การสกัดคำสำคัญที่เป็นกระแสและคำหยุดจากเพจเฟซบุ๊กภาษาไทยโดยใช้เอ็นแกรมแบบตัวอักษร

นายณัษฐพงษ์ อู่สิริมณีชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Extraction of Trend Keywords and Stop Words from Thai Facebook Pages using Character n-Grams

Mr. Nattapong Ousirimaneechai

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

หัวข้อวิทยานิพนธ์	การสกัดคำสำคัญที่เป็นกระแสและคำหยุดจากเพจเฟซบุ๊กภาษาไทยโดยใช้เอ็นแกรมแบบตัวอักษร
โดย	นายณัษฐพงษ์ อู่สิริมณีชัย
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ)

ณัษฐพงษ์ อู่สิริมณีชัย : การสกัดคำสำคัญที่เป็นกระแสและคำหยุดจากเพจเฟซบุ๊กภาษาไทยโดยใช้เอ็นแกรมแบบตัวอักษร. (
Extraction of Trend Keywords and Stop Words from Thai Facebook Pages using Character n-Grams) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

สื่อสังคมออนไลน์สามารถใช้วิเคราะห์พฤติกรรมของผู้คนในสังคมได้ โดยสื่อสังคมออนไลน์ที่คนไทยนิยมมากที่สุดคือเฟซบุ๊ก ดังนั้นถ้าเราสามารถวิเคราะห์พฤติกรรมของผู้คนในเฟซบุ๊กได้ก็จะสามารถเข้าใจพฤติกรรมของคนไทยส่วนใหญ่ในสังคมได้ ซึ่งหนึ่งในการวิเคราะห์พฤติกรรมของผู้คนนั้น เรามักจะวิเคราะห์ผ่านกระแสที่เกิดขึ้นในสังคม ว่าผู้คนในสังคมให้ความสนใจในกระแสนั้นอย่างไร จุดเริ่มต้นของกระแสคือเมื่อไหร่ เป็นต้น ซึ่งการวิเคราะห์กระแสนั้นสามารถทำได้ผ่านการวิเคราะห์คำสำคัญที่เกี่ยวข้องกับกระแสดังกล่าว แต่วิธีการที่ใช้ในการสกัดคำสำคัญในปัจจุบันนั้นจำต้องใช้เครื่องมือตัดคำภาษาไทย ซึ่งเครื่องมือในปัจจุบันถูกฝึกสอนด้วยคลังข้อมูลภาษาที่ไม่ได้รวมเอาข้อมูลประโยคที่พบในสื่อสังคมออนไลน์อย่างเฟซบุ๊กไว้ ผลจึงทำให้เครื่องมือตัดคำมีปัญหาเมื่อพบคำที่ไม่เป็นมาตรฐาน ส่งผลต่อประสิทธิภาพของการสกัดคำสำคัญ อีกทั้งวิธีสกัดคำสำคัญในปัจจุบันรองรับการสกัดคำสำคัญที่ความยาวคงที่เท่านั้น ทำให้วิทยานิพนธ์ฉบับนี้ได้พัฒนาวิธีการสกัดคำสำคัญที่เป็นกระแสโดยไม่ใช้เครื่องตัดคำ แต่เลือกใช้อัลกอริทึมเอ็นแกรมแบบตัวอักษรเข้ามาช่วย ซึ่งทำให้สามารถสกัดคำสำคัญที่มีความยาวแบบไม่คงที่ได้ และยังใช้ลักษณะของกระแสในการสร้างฐานข้อมูลคำหยุด และกรองเฉพาะคำที่เป็นกระแสออกมา โดยเมื่อเปรียบเทียบผลกับวิธีดั้งเดิมอย่างวิธี TF-IDF และวิธี TF พบว่าวิธีที่วิทยานิพนธ์นี้นำเสนอ ได้คะแนน F1 ที่ 0.402 ซึ่งดีกว่าวิธี TF-IDF ที่ได้คะแนน F1 ที่ 0.165 และวิธี TF ที่ได้คะแนน F1 ที่ 0.183 โดยวิธีที่วิทยานิพนธ์นี้นำเสนอเหมาะเป็นอย่างยิ่งสำหรับงานที่ต้องการคำสำคัญที่มีความยาวไม่คงที่ อย่างเช่นการหากระแสในสื่อสังคมออนไลน์เฟซบุ๊ก

สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2561	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6070188521 : MAJOR COMPUTER ENGINEERING

KEYWORD: Keyword Extraction, Word Segmentation, Character n-Grams

Nattapong Ousirimaneechai :

Extraction of Trend Keywords and Stop Words from Thai Facebook Pages using Character n-Grams. Advisor: Asst. Prof. Sukree Sinthupinyo, Ph.D.

Social media can be used to analyze the behavior of people in society, and we often analyze it through the trends in society. The trend analysis can be done through the analysis of keywords related to the trends. But the method used to extract the trend keywords requires Thai word segmentation tools, which are trained with a Thai corpus that does not include sentence information found on social media. As a result, the word segmentation tool has problems when segmenting non-standard words, and thus affecting the efficiency of keyword extraction. In addition, the keyword extraction method supports only the fixed length method. This thesis has developed a method for extracting keywords that are trends by using the character n-grams method instead of word segmentation methods. Which makes it possible to extract keywords that are not fixed in length. In addition, we used the trend characteristics to create the stop word database, then filtered only the words that are trends. By comparing the results with the traditional methods such as TF-IDF and TF methods, it was found that the method proposed by this thesis provided F1 score of 0.402 which is better than TF-IDF method with F1 score of 0.165 and TF method with F1 score of 0.183. Finally, the method presented in this thesis is especially suitable for tasks that require non-fixed length keywords, such as finding the trends on social media, Facebook.

Field of Study: Computer Engineering Student's Signature .....

Academic Year: 2018 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดีนั้น นอกจากตัวข้าพเจ้าแล้ว ยังมีบุคคลท่านอื่นที่คอยช่วยเหลือ ให้คำแนะนำ และเป็นกำลังใจในการทำวิทยานิพนธ์ฉบับนี้ขึ้นมา

ข้าพเจ้าขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ของข้าพเจ้า ที่คอยให้คำแนะนำในการทำวิจัยเป็นอย่างดี จนทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วง

ข้าพเจ้าขอขอบคุณ คณะกรรมการการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.นัททนิภานันท์ ผู้ช่วยศาสตราจารย์ ดร. ณัฐพงศ์ ชินธเนศ และผู้ช่วยศาสตราจารย์ ดร. เด่นดวง ประดับสุวรรณ ที่ให้คำแนะนำ และเสนอสิ่งที่ควรเพิ่มเติมลงในวิทยานิพนธ์ฉบับนี้จบเสร็จสมบูรณ์

ข้าพเจ้าขอขอบคุณ คุณพ่อ คุณแม่ และครอบครัว รวมถึงเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือ ให้คำแนะนำ และเป็นกำลังใจให้ข้าพเจ้าเสมอมา

ข้าพเจ้าขอขอบคุณ อาจารย์ทุกท่าน ที่ได้สั่งสอนข้าพเจ้ามา ซึ่งความรู้ แนวคิด และกระบวนการต่าง ๆ ที่ได้รับมา ล้วนมีประโยชน์ต่อวิทยานิพนธ์ฉบับนี้ทั้งสิ้น

ข้าพเจ้าขอขอบคุณ ทุนอุดหนุนการศึกษาอัจฉริยะคืนรัง จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ทำให้สนับสนุนให้ข้าพเจ้าได้เรียนโดยไม่มีค่าใช้จ่าย

สุดท้ายนี้ ข้าพเจ้าขอขอบคุณ นางสาวณภัทร ซิ้มสมบูรณ์ผล ที่คอยอยู่เคียงข้างข้าพเจ้า และเป็นกำลังใจให้ข้าพเจ้าเสมอมา จนทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้ได้ลุล่วง และสำเร็จการศึกษาในระดับมหาบัณฑิต จากคณะวิศวกรรมศาสตร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัยได้ในที่สุด

ณัษฐพงษ์ อู่สิริมณีชัย

สารบัญ

หน้า

..... ค

บทคัดย่อภาษาไทย ..... ค

..... ง

บทคัดย่อภาษาอังกฤษ ..... ง

กิตติกรรมประกาศ ..... จ

สารบัญ ..... ฉ

สารบัญตาราง ..... ฏ

สารบัญรูปภาพ ..... ณ

บทนำ 1

ที่มาและความสำคัญ 1

วัตถุประสงค์ 3

ขอบเขตการดำเนินงาน 3

ประโยชน์ที่คาดว่าจะได้รับ 3

วิธีการดำเนินงานวิจัย 4

ผลงานตีพิมพ์จากวิทยานิพนธ์ 4

ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์ 4

ทฤษฎีที่เกี่ยวข้อง 5

คำที่ไม่เป็นมาตรฐานตามพจนานุกรม 5

1. คำย่อ (Abbreviation) 5

- รูปย่อ (Short form) 5

- ตัวย่อ (Acronym) 6

2. คำที่สะกดผิด (Misspelling / Typing error) 6
- คำที่สะกดผิด (Misspelling) 6
- คำที่พิมพ์ผิด (Typing error) 6
3. คำที่ถูกตัดตัวอักษรออก (Punctuation omission/error) 7
4. คำแสลง (Non-dictionary slang) 7
5. การเล่นคำ (Wordplay) 7
6. คำที่หลบหลีกการจับผิด (Censor avoidance) 7
7. สัญรูปอารมณ์ (Emoticons) 8
การตัดคำ (Word Segmentation) 8
1. การตัดคำโดยใช้พจนานุกรม (Dictionary-based: DCB) 8
- การตัดคำแบบเลือกคำที่ยาวที่สุด (Longest Matching) [7] 8
- การตัดคำแบบเลือกคำที่เหมือนมากที่สุด (Maximum Matching) [9] 9
2. การตัดคำโดยใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning-based: MLB) 9
การแทนข้อความ (Text Representation) 10
1. ถุงคำ (Bag of Words: BoW) 10
2. การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (Term Frequency - Inverse Document Frequency: TF-IDF) 10
3. เวกเตอร์วันฮอต (One-hot Vector) 11
4. คำฝังตัว (Word Embedding) 11
การสกัดคำสำคัญ (Keyword Extraction) 12
1. เอ็นแกรม (n-Grams) 12
2. คำหยุด (Stop words) 13
3. การกำหนดน้ำหนักคำ (Term Weighting) 13

- การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (Term Frequency - Inverse Document Frequency, TF-IDF) 14
- การพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (Term Frequency, TF).... 14

การจำแนกข้อมูล (Classification) 15

1. แบบจำลองจากตัวจำแนกเชิงเส้น (Linear Classifiers) [21] 15
2. แบบจำลองจากซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) [22]16
3. แบบจำลองจากเพื่อนบ้านใกล้กัน k ตัว (K-Nearest Neighbor: KNN) [23]..... 17
4. แบบจำลองจากการเร่งความสามารถ (Boosting) [24] 17
5. แบบจำลองจากต้นไม้ตัดสินใจ (Decision Tree) [25] 18
6. แบบจำลองจากป่าสุ่ม (Random Forests) [26] 19
7. แบบจำลองจากโครงข่ายประสาทเทียม (Neural Networks) [28] 19

การแบ่งกลุ่มข้อมูล (Clustering) 20

1. การแบ่งกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering) [31] 21
2. การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering) [32] 21
   - วิธีข้อศอก (Elbow Method) [33] 22
3. การแบ่งกลุ่มจากโครงข่ายประสาทเทียม (Neural Networks) [29] 23
   - การแปลงข้อมูลแบบจัดการตนเอง (Self-Organizing Map: SOM) [34]..... 23
   - ทฤษฎีการสั่นพ้องแบบปรับได้ (Adaptive Resonance Theory: ART) [35] 25

งานวิจัยที่เกี่ยวข้อง 26

การตัดคำภาษาไทย 26

1. เลิร์นเล็กโต (LearnLexTo) เครื่องมือตัดคำภาษาไทยโดยใช้การเรียนรู้ของเครื่อง (Machine Learning-based) สำหรับการทำดรรชนีข้อความภาษาไทย..... 26

2. เล็กโตพลัส (LexToPlus) เครื่องมือตัดหน่วยคำ (Lexeme) ภาษาไทยและทำให้เป็นมาตราฐาน (Normalization) 27
3. การตัดคำที่สนใจผลลัพธ์หลายแบบโดยใช้โครงข่ายประสาทเทียมหน่วยความจำระยะสั้นแบบยาวแบบ 2 ทิศทาง (Bi-directional LSTM Neural Networks) 27
การจำแนกข้อความ 28
1. โครงข่ายประสาทเทียมคอนโวลูชัน (Convolutional Neural Network) สำหรับการจำแนก (Classification) ข้อความ 28
2. โครงข่ายประสาทเทียมคอนโวลูชัน สำหรับการสร้างแบบจำลอง (Modelling) ข้อความ 29
3. โครงข่ายประสาทเทียมคอนโวลูชันแบบลึก (Deep Convolution Neural Network) สำหรับวิเคราะห์อารมณ์ (Sentiment Analysis) ของข้อความขนาดสั้น 29
4. โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษร (Character-level Convolutional Neural Network) สำหรับการจำแนกข้อความ 29
5. โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษรกับความยาวข้อมูลแบบพลวัตสำหรับการจำแนกข้อความภาษาไทย 30
คำหยุด 30
1. คำหยุดในการประเมินค่าความสามารถในการอ่านข้อความภาษาไทย 31
2. การสร้างฐานข้อมูลคำหยุดแบบอัตโนมัติสำหรับงานด้านการค้นคืนสารสนเทศ 31
การสกัดคำสำคัญ 31
1. การสำรวจวิธีการสกัดความรู้ (Knowledge Extraction) ภาษาไทยสำหรับเครื่องมือและงานวิจัยเว็บเชิงหมายความ (Semantic Web) 32
2. การสกัดคำสำคัญภาษาไทยแบบอัตโนมัติจากคลังข้อมูลของข้อความที่ถูกจำแนกแล้ว 33
3. การสกัดคำสำคัญที่แสดงถึงเหตุการณ์สำคัญบนสื่อสังคมออนไลน์ทวิตเตอร์ (Twitter) ภาษาไทย 33
ขั้นตอนการสร้างระบบสกัดคำสำคัญที่เป็นกระแสและคำหยุด จากเพจเฟซบุ๊ก 35

ให้คำจำกัดความของกระแสที่ต้องการบนเฟซบุ๊ก 35
เก็บข้อมูลโพสต์จากเพจเฟซบุ๊ก 37
ทำความสะอาดข้อมูลโพสต์ 51
การทดลองใช้เครื่องมือตัดคำ deepcut และการวิเคราะห์ผลลัพธ์และข้อจำกัดที่เกิดขึ้น 51
แบ่งข้อความออกเป็นชุดของตัวอักษร (แกรม) 53
การวิเคราะห์ปัญหาจากการใช้อัลกอริทึมเอ็นแกรมแบบตัวอักษรแทนที่จะใช้เครื่องมือตัดคำ 54
นับความถี่ของแกรม 56
หาแกรมที่มีคุณสมบัติเป็นคำสำคัญของเพจ 56
หาแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ 57
รวมแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ กลับมาเป็นคำสำคัญ 58
สกัดคำหยุดออกจากคำสำคัญที่เป็นกระแส 59
สกัดคำสำคัญที่เป็นกระแสออกจากคำสำคัญ 60
การทดลอง และอภิปรายผล 61
การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุด 62
การวัดความถูกต้องของคำหยุด 66
การเพิ่มกฎของคำหยุด เพื่อเพิ่มประสิทธิภาพในการสกัดคำสำคัญ 68
การวัดความถูกต้องของคำสำคัญที่เป็นกระแส 68
การเปรียบเทียบผลเมื่อใช้วิธี TF-IDF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร 74
การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุดเมื่อใช้วิธี TF-IDF 74
การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF-IDF 78
การเปรียบเทียบผลเมื่อใช้วิธี TF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร 82
การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF 85

การเปรียบเทียบประสิทธิภาพของผลลัพธ์ของวิธีเอ็นแกรมแบบตัวอักษร กับวิธีที่ใช้เครื่องมือในการตัดคำอย่าง deepcut ได้แก่วิธี TF-IDF และวิธี TF 88
บทสรุปผลการวิจัย และข้อเสนอแนะ 100
สรุปผลการวิจัย 100
ข้อเสนอแนะ 101
ภาคผนวก ก ผลงานตีพิมพ์จากวิทยานิพนธ์ 102
บรรณานุกรม 109
ประวัติผู้เขียน 114

สารบัญตาราง

หน้า

ตารางที่ 1 ตารางแสดงตัวอย่างการใช้เอ็นแกรมแบบตัวอักษรโดยเลือก n ตั้งแต่ 1 ถึง 6 13

ตารางที่ 2 ตารางแสดงตัวอย่างของประโยคทั่วไปในภาษาอังกฤษ และประโยคที่ตัดคำหยุดออก 13

ตารางที่ 3 ตารางแสดงอันดับของเพจข่าวบนเฟซบุ๊กเรียงตามยอมผู้ติดตาม 36

ตารางที่ 4 ตารางแสดงจำนวนโพสต์ที่เก็บได้จากหน้าเว็บไซต์ของเพจบนเฟซบุ๊ก 37

ตารางที่ 5 ตารางแสดงตัวอย่างโพสต์ของเพจ Khaosod - ข่าวสด 38

ตารางที่ 6 ตารางแสดงตัวอย่างโพสต์ของเพจ Workpoint Entertainment 39

ตารางที่ 7 ตารางแสดงตัวอย่างโพสต์ของเพจ เรื่องเล่าเช้านี้ 40

ตารางที่ 8 ตารางแสดงตัวอย่างโพสต์ของเพจ Thairath 41

ตารางที่ 9 ตารางแสดงตัวอย่างโพสต์ของเพจ Ch7HD 42

ตารางที่ 10 ตารางแสดงตัวอย่างโพสต์ของเพจ Ch7HD (ต่อ) 43

ตารางที่ 11 ตารางแสดงตัวอย่างโพสต์ของเพจ ช่อง one31 44

ตารางที่ 12 ตารางแสดงตัวอย่างโพสต์ของเพจ ช่อง one31 (ต่อ) 45

ตารางที่ 13 ตารางแสดงตัวอย่างโพสต์ของเพจ Sanook News 46

ตารางที่ 14 ตารางแสดงตัวอย่างโพสต์ของเพจ GMM25Thailand 47

ตารางที่ 15 ตารางแสดงตัวอย่างโพสต์ของเพจ Thai PBS 48

ตารางที่ 16 ตารางแสดงตัวอย่างโพสต์ของเพจ Mono 29 49

ตารางที่ 17 ตารางแสดงตัวอย่างโพสต์ของเพจ Mono 29 (ต่อ) 50

ตารางที่ 18 ตารางแสดงความแตกต่างระหว่างข้อความก่อนทำความสะอาด กับหลังทำความสะอาด 51

ตารางที่ 19 ตารางแสดงข้อความที่ถูกตัดคำโดยใช้เครื่องมือ deepcut 52

ตารางที่ 20 ตารางแสดงตัวอย่างของผลลัพธ์ เมื่อใช้วิธีเอ็นแกรมแบบตัวอักษรโดยเลือก n = 5 53

ตารางที่ 21 ตารางแสดงตัวอย่างการใช้เอ็นแกรมแบบตัวอักษรโดยเลือก n ตั้งแต่ 1 ถึง 6 54

ตารางที่ 22 ตารางแสดงตัวอย่างของผลลัพธ์ เมื่อใช้วิธีเอ็นแกรมแบบตัวอักษรโดยเลือก n = 10.... 55

ตารางที่ 23 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน 62

ตารางที่ 24 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 120 วัน 62

ตารางที่ 25 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 90 วัน 63

ตารางที่ 26 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 60 วัน 63

ตารางที่ 27 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 180 วัน 64

ตารางที่ 28 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน 65

ตารางที่ 29 ตารางแสดงคำหยุดทั้งหมด 132 คำ ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562 66

ตารางที่ 30 ตารางแสดงคำหยุดทั้งหมด 103 คำ ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562 67

ตารางที่ 31 ตารางแสดงข้อมูลคำหยุดที่หายไปทั้งหมด 29 คำ เมื่อใช้ข้อมูลอ้างอิงย้อนหลัง ช่วง 60 วัน ถึง 120 วัน แทนช่วง 30 วัน ถึง 180 วัน 67

ตารางที่ 32 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแส โดยใช้เสียงข้างมากจากอาสาสมัครทั้ง 5 คน โดยหากคำใดเป็นกระแสที่มีเสียงโหวตเกินครึ่งจะมีสีน้ำเงิน (ขีดเส้นใต้) บ่งบอกอยู่ 71

ตารางที่ 33 ตารางสรุปคำสำคัญที่เป็นกระแส โดยสีน้ำเงิน (ขีดเส้นใต้) หมายถึง คำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) หมายถึง คำที่มีเสียงโหวต 2 จาก 5 เสียง ที่เป็นคำกำกวม 73

ตารางที่ 34 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน 75

ตารางที่ 35 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน 75

ตารางที่ 36 ตารางแสดงคำหยุดทั้งหมด 286 คำ โดยใช้วิธี TF-IDF ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562 76

ตารางที่ 37 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF-IDF โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 จาก 5 เสียง ที่เป็นคำกำกวม 78

ตารางที่ 38 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน 79

ตารางที่ 39 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม..... 79

ตารางที่ 40 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแสโดยใช้วิธี TF-IDF ด้วยฐานข้อมูลคำหยุดที่ใช้ข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน จนถึง 180 วัน..... 80

ตารางที่ 41 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 เสียง 81

ตารางที่ 42 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน 82

ตารางที่ 43 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน 82

ตารางที่ 44 ตารางแสดงคำหยุดทั้งหมด 110 คำ โดยใช้วิธี TF ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูล ก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562... 83

ตารางที่ 45 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแสโดยใช้วิธี TF โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 จาก 5 เสียง 84

ตารางที่ 46 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน..... 85

ตารางที่ 47 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม..... 85

ตารางที่ 48 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 เสียง 88

ตารางที่ 49 ตารางแสดงคำสำคัญที่เป็นกระแสทั้งหมดจากวิธีเอ็นแกรมแบบตัวอักษร วิธี TF-IDF และวิธี TF โดยพิจารณายูนิแกรม และไม่พิจารณายูนิแกรม รวม 5 วิธี 90

ตารางที่ 50 ตารางเปรียบเทียบจำนวนคำสำคัญที่เป็นกระแสทั้ง 5 วิธี 91

ตารางที่ 51 ตารางแสดงปริมาณคำหยุดของทั้ง 5 วิธี บนช่วงวันอ้างอิงย้อนหลังทั้ง 2 ช่วง..... 92

ตารางที่ 52 ตารางแสดงคำหยุดที่หลุดมาเป็นคำสำคัญที่เป็นกระแสทั้งหมดจากวิธีเอ็นแกรม แบบตัวอักษร วิธี TF-IDF และวิธี TF โดยพิจารณายูนิแกรม และไม่พิจารณายูนิแกรม รวม 5 วิธี 95

ตารางที่ 53 ตารางเปรียบเทียบจำนวนคำหยุดที่หลุดจากข้อมูลอ้างอิง ทั้ง 5 วิธี..... 98

ตารางที่ 54 ตารางเปรียบเทียบประสิทธิภาพโดยใช้หน่วยวัดต่าง ๆ 99

สารบัญรูปภาพ

หน้า

รูปภาพที่ 1 สมการการคำนวณค่า TF-IDF ของแต่ละคำ 10

รูปภาพที่ 2 รูปภาพตัวอย่างการแปลงคำเป็นเวกเตอร์วันฮอต 11

รูปภาพที่ 3 รูปภาพตัวอย่างคำที่คล้ายคลึงกันเมื่อพิจารณาการสร้างเวกเตอร์แบบคำฝังตัว 12

รูปภาพที่ 4 รูปภาพตัวอย่างของการใช้เอ็นแกรมแบบคำ 12

รูปภาพที่ 5 รูปภาพตัวอย่างกราฟที่แสดงความสัมพันธ์ระหว่างความถี่กับความมีคุณค่า (คำที่ปรากฏในไม่กี่เอกสาร) ซึ่งค่า TF-IDF สะท้อนถึงคำที่มีความสำคัญได้ 14

รูปภาพที่ 6 รูปภาพตัวอย่างการใช้ค่า TF ในการสกัดคำสำคัญ โดยลบคำหยุดออก 15

รูปภาพที่ 7 รูปภาพตัวอย่างการแบ่งข้อมูลออกเป็น 2 กลุ่มโดยใช้เส้นตรง 16

รูปภาพที่ 8 รูปภาพตัวอย่างการใช้เคอร์แนลทริกแปลงข้อมูลทางฝั่งซ้าย ที่ไม่สามารถใช้เส้นตรง ในการแบ่งได้ ไปเป็นข้อมูลทางฝั่งขวา ที่สามารถหาเวกเตอร์เส้นตรงมาแบ่งได้ 16

รูปภาพที่ 9 รูปภาพตัวอย่างการหากลุ่มของข้อมูลจุดสีดำ หากเราเลือก k = 3 ข้อมูลจุดสีดำจะเป็นข้อมูลในกลุ่มสามเหลี่ยม แต่หากเราเลือก k = 5 ข้อมูลจุดสีเขียวจะเป็นข้อมูลในกลุ่มสี่เหลี่ยม 17

รูปภาพที่ 10 รูปภาพตัวอย่างการนำผลลัพธ์จากโมเดลมาใช้ในการทำนายผลร่วมกัน 18

รูปภาพที่ 11 รูปภาพตัวอย่างการแบ่งข้อมูลออกเป็น 2 กลุ่มโดยการใช้ต้นไม้ตัดสินใจ 18

รูปภาพที่ 12 รูปภาพตัวอย่างแบบจำลองจากป่าสุ่ม โดยการนำผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจแต่ละต้นมาหาค่าเฉลี่ยเพื่อเป็นผลทำนายสุดท้าย 19

รูปภาพที่ 13 รูปภาพตัวอย่างแบบจำลองโครงข่ายประสาทเทียมที่มี 4 ชั้น 20

รูปภาพที่ 14 รูปภาพตัวอย่างลักษณะการแบ่งกลุ่มข้อมูลแบบลำดับชั้น 21

รูปภาพที่ 15 รูปภาพตัวอย่างการจัดแบ่งกลุ่มข้อมูลแบบเคมีน โดยกำหนดให้ k = 3 22

รูปภาพที่ 16 รูปภาพตัวอย่างการจัดแบ่งกลุ่มใหม่ของเคมีน หลังพบว่าจุดเซนทรอยด์เปลี่ยนแปลง 22

รูปภาพที่ 17 กราฟแสดงความสัมพันธ์ระหว่างจำนวนกลุ่มกับค่าความแปรปรวน เพื่อเลือกค่า k.... 23

รูปภาพที่ 18 รูปภาพตัวอย่างโครงข่ายประสาทเทียมแบบ SOM 24

รูปภาพที่ 19 รูปภาพแสดงการเปลี่ยนแปลงของแต่ละโหนดที่โดนปรับค่าถ่วงน้ำหนักให้เข้าใกล้ข้อมูล 24

รูปภาพที่ 20 รูปภาพตัวอย่างโครงข่ายประสาทเทียมแบบ ART 25

รูปภาพที่ 21 รูปภาพแสดงวิธีการเลือกจำนวนกลุ่มโดยใช้วิธีข้อศอก โดยแกนนอนแสดงจำนวนกลุ่ม 57

รูปภาพที่ 22 รูปภาพแสดงการพิจารณาตัวอักษร “ค” ว่าเป็นส่วนหนึ่งของคำสำคัญหรือไม่ 58

รูปภาพที่ 23 รูปภาพแสดงการสกัดคำสำคัญจากตัวอักษรที่สำคัญที่ติดกัน 59

รูปภาพที่ 24 รูปภาพแสดงลักษณะของคำหยุด เทียบกับ ลักษณะของคำสำคัญที่เป็นกระแส 59

รูปภาพที่ 25 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุด กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า 63

รูปภาพที่ 26 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุด กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า 65

รูปภาพที่ 27 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF-IDF กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า 75

รูปภาพที่ 28 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม กับจำนวนเดือนที่ใช้สร้างฐานข้อมูล โดยเส้นกราฟแทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า . 79

รูปภาพที่ 29 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF กับจำนวนเดือนที่ใช้ ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า 83

รูปภาพที่ 30 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม กับจำนวนเดือนที่ใช้สร้างฐานข้อมูล โดยเส้นกราฟแทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า 86

บทนำ

ที่มาและความสำคัญ

ในยุคของข้อมูลและสารสนเทศ การที่เรารู้จักศึกษาและทำความเข้าใจข้อมูลเพื่อตกผลึกข้อมูลเหล่านั้นให้เป็นความเข้าใจลึกซึ้ง (Insight) ย่อมสามารถช่วยเพิ่มความได้เปรียบทางด้านธุรกิจได้เป็นอย่างดี

เฟซบุ๊ก (Facebook) ถือเป็นหนึ่งในสื่อสังคมออนไลน์ (Social Media Platform) ที่ใหญ่ที่สุด และมีผู้ใช้งานเป็นประจำต่อวันเยอะที่สุด โดยข้อมูลต่าง ๆ ในเฟซบุ๊กถือเป็นคลังสมบัติข้อมูลดิบที่มีค่า ล่อตาล่อใจนักการตลาดทั้งหลายเป็นอย่างมาก เพราะข้อมูลเหล่านั้นสามารถใช้วิเคราะห์เพื่อบ่งบอกถึงความเข้าใจลึกซึ้งเกี่ยวกับพฤติกรรมของผู้ใช้งานอันได้แก่ ความสนใจ หรือความสัมพันธ์ของผู้ใช้งาน เป็นต้น รวมถึงยังสามารถใช้บ่งบอกถึงกระแส (Trend) ต่าง ๆ ในสังคม ที่กำลังเกิดขึ้นหรือจะเกิดขึ้นในอนาคตได้อีกด้วย

วิธีหนึ่งที่สามารถใช้ในการวิเคราะห์ข้อมูลจากสื่อสังคมออนไลน์ได้นั้น คือการวิเคราะห์ข้อมูลประโยคที่ผู้ใช้งานพิมพ์ไว้ในสื่อสังคมออนไลน์ เพื่อสกัดเอาความเข้าใจลึกซึ้งจากโพสต์เหล่านั้น เช่น เรื่องที่ต้องการสื่อสาร (Topic) อารมณ์ (Emotion) หรือข้อความสำคัญ (Keyword) ที่ผู้ใช้งานต้องการสื่อ เป็นต้น โดยยังมีอีกหลายวิธีที่สามารถใช้ในการวิเคราะห์ข้อมูลจากสื่อสังคมออนไลน์ได้ เช่น การวิเคราะห์ข้อมูลรูปภาพ หรือข้อมูลความสนใจของผู้ใช้งาน เป็นต้น แต่ทว่าการวิเคราะห์ข้อมูลในปัจจุบันนั้น ต้องมีการคำนึงถึงความเป็นส่วนตัวของผู้ใช้งานด้วย งานวิจัยนี้จึงนำเฉพาะข้อมูลที่ผู้ใช้งานเปิดเผยแบบสาธารณะ (Public) เท่านั้น มาใช้ในการวิเคราะห์

เพจ (Page) เป็นพื้นที่หนึ่งของเฟซบุ๊ก ที่เปิดโอกาสให้ผู้ใช้งานสามารถรังสรรค์เนื้อหาที่อยากเผยแพร่สู่สาธารณะได้ โดยก่อนจะสร้างเพจบนเฟซบุ๊กได้ ผู้สร้างจะต้องกำหนดหมวดหมู่ให้เพจเหล่านั้นก่อน เช่น ธุรกิจ แบรนด์ ชุมชน หรือ บุคคลสาธารณะ เป็นต้น โดยผู้ใช้งานทุกคนในเฟซบุ๊กสามารถเข้าถึงเนื้อหาของเพจแบบสาธารณะ รวมถึงสามารถกดชอบ (Like) เพจใด ๆ ก็ได้ เพื่อติดตามเนื้อหาได้อีกด้วย

ผู้สร้างเพจจะรังสรรค์เนื้อหาตามเวลาต่าง ๆ เรียกว่าโพสต์ (Post) โดยหนึ่งโพสต์อาจประกอบไปด้วย ข้อความ สัญรูปอารมณ์ (Emoticon) และแฮชแท็ก (Hashtag) ที่มักใช้เน้นคำสำคัญของข้อความ หรือจัดกลุ่มโพสต์ให้สามารถหาได้โดยง่าย และยังสามารถใส่รูปภาพนิ่ง รูปภาพเคลื่อนไหวสั้นๆ หรือวีดีโอได้อีกด้วย โดยทุกโพสต์นั้นจะมีเวลากำกับอยู่ ซึ่งจะทำให้โพสต์จะถูกเรียงตามเวลาจากปัจจุบันไปยังอดีตเป็นเส้นเวลา (Timeline) เพื่อให้ผู้ใช้งานสามารถค่อย ๆ เลื่อนเพื่อดูเนื้อหาตามเวลาได้ และในแต่ละโพสต์ ผู้ใช้งานยังสามารถกดชอบ แสดงความคิดเห็น (Comment) หรือส่งต่อ (Share) ให้เพื่อน ๆ ของตนเองดูได้ด้วย

เพจจึงเป็นหนึ่งในแหล่งข้อมูล ที่มักจะถูกเลือกนำมาวิเคราะห์กระแสต่าง ๆ ที่กำลังเกิดขึ้นในสังคม หรือคาดว่าจะเกิดขึ้นในอนาคตอันใกล้ เนื่องจากมีความเป็นสาธารณะ ทำให้ข่าวหรือเรื่องราวต่าง ๆ มักจะถูกเผยแพร่บนเพจ และผู้ใช้งานมักจะไปแสดงความคิดเห็น หรือส่งต่อให้เพื่อน ๆ ของตนดู จึงทำให้ข่าวแพร่กระจายออกไป เกิดเป็นกระแสขึ้นมา

งานวิจัยฉบับนี้ขอลงลึกที่การวิเคราะห์กระแสที่เกิดขึ้นบนสื่อสังคมออนไลน์ เนื่องจากหากเราสามารถรู้ถึงกระแสที่กำลังเกิดขึ้นในสื่อสังคมออนไลน์ได้เร็ว เราก็สามารถใช้ประโยชน์จากกระแสที่กำลังเกิดขึ้นได้ก่อน ซึ่งเป็นข้อได้เปรียบหนึ่งทางธุรกิจและการวิเคราะห์กระแสผ่านช่องทางอื่น ก็ทำได้ยากกว่ามาก ทำให้งานด้านการวิเคราะห์กระแสที่เกิดขึ้นบนสื่อสังคมออนไลน์นั้น เป็นงานที่จำเป็นและมีประโยชน์ต่อธุรกิจต่าง ๆ โดยตรง

กระแสต่าง ๆ ที่เกิดขึ้นบนสื่อสังคมออนไลน์ มักจะถูกวิเคราะห์ออกมาในรูปแบบของกลุ่มคำที่มักจะถูกพูดถึงหรือถูกพิมพ์ซ้ำอยู่บ่อย ๆ ในช่วงเวลาหนึ่ง หรือกล่าวคือ เราสามารถวิเคราะห์กระแสที่เกิดขึ้นบนสื่อสังคมออนไลน์ได้จากคำสำคัญ (Keyword) บางอย่างที่ถูกซ่อนอยู่ภายในประโยค หรือบางครั้งก็ถูกทำให้เห็นชัดขึ้นโดยผู้ใช้งาน โดยการใส่ป้ายกำกับ (Tag) เน้นคำเหล่านั้นให้เห็นเด่นชัดขึ้นมา ซึ่งคำที่ถูกเน้นในสื่อสังคมออนไลน์ ถูกเรียกว่าแฮชแท็ก (Hashtag) โดยคำสำคัญที่จะเป็นกระแสได้นั้น ต้องเป็นกลุ่มคำที่ถูกพูดถึงในหลากหลาย แต่ก็ต้องไม่ใช่คำธรรมดาทั่ว ๆ ไป โดยเราสามารถจำแนกได้จากการที่กลุ่มคำเหล่านั้นมักไม่เคยปรากฏขึ้นเลย แต่ในช่วงเวลาหนึ่งกลับปรากฏขึ้นมาบ่อยมาก และพอกระแสของคำสำคัญเหล่านั้นเริ่มหมด ก็จะปรากฏน้อยลงอย่างเห็นได้ชัด

การวิเคราะห์ข้อมูลประโยค และการสกัดคำสำคัญ (Keyword Extraction) เป็นหนึ่งในงานด้านการวิเคราะห์ข้อมูลภาษาธรรมชาติ (Natural Language Processing) ซึ่งหนึ่งในเครื่องมือพื้นฐานที่จำเป็นสำหรับการวิเคราะห์ข้อมูลประโยคคือเครื่องมือที่ใช้สำหรับการตัดคำ (Word Tokenizer) เพื่อช่วยให้คอมพิวเตอร์สามารถเข้าใจข้อมูลด้านภาษาได้ง่ายขึ้น

แต่ทว่า ข้อความบนสื่อสังคมออนไลน์ของคนไทย ไม่ได้ประกอบไปด้วยตัวอักษรภาษาไทยเท่านั้น แต่ยังมีตัวอักษรภาษาอังกฤษ หรือภาษาอื่น ๆ ปะปนมา รวมถึงยังมี ตัวเลข สัญรูปอารมณ์ และอักขระอื่น ๆ ด้วย นอกจากนี้ ด้วยการที่ผู้คนบนสื่อสังคมออนไลน์ต้องการแสดงออกซึ่งอารมณ์ของพวกเขาผ่านทางตัวอักษร ทำให้เกิดการแปลงข้อความให้ไม่ถูกต้องตามหลักภาษาอยู่บ่อยครั้ง ทำให้เกิดรูปแบบของคำไม่เป็นทางการต่าง ๆ [1] ที่ทำให้การวิเคราะห์ข้อมูลภาษาธรรมชาติ ทำงานได้ยากขึ้น และเกิดข้อผิดพลาดมากขึ้น อีกทั้งเครื่องมือพื้นฐานอย่างเครื่องมือสำหรับตัดคำ ก็มักไม่ทนทานต่อความผิดพลาดในลักษณะนี้ เช่น หนึ่งในเครื่องมือตัดคำภาษาไทยที่ไม่มีค่าใช้จ่ายในการใช้งาน เล็กซ์โตพลัส (LexToPlus) ซึ่งจัดทำโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) จึงเป็นที่นิยมใช้งานอย่างแพร่หลาย แต่ทว่าเล็กซ์โตพลัสนั้น เป็นเครื่องมือที่ใช้วิธีตัดคำโดยอิงคำตามพจนานุกรม ซึ่งไม่เหมาะสมกับการตัดคำบนสื่อสังคมออนไลน์ ที่มีการใช้ภาษาที่เพี้ยนไป

บ้าง เพื่อเล่นเสียง และสามารถสื่อถึงอารมณ์ต่าง ๆ ของผู้ใช้งานได้ อีกทั้งยังมีทั้งตัวอักษรภาษาอังกฤษ และตัวอักขระอื่น ๆ ด้วย ทำให้มีความผิดพลาดในการตัดคำมากกว่างานทางภาษาด้านอื่น ๆ ซึ่งทำให้การตัดคำให้ถูกต้องบนสื่อสังคมออนไลน์ยังคงเป็นงานที่ท้าทาย

การสกัดคำสำคัญ (Keyword Extraction) เป็นหนึ่งในวิธีที่การที่ต้องพึ่งพาเครื่องมือที่ใช้ตัดคำเพื่อใช้ในการสกัดคำสำคัญออกจากประโยคต่อไป โดยบางครั้งเรามักจะใช้ฐานข้อมูลคำหยุด (Stop words) สำหรับการสกัดคำสำคัญด้วย เนื่องจากคำหยุดนั้น เป็นคำที่เกิดขึ้นบ่อยในภาษา แต่ทว่าไม่ได้มีลักษณะของใจความสำคัญอยู่ในนั้น เช่นคำว่า “เลย” “ครับ” หรือ “ค่ะ” เป็นต้น เราจึงมักลบคำหยุดออกจากข้อความก่อนที่จะหาใจความสำคัญ แต่สำหรับภาษาไทยนั้น ฐานข้อมูลคำหยุด มีคำอยู่ค่อนข้างน้อย ซึ่งไม่เพียงพอต่อการใช้งาน

จากเหตุผลดังกล่าว การสกัดคำสำคัญสำหรับข้อมูลภาษาไทยในสื่อสังคมออนไลน์ ยังเป็นงานที่ท้าทาย งานวิจัยนี้จึงขอเสนอวิธีการใหม่ ที่สามารถสกัดคำสำคัญที่เป็นกระแส และคำหยุด จากเพจเฟซบุ๊กภาษาไทย โดยใช้อัลกอริทึมที่เรียบง่ายอย่าง เอ็นแกรม (n-Grams) มาประยุกต์ใช้เพื่อให้เราสามารถนำข้อมูลคำสำคัญที่เป็นกระแสเหล่านี้ไปวิเคราะห์ต่อเพื่อหาความเข้าใจลึกซึ้งของกระแสที่เกิดขึ้นต่อไป

วัตถุประสงค์

เพื่อคิดค้นวิธีการสกัดคำสำคัญที่เป็นกระแส สำหรับสื่อสังคมออนไลน์เฟซบุ๊กที่เป็นภาษาไทย โดยไม่พึ่งพิงเครื่องมือตัดคำ และฐานข้อมูลคำหยุดจากแหล่งข้อมูลอื่น โดยผลลัพธ์ที่ได้ใช้เป็นข้อมูลประกอบกับคนอ่านเพื่อทำความเข้าใจกระแสที่จะเกิดขึ้นบนสื่อสังคมออนไลน์เฟซบุ๊ก

ขอบเขตการดำเนินงาน

สนใจเฉพาะข้อมูลโพสต์จากเพจบนสื่อสังคมออนไลน์เฟซบุ๊กเท่านั้น และสนใจเฉพาะโพสต์ข้อความภาษาไทยเป็นหลัก อาจมีภาษาอื่นหรือตัวอักขระอื่นปนมาได้บางส่วน ซึ่งไม่ส่งผลเมื่อใช้คนมาอ่านคำสำคัญที่เป็นกระแสที่ระบบสกัดออกมาได้

ประโยชน์ที่คาดว่าจะได้รับ

1. ได้วิธีการที่สามารถสกัดคำสำคัญที่เป็นกระแส จากเพจเฟซบุ๊กภาษาไทยที่มีประสิทธิภาพ
2. ได้ฐานข้อมูลคำหยุดสำหรับขอบเขตงานด้านนี้ ที่สามารถนำไปใช้งานต่อยอดได้
3. สามารถนำวิธีที่วิจัย ไปปรับใช้เพื่อหาคำสำคัญที่เป็นกระแสในสื่อสังคมออนไลน์อื่น ๆ หรืองานด้านอื่น ๆ ที่มีความใกล้เคียงกัน
4. สามารถนำปัญหาที่พบ และวิธีแก้ไขไปปรับใช้ในงานลักษณะเดียวกันได้

วิธีการดำเนินงานวิจัย

1. ศึกษางานวิจัยที่เกี่ยวข้อง
2. ออกแบบฐานข้อมูล และสร้างเครื่องมือสำหรับเก็บข้อมูลโพสต์จากเพจเฟซบุ๊ก
3. ออกแบบวิธีที่ใช้ในการสกัดคำสำคัญที่เป็นกระแส และวิธีที่ใช้ในการวัดผล
4. เก็บรวบรวมข้อมูลโพสต์จากเพจเฟซบุ๊ก
5. ทดลองสกัดคำสำคัญที่เป็นกระแส และวัดผลการทดลอง
6. ตีพิมพ์ผลงานทางวิชาการ
7. สรุปผลการทดลองและจัดทำวิทยานิพนธ์
8. สอบวิทยานิพนธ์

ผลงานตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ถูกตีพิมพ์ในวารสารทางวิชาการ International Journal of Machine Learning and Computing 2018 Vol.8(6) เป็นบทความทางวิชาการในหัวข้อเรื่อง Extraction of Trend Keywords and Stop Words from Thai Facebook Pages using Character n-Grams จัดทำโดย Nattapong Ousirimaneechai, Sukree Sinthupinyo และถูกนำเสนอในงานประชุมวิชาการ 2018 2nd Asia Conference on Machine Learning and Computing ณ เมืองโฮจิมินห์ ประเทศเวียดนาม ในวันที่ 8 ธันวาคม 2561 ซึ่งแสดงในภาคผนวก ก

ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์

ในวิทยานิพนธ์ฉบับนี้ ได้มีการแบ่งเนื้อหาออกเป็น 5 บทย่อย ตามลำดับดังนี้

บทที่ 1: บทนำ กล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ ขอบเขตการดำเนินงาน ประโยชน์ที่คาดว่าจะได้รับ วิธีการดำเนินงานวิจัย และผลงานตีพิมพ์จากวิทยานิพนธ์ฉบับนี้

บทที่ 2: ทฤษฎีและงานวิจัยที่เกี่ยวข้อง กล่าวถึงทฤษฎีและแนวคิดที่เกี่ยวข้อง โดยประกอบด้วย การตัดคำ การสกัดคำสำคัญ และการแบ่งกลุ่มข้อมูลแบบเคมีน และอีกส่วนหนึ่งคืองานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์ฉบันนี้

บทที่ 3: ขั้นตอนการสร้างระบบสกัดคำสำคัญที่เป็นกระแสและคำหยุด จากเพจเฟซบุ๊ก กล่าวถึงกระบวนการต่าง ๆ ในการได้มาซึ่ง คำสำคัญที่เป็นกระแส และคำหยุด จากข้อมูลในเพจเฟซบุ๊ก

บทที่ 4: การทดลอง และอภิปรายผล กล่าวถึง ขั้นตอนการทดลอง การออกแบบตัววัดผลการทดลอง และผลการทดลอง

บทที่ 5: บทสรุปผลการวิจัย และข้อเสนอแนะ กล่าวถึง ข้อสรุปที่ได้จากการวิจัย จุดเด่นจุดด้อยของงานวิจัยฉบับนี้ รวมถึงข้อเสนอแนะต่าง ๆ สำหรับแนวทางการวิจัยในขั้นถัดไป

ทฤษฎีที่เกี่ยวข้อง

วิทยานิพนธ์ฉบับนี้ มีความเกี่ยวข้องกับทฤษฎีต่าง ๆ อยู่จำนวนหนึ่ง ซึ่งทฤษฎีที่เกี่ยวข้องนั้นเป็นความรู้พื้นฐานของวิทยานิพนธ์ฉบับนี้ และเป็นความรู้พื้นฐานสำหรับงานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ด้วย ซึ่งจากความรู้ดังกล่าว เราจะพบปัญหาและข้อจำกัดต่าง ๆ รวมถึงวิธีที่น่าจะสามารถแก้ปัญหาและข้อจำกัดเหล่านั้นได้ โดยมีหัวข้อดังนี้

1. คำที่ไม่เป็นมาตรฐานตามพจนานุกรม
2. การตัดคำ (Word Segmentation)
3. การแทนข้อความ (Text Representation)
4. การสกัดคำสำคัญ (Keyword Extraction)
5. การจำแนกข้อมูล (Classification)
6. การแบ่งกลุ่มข้อมูล (Clustering)

คำที่ไม่เป็นมาตรฐานตามพจนานุกรม

คำในสื่อสังคมออนไลน์ มักจะประกอบด้วยคำที่ไม่เป็นมาตราฐานตามพจนานุกรมจำนวนมาก ซึ่งส่งผลต่อการวิเคราะห์คำด้วยเครื่องมือที่อ้างอิงคำตามพจนานุกรม ซึ่งในภาษาอังกฤษมีคำที่ไม่เป็นมาตราฐานต่าง ๆ ซึ่งถูกจำแนกไว้ดังนี้ [1]

1. คำย่อ (Abbreviation)

   คือคำที่ถูกย่อเพื่อให้สามารถพิมพ์สื่อสารกันในสื่อสังคมออนไลน์ได้รวดเร็วขึ้นถูกแบ่งออกเป็น 2 แบบดังนี้

   - *รูปย่อ (Short form)*

     เกิดจากการย่อคำที่มักใช้บ่อย แต่ต้องการพิมพ์ให้สั้นลงโดยมีการคงลักษณะบางอย่างไว้ เช่นคำว่า “night” ถูกย่อเป็น “nite” หรือคำว่า “saying” ถูกย่อเป็น “sayin” ซึ่งบางครั้งอาจมีการใช้ตัวอักษรอื่นเช่นตัวเลขมาใช้ในการย่อคำ เช่น “great” ถูกย่อเป็น “gr8” หรือคำว่า “Kubernetes” ในแวดวงคอมพิวเตอร์ก็ถูกย่อเป็นคำว่า “K8s” เป็นต้น

     สำหรับภาษาไทยก็มีการย่อคำในลักษณะนี้เช่นกัน เช่นคำว่า “ครับ” เป็น “คับ” หรือคำว่า “เธอ” เป็น “เทอ” เป็นต้น โดยภาษาไทยมีการย่อด้วยตัวเลขเช่นกัน เช่นคำว่า “ฮ่าฮ่าฮ่า” ซึ่งแสดงเสียงหัวเราะ ถูกย่อด้วยคำว่า “555” ซึ่งออกเสียงคล้ายกันในภาษาไทย ซึ่งรูปย่อเหล่านี้ ยังไม่เป็นที่ยอมรับอย่างเป็นทางการสำหรับภาษาไทย จึงถือว่าเป็นคำที่เขียน/พิมพ์ผิดอยู่

- *ตัวย่อ (Acronym)*

วลีที่ถูกใช้บ่อยในสื่อสังคมออนไลน์ เช่นคำว่า “laugh out loud” มักถูกย่อเป็น “lol” หรือ “If I remember correctly” มักถูกย่อเป็น “iirc” เป็นต้น

สำหรับภาษาไทยก็มีการย่อคำในลักษณะนี้เช่นกัน แต่ย่อถึงระดับคำ เนื่องจากคำมีความยาวมาก เช่นคำว่า “กรุงเทพมหานคร” ถูกย่อเป็น “กทม.” หรือบางครั้งอาจใช้ “ฯ” ต่อท้ายแทนในการย่อแทน “.” ก็ได้เช่นกัน เช่นคำว่า “กรุงเทพมหานคร” ถูกย่อเป็น “กรุงเทพฯ” ซึ่งคำเหล่านี้หากถูกยอมรับอย่างกว้างขวางและเป็นเวลานาน ก็จะถูกยกให้เป็นตัวย่ออย่างเป็นทางการ ซึ่งในภาษาอังกฤษก็มีตัวย่อที่เป็นทางการเช่นกัน เช่นคำว่า “do not” ถูกย่อเป็น “don’t” เป็นต้น แต่ก็มีคำบางครั้งที่ถูกใช้บ่อย แต่ยังไม่ถูกยอมรับอย่างเป็นทางการ เช่นคำว่า “พรุ่งนี้” เป็น “พน.” เป็นต้น

2. คำที่สะกดผิด (Misspelling / Typing error)

คำที่สะกดผิด คือคำที่ผู้สะกดหรือผู้พิมพ์ไม่ได้ตั้งใจสะกดผิด โดยถูกแบ่งเป็น 2 ประเภท ดังนี้

- *คำที่สะกดผิด (Misspelling)*

คือคำที่เกิดขึ้นจากผู้สะกดจำตัวอักษรที่ใช้สะกดคำดังกล่าวได้ไม่ถูกต้อง เช่นคำว่า “ridiculous” มักถูกสะกดเป็น “rediculous" หรือคำในภาษาไทยอย่างคำว่า “กงสุล” มักถูกสะกดเป็นคำว่า “กงศุล” ซึ่งคำที่สะกดผิดบางคำ อาจถูกสะกดผิดโดยแพร่หลาย จนทำให้ผู้สะกดไม่รู้ตัวว่าตนเองสะกดคำดังกล่าวผิด

- *คำที่พิมพ์ผิด (Typing error)*

คือคำที่ผู้พิมพ์กดแป้นพิมพ์ผิดพลาดโดยบังเอิญ ซึ่งอาจจะไปโดนตัวอักษรที่อยู่ข้างเขียงกัน เป็นต้น เช่นคำว่า “would” อาจกลายเป็นคำว่า “wouls" เนื่องจากตัวอักษร “d” กับ “s” อยู่ติดกัน หรือในภาษาไทยคำว่า “สำเร็จ” อาจกลายเป็นคำว่า “สำเร็ต” เนื่องจากตัวอักษร “จ” กับ “ต” อยู่ติดกันบนแป้นพิมพ์ ซึ่งบางครั้งอาจจะทำให้ความหมายของคำเปลี่ยนเป็นอีกคำหนึ่ง ระบบจึงยากที่จะตรวจจับคำที่พิมพ์ผิดโดยไม่อ้างอิงความหมายจากบริบทโดยรอบ เช่น ผู้พิมพ์อาจต้องการพิมพ์คำว่า “ขอบคุณครับ” แต่ตัวอักษร “ข” กับ “ช” อยู่ติดกัน จึงอาจกลายเป็นประโยคว่า “ชอบคุณครับ” ซึ่งหากไม่มีบริบทอื่นใด ก็ไม่สามารถตีความได้ว่าผู้พิมพ์ พิมพ์ผิดหรือไม่ เป็นต้น

3. คำที่ถูกตัดตัวอักษรออก (Punctuation omission/error)

รูปย่อบางรูป อาจถูกละตัวอักษรที่ใช้ย่อออก เช่นคำว่า “do not” รูปย่อคือ “don’t” อาจถูกละตัวอักษร “’” อีกจนเหลือแค่ “dont" เป็นต้น ซึ่งสำหรับภาษาไทย ตัวอักษรลงท้ายรูปย่ออย่าง “.” และ “ฯ” ก็มักจะถูกละเช่นกัน เช่นคำว่า “กรุงเทพมหานคร” รูปย่อคือ “กทม.” หรือ “กรุงเทพฯ” ก็มักถูกละตัวอักษรย่อ เหลือแค่ “กทม” หรือ “กรุงเทพ” เป็นต้น

4. คำแสลง (Non-dictionary slang)

คือคำที่มีความหมายในบางสังคม แต่ไม่ถูกบรรจุอยู่ในพจนานุกรม ซึ่งรวมถึงวลีบางอย่าง ที่ไม่สามารถตีความหมายได้ตรงตัว เช่นคำว่า “Cool” อาจถูกพิมพ์เป็นคำแสลงว่า “Wicked” หรือวลีอย่าง “that was very good” อาจถูกพิมพ์แบบแสลงเป็น “that was well mint” หรือคำว่า “แอ๊บแบ๊ว” ในภาษาไทย ซึ่งเป็นคำที่ไม่มีในพจนานุกรม แต่คนในสังคมบางกลุ่มเข้าใจโดยทันทีว่าหมายถึงคำว่า “น่ารัก” เป็นต้น

5. การเล่นคำ (Wordplay)

การเล่นคำเป็นสิ่งที่เกิดขึ้นในสื่อสังคมออนไลน์ เนื่องจากการพิมพ์ติดต่อสื่อสารกันมักจะทำให้ไม่รู้ถึงอารมณ์ของผู้พิมพ์ ผู้พิมพ์จึงอาศัยการเล่นคำ เช่นการเล่นคำพ้องเสียงเพื่อแสดงออกทางอารมณ์ ซึ่งนับรวมถึงคำที่จงใจสะกดผิด (Intentional misspelling) ด้วย ตัวอย่างเช่นข้อความ “that was soooo great” โดยการทำให้คำว่า “so” มีตัวอักษร “o” เยอะขึ้นเพื่อทำให้ผู้อ่าน อ่านออกเสียงคำนี้ยาวขึ้น ซึ่งทำให้ผู้อ่านรู้สึกถึงอารมณ์ของคำนี้เยอะขึ้น ในภาษาไทยก็นิยมใช้ เช่นคำว่า “ชอบมากกกกก” โดยการทำให้คำว่า “มาก” มีตัวอักษร “ก” เยอะขึ้น เพื่อเน้นว่าชอบมากจริง ๆ หรือในบางครั้งอาจมีความหมายในแง่การประชดได้อีกด้วย หรือคำว่า “ถูกจัย” ก็เป็นการเล่นคำมากจากคำว่า “ถูกใจ” ซึ่งแสดงถึงสถานะทางสังคมของผู้พิมพ์ที่แตกต่างบนสื่อสังคมออนไลน์ เป็นต้น

6. คำที่หลบหลีกการจับผิด (Censor avoidance)

คำบางคำซึ่งเป็นคำหยาบ มักถูกกรองออกโดยระบบ ทำให้ผู้พิมพ์หลบหลีกการพิมพ์คำเหล่านั้นตรง ๆ โดยการจงใจสะกดผิด เช่นคำว่า “shit” กลายเป็นคำว่า “shlt” หรือในภาษาไทยคำว่า “มึง” กลายเป็นคำว่า “เมิง” เป็นต้น หรืออาจใช้การเล่นคำโดยการใส่ตัวอักษรที่ระบบใช้กรองคำหยาบแทนเพื่อล้อเลียนและทำให้เข้าใจคำเหล่านั้นได้ เช่นคำว่า “fuck” กลายเป็นคำว่า “f***” โดยหากพิมพ์คำว่า “fuck” จะถูกระบบแทนที่ด้วยตัวอักษร “****” เป็นต้น

7. สัญรูปอารมณ์ (Emoticons)

เป็นตัวอักษรที่ประกอบจากตัวอักษรหรืออักขระอื่น เพื่อแสดงความหมายบางอย่าง โดยมากใช้แทนที่รูปภาพ เช่นตัวอักษร “T^T” ต้องการแทนที่รูปภาพหน้าคนร้องไห้ โดยต้องการให้ผู้อ่านรับรู้อารมณ์ว่ากำลังเศร้าหรือกำลังร้องไห้ เป็นต้น ซึ่งจากการที่ถูกใช้อย่างแพร่หลายบนสื่อสังคมออนไลน์ ทำให้มีตัวอักขระสัญรูปอารมณ์แบบเฉพาะ ซึ่งมักถูกระบบแสดงผลตัวอักษรเหล่านี้ด้วยภาพขนาดเล็ก โดยต้องการใช้แทนสัญรูปอารมณ์แบบดั้งเดิมที่อาจมีความหลากหลายในการพิมพ์มากกว่า ซึ่งการใช้ตัวอักขระสัญรูปอารมณ์แบบเฉพาะทำให้คอมพิวเตอร์สามารถเข้าใจสัญรูปอารมณ์เหล่านี้ได้มากขึ้น

การตัดคำ (Word Segmentation)

เป็นการแบ่งข้อความเพื่อหาขอบเขตของแต่ละหน่วยคำ โดยสำหรับภาษาไทย ที่ไม่มีการใช้เครื่องหมายวรรคตอนในการแบ่งคำ ทำให้เราจำเป็นต้องมีเครื่องมือที่ใช้ในการตัดคำ ซึ่งสำหรับงานทางด้านคอมพิวเตอร์ การจัดการกับคำเป็นสิ่งจำเป็น โดยอัลกอริทึมที่ใช้ในการตัดคำ มักจะใช้พจนานุกรม เพื่อหาขอบเขตของคำ โดยมักจะมีการใช้กฎเพิ่มเติมที่ช่วยในการแบ่งขอบเขต เช่น กฎความยาวคำสูงสุด กฎจำนวนคำน้อยสุด เป็นต้น [2]

ซึ่งไม่ใช่แค่ภาษาไทยเท่านั้นที่จำเป็นต้องมีเครื่องมือที่ช่วยในการแบ่งข้อความ โดยมีงานวิจัยที่พัฒนาเครื่องมือที่ช่วยในการตัดคำบนภาษาอื่น ๆ เช่น ภาษาจีน และภาษาญี่ปุ่น [3], [4], [5] ซึ่งสามารถจำแนกวิธีที่ใช้ในการตัดคำออกเป็นเป็น 2 วิธีใหญ่ ๆ ได้แก่ [6]

1. การตัดคำโดยใช้พจนานุกรม (Dictionary-based: DCB)

วิธีนี้จะใช้คำที่อยู่ในพจนานุกรมมาอ้างอิงสำหรับการหาขอบเขตของคำ โดยสามารถประยุกต์ใช้คู่กับกฎเพิ่มเติมได้หลากหลายแบบ โดยวิธีที่นิยมได้แก่

- *การตัดคำแบบเลือกคำที่ยาวที่สุด (Longest Matching) [7]*

วิธีนี้จะพิจารณาคำจากซ้ายไปขวาของประโยค โดยพยายามตัดคำตามคำที่มีอยู่ในพจนานุกรม หากคำที่ตัดออกมา ยังเป็นส่วนหนึ่งของคำอื่นที่ยาวกว่า และยังสามารถตัดเป็นคำที่ยาวกว่าได้อยู่ ก็จะเลือกตัดเป็นคำที่ยาวกว่าก่อน แต่หากตัดคำไปจนจบประโยคแล้ว ปรากฏว่าตัวอักษรที่เหลืออยู่ในประโยคไม่เป็นคำที่อยู่ในพจนานุกรม จะยอมไล่ย้อนกลับไปเลือกคำที่ยาวน้อยกว่าในพจนานุกรมแทน จนสามารถตัดคำทั้งประโยคได้ แต่อย่างไรก็ตามวิธีนี้ก็ยังคงมีปัญหา หากมีคำที่ไม่พบในพจนานุกรมเลย [8]

- *การตัดคำแบบเลือกคำที่เหมือนมากที่สุด (Maximum Matching) [9]*

  วิธีนี้จะมีแนวคิดตรงข้ามกับการตัดคำแบบเลือกคำที่ยาวที่สุดที่ลองตัดคำเพียงไม่กี่รูปแบบ หากตัดได้สำเร็จทั้งประโยคตามเงื่อนไขก็จบ เนื่องจากหากเราเลือกคำที่ยาวที่สุดก่อน คำต่อ ๆ มามีโอกาสตัดผิดพลาดสูงมากเกินไป โดยวิธีนี้จะลองตัดคำทุกแบบที่เป็นไปได้ของประโยคนั้น แล้วเลือกคำตอบที่ทำให้จำนวนคำที่ตัดจากประโยคนั้นน้อยที่สุด เนื่องจากภาษาไทยมีคำผสมเป็นจำนวนมาก แต่แลกมาด้วยพลังที่ใช้ในการประมวลผลเพื่อตัดคำในแต่ละประโยค

นอกจากวิธีที่กล่าวข้างต้น ยังมีวิธีอื่น ๆ ที่สามารถใช้ได้ เช่น การตัดคำโดยใช้คำที่มีความยาวน้อยที่สุด (Shortest Matching) การตัดคำโดยใช้ความน่าจะเป็นทางสถิติ (Probabilistic Matching) และการตัดคำโดยใช้คุณลักษณะ (Rule-based Matching / Feature-based Matching) เป็นต้น [10]

2. การตัดคำโดยใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning-based: MLB)

การตัดคำรูปแบบนี้จะใช้การเรียนรู้ของเครื่องมาสร้างแบบจำลองที่บอกว่าตัวอักษรไหนของประโยคบ้าง ที่เป็นจุดเริ่มต้นของคำ โดยการจะสร้างแบบจำลองดังกล่าวได้จะต้องมีข้อมูลผลเฉลยจำนวนมาก ของประโยคตัวอย่าง และคำที่ถูกตัดเรียบร้อยแล้วเพื่อให้เครื่องจักรเรียนรู้ผลเฉลยเหล่านั้น และสร้างแบบจำลองการตัดคำได้ ซึ่งวิธีนี้ไม่จำเป็นต้องใช้ฐานข้อมูลคำจากพจนานุกรมเลย โดยวิธีการเรียนรู้ของเครื่องก็มีหลากหลายวิธีให้เลือกใช้สร้างแบบจำลอง เช่น นาอีฟเบย์ (Naive Bayes: NB) [11] ต้นไม้ตัดสินใจ (Decision Tree: DT) [12] ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) [13] คอนดิชันแนลแรนดอมฟิลด์ (Conditional Random Field: CRF) [14] และโครงข่ายประสาทเทียม (Artificial Neural Networks: NN) [15], [16] โดยความแม่นยำของวิธีการเรียนรู้ของเครื่องแปรผันตามปริมาณและความหลากหลายของข้อมูลผลเฉลย ซึ่งเป็นเรื่องยากที่เราจะสร้างผลเฉลยของประโยคบนสื่อสังคมออนไลน์ที่มีการใช้คำไม่เป็นทางการต่าง ๆ เปลี่ยนแปลงไปตามกระแสสังคม

การแทนข้อความ (Text Representation)

ในงานทางด้านการประมวลผลภาษาธรรมชาติ การประมวลผลตัวอักษร (String) ไม่สามารถกระทำได้กับในงานหลาย ๆ ประเภท โดยมากเรามักจะแปลงข้อความเหล่านั้นในรูปแบบอื่นก่อน แล้วจึงนำไปประมวลผลต่อไป โดยมีวิธีการแปลงข้อความดังต่อไปนี้

1. ถุงคำ (Bag of Words: BoW)

   เป็นการแปลงข้อความให้เป็นเวกเตอร์ โดยเปลี่ยนคำทั้งหมดที่มีในข้อความมาเป็นคำกับความถี่ของคำนั้น โดยเรียงตามลำดับของคำในพจนานุกรม ตัวอย่างเช่น ถ้าเรามีคำในพจนานุกรมดังต่อไปนี้ [กราบ, ของ, คุณครู, นักเรียน, สวัสดี] ดังนั้นข้อความ “นักเรียนสวัสดีคุณครูของนักเรียน” จะถูกแปลงเป็นเวกเตอร์ดังต่อไปนี้ [0, 1, 1, 2, 1] ซึ่งตีความได้ว่า คำว่า “กราบ” ไม่อยู่ในข้อความ คำว่า “ของ” มีอยู่ 1 คำในข้อความ คำว่า “คุณครู” มีอยู่ 1 คำในข้อความ คำว่า “นักเรียน” มีอยู่ 2 คำในข้อความ และคำว่า “สวัสดี” มีอยู่ 1 คำในข้อความ

2. การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (Term Frequency - Inverse Document Frequency: TF-IDF)

   การแทนข้อความด้วย TF-IDF เป็นวิธีการใช้ถุงคำรูปแบบหนึ่ง แต่แทนที่เราจะแทนที่แต่ละช่องของเวกเตอร์ด้วยความถี่ของคำ เราจะมีการคำนวณค่าความถี่ของคำ (Term Frequency: TF) โดยนับความถี่ของคำนั้นจากจำนวนของคำทั้งหมดในข้อความเหมือนถุงคำ และคำนวณค่าผกผันของจำนวนเอกสารที่คำนั้นปรากฏ (Inverse Document Frequency: IDF) โดยนับความถี่ว่าคำนั้นปรากฏอยู่ในกี่ข้อความ และนำจำนวนเอกสารทั้งหมด มาส่วนด้วยค่าความถี่ดังกล่าว แล้วนำไปเปลี่ยนให้อยู่ในรูปของค่า log โดยค่า TF-IDF จะมีค่าเท่ากับผลคูณของค่า TF กับค่า IDF ดังสมการใน*รูปภาพที่ 1* แล้วเมื่อคำนวณค่า TF-IDF ของแต่ละคำในแต่ละข้อความได้แล้ว เราก็สามารถสร้างเวกเตอร์ TF-IDF ของข้อความได้แบบเดียวกับถุงคำนั่นเอง

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$w_{i,j} =$ ค่า TF-IDF ของคำ i ในเอกสาร j

$tf_{i,j} =$ จำนวนการปรากฏคำ i ในเอกสาร j

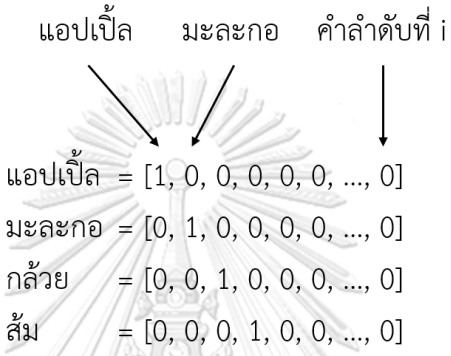
$df_i =$ จำนวนของเอกสารที่มีคำ i อยู่ในเอกสาร

$N =$ จำนวนเอกสารทั้งหมด

รูปภาพที่ 1 สมการการคำนวณค่า TF-IDF ของแต่ละคำ

3. เวกเตอร์วันฮอต (One-hot Vector)

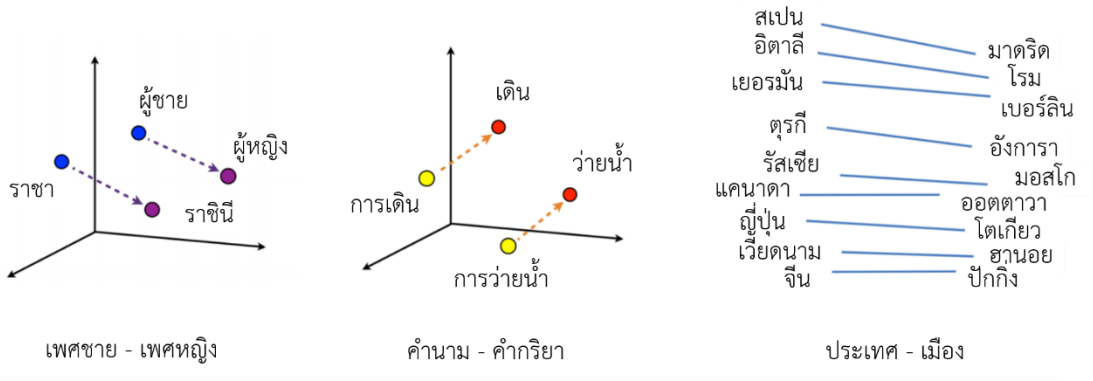
เป็นการแทนที่แต่ละคำด้วยเวกเตอร์หนึ่งหน่วย โดยจำนวนคำในพจนานุกรม จะมีค่าเท่ากับจำนวนมิติของเวกเตอร์วันฮอต กล่าวคือ ถ้าในพจนานุกรมมี 3 คำ เวกเตอร์วันฮอตก็จะมี - มิติ โดยแต่ละคำ จะถูกแทนด้วยเวกเตอร์ [1, 0, 0], [0, 1, 0], [0, 0. 1] ตามลำดับ ซึ่งมีข้อเสียคือขนาดของเวกเตอร์แต่ละตัวจะใหญ่มาก หากมีคำในพจนานุกรมจำนวนมาก จึงทำให้ไม่เป็นที่นิยม ดังตัวอย่างที่แสดงใน*รูปภาพที่ 2*



รูปภาพที่ 2 รูปภาพตัวอย่างการแปลงคำเป็นเวกเตอร์วันฮอต

4. คำฝังตัว (Word Embedding)

เป็นการแทนที่แต่ละคำด้วยเวกเตอร์คล้ายเวกเตอร์วันฮอต แต่แก้ไขปัญหาจำนวนมิติของเวกเตอร์วันฮอตที่มีจำนวนมากเกินไป โดยวิธีนี้สามารถกำหนดขนาดมิติของเวกเตอร์ได้เอง โดยการนิยามระยะห่างระหว่างเวกเตอร์ 2 ตัวด้วยค่าความแตกต่าง หาก 2 คำนี้แตกต่างกัน ค่าระยะห่างก็ควรจะมาก แต่หาก 2 คำนี้คล้ายกัน ค่าระยะห่างก็ควรจะน้อย โดยค่าความแตกต่างนี้ มีวิธีกำหนดได้หลากหลายแบบ ตัวอย่างที่นิยมเช่น คำที่สามารถสลับใช้ในบริบทเดียวกันได้ควรอยู่ใกล้กัน เช่นคำว่า “ชาย” กับ “หญิง” ควรใกล้กันมากกว่า “ชาย” กับ “ไป” เป็นต้น ดังตัวอย่างที่แสดงใน*รูปภาพที่ 3* โดยมากมักวิเคราะห์บริบทจากประโยคในฐานข้อมูลว่าคำนี้อยู่ตรงตำแหน่งไหนของประโยคบ้าง แล้วประโยคอื่นที่คล้ายกันที่คำอะไรที่แทนคำนี้ได้บ้าง เป็นต้น โดยเครื่องมือที่นิยมได้แก่ เวิร์ดทูเวก (word2vec) และโกลฟ (GloVe) เป็นต้น



รูปภาพที่ 3 รูปภาพตัวอย่างคำที่คล้ายคลึงกันเมื่อพิจารณาการสร้างเวกเตอร์แบบคำฝังตัว

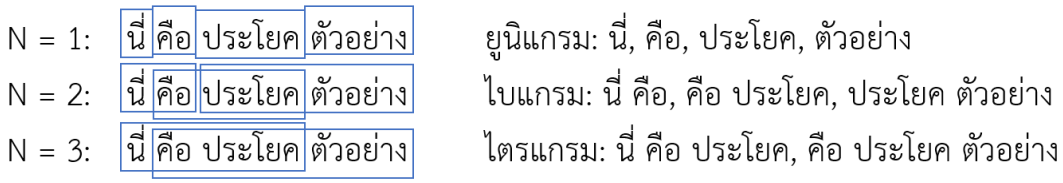
การสกัดคำสำคัญ (Keyword Extraction)

เป็นงานที่ต้องสามารถระบุ คำ หรือ วลี ที่มีความสำคัญจากเอกสารต่าง ๆ ได้อย่างอัตโนมัติ โดยการสกัดคำสำคัญมักจะถูกใช้สำหรับการค้นคืนสารสนเทศ (Information Retrieval, IR) รวมถึงถูกใช้ในงานต่าง ๆ อีกมากมายที่เกี่ยวข้องกับการประมวลผลภาษาธรรมชาติ (Natural Language Processing, NLP) เช่น การสรุปใจความสำคัญอัตโนมัติ และ การจัดการเอกสาร เป็นต้น [17]

โดยวิธีการในการสกัดคำสำคัญที่นิยมในปัจจุบัน ประกอบไปด้วยเรื่องที่เกี่ยวข้องดังนี้

1. เอ็นแกรม (n-Grams)

   เป็นอัลกอริทึมอย่างหนึ่ง ที่ใช้หลักการหั่นข้อความออกเป็นชุดตัวอักษร โดยมีจำนวนตัวอักษรในแต่ละชุดตามที่กำหนด [18] แต่เนื่องจากความสามารถของเครื่องมือตัดคำที่เพิ่มมากขึ้น ทำให้งานหลาย ๆ อย่างมักเริ่มที่หน่วยคำเลย จึงเป็นผลให้อัลกอริทึมนี้มักถูกเปลี่ยนเป็น การหั่นข้อความออกเป็นชุดของคำแทน ดังตัวอย่างใน*รูปภาพที่ 4* ซึ่งในงานวิจัยนี้ ไม่ได้ใช้งานเครื่องมือตัดคำ จึงจะใช้ความหมายดั้งเดิมของเอ็นแกรม แต่เพื่อความชัดเจนในด้านความหมาย งานวิจัยนี้จึงขอใช้คำว่า เอ็นแกรมแบบตัวอักษร (Character n-Grams) แทน เอ็นแกรม โดยแสดงตัวอย่างไว้ใน*ตารางที่ 1*



รูปภาพที่ 4 รูปภาพตัวอย่างของการใช้เอ็นแกรมแบบคำ

n = 1	n = 2	n = 3	n = 4	n = 5	n = 6
ก	กา	กาม	กามเ	กามเท	กามเทพ
า	าม	ามเ	ามเท	ามเทพ	
ม	มเ	มเท	มเทพ		
เ	เท	เทพ			
ท	ทพ				
พ					

ตารางที่ 1 ตารางแสดงตัวอย่างการใช้เอ็นแกรมแบบตัวอักษรโดยเลือก n ตั้งแต่ 1 ถึง 6

2. คำหยุด (Stop words)

คือคำที่ควรลบออกก่อนสกัดคำสำคัญ เนื่องจากคำหยุดนั้นปรากฏขึ้นบ่อยในเอกสาร แต่ไม่ค่อยแสดงถึงใจความสำคัญใด ๆ ของเอกสาร [19] ตัวอย่างเช่นคำว่า “การ” “ความ” “คือ” “ที่” “ซึ่ง” เป็นต้น ซึ่งคำเหล่านี้โดยส่วนใหญ่มักไม่ใช่ประเด็นที่สำคัญของเนื้อความ ดังตัวอย่างที่แสดงไว้ใน*ตารางที่ 2* นอกจากนี้คำบางคำ อาจนับเป็นคำหยุดหรือไม่นั้น มักขึ้นกับลักษณะงานที่ต้องการจะวิเคราะห์ด้วยเช่นกัน

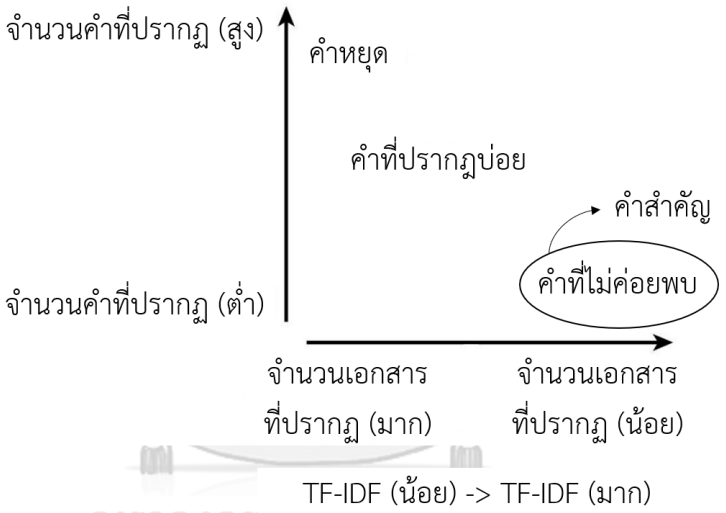
ตัวอย่างประโยคทั่วไป	ตัวอย่างประโยคที่ตัดคำหยุดออก
เว็บไซต์กูเกิ้ล เป็นเว็บไซต์ที่ใช้ค้นหาสิ่งต่าง ๆ จากคำสำคัญ	เว็บไซต์กูเกิ้ล, เว็บไซต์, ค้นหา, สิ่งต่าง ๆ, คำสำคัญ
การทำงานทำให้เหนื่อยใช่ไหม	ทำงาน, เหนื่อย, ใช่ไหม
ฉันชอบการอ่าน ฉันจึงอ่าน	ชอบ, อ่าน, อ่าน

ตารางที่ 2 ตารางแสดงตัวอย่างของประโยคทั่วไปในภาษาอังกฤษ และประโยคที่ตัดคำหยุดออก

3. การกำหนดน้ำหนักคำ (Term Weighting)

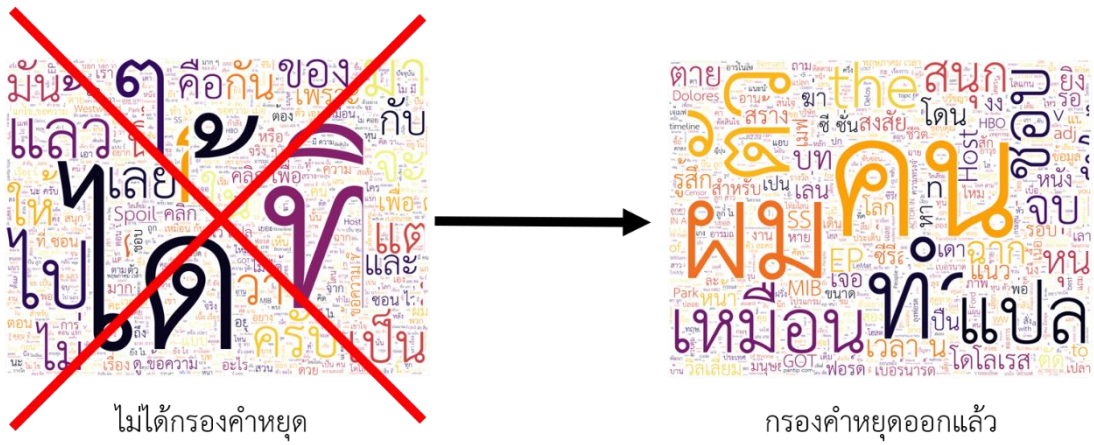
การกำหนดน้ำหนักคำนั้นมักจะถูกใช้ในการหาคำที่ปรากฏขึ้นบ่อย หรือ มีความโดดเด่นกว่าคำอื่น ๆ ในเอกสาร [19] โดยวิธีที่มักจะใช้ในการกำหนดน้ำหนักคำ ได้แก่

- *การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (Term Frequency - Inverse Document Frequency, TF-IDF)*
  โดยวิธีนี้หากจะนำมาหาคำสำคัญ มักจำเป็นต้องตัดคำ และใช้คู่กับอัลกอริทึมเอ็นแกรม เพื่อหาความถี่ของแกรม แล้วจำแนกแกรมออกมาเป็นคำสำคัญอีกที ซึ่งผลจากวิธีนี้จะทำให้คำที่น้ำหนักน้อยจะเป็นคำหยุด คำที่น้ำหนักกลาง ๆ จะเป็นคำทั่วไป และ คำที่น้ำหนักสูงจะเป็นคำสำคัญ ดังตัวอย่างที่แสดงใน*รูปภาพที่ 5* แต่ข้อจำกัดของวิธีนี้คือ มีการใช้หน่วยความจำแปรตามคำและจำนวนเอกสาร และการค้นหาว่ามีคำนั้น ๆ ในเอกสาร หากเอกสารมีคำจำนวนมาก ก็จะใช้เวลาประมวลผลนาน



รูปภาพที่ 5 รูปภาพตัวอย่างกราฟที่แสดงความสัมพันธ์ระหว่างความถี่กับความมีคุณค่า (คำที่ปรากฏในไม่กี่เอกสาร) ซึ่งค่า TF-IDF สะท้อนถึงคำที่มีความสำคัญได้

- *การพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (Term Frequency, TF)*
  เป็นวิธีที่เรียบง่ายและประมวลผลได้เร็วกว่าวิธีแรกมาก โดยการไม่คำนวณในส่วนของการหาจำนวนเอกสารที่มีคำนั้น ๆ ปรากฏอยู่ ซึ่งทำให้จำเป็นต้องมีการนำคำหยุดออกจากเอกสารก่อน ซึ่งหากมีฐานข้อมูลคำหยุดสำหรับงานทางด้านนั้นอยู่แล้ว วิธีนี้ก็สามารถสกัดคำสำคัญจากคำที่มีความถี่สูงได้เลย และใช้เวลาประมวลผลรวดเร็วกว่ามาก ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 6*



รูปภาพที่ 6 รูปภาพตัวอย่างการใช้ค่า TF ในการสกัดคำสำคัญ โดยลบคำหยุดออก

การจำแนกข้อมูล (Classification)

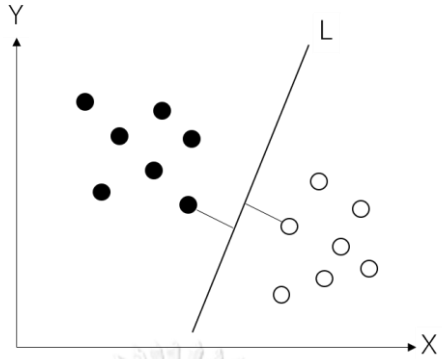
เมื่อเรามีการเก็บข้อมูลปริมาณมาก เราสามารถจำแนกข้อมูลของเราออกเป็นหมวดหมู่ต่าง ๆ ตามสิ่งที่เราต้องการนำไปใช้งาน เช่นหากเราต้องการหาภาพของสัตว์ เราจึงจำเป็นต้องจำแนกภาพที่เรามีก่อน เพื่อแยกภาพที่มีสัตว์และภาพที่ไม่มีสัตว์ออกจากกัน หรือบางครั้ง เราอาจต้องการจำแนกภาพสัตว์ ว่าเป็นสัตว์ชนิดใด ซึ่งเราอาจกำหนดหมวดหมู่ของภาพเป็น ภาพสุนัข ภาพแมว ภาพปลา และภาพอื่น ๆ เป็นต้น ในงานทางด้านการวิเคราะห์ภาษาธรรมชาติเอง ก็มีงานที่จำเป็นต้องจำแนกข้อมูลเช่นกัน เช่น การจำแนกอารมณ์จากข้อความ การจัดประเภทของจดหมายอิเล็กทรอนิกส์ เป็นต้น รวมถึงการสกัดคำสำคัญก็เป็นหนึ่งในงานด้านการจำแนกข้อมูลเช่นกัน

การจำแนกข้อมูลถูกจัดเป็นงานทางด้านการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Learning) [20] เนื่องจากการสร้างแบบจำลองที่สามารถจำแนกข้อมูลได้ จำเป็นต้องมีชุดข้อมูลฝึกสอน (Training set) ที่เป็นข้อมูลที่เป็นผลเฉลยของแต่ละหมวดหมู่ สำหรับสอนเครื่องจักรให้สามารถเข้าใจหมวดหมู่ต่าง ๆ ก่อน โดยจำนวนข้อมูล ที่ใช้สอนมีผลต่อความแม่นยำของแบบจำลอง ยิ่งมีข้อมูลเยอะและหลากหลาย แบบจำลองที่ได้จะยิ่งมีประสิทธิภาพ โดยมีวิธีสร้างแบบจำลองที่นิยมสำหรับการจำแนกข้อมูลดังนี้

1. แบบจำลองจากตัวจำแนกเชิงเส้น (Linear Classifiers) [21]

   เป็นแบบจำลองการจำแนกข้อมูลบนข้อมูล 2 กลุ่ม โดยใช้เส้นตรงในการแบ่ง โดยแบบจำลองนี้จะพยายามหาเส้นตรงที่ใช้แบ่งกลุ่มข้อมูลจากชุดข้อมูลฝึกสอน เมื่อได้เส้นตรงที่ใช้แบ่งกลุ่มข้อมูลแล้ว เราก็สามารถนำเส้นตรงนั้นมาทำนายข้อมูลใหม่ได้ ว่าอยู่ในกลุ่มใด ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 7* โดยมีแบบจำลองที่นิยมดังนี้ แบบจำลอง

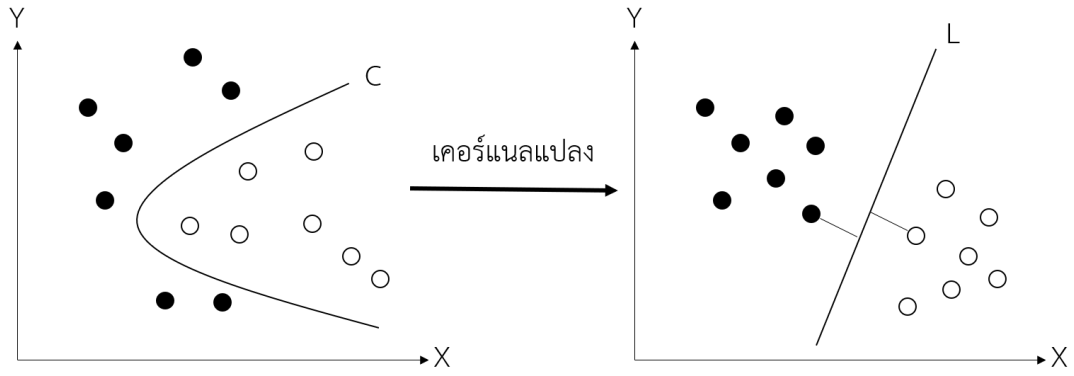
โลจิสติกส์ถดถอย (Logistic Regression) แบบจำลองตัวจำแนกเบย์อย่างง่าย (Naïve Bayes Classifier) และแบบจำลองเพอร์เซ็ปตรอน (Perceptron) เป็นต้น



รูปภาพที่ 7 รูปภาพตัวอย่างการแบ่งข้อมูลออกเป็น 2 กลุ่มโดยใช้เส้นตรง

2. แบบจำลองจากซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) [22]

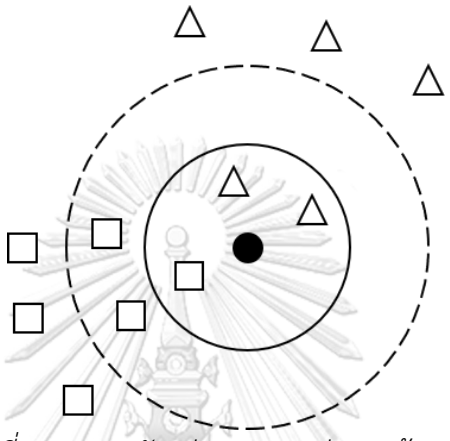
เป็นวิธีที่พัฒนามาจากตัวจำแนกเชิงเส้น โดยแบบจำลองนี้จะใช้เวกเตอร์ในการแบ่งข้อมูล ซึ่งหากเรามีข้อมูล n มิติ เวกเตอร์ที่ใช้ในการแบ่งข้อมูลก็จะมีขนาด n-1 มิติ และยังแก้ข้อจำกัดของแบบจำลองจากตัวจำแนกเชิงเส้น ที่ไม่สามารถแบ่งข้อมูลโดยตัวแบ่งไม่เป็นเส้นตรงได้ ซึ่งหากข้อมูลฝึกสอนไม่สามารถแบ่งได้โดยใช้เส้นตรง ก็จะได้ความแม่นยำของแบบจำลองต่ำนั่นเอง โดยในแบบจำลองนี้มีการใช้เคอร์แนลทริก (Kernel Trick) ที่พยายามหาฟังก์ชันที่ใช้แปลงข้อมูลก่อน เพื่อให้สามารถใช้เวกเตอร์เส้นตรงแบ่งข้อมูลที่ถูกแปลงแล้วได้ ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 8* ทำให้แบบจำลองนี้สามารถแก้ข้อจำกัดดังกล่าวได้ และยังมีการนำไปประยุกต์ใช้กับข้อมูลที่เราต้องการจำแนกมากกว่า 2 กลุ่ม โดยมีการใช้จำนวนเวกเตอร์เพิ่มขึ้นเพื่อแบ่งข้อมูลออกเป็นหลายกลุ่มอีกด้วย



รูปภาพที่ 8 รูปภาพตัวอย่างการใช้เคอร์แนลทริกแปลงข้อมูลทางฝั่งซ้าย ที่ไม่สามารถใช้เส้นตรงในการแบ่งได้ ไปเป็นข้อมูลทางฝั่งขวา ที่สามารถหาเวกเตอร์เส้นตรงมาแบ่งได้

3. แบบจำลองจากเพื่อนบ้านใกล้กัน k ตัว (K-Nearest Neighbor: KNN) [23]

แบบจำลองนี้จะใช้ชุดข้อมูลฝึกสอนในการทำนายข้อมูลใหม่ โดยดูว่าข้อมูลใหม่อยู่ใกล้กับข้อมูลฝึกสอน k ตัวใดมากที่สุด แล้วจะพิจารณาว่าในบรรดาข้อมูล k ตัวนั้นมีจำนวนข้อมูลในหมวดหมู่ใดมากที่สุด ก็จะได้ผลทำนายว่าข้อมูลใหม่เป็นข้อมูลในหมวดหมู่ดังกล่าว ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 9*



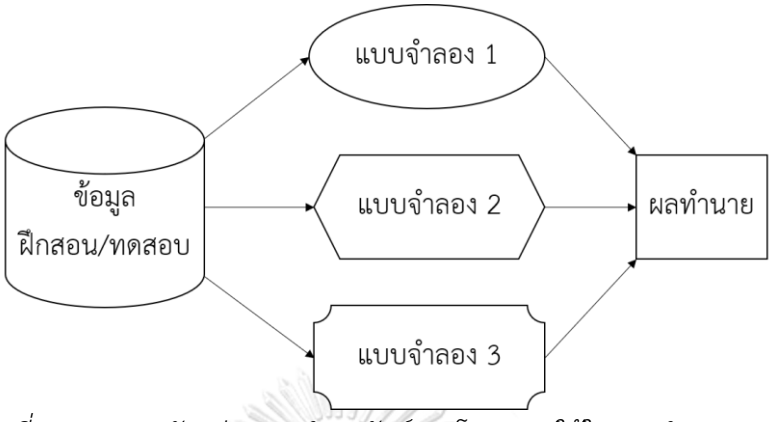
รูปภาพที่ 9 รูปภาพตัวอย่างการหากลุ่มของข้อมูลจุดสีดำ
หากเราเลือก k = 3 ข้อมูลจุดสีดำจะเป็นข้อมูลในกลุ่มสามเหลี่ยม
แต่หากเราเลือก k = 5 ข้อมูลจุดสีเขียวจะเป็นข้อมูลในกลุ่มสี่เหลี่ยม

ซึ่งแบบจำลองนี้ เราจำเป็นต้องนิยามระยะทางระหว่างข้อมูลใด ๆ เพื่อให้สามารถบอกได้ว่าข้อมูลแต่ละตัว อยู่ห่างกันเท่าใดนั่นเอง ซึ่งข้อเสียหลักของแบบจำลองนี้คือขนาดของแบบจำลองที่ได้จะมีขนาดแปรตามขนาดของชุดข้อมูลฝึกสอนที่ใช้ในการสร้างแบบจำลอง ทำให้ไม่เหมาะที่จะใช้กับชุดข้อมูลฝึกสอนปริมาณมาก ซึ่งแตกต่างจากซัพพอร์ตเวกเตอร์แมชชีนที่ขนาดของแบบจำลองที่ได้แปรตามจำนวนเวกเตอร์ที่ใช้แบ่งกลุ่มข้อมูลของแบบจำลอง ซึ่งทำให้ได้แบบจำลองที่มีขนาดเล็กกว่ามาก

4. แบบจำลองจากการเร่งความสามารถ (Boosting) [24]

เป็นแบบจำลองที่เกิดขึ้นโดยมีแนวคิดว่า เราสามารถพัฒนาแบบจำลองหลาย ๆ แบบจำลองที่ไม่ค่อยแม่นยำได้หรือไม่ โดยถ้านำพวกมันมาช่วยกันทำนายข้อมูล เราน่าจะได้แบบจำลองที่แม่นยำขึ้น โดยแบบจำลองนี้จะนำผลลัพธ์จากแบบจำลองอื่น ๆ มารวมกัน โดยอาจใช้ค่าเฉลี่ย หรือการโหวตเสียงข้างมากในการสรุปผลลัพธ์ โดยมีตัวถ่วงน้ำหนักว่าแบบจำลองใดน่าเชื่อถือกว่ากัน และพยายามปรับตัวถ่วงน้ำหนักดังกล่าวตามชุดข้อมูล

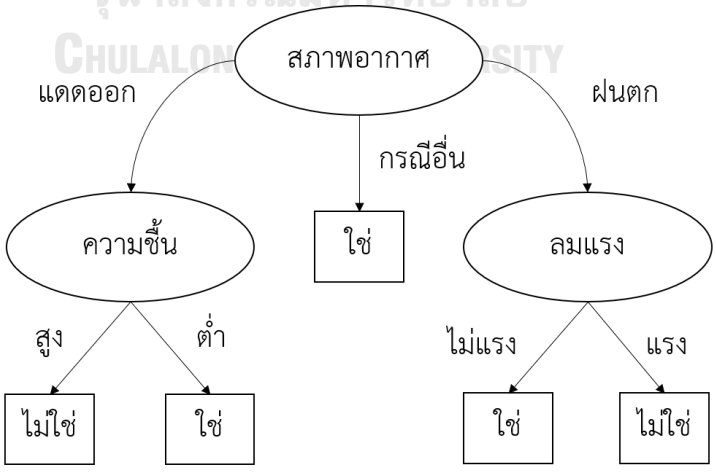
ฝึกสอน ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 10* ทำให้แบบจำลองนี้มีความซับซ้อนขึ้นแต่ได้ความแม่นยำสูงกว่าการใช้แบบจำลองที่ไม่แม่นยำเพียงแบบจำลองเดียวนั่นเอง



รูปภาพที่ 10 รูปภาพตัวอย่างการนำผลลัพธ์จากโมเดลมาใช้ในการทำนายผลร่วมกัน

5. แบบจำลองจากต้นไม้ตัดสินใจ (Decision Tree) [25]

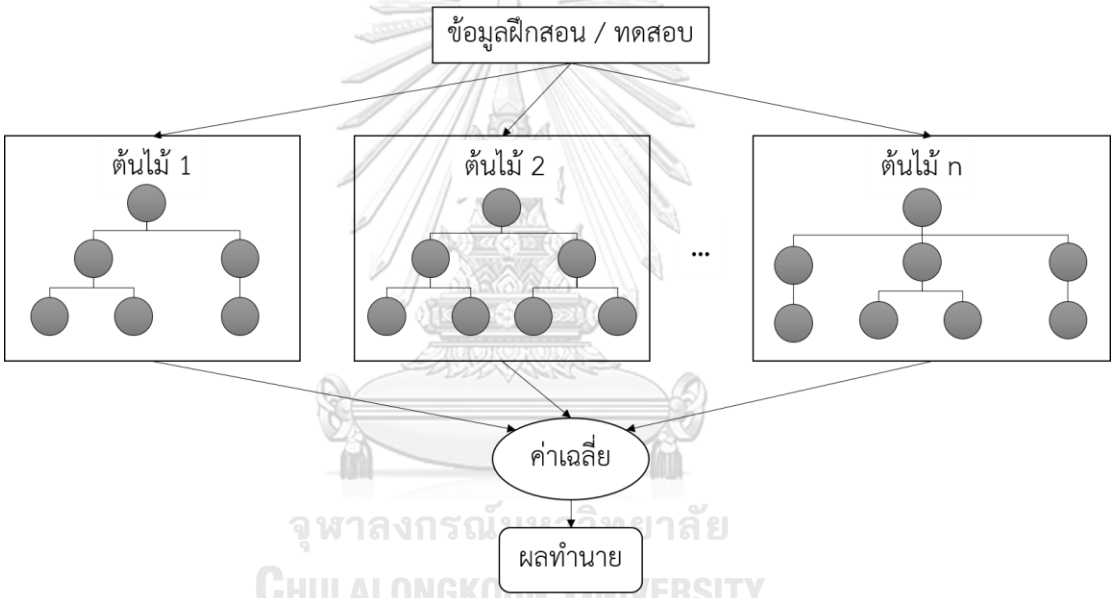
เป็นแบบจำลองที่อาศัยการตั้งเงื่อนไขให้กับข้อมูล เพื่อแบ่งข้อมูลออกเป็นกลุ่มย่อย ๆ ตามเงื่อนไข โดยแต่ละเงื่อนไขที่แบบจำลองตั้งจะพยายามให้แบ่งชุดข้อมูลฝึกสอนออกเป็น 2 กลุ่มที่สมาชิกในกลุ่มตรงกับกลุ่มของผลเฉลยมากที่สุด เพื่อลดจำนวนเงื่อนไขที่ต้องใช้ในการแบ่งข้อมูล โดยหากข้อมูลทั้งกลุ่มย่อยตรงกับกลุ่มของผลเฉลยแล้ว ก็จะหยุดสร้างเงื่อนไขที่ใช้แบ่งกลุ่ม ทำให้เราได้แบบจำลองที่มีลักษณะเป็นต้นไม้ของเงื่อนไขต่าง ๆ สำหรับจำแนกข้อมูล ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 11*



รูปภาพที่ 11 รูปภาพตัวอย่างการแบ่งข้อมูลออกเป็น 2 กลุ่มโดยการใช้ต้นไม้ตัดสินใจ

6. แบบจำลองจากป่าสุ่ม (Random Forests) [26]

เป็นแบบจำลองที่พยายามลดความพอดีเกินไป (Overfit) กับชุดข้อมูลฝึกสอน [27] ของแบบจำลองต้นไม้ตัดสินใจ เนื่องจากแบบจำลองต้นไม้ตัดสินใจมักได้ทำนายได้แม่นยำมากกับชุดข้อมูลฝึกสอน แต่ทำให้เงื่อนไขของต้นไม้เยอะมากเช่นกัน ซึ่งไม่เหมาะสำหรับไปใช้ทำนายข้อมูลใหม่ ๆ จึงทำให้เราพยายามตัดต้นไม้เหล่านั้นให้เตี้ยลง โดยการกำหนดชั้นความลึกสูงสุดของต้นไม้ แล้วสร้างต้นไม้เหล่านั้นหลายต้นแทน โดยผลลัพธ์ที่ได้จากแต่ละต้นจะถูกโหวตแบบถ่วงน้ำหนักคล้ายแบบจำลองเร่งความสามารถ จึงทำให้แบบจำลองมีลักษณะคล้ายป่า คือประกอบจากต้นไม้หลาย ๆ ต้นรวมกัน ซึ่งทำให้แบบจำลองมีความแม่นยำมากขึ้น และไม่พอดีเกินไปกับข้อมูลฝึกสอน ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 12*

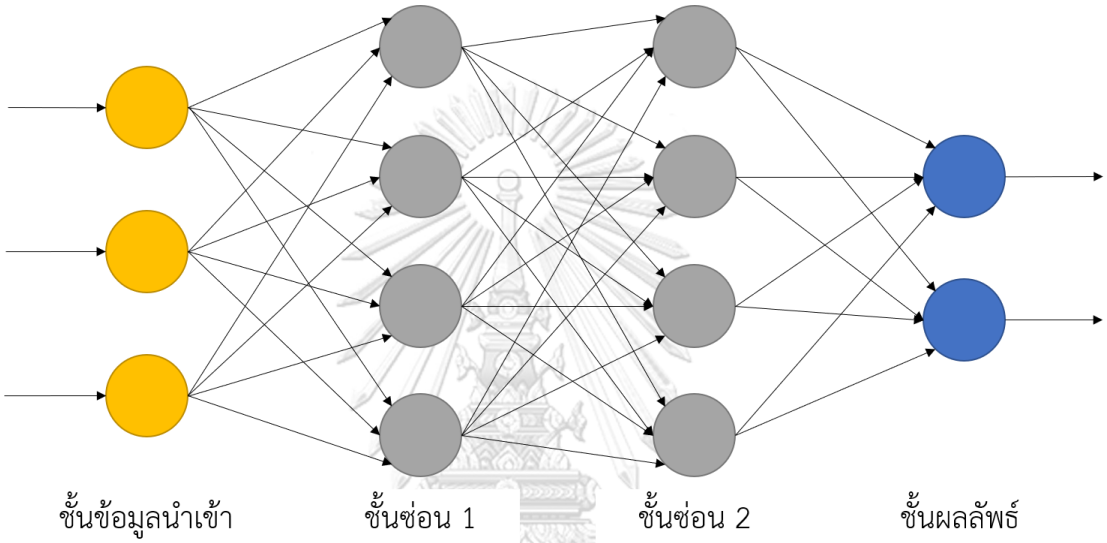


รูปภาพที่ 12 รูปภาพตัวอย่างแบบจำลองจากป่าสุ่ม
โดยการนำผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจแต่ละต้นมาหาค่าเฉลี่ยเพื่อเป็นผลทำนายสุดท้าย

7. แบบจำลองจากโครงข่ายประสาทเทียม (Neural Networks) [28]

แบบจำลองจากโครงข่ายประสาทเทียมนั้นสามารถนำมาใช้ในการจำแนกข้อมูลได้ โดยแบบจำลองดังกล่าวเลียนแบบการทำงานของเส้นใยประสาทของมนุษย์ โดยมีการแบ่งออกเป็นชั้น ๆ ในแต่ละชั้นประกอบไปด้วยโหนดของเพอร์เซ็ปตรอนที่ถูกถ่วงน้ำหนัก โดยชั้นแรกสุดเป็นชั้นของข้อมูลนำเข้า ที่ต้องการจะจำแนก และข้อมูลดังกล่าวจะถูกส่งไปยังชั้นถัดไปที่มีอาจทำหน้าที่ได้หลายอย่าง เช่น บีบอัดข้อมูล กรองข้อมูล หาลักษณะเฉพาะของข้อมูล สรุปข้อมูล เป็นต้น โดยชั้นสุดท้ายของโครงข่ายประสาทจะเป็นชั้นผลลัพธ์ที่ใช้

บอกว่าข้อมูลเป็นหมวดหมู่ใด โดยแบบจำลองนี้จะเรียนรู้โดยการปรับตัวถ่วงน้ำหนักของแต่ละชั้นตามความผิดพลาดที่เกิดขึ้นจากการทำนายชุดข้อมูลฝึกสอน จนได้แบบจำลองที่มีความแม่นยำเพียงพอ จึงหยุดการฝึกสอน โดยแสดงตัวอย่างแบบจำลองไว้ใน*รูปภาพที่ 13* โดยข้อเสียของแบบจำลองนี้ คือต้องการข้อมูลฝึกสอนปริมาณมาก ถึงจะทำให้แบบจำลองดังกล่าวเข้าใจลักษณะของข้อมูล และทำนายข้อมูลได้แม่นยำ เนื่องจากแบบจำลองมีความซับซ้อนและยืดหยุ่นสูง



รูปภาพที่ 13 รูปภาพตัวอย่างแบบจำลองโครงข่ายประสาทเทียมที่มี 4 ชั้น

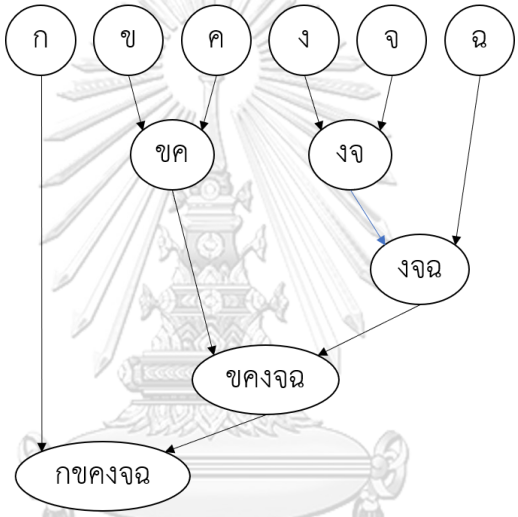
การแบ่งกลุ่มข้อมูล (Clustering)

ข้อมูลบางอย่างมีลักษณะเฉพาะตัว ซึ่งเราสามารถแบ่งกลุ่มข้อมูลตามลักษณะเฉพาะตัวเหล่านั้นได้ ซึ่งบางครั้งอาจทำให้เราเข้าใจข้อมูลที่เรามีมากยิ่งขึ้น เช่น ข้อมูลที่ได้รับจากเซ็นเซอร์ หากเรานำมาแบ่งกลุ่มข้อมูลดู อาจพบว่ามีข้อมูลบางค่ามีลักษณะแตกต่างกับข้อมูลอื่น ๆ ซึ่งทำให้เราสามารถพิจารณาข้อมูลเหล่านั้นได้ง่ายขึ้น โดยการแบ่งกลุ่มข้อมูลนี้สามารถนำมาประยุกต์กับงานทางด้านการจำแนกข้อมูลที่เราไม่รู้จำนวนหมวดหมู่ เช่นการจำแนกประเภทของผู้ใช้งาน ในตอนแรกเราอาจจะแบ่งผู้ใช้งานออกเป็นกลุ่ม ๆ ตามลักษณะของการใช้งาน แล้วจึงคิดชื่อกลุ่มของผู้ใช้งานขึ้นมา เพื่อใช้ในการจำแนกผู้ใช้งานในครั้งถัดไป เป็นต้น

การแบ่งกลุ่มข้อมูลจึงแตกต่างกับการจำแนกข้อมูลในด้านการสร้างแบบจำลอง เนื่องจากไม่จำเป็นต้องมีข้อมูลที่ถูกต้องของแต่ละกลุ่มก่อน ทำให้การแบ่งกลุ่มข้อมูลถูกจัดเป็นงานทางด้านการเรียนรู้ของเครื่องแบบไม่มีผู้สอน (Unsupervised Learning) [29] ซึ่งทำให้ไม่มีวิธีตายตัวในการแบ่งกลุ่ม [30] โดยมีวิธีที่นิยมสำหรับการแบ่งกลุ่มข้อมูลดังนี้

1. การแบ่งกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering) [31]

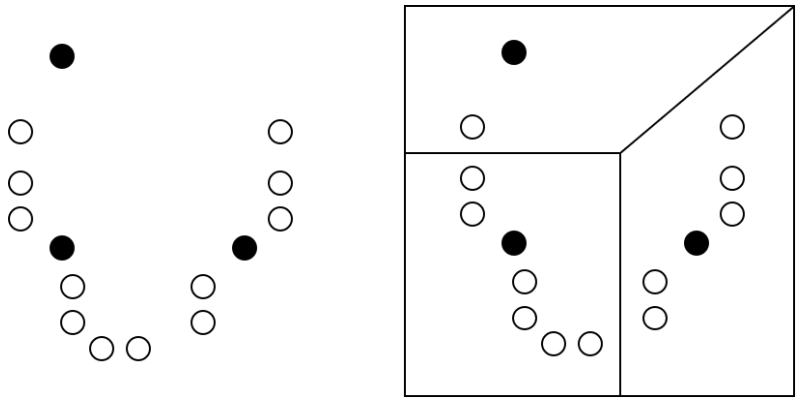
เป็นวิธีที่ใช้สำหรับการแบ่งข้อมูลออกเป็นลำดับชั้น กล่าวคือหากเรามีข้อมูล n ตัว ในชั้นแรกสุด เราจะแบ่งข้อมูลออกเป็น n กลุ่ม และในชั้นสุดท้ายเราจะรวมข้อมูลทุกกลุ่มเข้าด้วยกันเหลือ 1 กลุ่ม โดยในแต่ละชั้นย่อย เราจะพยายามรวมข้อมูลทีละ 2 กลุ่มเข้าด้วยกันโดยมีวิธีการรวมได้หลายวิธี เช่น ใช้ระยะห่างของจุดเซนทรอยด์ (Centroid) ของแต่ละกลุ่มที่ใกล้กันมากที่สุด เป็นต้น โดยวิธีนี้ เราสามารถเลือกได้ว่าเราอยากแบ่งข้อมูลออกเป็นกี่กลุ่ม เราก็หยุดการรวมกลุ่มที่ระดับชั้นที่มีจำนวนกลุ่มเท่ากับที่เราต้องการ ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 14*



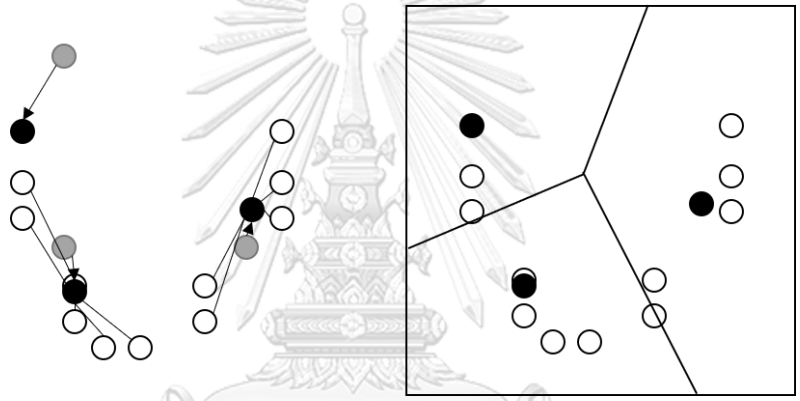
รูปภาพที่ 14 รูปภาพตัวอย่างลักษณะการแบ่งกลุ่มข้อมูลแบบลำดับชั้น

2. การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering) [32]

เป็นวิธีหนึ่งที่ใช้แบ่งกลุ่มข้อมูลออกเป็นกลุ่มย่อย ๆ โดยเราต้องกำหนดจำนวนกลุ่มที่ต้องการจะแบ่ง โดยการแบ่งกลุ่มข้อมูลแบบเคมีนนั้น จะใช้จุดเซนทรอยด์เป็นตัวแทนของกลุ่ม โดยเริ่มแรก เราจะสุ่มหยิบข้อมูลมา k ตัว แทนตัวแทนของกลุ่มและใช้เป็นจุดเซนทรอยด์ของกลุ่มตอนเริ่มต้นด้วย หลังจากนั้นเราจะพิจารณาข้อมูลแต่ละตัว ว่าอยู่ใกล้ตัวแทนกลุ่มไหนมากที่สุด ก็จะจัดให้อยู่กลุ่มนั้น เมื่อข้อมูลถูกจัดกลุ่มเสร็จแล้ว เราก็จะหาจุดเซนทรอยด์ของกลุ่มใหม่ และจัดข้อมูลลงกลุ่มใหม่ไปเรื่อย ๆ ซึ่งข้อมูลแต่ละตัวก็จะถูกเปลี่ยนกลุ่มไปเรื่อย ๆ จนกว่าจะไม่มีข้อมูลในกลุ่มใดถูกเปลี่ยนกลุ่มใหม่อีกแล้ว ก็จะจบกระบวนการ ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 15 และ 16* ซึ่งทำให้การสุ่มหยิบข้อมูลที่ใช้แทนตัวแทนของกลุ่มในขั้นตอนแรกมีผลต่อผลลัพธ์ที่ได้หลังจบกระบวนการเช่นกัน



รูปภาพที่ 15 รูปภาพตัวอย่างการจัดแบ่งกลุ่มข้อมูลแบบเคมีน โดยกำหนดให้ k = 3



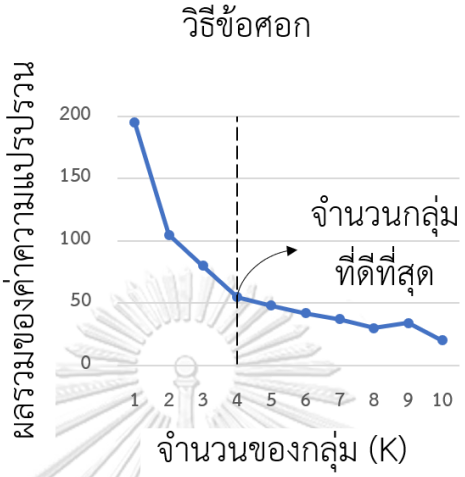
รูปภาพที่ 16 รูปภาพตัวอย่างการจัดแบ่งกลุ่มใหม่ของเคมีน หลังพบว่าจุดเซนทรอยด์เปลี่ยนแปลง

บางครั้ง เราก็ไม่รู้ว่าข้อมูลที่เรามีควรแบ่งออกเป็นกี่กลุ่ม ซึ่งในปัจจุบันก็มีวิธีที่ช่วยในการหาจำนวนกลุ่มของข้อมูล ซึ่งหนึ่งในวิธียอดนิยมก็คือ วิธีข้อศอก (Elbow Method)

- *วิธีข้อศอก (Elbow Method) [33]*

เป็นอัลกอริทึมหนึ่งที่ใช้ในการหาจำนวนกลุ่ม สำหรับการแบ่งกลุ่มข้อมูลแบบเคมีน เนื่องจากงานบางอย่าง เราไม่รู้ว่าข้อมูลควรถูกแบ่งออกเป็นกี่กลุ่มย่อย อัลกอริทึมนี้จึงทดลองแบ่งกลุ่มจาก 1 กลุ่ม 2 กลุ่ม 3 กลุ่ม ไปเรื่อย ๆ แล้วเริ่มสังเกตค่าความแปรปรวน เนื่องจากหากยังมีกลุ่มใดกลุ่มหนึ่ง สามารถแบ่งเป็นกลุ่มย่อย ๆ ได้อย่างมีนัยสำคัญ ค่าความแปรปรวนก็จะลดลงมาก แต่หากทำการแบ่งกลุ่มเพิ่มแล้ว กลุ่มย่อยที่แบ่งออกมานั้นมีค่าความแปรปรวนลดลงเพียงเล็กน้อย หมายความว่ากลุ่มย่อยดังกล่าว ไม่ได้มีนัยสำคัญ เราจึงควรยุบรวมกลุ่มนั้น กลับเป็นกลุ่มใหญ่มากกว่า ซึ่งจังหวะที่ค่าความแปรปรวนเริ่มเปลี่ยนจากลดลงมากมาเป็นลดลงเพียงเล็กน้อย จะทำให้กราฟค่าความแปรปรวนมีลักษณะหักคล้ายข้อศอก

และจุดหักนั้นเอง เป็นจุดที่แทนจำนวนกลุ่มที่เหมาะสม ดังตัวอย่างที่แสดงไว้ใน *รูปภาพที่ 17*



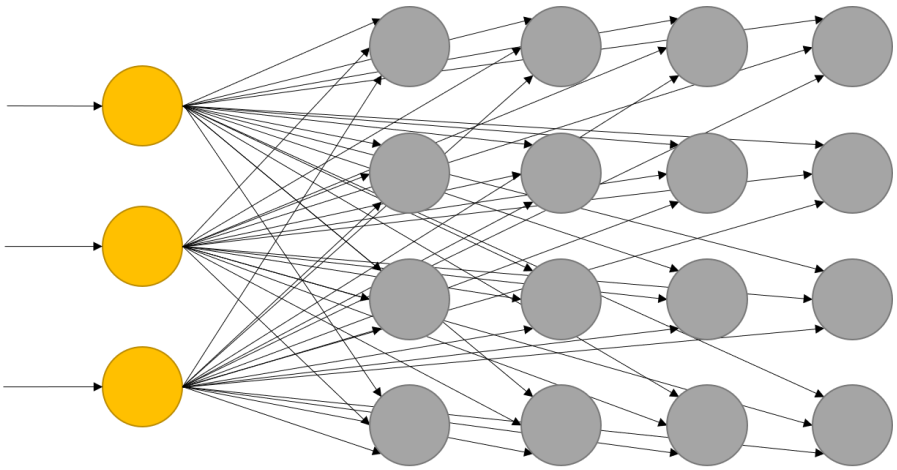
รูปภาพที่ 17 กราฟแสดงความสัมพันธ์ระหว่างจำนวนกลุ่มกับค่าความแปรปรวน เพื่อเลือกค่า k

3. การแบ่งกลุ่มจากโครงข่ายประสาทเทียม (Neural Networks) [29]

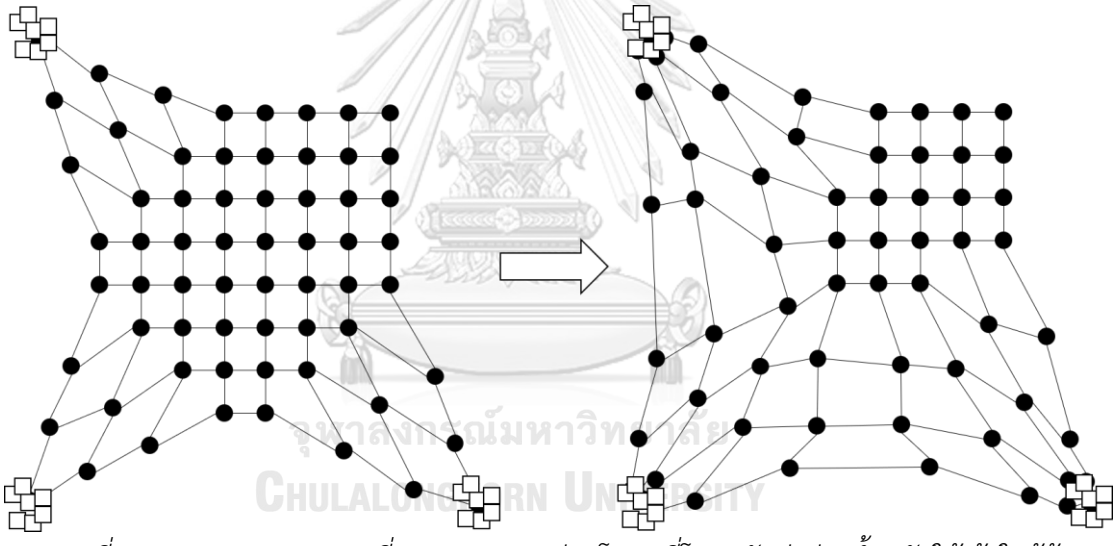
โดยปกติโครงข่ายประสาทเทียมมักนิยมนำมาใช้กับการจำแนกข้อมูล เนื่องจากการฝึกสอนโครงข่ายประสาทเทียมจำเป็นต้องมีผลเฉลยของข้อมูล แต่เราสามารถนำโครงข่ายประสาทเทียมมาใช้กับการแบ่งกลุ่มข้อมูลได้เช่นกัน เนื่องจากโครงข่ายประสาทเทียมมีความสามารถในการบีบอัดข้อมูลเพื่อลดจำนวนมิติของข้อมูล เราจึงนำความสามารถดังกล่าวมาประยุกต์ใช้ โดยวิธีที่นิยมได้แก่

- *การแปลงข้อมูลแบบจัดการตนเอง (Self-Organizing Map: SOM) [34]*

วิธีนี้จะสร้างโครงข่ายประสาทเทียมแบบตารางขึ้นมาในชั้นของข้อมูลผลลัพธ์ โดยลักษณะของโครงข่ายประสาทเทียมคือสามารถปรับค่าถ่วงน้ำหนักของแต่ละโหนดได้ ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 18* ผลทำให้ตารางดังกล่าวสามารถยืดหยุ่นได้ ต่อมาเราก็พยายามสอนโหนดในตารางดังกล่าวให้ลู่เข้าใกล้ข้อมูลที่มี โดยโหนดที่อยู่ใกล้กับข้อมูลจะโดนดึงเข้าหาข้อมูล โหนดที่อยู่ไกลก็จะถูกปรับแรงดึงให้เข้าใกล้ข้อมูลน้อยกว่าโหนดที่อยู่ใกล้ หลังจากการปรับค่าถ่วงน้ำหนัก เราสามารถใช้ความหนาแน่นของโหนดแทนแต่ละกลุ่มของข้อมูลได้ ดังตัวอย่างที่แสดงไว้ใน *รูปภาพที่ 19*



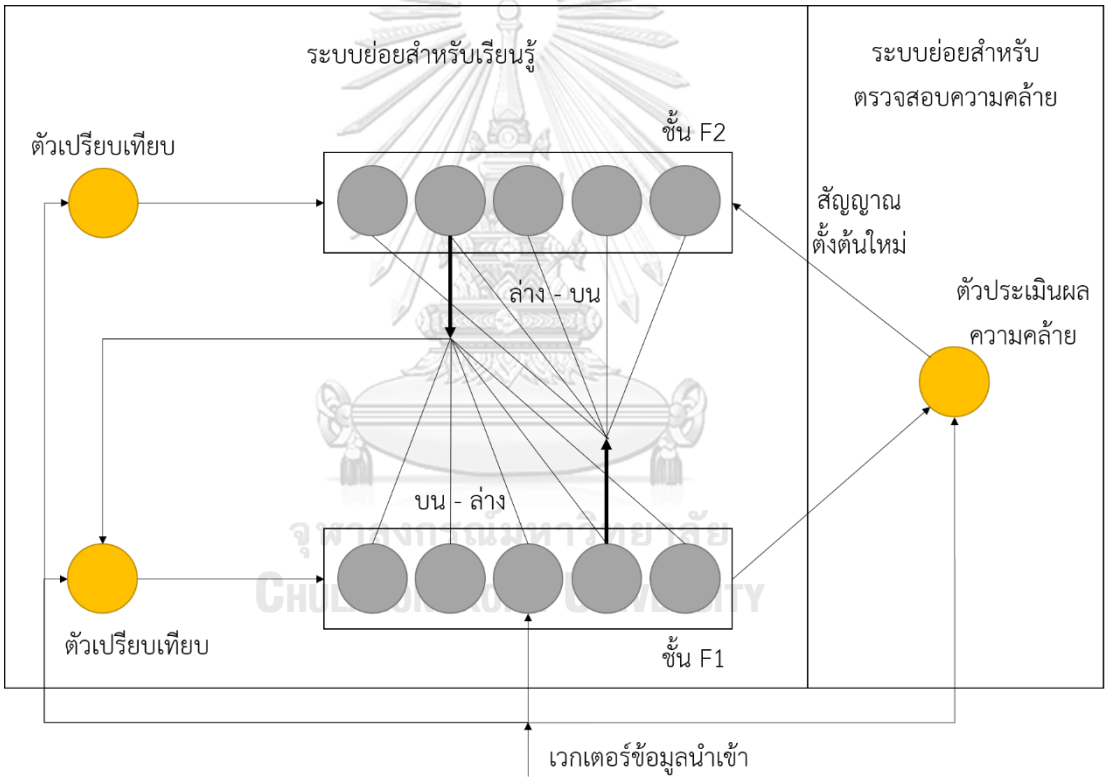
รูปภาพที่ 18 รูปภาพตัวอย่างโครงข่ายประสาทเทียมแบบ SOM



รูปภาพที่ 19 รูปภาพแสดงการเปลี่ยนแปลงของแต่ละโหนดที่โดนปรับค่าถ่วงน้ำหนักให้เข้าใกล้ข้อมูล

- *ทฤษฎีการสั่นพ้องแบบปรับได้ (Adaptive Resonance Theory: ART) [35]*

วิธีดังกล่าวเหมาะข้อมูลที่ต้องการจัดกลุ่มโดยไม่สนใจจำนวนกลุ่ม แต่สนใจว่าข้อมูลนี้คล้ายกับข้อมูลใด กล่าวคือวิธีนี้จะสร้างโครงข่ายประสาทเทียมที่ใส่ข้อมูล 2 ตัวเข้าไป แล้วให้โครงข่ายประสาทเทียมพยายามเปรียบเทียบเพื่อตอบว่าข้อมูลทั้ง 2 เหมือนหรือไม่เหมือนกัน โดยเราอาจฝึกสอนโมเดลโดยการนำข้อมูลเดียวกันมาใส่เพื่อบอกผลเฉลยว่าข้อมูลดังกล่าวเหมือนกัน และนำข้อมูลคนละตัวมาใส่เพื่อบอกว่าข้อมูลดังกล่าวแตกต่างกัน โดยเราสามารถปรับค่าความผิดพลาดที่ยอมรับได้ เพื่อให้เราแบ่งกลุ่มข้อมูลที่คล้ายกันได้ วิธีนี้จึงมักถูกนำไปใช้กับงานการตรวจจับใบหน้า โดยแสดงตัวอย่างของแบบจำลองไว้ใน*รูปภาพที่ 20*



รูปภาพที่ 20 รูปภาพตัวอย่างโครงข่ายประสาทเทียมแบบ ART

งานวิจัยที่เกี่ยวข้อง

วิทยานิพนธ์ฉบับนี้ มีความเกี่ยวข้องกับงานวิจัยในหัวข้อต่าง ๆ ซึ่งแสดงให้เห็นถึงปัญหาและข้อจำกัดต่าง ๆ รวมถึงบางงานวิจัยสามารถนำมาต่อยอดและพัฒนาให้สามารถแก้ไขข้อจำกัดที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ได้ โดยมีหัวข้อดังนี้

1. การตัดคำภาษาไทย
2. การจำแนกข้อมูลที่เกี่ยวข้องกับข้อความ
3. คำหยุด
4. การสกัดคำสำคัญ

การตัดคำภาษาไทย

งานวิจัยที่เกี่ยวกับการตัดคำภาษาไทยมีอยู่ปริมาณหนึ่ง ในที่นี้ขอยกตัวอย่างงานวิจัยที่มีความเกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ซึ่งเป็นงานวิจัยที่ทดลองใช้กระบวนการต่าง ๆ ในการตัดคำ ได้แก่ การตัดคำโดยใช้พจนานุกรม และการตัดคำโดยใช้วิธีการเรียนรู้ของเครื่อง ดังต่อไปนี้

1. เลิร์นเล็กโต (LearnLexTo) เครื่องมือตัดคำภาษาไทยโดยใช้การเรียนรู้ของเครื่อง (Machine Learning-based) สำหรับการทำดรรชนีข้อความภาษาไทย

   งานวิจัยของ C. Haruechaiyasak และคณะ [36] ได้ศึกษาวิธีการตัดคำภาษาไทยโดยใช้อัลกอริทึมคอนดิชันนอลแรนดอมฟิลด์ส (Conditional Random Fields: CRF) ซึ่งเป็นหนึ่งในแบบจำลองทางสถิติที่ให้ผลดีสำหรับการฝึกฝนแบบจำลองด้านภาษาธรรมชาติในเวลานั้น โดยข้อเสียของวิธีการเรียนรู้ของเครื่องคือจำเป็นต้องมีคลังข้อมูล (Corpus) การใช้งานภาษาที่เพียงพอ ถ้าเรามีข้อมูลด้านการใช้งานภาษาแค่เฉพาะด้านใดด้านหนึ่ง แบบจำลองที่ถูกฝึกฝนขึ้นมา ก็จะทำงานได้ดีเฉพาะด้านนั้น ๆ เท่านั้น ซึ่งถ้าหากเราต้องการให้แบบจำลองทำงานได้ดีในทุก ๆ ด้าน ก็จำเป็นต้องมีข้อมูลที่ใหญ่มากเพียงพอสำหรับใช้ในการฝึกฝน รวมถึงต้องใช้เวลาในการฝึกฝนแปรตามปริมาณข้อมูลที่ใช้ฝึกฝนอีกด้วย โดยหลังจากการศึกษาและทดลองใช้ในการตัดคำภาษาไทย ผู้วิจัยค้นพบว่าวิธีดังกล่าวยังไม่สามารถตัดคำได้ถึงหน่วยที่เล็กเพียงพอ จึงได้เสนอวิธีการใช้คลังพจนานุกรมเข้ามาช่วยตัดคำเพิ่มเติม แล้วให้ผลที่ดีกว่าเทียบกับการใช้คลังพจนานุกรมในการตัดคำอย่างเดียว และใช้อัลกอริทึมคอนดิชันนอลแรนดอมฟิลด์สอย่างเดียว

2. เล็กโตพลัส (LexToPlus) เครื่องมือตัดหน่วยคำ (Lexeme) ภาษาไทยและทำให้เป็นมาตราฐาน (Normalization)

งานวิจัยของ C. Haruechaiyasak และคณะ [37] ได้ศึกษาและพัฒนาเครื่องมือตัดคำใหม่จากเครื่องมือเดิม เลิร์นเล็กโต เนื่องด้วยความต้องการในการใช้เครื่องมือตัดคำในงานหลาย ๆ อย่าง รวมถึงงานทางด้านการวิเคราะห์ข้อมูลสื่อสังคมออนไลน์ ซึ่งมักจะมีความผิดพลาดทางภาษาที่ทำให้ประโยคมีการใช้คำที่เปลี่ยนแปลงไปตามบริบทต่าง ๆ ซึ่งจากปัญหานี้เอง ที่ทำให้เครื่องมือเดิม เลิร์นเล็กโต ที่ถูกพัฒนาด้วยวิธีการเรียนรู้ของเครื่องไม่สามารถแก้ไขข้อจำกัดนี้ได้ เนื่องจากไม่มีคลังข้อมูลที่เพียงพอ จึงทำให้นักวิจัยเลือกใช้วิธีการตัดคำโดยใช้คลังพจนานุกรมเป็นหลัก (Dictionary-based) และเสริมด้วยกฎทางภาษา (Rule-based) ซึ่งให้ผลที่ดีกว่าในการรับมือกับปัญหาเหล่านี้เทียบกับวิธีการเรียนรู้ของเครื่อง โดยข้อจำกัดของวิธีนี้คือจำเป็นต้องมีการเพิ่มคลังคำศัพท์ลงในพจนานุกรมให้เพียงพอ เพื่อให้สามารถรับมือกับคำใหม่ ๆ ที่เกิดขึ้นได้ แต่คลังคำศัพท์ที่มีอยู่ในปัจจุบันก็ยังไม่เพียงพอ และวิธีนี้ก็ยังไม่สามารถแก้ไขปัญหาความผิดพลาดทางภาษาอีกหลาย ๆ อย่างได้

3. การตัดคำที่สนใจผลลัพธ์หลายแบบโดยใช้โครงข่ายประสาทเทียมหน่วยความจำระยะสั้นแบบยาวแบบ 2 ทิศทาง (Bi-directional LSTM Neural Networks)

งานวิจัยของ T. Lapjaturapit และคณะ [38] ได้ศึกษาการตัดคำในหลายภาษา ได้แก่ จีน ญี่ปุ่น และไทย โดยพบว่าโครงข่ายประสาทเทียมแบบลึกกำลังเป็นที่นิยมในการแก้ปัญหาการตัดคำ โดยโครงข่ายประสาทเทียมหน่วยความจำระยะสั้นแบบยาว (Long Short-term Memory: LSTM) ได้รับความสนใจมากที่สุด เนื่องจากมีการใช้ลำดับของคำก่อนหน้ามาใช้ในการวิเคราะห์คำถัด ๆ ไปด้วย โดยเลือกเก็บคำที่มีความสำคัญในลำดับก่อนหน้าไว้วิเคราะห์ต่อ และที่ได้รับความสนใจรองลงมาคือ โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network: RNN) เนื่องจากมีการใช้ลำดับของคำในการฝึกสอนโมเดล โดยโครงข่ายประสาทเทียมแบบ LSTM ก็เป็นหนึ่งในโครงข่ายประสาทเทียมแบบวนซ้ำที่มีการพัฒนาเพิ่มเติมมา โดยได้มีการศึกษาเครื่องมือที่ใช้ในการตัดคำเพิ่มเติมทั้งแบบที่ใช้พจนานุกรมและแบบที่ไม่ใช้พจนานุกรม รวมถึงเครื่องที่ตัดคำที่ใช้โครงข่ายประสาทเทียมแบบลึก

ดีพคัต (Deepcut) เป็นหนึ่งในเครื่องมือที่ใช้โครงข่ายประสาทเทียมแบบลึกที่ฝึกสอนเพื่อใช้สำหรับตัดคำในภาษาไทยโดยใช้คลังข้อมูลภาษา BEST2009 [39] ซึ่งเป็นเครื่องมือตัดคำที่มีประสิทธิภาพสูงโดยมีค่า F1 ถึง 98.1% อีกทั้งเครื่องมือ Deepcut ได้

ถูกเปิดเป็นโอเพนซอร์ส (Opensource) บนเว็บไซต์กิตฮับ (Github) ทุกคนสามารถนำมาใช้งานได้

หลังจากที่ได้ศึกษาเครื่องมือตัดคำจำนวนมาก พบว่าเครื่องมือทั้งหมดได้ผลลัพธ์สุดท้ายเพียงแบบเดียว ซึ่งคำในภาษาต่าง ๆ บางคำนั้นเกิดจากการรวมกันของคำมากกว่า 1 คำ ในบางงานเราอาจต้องการให้คำเหล่านั้นโดนตัดออกเป็นคำย่อย ๆ แต่ในบางงานเราอาจจะต้องการให้คำเหล่านั้นไม่โดนตัดออกจากกัน งานวิจัยนี้จึงพัฒนาเครื่องมือตัดคำโดยใช้โครงข่ายประสาทเทียมแบบ LSTM โดยมีการประยุกต์ใช้ค่าขีดแบ่ง (Threshold) ตามความมั่นใจของแบบจำลอง โดยหากค่าขีดแบ่งสูง เราจะได้ผลลัพธ์เป็นคำที่โดนตัดเป็นคำย่อย ๆ มากขึ้น แต่หากค่าขีดแบ่งต่ำลงเรื่อง ๆ เราจะได้ผลลัพธ์เป็นคำ เป็นวลี หรือเป็นประโยค ทำให้เราสามารถเลือกปรับใช้ค่าขีดแบ่งได้ตามลักษณะงานที่เราต้องการ

การจำแนกข้อความ

งานวิจัยเกี่ยวกับการจำแนกข้อความมีปริมาณมาก ในที่นี้ขอยกตัวอย่างงานวิจัยที่มีความเกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ซึ่งเป็นงานวิจันที่มีการใช้โครงข่ายประสาทเทียมคอนโวลูชันในการจำแนกข้อความ โดยมีทั้งแบบระดับคำ และแบบระดับตัวอักษร ซึ่งเป็นการต่อยอดในเงื่อนไขของการไม่พึ่งพาเครื่องมือที่ใช้ในการตัดคำ ดังต่อไปนี้

1. โครงข่ายประสาทเทียมคอนโวลูชัน (Convolutional Neural Network) สำหรับการจำแนก (Classification) ข้อความ

   งานวิจัยของ Y. Kim และคณะ [40] ได้ศึกษาการจำแนกของข้อความ โดยเสนอวิธีการใช้แบบจำลองโครงข่ายประสาทเทียมคอนโวลูชัน (Convolutional Neural Network: CNN) ในระดับคำ โดยงานวิจัยนี้ ได้มีการตัดคำจากประโยค และแปลงคำให้เป็นเวกเตอร์ ก่อนจะนำเวกเตอร์ของคำทั้งประโยค ใส่ในแบบจำลอง โดยจะแบบจำลองดังกล่าว มีการวางชั้นตัวกรองเพื่อเลือกคำเพียงบางคำจากประโยค ทำให้เราได้เวกเตอร์ขนาดคงที่ ก่อนจะนำไปสู่ชั้นที่เชื่อมโยงโครงข่ายแบบเต็มรูปแบบ ทำให้แบบจำลองนี้สามารถรับประโยคที่มีความยาวเท่าใดก็ได้ ซึ่งแบบจำลองนี้ ได้ผลลัพธ์ดีกว่าแบบจำลองอื่น ๆ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสารทเทียมแบบอื่น ๆ ในการจำแนกข้อความไปถึง 4 ใน 7 ชุดข้อมูล โดยชุดข้อมูลดังกล่าวเป็นการจำแนกข้อความประเภทต่าง ๆ ได้แก่ รีวิวภาพยนตร์ด้านบวก รีวิวภาพยนตร์ด้านลบ รีวิวสินค้า และประโยคคำถาม

2. โครงข่ายประสาทเทียมคอนโวลูชัน สำหรับการสร้างแบบจำลอง (Modelling) ข้อความ

งานวิจัยของ N. Kalchbrenner และคณะ [41] ได้ศึกษาการสร้างแบบจำลองข้อความ โดยเสนออีกวิธีการหนึ่ง คือแบบจำลองโครงข่ายประสาทเทียมคอนโวลูชันแบบพลวัต (Dynamic Convolutional Neural Network: DCNN) ในระดับคำ โดยงานวิจัยนี้ได้มีการตัดคำจากข้อความ และแปลงคำให้เป็นเวกเตอร์ เหมือนงานวิจัยก่อนหน้า แต่ชั้นตัวกรองคำมีความแตกต่างจากเดิม โดยมีการกำหนดค่า K เพื่อรวมคำจำนวน K คำเข้าด้วยกัน คล้ายเอ็นแกรมของคำ โดยเราจำเป็นต้องมีการกำหนดค่า L คือจำนวนชั้นที่ใช้การรวมคำ และค่า S คือจำนวนคำที่เหลืออยู่ก่อนจะไปสู่ชั้นที่เชื่อมโยงโครงข่ายแบบเต็มรูปแบบ ทำให้เมื่อเรารู้จำนวนคำของประโยค ค่า L และ ค่า S เราจะสามารถคำนวณหาค่า K ได้ และนำไปสู่แบบจำลองดังกล่าว โดยแบบจำลองนี้ได้ผลลัพธ์ดีกว่า แบบจำลองอื่น ๆ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ เอนโทรปีมากที่สุด และโครงข่ายประสาทเทียมแบบหน่วงเวลา โดยชุดข้อมูลที่ใช้เป็นการจำแนกข้อความด้านอารมณ์จากรีวิวภาพยนตร์ และจากข้อความบนสื่อสังคมออนไลน์ทวิตเตอร์ (Twitter) และจำแนกของประเภทคำถามในภาษาอังกฤษ ทั้งนี้แบบจำลองนี้ให้ผลลัพธ์ใกล้เคียงกับแบบจำลองของ Y. Kim และคณะ โดยสามารถชนะในการจำแนกอารมณ์ของข้อความ

3. โครงข่ายประสาทเทียมคอนโวลูชันแบบลึก (Deep Convolution Neural Network) สำหรับวิเคราะห์อารมณ์ (Sentiment Analysis) ของข้อความขนาดสั้น

งานวิจัยของ C. N. d. Santos และคณะ [16] ได้ศึกษาการวิเคราะห์อารมณ์ของข้อความขนาดสั้น และเสนอการใช้โครงข่ายประสาทเทียมคอนโวลูชันจากตัวอักษรเป็นประโยค (Character to Sentence Convolutional Neural Network: CharSCNN) โดยแบบจำลองดังกล่าว จะมีการตัดข้อความออกเป็นคำก่อน แล้วจึงแปลงตัวอักษรแต่ละตัวของคำเป็นเวกเตอร์ แล้วนำกลับมารวมกันเป็นเมทริกซ์ของคำ ที่ประกอบไปด้วยเวกเตอร์ของแต่ละตัวอักษร จากนั้นจึงนำเมทริกซ์ดังกล่าว ไปใส่แบบจำลองโครงข่ายประสาทเทียมคอนโวลูชันแบบคำต่อไป จากวิธีการดังกล่าวทำให้ได้ผลลัพธ์ที่ดีกว่าการใช้เวกเตอร์ของคำเพียงอย่างเดียว ในการจำแนกอารมณ์จากข้อความรีวิวภาพยนตร์ และจากข้อความบนสื่อสังคมออนไลน์ทวิตเตอร์ ในภาษาอังกฤษ

4. โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษร (Character-level Convolutional Neural Network) สำหรับการจำแนกข้อความ

งานวิจัยของ X. Zhang และคณะ [15] ได้ศึกษาการจำแนกข้อความ โดยเสนอการใช้โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษร (Character-level Convolutional Neural Network: Char-CNN) ซึ่งแบบจำลองดังกล่าว รับข้อมูลนำเข้า

เป็นลำดับของเวกเตอร์ของตัวอักษรเลย โดยแบบจำลองที่เสนอนี้จำกัดความยาวของตัวอักษรที่ 1014 ตัวอักษร โดยไม่มีชั้นตัวกรองเพื่อกรองตัวอักษรออกให้สามารถรับตัวอักษรขนาดเท่าใดก็ได้ โดยหากมีตัวอักษรเกินมา แบบจำลองนี้จะตัดตัวอักษรที่เกินมาทิ้งทันที โดยแบบจำลองนี้มีชั้นกรองเพื่อรวมเวกเตอร์ของตัวอักษรเหล่านี้เป็นเมทริกซ์ขนาดคงที่ ก่อนจะนำไปสู่โครงข่ายชั้นที่เชื่อมโยงเครือข่ายแบบเต็มรูปแบบ โดยแบบจำลองดังกล่าวถูกนำไปทดสอบเทียบกับโครงข่ายประสาทเทียมคอนโวลูชันระดับคำและหน่วยความจำระยะสั้นแบบยาว (Long Short-term Memory: LSTM) โดยได้ผลลัพธ์ดีกว่าแบบจำลองอื่น 4 ใน 8 ของชุดข้อมูลทั้งหมด

5. โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษรกับความยาวข้อมูลแบบพลวัตสำหรับการจำแนกข้อความภาษาไทย

งานวิจัยของ T. Koomsubha และคณะ [42] ได้ศึกษาการจำแนกข้อความภาษาไทย โดยเสนอการใช้โครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษรแบบพลวัต (Dynamic Character-level Convolutional Neural Network: Dynamic Char-CNN) ซึ่งแบบจำลองที่เสนอมีพื้นฐานจากโครงข่ายประสาทเทียมคอนโวลูชันระดับตัวอักษร และมีการพัฒนาต่อยอดให้สามารถรับความยาวของตัวอักษรแบบพลวัตได้ โดยใช้ความรู้จากโครงข่ายประสาทเทียมคอนโวลูชันแบบพลวัต ซึ่งมีการกำหนดค่า K คล้ายในชั้นตัวกรอง เพื่อรวมตัวอักษรจำนวน K ตัวเข้าด้วยกัน คล้ายเอ็นแกรม และบีบอัดตัวอักษรทั้งหมด L ชั้นจนเหลือเวกเตอร์ขนาด S ตัว ก่อนจะนำไปสู่ชั้นเชื่อมโยงเครือข่ายแบบเต็มรูปแบบ โดยงานวิจัยดังกล่าวได้เปรียบเทียบผลลัพธ์กับแบบจำลองเดิมที่ไม่สามารถเพิ่มขนาดของข้อความได้ ในชุดข้อมูลข่าวในภาษาไทย พบว่าหากข้อความมีความยาวไม่เกิน 1014 ตัวอักษร แบบจำลองแบบเดิมให้ผลลัพธ์ที่ดีกว่า แต่หากข้อความมีตัวอักษรมากกว่านั้น แบบจำลองแบบพลวัตจะให้ผลลัพธ์ที่ดีกว่า แต่เมื่อเทียบกับโครงข่ายประสาทเทียมคอนโวลูชันระดับคำแบบพลวัต พบว่าแบบจำลองดังกล่าวให้ผลลัพธ์ที่ดีกว่าระดับอักษรแบบพลวัต

คำหยุด

ฐานข้อมูลคำหยุดเป็นหนึ่งในสิ่งที่สำคัญสำหรับงานด้านการสกัดคำสำคัญ จึงจำเป็นอย่างยิ่งที่ควรมีการเก็บฐานข้อมูลคำหยุด หรือพัฒนาระบบที่สามารถสร้างฐานข้อมูลคำหยุดได้อย่างอัตโนมัติ โดยในที่นี้ขอยกงานวิจัยที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ดังต่อไปนี้

1. คำหยุดในการประเมินค่าความสามารถในการอ่านข้อความภาษาไทย

   งานวิจัยของ P. Daowadung และคณะ [43] ได้ศึกษาคำหยุดว่ามีผลอย่างไรต่อความสามารถในการอ่านของเด็กประถม 1 ถึงประถม 6 หากถูกนำออกไปจากบทความ โดยมีการสร้างคำหยุดภาษาไทยจากใช้แบบจำลองทางสถิติควบคู่กับแบบจำลองทางสารสนเทศ เพื่อหาความน่าจะเป็นที่คำเหล่านั้นจะเป็นคำหยุด โดยได้มีการศึกษาถึงคำหยุดว่าสามารถเพิ่มประสิทธิภาพได้ในงานทางด้านการหาใจความสำคัญในคลังข้อมูล การจัดแบ่งประเภท (Classification) และงานทางด้านการค้นคืนสารสนเทศ (Information Retrieval) โดยคำหยุดนั้นขึ้นกับขอบเขตของงาน เช่นสำหรับงานด้านการท่องเที่ยว คำว่า “จาก” และ “ถึง” มีความสำคัญมาก ในขณะที่งานอีกหลาย ๆ ด้าน 2 คำนี้มักจะไม่มีความสำคัญจนถือว่าเป็นคำหยุด โดยผลสรุปว่าการนำคำหยุดออกไปจากบทความ จะทำนักเรียนประถมต้น อ่านจับใจความได้ยากขึ้น ในขณะที่นักเรียนประถมปลาย อ่านจับใจความได้ดีขึ้น

2. การสร้างฐานข้อมูลคำหยุดแบบอัตโนมัติสำหรับงานด้านการค้นคืนสารสนเทศ

   งานวิจัยของ R. T. Lo และคณะ [44] ได้เสนอหนึ่งในวิธีที่สามารถสร้างฐานข้อมูลคำหยุดได้ดีกว่าแบบเดิม โดยดั้งเดิมนั้น ฐานข้อมูลคำหยุด จะถูกสร้างจากการนับความถี่ของคำ (Term Frequency, TF) เนื่องจากคุณสมบัติของคำหยุด คือคำที่ปรากฏขึ้นบ่อยครั้ง ในภาษา ซึ่งเมื่อเราได้ความถี่ของแต่ละคำแล้ว เราก็กำหนดค่าขีดแบ่ง (Threshold) เพื่อใช้แบ่งแยกคำหยุด กับ คำที่ไม่ใช่คำหยุดออกจากกัน แต่ในงานวิจัยนี้พวกเขาได้ประยุกต์ใช้วิธีประเมินค่าที่มีชื่อว่า Kullback-Leibler divergence มาเสริมกับวิธีนับความถี่ของคำแบบเดิม ซึ่งสามารถให้ความน่าจะเป็นที่คำนั้นจะเป็นคำหยุดได้ดีกว่าวิธีดั้งเดิม แต่ทว่าวิธีนี้ความแม่นยำยังขึ้นกับลักษณะของภาษา และใช้เวลาในการประมวลผลค่อนข้างนาน

การสกัดคำสำคัญ

งานทางด้านการสกัดคำสำคัญมีประโยชน์ในการจำแนกข้อมูลต่าง ๆ รวมถึงสามารถหาคำสำคัญที่ใช้แทนกระแส หรือเหตุการณ์สำคัญบนสื่อสังคมออนไลน์อีกด้วย โดยขอยกตัวอย่างงานที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ดังต่อไปนี้

1. การสำรวจวิธีการสกัดความรู้ (Knowledge Extraction) ภาษาไทยสำหรับเครื่องมือและงานวิจัยเว็บเชิงหมายความ (Semantic Web)

ผลสำรวจของ P. Netisopakul และคณะ [45] ได้สำรวจงานทางด้านการสกัดคำสำคัญและการสกัดสารสนเทศภาษาไทย จากการสำรวจพบว่างานวิจัยทางด้านนี้ของไทยยังคงมีจำนวนน้อยเมื่อเทียบกับญี่ปุ่น โดยมีจำนวนงานวิจัยแตกต่างถึง 5-7 เท่า อีกทั้งงานวิจัยทางด้านนี้ส่วนใหญ่มาจากมหาวิทยาลัยไทยไม่กี่แห่ง และนักวิจัยคนไทยไม่กี่คน โดยนักวิจัยคนไทยหลัก 4 คน มีจำนวนงานวิจัยที่เกี่ยวข้องกับงานด้านการสกัดคำสำคัญและการสกัดสารสนเทศภาษาไทยในฐานข้อมูล scopus ถึง 68%

สำหรับเครื่องมือทางด้านการวิเคราะห์ภาษาธรรมชาติภาษาไทยนั้นค่อนข้างจำกัดมาก เนื่องจากเครื่องมือเด่น ๆ อย่าง GATE หรือ NLTK ไม่รองรับภาษาไทย จะมีเพียงเครื่องมือที่รองรับการตัดคำและการบอกชนิดของคำอย่าง PyThai, PyThaiNLP, WordCut, RDRPOStagger เป็นต้น ซึ่งมีความแม่นยำในการตัดคำอยู่ที่ประมาณ 90% แต่ก็ยังคงมีข้อจำกัดหลาย ๆ ด้าน

สำหรับงานวิจัย การสำรวจนี้แบ่งงานทางด้านการสกัดสารสนเทศ ออกเป็น 9 หมวด ได้แก่ การสกัดคำสำคัญหรือการสกัดหัวเรื่อง (Topic and Keyword Extraction), การสกัดคำศัพท์เฉพาะทาง (Terminology Extraction), การสกัดอนุกรมวิธาน (Taxonomy Extraction), การสกัดความสำพันธ์ทวิภาค (Binary Relation Extraction), การรู้จำชื่อที่มีเอกลักษณ์ (Named Entity Recognition), การเชื่อมต่อชื่อที่มีเอกลักษณ์ (Named Entity Linking), การติดป้ายความหมายของคำ (Supersense Tagging), การติดฉลากบทบาทเชิงความหมาย (Semantic Role Labeling) และ การสกัดเหตุการณ์ (Event Extraction) โดยมีการเปรียบเทียบงานวิจัยต่าง ๆ ว่ามีจำนวนมากน้อยเพียงใดในแต่ละด้าน และยังเทียบกับขอบเขต (Domain) ที่วิจัย เช่น สื่อสังคมออนไลน์ การแพทย์ และ การเกษตร เป็นต้น อีกทั้งได้มีการเปรียบเทียบวิธีที่ใช้ในงานวิจัย ไม่ว่าจะเป็น การเรียนรู้ของเครื่อง (Dictionary) รูปแบบ (Pattern) และ พจนานุกรม เป็นต้น

ผลคืองานวิจัยประมาณครึ่งหนึ่งของทั้งหมด อยู่ในหมวดหมู่การสกัดคำสำคัญ การสกัดหัวเรื่อง และการสกัดคำศัพท์เฉพาะทาง ซึ่งมีสัดส่วนใกล้เคียงกับญี่ปุ่น โดยเมื่อพิจารณาด้านสื่อสังคมออนไลน์แล้ว งานส่วนใหญ่มักจะเป็นด้านการสกัดคำสำคัญหรือการสกัดหัวเรื่อง โดยยังไม่มีงานทางด้านการสกัดเหตุการณ์มากนัก อีกทั้งวิธีที่มักจะถูกใช้ในงานเหล่านี้ มักจะเป็นการเรียนรู้ของเครื่อง หรือ ใช้รูปแบบในการจำแนกคำสำคัญออกมา

2. การสกัดคำสำคัญภาษาไทยแบบอัตโนมัติจากคลังข้อมูลของข้อความที่ถูกจำแนกแล้ว

งานวิจัยของ C. Haruechaiyasak และคณะ [46] ได้เสนออัลกอริทึมที่ใช้ในการสกัดคำสำคัญโดยไม่ใช้วิธีตัดคำ โดยใช้หลักการต่อสายอักขระ (String) จากความยาว 1 ขึ้นไปเรื่อย ๆ โดยระหว่างทางที่ต่อสายอักขระ ก็จะมีการตรวจสอบค่าการปรากฏของสายอักขระนี้ในจำนวนเอกสารทั้งหมด ว่าเกิน 10% หรือไม่ และตรวจสอบค่าเอนโทรปีของสายอักขระนี้กับเอกสารที่จำแนกแล้ว เพื่อดูว่าสายอักขระนี้ ปรากฏบนหลายเอกสารหรือไม่ หากมีค่าไม่เกิน 2.5 ถือว่าดี เพราะหากมากกว่านั้น คำเหล่านั้นน่าจะเป็นคำหยุดมากกว่าคำสำคัญ เพราะปรากฏขึ้นในเอกสารหลายหมวดหมู่มากเกินไป แล้วจึงนำสายอักขระดังกล่าวไปต่อความยาวเพิ่ม จนกระทั่งไม่มีสายอักขระที่ผ่านเงื่อนไข จึงเลิกต่อความยาว โดยยังมีการใช้แถวลำดับคำต่อท้าย (Suffix Array) ในการเพิ่มประสิทธิภาพในการค้นหาสายอักขระจากเอกสารอีกด้วย

หลังจากที่ได้เซตของสายอักขระที่ผ่านเงื่อนไขดังกล่าว จึงนำสายอักขระเหล่านั้นมารวมกันเป็นคำสำคัญ โดยหากสายอักขระ มีค่าการปรากฏเท่ากัน และสายอักขระนั้นทับซ้อนกันได้พอดี จึงจะนำมารวมกัน ผลคือเราจะได้คำสำคัญที่สามารถบอกหมวดหมู่ได้ว่ามาจากเอกสารหมวดหมู่ใดด้วย

แต่ข้อเสียของวิธีนี้คือมีการใช้พื้นที่หน่วยความจำค่อนข้างมาก และ เวลาประมวลผลที่นาน เนื่องจากเราต้องค่อย ๆ สร้างเซตของสายอักขระจากความยาว 1 ขึ้นไป และเปรียบเทียบกับเอกสารทั้งหมดที่เรามี รวมถึงต้องมีการจำแนกเอกสารไว้ก่อนอีกด้วย อีกทั้งการรวมกันของสายอักขระให้เป็นคำสำคัญ ก็มีโอกาสที่จะรวมกันแล้ว ได้คำสำคัญมากกว่า 1 แบบ หรือ รวมกันไม่ได้ ทำให้คำสำคัญหายไป จึงทำให้วิธีนี้ยังมีข้อจำกัดอยู่มาก

3. การสกัดคำสำคัญที่แสดงถึงเหตุการณ์สำคัญบนสื่อสังคมออนไลน์ทวิตเตอร์ (Twitter) ภาษาไทย

งานวิจัยของ A. Piyatumrong และคณะ [47] ได้เปรียบเทียบวิธีการ 5 วิธี ที่ออกแบบมาเพื่อใช้หาคำสำคัญที่แสดงถึงเหตุการณ์สำคัญ ได้แก่ แฮชแท็ก แฮชแท็ก5 (แทนแฮชแท็กที่ยาวตั้งแต่ 5 ตัวอักษร) ยูนิแกรม ไบแกรมแบบไม่ตัดคำหยุด และ ไบแกรมแบบตัดคำหยุด โดย ยูนิแกรม และ ไบแกรมในงานวิจัยนี้ คือ ชุดของคำที่มีความยาว 1 คำ และ 2 คำ ตามลำดับ โดยเปรียบเทียบทั้ง 5 วิธี โดยนำทั้ง 5 วิธีนี้ มาประมวลผลโดยใช้ การพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (Term Frequency, TF) และ การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (Term Frequency - Inverse Document Frequency, TF-IDF) ผลปรากฏว่า วิธีที่ได้

ค่า F1 สูงที่สุดคือ แฮชแท็ก5 ซึ่งมากกว่าแฮชแท็กปกติ และวิธีอื่น ๆ แต่เมื่อเราพิจารณาไบแกรมแบบตัดคำหยุด และ ไบแกรมแบบไม่ตัดคำหยุด วิธีทั้งคู่ได้ค่า F1 ใกล้เคียงกัน ซึ่งมากกว่ายูนิแกรม แต่เนื่องจากการใช้เพียงแค่แฮชแท็กไม่สามารถครอบคลุมคำสำคัญบางอย่างที่อยู่ในข้อความได้ จึงทำให้วิธีอย่าง ไบแกรม ดูเป็นทางเลือกที่เหมาะสมกว่า โดยหากพิจารณาวิธีที่ใช้ประมวลผลทั้ง 2 แบบ พบว่า การพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (TF) มีค่าไม่ต่างจาก การพิจารณาความถี่ของคำที่ปรากฏในเอกสารส่วนด้วยจำนวนของเอกสารที่คำนั้นปรากฏ (TF-IDF) แต่ประมวลได้เร็วกว่ามาก และใช้ทรัพยากรเครื่องน้อยกว่าด้วย งานวิจัยนี้จึงแนะนำให้ใช้เพียง การพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (TF) ก็เพียงพอแล้ว

โดยเมื่อพิจารณาถึงวิธีวัดผลของงานวิจัยนี้ พบว่ามีการวัดผลโดยพยายามให้ผลลัพธ์ที่ได้จากการวัดผลไม่ขึ้นกับบุคคลหรือพื้นฐานความรู้ดั้งเดิมของคนเพียงกลุ่มเดียว จึงได้วัดผลโดยอาศัยอาสาสมัครคนไทยมาตอบแบบสอบถามว่าคำที่ระบบสกัดออกมาได้นั้น เป็นคำที่ตนเองคิดว่าเป็นคำสำคัญที่แสดงถึงเหตุการณ์หรือไม่ เป็นจำนวนทั้งสิ้น 162 คน โดย 89% อยู่ในกลุ่มอายุ 20-40 ปี และ 74.6% มีการใช้งานสื่อสังคมออนไลน์มากกว่า 3 ชั่วโมงต่อวัน

ขั้นตอนการสร้างระบบสกัดคำสำคัญที่เป็นกระแสและคำหยุด จากเพจเฟซบุ๊ก

วิทยานิพนธ์ฉบับนี้ ได้นำเสนอขั้นตอนการสร้างระบบสกัดคำสำคัญที่เป็นกระแสและคำหยุดจากเพจเฟซบุ๊ก โดยมีการวิเคราะห์ปัญหาและข้อจำกัดที่เกิดขึ้นเมื่อทดลองใช้เครื่องมือตัดคำ โดยมีหัวข้อดังนี้

1. ก่อนเริ่มสร้างระบบสกัดคำสำคัญ
    a. ให้คำจำกัดความของกระแสที่ต้องการบนเฟซบุ๊ก
    b. เก็บข้อมูลโพสต์จากเพจเฟซบุ๊ก
    c. ทำความสะอาดข้อมูลโพสต์
    d. การทดลองใช้เครื่องมือตัดคำ deepcut และการวิเคราะห์ผลลัพธ์และข้อจากัดที่เกิดขึ้น
2. ขั้นตอนการสกัดคำสำคัญ
    a. แบ่งข้อความออกเป็นชุดของตัวอักษร (แกรม)
    b. การวิเคราะห์ปัญหาจากการใช้อัลกอริทึมเอ็นแกรมแบบตัวอักษรแทนที่จะใช้เครื่องมือตัดคำ
    c. นับความถี่ของแกรม
    d. หาแกรมที่มีคุณสมบัติเป็นคำสำคัญของเพจ
    e. หาแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ
    f. รวมแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ กลับมาเป็นคำสำคัญ
3. ขั้นตอนการจำแนกคำสำคัญที่เป็นกระแส และคำหยุดออกจากกัน
    a. สกัดคำหยุดออกจากคำสำคัญที่เป็นกระแส
    b. สกัดคำสำคัญที่เป็นกระแสออกจากคำสำคัญ

ให้คำจำกัดความของกระแสที่ต้องการบนเฟซบุ๊ก

เนื่องจากกระแสคือเรื่องที่ถูกพูดถึงบ่อย ๆ โดยคนจำนวนมากในช่วงเวลาสั้น ๆ ดังนั้นกระแสจึงมีหมวดหมู่จากกลุ่มคนที่พูดถึง เช่นหากเราต้องการรู้ว่ามือถือเครื่องไหนกำลังเป็นกระแส เราอาจต้องไปสำรวจดูว่าในกลุ่มคนที่พูดถึงเรื่องมือถือ เขาพูดถึงมือถือรุ่นไหนกันอยู่ เป็นต้น

งานวิจัยนี้ต้องการวิเคราะห์กระแสที่เกิดขึ้นทั่ว ๆ ไปในสังคมไทย จึงอาจกล่าวได้ว่า กระแสที่งานวิจัยนี้ต้องการ คือกระแสที่เกิดขึ้นจากข่าวต่าง ๆ ซึ่งสะท้อนถึงสิ่งที่กำลังเกิดขึ้นในสังคม โดยในงานวิจัยนี้เราจะนำข้อมูลเพจเฟซบุ๊กซึ่งเป็นพื้นที่สาธารณะบนเฟซบุ๊กที่ทุกคนสามารถเข้าถึงได้ มาใช้ในการวิเคราะห์กระแสดังกล่าว ทำให้งานวิจัยนี้จำเป็นต้องเลือกเพจเฟซบุ๊กที่เป็นเพจข่าวที่มียอด

ผู้ติดตามสูงที่สุดในไทย เพื่อให้รู้ว่าผู้คนในสังคมไทยที่ใช้สื่อสังคมออนไลน์เฟซบุ๊กนั้น กำลังสนใจอะไรอยู่ ผ่านทางเรื่องที่เพจข่าวเหล่านั้นโพสต์ออกมา อีกทั้งเพจข่าวยังสะท้อนกระแสที่เกิดขึ้นในสังคมได้ดี เนื่องจากหากเรื่องใดกำลังเป็นกระแสในสังคมอยู่ ก็จะมีข่าวเกี่ยวกับเรื่องเหล่านั้นเพิ่มขึ้นด้วยเช่นกัน โดยในงานวิจัยนี้ใช้ข้อมูลอ้างอิงยอดผู้ติดตามของเพจข่าวบนเฟซบุ๊กจากเว็บไซต์ Socialbakers (https://www.socialbakers.com) ซึ่งเป็นเว็บไซต์ที่มีการจัดอันดับเพจต่าง ๆ ที่อยู่บนสื่อสังคมออนไลน์เฟซบุ๊ก โดยในงานวิจัยนี้ได้เลือกเก็บข้อมูลโพสต์จากเพจข่าวไทยที่มียอดผู้ติดตามสูงที่สุด 10 อันดับแรกบนเฟซบุ๊ก โดยมีรายชื่อเพจแสดงไว้ใน*ตารางที่ 3*

อันดับ	ชื่อเพจ	ไอดีเพจ	ยอดผู้ติดตาม
1	Khaosod - ข่าวสด	khaosod	13 248 101
2	Workpoint Entertainment	workpoint	12 948 074
3	เรื่องเล่าเช้านี้	MorningNewsTV3	11 929 770
4	Thairath	thairath	10 965 794
5	Ch7HD	Ch7HD	10 802 256
6	ช่อง one31	one31Thailand	6 848 780
7	Sanook News	SanookNews	5 865 920
8	GMM25Thailand	GMM25Thailand	5 600 078
9	Thai PBS	ThaiPBSFan	4 784 047
10	Mono 29	Mono29TV	3 911 405

ตารางที่ 3 ตารางแสดงอันดับของเพจข่าวบนเฟซบุ๊กเรียงตามยอมผู้ติดตาม

หมายเหตุ ข้อมูลยอดผู้ติดตามเป็นข้อมูลเมื่อวันที่ 7 มิถุนายม พ.ศ. 2562

เก็บข้อมูลโพสต์จากเพจเฟซบุ๊ก

ในขั้นตอนการเก็บข้อมูลโพสต์นั้น เดิมเฟซบุ๊กมี API สำหรับเข้าถึงข้อมูลโพสต์ของเพจต่าง ๆ ผ่านทาง Facebook Graph API แต่เนื่องด้วยสถานการณ์ปัจจุบัน ที่เฟซบุ๊กมีปัญหากับการเปิดให้บุคคลภายนอกเข้าถึงข้อมูลส่วนบุคคลของผู้ใช้งานได้โดยไม่มีการควบคุมที่ดีพอ ทำให้เฟซบุ๊กตัดสินใจเปลี่ยนกฎระเบียบใหม่ในการเข้าถึงข้อมูลผ่าน Facebook Graph API ซึ่งทำให้การดึงข้อมูลโพสต์ของเพจที่เป็นสาธารณะย้อนหลังทำไม่ได้อีกต่อไป

งานวิจัยนี้จึงอาศัยการดึงข้อมูลโพสต์ผ่านทางหน้าเว็บไซต์ของเฟซบุ๊กโดยตรงแทน ซึ่งงานวิจัยนี้ทำได้โดยการเขียนสคริปต์ภาษา Python สำหรับดึงข้อมูลโพสต์ของเพจต่าง ๆ โดยอาศัยเครื่องมือชื่อ Beautiful Soup ซึ่งทำให้เราสามารถแปลงภาษา HTML ที่เว็บไซต์ใช้สำหรับแสดงผลให้เป็นข้อมูลที่สามารถนำไปใช้งานต่อได้ง่ายขึ้น โดยเราสามารถเลือกดึงข้อมูลเฉพาะส่วนที่เราสนใจออกมาจากภาษา HTML ได้ และงานวิจัยนี้ได้เลือกใช้ฐานข้อมูล MongoDB สำหรับเก็บข้อมูล เนื่องจากเป็นฐานข้อมูลที่ไม่เสียค่าใช้จ่าย ติดตั้งง่าย และมีความยืดหยุ่นสูง

งานวิจัยนี้ได้ทำการเก็บข้อมูลโพสต์ของทั้ง 10 เพจตั้งแต่ 1 มกราคม พ.ศ. 2561 จนถึง 31 มีนาคม พ.ศ. 2562 ตามเวลาประเทศไทย (กล่าวคือตั้งแต่ Unix timestamp วินาทีที่ 1514739600 จนถึงก่อน Unix timestamp วินาทีที่ 1554051600) โดยมีรายละเอียดแสดงไว้ใน*ตารางที่ 4 ถึง 17*

อันดับ	ชื่อเพจ	ไอดีเพจ	จำนวนโพสต์
1	Khaosod - ข่าวสด	khaosod	876
2	Workpoint Entertainment	workpoint	850
3	เรื่องเล่าเช้านี้	MorningNewsTV3	853
4	Thairath	thairath	874
5	Ch7HD	Ch7HD	253
6	ช่อง one31	one31Thailand	417
7	Sanook News	SanookNews	884
8	GMM25Thailand	GMM25Thailand	980
9	Thai PBS	ThaiPBSFan	786
10	Mono 29	Mono29TV	843

ตารางที่ 4 ตารางแสดงจำนวนโพสต์ที่เก็บได้จากหน้าเว็บไซต์ของเพจบนเฟซบุ๊ก

<u>หมายเหตุ</u> เนื่องจากข้อจำกัดในการเก็บข้อมูลจากหน้าเว็บไซต์ของเพจโดยตรง ซึ่งได้ข้อมูลโพสต์เพียงบางส่วนของเพจเท่านั้น เพราะเฟซบุ๊กมีการป้องกันไม่ให้สามารถย้อนดึงข้อมูลโพสต์ทั้งหมดได้

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	Live! ร่วมนับถอยหลัง เคาท์ดาวน์ส่งท้ายปีเก่า 2561 เข้าสู่ปีใหม่ 2562 ในงาน “AIS Bangkok Countdown 2019” พร้อมเฉลิมฉลองด้วยพลุ 6,130 นัด ที่หน้าลานหน้าศูนย์การค้าเซ็นทรัลเวิลด์ #ข่าวสดบันเทิง
2	2โจรอำมหิต ลวงวินมาฆ่า เหยื่อร้องขอชีวิต "อย่าทำผมเลย" สุดท้ายไม่รอด #ข่าวสด #ทุกทิศทั่วไทย
3	แป้งแพนเค้กใครว่าทำยาก!! เคยกินไหม "แพนเค้กข้าว" ทำเองได้ที่บ้าน #happydish #แพนเค้กข้าว
4	Live ! เกาะติดสถานการณ์ เจ้าหน้าที่เร่งเตรียมเจาะถ้ำ เพื่อระบายน้ำออกจากถ้ำหลวง #ข่าวสดไลฟ์สตรีม #khaosod #ทีมหมูป่า #ถ้ำหลวง
5	ข่าวสดท่องเที่ยว: หากคุณชื่นชอบผลไม้ ทั้งทุเรียน เงาะ ลองกอง ยังมีเวลาสำหรับวันว่างในช่วงสุดสัปดาห์ ไปเที่ยวเมืองระยอง ลิ้มรสบุฟเฟ่ต์ผลไม้แบบไม่อั้น แวะชมวัดโบราณพร้อมสิ่งศักดิ์สิทธิ์ของเมืองระยอง และพักโรงแรมที่ได้มาตรฐาน เพื่อการพักผ่อนอย่างเต็มที่ #ข่าวสดท่องเที่ยว #บุฟเฟ่ต์ผลไม้ #หมอนทอง
6	ถ้า "ตูนิเซีย" เล่นเหมือนนัดเจออังกฤษ จะถูก "เบลเยี่ยม" ฉีกกระจุยกระจายแน่! #ข่าวสดบอลโลก #ข่าวสดมติชนบอลโลก2018 #เชียร์ทีมชอบเมนต์ทีมใช่ #Worldcup2018 #ฟุตบอลโลก2018 #Russia2018
7	Live “อั้ม-พัชราภา” แจงกรณีจ่อฟ้องเว็บข่าว หลังบอกตนขึ้นศาลคดียักยอก 39 ล้านบาท! #ข่าวสดบันเทิง
8	Live ! บรรยากาศบวงสรวงละคร 2 เรื่อง "เงา" และ "ทีมล่าทรชน" นำทีมโดย พี่ฉอด-สายทิพย์ ,เอส-วรฤทธิ์ ,อั้ม-อธิชาติ ,เก้า-จิรายุ ,แป้ง-อรจิรา ,แพตตี้-อังศุมาลิน ,ออย-ธนา ,จ๋า-ณัฐฐาวีรนุช ,ยิปซี-คีรติ ฯลฯ ที่ จีเอ็มเอ็ม แกรมมี่ #ข่าวสดบันเทิง
9	Live ตร. คุมตัว เสก โลโซ ส่งศาลฝากขัง หลังถูกรวบคาบ้านเมื่อวันก่อน จากกรณียิงปืนขึ้นฟ้า 10 นัด #เสกเลโซ #10นัด #ส่งศาล #ข่าวสดไลฟ์สตรีม
10	เผื่อเมื่อคืนนี้ใครพลาด ชมอีกครั้งดอกไม้ไฟระยิบระยับริมแม่น้ำเจ้าพระยา ที่เอเชียทีค #NewYear2018

ตารางที่ 5 ตารางแสดงตัวอย่างโพสต์ของเพจ Khaosod - ข่าวสด

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	LIVE! มาร่วมชมงาน “สวดมนต์ข้ามปี ถวายพระราชกุศล เสริมสิริมงคลทั่วโลก ส่งท้ายปีเก่า วิถีไทย ต้อนรับปีใหม่วิธีธรรม พุทธศักราช ๒๕๖๒” ณ มณฑลพิธีท้องสนามหลวง
2	Live!! Victory BNK48 สงครามความน่ารักของเหล่า BNK48 ละลายหัวใจเหล่าโอตะด้วยกิจกรรมที่ทรมานใจน้องๆ #VictoryBNK48 #Workpoint23
3	กาแฟเขาช่อง ฉลองครบรอบ 59 ปี แจกทอง 59 บาท ! . เพียงแค่คุณเขียน ชื่อ-ที่อยู่ และเบอร์โทรศัพท์ ด้วยลายมือให้ชัดเจนใส่กระดาษ ติดลงบนฉลากผลิตภัณฑ์กาแฟ หรือ ครีมเทียม เขาช่อง ทุกชนิดทุกขนาด ยกเว้นซองสติ๊ก . ส่งมาที่ ตู้ ปณ.10 รังสิต ปทุมธานี 12130 ตั้งแต่วันนี้ – 12 ธ.ค.2561 . ฉีกซองมาก มีสิทธิ์ได้ทองมากนะจ๊ะ ติดตามรายละเอียดเพิ่มเติมได้ที่ Facebook Khao Shong Coffee Thailand #KhaoShongCoffee #กาแฟเขาช่อง #รสแท้กาแฟไทย
4	#ณัชชาแอนด์เดอะแก๊ง วันอาทิตย์นี้ ลุ้นระทึกไปกับเรื่องหมู หมู แต่! ไม่หมูอย่างที่คิดหล่ะสิ! รับชมความน่ารักไปกับพี่น้องแก๊งนี้กัน เช้าวันอาทิตย์ เวลา 8.30 น. #WorkPoint23
5	Live อายุน้อยร้อยล้าน ทำอย่างไรถึงประสบความสำเร็จในธุรกิจ ติดตามชมได้ที่นี่เลยค่ะ #อายุน้อยร้อยล้าน #Workpoint23 ติดตามรายการย้อนหลังได้ทาง WorkpointOfficial กดที่นี่ > https://goo.gl/NtpNjQ
6	Facebook Live บรรจงชงข่าว วันนี้ติดตามประเด็น - ตามทีมโรยตัวสำรวจพบปล่องที่ 3 ตัน! - แก๊งหนุ่ม เผย เคยเข้าถ้ำหลวง ทำสัญลักษณ์ ไว้ที่ทางตัน - รวบแล้ว! หื่น ใช้มีดจี้-ข่มขืนสาวท้อง 5 เดือน - เคลียร์ สิบล้อดั้มพ์ชนป้ายบอกทางแล้ว เสียหายล้านบาท . #WorkpointNews #บรรจงชงข่าว
7	พี่หอย... คือมันไม่ใช่อ่ะ #TheMaskSinger #Workpoint23
8	รัศมีแข ราชาแห่งแม่น้ำ... 'คิงคลองงง' อุ อุ อุ #ปริศนาฟ้าแลบ #Workpoint23 ========== ติดตามรายการย้อนหลังได้ทาง WorkpointOfficial กดที่นี่ > https://goo.gl/NtpNjQ
9	เสี่ยงแต่คงต้องยิ้มต้องสู้กันไป ไม่น่าเชื่อว่าพี่จ่อยจะเปลี่ยนแนว และไม่น่าเชื่อว่า กัปตัน แสตมป์ และ ผู้ช่วยกัปตันโอ๊ตจะ ใส่วิกแบบนี้!! น่ารักฝุดๆ #ทีมชาย #TheShow #ศึกชิงเวที #TheShowTH #Workpoint #กดเลข23
10	ไฟไหม้บ้านรับปีใหม่ เจ้าของเผยจุดธูปเทียนไหว้พระลืมดับ #ตลาดข่าวช่วงที่ 2

ตารางที่ 6 ตารางแสดงตัวอย่างโพสต์ของเพจ Workpoint Entertainment

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	(คลิป) แห่เที่ยวปีใหม่คึกคักทั่วไทย ชัยนาทปลุกเสกข้าวหลาม 4,500 กระบอกแจก - ลำปางยอดขายซีฟู้ดพุ่ง #เรื่องเล่าเช้านี้
2	[Live] เรื่องเล่าหน้าหนึ่ง วันที่ 21 ธ.ค.61 สรุปข่าวเด่นประจำวันจากเรื่องเล่าเช้านี้ และสื่อหลัก ก่อนไปตามชมกันเต็มๆได้ใน เรื่องเล่าเช้านี้ 6.00 - 8.00 น.ทางช่อง 3HD ติดตามเพิ่มเติมได้ที่ https://bit.ly/2T7Q5Zp และ https://bit.ly/2PQijpf
3	(คลิป) อุตุฯเตือนฝนตกยาวถึง 26 ก.ย. ชี้ปีนี้เหนือฝนหมดไว หนาวมาเร็ว #เรื่องเล่าเช้านี้
4	ชาวเน็ตแห่แชร์คลิป เครื่องบินสหรัฐฯ ปล่อยฝูงปลามหาศาลลงทะเลสาบห่างไกล #ต่างประเทศ #เรื่องเล่าเช้านี้
5	ตร.รวบสาวโพสต์แอบอ้างเป็นครอบครัวน้อง13คนติดในถ้ำ ขอรับบริจาคเงิน สารภาพไม่ตั้งใจจะเป็นมิจฉาชีพ แต่จะนำเงินมาช่วย พบมียอดบริจาคพันกว่าบาท นำไปซื้ออาหารมาแจก #เรื่องเล่าเช้านี้
6	[[LIVE] ฟังเสียง! ผบ.ทบ. ให้สัมภาษณ์ ระบุมีทหารร่วมค้นหา 13 ผู้สูญหายในถ้ำเพียงพอไม่จำเป็นต้องเพิ่มกำลัง #ถ้ำหลวง #เรื่องเล่าเช้านี้
7	‘โป๊ป - เบลล่า’ ควงคู่เปิดใจหลังแฟนคลับเชียร์ให้คบกัน พร้อมเผยเตรียมไปงานแฟนมีทติ้งที่ต่างประเทศ
8	ควรเลี่ยง! อุบัติเหตุบนทางหลวงหมายเลข 9 (บางปะอิน > บางพลี) กม. 34+300 รถกระบะยางแตก พลิกตะแคง เลี่ยงลง ถ.ลำลูกกา หรือ ถ.รังสิตนครนายก รถติดหนักท้ายแถว 16.30 น. ถึงด่านธัญบุรี #NewsUpdate #เรื่องเล่าเช้านี้
9	สด! ตำรวจคุมตัวครูปรีชา คาดว่าจะไปที่กองปราบปราม หลังถูกออกหมายจับ ผู้สื่อข่าว #เรื่องเล่าเช้านี้ รายงานสดจาก จ.กาญจนบุรี
10	สวัสดีปีใหม่ 2561  ขอให้ปีใหม่นำความสุขใหม่ ๆ ความสำเร็จ ความเบิกบาน ขอให้ปีใหม่เป็นปีที่วิเศษสำหรับแฟนข่าวทุกท่าน 

ตารางที่ 7 ตารางแสดงตัวอย่างโพสต์ของเพจ เรื่องเล่าเช้านี้

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	Live! ข่าวใส่ไข่ สดใหม่ ให้เยอะ
2	Thairath Talk แรงเปิดใจ ‘หมู อาซาว่า’ เสียงวิจารณ์การทำชุดนางงาม #ThairathTalk # เรื่องน่าขยี้แบบนี้ต้องคุย #ไทยรัฐ #interview #หมูอาซาว่า #Asava #นิ้งโศภิดา #มิสยูนิเวิร์ส2018 #missuniverse2018
3	ระทึก 😱 พายุไต้ฝุ่น ‘จ่ามี’ ถล่มเกาะโอกินาวา ประเทศญี่ปุ่น ความเร็วลมใกล้ศูนย์กลาง 216 กิโลเมตรต่อชั่วโมง พบบ้านเรือนพังเสียหาย บาดเจ็บ 17 ราย
4	☼ข่าวเช้าไทยรัฐ เล่าชัด ดูสบาย📺 -คิงคองยักษ์ ห้วยตึงเฒ่า แลนด์มาร์คใหม่ -ชื่นชมพนักงานรถไฟ ผลักชายขวางทางรถ -ชุดเฉพาะกิจพังงารวบแก๊งทวงหนี้โหด -วินจยย.เมินร่วมแกร็บแม้ได้เงินเพิ่ม
5	ข่าวด่วน มาแรง อัพเดททุกสถานการณ์โลก วันนี้ ✵เกาะติดปฏิบัติการ 13 ชีวิตต้องรอด การค้นหา 13 ชีวิต เด็กติดถ้ำหลวง ยังเดินหน้าต่อ แม้ฝนกระหน่ำ ล่าสุดทีมทหารสหรัฐสำรวจพื้นที่เจาะถ้ำ ด้านนายก เตรียมไปถ้ำหลวงให้กำลังใจผู้ปกครองและเจ้าหน้าที่พรุ่งนี้ ติดตาม #ไทยรัฐนิวส์โชว์ วันนี้ 20.10 น. #ไทยรัฐทีวี32
6	ข่าวด่วน มาแรง อัพเดททุกสถานการณ์โลก ในไทยรัฐนิวส์โชว์ วันนี้ ☆แก๊ง“ยันหว่าง” เหิมอีก ถ่ายคลิปแทงคน แก๊งยันหว่าง อุกอาจ บุกทำร้ายหนุ่มคาบ้าน อ้างถูกคู่กรณีหาเรื่องก่อนจึงตอบโต้ ☆โวยจับใบขับขี่ใส่กุญแจมือ ตร.แจงขัดขืน! ดราม่าสนั่น ยึดใบขับขี่ก่อนถูกใส่กุญแจมือ แจ้ง 2 ข้อหา ขับผิดเลนและขัดขวางเจ้าพนักงาน ด้านนครบาล ยันทำตามกฎหมาย ติดตามได้ใน #ไทยรัฐนิวส์โชว์ คืนนี้ 20.10 น. เป็นต้นไป ทาง #ไทยรัฐทีวี32
7	ละคร “บุพเพสันนิวาส” แรงจริง!! นายกฯ เลยให้ รมต. ท่องจำคนละบท หนังสือ 'จินดามณี' วธ.ชงใส่ ‘ชุดไทย’ ทำงาน
8	นาที 63 สโลวะเกีย ยิง แต่ล้ำหน้า! ดูสด #คิงส์คัพ ทาง #ไทยรัฐทีวี ช่อง 32 ⚽เว็บไซต์ thairath.co.th/tv/live ⚽ FB LIVE https://goo.gl/bjQ1dt ⚽ ยู ทู บ https://youtu.be/3Umw3xvAwZ0 #KingsCup2018 #บอลไทย #bnk48xchangsuek #ช้างศึก #ไทยรัฐ #ไทยรัฐทีวี32
9	อยากกินต้มเลือดหมูร้านนี้ต้อง รอคิวนะจ๊ะ
10	ในหลวง ร.10 ทรงมีพระราชดำรัส เนื่องในวารดิถีวันขึ้นปีใหม่ พุทธศักราช 2561 แก่ประชาชนชาวไทย

ตารางที่ 8 ตารางแสดงตัวอย่างโพสต์ของเพจ Thairath

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	[Live] 7 สีส่งสุขวิถีไทย รับปีใหม่วิถีธรรม ๒๕๖๒ 🙏 #Ch7HD #Ch7HDNews #7สีส่งสุขวิถีไทยรับปีใหม่วิถีธรรม #ปีใหม่2562
2	[LIVE] สด!! บรรยากาศหลังเกมฟุตบอลซูซูกิคัพ 2018 ไทย พบ สิงคโปร์ เช็คผลและตารางคะแนน >> https://www.bugaboo.tv/sport/affsuzukicup2018 #เชียร์บอลไทยกับช่อง7HD #ซูซูกิคัพ2018 #ทีมชาติไทย #Ch7HD
3	พุดดิ้งนมถั่วเหลือง สูตรเด็ดจากรายการ #VeryEasy มีส่วนผสมหลักมาจากนมถั่วเหลืองที่มีคุณสมบัติมากมาย และยังมีผลไม้ต่าง ๆ ที่นำมากินด้วยกันได้อย่างลงตัวสุด ๆ 🍊 ส้ม : มีวิตามินซีช่วยกระตุ้นการสร้างคอลลาเจน ช่วยให้ผิวพรรณสดใส และยังช่วยในระบบย่อยอาหารและระบบขับถ่ายทำงานดีขึ้นอีกด้วย 🥝 กีวี : ช่วยเสริมความแข็งแรงให้ระบบภูมิคุ้มกัน มีสารต้านอนุมูลอิสระ ช่วยฟื้นฟูเซลล์ผิว รักษาความอ่อนเยาว์ของผิวให้นานขึ้น แถมยังมีแคลอรีน้อยมาก 🍒 แก้วมังกร : ช่วยในการควบคุมระดับน้ำตาลในเลือด ช่วยลดระดับคอเลสเตอรอลชนิดที่ไม่ดีในเลือด และช่วยเพิ่มระดับคอเลสเตอรอลชนิดดีในร่างกาย 🍓 สตรอว์เบอร์รี : วิตามินซีสูง ช่วยป้องกันโรคหวัด ช่วยลดอาการภูมิแพ้ เป็นผลไม้ที่ให้พลังงานต่ำ อีกทั้งยังมีซูเปอร์ไฟเบอร์เพกทิน ที่มีส่วนช่วยในการลดระดับคอเลสเตอรอลได้ 🍒 มะม่วง : มีวิตามินเอและซี ที่มีส่วนช่วยให้ผิวพรรณเปล่งปลั่งสดใส นอกจากนี้ วิตามินซีกับสารเพคตินที่มีอยู่ในมะม่วง ยังสามารถช่วยลดระดับคอเลสเตอรอลชนิดไม่ดีในร่างกายได้อีกด้วย วิธีทำคลิก http://s.bugaboo.tv/399310 #Ch7HD
4	ผมรอให้คุณสารภาพรักผมไม่ไหวแล้ว... 😆 🤦 คุณเขตคิดแต่เรื่องดี ๆ จริ้งจริงงง #คุณเขตสายมโน ใครฟินขอให้ยกมือขึ้น! ✋ 😂 #เจ้าสาวจำยอม 💍 💐 ทุกวันศุกร์ 20.05 น. เสาร์ อาทิตย์ 20.15 น. ✔ #ช่อง7HD #กด35 LIVE สด Fanpage : @Ch7HD 📲 แอปพลิเคชั่น Ch7HD และเว็บไซต์ http://www.ch7.com/live ---------- #ละครย้อนหลัง รับชมได้ก่อนใคร ▶ https://minisite.bugaboo.tv/chaosaochamyom #Ch7HD #ไฮไลท์ละคร
5	#ไฮไลท์ละคร ท้องกับ 'ชีพ ชูชัย' หินเขียวดูหน้าพี่ชีพด้วยค่า 😂 😂 😆 #เล็บครุฑ 🤜 🎆 #ช่อง7HD ชมสด #กด35 ใครยังไม่จุใจ เช้านี้ไปชมย้อนหลังกันได้ที่ !! ▶ https://minisite.bugaboo.tv/lepkhrut เลยนะจ๊ะ #Ch7HD

ตารางที่ 9 ตารางแสดงตัวอย่างโพสต์ของเพจ Ch7HD

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
6	[LIVE] สนามข่าว 7 สี วันที่ 27 มิถุนายน 2561 - ปฏิบัติการค้นหา 13 ชีวิตในถ้ำหลวงฯ - กรมทรัพยากรธรณี ชี้ปล่องถ้ำหลวง-ขุนน้ำนางนอน - ร้องถูกโกงแชร์ทองสูญเงินกว่า 500 ล้านบาท - เร่งหาสาเหตุบีทีเอสขัดข้อง ทันข่าวช่อง 7 สี http://news.ch7.com #Ch7HDnews
7	รวมไฮไลท์มวยไทย7สี #ซูเปอร์สโลว์ #มวยไทย7สี #muaythai (3 มิถุนายน 2561) #ฉีกทุกความมันส์ อัพเดตโปรแกรม คลิปย้อนหลัง มวยไทย7 สี คลิก >> https://goo.gl/oFdXzC
8	#วันวิสาขบูชา สำหรับผู้ที่ต้องการทำบุญถวายของให้กับพระสงฆ์ หรือที่เรียกว่า การถวายสังฆทาน ควรเลือกซื้อข้าวของเครื่องใช้ที่มีประโยชน์ต่อพระสงฆ์อย่างแท้จริง 🙏✨ #BBTVCH7 #ช่อง7HDความสุขครบรสกดเบอร์เดียว35
9	#ที่นี่หมอชิต คืนนี้ พาสัมผัสบรรยากาศดินเนอร์สุดหรูเฉียดฟ้า พร้อมอัปเดทชีวิตพระเอกหน้าใสตลอดกาล 'นิว วงศกร' ห้ามพลาด !! เวลา 22.45 น. ทาง #ช่อง7HD #กด35 นะคะ 😂❤ #BBTVCH7 #ช่อง7HDความสุขครบรสกดเบอร์เดียว35
10	🎬✨ วันนี้ห้ามพลาด !! #เจาะประเด็นบันเทิง ทุกวันจันทร์ - ศุกร์ เวลา 16.00 น. ทาง ช่อง 7 HD กด 35 #BBTVCH7 #ช่อง7HDความสุขครบรสกดเบอร์เดียว35

ตารางที่ 10 ตารางแสดงตัวอย่างโพสต์ของเพจ Ch7HD (ต่อ)

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	Live คุยแซ่บshow 28 ธ.ค. 61 เปิดใจ “เมญ่า นนธวรรณ” จากนางงามสู่ความเป็นแม่!! พร้อมควงสามีและลูกชาย เปิดตัวที่แรก!! #คุยแซ่บshow #one31
2	3หนุ่มจากตระกูล "จิระอนันต์" มาชวนดูละคร #เลือดข้นคนจาง คืนนี้!! 20:45 น. ทาง #ช่องวัน31 . . บ้าจริงบรรยากาศนี้มันเหมือน นั่งฟังเจเจร้องเพลงผู้สาวขาเลาะ ประกบไอซ์ที่เล่นกีต้าร์ แล้วมีแจ็คกี้นั่งเขย่าน้ำอยู่ข้างๆ
3	เข้มข้นกว่าเดิม! แรงกว่าเดิม! เพื่อค้นหาสุดยอด "เชฟที่ดีที่สุด" #TOPCHEFTHAILAND ซีซั่น2 เริ่มวันอาทิตย์ที่ 7 ตุลาคมนี้ 18:20 น. #ช่องวัน31 อย่าลืม!! "กดไลก์เพจ" เพื่อติดตามละครข่าวสารและความบันเทิง ติดตามได้ที่ กลุ่ม one VARIETY - รายการวาไรตี้ช่องวัน กด -> http://bit.ly/2L8oGly ---------- ดูฟรี คมชัด ทั่วประเทศ ดูช่องวัน กดหมายเลข31 ชม Online สด ๆ ได้ทาง : http://bit.ly/2Kw2P7c ดูย้อนหลังได้ที่ : http://bit.ly/2OfZSKf ติดตามข่าวสารจากช่องวัน31 ดาวน์โหลดแอปพลิเคชั่น one31 : http://www.bit.ly/one31app Instagram : http://bit.ly/2OJODuB Twitter : http://bit.ly/2MiUwgK #one31 #ช่องวัน #ช่อง31 #ดูช่องวัน31 #กดหมายเลข31
4	#วันเมียแห่งชาติ คืนนี้ตอนอวสาน เมื่ออรุณาถามหัวใจตัวเองจริงๆ เธอจะเลือกใคร? ห้ามพลาด ดูสด มีเซอร์ไพรส์! #เมีย2018 #รักเลือกได้ คืนนี้เวลา 21.30 น. #ช่องวัน31 อย่าลืม!! "กดไลก์เพจ" เพื่อติดตามละครข่าวสารและความบันเทิง ติดตามได้ที่ กลุ่ม one LAKORN - ละครช่องวัน กด -> http://bit.ly/2w86kM2 ---------- ดูฟรี คมชัด ทั่วประเทศ ดูช่องวัน กดหมายเลข31 ชม Online สด ๆ ได้ทาง : http://bit.ly/2Kw2P7c ดูย้อนหลังได้ที่ : http://bit.ly/2OfZSKf ติดตามข่าวสารจากช่องวัน31 ดาวน์โหลดแอปพลิเคชั่น one31 : http://www.bit.ly/one31app Instagram : http://bit.ly/2OJODuB Twitter : http://bit.ly/2MiUwgK #one31 #ช่องวัน #ช่อง31 #ดูช่องวัน31 #กดหมายเลข31
5	เส้นทางพิชิตหนุ่มโสด... เคยไหม.. คบกันมานาน แต่พอหมดใจ ต่อให้เราจะทำดีสักแค่ไหน ถ้าเขาไม่รัก ก็คือไม่รัก.. ยังมีอยู่จริงใช่ไหม คนที่ดีๆที่จะอยู่เคียงข้างเราตลอดไป #รู้ไหมใครโสด2018 ทุกวันอาทิตย์ 2 ทุ่มตรง ทาง #ช่องวัน31 รับชม #ช่องone31 ได้ที่ทีวี หมายเลข 31 หรือดู Live ได้ 24 ชั่วโมง ทาง www.one31.net/live รับชมย้อนหลังได้ทาง > https://youtu.be/j_R1lEHtWQw

ตารางที่ 11 ตารางแสดงตัวอย่างโพสต์ของเพจ ช่อง one31

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
6	#จียอน กับ #อาร์เดอะสตาร์ เจอกันอีกครั้ง! #นี่แหล่ะความรัก StarWarสงครามดวงดาว เย็นนี้ 18:20 น. #ช่องวัน31
7	ถ่ายทอดสดการชั่งน้ำหนัก มวยรอบ MX CHAMPION
8	คุยแซ่บshow 14 มี.ค.61 - เปิดตัวสองบ่าวสุดฮ็อต “อีผิน อีแย้ม” ชีวิตจริงยิ่งกว่าละคร! - เปิดใจคุณพ่อสุดแซ่บ"สิงโต นำโชค" สั่งเมียปิดอู่ หลังมีลูกดก - Max หมูกะทะ” เปิดประสบการณ์ใหม่ของร้านหมูกระทะ! #คุยแซ่บshow #One31
9	LIVE #รักล้นๆคนเต็มบ้าน(บ้านสราญแลนด์) | 22 กุมภาพันธ์ 2561 | one31 #รักล้นๆคนเต็มบ้าน ... ฝากบ้านไว้กับพ่ออ๊อด รับชม #ช่องone31 ได้ที่ทีวี หมายเลข 31 หรือดู Live ได้ 24 ชั่วโมง ทาง www.one31.net/live #one31 Youtube Link >> https://youtu.be/GWPvQG7pils
10	เข็มขัดแชมป์ และ เงินรางวัลรวม อีก 600,000 บาท!! ใครจะได้ไปครอง? คืนนี้ลุ้นพร้อมกัน #MXMUAYXTREME Presented by #OMG เวลา 20:45 น. #ช่องวัน31 รับชม #ช่องวัน31 ได้ที่หมายเลข 31 หรือดู Live ได้ทาง www.one31.net/live

ตารางที่ 12 ตารางแสดงตัวอย่างโพสต์ของเพจ ช่อง one31 (ต่อ)

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	คำอวยพรปีใหม่ ความหมายดี เซฟหรือแชร์ให้คนที่คุณรักได้ในวันปีใหม่นี้ #สวัสดีปีใหม่2562 #HappyNewYear2019
2	#ทรัมป์โป๊ะแตก คู่แข่ง #มิ้งโป๊ะแตก
3	เรียงข่าวเล่าเรื่อง 17 ธันวาคม 2561 - สาวงามจาก "ฟิลิปปินส์" คว้ามงกุฎ "มิสยูนิเวิร์ส 2018" นิ้ง โศภิดา" เข้ารอบ 10 คนสุดท้าย - โหด! ลุงอัมพฤกษ์ถูกไม้เท้าตีลูกตาหลุด ซ้ำผลักลงน้ำ-เมียปัดทำ เอือมอีฉี่เรี่ยราด - พ่อแฟนสาว "บอล บางแก้ว" ปล่อยโฮ ไม่เชื่อลูกสำลักควันตายในบ้านหรู - ร่วมพูดคุย และ Chat สดๆกับรายการ ทาง VOOV sanook.com Channel - ** ดาวน์โหลด VOOV ได้ที่>> goo.gl/l049vb
4	ผู้ใช้ Facebook 50 ล้านบัญชีถูกแฮก บังคับล็อกอินใหม่
5	เรียงข่าวเล่าเรื่อง 28 มิถุนายน 2561 - สั่งปรับสาวโพสต์ขอเงิน ซื้ออาหารเลี้ยง จนท.ช่วย 13 ชีวิตติดถ้ำหลวง - “โอห์ม KPN” ถูกถล่มเละหลังโพสต์คิดต่าง การช่วยเหลือ 13 ชีวิตติดถ้ำหลวง - อาถรรพ์แชมป์เก่า! เกาหลีใต้ ช็อกโลกยิงเบิ้ลทดเจ็บ 2-0 เยี่ย เยอรมนี ตกรอบแรก - อินเดียครองแชมป์ ประเทศอันตรายที่สุดในโลกสำหรับผู้หญิง - ร่วมพูดคุย และ Chat สดๆกับรายการ ทาง VOOV sanook.com Channel - ** ดาวน์โหลด VOOV ได้ที่>> goo.gl/l049vb
6	เกิดมาเพิ่งเคยเห็น
7	คลื่นลูกใหม่กำลังมา Sanook Sport
8	ถอดหน้ากาก กล้วย-ยมทูต อึ้งกับความหล่อ The Mask Singer 4
9	สวยกว่ากล้องก็นางแบบชุดนี้ รวมภาพนางแบบซุปเปอร์โมเดล กับการลองกล้อง Sony A7 Mark III บนเรือหรู (อัลบั้ม)
10	19 เทคนิคง่ายๆ ที่ช่วยให้มือถือ Android ของเราทำงานเร็วและดีขึ้น

ตารางที่ 13 ตารางแสดงตัวอย่างโพสต์ของเพจ Sanook News

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	Live #เที่ยงข่าวใหญ่ 12.00-14.15 น. ประจำวันจันทร์ฐ่ 31 ธันวาคม 2561 เที่ยงนี้อยู่กับ หมอเอิง อังศ์วรา และ คุณแอน ทวีรัตน์ พร้อมด้วย ดีเจดาว ณัฐภัสสร และ ดีเจดาด้า วรินดาค่ะ
2	เจ้าหน้าที่นับพันคน บุกจับแรงงานต่างด้าวผิดกฎหมาย . #ข่าวใหญ่ไทยแลนด์ จันทร์-พฤหัสบดี 16.00-18.00 น. / ศุกร์ 16.00-18.20 น. ทางช่อง 25 #GMMnews #GMM25
3	บะหมี่เกี๊ยวหมูกรอบหมูแดง ที่หมูกรอบชิ้นใหญ่มากกกก พิเศษด้วยเส้นบะหมี่แบบกวางตุ้งเด็ดดวงขนาดนี้ เห็นแล้วน้ำลายไหล บะหมี่หัวโต หลังตลาดศรีย่าน #TEAMGIRLทีมกิน #TEAMGIRL #GMMพาเที่ยว #เที่ยวกันวันเสาร์ #GMM25
4	เล้งภูเขาไฟ...ระเบิดความแซ่บไซส์ยักษ์ ขายได้เดือนละ 60 ตัน ---------- #ข่าวใหญ่ไทยแลนด์ จันทร์-พฤหัสบดี 16.00-18.00 น. / ศุกร์ 16.00-18.20 น. ทางช่อง 25 #GMMnews #GMM25
5	ใครที่กำลังมี #ปัญหาของหัวใจ เข้ามาใกล้ๆ พี่อ้อยจะบอกวิธีแก้ไขให้ฟัง #ปัญหาของหัวใจ #GMM25
6	"อัจฉริยะ" เตือน "ทนายตั้ม" ต้องมีอุดมการณ์ ... สุดท้ายใช้คำว่าเละ! #ช็อตเด็ด #ข่าวใหญ่ไทยแลนด์ #GMMnews #GMM25
7	คนบางคน ไม่จำเป็นต้องรู้จักดีพอ ก็สมควรโดนเกลียดได้แล้วแหละ #สัมผัสรัตติกาล #GMM25 #viu /// ชมย้อนหลังทาง http://bit.ly/2u0twvB ///
8	สวัสดีวันปีใหม่ไทย สงกรานต์นี้อย่าลืมรดน้ำดำหัวผู้ใหญ่กันด้วยน้าา :) ขอให้แฟนๆช่อง GMM25 มีความสุขทุกวัน อยู่ด้วยกันทุกเวลานะคะ #สงกรานต์ #GMM25
9	ไปเที่ยวด้วยกันในฐานะแฟนครั้งแรกอะเนอะ มันก็เกร็งๆอะเนอะ 😂😂😂 #แหวนดอกไม้ #GMM25
10	น้องยังนอนไม่อิ่มเลยนะจ๊ะพี่จ๋า

ตารางที่ 14 ตารางแสดงตัวอย่างโพสต์ของเพจ GMM25Thailand

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	[Live] #เที่ยวไทยไม่ตกยุค 15.30 น. (31 ธ.ค. 61) เที่ยวไทยไม่ตกยุควันนี้ พาหนีความวุ่นวาย มาเที่ยวแบบเรียบง่าย ที่ปากน้ำประแส จ.ระยอง มาสูดอากาศบริสุทธิ์ให้ชุ่มปอด ล่องเรือชมธรรมชาติ ย้อนวันวานไปกับของโบราณสมัยคุณพ่อคุณแม่ยังเด็ก แล้วมาลิ้มรสเมนูเด็ด ***ชมออนไลน์ได้ทาง www.thaipbs.or.th/Live ***ชมย้อนหลัง www.thaipbs.or.th/Thailandintrend #ThaiPBS #ช่องหมายเลข3
2	ไปดูลีลาการการเล่นมายากลฟรี ที่จัดแสดงให้ดูในรถไฟใต้ดินของรัสเซีย #ThaiPBS #สีสันทันโลก
3	#รู้สู้ภัย วิเคราะห์ เกาะติดทุกสถานการณ์ด้านภัยต่าง ๆ กับผู้เชี่ยวชาญที่จะมาร่วมแบ่งปันความรู้ ทุกวันพฤหัสบดี เวลา 20.00-21.00 น. เริ่ม 27 ก.ย. นี้ 📌 สดเฉพาะออนไลน์ทาง Facebook www.facebook.com/ThaiPBSFan และ Youtube www.youtube.com/ThaiPBS #ThaiPBS
4	การทูตวัฒนธรรมไต้หวัน เครื่องมือยกระดับเศรษฐกิจ #ThaiPBS
5	เตือน! อย่าหลงเชื่อขอเงินบริจาค #ถ้ำหลวง จนท.ในพื้นที่เตือนประชาชน อย่าหลงเชื่อกลุ่มมิจฉาชีพ ย้ำ "ไม่มี" การเปิดรับการช่วยเหลือหรือเปิดขอรับบริจาคเงิน #ThaiPBS
6	พล.อ.ประยุทธ์ ยืนยัน การเยือนอังกฤษ -ฝรั่งเศส เป็นไปตามหน้าที่ ไม่มีการนัดใครเป็นการส่วนตัว แม้แต่การพบกับ "ทักษิณ" เพื่อต่อรอง ตกลงทางการเมือง ตามกระแส #ThaiPBS
7	ไก่ไร้หัว แต่ยังมีชีวิต ถูกส่งตัวมาโรงพยาบาลสัตว์คาดว่าอาจถูกของมีคม หรือถูกกัดมา #ThaiPBS #วันใหม่ไทยพีบีเอส
8	ย้อนรอยอโยธยา ข้าจักนำพา #ออเจ้า ไปรู้จัก "ท้าวทองกีบม้า (มารี กีมาร์)" กับชีวิตที่พลิกผัน #อยุธยาที่ไม่รู้จัก http://program.thaipbs.or.th/Ayutthaya #ThaiPBS
9	[LIVE] 07.00 - 07.30 น. #รู้เท่ารู้ทัน (12 มี.ค. 61) • เรียนรู้วิธีจัดการการนอน ปรับเปลี่ยนพฤติกรรมการนอนให้ถูกทาง เพื่อการนอนหลับที่สบายตัว ร่างกายแข็งแรง ไร้โรคภัย ดูออนไลน์ >> www.thaipbs.or.th/live ดูย้อนหลัง >> www.thaipbs.or.th/Rutan #ThaiPBS #ช่องหมายเลข3
10	บรรยากาศปีใหม่ที่ลานพระบรมราชานุสาวรีย์สามกษัตริย์ จ.เชียงใหม่ #HAPPYNEWYEAR #ThaiPBS

ตารางที่ 15 ตารางแสดงตัวอย่างโพสต์ของเพจ Thai PBS

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
1	[Live สด] ศึกมวยไทยระดับโลก “Mono 29 Topking World Series 2018” ซีซั่น 5 #TK27 Pattaya The Finals OBJOBJพบสุดยอดนักชกจากการแข่ง Thai Series และ World Series เข้มข้นไปกับทุกคู่การแข่งขัน OBJ พร้อมเชียร์ ชูเจริญ ดาบรันสารคาม นักชกไทย ให้คว้าชัยชนะ และก้าวขึ้นสู่บัลลังก์แชมป์ ! OBJ OBJ #MONO29 ถ่ายทอดสดจากแหลมบาลีฮาย เมืองพัทยา จังหวัดชลบุรี #เข้าชมฟรีตลอดงาน วันจันทร์ที่ 31 ธันวาคม ตั้งแต่เวลา 17.30 น. เป็นต้นไป OBJOBJ "เจ้าของสังเวียนมวยไทย ต้องเป็นของคนไทยเท่านั้น" OBJOBJ #MONO29 #TOPKING #TK27
2	แจกตั๋วหนังฟรี! พร้อม รีวิวหนังเข้าใหม่เรื่องไหนจะปัง เรื่องไหนจะพังดูได้เลย! หนังใหม่เข้าโรงสัปดาห์นี้ 1.Ten Years Thailand 2.I'm not here 3.Long Time No Sea 4.Journey รีวิวโดย บัณฑิต เทียนรัตน์ พิธีกรและนักข่าวสายภาพยนตร์ที่ทำงานในวงการนี้มาอย่างยาวนาน ติดตามรายการ Entertainment Now ได้ทุกวันจันทร์-ศุกร์ 10.40 น. และ 14.10 น. พิเศษสำหรับผู้ชมผ่านทางเพจMONO29 ก็สามารถร่วมสนุกเล่นเกมส์ผ่าน application mono29 ชิงตั๋วชมภาพยนตร์ ได้จนถึงเวลา 19.00 น. #MONO29 #EntertainmentNow
3	ให้กินดีๆ ก็ได้ไม่เห็นต้องถึงมีเรื่องกันขนาดนี้เลย !!! - Tai Chi O ไทเก๊ก หมัดเล็กเหล็กตันอากาศหนาวๆ แบบนี้เปิด MONO29 ดูกันฟินๆเถอะจ้าา #MONO29 #TAICHIO
4	🏆✨ศึกการแข่งขันตะกร้อไทยแลนด์ลีก ครั้งที่ 17 เพิ่งจบไปสดๆ ร้อนๆ ซึ่งกระแสตอบรับจากแฟนๆตะกร้อได้ให้ความสนใจ และสนับสนุนเป็นอย่างมาก 😊👍ทางโมโนกรุ๊ปก็จะยังเดินหน้าสนับสนุนกีฬาประจำชาติอย่างตะกร้อไทยต่อไป.. 🔥🔥 แล้วเรากลับมาพบกันได้ใหม่ในครั้งหน้า แน่นอนว่าต้องเดือด! มันส์! และยิ่งใหญ่กว่าเดิม! แฟนตะกร้อเตรียมตัวรอกันได้เลย. . .
5	วิ่งกันตั้งแต่เช้าเลยวันนี้ กับภาพยนตร์เรื่อง... "Freerunner เกรียน ซัด ฟัด"ตอนนี้เลยที่ช่อง MONO29 #Freerunner #MONO29

ตารางที่ 16 ตารางแสดงตัวอย่างโพสต์ของเพจ Mono 29

อันดับ	ข้อความโพสต์ที่ถูกเก็บในฐานข้อมูล
6	[Live สด] การแข่งขันบาสเกตบอล FIBA ASIA CUP 2021 รอบคัดเลือกโซนอาเซียน ทีมชาติไทย vs ทีมชาติมาเลเซีย 😆🎉🎉 ขอเสียงเชียร์ดังๆ เชียร์กัน ยาวไป! ยาวไป! 📣 ติดตามรับชม #MONO29 ถ่ายทอดสดรอบต่อไป ผ่านช่องโมโน หมายเลข 29 และ Facebook Live #MONO29 #เชียร์บาสไทย #BASKETBALLTHAILAND #ไทยแลนด์ปูนปูน
7	ว่ายน้ำอยู่ดีๆ สระน้ำก็กลายเป็นน้ำแข็ง !! Agents of S.H.I.E.L.D. ชี.ล.ด์. ทีมมหากาฬอเวนเจอร์ส ปี 1 ตอนนี้เลยที่ช่อง MONO29 #AgentsOfSHIELD #MONO29
8	คนสวยไม่ได้น่าอิจฉาเสมอไป สวยมักนก ตลกมักได้! เกิดขึ้นกับดาราสาวหลาย ๆ คน แต่จะมีใครกันบ้างนั้น ต้องติดตาม! ในรายการ Gossip 29 วันเสาร์ที่ 19 พฤษภาคมนี้ เวลา 14.15 น. ทางช่อง MONO29
9	เปิดรับความสนุกต้อนรับต้นปี วันหยุดเสาร์ - อาทิตย์ ที่ 10-11 กุมภาพันธ์นี้ ด้วยโปรแกรมหนังดีที่จะมาเติมความสุขถึงหน้าจอ Weekend Special ทางช่อง MONO29 วันเสาร์ที่ 10 กุมภาพันธ์นี้ เวลา 08.05 น. National Treasure ปฏิบัติการเดือด ล่าขุมทรัพย์สุดขอบโลก (ภาค1) เวลา 11.45 น. From Vegas to Macau II โคตรเซียนมาเก๊าเขย่าเวกัส 2 เวลา 15.00 น. Cast Away คนหลุดโลก เวลา 21.15 น. Maze Runner 2: The Scorch Trails สมรภูมิมอดไหม้ เวลา 23.45 น. Hidalgo ฮิดาลโก้ ฝ่านรกทะเลทราย วันอาทิตย์ที่ 11 กุมภาพันธ์นี้ เวลา 08.10 น. National Treasure : Book of Secrets ปฏิบัติการเดือดล่าบันทึกลับสุดขอบโลก (ภาค2) เวลา 11.45 น. Safe โคตรระห่ำ ทะลุรหัส เวลา 14.40 น. Knight and Day โคตรคนพยัคฆ์ร้าย กับหวานใจมหาประลัย เวลา 20.30 น. Underworld: Rise of the Lycans สงครามโค่นพันธุ์อสูร 3: ปลดแอกจอมทัพอสูร เวลา 23.20 น. The Ghost and the Darkness มัจจุราชมืด โหดมฤตยู ---------- รับชมผ่านทีวีดิจิตอล เคเบิ้ล ดาวเทียม หมายเลข 29 ช่องทางการรับข่าวสารเพิ่มเติม Instagram : Mono29TV
10	NBA เกมแรกของปีทางช่อง Mono Plus ฮอร์เน็ต บุกไปเยือน คลิปเปอร์ส และนี่คือจังหวะเข้าทำสวนงามของฮอร์เน็ต

ตารางที่ 17 ตารางแสดงตัวอย่างโพสต์ของเพจ Mono 29 (ต่อ)

ทำความสะอาดข้อมูลโพสต์

เนื่องจากข้อความในแต่ละโพสต์ ประกอบไปด้วย ตัวอักษร เว้นวรรค อักขระพิเศษ เว็บไซต์ และสัญรูปอารมณ์ (Emoticon) ซึ่งสำหรับการสกัดคำสำคัญในงานวิจัยนี้จะสนใจเพียงตัวอักษรเท่านั้น จึงจำเป็นต้องลบ เว้นวรรค อักขระพิเศษ เว็บไซต์ และสัญรูปอารมณ์ ออกจากข้อความ ดังตัวอย่างที่แสดงไว้ใน*ตารางที่ 18*

ข้อความก่อนทำความสะอาด	ข้อความหลังทำความสะอาด
ผมรอให้คุณสารภาพรักผมไม่ไหวแล้ว... 😆🐶 คุณเขตคิดแต่เรื่องดี ๆ จริ้งจริงงง #คุณเขตสายมโน ใครฟินขอให้ยกมือขึ้น! ✋😂 #เจ้าสาวจำยอม 💍👰 ทุกวันศุกร์ 20.05 น. เสาร์ อาทิตย์ 20.15 น. ✔ #ช่อง7HD #กด35 🔘 LIVE สด Fanpage : @Ch7HD 📲 แอปพลิเคชั่น Ch7HD และเว็บไซต์ http://www.ch7.com/live ---------- ---------- #ละครย้อนหลัง รับชมได้ก่อนใคร ▶ https://minisite.bugaboo.tv/chaosaochamyom #Ch7HD #ไฮไลท์ละคร	ผมรอให้คุณสารภาพรักผมไม่ไหวแล้ว คุณเขตคิดแต่เรื่องดีๆจริ้งจริงงงคุณเขตสายมโนใครฟินขอให้ยกมือขึ้น เจ้าสาวจำยอมทุกวันศุกร์2005นเสาร์อาทิตย์2015นช่อง7hdกด35livสดfanpagech7hdแอปพลิเคชั่นch7hdและเว็บไซต์ละครย้อนหลังรับชมได้ก่อนใครch7hdไฮไลท์ละคร

ตารางที่ 18 ตารางแสดงความแตกต่างระหว่างข้อความก่อนทำความสะอาด กับหลังทำความสะอาด

การทดลองใช้เครื่องมือตัดคำ deepcut และการวิเคราะห์ผลลัพธ์และข้อจำกัดที่เกิดขึ้น

หลังจากได้ข้อความที่ถูกทำความสะอาดแล้ว ผู้วิจัยได้ลองใช้เครื่องตัดคำ ดีพคัต (Deepcut) ซึ่งเป็นหนึ่งในเครื่องมือที่ใช้โครงข่ายประสาทเทียมแบบลึกที่ฝึกสอนเพื่อใช้สำหรับตัดคำในภาษาไทย โดยมีค่า F1 ถึง 98.1% ซึ่งถูกฝึกสอนโดยใช้ข้อมูล 90% จากคลังข้อมูลภาษา BEST2019 และทดสอบกับข้อมูล 10% ที่เหลือ อีกทั้งเครื่องมือ deepcut ได้ถูกเปิดเป็นโอเพนซอร์ส (Opensource) บนเว็บไซต์กิตฮับ (Github) ทุกคนสามารถนำมาใช้งานได้ โดยผู้วิจัยได้นำมาทดลองตัดคำกับข้อความในเพจเพื่อวิเคราะห์ว่าสามารถใช้กับคำต่าง ๆ ที่เกิดขึ้นบนสื่อสังคมออนไลน์เฟซบุ๊กได้หรือไม่ โดยแสดงตัวอย่างผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 19*

คลังข้อมูลภาษา BEST2009 [39] ซึ่งย่อมาจากเครื่องมือที่ใช้วัดประสิทธิภาพเพื่อเพิ่มประสิทธิภาพงานทางด้านกระบวนการทางภาษาที่เป็นมาตรฐาน (Benchmark for Enhancing the Standard of Thai language processing: BEST) ส่วน 2009 เป็นปี ค.ศ. ที่พัฒนาคลังข้อมูลภาษานี้ออกมา โดยคลังข้อมูลภาษานี้เก็บข้อความจาก 4 แหล่ง ได้แก่ บทความทั่วไป สารานุกรม ข่าวสาร

และ นวนิยาย ซึ่งไม่ได้มีข้อมูลข้อความในสื่อสังคมออนไลน์จึงอาจได้ความแม่นยำในการตัดคำลดลง หากนำมาใช้กับงานวิจัยนี้

ข้อความหลังทำความสะอาด	ข้อความที่ถูกตัดคำโดยใช้เครื่องมือ deepcut
ผมรอให้คุณสารภาพรักผมไม่ไหวแล้วคุณเขตคิดแต่เรื่องดีๆจริ้งจริงงงคุณเขตสายมโนใครฟินขอให้ยกมือขึ้นเจ้าสาวจำยอมทุกวันศุกร์2005นเสาร์อาทิตย์2015นช่อง7hdกด35liveสดfanpagech7hdแอปพลิเคชั่นch7hdและเว็บไซต์ละครย้อนหลังรับชมได้ก่อนใครch7hdไฮไลท์ละคร	ผม | รอ | ให้ | คุณสารภาพ | รัก | ผม | ไม่ | ไหว | แล้ว | คุณเขต | คิด | แต่ | เรื่อง | ดี | ๆ | จริ้ง | จริง | งง | คุณเขตสายมโน | ใคร | ฟิน | ขอ | ให้ | ยก | มือ | ขึ้น | เจ้า | สาว | จำยอม | ทุก | วัน | ศุกร์ | 2005 | นเสาร์ | อาทิตย์ | 2015 | นช่อง | 7 | hdกด | 35 | live | สดfanpagech | 7 | hd | แอปพลิเคชั่น | ch | 7 | hd | และ | เว็บไซต์ | ละคร | ย้อนหลัง | รับ | ชม | ได้ | ก่อน | ใคร | ch | 7 | hdไฮไลท์ | ละคร

ตารางที่ 19 ตารางแสดงข้อความที่ถูกตัดคำโดยใช้เครื่องมือ deepcut

เมื่อพิจารณาข้อความโพสต์จาก*ตารางที่ 19* พบว่า ข้อความมีการใช้คำที่ไม่เป็นมาตรฐาน อย่าง “จริ้งจริงงง” ซึ่งเป็นการเปลี่ยนคำว่า “จริง ๆ” ให้มีน้ำเสียงที่แสดงอารมณ์ประชดประชันมากยิ่งขึ้น อีกทั้งยังมีการเพิ่มตัวอักษร “ง” ต่อท้ายคำให้ยาวขึ้น เพื่อให้ผู้อ่านลากเสียงยาวตาม และมีการใช้คำว่า “มโน” ซึ่งเป็นคำนามที่มีอยู่ในพจนานุกรมไทย แปลว่า “ใจ” ซึ่งมักใช้เป็นคำผสมมากกว่าคำโดด เช่นคำว่า “มโนกรรม” แปลว่า “การกระทำทางใจ หรือ “มโนรม” แปลว่า “เป็นที่ชอบใจ” แต่คำว่า “มโน” ในสื่อสังคมออนไลน์อย่างเฟซบุ๊ก มักถูกใช้เป็นคำกริยา แปลว่า “ทึกทักขึ้นมาเอง” และยังมีการใช้คำว่า “ฟิน” ซึ่งมาจากคำว่า “ฟินาเล่” (Finale) ซึ่งในภาษาฝรั่งเศสแปลว่า “จบบริบูรณ์” โดยคำนี้ยังถูกเปลี่ยนความหมายเป็น “รู้สึกดี” ในสื่อสังคมออนไลน์อีกด้วย อีกทั้งโพสต์จากตารางนี้ยังมีการใช้ภาษาอังกฤษ ปนกับภาษาไทย และมีการใช้คำทับศัพท์ภาษาอังกฤษ อย่างคำว่า “ไฮไลท์” ซึ่งเป็นคำทับศัพท์มาจากคำว่า “Highlight” ที่แปลว่า “โดดเด่น” อีกด้วย

ผลจากการตัดคำโดยใช้เครื่องมือ deepcut ซึ่งเครื่องมือดังกล่าวไม่ได้ถูกฝึกสอนโดยใช้ฐานข้อมูลจากสื่อสังคมออนไลน์ ทำให้เครื่องมือดังกล่าวไม่สามารถตัดคำประโยคลักษณะนี้ให้ได้ผลดี โดยจะเห็นได้ว่า คำว่า “จริ้งจริงงง” ถูกตัดออกเป็น 3 คำ ได้แก่ “จริ้ง” “จริง” และ “งง” ส่วนคำว่า “มโน” ซึ่งไม่ได้ถูกใช้ทั่วไปในบริบทนี้บนคลังข้อมูลภาษา BEST2009 จึงทำให้คำดังกล่าวไม่ถูกตัดออกจากกัน เพราะถูกเครื่องมือเข้าใจว่าเป็นชื่อของบุคคล จึงกลายเป็นคำยาวว่า “คุณเขตสายมโน” เหมือนคำว่า “คุณสารภาพ” ซึ่งควรจะถูกตัดเป็นคำว่า “คุณ” และ “สารภาพ” ส่วนคำว่า “ฟิน” เครื่องมือตัดคำ deepcut สามารถตัดเป็นคำโดดได้ อาจเป็นเพราะอยู่ระหว่างคำว่า “ใคร” กับคำว่า

“ขอ” ซึ่งเป็นคำที่พบบ่อยในบทความทั่วไป จึงทำให้ความมั่นใจในการตัด 2 คำนี้ ออกมาเป็นคำโดดของเครื่องมือดังกล่าวมีค่าสูง ผลจึงทำให้คำว่า “ฟิน” ถูกตัดแยกออกมาเป็นอีกคำหนึ่งด้วยเช่นกัน

การทำความสะอาดโพสต์ ส่งผลให้เวลาและคำย่อ ไม่สามารถตีความได้ จึงทำให้มีความผิดพลาดเกิดขึ้น อย่าง “น.” ซึ่งย่อมาจาก “นาฬิกา” กลายเป็น “น” และเครื่องมือตัดคำ deepcut ไม่สามารถตัด “น” ออกมาเป็นอีกคำได้ จึงปรากฏคำว่า “นเสาร์” และ “นช่อง” ในผลลัพธ์ อีกทั้งตัวอักษรภาษาอังกฤษไม่ได้ถูกลบไปจากข้อความ และไม่มีเว้นวรรคช่วยแบ่งข้อความ จึงทำให้ผลลัพธ์ที่ได้เมื่อมีคำภาษาอังกฤษปนอยู่ในข้อความ มีความผิดเพี้ยนไปมาก เช่นคำว่า “สดfanpagech” ซึ่งควรถูกตัดเป็นคำว่า “สด” “fanpage” และ “ch” รวมถึงคำว่า “ไฮไลท์” เนื่องจากเป็นคำทับศัพท์ จึงทำให้เครื่องมือไม่แน่ใจ และเลือกที่จะตัดคำนี้ รวมอยู่กับคำภาษาอังกฤษ เป็นคำว่า “hdไฮไลท์”

จากผลลัพธ์นี้เอง ทำให้งานวิจัยนี้เห็นถึงปัญหาของเครื่องมือตัดคำที่ไม่ได้ถูกฝึกสอนโดยใช้คลังข้อมูลภาษาในสื่อสังคมออนไลน์ และด้วยข้อจำกัดของภาษาที่เกิดขึ้นใหม่ทุกวันในสื่อสังคมออนไลน์ การจะสร้างคลังข้อมูลภาษาที่มีประสิทธิภาพอาจเป็นเรื่องยาก ทำให้งานวิจัยนี้เสนออีกวิธีหนึ่งในการสกัดคำสำคัญที่เป็นกระแส โดยไม่ผ่านการใช้เครื่องมือตัดคำ มาเทียบกับการใช้เครื่องมือตัดคำ เพื่อเสนอเป็นอีกทางเลือกหนึ่งที่อาจจะได้ผลลัพธ์ที่ดีกว่า โดยงานวิจัยนี้เลือกใช้วิธีพิจารณาเอ็นแกรมแบบตัวอักษร (Character n-Grams) เข้ามาช่วยแก้ปัญหาแทนเครื่องมือตัดคำ

แบ่งข้อความออกเป็นชุดของตัวอักษร (แกรม)

ตอนนี้เรามีข้อมูลโพสต์ที่เก็บมาจากเพจข่าวทั้ง 10 เพจที่ถูกทำความสะอาดเรียบร้อยแล้ว และเราต้องการหาคำหรือวลี ที่มีใจความสำคัญของข้อความนั้น โดยที่เราจะไม่ใช้เครื่องมือตัดคำอย่าง deepcut เนื่องจากปัญหาที่พบกับผลลัพธ์ที่ได้จากตัวอย่างการตัดข้อความโพสต์ของเพจบนเฟซบุ๊ก

เราจึงเลือกใช้อัลกอริทึมเอ็นแกรมแบบตัวอักษร (Character n-Grams) โดยเอ็นแกรมแบบตัวอักษรเป็นการแบ่งข้อความออกเป็นส่วนย่อย ๆ เรียกว่าแกรม โดยแต่ละแกรมจะมีความยาวของตัวอักษรในแกรมทั้งหมด n ตัว โดยการแบ่งข้อความนั้น จะมีลักษณะค่อย ๆ หั่นข้อความออกเป็นแกรมยาว n ตัว ไปทีละตัวอักษรของข้อความ กล่าวคือ หากข้อความยาว k ตัวอักษร เราจะได้แกรมทั้งหมด k-n+1 แกรม โดยมีตัวอย่างของผลลัพธ์ที่ใช้วิธีเอ็นแกรมแบบตัวอักษร แสดงไว้ใน*ตารางที่ 20*

ข้อความหลังทำความสะอาด	แกรมที่เกิดขึ้นเมื่อเลือก n = 5
ผมรอให้คุณสารภาพรักผมไม่ไหวแล้ว	ผมรอใ | มรอให | รอให้ | อให้ค | ให้คุ | ห้คุณ | ้คุณส | คุณสา | ุณสาร | ณสารภ | สารภา | ารภาพ | รภาพร | ภาพรั | าพรัก | พรักผ | รักผม | ักผมไ | กผมไม | ผมไม่ | มไม่ไ | ไม่ไห | ม่ไหว | ่ไหวแ | ไหวแล | หวแล้ | วแล้ว

ตารางที่ 20 ตารางแสดงตัวอย่างของผลลัพธ์ เมื่อใช้วิธีเอ็นแกรมแบบตัวอักษรโดยเลือก n = 5

การวิเคราะห์ปัญหาจากการใช้อัลกอริทึมเอ็นแกรมแบบตัวอักษรแทนที่จะใช้เครื่องมือตัดคำ

หนึ่งในปัญหาที่งานวิจัยนี้จำเป็นต้องแก้คือ เมื่อเราใช้อัลกอริทึมอย่างเอ็นแกรมแบบตัวอักษรแทนการตัดคำแล้ว เราจะรวมแกรมเหล่านั้นกลับมาเป็นคำได้อย่างไร ซึ่งหากเป็นวิธีการสกัดคำสำคัญโดยทั่วไปที่ใช้การตัดคำ ก็สามารถใช้อัลกอริทึมอย่าง TF-IDF เพื่อหาคำสำคัญออกจากข้อความได้โดยง่าย ซึ่งหากเราเอาหลักการของ TF-IDF มาใช้กับเอ็นแกรมแบบตัวอักษรโดยตรง เราจะได้เอ็นแกรมแบบตัวอักษรที่น่าจะมีความสำคัญแทน โดยอ้างอิงจากเอ็นแกรมแบบตัวอักษรที่มีค่า TF-IDF สูง แต่ก็นำไปใช้ต่อได้ยาก

เมื่อเปรียบเทียบการรวมเอ็นแกรมแบบตัวอักษรให้กลายเป็นคำ ก็เหมือนกับการต่อจิ๊กซอว์ 2 แผ่นเข้าด้วยกัน โดยเราสามารถนำจิ๊กซอว์ที่ใกล้เคียงกันมาประกบเข้าด้วยกันได้ แต่ทว่าหากเรารวมเอ็นแกรมแบบตัวอักษร 2 เอ็นแกรมเข้าด้วยกันไปเรื่อย ๆ โดยไม่มีจุดอ้างอิงใด ๆ ย่อมเกิดคำเป็นคำใหม่ที่อาจจะไม่ได้มีอยู่ในข้อความจริง ๆ หรืออาจจะยาวจนไม่รู้จบก็ได้ เพราะเกิดวงวนของการต่อกันของเอ็นแกรมแบบตัวอักษรขึ้น ดังนั้นในการรวมเอ็นแกรมแบบตัวอักษรเข้าด้วยกันนั้น เราจำเป็นต้องใช้โพสต์ตั้งต้นของแต่ละแกรมเหล่านั้นมาอ้างอิงด้วย

อีกทั้งจำนวนตัวอักษรของแกรมก็มีผลต่อการรวม กล่าวคือยิ่งแต่ละแกรมมีตัวอักษรเยอะเท่าใด ก็แปลว่าสามารถต่อเอ็นแกรมแบบตัวอักษร 2 เอ็นแกรมเข้าด้วยกันด้วยความมั่นใจว่าข้อความเดิมมันน่าจะอยู่ติดกันได้มากเท่านั้น แต่ในมุมกลับกัน การเลือกตัวอักษรในแกรมที่มากเกินไป เป็นการจำกัดขอบเขตของคำที่คำสำคัญลง เพราะแต่ละแกรมที่เกิดขึ้นจะมีเอกลักษณ์ของตัวเองสูงมาก เนื่องจากเอ็นแกรมแบบตัวอักษรคือส่วนประกอบของคำที่ไปจับคู่กับคำอื่น ๆ อย่างมีลำดับอยู่ โดยแสดงตัวอย่างการใช้เอ็นแกรมแบบตัวอักษรโดยเลือก n แบบต่าง ๆ ไว้ใน*ตารางที่ 21*

n = 1	n = 2	n = 3	n = 4	n = 5	n = 6
ก	กา	กาม	กามเ	กามเท	กามเทพ
า	าม	ามเ	ามเท	ามเทพ	
ม	มเ	มเท	มเทพ		
เ	เท	เทพ			
ท	ทพ				
พ					

ตารางที่ 21 ตารางแสดงตัวอย่างการใช้เอ็นแกรมแบบตัวอักษรโดยเลือก n ตั้งแต่ 1 ถึง 6

จาก*ตารางที่ 21* จะเห็นได้ว่า หากเราเลือก n = 1 เราไม่สามารถรวมแกรมเข้าด้วยกันได้เลย แต่หากเราเลือก n = 2 แล้วสมมติมีเอ็นแกรมแบบตัวอักษรอีกตัวที่มีค่า TF-IDF สูง เช่น “าก” เราอาจจะรวมเข้ากับเอ็มแกรม “กา” ได้ไม่รู้จบ กล่าวคือเราจะได้คำว่า “กากากา...” และ “ากากาก...” แต่เมื่อเราพิจารณาค่า n ที่มากขึ้น การรวมแกรมก็จะทำได้ง่ายขึ้น เพราะแต่ละแกรมมีเอกลักษณ์ที่มากขึ้น แต่หากมากเกินไปแทนที่เราจะได้แกรมของคำที่ผสมกัน เราจะได้แกรมของกลุ่มคำที่ใหญ่ขึ้นแทน ซึ่งส่งผลให้เราสกัดคำสำคัญที่เล็กลงมาไม่ได้ โดยแสดงตัวอย่างไว้ใน*ตารางที่ 22*

ข้อความหลังทำความสะอาด	แกรมที่เกิดขึ้นเมื่อเลือก n = 10
ผมรอให้คุณสารภาพรักผมไม่ไหวแล้ว	ผมรอให้คุณ | มรอให้คุณส | รอให้คุณสา | อให้คุณสาร | ให้คุณสารภ | ห้คุณสารภา |้คุณสารภาพ | คุณสารภาพร | ุณสารภาพรั | ณสารภาพรัก | สารภาพรักผ | ารภาพรักผม | รภาพรักผมไ | ภาพรักผมไม | าพรักผมไม่ | พรักผมไม่ไ | รักผมไม่ไห |ักผมไม่ไหว | กผมไม่ไหวแ | ผมไม่ไหวแล | มไม่ไหวแล้ | ไม่ไหวแล้ว

ตารางที่ 22 ตารางแสดงตัวอย่างของผลลัพธ์ เมื่อใช้วิธีเอ็นแกรมแบบตัวอักษรโดยเลือก n = 10

จาก*ตารางที่ 22* หากคำว่า “สารภาพรัก” เป็นคำสำคัญของโพสต์นี้ เราจะไม่สามารถสกัดคำว่าสารภาพรักออกมาได้ เพราะคำว่าสารภาพรักอยู่ในแกรม “ณสารภาพรัก” และ “สารภาพรักผ” ซึ่งคำว่าสารภาพรักอาจไม่ได้เกิดขึ้นติดกับคำว่า “คุณ” และ “ผม” บ่อยมากพอที่จะทำให้แกรม “ณสารภาพรัก” และ “สารภาพรักผ” เป็นแกรมที่มีค่า TF-IDF สูงเพียงพอก็เป็นไปได้

จาก 2 เหตุผลข้างต้น เราจึงไม่ควรเลือก n ที่มีค่าน้อยมากเกินไปจนทำให้แกรมที่เกิดขึ้นไม่มีเอกลักษณ์ และไม่ควรเลือก n ที่มีค่ามากเกินไปเพราะจะทำให้แกรมมีเอกลักษณ์สูงเกินคำสำคัญที่เราต้องการจะสกัดออกมา ซึ่งจากการพิจารณาคำโดดในสื่อสังคมออนไลน์พบว่า โดยปกติคำโดดจะยาวโดยเฉลี่ยประมาณ 3 ถึง 4 ตัวอักษร ซึ่งคำหรือวลีที่สำคัญมักจะประกอบจากคำโดดตั้งแต่ 2 คำขึ้นไป ซึ่งหากเราไม่ต้องการให้แกรมมีเอกลักษณ์ที่มากจนเกินไป ก็ไม่ควรเลือกแกรมที่มีส่วนประกอบของคำที่ยาวเกิน 2 คำ ในงานวิจัยนี้จึงเลือกใช้ n = 5 มาใช้ในการแบ่งข้อความออกเป็นแกรมย่อย ๆ

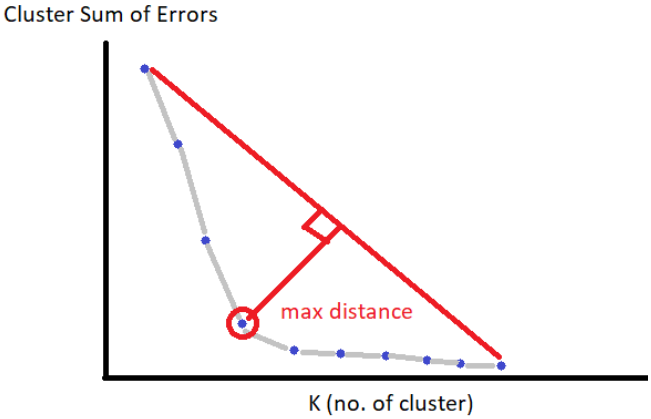
นับความถี่ของแกรม

เนื่องจากแกรมที่เราแบ่งออกมามีเอกลักษณ์ของการนำส่วนของคำมากกว่า 1 คำมาเชื่อมกัน ทำให้เราไม่สามารถแน่ใจได้ว่า วิธีอย่าง TF-IDF จะให้ผลดีกับแกรมเหล่านี้หรือไม่ อีกทั้ง TF-IDF ยังประมวลผลได้ช้า และแกรมที่แบ่งออกมามีจำนวนมากการการตัดคำอยู่หลายเท่า เราจึงพิจารณาเลือกใช้เพียงการนับความถี่ของแกรมหรือการหาค่า TF ของแต่แกรมแทนซึ่งประมวลผลได้เร็วกว่า โดยในการหาค่า TF ของแกรมนั้น เราได้หาค่า TF แบ่งตามวันที่และเพจ เพื่อให้เราสามารถใช้วิเคราะห์หาคำสำคัญและคำที่เป็นกระแสต่อไปได้

หาแกรมที่มีคุณสมบัติเป็นคำสำคัญของเพจ

ในตอนนี้เราต้องการหาแกรมที่โอกาสรวมเป็นคำสำคัญของแต่ละเพจ ซึ่งการที่แกรมเหล่านั้นจะมีโอกาสเป็นคำสำคัญได้ โดยแกรมที่เราคาดว่าน่าจะสำคัญ อาจจะมีค่า TF มากหรือน้อยก็ได้ แต่ก็ไม่ควรจะค่า TF น้อยเกินไป เพราะหากพิมพ์ถึงในประโยคน้อยเกินไปก็ดูจะสำคัญไม่มากพอ โดยหากเราเลือกหาค่าขีดแบ่ง (Threshold) บนค่า TF ค่าหนึ่ง เพื่อมาเลือกแกรมที่ดูสำคัญอาจจะเป็นไปได้ยาก เพราะจะทำให้คำว่ามากและน้อยอึงกับค่าเหล่านี้มากเกินไป เนื่องจากแต่ละเพจและแต่ละโพสต์มีลักษณะเฉพาะตัวที่แตกต่างกันออกไป การเลือกค่าขีดแบ่งเพียงค่าเดียวจึงทำได้ยาก

ในจุดนี้ อัลกอริทึมที่สามารถแบ่งกลุ่มแกรมออกเป็นกลุ่มย่อย ๆ ได้ โดยอ้างอิงค่า TF ของแต่ละแกรมอย่าง การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means Clustering) สามารถนำมาช่วยแก้ปัญหานี้ได้ เนื่องจากเราไม่จำเป็นต้องใช้ค่าขีดแบ่งที่คงที่ในการตัดแกรมที่ไม่น่าจะสำคัญทิ้งไป แต่ถึงแบบนั้นการแบ่งกลุ่มข้อมูลแบบเคมีน ก็ยังมีข้อจำกัด เนื่องจากเราจำเป็นต้องระบุจำนวนกลุ่มที่จะแบ่ง แต่เราไม่แน่ใจว่าเราควรแบ่งข้อมูลแกรมของเราออกเป็นกี่กลุ่มดี ซึ่งปัญหานี้สามารถแก้ไขได้โดยการใช้วิธีข้อศอก (Elbow Method) ซึ่งสามารถช่วยเราเลือกจำนวนกลุ่มของข้อมูลที่มีนัยสำคัญทางสถิติได้ โดยเราจะสร้างกราฟแสดงความสัมพันธ์ระหว่างจำนวนกลุ่ม กับค่าความแปรปรวน ซึ่งแปรตามผลรวมของระยะห่างระหว่างสมาชิกในกลุ่มกับจุดศูนย์กลางกลุ่ม โดยในช่วงแรกของการแบ่งกลุ่มข้อมูลจะมีค่าความแปรปรวนลดลงอย่างรวดเร็ว แต่พอเราพยายามแบ่งกลุ่มออกเป็นกลุ่มย่อย ๆ เพิ่มมากขึ้น จนมากเกินไป กล่าวคือเราแบ่งกลุ่มที่ค่อนข้างอยู่รวมกันอยู่แล้ว ออกเป็นกลุ่มย่อยอีก ค่าความแปรปรวนจะเริ่มไม่เปลี่ยนแปลง ซึ่งจำนวนกลุ่มสุดท้ายก่อนที่จะทำให้ค่าความแปรปรวนเริ่มไม่เปลี่ยนแปลง ก็คือจุดหักศอกของกราฟ ที่เราจะนำมาใช้เป็นจำนวนกลุ่มนั่นเอง โดยเพื่อลดความกำกวมของจุดหักศอกนี้ เราจะหาจุดที่มีค่าระยะห่างระหว่างจุดดังกล่าวกับเส้นตรงของจุดแรกกับจุดสุดท้ายที่มีระยะทางตั้งฉากที่มากที่สุด เพื่อใช้เป็นจำนวนกลุ่มที่เราจะแบ่ง ดัง*รูปภาพที่ 21*



รูปภาพที่ 21 รูปภาพแสดงวิธีการเลือกจำนวนกลุ่มโดยใช้วิธีข้อศอก โดยแกนนอนแสดงจำนวนกลุ่ม แกนตั้งแสดงค่าความแปรปรวน และจุดหักศอกจะมีค่าระยะทางจากจุดถึงเส้นมากสุด

โดยงานวิจัยนี้ นำแกรมของแต่ละเพจที่ถูกนับความถี่ไว้ในแต่ละวัน มาแบ่งกลุ่มแบบเคมีน ตั้งแต่ K = 1 ถึง K = 10 และนำมาพล็อตเป็นกราฟ เพื่อใช้วิธีข้อศอกในการเลือกค่า K ที่เหมาะสม สำหรับแบ่งกลุ่มของข้อมูล โดยหลังจากเราได้ค่า K ที่เหมาะสมแล้ว เราก็จะแบ่งแกรมออกเป็นทั้ง K กลุ่ม โดยแต่ละกลุ่มจะมีจุดศูนย์กลางของกลุ่ม คือค่า TF โดยเฉลี่ยของกลุ่มนั้น ซึ่งเราจะเลือกนำกลุ่มที่จุดศูนย์กลางกลุ่มมีค่าน้อยที่สุด ออกจากการคำนวณในลำดับถัดไป เนื่องจากเป็นกลุ่มที่ประกอบจากตัวอักษรที่ถูกพูดถึงน้อยเกินไป ซึ่งจะทำให้เราได้แกรมที่น่าจะมีความสำคัญของแต่ละเพจ ในแต่ละวันออกมา และในลำดับถัดไป เราจะเริ่มพิจารณาแกรมที่น่าจะมีความสำคัญในหลากหลายเพจ

หาแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ

ในขั้นตอนนี้ เราต้องการหาแกรมที่น่าจะรวมเป็นคำที่เป็นกระแสของวันนั้น โดยแกรมที่น่าจะเป็นกระแสนั้น ย่อมเป็นแกรมที่มีเพจข่าวพูดถึงหลาย ๆ เพจ ในวันเดียวกัน แต่เนื่องจากแต่ละเพจมีปริมาณการโพสต์ที่ไม่เท่ากัน ทำให้หากเรานำค่า TF ของแกรม ของแต่ละเพจมาเปรียบเทียบกัน จะไม่สมเหตุสมผล เราจึงไม่สามารถนำค่า TF เดิมของแต่ละแกรมมาเปรียบเทียบข้ามเพจได้

โดยเราจะพิจารณาจำนวนเพจที่แกรมเหล่านั้นถูกพูดถึงแทน (Document Frequency: DF) โดยจะทำให้วิธีนี้มีความแตกต่างจากการหาคำสำคัญด้วย TF-IDF เล็กน้อย เนื่องจากโดยปกติหากเราต้องการหาคำสำคัญด้วยวิธี TF-IDF เราจะตั้งต้นที่คำ และหาคำที่มีความถี่สูง แต่ปรากฏเพียงไม่กี่เอกสาร เพราะหากปรากฏหลายเอกสาร คำเหล่านั้นมักเป็นคำหยุดมากกว่าคำสำคัญ แต่ทว่าในงานวิจัยนี้ เราตั้งต้นที่แกรมซึ่งการเลือก n = 5 นั้นเป็นการตัดคำโดดส่วนมากซึ่งรวมถึงคำหยุดจำนวนหนึ่งออกไปแล้ว ทำให้เมื่อเราใช้วิธีอย่าง TF ที่มีการกรองแกรมที่ไม่สำคัญบางส่วนออกไปแล้ว เราจะได้แกรมที่มีโอกาสเป็นคำสำคัญอยู่แล้ว เพราะคำสองคำใด ๆ ที่เกิดติดกันและถูกพูดถึงบ่อย มีโอกาส

เป็นคำสำคัญกว่าคำทั่ว ๆ ไป ซึ่งการที่เราจำเป็นต้องหาค่า DF นั้น เพื่อหาคำที่เป็นกระแสของแต่ละวันออกจากคำสำคัญที่พบบ่อยเพียงแค่ในเพจตัวเอง

โดยเมื่อเราได้ค่า DF ของแต่ละแกรม เราจะนำวิธีอย่างการแบ่งกลุ่มข้อมูลแบบเคมีน และวิธีข้อศอก มาใช้ในการหาจำนวนกลุ่มเพื่อแบ่งกลุ่มแกรมเหล่านั้นเช่นเดิม เนื่องจากเราต้องการใช้ค่าขีดแบ่งในการแบ่งกลุ่มที่มีความสำคัญกับไม่มีความสำคัญออกจากกัน โดยหลังจากที่เราแบ่งกลุ่มของแกรมเสร็จเรียบร้อยแล้ว เราก็จะเลือกกลุ่มที่มีจุดศูนย์กลางกลุ่มที่แทนค่าเฉลี่ยของค่า DF ของกลุ่มนั้นที่มากที่สุดมาเป็นกลุ่มที่เราสนใจ เนื่องจากเป็นกลุ่มที่แทนแกรมที่ความสำคัญและยังถูกพูดถึงในหลาย ๆ เพจอีกด้วย แต่อย่างไรก็ดี แกรมเหล่านี้อาจไม่ใช่แกรมของคำสำคัญทั้งหมด อาจเป็นแกรมของคำหยุดก็เป็นไปได้ เนื่องด้วยการเลือก n = 5 ไม่สามารถกรองคำหยุดที่ยาวมากกว่า 4 ตัวอักษรได้ และแกรมเหล่านี้ย่อมปรากฏอยู่ในหลายเพจเช่นกัน จึงมีค่า DF สูงเหมือนกับแกรมที่เป็นกระแส ซึ่งเราจะทำการแยกคำสำคัญที่เป็นกระแสและคำหยุดออกจากกันในขั้นตอนถัด ๆ ไป

รวมแกรมที่มีคุณสมบัติเป็นคำสำคัญที่ไม่ขึ้นกับเพจใด ๆ กลับมาเป็นคำสำคัญ

ในขั้นตอนนี้ เราจะเหลือแกรมที่เราสนใจในแต่ละวันออกมาไม่มากแล้ว ซึ่งเราจะเริ่มรวมแกรมเหล่านี้กลับมาเป็นคำอีกครั้ง โดยเราอาจจะนำแกรมเหล่านี้มาต่อกันในลักษณะเดียวกับการต่อจิ๊กซอว์ย่อมทำได้ แต่อาจจะทำให้เกิดวงวนของการรวมแกรมขึ้น หรือเกิดคำใหม่ที่ไม่ได้มีอยู่จริงในโพสต์ขึ้น ตามที่ได้อธิบายไว้ใน “การวิเคราะห์ปัญหาจากการใช้วิธีเอ็นแกรมแบบตัวอักษรแทนที่จะใช้เครื่องมือตัดคำ” จึงจำเป็นต้องนำประโยคจากโพสต์เดิมมาใช้อ้างอิงในการรวมแกรมกลับมาเป็นคำ

โดยเมื่อเราตั้งต้นที่ข้อความในโพสต์ เราจะเริ่มไล่ไปทีละตัวอักษรของโพสต์ โดยเราต้องการรู้ว่าตัวอักษรนี้เป็นส่วนหนึ่งของคำสำคัญที่เกิดจากแกรมนั้นหรือไม่ เราก็จะพิจารณาเอ็นแกรมที่เกี่ยวข้องกับตัวอักษรนั้น ว่าเอ็นแกรมดังกล่าวมีอยู่กลุ่มของแกรมกลุ่มที่มีค่าเฉลี่ย DF สูงสุดหรือไม่ โดยแสดงตัวอย่างไว้ใน*รูปภาพที่ 22*

Global Keyword Grams	{('ด', 'ี', 'ค', 'ร', 'ั')}

Message	5 Character Grams
	('ส', 'ว', 'ั', 'ส', 'ด')
	('ว', 'ั', 'ส', 'ด', 'ี')
	('ั', 'ส', 'ด', 'ี', 'ค')
	('ส', 'ด', 'ี', 'ค', 'ร')
สวัสดีครับผม	('ด', 'ี', 'ค', 'ร', 'ั')
	('ี', 'ค', 'ร', 'ั', 'บ')
	('ค', 'ร', 'ั', 'บ', 'ผ')
	('ร', 'ั', 'บ', 'ผ', 'ม')

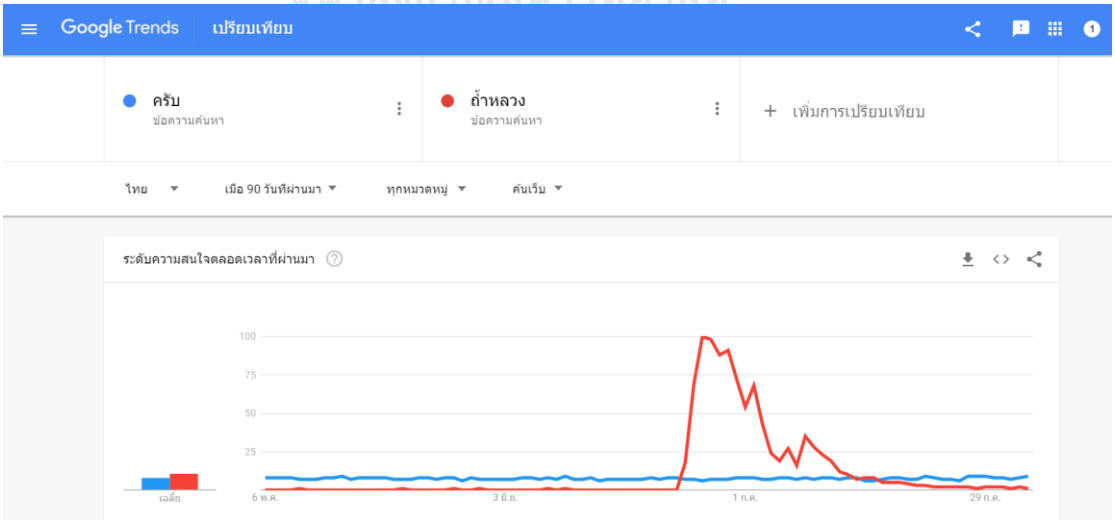
รูปภาพที่ 22 รูปภาพแสดงการพิจารณาตัวอักษร “ค” ว่าเป็นส่วนหนึ่งของคำสำคัญหรือไม่

จาก*รูปภาพที่ 22* พบว่ามีแกรม “ดีครั” ซึ่งอยู่ในแกรมที่ค่าเฉลี่ย DF สูงสุด (ในรูปเรียกแทนแกรมกลุ่มนี้ว่า Global Keyword Grams) จึงทำให้ตัวอักษร “ค” เป็นส่วนหนึ่งของคำสำคัญ โดยเมื่อเราพิจารณาครบทุกตัวอักษรของโพสต์แล้ว เราจะสามารถสกัดคำสำคัญที่เกิดจากการติดกันของตัวอักษรที่เป็นส่วนหนึ่งของคำสำคัญออกมาจากข้อความได้ ดังตัวอย่างที่แสดงไว้ใน*รูปภาพที่ 23*

รูปภาพที่ 23 รูปภาพแสดงการสกัดคำสำคัญจากตัวอักษรที่สำคัญที่ติดกัน

สกัดคำหยุดออกจากคำสำคัญที่เป็นกระแส

หลังจากเราสกัดคำสำคัญที่เป็นกระแสออกจากประโยคในแต่ละวันเรียบร้อยแล้ว แต่เนื่องจากคำสำคัญที่เราสกัดมาอาจมีคำหยุดประปนอยู่ด้วย เราจึงจำเป็นต้องแยกประเภทของคำทั้ง 2 ชนิดนี้ออกจากกันก่อน โดยเราจะใช้คุณสมบัติของความเป็นกระแสเข้ามาจำแนก กล่าวคือ คำที่เป็นกระแสนั้น มักจะปรากฏบ่อยจนเป็นกระแสอยู่เพียงไม่นาน เมื่อกระแสหมดลง คำเหล่านี้ก็จะไม่ปรากฏบ่อยเหมือนเดิม ซึ่งแตกต่างจากคำหยุดที่หากหลุดประปนออกมาก็จะไม่ขึ้นกับช่วงเวลา ซึ่งทำให้เราสามารถสร้างฐานข้อมูลของคำหยุดได้จากคุณสมบัตินี้นั่นเอง โดยมีตัวอย่างลักษณะดังกล่าวที่ใช้ข้อมูลอ้างอิงจากเว็บไซต์ Google Trends แสดงไว้ใน*รูปภาพที่ 24*



รูปภาพที่ 24 รูปภาพแสดงลักษณะของคำหยุด เทียบกับ ลักษณะของคำสำคัญที่เป็นกระแส โดยอ้างอิงข้อมูลจากฐานข้อมูลของ Google Trends

จาก*รูปภาพที่ 24* จะเห็นว่ากระแสของคำว่า “ถ้ำหลวง” เดิมแทบไม่ปรากฏเลย เทียบกับคำว่า “ครับ” ซึ่งเป็นคำลงท้ายประโยค มักพบอยู่เสมอ แต่เมื่อคำว่า “ถ้ำหลวง” กลายเป็นกระแสขึ้นมา ความถี่ของคำนี้ก็เพิ่มขึ้นอย่างเห็นได้ชัด แต่พอกระแสเริ่มลดลง ความถี่ของคำนี้ก็จะลดลงจนกลับไปปรากฏเท่าเดิม แต่คำว่า “ครับ” มีจำนวนการปรากฏที่ค่อนข้างคงที่เสมอในทุกช่วงเวลา

ซึ่งจากคุณสมบัตินี้เอง เราจึงนำคำสำคัญทั้งหมดที่สกัดออกมาได้ในแต่ละวัน มาพิจารณาสร้างฐานข้อมูลของคำหยุด โดยหากคำสำคัญใดปรากฏขึ้นโดยไม่อิงกับช่วงเวลา ก็จะนำมาเพิ่มในฐานข้อมูลคำหยุด โดยในงานวิจัยนี้ เลือกพิจารณาจากการที่คำสำคัญเหล่านั้น ยังคงปรากฏอยู่แม้ว่าจะผ่านมาเกิน 1 เดือนแล้ว กล่าวคือ พบคำสำคัญเดียวกันเคยปรากฏในช่วงก่อนหน้านั้น 1 เดือน แต่ไม่ถึง 6 เดือน เพื่อป้องกันการเลือกคำหยุดผิดเพราะคำนั้นกลับมาเกิดเป็นกระแสใหม่อีกครั้ง

สกัดคำสำคัญที่เป็นกระแสออกจากคำสำคัญ

หลังจากที่เรามีฐานข้อมูลคำหยุดแล้ว คำสำคัญในแต่ละวันที่เหลืออยู่ ซึ่งไม่อยู่ในฐานข้อมูลคำหยุดที่เราสร้างขึ้นมา ย่อมเป็นคำสำคัญที่เป็นกระแสที่เกิดขึ้นจริง โดยเราสามารถเลือกกรองคำสำคัญที่เป็นกระแสประจำวันออกได้ หากเราต้องการหากระแสที่เกิดขึ้นมากกว่า 1 วัน โดยพิจารณาว่ามีคำสำคัญดังกล่าวในช่วงเวลาใกล้เคียงกันหรือไม่

โดยในงานวิจัยนี้ ได้เลือกใช้ตัวกรองดังกล่าว โดยคำสำคัญที่เป็นกระแสที่เกิดขึ้น ต้องเคยเกิดขึ้นแล้วย้อนหลังตั้งแต่ 1 วัน ถึง 10 วัน ก่อนหน้า ถึงจะถูกนำมาคิดมาเป็นคำสำคัญที่เป็นกระแส

การทดลอง และอภิปรายผล

ในงานวิจัยนี้ ได้มีการทดลองโดยใช้ข้อมูลโพสต์จากเพจข่าวบนเฟซบุ๊กทั้งหมด 10 เพจ ซึ่งเป็นเพจที่มีผู้ติดตามสูงที่สุด 10 อันดับแรกในประเทศไทย โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม พ.ศ. 2561 จนถึง 31 มีนาคม พ.ศ. 2562 ตามเวลาประเทศไทย มาผ่านกระบวนการทำข้อความสะอาดโพสต์ การใช้เอ็นแกรมแบบตัวอักษรเพื่อจำแนกแกรมที่มีความสำคัญออกมาในแต่ละวัน และรวมแกรมเหล่านั้นกลับมาเป็นคำสำคัญ โดยมีการใช้คุณสมบัติของความเป็นกระแสที่ขึ้นกับช่วงเวลา เพื่อมาสร้างฐานข้อมูลคำหยุด และสกัดเอาคำสำคัญที่เป็นกระแสออกจากคำหยุด โดยพบว่ามีตัวแปรหลายตัวที่สามารถปรับได้ ซึ่งอาจเกี่ยวเนื่องกับผลลัพธ์และประสิทธิภาพของงานวิจัยครั้งนี้ โดยมีหัวข้อดังนี้

1. การทดลองที่เกี่ยวข้องกับวิธีที่วิทยานิพนธ์ฉบับนี้ได้นำเสนอ
    a. การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุด
    b. การวัดความถูกต้องของคำหยุด
    c. การเพิ่มกฎของคำหยุด เพื่อเพิ่มประสิทธิภาพในการสกัดคำสำคัญ
    d. การวัดความถูกต้องของคำสำคัญที่เป็นกระแส
2. การทดลองที่เกี่ยวข้องกับวิธีที่ใช้เปรียบเทียบผลลัพธ์ - วิธี TF-IDF
    a. การเปรียบเทียบผลเมื่อใช้วิธี TF-IDF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร
    b. การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุดเมื่อใช้วิธี TF-IDF
    c. การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF-IDF
3. การทดลองที่เกี่ยวข้องกับวิธีที่ใช้เปรียบเทียบผลลัพธ์ - วิธี TF
    a. การเปรียบเทียบผลเมื่อใช้วิธี TF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร
    b. การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF
4. การเปรียบเทียบผลลัพธ์
    a. การเปรียบเทียบประสิทธิภาพของผลลัพธ์ของวิธีเอ็นแกรมแบบตัวอักษร กับวิธีที่ใช้เครื่องมือในการตัดคำอย่าง deepcut ได้แก่วิธี TF-IDF และวิธี TF

การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุด

ในขั้นตอนนี้ เราอยากทราบว่า ปริมาณของโพสต์ที่เรานำมาสร้างเป็นฐานข้อมูลคำหยุดนั้น ทำให้เราได้ฐานข้อมูลคำหยุดเพิ่มขึ้นจริงหรือไม่ และมีลักษณะการเพิ่มขึ้นแบบไหน อีกทั้งจำนวนวันก่อนหน้าที่เราใช้เพื่อพิจารณาว่าคำนั้นเคยปรากฏมาก่อนแล้วหรือไม่ ซึ่งทำให้เราจำแนกคำหยุดได้นั้น มีผลอย่างไรต่อปริมาณคำหยุดที่เราได้อย่างไร

ในการทดลองนี้ จึงมีการทดลองปรับตัวแปรของจำนวนเดือนที่เรานำมาสร้างฐานข้อมูลคำหยุด ตั้งแต่ 3 เดือน 6 เดือน 9 เดือน 12 เดือน และ 15 เดือน และยังเปรียบเทียบด้วยจำนวนวันที่ใช้อ้างอิงตั้งแต่ 30-60 วัน 30-90 วัน 30-120 วัน และ 30-150 วัน โดยมีแสดงอยู่ใน*ตารางที่ 23 ถึง 26*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	132 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	105 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	90 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	56 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	16 คำ

ตารางที่ 23 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	111 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	99 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	84 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	55 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	16 คำ

ตารางที่ 24 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 120 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	80 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	65 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	50 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	34 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	16 คำ

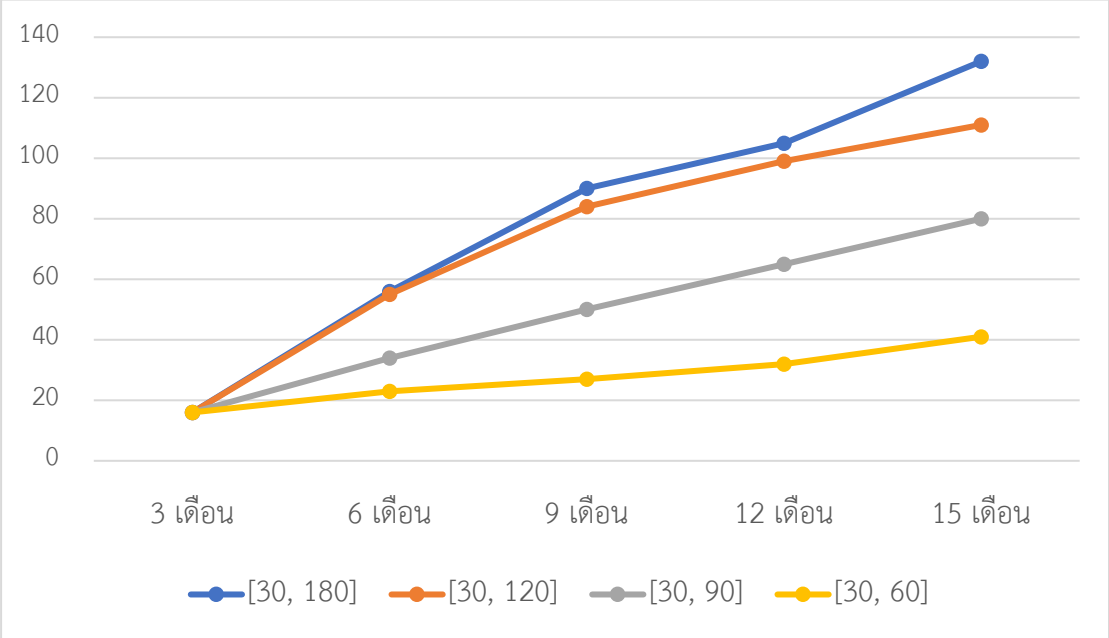
ตารางที่ 25 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน

เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 90 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	41 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	32 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	27 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	23 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	16 คำ

ตารางที่ 26 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน

เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 60 วัน



รูปภาพที่ 25 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุด กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

จาก*ตารางที่ 23 ถึง 26* เรานำข้อมูลมาแสดงเป็นกราฟดัง*รูปภาพที่ 25* ซึ่งจะเห็นได้ว่า ยิ่งเรามีข้อมูลโพสต์มาก เราจะได้จำนวนของคำหยุดในฐานข้อมูลเรามากขึ้นตามลำดับ โดยจะเห็นได้ว่าเส้นกราฟเป็นลักษณะของกราฟโค้งที่ลู่เข้า ซึ่งแสดงให้เห็นว่า เมื่อเริ่มต้นสร้างฐานข้อมูลคำหยุด เราจะพบคำหยุดปะปนออกมามากมาย แต่เมื่อเรามีข้อมูลสะสมมากพอ เราจะเริ่มพบคำหยุดใหม่น้อยลง ๆ ซึ่งแสดงว่าฐานข้อมูลคำหยุดของเรานั้นเพียงพอให้เรากรองคำสำคัญที่เป็นกระแสออกมาแล้ว

อีกประเด็นหนึ่งที่ส่งผลต่อฐานข้อมูลคำหยุดก็คือ ตัวแปรจำนวนวันที่ใช้ในการอ้างอิงลักษณะของคำหยุด เนื่องจากปริมาณโพสต์ต่อวันที่เราเก็บจากเฟซบุ๊กค่อนข้างน้อยจากข้อจำกัดที่เฟซบุ๊กป้องกันเราเข้าถึงข้อมูลทั้งหมดของเพจนั้น ทำให้การที่เราจะพบคำหยุดเดิม ปรากฏอีกครั้งในช่วงเวลาสั้น ๆ นั้นเป็นไปได้ยาก โดยเมื่อเพิ่มจำนวนวันไปเรื่อย ๆ จะพบว่า ปริมาณคำหยุดที่เราได้เริ่มลดลง และมีโอกาสที่คำหยุดดังกล่าวจะเป็นกระแสที่กลับมาเกิดซ้ำมากขึ้น ซึ่งจำนวนวันที่ให้ผลดีมากที่สุดคือช่วง ตั้งแต่ 30 วัน จนถึง 120 วันก่อนหน้า

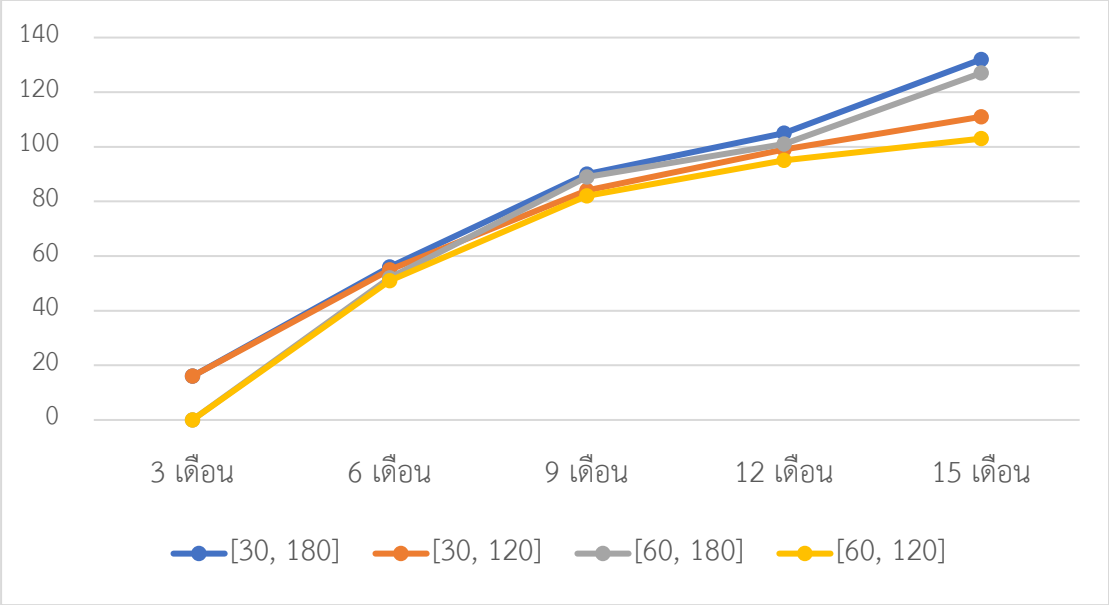
นอกจากนี้ เรามีการทดลองขยายระยะเวลาของการเกิดซ้ำของคำหยุดจากเดิม 30 วัน เพื่อป้องกันกระแสที่อยู่นานกว่า 1 เดือน โดยเพิ่มเป็น 60 วัน โดยมีการทดลองด้วยจำนวนวันที่ใช้อ้างอิงตั้งแต่ 60-120 วัน และ 60-180 วัน ดัง*ตารางที่ 27 และ 28*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	127 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	101 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	89 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	52 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	0 คำ

ตารางที่ 27 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	103 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	95 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	82 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	51 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	0 คำ

ตารางที่ 28 ตารางแสดงปริมาณคำหยุดตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน



รูปภาพที่ 26 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุด กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

จาก*ตารางที่ 27 และ 28* เมื่อแสดงข้อมูลเป็นกราฟดัง*รูปภาพที่ 26* จะเห็นได้ว่า การใช้ข้อมูลอ้างอิงย้อนหลังตั้งแต่ 60 วัน ถึง 120 วัน ในการพิจารณาคำหยุด ให้ผลใกล้เคียงกับการใช้ข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน ถึง 120 วัน ซึ่งแม้ว่าตอนแรกจะเห็นว่ามีปริมาณคำหยุดน้อยกว่าก็จริง แต่เมื่อใช้ข้อมูลโพสต์เพิ่มขึ้นเรื่อย ๆ ข้อมูลคำหยุดก็กลับมาใกล้เคียงกัน ซึ่งหากใช้ช่วงเวลาดังกล่าว ก็สามารถลดการลบคำสำคัญที่เป็นกระแสที่เกิดเป็นกระแสยาวนานกว่า 1 เดือน โดยไม่สูญเสียปริมาณคำหยุดที่มากเกินไป ในขณะเดียวกัน การใช้ข้อมูลอ้างอิงเพิ่มขึ้นเป็นย้อนหลังตั้งแต่ 60 วัน ถึง 180 วัน ได้ปริมาณคำหยุดเพิ่มขึ้นไม่แตกต่างกันอย่างมีนัยสำคัญ แต่อาจส่งผลถึงคำสำคัญที่เป็นกระแสที่หมดกระแสไปแล้ว แต่กลับมาเป็นกระแสอีกครั้งได้ ซึ่งทำให้การเลือกช่วงข้อมูลอ้างอิงย้อนหลังตั้งแต่ 60 วัน ถึง 120 วัน ดูเป็นทางเลือกที่ปลอดภัยกว่า

การวัดความถูกต้องของคำหยุด

ในการวัดความถูกต้องของคำหยุดนั้น ในงานวิจัยนี้ ได้ใช้อาสาสมัครที่ผู้ติดตามข่าวสารสังคมไทยจำนวน 5 คน มาร่วมกันโหวตเสียงข้างมาก ของคำหยุดทั้งหมด 132 คำ ที่ระบบสกัดออกมาได้ โดยใช้ช่วงข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน ถึง 180 วัน โดยมีหลักการว่า คำใดไม่ใช่คำที่หมายถึงกระแสในช่วง 1 มกราคม 2561 ถึง 31 มีนาคม 2562 ถือว่าเป็นคำหยุด โดยแสดงผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 29*

0800น | 0นทาง | 0นทางช่อง | facebook | liveเ | liveเรื่อง | liveส | liveสด | news1 | official | ookcom | thailand | เคราะห์ | เช้านี้ | เชียร์ | เดียว | เดือด | เที่ย | เที่ยว | เปิดใจ | เพราะ | เพื่อ | เมือง | เริ่ม | เรียน | เรื่อ | เรื่อง | เรื่องเล่าเ | เล่าเรื่อง | เวลา1 | เวลา2 | เวลา20 | เสาร์ | เสียง | เสียชีวิต | เหลือ | แลนด์ | ได้ในรายการ | ได้ทาง | ได้ที่ | การเมือง | การณ์ | กำลัง | ข่าวเ | ข่าวเด | ข่าวเที่ย | ข่าวเรื่องเล่าเ | ข่าวเล่าเรื่อง | ข่าวส | ข่าวสาร | ข่าวสารเพิ่มเติม | ครอบครัว | ครั้ง | ความเ | ความส | คืนนี้ | งข่าวส | งช่อง | จันทร์ | ชีวิต | ตำรวจ | ติดตาม | ติดตามเรื่อง | ติดตามได้ที่ | ติดตามร | ติดตามรับชม | ติดตามราย | ติดตามรายการ | ติดตามรายการย้อนหลัง | ติดตามรายการย้อนหลังได้ทาง | ติดตามรายการย้อนหลังได้ทางw | ถ่ายทอดสด | ทั่วไทย | ทางช่อง | ที่จะ | ที่นี่ | ทุกวัน | ทุกวันจันทร์ | ทุกวันจันทร์- | นเป็น | นข่าว | นข่าวส | นทางช่อง | นนี้เ | นประจำวัน | นหลัง | นอย่าง | บันเทิง | ประเด็น | ประเทศ | ประจำ | ประจำวัน | พบกับ | พร้อม | พฤหัสบดี | พ่อแม่ | พูดคุย | ภาพยนตร์ | ยชีวิต | ย้อนหลัง | ย้อนหลังได้ท | ร์ที่ | รย้อนหลัง | รับชม | รับชมย้อนหลัง | รายการ | รายการเรื่อง | รายการย้อนหลัง | รื่อง | วันเสาร์ | วันจันทร์ | วันที่ | วันที่2 | วันนี้ | วันพุธ | วันอังคาร | วันอังคารที่ | สามารถ | หน้าจอ | หน้าที่ | หนึ่ง | หนุ่ม | หมายเลข | หมายเลข2 | อนหลั | อย่าง | อย่างไ | ออนไลน์ | อังคาร | อาทิตย์ |ันที่ | านการ

ตารางที่ 29 ตารางแสดงคำหยุดทั้งหมด 132 คำ ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562

จากคำหยุดทั้งหมด 132 คำ พบว่าอาสาสมัครทั้ง 5 คน ให้ความเห็นตรงกันทั้งหมด 131 คำ จะมีแค่คำว่า “การเมือง” เท่านั้น ที่พบว่ามีเสียง 4 ใน 5 ให้ความเห็นว่าเป็นคำที่เป็นกระแส โดยอาสาสมัครคนที่บอกว่าคำนี้เป็นคำหยุด ให้เหตุผลว่า คำดังกล่าวไม่ได้ระบุชัดว่าเป็นกระแสเกี่ยวกับการเมืองด้านใด อีกทั้งข่าวการเมืองก็มีอยู่เสมอ เหมือนข่าวกีฬา เพียงแต่ว่าประเทศไทยไม่ได้พูดถึงการเมืองมานานจนคำนี้อาจมองว่าเป็นกระแสในช่วงนี้ก็เป็นได้

ผู้วิจัยได้เห็นถึงประเด็นดังกล่าว จึงตรวจสอบผลลัพธ์ที่ได้ โดยใช้ช่วงข้อมูลอ้างอิงย้อนหลังที่ปลอดภัยที่สุด คือช่วง 60 วัน ถึง 120 วัน ซึ่งเป็นช่วงที่ลดโอกาสเจอคำที่เป็นกระแสที่เกิดขึ้นยาวนาน หรือคำที่เป็นกระแสแต่หมดกระแสไปแล้ว แต่กลับมาเป็นกระแสใหม่อีกครั้ง โดยแสดงผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 30*

0นทาง | 0นทางช่อง | liveเ | liveเรื่อง | liveส | news1 | official | thailand | เคราะห์ | เช้านี้ | เชียร์ | เดียว | เดือด | เที่ย | เที่ยว | เปิดใจ | เพราะ | เพื่อ | เมือง | เริ่ม | เรียน | เรื่อง | เรื่องเล่าเ | เล่าเรื่อง | เวลา1 | เวลา2 | เวลา20 | เสาร์ | เสียง | เสียชีวิต | เหลือ | ได้ในรายการ | ได้ทาง | ได้ที่ | การณ์ | กำลัง | ข่าวเ | ข่าวเด | ข่าวเที่ย | ข่าวเรื่องเล่าเ | ข่าวเล่าเรื่อง | ข่าวส | ข่าวสาร | ครั้ง | ความเ | คืนนี้ | งข่าวส | งช่อง | จันทร์ | ชีวิต | ตำรวจ | ติดตาม | ติดตามได้ที่ | ติดตามรับชม | ติดตามราย | ติดตามรายการ | ติดตามรายการย้อนหลัง | ติดตามรายการย้อนหลังได้ทาง | ถ่ายทอดสด | ทั่วไทย | ทางช่อง | ที่นี่ | ทุกวัน | ทุกวันจันทร์ | ทุกวันจันทร์- | นเป็น | นข่าว | นข่าวส | นนี้เ | นประจำวัน | นหลัง | นอย่าง | บันเทิง | ประจำวัน | พบกับ | พร้อม | พฤหัสบดี | พ่อแม่ | พูดคุย | ภาพยนตร์ | ยชีวิต | ย้อนหลัง | รับชม | รายการ | รายการย้อนหลัง | รื่อง | วันเสาร์ | วันจันทร์ | วันที่ | วันที่2 | วันนี้ | วันพุธ | หน้าจอ | หน้าที่ | หนึ่ง | หนุ่ม | หมายเลข | อนหลั | อย่าง | อังคาร | อาทิตย์ |ันที่ | านการ

ตารางที่ 30 ตารางแสดงคำหยุดทั้งหมด 103 คำ ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562

0800น | facebook | liveสด | ookcom | เรื่อ | แลนด์ | การเมือง | ข่าวสารเพิ่มเติม | ครอบครัว | ความส | ติดตามเรื่อง | ติดตามร | ติดตามรายการย้อนหลังได้ทางw | ที่จะ | นทางช่อง | ประเด็น | ประเทศ | ประจำ | ย้อนหลังได้ท | ร์ที่ | รย้อนหลัง | รับชมย้อนหลัง | รายการเรื่อง | วันอังคาร | วันอังคารที่ | สามารถ | หมายเลข2 | อย่างไ | ออนไลน์ |

ตารางที่ 31 ตารางแสดงข้อมูลคำหยุดที่หายไปทั้งหมด 29 คำ เมื่อใช้ข้อมูลอ้างอิงย้อนหลังช่วง 60 วัน ถึง 120 วัน แทนช่วง 30 วัน ถึง 180 วัน

จาก*ตารางที่ 30* ไม่พบคำว่า “การเมือง” แล้ว แต่ก็มีคำอื่นที่หายไปรวมทั้ง 29 คำ ดัง*ตารางที่ 31* ซึ่งผู้วิจัยคาดว่า หากเราไม่มีข้อจำกัดเรื่องข้อมูลโพสต์ที่สามารถเก็บได้จากเพจเฟซบุ๊ก เราน่าจะมีข้อมูลโพสต์ที่เพียงพอทำให้ คำหยุดอื่น ๆ ที่ไม่ใช่คำว่า “การเมือง” มีโอกาสถูกตรวจพบมากขึ้น

การเพิ่มกฎของคำหยุด เพื่อเพิ่มประสิทธิภาพในการสกัดคำสำคัญ

ผู้วิจัยได้วิเคราะห์ผลของคำสำคัญที่เป็นกระแส และพบว่ายังมีข้อมูลส่วนหนึ่ง มีการปรากฏคำหยุดที่เกี่ยวเนื่องกับเดือน เช่น “มีนาคม” “มีนาคม2562” “มีค62” เป็นต้น ซึ่งฐานข้อมูลคำหยุดที่เราสกัดมาได้นั้น โดยการใช้ข้อมูลอ้างอิงย้อนหลังเป็นหน่วยเดือน ไม่สามารถสกัดคำหยุดเหล่านี้ได้ จึงจำเป็นที่จะต้องมีการเพิ่มกฎ เพื่อสกัดคำหยุดเหล่านี้ออกมาด้วย โดยจะมีการนำคำสำคัญที่เป็นกระแส ที่ภายในข้อความประกอบด้วยชื่อเดือนออก เพื่อให้เหลือแค่กระแสหลักที่เกิดขึ้นจริง

การวัดความถูกต้องของคำสำคัญที่เป็นกระแส

ในการวัดความถูกต้องของคำสำคัญที่เป็นกระแสนั้น ในงานวิจัยนี้ ได้ใช้อาสาสมัครที่ผู้ติดตามข่าวสารสังคมไทยจำนวน 5 คน มาร่วมกันโหวตเสียงข้างมาก ของคำสำคัญที่เป็นกระแสที่ระบบสกัดมาได้ทั้งหมด โดยหากวันไหนไม่มีคำสำคัญที่ระบบสกัดมาได้เลย ก็จะไม่ถูกแสดงให้อาสาสมัครเห็น โดยข้อมูลที่นำมาวัดความถูกต้องนั้น ใช้ฐานข้อมูลคำหยุดที่อ้างอิงช่วงข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน ถึง 180 วัน เนื่องจากสามารถสกัดคำหยุดได้มากที่สุด โดยมีหลักการว่า คำใดที่สื่อถึงเรื่องราวที่เป็นกระแสในช่วง 1 มกราคม 2561 ถึง 31 มีนาคม 2562 ถือว่าเป็นคำที่เป็นกระแส โดยหากคำดังกล่าวมีส่วนผสมของคำหยุด ให้เลือกยืนยันคำที่เป็นกระแสแค่บางส่วนของคำได้ โดยแสดงผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 32 และ 33*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	เลือก | จัดตั้งรัฐบาลเรื่อง | รัฐบาล | รายการเล | เลือกตั้ง | ตั้งรัฐบาล | เพื่อไทย | เลือกตั้ง62 | จัดตั้งรัฐบาล | การเล | ติดตามข่าวสารการเลือกตั้ง62 | การเลือกตั้งเลือกตั้ง | ประชา | การเลือกตั้ง | หมายเ
2019-03-26	เลือก | เลือกตั้ง | เลือกตั้ง62เลือกตั้ง | เพื่อไทย | เลือกตั้ง62 | คะแนน | เลือกตั้งเรื่อง | เลือกตั้ง62เลือกตั้ง62 | คะแนนเสียง | ประชา | งประชา | รเลือกตั้ง | คะแนนเลือกตั้ง
2019-03-25	เลือกตั้ง62thailandelection2019 | คะแนนเลือกตั้ง62 | election62thailandelection2019 | เลือกตั้งเรื่อง | นับคะแนน | นับคะแนนเลือกตั้ง | เลือกตั้ง62 | คะแนน | คะแนนเสียง | ประชา | การเลือกตั้ง | งประชา | เลือกตั้ง62เลือกตั้ง62thailandelection2019 | เลือกตั้ง62thai | ทางการ | election | นเลือกตั้ง62 | คะแนนเลือกตั้ง | thailandelection2019 | เลือกตั้ง | การเลือกตั้ง62 | เลือกตั้ง62เรื่อง | เสียงประชา | รเลือกตั้ง

วันที่	คำสำคัญที่เป็นกระแส
2019-03-24	เลือก | เกาะติด | สิทธิ | เลือกตั้ง | เพื่อไทย | เลือกตั้ง62 | สิทธิเลือกตั้ง | สิทธิ์ | เลือกตั้ง62เรื่อง | เลือกตั้ง62เลือกตั้ง62 | ประชา
2019-03-23	เลือก | เลือกตั้ง | เลือกตั้ง62 | ประชา | รเลือกตั้ง
2019-03-22	เลือก | สิทธิ | เลือกตั้ง | เลือกตั | เลือกตั้ง62 | ประชา
2019-03-21	แลนด์ทางช่อง | แลนด์วันนี้ | เชียง | นรายการ | สิทธิ์ | ติดตามเพิ่มเติมได้ที่ | เป็นก | ครั้งข่าวส | รับชมผ่านช่องท | ธนาธร | ประชาชน | เลือกตั้ง62 | liveข | รับชมย้อนหลังได้ทางw | มหมาย | รับชมผ่าน | เรื่องเล่าเช้านี้ | liveคุยกับ | twitter | เป็นเ | ออนไลน์wwwfacebookcom/ | ได้ที่นี่ | รับชมได้ทาง | newsupdate | เรื่องเล่าเช้านี้เลือกตั้ง62เรื่องเล่าเช้านี้เลือกตั้ง62 | งประเด็น | ต้องม | เมื่อ | หมายเ | ได้ในเรื่องเล่าเช้านี้ | liveค | มหมายเลข2 | ทุกทิศทั่วไทย | อย่างไร | เลือกตั้ง | ช่องท | เพิ่ม | งประเทศ | เรื่องเล่าเช้านี้เลือกตั้ง62 | facebookเดียว | งข่าว | liveคุย | liveสรุปข่าวเด่นประเด็น | งเรื่อง
2019-03-20	สรุปข่าวเด่นประเด็น | ความเคลื่อนไหว | นรายการ | ไปตามชมกัน | รับชมผ่านช่องท | ารเพื่อ | เลือกตั้ง62 | รับชมกัน | เหนือ | liveข | รับชมย้อนหลังได้ทางw | พูดคุยในประเด็น | 0800นทางช่อง | มหมาย | ประชา | รับชมผ่าน | เรื่องเล่าเช้านี้ | twitter | ได้ที่นี่เรื่องเล่าเช้านี้ | ายได้ | รับชมได้ทาง | liveเทียงข่าว | facebooktwitterofficial | เลือกตั ⌈พากษ | newsupdate | จ้าาา | ได้ในเรื่องเล่าเช้านี้ | หมายเ | เลือก | มหมายเลข2 | ติดตามข่าวสาร | เล่าเ | รตอวัชรินทร์ | ดประเทศไทย | เลือกตั้ง | ช่องท | ไปตาม | งเรื่อง
2019-03-19	liveด | ติดตามข่าวสาร | เลือกตั้ง | facebooktwitterofficial | เลือกตั้ง62 | เทียงข่าว | พูดคุยในประเด็น | รับชมย้อนหลังได้ทางw | liveข | ประชา | newsupdate | twitter
2019-03-18	เลือก | เลือกตั้ง
2019-03-17	เลือกตั้ง
2019-03-16	เลือกตั้ง
2019-03-15	เลือก | เลือกตั้ง
2019-03-14	เลือกตั้ง62 | เลือกตั้ง
2019-03-12	เป็นเ | ติดตามข่าวสาร | เลือกตั้ง | วันนี้พบกับ | นที่1 | newsupdate | เช้านี้ทางช่อง | ที่นี่เรื่อง

วันที่	คำสำคัญที่เป็นกระแส
2019-03-10	เสาร์อาทิตย์
2019-03-09	เสาร์อาทิตย์ | มเวลา
2019-03-08	newsupdate | ไทยรักษาชาติ
2019-03-03	เสาร์อาทิตย์
2019-02-14	าพันธ
2019-02-07	ประจำวันพ
2019-02-01	พร้อมห
2018-12-31	ปีใหม่
2018-12-28	ข่าวเช้านี้
2018-12-27	งวันที่2 | สรุปข่าวเด่นประ | ข่าวเช้านี้ | ปีใหม่ | ที่นี
2018-12-26	สรุปข่าวเด่นประ
2018-12-25	งวันที่2
2018-12-19	ได้ที่เรื่องเล่าเ
2018-12-18	ได้ที่เรื่องเล่าเ
2018-12-14	ชมกัน | ได้ใน
2018-12-13	ได้ใน
2018-12-07	ได้ใน
2018-09-28	รายการเพื่อ | เพื่อติดตาม
2018-09-27	หมายเลข3 | เพื่อติดตาม | แชมป์ | คุยกับ
2018-09-26	นข่าวเ
2018-09-24	นข่าวเพิ่มเติมได้ที่
2018-09-22	งความ | อกจาก
2018-09-11	มข่าว
2018-09-07	แอปพลิเคชั | ย้อนหลังได้ที่ติดตามข่าว | มได้ที่นี่ | เพื่อติดตาม | นช่อง
2018-09-06	Newsupdate
2018-08-29	ย้อนหลังได้ทาง
2018-06-28	ถ้ำหลวงหมูป่า13ชีวิต | หมูป่าติดถ้ำ | 13ชีวิตถ้ำหลวง | ในถ้ำขุนน้ำนางนอนถ้ำหลวง ทีมหมูป่า | ถ้ำหลวงข | ติดถ้ำหลวงขุนน้ำนางนอน | 13ชีวิตทีมหมูป่า | ค้นหา | เกาะติดปฏิบัติการค้นหาทีมหมูป่า | ค้นหา13ชีวิต | เชียงราย | ถ้ำหลวงขุนน้ำนาง

วันที่	คำสำคัญที่เป็นกระแส
	นอน | ติดถ้ำหลวง | ค้นหา13ชีวิตติดถ้ำหลวง | ในถ้ำหลวง | ทีมหมูป่าถ้ำหลวง | นถ้ำหลวง | ถ้ำหลวงเรื่อง | หน่วยซีล | 13ชีวิต | เกาะติด | 13ชีวิตติดถ้ำหลวงเรื่อง | ถ้ำหลวง | ช่วยเหลือ | 13ชีวิตติดถ้ำหลวง | ในถ้ำหลวงขุนน้ำนางนอน | หมูป่า | เจ้าหน้าที่
2018-06-27	ในถ้ำ | ถ้ำหลวง | เกาะติด | การค้นหา | ในถ้ำหลวง | ค้นหา | 13ชีวิตติดถ้ำหลวง | ปฏิบัติ | 13ชีวิต
2018-06-26	ระดับ | พร้อมก | กับข่าวส | หน่วย | ค้นหา | นิวส์ | กำลังใจ | เชียงราย | ฟุตบอล | กับข่าวเ | าแล้ว | หน่วยซีล | ยังมี | ลื่อน | 13ชีวิต | ถ้ำหลวง | งข่าวเล่าเรื่อง | งข่าว | เจ้าหน้าที่ | วาม
2018-06-25	พิเศษ | หั่นศพ | ฟุตบอล | พิเศษวันนี้ | ฆ่าหั่นศพ | cup20
2018-06-20	ประหาร
2018-06-12	เรื่องเ
2018-05-16	ย้อนหลังได้
2018-05-11	ได้เลย
2018-03-27	ายการ
2018-03-23	อร่อย | ซิงเกิ | รถทัวร์ | channel
2018-03-22	ชมย้อนหลัง | บุพเพสันนิวาส | 0ล้าน | จังหวัด | channel | /live
2018-03-21	รายการย้อนหลังได้ที่
2018-03-18	/live
2018-03-08	ห้ามพลาด
2018-02-20	ย้อนหลังได้ทาง
2018-02-10	ได้ทางw
2018-02-04	ย้อนหลังได้ที่นี่
2018-01-28	ย้อนหลังได้ที่นี่

ตารางที่ 32 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแส โดยใช้เสียงข้างมากจากอาสาสมัครทั้ง 5 คน โดยหากคำใดเป็นกระแสที่มีเสียงโหวตเกินครึ่งจะมีสีน้ำเงิน (ขีดเส้นใต้) บ่งบอกอยู่

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	*เลือก* | จัดตั้งรัฐบาลเรื่อง | *รัฐบาล* | เลือกตั้ง | ตั้งรัฐบาล | เพื่อไทย | เลือกตั้ง62 | จัดตั้งรัฐบาล | ติดตามข่าวสารการเลือกตั้ง62 | การเลือกตั้งเลือกตั้ง | *ประชา* | การเลือกตั้ง
2019-03-26	*เลือก* | เลือกตั้ง | เลือกตั้ง62เลือกตั้ง | เพื่อไทย | เลือกตั้ง62 | *คะแนน* | เลือกตั้งเรื่อง | เลือกตั้ง62เลือกตั้ง62 | คะแนนเสียง | *ประชา* | งประชา | รเลือกตั้ง | คะแนนเลือกตั้ง
2019-03-25	เลือกตั้ง62thailandelection2019 | คะแนนเลือกตั้ง62 | election62thailandelection2019 | เลือกตั้งเรื่อง | นับคะแนน | นับคะแนนเลือกตั้ง | เลือกตั้ง62 | *คะแนน* | คะแนนเสียง | *ประชา* | การเลือกตั้ง | งประชา | เลือกตั้ง62เลือกตั้ง62thailandelection2019 | เลือกตั้ง62thai | election | นเลือกตั้ง62 | คะแนนเลือกตั้ง | thailandelection2019 | เลือกตั้ง | การเลือกตั้ง62 | เลือกตั้ง62เรื่อง | เสียงประชา | รเลือกตั้ง
2019-03-24	*เลือก* | สิทธิ | เลือกตั้ง | เพื่อไทย | เลือกตั้ง62 | สิทธิเลือกตั้ง | สิทธิ์ | เลือกตั้ง62เรื่อง | เลือกตั้ง62เลือกตั้ง62 | *ประชา*
2019-03-23	*เลือก* | เลือกตั้ง | เลือกตั้ง62 | *ประชา* | รเลือกตั้ง
2019-03-22	*เลือก* | สิทธิ | เลือกตั้ง | *เลือกตั* | เลือกตั้ง62 | *ประชา*
2019-03-21	สิทธิ์ | ธนาธร | *ประชาชน* | เลือกตั้ง62 | เรื่องเล่าเช้านี้เลือกตั้ง62เรื่องเล่าเช้านี้เลือกตั้ง62 | เลือกตั้ง | เรื่องเล่าเช้านี้เลือกตั้ง62
2019-03-20	เลือกตั้ง62 | รตอวัชรินทร์ | เลือกตั้ง
2019-03-19	เลือกตั้ง | เลือกตั้ง62
2018-03-18	*เลือก* | เลือกตั้ง
2019-03-17	เลือกตั้ง
2019-03-16	เลือกตั้ง
2019-03-15	*เลือก* | เลือกตั้ง
2019-03-14	เลือกตั้ง62 | เลือกตั้ง
2019-03-12	เลือกตั้ง
2019-03-08	ไทยรักษาชาติ
2018-12-31	ปีใหม่
2018-12-27	ปีใหม่

วันที่	คำสำคัญที่เป็นกระแส
2018-06-28	ถ้ำหลวงหมูป่า13ชีวิต | หมูป่าติดถ้ำ | 13ชีวิตถ้ำหลวง | ในถ้ำขุนน้ำนางนอนถ้ำหลวง ทีมหมูป่า | ถ้ำหลวงข | ติดถ้ำหลวงขุนน้ำนางนอน | 13ชีวิตทีมหมูป่า | *ค้นหา* | เกาะติดปฏิบัติการค้นหาทีมหมูป่า | ค้นหา13ชีวิต | เชียงราย | ถ้ำหลวงขุนน้ำนางนอน | ติดถ้ำหลวง | ค้นหา13ชีวิตติดถ้ำหลวง | ในถ้ำหลวง | ทีมหมูป่าถ้ำหลวง | นถ้ำหลวง | ถ้ำหลวงเรื่อง | หน่วยซีล | 13ชีวิต | 13ชีวิตติดถ้ำหลวงเรื่อง | ถ้ำหลวง | *ช่วยเหลือ* | 13ชีวิตติดถ้ำหลวง | ในถ้ำหลวงขุนน้ำนางนอน | หมูป่า | *เจ้าหน้าที่*
2018-06-27	ในถ้ำ | ถ้ำหลวง | *การค้นหา* | ในถ้ำหลวง | *ค้นหา* | 13ชีวิตติดถ้ำหลวง | 13ชีวิต
2018-06-26	*ค้นหา* | *กำลังใจ* | เชียงราย | ฟุตบอล | หน่วยซีล | 13ชีวิต | ถ้ำหลวง | *เจ้าหน้าที่*
2018-06-25	หั่นศพ | ฟุตบอล | ฆ่าหั่นศพ | cup20
2018-06-20	ประหาร
2018-03-22	บุพเพสันนิวาส

ตารางที่ 33 ตารางสรุปคำสำคัญที่เป็นกระแส โดยสีน้ำเงิน (ขีดเส้นใต้) หมายถึง คำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) หมายถึง คำที่มีเสียงโหวต 2 จาก 5 เสียง ที่เป็นคำกำกวม

จาก*ตารางที่ 32* แสดงให้เห็นว่า ยังคงมีคำหยุดหลงเหลืออยู่จำนวนหนึ่ง โดยคำหยุดเหล่านั้นจะปรากฏชัด ในวันที่ไม่มีกระแสเกิดขึ้น ซึ่งหากเรามีข้อมูลโพสต์เพิ่มมากขึ้น ก็อาจจะกรองคำหยุดเหล่านี้ออกไปได้ อีกทั้งยังพบคำหยุดติดอยู่กับคำสำคัญที่เป็นกระแส ซึ่งเป็นเรื่องยากในการกรองออก เนื่องจากการพบคำในลักษณะดังกล่าว แสดงให้เห็นว่าคำหยุดเหล่านั้น มักปรากฏติดกับคำที่เป็นกระแส ซึ่งบางครั้งคำเหล่านั้นอาจจะเป็นส่วนหนึ่งของกระแสที่เกิดขึ้น

อีกทั้ง เมื่อเราวิเคราะห์คำที่มีเสียงโหวตเกือบครึ่งหนึ่ง จะพบว่าคำเหล่านั้นมีลักษณะเป็นคำหยุดโดด ๆ ที่หากปรากฏอยู่ในวันอื่น จะไม่เกี่ยวข้องกับคำที่เป็นกระแส แต่หากปรากฏในสถานการณ์ที่มีกระแสที่เกี่ยวข้องกับคำเหล่านี้ คำเหล่านี้จะช่วยขยายเรื่องราวของกระแสให้ชัดเจนมากยิ่งขึ้น อย่างเช่นคำว่า “เลือก” “รัฐบาล” “ประชาชน” และ “คะแนน” ซึ่งช่วยขยายความการเลือกตั้งที่เกิดขึ้นได้ ว่าเกี่ยวข้องกับการจัดตั้งรัฐบาล โดยผู้มีสิทธิเลือกตั้งคือประชาชน อีกทั้งวันดังกล่าวอาจเป็นวันนับคะแนนเสียงเลือกตั้ง เป็นต้น หรืออย่างคำว่า “ค้นหา” “ช่วยเหลือ” “กำลังใจ” และ “เจ้าหน้าที่” ซึ่งดูเป็นคำหยุดโดด ๆ แต่เมื่อประกอบกับกระแสถ้ำหลวงที่เกิดขึ้น สามารถขยายเรื่องราวในวันนั้นได้ ว่ามีการค้นหาและช่วยเหลือผู้รอดชีวิตที่ติดถ้ำหลวง และมีการส่งกำลังใจไปให้ผู้ประสบภัย และอาจมีข่าวที่เกี่ยวข้องกับเจ้าหน้าที่ที่ปฏิบัติงาน เป็นต้น

การเปรียบเทียบผลเมื่อใช้วิธี TF-IDF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร

เนื่องจากงานวิจัยนี้ ได้เสนอวิธีการที่ใช้ในการสกัดคำสำคัญที่เป็นกระแส โดยวิธีเอ็นแกรมแบบตัวอักษรแทนเครื่องมือตัดคำ เนื่องจากข้อจำกัดด้านข้อมูลฝึกสอน หากใช้เครื่องมือตัดคำที่ใช้การเรียนรู้ของเครื่อง หรือขาดคลังคำศัพท์บนสื่อสังคมออนไลน์ หากใช้เครื่องมือตัดคำแบบพจนานุกรม โดยเมื่อทดลองใช้เครื่องมือตัดคำ deepcut ที่เป็นเครื่องมือที่ใช้ข้อมูลฝึกสอน จากคลังข้อมูลภาษา BEST2009 ซึ่งไม่มีข้อมูลฝึกสอนเกี่ยวกับข้อความบนสื่อสังคมออนไลน์เฟซบุ๊ก พบว่ามีไม่สามารถแบ่งข้อความที่ไม่ใช่ข้อความมาตรฐานออกจากกันได้

โดยเพื่อเป็นการเปรียบเทียบ วิธีที่งานวิจัยนี้ได้คิดค้นขึ้น กับวิธีดั้งเดิม ว่ามีประสิทธิภาพแตกต่างกันอย่างไร ในงานวิจัยนี้จึงได้นำเครื่องมือตัดคำ deepcut มาใช้คู่กับวิธีการสกัดคำสำคัญอย่าง TF-IDF เพื่อให้ได้คำสำคัญของแต่ละเพจในแต่ละวันออกมา หลังจากได้คำสำคัญดังกล่าว เราก็ทำการคัดเลือกคำสำคัญนั้นในลักษณะเดียวกันกับงานวิจัยนี้ โดยการหาค่า DF ของแต่ละคำสำคัญจากจำนวนเพจข่าว เพื่อใช้การแบ่งกลุ่มแบบเคมีน และวิธีข้อศอกมาคัดเลือกคำสำคัญที่เป็นกระแสของวันนั้นออกมา และนำมาสร้างฐานข้อมูลคำหยุด เพื่อจำแนกคำสำคัญที่เป็นกระแสออกมา

ในการใช้ค่า TF-IDF โดยที่เราต้องการคำสำคัญที่มีจำนวนคำมากกว่า 1 คำ สามารถทำได้โดยการหาค่า TF-IDF บนแกรมของคำแทน ซึ่งโดยทั่วไปคำสำคัญที่เป็นกระแส มักมีความยาวตั้งแต่ 1 คำ จนถึง 3 คำ ในงานวิจัยนี้จึงทำการตัดคำโดยใช้เครื่องมือ deepcut แล้วจึงนำคำที่ตัดออกมานั้นไปผ่านกระบวนการเอ็นแกรมแบบตัวอักษร เพื่อให้ได้ ยูนิแกรม ไบแกรม และไตรแกรม แล้วจึงนำแกรมทั้ง 3 ชนิดไปหาค่า TF-IDF เพื่อจำแนกแกรมที่มีความสำคัญออกมา

การทดลองปรับตัวแปรที่เกี่ยวข้องกับฐานข้อมูลคำหยุดเมื่อใช้วิธี TF-IDF

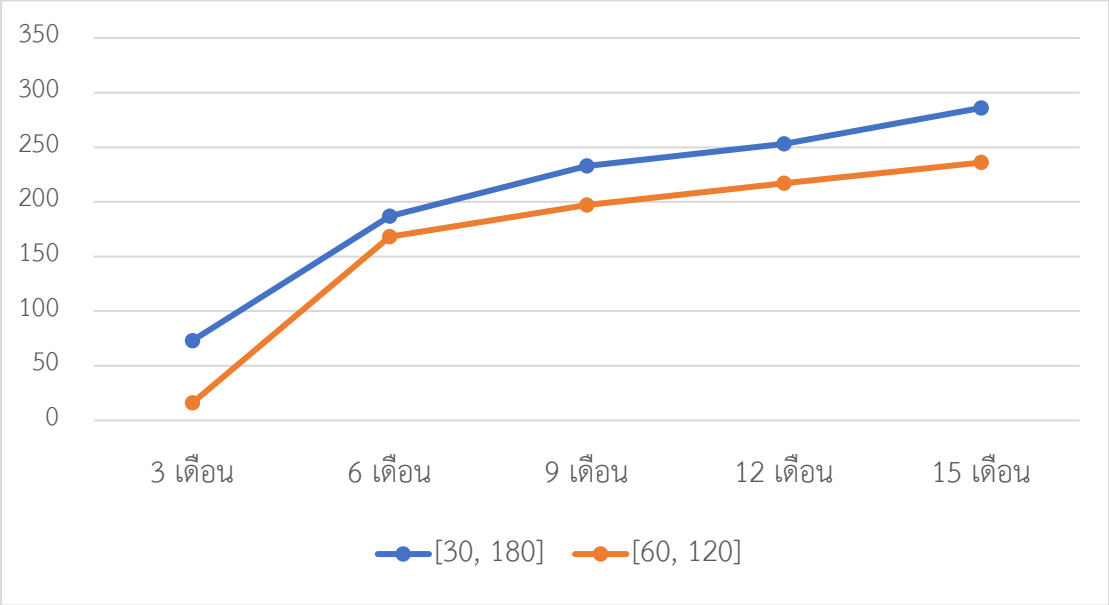
จากความรู้ที่เราเคยทดลองกับวิธีอย่างเอ็นแกรมแบบตัวอักษรไปแล้วนั้น พบว่าช่วงข้อมูลอ้างอิงที่เหมาะสมที่สุดคือ วันย้อนหลังตั้งแต่ 60 วัน จนถึง 120 วัน แต่หากเราต้องการปริมาณคำหยุดที่เพิ่มมากขึ้น โดยยอมให้มีโอกาสเจอคำที่เป็นกระแสที่นานกว่า 1 เดือน หรือกลับเกิดซ้ำได้นั้น เราสามารถใช้ช่วงวันย้อนหลังตั้งแต่ 30 วัน จนถึง 180 วัน ได้ และเพื่อยืนยันว่าเราสามารถนำช่วงดังกล่าว มาใช้กับวิธีอย่าง TF-IDF ได้ ในงานวิจัยนี้จึงได้มีการเปรียบเทียบทั้งช่วงข้อมูลอ้างอิงทั้ง 2 ช่วงนี้อีกครั้ง และนำมากรองคำหยุดออกเพื่อหาคำสำคัญที่เป็นกระแส โดยมีการใช้กฎคำหยุดที่เกี่ยวข้องกับเดือนเพิ่มเติม และสกัดกระแสที่เกิดมาแล้วมากกว่า 1 วัน เหมือนอย่างวิธีเอ็นแกรมแบบตัวอักษร โดยแสดงผลลัพธ์ที่ไว้ใน*ตารางที่ 34 และ 35*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	286 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	253 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	233 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	187 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	73 คำ

ตารางที่ 34 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	236 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	217 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	197 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	168 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	16 คำ

ตารางที่ 35 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน



รูปภาพที่ 27 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF-IDF กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

เมื่อนำข้อมูลจาก*ตารางที่ 34 ถึง 35* มาแสดงเป็นกราฟดัง*รูปภาพที่ 27* พบว่า ลักษณะกราฟของปริมาณคำหยุดโดยใช้วิธี TF-IDF มีลักษณะเดียวกับวิธีอย่างเอ็นแกรมแบบตัวอักษร แต่ปริมาณคำหยุดที่พบมีเยอะกว่ามาก เนื่องจากเรามีการใช้ทั้ง ยูนิแกรม ไบแกรม และไตรแกรม จึงทำให้ปริมาณคำที่ได้มีมากกว่าวิธีอย่างเอ็นแกรมแบบตัวอักษร ซึ่งแสดงว่าเราสามารถใช้ช่วงอ้างอิงดังกล่าวเหมือนวิธีอย่างเอ็นแกรมแบบตัวอักษรได้ โดยเพื่อให้เห็นภาพฐานข้อมูลของคำหยุด โดยใช้วิธี TF-IDF ผู้วิจัยจึงเลือกช่วงอ้างอิงย้อนหลัง 30 วัน จนถึง 180 วัน ของข้อมูล 15 เดือน จำนวน 286 คำ มาแสดงใน *ตารางที่ 36*

1	2	3	4	5	7	10	13	14	18	19	20	21	23	25	26	27	28	30	31	61
2018	2561	2562	-	live	เกม	เขา	เข้า	เจอ	เจ้าของ	เจาะ	เช้า	เช้า นี้	เชียร์	เด็ก	เด็ด	เดิน	เดียว	เต็ม	เตรียม	เท่า
เป็น	เปิด	เปิดใจ	เผย	เพลง	เพื่อ	เมื่อ	เรา	เริ่ม	เรื่อง	เลข	เลย	เลย นะ	เล่า	เลือก	เวลา	เสาร์	เสีย	เสียง	เห็น	แข่งขัน
แต่	แบบ	แบบ นี้	แฟน	แรก	แล้ว	และ	แสน	โดน	โดย	โรง	โลก	ใคร	ใช้	ใด	ใน	ให้	ใหญ่	ใหม่	ได้	ได้ ทาง
ได้ ที่	ไทย	ไป	ไม่	ไม่ ได้	ไม่ มี	ไหน	ก็	กด	กลับ	กลาง	กว่า	ก่อน	กัน	กัน ได้	กับ	การ	กำลังใจ	กำลังใจ ให้	ขนาด	ขอ
ของ	ข่าว	ขึ้น	คน	ครั้ง	ครั้ง แรก	ครั้ง ที่	คลิป	ความ	คะ	ค่ะ	คำ	คิด	คืน	คืน นี้	คือ	คุณ	คู่	ฆ่า	ง	งาน
จริง	จะ	จัด	จันทร์	จันทร์ ที่	จับ	จาก	จีน	ฉาก	ชน	ชม	ชม ได้	ชม ย้อนหลัง	ชวน	ช่วย	ช่อง	ชาว	ชิง	ชีวิต	ชื่อ	ญี่ปุ่น
ด้วย	ดัง	ดี	ดี ๆ	ดู	ต่อ	ต้อง	ตอน	ตัว	ตาม	ตาย	ตำรวจ	ติด	ติดตาม	ถึง	ถูก	ทั้ง	ทาง	ทำ	ทำ ให้	ที่
ที่ 1	ที่ 20	ที่ จะ	ที่ นี่	ทีม	ที่สุด	ทุก	ทุก วัน	น	นะ	นะ คะ	นัก	น่า	นาย	นายก	นายก ฯ	นำ	น้ำ	นี่	นี้	นี้ เวลา
บน	บันเทิง	บ้าน	ประจำ	ประจำ วัน	ประจำ วัน จันทร์	ประจำ วัน พฤหัสบดี	ประจำ วัน ศุกร์	ปิด	ปี	ผ่าน	ผู้	พบ	พร้อม	พฤหัสบดี	พฤหัสบดี ที่	พลาด	พ่อ	พา	พิเศษ	พิธี
พี่	พุธ	ฟัง	ภัย	มา	มาก	มี	ยอด	ย้อนหลัง	ย้อนหลัง ได้	ยัง	ยิง	รถ	รวม	ร่วม	ร้อง	รอบ	รัก	รับ	รับ ชม	รายการ
รุ่น	รู้	ลง	ละคร	ล้าน	ลูก	วัน	วัน จันทร์	วัน จันทร์ ที่	วัน ที่	วัน นี้	วัน พฤหัสบดี	วัน พฤหัสบดี ที่	วัน พุธ	วัน ศุกร์	วัน ศุกร์ ที่	วัน อังคาร	วัน อาทิตย์	ว่า	ศุกร์	ศุกร์ ที่
ส่ง	สด	สด ๆ	สนุก	สอบ	สัก	สามารถ	สามารถ รับ	สาว	สำหรับ	สี	สุข	สุด	หน้า	หนุ่ม	หมายเลข	หรือ	หลัง	หา	ห้าม	ห้าม พลาด
อย่า	อย่าง	อยู่	ออก	ออนไลน์	อะไร	อากาศ	อาทิตย์	อีก	ฯ	ๆ	ๆ ได้	ๆ ที่								

ตารางที่ 36 ตารางแสดงคำหยุดทั้งหมด 286 คำ โดยใช้วิธี TF-IDF ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562

จากตารางแสดงคำหยุดทั้งหมด 286 คำ พบว่ามีคำหยุดบางคำ มีลักษณะเป็นคำสำคัญที่เป็นกระแสที่กลับมาเกิดขึ้นใหม่อีกครั้ง เช่นคำว่า “จีน” “ญี่ปุ่น” หรือ “นายก” เป็นต้น แต่โดยข้อมูลโดยส่วนมากยังคงเป็นข้อมูลที่ไม่มีใจความสำคัญที่เกี่ยวข้องกับกระแสแต่อย่างใด แต่เนื่องด้วยปริมาณคำหยุดที่ได้นั้นมากกว่าการใช้ช่วงอ้างอิงย้อนหลัง 60 วัน จนถึง 120 วัน ถึง 50 คำ งานวิจัยนี้จึงเลือกใช้ฐานข้อมูลคำหยุดนี้ในการสกัดคำสำคัญ โดยมีผลลัพธ์ที่ได้แสดงไว้ใน*ตารางที่ 37*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	เลือกตั้ง
2019-03-26	เลือกตั้ง
2019-03-25	*ผล* | *ทั่วไทย* | ครอบครัว | *คะแนน* | 62 | เลือกตั้ง
2019-03-24	เลือกตั้ง 62 | *เขต* | เลือกตั้ง | 62
2019-03-23	เลือกตั้ง
2019-03-22	เลือกตั้ง 62 | 62
2019-03-21	เลือกตั้ง
2019-03-18	เลือกตั้ง
2019-03-15	15
2019-03-14	เลือกตั้ง
2019-03-11	62
2019-02-04	ต้อนรับ
2018-12-04	live สด
2018-09-26	29
2018-09-22	เอง
2018-09-21	ภาพยนตร์ | เพิ่มเติม
2018-08-27	ทาง ช่อง
2018-08-24	ทาง ช่อง
2018-06-28	ทีม หมู | ค้นหา 13 | ถ้ำหลวง | 13 ชีวิต | *ป่า* | *ถ้ำ* | *บอล* | *หมู* | หมู ป่า | ทีม หมู ป่า | *รอด* | *ค้นหา*
2018-06-27	13 ชีวิต | *ค้นหา*
2018-06-26	*ถ้ำ* | *ค้นหา*
2018-06-25	อังคาร

วันที่	คำสำคัญที่เป็นกระแส
2018-06-02	นั้น
2018-03-27	สู่
2018-03-23	สุดท้าย | ทัวร์ | ศพ | ลุ้น
2018-03-20	ข้อ | ผล
2018-03-14	หรือ ไม่
2018-02-16	นทาง

ตารางที่ 37 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF-IDF โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 จาก 5 เสียง ที่เป็นคำกำกวม

จากข้อมูลพบว่า วิธี TF-IDF โดยใช้เครื่องมือตัดคำ deepcut ให้ผลลัพธ์ใกล้เคียงกับวิธีเอ็นแกรมแบบตัวอักษร โดยจะพบว่า ข้อดีของวิธีนี้คือ ผลลัพธ์ที่ได้จะได้ออกมาเป็นคำของแกรม ซึ่งจะพบว่าแต่ละคำจะไม่มีเศษของตัวอักษรของคำข้างเคียงติดมาด้วยเหมือนวิธีอย่างเอ็นแกรมแบบตัวอักษร และแกรมความยาว 1 คำนั้น บางครั้งให้ผลลัพธ์เป็นคำหยุดที่มีความเกี่ยวข้องกับกระแส แต่เนื่องจากความยาวน้อยเกินไป หากอยู่โดด ๆ โดยไม่มีคำอื่นปรากฏด้วย ก็ไม่สามารถตีความกระแสออกมาได้ แต่เนื่องจากเราพิจารณาแกรมความยาว 1 ถึง 3 คำ เท่านั้น ทำให้เราไม่พบคำสำคัญที่ยาวกว่านี้ เช่น “ค้นหา13ชีวิตติดถ้ำหลวง” เหมือนวิธีอย่างเอ็นแกรมแบบตัวอักษร ที่เราได้มีการรวมแกรมของตัวอักษรกลับไปเป็นคำสำคัญแล้ว ทำให้ไม่มีข้อจำกัดของความยาวนั่นเอง

การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF-IDF

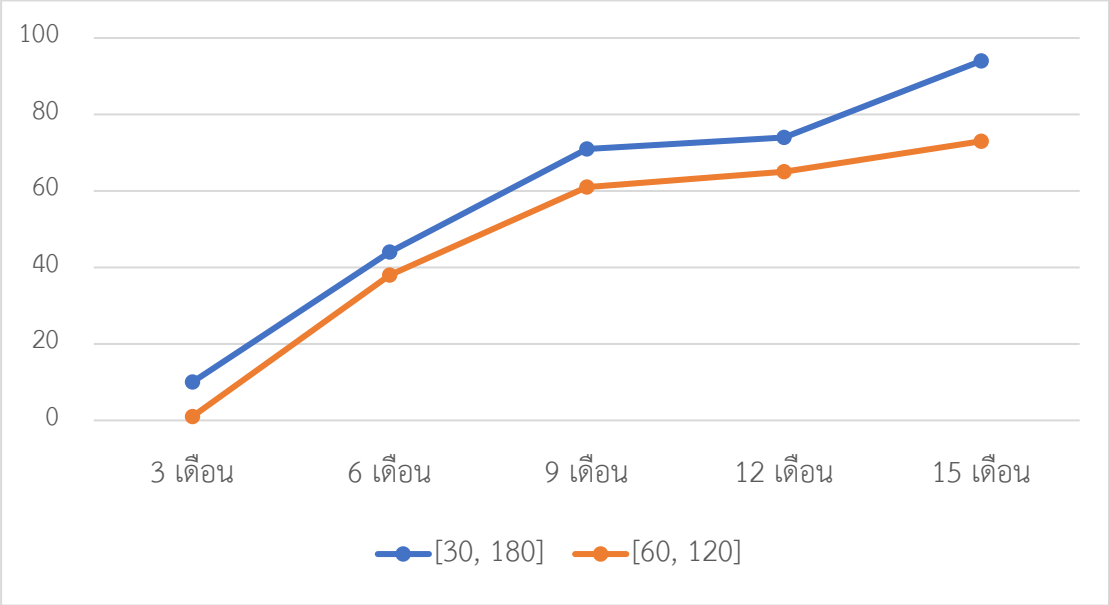
ในการทดลองนี้ เราอยากพบเห็นผลลัพธ์ว่า ถ้าหากเราไม่พิจารณา ยูนิแกรม จะส่งผลอย่างไรต่อฐานข้อมูลคำหยุด และคำสำคัญที่เป็นกระแสที่สามารถสกัดออกมาได้ โดยในขั้นตอนของการหาค่า TF-IDF เราจะนำเพียง ไบแกรม และไตรแกรม ไปหาค่า TF-IDF และใช้วิธีการแบ่งกลุ่มแบบเคมีนและวิธีข้อศอก ในการหาคำสำคัญของแต่ละเพจในแต่ละวัน แล้วจึงนำคำสำคัญเหล่านั้น มาหาค่า DF เพื่อหาว่าคำสำคัญใดปรากฏขึ้นในหลายเพจ โดยใช้วิธีแบ่งกลุ่มแบบเคมีน และวิธีข้อศอก แล้วจึงนำคำสำคัญที่น่าจะเป็นกระแสในแต่ละวัน มาสร้างฐานข้อมูลคำหยุด และกรองคำหยุดออก เพื่อให้ได้คำสำคัญที่เป็นกระแส โดยมีการเพิ่มประสิทธิภาพการกรองด้วยกฎของคำหยุดที่เกี่ยวข้องกับเดือน และเลือกคำที่เป็นกระแสที่เกิดมาอย่างน้อย 1 วัน โดยมีผลการทดลองแสดงไว้ใน*ตารางที่ 38 และ 39*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	94 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	74 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	71 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	44 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	10 คำ

ตารางที่ 38 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	73 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	65 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	61 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	38 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	1 คำ

ตารางที่ 39 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน



รูปภาพที่ 28 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม กับจำนวนเดือนที่ใช้สร้างฐานข้อมูล โดยเส้นกราฟแทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

จาก*ตารางที่ 38 ถึง 39* เมื่อนำมาแสดงเป็นกราฟดัง*รูปภาพที่ 28* พบว่า ลักษณะของกราฟมีความใกล้เคียงกับทั้งวิธี TF-IDF ที่พิจารณายูนิแกรม และวิธีอย่างเอ็นแกรมแบบตัวอักษร แต่มีปริมาณคำหยุดลดลงอย่างเห็นได้ชัดเมื่อเทียบกับวิธี TF-IDF ที่พิจารณายูนิแกรม ซึ่งเป็นเพราะการเลือกใช้แค่ ไบแกรม และไตรแกรม นั้นเป็นการกรองคำหยุดโดยใช้ลักษณะลำดับการติดกันของคำมาช่วย ทำให้แกรมที่มีความถี่สูงเท่านั้น ที่มีโอกาสที่จะมีค่า TF-IDF เยอะ โดยคำหยุดที่หลงเหลือมักเป็นคำหยุดที่มีค่าของ DF เยอะเมื่อเปรียบเทียบกับหลายเพจ แต่โดนกรองออกด้วยช่วงเวลานั่นเอง เพื่อให้เห็นภาพ จึงขอแสดงคำหยุดทั้ง 94 คำ ที่เป็นปริมาณคำหยุดที่มากที่สุด เนื่องจากใช้ช่วงอ้างอิงที่กว้างตั้งแต่ ย้อนหลัง 30 วัน จนถึง 180 วัน ไว้ใน*ตารางที่ 40*

เข้า ไป | เช้า นี้ | เต็ม ๆ | เท่า นั้น | เลย นะ | เสีย ชีวิต | แบบ นี้ | แฟน ๆ | ใคร จะ | ใน รายการ | ได้ เลย | ได้ ใน | ได้ ทาง | ได้ ที่ | ไป ด้วย | ไป ด้วย กัน | ไป พร้อม | ไม่ ได้ | ไม่ มี | กลับ มา | กัน ได้ | กัน นะ | การ เมือง | คน ที่ | ครั้ง ที่ | คืน นี้ | จะ เป็น | จะ ทำ | จะ ทำ ให้ | จะ มี | จันทร์ - | จันทร์ ที่ | ชม ได้ | ชม ย้อนหลัง | ชม สด | ด้วย กัน | ดี ๆ | ดู กัน | ต้อง มา | ติด อยู่ | ติดตาม ชม | ติดตาม รายการ | ถ่ายทอด สด | ทาง ช่อง | ทำ ให้ | ทำ งาน | ที่ 20 | ที่ จะ | ที่ นี่ | ทุก คน | ทุก วัน | ทุก วัน เสาร์ | ทุก วัน พุธ | ทุก วัน อาทิตย์ | นทาง ช่อง | นะ คะ | นายก ฯ | นี้ เวลา | นี้ จะ | นี้ พบ | นี้ พบ กับ | ประจำ วัน | ประจำ วัน จันทร์ | ปลอด ภัย | ผู้ หญิง | พบ กับ | มา แล้ว | มา ดู | ย้อน หลัง | ย้อนหลัง ได้ | ย้อนหลัง ได้ ทาง | ยัง ไง | ยัง ไม่ | รับ ชม | ราย ได้ | ลูก สาว | วัน เสาร์ | วัน จันทร์ | วัน จันทร์ ที่ | วัน ที่ | วัน นี้ | วัน นี้ จะ | วัน นี้ พบ | วัน พุธ | วัน ศุกร์ | วัน อังคาร | วัน อาทิตย์ | สด ๆ | หรือ ไม่ | ห้าม พลาด | อย่าง ไร | อาทิตย์ เวลา | ๆ กับ | ๆ ที่

ตารางที่ 40 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแสโดยใช้วิธี TF-IDF ด้วยฐานข้อมูลคำหยุดที่ใช้ข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน จนถึง 180 วัน

จาก*ตารางที่ 40* พบคำว่าคำส่วนใหญ่ไม่อยู่ในฐานข้อมูลคำหยุดเมื่อพิจารณาคู่กับยูนิแกรม ซึ่งอาจเป็นพบคำดังกล่าวมีความถี่ไม่มากพอ จึงถูกกรองออกไปก่อนด้วยค่า TF-IDF แต่เมื่อไม่พิจารณายูนิแกรม ทำให้มีหลายคำที่กลับมาปรากฏอีกครั้ง และยังปรากฏคำว่า “การ เมือง” ซึ่งเป็นคำหยุดที่มีความกำกวม เหมือนตอนพิจารณาคำหยุดของวิธีเอ็นแกรมแบบตัวอักษร ซึ่งเมื่อนำคำหยุดเหล่านี้ไปกรองคำสำคัญที่เป็นกระแสออกมา โดยแสดงผลลัพธ์ไว้ใน*ตารางที่ 41*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	จัดตั้ง รัฐบาล | ไป กับ | ผล การ | ไม่ เป็น | live คน | thailandelection 2019 | ผล เลือกตั้ง | เพื่อ ไทย | อนาคต ใหม่ | เลือกตั้ง 62

วันที่	คำสำคัญที่เป็นกระแส
2019-03-26	บัตร เลือกตั้ง | ข่าว เด่น | ผล การ | สรุป ข่าว เด่น | สรุป ข่าว | พรรค อนาคต | อนาคต ใหม่ | เลือกตั้ง 62
2019-03-24	บัตร เลือกตั้ง | เลือกตั้ง 62
2019-03-23	เลือกตั้ง ล่วงหน้า | กัน ค่ะ | ที่ มี | เลือกตั้ง 62
2019-03-22	ที่ 2 | ข่าว เด่น | มา ทำ | เลือกตั้ง 62
2019-03-21	ครั้ง แรก | ข่าว เด่น | เลือกตั้ง 62
2019-03-20	เลือกตั้ง 62
2019-03-18	เลือกตั้ง ล่วงหน้า
2019-03-14	เลือกตั้ง 62 | ไป กับ
2018-12-31	ปี ใหม่
2018-12-27	2 ขวบ
2018-12-04	live สด
2018-09-27	2 คน | ติดตาม ละคร
2018-08-30	กด หมายเลข
2018-06-28	ผู้ สูญหาย | 13 ชีวิต | *ชีวิต ติด* | จาก ถ้ำ | จาก ถ้ำหลวง | *ออก จาก* | ที่ ถ้ำหลวง | 13 ชีวิต ติด | สูบ น้ำ | ออก จาก ถ้ำ
2018-06-27	ผู้ สูญหาย | ตัว เอง | หน่วย ซีล | ค้นหา 13 | 13 ชีวิต | ค้นหา 13 ชีวิต | ใน ถ้ำ | ถ้ำ หลวง | สูบ น้ำ
2018-06-26	ไม่ ใช่ | *การ ค้นหา*
2018-06-25	อังคาร เวลา
2018-06-24	รับ สารภาพ | คำ รับ | คำ รับ สารภาพ
2018-06-19	จะ อยู่ | อังคาร เวลา
2018-06-12	- อังคาร | - อังคาร เวลา | อังคาร เวลา | จันทร์ - อังคาร
2018-06-11	ทุก วัน จันทร์
2018-06-04	วัน จันทร์ - | ได้ ทุก
2018-03-22	30 ล้าน | หวย 30 | หวย 30 ล้าน
2018-03-11	กับ การ

ตารางที่ 41 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF-IDF ที่ไม่พิจารณายูนิแกรม โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 เสียง

จาก*ตารางที่ 41* พบว่า คำกำกวมลดลงอย่างมาก และปรากฏคำที่เป็นกระแสมากขึ้นกว่าการพิจารณายูนิแกรม โดยมีการปรากฏคำที่เป็นกระแสใหม่ ๆ เช่น คำว่า “เลือกตั้ง ล่วงหน้า” หรือ “หวย 30 ล้าน” ซึ่งไม่ปรากฏมาก่อนทั้งจากวิธีเอ็นแกรมแบบตัวอักษร และวิธี TF-IDF ที่พิจารณายูนิแกรม เนื่องจากเดิมคำเหล่านี้มีค่าความถี่ต่ำ มักจะโดนกรองออกเมื่อพิจารณาคู่กับยูนิแกรม แต่เมื่อไม่มียูนิแกรม ทำให้คำเหล่านี้หลุดลอดและปรากฏออกมาเป็นคำสำคัญที่เป็นกระแส

การเปรียบเทียบผลเมื่อใช้วิธี TF ด้วยเครื่องตัดคำ deepcut แทนวิธีเอ็นแกรมแบบตัวอักษร

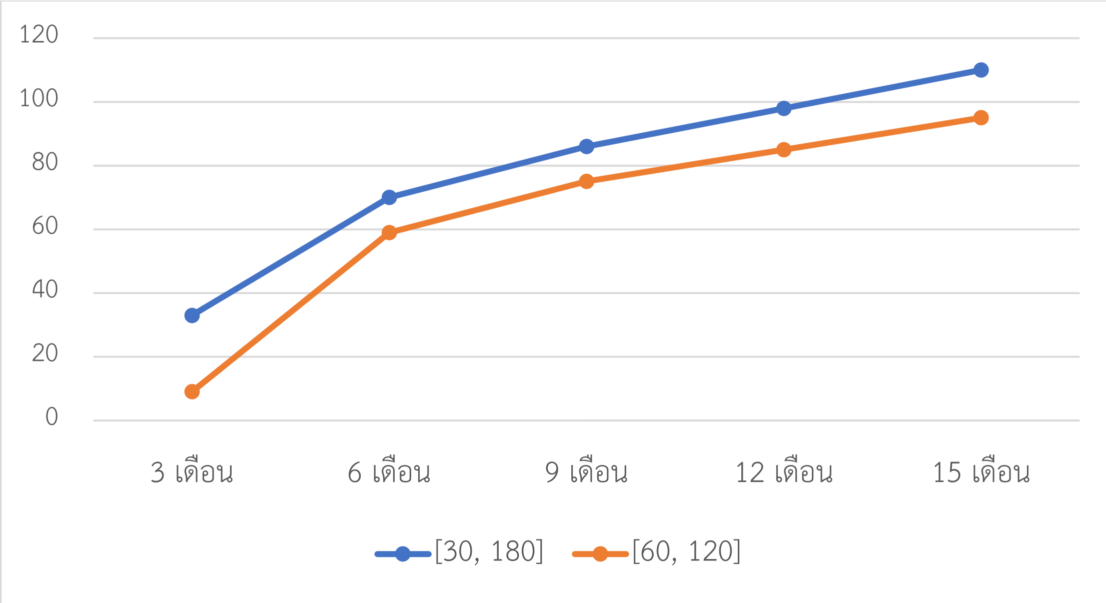
ในการทดลองนี้ เราอยากทดสอบว่าหากเราไม่ใช้วิธีการอย่าง TF-IDF หาคำสำคัญของแต่ละเพจ แต่จะใช้วิธีอย่าง TF หาคำที่มีความถี่สูงเหมือนวิธีเอ็นแกรมแบบตัวอักษร แต่เปลี่ยนไปใช้เอ็นแกรมของคำ โดยไม่มีการรวมแกรมเข้าด้วยกัน และเมื่อเราได้แกรมที่มีค่า TF สูงแล้ว เราก็กรองแกรมเหล่านั้นโดยใช้วิธีการแบ่งกลุ่มแบบเคมีน และวิธีข้อศอก แบบเดียวกับวิธีเอ็นแกรมแบบตัวอักษร และหาค่า DF ของแต่และแกรมจากข้อมูลหลายเพจ เพื่อหาแกรมที่เป็นกระแส และนำมาสร้างฐานข้อมูลคำหยุด เพื่อวิเคราะห์คำสำคัญที่เป็นกระแสที่ได้ออกมาจากวิธีนี้ โดยแสดงผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 42 และ 43*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	110 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	98 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	86 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	70 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	33 คำ

ตารางที่ 42 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	95 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	85 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	75 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	59 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	9 คำ

ตารางที่ 43 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ตามจำนวนเดือน
เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน



รูปภาพที่ 29 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF กับจำนวนเดือนที่ใช้ในการสร้างฐานข้อมูลคำหยุด โดยแต่ละเส้นกราฟ แทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

จาก*ตารางที่ 42 ถึง 43* เมื่อเรานำมาแสดงเป็นกราฟดัง*รูปภาพที่ 29* พบว่า กราฟมีลักษณะความสัมพันธ์แบบเดียวกับวิธีอย่าง TF-IDF หรือวิธีเอ็นแกรมแบบตัวอักษร โดยงานวิจัยนี้จะทดลองนำคำหยุดที่ได้จากข้อมูลอ้างอิงย้อนหลัง 30 วัน จนถึง 180 วัน ที่มีปริมาณคำหยุดทั้งหมด 110 คำ ไปใช้ในการสกัดคำสำคัญที่เป็นกระแสออกมา โดยยังคงมีการเพิ่มกฎของคำหยุดที่เกี่ยวข้องกับเดือนและเลือกคำที่เป็นกระแสอย่างน้อย 1 วัน เหมือนวิธีก่อน ๆ โดยแสดงผลไว้ใน*ตารางที่ 44 และ 45*

1 | 2 | 3 | 4 | 7 | 8 | 13 | 28 | 31 | 2561 | 2562 | - | live | เข้า | เช้า | เด็ก | เป็น | เปิด | เพื่อ | เรื่อง | เลย | เวลา | เสาร์ | แต่ | แบบ | แล้ว | และ | โลก | ใคร | ใน | ให้ | ใหม่ | ได้ | ได้ ทาง | ได้ ที่ | ไทย | ไป | ไม่ | ก็ | กด | กัน | กับ | การ | ของ | ข่าว | คน | ความ | คะ | คืน | คุณ | งาน | จะ | จาก | ชม | ช่อง | ชีวิต | ด้วย | ดี | ดู | ต้อง | ตอน | ตัว | ติดตาม | ถึง | ถูก | ทาง | ทำ | ที่ | ที่ นี่ | ทุก | ทุก วัน | น | นทาง | นะ | นะ คะ | นัก | นี่ | นี้ | ประจำ | ประจำ วัน | ปี | ผู้ | พบ | พร้อม | พุธ | มา | มี | ย้อนหลัง | ยัง | รถ | ร่วม | รับ | รับ ชม | ร้าน | รายการ | ลูก | วัน | วัน นี้ | วัน พุธ | ว่า | สด | สาว | สุด | หนุ่ม | หมายเลข | หรือ | หลัง | อยู่ | ออก | ๆ

ตารางที่ 44 ตารางแสดงคำหยุดทั้งหมด 110 คำ โดยใช้วิธี TF ที่ถูกสร้างขึ้นโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน โดยใช้ข้อมูลโพสต์ตั้งแต่ 1 มกราคม 2561 ถึง 31 มีนาคม 2562

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	จัดตั้ง รัฐบาล | *รัฐบาล* | 62 | เลือกตั้ง 62 | เลือกตั้ง | *จัดตั้ง*
2019-03-26	เลือกตั้ง | เลือกตั้ง 62 | 62
2019-03-25	*คะแนน* | เลือกตั้ง | เลือกตั้ง 62 | 62
2019-03-24	เลือกตั้ง | เลือกตั้ง 62 | 62
2019-03-23	เลือกตั้ง | เลือกตั้ง 62 | 62
2019-03-22	เลือกตั้ง
2019-03-21	ก่อน | เลือกตั้ง
2019-03-20	เลือกตั้ง | 62 | เลือกตั้ง 62
2019-03-18	เลือกตั้ง
2019-02-04	จันทร์ | วัน จันทร์ | วัน จันทร์ ที่ | จันทร์ ที่
2018-12-20	วัย
2018-12-13	ได้ ใน
2018-06-29	*ทีม*
2018-06-28	*ถ้ำ* | *ค้นหา* | ถ้ำหลวง | *น้ำ* | *ติด* | *ทีม* | 13 ชีวิต
2018-06-27	*ทีม*
2018-06-26	*ค้นหา* | ถ้ำหลวง | ก่อน | *น้ำ* | *ติด* | *ทีม*
2018-06-25	*ติด* | ต่อ
2018-05-11	ไหน | ได้ เลย
2018-03-08	ห้าม | ห้าม พลาด | พลาด
2018-02-04	ย้อนหลัง ได้
2018-01-28	ย้อนหลัง ได้

ตารางที่ 45 ตารางแสดงข้อมูลคำสำคัญที่เป็นกระแสโดยใช้วิธี TF โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 จาก 5 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 จาก 5 เสียง

จาก*ตารางที่ 44* พบว่า คำหยุดส่วนใหญ่มักเป็นคำโดด และคำที่เป็นกระแสที่พบ พบว่ามีปริมาณน้อยกว่าวิธีอย่าง TF-IDF อย่างเห็นได้ชัด ซึ่งเกี่ยวข้องกับความถี่ของคำที่เป็นกระแสดังกล่าว หากความถี่ของคำเหล่านั้นไม่มากเทียบพอ ก็จะโดนกรองออกในชั้นการแบ่งกลุ่มของเคมีน จึงทำให้กระแสที่ออกมามักเป็นคำที่ปรากฏในข่าวค่อนข้างมากในแต่ละวัน

การทดลองปรับความยาวเริ่มต้นของแกรมเมื่อใช้วิธี TF

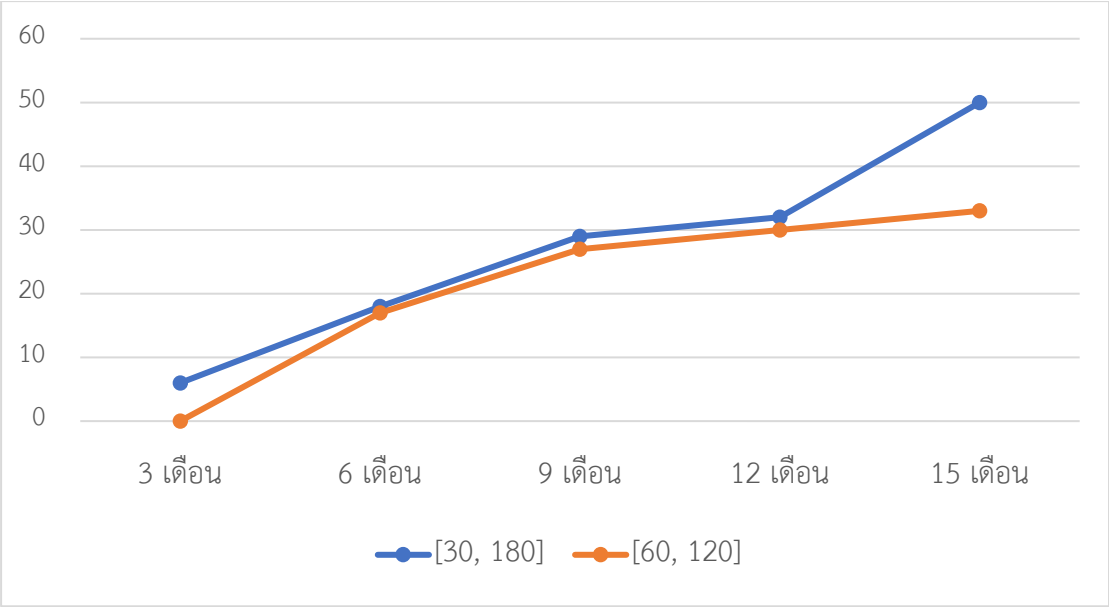
ในการทดลองนี้เราอยากทราบว่า หากเราไม่พิจารณายูนิแกรม ผลที่ได้รับจะแตกต่างกันหรือไม่ และมีโอกาสที่เราจะได้คำสำคัญที่เป็นกระแสเพิ่มขึ้นเหมือนอย่างวิธีอย่าง TF-IDF หรือไม่ โดยในขั้นตอนของการหาค่า TF เราจะนำเพียงไบแกรม และไตรแกรม ไปหาค่า TF เท่านั้น และใช้วิธีการแบ่งกลุ่มแบบเคมีน และวิธีข้อศอก ในการหาคำสำคัญของแต่ละเพจในแต่ละวัน แล้วจึงนำคำสำคัญเหล่านั้น มาหาค่า DF เพื่อหาว่าคำสำคัญใดปรากฏขึ้นในหลายเพจ โดยใช้วิธีแบ่งกลุ่มแบบเคมีน และวิธีข้อศอก แล้วจึงนำคำสำคัญที่น่าจะเป็นกระแสในแต่ละวัน มาสร้างฐานข้อมูลคำหยุด และกรองคำหยุดออก เพื่อให้ได้คำสำคัญที่เป็นกระแส โดยมีการเพิ่มประสิทธิภาพการกรองด้วยกฎของคำหยุดที่เกี่ยวข้องกับเดือน และเลือกคำที่เป็นกระแสที่เกิดมาอย่างน้อย 1 วัน โดยมีผลการทดลองแสดงไว้ใน *ตารางที่ 46 และ 47*

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	50 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	32 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	29 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	18 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	6 คำ

ตารางที่ 46 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 30 วัน แต่ไม่เกิน 180 วัน

วันที่เริ่มต้น	วันที่สิ้นสุด	จำนวนเดือน	ปริมาณคำหยุด
1 มกราคม 2561	31 มีนาคม 2562	15 เดือน	33 คำ
1 มกราคม 2561	31 ธันวาคม 2561	12 เดือน	30 คำ
1 มกราคม 2561	31 กันยายน 2561	9 เดือน	27 คำ
1 มกราคม 2561	30 มิถุนายม 2561	6 เดือน	17 คำ
1 มกราคม 2561	31 มีนาคม 2561	3 เดือน	0 คำ

ตารางที่ 47 ตารางแสดงปริมาณคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม ตามจำนวนเดือน เมื่อเราจำแนกคำหยุดโดยอ้างอิงข้อมูลก่อนหน้า 60 วัน แต่ไม่เกิน 120 วัน



รูปภาพที่ 30 กราฟแสดงความสัมพันธ์ระหว่างจำนวนคำหยุดด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม กับจำนวนเดือนที่ใช้สร้างฐานข้อมูล โดยเส้นกราฟแทนประเภทของการอ้างอิงจำนวนวันก่อนหน้า

จาก*ตารางที่ 46 ถึง 47* เมื่อนำมาแสดงเป็นกราฟดัง*รูปภาพที่ 30* พบว่า ลักษณะของกราฟมีความใกล้เคียงกับทั้งวิธี TF ที่พิจารณายูนิแกรม และวิธีอย่างเอ็นแกรมแบบตัวอักษรแต่มีปริมาณคำหยุดลดลงอย่างเห็นได้ชัดเมื่อเทียบกับวิธี TF ที่พิจารณายูนิแกรม ซึ่งเป็นลักษณะเดียวกับวิธีอย่าง TF-IDF เพราะไบแกรม และไตรแกรม ที่มีค่าTF สูง และมีค่า DF สูง มีปริมาณน้อยกว่าแบบยูนิแกรม จึงทำให้ฐานข้อมูลคำหยุดที่ได้มีปริมาณน้อยลงไปด้วย และเมื่อนำฐานข้อมูลคำหยุดทั้ง 50 คำ จากการใช้ข้อมูลอ้างอิงย้อนหลังตั้งแต่ 30 วัน จนถึง 180 วัน ไปสกัดคำสำคัญที่เป็นกระแส โดยแสดงผลลัพธ์ที่ได้ไว้ใน*ตารางที่ 48*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	ติดตาม ข่าวสาร | มา แล้ว | thailandelection 2019 | *ทิศ ทั่วไทย* | เพื่อ ไทย | เลือกตั้ง 2562 | จะ มี | ย้อนหลัง ได้ ทาง | การ เลือกตั้ง | 62 thailandelection 2019 | เลือกตั้ง 62 | ตั้ง รัฐบาล | เลือกตั้ง ครั้ง นี้ | อนาคต ใหม่ | ไป ตาม | จัดตั้ง รัฐบาล | live เรื่อง | ครั้ง นี้ | ทำ ให้ | 62 thailandelection | เลือกตั้ง 62 thailandelection | พรรค เพื่อไทย | เลือกตั้ง ครั้ง | ทุก ทิศ | ใน การ | ได้ ที่ นี่ | น่า รัก | ไม่ มี | *ทุก ทิศ ทั่วไทย*
2019-03-26	ได้ ที่ นี่ | บัตร เลือกตั้ง | ผล การ เลือกตั้ง | thailandelection 2019 | มา แล้ว | เลือกตั้ง 62 | ใน การ | อนาคต ใหม่ | ผล การ | ไม่ มี | ย้อนหลัง ได้ ทาง | การ

	เลือกตั้ง
2019-03-25	thailandelection 2019 | เลือกตั้ง 62 | อนาคต ใหม่ | ผล การ | คะแนน เลือกตั้ง | ผล คะแนน | ผล คะแนน เลือกตั้ง | การ เลือกตั้ง
2019-03-24	ใช้ สิทธิ | บัตร เลือกตั้ง | สิทธิ เลือกตั้ง | 62 เลือกตั้ง | เขต เลือกตั้ง | นับ คะแนน | ใช้ สิทธิ์ | เลือกตั้ง 62 | ลง คะแนน | อนาคต ใหม่ | เลือกตั้ง 2562 | เลือกตั้ง 62 เลือกตั้ง | 24 มี | การ เลือกตั้ง
2019-03-23	โค้ง สุดท้าย | ได้ ที่ นี่ | เลือกตั้ง 62 | อนาคต ใหม่ | การ เลือกตั้ง
2019-03-22	เลือกตั้ง 62
2019-03-21	เลือกตั้ง 62
2019-03-20	ค 62 | เลือกตั้ง 62
2019-03-19	ได้ ที่ นี่ | เลือกตั้ง 62 | วัน อังคาร | ย้อนหลัง ได้ ทาง
2019-03-18	เลือกตั้ง 62 | เลือกตั้ง ล่วงหน้า
2019-03-17	เลือกตั้ง 62
2019-03-16	เลือกตั้ง 62
2019-03-15	เลือกตั้ง 62
2019-03-14	เลือกตั้ง 62
2019-03-12	เลือกตั้ง 62 | อยู่ ใน | เกิด ขึ้น | ย้อนหลัง ได้ ทาง
2019-03-03	เสาร์ อาทิตย์
2019-03-01	ที่ 1
2019-01-27	อาทิตย์ เวลา
2018-12-31	ปี ใหม่
2018-12-27	2 ขวบ | ปี ใหม่
2018-12-26	ปี ใหม่
2018-09-28	ที่ มา
2018-09-21	อย่า ลืม
2018-06-28	13 ชีวิต | หมู ป่า | ทีม หมู ป่า | ทีม หมู | ถ้ำ หลวง
2018-06-27	หน่วย ซีล | หมู ป่า | ค้นหา 13 | ใน ถ้ำ หลวง | ใน ถ้ำ | ถ้ำ หลวง | 13 ชีวิต
2018-06-26	ค้นหา 13 | ใน ถ้ำ หลวง | *อยู่ ใน* | อยู่ ใน ถ้ำ | *ติด อยู่ ใน* | *ติด อยู่* | ใน ถ้ำ | 13 ชีวิต | ถ้ำ หลวง
2018-06-25	ฆ่า หั่น ศพ | ฆ่า หั่น | หั่น ศพ

2018-06-24	ฆ่า หั่น ศพ | ฆ่า หั่น | หั่น ศพ
2018-06-21	ทุก วัน พุธ | พฤหัสบดี เวลา
2018-06-20	ใน รายการ | โทษ ประหาร
2018-06-19	ใน รายการ
2018-06-13	ความ จริง
2018-06-02	ชม ได้
2018-05-11	ได้ เลย
2018-03-27	ราย ได้ | ข่าว ค่ำ
2018-03-25	ค 61
2018-03-24	ย้อนหลัง ได้ ทาง | ค 61
2018-03-23	รถ ทัวร์

ตารางที่ 48 ตารางสรุปคำสำคัญที่เป็นกระแสด้วยวิธี TF ที่ไม่พิจารณายูนิแกรม โดยสีน้ำเงิน (ขีดเส้นใต้) คือคำที่มีเสียงโหวตตั้งแต่ 3 เสียง และสีเขียว (ตัวเอียง) คือคำที่มีเสียงโหวต 2 เสียง

จาก*ตารางที่ 48* พบว่า วิธี TF ที่ไม่พิจารณายูนิแกรม พบคำที่เป็นกระแสมากกว่าแบบพิจารณายูนิแกรม แต่แลกกับการไม่พบคำโดดที่เป็นกระแส โดยยังพบกระแสเพิ่มเติมมาลักษณะเดียวกับการใช้ TF-IDF แต่มีบางกระแสที่แตกต่างกัน เช่นวิธีอย่าง TF-IDF สามารถสกัดกระแสของ “หวย 30 ล้าน” ได้ แต่วิธี TF สามารถสกัดกระแสอย่าง “ฆ่า หั่น ศพ” หรือ “โทษ ประหาร” ได้

การเปรียบเทียบประสิทธิภาพของผลลัพธ์ของวิธีเอ็นแกรมแบบตัวอักษร กับวิธีที่ใช้เครื่องมือในการตัดคำอย่าง deepcut ได้แก่วิธี TF-IDF และวิธี TF

จากผลการทดลองสกัดคำสำคัญที่เป็นกระแส ทั้งวิธีเอ็นแกรมแบบตัวอักษร และวิธีที่ใช้เครื่องมือตัดคำอย่างวิธี TF-IDF และวิธี TF ทุกวิธีสามารถจำแนกคำสำคัญได้แตกต่างกัน โดยผลลัพธ์ของคำที่เป็นกระแสที่ได้เสียงโหวตจากอาสาสมัครตั้งแต่ 3 เสียง จาก 5 เสียง แสดงไว้ใน*ตารางที่ 49*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	62thailandelection | 62thailandelection2019 | thailandelection2019 | เพื่อไทย | เลือกตั้ง | เลือกตั้ง62 | เลือกตั้งครั้งนี้ | การเลือกตั้ง | การเลือกตั้ง62 | จัดตั้งรัฐบาล | ตั้งรัฐบาล | ผลเลือกตั้ง | พรรคเพื่อไทย | อนาคตใหม่
2019-03-26	thailandelection2019 | เพื่อไทย | เลือกตั้ง | เลือกตั้ง62 | การเลือกตั้ง | คะแนน

วันที่	คำสำคัญที่เป็นกระแส
	เลือกตั้ง | คะแนนเสียง | บัตรเลือกตั้ง | ผลการเลือกตั้ง | พรรคอนาคต | อนาคตใหม่
2019-03-25	election | election62thailandelection2019 | thailandelection2019 | เลือกตั้ง | เลือกตั้ง62 | เลือกตั้ง62thai | เลือกตั้ง62thailandelection2019 | เสียงประชา | การเลือกตั้ง | การเลือกตั้ง62 | คะแนนเลือกตั้ง | คะแนนเลือกตั้ง62 | คะแนนเสียง | นับคะแนน | นับคะแนนเลือกตั้ง | ผลคะแนน | ผลคะแนนเลือกตั้ง | อนาคตใหม่
2019-03-24	62เลือกตั้ง | เขตเลือกตั้ง | เพื่อไทย | เลือกตั้ง | เลือกตั้ง2562 | เลือกตั้ง62 | ใช้สิทธิ | ใช้สิทธิ์ | การเลือกตั้ง | นับคะแนน | บัตรเลือกตั้ง | ลงคะแนน | สิทธิ | สิทธิ์ | สิทธิเลือกตั้ง | อนาคตใหม่
2019-03-23	เลือกตั้ง | เลือกตั้ง62 | เลือกตั้งล่วงหน้า | การเลือกตั้ง | อนาคตใหม่
2019-03-22	เลือกตั้ง | เลือกตั้ง62 | สิทธิ
2019-03-21	เลือกตั้ง | เลือกตั้ง62 | ธนาธร | สิทธิ์
2019-03-20	เลือกตั้ง | เลือกตั้ง62 | รตอวัชรินทร์
2019-03-19	เลือกตั้ง | เลือกตั้ง62
2019-03-18	เลือกตั้ง | เลือกตั้งล่วงหน้า
2019-03-17	เลือกตั้ง | เลือกตั้ง62
2019-03-16	เลือกตั้ง | เลือกตั้ง62
2019-03-15	เลือกตั้ง | เลือกตั้ง62
2019-03-14	เลือกตั้ง | เลือกตั้ง62
2019-03-12	เลือกตั้ง | เลือกตั้ง62
2019-03-08	ไทยรักษาชาติ
2018-12-31	ปีใหม่
2018-12-27	ปีใหม่
2018-12-26	ปีใหม่
2018-06-28	13ชีวิต | 13ชีวิตติด | 13ชีวิตติดถ้ำหลวง | 13ชีวิตถ้ำหลวง | 13ชีวิตทีมหมูป่า | เกาะติดปฏิบัติการค้นหาทีมหมูป่า | เชียงราย | ในถ้ำขุนน้ำนางนอนถ้ำหลวงทีมหมูป่า | ในถ้ำหลวง | ในถ้ำหลวงขุนน้ำนางนอน | ค้นหา13 | ค้นหา13ชีวิต | ค้นหา13ชีวิตติดถ้ำหลวง | จากถ้ำ | จากถ้ำหลวง | ติดถ้ำหลวง | ติดถ้ำหลวงขุนน้ำนางนอน | ถ้ำหลวง | ถ้ำหลวงขุนน้ำนางนอน | ถ้ำหลวงหมูป่า13ชีวิต | ที่ถ้ำหลวง | ทีมหมู | ทีม

วันที่	คำสำคัญที่เป็นกระแส
	หมูป่า | ทีมหมูป่าถ้ำหลวง | ผู้สูญหาย | สูบน้ำ | หน่วยซีล | หมูป่า | หมูป่าติดถ้ำ | ออกจากถ้ำ
2018-06-27	13ชีวิต | 13ชีวิตติดถ้ำหลวง | ในถ้ำ | ในถ้ำหลวง | ค้นหา13 | ค้นหา13ชีวิต | ถ้ำหลวง | ผู้สูญหาย | สูบน้ำ | หน่วยซีล
2018-06-26	13ชีวิต | เชียงราย | ในถ้ำ | ในถ้ำหลวง | ค้นหา13 | ถ้ำหลวง | ฟุตบอล | หน่วยซีล | อยู่ในถ้ำ
2018-06-25	cup20 | ฆ่าหั่น | ฆ่าหั่นศพ | ฟุตบอล | หั่นศพ
2018-06-24	ฆ่าหั่น | ฆ่าหั่นศพ | หั่นศพ
2018-06-20	โทษประหาร | ประหาร
2018-03-22	30ล้าน | บุพเพสันนิวาส | หวย30 | หวย30ล้าน

ตารางที่ 49 ตารางแสดงคำสำคัญที่เป็นกระแสทั้งหมดจากวิธีเอ็นแกรมแบบตัวอักษร วิธี TF-IDF และวิธี TF โดยพิจารณายูนิแกรม และไม่พิจารณายูนิแกรม รวม 5 วิธี

วันที่	คำสำคัญที่เป็นกระแส	เอ็นแกรมแบบตัวอักษร	TF-IDFมียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
2019-03-27	14	6	1	6	3	14
2019-03-26	11	5	1	4	2	6
2019-03-25	18	14	1	0	2	7
2019-03-24	16	6	2	2	2	13
2019-03-23	5	2	1	2	2	3
2019-03-22	3	3	1	1	1	1
2019-03-21	4	3	1	1	1	1
2019-03-20	3	3	0	1	2	1
2019-03-19	2	2	0	0	0	1
2019-03-18	2	1	1	1	1	2
2019-03-17	2	1	0	0	0	1
2019-03-16	2	1	0	0	0	1
2019-03-15	2	1	0	0	0	1
2019-03-14	2	2	1	1	0	1

วันที่	คำสำคัญที่เป็นกระแส	เอ็นแกรมแบบตัวอักษร	TF-IDFมียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
2019-03-12	2	1	0	0	0	1
2019-03-08	1	1	0	0	0	0
2018-12-31	1	1	0	1	0	1
2018-12-27	1	1	0	0	0	1
2018-12-26	1	0	0	0	0	1
2018-06-28	30	19	6	8	2	5
2018-06-27	10	5	1	8	0	7
2018-06-26	9	5	0	0	1	6
2018-06-25	5	4	0	0	0	3
2018-06-24	3	0	0	0	0	3
2018-06-20	2	1	0	0	0	1
2018-03-22	4	1	0	3	0	0
รวม	155	89	17	39	19	82

ตารางที่ 50 ตารางเปรียบเทียบจำนวนคำสำคัญที่เป็นกระแสทั้ง 5 วิธี

จาก*ตารางที่ 49* เรานำผลลัพธ์จากทุกวิธีมาเปรียบเทียบกันดังที่แสดงไว้ใน*ตารางที่ 50* พบว่าวิธีเอ็นแกรมแบบตัวอักษรให้คำสำคัญที่เป็นกระแสได้เยอะที่สุด โดยวิธี TF แบบไม่มียูนิแกรมมีจำนวนคำน้อยกว่าอยู่เล็กน้อย โดยจำนวนคำมีความแตกต่างกันในแต่ละวันอย่างชัดเจน และเมื่อเปรียบเทียบการพิจารณายูนิแกรม และไม่พิจารณายูนิแกรมพบว่า ถ้าไม่พิจารณายูนิแกรม จะให้คำสำคัญที่เป็นกระแสเยอะกว่าอย่างเห็นได้ชัดทั้งวิธี TF-IDF และ TF

โดยเมื่อพิจารณาปริมาณคำหยุดจากทั้ง 5 วิธีพบว่าวิธี TF แบบไม่มียูนิแกรมมีจำนวนข้อมูลคำหยุดน้อยที่สุด และรองลงมาคือ TF-IDF แบบไม่มียูนิแกรม ซึ่งแสดงให้เห็นว่าคำหยุดส่วนใหญ่ถูกกรองออกด้วยวิธีเอ็นแกรมแบบคำ และใช้การแบ่งกลุ่มแบบเคมีนและวิธีข้อศอกแทนการใช้ค่าขีดแบ่งโดยเมื่อมีการคิดยูนิแกรมเข้ามา ปริมาณคำหยุดก็จะเพิ่มขึ้นอย่างชัดเจน ดัง*ตารางที่ 51*

วันที่ใช้อ้างอิงย้อนหลัง	เอ็นแกรมแบบตัวอักษร	TF-IDF มียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
30 – 180 วัน	132 คำ	286 คำ	94 คำ	110 คำ	50 คำ
60 – 120 วัน	103 คำ	236 คำ	73 คำ	95 คำ	30 คำ

ตารางที่ 51 ตารางแสดงปริมาณคำหยุดของทั้ง 5 วิธี บนช่วงวันอ้างอิงย้อนหลังทั้ง 2 ช่วง

และเพื่อพิจารณาค่าความผิดพลาดที่เกิดขึ้นในการกรองคำหยุดของทั้ง 5 วิธี จึงนำผลลัพธ์ของคำสำคัญที่เป็นกระแสที่ถูกอาสาสมัครโหวตตั้งแต่ 2 เสียงลงมา จาก 5 เสียง มารวมกัน โดยแสดงผลลัพธ์ไว้ใน*ตารางที่ 52*

วันที่	คำสำคัญที่เป็นกระแส
2019-03-27	62 | liveเรื่อง | liveคน | เลือก | ในการ | ได้ที่นี่ | ไปกับ | ไปตาม | ไม่เป็น | ไม่มี | การเล | ครั้งนี้ | จะมี | ติดตามข่าวสาร | ทำให้ | ทิศทั่วไทย | ทุกทิศ | ทุกทิศทั่วไทย | น่ารัก | ประชา | ผลการ | มาแล้ว | ย้อนหลังได้ทาง | รัฐบาล | รายการเล | หมายเ
2019-03-26	62 | เลือก | ในการ | ได้ที่นี่ | ไม่มี | ข่าวเด่น | คะแนน | งประชา | ประชา | ผลการ | ผลการ | มาแล้ว | ย้อนหลังได้ทาง | สรุปข่าว | สรุปข่าวเด่น
2019-03-25	62 | ครอบครัว | คะแนน | งประชา | ทางการ | ประชา | ผลการ
2019-03-24	62 | 24 มี | เกาะติด | เลือก | ประชา
2019-03-23	62 | เลือก | โค้งสุดท้าย | ได้ที่นี่ | กันค่ะ | ที่มี | ประชา
2019-03-22	62 | เลือก | เลือกตั | ข่าวเด่น | ที่2 | ประชา | มาทำ
2019-03-21	facebookเดียว | liveข | liveค | liveคุย | liveคุยกับ | liveสรุปข่าวเด่นประเด็น | newsupdate | twitter | เชียง | เป็นเ | เป็นก | เพิ่ม | เมื่อ | เรื่องเล่าเช้านี้ | แลนด์ทางช่อง | แลนด์วันนี้ | ได้ในเรื่องเล่าเช้านี้ | ได้ที่นี่ | ก่อน | ข่าวเด่น | ครั้งแรก | ครั้งข่าวส | งเรื่อง | งข่าว | งประเด็น | งประเทศ | ช่องท | ต้องม | ติดตามเพิ่มเติมได้ที่ | ทุกทิศทั่วไทย | นรายการ | ประชาชน | มหมาย | มหมายเลข2 | รับชมได้ทาง | รับชมผ่าน | รับชมผ่านช่องท | รับชมย้อนหลังได้ทางw | หมายเ | อย่างไร | ออนไลน์wwwfacebookcom/
2019-03-20	62 | 0800นทางช่อง | facebooktwitterofficial | liveเที่ยงข่าว | liveข | newsupdate | twitter | เรื่องเล่าเช้านี้ | เล่าเ | เลือก | เลือกตั | เหนือ | ได้ในเรื่องเล่าเช้านี้ | ได้ที่นี่เรื่องเล่าเช้านี้ | ไปตาม | ไปตามชมกัน | ค62 | ความเคลื่อนไหว |

วันที่	คำสำคัญที่เป็นกระแส
	งเรื่อง | จ้าาา | ช่องท | ดประเทศไทย | ติดตามข่าวสาร | นรายการ | ประชา | พูดคุยในประเด็น | มหมาย | มหมายเลข2 | รับชมได้ทาง | รับชมกัน | รับชมผ่าน | รับชมผ่านช่องท | รับชมย้อนหลังได้ทางw | สรุปข่าวเด่นประเด็น | หมายเ | ายได้ | ารเพื่อ ‍พากษ
2019-03-19	facebooktwitterofficial | liveข | liveด | newsupdate | twitter | เที่ยงข่าว | ได้ที่นี่ | ติดตามข่าวสาร | ประชา | พูดคุยในประเด็น | ย้อนหลังได้ทาง | รับชมย้อนหลังได้ทางw | วันอังคาร
2019-03-18	เลือก
2019-03-15	15 | เลือก
2019-03-14	ไปกับ
2019-03-12	newsupdate | เกิดขึ้น | เช้านี้ทางช่อง | เป็นเ | ติดตามข่าวสาร | ที่นี่เรื่อง | นที่1 | ย้อนหลังได้ทาง | วันนี้พบกับ | อยู่ใน
2019-03-11	62
2019-03-10	เสาร์อาทิตย์
2019-03-09	เสาร์อาทิตย์ | มเวลา
2019-03-08	newsupdate
2019-03-03	เสาร์อาทิตย์
2019-03-01	ที่1
2019-02-14	าพันธ
2019-02-07	ประจำวันพ
2019-02-04	จันทร์ | จันทร์ที่ | ต้อนรับ | วันจันทร์ | วันจันทร์ที่
2019-02-01	พร้อมห
2019-01-27	อาทิตย์เวลา
2018-12-28	ข่าวเช้านี้
2018-12-27	2ขวบ | ข่าวเช้านี้ | งวันที่2 | ที่นี | สรุปข่าวเด่นประ
2018-12-26	สรุปข่าวเด่นประ
2018-12-25	งวันที่2
2018-12-20	วัย
2018-12-19	ได้ที่เรื่องเล่าเ

วันที่	คำสำคัญที่เป็นกระแส
2018-12-18	ได้ที่เรื่องเล่าเ
2018-12-14	ชมกัน | ได้ใน
2018-12-13	ได้ใน
2018-12-07	ได้ใน
2018-12-04	livสด
2018-09-28	เพื่อติดตาม | ที่มา | รายการเพื่อ
2018-09-27	2คน | เพื่อติดตาม | แชมป์ | คุยกับ | ติดตามละคร | หมายเลข3
2018-09-26	29 | นข่าวเ
2018-09-24	นข่าวเพิ่มเติมได้ที่
2018-09-22	เอง | งความ | อกจาก
2018-09-21	เพิ่มเติม | ภาพยนตร์ | อย่าลืม
2018-09-11	มข่าว
2018-09-07	เพื่อติดตาม | แอปพลิเคชั | นช่อง | มได้ที่นี่ | ย้อนหลังได้ที่ติดตามข่าว
2018-09-06	newsupdate
2018-08-30	กดหมายเลข
2018-08-29	ย้อนหลังได้ทาง
2018-08-27	ทางช่อง
2018-08-24	ทางช่อง
2018-06-29	ทีม
2018-06-28	เกาะติด | เจ้าหน้าที่ | ค้นหา | ช่วยเหลือ | ชีวิตติด | ติด | ถ้ำ | ทีม | น้ำ | บอล | ป่า | รอด | หมู | ออกจาก
2018-06-27	13ชีวิต | เกาะติด | การค้นหา | ค้นหา | ตัวเอง | ทีม | ปฏิบัติ
2018-06-26	เจ้าหน้าที่ | ไม่ใช่ | ก่อน | กับข่าวเ | กับข่าวส | การค้นหา | กำลังใจ | ค้นหา | ค้นหา | งข่าว | งข่าวเล่าเรื่อง | ติด | ติดอยู่ | ติดอยู่ใน | ถ้ำ | ทีม | น้ำ | นิวส์ | พร้อมก | ยังมี | ระดับ | ลื่อน | วาม | หน่วย | อยู่ใน | าแล้ว
2018-06-25	ต่อ | ติด | พิเศษ | พิเศษวันนี้ | อังคาร | อังคารเวลา
2018-06-24	คำรับ | คำรับสารภาพ | รับสารภาพ
2018-06-21	ทุกวันพุธ | พฤหัสบดีเวลา
2018-06-20	ในรายการ

วันที่	คำสำคัญที่เป็นกระแส
2018-06-19	จะอยู่ | อังคารเวลา
2018-06-13	ความจริง
2018-06-12	-อังคาร | -อังคารเวลา | เรื่องเ | จันทร์-อังคาร | อังคารเวลา
2018-06-11	ทุกวันจันทร์
2018-06-04	ได้ทุก | วันจันทร์-
2018-06-02	ชมได้ | นั้น
2018-05-16	ย้อนหลังได้
2018-05-11	ได้เลย | ไหน
2018-03-27	ข่าวค่ำ | รายได้ | สู่ | ายการ
2018-03-25	ค61
2018-03-24	ค61 | ย้อนหลังได้ทาง
2018-03-23	channel | ซิงเกิ | ทัวร์ | รถทัวร์ | ลุ้น | ศพ | สุดท้าย | อร่อย
2018-03-22	/live | 0ล้าน | channel | จังหวัด | ชมย้อนหลัง
2018-03-21	รายการย้อนหลังได้ที่
2018-03-20	ข้อ | ผล
2018-03-18	/live
2018-03-14	หรือไม่
2018-03-11	กับการ
2018-03-08	พลาด | ห้าม | ห้ามพลาด
2018-02-20	ย้อนหลังได้ทาง
2018-02-16	นทาง
2018-02-10	ได้ทางw
2018-02-04	ย้อนหลังได้ | ย้อนหลังได้ที่นี่
2018-01-28	ย้อนหลังได้ | ย้อนหลังได้ที่นี่

ตารางที่ 52 ตารางแสดงคำหยุดที่หลุดมาเป็นคำสำคัญที่เป็นกระแสทั้งหมดจากวิธีเอ็นแกรมแบบตัวอักษร วิธี TF-IDF และวิธี TF โดยพิจารณายูนิแกรม และไม่พิจารณายูนิแกรม รวม 5 วิธี

วันที่	คำหลุดที่หลุดออกมา	เอ็นแกรมแบบตัวอักษร	TF-IDFมียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
2019-03-27	26	6	0	4	1	15
2019-03-26	15	5	0	4	1	6
2019-03-25	7	4	2	0	1	1
2019-03-24	5	3	1	0	1	1
2019-03-23	7	2	0	2	1	2
2019-03-22	7	3	1	3	0	0
2019-03-21	41	38	0	2	1	0
2019-03-20	38	36	0	0	1	1
2019-03-19	13	10	0	0	0	3
2019-03-18	1	1	0	0	0	0
2019-03-15	2	1	1	0	0	0
2019-03-14	1	0	0	1	0	0
2019-03-12	10	7	0	0	0	3
2019-03-11	1	0	1	0	0	0
2019-03-10	1	1	0	0	0	0
2019-03-09	2	2	0	0	0	0
2019-03-08	1	1	0	0	0	0
2019-03-03	1	1	0	0	0	1
2019-03-01	1	0	0	0	0	1
2019-02-14	1	1	0	0	0	0
2019-02-07	1	1	0	0	0	0
2019-02-04	5	0	1	0	4	0
2019-02-01	1	1	0	0	0	0
2019-01-27	1	1	0	0	0	1
2018-12-28	1	1	0	0	0	0
2018-12-27	5	4	0	1	0	1
2018-12-26	1	1	0	0	0	0
2018-12-25	1	1	0	0	0	0

วันที่	คำหลุดที่หลุดออกมา	เอ็นแกรมแบบตัวอักษร	TF-IDFมียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
2018-12-20	1	0	0	0	1	0
2018-12-19	1	1	0	0	0	0
2018-12-18	1	1	0	0	0	0
2018-12-14	2	2	0	0	0	0
2018-12-13	1	1	0	0	1	0
2018-12-07	1	1	0	0	0	0
2018-12-04	1	0	1	1	0	0
2018-09-28	3	2	0	0	0	1
2018-09-27	6	4	0	2	0	0
2018-09-26	2	1	1	0	0	0
2018-09-24	1	1	0	0	0	0
2018-09-22	3	2	1	0	0	0
2018-09-21	3	0	2	0	0	1
2018-09-11	1	1	0	0	0	0
2018-09-07	5	5	0	0	0	0
2018-09-06	1	1	0	0	0	0
2018-08-30	1	0	0	1	0	0
2018-08-29	1	1	0	0	0	0
2018-08-27	1	0	1	0	0	0
2018-08-24	1	0	1	0	0	0
2018-06-29	1	0	0	0	1	0
2018-06-28	14	2	6	2	5	0
2018-06-27	7	4	1	1	1	0
2018-06-26	26	15	2	1	5	3
2018-06-25	6	2	1	1	2	0
2018-06-24	3	0	0	3	0	0
2018-06-21	2	0	0	0	0	2
2018-06-20	1	0	0	0	0	1

วันที่	คำหลุดที่หลุดออกมา	เอ็นแกรมแบบตัวอักษร	TF-IDFมียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
2018-06-19	2	0	0	2	0	1
2018-06-13	1	0	0	0	0	1
2018-06-12	5	1	0	4	0	0
2018-06-11	1	0	0	1	0	0
2018-06-04	2	0	0	2	0	0
2018-06-02	2	0	1	0	0	1
2018-05-16	1	1	0	0	0	0
2018-05-11	2	1	0	0	2	1
2018-03-27	4	1	1	0	0	2
2018-03-25	1	0	0	0	0	1
2018-03-24	2	0	0	0	0	2
2018-03-23	8	4	4	0	0	1
2018-03-22	5	5	0	0	0	0
2018-03-21	1	1	0	0	0	0
2018-03-20	2	0	2	0	0	0
2018-03-18	1	1	0	0	0	0
2018-03-14	1	0	1	0	0	0
2018-03-11	1	0	0	1	0	0
2018-03-08	3	1	0	0	3	0
2018-02-20	1	1	0	0	0	0
2018-02-16	1	0	1	0	0	0
2018-02-10	1	1	0	0	0	0
2018-02-04	2	1	0	0	1	0
2018-01-28	2	1	0	0	1	0
รวม	346	199	34	39	34	54

ตารางที่ 53 ตารางเปรียบเทียบจำนวนคำหยุดที่หลุดจากข้อมูลอ้างอิง ทั้ง 5 วิธี

จาก*ตารางที่ 52* เรานำข้อมูลของทุกวิธีมาเปรียบเทียบกัน ดัง*ตารางที่ 53* พบว่าวิธีเอ็นแกรมแบบตัวอักษรมีจำนวนคำหยุดที่หลุดเยอะกว่าวิธีอื่นอย่างชัดเจน โดยทั้งวิธี TF-IDF และวิธี TF มีปริมาณคำหยุดที่หลุดใกล้เคียงกัน โดยพบว่าหากเป็นวิธีที่ไม่พิจารณายูนิแกรม จะมีปริมาณคำหยุดที่หลุดสูงกว่าแบบพิจารณายูนิแกรม

จากการเปรียบเทียบแสดงให้เห็นว่า แม้ว่าวิธีเอ็นแกรมแบบตัวอักษรจะให้ปริมาณคำสำคัญที่เป็นกระแสได้มากกว่าวิธีอื่น อีกทั้งยังไม่จำกัดความยาวของคำสำคัญที่สกัดได้ โดยแม้จะมีการสร้างฐานข้อมูลคำหยุดแล้ว แต่ก็ยังคงมีคำหยุดที่หลุดที่เยอะกว่าวิธีอื่น ๆ อย่างเห็นได้ชัด

โดยเมื่อเรานำหน่วยวัดต่าง ๆ มาพิจารณาประสิทธิภาพของแต่ละวิธี ดัง*ตารางที่ 54*

หน่วยวัด	เอ็นแกรม แบบตัวอักษร	TF-IDF มียูนิแกรม	TF-IDF ไม่มียูนิแกรม	TF มียูนิแกรม	TF ไม่มียูนิแกรม
ผลบวกจริง	89	17	39	19	82
ผลบวกปลอม	199	34	39	34	54
ผลลบปลอม	66	138	116	136	73
ความเที่ยง	0.309	0.333	0.500	0.358	0.603
ความไว	0.574	0.110	0.252	0.123	0.529
คะแนน F1	0.402	0.165	0.335	0.183	0.564

ตารางที่ 54 ตารางเปรียบเทียบประสิทธิภาพโดยใช้หน่วยวัดต่าง ๆ

จาก*ตารางที่ 54* พบว่าวิธี TF แบบไม่พิจารณายูนิแกรมได้คะแนน F1 มากที่สุด และยังคงมีค่าความเที่ยงมากที่สุดด้วย โดยวิธีที่ได้คะแนน F1 รองลงมาคือวิธีเอ็นแกรมแบบตัวอักษร และวิธีนี้ยังมีค่าความไวมากที่สุด โดยวิธีเอ็นแกรมแบบตัวอักษรนั้น แม้ว่าจะให้ผลบวกจริงมาก แต่เนื่องจากได้ผลปลอมมากเช่นกัน ทำให้ค่าความเที่ยงที่ได้นั้น น้อยกว่าทุกวิธี เมื่อเราพิจารณาวิธี TF-IDF และ TF จะพบว่า เมื่อเรานำยูนิแกรมเข้ามาพิจารณา จะส่งผลให้คะแนน F1 น้อยลงอย่างชัดเจน โดยวิธี TF-IDF แบบมียูนิแกรม ได้คะแนน F1 น้อยที่สุด

บทสรุปผลการวิจัย และข้อเสนอแนะ

สรุปผลการวิจัย

วิทยานิพนธ์ฉบับนี้ ได้เสนอวิธีการหาคำสำคัญที่เป็นกระแสและคำหยุดจากเพจเฟซบุ๊กภาษาไทยโดยใช้เอ็นแกรมแบบตัวอักษร ซึ่งไม่จำเป็นต้องใช้เครื่องมือในการตัดคำ เนื่องจากเครื่องมือตัดคำภาษาไทยนั้น จำเป็นต้องมีคลังข้อมูลภาษา เช่น พจนานุกรม หรือข้อมูลประโยคที่มีผู้เชี่ยวชาญแบ่งคำไว้แล้ว เป็นต้น เพื่อให้เครื่องมือ หรือแบบจำลองสามารถเรียนรู้และตัดคำได้ถูกต้อง โดยคลังข้อมูลภาษาที่มีผู้เชี่ยวชาญจัดทำไว้ล่าสุดคือ BEST2009 ซึ่งไม่มีข้อมูลภาษาที่เกี่ยวข้องกับสื่อสังคมออนไลน์ ทำให้เมื่อเราเครื่องมือที่ฝึกสอนผ่านคลังข้อมูลภาษา BEST2009 อย่าง deepcut ซึ่งเป็นเครื่องมือที่ใช้วิธีการเรียนรู้ของเครื่องและเป็นเครื่องมือที่มีประสิทธิภาพสูงที่สุดที่สามารถใช้ฟรีได้มาใช้กับประโยคบนสื่อสังคมออนไลน์จะพบว่า เครื่องมือดังกล่าวมียังคงมีข้อผิดพลาดหลายจุด เมื่อพบเจอกับคำที่ไม่เป็นมาตรฐาน และคำภาษาไทยปนภาษาอังกฤษ

โดยวิทยานิพนธ์ฉบับนี้ ได้นำวิธีที่เสนอมาเปรียบเทียบกับวิธีทั่วไป ได้แก่ วิธีการสกัดคำสำคัญโดยใช้ค่า TF-IDF โดยใช้ฐานข้อมูลคำหยุดประกอบ และวิธีการสกัดคำสำคัญโดยใช้ค่า TF โดยใช้ฐานข้อมูลคำหยุดประกอบ ซึ่งทั้ง 2 วิธีนี้จำเป็นต้องใช้เครื่องมือตัดคำ deepcut ซึ่งเมื่อเครื่องมือตัดคำเกิดความผิดพลาดขึ้น ทำให้คำที่ควรจะยาว 1 คำ อาจถูกแบ่งเป็น 2 คำ หรืออาจจะอยู่รวมกับคำอื่น ทำให้เมื่อพิจารณาโดยใช้ค่า TF หรือค่า TF-IDF จะได้ผลลัพธ์ของคำนั้นผิดพลาดไป จนอาจจะไม่ถูกจัดเป็นคำสำคัญ อีกทั้งวิธีการสกัดคำสำคัญโดยใช้ค่า TF หรือค่า TF-IDF มีข้อจำกัดในเรื่องความยาวของคำสำคัญที่สกัดออกมา ซึ่งแก้ไขได้โดยใช้วิธี
เอ็นแกรมแบบคำ ซึ่งเป็นการขยายความยาวแบบคงที่ และส่งผลต่อประสิทธิภาพและเวลาที่ใช้ในการประมวลผล วิทยานิพนธ์ชิ้นนี้จึงต้องการแก้ไขข้อจำกัดดังกล่าว

อีกทั้งวิทยานิพนธ์ฉบับนี้ ยังได้นำเสนอวิธีการสกัดคำสำคัญที่เป็นกระแสจากเพจเฟซบุ๊ก โดยการเปรียบเทียบคำสำคัญของแต่ละเพจ ตามช่วงเวลา เพื่อคัดแยกคำหยุดออกมา และสร้างเป็นฐานข้อมูลคำหยุดขึ้น เพื่อใช้ในการกรองคำหยุดออกจากคำสำคัญที่เป็นกระแส ซึ่งเราสามารถเปลี่ยนขอบเขตของกระแสได้ตามเพจเฟซบุ๊กที่เราเลือกมาวิเคราะห์อีกด้วย

จากผลลัพธ์ที่ได้จากการทดลองพบว่า วิธีที่วิทยานิพนธ์ฉบับนี้ได้นำเสนอนั้น สามารถแก้ไขข้อจำกัดด้านความยาวของคำสำคัญที่สกัดออกมาได้สำเร็จ และสามารถคัดแยกคำสำคัญที่เป็นกระแสและคำหยุดออกจากกันได้ โดยมีประสิทธิภาพดีกว่าวิธี TF-IDF และวิธี TF ที่แก้ไขข้อจำกัดด้านความยาวโดยการพิจารณา ยูนิแกรม ไบแกรม และไตรแกรมของคำ แต่เมื่อเรานำยูนิแกรมออกซึ่งเป็นการไม่พิจารณาคำสำคัญที่เป็นคำโดด จะพบว่าวิธี TF แบบพิจารณาแค่ไบแกรม และไตรแกรม ได้ผลลัพธ์ที่ดีกว่าวิธีที่วิทยานิพนธ์ฉบับนี้ได้นำเสนอ แต่ส่งผลให้เราไม่พบคำสำคัญที่เป็นกระแสที่ยาว 1 คำแทน

โดยข้อจำกัดของวิทยานิพนธ์ฉบับนี้คือ ผลลัพธ์ที่ได้ยังจำเป็นต้องใช้คนในการทำความเข้าใจกระแสที่เกิดขึ้น เนื่องจากคำสำคัญที่สกัดออกมาได้อาจมีตัวอักษร และคำข้างเคียงปะปนมาด้วย จึงยังไม่เหมาะกับระบบที่เป็นอัตโนมัติล้วน อีกทั้งประสิทธิภาพที่ได้แปรผันกับปริมาณคำหยุดที่สร้างขึ้น ซึ่งแสดงให้เห็นว่าหากเราเก็บข้อมูลโพสต์ได้เยอะ ก็สามารถนำมาสร้างฐานข้อมูลคำหยุดได้มากขึ้น

ข้อเสนอแนะ

วิธีที่วิทยานิพนธ์ฉบับนี้นำเสนอนั้น เหมาะอย่างยิ่งสำหรับงานที่ต้องการสกัดคำสำคัญโดยไม่จำกัดความยาว ซึ่งมีคนคอยทำความเข้าใจผลลัพธ์ที่ได้จากวิธีที่วิทยานิพนธ์ฉบับนี้นำเสนอ โดยเฉพาะงานด้านการหากระแสบนสื่อสังคมออนไลน์ เนื่องจากกระแสที่เกิดขึ้นบนสื่อสังคมออนไลน์นั้นมักจะถูกนำไปใช้ต่อเพื่อวิเคราะห์พฤติกรรมของผู้คนในสังคม หรือใช้สำหรับหาสาเหตุหรือเวลาที่เริ่มต้นของกระแส ซึ่งจำเป็นต้องใช้คำสำคัญปริมาณมากที่เกี่ยวข้องกับกระแสดังกล่าวในการวิเคราะห์ ซึ่งวิธีที่วิทยานิพนธ์ฉบับนี้นำเสนอนั้นให้ผลลัพธ์ที่ดี โดยได้คำที่เกี่ยวข้องกับกระแสมากที่สุดเมื่อเทียบกับวิธีอื่น ๆ โดยได้ทั้งคำที่ยาวและสั้น ซึ่งแตกต่างจากวิธีอื่น ๆ ที่ต้องกำหนดความยาวของคำสำคัญก่อน อีกทั้งยังสามารถสกัดคำที่เกี่ยวข้องกับกระแสที่แม้ว่าจะเป็นคำหยุด แต่ก็สามารถใช้ประกอบความเข้าใจของกระแสนั้นได้เป็นอย่างดี

ตัวอย่างเช่นกระแสเกี่ยวกับถ้ำหลวงนั้น วิธีนี้สามารถสกัดคำสำคัญที่เป็นกระแสได้ทั้งคำสั้น ๆ อย่าง “ถ้ำหลวง” “หมูป่า” หรือ “13ชีวิต” และคำยาว ๆ อย่าง “ค้นหา13ชีวิตติดถ้ำหลวง” หรือ “ติดถ้ำหลวงขุนน้ำนางนอง” เป็นต้น อีกทั้งยังคงมีคำอย่าง “ค้นหา” “ช่วยเหลือ” “เจ้าหน้าที่” หรือ “กำลังใจ” ที่แม้จะไม่เป็นคำสำคัญแต่สามารถช่วยให้เราเห็นภาพกระแสที่เกิดขึ้นได้ชัดเจนมากยิ่งขึ้น

สำหรับผู้ที่ต้องการวิจัยในขั้นถัดไป งานวิจัยชิ้นนี้ยังคงมีปัญหาที่เกิดขึ้นหลายอย่าง ได้แก่ ปริมาณโพสต์ของเพจเฟซบุ๊กที่ถูกจำกัด ทำให้ไม่สามารถได้ข้อมูลที่ครบถ้วนเพียงพอสำหรับวิเคราะห์กระแสได้แม่นยำ ซึ่งส่งผลเกี่ยวเนื่องถึงปริมาณคำสำคัญที่ไม่ถูกกรองเป็นคำหยุดที่มีเป็นปริมาณมาก ทำให้ค่าความเที่ยงลดลง และปัญหาคำสำคัญที่สกัดได้นั้น อาจไม่ได้เป็นคำที่สมบูรณ์ แต่เป็นคำที่ประกอบด้วยตัวอักษรอื่นซึ่งเป็นส่วนหนึ่งของคำข้างเคียง เนื่องจากวิธีการรวมเอ็นแกรมแบบตัวอักษรที่อ้างอิงข้อมูลโพสต์ต้นทางนั้น หากคำข้างเคียงมีตัวอักษรอื่นที่เป็นส่วนหนึ่งของคำสำคัญ ก็ทำให้ถูกสกัดออกมาได้ ดังเช่น “ถ้ำหลวงข” จะเห็นได้ว่าเป็นคำที่เกิดขึ้นเพราะคำสำคัญที่เป็นกระแสอย่าง “ถ้ำหลวงขุนน้ำนางนอน” ถูกสกัดออกมาได้ แต่โพสต์ต้นทางอาจเป็น “...ถ้ำหลวงขอขมาสิ่งศักดิ์สิทธิ์...” ซึ่งทำให้เกิดคำว่า “ถ้ำหลวงข” ขึ้น ซึ่งไม่ค่อยมีผลกระทบสำหรับผู้นำไปงานที่เป็นคน แต่มีผลกระทบต่อระบบที่เป็นอัตโนมัติ ซึ่งจำเป็นต้องแก้ไขปัญหานี้ต่อไป

ภาคผนวก ก
ผลงานตีพิมพ์จากวิทยานิพนธ์

International Journal of Machine Learning and Computing, Vol. 8, No. 6, December 2018

Extraction of Trend Keywords and Stop Words from Thai Facebook Pages Using Character n-Grams

Nattapong Ousirimaneechai and Sukree Sinthupinyo

Abstract—**In the era of data and information, insight of user's behavior such as trend is normally used in real-time marketing for improvement of gross profit, therefore, it is beneficial to know the trend in social media. Word tokenization and stop words list are the conventional method for keyword extraction task, however for Thai language in social media platform, there are still no efficient word tokenization tools and stop words list to extract trend from platform such as Facebook. Therefore, in this research, we propose an algorithm that require no word tokenization tools and external stop words list for the purpose of *Trend Keywords* extraction. The core idea is using Character n-Grams, instead of Word n-Grams, to tokenize, process, and combine n-Grams into keyword. After that we identified *Trend Keywords* from other keywords by using our algorithm to generate stop words list for filtering out stop words. For the evaluation of result, we use human to classify the retrieved *Trend Keywords* and compare them with *Trend Keywords* from baseline method. As a result, our algorithm can identify more keyword than baseline method. Finally, the precision of generated stop words list is 97.6%, and the precision of *Trend Keywords* is 40% with the used of 1-month generated stop words list. Furthermore, by using 2-months generated stop words list, the precision can be increased to 44% by consuming more processing time for list of stop words.**

Index Terms—**Information retrieval, keyword extraction, social media mining, stop words.**

I. Introduction

Facebook is one of the largest social media platforms. It allows users to set up *Facebook page* which is a public profile created for a specific purpose. Facebook pages can be divided into many categories, such as business, news, celebrity, brands, and organization. The amount of public information generated and provided by these pages is enormous and their contents are up-to-date; therefore, if we can extract and analyze *Trend Keywords*, which are appeared globally across multiple public pages and still appear for a duration of time, it will be possible to find global trends, events, or even behavior of the mass.

The solutions for keywords extraction in Thai language still have plenty rooms for improvement. Most of the solutions require external tools, which is word tokenization, or external database, such as training data, Thai word corpus, and stop words list. However, due to the complexity of Thai language, most of the tools are not robust enough.

The most complicacy part of Thai word tokenization is sentence segmentation. To begin with, word separation does not present between words in Thai sentence and the alphabet itself also not having their own meaning, so reliable word tokenization tool is significantly crucial. Moreover, the present official tokenization tools do have limitation to use with social media like Facebook due to occurrence of the newly adapt and slang word out of official dictionary.

For tokenization of Thai word, there is a corpus by NECTEC called BEST2010 [1], which provided data like dictionary, segmented sentences or part of speech of each word in sentences, which is the only one widely available for Thai corpus. However, this corpus is not up-to-dated.

For stop words lists, there are only small lists available. The biggest one consists of only 115 words which is not nearly enough for social media information analysis.

From reasons stated above, we decided to develop our own algorithm that require no training data or external tools. The objective of this algorithm is to extract *Trend Keywords* from Thai Facebook pages using Character n-Grams instead of word tokenization method.

To test the algorithm, we decided to collect Facebook posts from several public pages and analyzed them, using our developed algorithm, to find *Trend Keywords* during specific time periods.

II. Related Work

A. Finding Keywords

For keyword extraction, n-Grams and stop words should be stated and clarified before.

n-Grams [2] is an algorithm for slicing a message into small groups, called gram, which has constant length. n-Grams can be used to slice sentences and phrases by word or character called *Word n-Grams* and *Character n-Grams* respectively. If the message is sliced into 1-length gram called unigram. In the same way, we call bigram for 2-lengths gram and trigram for 3-lengths.

Stop words [3] are words which should be filtered out before finding other significant keywords in documents, because these stop words can appear frequently but provide no specific or important meaning. For example, common word such as "the" and "and" are considered as stop words.

Keyword Extraction [4] is described as an automatic task that can identify important terms which best represent the content of a document.

To find keywords, one of the most popular algorithm in Keyword Extraction is TF-IDF [3], TF stands for Term Frequency and IDF stands for Inverse Document Frequency. TF-IDF is TF value multiply by IDF value of the same word, so the word with high frequency and appearing in less

Manuscript received August 29, 2018; revised October 20, 2018.

The authors are with Chulalongkorn University, Thailand (e-mail: 6070188521@student.chula.ac.th, sukree.s@chula.ac.th).

doi: 10.18178/ijmlc.2018.8.6.750

documents is the keyword. On the other hand, the word with low frequency or appearing in almost documents is not the keyword.

TF-IDF with Word n-Grams can be used to find multiple-word keyword. For an example, to find keywords with 1-to-3-length, all unigram, bigram and trigram will be applied with TF-IDF, then grams with high TF-IDF value are keyword. By this method, stop words with low IDF value will be excluded from the result keyword list, but also consumed more computing power and memory. If the stop words database is provided, the calculation process will be reduced to only TF method without finding IDF value. Unfortunately, the stop word is unique and diverse across categories. [5] In some categories, ex. travel, the seems like stop word "from" and "to" will be not excluded from the keyword list according to the meaning of the location, which these words are needed, however not in others. For each specific category, stop word is not easy to come by from the accessible database, and general stop words database is not enough to categorize keyword from off-list stop words.

According from the reason above, stop word extraction task is an important task. One's solution to extract stop word is filtering by high TF or low IDF value. There are plenty room to improve such as R. T. Lo, *et al.* [6] proposed a method to build a stop words list by using Kullback-Leibler divergence on the lexicon file and select L top rank words, where L is a parameter. The result show that the new method provided higher average precision than the conventional method of using TF, term frequencies, on the corpus and perform thresholding to select stop words.

B. Finding Trend Keywords

Trend Keyword has some properties that are different from normal keyword. *Trend Keyword* should be appeared in most of all documents with high frequency like stop word, but it has a specific meaning. From this property, *Trend Keyword* must be separated apart from the extracted stop word list.

There is some similar research about this topic in Thai Language. A. Piyatumrong, *et al.* [7] want to find the keyword that can represent event in social using Twitter data. They used a word tokenization tool, LexToPlus, to tokenize Thai twitter message into a set of unigrams, bigrams with stop words removal and a set of bigrams without stop words removal, then comparing with non-required word tokenization tool algorithm that using hashtag and hashtag5, which have more than 5 characters. From 5 groups, both of only TF and TF-IDF method were applied to extract event keyword. The output result of TF and TF-IDF were similar event keyword without any statistical significant difference, so it was suggested that the IDF value could not classify event keyword from stop word. According to no different results, the only TF method were used for further analysis due to implementation, memory and processing process. At the end, across all method, hashtag5 shown the best F1 score. However, the limitation of hashtag5 was that it can only identified event keywords specified in hashtag and missed other important keywords in the content. In addition, they focused on the extracted event keywords, and founded a number of intersection between event keywords with each method is low, so it means that, the extracted event keyword cannot be relied on only one specific mentioned method.

C. Word Tokenization

Thai word tokenization tools are developed by NECTEC research organization. At first, they developed a tool named LexTo that is a dictionary-based word tokenization using longest matching, but it is not robust for words that not have in a dictionary, so the precision of this tool is very low in this case. After that, they developed a tool named TLex that is a machine learning-based word tokenization using conditional random fields and use the training data from BEST corpus that officially segmented. The precision of this tool is impressive, but some segmented words should be segmented into a smaller word, so they used dictionary-based word tokenization to help after segmented by TLex and called LearnLexTo [8]. Unfortunately, this tool still has its own limitation in some categories, ex. social media, due to intentional spelling error, usually found in social media, is generating new words and new sentences, which are not in BEST corpus, and LearnLexTo is not trained by these intentional spelling error sentences, so NECTEC decided to develop new tool from LexTo named LexToPlus [9]. With LexToPlus, user can access to add any new words to dictionary, and the method itselfs can be applied with insertion type of intention spelling error, so it will be practically used, if the up-to-date social media dictionary are provided.

D. Finding Keywords without Word Tokenization

A research of C. Haruechaiyasak, *et al.* [10] proposed an algorithm to extract keyword from categorized documents by not using word tokenization. The concept of algorithm is adding character to string and checking TF and IDF value to classify a keyword. Begin with finding TF value of all 1-character length string, then add one more character that TF value pass the threshold, the rest will be rejected. Before adding a character, each string will be check whether it is keyword or not by thresholding IDF value. The keyword string will be collected, then all of the string will be sent for the next TF value calculating. These process goes on until none of the string pass through TF value thresholding. Finally, all of the keyword list will be collected.

However, by using a data structure named suffix array to optimize the calculation process for each string contain in the documents, but the optimization is not sufficient. At the end, the process loop is terminated before completion, and getting the partially keyword, so the combination of keyword on the list is required. Moreover, it is suggested that, combining between the partial keyword under the condition that both of keyword are overlapped with the same exact TF value. Although the result is quite impress, some keywords are still lost and some in the result list are wrong.

III. Methodology

In Thai language, words are not separated by space, instead, the space is used to separate sentences. However, in social media platform, such as Facebook, the format is unrestricted, as shown in Fig. 1. It is also possible to incorporate two or more type of character, such as Thai, English, emoticon,

number, special character, and hashtag.

Fig. 1. A sample of Thai Facebook post.

A. Preprocessing of Original Posts

For each post collected from Facebook public pages, the post is split by space into smaller separated messages. Special characters, which are not Thai character, English Character or number, at the beginning and the end of each message are then removed by left-right message stripping. English characters are changed to lowercase, hyperlinks are removed, every split message are joined back, and finally, remove all space. The result is shown in Fig. 2.

อยู่ที่บ้านก็ทำบุญได้แค่ปลายนิ้วแถมยังเลือกได้อีกว่าจะบริจาคให้กับมูลนิธิไหนหรือจะอุดหนุนของที่ระลึกก็ได้อีกด้วยอีกทางเลือกของการทำบุญยุคนี้Pantip.comฟีเจอร์ใหม่ๆพร้อมใช้บนแอปแท้จากPantipเท่านั้นโหลดเลย

Fig. 2. After preprocessing of the example post.

B. Finding Local Keyword Grams of Each Pages

After preprocessing, posts from a specific page are selected. Each post is tokenized into multiple grams of 5 character then grouped by date of original post. In each group, we count the number of time that each gram appears on that specific date and increment by the number of time it appeared on the day before and the day after. We give an example of Gram counting in Fig. 3.

	Post	Grams of 5 Characters	After Increment
Previous Day:	หวัดดี	{หวัดด: 1, วัดดี: 1}	
Today:	สวัสดี	{สวัสด: 1, วัสดี: 1}	{สวัสด: 1, วัสดี: 2}
Next Day:	ถถถถถ	{ถถถถ: 2}	

Fig. 3. Gram counting

After the process described in the above paragraph, we then have the number of frequency for each gram in a specific date. We used K-Means Clustering [11] to cluster the counted grams into k cluster, where k is the optimal number of cluster determined by Elbow Method [12] using maximum point-to-line distance function [13] as shown in Fig. 4.

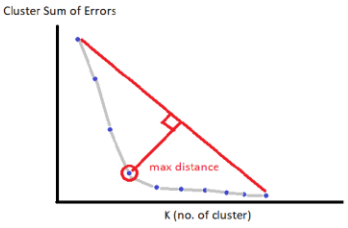


Fig. 4. Elbow Method using maximum point-to-line distance function.

Rank of each cluster is determined by mean of that cluster. The highest rank cluster is the cluster with the highest mean and the lowest rank cluster is the cluster with the lowest mean. The lowest rank cluster is discarded. The remaining clusters are kept because they have the probability to contain both keywords and stop words. We defined the remaining clusters as *Local Keyword Grams* of page.

C. Finding Global Keyword Grams

In each date, we put together the *Local Keyword Grams* from every pages and summing the frequency of the same gram. We then use again K-Means Clustering. The rank of each cluster determines by mean. We kept only the highest rank cluster and define it as *Global Keyword Grams*. The rest of the clusters are discarded as they are more likely to represent local keywords.

D. Finding Global Keywords in Each Post

To find *Global Keywords* that appear in a post, first, we need to obtain *Global Keyword Character* by evaluating every character in the post. For each character in each position, there are 5 grams which contain that character, except for the first four and the last four character of the post. Considering the set of 5 grams for a character, if there are one or more grams that match with *Global Keyword Grams*, we consider that character to be a *Global Keyword Character*. An example of *Global Keyword Character* is given in Fig. 5.

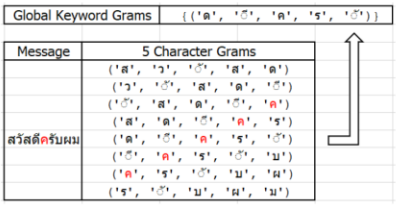


Fig. 5. Example of *Global Keyword Character*.

After that, we replace non-*Global Keyword Character* with space and split the whole post by space, so we can obtain *Global Keywords* for this post as shown in Fig. 6.

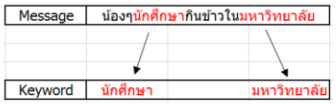


Fig. 6. Extraction of *Global Keywords* after identification of *Global Keyword Characters*.

Unfortunately, due to the nature of stop words that tend to appear in high frequency, the obtained *Global Keywords* is still containing stop words which need to be removed.

E. Building a Stop Words List from Global Keywords

After we obtain *Global Keywords* from every day of interest, it is possible to retrieve *Trend Keywords* and stop words. *Trend Keywords* appear during a specific period of days during the event, but stop words are appearing all the time. Therefore, we can set up a condition to distinguish *Trend Keywords* from stop words. Considering each *Global Keyword*, if the keyword appears during the period of 35-55

days prior to the day of interest, there's a high probability that it is a stop word. We retrieve all keywords that met with this condition and use them to build a stop words list. We can also use this process to retrieve stop words from each day and add them together.

F. Finding Trend Keywords

After we have the stop words list, we can determine if a *Global Keyword* is a *Trend Keyword*. Considering each remaining *Global Keywords* that is not stop words, if the keyword appears on 1-10 days prior to the day of interest, we can consider it as a *Trend Keyword*. The remaining keyword can be considered as a *Global Keyword* of that specific day.

IV. Parameter Discussions

A. Number of Character Grams

To choose the number of character in a gram, we must consider the number of character in Thai word. From the work of A. Piyatumrong which we discussed in related work, the average length of Thai word is 5 characters. Therefore, we decided to use 5-character n-grams in our algorithm. It is important to note that the character length of gram can be changed and is considered as a tradeoff. For example, 3 characters gram gives a high detection sensitivity of both keywords and stop words, on the contrary, 7 characters gram can only detect long words (7 characters or more) which are more likely to be keywords than stop words.

B. Range of Day Used to Weight Local Keyword Grams

For weighting of *Local Keyword Grams* in a specific day, we must consider the nature of trends and events that normally appear on social media for only a certain period. Therefore, we decided to weight each *Local Keyword Grams*, by summing the appearing time on that day, the day prior, and the day after. If we use bigger range, stop words which appear nearly all the time will gain more weight than the actual keywords.

C. Clustering Method

To cluster keywords by appearing frequency, we decided to use K-Means Clustering method which is computationally inexpensive and sufficient. To choose the optimal number if clustering (K), we used elbow method with point to line distance function to select K. With this method, we can obtain the number of clustering which is suitable for removing non-keyword grams. If the number of cluster is more than K, *Local Keyword Grams* will be further divided and this result in the vagueness between keyword grams and non-keyword grams.

D. Ranking of Global Keyword Grams

We rank each cluster by its mean. Cluster with the highest mean has the highest rank, and cluster with the lowest mean has the lowest rank. We chose keywords gram in the highest rank cluster as *Global Keyword Grams*, due to the high frequency of appearance which indicate trends and events.

E. Range of Day Used to Identify Stop Words

In our work, we chose the period of 35-55 days prior to the day of interest as a range for stop words identification. It is possible that there is a specific trend that occur on the same date of every month, so we chose the period back of more than 1-month but not reaching 2-months. The range is flexible, and it is possible that there are better range than the one we use.

F. Range of Day Used to Identify Trend Keywords

Some *Trend Keywords* appear for only a short duration of time, so we chose to consider each *Global Keywords* in the range of 1-10 days prior to the day of interest. If the *Global Keywords* appear in that range, it can be considered as *Trend Keywords*. Again, this range is flexible.

G. Number of Sample Pages

The number of chosen pages is very important because we must consider both the amount of post per page and the diversity of pages.

Concerning the amount of pages and posts per page, if we have a higher volume of sample, the keywords are easier to retrieve but the process of n-grams splitting, and counting will take a long time.

Moreover, the sampled pages must be diverse enough to make sure that the retrieved *Trend Keyword* is not keyword specific to a category of page.

In this work, we selected 10 sample pages from different categories which are news, brand, business, and organization.

V. Baseline Methodology

In social media platform such as Facebook, hashtag is a way to represent the subject of each post. Therefore, we chose hashtag as the baseline for keyword finding method

A. Finding Hashtags from Each Facebook Page

On the day of interest, hashtags of each post of the page are retrieved. These extracted hashtags are the equivalent of page *Local Keywords* of our algorithm.

B. Finding Global Keyword from Hashtags

The *Local Keywords* retrieved from hashtags can be considered as *Global Keywords* if they appeared on 2 or more pages.

C. Finding Trend Keyword from Global Keyword

We use same methodology, the *Global Keywords* that same appeared in range of 1-10 days ago are *Trend Keywords*.

VI. Result Discussions

In the experiment, we sampled 10 Facebooks pages, namely Checkbait, Brandbuffet, Investerest, Kapookdotcom, Longtunman, Pantipdotcom, Thairath, Thematterco, Themomentumco, and Underbedstar. Posts are collected from 2018/01/01 to 2018/07/07. We focus on the event and keyword trend of Thailand Cave Rescue event from 2018/06/28 to 2018/07/02.

We use human to classify the obtained result into different categories of true *Trend Keywords*, true stop words, and ambiguous words which cannot be clearly classified.

A. Stop Words

From our algorithm, we built a stop words list using posts

from 2018/06/01 to 2018/07/02. The list contains 167 words as shown in Fig. 7.

ไม่ใช่	ายที่	เกี่ยว	กับการ	งหลัง	เกาหลีเหนือ	ความเ	เรียน
ต้องการ	ความส	มากที่สุด	ีที่	มีญหา	ก่าลัง	ีเห็น	ระดับ
ในโลก	น่าจะ	อย่าง	ภายใน	มีความส	บนั้น	ไม่มี	เพราะ
รายได้	เหล่า	ประเทศ	ครั้งเพื่อ	เรียก	วันนี้	ความเป็น	พร้อม
บทความ	ติการ	ีความห	เตอร์	ประกาศ	ที่มี	แบบนี้	ท่าให้เ
นเรื่อง	เข้ม	ประจำ	ติดตาม	เพิ่มเ	เจ้าของ	บริษัท	านั้น
เพื่อ	ประเทศไทย	านแล้ว	นอยู่	ท่างาน	ท่าให้	ดูเพิ่มเติม	มได้
เป็นค	ปัจจุบัน	ลงทุน	เลือก	นเป็น	านที่	เข้าใจ	ตัวเอง
ต่างๆ	ีความเ	ศาสตร์	เกาหลี	หลายคน	สร้าง	กลุ่ม	เหนือ
เป็นเ	ไม่ได้	ยังไม่	กลายเป็น	ก็เป็น	นแต่ว	ันที่	อาจจะ
เกิดขึ้น	ันนี้	าเป็น	าเป็นเ	คนที่	จะเป็น	ความส	เรื่อง
องที่	ีต้อง	เพิ่มเพื่อ	ทุกคน	เป็นเรื่องที่	นอย่าง	เป็น	เพิ่ม
อกจาก	ตอนนี้	งเป็นค	เหล่าน	หรือ	ด้วยเ	เรื่องการ	ที่เรา
ชื่อง	นอย่างไร	จากการ	ะเป็น	ที่มา	งความ	ันการ	ยู่น่า
ดั้งแ	สามารถ	เมื่อ	ได้รับ	นความ	ลงทุนศาสตร์	เป็นเรื่อง	เรื่องนี้
เท่านั้น	ยเป็น	ที่จะ	เพิ่มเติม	เหมือน	าเป็นเรื่อง	ที่เร	อย่างไร
ชีวิต	ธุรกิจ	วันที่	ดั้งแต่	ีความ	เดียว	องการ	ีงที่
มีความ	สังคม	งการเ	นเข้า	เรื่อ	ที่สุด	มีการ	ีนั้น
สำหรับ	ว่าจะ	ให้เรา	ต้องเ	เปลี่ยน	ล่าสุด	การณ์	ออกมา
วงการ	ในการ	เหล่านี้	เสียง	ข่อมูล	หนึ่ง	เพื่อน	ที่เพ
เชื่อ	เมือง	ครั้ง	งเป็น	งความเ	พิ่มพ์	ีความส	

Fig. 7. Result of stop words list. *Trend Keywords* that were misidentified as stop words are highlighted in red.

The result shown that the list has a precision of 97.6% and left 2.4% as false discovery rate such as "ประเทศไทย" (Thailand) and "เกาหลีเหนือ" (South Korea) which should be considered as *Trend Keywords*.

In addition, we gathered more stop words to improve our algorithm precision. The stop words list was gathered from 2018/05/01 to 2018/07/02 which extend to 229 words. As shown in Fig. 8, misidentified words remain the same.

ตัวเอง	ปัจจุบัน	ทุกคน	ยเป็น	นนี้	ภายใน	วันที่	ลูกค้า
านั้น	เหตุการณ์	เพิ่ม	ประจำ	เข้าไป	ation	ันนี้	เรื่อง
ีตตาม	ที่มา	เพิ่มเ	าเป็น	จากการ	จำนวน	ข้อมูล	วิทยา
นความ	งความ	ออกมา	เพื่อ	ต้องการ	ความเป็น	ีที่	นอย่าง
เป็นเรื่องที่	เกาหลี	อกจาก	ีความเ	เลือก	เจ้าของ	เมื่อเ	ไม่ได้
สร้าง	ที่จะ	ายที่	บริษัท	กว่าเ	นเรื่อง	ในโลก	แล้วเ
จริงๆ	หนังสือลงทุน	หรือก	เหล่าน	ในการ	เพื่อน	เพิ่มเติม	ีความห
เกี่ยว	มีความส	าที่	เกิดขึ้น	ประกาศ	พวกเขา	พร้อม	เปลี่ยน
ันการ	ท่าให้	ใหม่ๆพร้อม	มากที่สุด	ก็เป็น	ที่เราได้	ตั้งแต่	นความเ
ต่างๆ	เข้ม	เรื่องการ	ดอนนี้	ีตได้	คนที่	เท่านั้น	ท่าให้เ
ก่าลัง	นนั้น	วิทยาศาสตร์	เป็นเรื่อง	ความเ	ประเทศ	ศาสตร์	ีต้อง
มีการ	ประเทศไทย	กลายเป็น	เรื่องนี้	เกี่ยว	อความ	น่าจะ	เหล่านี้
ีเห็น	นเรื่อ	ะเป็น	ในประเทศ	ไม่ใช่	ไม่มี	นเป็น	ให้เรา
านที่	ไม่ต้อง	นหนึ่งใน	หลายค	นเข้า	อย่างใร	งหลัง	อย่าง
ความน	จะเป็น	เหนือ	เสียง	เป็นเ	โซเชียล	ีงใน	งการเ
งความเ	หนังสือ	เดียวที่	เหมือน	การณ์	น่าสนใจ	...อ่านต่อ	ีความเ
ีความส	ชีวิต	อย่างที่	ที่เพ	เมื่อความเ	การลงทุน	ีงที่	เดียว
ดูเพิ่มเติม	ล่าสุด	สังคม	โครงการ	านแล้ว	มได้	นหนึ่ง	เป็นค
เชื่อ	เพิ่มเพื่อ	ระหว่าง	ด้วยก	นประเทศ	าเป็น	ที่เรา	ติดตาม
ธุรกิจ	ท่างาน	เรียน	ีปี	ที่เร	สิ่งที่	สามารถ	งเป็น
วงการ	ระดับ	องที่	รายได้	นอยู่	สำหรับ	ความร	งเป็นค
ครั้งเพื่อ	หลังจาก	มีความ	เรียก	วันนี้	ปัญหา	ด้วยเ	เข้าใจ
แบบนี้	นอย่างไร	ลงทุน	เมือง	หนึ่ง	ลงทุนศาสตร์	ต้อง	ความรู้
อาจจะ	หลายคน	องการ	ได้รับ	ติการ	ว่าจะ	ีเป็นอย่าง	าเป็นเรื่อง
เชื่อ	เพราะ	ผู้นำ	เปิดเ	ครั้ง	ตั้งแ	สร้างความเ	งที่
เพื่อง	ันนั้น	ที่มี	ีงที่	กับการ	เหตุ	ยังไม่	เรื่อ
ีเป็นเพื่อ	ความส	มนุษย	บทความ	ีเป็น	ที่สุด	ความห	เตอร์
ทุกถึง	สร้างความ	นแล้ว	กลุ่ม	เกาหลีเหนือ	เหล่า	เป็นเ	ีความ
ได้ที่	ประเทศไ	นที่	พิมพ์	ชื่อง			

Fig. 8. Result of 2-months generated stop words list.

B. Trend Keywords

Fig. 9 shows *Trend Keywords* from baseline method, Fig 10 shows *Trend Keywords* from our algorithm classified by generated stop words from 1-month data, and Fig 11 shows *Trend Keywords* from our algorithm classified by generated stop words from 2-months data. True *Trend Keywords* are highlighted in red, true stop words are highlighted in blue, and the words in black are ambiguous words.

Date	Trend Keywords			
2018-06-28	ถ้าหลวง	พาทีมหมูป่าเผชิญป่าน		
2018-06-29	ถ้าหลวง			
2018-06-30	ถ้าหลวง	ทีมหมูป่า		
2018-07-01	ถ้าหลวง			
2018-07-02	ถ้าหลวง	13ชีวิตต้องรอด	หาทีมหมูป่าคลับป่าน	ทีมหมูป่า

Fig. 9. The result of baseline algorithm.

Date	Trend Keywords						
2018-06-28	ลงประ	หลายๆ	ในถ้ำหลวง	มียนายน	ๆที่	13ชีวิตถ้ำหลวง	นที่	ต้องร
	ถ้ำหลวง	13ชีวิต	เจ้าหน้าที่	หน้าที่	หน่วย	หมูป่า	วันที่2	
2018-06-29	เข้าไป	เรื่องความ	ๆที่	นที่	ทีมหมูป่าคลับป่าน		นางนอน	หน้าที่
	ได้ที่	หมูป่า	พิเศษ	กลายเป็นค	เป็น	หนึ่งใน	โอกาส	ประเด็น
		ลงทุนศาสตร์วันนี้	ถ้ำหลวง	ทรัพย์	การไป	ประมาณ	หลางประเทศ	งเพื่อน
	ที่พ	13ชีวิตติดถ้ำหลวง	วยอาก	นส่วน	หรือๆ	อื่นๆ	วมเป็น	
	ศึกษา	มียนายน	ติดถ้ำหลวง	คลับป่าน	ันที่	นโหญ่	สำรวจ	ีแล้ว
	ได้ยย	นหนึ่ง	นล่าน	นการ	นถ้ำหลวง	เป็นย	านการ	เป็น
	ในถ้ำหลวง	ที่จพ	มากกว่า	หนึ่งสื่อ	ทีมหมูป่า	ีเป็น	จนถึง	แล้วท
	เจ้าหน้าที่	ราคาก	เป็นส	นตอนนี้	ชีวิตต	พวกเขา	ระหว่าง	13ชีวิต
2018-06-30	หมูป่า	เรื่องหนึ่ง						
2018-07-01	ๆที่	ูป่วิว	ชีวิตอย่าง	เป็นย	บย่าง	ทีมหมูป่าคลับป่าน		ภารกิจ
	หน้าที่	หมูป่า	เยียงราย	ถ้ำหลวง	13ชีวิตต่อง	เรื่องการณ์	แต่ก	คลับป่าน
	ต่อง	ด้วยก	นการ	ในถ้ำ	นถ้ำหลวง	13ชีวิตต	านการ	ในถ้ำหลวง
		หนึ่งสื่อ	ทีมหมูป่า	ต้นหา	ชีวิตต	ทีมหมูป่าคลับป่านวันนี้	13ชีวิต	
2018-07-02	ูป่วิว	ีงที่	หน้าที่	หมูป่า	ข่าวดีถ้ำหลวง	ปลอดภัย	เยียงราย	ถ้ำหลวง
	นแล้วท	ต่อง	ในถ้ำ	นถ้ำหลวง	ในถ้ำหลวง	ทั้ง13ชีวิตในถ้ำหลวง		พยาบาล
	ทีมหมูป่า	ข่าวดี	เจ้าหน้าที่	เตรียม	13ชีวิตถ้ำหลวงเยียงราย	13ชีวิต		

Fig. 10. The result of our algorithm classified by 1-month generated stop words list.

Date	Trend Keywords						
2018-06-28	ลงประ	หลายๆ	ในถ้ำหลวง	มียนายน	ๆที่	13ชีวิตถ้ำหลวง	นที่	ต้องร
	ถ้ำหลวง	13ชีวิต	เจ้าหน้าที่	หน้าที่	หน่วย	หมูป่า	วันที่2	
2018-06-29		เรื่องความ	ๆที่		ทีมหมูป่าคลับป่าน		นางนอน	หน้าที่
		หมูป่า	พิเศษ	กลายเป็นค		หนึ่งใน	โอกาส	ประเด็น
		ลงทุนศาสตร์วันนี้	ถ้ำหลวง	ทรัพย์	การไป	ประมาณ	หลางประเทศ	งเพื่อน
	ที่พ	13ชีวิตติดถ้ำหลวง	วยอาก	นส่วน			อื่นๆ	วมเป็น
	ศึกษา	มียนายน	ติดถ้ำหลวง	คลับป่าน	ันที่	นโหญ่	สำรวจ	ีแล้ว
	ได้ยย		นล่าน	นการ	นถ้ำหลวง	เป็นย	านการ	เป็น
	ในถ้ำหลวง	ที่จพ	มากกว่า		ทีมหมูป่า	ีเป็น	จนถึง	แล้วท
	เจ้าหน้าที่	ราคาก	เป็นส	นตอนนี้	ชีวิตต			13ชีวิต
2018-06-30	หมูป่า	เรื่องหนึ่ง						
2018-07-01	ๆที่	ูป่วิว	ชีวิตอย่าง	เป็นย	บย่าง	ทีมหมูป่าคลับป่าน		ภารกิจ
	หน้าที่	หมูป่า	เยียงราย	ถ้ำหลวง	13ชีวิตต่อง	เรื่องการณ์	แต่ก	คลับป่าน
	ต่อง		นการ	ในถ้ำ	นถ้ำหลวง	13ชีวิตต	านการ	ในถ้ำหลวง
		ทีมหมูป่า	ต้นหา	ชีวิตต	ทีมหมูป่าคลับป่านวันนี้		13ชีวิต	
2018-07-02	ูป่วิว	ีงที่	หน้าที่	หมูป่า	ข่าวดีถ้ำหลวง	ปลอดภัย	เยียงราย	ถ้ำหลวง
	นแล้วท	ต่อง	ในถ้ำ	นถ้ำหลวง	ในถ้ำหลวง	ทั้ง13ชีวิตในถ้ำหลวง		พยาบาล
	ทีมหมูป่า	ข่าวดี	เจ้าหน้าที่	เตรียม	13ชีวิตถ้ำหลวงเยียงราย	13ชีวิต		

Fig. 11. The result of our algorithm classified by 2-months generated stop words list.

The precision of baseline is 100%, which mean that all keywords extracted using baseline method are *Trend Keywords*. However, a lot of keywords are missed and cannot be found by this method.

Our algorithm can extract a larger list of keywords, but it still contains stop words and some of the keywords can also be considered as just *Local Keywords*. As the result, our algorithm gave a precision of 40% with 1-month data. Consecutively, the precision is shown up to 44% with 2-months data, and the result can be more accurate depends on size of stop words list.

VII. Limitation and Future Work

A. Cleaning Noise from Character n-Grams

Algorithm suggested in this work, sometimes, produce keyword character with noise. The example of noise is given in Fig. 12. Some keywords obtained are not a complete word or contain extra character.

Message	Keyword with noise
ผมช่อบบอล เตะชอบบอล เขาปลอบบอล	อบบอล

Word Tokenized Message			Keyword
ผม	ช่อบ	บอล	
เตะ	ชอบ	บอล	บอล
เขา	ปลอบ	บอล	

Fig. 12. Example of noise from Character n-Grams.

The issue may be fixed with spell-correction algorithm of Thai language.

B. Getting Keyword Grams

In our work, we used K-Means Clustering as the main method to identity *Keyword Grams*, however there is still room for improvement.

The difficult point is that the frequency of keyword appearing is vary in each page, so the conventional thresholding method to determine keyword is not suitable. New solution is needed.

C. Validation Method

The problem of validation method in this work is the insufficiency of robustness, since we use human validation to classify the output of the algorithm because we cannot find a standard testing set to evaluate our algorithm.

VIII. Conclusion

In this work, we propose an algorithm which use Character n-Grams to extract keywords and stop words without using word tokenization or other training data set and corpus. However, the limitation of the proposed algorithm is the noise which is mistakenly identified as keyword. Moreover, the precision of the result depends heavily on generated stop words list, which mean if we can improve the stop words list, the precision of the algorithm can also be improved.

Acknowledgment

Nattapong Ousirimaneechai, author, would like to thank Tanawat Chansophonkul, Kidthiphutn Tejakumput for reviewing and editing the content of this paper, and Napat Simsomboonphol for support.

This research is supported by the 90th Anniversary of Chulalongkorn University, Rachadapisek Sompote Fund.

References

[1] NECTEC. (December 2009). BEST2010. *NECTEC*. [Online] Available: http://thailang.nectec.or.th/downloadcenter/

[2] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," *Ann Arbor Mi*, vol. 48113, no. 2, pp. 161-175, 1994.

[3] A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, Stanford University, California, Cambridge University Press, 2011, ch. 1, pp. 1–17.

[4] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An Overview of Graph-Based Keyword Extraction Methods and Approaches," *Journal of Information and Organizational Sciences*, vol. 39, no. 1, pp. 1–20, 2015.

[5] P. Daowadung and Y. H. Chen, "Stop word in readability assessment of Thai text," in *Proc. 2012 IEEE 12th International Conference on Advanced Learning Technologies*, July 2012, pp. 497-499.

[6] R. T. W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, pp. 17-24, 2005.

[7] A. Piyatumrong, C. Sangkeettrakarn, C. Haruechaiyasak, and A. Kongthon, "Finding Key Terms Representing Events from Thai Twitter," in *Proc. International Symposium on Natural Language Processing*, February 2016, pp. 73-87.

[8] C. Haruechaiyasak, S. Kongyoung, and C. Damrongrat, "LearnLexTo: a machine-learning based word segmentation for indexing Thai texts," in *Proc. the 2nd ACM workshop on Improving non English Web Searching*, 2008, pp. 85-88.

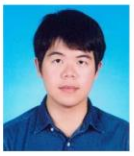
[9] C. Haruechaiyasak, and A. Kongthon, "LexToPlus: a Thai lexeme tokenization and normalization tool," in *Proc. the 4th Workshop on South and Southeast Asian Natural Language Processing*, 2013, pp. 9–16.

[10] C. Haruechaiyasak, P. Srichaivattana, S. Kongyoung, and C. Damrongrat, *Automatic Thai Keyword Extraction from Categorized Text Corpus*, 2008.

[11] J. A. Hartigan, and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[12] T. M. Kodinariya, and P. R. Makwana, "Review on determining number of cluster in K-Means clustering," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, pp. 90-95, 2013.

[13] A. Perera. (October 2017). Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach. *Linkedin*. [Online]. Available: https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera/



Nattapong Ousirimaneechai was born in 1995 in Bangkok, Thailand. He received his bachelor's degree in computer engineering from Chulalongkorn University in 2017. Now he is a master student in Chulalongkorn University. His research interests are information retrieval, natural language processing, machine learning, social network analysis and social network mining

Sukree Sinthupinyo was born in 1975 in Bangkok, Thailand. He received his bachelor's, master's, and doctoral degree from the Department of Computer Engineering, Chulalongkorn University. Now he is working as a lecturer at the same department. His research interests are artificial intelligence, machine learning, innovation, social network analysis and social network mining.

บรรณานุกรม

1. Clark, E. and K. Araki, *Text normalization in social media: progress, problems and applications for a pre-processing system of casual English.* Procedia-Social and Behavioral Sciences, 2011. **27**: p. 2-11.
2. Sornlertlamvanich, V., *Word segmentation for Thai in machine translation system*. 1993.
3. Khoo, C.S., Y. Dai, and T. Ee Loh, *Using statistical and contextual information to identify two-and three-character words in Chinese text.* Journal of the American Society for Information Science and Technology, 2002. **53**(5): p. 365-377.
4. Kudo, T., K. Yamamoto, and Y. Matsumoto. *Applying conditional random fields to Japanese morphological analysis*. in *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004.
5. Peng, F., F. Feng, and A. McCallum. *Chinese segmentation and new word detection using conditional random fields*. in *Proceedings of the 20th international conference on Computational Linguistics*. 2004. Association for Computational Linguistics.
6. Haruechaiyasak, C., S. Kongyoung, and M. Dailey. *A comparative study on thai word segmentation approaches*. in *2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*. 2008. IEEE.
7. Poowarawan, Y. *Dictionary-based thai syllable separation*. in *Proc. Ninth Electronics Engineering Conference (EECON-86), Thailand*. 1986.
8. Haruechaiyasak, C., et al. *A collaborative framework for collecting thai unknown words from the web*. in *Proceedings of the COLING/ACL on Main conference poster sessions*. 2006. Association for Computational Linguistics.
9. Sornlertlamvanich, V., *Word segmentation for Thai in machine translation system.* Machine Translation, NECTEC, 1993: p. 556-561.
10. Brent, M.R., *An efficient, probabilistically sound algorithm for segmentation and*

word discovery. Machine Learning, 1999. **34**(1-3): p. 71-105.

11. McCallum, A. and K. Nigam. *A comparison of event models for naive bayes text classification*. in *AAAI-98 workshop on learning for text categorization*. 1998. Citeseer.
12. Beeferman, D., A. Berger, and J. Lafferty, *Statistical models for text segmentation.* Machine learning, 1999. **34**(1-3): p. 177-210.
13. Joachims, T. *Text categorization with support vector machines: Learning with many relevant features*. in *European conference on machine learning*. 1998. Springer.
14. Lafferty, J., A. McCallum, and F.C. Pereira, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* 2001.
15. Zhang, X., J. Zhao, and Y. LeCun. *Character-level convolutional networks for text classification*. in *Advances in neural information processing systems*. 2015.
16. Dos Santos, C. and M. Gatti. *Deep convolutional neural networks for sentiment analysis of short texts.* in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014.
17. Beliga, S., A. Meštrović, and S. Martinčić-Ipšić, *An overview of graph-based keyword extraction methods and approaches.* Journal of information and organizational sciences, 2015. **39**(1): p. 1-20.
18. Cavnar, W.B. and J.M. Trenkle, *N-gram-based text categorization.* Ann arbor mi, 1994. **48113**(2): p. 161-175.
19. Leskovec, J., A. Rajaraman, and J.D. Ullman, *Mining of massive datasets*. 2014: Cambridge university press.
20. Alpaydin, E., *Introduction to machine learning*. 2009: MIT press.
21. Yuan, G.-X., C.-H. Ho, and C.-J. Lin, *Recent advances of large-scale linear classification.* Proceedings of the IEEE, 2012. **100**(9): p. 2584-2603.
22. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.
23. Altman, N.S., *An introduction to kernel and nearest-neighbor nonparametric regression.* The American Statistician, 1992. **46**(3): p. 175-185.
24. Zhou, Z.-H., *Ensemble methods: foundations and algorithms*. 2012: Chapman

and Hall/CRC.

25. Rokach, L. and O.Z. Maimon, *Data mining with decision trees: theory and applications*. Vol. 69. 2008: World scientific.
26. Ho, T.K. *Random decision forests*. in *Proceedings of 3rd international conference on document analysis and recognition*. 1995. IEEE.
27. Friedman, J., T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Vol. 1. 2001: Springer series in statistics New York.
28. Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural networks, 2015. **61**: p. 85-117.
29. Hinton, G.E., T.J. Sejnowski, and T.A. Poggio, *Unsupervised learning: foundations of neural computation*. 1999: MIT press.
30. Estivill-Castro, V., *Why so many clustering algorithms: a position paper.* SIGKDD explorations, 2002. **4**(1): p. 65-75.
31. Rokach, L. and O. Maimon, *Clustering methods*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 321-352.
32. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. **28**(1): p. 100-108.
33. Kodinariya, T.M. and P.R. Makwana, *Review on determining number of Cluster in K-Means Clustering.* International Journal, 2013. **1**(6): p. 90-95.
34. Kohonen, T., *Self-organized formation of topologically correct feature maps.* Biological cybernetics, 1982. **43**(1): p. 59-69.
35. Grossberg, S., *Competitive learning: From interactive activation to adaptive resonance.* Cognitive science, 1987. **11**(1): p. 23-63.
36. Haruechaiyasak, C., S. Kongyoung, and C. Damrongrat. *LearnLexTo: a machine-learning based word segmentation for indexing Thai texts*. in *Proceedings of the 2nd ACM workshop on Improving non english web searching*. 2008. ACM.
37. Haruechaiyasak, C. and A. Kongthon. *LexToPlus: A thai lexeme tokenization and normalization tool*. in *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing*. 2013.
38. Lapjaturapit, T., K. Viriyayudhakom, and T. Theeramunkong. *Multi-Candidate*

Word Segmentation using Bi-directional LSTM Neural Networks. in *2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES)*. 2018. IEEE.

39. Kosawat, K., et al. *BEST 2009: Thai word segmentation software contest*. in *2009 Eighth International Symposium on Natural Language Processing*. 2009. IEEE.
40. Kim, Y., *Convolutional neural networks for sentence classification.* arXiv preprint arXiv:1408.5882, 2014.
41. Kalchbrenner, N., E. Grefenstette, and P. Blunsom, *A convolutional neural network for modelling sentences.* arXiv preprint arXiv:1404.2188, 2014.
42. Koomsubha, T. and P. Vateekul. *A character-level convolutional neural network with dynamic input length for Thai text categorization*. in *2017 9th International Conference on Knowledge and Smart Technology (KST)*. 2017. IEEE.
43. Daowadung, P. and Y.-H. Chen. *Stop Word in Readability Assessment of Thai Text*. in *Advanced Learning Technologies (ICALT), 2012 IEEE 12th International Conference on*. 2012. IEEE.
44. Lo, R.T.-W., B. He, and I. Ounis. *Automatically building a stopword list for an information retrieval system*. in *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*. 2005.
45. Netisopakul, P. and G. Wohlgenannt, *A Survey of Thai Knowledge Extraction for the Semantic Web Research and Tools.* IEICE TRANSACTIONS on Information and Systems, 2018. **101**(4): p. 986-1002.
46. Haruechaiyasak, C., et al., *Automatic thai keyword extraction from categorized text corpus.* 2008.
47. Piyatumrong, A., et al. *Finding Key Terms Representing Events from Thai Twitter*. in *International Symposium on Natural Language Processing*. 2016. Springer.

ประวัติผู้เขียน

ชื่อ-สกุล	นายณัษฐพงษ์ อู่สิริมณีชัย
วัน เดือน ปี เกิด	23 พฤษภาคม 2538
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ปัจจุบัน	73/126 หมู่ 1 ถนนจอมทอง แขวงจอมทอง เขตจอมทอง กรุงเทพมหานคร 10150
ผลงานตีพิมพ์	Extraction of Trend Keywords and Stop Words from Thai Facebook Pages using Character n-Grams