การวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์

ANALYSIS AND PREDICTION OF TEMPORAL TWITTER POPULARITY

รัฐสิทธิ์ เสริมสัย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2562

การวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์

รัฐสิทธิ์ เสริมสัย

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2562

ANALYSIS AND PREDICTION OF TEMPORAL TWITTER POPULARITY

RATTASIT SERMSAI

A Thesis Submitted in partial Fulfillment of Requirements

for MASTER OF SCIENCE (Information Technology)

Faculty of Science Srinakharinwirot University

2019

ปริญญานิพนธ์

เรื่อง

การวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์

ของ

รัฐสิทธิ์ เสริมสัย

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

---------- คณบดีบัณฑิตวิทยาลัย

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณะกรรมการสอบปากเปล่าปริญญานิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... ประธาน

(ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(อาจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วราภรณ์ วิยานนท์)

ชื่อเรื่อง	การวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์
ผู้วิจัย	รัฐสิทธิ์ เสริมสัย
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2562
อาจารย์ที่ปรึกษา	อาจารย์ ดร. ศิริสรรพ เหล่าหะเกียรติ

ทวิตเตอร์เป็นหนึ่งในโซเชียลเน็ตเวิร์คที่ได้รับความนิยมสูง มีผู้ใช้งานทวิตเตอร์ทั่วโลกเป็นจำนวนมาก ทำให้ทวิตเตอร์มีข้อมูลหมุนเวียนเป็นจำนวนมาก และทวิตเตอร์ยังเปิดช่องทางให้สามารถเข้าถึงข้อมูลกิจกรรมที่เกิดขึ้นบนทวิตเตอร์ได้ ทำให้สามารถเก็บข้อมูลจากทวิตเตอร์เพื่อนำมาวิเคราะห์และทำนายความนิยมของการโพสต์ข้อความบนทวิตเตอร์ ที่ได้รับความนิยมจากผู้ติดตาม (follower) ซึ่งในงานวิจัยนี้จะวัดความนิยมจากจำนวนการแสดงความคิดเห็น (reply) และจำนวนการโพสต์ซ้ำ (retweet) งานวิจัยนี้จะนำเสนอวิธีการทำทายผลความนิยมโพสต์โดยใช้เทคนิค dynamic time warping ในการหาระยะทางระหว่างชุดข้อมูล เพื่อทำการจัดกลุ่มข้อมูลและหาค่าเฉลี่ยของแต่ละกลุ่ม ซึ่งจะนำไปใช้ในการทำนายความนิยมของโพสต์ต่อไป ในการวิจัยนี้ใช้ข้อมูลจากทวิตเตอร์ของสำนักข่าวที่โพสต์ข้อความเป็นภาษาอังกฤษ จำนวน 4 สำนักข่าว ได้แก่ CNN, The New York Times, Fox News และ The Washington Post

คำสำคัญ : ทวิตเตอร์, อนุกรมเวลา, การเรียนรู้ของเครื่อง

Title	ANALYSIS AND PREDICTION OF TEMPORAL TWITTER POPULARITY
Author	RATTASIT SERMSAI
Degree	MASTER OF SCIENCE
Academic Year	2019
Thesis Advisor	Dr. Sirisup Laohakiat

Twitter is one of the most popular social networks with millions of users every day around the world. The analysis and prediction of the popularity of Twitter posts have been one of most widely studied topics be cause they allow us to uncover the patterns and trends of collective interest of Twitter users towards each post. In this study, we present a novel method of analyzing and predicting a temporal profile of Twitter popularity, including both retweets and replies, using dynamic time warping (DTW) by identifying similarities between the temporal profiles and their popularity. Then, similar temporal profiles were grouped together using sequential clustering and the centroids of each cluster were determined by calculating the Barycenter of each cluster. These centroids are used as popularity profile templates for the popularity prediction of a new post. The proposed method tested real Twitter posts obtained from international Twitter news channels. The experimental results showed that the proposed method outperformed the existing method which was based on a exponential model.

Keyword : Twitter, Temporal analysis, Dynamic time warping, DTW barycenter averaging, Sequential clustering

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ ด้วยความกรุณาช่วยเหลือแนะนำอย่างดียิ่งจาก ดร. ศิริสรรพ เหล่าหะเกียรติ อาจารย์ที่ปรึกษางานวิจัย ที่ได้ให้คำปรึกษาแนะนำ และให้ข้อคิดเห็นต่างๆ ในการวิจัยตลอดจนแก้ไขข้อบกพร่องต่างๆ มาโดยตลอดผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูง ขอกราบขอบพระคุณอาจารย์และกรรมการบริหารหลักสูตรสาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้กรุณาให้ความรู้ในด้านต่างๆ แก่ผู้วิจัย จึงเป็นที่มาของปริญญานิพนธ์ฉบับนี้

คุณค่าและประโยชน์อันพึงมีจากการศึกษาวิจัยนี้ ผู้วิจัยขอน้อมบูชาพระคุณบิดามารดาและบูรพาจารย์ทุกท่านที่ได้อบรมสั่งสอนวิชาความรู้ และให้ความเมตตาแก่ผู้วิจัยมาโดยตลอดเป็นกำลังใจสำคัญที่ทำให้การศึกษาวิจัยฉบับนี้สำเร็จลุล่วงได้ด้วยดี

รัฐสิทธิ์ เสริมสัย

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

ปัจจุบันกระแสการใช้งานโซเชียลเน็ตเวิร์คในสังคมมีอัตราการเติบโตที่สูงขึ้นในทุกๆปี ไม่ว่าจะเป็นเฟซบุ๊ค ทวิตเตอร์ ยูทิวบ์ รวมทั้งการใช้งานโทรศัพท์สมาร์ตโฟนนั้น ก็มีผลกระทบโดยตรงต่อการใช้งานที่มากขึ้นเนื่องจากในปัจจุบันพฤติกรรมของเด็กจนถึงผู้ใหญ่นั้น มักจะมีโทรศัพท์สมาร์ตโฟนที่พกพาในชีวิตประจำวัน รวมทั้งหลากหลายองค์กรไม่ว่าจะขนาดใหญ่ ขนาดกลาง และขนาดเล็ก ก็เริ่มหันมาใช้โซเชียลเน็ตเวิร์คเพื่อสร้างภาพลักษณ์ขององค์กร รวมทั้งยังนำมาแบ่งปันกิจกรรมขององค์กร หรือหน่วยงานผ่านโซเชียลเน็ตเวิร์คอย่างแพร่หลาย โดยที่ในขณะนี้เริ่มมีการปรับตัวมากยิ่งขึ้น และมีแนวโน้มว่าจะเป็นมีอัตราการเติบโตที่ขยายเพิ่มขึ้นอย่างต่อเนื่อง รวมทั้งการใช้งานดังกล่าวนี้ นอกจากจะลดต้นทุนในการโฆษณาแล้วยังเป็นการช่วยสร้างการรับรู้ให้แก่สาธารณชนได้ในวงกว้าง

ทวิตเตอร์ก็เป็นหนึ่งในโซเชียลเน็ตเวิร์คที่มีผู้ใช้งานจำนวนมาก เป็นช่องทางการสื่อสารระหว่างผู้โพสต์ข้อความและผู้ติดตามที่มีความสะดวกรวดเร็ว โดยผู้ใช้สามารถโพสต์ข้อความยาวไม่เกิน 280 ตัวอักษร ว่าตนเองกำลังทำอะไรอยู่ หรือแจ้งข่าวสารต่างๆ โดยเรียกการโพสต์ข้อความนี้ว่าการ ทวีต (tweet) ซึ่งแปลว่าเสียงนกร้อง ทำให้มีการโพสต์ข้อความบนทวิตเตอร์เป็นจำนวนมากเพื่อจะแจ้งให้ผู้ติดตามรับรู้ว่ากำลังทำอะไรอยู่ หรือแจ้งข่าวต่างๆ ซึ่งในการโพสต์แต่ละครั้ง ก็จะมีจำนวนผู้ติดตามให้ความสนใจโพสต์นั้นๆ แต่ละโพสต์แตกต่างกัน โดยวัดจากการที่ผู้ติดตามมาตอบกลับ (reply) และโพสต์ข้อความซ้ำ (retweet) ผู้วิจัยเล็งเห็นว่าในการโพสต์ข้อความบนทวิตเตอร์แต่ละครั้ง สามารถนำมาศึกษาวิเคราะห์และทดลองทำการทำนายความสนใจของผู้ติดตามทวิตเตอร์ได้ และประเด็นการทำนายความนิยมของทวิตเตอร์ ก็เป็นประเด็นที่ได้รับความสนใจจากนักวิจัยจำนวนมาก อีกทั้งทวิตเตอร์ยังเปิดช่องทางให้สามารถเก็บข้อมูลกิจกรรมต่างๆ บนทวิตเตอร์ได้อีกด้วย ทำให้สามารถเก็บข้อมูลเพื่อมาทำงานวิจัยได้

จากความเป็นมาและความสำคัญของปัญหาข้างต้น จะเห็นได้ถึงความนิยมในการใช้ทวิตเตอร์จากผู้คนทั่วโลก ผู้วิจัยจึงได้เลือก ทวิตเตอร์ของสำนักข่าว ในการศึกษาวิเคราะห์และทดลองทำนายความสนใจของผู้ติดตาม โดยใช้เทคนิค dynamic time warping ในการทำนาย เพื่อนำเสนอเทคนิคใหม่ที่สามารถนำไปประยุกต์ใช้กับการทำนายในรูปแบบอื่นๆ ต่อไป

1.2 วัตถุประสงค์ของโครงการวิจัย

เพื่อทำการศึกษาวิเคราะห์และทดลองทำนายความนิยมของผู้ติดตาม ที่มีต่อการโพสต์ข้อความบนทวิตเตอร์ โดยใช้เทคนิค dynamic time warping

1.3 ขอบเขตของโครงการวิจัย

1. ใช้ข้อมูลในการวิเคราะห์และทำนายความนิยมของโพสต์ จากทวิตเตอร์ของสำนักข่าวจำนวนสี่สำนักข่าวได้แก่ CNN(@CNN), The New York Times (@nytimes), Fox News (@FoxNews) และ Washington Post (@washingtonpost)

2. ข้อมูลที่ใช้ในงานวิจัยคือ ข้อมูลที่เก็บจากทวิตเตอร์ของสำนักข่าวทั้งสี่สำนัก เป็นเวลาหนึ่งสัปดาห์ ใช้เทคนิค dynamic time warping และ dynamic time warping barycenter averaging ในการวิเคราะห์ข้อมูล และใช้เทคนิค sequential clustering ในการจัดกลุ่มข้อมูล

3. ใช้ค่า root mean square error ค่า correlation coefficient และค่า r2 scoreในการวัดความแม่นยำในการทำนายผลความนิยม

4. ใช้เทคนิคการทำนายแบบ exponential เป็นตัวเปรียบเทียบประสิทธิภาพกับเทคนิคการทำนายแบบใช้ dynamic time warping

1.4 วิธีการดำเนินการวิจัย

1. การทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2. จัดเตรียมเครื่องคอมพิวเตอร์และโปรแกรมสำหรับใช้ในงานวิจัย

3. เก็บข้อมูลจากทวิตเตอร์ของสำนักข่าวทั้งสี่สำนัก เป็นเวลาหนึ่งสัปดาห์

4. เปลี่ยนแปลงรูปแบบข้อมูลให้อยู่ในรูปแบบที่เหมาะสม เพื่อทำไปใช้ในการวิเคราะห์และทำนายความนิยม

5. ทำการสร้างเทมเพลตของข้อมูลอนุกรมเวลาของแต่ละสำนักข่าวโดยใช้เทคนิค dynamic time warping และเทคนิค sequential clustering

6. ทำการศึกษาวิเคราะห์และทดลองทำนายผลความนิยมของโพสต์บนทวิตเตอร์โดยใช้การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง

7. ทำการศึกษาวิเคราะห์และทดลองทำนายผลความนิยมของโพสต์บนทวิตเตอร์โดยใช้การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax

8. ทำการศึกษาวิเคราะห์และทดลองทำนายผลความสนใจของโพสต์บนทวิตเตอร์โดยใช้แบบจำลอง exponential

9. นำผลลัพธ์ที่ได้มาวิเคราะห์และเปรียบเทียบประสิทธิภาพในการทำนายผล

10. ประเมินผลและสรุปงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

ประโยชน์จากงานวิจัยการวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์โดยใช้ dynamic time warping

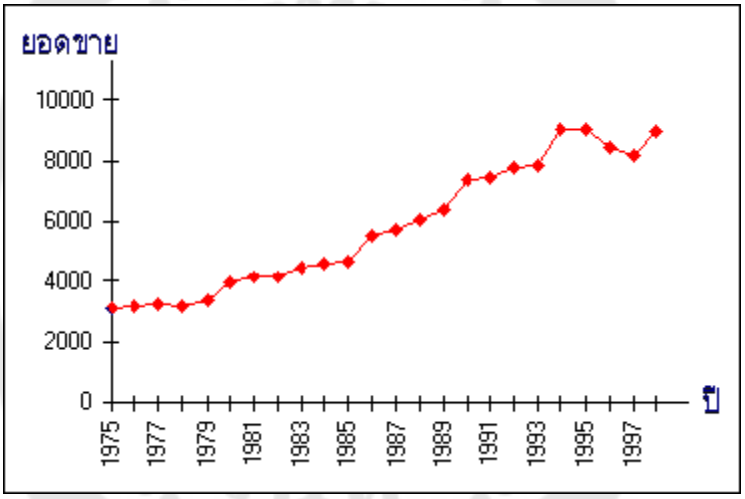
1. ได้รับความรู้เกี่ยวเทคนิคทำนายความนิยมเชิงเวลา โดยใช้ dynamic time warping

2. ได้รับความรู้ในการทำนายความนิยมของผู้ติดตาม ต่อข้อความบนทวิตเตอร์ ซึ่งสามารถนำไปต่อยอดงานวิจัย หรือไปประยุกต์ใช้ในการทำนายผลข้อมูลที่มีลักษณะข้อมูลเป็นแบบอนุกรมเวลาได้ต่อไป

บทที่ 2
แนวคิด ทฤษฎี เทคโนโลยี และงานวิจัยที่เกี่ยวข้อง

2.1 อนุกรมเวลา (time series data)

อนุกรมเวลา คือ ชุดของข้อมูลที่เก็บรวบรวมตามระยะเวลาเป็นช่วงๆ อย่างต่อเนื่องกัน เช่น ข้อมูลยอดขายสินค้าที่เก็บรวบรวมต่อเนื่องกันไปเป็นระยะเวลาหลายๆ เดือน ข้อมูลรายได้ประชาชาติปีต่างๆ ที่เก็บรวบรวมต่อเนื่องกันไปเป็นระยะเวลาหลายๆ ปี เป็นต้น ข้อมูลอนุกรมเวลาอาจอยู่ในลักษณะที่เป็นข้อมูลรายปี รายไตรมาส หรือรายเดือนก็ได้ ทั้งนี้ขึ้นอยู่กับความเหมาะสมในการนำไปใช้ประโยชน์



ภาพประกอบ 1 ตัวอย่างข้อมูลอนุกรมเวลา

2.2 Dynamic time warping [1]

dynamic time warping เป็นอัลกอริทึมที่ใช้เปรียบเทียบความคล้ายกันระหว่างอนุกรมเวลา 2 ชุด โดยทั่วไป dynamic time warping เป็นวิธีที่ทำให้คอมพิวเตอร์สามารถหาการจับคู่ที่เหมาะสมของลำดับสองชุดได้ แบบไม่ได้นำค่าที่เวลาเดียวกันมาเปรียบเทียบกันตรงๆ แบบ euclidean distance แต่จะทำการบิดเบือนหาจุดเวลาที่เหมาะสมในการเปรียบเทียบแทน โดยผลลัพธ์ที่ได้ จะให้เป็นค่าระยะทางที่แตกต่างกันระหว่าง 2 อนุกรมเวลา ตัวอย่างเช่น ให้ $A = (a_1,...,a_t)$ และ $B = (b_1,...,b_t)$ เป็น อนุกรรมเวลา ให้ δ เป็นระยะทางระหว่างอนุกรมเวลา ถ้าใช้ euclidean distance ในการเปรียบเทียบอนุกรม A และ B จะได้ดัง (1)

$$D(A,B) = \sqrt{\delta(a_1 b_1)^2 + \cdots + \delta(a_t b_t)^2} \quad (1)$$

จาก (1) จะเห็นว่า euclidean distance เป็นการเปรียบเทียบอนุกรมเวลาแบบตรงไปตรงมา ไม่ยืดหยุ่น ซึ่งในความเป็นจริง อนุกรมเวลาที่นำมาเปรียบเทียบกันเกือบทั้งหมดจะมีจุดเวลาที่ไม่เท่ากัน ซึ่งถ้าใช้ euclidean distance ในการเปรียบเทียบจะดูไม่สมเหตุสมผล ตัวอย่างเช่น อนุกรมเวลา X = (a,b,a,a) และ Y = (a,a,b,a) ถ้าใช้ euclidean distance ในการเปรียบเทียบ อนุกรมเวลา X,Y จะดูไม่สมเหตุสมผลเป็นอย่างมาก แต่ถ้าเปลี่ยนไปใช้ dynamic time warping ในการเปรียบเทียบ จะดูสมเหตุสมผลมากขึ้น เนื่องจาก dynamic time warping จะไม่ทำการเปรียบเทียบอนุกรมเวลาทั้งสอง ณ จุดเวลาเดียวกันตรงๆ แต่จะบิดเบือนจุดเวลาให้เหมาะสม ณ จุดนั้นๆ แทน

วิธีการคำนวณ dynamic time warping ให้ A,B เป็นอนุกรมเวลาที่ต้องการเปรียบเทียบ เริ่มจากการสร้างเมตริกระยะทาง D ของอนุกรมเวลา 2 อนุกรมคือ A และ B ซึ่งจะมีขนาด m และ n จุดข้อมูล ตามลำดับ ตาม (2)

$$\delta_{i,j} = (a_i - b_j)^2 \quad (2)$$

ทำการค้นหาระยะทางสะสมที่น้อยที่สุดจาก $D_{0,0}$ ถึง $D_{m,n}$ ซึ่งสามารถใช้ dynamic time warping ช่วยในการค้นหาคู่เปรียบเทียบ ดังนั้นฟังก์ชันที่ใช้ในการค้นหาจึงถูกเปลี่ยนเป็นสมการรีเคอร์ซีฟตาม (3)

$$D(A_i, B_j) = \delta(a_i, b_j) + \min \begin{Bmatrix} D(A_{i-1}, B_{j-1}) \\ D(A_i, B_{j-1}) \\ D(A_{i-1}, B_j) \end{Bmatrix} \quad (3)$$

โดยที่ $D(A_i,B_j)$ จะเป็นระยะสะสม ของค่าระยะที่น้อยที่สุดที่ตำแหน่ง (i,j) หลังจากการคำนวณ m x n ครั้ง $D(A_m,B_n)$ จะมีค่าเป็นกำลังสองของค่าความเหมือนระหว่างทั้งสองอนุกรม ส่วนการปรับแนวจะได้จากเส้นทางการวอร์ป (warping path) ซึ่งจะจับคู่ระหว่างจุดข้อมูลที่ถูกเลือกมาคำนวณระยะทางสะสมที่น้อยที่สุดรายละเอียดเพิ่มสามารถศึกษาได้จาก [2]

2.3 Dynamic time warping barycenter averaging: DBA [1]

dynamic time warping barycenter averaging เป็นการหาค่าเฉลี่ยแต่จะจุดการเปรียบเทียบ จากเส้นทางการเดินด้วยวิธี dynamic time warping ซึ่งก็คือค่าเฉลี่ยของตัวเลขในสมการ (3) เมื่อทำการหาค่าเฉลี่ยในทุกๆจุด ผลลัพธ์ที่ได้จะเป็น อนุกรมเวลาใหม่ที่ได้จากการคิดค่าเฉลี่ยด้วยวิธี dynamic time warping barycenter averaging ซึ่งวิธีการดังนี้

1. คิดคำนวณระยะทางแต่ละจุดของอนุกรมเวลา ด้วยวิธี dynamic time warping และทำการเฉลี่ยจุดเปรียบเทียบเส้นทางการเดินเก็บไว้

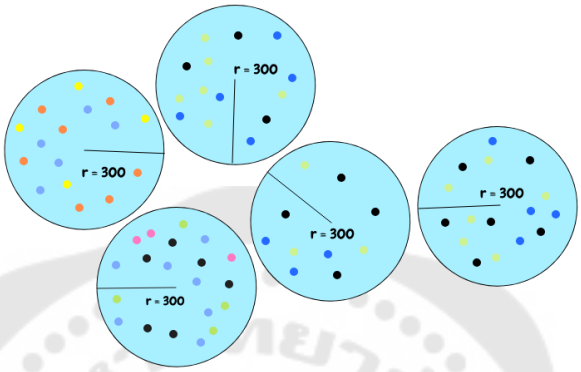
2. ทำการปรับปรุงค่าระยะทางแต่ละจุดโดยใช้ค่าเฉลี่ยจากขั้นตอนที่ 1 ในทุกๆจุดการเดิน อัลกอริทึมการคิด dynamic time warping barycenter averaging แสดงได้ดังนี้

```
Algorithm DBA
Require: C = (C1, ... ,Ct') the initial average sequence
Require: S1 =  (s1 , ... ,s1t) S the 1st sequence to average

Require: Sn = (sn1 , ... ,snt) S the nth sequence to average
Let T be the length of sequences
Let assocTab be a table of size T' containing in each cell a set of
coordinates associated to each coordinate of C
Let m[T,T] be a temporary DTW (cost,path) matrix
   assocTab ←[∅, ... ,∅]
   for seq in S do
      m←DTW(C,seq)
      i←T'
      j←T
   while i⩾1 and j⩾1 do
      assocTab[i] ←assocTab[i] ∪ seqj
      (i,j) ←second(m[i,j])
   end while
end for
for i=1to T do
   Ci’ barycenter(assocTab[i])
end for
return C'
```

2.4 Sequential clustering

Sequential clustering เป็นการจัดกลุ่มที่ไม่ซับซ้อน ตรงไปตรงมา จำนวนกลุ่มที่แบ่งด้วยวิธีนี้จะไม่สามารถกำหนดได้ จำนวนกลุ่มขึ้นจะเพิ่มขึ้นเรื่อยๆ ขึ้นอยู่กับข้อมูลและเงื่อนไขที่กำหนด ตัวอย่างเช่น การกำหนดกลุ่มของข้อมูลอนุกรมเวลา แต่ละอนุกรมที่อยู่กลุ่มเดียวกันจะต้องมีระยะทาง (distance) ห่างกันไม่เกิน 300 เมื่อเปรียบเทียบกับจุดศูนย์กลางด้วยวิธี

dynamic time warping วิธีการจัดกลุ่มคือ จะต้องทำการเปรียบเทียบทุกๆ อนุกรมเวลา ถ้าระยะทางน้อยกว่า 300 ก็จัดให้อยู่กลุ่มเดียวกัน แต่ถ้ามากกว่า 300 ก็ให้ตั้งเป็นกลุ่มใหม่ไปเรื่อยๆ ตัวอย่างแสดงดังภาพประกอบ 2



ภาพประกอบ 2 การจัดกลุ่มด้วยวิธี sequential clustering

2.5 Exponential regression [3]

Exponential regression เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปร 2 ตัวขึ้นไป ซึ่งได้แก่ตัวประมาณการ (predictor, x) และตัวตอบสนอง (response, y) ในขั้นตอนการทำ regression ต้องมีการเก็บจำนวนข้อมูลตัวอย่างจำนวนมากพอ นั้นคือ มี x และ y ที่มีความสัมพันธ์กันหลายๆ ครั้ง เพื่อนำมาหาสมการความสัมพันธ์ สมการ exponential จะอยู่ในฟอร์ม (4)

$$y = ae^{kx}, a > 0 \quad (4)$$

เมื่อได้สมการความสัมพันธ์แล้ว จะสามารถแทนค่าเพื่อหาผลลัพธ์การประเมินตามสมการได้ รายละเอียดเพิ่มเติม [4]

2.6 **ความคลาดเคลื่อนเฉลี่ยกำลังสอง** (root mean square error : RMSE) [5]

ความคลาดเคลื่อนเฉลี่ยกำลังสอง เป็นวิธีการวัดความคลาดเคลื่อนจากค่าการทำนายจากแบบจำลอง เทียบกับค่าจริงที่เกิดขึ้น หากค่า root mean square error มีค่าน้อยแสดงว่า

แบบจำลองสามารถพยากรณ์ค่าได้ใกล้เคียงกับค่าจริง ดังนั้นหากค่า root mean square error มีค่าเท่ากับศูนย์ หมายความว่า ไม่เกิดความคลาดเคลื่อนในแบบจำลองนี้เลย ค่า root mean square error สามารถคำนวณได้ดัง (5)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2} \quad (5)$$

โดยที่

y_i คือ ค่าจริงของข้อมูล

y_i' คือ ค่าที่ได้จากการทำนาย

n คือ จำนวนตัวอย่างข้อมูลที่ใช้ในการทดสอบแบบจำลอง

2.7 Correlation coefficient

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป (หรือข้อมูล 2 ชุดขึ้นไป) ตัวอย่างเช่น การหาความสัมพันธ์ระหว่างยอดขายกับจำนวนลูกค้า ความสัมพันธ์ระหว่างอายุกับจำนวนเวลาในการใช้มือถือ ความสัมพันธ์ระหว่างความเร็วอินเตอร์เน็ตกับความยาวสายไฟเบอร์ออฟติก เป็นต้น โดยมีค่า correlation coefficient หรือ ค่าสัมประสิทธิ์สหสัมพันธ์ เป็นตัวบ่งชี้ถึงความสัมพันธ์นี้ ซึ่งค่าสัมประสิทธิ์สหสัมพันธ์นี้ จะมีค่าอยู่ระหว่าง -1.0 ถึง +1.0 ซึ่งหากมีค่าใกล้ -1.0 นั้นหมายความว่าตัวแปรทั้งสองตัวมีความสัมพันธ์กันอย่างมากในเชิงตรงกันข้าม หากมีค่าใกล้ +1.0 นั้นหมายความว่า ตัวแปรทั้งสองมีความสัมพันธ์กันโดยตรงอย่างมาก และหากมีค่าเป็น 0 นั้นหมายความว่า ตัวแปรทั้งสองตัวไม่มีความสัมพันธ์ต่อกัน [6] ซึ่งการคำนวณ correlation coefficient สามารถคำนวณได้จาก (6)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (6)$$

โดยที่

r คือ ค่าสัมประสิทธิ์สหสัมพันธ์

x_i คือ ค่า x ที่ตำแหน่ง i

y_i คือ ค่า y ที่ตำแหน่ง i

$\bar{x}$ คือ ค่าเฉลี่ยของชุดข้อมูล x

$\bar{y}$ คือ ค่าเฉลี่ยของชุดข้อมูล y

2.8 Softmax

Softmax เป็นการคำนวณการแจกแจงความน่าจะเป็นของเหตุการณ์ในช่วงเหตุการณ์ที่แตกต่างกัน โดยทั่วไปแล้วการจะคำนวณความน่าจะเป็นของคลาสเป้าหมายแต่ละคลาสที่เป็นไปได้ทั้งหมด หลังจากนั้นความน่าจะเป็นที่คำนวณได้จะมีประโยชน์สำหรับการกำหนดคลาสเป้าหมายสำหรับอินพุตที่กำหนด ข้อดีของ softmax คือช่วงความน่าจะเป็นของเอาต์พุต ช่วงจะเป็น 0 ถึง 1 และผลรวมของความน่าจะเป็นทั้งหมดจะเท่ากับ 1 หาก softmax ที่ใช้สำหรับการจำแนกหลายแบบจะส่งคืนความน่าจะเป็นของแต่ละคลาส และคลาสเป้าหมายจะมีความน่าจะเป็นสูง

ในการทำควณ softmax จะทำการคำนวณ ค่า exponential ของอินพุตที่กำหนด และคำนวณผลรวมค่า exponential ของค่าอินพุตทั้งหมด จากนั้นนำไปหาอัตราส่วนระหว่างค่า exponential ของอินพุตกับผลรวมของค่า exponential ผลลัพธ์จะได้เป็นค่า softmax ของกลุ่มอินพุตนั้นๆ สามารถคำนวณได้จากสมการที่ (7)

$$softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{7}$$

โดยที่

$x_{i\ldots j}$ คือ ค่าอินพุตที่ i ถึง j

2.9 R2 score

ค่า r2 score คือตัวสถิติที่ใช้วัดว่าตัวแบบคณิตศาสตร์ที่ได้นี้มีความสมรูปกับข้อมูลมากน้อยอย่างไร หรือรู้จักกัน ในอีกความหมายหนึ่งว่าเป็น ค่าสัมประสิทธิ์แสดงการตัดสินใจ (Coefficient of Determination) หรือ ค่าสัมประสิทธิ์แสดง การตัดสินใจเชิงซ้อน (Coefficient of Multiple Determination) สำหรับการวิเคราะห์การถดถอยแบบพหุคูณ (Multiple Regression) โดยค่า r2 score จะอยู่ระหว่าง 0 ถึง 1 ซึ่งเข้าใกล้ 1 ยิ่งดี สามารถคำนวณได้จากสมการที่ (8)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{8}$$

โดยที่

SS_{res} คือ ผลรวมค่ากำลังสองที่ได้จากการทำนาย

SS_{tot} คือ ผลรวมค่ากำลังสองที่ได้จากเส้นค่าเฉลี่ย

2.10 ทบทวนวรรณกรรม

การทบทวนวรรณกรรมของงานวิจัยนี้ ได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายความนิยมของทวิตเตอร์ มีรายละเอียด ดังต่อไปนี้

2.10.1 วารสารงานวิจัยเรื่อง Analyzing and predicting news popularity on Twitter โดย Bo Wu และคณะ (2015) [3]

งานวิจัยนี้นำเสนอการทำนาย popularity ของสำนักข่าว โดยใช้การเก็บข้อมูลจากสำนักข่าว 33 สำนัก ที่มีผู้ติดตามเป็นจำนวนมาก ระยะเวลาในการเก็บข้อมูล 4 เดือน ซึ่งการวัดค่า popularity วัดจากจำนวนการ retweet ของโพสต์นั่นๆ วิธีการสร้าง model ที่ใช้ในการทำนาย popularity คือ

1. Micro-scope popularity prediction คือ model ผู้ติดตามได้รับข้อมูลข่าวจากสำนักข่าวโดยตรงเลย ไม่ผ่านผู้ติดตามคนอื่น ทำให้สามารถทำนาย popularity จากแหล่งข้อมูลที่แน่ใจได้เลยหลังจากสำนักข่าวโพสต์ข้อความ

2. Macro-scope popularity prediction คือ model การทำนาย popularity จากการรับข้อมูลจากต้นทางถึงปลายทางมีจำนวนโหนด มากกว่า 1 โหนด

โดยการทำนายจะใช้ทั้ง 2 model ในการทำนายร่วมกัน การทำนายจะทำนาย popularity แต่ละวินาทีหลังจากสำนักข่าวโพสต์ข้อความ ในงานวิจัยนี้เปรียบเทียบประสิทธิภาพการทำนายกับ regression prediction model

2.10.2 วารสารงานวิจัยเรื่อง Modeling the infectiousness of Twitter hashtags โดย Jonathan Skaza และคณะ (2016) [6]

งานวิจัยนี้เป็นการศึกษาการทำนาย popularity ของ hashtags ที่ใช้ในทวิตเตอร์ของผู้ที่ใช้งานในเมือง New York City และ San Francisco ประเทศสหรัฐอเมริกา โดยจะนำเสนอการประมาณค่าอัตราการใช้ hashtags และการนำกลับมาใช้ใหม่ ของ hashtags จำนวนหลายร้อย ซึ่งวิธีการสร้าง model จะใช้วิธีการ bayesian markov chain monte carlo (MCMC) และ SIR-type ในการสร้าง model ซึ่งในงานวิจัยนี้จะสร้าง model การทำนาย 2 แบบ คือ dynamical model และ statistical model

2.10.3 วารสารงานวิจัยเรื่อง Fuzzy clustering of time series data using dynamic time warping distance โดย Hesam Izakian และคณะ (2015) [7]

งานวิจัยนี้เป็นการศึกษาเกี่ยวกับการจัดกลุ่มของอนุกรมเวลา (time series) โดยใช้ dynamic time warping ในการหาระยะทางระหว่างอนุกรมเวลา ที่เวลาไม่เท่ากันเพื่อหาวิธีที่ดีในการจัดกลุ่มอนุกรมเวลา ข้อมูลอนุกรมที่ใช้นำมาจากออนไลน์ที่ UCR time series (www.cs.ucr.edu/eamonn/time_series_data/) ซึ่งความยาวและจำนวนของข้อมูลมีไม่เท่ากัน นำข้อมูลมาใช้ในการวิธีทั้งหมด 8 ชุดข้อมูล ในงานวิจัยนี้ใช้วิธีการจัดกลุ่มรวมกับการใช้ dynamic time warping 3 แบบ ได้แก่

1. Fuzzy C-means clustering
2. Fuzzy C-medoids clustering of time series
3. Hybrid fuzzy C-means and fuzzy C-medoids for time series clustering

ในการจัดกลุ่มแต่ละแบบจะใช้ dynamic time warping ในการหาจุดศูนย์กลางของกลุ่ม นอกจากนี้ในงานวิจัยยังนำวิธีการจัดกลุ่มแบบ fuzzy C-means clustering โดยใช้ euclidean ในการหาระยะทาง เพื่อเปรียบผลการจัดกลุ่ม 3 วิธีที่นำเสนอในงานวิจัยนี้ ผลการเปรียบเทียบวิธีการจัดกลุ่ม วิธีที่ให้ค่าความถูกต้องมากที่สุดคือ hybrid fuzzy C-means and fuzzy C-medoids for time series clustering

2.10.4 บทความวิจัยเรื่อง Predicting Twitter hashtags popularity level โดย Shing H. Doong (2016) [8]

งานวิจัยเรื่องนี้นำเสนอการทำนายระดับความนิยม hashtags ของทวิตเตอร์ ข้อมูลที่ใช้คือข้อมูลการโพสต์ข้อความจำนวน 18 ล้านครั้ง มี hashtags จำนวน 748,000 hashtags ซึ่งได้จากการใช้ streaming API ของทวิตเตอร์ ในการเก็บข้อมูลแบบสุ่ม จากกลุ่มข้อมูลของทวิตเตอร์ตั้งแต่วันที่ 13 พฤศภาคม ค.ศ. 2005 ถึง วันที่ 2 มิถุนายม ค.ศ. 2015 โดยเลือกแต่ข้อความที่เป็นภาษาอังกฤษเท่านั้น ความนิยมของ hashtags จะนับจากจำนวนผู้ติดตาม และจำนวนการโพสต์ที่ใช้ hashtags นั้นๆ ในการสกัดหาตัวแปรที่จะทำไปใช้ในการทำนาย งานวิจัยนี้ จะทำการจำกัดข้อมูลโดยการนับจำนวน hashtags ถ้า hashtags ใดมีจำนวนน้อยกว่า 200 จะตัด hashtags นั้นออก จะทำให้เหลือจำนวน hashtags อยู่เพียง 3331 hashtags จากนั้นนำ hashtags ที่เหลือมาจัดระดับโดยใช้จำนวน hashtags ในการแบ่งระดับซึ่งจะแบ่งได้ 3 ระดับคือ low medium high และทำการสกัดตัวแปรที่จะใช้ในการทำนายอีก 18 ตัวแปร ซึ่งสกัดจากข้อมูลที่สุ่มมาจากทวิตเตอร์ รวมเป็น 19 ตัวแปรที่จะใช้ในการทำนาย

งานวิจัยนี้จะใช้ random forest (RF) ในการทำนายระดับความนิยมของ hashtags และใช้ตัวแปรอีก 19 ตัวที่ได้จากสกัดในขึ้นตอนก่อนหน้าเป็นตัวแปรในการสร้าง model การทำนาย ในการวัดความถูกต้องจะใช้ผลรวมของค่า recall ของแต่ละระดับ

2.10.5 บทความวิจัยเรื่อง The pulse of news in social media: forecasting popularity โดย Roja Bandari และคณะ(2012) [9]

บทความวิจัยนี้นำเสนอการทำนาย popularity ของข่าวบนทวิตเตอร์ การเก็บข้อมูลจะใช้ API ที่ชื่อว่า Feedzilla ใช้เวลาในการเก็บข้อมูลเป็นเวลา 7 วัน ตั้งแต่วันที่ 8 สิงหาคม ค.ศ. 2011 ถึง 16 สิงหาคม ค.ศ. 2011 ข้อมูลโพสต์ที่ได้ มีจำนวน 44,000 โพสต์ จากนั้นนำข้อมูลไปจัดกลุ่มเพื่อสกัดตัวแปรไปให้ในการทำนาย popularity บทความวิจัยนี้ใช้วิธีการทำนาย 2 แบบคือ

1. regression จะใช้ 2 เทคนิคได้แก่ linear regression และ svm regression

2. classification จะใช้ 4 เทคนิคด้วยกันได้แก่ bagging, j48 decision trees, svm และ naive bayes

โดยสรุปการทำนายแบบ classification ด้วยเทคนิค bagging จะให้ค่าความถูกต้องมากที่สุด

บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีจุดประสงค์ เพื่อทำการศึกษาวิเคราะห์และทดลองทำนายความนิยมของผู้ติดตาม ที่มีต่อการโพสต์ข้อความบนทวิตเตอร์ โดยใช้เทคนิค dynamic time warping ซึ่งในบทนี้จะได้กล่าวถึงสาระสำคัญเกี่ยวกับวิธีการดำเนินการศึกษาค้นคว้า คือ กลุ่มเป้าหมายข้อมูลที่ใช้ในการวิจัย เครื่องมือที่ใช้ในงานวิจัย การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล การทำนายความนิยมที่มีต่อการโพสต์บนทวิตเตอร์และการวัดผลความถูกต้องของการทำนายความนิยมเชิงเวลาตามลำดับ ดังนี้

3.1 กลุ่มเป้าหมายข้อมูลที่ใช้ในการวิจัย

ตามจุดประสงค์งานวิจัย มุ่งเน้นการวิเคราะห์และทำนายความนิยมของการโพสต์บนทวิตเตอร์ ผู้วิจัยจึงเลือก กลุ่มเป้าหมายข้อมูลที่ใช้ในการวิจัยคือทวิตเตอร์ของสำนักข่าวที่โพสต์ข้อมูลข่าวเป็นภาษาอังกฤษ และมีผู้ติดตามเกิน 5 ล้านคน เนื่องจากจะได้ข้อมูลจากกลุ่มตัวอย่างที่คล้ายคลึงกันคือผู้ติดตามข่าว และที่ต้องมีผู้ติดตามเกิน 5 ล้านคน เพราะจะได้มีตอบโพสต์และการโพสต์ซ้ำในปริมาณมากพอที่จะทำไปวิเคราะห์และทำนายได้ ผู้วิจัยจึงเลือกใช้ข้อมูลจากทวิตเตอร์ของสำนักข่าว 4 สำนักได้แก่

1. สำนักข่าว CNN มีผู้ติดตามจำนวน 40.7 ล้านคน

ภาพประกอบ 3 หน้าทวิตเตอร์ของสำนักข่าว CNN

2. สำนักข่าว Fox News มีผู้ติดตามจำนวน 18.2 ล้านคน

ภาพประกอบ 4 หน้าทวิตเตอร์ของสำนักข่าว Fox News

3. สำนักข่าว The New York Times มีผู้ติดตามจำนวน 42.3 ล้านคน

ภาพประกอบ 5 หน้าทวิตเตอร์ของสำนักข่าว The New York Times

4. สำนักข่าว The Washington Post มีผู้ติดตามจำนวน 13 ล้านคน

ภาพประกอบ 6 หน้าทวิตเตอร์ของสำนักข่าว The Washington Post

3.2 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้งานวิจัยการวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์โดยใช้ dynamic time warping ได้แก่

1. โปรแกรมภาษา python คือชื่อภาษาที่ใช้ในการเขียนโปรแกรมภาษาหนึ่ง ซึ่งถูกพัฒนาขึ้นมาโดยไม่ยึดติดกับแพลตฟอร์ม กล่าวคือสามารถรันภาษ python ได้ทุกแพลตฟอร์มนั้นเอง เนื่องจากภาษา python เป็นภาษาสคริปต์ ทำให้ใช้เวลาในการเขียนและคอมไพล์ไม่มาก ไวยากรณ์ของภาษา python ได้กำจัดการใช้สัญลักษณ์ที่ใช้ในการแบ่งบล็อกของโปรแกรม ใช้การย่อหน้าแทน ทำให้สามารถอ่านโปรแกรมที่เขียนได้ง่าย นอกจากนี้โปรแกรมภาษา python ยังมีไลบรารีสำหรับใช้ในการเก็บข้อมูลจากทวิตเตอร์ ใช้ในการวิเคราะห์ข้อมูล ร่วมถึงใช้ในการทำนายความนิยมเตรียมพร้อมไว้อยู่แล้ว ผู้จัดทำจึงเลือกใช้โปรแกรมภาษา python ในการทำวิจัยในครั้งนี้

2. Jupyter notebook คือโปรแกรมที่ใช้ในการเขียน รัน และแสดงผลลัพธ์ โค้ดภาษา python ที่ผู้วิจัยได้เขียนในการทำงานวิจัย jupyter notebook สามารถที่จะรันโค้ดทีละบล็อคซึ่งจะแสดงผลลัพธ์ทันที ทำให้สะดวกในการเขียนโค้ดเป็นอย่างมาก นอกจากนี้ยังสามารถแสดงผลลัพธ์ของโค้ดให้เป็นรูปแบบของกราฟได้ และแสดงข้อมูลในรูปแบบตารางได้อีกด้วย

```
In [4]: cnn_response_timeseries= pd.read_csv('./data/cnn_response_timeseries.csv')
        cnn_response_timeseries.head(5)
```

Out[4]:

	id	minute	counts
0	1039038561557925888	25609352	1
1	1039038561557925888	25609353	12
2	1039038561557925888	25609354	16
3	1039038561557925888	25609355	21
4	1039038561557925888	25609356	30

ภาพประกอบ 7 ตัวอย่างผลลัพธ์ที่แสดงบน jupyter notebook

3. Amazon Elastic Compute Cloud (Amazon EC2) คือผู้ให้บริการคลาวด์ ผู้วิจัยจะทำการติดตั้ง jupyter notebook และโปรแกรมภาษา python ไว้บนคลาวด์ เพื่อใช้เป็นเครื่องคอมพิวเตอร์สำหรับรันโค้ดต่างๆ ในงานวิจัย ซึ่งในงานวิจัยจำเป็นจะต้องรันโค้ดภาษา python ต่อเนื่องเป็นเวลานาน จึงเลือกที่จะใช้คลาวด์เนื่องจากมีความเสถียรสูง และสะดวกในการใช้งาน สามารถเข้าใช้งานได้ทุกที่ที่มีอินเตอร์เน็ต

3.3 การเก็บข้อมูล

ในงานวิจัยนี้จะใช้ข้อมูลจากทวิตเตอร์ของสำนักข่าว CNN, Fox News, The New York Times และThe Washington Post ในการเก็บข้อมูลนั้น จะต้องเก็บข้อมูลกิจกรรมที่เกิดขึ้นบนทวิตเตอร์ของสำนักข่าวทั้ง 4 สำนัก เป็นระยะเวลาหนึ่งสัปดาห์ การเก็บข้อมูลนั้น ทวิตเตอร์จะมีช่องทางสำหรับเก็บข้อมูลจัดเตรียมไว้ให้ คือช่องทางที่มีชื่อว่า Filter real time tweets ซึ่งวิธีการใช้งานค่อนข้างซับซ้อนใช้งานยาก จึงมีผู้พัฒนา ไลบรารีภาษา python ที่ชื่อว่า Tweepy ที่ใช้สำหรับเก็บข้อมูลกิจกรรมบนทวิตเตอร์ให้ใช้งานง่ายขึ้น ในการเก็บข้อมูลจะมีวิธีการดังต่อไปนี้

1. เปลี่ยนชื่อทวิตเตอร์ของสำนักข่าวเป็นทวิตเตอร์ id เนื่องจากการเก็บข้อมูลโดยใช้ไลบรารี Tweepy จะใช้ชื่อผู้ใช้งานทวิตเตอร์เป็นตัวเลขเพื่อใช้ในการเก็บข้อมูล ซึ่งวิธีการเปลี่ยนจากชื่อเป็น id นั้นสามารถเปลี่ยนได้ที่ http://gettwitterid.com โดยใส่ชื่อทวิตเตอร์เข้าไป จากนั้นกดปุ่ม GET USER ID เว็บไซต์จะแสดง id ให้เห็นด้านล่าง ตัวอย่าง ดังภาพประกอบ 8

ภาพประกอบ 8 เมนูการเปลี่ยนชื่อเป็น id

จะต้องทำการเปลี่ยนจากชื่อทวิตเตอร์เป็น id ทั้งสี่สำนักข่าว ผลลัพธ์ที่ได้แสดงในตาราง 1

ตาราง 1 ตางรางแสดงชื่อและ id ของทวิตเตอร์แต่ละสำนักข่าว

Full Name	Screen Name	Twitter User ID
CNN	CNN	759251
The New York Times	nytimes	807095
Fox News	FoxNews	1367531
The Washington Post	washingtonpost	2467791

2. เขียนโปรแกรมเก็บข้อมูลจากทวิตเตอร์ของสำนักข่าวทั้ง 4 สำนัก สิ่งที่จำเป็นที่จะต้องมีคือ consumer key (api key), consumer secret (api secret), access token และ access token secret เพื่อใช้ยืนยันตัวตนในการเก็บข้อมูลทวิตเตอร์ สามารถขอได้ที่ https://apps.twitter.com ในการเก็บข้อมูลจะแยกการรันโปรแกรม 1 โปรแกรมต่อ 1 สำนักข่าว เพื่อให้ง่ายต่อการแยกข้อมูล ดังนั้น การเก็บข้อมูล 4 สำนักข่าวจะต้องรันโปรแกรม 4 โปรแกรมพร้อมกัน ตัวอย่างโค้ดการเก็บข้อมูลจากสำนักข่าว CNN จะแสดงดังได้ดังนี้

```
import tweepy
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream, API
consumer_key = "key"
consumer_secret = "key"
access_token = "key"
access_token_secret = "key"
auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
class CustomStreamListener(tweepy.StreamListener):
    def on_data(self, data):
        with open('./cnn.json', 'a') as f:
            f.write(data)
            return True
def start_stream():
    while True:
        try:
            sapi = tweepy.streaming.Stream(auth,
CustomStreamListener(api))
```

```
            sapi.filter(follow=['759251'])
        except:
            continue
start_stream()
```

ซึ่งในงานวิจัยนี้จะให้เวลาในการเก็บข้อมูลเป็นเวลา 1 สัปดาห์

3.4 วิเคราะห์ข้อมูล

3.4.1 จัดกลุ่มกิจกรรมข้อมูลจากทวิตเตอร์

รูปแบบของข้อมูลที่ได้จากทวิตเตอร์ ในการเก็บข้อมูลตามหัวข้อที่ 3.3 ข้อ 2 ข้อมูลที่ได้จะอยู่ในรูปแบบ คีย์-แวลู ที่เรียกว่า javascript object notation (json) โดยกิจกรรมที่เกิดขึ้นบนทวิตเตอร์ 1 กิจกรรมจะมีข้อมูลอยู่ในรูปแบบ json แสดงตัวอย่าง ดังนี้

```
{
  "created_at":"Wed Nov 14 08:27:37 +0000 2018",
  "id":1062623071293517824,
  "id_str":"1062623071293517824",
  "text":"RT @washingtonpost: U.S. military could lose a war to
China or Russia, commission warns https:\/\/t.co\/fiOUkUnufV",
  "source":"\u003ca
href=\"http:\/\/twitter.com\/#!\/download\/ipad\"
rel=\"nofollow\"\u003eTwitter for iPad\u003c\/a\u003e",
....
  "possibly_sensitive":false,
  "filter_level":"low",
  "lang":"en",
  "timestamp_ms":"1542184057435"
}
```

เนื่องจากในงานวิจัยนี้สนใจกิจกรรม 3 กิจกรรมคือ การโพสต์ข้อความ (tweet) การตอบข้อความ (reply) และการโพสต์ข้อความซ้ำ (retweet) ซึ่งแต่ละกิจกรรมจะถูกเก็บปะปนกันอยู่ในไฟล์ข้อมูลที่ได้จากหัวข้อที่ 3.3 ข้อ 2 (cnn.json, foxnews.json, nytimes.json, washingtonpost.json) จะต้องทำการจัดกลุ่มกิจกรรมแต่ละอย่าง โดยวิธีการจัดกลุ่มจะต้องจัดตามค่าของ คีย์-แวลู โดยอ้างอิงความหมายแต่ละคู่ข้อมูลจาก เว็บไซต์ทวิตเตอร์ ซึ่งจำนวนแต่ละกิจกรรมแสดงได้ตามตาราง 2

ตาราง 2 แสดงจำนวนกิจกรรมที่เกิดขึ้นบนทวิตเตอร์ของแต่ละสำนักข่าว

ชื่อสำนักข่าว	tweet	reply	retweet
CNN	855	69,667	147,404
The New York Times	602	51,759	166,330
Fox News	1,100	132,092	193,871
The Washington Post	1,001	63,459	95,983

3.4.2 **สร้างข้อมูลอนุกรมเวลา** (time series data)

เนื่องจากในการวิเคราะห์และการทำนายความนิยมเชิงเวลาโดยใช้ dynamic time warping จะต้องใช้ข้อมูลที่อยู่ในรูปแบบอนุกรมเวลา ตัวอย่างเช่น [3,5,6,8,10,15,30] โดยจะสร้างจากข้อมูลที่เก็บมา กำหนดให้ 1 อนุกรมเวลา คือ 1 โพสต์ ตัวอย่างเช่น สำนักข่าว CNN มีการโพสต์ข้อความบนทวิตเตอร์ 855 ครั้ง จำนวนอนุกรมเวลาที่สร้างก็คือ 855 อนุกรม ระยะเวลาห่างกันแต่ละจุดคือ 1 นาที ตัวเลขในแต่ละจุดคือ ผลรวมของการตอบโพสต์ (reply) บวกกับการโพสต์ข้อความซ้ำ (retweet) สามารถเขียนโค้ดโปรแกรมการสร้างข้อมูลอนุกรมเวลา ได้ดังต่อไปนี้

```
id = 0
for i in range(len(foxnews_response_group)):
    if (foxnews_response_group['id'].values[i] == id):

        foxnews_response_group['counts'].values[i] +=
        foxnews_response_group['counts'].values[i-1]

    id = foxnews_response_group['id'].values[i]

foxnews_response_timeseries =
(foxnews_response_group.groupby('id')['minute', 'counts']
                       .apply(lambda x: x.set_index('minute')
                       .reindex(np.arange(x['minute']
                       .min(), x['minute'].max() + 1)))
                       .ffill()
                       .astype(int)
                       .reset_index())
```

3.4.3 จัดกลุ่มอนุกรมเวลา

จากหัวข้อที่ 3.4.2 จะได้อนุกรมเวลาของโพสต์ แต่ละสำนักข่าว ผู้วิจัยต้องการจัดกลุ่มอนุกรมเวลาของแต่ละสำนักข่าว ที่มีความคล้ายคลึงกันให้อยู่ในกลุ่มเดียวกัน ซึ่งความคล้ายคลึงนั้นจะใช้วิธี dynamic time warping เป็นตัววัดระยะทาง (distance) ถ้าระยะทางมีค่าน้อยแสดงว่ามีความคล้ายคลึงกัน และใช้วิธี sequential clustering ในการจัดกลุ่มอนุกรมเวลา จะมีวิธีการจัดกลุ่มดังต่อไปนี้

1. นำอนุกรมเวลา 2 อนุกรมแรกของชุดข้อมูล มาเปรียบเทียบความคล้ายคลึงกันโดยใช้วิธี dynamic time warping ซึ่งในการคำนวณระยะทางจะใช้ไลบรารีที่ชื่อว่า dtaidistance เป็นตัวช่วยในการคำนวณระยะทาง

2. ถ้าระยะทาง ระหว่าง 2 อนุกรมเวลาน้อยกว่า 300 จะถือว่ามีความคล้ายคลึงกันจะจัดอยู่ในกลุ่มด้วยกัน จากนั้นทำการหาค่าเฉลี่ยของอนุกรมเวลาทั้งสองโดยใช้วิธี dynamic time warping barycenter averaging เพื่อเป็นอนุกรมเวลาประจำกลุ่ม โดยในการคำนวณค่าเฉลี่ยจะใช้ไลบรารีที่ชื่อว่า tslearn barycenters DTWBarycenterAveraging

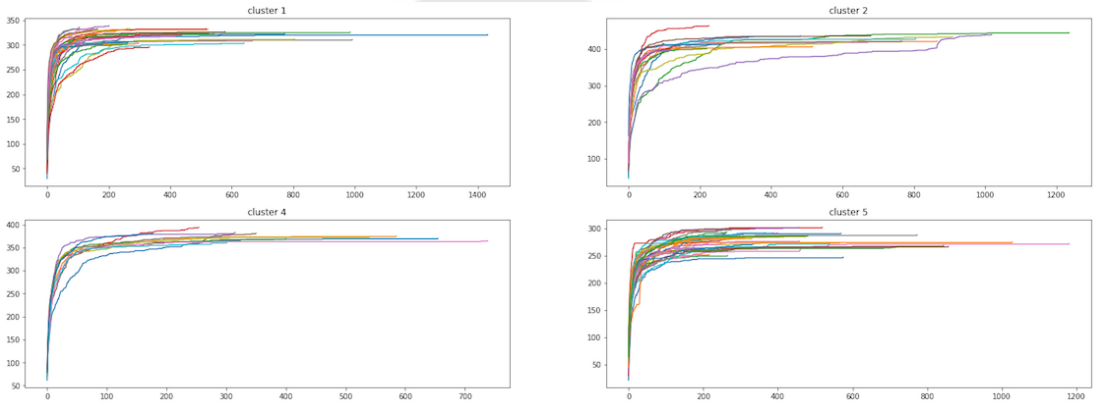
3. ถ้าระยะทางระหว่าง 2 อนุกรมเวลามากกว่า 300 จะถือว่าอนุกรมเวลา อยู่คนละกลุ่มกัน ให้อนุกรมเวลาแยกเป็นกลุ่มใหม่

4. นำอนุกรมเวลาที่เหลือ มาเปรียบเทียบความคล้ายคลึงกับกลุ่มอนุกรมเวลาก่อนหน้านี้ให้ครบทุกกลุ่ม โดยใช้เงื่อนไขตามข้อ 2 และ ข้อ 3 จนครบทุกอนุกรมเวลาของแต่ละสำนักข่าว ซึ่งสามารถเขียนโค้ดโปรแกรมการจัดกลุ่มอนุกรมเวลา ได้ดังต่อไปนี้

```
r = 300
c_id = 1

for i in range(len(series_test) -1):
    i += 1
    series = series_test.at[i,'counts']
    new = True

    for j in range(len(cluster)):
        d = dtw.distance_fast(series,cluster.at[j,'centroid'])
        if(d <= r):
            c_new_list =

to_time_series_dataset([series,cluster.at[j,'centroid']])
            bary = dtw_barycenter_averaging(c_new_list)
            c_new = list(map(float,bary))
            c_new = np.array(c_new)
            cluster['centroid'].values[j] = c_new
```

```
            series_test['cluster_id'].values[i] =
cluster['cluster_id'].values[j]
            new = False
            break
    if(new):
        c_id += 1
        cluster.loc[len(cluster)]=[c_id,series_test.at[i,'counts']]
        series_test['cluster_id'].values[i] = c_id
```

สามารถแสดงกราฟตัวอย่างของกลุ่มอนุกรมเวลาได้ดังภาพประกอบ 9



ภาพประกอบ 9 ตัวอย่างการจัดกลุ่มอนุกรมเวลา

ผลลัพธ์การจัดกลุ่มอนุกรมเวลาของแต่ละสำนักข่าว เป็นไปตามตาราง 3

ตาราง 3 ตารางแสดงจำนวนกลุ่มของอนุกรมเวลา แต่ละสำนักข่าว

ชื่อสำนักข่าว	จำนวนกลุ่มอนุกรมเวลา
CNN	57
The New York Times	99
Fox News	75
The Washington Post	98

3.5 ทำนายความนิยมเชิงเวลา

การทำนายความนิยมเชิงเวลาในงานวิจัยนี้ จะเป็นการทำนายความนิยมของแต่ละโพสต์ที่สำนักข่าวโพสต์ไว้บนทวิตเตอร์ โดยค่าความนิยมที่จะทำนายคือ จำนวนการตอบโพสต์บวกด้วยจำนวนการโพสต์ซ้ำ จะใช้ข้อมูล 7 ชั่วโมงในการทำนาย จะแบ่งการทำนายเป็น 2 กรณี กรณีที่แรกใช้ข้อมูล 1 ชั่วโมงในการหา distance ระหว่างอนุกรมเวลากับเทมเพลต และทำการทำนายต่อไปอีก 6 ชั่วโมง กรณีที่สอง จะใช้ข้อมูล 2 ชั่วโมงในการหา distance ระหว่างอนุกรมเวลากับเทมเพลต และทำการทำนายต่อไปอีก 5 ชั่วโมง ซึ่งในงานวิจัยนี้จะใช้วิธีการทำนาย 3 แบบ คือ

1. การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง
2. การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax
3. ทำนายความนิยมโดยใช้แบบจำลอง exponential

เพื่อใช้ในการเปรียบเทียบประสิทธิภาพในการทำนายความนิยมเชิงเวลา แต่ละวิธีจะมีขั้นตอนดำเนินการ ดังต่อไปนี้

3.5.1 การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง

จากการจัดกลุ่มของอนุกรมเวลาในหัวข้อที่ 3.4 เราจะได้กลุ่มของอนุกรมเวลาของแต่ละสำนักข่าว ในขั้นตอนนี้เราจะนำกลุ่มของอนุกรมเวลามาสร้าง เทมเพลตโมเดลเพื่อใช้ในการทำนายความนิยมเชิงเวลา

$$x(t) = \frac{(d_1^{-1}c_1(t)+d_2^{-1}c_2(t)+\cdots+d_i^{-1}c_i(t))}{(d_1^{-1}+d_2^{-1}+\cdots+d_i^{-1})} \quad (9)$$

โดยที่

x(t) คือ ค่าความนิยม ณ จุดเวลา t

d_i คือ ค่าระยะทาง (distance) ระหว่างแต่ละอนุกรมเวลากับ เทมเพลตโมเดลที่ i

$c_i(t)$ คือ ค่าความนิยม ณ จุดเวลา t ของเทมเพลตโมเดลที่ i

โค้ดโปรแกรมการทำนายความนิยมสามารถเขียนได้ดังนี้

```
def predict_func(p):
    distance = p['distance']
    sum_distance = p['sum_dist']
    predict_array = []
    for n in range(len(temp_test.at[0,'template'])):
        predict_point = 0
        for i in range(len(distance)):
            t_p = temp_test.at[i,'template']
            predict_point += (distance[i] * t_p[n])
        predict_array.append( (predict_point/sum_distance) )
    p['predict'] = np.array( predict_array )
    return p
test_data = test_data.apply(predict_func,axis=1)
```

3.5.2 การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax

จากการจัดกลุ่มของอนุกรมเวลาในหัวข้อที่ 3.4 เราจะได้กลุ่มของอนุกรมเวลาของแต่ละสำนักข่าว ในขั้นตอนนี้เราจะนำกลุ่มของอนุกรมเวลามาสร้าง เทมเพลตโมเดลเพื่อใช้ในการทำนายความนิยมเชิงเวลา

$$x(t) = \frac{\sum_i \left(e^{(D_i)}\right)^{-1} . c_i(t)}{\sum_i (e^{(D_i)})^{-1}} \quad (10)$$

โดยที่

x(t) คือ ค่าความนิยม ณ จุดเวลา t

d_i คือ ค่าระยะทาง (distance) ระหว่างแต่ละอนุกรมเวลากับ เทมเพลตโมเดลที่ i

$c_i(t)$ คือ ค่าความนิยม ณ จุดเวลา t ของเทมเพลตโมเดลที่ i

โค้ดโปรแกรมการทำนายความนิยมสามารถเขียนได้ดังนี้

```
def predict_func(p):
    expo = p['expo']
    sum_expo = p['sum_expo']
    predict_array = []
    for n in range(len(temp_test.at[0,'template'])):
        predict_point = 0
        for i in range(len(expo)):
            t_p = temp_test.at[i,'template']
            predict_point += (expo[i] * t_p[n])
        predict_array.append( (predict_point/sum_expo) )
    p['predict'] = np.array( predict_array )
    return p
test_data = test_data.apply(predict_func,axis=1)
```

3.5.3 ทำนายความนิยมโดยใช้แบบจำลอง exponential

ในการทำนายโดยใช้แบบจำลอง exponential มีวิธีการทำดังนี้

1. ใช้ไลบรารีที่ชื่อว่า curve_fit ในการสร้างโมเดลเพื่อให้ทำนายผล ตามสมการ exponential (11)

$$y_i = m(1 - e^{(-\mathrm{x}.i)}) \tag{11}$$

2. การทำนายความนิยม จะทำการแทนค่าเวลาในสมการตามข้อ 1 เพื่อทำนายผลโค้ดโปรแกรมสามารถเขียนได้ดังนี้

```
def predict(curve):
    m,t = curve['curve']
    predict = []
    for i in time:
        y = m * (1 - np.exp(-t*i))
        predict.append(y)
    curve['predict'] = np.array(predict)
    return curve
train_data = train_data.apply(predict, axis=1)
```

3.6 การวัดผลความถูกต้องของการทำนายความนิยมเชิงเวลา

ในงานวิจัยนี้จะใช้ค่าเฉลี่ย root mean square error (RMSE) ค่าเฉลี่ยของ correlation coefficient และค่า r2 score เป็นตัวัดประสิทธิภาพความแม่นยำในการทำนาย ซึ่งแต่ละวิธีการมีขั้นตอนหาคำนวณ ดังนี้

3.6.1 การหาค่า root mean square error (RMSE)

ในงานวิจัยนี้จะใช้ library code จาก sklearn ในการหาค่า RMSE สามารถเขียนโค้ดเพื่อหาค่า RMSE ได้ดังนี้

```
def rmse(p):
    return sqrt(mean_squared_error(p['test'], p['predict']))
test_data['rmse'] = test_data.apply(rmse,axis=1)
```

3.6.2 การหาค่า correlation coefficient

ในงานวิจัยนี้จะใช้ library code จาก scipy ในการหาค่า correlation coefficient สามารถเขียนโค้ดเพื่อหาค่า correlation coefficient ได้ดังนี้

```
def coeff(p):
    r,p_v= stats.pearsonr(p['predict'],p['test'])
    p['coeff'] = r
    return p
```

3.6.3 การหาค่า r2 score

ในงานวิจัยนี้จะใช้ library code จาก sklearn ในการหาค่า r2 score สามารถเขียนโค้ดเพื่อหาค่า r2 score ได้ดังนี้

```
def r2(p):
    return r2_score(p['test'],p['predict'])
test_data['r2_score'] = test_data.apply(r2,axis=1)
```

3.7 สิ่งที่จะทำต่อไปให้อนาคต

1. ทำการแบ่งเวลาในการทดสอบความถูกต้องให้มีหลายกรณีมากขึ้น เช่น แบ่ง 3 ชั่วโมง สำหรับหา distance อีก 1 ชั่วโมงสำหรับการทดสอบการทำนาย

2. ใช้วิธีการวัดความถูกต้องในการทำนายแบบอื่น เพิ่มเติมจากวิธี root mean square error และ correlation coefficient

3. หาวิธีการกำหนด รัศมี ที่เหมาะสมให้กับวิธีการจัดกลุ่มแบบ sequential clustering

4. ศึกษางานวิจัยที่เกี่ยวข้องกับการทำนายความนิยมของทวิตเตอร์เพิ่มเติม

บทที่ 4
ผลการดำเนินงานวิจัย

ในงานวิจัยนี้มีจุดประสงค์ เพื่อทำการศึกษาวิเคราะห์และทดลองทำนายความนิยมของผู้ติดตาม ที่มีต่อการโพสต์ข้อความบนทวิตเตอร์ โดยใช้เทคนิค dynamic time warping ใช้ข้อมูลในการวิเคราะห์และทำนายความนิยมของโพสต์ จากทวิตเตอร์ของสำนักข่าวจำนวนสี่ สำนักข่าวได้แก่ CNN(@CNN), The New York Times (@nytimes), Fox News (@FoxNews) และ Washington Post (@washingtonpost) ใช้เวลาในการเก็บข้อมูลจากทั้งสี่สำนักข่าว เป็นเวลาหนึ่งสัปดาห์ จากนั้น ใช้เทคนิค dynamic time warping และ dynamic time warping barycenter averaging ในการวิเคราะห์ข้อมูล และใช้เทคนิค sequential clustering ในการจัดกลุ่มข้อมูล เมื่อจัดกลุ่มข้อมูลแล้ว จะทำกลุ่มข้อมูลไปทำนายความนิยมของแต่ละโพสต์ โดยใช้วิธีการทำนายความนิยม 3 วิธี ได้แก่ การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax และการทำนายความนิยมโดยใช้แบบจำลอง exponential ในการวัดความแม่นยำในการทำนายผลความนิยมจะใช้ ค่า RMSE, correlation coefficient และ r2 score เป็นตัววัดประสิทธิภาพ

การแบ่งข้อมูลเพื่อใช้ในการทำนายความนิยม จะแบ่งเป็น 2 แบบดังนี้

1. แบ่ง 1 ชั่วโมงของแต่ละโพสต์ เพื่อหาระยะทางระหว่างโพสต์กับเทมเพลต และทำนายความนิยมต่อไปอีก 6 ชั่วโมง
2. แบ่ง 2 ชั่วโมงของแต่ละโพสต์ เพื่อหาระยะทางระหว่างโพสต์กับเทมเพลต และทำนายความนิยมต่อไปอีก 5 ชั่วโมง

แต่ละวิธีการทำนายจะมีผลการดำเนินงานดังต่อไปนี้

4.1 ผลการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง

จากการเก็บข้อมูลจากสำนักข่าวทั้งสี่สำนัก และนำข้อมูลมาจัดกลุ่มแล้วเรียบร้อยแล้ว จากนั้นนำข้อมูลโพสต์ ของแต่ละสำนักข่าวมาตัดช่วงเวลาเป็น 1 และ 2 ชั่วโมงเพื่อทำการหาระยะทางระหว่างเทมเพลตกับแต่ละโพสต์ และทำการหาค่าเฉลี่ยเพื่อทำนายผลความนิยม จากนั้นจะทำการทำนายความนิยมต่ออีก 6 และ 7 ชั่วโมงตามลำดับ เมื่อได้ผลการทำนายแล้วจะใช้ ค่า RMSE, correlation coefficient และ r2 score ในการวัดประสิทธิภาพ ซึ่งจะได้ผลตามตาราง 4 และ 5

ตาราง 4 แสดงผลลัพธ์ของวิธีการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง ใช้ 1 ชั่วโมงในการหาระยะทางระหว่างเทมเพลต

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	33.47	0.53	0.96
The New York Times	27.74	0.88	0.96
Fox News	64.79	0.33	0.96
The Washington Post	25.35	0.51	0.94

ตาราง 5 แสดงผลลัพธ์ของวิธีการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทางใช้ 2 ชั่วโมงในการหาระยะทางระหว่างเทมเพลต

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	27.91	0.20	0.93
The New York Times	23.36	0.84	0.94
Fox News	54.45	0.43	0.94
The Washington Post	20.53	0.80	0.91

4.2 ผลการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax

วิธีการนี้จะต่างจากวิธีที่ 4.1 ตรงที่เปลี่ยนจากการใช้ค่าเฉลี่ยมาเป็นใช้ softmax function แทน ซึ่งผลลัพธ์ได้แสดงที่ตาราง 6 และ 7

ตาราง 6 แสดงผลลัพธ์ของวิธีการเฉลี่ยเทมเพลตแบบใช้ softmax ใช้ 1 ชั่วโมงในการหาระยะทางระหว่างเทมเพลต

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	13.60	0.66	0.94
The New York Times	11.17	0.91	0.94
Fox News	25.82	0.74	0.92
The Washington Post	11.26	0.61	0.85

ตาราง 7 แสดงผลลัพธ์ของวิธีการเฉลี่ยเทมเพลตแบบใช้ softmaxใช้ 2 ชั่วโมงในการหาระยะทางระหว่างเทมเพลต

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	9.09	0.31	0.87
The New York Times	8.09	0.88	0.91
Fox News	19.24	0.79	0.81
The Washington Post	6.75	0.90	0.81

4.3 ผลการทำนายความนิยมโดยใช้แบบจำลอง exponential

วิธีการทำนายนี้จะใช้เพื่อเปรียบเทียบประสิทธิภาพการทำนายความนิยม กับวิธีที่ 4.1 และ 4.2 ซึ่งผลลัพธ์ได้แสดงที่ตาราง 8 และ 9

ตาราง 8 แสดงผลลัพธ์ของวิธีการใช้แบบจำลอง exponential ในช่วงเวลาที่ 2 ถึง 7 ชั่วโมง

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	77.88	-1.65	0.28
The New York Times	101.41	0.08	0.33
Fox News	103.58	-0.97	0.24
The Washington Post	36.45	-0.86	0.32

ตาราง 9 แสดงผลลัพธ์ของวิธีการใช้แบบจำลอง exponential ในช่วงเวลาที่ 3 ถึง 7 ชั่วโมง

ชื่อสำนักข่าว	RMSE	R^2	Coeff.
CNN	63.94	-2.66	0.23
The New York Times	90.87	0.14	0.25
Fox News	85.87	-0.73	0.23
The Washington Post	55.08	-0.40	0.27

จากผลลัพธ์ในตารางที่ผ่านมา สามารถนำมารวมตารางเพื่อง่ายต่อการเปรียบเทียบข้อมูลได้ดังนี้

ตาราง 10 แสดงผลลัพธ์การทำนายความนิยมที่ช่วงเวลา 2 ถึง 7 ชั่วโมง

ชื่อสำนักข่าว	DTW avg.			DTW softmax			exponential		
	RMSE	R^2	Coeff.	RMSE	R^2	Coeff.	RMSE	R^2	Coeff.
CNN	33.47	0.53	0.96	13.60	0.66	0.94	77.88	-1.65	0.28
The New York Times	27.74	0.88	0.96	11.17	0.91	0.94	101.41	0.08	0.33
Fox News	64.79	0.33	0.96	25.82	0.74	0.92	103.58	-0.97	0.24
The Washington Post	25.35	0.51	0.94	11.26	0.61	0.85	36.45	-0.86	0.32

ตาราง 11 แสดงผลลัพธ์การทำนายความนิยมที่ช่วงเวลา 3 ถึง 7 ชั่วโมง

ชื่อสำนักข่าว	DTW avg.			DTW softmax			exponential		
	RMSE	R^2	Coeff.	RMSE	R^2	Coeff.	RMSE	R^2	Coeff.
CNN	27.91	0.20	0.93	9.09	0.31	0.87	63.94	-2.66	0.23
The New York Times	23.36	0.84	0.94	8.09	0.88	0.91	90.87	0.14	0.25
Fox News	54.45	0.43	0.94	19.24	0.79	0.81	85.87	-0.73	0.23
The Washington Post	20.53	0.80	0.91	6.75	0.90	0.81	55.08	-0.40	0.27

บทที่ 5
สรุปผลการวิจัย อภิปายผล และข้อเสนอแนะ

ในการวิจัยเรื่องการวิเคราะห์และทำนายความนิยมเชิงเวลาของทวิตเตอร์ ผู้วิจัยได้ทำการวัดประสิทธิภาพของแต่ละวิธีการทำนายความนิยม โดยใช้ค่า root mean square error (RMSE) ค่า correlation coefficient และค่า r2 score เป็นเกณฑ์ในการวัดประสิทธิภาพ จากผลที่ได้ดำเนินงานแล้ว สามารถสรุปผลการดำเนินงานได้ดังต่อไปนี้

5.1 สรุปผลการวิจัย

ในงานวิจัยนี้นำเสนอการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง วิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax และทำนายความนิยมโดยใช้แบบจำลอง exponential โดยแบ่งช่วงเวลาการทำนายความนิยมออกเป็น 2 แบบคือ ใช้ข้อมูล 1 ชั่วโมงในการหาระยะทางระหว่างแต่ละโพสต์กับเทมเพลตและทำนายต่อไปอีก 6 ชั่วโมง แบบที่สองคือ ใช้ข้อมูล 2 ชั่วโมงในการหาระยะทางระหว่างโพสต์กับเทมเพลตและทำนายต่อไปอีก 5 ชั่วโมง ผลการวัดประสิทธิภาพของแต่ละแบบเป็นดังนี้

1. ค่า rmse ยิ่งมีค่าน้อยเข้าใกล้ 0 แสดงว่าประสิทธิภาพในการทำนายมีความแม่นยำมีความผิดพลาดน้อย ผลของค่า rmse แสดงในตารางที่ 10 และ 11 ชี้ให้เห็นว่าวิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax ให้ค่า rmse ที่น้อยที่สุดเมื่อเทียบผลของแต่ละสำนักข่าว วิธีการที่ให้ค่าน้อยรองลงมาคือ วิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง และตามด้วยทำนายความนิยมโดยใช้แบบจำลอง exponential

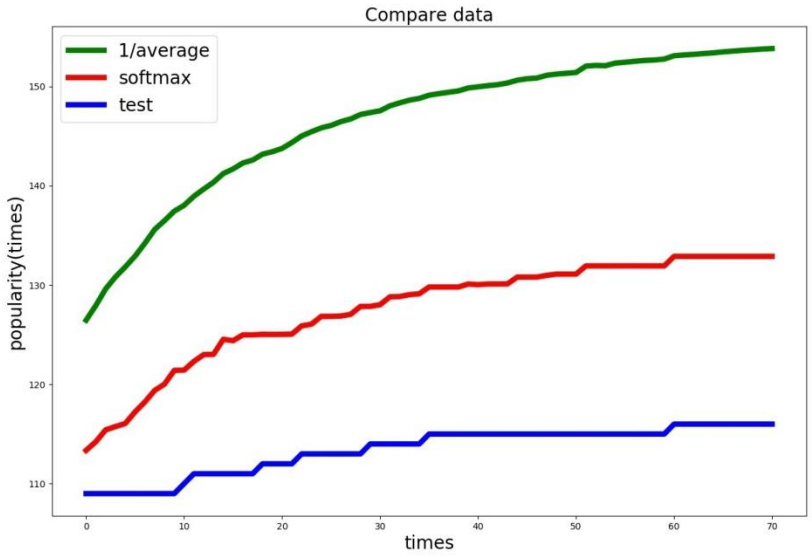
2. ค่า r2 score มีค่ามากที่สุดเท่ากับ 1 ถ้าค่าที่ได้เข้าใกล้ 1 แสดงว่าวิธีการทำนายนั้นแม่นยำ ผลของค่า r2 score แสดงในตารางที่ 10 และ 11 ชี้ให้เห็นว่าวิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax ให้ค่า r2 score เข้าใกล้ 1 มากที่สุด รองลงมาคือ การทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง และสุดท้ายคือ วิธีการทำนายความนิยมโดยใช้แบบจำลอง exponential

3. ค่า correlation coefficient มีค่ามากที่สุดเท่ากับ 1 ถ้าค่าที่ได้เข้าใกล้ 1 แสดงว่าวิธีการทำนายนั้นแม่นยำ ผลของค่า correlation coefficient แสดงในตารางที่ 10 และ 11 ชี้ให้เห็นว่าวิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง ให้ค่า

correlation coefficient เข้าใกล้ 1 มากที่สุด รองลงมาคือ วิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax และตามด้วยวิธีการทำนายความนิยมโดยใช้แบบจำลอง exponential

5.2 อภิปรายผล

จากการทดลองวิธีการทำนายความนิยมทวิตเตอร์ของแต่ละสำนักข่าว ด้วยวิธีทั้ง 3 แบบ จะพบว่าวิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax เป็นวิธีที่ให้ประสิทธิภาพในการทำนายแม่นยำที่สุด เมื่อเปรียบเทียบกับอีก 2 วิธี โดยพิจารณาจากค่า rmse และค่า r2 score แต่วิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ส่วนกลับของระยะทาง จะให้ค่า correlation coefficient ที่ดีที่สุด เนื่องมาจาก วิธีการใช้สัมประสิทธิ์ค่าเฉลี่ยส่วนกลับของระยะทาง จะมีค่าเฉลี่ยการทำนายมาจากทุกๆ เทมเพลต ซึ่งทำให้ profile ของการทำนายผล มีลักษณะคล้ายกับ profile ของค่า test จึงทำให้ค่า correlation coefficient ออกมาสูง ทั้งที่ค่าความผิดพลาดของการทำนายมีค่ามาก ส่วนวิธีการแบบค่าเฉลี่ย softmax จะมีค่าเฉลี่ยการทำนายมาจากเพียงบางเทมเพลต เนื่องจากค่าสัมประสิทธิ์ของ template บางเทมเพลต เข้าใกล้ศูนย์ จากวิธีการเฉลี่ย softmax ทำให้ profile ของการทำนายผลมีลักษณะแตกต่างไปจาก profile ของค่า test จึงทำให้ค่า correlation coefficient ออกมามีค่าน้อยกว่าแบบวิธีใช้สัมประสิทธิ์ค่าเฉลี่ยส่วนกลับของระยะทาง ทั้งที่ค่า rmse น้อยกว่า ดูตัวอย่างได้จากภาพประกอบ 10



ภาพประกอบ 10 แสดง profile ของแต่ละข้อมูล

จากจุดประสงค์ของงานวิจัยที่ต้องการทำนายความนิยมเป็นหลัก ทำให้ค่าที่นำมาพิจารณาประสิทธิภาพของการทำนายคือค่า rmse และค่า r2 score เมื่อดูจากผลการทดลองแล้ว วิธีการทำนายความนิยมด้วยการเฉลี่ยเทมเพลตแบบใช้ softmax เป็นวิธีที่ให้ประสิทธิภาพในการทำนายแม่นยำที่สุดในงานวิจัยนี้

5.3 ข้อเสนอแนะ

1. ในงานวิจัยนี้ใช้เวลาในการเก็บข้อมูลทวิตเตอร์จากสำนักข่าว 4 สำนัก เป็นเวลา 1 สัปดาห์ อาจจะเป็นเวลาที่สั้นเกินไป ถ้าเพิ่มเวลาในการเก็บข้อมูลให้นานขึ้น ค่าความถูกต้องในการทำนายอาจจะดีกว่านี้ เนื่องจากมีจำนวนโพสต์ที่นำมาจัดกลุ่มเทมเพลตมีมากขึ้น

2. ใช้ข้อมูลทวิตเตอร์จากแหล่งอื่นๆ ที่ไม่ใช้สำนักข่าว มาทดลองใช้ในการทำนายความนิยมด้วยวิธีที่งานวิจัยนี้นำเสมอเพื่อเปรียบเทียบผลลัพธ์ว่าเป็นไปในทิศทางเดียวกันหรือไม่

3. ในงานวิจัยนี้ใช้พารามิเตอร์จำนวน 2 ค่า คือการตอบโพสต์ (reply) และการโพสต์ซ้ำ (retweet) ในการทำนายความนิยม ซึ่งถ้าใช้พารามิเตอร์อื่นๆ เข้ามาใช้ในการทำนาย ผลการทำนายน่าจะมีความแม่นยำมากกว่านี้

4. ในงานวิจัยนี้ได้ทำการเปรียบประสิทธิภาพในการทำนายของแต่ละวิธีการ ทำให้รู้ถึงวิธีการทำนายที่มีประสิทธิภาพดีที่สุด แต่ยังไม่ได้ทำการเปรียบเทียบความแม่นยำในการทำนายของแต่ละสำนักข่าว ถ้ามีการเปรียบเทียบความแม่นยำในการทำนายของแต่ละสำนักข่าว จะทำให้เห็นผลการทำนายในมุมอื่นๆ ที่น่าสนใจเพื่อนำไปต่อยอดเป็นงานวิจัยเพิ่มเติมหรือทำไปใช้ประโยชน์ในด้านอื่นๆ ได้

บรรณานุกรม

1. Petitjean, F., A. Ketterlin, and P. Gançarski, *A global averaging method for dynamic time warping, with applications to clustering.* Pattern Recognition, 2011. **44**(3): p. 678-693.
2. Rath, T.M. and R. Manmatha. *Word image matching using dynamic time warping*. in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*. 2003. IEEE.
3. Wu, B. and H. Shen, *Analyzing and predicting news popularity on Twitter.* International Journal of Information Management, 2015. **35**(6): p. 702-711.
4. Zou, C.C., D. Towsley, and W. Gong, *Email virus propagation modeling and analysis.* Department of Electrical and Computer Engineering, Univ. Massachusetts, Amherst, Technical Report: TR-CSE-03-04, 2003.
5. Chai, T. and R.R. Draxler, *Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature.* Geoscientific model development, 2014. **7**(3): p. 1247-1250.
6. Skaza, J. and B. Blais, *Modeling the infectiousness of Twitter hashtags.* Physica A: Statistical Mechanics and its Applications, 2017. **465**: p. 289-296.
7. Izakian, H., W. Pedrycz, and I. Jamal, *Fuzzy clustering of time series data using dynamic time warping distance.* Engineering Applications of Artificial Intelligence, 2015. **39**: p. 235-244.
8. Doong, S.H. *Predicting Twitter Hashtags Popularity Level*. in *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 2016. IEEE.
9. Bandari, R., S. Asur, and B.A. Huberman, *The pulse of news in social media: Forecasting popularity.* ICWSM, 2012. **12**: p. 26-33.

ประวัติผู้เขียน

ชื่อ-สกุล	Rattasit Sermsai
วัน เดือน ปี เกิด	02 July 1987
สถานที่เกิด	Chanthaburi
วุฒิการศึกษา	Srinakharinwirot University
ที่อยู่ปัจจุบัน	37/9 M.14 T.Tungbenja A.Thamai Chanthaburi 22170