดต่อกับฐานข้อมูล 2 อย่างได้แก่ การส่งความคิดเห็นจากผู้ใช้งานบนเว็บแอปพลิเคชันไปให้ฐานข้อมูล การประมวลผลคะแนนและความรู้สึกที่มีต่อภาพยนตร์และนำมาแสดงผล ฟีเจอร์ต่าง ๆ ในระบบแนะนำจะถูกออกแบบให้มี 3 ฟีเจอร์หลัก ๆ

1) แนะนำภาพยนตร์โดยอ้างอิงจากคะแนน (movie recommendation with score criteria) ภาพยนตร์ที่ถูกแนะนำจะถูกเรียงลำดับตามคะแนนที่ได้โดยจะนำมาเรียงกันและจะแนะนำ 10 เรื่องที่มีคะแนนดีกว่าหนังเรื่องปัจจุบันที่ผู้ใช้งานกำลังรับชม และจะทำให้ผู้ใช้งานสามารถเลือกได้ว่าจะใช้เกณฑ์คะแนนในแง่มุมไหนเป็นตัวที่เรียงลำดับภาพยนตร์ที่แนะนำ
2) แนะนำภาพยนตร์โดยอ้างอิงจากชนิดของภาพยนตร์ (movie recommendation with type criteria) จะทำงานด้วยหลักการเดียวกับข้อแรก แต่จะเพิ่มกฎเกณฑ์ในการคัดสรร
3) การรวบรวมหนังที่ชื่นชอบ (movie favourite collection) ระบบนี้จะเก็บข้อมูลการรับชมภาพยนตร์ได้ว่าผู้ใช้งานได้ดูเรื่องอะไร และเก็บสถิติว่าดูหนังประเภทไหนไปแล้วบ้างเพื่อนำมาปรับปรุงระบบที่ใช้แนะนำภาพยนตร์

3.4 การพัฒนา

3.4.1 การเก็บข้อมูล

ข้อความเกี่ยวกับคำวิจารณ์ภาพยนตร์เป็นข้อความมาจากเว็บไซต์ที่มีความน่าเชื่อถือ เช่น กระทู้ที่พูดคุยเกี่ยวกับภาพยนตร์, ข้อความจากโซเชียลมีเดียที่วิจารณ์ภาพยนตร์ เป็นต้น ในการดึงข้อมูลจากเว็บไซต์จะใช้ไลบรารีสำหรับดึงข้อมูลจากเว็บไซต์ (web scraping) ร่วมกับส่วนต่อประสานโปรแกรมประยุกต์ (application programming interface: API) ของเว็บไซต์นั้น ๆ

โปรแกรม 3.3 การดึงข้อมูลจากแฟนเพจของโซเชียลมีเดีย

```
url = 'https://graph.facebook.com/v2.10/' + user_id \
+ '/posts?access_token=' + access_token
while(True):
  json_res = urllib.request.urlopen(url)
  json_obj = json_res.read().decode('utf-8')
  data = json.loads(json_obj)
  post.extend(data['data'])
  c = c + 1
  try:
      url = data['paging']['next']
  except:
      break
```

ข้อความที่ถูกดึงมาเหล่านั้นถูกเก็บไว้ในรูปแบบไฟล์เจสันและใช้เป็นชุดข้อมูลไว้ใช้ฝึกในโมเดลทั้งสองโมเดลต่อไป

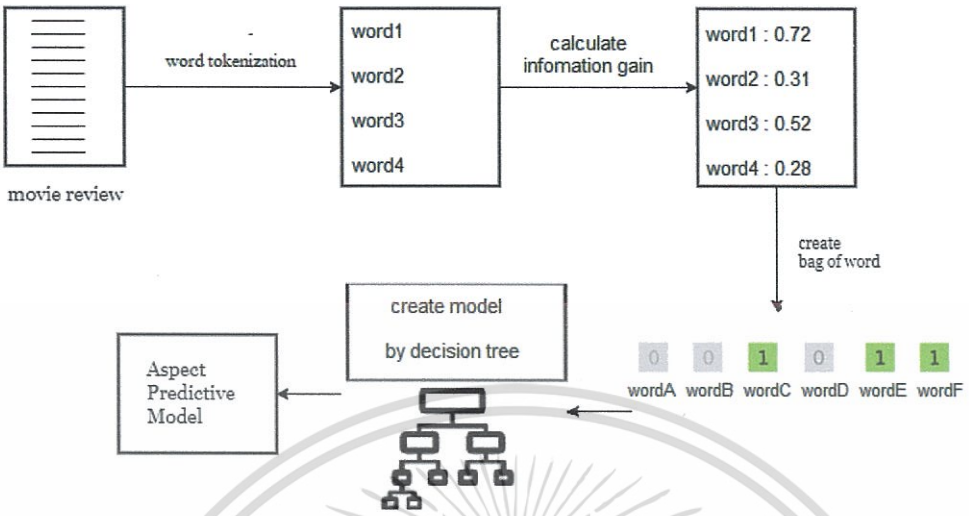
ตัวอย่าง 3.1 การเก็บไฟล์ข้อมูลในรูปแบบ JSON file

```
{
        "movie_name": "Life of Pi",
        "topic_id": 13082688,
        “source”: “www.pantip.com”,
        "text": "Life of Pi คือผลงานเรื่องล่าสุดของผู้กำกับ Ang Lee ที่ดัดแปลงมาจากวรรณกรรม
        สัญชาติแคนาดาชื่อเดียวกันของ Yann Martel
        หนังเล่าเรื่องของ Pi Patel ชายชาวอินเดีย ที่ต้องเล่าถึงเหตุการณ์สำคัญที่สุดในชีวิตของเขา ใน
        ฐานะของผู้รอดชีวิตเพียงคนเดียวจากเหตุการณ์เรืออับปางกลางมหาสมุทรแปซิฟิก ให้นักเขียน
        หนุ่มผู้กำลังมองหาเรื่องราวที่จะนำความศรัทธากลับคืนมา
        เรื่องราวชั้นดีของ Martel เปิดโอกาสให้ผู้รับสาร ตีความ ขบคิด วิพากษ์ วิจารณ์ ได้อย่าง
        หลากหลายตามแต่เครื่องมือของแต่ละคน ไม่ว่าจะเป็น จิตวิทยา ศาสนา ปรัชญา สังคมวิทยา
        หรือ จริยธรรม ซึ่ง Ang Lee ก็ได้บอกเป็นนัย ผ่านตัวละคร Pi ที่บอกกับนักเขียนหนุ่มว่า
        “เมื่อคุณรับรู้เรื่องราวเหล่านี้ของผมแล้ว มันก็จะกลายเป็นเรื่องของคุณ”
        "
}
```

3.4.1.1 การกำหนดคำตอบของแง่มุมและความรู้สึกของข้อความ

ในการศึกษาครั้งนี้จะเก็บข้อมูลจากเว็บไซต์ที่มีเนื้อหาเกี่ยวกับการวิจารณ์ภาพยนตร์ โดยทำการสุ่มเลือกข้อความออกมาเป็นลักษณะของประโยคซึ่งมีทั้งประโยคความเดียว ประโยคความซ้อน และประโยคความรวมเก็บเป็นชุดข้อมูล (dataset) และให้อาสาสมัครจำนวน 10 คน ซึ่งทั้งหมดใช้ภาษาไทยเป็นภาษาแม่ (native language) และเป็นภาษาหลักที่ใช้ในการสื่อสารในชีวิตประจำวัน อ่านประโยคและทำการตอบคำถามว่าประโยคดังกล่าวกล่าวถึงแง่มุมใดของภาพยนตร์และพิจารณาว่าประโยคดังกล่าวแสดงการตำหนิหรือชื่นชม คำตอบที่อาสาสมัครเลือกมากที่สุดอย่างเป็นเอกฉันท์จะถือว่าเป็นคำตอบแง่มุมและความรู้สึกของข้อความนั้น ๆ

3.4.2 การฝึกจำแนกแง่มุมภาพยนตร์ (Aspect trainer)



รูป 3.8 การทำงานของการฝึกจำแนกแง่มุมภาพยนตร์

ในส่วนของการฝึกจำแนกแง่มุมภาพยนตร์ (Aspect trainer) จะเริ่มจากการนำข้อมูลบทวิจารณ์ภาพยนตร์มาใช้ซึ่งจะถูกเก็บมาอยู่ในรูปแบบของเจสัน (JSON) หลังจากนั้นทำการตัดคำจากประโยคในทุก ๆ ไฟล์ของบทวิจารณ์ด้วยไลบรารี ไพไทยเอ็นแอลพี (Pythainlp) และดึงเฉพาะคำภาษาไทยออกมา จากตัวอย่าง 3.1 จะเห็นได้ว่าส่วนบทวิจารณ์จะถูกเก็บที่ “text” คีย์ของทุก ๆ ไฟล์ จึงต้องใช้คำสั่งวนซ้ำเพื่อรวบรวมข้อความบทวิจารณ์จากทุก ๆ ไฟล์

โปรแกรม 3.4 การใช้คำสั่งวนซ้ำเพื่อเก็บข้อมูลบทวิจารณ์จากทุกๆไฟล์

```
dirName = 'document/pantip_data/'

for fileName in os.listdir(dirName):
  data =
json.loads(open(dirName+fileName,encoding="utf8").read())
  documents.append(data["text"])

for document in documents:
  for line in document.split('\n'):
      if(line != ""):
          sentences.append(line)
```

เมื่อได้ข้อความวิจารณ์จากทุกไฟล์ ขั้นตอนต่อไปจะใช้แพ็กเกจไพไทยเอ็นแอลพี (Pythalnlp package) เพื่อทำการแยกคำภาษาไทยและรวบรวมคำทั้งหมดเพื่อคัดเลือกคำที่มีความหมายสื่อ

ถึงองค์ประกอบของภาพยนตร์ หลังจากที่ได้คำทั้งหมดที่ต้องการ บันทึกค่าเหล่านั้นลงในไฟล์พิกเกิล (pickle file) นำไปใช้ต่อในส่วนของโมเดลทำนายแง่มุมภาพยนตร์ (Aspect predictive model)

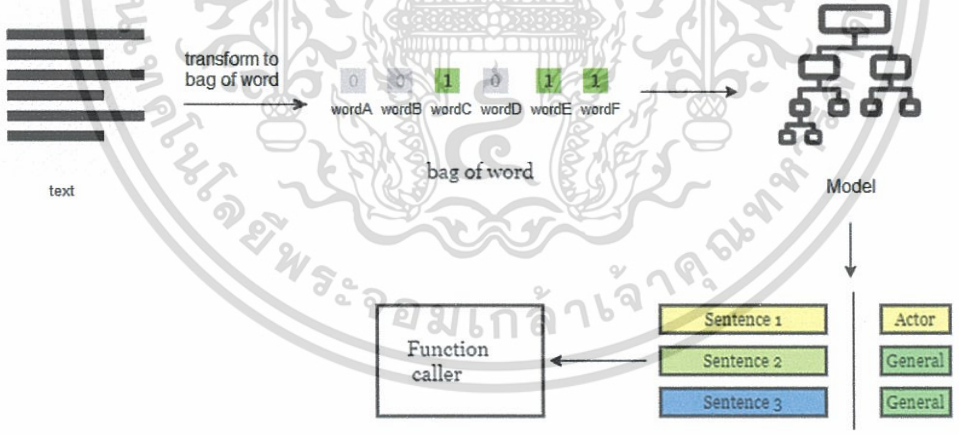
โปรแกรม 3.5 การใช้คำสั่งในแพ็กเกจไพไทยเอ็นแอลพี (Pythainlp package) เพื่อแยกคำภาษาไทย

```
for text in sentences:
  word_in_text = word_tokenize(text,engine="mm")
  for word in word_in_text:
       word_all.append(word)
```

โปรแกรม 3.6 การบันทึกค่าลงในไฟล์พิกเกิล (pickle file)

```
save_documents = open("aspect_model.pickle","wb")
pickle.dump(aspect_classifier, save_documents)
save_documents.close()
```

3.4.3 โมเดลจำแนกแง่มุมภายนตร์ (Aspect predictive model)



รูป 3.9 การทำงานของโมเดลจำแนกแง่มุมภาพยนตร์ (Aspect predictive model)

ในส่วนของโมเดลจำแนกแง่มุมภาพยนตร์ (Aspect predictive model) นั้นจะทำการดึงค่าจากไฟล์พิกเกิลมาใช้หลังจากนั้นจะใช้อัลกอริธึมต้นไม้ตัดสินใจ (Decision tree) ในการสร้างโมเดลทำนายองค์ประกอบของภาพยนตร์ (Aspect) ซึ่งโมเดลนี้จะสามารถรับความคิดเห็นของภาพยนตร์เป็นข้อมูลขาเข้าและจะต้องสามารถคืนค่าออกมาเป็นองค์ประกอบภาพยนตร์อันประกอบไปด้วย นักแสดง, เสียง, บท, ภาพ และภาพรวม โดยในที่นี้จะสามารถตอบได้หลายองค์ประกอบใน 1 ความคิดเห็น

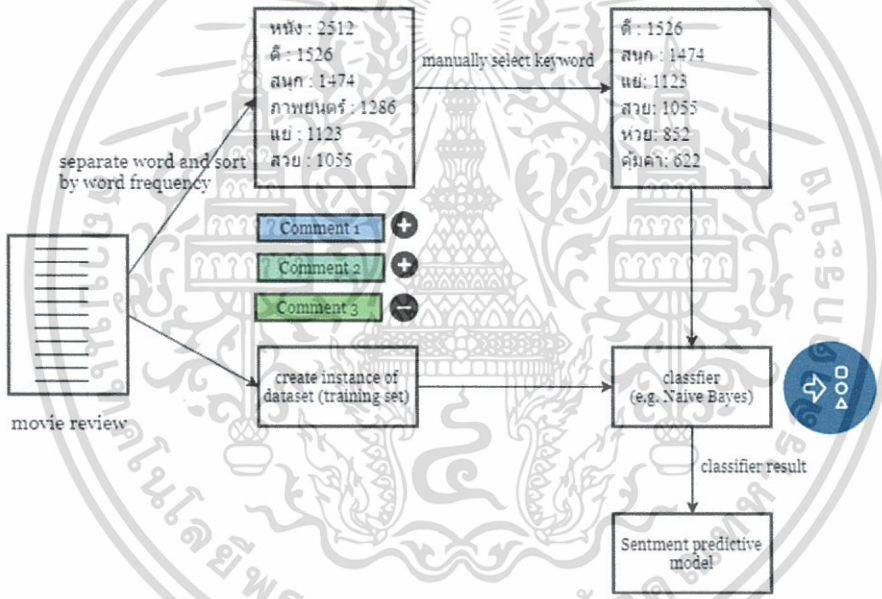
โปรแกรม 3.7 ตัวอย่างการเรียกใช้ Aspect predictive model

```
print(seperator("ภาพและเสียงของภาพยนตร์เรื่องนี้จัดได้ว่าอลังการมากๆ"))
print(seperator("นักแสดงเล่นได้ห่วยที่สุด ไม่เคยเจอใครแย่แบบนี้มาก่อน"))
```

ตัวอย่าง 3.2 ตัวอย่างผลลัพธ์ที่ได้จากการเรียก Aspect predictive model

```
{'general'}
{'actor'}
```

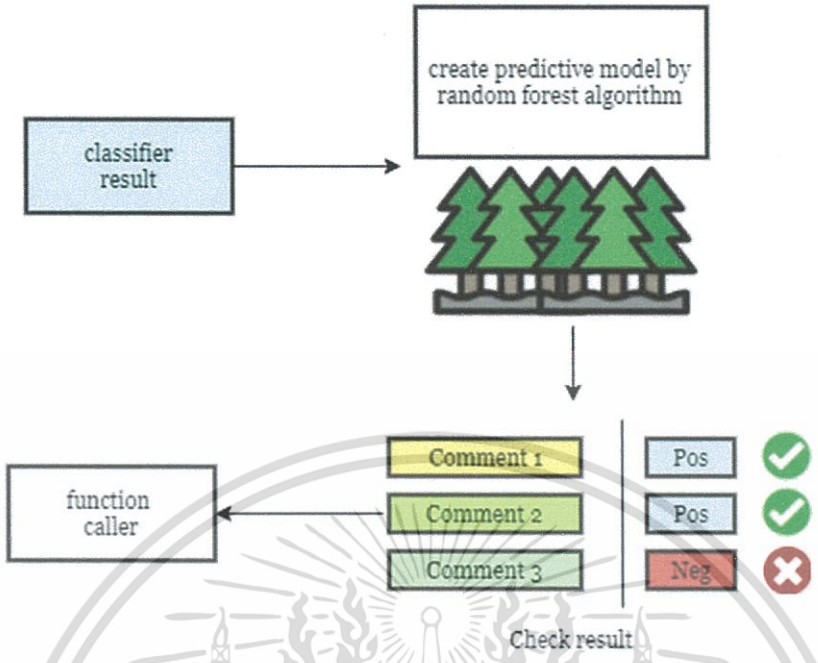
3.4.4 การฝึกจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment trainer)



รูป 3.10 การทำงานของการฝึกจำแนกความรู้สึกที่มีต่อภาพยนตร์ (sentiment trainer)

ในส่วนของการฝึกจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment trainer) จะเริ่มจากการนำข้อมูลบทวิจารณ์ภาพยนตร์มาใช้ซึ่งจะถูกเก็บมาอยู่ในรูปแบบของเจสัน (JSON) หลังจากนั้นทำการตัดคำจากประโยคในทุก ๆ ไฟล์ของบทวิจารณ์ด้วยไลบรารี ไพไทยเอ็นแอลพี (Pythainlp) และดึงเฉพาะคำภาษาไทยออกมา จากนั้นใช้การกำหนดหน้าที่ของคำและเลือกคำสำคัญที่จะมาใช้เป็นฟีเจอร์ซึ่งจะใช้คำที่เป็นคำคุณศัพท์และคำกริยาวิเศษณ์ ซึ่งแบ่งออกเป็นความรู้สึกที่เชิงบวกและความรู้สึกเชิงลบ ในส่วนนี้จะใช้ตัวจำแนก (classifier) ทั้งหมด 7 ตัวได้แก่ Naïve Bays, Multinomial Naïve Bays, Bernoulli Naïve Bays, Logistic regression, Linear support vector classification, Stochastic gradient descent และ Nu-Support vector classification

3.4.5 โมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment Predictive Model)



รูป 3.11 การทำงานของโมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model)

ในส่วนของโมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model) จะสร้างโมเดลทำนายออกมาซึ่งจะใช้ผลจากการฝึกเพื่อให้ได้โมเดลจำแนก (classifier) โดยจะประกอบไปด้วยตัวจำแนก (classfier) หลายตัวจึงทำให้สร้างฟังก์ชันที่เรียกตัวจำแนกทั้งหมดแล้วคำนวณความรู้สึกที่มีต่อภาพยนตร์แล้วรวบรวมผลจาก แต่ละตัวแล้วใช้หลักการในการโหวตกันหาเสียงส่วนใหญ่ว่าผลลัพธ์ที่ได้จากตัวจำแนกควรออกเป็นอย่างไร เพราะฉะนั้นผลลัพธ์ที่ได้จากตัวจำแนกจะมีคำตอบ 2 ประเภท อย่างแรกคือความรู้สึกที่มีต่อภาพยนตร์และอันที่สองคือร้อยละของตัวจำแนกที่เลือกตอบความรู้สึกดังกล่าว

โปรแกรม 3.8 แสดงตัวอย่างการเรียกใช้ฟังก์ชัน Sentiment predictive model

```
print(sentiment('นักแสดงแสดงได้ดีมาก โดยเฉพาะนางเอก'))
```

โปรแกรม 3.9 แสดงตัวอย่างผลลัพธ์ของ Sentiment predictive model

```
('pos', 1.0)
```

3.4.6 ฟังก์ชันคอลเลอร์ (Function caller)

สิ่งที่ฟังก์ชันคอลเลอร์ทำจะมีการเรียกฟังก์ชันจากทั้งสองโมเดลได้แก่ โมเดลจำแนกแง่มุมภาพยนตร์และโมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ เพื่อนำมาหาผลลัพธ์จากฐานข้อมูล มาให้คำตอบในแง่มุมภาพยนตร์และความรู้สึกที่มีต่อภาพยนตร์ หลังจากนั้นจะทำการคำนวณคะแนน ให้แก่ภาพยนตร์แต่ละเรื่องโดยแบ่งเป็นคะแนนออกเป็น 5 ส่วนตามแง่มุมของภาพยนตร์ แต่ละแง่มุมจะมีคะแนนต่ำสุดที่ 0 คะแนนและคะแนนสูงสุดที่ 10 คะแนน โดยใช้สูตรคณิตศาสตร์ 2 รูปแบบ

1) คะแนนในแต่ละแง่มุมใดๆ เมื่อมีความคิดเห็นน้อยกว่า 25 ความคิดเห็น

$$\text{คะแนน} = 5 + 0.2 \times (\text{จำนวนความคิดเห็นเชิงบวก} - \text{จำนวนความคิดเห็นเชิงลบ}) \quad (3.1)$$

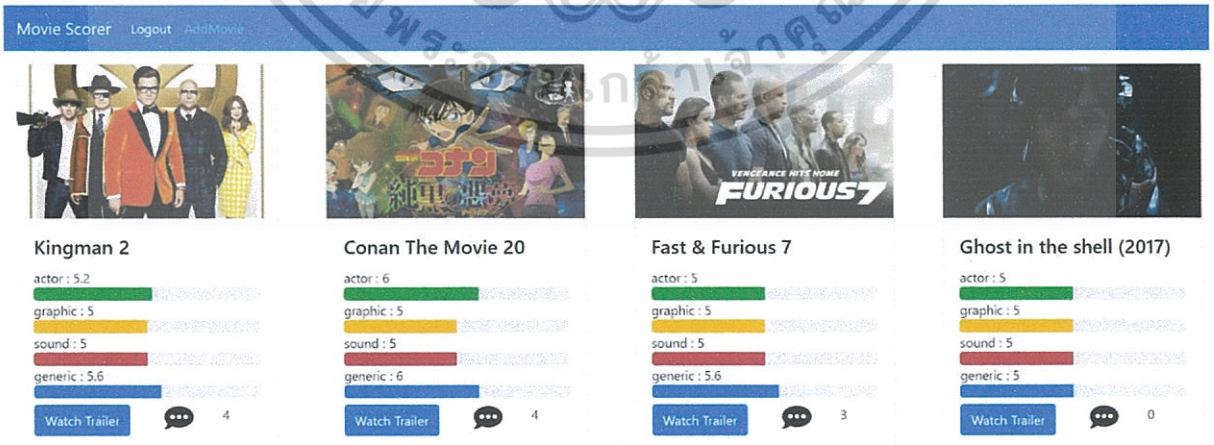
2) คะแนนในแต่ละแง่มุมใดๆ เมื่อมีความคิดเห็นมากกว่า 25 ความคิดเห็น

$$\text{คะแนน} = 5 + 5 \times \frac{\text{จำนวนความคิดเห็นเชิงบวก}}{\text{จำนวนความคิดเห็นทั้งหมด}} - 5 \times \frac{\text{จำนวนความคิดเห็นเชิงลบ}}{\text{จำนวนความคิดเห็นทั้งหมด}} \quad (3.2)$$

ทั้งนี้ความคิดเห็นทั้งเชิงบวกและเชิงลบจะต้องได้คำตอบจากตัวจำแนกในโมเดลจำแนกความรู้สึกที่มีต่อภาพยนตร์ (sentiment predictive model) ให้คำตอบมาในทิศทางเดียวกันเป็นจำนวน 5 ตัวเป็นอย่างน้อย

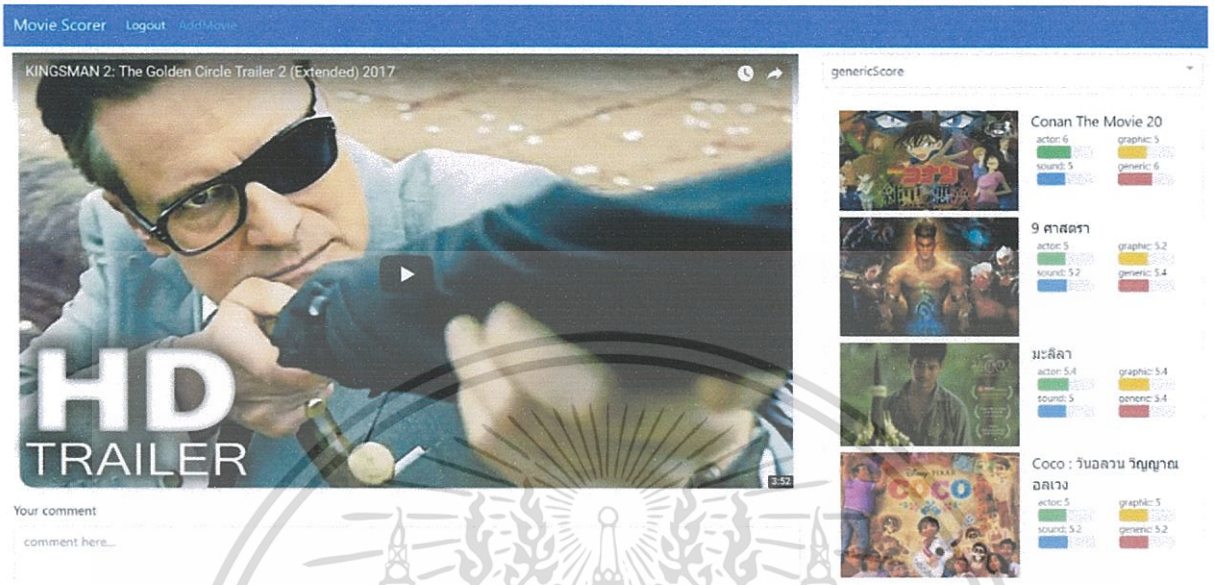
3.4.7 ระบบแนะนำ (recommendation system)

ระบบจะมีหน้าหลัก ๆ ทั้งหมด 2 หน้า หน้าแรกจะทำให้แสดงออกเป็นภาพยนตร์ทั้งหมดที่มีในระบบและแสดงคะแนนในแต่ละแง่มุมของภาพยนตร์เรื่องนั้น และจะแสดงจำนวนความคิดเห็นที่มีต่อภาพยนตร์เรื่องนั้น



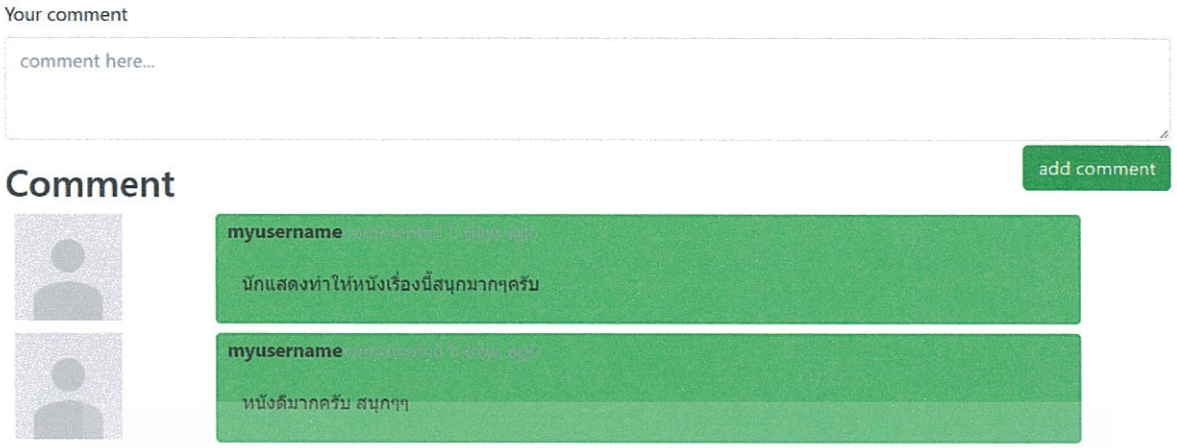
รูป 3.12 หน้าหลักแสดงคะแนนของแต่ละภาพยนตร์

เมื่อผู้ใช้งานกดที่ปุ่ม “watch trailer” จะนำไปสู่หน้าที่ 2 ของระบบซึ่งจะถูกแบ่งออกเป็น 4 ส่วนหลัก ๆ



รูป 3.13 หน้าเว็บเมื่อกดเข้าไปที่ปุ่ม “watch trailer”

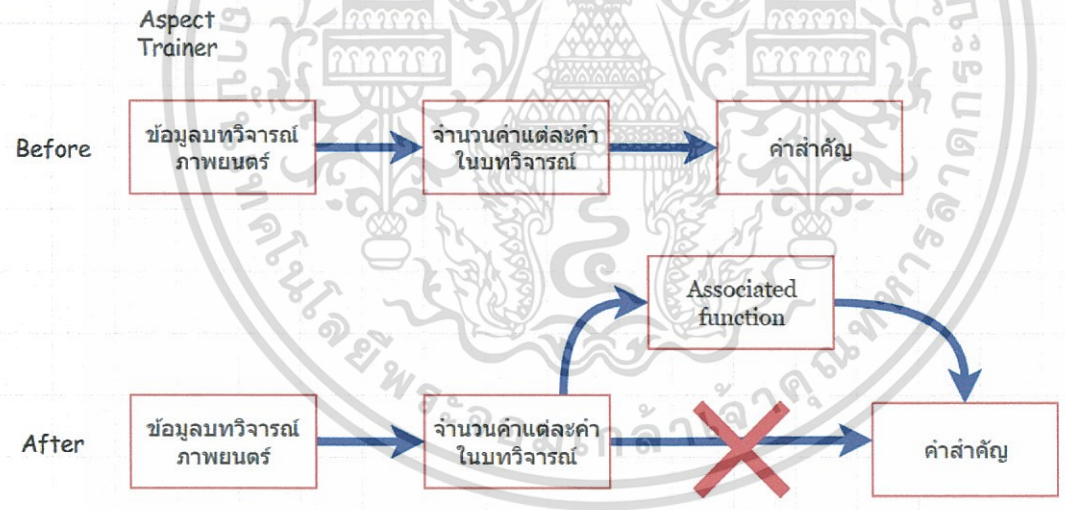
1) ส่วนการแสดงภาพยนตร์
2) ส่วนแนะนำภาพยนตร์ จะเป็นส่วนที่นำภาพยนตร์ที่มีคะแนนใกล้เคียงกันออกมาโดยเรียงตามเกณฑ์ที่เลือกไว้ด้านบน ในภาพจะเรียงจากคะแนนด้านทั่วไปของภาพยนตร์
3) ส่วนเลือกเกณฑ์ จะมีลักษณะเป็น dropdown อยู่ด้านบนเหนือส่วนแนะนำภาพยนตร์ ส่วนนี้จะทำหน้าที่ในการรับการเปลี่ยนแปลงเกณฑ์การเรียงคะแนนจากผู้ใช้งาน
4) ส่วนแสดงความคิดเห็น จะเป็นส่วนที่แสดงความคิดเห็นในภาพยนตร์แต่ละเรื่องจากผู้ใช้งานทุกคน โดยที่กรอบสีของความคิดเห็นจะขึ้นอยู่กับว่าเป็นความคิดเห็นเชิงบวกหรือเชิงลบในภาพยนตร์ถ้าเป็นความคิดเห็นเชิงบวกจะแสดงกรอบสีเขียว และความคิดเห็นเชิงลบจะแสดงกรอบสีแดง



รูป 3.14 ความคิดเห็นแง่บวกจะแสดงสีเขียว

3.5 การทดลองและปรับเปลี่ยนการพัฒนา

3.5.1 การเพิ่มฟังก์ชันความสัมพันธ์ (Associated function)



รูป 3.15 การทำงานของการฝึกจำแนกแง่มุมภาพยนตร์ร่วมกับฟังก์ชันความสัมพันธ์

ฟังก์ชันความสัมพันธ์ (Associated function) เป็นฟังก์ชันที่จะสามารถจับกลุ่มคำที่มักจะเกิดขึ้นพร้อมกันได้ โดยอาศัยหลักการการหากฎของความสัมพันธ์ (association rules) เช่น คำว่าพระเอกมักจะมากับคำว่าแสดง เมื่อได้คำที่มักเกิดขึ้นพร้อมกันให้บันทึกทั้ง 2 ค่าลงไปในคำสำคัญด้วย ทั้งนี้เพื่อ

เพิ่มคำที่อยู่ในถุงคำ (bag of word) ของส่วนโมเดลจำแนกแง่มุมของภาพยนตร์ (Aspect predictive model) และทำให้สามารถตรวจจับประโยคที่มีความหมายสื่อถึงองค์ประกอบของฉากได้ดียิ่งขึ้น

ตัวอย่าง 3.3 ตัวอย่างผลลัพธ์กลุ่มคำที่มักจะเกิดขึ้นพร้อมกันจากการใช้ associated function

antecedents	consequents	support	confidence	lift
{'หนัง'}	{'คน'}	0.289285	0.067901	0.826623
{'หนัง'}	{'ฉาก'}	0.289285	0.055555	0.987654
{'ชอบ'}	{'หนัง'}	0.045238	0.434210	1.500974
{'หนัง'}	{'ดี'}	0.289285	0.048353	1.342720
{'แนว'}	{'หนัง'}	0.015476	0.711538	2.459639

3.5.2 การเพิ่มตัวตัดประโยค (Sentence cutter)

ตัวตัดประโยค (Sentence cutter) มีหน้าที่ในการตัดบทความขนาดยาวให้มีหน่วยที่เล็กลงเหลืออยู่ในรูปของประโยคแทนโดยการตัดประโยคจะอ้างอิงจากไวยากรณ์ภาษาไทย ทั้งนี้ตัวตัดประโยคยังสามารถนำไปใช้กับการตัดความคิดเห็นที่เป็นข้อมูลนำเข้าในส่วนของ Sentiment predictive model เพื่อทำการแยกประโยคและใส่คำตอบของ Sentiment predictive model ลงในแต่ละประโยคได้

ตัวอย่าง 3.4 ตัวอย่างข้อความวิจารณ์ภาพยนตร์

ด้วยโปรดักชั่นทุ่มทุนสร้างและผลสืบเนื่องจากการแก้งานหลังการถ่ายทำที่เพิ่มขึ้น หลังจากที่ Zack Snyder ผู้กำกับคนก่อนได้ถอนตัวออกไปกลางคัน ตัวหนังบอกเล่าเหตุการณ์หลังจากสิ่งที่เกิดขึ้นใน Batman v Superman: Dawn of Justice (2016) โดยเพิ่มรายละเอียดจากภาพยนตร์ภาคแยกของแต่ละตัวละคร และกล่าวถึงพื้นเพและปมในใจของสามตัวละครที่ยังไม่มีภาพยนตร์ภาคหลักเป็นของตัวเอง ตัวหนังยังคงงานภาพที่สวยสดและฉากแอคชั่นสุดมันส์เช่นเคย แต่บทหนังกลับดูง่ายและธรรมดาไปมาก ประเด็นทางการเมืองและสัญญะความศรัทธาของมนุษย์แทบจางหายไปหมดสิ้นหลังจากเปลี่ยนผู้กำกับคนใหม่เป็น Joss Whedon แต่กระนั้นแล้วการนำเสนอคาแรคเตอร์และการเกลี่ยบทให้กับแต่ละตัวละครทำออกมาได้ค่อนข้างดีทีเดียว

ตัวอย่าง 3.5 ตัวอย่างผลลัพธ์ของการตัดประโยคจากตัวอย่าง 3.4 ที่ถูกต้อง

- ด้วยโปรดักชั่นทุ่มทุนสร้างและผลสืบเนื่องจากการแก้งานหลังการถ่ายทำที่เพิ่มขึ้น หลังจากที่ Zack Snyder ผู้กำกับคนก่อนได้ถอนตัวออกไปกลางคัน
- ตัวหนังบอกเล่าเหตุการณ์หลังจากสิ่งที่เกิดขึ้นใน Batman v Superman Dawn of Justice 2016 โดยเพิ่มรายละเอียดจากภาพยนตร์ภาคแยกของแต่ละตัวละคร และกล่าวถึงพื้นเพและปมในใจของสามตัวละครที่ยังไม่มีภาพยนตร์ภาคหลักเป็นของตัวเอง
- ตัวหนังยังคงงานภาพที่สวยสดและฉากแอคชั่นสุดมันส์เช่นเคย แต่บทหนังกลับดูง่ายและธรรมดาไปมาก
- ประเด็นทางการเมืองและสัญญะความศรัทธาของมนุษย์แทบจางหายไปหมดสิ้นหลังจากเปลี่ยนผู้กำกับคนใหม่เป็น Joss Whedon
- แต่กระนั้นแล้วการนำเสนอคาแรคเตอร์และการเกลี่ยบทให้กับแต่ละตัวละครทำออกมาได้ค่อนข้างดีทีเดียว

3.5.3 การตรวจสอบคำสำคัญด้วยเอ็นแกรม (N-grams detection)

ในกรณีที่คำสำคัญนั้นเกิดจากการรวมของคำมากกว่า 1 คำ การใช้ไลบรารีตัดคำจะตัดคำออกเป็นส่วนย่อยทำให้ไม่สามารถรวมให้เป็นคำสำคัญในกรณีนี้ได้ ดังนั้นการแก้ปัญหาสามารถแก้ได้ด้วยการการตรวจสอบคำด้วยเอ็นแกรม (N-grams) เพื่อรวมคำไว้เป็นกลุ่มและตรวจสอบว่าเป็นคำสำคัญหรือไม่ ระบบนี้จะใช้เอ็นแกรมสำหรับรวมคำตั้งแต่ 1 คำ (unigram) ไปจนถึง 5 คำ (five-gram) อาทิ หากต้องการตรวจสอบว่าข้อความมีคำว่า "ไม่น่าสนใจ" หรือไม่ ถ้าใช้แค่ไลบรารีตัดคำเพียงอย่างเดียวตัดข้อความว่า "หนังเรื่องนี้ไม่น่าสนใจ" จะได้ยูนิแกรม (unigram) เพียงอย่างเดียวและได้เป็น หนัง-เรื่อง-นี้-ไม่-น่า-สนใจ ซึ่งทำให้ตรวจสอบไม่ได้ว่าข้อความนี้มีคำว่า "ไม่น่าสนใจ" อยู่ในนั้น ถ้าหากใช้ไตรแกรม (trigram, 3-gram) จะได้เป็น หนังเรื่องนี้-เรื่องนี้ไม่-นี้ไม่น่า-ไม่น่าสนใจ ซึ่งจะทำให้ตรวจสอบได้ว่าข้อความนี้มีคำว่า "ไม่น่าสนใจ"

3.5.4 การเลือกคำสำคัญโดยอาศัยการคำนวณค่าเกนความรู้ (information gain)

อีกวิธีในการคัดเลือกคำที่นำมาใช้เป็นคำสำคัญ (keyword) ก็คือการคำนวณเพื่อหาค่า ค่าเกนความรู้ (information gain) ซึ่งเป็นการวัดว่า คำใดบ้างที่นำมาจำแนกข้อมูลแล้วเกิดประสิทธิภาพดี โดยนำข้อความที่ให้อาสาสมัครกำหนดคำตอบของแง่มุมและความรู้สึกจากหรือจากหัวข้อ 3.4.1.1 มาตัดคำและนับว่าแต่ละคำปรากฏในข้อความที่มีแง่มุมหรือความรู้สึกต่าง ๆ มีมากน้อยเพียงใด และทำการคำนวณค่าเกนความรู้เพื่อใช้พิจารณาว่าคำใดควรนำมาใช้เป็นคำสำคัญ

ตาราง 3.1 ตัวอย่างการแสดงจำนวนคลาสของแง่มุม (aspect) ของแต่ละคำ

คำ	นักแสดง	เสียง	บท	ภาพ/ กราฟิก	ภาพรวม	ไม่มี	Information gain
เพลง	0	25	0	0	0	1	0.71
เวียร์	4	0	0	0	0	0	0.72
คน	6	0	11	3	10	38	0.58
หนัง	10	15	43	13	45	71	0.35

บทที่ 4

การทดลองและผลการทดลอง

4.1 การตัดประโยค

เริ่มจากการนำข้อความที่เก็บมาจากเว็บไซต์ต่าง ๆ มาลบข้อความที่เห็นว่าไม่มีนัยสำคัญในการวิเคราะห์หรือข้อความที่อาจทำให้การวิเคราะห์ข้อมูลคาดเคลื่อน เช่น ยูอาร์แอล (universal resource locator, URL), คำวิบัติโดยการลากอักขระตัวท้ายให้ยาว เป็นต้น หลังจากนั้นให้แยกข้อความด้วยช่องว่าง (white space) หรืออักขระขึ้นบรรทัดใหม่ (new line) ให้กลายเป็นลิสต์ของข้อความสั้น ๆ

นำข้อความสั้น ๆ เหล่านั้นมาตัดคำ และหาชนิดของคำ (POS tag) นำชนิดของคำมาพิจารณาว่าเป็นประโยคหรือไม่ด้วยวิธีการแปลง POS tag ให้เป็นชนิดของคำตามที่พจนานุกรมฉบับราชบัณฑิตยสถานกำหนดดังต่อไปนี้

1) คำนาม ได้แก่ NPRP, NCMN, NONM, CNIT, CLTV เป็นต้น
2) คำสรรพนาม ได้แก่ PPRS, PDMN, PDEF, PIND, PNTR, PREL เป็นต้น
3) คำกริยา ได้แก่ VACT, VSTA, XVBM, XVAM, XVMM, ADVN เป็นต้น
4) คำวิเศษณ์ ได้แก่ VATT, DDAN, DDAC, DDBQ, DDAQ, DIAC เป็นต้น
5) คำสันธาน ได้แก่ JCRG, JCMP, JSBR เป็นต้น
6) คำบุพบท ได้แก่ RPRE
7) คำอุทาน ได้แก่ INT

ประโยคจะประกอบไปด้วยภาคประธานและภาคแสดง ภาคประธานจะประกอบด้วยคำหลักคือคำนามหรือคำสรรพนาม ส่วนภาคแสดงจะประกอบไปด้วยคำหลักคือคำกริยา ดังนั้นข้อความที่เป็นประโยคได้จะต้องมีคำนามหรือคำสรรพนาม และคำกริยา หากข้อความใดไม่ใช่ประโยคจะนำไปต่อกับข้อความต่อไปและทำการตรวจสอบจนกว่าจะเป็นประโยค

ตัวอย่าง 4.1 ข้อความที่ต้องการตัดประโยค

จากพล็อตเรื่องก็จะเห็นว่าทางสตูดิโอโพนอคยังคงความเป็นจิบลิไว้เหมือนเดิม ทั้งเรื่องของบท งานภาพ ลายเส้นที่เราคุ้นเคย ตอนจบของเรื่องก็ยังคงสวยงามแฮปปี้ตามสูตรสำเร็จเหมือนเดิม เพียงแค่ต้องไปดูเรื่องราวระหว่างทางที่กว่าจะไปถึงจุดจบนั้นเป็นอย่างไร ซึ่งก็ไม่ทำให้ผิดหวัง เพราะเรื่องราวของโลกมนุษย์กับโลกเวทย์มนตร์นั้นก็ทำให้เราเข้าถึงง่าย แถมยังมีความล้ำสมัยกว่าหนังเวทย์มนตร์เรื่องอื่นๆ

นอกจากจะได้เสพความสดใสจากงานภาพงามๆ แล้ว เชื่อว่าทาสแมวทั้งหลายต้องตกหลุมรักเจ้าแมวนามว่าทิปแน่นอน ส่วนจะตกหลุมรักเพราะอะไรต้องไปดูเอง และที่สำคัญที่เห็นได้ชัดจากแอนิเมชั่นเรื่องนี้ก็คือ ยังคงแนวคิดเกี่ยวกับธรรมชาติไว้ไม่เปลี่ยนแปลง เพราะสิ่งมีชีวิตทุกชนิดล้วนก่อเกิดมาจากธรรมชาติ จะดัดแปลง ปรับเปลี่ยนอย่างไรก็หนีความจริงที่เกิดจากธรรมชาติไม่ได้

ตัวอย่าง 4.2 ผลลัพธ์ของการตัดประโยคจากตัวอย่าง 4.1

- จากพล็อตเรื่องก็จะเห็นว่าทางสตูดิโอโพนอคยังคงความเป็นจิบลิไว้เหมือนเดิม ทั้งเรื่องของบทงานภาพ
- ลายเส้นที่เราคุ้นเคย
- ตอนจบของเรื่องก็ยังคงสวยงามแฮปปี้ตามสูตรสำเร็จเหมือนเดิม
- เพียงแค่ต้องไปดูเรื่องราวระหว่างทางที่กว่าจะไปถึงจุดจบนั้นเป็นอย่างไร
- ซึ่งก็ไม่ทำให้ผิดหวัง
- เพราะเรื่องราวของโลกมนุษย์กับโลกเวทย์มนตร์นั้นก็ทำให้เราเข้าถึงง่าย
- แถมยังมีความล้ำสมัยกว่าหนังเวทย์มนตร์เรื่องอื่นๆ
- นอกจากจะได้เสพความสดใสจากงานภาพงามๆ แล้ว
- เชื่อว่าทาสแมวทั้งหลายต้องตกหลุมรักเจ้าแมวนามว่าทิปแน่นอน
- ส่วนจะตกหลุมรักเพราะอะไรต้องไปดูเอง
- และที่สำคัญที่เห็นได้ชัดจากแอนิเมชั่นเรื่องนี้ก็คือ
- ยังคงแนวคิดเกี่ยวกับธรรมชาติไว้ไม่เปลี่ยนแปลง เพราะสิ่งมีชีวิตทุกชนิดล้วนก่อเกิดมาจากธรรมชาติ จะดัดแปลง
- ปรับเปลี่ยนอย่างไรก็หนีความจริงที่เกิดจากธรรมชาติไม่ได้

จากตัวอย่าง 4.2 จะเห็นได้ว่าการตัดประโยคด้วยวิธีนี้ยังมีข้อความบางข้อความยังไม่เป็นประโยค เป็นเพียงแค่ภาคแสดงเท่านั้น เพราะภาคแสดงที่มีกรรมซึ่งเป็นคำนามอยู่ด้วย มีเพียงแค่ประโยค

ความเดียวที่ตัดได้ถูกต้อง ส่วนประโยคที่มีใจความ 2 ใจความขึ้นไปได้แก่ ประโยคความซ้อน และประโยคความรวมยังทำได้ไม่ดีนัก

4.2 การทดสอบโมเดลวิเคราะห์บทวิจารณ์ภาพยนตร์

ในหัวข้อนี้จะกล่าวถึงการแบ่งข้อมูลสำหรับฝึก (training set) และข้อมูลสำหรับทดสอบ (test set) เพื่อนำส่วนของข้อมูลสำหรับฝึกไปใช้ฝึก และการทดสอบความแม่นยำของโมเดล ได้แก่ โมเดลที่ใช้ทำนายแง่มุมของภาพยนตร์ (Aspect predictive model) และโมเดลที่ใช้ทำนายความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model) รวมถึงการทดสอบโดยรวมของทั้งสองโมเดลที่ผ่านการเรียกด้วยฟังก์ชันคอลเลอร์ (Function caller)

4.2.1 การแบ่งข้อมูล

จากการที่ให้อาสาสมัครทำการกำหนดคำตอบของแง่มุมภาพยนตร์และความรู้สึกที่มีต่อภาพยนตร์ จากข้อความทั้งหมด 500 อินสแตนซ์ (instances) ผู้จัดทำได้ทำการคัดข้อความที่อาสาสมัครเห็นว่าเป็นข้อความที่ไม่ได้พูดถึงแง่มุมหรือความรู้สึกแบบใดเลย รวมถึงข้อความที่อาสาสมัครให้เห็นด้วยกับคำตอบที่มากที่สุดแต่น้อยกว่า 7 ใน 10 ออกไป เหลือประมาณ 400 อินสแตนซ์ โดยแบ่ง 300 อินสแตนซ์ ไว้สำหรับใช้เป็นชุดข้อมูลสำหรับการฝึก (training set) และอีก 100 อินสแตนซ์เป็นชุดข้อมูลที่ใช้ทดสอบ (test set) นำชุดข้อมูลสำหรับฝึกไปโมเดลในส่วนของสำหรับสร้างโมเดลจำแนกแง่มุมและโมเดลจำแนกความรู้สึกเพื่อให้ได้โมเดลจำแนกออกมา จากนั้นความแม่นยำของโมเดลทั้ง 2 โมเดล ได้แก่ โมเดลที่ใช้ทำนายแง่มุมของภาพยนตร์ (Aspect predictive model) และโมเดลที่ใช้ทำนายความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model) ด้วยชุดข้อมูลสำหรับทดสอบ

ตาราง 4.1 แสดงจำนวนคำตอบของแง่มุมภาพยนตร์ (aspect) และคำตอบของความรู้สึกที่มีต่อภาพยนตร์ (sentiment) จากของชุดข้อมูลสำหรับฝึก (training set)

ความรู้สึก / แง่มุมของภาพยนตร์	แง่บวก	แง่ลบ	รวม
นักแสดง	41	6	47
เสียง	21	4	25
ภาพ/กราฟิก	39	10	49
บท/เนื้อหา	34	19	53
ภาพรวม	106	22	128
รวม	239	61	300

ตาราง 4.2 แสดงจำนวนคำตอบของแง่มุมภาพยนตร์ (aspect) และคำตอบของความรู้สึกที่มีต่อภาพยนตร์ (sentiment) จากของชุดข้อมูลสำหรับทดสอบ (test set)

ความรู้สึก / แง่มุมของภาพยนตร์	แง่บวก	แง่ลบ	รวม
นักแสดง	11	5	16
เสียงประกอบ	6	1	7
ภาพ/กราฟิก	11	2	13
บท/เนื้อหา	18	8	26
ภาพรวม	22	16	38
รวม	68	32	100

4.2.2 การทดสอบโมเดลจำแนกแง่มุมของภาพยนตร์

การทดสอบความแม่นยำของการแยกประเภทของแง่มุมภาพยนตร์ (Aspect predictive model) สามารสรุปการทดสอบได้อัตราความแม่นยำเป็นร้อยละ 77 เมื่อนำมาแสดงเป็นคอนฟิวชันเมทริกซ์ตามตารางที่ 4.3 โดยให้ในแถวเป็นคลาสที่เป็นคำตอบจริง (actual) แนวคอลัมน์เป็นคลาสที่ทำนายมาได้ (predicted)

ตาราง 4.3 คอนฟิวชันเมทริกซ์แสดงผลการทดสอบเมื่อจำแนกแง่มุมของภาพยนตร์ด้วยโมเดลจำแนกแง่มุมของภาพยนตร์

predicted / actual	นักแสดง	เสียงประกอบ	ภาพ/กราฟิก	บท/เนื้อหา	ภาพรวม
นักแสดง	12	1	0	0	3
เสียงประกอบ	0	6	0	0	1
ภาพ/กราฟิก	0	0	11	0	2
บท/เนื้อหา	2	0	2	13	9
ภาพรวม	0	0	0	3	35

จากคอนฟิวชันเมทริกซ์ที่แสดงในตารางที่ 4.3 สามารถอธิบายประสิทธิภาพในการทำนายของแต่ละคลาสได้ อาทิเช่น จำนวนข้อมูลที่ทดสอบมีแง่มุมนักแสดงทั้งหมด 16 อินสแตนซ์ โมเดลทำนายออกมาเป็นแง่มุมนักแสดง เสียงประกอบ ภาพ/กราฟิก บท/เนื้อหาและภาพรวมเป็น 12, 1,

0, 0 และ 3 อินสแตนซ์ ตามลำดับ จากตารางที่ 4.3 พบว่า การจำแนกแง่มุมที่เป็นเรื่องเกี่ยวกับภาพรวมทำนายถูกมากที่สุดแต่ก็ทำนายผิดมากที่สุดเช่นกัน ส่วนแง่มุมเกี่ยวกับเสียงทำนายผิดน้อยที่สุดคือทำนายออกมาเป็นแง่มุมเกี่ยวกับเสียง 7 อินสแตนซ์ ซึ่งถือว่าทำนายถูกและทำนายผิดเป็นแง่มุมภาพรวมเพียงแค่ 1 อินสแตนซ์เท่านั้น

4.2.3 การทดสอบโมเดลจำแนกประเภทของความรู้สึกที่มีต่อภาพยนตร์

การทดสอบความแม่นยำของการแยกประเภทของความรู้สึกที่มีต่อภาพยนตร์ (Sentiment predictive model) สามารสรุปการทดสอบการทดสอบได้อัตราความแม่นยำเป็นร้อยละ 77 เมื่อนำมาแสดงเป็นคอนฟิวชันเมทริกซ์ตามตารางที่ 4.4 โดยให้ในแถวเป็นคลาสที่เป็นคำตอบจริง (actual) แนวคอลัมน์เป็นคลาสที่ทำนายมาได้ (predicted)

ตาราง 4.4 คอนฟิวชันเมทริกซ์แสดงผลการทดสอบเมื่อแยกประเภทของความรู้สึกด้วยโมเดลจำแนกประเภทของความรู้สึกที่มีต่อภาพยนตร์

predicted / actual	แง่บวก	แง่ลบ
แง่บวก	68	0
แง่ลบ	23	9

สำหรับคอนฟิวชันเมทริกซ์ที่ปรากฏในตารางที่ 4.4 พบว่าในการทำนายคลาสที่เป็นความรู้สึกแง่บวกทำได้ดี แต่ในแง่ลบนั้นมีความแม่นยำที่ค่อนข้างต่ำ ซึ่งสืบเนื่องจากจำนวนชื่อใช้ฝึกโมเดลมีจำนวนคลาสที่เป็นแง่ลบมีจำนวนน้อยส่งผลให้การทำนายข้อความที่เป็นแง่ลบมีประสิทธิภาพที่ไม่ดี

4.2.4 ผลการทดสอบด้วยฟังก์ชันคอลเลอร์

การทดสอบความแม่นยำของการแยกประเภทของแง่มุมมารสรุปการทดสอบการทดสอบได้อัตราความแม่นยำเป็นร้อยละ 61 เมื่อนำมาแสดงเป็นคอนฟิวชันเมทริกซ์ตามตารางที่ 4.5 โดยให้ในแถวเป็นคลาสที่เป็นคำตอบจริง (actual) แนวคอลัมน์เป็นคลาสที่ทำนายมาได้ (predicted)

ตาราง 4.5 คอนฟิวชันเมทริกซ์แสดงผลการทดสอบด้วยฟังก์ชันคอลเลอร์

predicted / actual	นักแสดง - แง่บวก	นักแสดง - แง่ลบ	เสียง - แง่บวก	เสียง - แง่ลบ	บท-แง่บวก	บท-แง่ลบ	ภาพ-แง่บวก	ภาพ-แง่ลบ	ภาพรวม-แง่บวก	ภาพรวม-แง่ลบ
นักแสดง แง่บวก	7	0	0	0	1	0	0	0	3	0
นักแสดง แง่ลบ	4	1	0	0	0	0	0	0	0	0
เสียง แง่บวก	0	0	5	0	0	0	0	0	1	0
เสียง แง่บวก	0	0	0	1	0	0	0	0	0	0
บท แง่บวก	0	0	0	0	10	0	0	0	1	0
บท แง่ลบ	0	0	0	0	1	0	0	0	0	1
ภาพ แง่บวก	1	0	0	0	1	0	11	0	5	0
ภาพ แง่ลบ	1	0	0	0	0	0	1	1	3	1
ภาพรวม แง่บวก	0	0	0	0	0	0	1	0	21	0
ภาพรวม แง่ลบ	0	0	0	0	0	0	2	0	10	4

จากผลการทดสอบความแม่นยำจากคอนฟิวชันเมทริกซ์ในตารางที่ 4.5 พบว่าการใช้ฟังก์ชันคอลเลอร์เพื่อทำนายแง่มุมและความรู้สึกพร้อมกับในบางคลาสทำนายไม่ถูกต้องได้แก่ ข้อความที่กล่าวถึงบทหรือเนื้อหาในแง่ลบ สาเหตุมาจากการทดสอบการทำนายความรู้สึกในแง่ลบนั้นมีประสิทธิภาพที่ไม่ดี ซึ่งดูได้จากคอนฟิวชันเมทริกซ์ในตารางที่ 4.4

บทที่ 5

บทสรุป

5.1 อุปสรรคและปัญหา

1) ข้อมูลเชิงลบ (negative sentiment) ที่นำมาใช้ฝึกโมเดลในส่วนของการฝึกจำแนกความรู้สึกที่มีต่อภาพยนตร์มีน้อยกว่าข้อมูลเชิงบวก (positive sentiment) ทำให้มีผลต่อการตรวจสอบความแม่นยำของระบบ
2) การตัดประโยคยังทำได้ไม่ดีนักในกรณีที่เป็นประโยคความซ้อนและประโยคความรวม
3) ข้อความบางข้อความเป็นเพียงแค่วลีก็สามารถบอกแง่มุมและความรู้สึกได้เช่น “ภาพสวย” แต่เมื่อต้องใช้การตัดประโยคทำให้ต้องนำไปรวมกับข้อความอื่นเพื่อให้กลายเป็นประโยคอาจทำให้ประสิทธิภาพในการให้คะแนนลดน้อยลง
4) ข้อความที่มีเนื้อหาประชด เช่น “หนังเรื่องนี้ผู้กำกับยังคงรักษามาตรฐานไว้ดังเดิม” สามารถตีความได้ทั้งแง่บวกและแง่ลบ ซึ่งอาจทำให้การจำแนกประเภทเกิดความผิดพลาดได้

5.2 แนวทางการพัฒนา

1) เพิ่มคำสำคัญเพื่อให้จำแนกแง่มุมและความรู้สึกได้ประสิทธิภาพดียิ่งขึ้น
2) ปรับเปลี่ยนอัลกอริทึมใหม่ในการตัดข้อความ ควรสามารถแยกวลีได้และสามารถแยกประโยคได้โดยไม่ต้องอาศัยแบ่งด้วยช่องว่าง (whitespace) หรือการขึ้นบรรทัดใหม่ (newline)
3) ปรับเปลี่ยนโมเดลให้แยกข้อความที่ไม่ได้กล่าวถึงแง่มุมหรือความรู้สึกเลยได้ดียิ่งขึ้น
4) ปรับน้ำหนักการให้คะแนนของผู้ชมไม่เท่ากัน เช่น ผู้ที่เป็นนักวิจารณ์มืออาชีพอาจให้คะแนนที่มีน้ำหนักมากกว่า
5) พัฒนาส่วนต่อประสานกับผู้ใช้ (User Interface) ให้น่าใช้มากยิ่งขึ้น

บรรณานุกรม

ชูชาติ หฤไชยะศักดิ์. 2556. **การประมวลผลภาษาธรรมชาติ เทคนิคการสืบค้นสารสนเทศและทำเหมืองข้อความ.** [Online]. Available: http://www2.it.kmutnb.ac.th/teacher/FileDL/178255420125.pdf

ราชบัณฑิตยสถาน. 2554. **คำชี้แจงหลักการจัดทำและวิธีใช้พจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. 2554.** กรุงเทพฯ : สำนักงานราชบัณฑิตยสภา.

วรรณพงษ์ ภัททิยไพบูลย์. 2560. **คู่มือการใช้งาน PyThaiNLP 1.5.** [Online] Available : https://github.com/PyThaiNLP/pythainlp.

วิจินตน์ ภาณุพงศ์. 2520. **โครงสร้างภาษาไทย : ระบบไวยากรณ์.** กรุงเทพฯ : มหาวิทยาลัยรามคำแหง.

Bird, S. Klein, E. and Loper, E. 2009. **Natural Language Processing with Python.** California : O'Reilly Media, Inc.

Google Inc. 2018. **Documentation | Firebase.** [Online] Available : https://firebase.google.com/docs/reference.

Facebook Inc. 2018. **React – Docs.** [Online] Available : https://reactjs.org/docs.

Li, Z. Feng, J. and Xiao-Yan, Z. 2009. **Movie Review Mining and Summarization.** Beijing : Tsinghua University.

Pannala, N. Nawarathna, C. and Jayakody, J. 2016. **Supervised Learning Based Approach to Aspect Base Sentiment Analysis.** Malabe : Sri Lanka Institute of Information Technology.

Python Software Foundation. 2017. **Python 3.5 documentation.** [Online] Available : https://docs.python.org/3.5.

Roger, S. and Girolami, M. 2012. **A First Course in Machine Learning.** Florida : CRC Press.

Sornlertlamvanich, V. and Phantachat, W. 2015. **Thai Part-of-Speech Tagged Corpus.** Bangkok : National Electronics and Computer Technology Center.