

แบบจำลองคืนหาหัวข้อจากความคิดเห็นขนาดเล็ก  
ที่มีเวลาเป็นปัจจัยประกอบ

**TOPIC MODELING FOR PUBLIC OPINION  
USING TIMESTAMP-BASED**

นาย สรวิศ ยินดีอนันต์

**SORAVIT YINDEEANANTA**

นาย อภิพล ด้วงเพียร

**APIPOL DUANGPHIAN**

ปริญญาบัณฑิตนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์เชิงธุรกิจ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 1 ปีการศึกษา 2565

แบบจำลองคืนหาหัวข้อจากความคิดเห็นขนาดเล็ก  
ที่มีเวลาเป็นปัจจัยประกอบ

TOPIC MODELING FOR PUBLIC OPINION  
USING TIMESTAMP-BASED

โดย

นาย สรวิศ ยินดีอนันต์

นาย อภิพล ด้วงเพียร

อาจารย์ที่ปรึกษา

ดร. นนท์ คณึงสุขเกยม

ปริญนานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์เชิงธุรกิจ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 1 ปีการศึกษา 2565

**TOPIC MODELING FOR PUBLIC OPINION  
USING TIMESTAMP-BASED**

**SORAVIT YINDEEANANTA**

**APIPOL DUANGPHIAN**

**A PROJECT SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENT FOR THE DEGREE OF  
BACHELOR OF SCIENCE PROGRAM IN DATA SCIENCE AND BUSINESS  
ANALYTICS  
SCHOOL OF INFORMATION TECHNOLOGY  
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG**

**1/2022**

**COPYRIGHT 2022**

**SCHOOL OF INFORMATION TECHNOLOGY**

**KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG**

# ใบรับรองปริญญาบัณฑ์ ประจำปีการศึกษา 2565

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง แบบจำลองค้นหาหัวข้อจากความคิดเห็นขนาดเล็กที่มีเวลาเป็นปัจจัย  
ประกอบ

**TOPIC MODELING FOR PUBLIC OPINION USING  
TIMESTAMP-BASED**

## ผู้จัดทำ

- นาย สรวิศ ยินดีอนันต์ รหัสนักศึกษา 62070277
- นาย อภิพล ด้วงเพียร รหัสนักศึกษา 62070285

  
อาจารย์ที่ปรึกษา  
( ดร. นนท์ คณึงสุขแคนਮ )

# ใบรับรองโครงการ (PROJECT)

เรื่อง

แบบจำลองค้นหาหัวข้อจากความคิดเห็นขนาดเล็ก  
ที่มีเวลาเป็นปัจจัยประกอบ

**TOPIC MODELING FOR PUBLIC OPINION USING  
TIMESTAMP-BASED**

นาย สรวิศ ยินดีอนันต์ รหัสนักศึกษา 62070277

นาย อภิพล ด้วงเพียร รหัสนักศึกษา 62070285

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด  
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ  
การศึกษาวิชาโครงการ หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาการข้อมูลและการวิเคราะห์เชิงธุรกิจ  
ภาคเรียนที่ 1 ปีการศึกษา 2565

สรวิศ ยินดีอนันต์

(นาย สรวิศ ยินดีอนันต์)

อภิพล ด้วงเพียร

(นาย อภิพล ด้วงเพียร)

หัวข้อโครงการ	แบบจำลองคืนหาหัวข้อจากความคิดเห็นขนาดเล็กที่มีเวลาเป็นปัจจัยประกอบ			
นักศึกษา	นาย สรวิศ ยินดีอนันต์	รหัสนักศึกษา	62070277	
	นาย อภิพล ดวงเพียร	รหัสนักศึกษา	62070285	
ปริญญา	วิทยาศาสตรบัณฑิต			
สาขาวิชา	วิทยาการข้อมูล และการวิเคราะห์เชิงธุรกิจ			
ปีการศึกษา	2565			
อาจารย์ที่ปรึกษา	ดร. นนท์ คงสุขเกغم			

## บทคัดย่อ

วิทยานิพนธ์ฉบับนี้มุ่งเน้นการนำเสนอการศึกษาการสร้างแบบจำลองหัวข้อ โดยเราได้ทำการศึกษาการสร้างแบบจำลองหัวข้อทั้งหมด 3 แบบ ได้แก่ Latent Dirichlet Allocation (LDA), Gibbs Sampling for Dirichlet Multinomial Mixture (GSDMM) และ Non-Negative Matrix Factorization (NMF) เพื่อเปรียบเทียบการหาหัวข้อซ่อนเร้นจากชุดข้อมูลสาระแตกต่าง ๆ ได้แก่ ชุดข้อมูล AG News และ Twitter COVID-19 และประเมินผลลัพธ์หัวข้อที่ได้จากแบบจำลองด้วยค่า Topic Coherence เพื่อทำการหาจำนวนหัวข้อที่เหมาะสมในแต่ละโมเดล ซึ่งผลลัพธ์ในการทดลองนี้พบว่าแบบจำลอง NMF ที่กำหนดจำนวนหัวข้อ 7 หัวข้อ ให้ผลลัพธ์ค่าความสอดคล้องดีที่สุดสำหรับข้อมูล AG News และแบบจำลอง GSDMM ที่กำหนดจำนวนหัวข้อ 3 หัวข้อให้ผลลัพธ์ค่าความสอดคล้องดีที่สุดเมื่อใช้ชุดข้อมูล Twitter COVID-19

ในงานวิจัยครั้งนี้ได้ไปผู้วิจัยจัดเก็บชุดข้อมูลข้อความขนาดสั้นจากทวิตเตอร์ด้วย Tweets API ของ Twitter Developers ด้วย Hashtag ยอดนิยมระหว่างวันที่ 3 – 9 พฤษภาคม พ.ศ. 2565 จำนวน 10 Tags มาใช้เป็นชุดข้อมูลแทนชุดข้อมูลที่จัดเก็บผ่านเว็บแอปพลิเคชันที่ผู้วิจัยจัดทำขึ้นมาเองที่มีข้อมูลไม่เพียงพอ เพื่อนำมาแบ่งเป็นส่วน เพื่อใช้สำหรับการทดลองเชิงเปรียบเทียบระหว่างการใช้ Sliding Window และ Expanding Window รวมกับการใช้และไม่ใช้ปัจจัยเวลาประกอบกับแบบจำลอง เพื่อนำมาเปรียบเทียบความแตกต่างของผลลัพธ์หัวข้อที่ได้ด้วยค่า Topic Coherence ซึ่งผลลัพธ์ในการ

ทดลองครั้งนี้พบว่าการทดลองแบบ Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบให้ผลลัพธ์ที่ดีที่สุด แต่ผลลัพธ์ของการทดลองแบบใช้เวลาเป็นปัจจัยประกอบยังได้เปรียบในด้านของผลลัพธ์หัวข้อที่สามารถตีความได้มากกว่า อีกทั้งยังประเมินความเชื่อมโยงได้ง่ายกว่าการทดลองแบบไม่ใช้เวลา

<b>PROJECT TITLE</b>	TOPIC MODELING FOR PUBLIC OPINION USING TIMESTAMP-BASED	
<b>STUDENT</b>	MR. SORAVIT YINDEEANANTA	STUDENT ID : 62070277
	MR. APIPOL DUANGPHIAN	STUDENT ID : 62070285
<b>DEGREE</b>	BACHELOR OF SCIENCE	
<b>PROGRAM</b>	DATA SCIENCE AND BUSINESS ANALYTICS	
<b>ACADEMIC YEAR</b>	2565	
<b>ADVISOR</b>	DR. NONT KANUNGSUKKASEM	

## ABSTRACT

This thesis focuses on presenting the study of Topic Modeling. We had preliminary studied three topic modeling including Latent Dirichlet Allocation (LDA), Gibbs Sampling for Dirichlet Multinomial Mixture (GSDMM), and Non-Negative Matrix Factorization (NMF) to compare their latent topics extracted from different public datasets, i.e., AG News and Twitter COVID-19. We evaluated those extracted topics by Topic Coherence measures which are also adopted to find the optimal number of topics for each topic modeling. The experiment results show that NMF topic modeling gives the best results for the AG News dataset with 7 topics and GSDMM topic modeling gives the best result for the Twitter COVID-19 dataset.

Afterwards, we collect a dataset of the small opinions from Twitter using Tweets API of Twitter Developers with 10 popular hashtags in Thailand instead of our insufficiently collected dataset from our web-application, to be divided into portions for the comparative study between using sliding window and expanding window combining with and without considering the time factor into the model. The results are then compared by their topic coherences. The experiment results show that the sliding window topic modeling without using the time factor gives the best results, but the results using the time factor still take advantage of more interpretable topic results than without taking the time factor.

## กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ได้ด้วยความกรุณาของอาจารย์ที่ปรึกษาปริญญาด้านนี้ ได้แก่ ดร. ชยานนท์ ทรัพย์อาภา และ ดร. นนท์ คันธสุขเกย์ม ที่ได้ให้คำปรึกษา ซึ่งแนะนำใน การศึกษาค้นคว้าเกี่ยวกับทวิจัยฉบับนี้จนสำเร็จลุล่วงด้วยดี

ขอขอบพระคุณคณาจารย์คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณ พหาราดgrade ฯ ท่านที่เคยสั่งสอน อบรมความรู้ และแนวคิดที่เป็นประโยชน์ ที่สามารถนำมาใช้ เพื่อพัฒนาต่อยอดในการทำงานในอนาคต

สรวิษ ยินดีอนันต์

อภิพล ด้วงเพียร

# สารบัญ

บทคัดย่อ .....	I
ABSTRACT .....	III
กิตติกรรมประกาศ.....	IV
สารบัญ .....	V
สารบัญรูป .....	VII
สารบัญตาราง .....	XVII
บทที่ 1.....	1
บทนำ.....	1
1.1 ที่มาและความสำคัญ.....	1
1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา .....	4
1.3 ขอบเขตการพัฒนาโครงการ .....	5
1.4 ขั้นตอนการดำเนินงาน .....	7
1.5 ประโยชน์ที่คาดว่าจะได้รับ .....	8
บทที่ 2.....	9
การทบทวนวรรณกรรมที่เกี่ยวข้อง .....	9
2.1 ทฤษฎีที่เกี่ยวข้อง .....	9
2.2 เทคนิค หรือเครื่องมือที่ใช้ .....	20
2.3 งานวิจัยที่เกี่ยวข้อง .....	27

## สารบัญ (ต่อ)

บทที่ 3.....	38
วิธีการดำเนินการวิจัย .....	38
3.1 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 1 .....	38
3.2 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 2 .....	43
บทที่ 4.....	50
ผลการดำเนินงานเบื้องต้น.....	50
4.1 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 1 .....	50
4.2 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 2 .....	71
บทที่ 5.....	101
สรุปผลการวิจัยและข้อเสนอแนะ .....	101
5.1 สรุปผลการวิจัย .....	101
5.2 ปัญหาและอุปสรรคในงานวิจัย .....	103
บรรณานุกรม.....	105
ภาคผนวก.....	108
ภาคผนวก ก. ผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธีที่แตกต่างกัน .....	109
ภาคผนวก ข. ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธีที่แตกต่างกัน .....	134

# สารบัญ

รูปที่ 2.1 ตัวอย่างของ Bag of Words .....	9
รูปที่ 2.2 การแบ่งคำแบบ unigrams .....	10
รูปที่ 2.3 การแบ่งคำแบบ bigrams .....	10
รูปที่ 2.4 ตัวอย่างการกระจายของคำและหัวข้อในเอกสาร Seeking Life's Bare (Genetic) Necessities ..	11
รูปที่ 2.5 PGM ของแบบจำลอง Latent Dirichlet Allocation .....	12
รูปที่ 2.6 PGM ของแบบจำลอง Dirichlet Multinomial Mixture .....	14
รูปที่ 2.7 การแยกองค์ประกอบของ NMF.....	15
รูปที่ 2.8 กระบวนการหาค่าความสอดคล้อง .....	14
รูปที่ 2.9 ตัวอย่างการแสดงผลลัพธ์ทั้งก่อน – หลังจากใช้ BERT ของ Google .....	19
รูปที่ 2.10 ตัวอย่างการใช้งาน BERT สำหรับงาน Sentiment Analysis.....	16
รูปที่ 2.11 ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลด้วย K-Means Clustering .....	20
รูปที่ 2.12 ตัวอย่างการตัดคำจากประโยค .....	21
รูปที่ 2.13 ตัวอย่างการลดครุ่ปคำโดยใช้ Stemmer แบบ Porter.....	21
รูปที่ 2.14 ตัวอย่างการใช้งาน thai2fit ในการแปลงประโยคให้อยู่ในรูปเวกเตอร์ .....	22
รูปที่ 2.15 ตัวอย่างการจัดการข้อมูลโดยการใช้ Pandas .....	24
รูปที่ 2.16 ตัวอย่างการใช้ UMAP ในการจัดกลุ่มข้อมูล .....	25
รูปที่ 2.17 ตัวอย่างผลลัพธ์ของมุนระหว่างเวกเตอร์ที่มีค่าเข้าใกล้ 0 องศา .....	26
รูปที่ 2.18 ตัวอย่างผลลัพธ์ของมุนระหว่างเวกเตอร์ที่มีค่าเข้าใกล้ 90 องศา .....	26
รูปที่ 2.19 การเปรียบเทียบประสิทธิภาพของแบบจำลองหัวข้อสำหรับข้อความสั้น .....	28
รูปที่ 2.20 PGM ของ Location Aware Topic Model (LATM) .....	29
รูปที่ 2.21 PGM ของ Topic Over Time (TOT).....	30
รูปที่ 2.22 ผลลัพธ์หัวข้อจากชุดข้อมูลข้อความจากอีเมลส่วนตัวแบบจำลอง TOT และ LDA.....	31

## สารบัญรูป (ต่อ)

รูปที่ 2.23 ผลลัพธ์หัวข้อจากชุดข้อมูลแต่งน้อยนายนายประจำปีแบบจำลอง TOT และ LDA .....	32
รูปที่ 2.24 PGM ของ Continuous Dynamic Topic Model .....	33
รูปที่ 2.25 ผลลัพธ์การวัดผลแบบจำลอง cDTM แต่ละแบบจำลอง .....	34
รูปที่ 2.26 การทำงานของแบบจำลอง BERTopic .....	35
รูปที่ 2.27 PGM ของแบบจำลอง Dynamic Topic Model .....	36
รูปที่ 2.28 ตัวอย่างผลลัพธ์หัวข้อจากแบบจำลอง Dynamic Topic Model .....	37
รูปที่ 2.29 การเปรียบเทียบแบบจำลองหัวข้ออื่นกับ DTM .....	37
รูปที่ 3.1 ตัวอย่างหน้าแอปพลิเคชันที่ใช้ในการเก็บข้อมูล .....	40
รูปที่ 3.2 การสร้างแบบจำลองหัวข้อด้วยวิธี Sliding Window .....	46
รูปที่ 3.3 การสร้างแบบจำลองหัวข้อด้วยวิธี Expanding Window .....	47
รูปที่ 4.1 ตัวอย่างชุดข้อมูล AG News .....	50
รูปที่ 4.2 ตัวอย่างชุดข้อมูล Twitter COVID-19 .....	51
รูปที่ 4.3 ตัวอย่างข้อความ Tweets .....	51
รูปที่ 4.4 ตัวอย่างข้อมูลที่เก็บลงฐานข้อมูล .....	51
รูปที่ 4.5 ตัวอย่างชุดข้อมูล AG News หลังผ่านการเตรียมข้อมูล .....	52
รูปที่ 4.6 ตัวอย่างชุดข้อมูล AG News หลังการทำ Bag-of-Words .....	53
รูปที่ 4.7 ผลลัพธ์จากแบบจำลอง LDA ด้วยชุดข้อมูล AG News .....	54
รูปที่ 4.8 ผลลัพธ์จากแบบจำลอง GSDMM ด้วยชุดข้อมูล AG News .....	55
รูปที่ 4.9 ผลลัพธ์จากแบบจำลอง NMF ด้วยชุดข้อมูล AG News .....	55
รูปที่ 4.10 ผลลัพธ์จากแบบจำลอง LDA ด้วยชุดข้อมูล Twitter COVID-19 .....	56
รูปที่ 4.11 ผลลัพธ์จากแบบจำลอง GSDMM ด้วยชุดข้อมูล Twitter COVID-19 .....	56
รูปที่ 4.12 ผลลัพธ์จากแบบจำลอง NMF ด้วยชุดข้อมูล Twitter COVID-19 .....	57

## สารบัญรูป (ต่อ)

<b>รูปที่ 4.13</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA ด้วยชุดข้อมูล AG News .....	59
<b>รูปที่ 4.14</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA ด้วยชุดข้อมูล Twitter COVID-19 .....	60
<b>รูปที่ 4.15</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง GSDMM ด้วยชุดข้อมูล AG News .....	61
<b>รูปที่ 4.16</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA ด้วยชุดข้อมูล Twitter COVID-19 .....	62
<b>รูปที่ 4.17</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง NMF ด้วยชุดข้อมูล AG News .....	63
<b>รูปที่ 4.18</b> กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง NMF ด้วยชุดข้อมูล Twitter COVID-19 .....	64
<b>รูปที่ 4.19</b> กราฟเปรียบเทียบค่าความสอดคล้อง U_MASS ของแต่ละแบบจำลองด้วยชุดข้อมูล AG News .....	65
<b>รูปที่ 4.20</b> กราฟเปรียบเทียบค่าความสอดคล้อง C_V ของแต่ละแบบจำลองด้วยชุดข้อมูล AG News ..	65
<b>รูปที่ 4.21</b> กราฟเปรียบเทียบค่าความสอดคล้อง C_UCI ของแต่ละแบบจำลองด้วยชุดข้อมูล AG News ..	66
<b>รูปที่ 4.22</b> กราฟเปรียบเทียบค่าความสอดคล้อง C_NPMI ของแต่ละแบบจำลองด้วยชุดข้อมูล AG News ..	66
<b>รูปที่ 4.23</b> กราฟเปรียบเทียบค่าความสอดคล้อง U_MASS ของแต่ละแบบจำลองด้วยชุดข้อมูล Twitter COVID-19 .....	67

## สารบัญรูป (ต่อ)

รูปที่ 4.24 กราฟเปรียบเทียบค่าความสอดคล้อง C_V ของแต่ละแบบจำลองด้วยชุดข้อมูล Twitter COVID-19.....	68
รูปที่ 4.25 กราฟเปรียบเทียบค่าความสอดคล้อง C_UCI ของแต่ละแบบจำลองด้วยชุดข้อมูล Twitter COVID-19.....	68
รูปที่ 4.26 กราฟเปรียบเทียบค่าความสอดคล้อง C_NPMI ของแต่ละแบบจำลองด้วยชุดข้อมูล Twitter COVID-19.....	69
รูปที่ 4.27 ตัวอย่างข้อมูลความคิดเห็นจาก Twitter.....	71
รูปที่ 4.28 ตัวอย่างข้อมูลความคิดเห็นหลังจากผ่านกระบวนการเตรียมข้อมูลแล้ว.....	71
รูปที่ 4.29 ตัวอย่างข้อมูลที่แปลงเป็นภาษาเตอร์ด้วย Sentence Transformer.....	72
รูปที่ 4.30 ตัวอย่างข้อมูลหลังจากลดมิติและเพิ่มปัจจัยเวลา.....	72
รูปที่ 4.31 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	74
รูปที่ 4.32 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	75
รูปที่ 4.33 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	76
รูปที่ 4.34 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	77
รูปที่ 4.35 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	78
รูปที่ 4.36 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	79

## สารบัญ (ต่อ)

รูปที่ 4.37 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	80
รูปที่ 4.38 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	81
รูปที่ 4.39 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	82
รูปที่ 4.40 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	83
รูปที่ 4.41 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	84
รูปที่ 4.42 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	85
รูปที่ 4.43 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	86
รูปที่ 4.44 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	87
รูปที่ 4.45 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	88
รูปที่ 4.46 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	89
รูปที่ 4.47 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนกภาพ Heatmap.....	91

## สารบัญรูป (ต่อ)

รูปที่ 4.48 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	92
รูปที่ 4.49 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	93
รูปที่ 4.50 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	94
รูปที่ 4.51 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ	96
รูปที่ 4.52 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ	97
รูปที่ 4.53 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ	98
รูปที่ 4.54 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ	99
รูปที่ ก.1 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)	110
รูปที่ ก.2 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)	111
รูปที่ ก.3 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)	112
รูปที่ ก.4 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)	113

## สารบัญรูป (ต่อ)

รูปที่ ก.5 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	114
รูปที่ ก.6 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	115
รูปที่ ก.7 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	116
รูปที่ ก.8 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	117
รูปที่ ก.9 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	118
รูปที่ ก.10 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	119
รูปที่ ก.11 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	120
รูปที่ ก.12 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	121
รูปที่ ก.13 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	122
รูปที่ ก.14 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	123
รูปที่ ก.15 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	124

## สารบัญรูป (ต่อ)

รูปที่ ก.16 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	125
รูปที่ ก.17 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	126
รูปที่ ก.18 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	127
รูปที่ ก.19 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	128
รูปที่ ก.20 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	129
รูปที่ ก.21 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	130
รูปที่ ก.22 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	131
รูปที่ ก.23 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1).....	132
รูปที่ ก.24 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2).....	133
รูปที่ ข.1 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap.....	135
รูปที่ ข.2 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap.....	136

## สารบัญรูป (ต่อ)

รูปที่ ข.3 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	137
รูปที่ ข.4 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	138
รูปที่ ข.5 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	139
รูปที่ ข.6 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	140
รูปที่ ข.7 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	141
รูปที่ ข.8 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	142
รูปที่ ข.9 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	143
รูปที่ ข.10 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	144
รูปที่ ข.11 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	145
รูปที่ ข.12 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap	146

## สารบัญตาราง

ตารางที่ 4.1 ค่าความสอดคล้องจากการคำนวณแต่ละวิธีของแบบจำลอง LDA, GSDMM และ NMF เมื่อใช้ชุดข้อมูล AG News.....	70
ตารางที่ 4.2 ค่าความสอดคล้องจากการคำนวณแต่ละวิธีของแบบจำลอง LDA, GSDMM และ NMF เมื่อใช้ชุดข้อมูล Twitter COVID-19 .....	70
ตารางที่ 4.3 สรุปผลลัพธ์การค้นหาหัวข้อจากการทดลองที่แตกต่างกัน .....	73
ตารางที่ 4.4 สรุปผลการทดลองการประเมินความเชื่อมโยงของหัวข้อด้วย Cosine Similarity .....	93
ตารางที่ 4.5 ผลลัพธ์ค่าเฉลี่ยของค่า Topic Coherences ที่ได้จากการประเมินผลลัพธ์หัวข้อ .....	114

## บทที่ 1

### บทนำ

#### 1.1 ที่มาและความสำคัญ

เมื่อก่อนเทคโนโลยียังไม่มีการใช้อย่างแพร่หลาย จึงทำให้การเก็บข้อมูลของผู้บริโภคของบริษัทส่วนใหญ่มักได้มาจากแบบสอบถาม การสังเกต และการวิจัยข้อมูล ซึ่งต้องใช้ระยะเวลาที่ยาวนานในการเก็บรวบรวมข้อมูลต่าง ๆ จากผู้บริโภค กว่าจะสามารถนำมาทำสรุป อาจทำให้พฤติกรรมความต้องการของผู้บริโภคเปลี่ยนไปแล้ว แต่ในทางกลับกันเมื่อเวลาผ่านไปเทคโนโลยีเริ่มถูกใช้กันอย่างแพร่หลายมากขึ้น ผู้คนส่วนใหญ่หันมาสเปล์ล์ออนไลน์กันมากขึ้น จึงทำให้การตลาดของหลาย ๆ บริษัทมีความยากขึ้น เนื่องจากการตลาดยุคใหม่นี้ต้องยึดผู้บริโภคเป็นหลัก นั่นจึงทำให้เกิดการติดตามความต้องการ ความรู้สึก และทัศนคติของผู้บริโภคที่มีต่อแบรนด์ และสินค้าต่าง ๆ บนสื่อออนไลน์ หรือที่เรียกว่าการฟังเสียงของผู้บริโภคบนสื่อสังคมออนไลน์ (Social Listening) [1]

การฟังเสียงของผู้บริโภคบนสื่อสังคมออนไลน์ คือ การเก็บข้อมูลของผู้บริโภคที่อยู่บนสื่อออนไลน์ อาทิเฟสบุ๊ค (Facebook), ทวิตเตอร์ (Twitter), อินสตราแกรม (Instagram) และ ยูทูป (YouTube) เพื่อให้เราสามารถทราบได้ว่าใครกำลังพูดถึงสินค้า และบริการของแบรนด์เราบ้าง รวมถึงผู้คนที่เข้ามาแสดงความคิดเห็นบนสื่อออนไลน์ แม้ว่าข้อความเหล่านั้นจะไม่ได้มีการกล่าวถึงโดยตรง [2] เช่น ทางบริษัท A เห็นการสนทนาระหว่างผู้คนบนสื่อออนไลน์ที่กำลังมองหาแนวทางที่ผลิตภัณฑ์หรือบริการของบริษัทสามารถให้กับพวกเขาได้ ทำให้บริษัทสามารถนำมาระบุแผนเพื่อตอบสนองความต้องการของผู้บริโภคตามแนวโน้มความคิดที่กำลังเป็นที่นิยม ไม่ว่าจะเป็นการปรับกลยุทธ์ทางการตลาด หรือปรับปรุงผลิตภัณฑ์ของบริษัท เพื่อที่จะสามารถตอบสนองความต้องการของผู้บริโภคได้อย่างเหมาะสม โดยข้อมูลที่เราสามารถนำมาวิเคราะห์ได้มีมากmany เช่น ข้อมูลบริษัทหรือแบรนด์ของเรา, ข้อมูลสินค้าและบริการ, ข้อมูลความคิดเห็นจากผู้มีอิทธิพลที่มีอิทธิพลต่อความคิดและพฤติกรรมของผู้บริโภค (Key Opinion Leader), ข้อมูลบริษัทหรือแบรนด์ของคู่แข่ง และแบรนด์ที่กำลังถูกพูดถึง

เป็นต้น ดังนั้น การฟังเสียงของผู้บริโภคบนสื่อสังคมออนไลน์จึงเป็นสิ่งที่สามารถสร้างโอกาสให้กับบริษัท นอกจากนี้องค์กรภาครัฐ หรือหน่วยงานต่าง ๆ ยังสามารถนำเทคนิคการฟังเสียงของผู้บริโภคบนสื่อสังคมออนไลน์ในการตรวจจับแนวความคิด และปัญหาที่เกิดขึ้นในปัจจุบัน เพื่อนำมาวางแผนในการพัฒนาโครงการสาธารณะ ให้ประชาชนในประเทศมีคุณภาพชีวิตที่ดีขึ้น

เนื่องจากความคิดเห็นของผู้คนบนสื่อออนไลน์นั้น เกิดขึ้นต่างเวลา และสถานที่ ซึ่งแพลตฟอร์มนั้นไม่ได้เปิดเผยถึงตำแหน่งของผู้ใช้งานอย่างละเอียด [3] จึงอาจมีการปลอมแปลงข้อมูลเกิดขึ้นได้ นอกจากนี้ยังมีมิติของข้อมูล และความคิดเห็นในปริมาณมาก จึงทำให้ยากต่อการสรุปหัวข้อที่ถูกพูดถึงภายในชุดข้อความนั้นด้วยความสามารถของมนุษย์ ซึ่งในชุดข้อความนั้นอาจไม่ได้มีเพียงหัวข้อเดียว นั่นจึงเป็นเหตุผลที่ว่าทำไมเราถึงต้องทำการสกัดหัวข้อซ่อนเร้นที่ซ่อนอยู่ในชุดข้อความด้วยอัลกอริทึมประเภทการสร้างแบบจำลองหัวข้อ

การสกัดหัวข้อจำลองเป็นที่จับตาของสังคมอย่างมาก เช่น ความนิยมบนทวิตเตอร์ (Twitter) สามารถทำให้ผู้อ่านทราบถึงความต้องการ หรือสิ่งที่กำลังเป็นที่สนใจของสังคมในการอกรอบระยะเวลา ณ ขณะนั้น แต่การสกัดหัวข้อนี้เป็นเพียงการคัดกรองข้อความในช่วงเวลาที่ใกล้เคียงกัน และไม่ได้นำตำแหน่งของผู้เผยแพร่ข้อมูลมาพิจารณาด้วย ทางผู้จัดทำโครงงานจึงมีแนวคิดที่จะสร้างแบบจำลองในการสกัดหัวข้อ (Topic Modeling) โดยนำเวลา และสถานที่เข้ามาเป็นปัจจัยประกอบ ซึ่งอาจจะทำให้ค้นพบหัวข้อที่ตรงกับความเป็นจริงมากขึ้น เช่น หากมีข้อความว่า “ถนนมีดมากเลย” ในพื้นที่สาธารณะ A เป็นประจำในช่วงเวลา 2 ทุ่ม ในส่วนของระบบการนำเสนอความนิยมจะไม่สกัดหัวข้อนี้ออกจากข้อความจะมีปริมาณมากพอ ซึ่งถ้าหากพิจารณาจากช่วงเวลา และสถานที่แล้ว การสกัดหัวข้อควรจะแสดงเป็น “ถนนมีดในพื้นที่ A” เมื่อได้หัวข้อดังกล่าวแล้วผู้ที่นำข้อมูลไปวิเคราะห์ ก็จะทราบถึงปัญหาสังคมที่เกิดขึ้นในพื้นที่ และอาจจะนำไปสู่การเปลี่ยนแปลงทางสังคมต่อไปในอนาคต

ในการแก้ไขปัญหาข้างต้น เราได้มีการนำเสนอเทคนิคที่เป็นที่นิยมอย่าง Latent Dirichlet Allocation (LDA) [4], Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) [5] และ Non-Negative Matrix Factorization (NMF) [6] มาใช้ในการสร้างแบบจำลองหัวข้อ (Topic Modeling) สำหรับการสรุปสาระสำคัญจากข้อความทั้งข้อความปริมาณปานกลางและข้อความปริมาณน้อย [7], [8]

รวมถึงมีการนำเสนอแนวทางการสร้างแบบจำลองหัวข้อที่มีการคำนึงถึงปัจจัยเวลา [9] และสถานที่ [10] ซึ่งมีบทบาทสำคัญสำหรับความสามารถในการจำแนกหัวข้อจากข้อมูลที่เกิดจากปัจจัยที่แตกต่างกัน

ด้วยปัจจัยที่กล่าวข้างต้น ทางผู้จัดทำโครงการจึงได้ทำการวางแผนการศึกษาในวิชาโครงการ 1 ออกเป็น 2 ส่วน คือ 1) ศึกษาการนำเทคนิคที่นิยมต่อการใช้งานทั้ง Latent Dirichlet Allocation (LDA), Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) และ Non-Negative Matrix Factorization (NMF) โดยใช้ข้อมูลทุกดิจิทัล จากแหล่งข้อมูลที่ต่างกันมาใช้ในการจัดทำแบบจำลองหัวข้อ (Topic Modeling) และ 2) ทำการเก็บข้อมูลโดยสร้างเว็บแอปพลิเคชันที่ทางผู้จัดทำการออกแบบขึ้นเพื่อการรวบรวมข้อมูลความคิดเห็นแบบเสียง เวลา สถานที่ (ประกอบด้วย ละติจูด และลองจิจูด) และการตอบสนองทางอารมณ์ต่อความคิดเห็นเพื่อนำไปใช้เป็นชุดข้อมูลในการสร้างแบบจำลองที่คำนึงถึงเวลา และสถานที่ต่อไปในวิชาโครงการ 2

เนื่องจากโรคระบาดไวรัส COVID-19 และฝุ่น PM 2.5 ที่กลับมาเมื่อค่าสูงอีกครั้งในระยะหลัง ทำให้ผู้วิจัยไม่สามารถออกไปจัดเก็บข้อมูลคนอื่นจากนอกสถานที่ได้ จึงทำได้เพียงส่ง URL ของเว็บแอปพลิเคชันให้กับคนใกล้ชิด และทำการกระจายผ่านสื่อสังคมออนไลน์ของผู้จัดทำเท่านั้น จึงส่งผลให้ข้อมูลที่เก็บได้ผ่านเว็บแอปพลิเคชันมิໄม่เพียงพอต่อการนำมาใช้งานต่อในวิชาโครงการ 2 จึงทำให้ทางผู้จัดทำโครงการได้ทำการวางแผนการศึกษาออกแบบ 2 ส่วน คือ 1) จัดเตรียมข้อมูลข้อความขนาดสั้นจากสื่อออนไลน์ด้วย Hashtag ยอดนิยมจำนวน 10 Tags มาใช้เป็นชุดข้อมูลสำหรับการจัดทำแบบจำลองหัวข้อ (Topic Modeling) แทนชุดข้อมูลดังกล่าว ทั้งแบบ Sliding Window และ Expanding Window โดยที่มีทั้งใช้ปัจจัยเวลาประกอบ และแบบไม่นำปัจจัยเวลาเข้ามาประกอบ และ 2) เปรียบเทียบความแตกต่างของผลลัพธ์หัวข้อที่ได้จากการทดลองทั้ง 4 แบบด้วยค่า Topic Coherence

## 1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

### 1.2.1 การทดลองที่ 1

#### 1.2.1.1 ความมุ่งหมาย

เพื่อรับร่วมข้อมูลความคิดเห็นแบบเสียง เวลา สถานที่ (ประกอบด้วย ละติจูด และ ลองจิจูด) และการตอบสนองทางอารมณ์ต่อความคิดเห็น เพื่อนำไปใช้เป็นชุดข้อมูลในการสร้างแบบจำลองหัวข้อ (Topic Modeling) ที่มีเวลา และสถานที่ร่วมในการวิเคราะห์เพื่อสรุปหัวข้อ และสาระสำคัญของข้อมูล ทั้งนี้ข้อมูลการตอบสนองทางอารมณ์ต่อความคิดเห็นนั้นจะทำการรวบรวมไว้เพื่อการวิจัยต่อยอดในอนาคตซึ่งไม่อยู่ในขอบเขตงานของโครงการนี้

#### 1.2.1.2 วัตถุประสงค์

- 1) เพื่อศึกษาการสร้างแบบจำลองหัวข้อที่มีกลไกการทำงานที่แตกต่างกัน
- 2) เพื่อการทดลองเชิงปริยบเทียบระหว่างแบบจำลอง Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF) และ Gibb Sampling for Dirichlet Multinomial Mixture (GSDMM)
- 3) เพื่อการทดลองเชิงปริยบเทียบระหว่างแบบจำลอง LDA, NMF และ GSDMM เมื่อใช้ชุดข้อมูลแตกต่างกัน
- 4) เพื่อการทดลองเชิงปริยบเทียบระหว่างแบบจำลอง LDA, NMF และ GSDMM เมื่อใช้ชุดข้อมูลแตกต่างกัน

### 1.2.2 การทดลองที่ 2

#### 1.2.2.1 ความมุ่งหมาย

เพื่อรับร่วมข้อมูลความคิดเห็นผ่าน Hashtag ที่เป็นที่นิยมจากสื่อออนไลน์อย่าง ทวิตเตอร์ (Twitter) เป็นจำนวนมาก เพื่อนำไปใช้เป็นชุดข้อมูลในการสร้างแบบจำลองหัวข้อ (Topic Modeling) ด้วยวิธี Sliding Window และ Expanding Window ทั้งแบบที่มีปัจจัยเวลาประกอบ และไม่มีปัจจัยเวลาเข้ามาประกอบ แทนข้อมูลที่เก็บรวบรวมผ่านเว็บแอปพลิเคชันที่มีข้อมูลไม่เพียงพอ

### 1.2.2.2 วัตถุประสงค์

- 1) เพื่อศึกษาการนำวิธี Sliding Window และ Expanding Window มาใช้ในการสร้างแบบจำลองหัวข้อ
- 2) เพื่อศึกษาการสร้างแบบจำลองหัวข้อทั้งแบบที่ใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ
- 3) เพื่อเปรียบเทียบผลลัพธ์หัวข้อ ความเชื่อมโยงหรือการเปลี่ยนไปของหัวข้อตลอดช่วงเวลาที่ทำการทดลอง และความสามารถในการตีความของหัวข้อจากการสร้างแบบจำลองด้วยวิธีระหว่าง Sliding Window และ Expanding Window
- 4) เพื่อเปรียบเทียบผลลัพธ์หัวข้อ ความเชื่อมโยงหรือการเปลี่ยนไปของหัวข้อตลอดช่วงเวลาที่ทำการทดลอง และความสามารถในการตีความของหัวข้อจากการสร้างแบบจำลองด้วยวิธีระหว่าง Sliding Window และ Expanding Window เมื่อใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ

## 1.3 ขอบเขตการพัฒนาโครงการ

ขอบเขตของงานวิจัยฉบับนี้ออกเป็น 4 ส่วน คือ การศึกษาการใช้งานและการสร้างแบบจำลองการสร้างหัวข้อ การเก็บรวบรวมข้อมูล การศึกษาและหาวิธีประเมินประสิทธิภาพแบบจำลองการสร้างหัวข้อ และขอบเขตของการออกแบบการทดลอง

### 1.3.1 การศึกษาการใช้งานและการสร้างแบบจำลองการสร้างหัวข้อ

โดยการสร้างแบบจำลองหัวข้อที่ทางผู้จัดทำโครงการได้นำมาศึกษาการใช้งาน คือ Latent Dirichlet Allocation (LDA), Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) และ Non-Negative Matrix Factorization (NMF) และการสร้างแบบจำลองหัวข้อด้วยแบบจำลองประเภท K-Mean Clustering ด้วยวิธีที่แตกต่างกัน ซึ่งประกอบด้วย Sliding Window และ Expanding Window ด้วยการทดลองทั้งแบบใช้และไม่ใช้เวลาเป็นปัจจัยประกอบสำหรับการสร้างแบบจำลอง

### 1.3.2 การเก็บรวบรวมข้อมูล

ทางผู้จัดทำโครงการได้นำข้อมูลสาระน่าอ่านอย่าง AG News และ Twitter COVID-19 มาใช้เพื่อสร้างแบบจำลองหัวข้อ อีกทั้งยังสร้างเว็บแอปพลิเคชันเพื่อรับรวมข้อมูลความคิดเห็น เวลา สถานที่ และการตอบสนองทางอารมณ์ต่อความคิดเห็นเพื่อนำไปใช้เป็นอีกหนึ่งชุดข้อมูล รวมถึงมีการดึงข้อมูลความคิดเห็นจาก Twitter ด้วย Hashtag ที่เป็นที่นิยมเพื่อนำไปใช้ในการสร้างและพัฒนาแบบจำลองต่อไปในอนาคต

### 1.3.3 การศึกษาและหาวิธีประเมินประสิทธิภาพแบบจำลองการสร้างหัวข้อ

ศึกษาการประเมินประสิทธิภาพของแบบจำลองการสร้างหัวข้อแบบ Topic Coherence ที่ประกอบด้วยวิธีการคำนวณหาค่าโดยรูปแบบ และศึกษาแนวทางการประยุกต์ใช้งาน หากแบบจำลองมีการนำเวลาเข้ามาเป็นส่วนประกอบร่วมกับข้อมูลความ

### 1.3.4 ขอบเขตของการออกแบบการทดลอง

- 1) เปรียบเทียบความแตกต่างของการสร้างแบบจำลองหัวข้อ LDA, NMF และ GSDMM
- 2) เปรียบเทียบความแตกต่างของการสร้างแบบจำลองหัวข้อ LDA, NMF และ GSDMM เมื่อใช้ชุดข้อมูลแตกต่างกัน
- 3) เปรียบเทียบความแตกต่างเชิงประสิทธิภาพโดยการหาค่า Topic Coherences ของผลลัพธ์หัวข้อจากการสร้างแบบจำลองหัวข้อ LDA, NMF และ GSDMM เมื่อใช้ชุดข้อมูลแตกต่างกัน
- 4) เปรียบเทียบความแตกต่างของผลลัพธ์หัวข้อที่ได้ระหว่างทดลองแบบ Sliding Window และ Expanding Window และเมื่อใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ
- 5) เปรียบเทียบความแตกต่างของความซึ่อมโยงของผลลัพธ์หัวข้อที่ได้จากการหาค่า Cosine Similarity ซึ่งได้จากการทดลองแบบ Sliding Window และ Expanding Window และเมื่อใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ

- 6) เปรียบเทียบความแตกต่างของความสามารถในการตีความของผลลัพธ์หัวข้อในแต่ละวันของช่วงเวลาที่พิจารณาที่ได้จากการหาค่า Topic Coherences ซึ่งได้จากการทดลองแบบ Sliding Window และ Expanding Window และเมื่อใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ

## 1.4 ขั้นตอนการดำเนินงาน

### 1.4.1 ศึกษาการสร้างแบบจำลองหัวข้อ

1. ศึกษาวิธีการเตรียมข้อมูลหัวข้อความที่เหมาะสมสำหรับใช้สร้างแบบจำลองหัวข้อ
2. ศึกษาการสร้างแบบจำลองหัวข้อแต่ละรูปแบบ รวมถึงการสร้างแบบจำลองหัวข้อแบบคำนึงถึงปัจจัยสถานที่และเวลา จากบทวิจัยที่มีอยู่แล้ว

### 1.4.2 การเก็บข้อมูลเพื่อนำมาใช้เป็นข้อมูลในการวิจัย

ทางผู้จัดทำโครงการจะมีการเก็บข้อมูลทุกชิ้นจากเว็บไซต์ AG News, COVID-19 Twitter และเก็บข้อมูลหัวข้อความคิดเห็นจาก Twitter โดยข้อมูลที่เก็บมานั้นมี Attribute ดังนี้

1. ข้อมูลหัวข้อความคิดเห็น
2. ข้อมูลเวลาที่แสดงความคิดเห็น
3. ข้อมูล Hashtag ที่ทำการดึงข้อมูล

### 1.4.3 ฝึกฝนการใช้โมเดล

1. ฝึกฝนการใช้ Toolkit ในการประมวลภาษาบนภาษาไทยให้อยู่ในรูปภาษาต่างๆ อาทิเช่น Natural Language Toolkit (NLTK) และ PyThaiNLP เป็นต้น
2. ฝึกฝนการใช้ Gensim ในการสร้างแบบจำลองหัวข้อ และประมวลผลภาษาธรรมชาติ

### 1.4.4 ทดสอบแบบจำลองการสร้างหัวข้อ และประเมินประสิทธิภาพของแบบจำลอง

### 1.4.5 สรุปผลการเปรียบเทียบแบบจำลองหัวข้อที่ทดลองสร้างด้วยปัจจัยที่แตกต่างกัน

## 1.5 ประโยชน์ที่คาดว่าจะได้รับ

1.5.1 ได้รับความรู้ และความเข้าใจด้านการวิเคราะห์ข้อความและการสร้างแบบจำลองหัวข้อ (Topic Modeling)

1.5.2 เข้าใจวิธีการทำงานของแบบจำลองการสร้างหัวข้อแบบ LDA, GSDMM และ NMF

1.5.3 ได้รับความรู้ และความเข้าใจจากการศึกษาหลักการประเมินประสิทธิภาพของแบบจำลอง โดยใช้ Topic Coherence

1.5.4 สามารถสำรวจอารมณ์ ความคิดเห็น และแนวโน้มพฤติกรรมของบุคคลในแต่ละสถานที่ และช่วงเวลาได้อย่างรวดเร็ว แม่นยำ และสะท้อนสนาญมากขึ้น

1.5.5 ทราบถึงแนวคิดในการพัฒนาแบบจำลองหัวข้อด้วยวิธีที่ต่างจากแบบจำลองหัวข้อแบบดั้งเดิม

## บทที่ 2

### การทบทวนวรรณกรรมที่เกี่ยวข้อง

#### 2.1 ทฤษฎีที่เกี่ยวข้อง

##### 2.1.1 Bag of Words

เป็นหนึ่งในทฤษฎีพื้นฐานในเรื่องของการประมวลภาษาธรรมชาติที่ใช้สำหรับการแบ่งคำในชุดข้อความให้อยู่ในรูปแบบเวกเตอร์ หรือที่เรียกว่า กระเบื้องของคำ โดย Bag of Words จะนำเสนอเมทริกซ์ของคำที่ปรากฏอยู่ในชุดข้อความ และจะแสดงผลลัพธ์เป็นการนับความถี่ของคำที่เกิดขึ้นในข้อความนั้น โดยการแปลงข้อมูลข้อความให้เป็นเวกเตอร์ด้วยวิธี Bag of Words นั้นจะไม่คำนึงถึงลำดับ ไวยากรณ์ และความหมายของคำ ยกตัวอย่างเช่น ชุดข้อความทั้ง 3 ข้อความอย่าง the cat sat, the cat sat in the hat และ the cat with the hat จะสามารถแบ่งคำที่แตกต่างกันออกมายได้ทั้งหมด 6 คำ ได้แก่ the, cat, sat, in, hat และ with และจะแสดงความถี่ของคำแต่ละคำตามที่ปรากฏ ดังรูปที่ 2.1

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

รูปที่ 2.1 ตัวอย่างของ Bag of Words

อีกทั้งยังสามารถกำหนดจำนวนคำที่ต้องการแบ่งได้ เช่น กัน อย่าง ข้อความ This is a sentence. หากเรากำหนดจำนวนการแบ่งคำไว้จำนวน 1 คำ (unigram) จะสามารถแบ่งคำออกเป็น This, is, a, sentence ดังรูปที่ 2.2 หรือหากกำหนดไว้ที่จำนวน 2 คำ (bigrams) จะสามารถแบ่งคำออกมาเป็น This is, is a, a sentence ดังรูปที่ 2.3



รูปที่ 2.2 การแบ่งคำแบบ unigrams



รูปที่ 2.3 การแบ่งคำแบบ bigrams

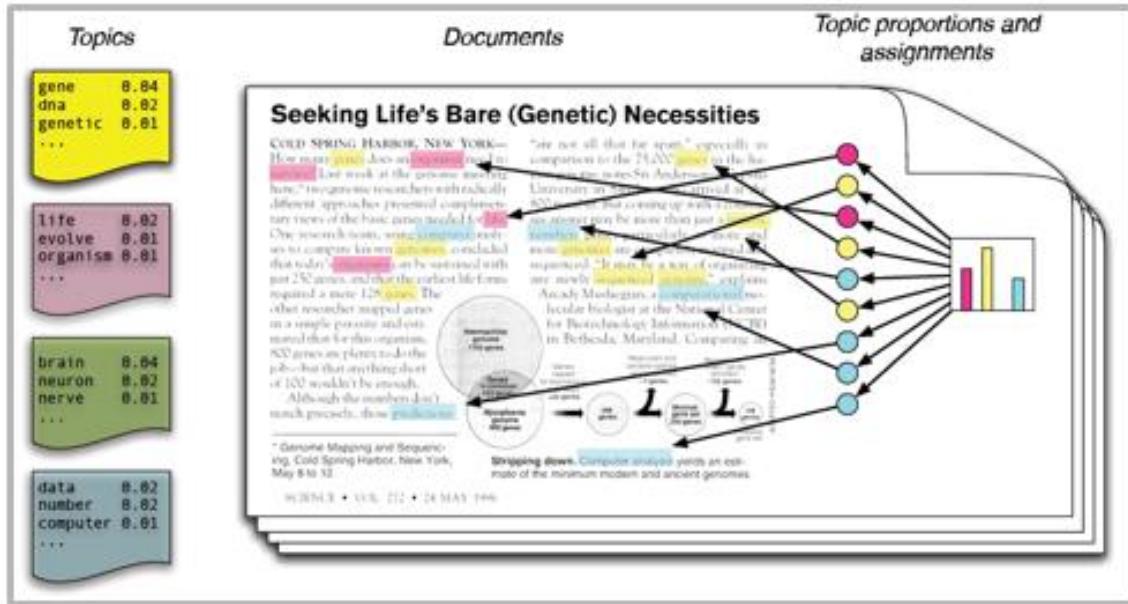
### 2.1.2 การสร้างแบบจำลองหัวข้อ (Topic Modeling)

การสร้างแบบจำลองหัวข้อ (Topic Modeling) เป็นการสร้างแบบจำลองความน่าจะเป็น (Probabilistic Modeling) เพื่อใช้ในการหาหัวข้อซ่อนเร้นในเอกสารจากชุดข้อมูลเอกสารทั้งหมด โดยทฤษฎีนี้สามารถทำให้เรารู้ความหมาย หรือเนื้อหาในภาพรวมของเอกสารชุดหนึ่ง ว่ากำลังกล่าวถึงอะไร โดยไม่ต้องใช้ทรัพยากรมนุษย์ อีกทั้งยังประหยัดเวลาในการหาหัวข้อจากชุดเอกสารจำนวนมากอีกด้วย

### 2.1.3 Latent Dirichlet Allocation (LDA)

เป็นแบบจำลองความน่าจะเป็น (Probabilistic Modeling) ตัวหนึ่งที่ใช้ในการสร้างแบบจำลองหัวข้อ (Topic Modelling) หรือการหาหัวข้อที่ซ่อนอยู่ภายในเอกสาร ซึ่งเป็นสิ่งที่เราต้องการดึงออกมาจากเอกสาร แต่มันเป็นหัวข้อที่คอมพิวเตอร์ไม่เห็นมันตรง ๆ ซึ่งมันจะทำการดึงเอาคำที่สามารถเป็นหัวข้อได้ภายในเอกสารแยกออกจากเป็นกลุ่ม ๆ โดยเราจะต้องเป็นคนกำหนดเองว่ากลุ่มนี้เป็นกลุ่มคำที่กำลังถือถึงอะไร เช่น กลุ่มคำที่ 1 ที่ประกอบด้วยคำว่า sport,

team, referee และ player กำลังกล่าวถึงเรื่องกีฬา กลุ่มคำที่ 2 ที่ประกอบด้วยคำว่า computer, hardware และ keyboard กำลังกล่าวถึงเรื่องคอมพิวเตอร์ เป็นต้น

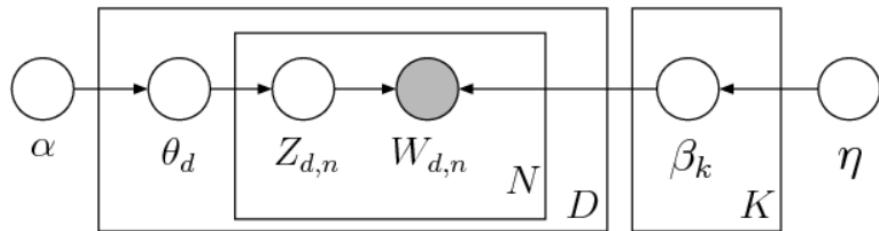


รูปที่ 2.4 ตัวอย่างการกระจายของคำและหัวข้อในเอกสาร Seeking Life's Bare (Genetic) Necessities

จากรูปที่ 2.4 จะสามารถอธิบายแนวคิดของ LDA ได้ว่าในส่วนของ Topics หรือส่วนทางด้านซ้ายของภาพ จะบอกถึงหัวข้อที่ซ้อนเร้นภายในเอกสารนี้ ซึ่งเราสามารถกำหนดจำนวนหัวข้อเองได้ด้วยการกำหนด Hyperparameter ไว้ล่วงหน้าสำหรับการค้นหาหัวข้อ ต่อมาในส่วนของ Topic Proportion and Assignments หรือแผนภูมิแท่งทางด้านขวา กล่าวได้ว่าในแต่ละเอกสารมีแต่ละหัวข้อ (Topic) หรือหัวข้อจากทางด้านซ้ายอยู่ด้วยความน่าจะเป็นเท่าไร โดยการกระจายของความน่าจะเป็น (Probability Distribution) นี้เป็นการกระจายแบบไม่ต่อเนื่อง และเมื่อได้ผลลัพธ์ออกมาแล้ว จะเป็นหน้าที่ของเราที่จะเป็นคนกำหนดต่อว่าหัวข้อนี้เกี่ยวกับอะไร อย่างเช่นกลุ่ม Topic สีเหลืองที่ประกอบด้วยคำว่า gene, dna และ genetics จะต้องมีความเกี่ยวข้องกับ Genetics หรือแม้แต่กลุ่ม Topic สีฟ้าที่ประกอบด้วยคำว่า data, number และ computer จะต้องมีความเกี่ยวข้องกับ Computer Science เป็นต้น

แบบจำลอง LDA ถูกพัฒนาโดย Blei et al. [4] ด้วยการอิงความรู้มาจากการจำลองที่ถูกคิดกันมาก่อนหน้าอย่าง Naive Bayes, Mixture of Unigram, และ pLSA โดยแบบจำลอง LDA มีสมมติฐานทั้งหมด 3 ข้อ คือ 1) ลำดับของคำไม่มีความสำคัญ 2) ลำดับของเอกสารไม่มีความสำคัญ และ 3) จำนวนหัวข้อเป็นค่าคงที่ที่ถูกกำหนดไว้ก่อนหน้าแล้ว

ซึ่งจากสมมติฐานดังกล่าว Blei et al. จึงได้เสนอ Probabilistic Graphical Model ของ LDA ไว้ดังรูปที่ 2.5



รูปที่ 2.5 PGM ของแบบจำลอง Latent Dirichlet Allocation

จากรูปที่ 2.5 สามารถอธิบายความหมายของตัวแปรได้ดังนี้

$\alpha$  คือ ตัวแปรควบคุมการกระจายตัวของ  $\theta$

$\theta_d$  คือ การแจกแจงหัวข้อในแต่ละเอกสาร  $d$

$Z_{d,n}$  คือ หัวข้อของคำที่  $n$  ในเอกสาร  $d$

$W_{d,n}$  คือ คำที่ปรากฏในเอกสาร  $d$  ทั้งหมด  $n$  คำ

$\beta_k$  คือ การแจกแจงของคำในแต่ละหัวข้อ  $k$

$\eta$  คือ ตัวแปรควบคุมการกระจายตัวของ  $\beta$

และสามารถนำมารีเขียนเป็นสมการได้ดังนี้

$$P(\beta, \theta, z, w | \alpha, \eta) =$$

$$\prod_{k=1}^K P(\beta_k | \eta) \prod_{d=1}^D [P(\theta_d | \alpha) (\prod_{n=1}^{N_d} P(z_{d,n} | \theta_d) P(w_{d,n} | \beta, z_{d,n}))] \quad (2.1)$$

เมื่อได้สมการแล้วจึงนำมาฝึกสอน และปรับค่าตัวแปรภายในแบบจำลองหัวข้อ LDA ซึ่งการประเมินความน่าจะเป็นสามารถทำได้โดยใช้อัลกอริทึมในการอนุมานได้หลายอัลกอริทึม เช่น Variational Approximation, Expectation Propagation, Laplace Approximation และ Gibbs Sampling หรือ Markov Chain Monte Carlo (MCMC)

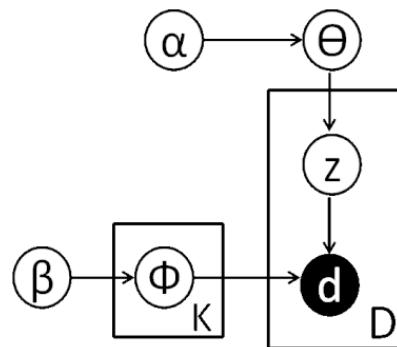
แบบจำลอง LDA เป็นแบบจำลองแบบ Generative ซึ่งหมายความว่าเมื่อได้เรียนรู้ความน่าจะเป็นจากชุดเอกสารแล้ว แบบจำลองจะสามารถสร้างเอกสารใหม่ได้จากการสุ่ม Topic และสุ่มคำจากใน Topic นั้น และกระบวนการสร้างเอกสารเหล่านี้จะเรียกว่า Generative Process สามารถอธิบายได้ดังนี้

1. ในแต่ละ Topic ( $k$ ) จากทั้งหมด  $K$  topic(s) จะสุ่ม  $\beta_k$  จาก Dirichlet Distribution ที่ควบคุมการกระจายตัว โดย  $\eta$
2. ในแต่ละ Document ( $d$ ) จากทั้งหมด  $D$  document(s) สุ่ม  $\theta_d$  จาก Dirichlet Distribution ที่ควบคุมการกระจายตัวโดย  $\alpha$
3. แต่ละคำ  $W$  จากคำทั้งหมด  $N$  คำ ในชุดเอกสาร  $d$ 
  - a. สุ่ม Topic  $Z_{d,n}$  มาจาก Multinomial Distribution  $\theta_d$
  - b. สุ่มคำ  $W_{d,n}$  มาจาก Multinomial Distribution  $\beta_{Z_{d,n}}$

#### **2.1.4 Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)**

Yin and Wang [5] เสนอการนำอัลกอริทึม Gibbs Sampling ไปใช้สำหรับการสร้างแบบจำลอง Dirichlet Multinomial Mixture ในการจัดกลุ่มข้อมูลของข้อความขนาดสันนิhood โดยผู้จัดทำหนังสือ Movie Group Process ซึ่งมีกระบวนการแบบเดียวกับ GSDMM สำหรับใช้ในการอธิบายลักษณะของแบบจำลอง GSDMM เพื่อให้สามารถทำความเข้าใจแบบจำลองได้ง่ายยิ่งขึ้น โดยสามารถสรุปกระบวนการดังกล่าวและความเชื่อมโยงกับการจัดกลุ่มข้อมูลข้อความขนาดสันนิhood ได้ด้วยตัวอย่างเชิงเปรียบเทียบ ดังนี้

- ให้นักเรียนทั้งหมด D คน (ชุดเอกสารทั้งหมด) นั่งโดยจำนวน K โดยแบบสุ่ม
  - โดยที่นักเรียนแต่ละคน (เอกสารแต่ละฉบับ) ลูกขอให้เขียนรายการภพยนตร์ที่ตนเองสนใจ / ชื่นชอบ (คำที่อยู่ในเอกสาร) ลงในกระดาษ โดยมีเงื่อนไขว่าจะต้องเป็นรายการสั้น ๆ เท่านั้น
  - ซึ่งมีป้าหมายคือการจัดกลุ่มนักเรียน (เอกสาร) ที่มีความสนใจในหนังเรื่องเดียวกันได้นั่งโดยตัวเดียวกัน
  - ในการกระทำเช่นนี้ นักเรียนจึงต้องเลือกโดยใหม่ โดยต้องคำนึงถึงกฎสองข้อ ดังนี้
    - ในกรณีที่ไม่มีโดยที่มีความสนใจคล้ายกัน และในกรณีที่ต้องลดจำนวนหัวข้อให้น้อยลง ให้เลือกโดยที่มีนักเรียนมากกว่า
    - เลือกโดยที่มีนักเรียนที่ให้ความสนใจภพยนตร์แบบเดียวกัน
- เมื่อกระบวนการจัดกล่าวดำเนินต่อไป โดยบางโดยที่รีอกลุ่มบางกลุ่มจะมีขนาดใหญ่ขึ้น และบางกลุ่มอาจหายไป โดยกลุ่มที่เหลืออยู่จะเป็นกลุ่มที่มีความสนใจหรือลักษณะใกล้เคียงกัน ซึ่งผู้วิจัยได้กล่าวว่า Movie Group Process มีการทำงานแบบเดียวกับอัลกอริทึม Gibbs Sampling ในการสร้างแบบจำลอง Dirichlet Multinomial Mixture ซึ่งเป็นแบบจำลองทางสถิติสำหรับการสร้างเอกสาร ดัง PGM ของแบบจำลอง ในรูปที่ 2.6



รูปที่ 2.6 PGM ของแบบจำลอง Dirichlet Multinomial Mixture

จากรูปที่ 2.6 สามารถอธิบายความหมายของตัวแปรได้ดังนี้

$\alpha$  กือ ตัวแปรควบคุมการกระจายตัวของ  $\theta$

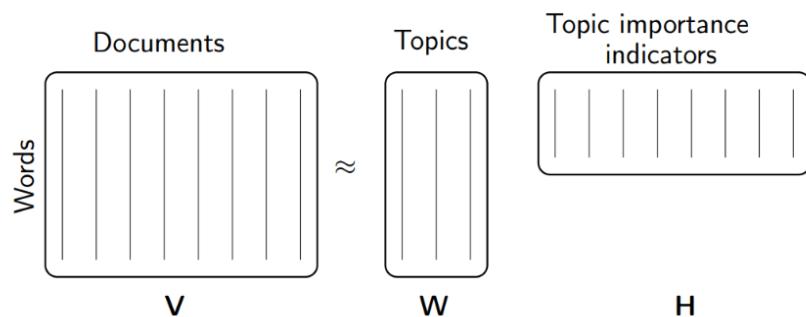
$\beta$  กือ ตัวแปรควบคุมการกระจายตัวของ  $\varphi$

$\theta$	คือ ความน่าจะเป็นของหัวข้อในแต่ละเอกสาร $d$
$z$	คือ หัวข้อของเอกสาร $d$
$d$	คือ เอกสาร $d$
$\varphi$	คือ ความน่าจะเป็นของหัวข้อ $k$ ของเอกสาร $d$

จะเห็นว่า PGM ของ Dirichlet Multinomial Mixture มีความคล้ายกับ PGM ของ LDA ซึ่งทั้งสองแบบจำลองต่างกันที่ LDA ต้องการจำนวนหัวข้อที่ต้องกำหนดไว้ล่วงหน้า ส่วน GSDMM ต้องการจำนวนหัวข้อมากที่สุดที่ต้องการ และจะทำการสร้างแบบจำลองด้วยจำนวนหัวข้อดังกล่าว ซึ่งจำนวนหัวข้ออาจลดลงตามกระบวนการของแบบจำลอง

### 2.1.5 Non-Negative Matrix Factorization (NMF)

Non-Negative Matrix Factorization หรือ NMF [6] คือวิธีการทางสถิติที่ช่วยในการลดมิติของชุดข้อมูล และเป็นแบบจำลองหัวข้ออนเรียนที่ใช้ในการหาหัวข้อที่มีใจความสำคัญจากชุดเอกสารที่ป้อนเข้าแบบจำลอง โดย NMF จะทำการแยกองค์ประกอบของชุดเอกสาร โดยการแทนเอกสารด้วยเมตริกซ์  $V$  และแยกองค์ประกอบออกเป็นสองเมตริกซ์ดังนี้



รูปที่ 2.7 การแยกองค์ประกอบของ NMF

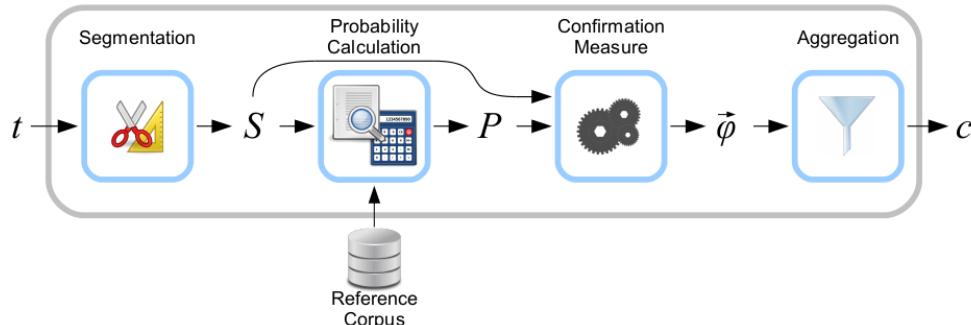
### จากรูปที่ 2.7 สามารถอธิบายเมทริกซ์แต่ละชุดได้ดังนี้

- W      กือ เมทริกซ์ที่แสดงถึงน้ำหนักของแต่ละหัวข้อที่ปรากฏอยู่ในเอกสาร หรือ เมทริกซ์ที่แสดงถึงความสัมพันธ์ระหว่างหัวข้อ กับเอกสาร
- H      กือ เมทริกซ์ที่แสดงถึงความน่าจะเป็นของหัวข้อจากคำต่าง ๆ ในชุดข้อมูล หรือ เมทริกซ์ที่แสดงถึงความสัมพันธ์ของคำกับหัวข้อ โดย NMF จะทำการอนุமานค่าภายในเมทริกซ์ W และเมทริกซ์ H เพื่อหาหัวข้อซ่อนเร้นที่อยู่ภายใต้เอกสาร

#### 2.1.6 Topic Coherence

Topic Coherence เป็นการวัดความสอดคล้องกันของหัวข้อ โดยการวัดคะแนนความสอดคล้อง แต่ละหัวข้อ โดยการวัดระดับความคล้ายคลึงกันของคำที่มีคะแนนสูงในหัวข้อนั้น โดยการวัดรูปแบบนี้จะเป็น การวัดว่าหัวข้อที่หาออกมายังไงเป็นหัวข้อที่มีความหมาย สามารถตีความได้จริง หรือเป็นหัวข้อที่ไม่สามารถตีความได้เกิดจากการอนุมานทางสถิติ

Röder et al. [11] ได้นำเสนอโครงสร้างกระบวนการ สำหรับใช้ในการประเมิน แบบจำลองหัวข้อซึ่งเป็นกระบวนการหลักที่ไลบรารี Gensim ใช้ในการประเมิน โดยมีกระบวนการหลักสี่ขั้นตอนดังรูปที่ 2.8



รูปที่ 2.8 กระบวนการหาค่าความสอดคล้อง

### จากรูปที่ 2.8 สามารถอธิบายความหมายของตัวแปรได้ดังนี้

- $t$       กือ หัวข้อที่ได้จากการอนุமานของแบบจำลองหัวข้อ

- S** กือ หัวข้อที่ถูกแบ่งออกเป็นส่วนๆ
- P** กือ ความน่าจะเป็นที่คำนวณ โดยกระบวนการ Probability Calculation
- φ** กือ เวกเตอร์ของคะแนนความสอดคล้องที่คำนวณจาก Confirmation Measure
- C** กือ ค่าความสอดคล้องที่ผ่านกระบวนการรวมมาเป็นค่าสุดท้าย
- แต่ละกระบวนการดังรูปที่ 2.8 มีการทำงานเบื้องต้นดังนี้
- 1) Segmentation เป็นการแบ่งคำที่สื่อถึงความหมายของหัวข้อออกเป็นหลายส่วนด้วย การจับคู่คำและตั้งสมมติฐานว่าคุณภาพของการหาหัวข้อต่างกัน
  - 2) Probability Calculation เป็นกระบวนการประมาณค่าความน่าจะเป็นในการประกูชื่นของคำที่อยู่ในหัวข้อที่ทำการพิจารณา ซึ่งเป็นกระบวนการที่แสดงให้เห็นถึงความสำคัญของหัวข้อ
  - 3) Confirmation Measure เป็นกระบวนการคำนวณค่าความสอดคล้องระหว่างคำที่ทำการแบ่งและจับคู่ในกระบวนการ Segmentation โดยผู้วิจัยได้นำเสนอหลักการคำนวณค่าความสอดคล้องในกระบวนการนี้ ซึ่งมีหลักการอยู่ทั้งหมด 3 แบบด้วยกัน ประกอบด้วย

a)  $C_{UCI}$  เสนอโดย Newman et al. [11] โดยมีหลักการคือการประเมินค่าความสอดคล้องของหัวข้อด้วยการเปรียบเทียบกับคะแนนที่มนุษย์ได้ให้ไว้เป็นการคำนวณ โดยใช้หลักการ pointwise mutual information (PMI) สามารถคำนวณได้ด้วยสมการ ดังนี้

$$C_{UCI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N PMI(w_i, w_j) \quad (2.2)$$

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i) \cdot P(w_j)} \quad (2.3)$$

b)  $C_{UMass}$  เสนอโดย Mimmo et al. [11] โดยเป็นการประเมินความสอดคล้องโดยการ จัดอันดับคุณภาพการหาหัวข้อจากแบบจำลอง สามารถคำนวณได้ด้วยสมการดังนี้

$$C_{UMass} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^N \sum_{j=1}^{i-1} \log \frac{P(w_i, w_j) + \epsilon}{P(w_j)} \quad (2.4)$$

c)  $C_{NPMI}$  เสนอโดย Aletras and Stevenson [12] โดยเป็นการคำนวณคะแนนความสอดคล้อง Normalized PMI (NPMI) ซึ่งพัฒนามาจาก PMI โดยมีแนวคิดคือ การพิจารณาจากบริบทโดยรอบของคำที่อยู่อันดับต้นในแต่ละหัวข้อ สามารถคำนวณได้ด้วยสมการดังนี้

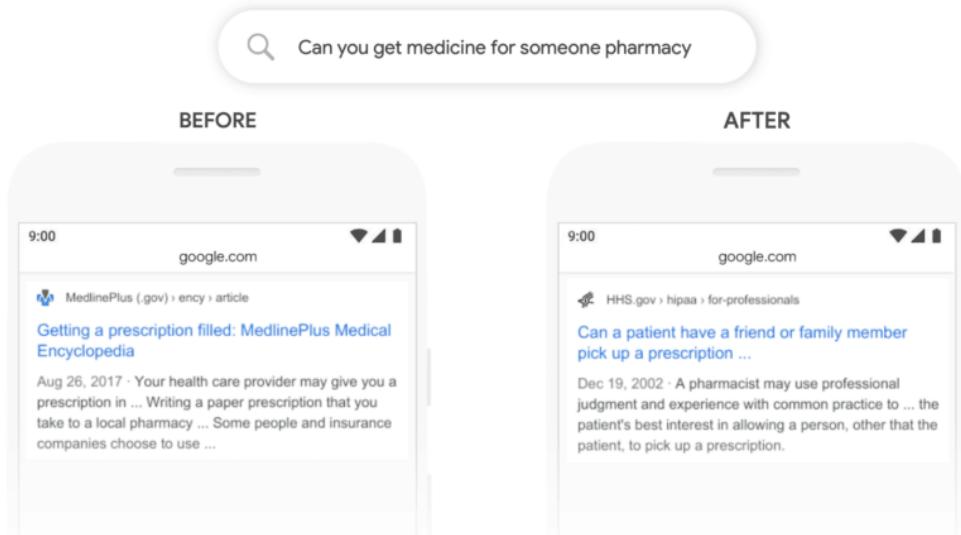
$$C_{NPMI} = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N NPMI(w_i, w_j) \quad (2.5)$$

$$NPMI(w_i, w_j) = \frac{PMI(w_i, w_j)}{-\log(P(w_i, w_j)) + \epsilon} \quad (2.6)$$

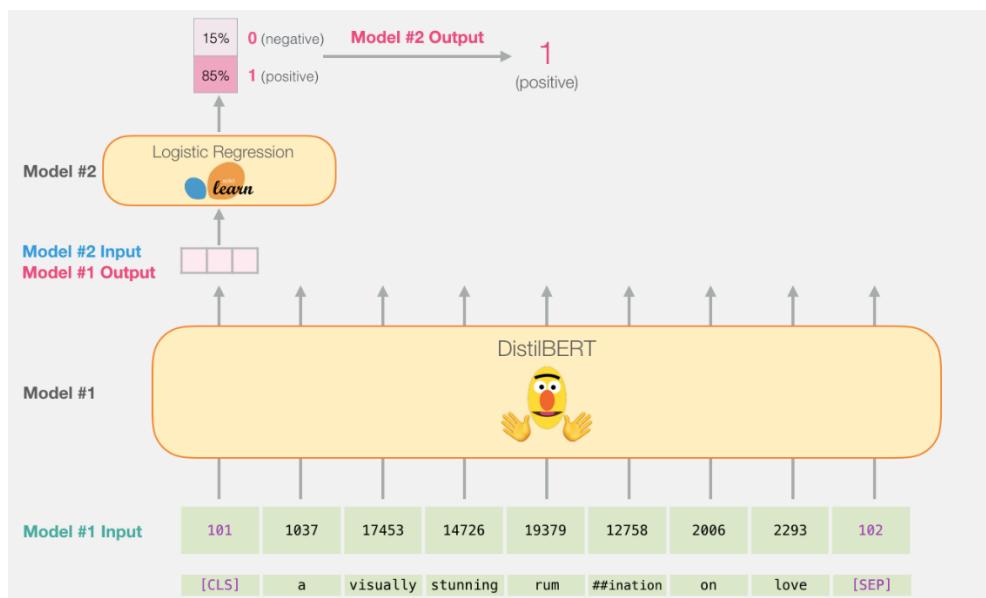
- 4) Aggregation เป็นการรวมคะแนนความสอดคล้องซึ่งเป็นผลลัพธ์จากการคำนวณจากกระบวนการก่อนหน้าและรวมให้ เป็นค่าเดียวโดยการหาค่าเฉลี่ย หรือ มัธยฐาน

### 2.1.7 BERT

BERT [13] ย่อมาจาก Bidirectional Encoder Representations from Transformer ซึ่งพัฒนาโดย Google เป็นแบบจำลองที่อาศัยการทำงานของ Neural Network แบบ Transformer ซึ่งมีจุดเด่นคือกลไกที่เรียกว่า Attention ที่ให้ความสามารถในการเรียนรู้ความสัมพันธ์ในเชิงความหมายและบริบทของคำจากชุดข้อมูล ทำให้ BERT เป็นแบบจำลองที่มีประสิทธิภาพและเป็นที่นิยมในการนำมาใช้สำหรับงานเกี่ยวกับ NLP (Natural Language Processing) ตามตัวอย่างการแสดงผลลัพธ์หลังการใช้ BERT ของ Google ดังรูปที่ 2.9 เมื่องจาก BERT มีความสามารถในการฝึกสอนแบบจำลองล่วงหน้าหรือเทคนิคที่เรียกว่า Transfer Learning ทำให้สามารถนำแบบจำลองที่ฝึกสอนล่วงหน้ามาแล้ว ไปฝึกสอนแบบจำลองเพิ่มเติมด้วยข้อมูลที่ใช้สำหรับงานที่ต้องการ งานที่ BERT สามารถนำไปใช้งานต่ออุดได้ เช่น เช่น Multiclass Classification, Sentences Prediction และ Sentiment Analysis ดังรูปที่ 2.10 เป็นตัวอย่างการทำางานของการนำเข้าข้อมูลข้อความ ทำการประมวลผลผ่าน DistilBERT และได้ผลลัพธ์เป็นเวกเตอร์เพื่อไปฝึกสอนแบบจำลองสำหรับการจำแนกอารมณ์ Logistic Regression



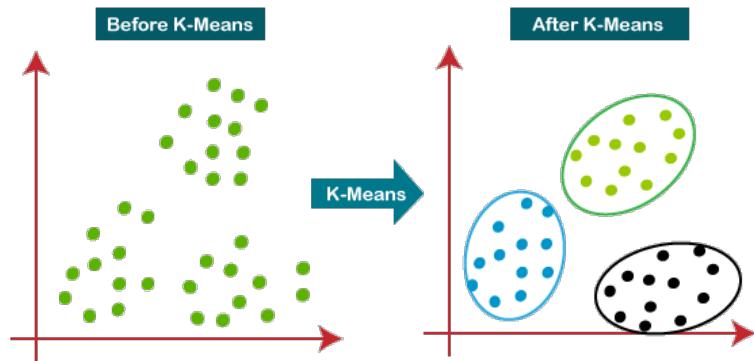
รูปที่ 2.9 ตัวอย่างการแสดงผลลัพธ์ทั้งก่อน – หลังจากใช้ BERT ของ Google



รูปที่ 2.10 ตัวอย่างการใช้งาน BERT สำหรับงาน Sentiment Analysis

### 2.1.8 K-Means

การจัดกลุ่มแบบ K-Means เป็นวิธีที่ใช้สำหรับการวิเคราะห์การจัดกลุ่ม และถูกจัดอยู่ในกลุ่มของ Unsupervised Learning ซึ่งมีจุดมุ่งหมายเพื่อแบ่งชุดข้อมูลทั้งหมด  $n$  รายการออกเป็น  $k$  กลุ่ม โดยที่ข้อมูลมีค่าเฉลี่ยใกล้เคียงกับจุดศูนย์กลางของคลัสเตอร์ (Centroid) ใหม่ที่สุด จะถูกจัดให้อยู่ในกลุ่มดังกล่าว ดังรูปที่ 2.11



รูปที่ 2.11 ผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลด้วย K-Means Clustering

## 2.2 เทคนิค หรือเครื่องมือที่ใช้

### 2.2.1 Python

เป็นภาษาโปรแกรมที่พัฒนาขึ้นด้วยจุดประสงค์ในการใช้งานที่หลากหลาย Python ประกอบด้วย ไลบรารีสำหรับความสะดวกในการทำงานหลากหลายแขนง โดยเฉพาะในด้าน Data Science ที่มีเครื่องมือที่ช่วยอำนวยความสะดวกอย่าง Jupyter Notebook ซึ่งช่วยให้สามารถเขียนคำสั่งและแสดงผลได้ง่ายมากขึ้น อีกทั้งยังมีไลบรารีที่ช่วยในการทำงานได้ทั้งกระบวนการอย่าง Pandas ที่ช่วยในเรื่องของการจัดการและแสดงข้อมูลให้อยู่ในรูปแบบของตาราง Numpy ที่ช่วยในเรื่องของการคำนวณที่มีความซับซ้อน และการจัดการข้อมูลในรูปแบบอาเรย์ Matplotlib ที่ช่วยในเรื่องของการแสดงผลข้อมูลให้อยู่ในแบบรูปภาพ และ Scikit-learn ที่ช่วยในเรื่องของการสร้างแบบจำลองการเรียนรู้ของเครื่องจักร (Machine Learning) รวมถึงการประเมินประสิทธิภาพของแบบจำลอง เป็นต้น

## 2.2.2 เครื่องมือที่ใช้ในขั้นตอน Data Preparation

### 2.2.2.1 NLTK

NLTK [11] หรือ Natural Language Toolkit เป็นไลบรารีชุดเครื่องมือที่อำนวยความสะดวกในการสร้างโปรแกรมเพื่อการประมวลผลภาษาธรรมชาติในรูปแบบภาษาอังกฤษ โดย NLTK มีคลังข้อมูลภาษาามากกว่า 50 คลังข้อมูล รวมถึงชุดเครื่องมือที่ใช้ในการประมวลผลทางภาษา เช่น การตัดคำ (Tokenization) โดยมีตัวอย่างดังรูปที่ 2.12 คือการตัดคำจากประโยค "I love Thailand" ออกเป็นสามคำ การตัดทอนลดรูปคำ (Stemming) ตามตัวอย่างดังรูปที่ 2.13 จะเห็นได้ว่ามีการลดรูปคำจากคำว่า connection เป็นคำว่า connect การแปลงคำ โดยอิงจากคำที่เกี่ยวข้องกัน (Lemmatization) การติดป้ายคำ (Tagging) และการวิเคราะห์โครงสร้างประโยค (Parsing)

```
from nltk.tokenize import word_tokenize

sentence = "I love Thailand"
print(word_tokenize(sentence))

['I', 'love', 'Thailand']
```

รูปที่ 2.12 ตัวอย่างการตัดคำจากประโยค

```
porter = PorterStemmer()
print(porter.stem("connection"))

connect
```

รูปที่ 2.13 ตัวอย่างการลดรูปคำโดยใช้ Stemmer แบบ Porter

### 2.2.2.2 PyThaiNLP

PyThaiNLP [14] เป็นไลบรารีชุดเครื่องมือที่อำนวยความสะดวกในการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ในรูปแบบภาษาไทย ประกอบด้วยเครื่องมือการประมวลผลข้อมูลคล้ายกับ NLTK และเครื่องมือการแปลงข้อมูลข้อความให้อยู่ในรูปแบบเวกเตอร์อย่าง thai2fit เนื่องจากภาษาไทยมีลักษณะการพิมพ์ที่ติดกันไม่มีการเว้นวรรค อิกทั้งยังมีลักษณะไวยากรณ์ของภาษาที่ต่างกัน จึงไม่สามารถใช้ word2vec ได้ โดยมีตัวอย่างการแปลงข้อมูลจากประโดยกเป็นเวกเตอร์ดังรูปที่ 2.14 โดยมีกระบวนการเบื้องต้นคือเครื่องมือ Word Vector ของไลบรารี PyThaiNLP จะทำการตัดคำ (Tokenize) และแปลงเป็นเวกเตอร์โดยการใช้แบบจำลอง thai2fit โดยไลบรารี PyThaiNLP เปิดให้ทุกคนสามารถใช้ได้โดยไม่มีค่าใช้จ่าย และยังมีการพัฒนาอยู่จนถึงปัจจุบัน

```
sentence = 'ฉันรักประเทศไทย'
sentence_vector = wv.sentence_vectorizer(sentence)
print(sentence_vector)

[[ 1.71394671e-01 -2.43284330e-01 -1.41679967e-02  2.96576001e-01
-2.22379004e-01 -7.44023304e-02 -5.33566376e-03  9.41766674e-03
-1.74809662e-01 -1.44934667e-01 -3.31633329e-01 -1.30679997e-01
1.81364005e-01  1.67169669e-01 -2.02283661e-01 -1.57379980e-02
-7.14666670e-02  2.12028998e-01  4.89439977e-02 -2.99740005e-02
-1.26108664e-01  2.74677332e-01  9.74936659e-02  6.31487002e-01
-3.15139999e-01  4.49893996e-01  1.27577665e-01 -1.58133171e-03
-2.14869662e-01 -5.12753278e-02 -1.92380051e-02 -2.24013329e-01]
```

รูปที่ 2.14 ตัวอย่างการใช้งาน thai2fit ในการแปลงประโดยกให้อยู่ในรูปเวกเตอร์

### 2.2.2.3 Sentence Transformers

Sentence Transformers [15] เป็นเฟรมเวิร์กของ Python ที่ใช้สำหรับการฝังข้อความในรูปแบบของคำ หรือประโดยก และรูปภาพที่ทันสมัยตัวหนึ่ง การฝังสามารถรองรับภาษามากกว่า 100 ภาษา และยังใช้งานได้จำกันงานทั่วไป อย่างเช่น การค้นหาเชิงความหมาย (Semantic Search) และการหาความคล้ายคลึงของข้อความ เป็นต้น

โดยเฟรมเวิร์กตัวนี้ถูกสร้างขึ้นด้วย PyTorch และ Hugging Face Transformers อิกทั้งยังผ่านการฝึกสอนแบบจำลองเป็นจำนวนมาก เพื่อรองรับงานที่หลากหลาย และยังง่ายต่อการปรับแต่งแบบจำลองของตัวเอง

จากการสร้างแบบจำลองภาษา BERT ซึ่งได้เพิ่มประสิทธิภาพของการสร้างแบบจำลองทางภาษาอย่างมีนัยสำคัญ แต่ BERT ยังมีข้อจำกัดคือไม่สามารถสร้างเวกเตอร์ที่แสดงแทนข้อความแต่ละประโยคที่เป็นอิสระต่อกันได้ จึงมีการคิดค้น Sentence-BERT (SBERT) ซึ่งเป็นวิธีแบบจำลองที่ใช้ในการแปลงข้อมูลข้อความหรือประโยค ให้อยู่ในรูปแบบเวกเตอร์มิติขนาดคงที่ โดยอาศัยแบบจำลอง BERT ที่ฝึกสอนล่วงหน้า จึงทำให้ข้อจำกัดดังกล่าวเป็นไปได้โดยการเพิ่มชั้นการทำงาน Pooling หลังจากได้ผลลัพธ์จากแบบจำลอง BERT ทำให้ได้เวกเตอร์ขนาดคงที่โดยที่บังคับคุณสมบัติและประสิทธิภาพเช่นเดียวกับ BERT สำหรับการทำงานอื่น ๆ ต่อไป

#### 2.2.2.4 Pandas

Pandas เป็นไลบรารีสำคัญที่ถูกเขียนขึ้นด้วยภาษา Python ที่ถูกนำมาใช้กันอย่างแพร่หลายสำหรับการวิเคราะห์ข้อมูล และการอบรมแบบจำลอง (Machine Learning)

โดย Pandas ช่วยให้การทำงานกับข้อมูลที่ซับซ้อน และต้องใช้เวลานานเป็นเรื่องง่าย ซึ่งข้อมูลในที่นี้หมายถึง การทำความสะอาดข้อมูล (Data Cleansing) การปรับข้อมูลให้เป็นมาตรฐาน (Normalization) การแสดงผลข้อมูลด้วยกราฟ (Data Visualization) และอื่น ๆ อีกมากมาย

```
In [38]: import pandas as pd
one = pd.DataFrame({
    'Name': ['Amber', 'Jack', 'Brown', 'Smith', 'Young'],
    'subject_id':['sub1','sub2','sub4','sub6','sub5'],
    'Marks_scored':[93,90,82,64,71]},
    index=[1,2,3,4,5])
two = pd.DataFrame({
    'Name': ['Ben', 'Cole', 'Sam', 'Tom', 'Martial'],
    'subject_id':['sub2','sub4','sub3','sub6','sub5'],
    'Marks_scored':[96,88,73,77,81]},
    index=[1,2,3,4,5])
print (pd.concat([one,two]))
```

	Name	subject_id	Marks_scored
1	Amber	sub1	93
2	Jack	sub2	90
3	Brown	sub4	82
4	Smith	sub6	64
5	Young	sub5	71
1	Ben	sub2	96
2	Cole	sub4	88
3	Sam	sub3	73
4	Tom	sub6	77
5	Martial	sub5	81

รูปที่ 2.15 ตัวอย่างการจัดการข้อมูลโดยการใช้ Pandas

จากรูปที่ 2.15 ได้มีการจัดการกับข้อมูลจากทั้ง 2 กลุ่มข้อมูล (Dataframe) ผ่านการใช้คำสั่ง .concat() ที่มีคุณสมบัติในการรวมข้อมูลจากกลุ่มข้อมูล 2 กลุ่มมาไว้ในกลุ่มเดียวกัน

### 2.2.3 เครื่องมือที่ใช้ในขั้นตอน Modeling

#### 2.2.3.1 Gensim

Gensim [16] เป็นไลบรารีสำหรับการสร้างแบบจำลองหัวข้อ (Topic Modeling) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing) โดย Gensim ถูกนำมาใช้กันอย่างแพร่หลาย ทั้งฟังก์ชันที่ใช้สำหรับการเตรียมข้อมูลอย่าง Word2Vec [12] และ ฟังก์ชันการสร้างแบบจำลองหัวข้ออย่าง Latent Dirichlet Allocation (LDA) หรือ Non-Negative Matrix Factorization (NMF)

#### 2.2.3.2 Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)

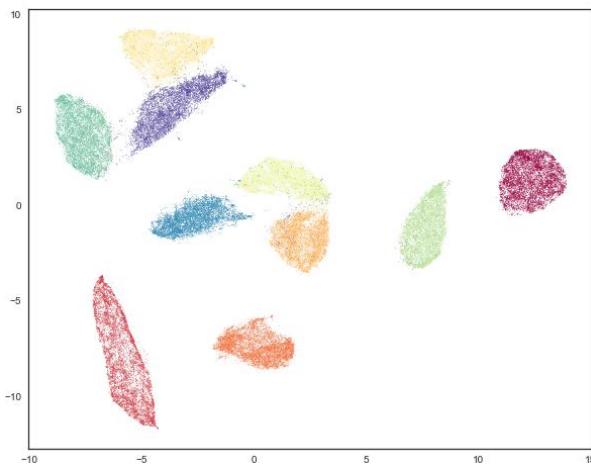
GSDMM หรือ Gibbs Sampling Dirichlet Multinomial Mixture [5] เป็นไลบรารีที่อ่านว่าความสะดวกในการสร้างแบบจำลอง Dirichlet Multinomial Mixture ด้วยอัลกอริทึม Gibbs Sampling โดยสามารถสร้างแบบจำลองตามกระบวนการ Movie Group Process ตามที่ผู้วิจัยได้นำเสนอไว้ดังข้อที่ 2.1.4

### 2.2.3.3 Scikit-Learn

Scikit-Learn [17] หรือ Sklearn เป็นไลบรารีที่มีประโยชน์ที่สุดสำหรับการอบรมแบบจำลอง โดยที่ไลบรารีนี้มีการเลือกเครื่องมือที่มีประสิทธิภาพสำหรับการเรียนรู้ของเครื่องและการสร้างแบบจำลองทางสถิติ ทั้งการจำแนกประเภท (Classification) การ回帰 (Regression) การจัดกลุ่ม (Clustering) และการลดมิติของข้อมูล (Dimensional Reduction) ซึ่งไลบรารีนี้ถูกเขียนด้วยภาษา Python เป็นส่วนใหญ่

### 2.2.3.4 UMAP

UMAP เป็นเทคนิคการลดขนาดของข้อมูลแบบใหม่ที่สามารถใช้เพื่อแสดงรูปแบบของการจัดกลุ่มในข้อมูลที่มีมิติสูงคล้ายคลึงกับ t-SNE (t-Distributed Stochastic Neighbor Embedding) ดังรูปที่ 2.16 นอกจากนี้ UMAP ยังสามารถลดขนาดข้อมูลที่ไม่เป็นเชิงเส้นในมิติที่คนสามารถแยกแยะได้อีกด้วย



รูปที่ 2.16 ตัวอย่างการใช้ UMAP ในการจัดกลุ่มข้อมูล

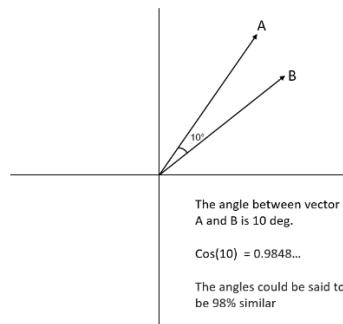
## 2.2.4 เครื่องมืออื่น ๆ

### 2.2.4.1 Cosine Similarity

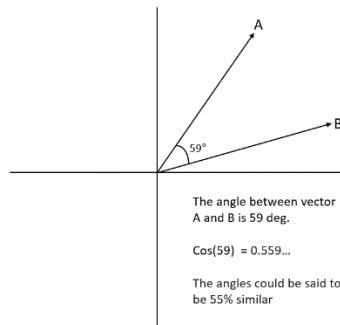
Cosine Similarity เป็นมาตรการวัดที่หาปริมาณความคล้ายคลึงระหว่างเวกเตอร์ตั้งแต่ 2 ตัวขึ้นไป โดยเป็นโคไซน์ (Cosine) ของมุนหมายระหว่างเวกเตอร์ที่ไม่ใช่สูนย์ซึ่งจะถูกอธิบายทางคณิตศาสตร์ได้ดังสมการที่ 6

$$\text{Similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} \quad (2.7)$$

ความคล้ายคลึงกันของโโคไซน์จะมีค่าอยู่ระหว่าง 0 และ 1 โดยในกรณีที่มุนระหว่างเวกเตอร์ทั้งสองมีค่า 90 องศา ความคล้ายคลึงของโโคไซน์จะมีค่าเป็น 0 ซึ่งหมายความว่าเวกเตอร์ทั้งสองตั้งฉากกัน และหากความคล้ายคลึงของโโคไซน์เข้าใกล้ 1 มากขึ้น จะส่งผลให้มุนระหว่างเวกเตอร์ A และ B ทั้งสองจะเล็กลง



รูปที่ 2.17 ตัวอย่างผลลัพธ์ของมุนระหว่างเวกเตอร์ที่มีค่าเข้าใกล้ 0 องศา



รูปที่ 2.18 ตัวอย่างผลลัพธ์ของมุนระหว่างเวกเตอร์ที่มีค่าเข้าใกล้ 90 องศา

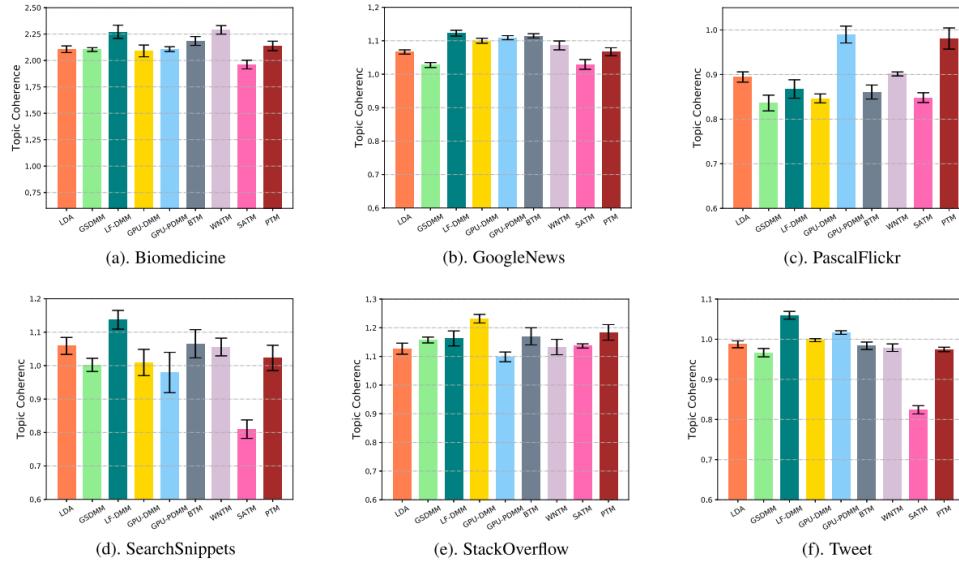
จากรูปที่ 2.17 มุนระหว่างเวกเตอร์ A และ B มีค่า 10 องศาจึงทำให้ค่าความคล้ายคลึงของโโคไซน์มีค่าประมาณ 0.9848 ซึ่งมีค่าเข้าใกล้ 1 และหากมุนระหว่างเวกเตอร์ทั้งสองมีค่าสูงมากเท่าไหร่ จะทำให้ค่าความคล้ายคลึงของโโคไซน์เข้าใกล้ 0 มากขึ้น ดังรูปที่ 2.18 ตามที่ได้กล่าวไปข้างต้น

## 2.3 งานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ทำการศึกษาและค้นคว่างานวิจัยที่มีความเกี่ยวข้องกับการสร้างแบบจำลองหัวข้อกับข้อมูล และข้อมูลที่มีปัจจัยเวลาและสถานที่ประกอบ เพื่อเป็นต้นแบบในการวิจัยและพัฒนาการสร้างแบบจำลอง

### 2.3.1 งานวิจัยที่ศึกษาการสร้างแบบจำลองหัวข้อ

- Qiang et al. [7] ได้ทำการศึกษารอบรวมการสร้างแบบจำลองหัวข้อแต่ละแบบที่ใช้ในการหาหัวข้อซ่อนเร้นจากข้อความขนาดสั้น โดยผู้วิจัยได้อธิบายถึงลักษณะของแบบจำลองหัวข้อแต่ละแบบ มีการศึกษาเกี่ยวกับการพัฒนาของหัวข้อและปัญหา โดยมีการกล่าวถึงปัญหาของ แบบจำลองหัวข้อแบบดั้งเดิม เช่น Probabilistic Latent Semantic Analysis (pLSA) และ Latent Dirichlet Allocation (LDA) ในด้านประสิทธิภาพที่ลดลงเมื่อนำมาใช้งานกับข้อความขนาดสั้น เนื่องจากข้อความขนาดสั้นมีคำที่เกิดขึ้นซ้ำในแต่ละเอกสารเป็นจำนวนมากโดย Qiang et al. ได้ทำการรอบรวมและเปรียบเทียบแบบจำลองหัวข้อสำหรับข้อความขนาดสั้น ที่ถูกคิดค้นขึ้นเพื่อแก้ปัญหาดังกล่าว โดยแบ่งเป็นสามประเภทคือ 1) Dirichlet Multinomial Mixture Model มีแนวคิดคล้ายกับ LDA แต่เป็นการกำหนดหัวข้อให้กับแต่ละเอกสารต่างจาก LDA ที่เป็นการกำหนดหัวข้อให้กับคำที่ปรากฏ ผู้วิจัยรอบรวมอัลกอริทึมประเภทนี้มาสี่แบบคือ GSDMM, LF-DMM, GPU-DMM, และ GPU-PDMM 2) Global Word Co-occurrences Methods มีแนวคิดคือคำที่อยู่ใกล้กัน จะมีความเกี่ยวข้องกันในเชิงความหมาย ผู้วิจัยได้รอบรวมอัลกอริทึมประเภทนี้มาสองแบบคือ BTM และ WNTM 3) Self-Aggregation Based Methods ใช้แนวคิดในการลดการกระจายของข้อความสั้น โดยการรวมข้อความสั้นเข้าด้วยกันเป็นเอกสารใหม่แล้วนำเข้าแบบจำลองหัวข้อ โดยผู้วิจัยได้รอบรวมอัลกอริทึมประเภทนี้มาสองแบบคือ SATM และ PTM โดยทางผู้วิจัยได้นำแบบจำลองที่ศึกษามาทดลอง เพื่อนำมาใช้ในการหาหัวข้อกับชุดข้อมูลที่ต่างกัน ทกชุดข้อมูล และทำการประเมินประสิทธิภาพโดยใช้ตัววัดเป็นคะแนนความสอดคล้อง Topic Coherence ดังรูปที่ 2.19

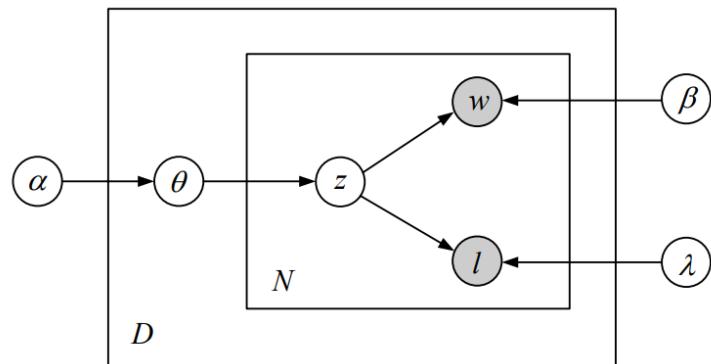


รูปที่ 2.19 การเปรียบเทียบประสิทธิภาพของแบบจำลองหัวข้อสำหรับข้อความสั้น

### 2.3.2 งานวิจัยที่ศึกษาการสร้างแบบจำลองหัวข้อที่มีเวลาและสถานที่เป็นปัจจัยประกอบ

- Alghamadhi et al. [9] ได้ศึกษาเกี่ยวกับการสร้างแบบจำลองหัวข้อโดยแบ่งออกเป็นสอง ประเภท คือ Topic Modeling ประกอบด้วย Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM) โดยได้ทำการวิเคราะห์จุดเด่นและข้อสังเกตในการใช้แบบจำลองหัวข้อ แบบต่างๆ และอีกประเภทคือ Topic Evolution Models ซึ่งมีหลักการสร้างแบบจำลองโดย คำนึงถึง ความสำคัญของปัจจัยเวลา เช่น Topic over Time, Dynamic Topic Models ซึ่งผู้วิจัยได้ทำการ รวบรวมการสร้างแบบจำลองหัวข้อ ที่คำนึงถึงเวลาแต่ละรูปแบบซึ่งมีการ ใช้ตัวแปรเวลาต่างรูปแบบกัน เช่น Topic over Time จะใช้เวลาที่เป็นตัวแปรแบบต่อเนื่อง (Continuous) และ Dynamic Topic Models และ Multiscale Topic Tomography มีการคำนึงถึง ปัจจัยเวลาเป็นตัวแปรแบบไม่ต่อเนื่อง (Discrete) ผู้วิจัยจึงได้ทำการเปรียบเทียบการสร้าง แบบจำลองหัวข้อสองรูปแบบดังกล่าวว่า Topic Evolution Models มีความแม่นยำมากกว่าเมื่อ เนื้อหาและหัวข้อมีการเปลี่ยนแปลงและเติบโตไปตามเวลา
- Wang et al. [18] ได้เสนอ Location Aware Topic Model (LATM) ซึ่งเป็นส่วนที่ต่อขอดมาจาก แบบจำลองหัวข้อในปัจจุบัน โดยเฉพาะ Latent Dirichlet Allocation

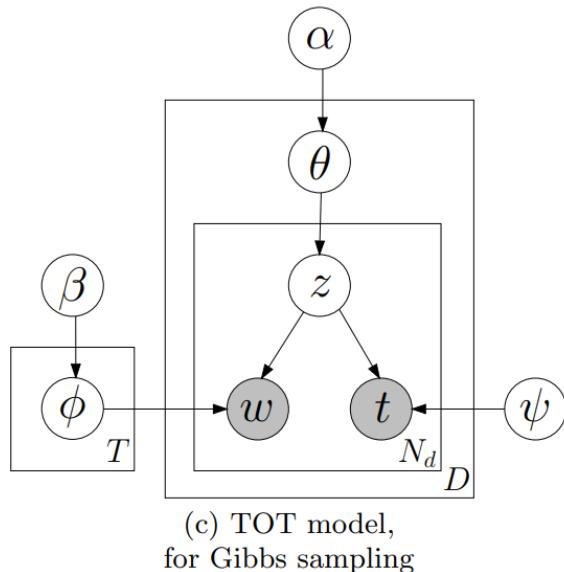
(LDA) โดยการสร้างแบบ จำลองหัวข้อที่มีสถานที่เป็นปัจจัยประกอบที่ Wang et al. นำเสนอเป็นการเพิ่มความสามารถ ของแบบจำลองหัวข้อที่เป็นที่นิยมอย่าง Latent Dirichlet Allocation (LDA) ใน การเรียนรู้ความ สัมพันธ์ระหว่างคำและสถานที่จากชุด ข้อมูลเอกสาร โดยมีการตั้งสมมติฐานว่าแต่ละคำมี ความสัมพันธ์กับแต่ละสถานที่ และ ได้เสนอ Probabilistic Graphical Model ไว้ดังรูปที่ 2.20 ทั้งนี้ทางผู้วิจัยได้มีการ ระบุถึงปัญหาของการระบุสถานที่สำหรับคำแต่ละคำ เมื่อคำดังกล่าวมีความหมาย สัมพันธ์กับสถานที่มากกว่าหนึ่งสถานที่ และปัญหาของสถานที่ที่เป็นสถานที่ภายใน สถานที่ใหญ่ เช่น บักกิ่ง เป็นเมืองหนึ่งในประเทศจีน จึงเป็นข้อจำกัดที่แบบจำลอง หัวข้อนี้ต้องมีการจำกัดขอบเขตระดับของสถานที่



รูปที่ 2.20 PGM ของ Location Aware Topic Model (LATM)

- Wang and McCallum [19] ได้เสนอแบบ จำลองหัวข้อที่มีการต่อยอดมากจาก แบบ จำลองหัวข้อ Latent Dirichlet Allocation คือ Topic Over Time (TOT) ซึ่ง มี ความสามารถในการสร้างแบบ จำลองที่ค้นหาการกระจายของเวลาไปพร้อมกับแบบ แผนการเกิดขึ้นของคำภายในชุดข้อมูล นำมาซึ่งความสามารถในการค้นหาหัวข้อ และ ช่วงเวลาที่พบหัวข้อนั้นๆ กล่าวคือทำให้สามารถทำความเข้าใจหัวข้อได้มากขึ้น สำหรับแบบ จำลอง TOT การค้นหาหัวข้อ่อนเร้นจะมีอิทธิพลมาจากการ และแบบ แผนการเกิดขึ้นของคำ และเหตุผลของการหลีกเลี่ยงการสร้างแบบ จำลองบน สมมติฐาน Markov ทำให้ TOT มีความสามารถในการทำนายช่วงเวลาที่เกิดขึ้น เอกสารเมื่อให้เอกสารที่ไม่มีเวลาประกอบ หรือแม้แต่ทำนายหัวข้อที่เกิดขึ้น เมื่อ กำหนดช่วงเวลาให้กับแบบ จำลอง โดยที่หัวข้อแต่ละหัวข้อของแบบ จำลอง TOT จะมี

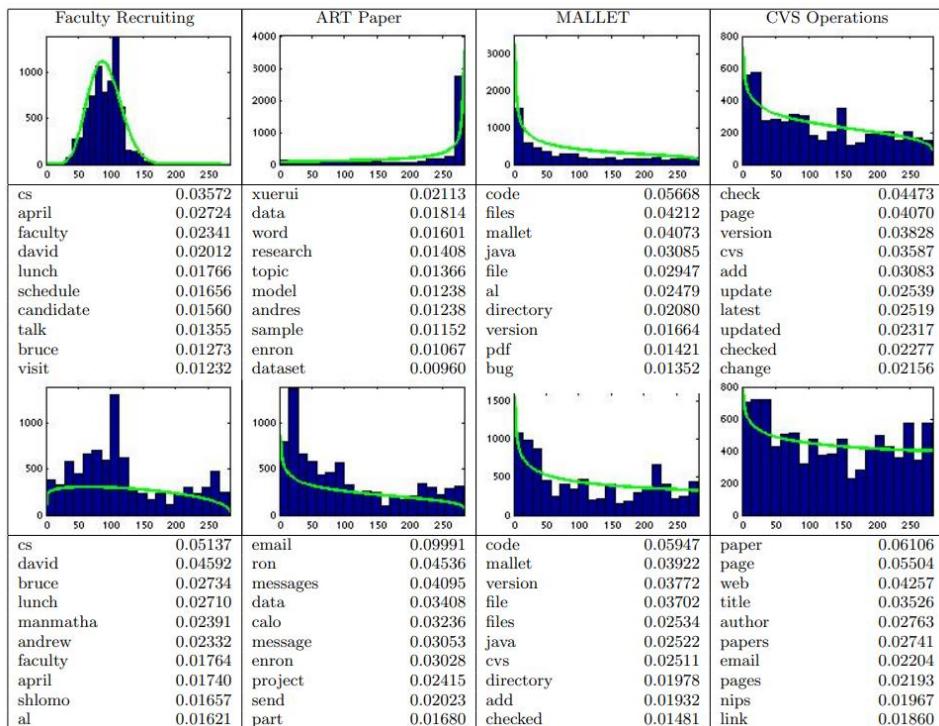
ความเกี่ยวข้องกับการกระจายของเวลา โดย TOT เลือกใช้การกระจายแบบ Beta นอกจานี้ในกระบวนการสร้างเอกสาร หัวข้อยังมีหน้าที่ในการสู่มสร้างคำและเวลา สำหรับคำนั้นๆอีกด้วย ผู้วิจัยได้กล่าวถึงความสามารถที่แบบจำลองหัวข้อ LDA ไม่มี และยกตัวอย่างประโยชน์เมื่อใช้แบบจำลอง TOT เช่นแบบจำลองหัวข้อ LDA ไม่สามารถสร้างความแตกต่างในแบ่งเวลาให้กับหัวข้อ Mexican-American War กับ World War I ได้ ทั้งที่ทั้งสองเหตุการณ์เกิดห่างกันถึง 70 ปี หรือกล่าวได้อีกแบบคือ แบบจำลองหัวข้อปกติ พลาดโอกาสในการอธิบายหัวข้อที่เกิดขึ้น ได้อย่างละเอียดมาก ขึ้น ผู้วิจัยได้อธิบายแบบจำลองในรูปแบบ Probabilistic Graphical Model ไว้ดังรูปที่ 2.21 ผู้วิจัยกล่าวว่าแบบจำลอง TOT ต่างจากแบบจำลองหัวข้อที่มีเวลาเป็นปัจจัย ประกอบรูปแบบอื่นคือ TOT คำนึงถึงปัจจัยเวลาเป็นตัวแปรต่อเนื่องซึ่งทำให้มีปัญหาในด้านของการเลือกช่วงเวลาที่ต้องการสร้างแบบจำลอง รวมถึงการไม่ตั้งสมมติฐาน Markov Assumption



รูปที่ 2.21 PGM ของ Topic Over Time (TOT)

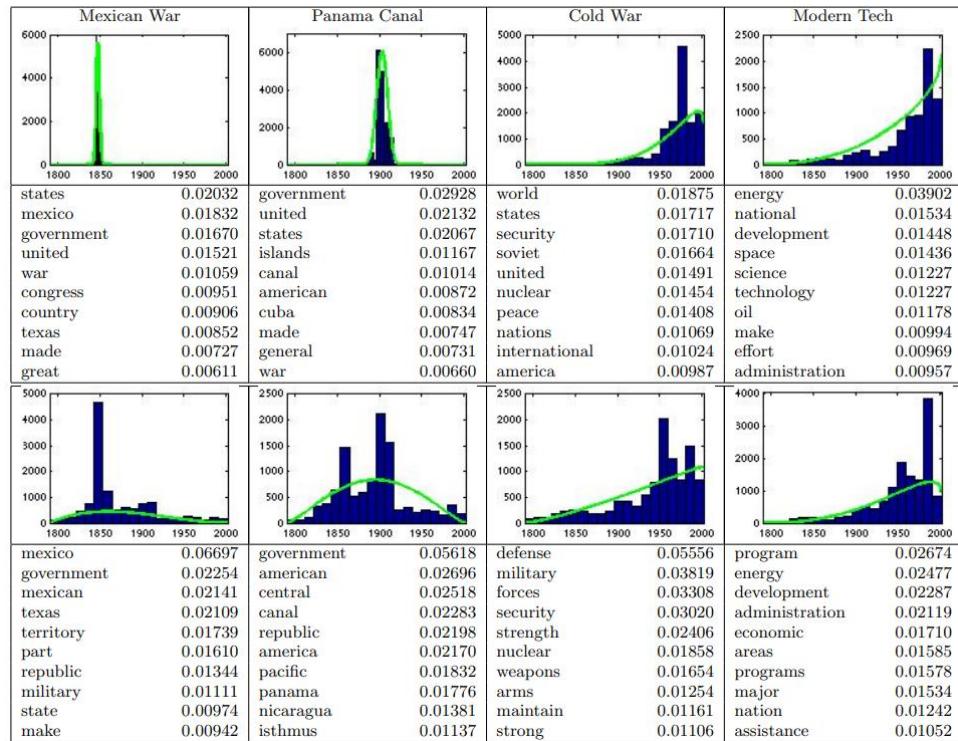
ผู้วิจัยได้ทำการเสนอผลลัพธ์ของการวิจัยด้วยการใช้ข้อมูลสามชุด ในการสร้างแบบจำลองหัวข้อ TOT และ LDA เพื่อเปรียบเทียบผลลัพธ์หัวข้อที่ได้จากการทั้งสองแบบจำลอง สำหรับแบบจำลอง LDA การกระจายของเวลาได้ทำการประมาณค่าในภายหลังจากการสร้างแบบจำลองแล้ว ผู้วิจัยกำหนดจำนวนหัวข้อไว้ที่ 50 สำหรับทั้งสามชุดข้อมูล ประกอบไปด้วย ข้อความจากอีเมลส่วนตัวของผู้วิจัยตลอดระยะเวลา 9

เดือน ตั้งแต่เดือนมกราคมถึงกันยายนปี 2004 รวมทั้งข้อความที่ส่งและได้รับประกอบด้วยอีเมลล์จำนวน 13,300 ฉบับ โดยผลลัพธ์ที่ได้พบตัวอย่างหัวข้อเช่น Faculty Recruiting ซึ่งแบบจำลอง TOT สามารถระบุหัวข้อที่ได้อย่างชัดเจนในว่าอยู่ช่วงๆ ในไม้ผล ต่างจากแบบจำลอง LDA ที่ระบุว่ามีอยู่ตลอดทุกช่วงเวลา และมีตัวอย่างผลลัพธ์ดังรูปที่ 2.22 โดยแควรบันเป็นผลลัพธ์จากแบบจำลอง TOT และถ้าล่างเป็นผลลัพธ์ของ LDA



รูปที่ 2.22 ผลลัพธ์หัวข้อจากชุดข้อมูลข้อความจากอีเมลส่วนตัวแบบจำลอง TOT และ LDA

ข้อมูลการแฉลงนโยบายประจำปีต่อรัฐสภาของประธานาธิบดีสหรัฐอเมริกาตลอดระยะเวลา 200 ปี ซึ่งครอบคลุมประวัติศาสตร์ทั้งหมดของสหรัฐอเมริกา ตัวอย่างหัวข้อที่ได้จากแบบจำลอง TOT เช่น Mexican-American War จะเห็นได้ว่าเกิดในช่วงก่อนปี 1850 แต่สำหรับแบบจำลอง LDA หัวข้อมีลักษณะกระจายอยู่ตลอดระยะเวลา ประวัติศาสตร์สหรัฐอเมริกา และมีตัวอย่างผลลัพธ์ดังรูปที่ 2.23 โดยแควรบันเป็นผลลัพธ์จากแบบจำลอง TOT และถ้าล่างเป็นผลลัพธ์ของ LDA



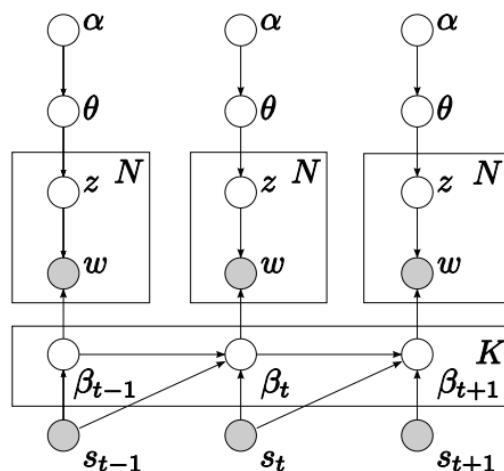
รูปที่ 2.23 ผลลัพธ์หัวข้อจากชุดข้อมูลแกล้งน้อยบายประจำปี แบบจำลอง TOT

และ LDA

ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง TOT และ LDA ผู้วิจัยได้เลือกใช้ค่า KL Divergence เป็นเกณฑ์ในการวัดความแตกต่างของการกระจาย โดยผลลัพธ์คือ TOT มีประสิทธิภาพที่ดีกว่า LDA กล่าวคือมีความสามารถในการค้นหาหัวข้อที่ดีกว่า นอกจากนี้ผู้วิจัยยังมีการทดลองการใช้งานแบบจำลองสำหรับการทำนายปีของเอกสาร เมื่อเปรียบเทียบด้วยค่า Accuracy TOT มีประสิทธิภาพที่ดีกว่า LDA เกือบสองเท่า

- Blei et al. [20] ได้พัฒนา Continuous Time Dynamic Topic Models (cDTM) ที่ใช้ Brownian Motion ในการสร้างแบบจำลองเพื่อค้นหาหัวข้อซ่อนเร้นที่อยู่ภายในชุดเอกสาร โดยหัวข้อในที่นี้หมายถึงรูปแบบของคำที่เกิดขึ้นอย่างมีแบบแผนและมีพัฒนาการตลอดการเรียงตัวของชุดเอกสาร cDTM เป็นแบบจำลองที่พัฒนามาเพื่อต่อ ยอดจากแบบจำลอง Discrete Dynamic Topic Model (dDTM) ที่มีลักษณะคือชุดเอกสารถูกแบ่งเท่า ๆ กันออกเป็นกลุ่มย่อยตามลำดับเวลา และตั้งสมมติฐานว่ามีการพัฒนาของหัวข้อในแต่ละกลุ่ม ซึ่งหมายความว่าตัวแปรเวลาสำหรับ dDTM จะเป็นแบบไม่ต่อเนื่องและสามารถสร้างแบบจำลองได้ด้วยความละเอียดของระดับเวลาเดียว

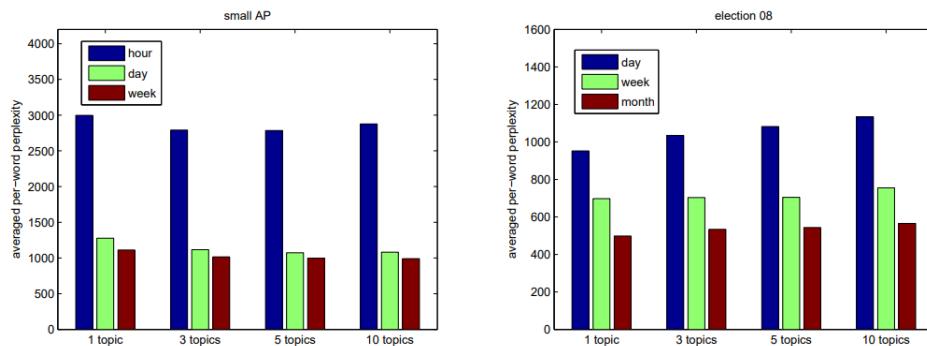
เท่านั้น นอกจากนี้ยังมีปัญหาของความต้องการของทรัพยากรในการสร้างแบบจำลองที่สูงอีกด้วย จึงมีการพัฒนา cDTM โดยการเปลี่ยนตัวแปรเวลาให้เป็นตัวแปรต่อเนื่อง ทำให้การสร้างแบบจำลองสามารถทำได้บนความละเอียดของระดับเวลาได้ ได้ และมีการใช้ Brownian motion ใน การตรวจจับพัฒนาการของหัวข้อตลอดชุดข้อมูลเอกสาร นอกจากนี้การให้เวลาเป็นตัวแปรแบบต่อเนื่องยังทำให้ปัญหาด้านความต้องการทรัพยากรลดลงอย่างมาก โดยโครงสร้างของแบบจำลอง cDTM สามารถอธิบายเป็น PGA [ให้ดังรูปที่ 2.24 ผู้วิจัยกล่าวว่า cDTM ต่างจากแบบจำลองหัวข้ออื่นๆ ที่มีการใช้เวลาเป็นปัจจัยประกอบอย่าง TOT และ DMM โดยที่ทั้งสองแบบจำลองสมมติให้หัวข้อซ่อนเร้นนั้นมีลักษณะคงที่ และใช้เวลาในการค้นหาและอนุญาตได้อย่างมีประสิทธิภาพมากขึ้น ต่างจาก cDTM ที่ทำการอนุญาตการพัฒนาของหัวข้อในชุดข้อมูล]



รูปที่ 2.24 PGM ของ Continuous Dynamic Topic Model

ผู้วิจัยทำการทดลองโดยการใช้ชุดข้อมูลสองชุดประกอบด้วยชุด AP ซึ่งเป็นข้อมูลหัวข่าวที่มีการระบุเวลาด้วยความละเอียดหลักชั่วโมง ตั้งแต่วันที่ 1 พฤษภาคม 1988 ถึง 30 มิถุนายน 1988 และชุด Election 08 ข้อมูลเกี่ยวกับการเลือกตั้งประธานาธิบดีสหรัฐอเมริกาในช่วงปี 2008 ด้วยความละเอียดเวลาหลักวัน ตั้งแต่วันที่ 27 กุมภาพันธ์ 2007 ถึง 22 กุมภาพันธ์ 2008 และสำหรับชุด Election ตั้งแต่ 26 เมษายน 2007 ถึง 22 กุมภาพันธ์ 2008 ผู้วิจัยได้ทำการสร้างแบบจำลองด้วยความละเอียดเวลาหลักรายระดับและเพื่อให้สามารถเบริร์ยนเทียบแบบจำลองได้ ผู้วิจัยเลือกวัดผลด้วยค่า Per-Word Perplexity โดยการคำนวณ Per-Word Perplexity ของข้อมูลปัจจุบัน ด้วยการใช้

ชุดข้อมูลก่อนหน้า หรือข้อมูลในอดีต สำหรับข้อมูลชุด AP ผู้วิจัยทำนายตั้งแต่ 15 พฤษภาคม 1988 ถึง 29 พฤษภาคม 1988 และสำหรับชุด Election ตั้งแต่ 26 เมษายน 2007 ถึง 22 กุมภาพันธ์ 2008 โดยได้ผลลัพธ์ของค่า per-word perplexity ของแบบจำลองที่สร้างโดยกำหนดจำนวนหัวข้อ 1, 3, 5, และ 10 หัวข้อตามลำดับดังรูปที่ 2.25

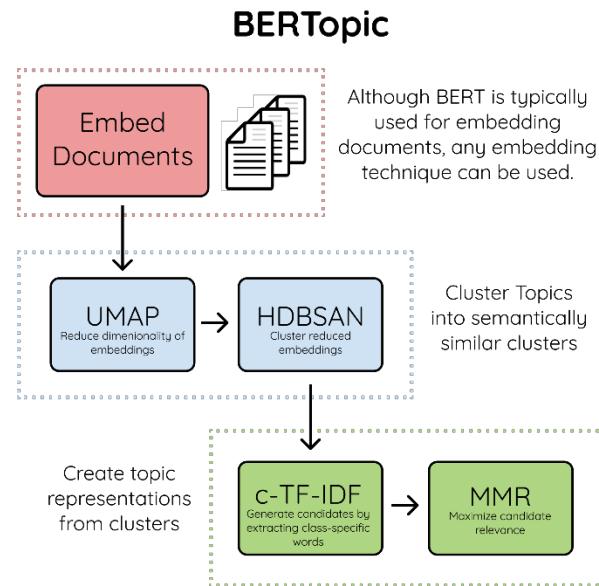


รูปที่ 2.25 ผลลัพธ์การวัดผลแบบจำลอง cDTM แต่ละแบบจำลอง

ผู้วิจัยได้สุ่มตัวอย่างผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยชุดข้อมูล Election 08 ที่ความละเอียดระดับสัปดาห์ ทำการสุ่มหัวข้อทุก ๆ สองเดือน โดยในช่วงเริ่มต้นการเลือกตั้งปี 2007 ได้หัวข้อ Healthcare เมื่อมีการแข่งขันทางการเมืองมากขึ้นในปี 2008 หัวข้อที่ได้จะเกี่ยวกับผู้ลังสมัครเลือกตั้งและมีการเปลี่ยนแปลงอย่างรวดเร็ว นอกจากนี้ผู้วิจัยยังได้ทำการทดลองการใช้แบบจำลอง cDTM สำหรับการทำนายเวลาที่เกิดขึ้นของเอกสารแต่ละชุด โดยมีแนวทางการทำนายคือแบบ Flat คือการให้แบบจำลองต่างความละเอียดเวลาทำงานเวลาให้ดีที่สุด และแบบ Hierarchical หรือการทำนายโดยเริ่มจากแบบจำลองที่มีความละเอียดเวลามากแล้วเพิ่มความละเอียดตามลำดับ เช่น ในการทำนายวัน ต้องทำนายเดือนที่ดีที่สุด จากนั้นทำนายสัปดาห์ที่ดีที่สุดจากเดือนที่ได้ ถึงจะทำนายวันที่ดีที่สุดจากแบบจำลอง โดยพบว่าการทำนายแบบ Hierarchical ได้ผลดีที่สุด

- Maarten Grootendorst [21] เสนอแนวทางการสร้างแบบจำลองหัวข้อแบบใหม่ โดยอาศัยโครงสร้างของแบบจำลองที่ทันสมัยและให้ประสิทธิภาพสูงอย่าง BERT ในการแปลงข้อมูลข้อความเอกสารเป็นเวกเตอร์ที่สามารถมีคุณสมบัติในการแสดงความสัมพันธ์เชิงบริบทและความหมายของคำในข้อมูลและมีการลดมิติข้อมูลด้วยเทคนิค UMAP BERTopic เลือกใช้แบบจำลองประเภท Clustering เช่น K-Mean

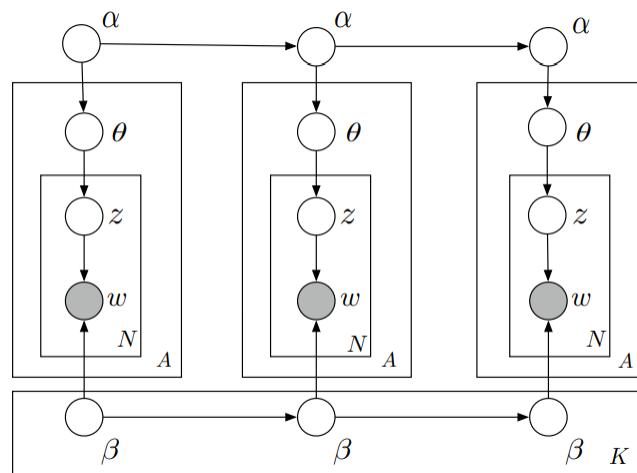
Clustering, DBScan และ HDBScan ในการจัดกลุ่มเอกสารเข้าเป็นหัวข้อเดียวกัน ต่างจากการสร้างแบบจำลองหัวข้อแบบดั้งเดิม เช่น LDA, LSA หรือ PLSA ที่ใช้วิธีเชิงสถิติมากกว่า BERTopic ใช้เทคนิค cTF-IDF ให้การสกัดค่าสถิติของคำในแต่ละหัวข้อ เพื่อประกอบคำนั้น ๆ เป็นหัวข้อที่พบรากการสร้างแบบจำลองหัวข้อ โดยมีตัวอย่างขั้นตอนการทำงานของ BERTopic ดังรูปที่ 2.21 ผู้วิจัยได้ทำการทดลองสร้างแบบจำลองหัวข้อต่างประเภท เช่น LDA, NMF, Top2Vec และ BERTopic ด้วยข้อมูลที่แตกต่างกันสามชุดประกอบด้วย 20 NewsGroups, BBC News และ Trump เพื่อเปรียบเทียบประสิทธิภาพด้วยค่า Topic Coherences พบว่า BERTopic ได้ค่า Coherences ที่สูงลำดับทุกชุดข้อมูล ได้ประสิทธิภาพดีที่สุดลำดับชุดข้อมูล Trump และประสิทธิภาพสามารถแบ่งขั้นกับแบบจำลองอื่นได้ลำดับชุดข้อมูลอิกซ์โซนชุดทึ้งนี้ประสิทธิภาพของการสร้างแบบจำลอง BERTopic ขึ้นอยู่กับการเลือกใช้โครงสร้างแบบจำลอง BERT สำหรับการ Embed ข้อมูลอิกด้วย ผู้วิจัยสรุปว่า BERTTopic มีความเสถียรและสามารถให้ประสิทธิภาพได้ดีกว่าการสร้างแบบจำลองหัวข้อแบบดั้งเดิม



รูปที่ 2.26 การทำงานของแบบจำลอง BERTopic

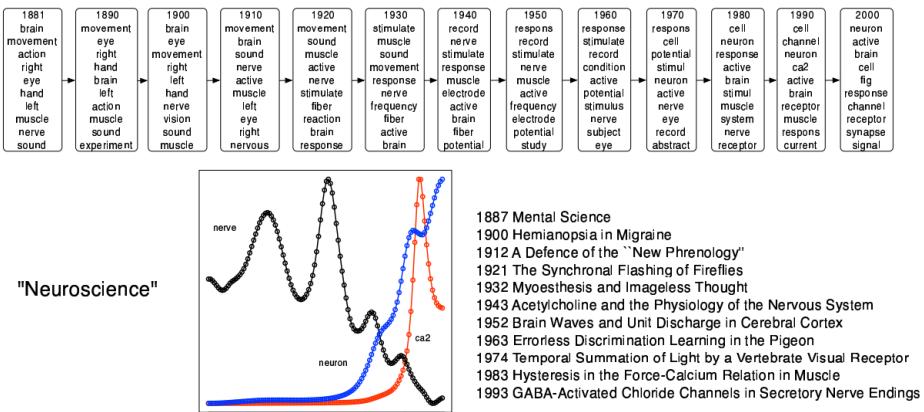
- Blei and Lafferty [22] ได้เสนอ Dynamic Topic Models เป็นแบบจำลองในการวิเคราะห์พัฒนาการของหัวข้อเมื่อเวลาที่เกิดขึ้นของชุดข้อมูลเปลี่ยนแปลงไป ซึ่งเป็นข้อจำกัดของแบบจำลองหัวข้อแบบดั้งเดิม ผู้วิจัยได้กล่าวถึงลักษณะธรรมชาติของข้อมูลที่มีการพัฒนาและเปลี่ยนไปอยู่ตลอดเวลา ทำให้เป็นที่มาของการสร้าง

แบบจำลอง Dynamic Topic Model ซึ่งเป็นแบบจำลองหัวข้อที่มีการต่อஇອดณาจากแบบจำลอง LDA มีการคำนึงถึงปัจจัยเวลาเป็นตัวแปรไม่ต่อเนื่อง และใช้สมมติฐานว่าหัวข้อที่อู้ฟู่ในเวลาแต่ละช่วงมีการเข้ามายิงและพัฒนามาจากหัวข้อจากช่วงเวลาก่อนหน้า โดยช่วงเวลาในที่นี่เป็นช่วงแบบไม่ต่อเนื่องที่ต้องกำหนด เช่น เดือนหรือวัน เพื่อการสำรวจพัฒนาการและการเปลี่ยนไปของหัวข้อมือเวลาผ่านไป โดยมีการอธิบายโครงสร้างของแบบจำลองด้วย PGM ดังรูปที่ 2.27



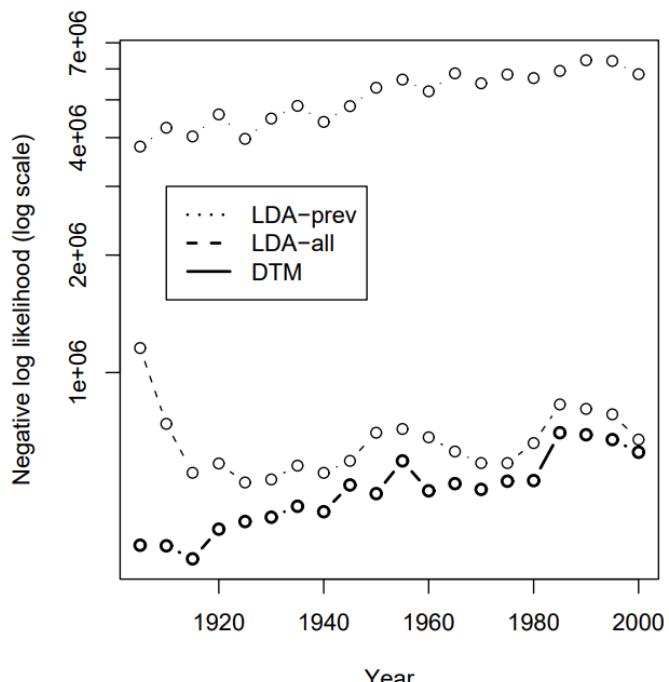
รูปที่ 2.27 PGM ของแบบจำลอง Dynamic Topic Model

ผู้วิจัยทำการทดลองโดยใช้ชุดข้อมูลบทความวิทยาศาสตร์ตลอด 100 ปีที่ได้จากการ OCR หรือการสกัดข้อความจากเอกสารกระดาษเข้าคอมพิวเตอร์ ขนาดข้อมูลประกอบด้วยคำทั้งหมด ประมาณ 7.5 ล้านคำ และทำการสร้าง แบบจำลอง Dynamic Topic Model ด้วยการกำหนดจำนวนหัวข้อ 20 หัวข้อ โดยมีตัวอย่างผลลัพธ์หัวข้อ Neuroscience ดังรูปที่ 2.27 ซึ่งมีการเปลี่ยนไปตามเวลา



รูปที่ 2.28 ตัวอย่างผลลัพธ์หัวข้อจากแบบจำลอง Dynamic Topic Model

สำหรับการวัดผลแบบจำลองในเชิงตัวเลข ผู้วิจัยให้แบบจำลองทำการทำงานหัวข้อของปีลักษณะโดยให้ข้อมูลนำเข้าเป็นบทความจากปีก่อนหน้า และทำการเปรียบเทียบประสิทธิภาพกับแบบจำลองหัวข้ออื่นด้วยการทำงานแบบเดียวกันดังรูปที่ 2.29 ผู้วิจัยพบว่า Dynamic Topic Model ทำได้ดีเมื่อเทียบกับแบบจำลองอื่น



รูปที่ 2.29 การเปรียบเทียบแบบจำลองหัวข้ออื่นกับ DTM

## บทที่ 3

### วิธีการดำเนินการวิจัย

ผู้วิจัยได้ทำการแบ่งการดำเนินงานวิจัยออกเป็นสองส่วน ส่วนแรกเป็นการศึกษากระบวนการ การทำงาน ของแบบจำลองหัวข้อช้อนเร้นกับข้อมูลข้อความขนาดปกติ และข้อมูลข้อความขนาดเล็ก ได้แก่ Latent Dirichlet Allocation(LDA), Gibbs Sampling for Dirichlet Multinomial Mixture, และ Non-negative Matrix Factorization เครื่องมือทางสถิติที่ใช้ในการวัดประสิทธิภาพอย่าง Topic Coherence รวมถึงการปรับค่าตัวแปรของแบบจำลอง และส่วนที่สองคืองานวิจัยที่ต่อเนื่องมาจาก งานวิจัยหลัก โดยจะเป็นขั้นตอน การทำงานต่อยอดกับข้อมูล ข้อความขนาดสั้นภาษาไทยจากความ คิดเห็นสาธารณะที่เก็บจากแพลตฟอร์มทวิตเตอร์ (Twitter) รวมถึงทดลองการพิจารณา ถึงปัจจัยเวลา ในการหาหัวข้อช้อนเร้นจากข้อมูล

#### 3.1 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 1

##### 3.1.1 การทำความเข้าใจธุรกิจ (Business Understanding)

ปัจจุบันมีการทำ Social Listening จากหลากหลายช่องทาง จากความคิดเห็นของผู้คน ที่เกิดขึ้น อย่างต่อเนื่องจากต่างสถานที่ และเวลา แต่ข้อมูลที่ได้จากการรวบรวมข้อมูลวิธี ดังกล่าวไม่สามารถระบุตำแหน่งได้อย่างชัดเจน ทั้งยังอาจเกิดการปลอมแปลงได้

วิทยานิพนธ์ฉบับนี้ต้องการรวบรวมข้อมูลความคิดเห็นความรู้สึกของสิ่งที่เกิดขึ้นหรือ ได้สมผัสของผู้คน ณ เวลาและสถานที่ต่าง ๆ และทำการหาหัวข้อช้อนเร้น จากข้อมูลความ คิดเห็นดังกล่าว เพื่อระบุแนวความคิดของผู้คนโดยพิจารณาถึงปัจจัย ซึ่งนำไปสู่การสำรวจ พฤติกรรมของผู้คนต่อสถานที่ ได้อย่างมีประสิทธิภาพมากขึ้น

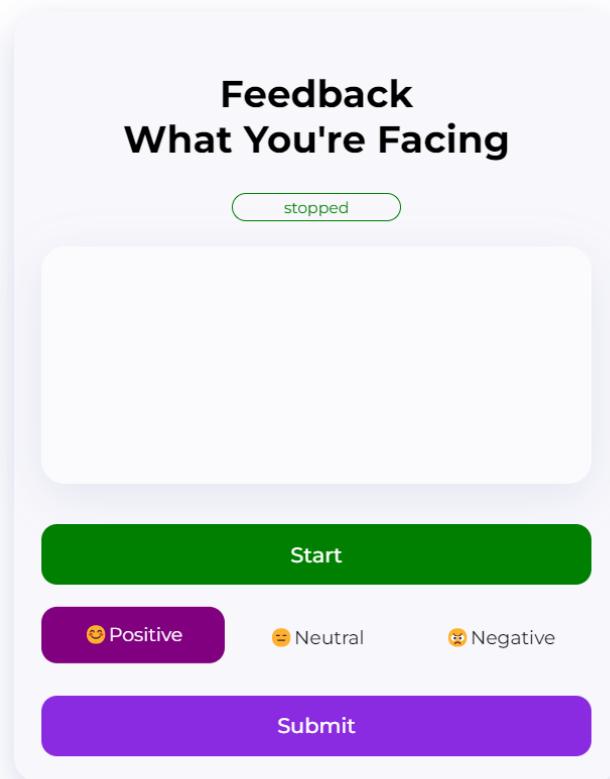
### 3.1.2 การทำความเข้าใจข้อมูล (Data Understanding)

ชุดข้อมูลที่นำมาใช้ในงานวิจัยนี้ ประกอบด้วย

- AG News [23] เป็นข้อมูลที่รวบรวมบทความจากแหล่งข่าวมากกว่า 2000 แห่ง จำนวน 196,000 Records โดยมีข้อมูลเนื้อข่าวเป็นหลัก
- Twitter COVID-19 [24] ข้อมูลทวีตที่มีความเกี่ยวข้องกับวิกฤต COVID-19 ประกอบด้วย ข้อมูลคือ
  - a) tweet\_id ไอดีของทวีต
  - b) date วันที่ทวีต
  - c) time เวลาที่มีการทวีต
  - d) lang ภาษาของทวีต
  - e) country\_code ตัวอักษรแสดงประเทศที่ทวีตถูกเขียน

ซึ่งจะต้องมีการนำ tweet\_id ไปใช้ในการดึงข้อมูลจาก Twitter โดยผ่านทางชุดเครื่องมือ Social Media Mining Toolkit (SMMT) [25] ในการดึงข้อมูลด้วย Twitter API ซึ่งข้อมูลส่วนที่นำมาใช้หลังจากการดึงมาแล้วคือเนื้อหาของ Tweets

- ข้อมูลที่ทำการเก็บผ่านแอปพลิเคชันที่จัดทำขึ้น เพื่อเก็บข้อมูลความคิดเห็นของผู้คน โดยมีหน้าอินเตอร์เฟสเบื้องต้นดังรูปที่ 3.1



รูปที่ 3.1 ตัวอย่างหน้าแอปพลิเคชันที่ใช้ในการเก็บข้อมูล

โดยข้อมูลที่ทำการเก็บมีดังนี้

- a) Latitude (ละติจูด)
  - b) Longitude (ลองจิจูด)
  - c) Timestamp (ໄວລາ)
  - d) Voice Transcript (ข้อความคิดเห็นของผู้พูด)
  - e) Sentiment (อารมณ์ของผู้พูด)

### 3.1.3 การเตรียมข้อมูล (Data Preparation)

#### 1) การทำความสะอาดข้อมูล (Data Cleaning)

ทำการตัดคำชื่อของอักขระพิเศษ (Stopwords) รวมถึงตัวอักษรพิเศษ จากข้อมูลเพื่อไม่ให้เป็นการรบกวนและทำให้การสร้างแบบจำลองมีข้อผิดพลาดจากลักษณะของ Stopwords และอักษรพิเศษที่มีซ้ำกันอยู่ในข้อความ

#### 2) การตัดคำ (Word Tokenization)

การตัดคำเป็นการตัดคำออกให้อยู่ในรูปแบบรายการของคำ จากข้อมูลตัวอักษร หรือคำที่อยู่ในลักษณะข้อความขนาดใหญ่หรือประโยค

#### 3) การแปลงข้อมูล (Bag-of-Word Transformation)

การแปลงข้อมูลให้อยู่ในรูปแบบเวกเตอร์กระเพาคำ (Bag of Word) เพื่อที่จะสามารถนำไปประมวลผลและฝึกสอนแบบจำลองได้

จากชุดข้อมูลทุกชุดที่วิจัยฉบับนี้เลือกใช้เป็นชุดข้อมูลภาษาอังกฤษ จึงมีการเลือกใช้ตัวตัดคำที่เป็นที่นิยมโดยการใช้เครื่องมือจากไลบรารี NLTK เพื่อลดรูปคำให้อยู่ในแบบพื้นฐาน

### 3.1.4 การสร้างแบบจำลอง (Modeling)

ผู้จัดได้ฝึกฝนการใช้งานแบบจำลองหัวข้อและศึกษาผลลัพธ์ของการสร้างแบบจำลองหัวข้อ โดยประกอบด้วยแบบจำลอง Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) โดยใช้ไลบรารี Gensim และแบบจำลอง Gibbs Sampling for Dirichlet Multinomial Mixture (GSDMM) ในการสร้างแบบจำลองหัวข้อ Latent Dirichlet Allocation (LDA) โดยการสร้างจากคลาส LdaModel ของไลบรารี Gensim โดยการสร้างแบบจำลอง LDA ผู้จัดทำได้ตั้งค่าหัวข้อที่ต้องการจากการสร้างแบบจำลองเริ่มต้นไว้ที่ 5 หัวข้อ เนื่องจากค่าพื้นฐานถูกตั้งไว้ที่ 100 หัวข้อ ยกต่อการทำความสะอาดข้อความ เช่น การนำคำ Alpha และ ETA ไว้เป็น Auto ซึ่งหมายความว่าแบบจำลองจะทำเรียนรู้การกระจายแบบ Dirichlet ของหัวข้อและคำแบบไม่สมมาตร โดยอัตโนมัติ

ในการสร้างแบบจำลอง Gibbs Sampling for Dirichlet Mixture Model (GSDMM) ทำการสร้างจากคลาส Movie Group Process ของไลบรารี GSDMM โดยผู้จัดทำตั้งค่าแบบจำลอง GSDMM ไว้เป็นค่าพื้นฐานทั้งหมด โดยมีรายละเอียดเบื้องต้นคือจำนวนหัวข้อที่ต้องการคือ 8 หัวข้อ การตั้งค่า Alpha และ Beta ซึ่งเป็นตัวควบคุมการกระจายแบบ Dirichlet ของแบบจำลอง DMM ไว้ที่ 0.1 ตามค่าพื้นฐานที่ผู้จัดทำตั้งไว้

ในการสร้างแบบจำลอง Non-Negative Matrix Factorization (NMF) ทำการสร้างจากคลาส NMF ของไลบรารี Gensim โดยการผู้จัดทำได้ทำการตั้งค่าจำนวนหัวข้อไว้ที่ 5 หัวข้อ เนื่องจากค่าพื้นฐานถูกตั้งไว้ที่ 100 หัวข้อ ทำให้ยากต่อการทำความเข้าใจและตีความ โดยมุ่ยย์

### 3.1.5 การประเมินผลแบบจำลอง (Evaluation)

ในการประเมินผลการสร้างแบบจำลองการสร้างแบบจำลองหัวข้อแบบ LDA, GSDMM และ NMF ด้วยผู้จัดทำได้ทำการตั้งค่าจำนวนหัวข้อที่แตกต่างกันคือชุด AG News และ Twitter COVID-19 และทำการประเมินผลแบบจำลองหัวข้อโดยการใช้เครื่องมือสถิติจากคลาส CoherenceModel ของไลบรารี Gensim ในการประเมินคะแนนความสอดคล้องของหัวข้อที่ได้จากการสร้างแบบจำลองหัวข้อโดยมีวิธีการคำนวณค่าความสอดคล้องที่แตกต่างกัน (UMASS, UCI, NPMI และ C\_V)

ในการทดลองผู้จัดทำได้ทำการเปลี่ยนค่าจำนวนหัวข้อที่กำหนดสำหรับแบบจำลอง และทำการประเมินค่าความสอดคล้อง โดยค่าจำนวนหัวข้อที่เลือกใช้การทดลองสร้างแบบจำลองหัวข้อคือ 2 หัวข้อจนถึง 9 หัวข้อ โดยเพิ่มขึ้นทีละหนึ่ง เนื่องจากเมื่อทดลองเพิ่มจำนวนหัวข้อขึ้นไป ค่า Coherences มีแนวโน้มที่จะเพิ่มขึ้นอย่างต่อเนื่อง จึงเลือกทำการทดลองในช่วงจำนวนหัวข้อที่เห็นการเปลี่ยนแปลงของค่า Coherences ได้อย่างชัดเจนที่สุด เพื่อที่จะใช้ Elbow Techniques ในการสังเกตการณ์จำนวนหัวข้อที่เหมาะสมสำหรับการค้นหาจำนวนหัวข้อที่เหมาะสมสำหรับแต่ละแบบจำลองและชุดข้อมูล หรือหมายความว่าเป็นจำนวนหัวข้อที่ทำให้หัวข้อที่ได้จากการสกัดหัวหัวข้อนั้นมีความสอดคล้องกัน สามารถตีความได้โดยมุ่ยย์

### 3.2 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 2

จากวัตถุประสงค์ของการค้นหาหัวข้อประกอบกับการพิจารณาปัจจัยเวลาและสถานที่ ปรากฏว่าข้อมูลส่วนที่ทำการเก็บด้วยแอปพลิเคชันนั้นมีจำนวนไม่เพียงพอต่อการสร้างแบบจำลองหัวข้อที่จะสามารถตีความและจำแนกเนื้อหาได้ ทำให้ต้องมีการเปลี่ยนทิศทางการทดลอง โดยยังคงเป้าหมายการสำรวจความคิดเห็นของผู้คนผ่านโซเชียลมีเดียแทน ผู้วิจัยเลือก Twitter เป็นแหล่งข้อมูลสมบัติจากผู้แสดงความคิดเห็นอย่างเช่น ข้อความ เวลาและสถานที่ตามที่ต้องการ นอกเหนือไป之外 ได้ทำการศึกษางานวิจัยเพิ่มเติม ได้พบกับวิธีการสร้างแบบจำลองหัวข้อที่ทันสมัยโดยอาศัยแบบจำลองภาษาที่เป็นที่นิยมอย่าง BERT เข้ามาช่วย ประกอบกับเทคนิคการพิจารณาปัจจัยเวลาในการสร้างแบบจำลองหัวข้อเข้าด้วยกัน เพื่อค้นหาพัฒนาการของหัวข้อภายในชุดข้อมูล ผู้วิจัยจึงได้สนใจที่จะทดลองต่อข้อเสนอแบบจำลองดังกล่าวด้วยวัตถุประสงค์ที่ต่างออกไปคือ การสำรวจการเปลี่ยนไป การเกิดขึ้นใหม่ หรือพัฒนาการของหัวข้อเมื่อเวลาผ่านไป และการค้นหาวิธีการทดลองที่ให้ผลลัพธ์หัวข้อที่มีความสามารถในการตีความและสอดคล้องกัน ได้สูงที่สุด

#### 3.2.1 การทำความเข้าใจข้อมูล (Data Understanding)

ข้อมูลความคิดเห็นสาธารณะที่ได้ทำการรวบรวมโดยการใช้แอปพลิเคชันที่ผู้จัดทำได้สร้างขึ้นโดยเฉพาะ ซึ่งมีข้อมูลที่เก็บคือ

- ข้อมูลความคิดเห็น
- ข้อมูลสถานที่ (ละติจูด, ลองจิจูด)
- ข้อมูลเวลาที่แสดงความคิดเห็น
- ข้อมูลการตอบสนองทางอารมณ์ต่อความคิดเห็นนั้น ๆ

ส่วนข้อมูลความคิดเห็นจาก Twitter จากการใช้ Tweets API ของ Twitter Developers และหา Hashtag ยอดนิยมที่สูงระหว่างวันที่ 3 – 9 พฤษภาคม 2022 จากเว็บไซต์ Getdaytrends จำนวน 10 แท็ก ซึ่ง Hashtag ดังกล่าวมีเนื้อหาที่แตกต่างกันค่อนข้างชัดเจน มีช่วงเวลาที่ถูกพูดถึงแตกต่างกัน ทำให้มีความเหมาะสมในการนำมาทดลองสร้างแบบจำลองเพื่อค้นหาหัวข้อที่มีเวลาเป็นปัจจัยประกอบ เพื่อใช้ในการระบุเนื้อหาของความคิดเห็นที่แตกต่างกันตามแต่ละแท็ก ประกอบด้วย

#เด็ก18岁่าเด็ก13 #จันทร์ปราภา #ลอยกระฟง #PunBNK48 #คุณชายEP12 #MILLI #วาคนด้า  
งเจริญ #راكแก้วep6 #4EVE และ #ยูครอน ซึ่งมีข้อมูลที่เก็บคือ

- ข้อมูลความคิดเห็น
- ข้อมูลเวลาที่ทวีต
- ข้อมูลหัวข้อที่กล่าวถึง
- ข้อมูลสถานที่ (ละติจูด, ลองจิจูด)

และเมื่อทำการพิจารณาคุณลักษณะของข้อมูลพบว่าข้อมูลสถานที่พบน้อยมากหรือ  
หมายความว่าผู้แสดงความคิดเห็นไม่ประสงค์ที่จะให้ข้อมูลสถานที่สำหรับการแสดงความ  
คิดเห็นบน Twitter จึงพิจารณาตัดสถานที่ออกจากกระบวนการนำทางดลง

### 3.2.2 การเตรียมข้อมูล (Data Preparation)

#### 1) การทำความสะอาดข้อมูล (Data Cleaning)

เนื่องจากข้อมูลความคิดเห็นที่ดึงออกมาจาก Twitter มีข้อความที่ไม่เป็นแบบ  
แผน เราจึงนำข้อมูลดังกล่าวมาตัดคำ (Tokenization) และทำการลบคำที่ไม่ช่วยใน  
การสื่อความหมายของข้อความในภาพรวม (Stopwords) ตัวเลข อีโมจิหรือไอคอน  
ตัวอักษรพิเศษ และตัวอักษรภาษาเยียงกุยออกจากข้อความไป จากนั้นนำคำที่  
เหลืออยู่ในข้อความมาต่อกันให้เป็นข้อความอีกครั้ง

#### 2) การแปลงให้อยู่ในรูปเวกเตอร์ (Sentences Transformation)

ทำการแปลงชุดข้อมูลที่ได้ทั้งหมดให้อยู่ในรูปเวกเตอร์ เพื่อให้สามารถนำไป  
ประมวลผลต่อได้ โดยเลือกใช้วิธีการแปลงข้อมูลแบบ Sentence Embedding ซึ่ง  
เป็นเทคนิคที่ทันสมัย มีประสิทธิภาพสูงในการสร้างเวกเตอร์ที่สามารถแสดง  
ความสัมพันธ์เชิงบริบทและความหมายของข้อความ และสามารถนำมาใช้ในงาน  
ด้าน Clustering ได้ ผลลัพธ์ของการแปลงข้อมูลข้อความหนึ่งประจำให้อยู่ในรูป  
เวกเตอร์ด้วยเทคนิคดังกล่าวจะได้ออกมาเป็นเวกเตอร์ที่มีขนาดมิติคงที่ 768  
คุณสมบัติ มาจากจำนวนพารามิเตอร์ของ Embedding Layer ภายในแบบจำลอง

BERT ขนาดพื้นฐาน และสำหรับการใช้งานเทคนิค Sentence Embedding จะต้องมีการเลือกใช้แบบจำลองทางภาษาที่มีการฝึกสอนล่วงหน้ามา ผู้วิจัยเลือกใช้แบบจำลอง simcse-model-roberta-base-thai เนื่องจากมีประสิทธิภาพสูงเมื่อเทียบกับแบบจำลองอื่น ๆ ข้างต้นจาก benchmark [26] โดยทำการเรียกใช้ผ่านไลบรารี sentence\_transformer

### 3) การลดมิติข้อมูล (Dimensionality Reduction)

เนื่องจากปัจจัยมิติของเวกเตอร์มีขนาดสูง ทำให้ใช้ระยะเวลาในการสร้างแบบจำลองนาน ผู้วิจัยจึงเลือกทำการลดมิติข้อมูล โดยการใช้เทคนิค UMAP ซึ่งทำการตัดต่อโดยให้  $n\_components$  คือ 19 หรือการแปลงเป็นเวกเตอร์ที่มี 19 มิติ เนื่องจากปัจจัยด้านประสิทธิภาพในการประมวลผล นอกจากนี้ผู้วิจัยได้ทดลองแปลงเป็นเวกเตอร์ขนาด 9 มิติ แล้วพบว่าผลลัพธ์หัวข้อสามารถตีความได้ยาก เป็นเหตุผลให้เลือกใช้ขนาดมิติปัจจุบัน

### 4) การเพิ่มปัจจัยเวลาเป็นปัจจัยประกอบ

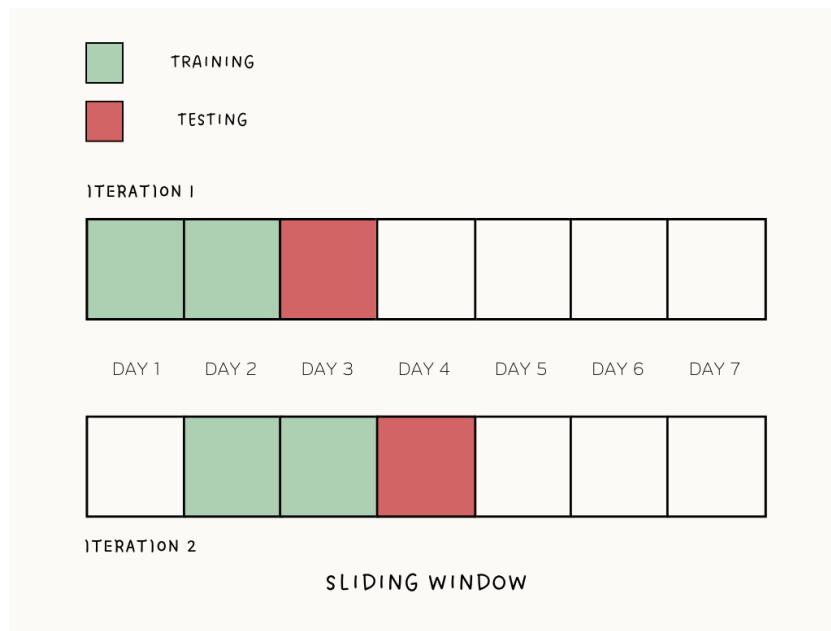
ส่วนของการเพิ่มปัจจัยเวลาเป็นปัจจัยประกอบ ผู้วิจัยเลือกที่จะแปลงข้อมูลเวลาให้อยู่ในมาตรฐาน Unix Timestamp และได้เพิ่มเป็นอีกคุณสมบัติ (feature) หนึ่งของเวกเตอร์ หลังจากนั้นทำการ Normalize เวกเตอร์ที่เพิ่มปัจจัยเวลาเพื่อให้ช่วงข้อมูลอยู่ในช่วงเดียวกัน

#### 3.2.3 การสร้างแบบจำลอง (Modeling)

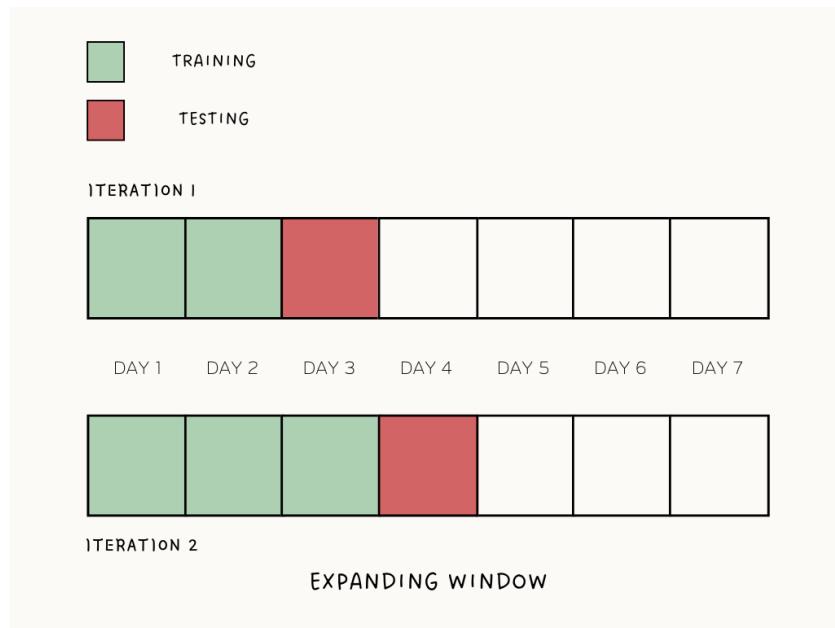
จาก BERTopic [21] ที่มีความสามารถในการสร้างแบบจำลองหัวข้อที่มีประสิทธิภาพรวมถึงการสร้างแบบจำลองหัวข้อแบบ Dynamic Topic Model [22] ที่มีความสามารถในการพิจารณาเวลาเป็นปัจจัยประกอบเพื่อสำรวจพัฒนาการของหัวข้อตลอดชุดข้อมูล เนื่องด้วยวิธีที่ BERTopic สร้างแบบจำลองหัวข้อคือการใช้ข้อมูลฝึกสอน และใช้แบบจำลองสำหรับการค้นหาหัวข้อสำหรับชุดข้อมูลที่ไม่เคยพบมาก่อน ซึ่งเทียบเท่ากับวิธีการทดลองแบบ Expanding Window รวมถึงมีการพิจารณาเวลาในช่วงที่คาดหวังระดับเดือนถึงปี และสมมติฐานว่าหัวข้อในแต่ละช่วงเวลาพัฒนามาจากช่วงเวลา ก่อนหน้า วัตถุประสงค์ดังกล่าวต่างกับการทดลองใน

งานวิจัยฉบับนี้ ผู้วิจัยต้องการสำรวจการเปลี่ยนไปของหัวข้อในระยะสั้นระดับวัน และเชื่อว่า หัวข้อมีการเปลี่ยนไปและสามารถเกิดขึ้นใหม่ได้สำหรับแต่ละช่วงเวลา ทำให้นำมาซึ่งการทดลองเปรียบเทียบวิธีการสร้างแบบจำลองที่แตกต่างกันออกໄປ ผู้วิจัยทำการพัฒนาต่อขอดมา จาก BERTopic โดยการใช้วิธีการแปลงเป็นเวกเตอร์โดยอาศัยแบบจำลอง BERT การลดมิติ ข้อมูลด้วย UMAP การแบบจำลองประเภท Clustering ในการจัดกลุ่มหัวข้อ เช่นเดียวกับ BERTopic และต้องการพิจารณาค่าด้วยวิธีที่แตกต่างระหว่างวิธี Sliding Window ซึ่งเป็นวิธีที่ผู้วิจัยคาดว่าจะมีความสามารถในการค้นหาและให้ผลลัพธ์หัวข้อสำหรับระยะเวลาสั้น ได้ดีกว่า และ Expanding Window ซึ่งคล้ายกับวิธีของ BERTopic รวมถึงพิจารณาเปรียบเทียบ ความสำคัญของปัจจัยเวลาเมื่อทดลองด้วยวิธีดังกล่าว ด้วยการทดลองใช้และไม่ใช้เวลาเป็นปัจจัยประกอบ การทดลองทั้งหมดข้างต้นจะนำໄປสู่การค้นพบความแตกต่างของผลลัพธ์ที่ได้จากการทดลอง และวิธีการสร้างแบบจำลองที่ให้ผลลัพธ์ดีที่สุด

ผู้วิจัยทำการสร้างแบบจำลองหัวข้อในเชิงเปรียบเทียบทั้งแบบใช้เวลาเป็นปัจจัยประกอบ และไม่ใช้เวลาเป็นปัจจัยประกอบ ซึ่งเป็นการเพิ่มคุณสมบัติเข้าไปในเวกเตอร์ต่างกัน BERTopic และด้วยวิธีที่ต่างกันสองวิธีประกอบด้วยวิธี Sliding Window ดังรูปที่ 3.2 และวิธี Expanding Window ดังรูปที่ 3.3 เพื่อทดสอบประสิทธิภาพและผลลัพธ์ที่แตกต่างกัน



รูปที่ 3.2 การสร้างแบบจำลองหัวข้อด้วยวิธี Sliding Window



รูปที่ 3.3 การสร้างแบบจำลองหัวข้อด้วยวิธี Expanding Window

โดยแบบจำลองที่เลือกใช้ในการค้นหาหัวข้อคือ K-Means Clustering เนื่องจากง่ายต่อการทำความเข้าใจและการทดลอง โดยสำหรับแต่ละเทคนิคข้างต้นที่จะใช้ในการสร้างแบบจำลอง ตลอดช่วงเวลาที่ทำการทดลองสำหรับการทดลองแบบ Sliding Window จะแบ่งเป็นข้อมูลชุด Train สำหรับการฝึกสอนแบบจำลองจำนวน 2 วัน และชุด Test 1 วัน ดังรูปที่ 3.2 เพื่อการทดสอบและค้นหาหัวข้อโดยใช้แบบจำลองที่ได้จากการฝึกสอนแล้ว และสำหรับวิธี Expanding Window การแบ่งข้อมูลชุด Train จะเริ่มที่สองวันแรกและเพิ่มขึ้นรอบละหนึ่งวัน และชุด Test 1 วัน ไปตามลำดับเวลา ดังรูปที่ 3.3 โดยการสร้างแบบจำลองจะทำการกำหนดค่า K ไว้ที่ 10 ซึ่งมาจากจำนวน Hashtag ที่ทำการเก็บซึ่งมีเนื้อหาที่แตกต่างกัน เพื่อการตีความและความเข้าใจได้ยามากขึ้น สำหรับพารามิเตอร์อื่นๆ ของแบบจำลองจะให้เป็นค่าพื้นฐาน

### 3.2.4 การเปรียบเทียบและการประเมินผล (Comparison & Evaluation)

ในส่วนของการประเมินผลผลลัพธ์ที่ได้จากการค้นหาหัวข้อ โดยใช้แบบจำลอง K-Means Clustering จะแบ่งออกเป็นสองส่วนด้วยกัน คือส่วนของการสำรวจและเปรียบเทียบการเปลี่ยนแปลงของผลลัพธ์หัวข้อที่ได้ระหว่างช่วงเวลาที่ทำการทดลอง และส่วนการเปรียบเทียบประเมินความสามารถในการตีความได้ของหัวข้อผลลัพธ์

#### 3.2.4.1 การสำรวจความเชื่อมโยงของหัวข้อด้วย Cosine Similarity

ในส่วนของการสำรวจและเปรียบเทียบลักษณะการเปลี่ยนแปลงของหัวข้อ ผู้วิจัยเลือกใช้ค่า Cosine Similarity เพื่อหาความคล้ายคลึงระหว่างคำที่เป็นองค์ประกอบของหัวข้อซึ่งชุดคำดังกล่าวจะได้มาจากการใช้เทคนิค TF-IDF ในการหาค่าสถิติของคำภายในหัวข้อหรือกึ่อแต่ละ Cluster ซึ่งการสำรวจความคล้ายคลึง จะนำไปสู่การทราบพัฒนาการของหัวข้อที่อาจมีการพัฒนาหรือเปลี่ยนผ่านไปตามช่วงเวลา ซึ่งจะทำโดยการคำนวณค่า Cosine Similarity ของหัวข้อระหว่างวัน

#### 3.2.4.2 การประเมินความสามารถในการตีความของหัวข้อด้วย Topic Coherences

ส่วนของการประเมินประสิทธิภาพหรือความสามารถในการตีความได้ของหัวข้อ จะทำโดยการใช้ค่า Topic Coherences โดยการใช้เครื่องมือจากคลาส CoherenceModel ของไลบรารี Gensim ในการประเมินความสอดคล้องของหัวข้อที่ได้จากการค้นหาหัวข้อในระหว่างช่วงเวลาที่ทำการทดลอง และระหว่างวิธีการสร้างแบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพของวิธีการค้นหาหัวข้อซ่อนเร้น โดยมีวิธีการคำนวณค่าความสอดคล้องที่แตกต่างกัน (UMASS, UCI, NPMI, C\_V)

ในงานวิจัยฉบับนี้จะมีการทดลอง 4 การทดลองเพื่อเปรียบเทียบผลลัพธ์หัวข้อ ทั้งในด้านการสำรวจและเปรียบเทียบลักษณะการเปลี่ยนแปลงของหัวข้อด้วยค่า Cosine Similarity และการเปรียบเทียบความสามารถในการตีความของหัวข้อด้วย ค่า Topic Coherences ระหว่างการทดลองแต่ละแบบ ประกอบด้วย

- การทดลองแบบที่ 1 ไม่ใช้เวลาเป็นปัจจัยประกอบ และสร้างแบบจำลองแบบ Sliding Window
- การทดลองแบบที่ 2 ไม่ใช้เวลาเป็นปัจจัยประกอบ และสร้างแบบจำลองแบบ Expanding Window
- การทดลองแบบที่ 3 ใช้เวลาเป็นปัจจัยประกอบ และสร้างแบบจำลองแบบ Sliding Window
- การทดลองแบบที่ 4 ใช้เวลาเป็นปัจจัยประกอบ และสร้างแบบจำลองแบบ Expanding Window

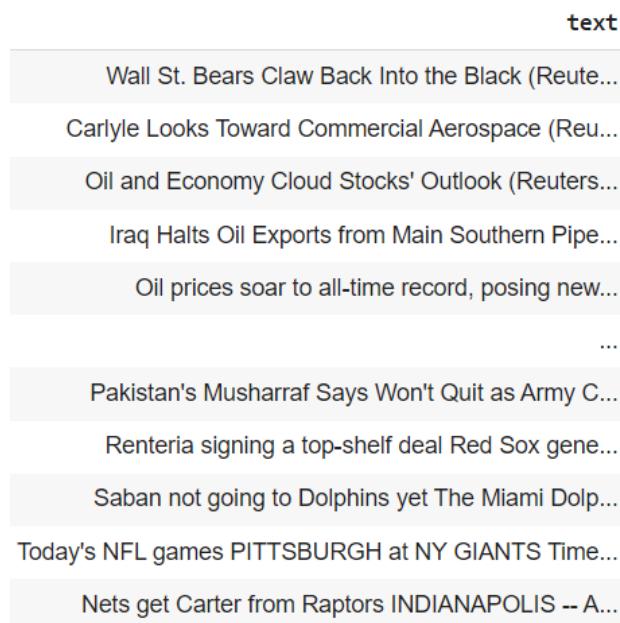
## บทที่ 4

### ผลการดำเนินงาน

#### 4.1 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 1

##### 4.1.1 ผลลัพธ์จากขั้นตอนที่ 3.1.2

จากการดำเนินงานขั้นตอนที่ 3.1.2 เป็นการนำข้อมูล AG News และ Twitter COVID-19 มาใช้งาน ซึ่งผู้วิจัยได้เริ่มจากการนำข้อมูล AG News มาใช้งานโดยการนำข้อมูลเข้าโดยใช้ไลบรารี datasets และได้ผลลัพธ์ดังรูปที่ 4.1



รูปที่ 4.1 ตัวอย่างชุดข้อมูล AG News

สำหรับการนำเข้าข้อมูล Twitter COVID-19 ชุดข้อมูลจะประกอบด้วย tweet\_id, date, time, และ lang จะเห็นได้ว่าทางผู้รวมรวมข้อมูลทำการรวม tweet\_id ที่เกี่ยวกับวิกฤต COVID-19 ไว้ดังรูปที่ 4.2 โดยให้ทางผู้ที่นำໄปใช้ทำการคัดข้อมูลเอง ทางผู้วิจัยได้ทำการใช้ชุดเครื่องมือ SMMT ซึ่งเป็นชุดเครื่องมือที่อำนวยความสะดวกในการรวบรวมข้อมูลจาก

แพลตฟอร์มสื่อสังคมออนไลน์อย่าง Twitter โดยผู้วิจัยได้ทำการคัดเลือก Tweets ที่เป็นภาษาอังกฤษเท่านั้นในการนำมาประมวลผลโดยมีตัวอย่างข้อมูลที่ได้ทำการดึงแล้วดังรูปที่ 4.3

	tweet_id	date	time	lang
0	1505031523711344643	2022-03-19	04:00:52	en
1	1505031527251394561	2022-03-19	04:00:52	en
2	1505031528551948293	2022-03-19	04:00:53	en
3	1505031529117982724	2022-03-19	04:00:53	en
4	1505031530783121408	2022-03-19	04:00:53	en

รูปที่ 4.2 ตัวอย่างชุดข้อมูล Twitter COVID-19

text
RT @CDCgov: New @CDCMMWR study finds #COVID19 vaccines continued to be highly effective at protecting adults from being put on a ventilator...
RT @BeingCharisBlog: CDC: "A person with any of the medical conditions listed below is more likely to get very sick with COVID-19. If you h...
RT @USMortality: WOW look at Vietnam! Basically no COVID-19 Cases AT ALL, until they started VAXXING!!! https://t.co/l5K8QfnGJV
RT @AppSame: 997,136#Covid19 deaths Biden is a week away from 1,000,000. That's 600,000 on his watch and just 20,000 away from the total nu...
RT @AppSame: 997,136#Covid19 deaths Biden is a week away from 1,000,000. That's 600,000 on his watch and just 20,000 away from the total nu...
...
RT @naearthiive: Announcement from #1BLOODFOR28thEARTHDAY \nDue to COVID-19 infection and its complications, I am not allowed for any outdoo...
Still one-way masking: inside retail spaces and in federal buildings. #MaskUp for those who can't get vaccinated.... https://t.co/O4txk9JLcN
RT @MarkusMannheim: About 15.4% of Canberrans — almost 1 in 6 of us — have tested positive for COVID-19 at some point during the pandemic....
RT @terminalcwo: We should all rally against this, but I know that won't happen. Some of you champion the suppression of legal rights and t...
RT @uche_blackstock: Now would be a great time to hear from local, state and federal officials about which mitigation policies they're putt...

รูปที่ 4.3 ตัวอย่างข้อความ Tweets

ทางผู้วิจัยได้ทำการเก็บข้อมูลผ่านทางแอปพลิเคชัน โดยมีตัวอย่างข้อมูลที่เก็บไว้ในฐานข้อมูลดังรูปที่ 4.4

longitude #	latitude #	fileName A	text A	timestamp □	sentiment A
100.5999	13.7842	audio-1651238922515.mpeg	แสดงเวลาทุกเช้านั่งเฉยๆแล้ว...	2022-04-29T13:28:42.515Z	Neutral
100.7328	13.9876	audio-1651230620544.mpeg	อย่าเขยบแผลเมย์เดียวสมอ Eck	2022-04-29T11:10:20.544Z	Positive
100.5017651	13.7563309	audio-1651229808729.mpeg	I love you มองไปเลีย	2022-04-29T10:56:48.729Z	
100.5017651	13.7563309	audio-1651229765461.mpeg	รักใจอาจไม่เดี๋ยวไปศรีนครินทร์	2022-04-29T10:56:05.461Z	
100.7328	13.9876	audio-1651229692712.mpeg	รักที่ไปตกใจหลังจากหน้าต่าง	2022-04-29T10:54:52.712Z	Positive
100.5017651	13.7563309	audio-1651229671964.mpeg	รักที่อาจไม่เดี๋ยวไปศรีนครินทร์	2022-04-29T10:54:31.964Z	
100.7259136	13.7326333	audio-1651222788471.mpeg	รัก	2022-04-29T08:59:48.471Z	
100.7258535375824	13.73257039742445	audio-1651143988011.mpeg		2022-04-28T11:06:28.011Z	Neutral
100.6239744	13.8051584	audio-1651073492036.mpeg	สวัสดิ์รักที่เจ้มส่องสดๆไปเลย...	2022-04-27T15:31:32.037Z	Positive
100.7640189874471	13.72758139399204	audio-1651052358033.mpeg		2022-04-27T09:39:18.033Z	
100.7661741	13.7358729	audio-1650602365419.mpeg	รักที่ของไม่ได้อยู่เบื้องรักแม่น้ำ...	2022-04-22T04:39:25.419Z	
100.7266323	13.7677636	audio-1650551970592.mpeg		2022-04-21T14:39:30.592Z	Neutral

รูปที่ 4.4 ตัวอย่างข้อมูลที่เก็บลงฐานข้อมูล

#### 4.1.2 ผลลัพธ์จากขั้นตอนที่ 3.1.3

จากการดำเนินงานขั้นตอนที่ 3.1.3 เป็นการเตรียมข้อมูล ซึ่งข้อมูลที่ผู้วิจัยได้ทำการเลือกมาฝึกฝนใช้งานแบบจำลองหัวข้อจำเป็นต้องมีการเตรียมข้อมูลเบื้องต้น โดยผู้วิจัยได้ทำการตัดคำเข้ามาร่วมที่อยู่ในข้อความ, ทำการตัดคำ (Tokenization), ลดรูปคำให้อยู่ในรูปเดิม (Lemmatization) และทำการเพิ่มคำที่อยู่ติดกันสองคำเข้าไปในชุดข้อมูลหรือการทำ Bi-gram โดยมีผลลัพธ์เบื้องต้นดังรูปที่ 4.5

```
[ 'wall',
  'st',
  'bears',
  'claw',
  'back',
  'black',
  'reuters',
  'reuters',
  'short',
  'sellers',
  'wall',
  'street',
  'dwindling',
  'band',
  'ultra',
  'cynics',
  'seeing',
  'green',
  'wall_st',
  'reuters_reuters',
  'wall_street']
```

รูปที่ 4.5 ตัวอย่างชุดข้อมูล AG News หลังผ่านการเตรียมข้อมูล

ในการสร้างแบบจำลองหัวข้อซึ่งเป็นแบบจำลองทางสถิติ จึงมีความจำเป็นที่จะต้องแปลงข้อมูลข้อความให้อยู่ในรูปแบบเวกเตอร์ หรือรูปแบบ Bag of Word ซึ่งผู้วิจัยได้ทำการแปลงข้อมูลโดยใช้เครื่องมือจากไลบรารี Gensim มาใช้งาน โดยมีผลลัพธ์ดังรูปที่ 4.6

`[(8, 2),  
(9, 1),  
(52, 3),  
(63, 1),  
(64, 1),  
(65, 1),  
(66, 1),  
(67, 1),  
(68, 1),  
(69, 1),  
(70, 1),  
(71, 1),  
(72, 2),  
(73, 2),  
(74, 1),  
(75, 1),  
(76, 1),  
(77, 1),  
(78, 2),  
(79, 1),  
(80, 1),  
(81, 1),  
(82, 1),  
(83, 2),  
(84, 1)],`

รูปที่ 4.6 ตัวอย่างชุดข้อมูล AG News หลังการทำ Bag-of-Words

#### 4.1.3 ผลลัพธ์จากขั้นตอนที่ 3.1.4

จากการดำเนินงานขั้นตอนที่ 3.1.4 เป็นการสร้างแบบจำลองหัวข้อ ผู้วิจัยได้ทำการเลือกแบบจำลองที่เป็นที่นิยมสำหรับการทำแบบจำลองหัวข้อสำหรับข้อมูลข้อความซึ่งประกอบด้วย LDA, GSDMM, และ NMF โดยใช้ชุดข้อมูลสองชุดคือ AG News และ Twitter Covid-19 โดยการทดสอบผู้วิจัยได้ทำการตั้งค่าแบบจำลองด้วยค่าพื้นฐาน ยกเว้นแบบจำลอง LDA และ NMF ที่ทำการตั้งค่าหัวข้อไว้ที่ 5 หัวข้อเนื่องจากค่าพื้นฐานอยู่ที่ 100 หัวข้อ ซึ่งยากต่อการจับใจความและแสดงผล

ผลลัพธ์การสร้างแบบจำลองหัวข้อด้วยชุดข้อมูล AG News

0	1	2	3	4
new	oil	game	said	iraq
company	reuters	new	reuters	afp
said	new	one	president	said
inc	year	quot	search	talks
reuters	york	two	government	united
corp	new_york	first	people	india
million	prices	season	google	world
microsoft	dollar	night	security	cup
software	percent	last	palestinian	union
com	stocks	coach	minister	south
billion	friday	time	officials	test
deal	wednesday	team	leader	two
fullquote	points	back	election	european
business	thursday	year	killed	iraqi

รูปที่ 4.7 ผลลัพธ์จากแบบจำลอง LDA ด้วยชุดข้อมูล AG News

0	1	2	3	4	5	6	7
new	said	iraq	reuters	game	world	oil	england
microsoft	reuters	said	said	new	cup	prices	league
software	new	reuters	inc	season	olympic	reuters	champions
internet	space	president	company	night	first	stocks	united
company	bush	minister	new	first	athens	new	test
search	nuclear	killed	fullquote	two	gold	oil_prices	first
said	quot	two	corp	sox	one	dollar	new
music	china	iraqi	profit	team	open	said	cup
service	president	baghdad	million	win	final	percent	arsenal
quot	iran	prime	quarter	red	team	york	win

รูปที่ 4.8 ผลลัพธ์จากแบบจำลอง GSDMM ค่าวิชุดข้อมูล AG News

0	1	2	3	4
reuters	new	said	quot	reuters
fullquote	oil	afp	year	new
stocks	prices	president	first	two
com	york	company	world	iraq
http	new_york	monday	one	reuters_reuters
www	oil_prices	tuesday	google	game
href	stocks	thursday	last	three
ticker	crude	wednesday	search	first
target	record	million	time	sunday
investor	barrel	officials	microsoft	win
aspx	high	government	inc	baghdad

รูปที่ 4.9 ผลลัพธ์จากแบบจำลอง NMF ค่าวิชุดข้อมูล AG News

### ผลลัพธ์จากการสร้างแบบจำลองด้วยชุดข้อมูล Twitter COVID-19

0	1	2	3	4
covid	https	breaking	drericding	https
fauci	new	additional	covid19	covid
https	covid	yet	need	state
restrictions	cases	infected	know	coronavirus_cases
may	coronavirus	want	covid19_hospitalizatio	bnodesk
surge	covid19	stop	drericding_need	hospital
says	last	covid	goto_hospital	hospitalizations
institute	week	covid19	infected_yet	ivermectin
fauci_says	increase	health	want_safe	largest
drsimonegold	reports	hear	know_cross	reduce
additional_surge	york	amp	hospitalizatio	date
restrictions_stop	reports_new	people	goto	trial
may_necessary	never	federal	china	number
drsimonegold_breaking	increase_last	would	two	hospitalizations_largest

รูปที่ 4.10 ผลลัพธ์จากแบบจำลอง LDA ด้วยชุดข้อมูล Twitter COVID-19

0	1	2	3	4	5	6	7
https	covid	covid	covid	https	china	covid	covid
covid	https	new	https	covid	deaths	https	high
covid19	cases	stop	new	covid19	first	ivermectin	number
coronavirus	covid19	surge	fauci	positive	two	largest	really
people	deaths	says	covid19	yet	covid19	reduce	care
amp	new	breaking	back	infected	year	hospitalizations	patients
vaccine	last	fauci	never	know	covid	trial	take
pandemic	coronavirus	necessary	going	safe	breaking	date	one
new	pushing	restrictions	say	hospital	https	hospitalizations_largest	take_care
health	far	fauci_says	mandates	drericding	deaths_year	trial_date	days

รูปที่ 4.11 ผลลัพธ์จากแบบจำลอง GSDMM ด้วยชุดข้อมูล Twitter COVID-19

0	1	2	3	4
china	new	covid	china	https
test	cases	number	deaths	covid
hours	covid	one	first	ivermectin
woman	last	high	breaking	hospitalizations
turn	covid19	patients	year	largest
chicken	says	care	deaths_year	reduce
swab	coronavirus	take	covid19	date
roast	week	days	drericding	trial
roast_chicken	reports	really	confirmed	hospitalizations_largest
swab_test	coronavirus_cases	take_care	outbreaks	trial_date
queue	fauci	talking	grappling	wsj
queues	bnodesk	asking	china_grappling	covid19
realises_queue	reports_new	armys	sustained	comments
realises	state	asking_armys	china_confirmed	peterhotez

#### รูปที่ 4.12 ผลลัพธ์จากแบบจำลอง NMF ด้วยชุดข้อมูล Twitter COVID-19

จากผลลัพธ์การทดลองสร้างแบบจำลองทั้งสามแบบจำลองด้วยชุดข้อมูล AG News ดังรูปที่ 4.7 ถึง 4.9 จะเห็นได้ว่าในแต่ละแบบจำลอง มีการอนุมานคำในแต่ละหัวข้อที่ซ้ำกัน เช่น คำว่า reuters ที่หมายถึงสำนักข่าว ปรากฏอยู่ในหลายหัวข้อของแต่ละแบบจำลอง และจากการตีความแบบจำลอง LDA มีแนวโน้มที่จะหัวข้อที่สามารถตีความหรือเข้าใจได้โดยมนุษย์มากกว่าแบบจำลองอื่น ยกตัวอย่าง เช่น หัวข้อนี้ อาจหมายถึงการซื้อขายธุรกิจเทคโนโลยี จากการพิจารณาคำในหัวข้อ เช่น company, corp, million, deal, software, microsoft หรือ business ซึ่งต่างจากแบบจำลอง NMF ที่มีแนวโน้มที่จะให้ผลลัพธ์หัวข้อที่สามารถตีความได้ยาก ยกตัวอย่าง เช่นหัวข้อที่สองของแบบจำลอง NMF ให้ผลลัพธ์คำให้หัวข้อ เช่น monday, tuesday, thursday และ wednesday เป็นต้น ซึ่งคำดังกล่าวคือวัน ซึ่งไม่ให้ความหมายเกี่ยวกับหัวข้อที่ค้นหามาได้ หรือตีความได้ยากโดยมนุษย์

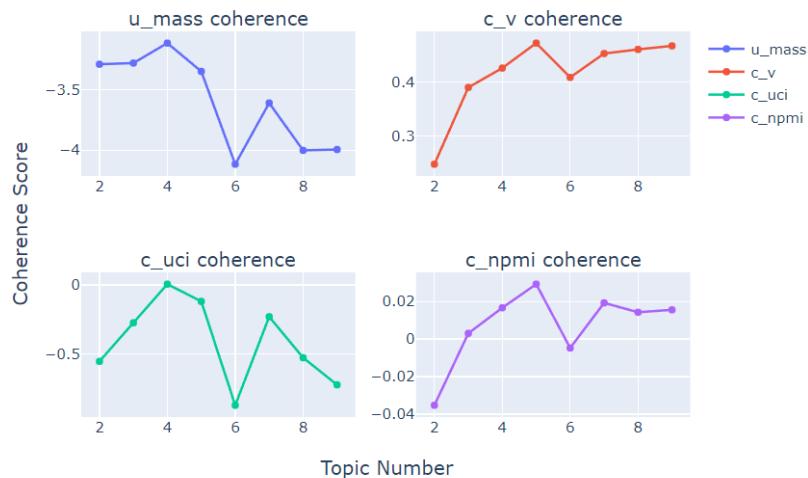
จากผลลัพธ์การสร้างแบบจำลองทั้งสามแบบจำลองด้วยชุดข้อมูล Twitter COVID-19 ดังรูปที่ 4.10 ถึง 4.12 ผลลัพธ์หัวข้อของแต่ละแบบจำลอง พบว่ามีลักษณะหัวข้อที่ยากต่อการตีความได้โดยมนุษย์ มีการปรากฏคำว่า covid และ covid19 ซ้ำอยู่ในหลายหัวข้อ เช่น แบบจำลอง GSDMM ซึ่งประกอบด้วยคำว่า covid อยู่ในทุกหัวข้อ และมีการปรากฏของคำที่ไม่มีความหมาย เช่น wsj ซึ่งปรากฏอยู่ในหัวข้อที่ 4 จากแบบจำลอง NMF

#### 4.1.4 ผลลัพธ์จากขั้นตอนที่ 3.1.5

ผลลัพธ์จากการดำเนินงานขั้นตอนที่ 3.1.5 การประเมินความเชื่อมโยงของหัวข้อด้วย Topic Coherence หลังจากการสร้างแบบจำลองหัวข้อและการแสดงผลลัพธ์ไว้ดังหัวข้อที่ 4.1.3 ทีมผู้วิจัยได้เลือกใช้วิธีการคำนวณหาค่าความสอดคล้องของหัวข้ออย่าง Topic Coherence ประกอบด้วยวิธีการคำนวณที่แตกต่างกัน 4 วิธี ได้แก่ UMASS, UCI, NPMI และ CV โดยผู้วิจัยทำการทดลองวัดความสอดคล้องของหัวข้อเมื่อจำนวนหัวข้อที่กำหนดสำหรับแบบจำลองเปลี่ยนแปลงไป โดยจะมีค่าตั้งแต่ 2 ถึง 9 หัวข้อซึ่งเพิ่มขึ้นขั้นละหนึ่งตามลำดับ และสังเกตการณ์ผลลัพธ์ค่าความสอดคล้องทั้งสี่วิธีที่เปลี่ยนแปลงไปสำหรับแต่ละแบบจำลอง เพื่อทำการหาจำนวนหัวข้อที่เหมาะสมที่สุดสำหรับการหาหัวข้อจากชุดข้อมูล

**ผลลัพธ์การเปลี่ยนแปลงค่าความสอดคล้องหัวข้อที่ได้จากแบบจำลอง LDA  
เมื่อมีการปรับจำนวนหัวข้อของแบบจำลอง**

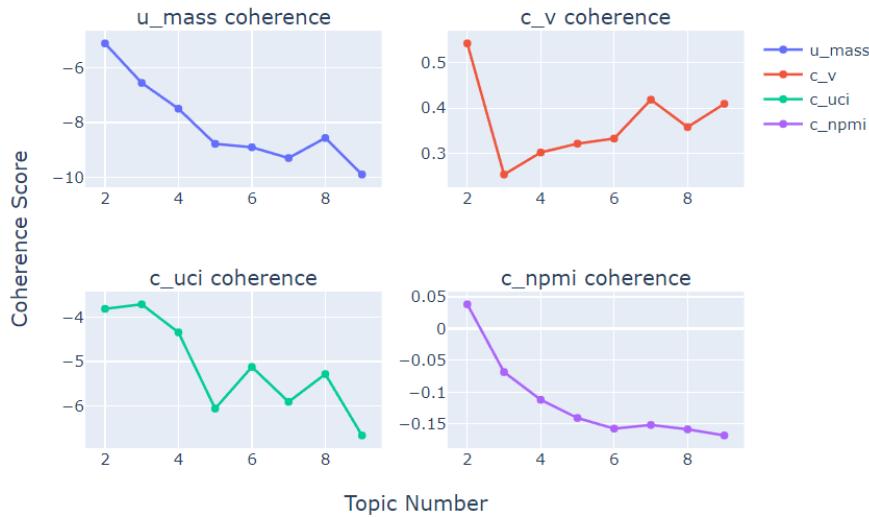
LDA Topic Coherence Comparison with AG News Dataset



**รูปที่ 4.13** กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA  
ด้วยชุดข้อมูล AG News

จากการแสดงค่าความสอดคล้องที่เปลี่ยนแปลงไปเมื่อปรับค่าจำนวนหัวข้อที่กำหนดตัวหรับแบบจำลอง LDA จะเห็นว่า ในชุดข้อมูล AG News ค่าความสอดคล้องทุกแบบเพิ่มขึ้นจนหยุดที่ค่าจำนวนหัวข้อเท่ากับ 5 ดังรูปที่ 4.13 ซึ่งทางผู้จัยได้ใช้วิธีในการเลือกค่าจำนวนหัวข้อที่เหมาะสมคือ Elbow Technique ทำให้ได้ค่าจำนวนหัวข้อที่เหมาะสมที่สุดสำหรับแบบจำลอง LDA ในชุดข้อมูล AG News คือ 5 หัวข้อ

LDA Topic Coherence Comparison with Tweets Dataset



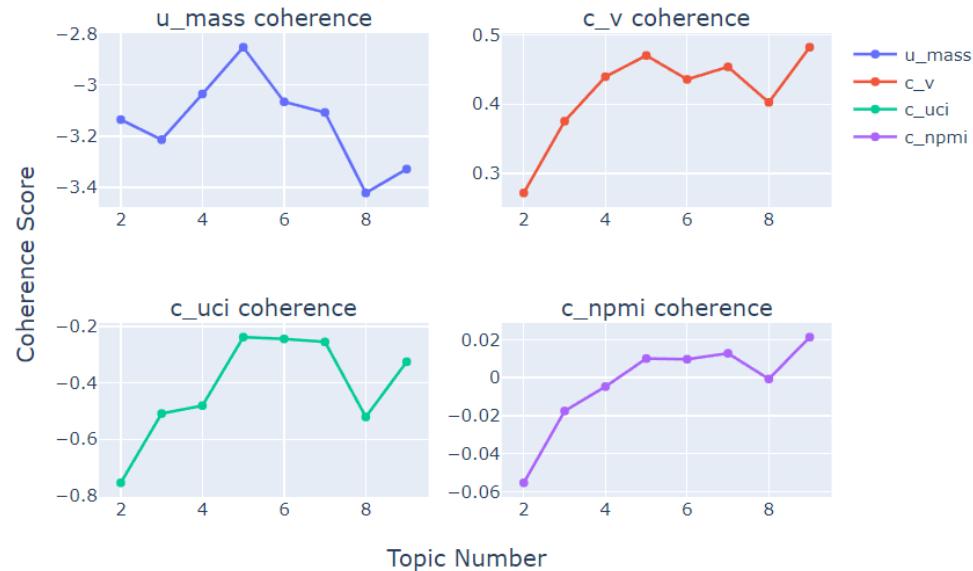
รูปที่ 4.14 กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA

ด้วยชุดข้อมูล Twitter COVID-19

สำหรับชุดข้อมูล Twitter COVID-19 จะเห็นว่า ค่าความสอดคล้องทุกแบบลดลงอย่างเห็นได้ชัด แต่มetric C\_V ที่เพิ่มขึ้นหลังจากลดลงอย่างมากเมื่อจำนวนหัวข้อที่กำหนดมีค่าเท่ากับ 3 ดังรูปที่ 4.14 และเมื่อทำการเลือกค่า K ด้วย Elbow Technique จะได้ค่า K ที่ 5 จากการลดลงของค่าความสอดคล้องแต่ละแบบที่ลดลงอย่างมากและค่อย ๆ น้อยลงเมื่อค่า K มีค่าเท่ากับ 5

ผลลัพธ์การเปลี่ยนแปลงค่าความสอดคล้องหัวข้อที่ได้จากแบบจำลอง GSDMM เมื่อทำการปรับจำนวนหัวข้อของแบบจำลอง

GSDMM Topic Coherence Comparison with AG News Dataset

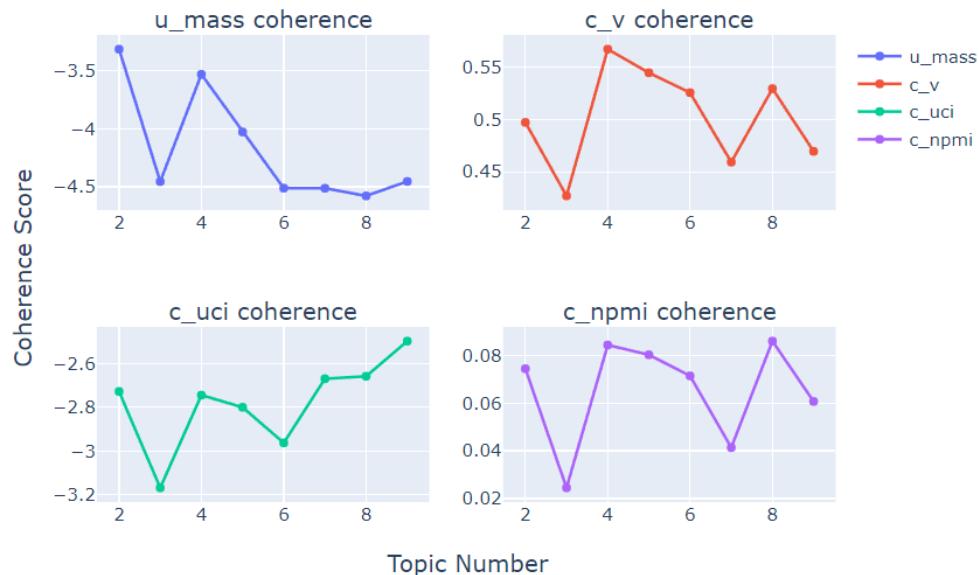


รูปที่ 4.15 กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง GSDMM

ด้วยชุดข้อมูล AG News

จากราฟที่แสดงผลลัพธ์ค่าความสอดคล้องที่เปลี่ยนแปลงไปสำหรับแบบจำลอง GSDMM ด้วยชุดข้อมูล AG News จะเห็นได้ว่าช่วงค่าจำนวนหัวข้อตั้งแต่ 2 จนถึง 5 ค่าความสอดคล้องของหัวข้อทุกแบบเพิ่มขึ้นอย่างเห็นได้ชัดและเพิ่มขึ้นช้าลงหลังค่าจำนวนหัวข้อเท่ากับ 5 หัวข้อดังรูปที่ 4.15 เมื่อทำการเลือกค่า K ด้วยวิธี Elbow Technique จะได้ค่า K ที่เหมาะสมสำหรับชุดข้อมูล AG News คือ 5 หัวข้อ

### GSDMM Topic Coherence Comparison with Tweets Dataset



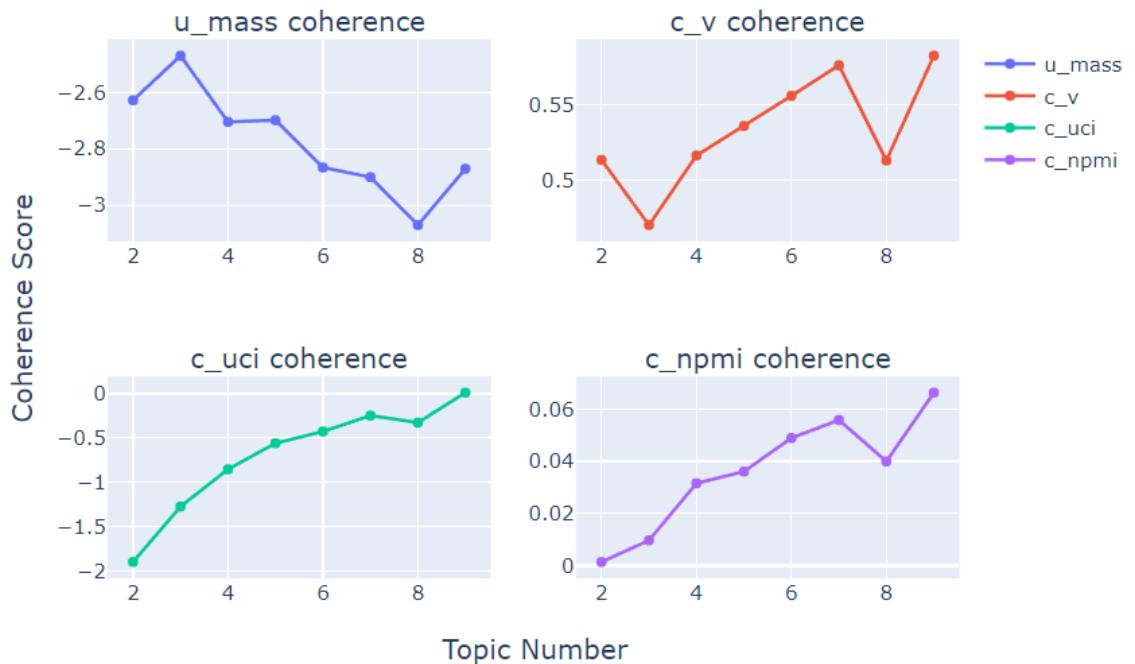
รูปที่ 4.16 กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง LDA

ด้วยชุดข้อมูล Twitter COVID-19

ส่วนชุดข้อมูล Twitter COVID-19 จะเห็นได้ว่าค่าความสอดคล้องหัวข้อแบบ C\_V และ C\_NPMI เพิ่มขึ้นอย่างเห็นได้ชัดในช่วงที่เปลี่ยนจาก 4 หัวข้อเป็น 5 หัวข้อ ส่วนค่าความสอดคล้องแบบ UMASS และ C\_UCI ไม่มีรูปแบบการเปลี่ยนแปลงที่ชัดเจนดังรูปที่ 4.16 ดังนั้นเมื่อทำการเลือกจำนวนหัวข้อที่เหมาะสมสมที่สุดด้วย Elbow Technique จำนวนหัวข้อที่เหมาะสมที่สุดสำหรับชุดข้อมูล Twitter COVID-19 คือ 4 หัวข้อ

ผลลัพธ์การเปลี่ยนแปลงค่าความสอดคล้องหัวข้อที่ได้จากแบบจำลอง NMF เมื่อมีการปรับ  
จำนวนหัวข้อของแบบจำลอง

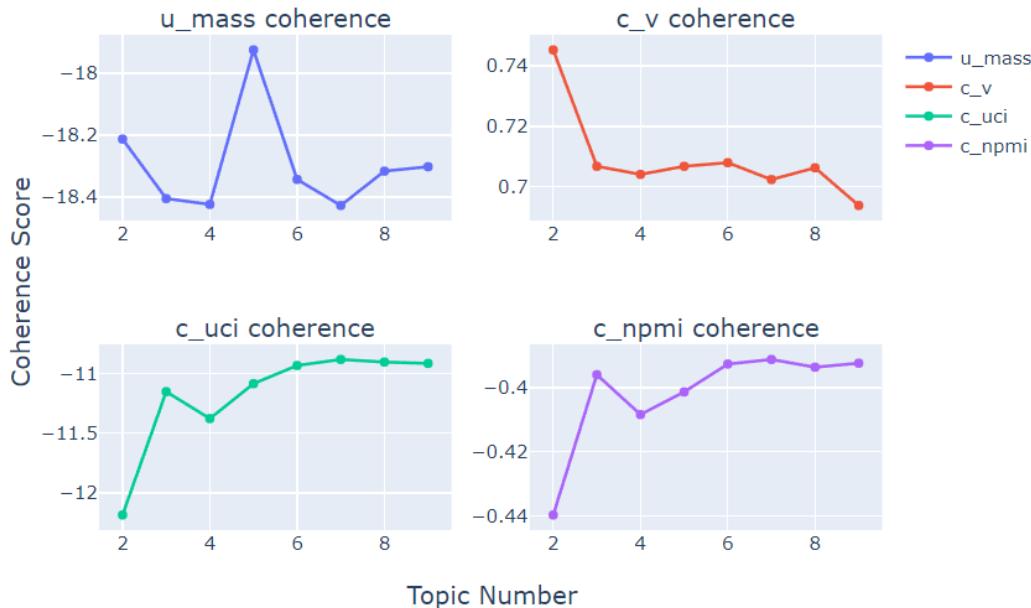
NMF Topic Coherence Comparison with AG News Dataset



รูปที่ 4.17 กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง NMF  
ด้วยชุดข้อมูล AG News

จากการที่แสดงผลลัพธ์ค่าความสอดคล้องที่เปลี่ยนแปลงไปสำหรับแบบจำลอง NMF ด้วยชุดข้อมูล AG News ดังรูปที่ 4.17 จะเห็นได้ว่าค่าความสอดคล้อง U\_MASS มี จุดสูงสุดที่ค่าจำนวนหัวข้อ 3 ส่วนค่าความสอดคล้องแบบอื่นเพิ่มขึ้นเรื่อยๆ ตามจำนวนหัวข้อ ที่เพิ่มขึ้น โดยเพิ่มขึ้นช้าลงหรือลดลงเมื่อจำนวนหัวข้อมีเท่ากับ 7 เมื่อทำการเลือกจำนวนหัวข้อ ด้วยวิธี Elbow Technique จะได้ค่า K ที่เหมาะสมสำหรับชุดข้อมูล AG News คือ 7 หัวข้อ

### NMF Topic Coherence Comparison with Tweets Dataset



รูปที่ 4.18 กราฟแสดงการเปลี่ยนแปลงของค่าความสอดคล้องจากแบบจำลอง NMF

ตัวชุดข้อมูล Twitter COVID-19

ส่วนชุดข้อมูล Twitter COVID-19 จะเห็นได้ว่าค่าความสอดคล้องหัวข้อแบบ C\_UCI และ C\_NPMI เพิ่มขึ้นอย่างเห็นได้ชัดและลดลงเมื่อมีจำนวนหัวข้อเท่ากับ 3 ส่วนค่าความสอดคล้องแบบ U\_MASS และ C\_V มีลักษณะลดลงซึ่งไม่สามารถพิจารณาได้ดังรูปที่ 4.18 ทำให้มีการทำการเลือกจำนวนหัวข้อที่เหมาะสมที่สุดด้วย Elbow techniques จำนวนหัวข้อที่เหมาะสมที่สุดสำหรับชุดข้อมูล Twitter COVID-19 คือ 3 หัวข้อ

ผลลัพธ์การเปรียบเทียบการเปลี่ยนไปของค่าความสอดคล้องที่ได้จากแบบจำลองทั้งสาม  
แบบจำลอง ด้วยชุดข้อมูล AG News

u\_mass Topic Coherence Comparison with AG News Dataset



รูปที่ 4.19 กราฟเปรียบเทียบค่าความสอดคล้อง U\_MASS ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล AG News

c\_v Topic Coherence Comparison with AG News Dataset



รูปที่ 4.20 กราฟเปรียบเทียบค่าความสอดคล้อง C\_V ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล AG News

c\_uci Topic Coherence Comparison with AG News Dataset



รูปที่ 4.21 กราฟเปรียบเทียบค่าความสอดคล้อง C\_UCI ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล AG News

c\_npmi Topic Coherence Comparison with AG News Dataset

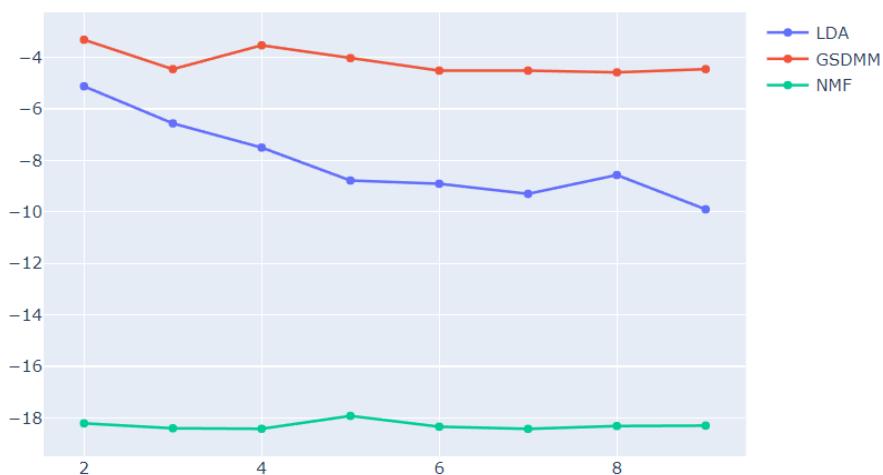


รูปที่ 4.22 กราฟเปรียบเทียบค่าความสอดคล้อง C\_NPMI ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล AG News

จากการแสดงการเปรียบเทียบค่าความสอดคล้องที่เปลี่ยนแปลงไปในแต่ละแบบจำลองโดยใช้ชุดข้อมูล AG News ดังรูปที่ 4.19 ถึง 4.22 จะเห็นได้ว่าในการคำนวณค่าความสอดคล้องแบบ U\_MASS, C\_V และ C\_NPMI ค่าความสอดคล้องของแบบจำลอง NMF มีค่ามากกว่าแบบจำลองอื่นอย่างเห็นได้ชัดในทุกๆ การเปลี่ยนจำนวนหัวข้อสำหรับแบบจำลอง และมีรูปแบบเพิ่มขึ้นอย่างต่อเนื่องในการคำนวณแบบ C\_UCI และทั้งสามแบบจำลองมีการเพิ่มขึ้นของค่าความสอดคล้องเมื่อจำนวนหัวข้อเพิ่มขึ้น ซึ่งเป็นไปในทิศทางเดียวกัน

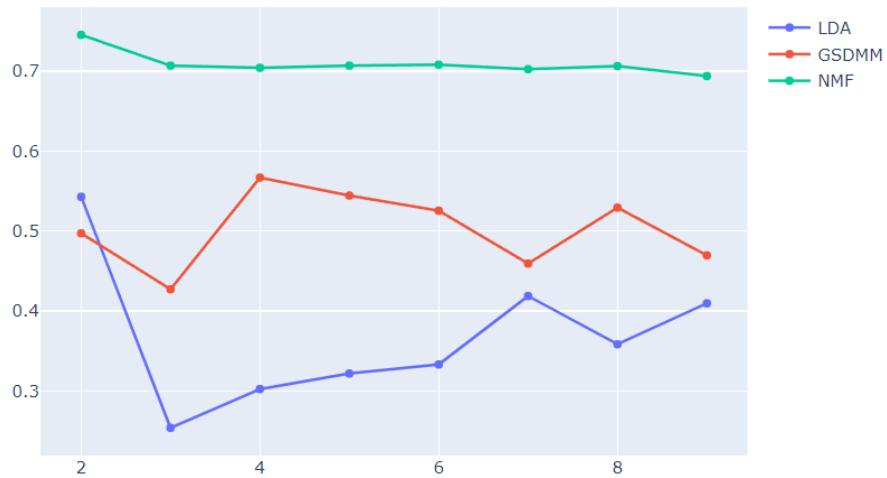
### ผลลัพธ์การเปรียบเทียบการเปลี่ยนไปของค่าความสอดคล้องที่ได้จากแบบจำลองทั้งสามแบบจำลอง ด้วยชุดข้อมูล Twitter COVID-19

u\_mass Topic Coherence Comparison with Tweets Dataset



รูปที่ 4.23 กราฟเปรียบเทียบค่าความสอดคล้อง U\_MASS ของแต่ละแบบจำลอง ด้วยชุดข้อมูล Twitter COVID-19

c\_v Topic Coherence Comparison with Tweets Dataset



รูปที่ 4.24 กราฟเปรียบเทียบค่าความสอดคล้อง C\_V ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล Twitter COVID-19

c\_uci Topic Coherence Comparison with Tweets Dataset



รูปที่ 4.25 กราฟเปรียบเทียบค่าความสอดคล้อง C\_UCI ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล Twitter COVID-19

c\_npmi Topic Coherence Comparison with Tweets Dataset



รูปที่ 4.26 กราฟเปรียบเทียบค่าความสอดคล้อง C\_NPMI ของแต่ละแบบจำลอง  
ด้วยชุดข้อมูล Twitter COVID-19

จากราฟแสดงการเปรียบเทียบค่าความสอดคล้องที่เปลี่ยนแปลงไปในแต่ละแบบจำลองโดยใช้ชุดข้อมูล AG News ดังรูปที่ 4.23 ถึง 4.26 จะเห็นได้ว่าค่าความสอดคล้องของแต่ละแบบจำลองจากการคำนวณทุกแบบมีค่าต่างกันอย่างเห็นได้ชัด โดยแบบจำลอง GSDMM มีค่ามากที่สุดจากการคำนวณแบบ U\_MASS, C\_UCI และ C\_NPMI และจากการคำนวณทุกแบบ ค่าความสอดคล้องมีลักษณะคงที่หรือลดลงเมื่อจำนวนหัวข้อเปลี่ยนแปลงไปโดยมีแบบจำลอง LDA แบบจำลองเดียวที่มีค่าความสอดคล้องเพิ่มขึ้นหรือลดลงอย่างเห็นได้ชัด

ผลลัพธ์ค่าความสอดคล้องของแต่ละแบบจำลองเมื่อกำหนดจำนวนหัวข้อที่เหมาะสมให้กับแบบจำลอง

**ตารางที่ 4.1** ค่าความสอดคล้องจากการคำนวณแต่ละวิธีของแบบจำลอง LDA, GSDMM และ NMF  
เมื่อใช้ชุดข้อมูล AG News

	U_MASS	C_V	C_UCI	C_NPMI
LDA	-3.34707	0.47138	-0.12040	0.02924
GSDMM	-2.85268	0.47002	-0.23802	0.01009
NMF	-2.90027	0.57599	-0.24815	0.05574

จากตารางที่ 4.1 จะเห็นว่าค่าความสอดคล้องแบบ U\_MASS จากแบบจำลอง GSDMM มีค่ามากที่สุดจากแบบ สำหรับค่าความสอดคล้องแบบ C\_NPMI และ C\_V แบบจำลอง NMF ให้ผลลัพธ์ดีที่สุดซึ่งมากกว่าแบบจำลอง GSDMM เพียงเล็กน้อย สำหรับการคำนวณแบบ C\_UCI แบบจำลอง LDA ให้ผลลัพธ์ดีที่สุด

**ตารางที่ 4.2** ค่าความสอดคล้องจากการคำนวณแต่ละวิธีของแบบจำลอง LDA, GSDMM และ NMF  
เมื่อใช้ชุดข้อมูล Twitter COVID-19

	U_MASS	C_V	C_UCI	C_NPMI
LDA	-8.78082	0.32171	-6.06184	-0.14055
GSDMM	-3.52862	0.56680	-2.74439	0.08435
NMF	-18.40512	0.70681	-11.15197	-0.39591

สำหรับผลลัพธ์ค่าความสอดคล้องของหัวข้อที่สกัดหรือจัดกลุ่มได้ด้วยชุดข้อมูล COVID-19 Twitter จากแต่ละแบบจำลอง โดยใช้จำนวนหัวข้อที่เหมาะสมที่สุด สามารถนำมาเปรียบเทียบกันได้โดยตารางที่ 4.2 โดยจะเห็นว่าแบบจำลอง GSDMM ให้ผลลัพธ์ดีที่สุด สำหรับการคำนวณสามแบบคือ U\_MASS, C\_UCI และ C\_NPMI และสำหรับการคำนวณแบบ C\_V แบบจำลอง NMF ให้ผลลัพธ์ดีที่สุด

#### 4.2 รายละเอียดการทำงานแต่ละขั้นตอนของการทดลองที่ 2

#### 4.2.1 ผลลัพธ์จากขั้นตอนที่ 3.2.1

จากการดำเนินงานขั้นตอนที่ 3.2.1 เป็นการนำข้อมูลความคิดเห็นจาก Twitter มาใช้งาน โดยผู้วิจัยนำข้อมูลความคิดเห็นจาก Social Media อย่าง Twitter มาใช้งานโดยการใช้ไลบรารี Tweepy ประกอบกับ Twitter Developer API ในการเข้าถึงข้อมูล และทำการเก็บเฉพาะความคิดเห็นภาษาไทยโดยหาจาก Hashtag จำนวน 10 หัวข้อด้วยกัน จากนั้นทำการเก็บข้อมูลข้อความคิดเห็น เวลา และ Hashtag ที่ทำการคั่ง ได้ผลลัพธ์ดังรูปที่ 4.27

		text	timestamp	hashtag
0	รับกลับมาร 4EVE น่า ❤️ ☺️ 🌸 ภ-คาก็ขอชั่ว 3,000-4,000 (400)ภ-ชั่ว 2,500-1,200 (350)ภ-หวานไปอธิบายเค้าได้ใจที่ #ชีรุ่งนพสินรักบดบัง/or- กดในไลค์สิเน็ปเป็นบันดาลค่ากากฯค่าหูไม่เย็บ นะครับ สื่อสารทางไอล์เก็บด์ ไม่ต้องจำนาดจำกัดนะ ❌ X/#4EVE #XOXOentertainment #Tpop <a href="https://t.co/mjyPn29Var">https://t.co/mjyPn29Var</a>		2022-11-03 16:12:47	#4EVE
1		ฉันไปปป เพื่อเจ้ม #PunBNK48 <a href="https://t.co/Y008RnM2MI">https://t.co/Y008RnM2MI</a>	2022-11-03 16:18:25	#PunBNK48
2	รบกวนทางค่ายพิจารณา Benefits สำหรับคนเดียวของ 4EVE มีกรอบแนะ ตัวนี้ใหญ่ไปสุดเคย์เรียบให้ทุกคนมีผลเบบะ แล้วลิฟท์นี้มี มีกันกับอะนาก อยากรู้ให้ลองเชื้อใจกับเครื่องนั้นนะ คุณเดียวคนแรกน้ำเสียงนุ่มนวลมากๆ 😊 💕 💚 #PunBNK48 <a href="https://t.co/lBqgk8pDwA">https://t.co/lBqgk8pDwA</a>		2022-11-03 16:22:47	#4EVE
3		น้องนาหะขอันงนี่มีภูมิใจเลย 😊 💕 💚 #PunBNK48 <a href="https://t.co/XDs8azbDG0">https://t.co/XDs8azbDG0</a>	2022-11-03 16:23:52	#PunBNK48
4		น้องนาหะขอันงนี่มีภูมิใจเลย 😊 💕 💚 #PunBNK48 #TheChesseSisters_OT #NowIrene <a href="https://t.co/noj1VGRSD">https://t.co/noj1VGRSD</a>	2022-11-03 16:25:04	#PunBNK48

รูปที่ 4.27 ตัวอย่างข้อมูลความคิดเห็นจาก Twitter

#### 4.2.2 ผลลัพธ์จากขั้นตอนที่ 3.2.2

จากการดำเนินงานขั้นตอนที่ 3.2.2 เป็นการจัดเตรียมข้อมูลที่ผู้ใช้ได้ทำการเลือกมา ฝึกฝนใช้งานการสร้างแบบจำลองหัวข้อ ข้อมูลดังกล่าวจำเป็นต้องมีการเตรียมข้อมูลเบื้องต้น โดยผู้ใช้ได้ตัดคำ (Tokenization) และทำการลบคำที่ไม่ช่วยในการสื่อความหมายของข้อความ ในภาษาธรรม (Stopwords) ตัวเลข อิมิจิหรือไอคอน ตัวอักษรพิเศษ และตัวอักษรภาษาเยอรมัน ออกจากข้อความไป โดยมีผลลัพธ์เบื้องต้นดังรูปที่ 4.28

[ 'รักบันดัลร่าภาคโภโน้มแน่ไปตีริวะเค้าได้ที่รีวะนิลับบันดัลรักรถไม่โน่ได้คืนเป็นแบนน์ค่าเคาร์บันไม่ยอมนะจะได้เรียบมานะจ่อได้เลียค่าบันจี้น้อยมัดจักก่อนค่ะ' ,  
'จัดไปปไปเพื่ออาบป' ,  
'บันบานดัลร่าภาคพิราภานลับบันดูเดียวของอึกรอบบนະส่วนใหญ่ปุ่สเดือร์บานให้ทุกบันดัลเรียลະแลวสิทธิ์อื้นเมินกันอึกเยอะนาโกยกาโน่โลลงช้อใจกันอึกครั้งนะจะตอนเดียวครอบแรกกัน  
สำคัญมาเจริญปะ' ,  
'บ้องนาหรืออ่วงปุ่ยลูบปะ' ,  
'บ้องนาบานราก' ,  
'บ้องนาแจ๊อซ' ,  
'ปีบุณย์ดัลปีนีเริ่มมากกอกกรรมบ้องมันบีกูนนบ' ,  
'ขอบคุณลูบมีนีการหื้ดดุดมนูบีเขมมากกอกกรรมบ้องมันทพารรรรรบีกูนบ' ,  
'อยุยนี่ยักดูอัวน์ทะจะทกอก' ,  
'เรยวะบันดูที่บังปีกูนต่องไว้ปะ' ]

รูปที่ 4.28 ตัวอย่างข้อมูลความคิดเห็นหลังจากผ่านกระบวนการเตรียมข้อมูลแล้ว

เมื่อได้ข้อมูลที่ผ่านการเตรียมขึ้นแรกแล้ว จานนี้จะนำข้อมูลดังกล่าวมาแปลงเป็นเวกเตอร์โดยใช้ไลบรารี Sentence Transformer และใช้แบบจำลองฝึกสอนล่วงหน้า simcse-model-roberta-base-thai ที่ทำการฝึกสอนด้วยข้อมูล Wikipedia ภาษาไทยมาแล้ว โดยมีตัวอย่างข้อมูลที่ทำการแปลงแล้วดังรูปที่ 4.29

```
array([ 2.32386380e-01, -5.35174310e-01,  1.50577575e-01,  7.51033798e-02,
       -1.14353418e+00,  2.23583415e-01,  4.91340429e-01, -1.74547538e-01,
       2.79313356e-01,  9.01893973e-01, -2.76651949e-01, -3.35553706e-01,
      1.00030553e+00, -9.97770190e-01,  3.00644845e-01, -1.08471417e+00,
     -4.51268941e-01, -6.04983389e-01,  2.73692496e-02,  5.34513414e-01,
     -6.31842464e-02,  6.21075332e-01, -1.30175814e-01, -8.57994080e-01,
    -3.68137747e-01,  3.02926689e-01, -3.84197265e-01,  5.30498862e-01,
     1.20816457e+00,  3.93316239e-01, -6.55789495e-01, -8.48775613e-04,
    -2.07952768e-01, -3.34227115e-01, -1.94593802e-01,  1.03142297e+00,
   -7.55951330e-02, -4.05274153e-01, -1.78186095e+00,  1.28068089e+00,
    7.35838413e-02,  2.56273821e-02,  4.86057878e-01,  9.19095874e-01,
   -2.51976084e-02, -2.78663938e-03, -2.49613628e-01,  2.84251124e-01,
    1.51897445e-01, -5.74075997e-01, -9.98214841e-01,  1.01612735e+00,
    8.38984728e-01,  3.25914025e-01,  8.46830368e-01,  1.46987116e+00,
    3.92683238e-01,  9.63376343e-01,  5.09619176e-01,  3.02076280e-01,
    2.74990946e-02,  1.01652217e+00,  8.16176116e-01,  1.69731092e+00,
    1.28005922e+00,  9.09315646e-02,  3.01153928e-01, -2.95410395e-01,
   -1.03146410e+00,  5.95290661e-01, -5.79415262e-01, -7.09930956e-01,
   -1.04802656e+00, -1.04895175e+00,  1.07811987e+00, -4.34260070e-01,
    7.96479046e-01, -4.32267375e-02, -1.90475613e-01, -1.59328103e+00,
   -1.79145977e-01, -8.81750524e-01,  5.36710083e-01, -4.56241667e-01,
    1.16830945e-01, -6.38084531e-01,  1.52563894e+00, -9.85301316e-01,
```

รูปที่ 4.29 ตัวอย่างข้อมูลที่แปลงเป็นเวกเตอร์ด้วย Sentence Transformer

เมื่อได้ข้อมูลที่อยู่ในรูปแบบเวกเตอร์เรียบร้อยแล้ว แต่ยังคงใช้เวลาในการสร้างแบบจำลอง เนื่องจากข้อมูลดังกล่าวมีจำนวนมิติสูง ผู้วิจัยจึงทำการลดมิติข้อมูลโดยการใช้เทคนิค UMAP สร้างคุณสมบัติที่สำคัญจำนวน 19 คุณสมบัติ และเพิ่มตัวแปรเวลาในมาตรฐาน Unix Timestamp สำหรับการทดลองที่ต้องการพิจารณาปัจจัยเวลาร่วมด้วย โดยมีตัวอย่างข้อมูลดังรูปที่ 4.30

```
array([ 8.47184467e+00,  5.18076372e+00,  4.09976339e+00,  5.11178255e+00,
       4.69777584e+00,  5.00757408e+00,  5.59142208e+00,  4.94909811e+00,
       6.73089218e+00,  3.79925489e+00,  5.58387661e+00,  3.35612941e+00,
      4.59946346e+00,  4.91808510e+00,  4.79168558e+00,  4.43590498e+00,
      4.96245003e+00,  5.11414909e+00,  4.53396654e+00,  1.66749197e+09])
```

รูปที่ 4.30 ตัวอย่างข้อมูลหลังจากการลดมิติและเพิ่มปัจจัยเวลา

นอกจากนี้สำหรับการทดลองแบบคำนึงถึงปัจจัยเวลาหลังจากการเพิ่มปัจจัยเวลาแล้ว ยังมีการทำ Normalization เพิ่มเติมเพื่อให้ช่วงของข้อมูลเป็นช่วงเดียวกัน

#### 4.2.3 ผลลัพธ์จากขั้นตอนที่ 3.2.3

จากการดำเนินงานขั้นตอนที่ 3.2.3 เป็นการสร้างแบบจำลองหัวข้อ ผู้วิจัยได้ทำการสร้างแบบจำลองโดยการใช้แบบจำลอง K-Means Clustering โดยใช้ชุดข้อมูลข้อความคิดเห็นจาก Twitter ที่ผ่านการจัดเตรียมข้อมูลแล้ว โดยนำมาสร้างแบบจำลองด้วยวิธีที่ต่างกันสองวิธี คือ Sliding Window และ Expanding Window และยังมีการทดลองเปรียบเทียบระหว่างการสร้างแบบจำลองด้วยสองวิธีดังกล่าวแบบใช้เวลาเป็นปัจจัยประกอบ โดยจำนวน Cluster หรือค่า K ที่ตั้งค่าคือ 10 อ้างอิงจากจำนวน Hashtag ของข้อมูลและเพื่อความสะดวกต่อการศึกษาความหัวข้อที่มีเนื้อหาแตกต่างกัน และในการค้นหาคำซึ่งเป็นส่วนประกอบของหัวข้อทำ จะใช้เทคนิค TF-IDF ในการหาค่าสถิติของคำที่พบในแต่ละหัวข้อ หรือ Cluster โดยมีการสรุปลักษณะของผลลัพธ์หัวข้อจากการทดลองดังตารางที่ 4.3 และสามารถดูผลลัพธ์หัวข้ออื่น ๆ ได้จากภาคผนวก ก. (หน้าที่ 109)

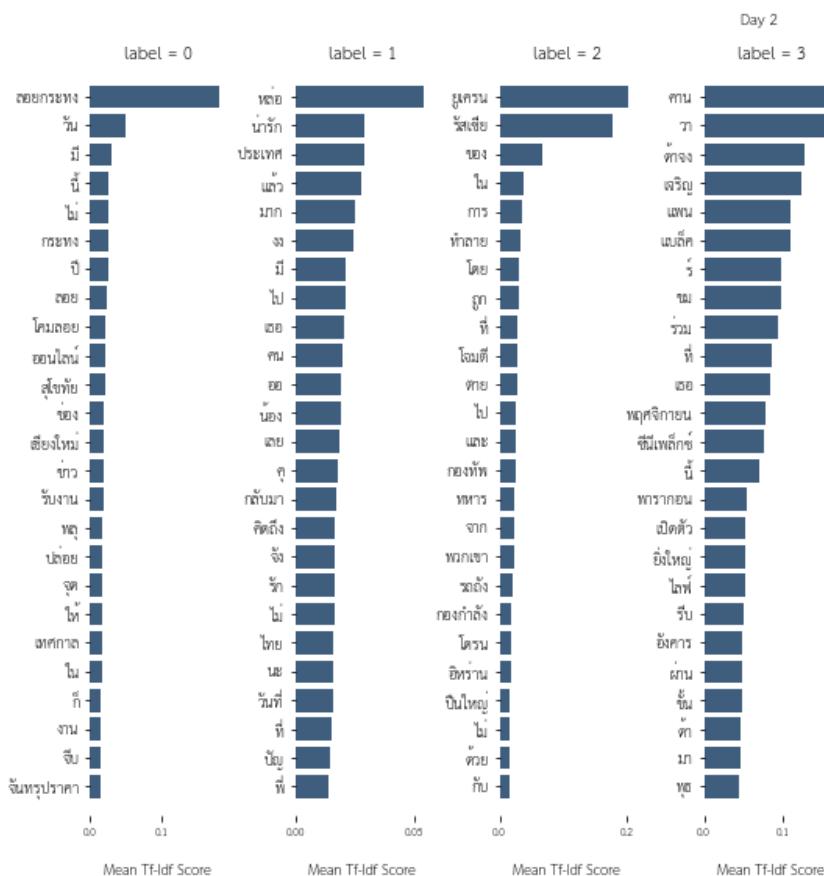
ตารางที่ 4.3 สรุปผลลัพธ์การค้นหาหัวข้อจากการทดลองที่แตกต่างกัน

รูปแบบการทดลอง	คุณลักษณะ
การทดลองแบบ Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ	แบบจำลองค้นหาหัวข้อที่สามารถตีความได้ พบรูปแบบใหม่ไปของหัวข้อในแต่ละช่วงเวลา การเกิดขึ้นใหม่ของหัวข้อ แต่ก็พบหัวข้อที่ตีความไม่ได้อยู่ เช่นกัน จำนวนหัวข้อที่พบอยู่ในช่วง 9-10 หัวข้อ
การทดลองแบบ Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ	แบบจำลองค้นหาหัวข้อที่สามารถตีความได้ มีการพบการเปลี่ยนไปของหัวข้อในแต่ละช่วงเวลา การเกิดขึ้นใหม่ของหัวข้อ แต่ก็พบหัวข้อที่ตีความไม่ได้อยู่ เช่นกัน จำนวนหัวข้อที่พบอยู่ในช่วง 9-10 หัวข้อ
การทดลองแบบ Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ	แบบจำลองค้นหาหัวข้อที่สามารถตีความได้ มีการพบการเปลี่ยนไปของหัวข้อ การเกิดขึ้นใหม่ของหัวข้อ พบรูปจำนวนหัวข้อที่ตีความไม่ได้น้อยกว่าการทดลองแบบอื่น พบจำนวนหัวข้อที่ค้นหาได้จากแบบจำลองน้อยกว่าการทดลองแบบใช้เวลา และมีลักษณะค้นพบมากขึ้นเรื่อย ๆ เมื่อค้นหาไปตามเวลา

ตารางที่ 4.3 (ต่อ) สรุปผลลัพธ์การค้นหาหัวข้อจากการทดลองที่แตกต่างกัน

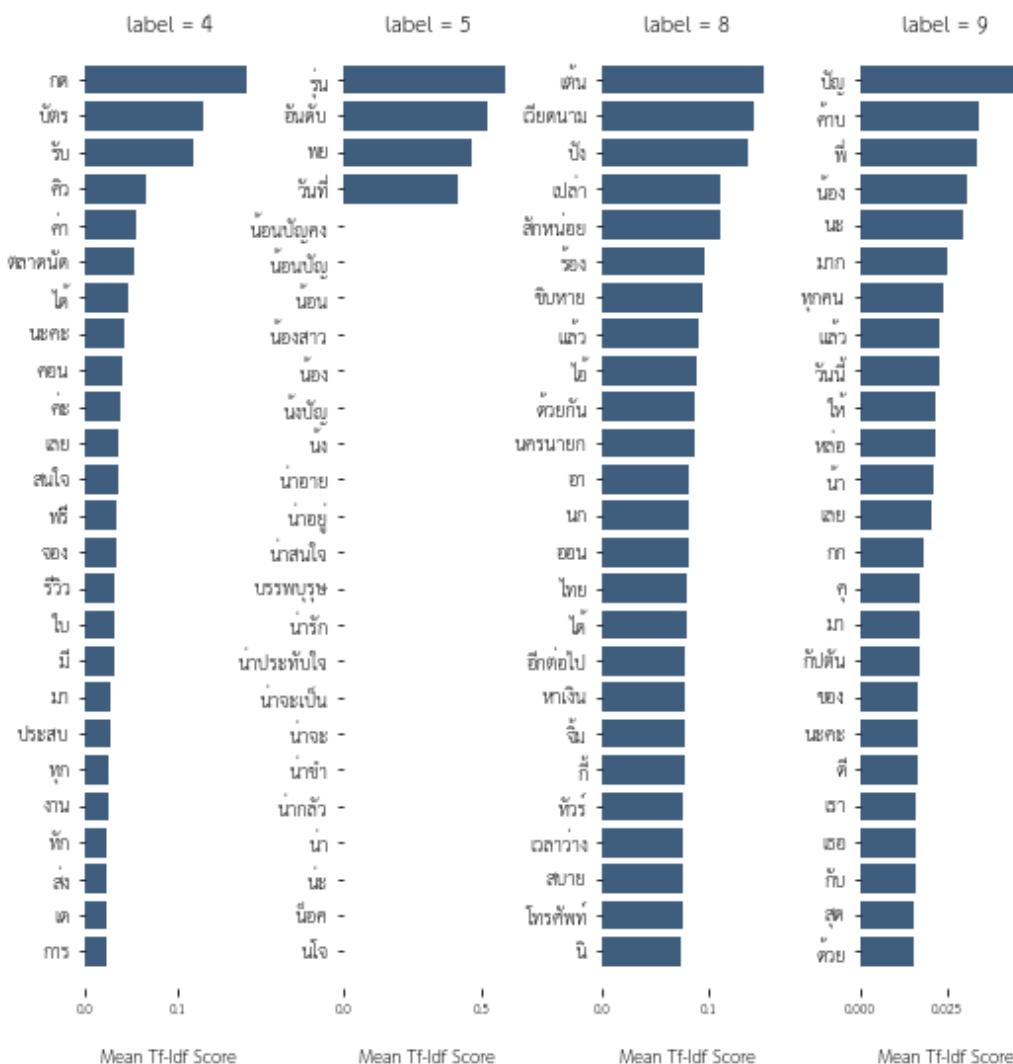
รูปแบบการทดลอง	คุณลักษณะ
การทดลองแบบ Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ	แบบจำลองก้นหาหัวข้อที่สามารถตีความได้มีการพับการเปลี่ยนไปของหัวข้อ การเกิดขึ้นใหม่ของหัวข้อ พบจำนวนหัวข้อที่ตีความไม่ได้น้อยที่สุด พบจำนวนหัวข้อที่ก้นหาได้จากแบบจำลองน้อยที่สุดเทียบกับแบบอื่น จำนวนหัวข้อที่พบทคลอดการทดลองมีจำนวน 5-7 หัวข้อเท่านั้น

ตัวอย่างผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ



รูปที่ 4.31 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window

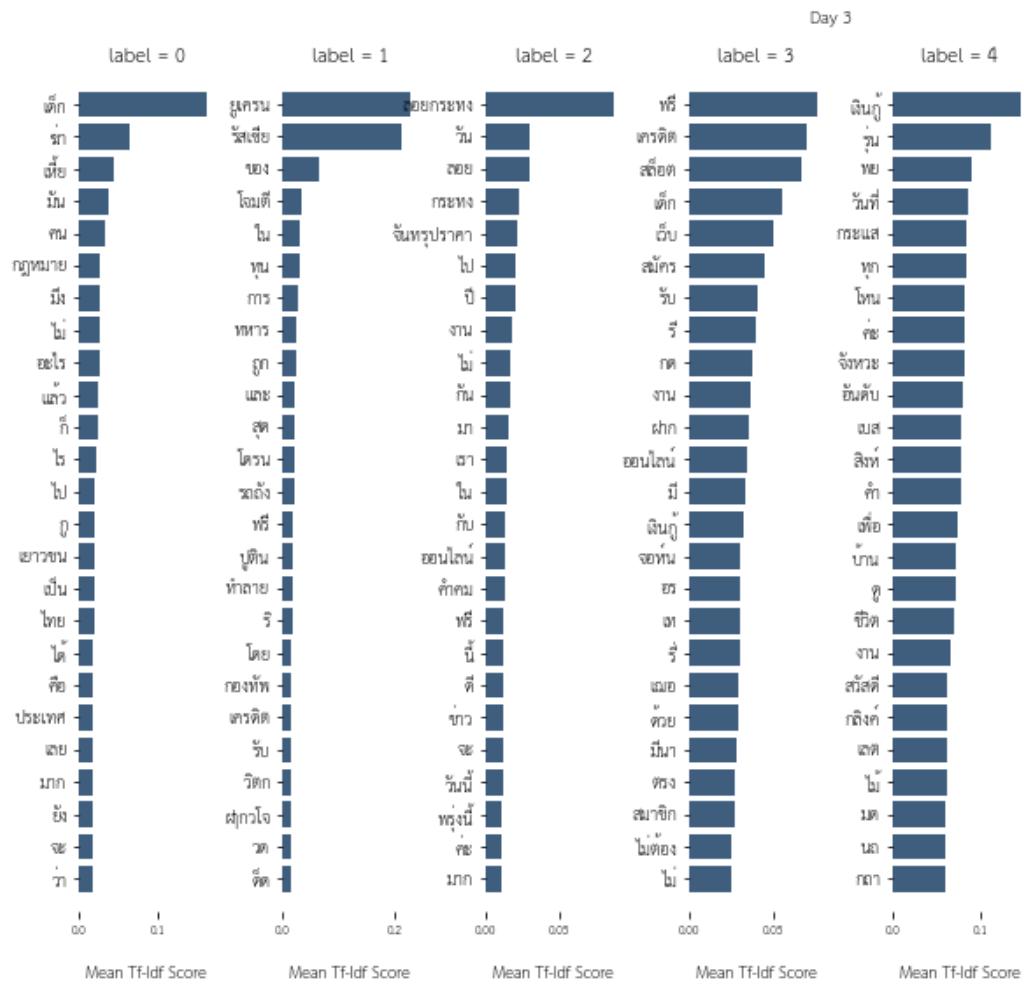
## โดยไม่ใช่เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ 4.32 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window

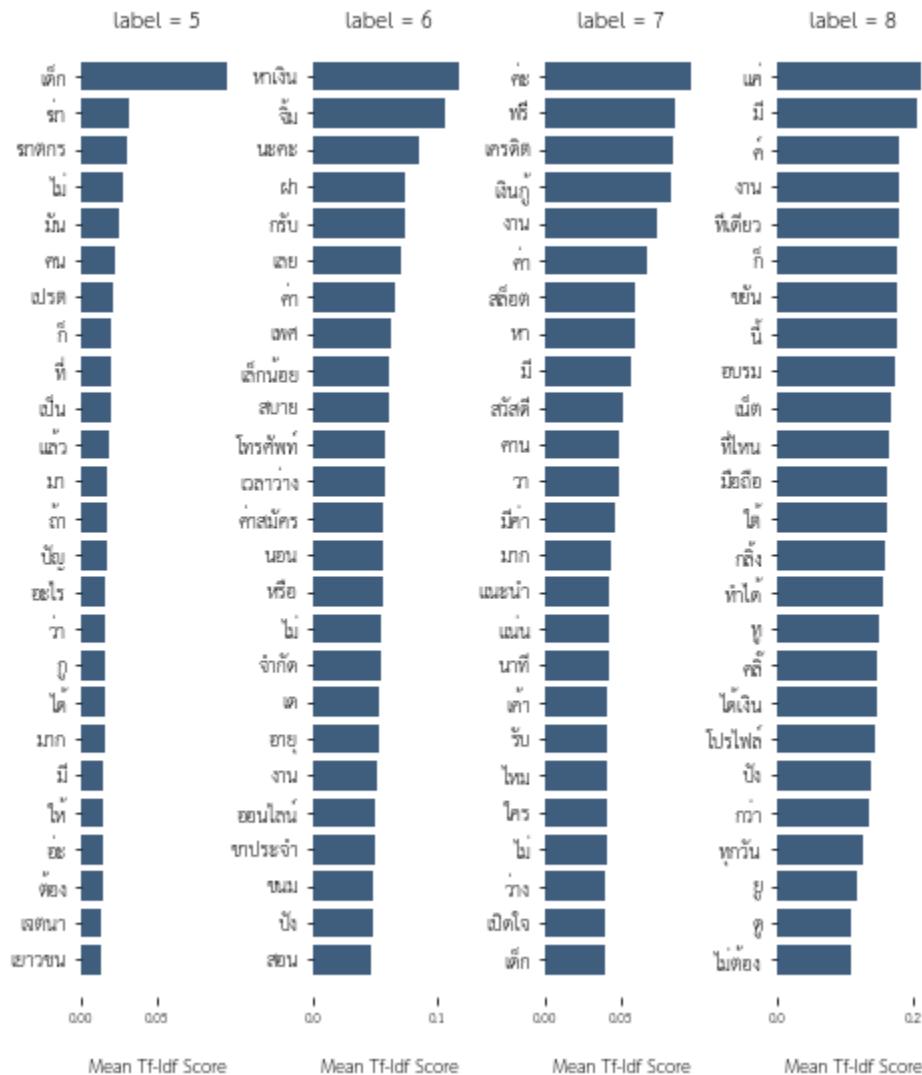
โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ 4.31 และ 4.32 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สอง (วันที่ 6 พฤศจิกายน พ.ศ. 2565) พบว่าลักษณะหัวข้อมีลักษณะคล้ายกับวันแรก มีบางหัวข้อที่หายไป เช่นวันที่หนึ่งหัวข้อที่ 9 หัวข้อที่มีลักษณะเดียวกันนี้ไม่พบในวันที่สอง รวมถึงหัวข้อเกี่ยวกับ ขันหมุ่ปราการ์ไม่พบเช่นกัน สำหรับวันที่สอง หัวข้อที่ทำการหมายไม่ครบ 10 หัวข้อ ซึ่งอาจเกิดจากเนื้อหาของช่วงวันดังกล่าวอาจไม่สอดคล้องกับข้อมูลของวันที่ใช้ฝึกสอนแบบจำลอง



รูปที่ 4.33 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

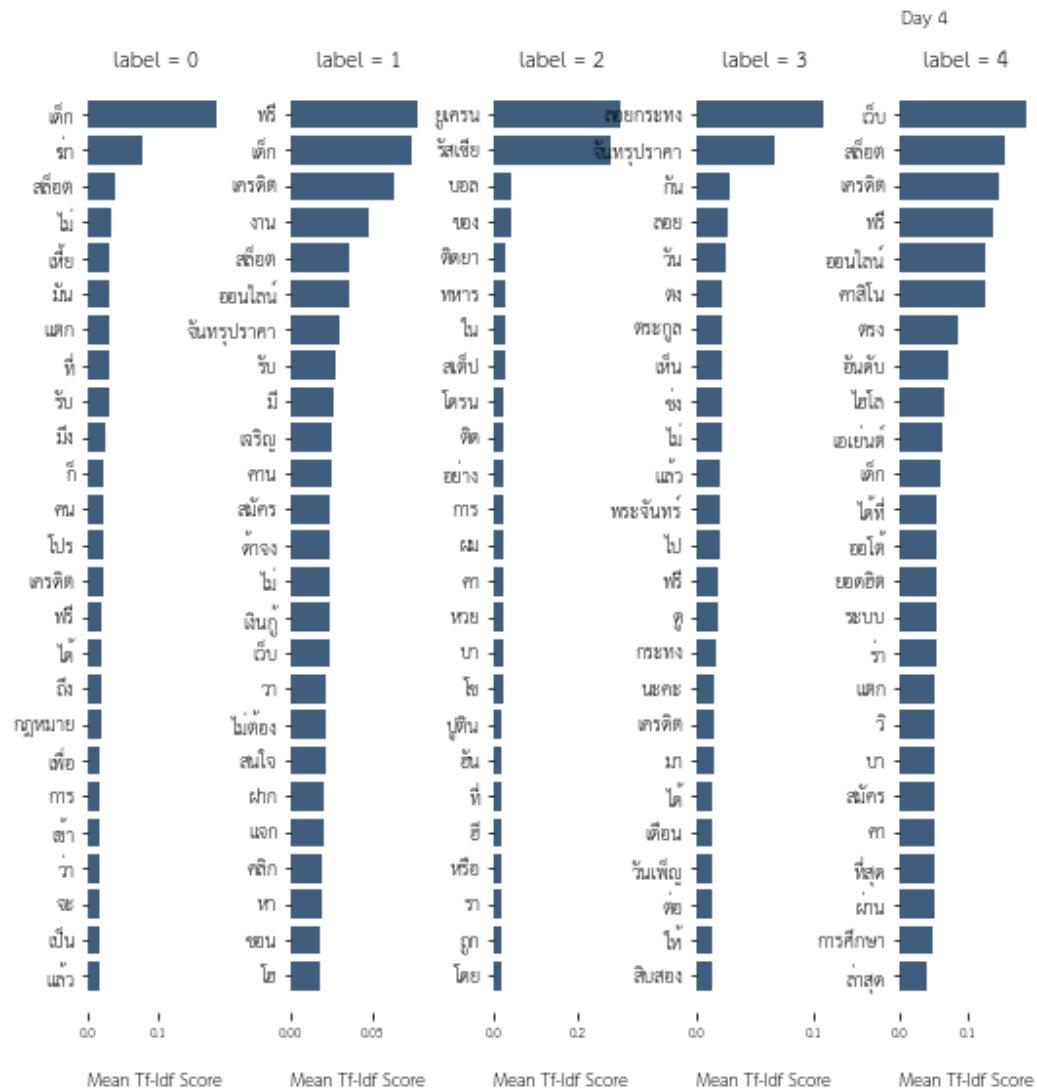


รูปที่ 4.34 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window

## โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

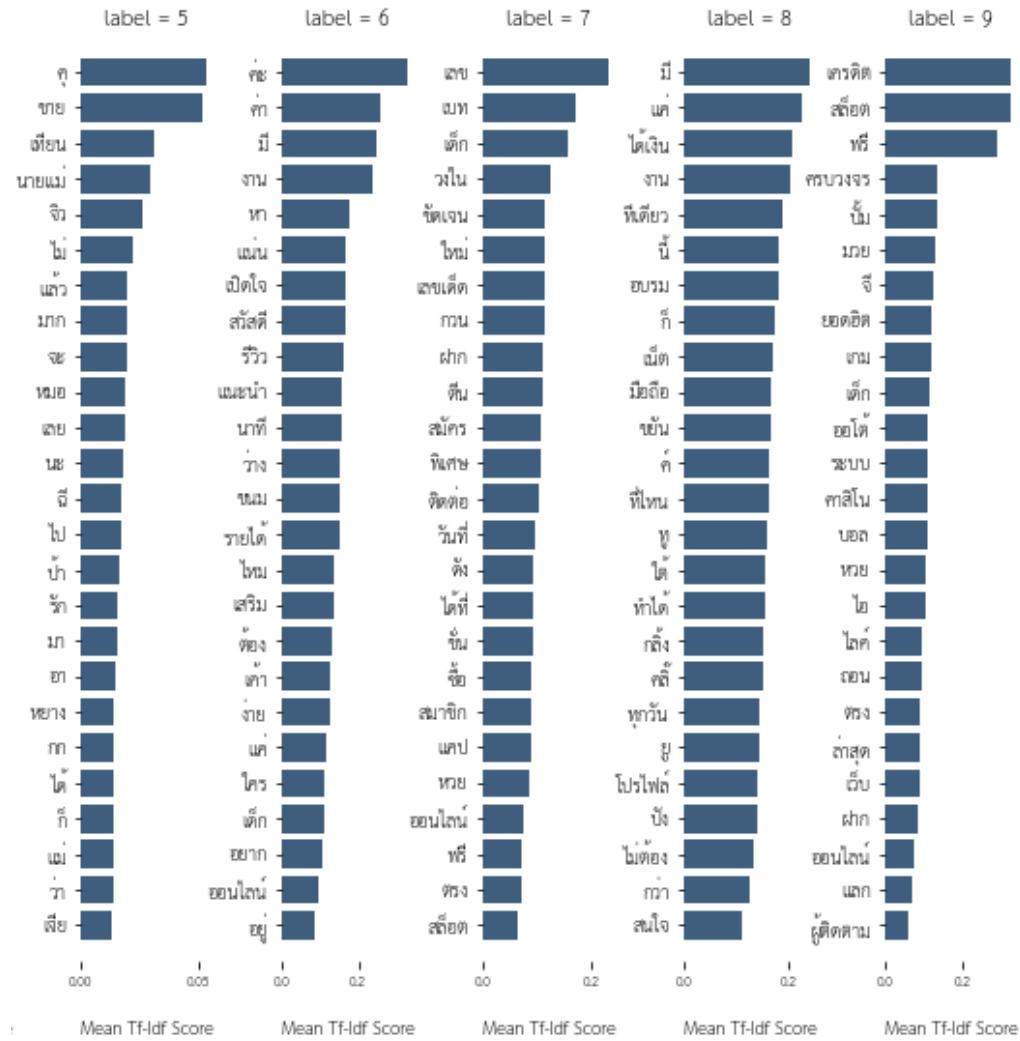
จากรูปที่ 4.33 และ 4.34 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สาม (วันที่ 7 พฤษภาคม พ.ศ. 2565) พบว่ามีหัวข้อใหม่เพิ่มขึ้นมาบางหัวข้อ เช่น หัวข้อ 0 ที่อาจตีความได้ว่า คือหัวข้อเกี่ยวกับข่าวเด็ก หัวข้อ 6, 7 และ 8 ที่มีส่วนประกอบของหัวข้อที่อาจตีความได้ว่า เกี่ยวกับเงินกู้ และการหางาน รวมถึงส่วนอื่นที่ไม่สามารถตีความได้ หรืออาจเกิดจากเนื้อหา ข้อความที่มีข้อความคิดเห็น Spam ที่ไม่เกี่ยวข้องเป็นจำนวนมาก

ตัวอย่างผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ



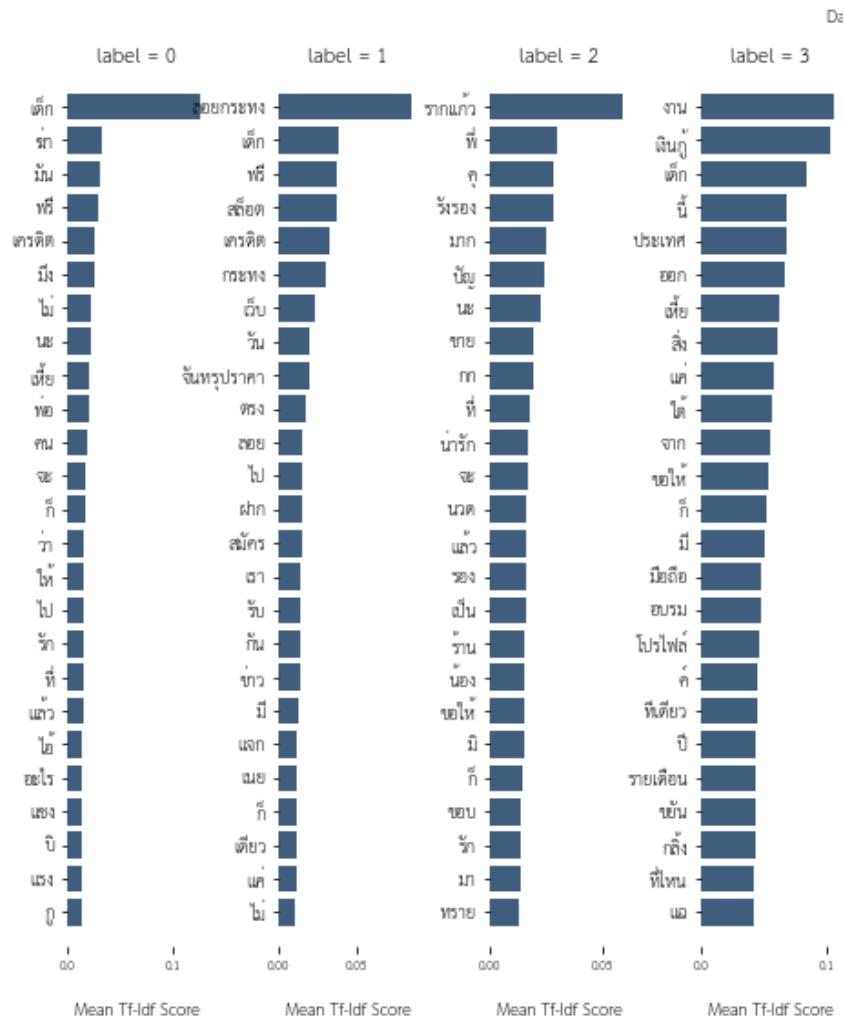
รูปที่ 4.35 ผลลัพธ์การค้นหาหัวข่าววันที่สี่จากการทดลองแบบ Expanding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



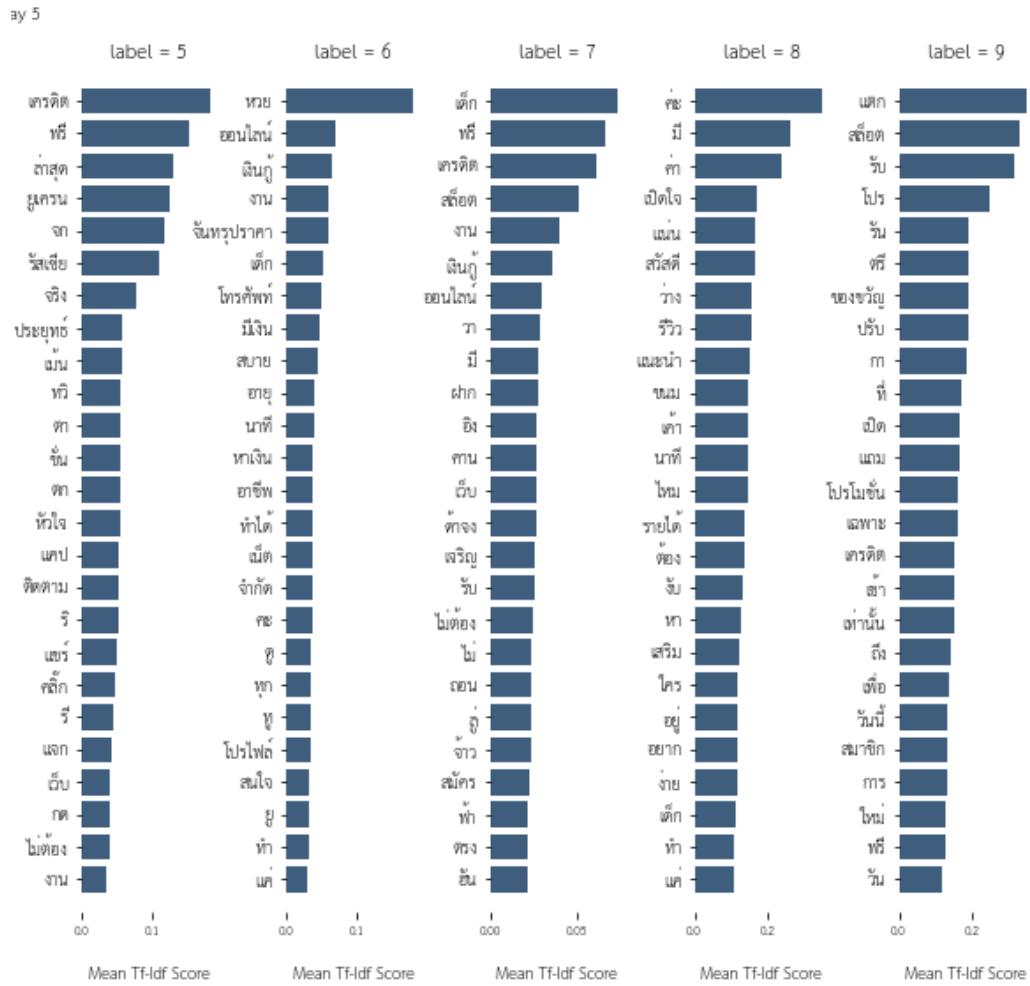
รูปที่ 4.36 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window

## โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)



รูปที่ 4.37 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

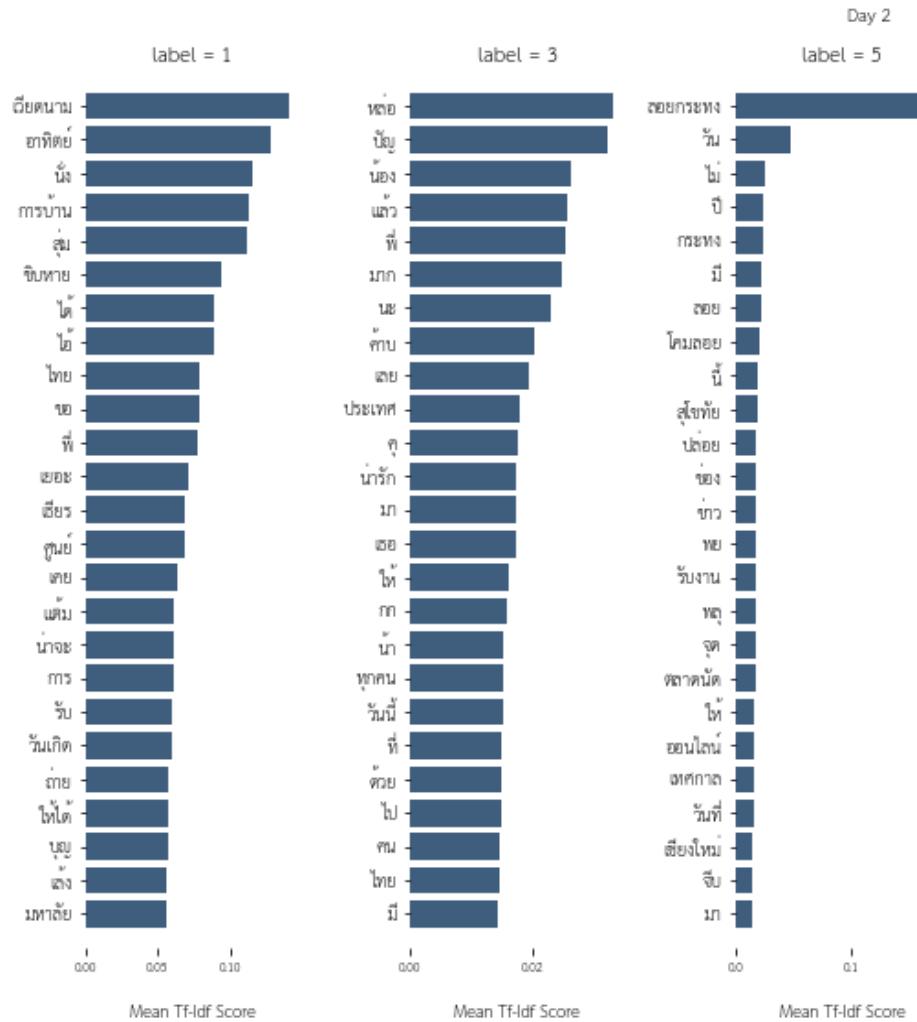


รูปที่ 4.38 ผลลัพธ์การกันหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window

## โดยไม่ใช่เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

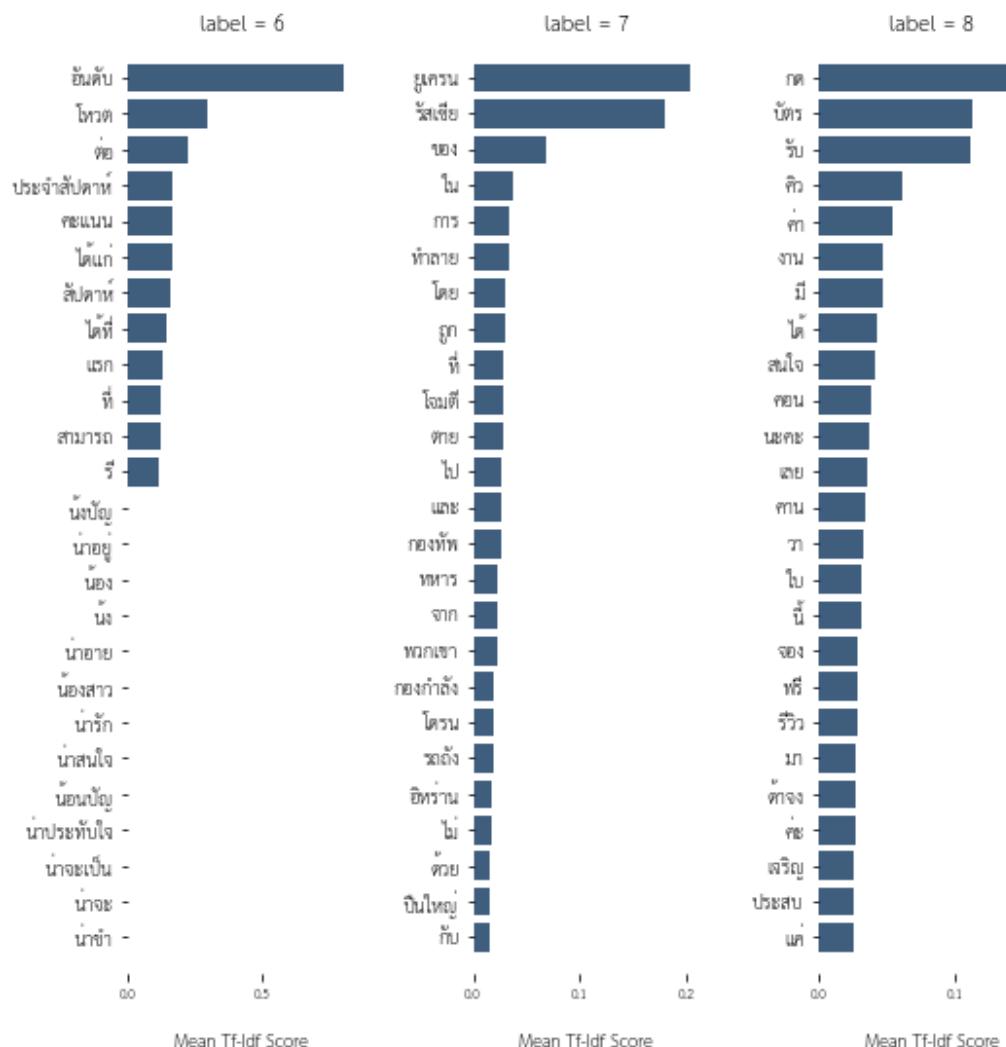
จากรูปที่ 4.37 และ 4.38 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่ 9 (วันที่ 9 พฤษภาคม 2565) พบหัวข้อที่มีลักษณะคล้ายกับของวันก่อนหน้าอย่าง หัวข้อเกี่ยวกับการมาตกรรมเด็ก หัวข้อเกี่ยวกับประเพณีลอยกระทง และหัวข้อเกี่ยวกับสังคมระหว่างยุเครนกับรัสเซีย แต่มีลักษณะเปลี่ยนไปจากการที่หัวข้อมีการผสมด้วยคำที่เกี่ยวกับการพนันมากขึ้น อาจหมายถึงการใช้ Hashtag ในการโปรโมทเงินกู้ออนไลน์หรือการพนัน นอกจากนี้ยังมีหัวข้อใหม่ที่เห็นได้ชัดคือรากแก้ว หรือกระรากแก้ว ซึ่งอาจเชื่อมโยงกับการที่วันดังกล่าวเป็นวันชาติ กองที่ 6 ของประเทศไทย

ตัวอย่างผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ



รูปที่ 4.39 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window

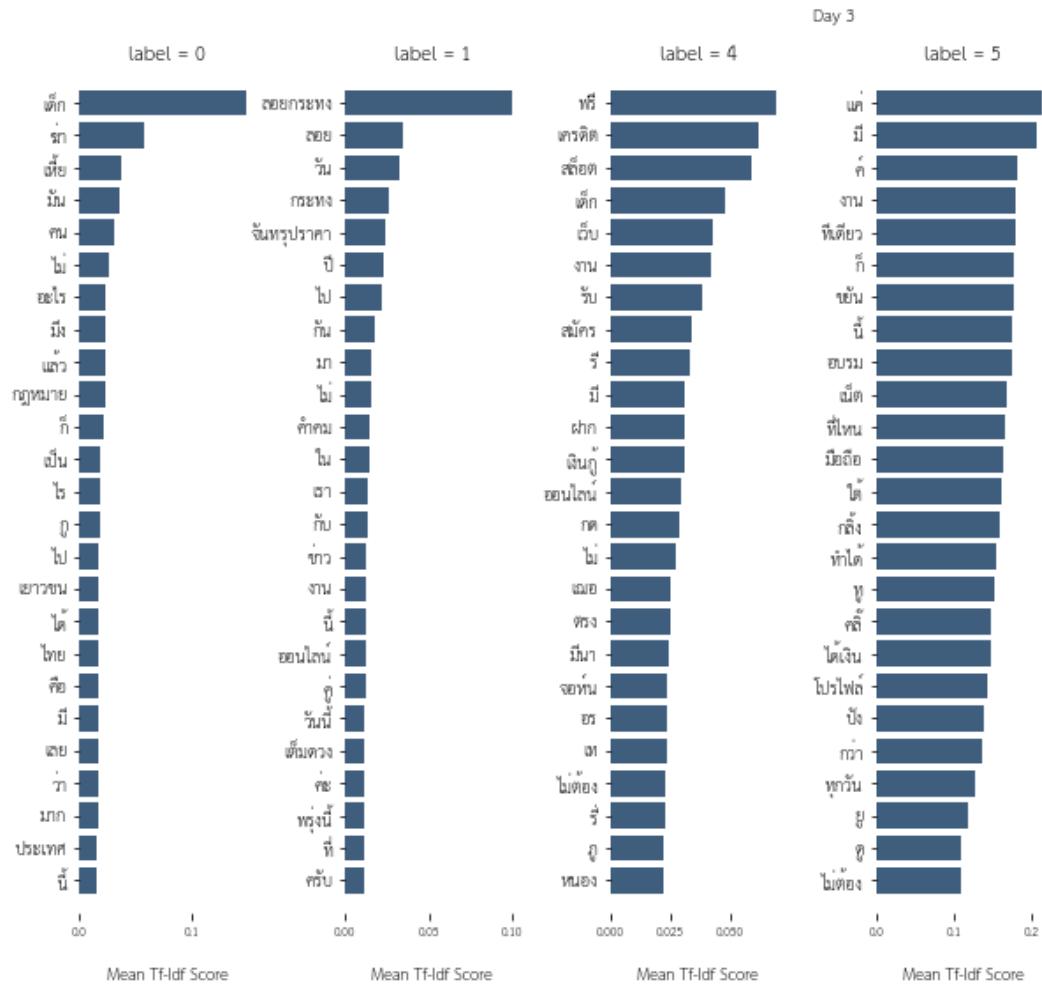
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ 4.40 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Sliding Window

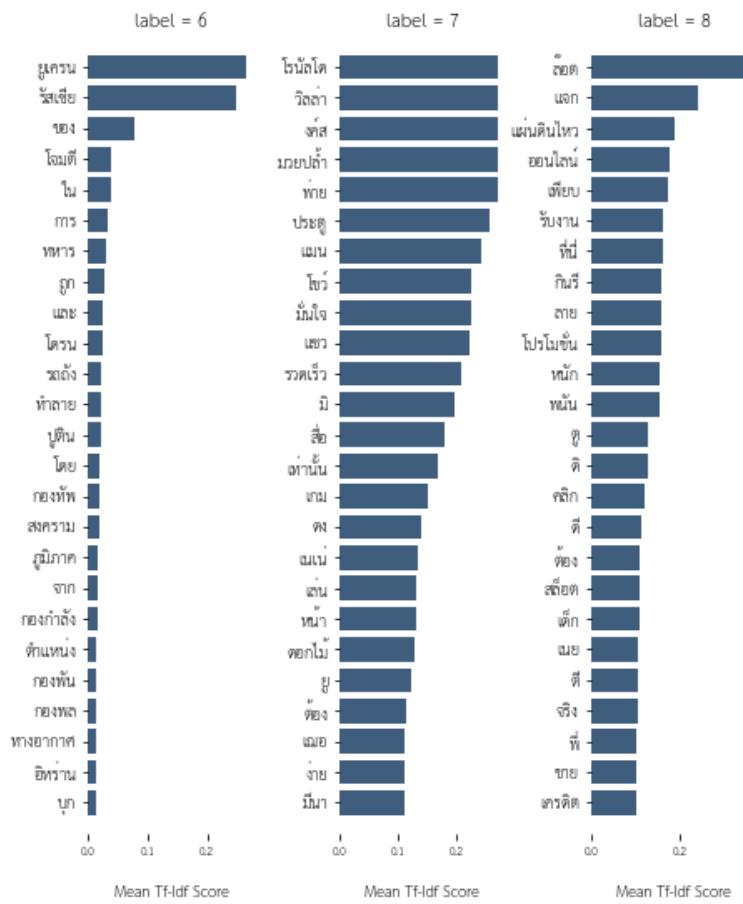
โดยใช้เวลาเป็นปีจัชประกอบ (ส่วนที่ 2)

จากรูปที่ 4.39 และ 4.40 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สอง (วันที่ 6 พฤษภาคม พ.ศ. 2565) พบว่าอย่างหัวข้อที่ 3 (BNK 48) และหัวข้อที่ 5 (ประเพณีลอดยกระดง) ยังคงมีการกล่าวถึงในวันที่สอง อีกทั้งยังมีบางหัวข้อที่มีการกล่าวถึงในวันที่หนึ่ง แต่ไม่ถูกกล่าวถึงในวันที่สองหายไป อย่าง Black Panther รวมถึงหัวข้อเกี่ยวกับจันทรุปราคาก็ไม่พบ เช่นกัน แต่มีการพนกรากล่าวถึงเรื่องใหม่อย่างเรื่องสังคมระหว่างยุเครน กับรัสเซีย โดยสำหรับวันที่สอง หัวข้อที่ทำการขยายคงมีไม่ครบ 10 หัวข้อ เช่นเดียวกับของวันที่หนึ่ง



รูปที่ 4.41 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window

## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

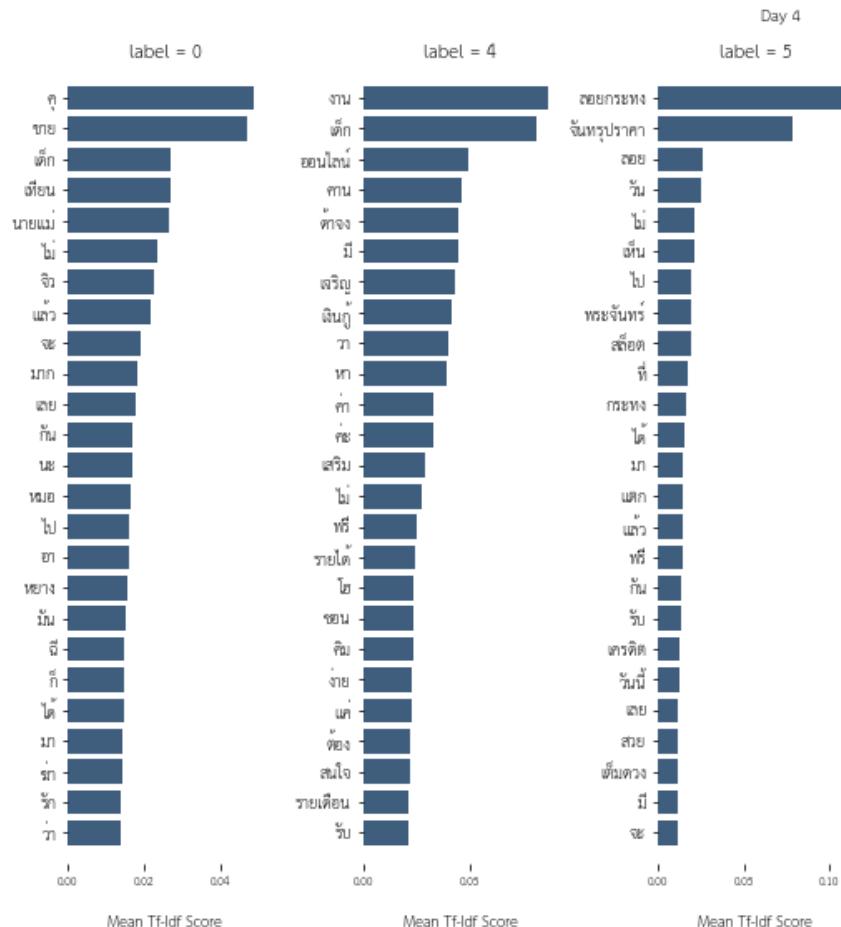


รูปที่ 4.42 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Sliding Window

โดยใช้เวลาเป็นปีจังประกอบ (ส่วนที่ 2)

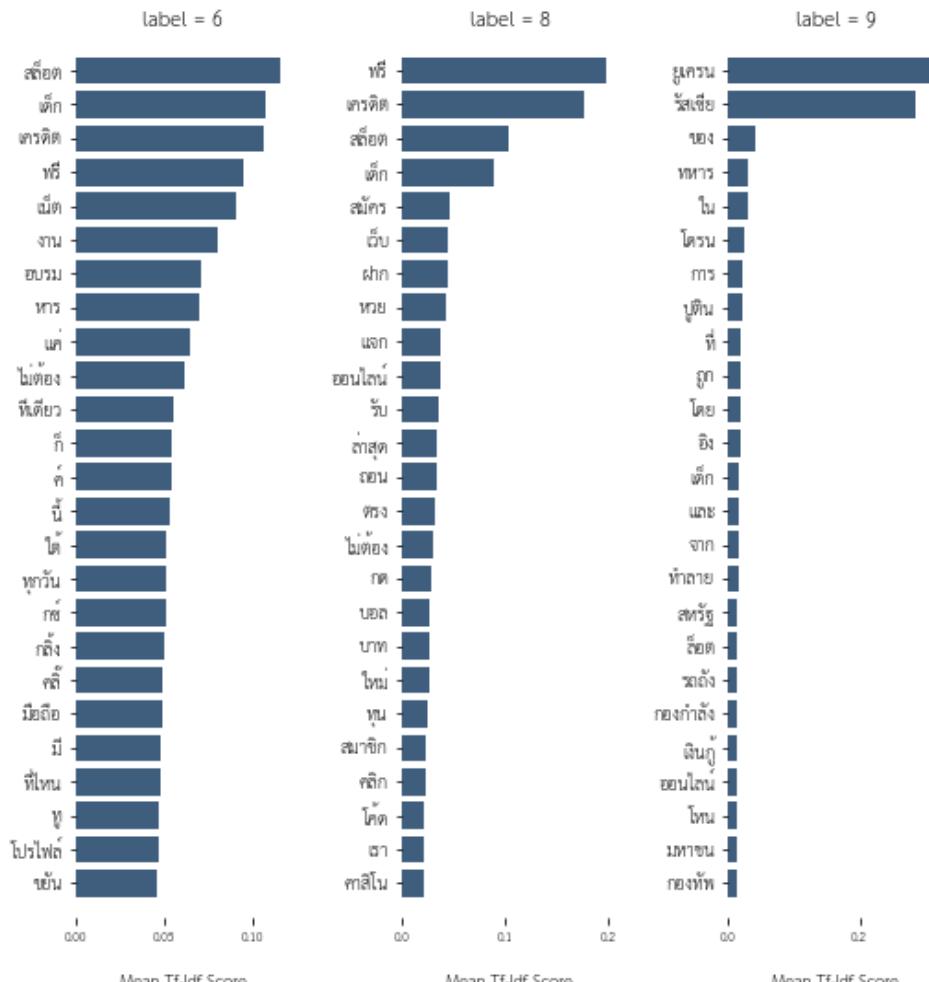
จากรูปที่ 4.41 และ 4.42 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สาม (วันที่ 7 พฤศจิกายน พ.ศ. 2565) พบว่าอย่างหัวข้อที่ 1 (loydkratong) และหัวข้อที่ 6 (singkram rahwawang yucren กับรัสเซีย) ยังคงมีการกล่าวถึงในวันที่สาม อีกทั้งยังมีบางหัวข้อที่มีการกล่าวถึงในวันที่สอง แต่ไม่ถูกกล่าวถึงในวันที่สามอย่าง BNK48 แต่มีการพนgrammer กล่าวถึงเรื่องใหม่อย่างเรื่องของโronal โดกับวิลล่า ซึ่งอยู่ในช่วงเดียวกับวันที่มีเกมการแข่งขันระหว่างแม่นแซสเตอร์ ยูไนเต็ด กับ แอสตัน วิลล่า จากหัวข้อที่ 7 และเช่นเดียวกับคดีที่เด็กอายุ 18 ปีฆ่ากรรมเด็กอายุ 13 ปี ที่เกิดในช่วงเวลาเดียวกัน ดังที่ผลลัพธ์ได้แสดงออกมากจากหัวข้อที่ 0 อีกทั้งยังมีเรื่องที่ถูกกลับมาพูดถึงอีกครั้งอย่าง จันทรุปราสาท ที่ถูกกล่าวถึงในวันที่หนึ่ง โดยสำหรับวันที่สาม หัวข้อที่ทำการหายังคงมีไม่ครบ 10 หัวข้อ เช่นเดียวกับสองวันก่อนหน้า

ตัวอย่างผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ



รูปที่ 4.43 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window

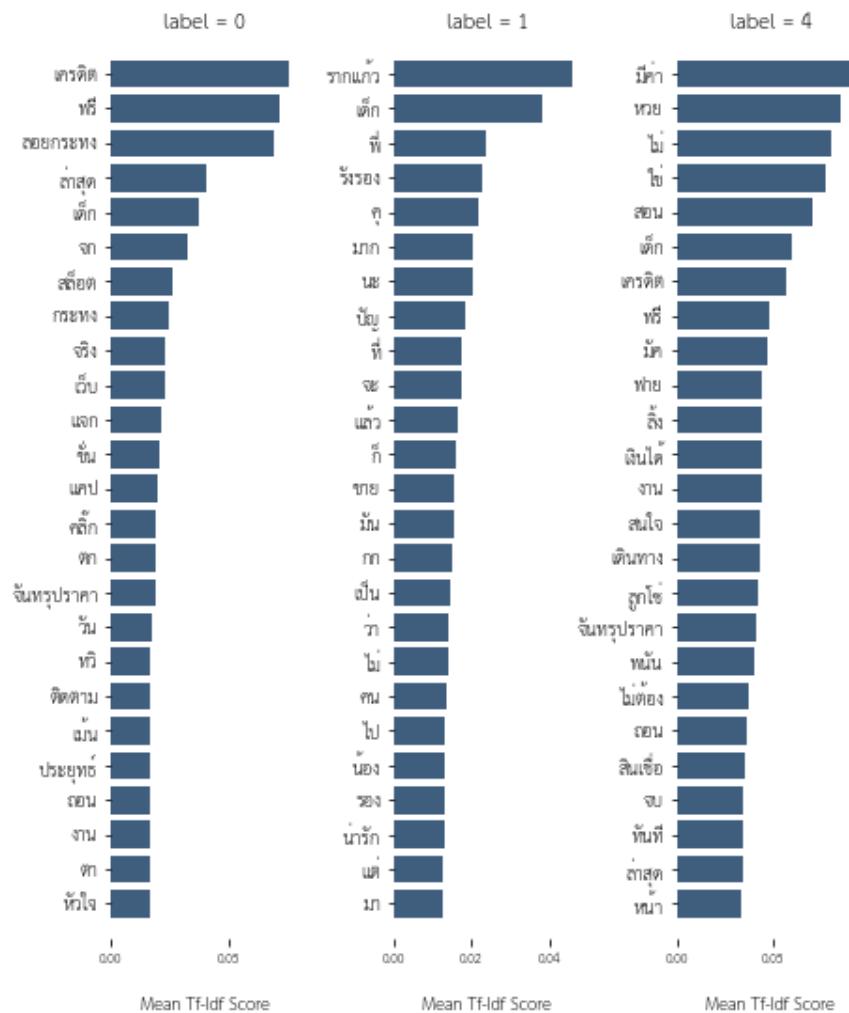
โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



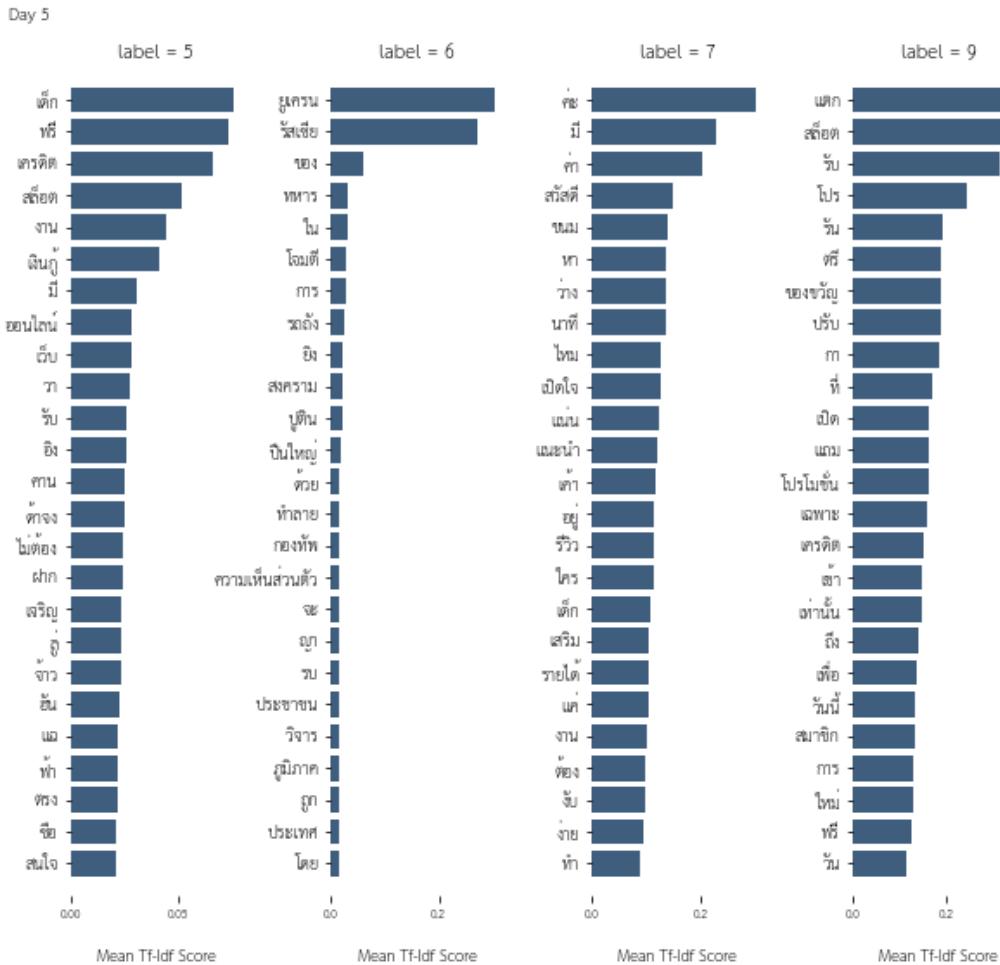
รูปที่ 4.44 ผลลัพธ์การกันไฟหัวข้อวันที่สี่จากการทดลองแบบ Expanding Window

โดยใช้เวลาเขียนใจจักรภพอฯ (ส่วนที่ 2)

จากรูปที่ 4.43 และ 4.44 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สี่ (วันที่ 8 พฤษภาคม พ.ศ. 2565) พบหัวข้อที่เห็นได้ชัดถ่ายกับวันก่อนหน้าอย่างหัวข้อที่กล่าวถึงประเพณีลอยกระทง รวมถึงจันทรุปราคาในหัวข้อที่ 5 หัวข้อที่เกี่ยวกับสกุลกรรมระหว่างผู้คนกับรัศมีเชิงในหัวข้อที่ 9 และหัวข้อที่เกี่ยวกับนิรภัย รวม และ โฆษณาออนไลน์ในหัวข้อที่ 4, 6 และ 8



รูปที่ 4.45 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ 4.46 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window

โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ 4.45 และ 4.46 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่ 9 พฤษภาคม พ.ศ. 2565) พบหัวข้อที่เห็นได้ชัดอย่างหัวข้อที่กล่าวถึงการพนัน โฆษณาออนไลน์ เป็นส่วนมาก โดยถูกพบในหัวข้อที่ 0, 4, 5 และ 9 อีกทั้งยังมีหัวข้อที่กล่าวถึงผลกระทบแก่ใน หัวข้อที่ 2 และยังคงมีการกล่าวถึงสังคมระหว่างประเทศนับร้อยเชียในหัวข้อที่ 6 เช่นเดียวกับที่ ผ่านมา

#### 4.2.4 ผลลัพธ์จากขั้นตอนที่ 3.2.4

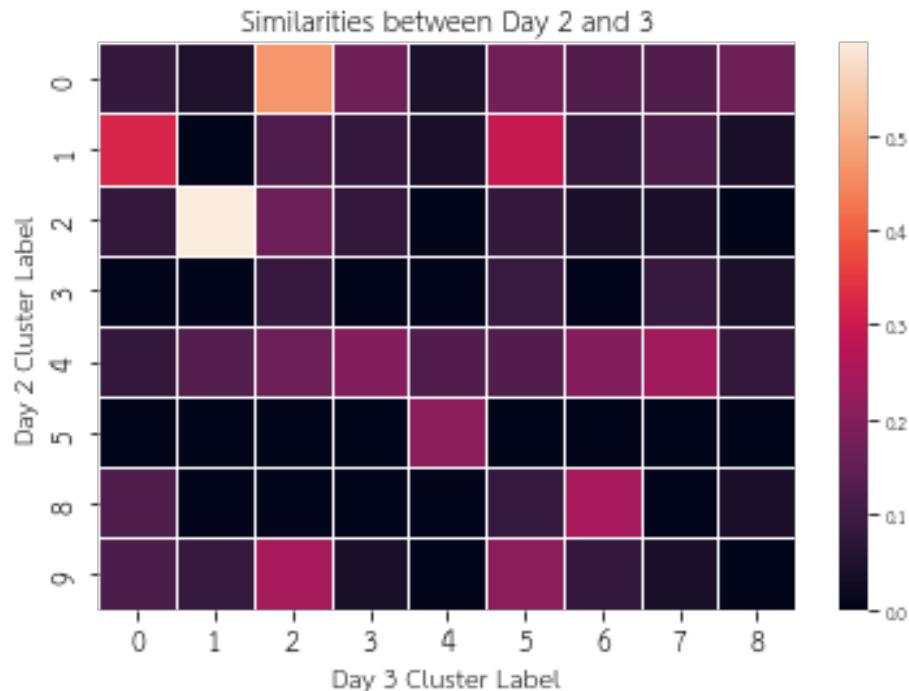
##### 4.2.4.1 ผลลัพธ์การทดลองย่อยที่ 3.2.4.1

จากการดำเนินงานขั้นตอนที่ 3.2.4.1 เป็นการประเมินความเชื่อมโยงและความสามารถในการตีความหัวข้อ ในส่วนแรกจะเป็นการประเมินความเชื่อมโยงจากหัวข้อโดยการใช้ค่า Cosine Similarity ในการสำรวจการเปลี่ยนไป และความเชื่อมโยงของหัวข้อระหว่างวัน โดยมีตารางสรุปผลการทดลองดังตารางที่ 4.4 และสามารถดูผลลัพธ์การประเมินความเชื่อมโยงของหัวข้ออื่น ๆ ได้จากภาคผนวก ข. (หน้าที่ 134)

**ตารางที่ 4.4** สรุปผลการทดลองการประเมินความเชื่อมโยงของหัวข้อด้วย Cosine Similarity

รูปแบบการทดลอง	ผลการทดลองที่พบ
การทดลองแบบ Sliding Window แบบไม่ใช้เวลา เป็นปัจจัยประกอบ	พบความเชื่อมโยงของหัวข้อระหว่างวัน แสดงให้เห็นการเปลี่ยนแปลงของหัวข้อ มีความเชื่อมโยงของหัวข้อระหว่างวันมากกว่าการทดลองแบบใช้เวลา และพบหัวข้อบางหัวข้อที่เชื่อมโยงกับหัวข้อวันถัดไปมากกว่าหนึ่งหัวข้อ
การทดลองแบบ Expanding Window แบบไม่ใช้เวลา เป็นปัจจัยประกอบ	พบความเชื่อมโยงของหัวข้อระหว่างวัน แสดงให้เห็นการเปลี่ยนแปลงของหัวข้อ มีความเชื่อมโยงของหัวข้อระหว่างวันมากกว่าการทดลองแบบใช้เวลา และพบหัวข้อหลายหัวข้อที่เชื่อมโยงกับหัวข้อวันถัดไปมากกว่าหนึ่งหัวข้อ
การทดลองแบบ Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ	พบความเชื่อมโยงของหัวข้อระหว่างวัน แสดงให้เห็นการเปลี่ยนแปลงของหัวข้อ มีความเชื่อมโยงของหัวข้อระหว่างวันน้อยแต่สามารถตีความได้ชัดเจน หัวข้อส่วนใหญ่ถ้ามีความเชื่อมโยง จะเชื่อมโยงกับหัวข้อใดหัวข้อหนึ่งของวันถัดไป
การทดลองแบบ Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ	พบความเชื่อมโยงของหัวข้อระหว่างวัน แสดงให้เห็นการเปลี่ยนแปลงของหัวข้อ มีความเชื่อมโยงของหัวข้อระหว่างวันมากกว่าการทดลอง Sliding Window ชัดเจน

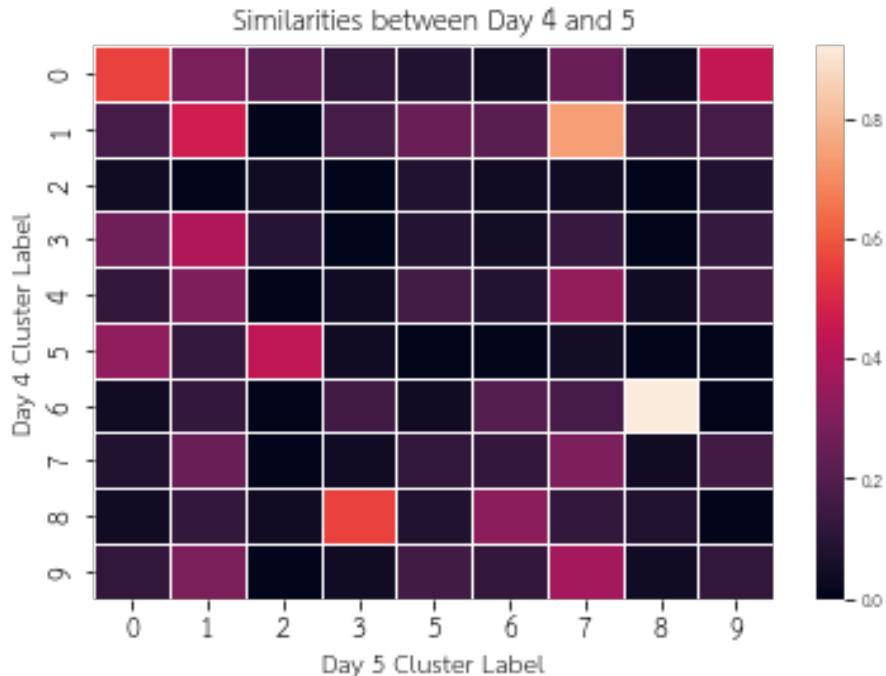
ตัวอย่างผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบไม่ใช้วремาเป็นปัจจัยประกอบ



รูปที่ 4.47 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช่วремาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ 4.52 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สอง (วันที่ 6 พฤศจิกายน 2565) และ วันที่สาม (วันที่ 7 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณาไม่เปลี่ยนแค่หัวข้อที่ 0, 1, และ 2 เท่านั้นที่มีความคล้ายกับหัวข้อกับวันถัดไปอย่างเห็นได้ชัด ยกตัวอย่างเช่นหัวข้อที่ 0 คล้ายกับหัวข้อที่ 2 ของวันถัดไป ซึ่งพบว่าเป็นหัวข้อที่เกี่ยวกับ lobbying กระทรวง แต่มีเนื้อหาที่เปลี่ยนไปเล็กน้อย และหัวข้อที่ 2 คล้ายกับหัวข้อที่ 1 ของวันถัดไป ซึ่งเป็นเนื้อหาที่มีการพูดถึงอยู่ตลอดเวลาอย่างสูงในรายงานระหว่างยุคเศรษฐกิจ萧条 หัวข้ออื่นนอกจากหัวข้อข้างต้นจะเห็นได้ว่ามีความคล้ายกับหัวข้ออื่นของวันถัดไปค่อนข้างน้อย ซึ่งหัวข้อคังก์ล่าวมีหัวข้อที่มีเนื้อหาเกี่ยวกับ ภาระน้ำ วงศ์ BNK48 และอื่นๆที่สามารถติดตามได้มาก

ตัวอย่างผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ

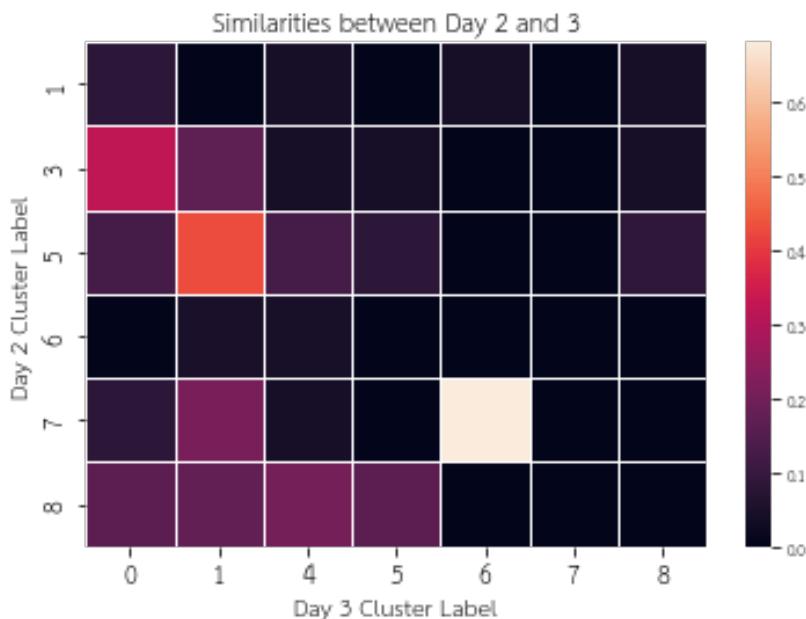


รูปที่ 4.48 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ 4.48 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สี่ (วันที่ 8 พฤศจิกายน 2565) และ วันที่ห้า (วันที่ 9 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณา มีหัวข้อที่คล้ายกับวันถัดไปอย่างเห็นได้ชัด อย่าง หัวข้อที่ 0, 1, 5, 6, และ 8 โดยมีเนื้อหาเบื้องต้น เช่น หัวข้อที่ 0 เกี่ยวกับการมาตรฐานเด็กซึ่งบังเกี่ยวโยงต่อเนื่องไปยังวันถัดไปในหัวข้อที่ 0 หัวข้อที่ 1 ที่มีการสมของคำหลากหลาย จึงทำให้ติดความได้ยาก คล้ายกับหัวข้อที่ 7 ของวันถัดไป หัวข้อที่ 5 และ หัวข้อที่ 2 ของวันถัดไป ซึ่งอาจหมายถึงผลกระทบแก้วที่ถูกกล่าวถึงอย่างชัดเจนในวันที่ห้า (วันที่ 9 พฤศจิกายน 2565) หัวข้อที่ 6 เกี่ยวกับการหารายได้เสริมที่มีการพูดถึงต่อเนื่องไปจนถึงวันถัดไปในหัวข้อที่ 8 และหัวข้ออื่น ๆ อย่างหัวข้อที่ 2 เกี่ยวกับสังคม ระหว่างยุคเรนและรัสเซีย ไม่พูดเป็นหัวข้อในวันถัดไป รวมถึงหัวข้อที่ 3 ที่มีเนื้อหาเกี่ยวกับ lobbying ที่มีอยู่ในวันถัดไปหัวข้อ 1 แต่มีความคล้ายกันน้อย เนื่องจากคำที่เป็นองค์ประกอบ แตกต่างกัน

จากผลลัพธ์การหาความคล้ายคลึงของหัวข้อด้วยค่า Cosine Similarity ตลอด 5 วันที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ จะเห็นว่ามีการแสดงเนื้อหาเป็นหัวข้อได้ชัดเจนหลายหัวข้อ เช่นเดียวกับวิธี Sliding Window แต่ยังมีหัวข้อที่มีความไม่ได้ปะปนมา รวมถึงความเชื่อมโยงของหัวข้อระหว่างวันในการทดลองแบบ Expanding Window ถูกพบมากกว่าแบบ Sliding Window เล็กน้อย และยังมีการพบหัวข้อใหม่อีก เช่น หัวข้อที่เกี่ยวกับการมาตรฐานเด็กในวันที่สาม, หัวข้อที่เกี่ยวกับผลกระทบแก้วในวันที่ห้า ส่วน หัวข้อที่เกี่ยวกับระหว่างยุเครนและรัสเซียที่เกิดขึ้นมาตลอด ไม่พบในวันที่ห้า ต่างจากการทดลองแบบ Sliding Window

ตัวอย่างผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ

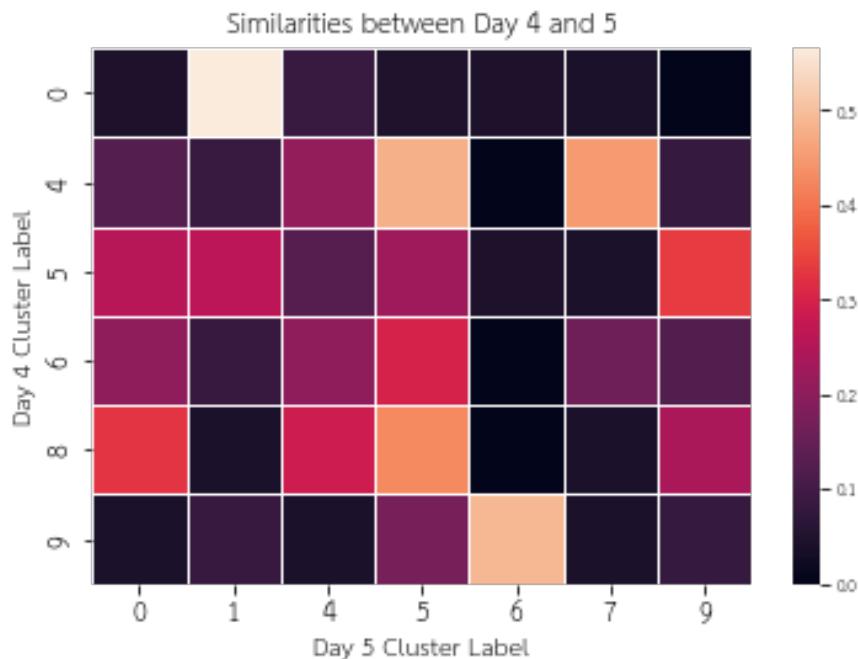


รูปที่ 4.49 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ 4.49 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สอง (วันที่ 6 พฤศจิกายน พ.ศ. 2565) และ วันที่สาม (วันที่ 7 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 7 มีความคล้ายกับหัวข้อที่ 6 ของวันถัดไปมากที่สุด โดยที่มีหัวข้อที่ 3 กับหัวข้อที่ 0 ของวันถัดไป และมี

หัวข้อที่ 5 กับหัวข้อที่ 1 ของวันถัดไป ที่มีความคล้ายคลึงกันของหัวข้อรองลงมา ซึ่งเมื่อดูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นว่าเนื้อหาที่เปลี่ยนไปพอสมควร แต่คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อยังคงเดิมเช่น หัวข้อเกี่ยวกับสังคมรัฐบาลยุเครนกับรัสเซีย จะเห็นว่า ยุเครน และ รัสเซีย ยังคงเป็นคำที่ถูกกล่าวถึงบ่อย ๆ หรือคำว่า ลอยกระทง ที่ถูกพบในหัวข้อประเพณีลอยกระทง นอกจากหัวข้อที่ได้กล่าวไปข้างต้นจะเห็นได้ว่ามีความคล้ายกับหัวข้ออื่นของวันถัดไปค่อนข้างน้อย ประกอบด้วย หัวข้อ 1, 6 และ 8 ซึ่งไม่สามารถตีความหัวข้อดังกล่าวได้

ตัวอย่างผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วย Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ



รูปที่ 4.50 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี

Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ 4.50 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สี่ (วันที่ 8 พฤศจิกายน พ.ศ. 2565) และ วันที่ห้า (วันที่ 9 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 0 กับหัวข้อที่ 1 ของวันถัดไป หัวข้อที่ 4 กับหัวข้อที่ 5 ของวันถัดไป และหัวข้อที่ 9 กับหัวข้อที่ 6 ของวันถัดไปมากที่สุด ซึ่งเมื่อดูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็น

ได้จากรูปที่ 4.43 และ 4.44 ว่า คำที่เป็นองค์ประกอบหลักที่สื่อถึงความหมายของหัวข้อยังคงเดิม เช่น หัวข้อที่เกี่ยวกับโฆษณาออนไลน์ จะเห็นว่า งาน เด็ก และ ออนไลน์ และหัวข้อที่เกี่ยวกับสังคมระหว่างยุเครนกับรัสเซีย จะเห็นว่า ยุเครน และ รัสเซีย เป็นคำที่ถูกกล่าวถึงบ่อยในหัวข้อดังกล่าว แต่หัวข้อที่ 0 ไม่สามารถนำมาตีความหมายด้วยคำเหล่านี้ได้

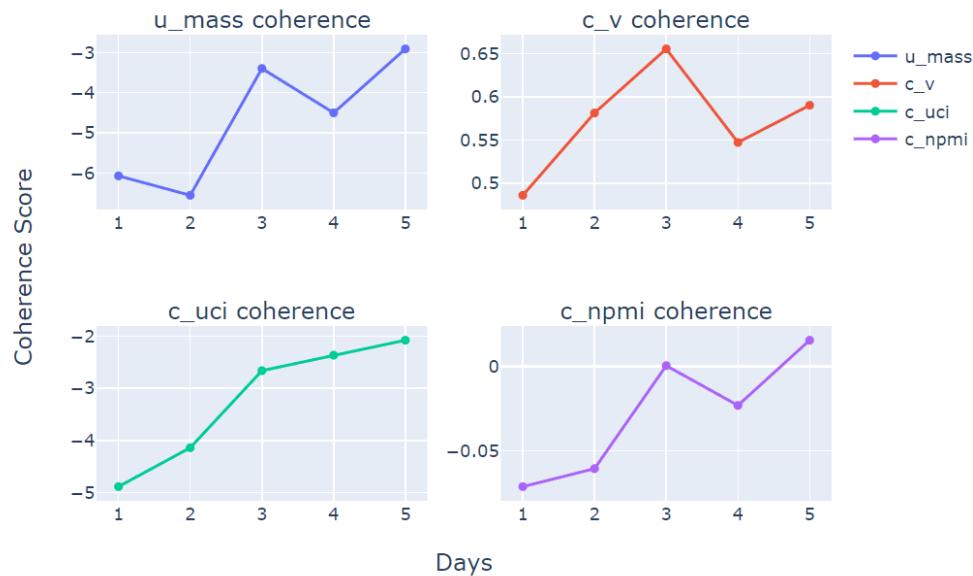
จากผลลัพธ์การหาความคล้ายคลึงของหัวข้อด้วยค่า Cosine Similarity ตลอด 5 วันที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ จะเห็นว่ามีการแสดงเนื้อหาเป็นหัวข้อ ได้ชัดเจนหลายหัวข้อ แต่ยังมีหัวข้อที่ตีความไม่ได้ปะปนมา เช่นเดียวกับวิธี Sliding Window รวมถึงความเชื่อมโยงของหัวข้อระหว่างวันในการทดลองแบบ Expanding Window ถูกพบมากกว่าแบบ Sliding Window เนื่องจากชุดข้อมูลฝึกสอนใช้วิธีการสะสมชุดข้อมูลตั้งแต่วันแรกจนถึงวันปัจจุบัน จึงทำให้ตัวแบบจำลองมีการเรียนรู้จากหัวข้อที่ถูกกล่าวถึงในวันก่อน ๆ หน้าต่อหน้า และยังมีการpubหัวข้อใหม่อよ่างเช่นหัวข้อที่เกี่ยวกับการซื้อบัตรคอนเสิร์ตในวันที่หนึ่ง และหัวข้อที่เกี่ยวกับโฆษณาออนไลน์ในวันที่หลัง ส่วนหัวข้อที่เกี่ยวกับสังคมระหว่างยุเครนและรัสเซียที่เกิดขึ้นมาตลอด ไม่พบในความคล้ายคลึงของหัวข้อเพียงช่วงเดียว (วันที่หนึ่ง - วันที่สอง) เช่นเดียวกับวิธี Sliding Window

#### 4.2.4.2 ผลลัพธ์การทดลองอย่างที่ 3.2.4.2

จากการดำเนินงานขั้นตอนที่ 3.2.4.2 เป็นการประเมินความสามารถในการตีความของหัวข้อด้วย Topic Coherences สำหรับส่วนที่สองของการประเมินคือการเปรียบเทียบความสามารถในการตีความของผลลัพธ์ด้วยค่า Topic Coherences ของหัวข้อที่ได้จากการทดลองด้วยวิธี Sliding Window และ Expanding Window ทั้งแบบใช้และไม่ใช้เวลาเป็นปัจจัยประกอบด้วยค่า Topic Coherences ซึ่งประกอบด้วยวิธีการคำนวณที่แตกต่างกัน 4 วิธี ได้แก่ U\_MASS, C\_UCI, C\_NPMI และ C\_V โดยผู้วิจัยจัดการหาค่า Topic Coherences ของหัวข้อจากทุกวันทดลองระยะเวลาของข้อมูลที่ทำการทดลองแต่ละแบบ และนำผลลัพธ์มาเปรียบเทียบเพื่อหาวิธีการสร้างแบบจำลองที่ให้ผลลัพธ์ที่ดีที่สุด

ผลลัพธ์การประเมินความสามารถในการตีความของหัวข้อด้วยค่า Topic Coherences จากการสร้างแบบจำลองด้วยวิธี Sliding Window และไม่ใช้เวลาเป็นปัจจัยประกอบ

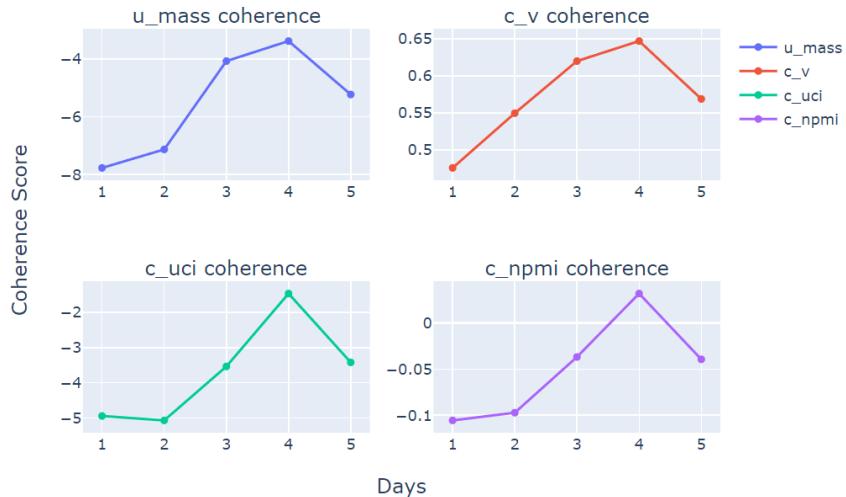
Normal Sliding Topic Coherence



รูปที่ 4.67 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Sliding Window  
แบบไม่ใช้เวลาเป็นปัจจัยประกอบ

ผลลัพธ์การประเมินความสามารถในการตีความของหัวข้อด้วยค่า Topic Coherences จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ

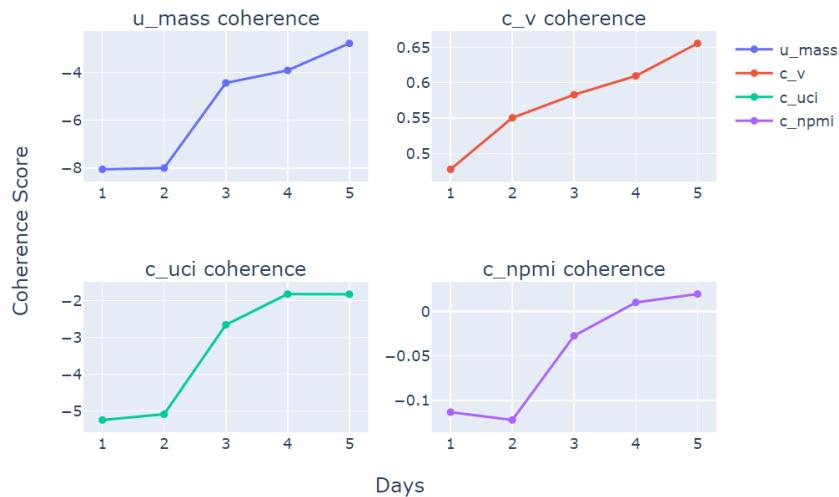
Normal Expanding Topic Coherence



รูปที่ 4.68 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ

ผลลัพธ์การประเมินความสามารถในการตีความของหัวข้อด้วยค่า Topic Coherences จากการสร้างแบบจำลองด้วยวิธี Sliding Window และใช้เวลาเป็นปัจจัยประกอบ

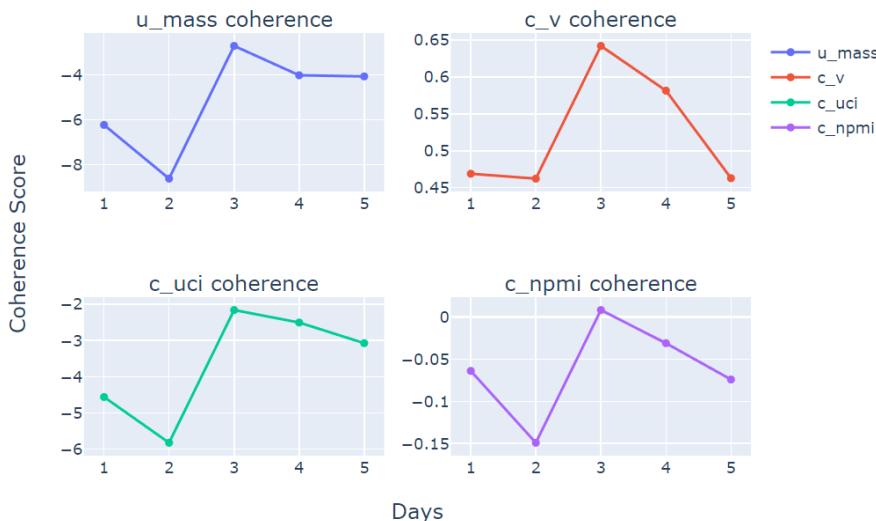
Time-related Sliding Topic Coherence



รูปที่ 4.69 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Sliding Window  
แบบใช้เวลาเป็นปัจจัยประกอบ

ผลลัพธ์การประเมินความสามารถในการตีความของหัวข้อด้วยค่า Topic Coherences จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ

Time-related Expanding Topic Coherence



รูปที่ 4.70 ผลลัพธ์การเปรียบเทียบค่า Topic Coherences ของหัวข้อจากวิธี Expanding Window

แบบใช้เวลาเป็นปัจจัยประกอบ

จากผลลัพธ์ของการหาค่า Topic Coherences ดังรูปที่ 4.67 ถึง 4.70 จะเห็นได้ว่าค่า Topic Coherences ทั้งสี่แบบจากทุกการทดลองได้ผลลัพธ์อยู่ในช่วงที่ใกล้เคียงกัน โดยเมื่อหาค่าเฉลี่ยของ Topic Coherences ของผลลัพธ์หัวข้อแล้วได้ผลลัพธ์ดังตารางที่ 4.5 สำหรับค่า Topic Coherence แบบ U\_MASS การทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบให้ค่าโดยรวมสูงที่สุด สำหรับแบบ C\_V การทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบให้ผลลัพธ์สูงที่สุด แต่มากกว่าแบบอื่นเพียงเล็กน้อยเท่านั้น สำหรับแบบ C\_UCI การทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบให้ผลลัพธ์สูงที่สุด แต่ช่วงของค่าลดลงระหว่างที่พิจารณาต่างกับการทดลองอื่นเพียงเล็กน้อย สำหรับแบบ C\_NPMI การทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบให้ผลลัพธ์ที่ดีที่สุด แต่ก็ต่างกับการทดลองอื่นเพียงเล็กน้อยเช่นกัน

**ตารางที่ 4.5** ผลลัพธ์ค่าเฉลี่ยของค่า Topic Coherences ที่ได้จากการประเมินผลลัพธ์หัวข้อ

รูปแบบการทดลอง	ประเกต Coherences			
	U_MASS	C_V	C_UCI	C_NPMI
การทดลองแบบ Sliding Window แบบไม่ใช้เวลา	-4.688109	0.572075	-3.225898	-0.027669
การทดลองแบบ Expanding Window แบบไม่ใช้เวลา	-5.513630	0.572126	-3.686523	-0.049449
การทดลองแบบ Sliding Window แบบใช้เวลา	-5.434477	0.575224	-3.328747	-0.046541
การทดลองแบบ Expanding Window แบบใช้เวลา	-5.134367	0.523629	-3.624031	-0.062051

## บทที่ 5

### สรุปผลการวิจัยและข้อเสนอแนะ

#### 5.1 สรุปผลการวิจัย

##### 5.1.1 สรุปผลการทดลองที่ 1

จากการทดลองสร้างแบบจำลองหัวข้อด้วยชุดข้อมูล AG News พบร่วมแบบจำลองที่สามารถให้ผลลัพธ์หัวข้อที่ตีความได้ง่ายที่สุดคือแบบจำลอง LDA และ ชุดข้อมูล Twitter Covid-19 ได้พบว่าผลลัพธ์การสร้างแบบจำลองเพื่อค้นหาหัวข้อซ่อนเร้นนั้นสามารถตีความได้ยาก ทุกแบบจำลองให้ผลลัพธ์ที่มีลักษณะคำชา็กันอยู่หลายหัวข้อ ต่อมามีอีกด้วยการทำทดลองเปลี่ยนค่าจำนวนหัวข้อสำหรับแบบจำลองกับทุกแบบจำลองและชุดข้อมูล เพื่อค้นหาจำนวนหัวข้อที่เหมาะสมกับแบบจำลองและชุดข้อมูลพบว่าสำหรับชุดข้อมูล AG News หัวข้อที่เหมาะสมกับแบบจำลอง LDA และ GSDMM คือ 5 หัวข้อ แบบจำลอง NMF คือ 7 หัวข้อ และเมื่อใช้ชุดข้อมูล Twitter COVID-19 จะได้หัวข้อที่เหมาะสมกับแบบจำลอง LDA คือ 5 หัวข้อ สำหรับแบบจำลอง GSDMM คือ 4 หัวข้อ และแบบจำลอง NMF คือ 3 หัวข้อ เมื่อทำการเปรียบเทียบทั้งสามแบบจำลองกับการคำนวณค่าความสอดคล้องทุกๆแบบ จะเห็นว่า เมื่อใช้ชุดข้อมูล AG News ค่าความสอดคล้องของหัวข้อจากทุกแบบจำลองเปลี่ยนไปในทิศทางเดียวกันเมื่อจำนวนหัวข้อเพิ่มขึ้น แต่เมื่อใช้ชุดข้อมูล Twitter COVID-19 ค่าความสอดคล้องแต่ละแบบจำลองมีการเปลี่ยนไปแตกต่างกัน โดยจะเห็นได้ชัดในจากผลลัพธ์ของแบบจำลอง LDA จากการคำนวณแบบ C\_V และ C\_UCI ที่มีค่าความสอดคล้องเพิ่มขึ้น และลดลง ในขณะที่แบบจำลองอื่นมีค่าเพิ่มขึ้นเพียงเล็กน้อย เมื่อทำการทดลองสร้างแบบจำลองด้วยจำนวนหัวข้อที่เหมาะสมแล้วประเมินค่าความสอดคล้องทุกแบบและเปรียบเทียบกันสำหรับชุดข้อมูล AG News จะพบว่าค่าความสอดคล้องของแบบจำลอง LDA มากที่สุดเมื่อใช้การคำนวณแบบ C\_V และ C\_NPMI ซึ่งสรุปจากการเปรียบเทียบค่าความสอดคล้องได้ว่าในการทดลองสร้างแบบจำลองหัวข้อซ่อนเร้นด้วยชุดข้อมูล AG News แบบจำลอง NMF ให้

ผลลัพธ์ที่ดีที่สุด โดยมีจำนวนหัวข้อ 7 หัวข้อ สำหรับชุดข้อมูล Twitter COVID-19 จะพบว่าค่าความสอดคล้องของแบบจำลอง GSDMM มีค่ามากที่สุดเมื่อคำใช้การคำนวณแบบ U\_MASS, C\_UCI และ C\_NPMI โดยสามารถสรุปจากการเปรียบเทียบค่าความสอดคล้องได้ว่า แบบจำลอง GSDMM ที่ใช้จำนวนหัวข้อ 4 หัวข้อ ให้ผลลัพธ์คะแนนความสอดคล้องดีที่สุดในการสร้างแบบจำลองหัวข้อ โดยใช้ชุดข้อมูล Twitter COVID-19

### 5.1.2 สรุปผลการทดลองที่ 2

จากผลการทดลองสร้างแบบจำลองหัวข้อด้วยวิธีที่ต่างกันสองวิธีคือ Sliding Window และ Expanding Window ซึ่งมีการใช้ข้อมูลที่ใช้และไม่ใช้เวลาเป็นปัจจัยประกอบเพื่อสร้างแบบจำลอง K-Mean Clustering สำหรับการค้นหาหัวข้อจากข้อมูลข้อความคิดเห็นจาก Twitter พบว่าหัวข้อผลลัพธ์ที่ได้มีเนื้อหาที่รวมกันเป็นหัวข้อที่สามารถตีความได้ แต่ยังพบหัวข้อที่มีค่าที่ประปนไม่สามารถตีความได้ จะเห็นว่าผลลัพธ์ที่ได้จากวิธี Sliding Window ให้ผลลัพธ์การค้นหาหัวข้อไม่ต่างกับวิธี Expanding Window มากนัก มีลักษณะของการพบหัวข้อที่เกิดขึ้นใหม่และหายไปในแต่ละวัน มีค่าส่วนประกอบของหัวข้อที่เปลี่ยนแปลงไป แต่สำหรับการทดลองที่ใช้เวลาเป็นปัจจัยประกอบพบว่าผลลัพธ์การค้นหาหัวข้อมีหัวข้อที่ตีความได้ยากน้อยกว่าการทดลองแบบไม่ใช้เวลาเป็นปัจจัยประกอบซึ่งเป็นข้อได้เปรียบที่เห็นได้ชัด และยังพบว่าจำนวนหัวข้อที่ได้จะน้อยกว่าจำนวนหัวข้อของ การทดลองแบบไม่ใช้เวลาเป็นปัจจัยประกอบอยู่พอสมควรซึ่งอาจอนุมานได้ว่าช่วงเวลาที่ทำการค้นหาหัวข้อมีเนื้อหาของข้อมูลต่างจากช่วงเวลาของข้อมูลที่ทำการฝึกสอนรวมถึงการใช้เวลาทำให้การจับกลุ่มของ Cluster เปลี่ยนแปลงไป และหลังจากการประเมินการเปลี่ยนไปของหัวข้อด้วยค่า Cosine Similarity ของผลลัพธ์หัวข้อที่ได้ระหว่างวันที่พิจารณา พบว่าการทดลองด้วยวิธี Expanding Window มีความเชื่อมระหว่างหัวข้อระหว่างวันมากกว่าการทดลองด้วยวิธี Sliding Window ถ้าทั้งการประเมินความเชื่อมโดยของหัวข้อสำหรับการทดลองที่ใช้เวลาเป็นปัจจัยประกอบทำได้ง่ายกว่าการทดลองแบบไม่ใช้เวลาจากความเชื่อมโดยที่เห็นได้ชัด และส่วนใหญ่เชื่อมโดยกับหัวข้อใดหัวข้อหนึ่งของวันถัดไป และในด้านการประเมินความสามารถในการตีความของผลลัพธ์หัวข้อด้วยค่า Topic Coherences พบว่าช่วงของค่าเฉลี่ย Coherences แต่ละแบบที่ได้จากการ

ทดลองทุกแบบนั้นที่ใกล้เคียงกัน แต่การทดลองแบบ Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบให้ผลลัพธ์ดีที่สุด ซึ่งจากการทดลองทั้งหมดสามารถสรุปได้ว่าถึงแม้ผลลัพธ์ในการวัดความสามารถในการตีความของหัวข้อด้วย Coherences การทดลองแบบ Sliding Window แบบไม่ใช้เวลาประกอบจะได้ผลลัพธ์ดีที่สุด แต่ผลลัพธ์ของการทดลองแบบใช้เวลาเป็นปัจจัยประกอบยังไบเบริยนในด้านของผลลัพธ์หัวข้อที่สามารถตีความได้นากกว่า อีกทั้งยังประเมินความเชื่อมโยงได้ง่ายกว่าการทดลองแบบไม่ใช้เวลา

## 5.2 ปัญหาและอุปสรรคในงานวิจัย

- ความไม่แม่นยำของข้อมูลข้อความที่ทำการเก็บผ่าน Application โดยผู้วิจัยได้ทำการเรียกใช้เครื่องมือเพื่อรูปแบบเว็บбрауз์สำหรับการสักดิ้นและแปลงข้อมูลเสียงซึ่งเป็นคำพูดให้อยู่ในรูปข้อความซึ่งเครื่องมือดังกล่าวมีความแม่นยำต่ำ ทำให้ข้อมูลเสียงที่ทำการเก็บลงฐานข้อมูลจำเป็นต้องมีการนำไปประมวลผลเพื่อเปลี่ยนข้อความเสียงให้เป็นตัวอักษรอีกรอบ
- เครื่องมือแปลงข้อความเสียงจากbrauzerไม่ทำงาน โดยเมื่อทำการเปิดเว็บแอปพลิเคชันจากต่างอุปกรณ์ เช่น การเปิดในอุปกรณ์โทรศัพท์ที่เป็นระบบปฏิบัติการ Android และการเปิดในระบบปฏิบัติการ Apple บริษัทจะทำการเลือกเครื่องมือเพื่อรูปฐานต่างกันซึ่งในบางบริษัท เครื่องมือไม่สูญเสียข้อมูล ทำให้ข้อมูลข้อความตัวอักษรไม่ปรากฏเมื่อทำการเก็บลงฐานข้อมูล ทำให้ต้องมีการแปลงข้อความเสียงภายหลังอีกรอบ
- ใช้ระยะเวลาในการขอสิทธิ API เนื่องจากการนำชุดข้อมูล Twitter COVID-19 มาใช้ทางผู้ที่รวบรวมข้อมูลทำการรวบรวมข้อมูลเฉพาะ tweet\_id เท่านั้นซึ่งดำเนินต่อไปมีการนำไปดึงข้อมูลอีกรอบผ่านชุดเครื่องมือซึ่งมีความจำเป็นต้องใช้ API Credential ซึ่งกระบวนการขอ API ใช้เวลานานและจำเป็นต้องให้ข้อมูลกับทาง Twitter หลายครั้งกว่าจะได้รับการอนุมัติ
- การเก็บข้อมูลทำได้ยาก เนื่องจากในช่วงที่ทำการสร้างแอปพลิเคชันและการเก็บข้อมูล เป็นช่วงที่มีการระบาดของโรค COVID-19 ทำให้เป็นการเรียนการสอนจากที่บ้าน จึงยากต่อการเก็บข้อมูลจากบริเวณมหาวิทยาลัย

- ข้อจำกัดในการดึงข้อมูลจาก Twitter เนื่องจาก Twitter Developer API ให้สิทธิในการดึงข้อมูลย้อนหลังได้ 7 วันเท่านั้นทำให้ข้อมูลที่ดึงมาใช้ในการทดลองมีระยะเวลาที่สั้น ซึ่งอาจทำให้เห็นแนวโน้มการพูดถึงหัวข้อต่าง ๆ ได้อย่างชัดเจนยิ่งขึ้น
- ลักษณะความคิดเห็นจาก Twitter เนื่องจากมีการดึงข้อมูลจาก Hashtag และใน Hashtag ที่เป็นที่นิยมจะมีการโปรโมทเนื้อหาที่ไม่เกี่ยวกับ Hashtag
- ข้อจำกัดในการ Tokenize ภาษาไทย เนื่องจากในภาษาไทยยังมีปัญหาในการตัดคำ เช่น “ร์” และ อื่น ๆ ที่มีการปรากฏขึ้นในหัวข้อ ทำให้เกิดการระบุกวนต่อการตีความได้
- การวัดผลเปรียบเทียบทำได้ยาก เนื่องจากแบบจำลองและปัญหาที่ต้องการแก้คือ Topic Modeling ซึ่งเป็นแบบจำลองประเภท Unsupervised การวัดค่าที่ดีที่สุดนอกจาก Topic Coherences คือ การตีความโดยมนุษย์ซึ่งทำได้ยาก ทำให้การวัดผลประสิทธิภาพการค้นหาหัวข้อของแบบจำลองทำได้ในภายใต้ข้อจำกัดเท่านั้น

## បរចណានុករម

- [1] “**Social Listening គីអ៊ូខាងវិរោះ**”, Accessed: Oct. 02, 2022. [Online]. Available:  
<https://wisesight.com/news/social-listening/>
- [2] “**Social Listening គីអ៊ូខាងវិរោះ និងធានាថ្មីនូវការងារដែលស្វែងរកនូវបន្ទីរិភីក**,” STEPS Academy.  
<https://stepstraining.co/analytics/what-is-social-listening> (accessed Nov. 04, 2022).
- [3] Twitter, “**Twitter API.**” <https://developer.twitter.com/en/docs/twitter-api>  
 (accessed Nov. 01, 2022).
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “**Latent dirichlet allocation,**”  
*Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [5] J. Yin and J. Wang, “**A dirichlet multinomial mixture model-based approach for short text clustering,**” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 233–242.
- [6] R. Zhao and V. Y. F. Tan, “**Online nonnegative matrix factorization with outliers,**” *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 555–570, 2016.
- [7] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, “**Short text topic modeling techniques, applications, and performance: a survey,**” *IEEE Trans Knowl Data Eng*, 2020.
- [8] X. Cheng, X. Yan, Y. Lan, and J. Guo, “**Btm: Topic modeling over short texts,**” *IEEE Trans Knowl Data Eng*, vol. 26, no. 12, pp. 2928–2941, 2014.
- [9] R. Alghamdi and K. Alfalqi, “**A survey of topic modeling in text mining,**” *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, vol. 6, no. 1, 2015.
- [10] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsoutsouliklis, “**Discovering geographical topics in the twitter stream,**” in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 769–778.
- [11] E. Loper and S. Bird, “**Nltk: The natural language toolkit,**”  
*arXiv preprint cs/0205028*, 2002.

## បរចាំនាច្បារ (៩)

- [12] M. Röder, A. Both, and A. Hinneburg, “**Exploring the space of topic coherence measures,**” in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.
- [13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,**” *CoRR*, vol. abs/1810.04805, 2018, [Online]. Available: <http://arxiv.org/abs/1810.04805>
- [14] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai, “**PyThaiNLP: Thai Natural Language Processing in Python.**” Zenodo, Jun. 2016. doi: 10.5281/zenodo.3519354.
- [15] N. Reimers and I. Gurevych, “**Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,**” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Nov. 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084>
- [16] R. Rehurek and P. Sojka, “**Gensim—python framework for vector space modelling,**” *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.
- [17] F. Pedregosa *et al.*, “**Scikit-learn: Machine Learning in Python,**” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [18] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, “**Mining geographic knowledge using location aware topic model,**” in *Proceedings of the 4th ACM workshop on Geographical information retrieval*, 2007, pp. 65–70.
- [19] X. Wang and A. McCallum, “**Topics over time: a non-markov continuous-time model of topical trends,**” in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 424–433.

## បររណានុករម (៩)

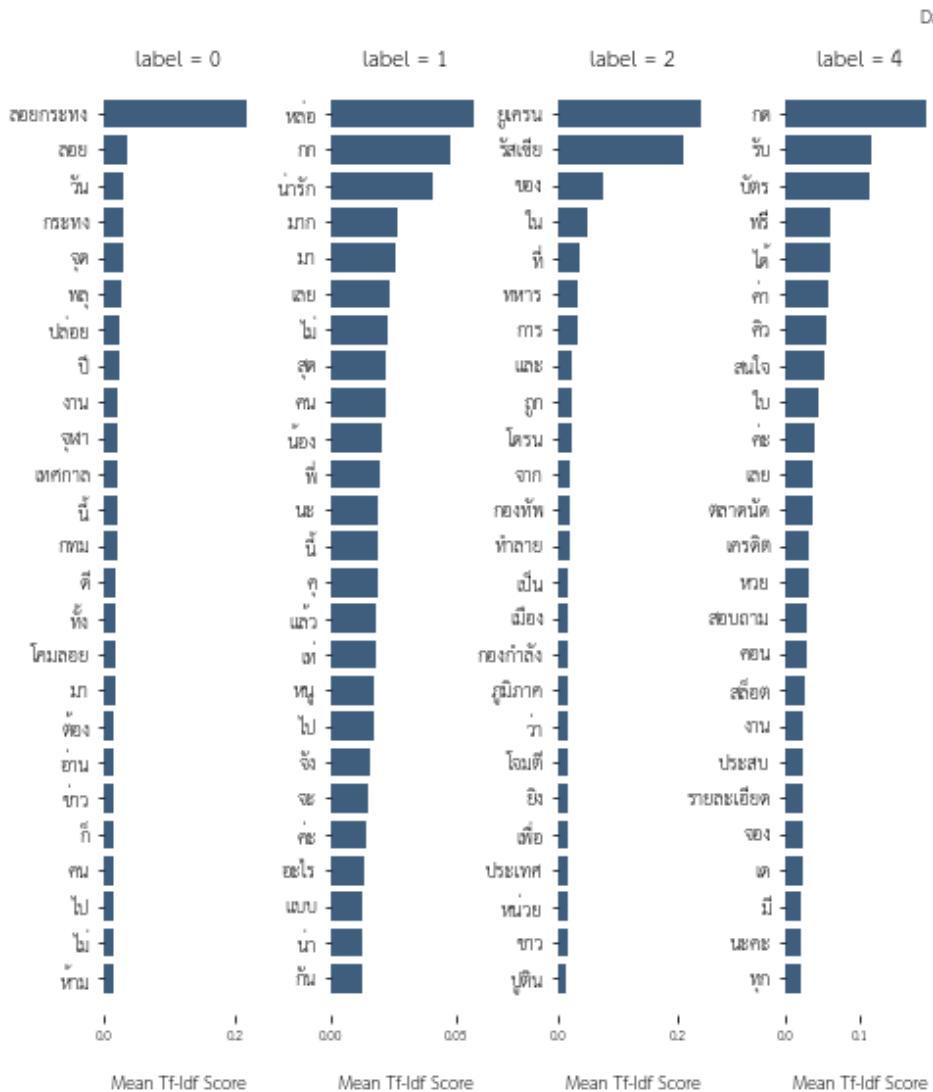
- [20] C. Wang, D. Blei, and D. Heckerman, “**Continuous time dynamic topic models,**” *arXiv preprint arXiv:1206.3298*, 2012.
- [21] M. Grootendorst, “**BERTopic: Neural topic modeling with a class-based TF-IDF procedure,**” *arXiv preprint arXiv:2203.05794*, 2022.
- [22] D. M. Blei and J. D. Lafferty, “**Dynamic topic models,**” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 113–120.
- [23] X. Zhang, J. J. Zhao, and Y. LeCun, “**Character-level Convolutional Networks for Text Classification,**” in *NIPS*, 2015.
- [24] J. M. Banda *et al.*, “**A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research—An International Collaboration,**” *Epidemiologia*, vol. 2, no. 3, pp. 315–324, 2021, doi: 10.3390/epidemiologia2030024.
- [25] R. Tekumalla and J. M. Banda, “**Social media mining toolkit (SMMT),**” *Genomics Inform*, vol. 18, no. 2, 2020.
- [26] “**Thai-Sentence-Vector-Benchmark,**” Jul. 09, 2022. <https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark> (accessed Oct. 20, 2022).

## ภาคผนวก

### ภาคผนวก ก.

ผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธีที่แตกต่างกัน

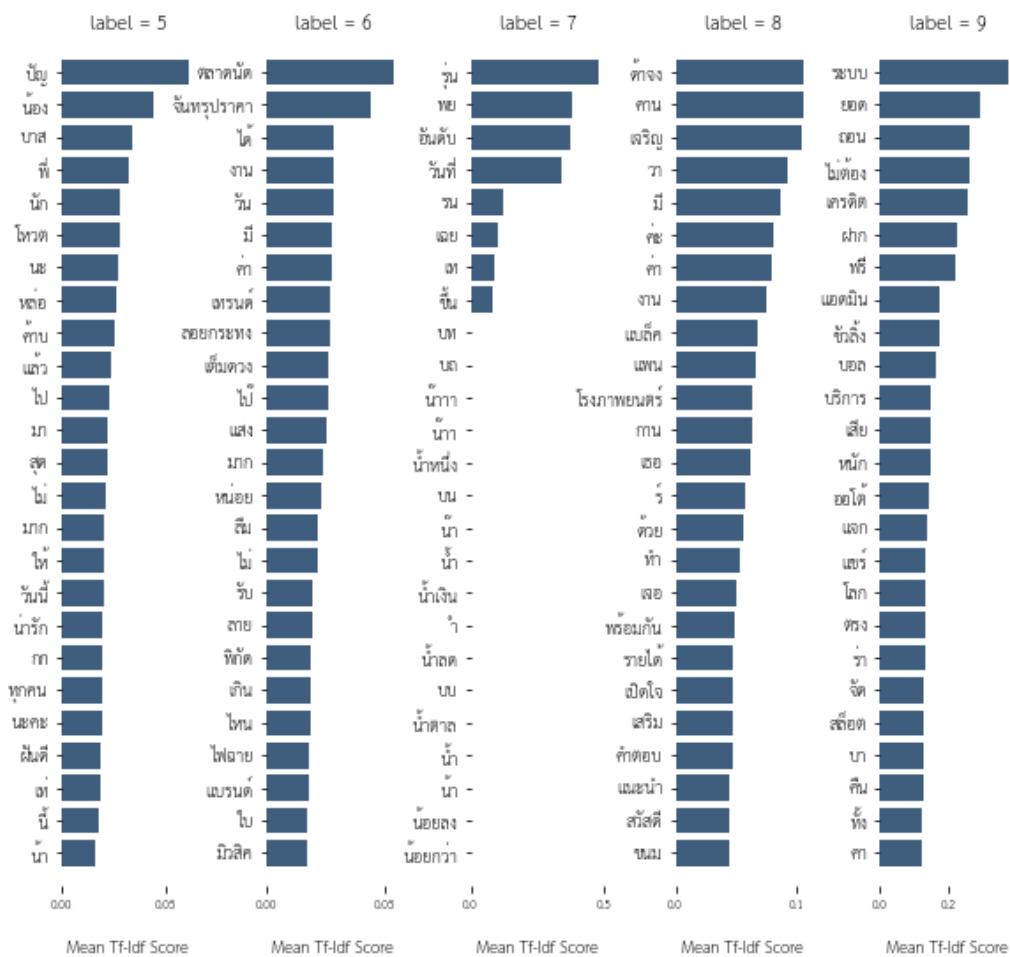
ผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัย  
ประกอบ



รูปที่ ก.1 ผลลัพธ์การค้นหาหัวข้อที่หนึ่งจากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

y 1

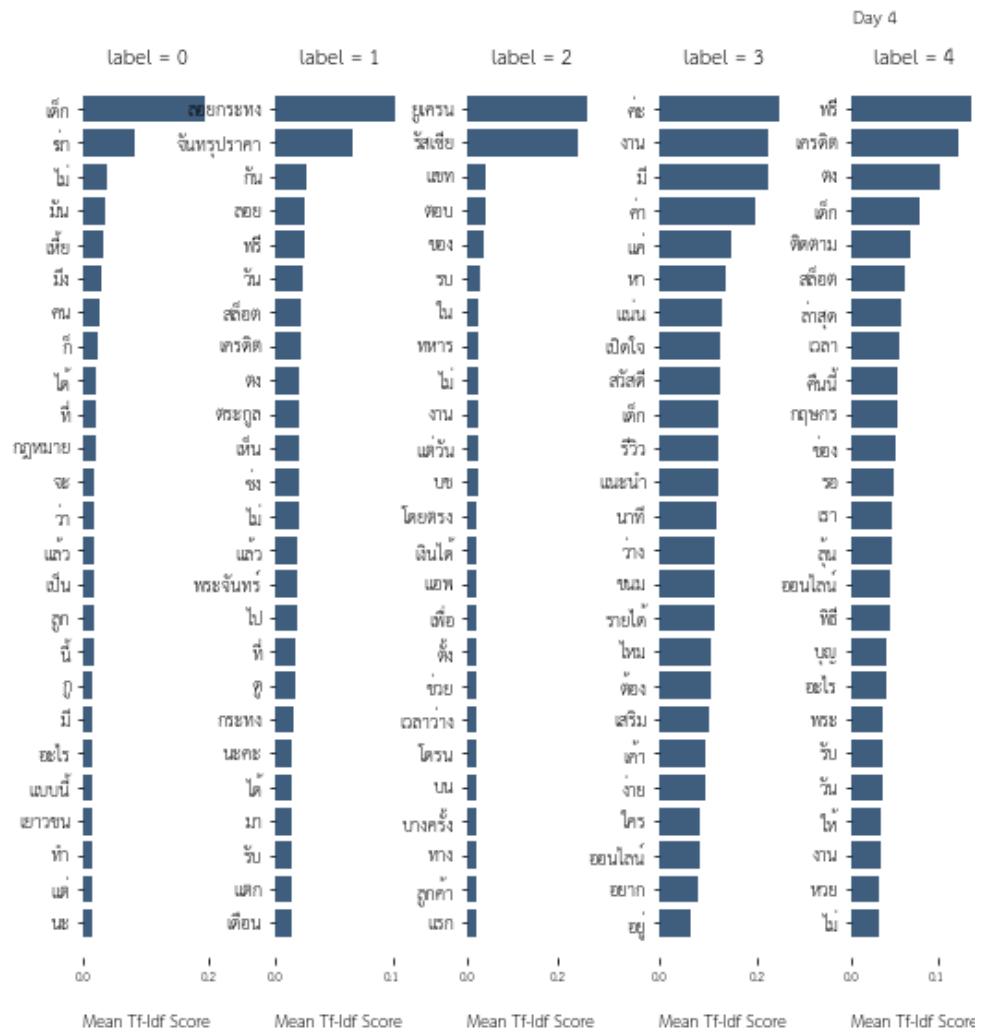


รูปที่ ก.2 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

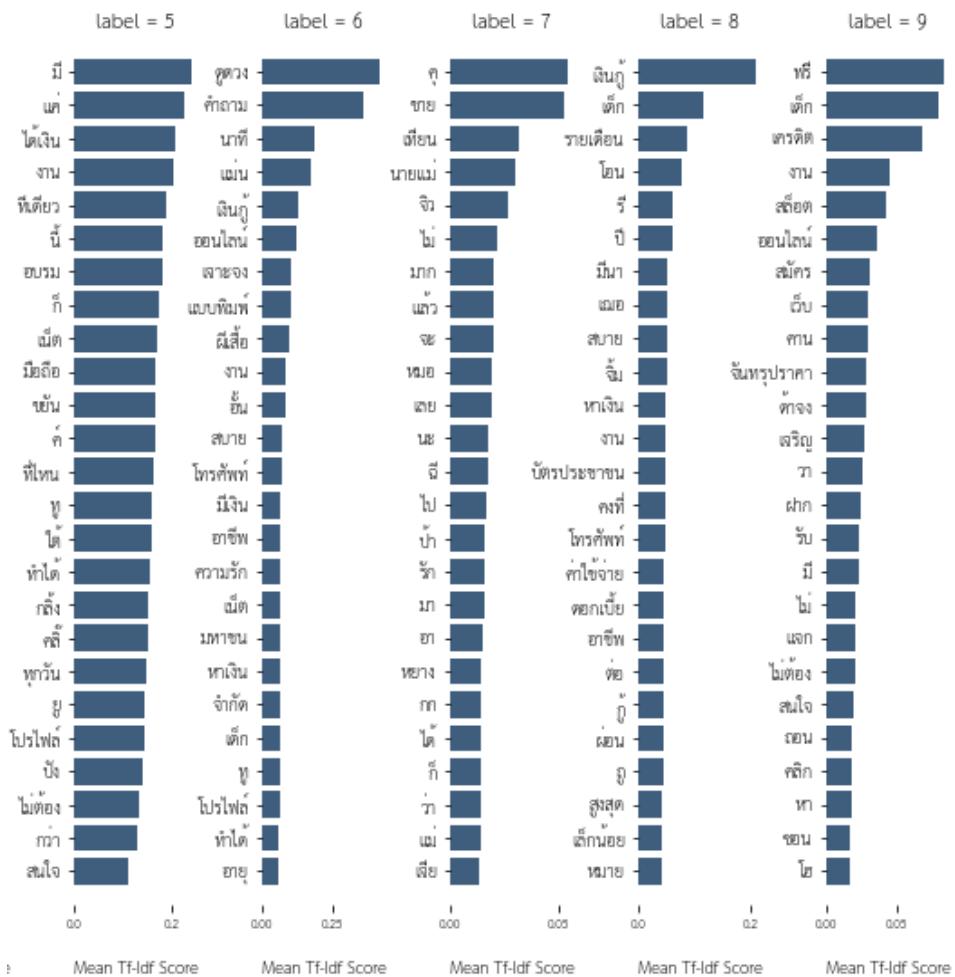
จากรูปที่ ก.1 และ ก.2 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่หนึ่ง (วันที่ 5

พฤษภาคม พ.ศ. 2565) ลักษณะเบื้องต้นของหัวข้อที่พบคือ หัวข้อที่ 0 อาจสามารถอนุมานได้ว่าเป็นหัวข้อของวันลองยกกระทง หัวข้อที่ 6 ที่อาจหมายถึงเงินทรุปร้าภา เนื่องจากมีคำเป็นส่วนประกอบของหัวข้อเป็นอันดับต้น ๆ



รูปที่ ก.3 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window

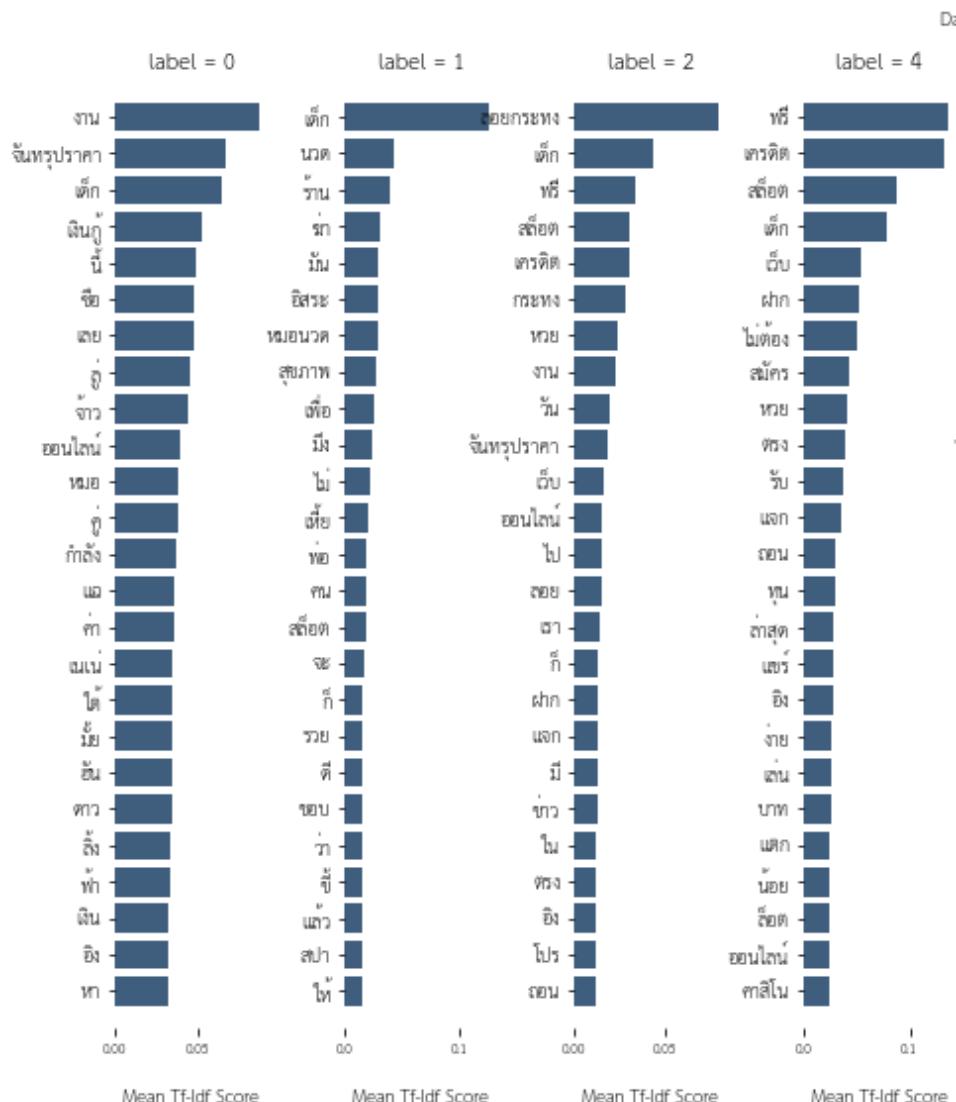
โดยไม่ใช้เวลาปืนปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.4 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

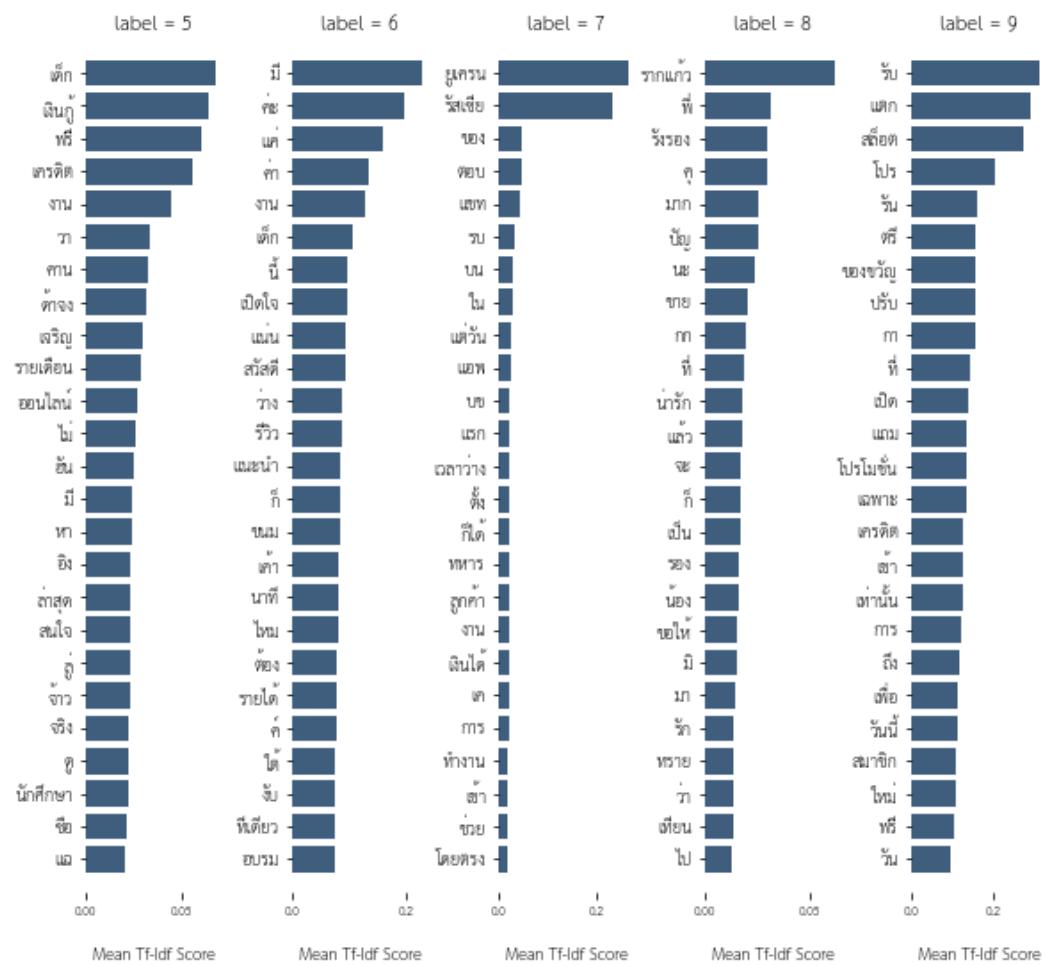
จากรูปที่ ก.3 และ ก.4 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สี่ (วันที่ 8 พฤศจิกายน พ.ศ. 2565) มีการกล่าวถึงหัวข้อเกี่ยวกับจันทรุปราคາอีกครั้งในหัวข้อที่ 1 ส่วนหัวข้ออื่น ๆ ประกอบด้วยเนื้อหาเกี่ยวกับการพนัน ทำงาน และครอบครัวเป็นจำนวนมาก เช่นเดียวกับวันที่สาม (วันที่ 7 พฤศจิกายน พ.ศ. 2565)



รูปที่ ก.5 ผลลัพธ์การค้นหาหัวข้อวันที่ ห้าจากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

by 5

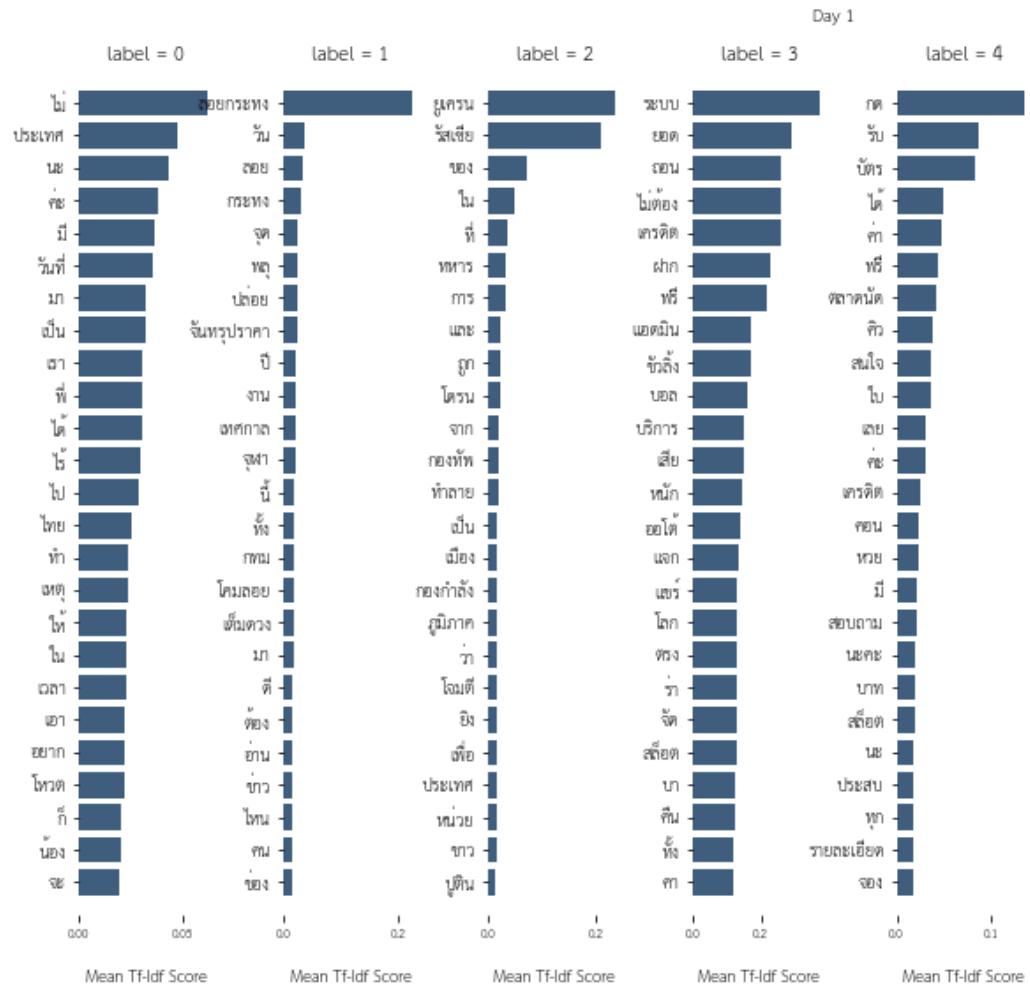


รูปที่ ก.6 ผลลัพธ์การค้นหาหัวข้อวันที่หลังจากการทดลองแบบ Sliding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

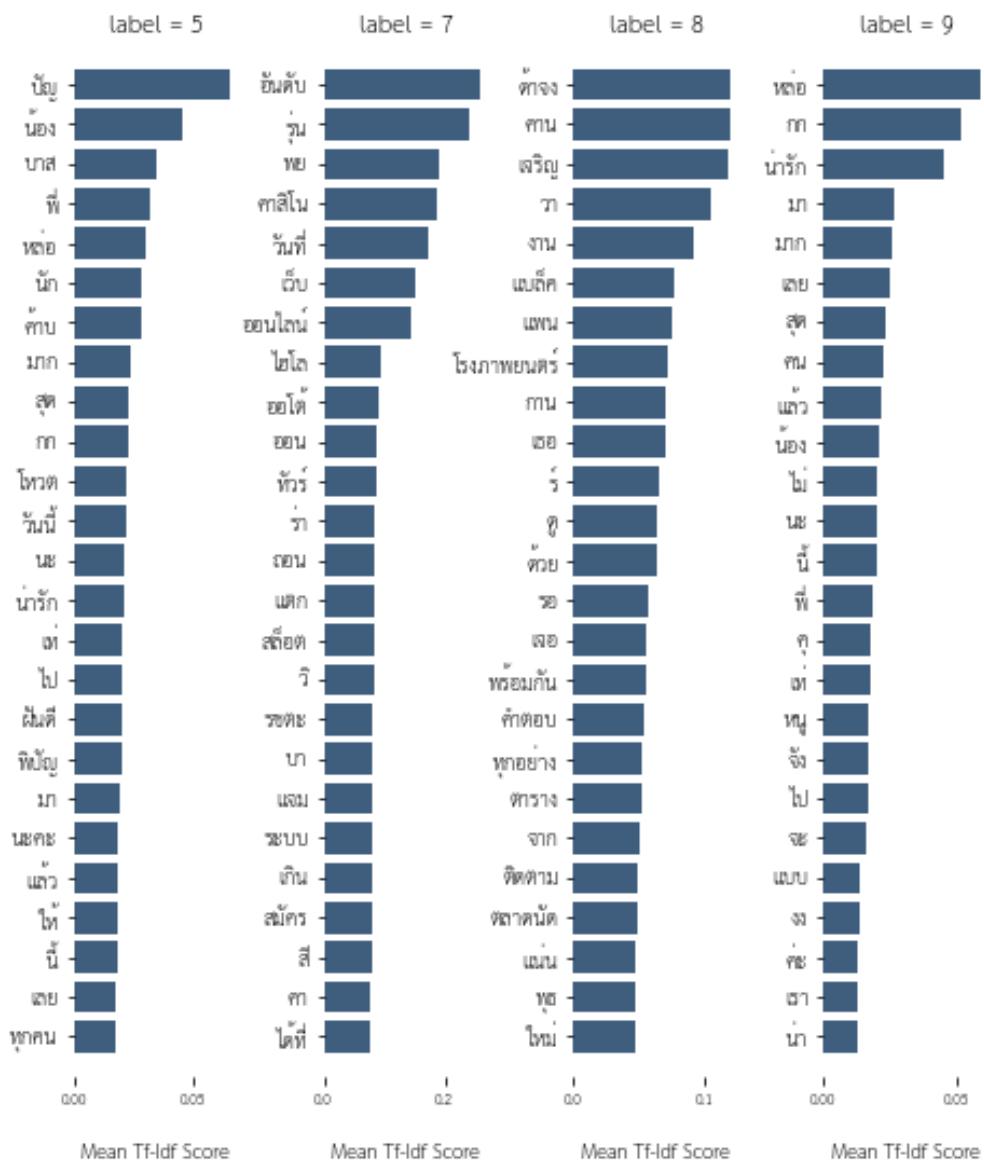
จากรูปที่ ก.๕ และ ก.๖ เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สี่ (วันที่ 9 พฤษภาคม พ.ศ. 2565) มีหัวข้อจันทรุปราคากาย่าเช่นเดิมแต่ต่างจากวันที่สี่ โดยที่หัวข้อดังกล่าวประกอบด้วย คำที่เกี่ยวกับเงินกู้ออนไลน์ซึ่งอาจดึงความได้มาจากการโปรโมทเงินกู้ออนไลน์ผ่าน Hashtag จันทรุปราคากายา และมีการพับหัวข้อใหม่อีกหนึ่งหัวข้อที่ 8 ซึ่งเกิดขึ้นมาอาจจะดึงความได้มาจากการพูดถึงผลกระทบเรื่องราภัยแก้ว

ผลลัพธ์ที่ได้จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ



รูปที่ ก.7 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window

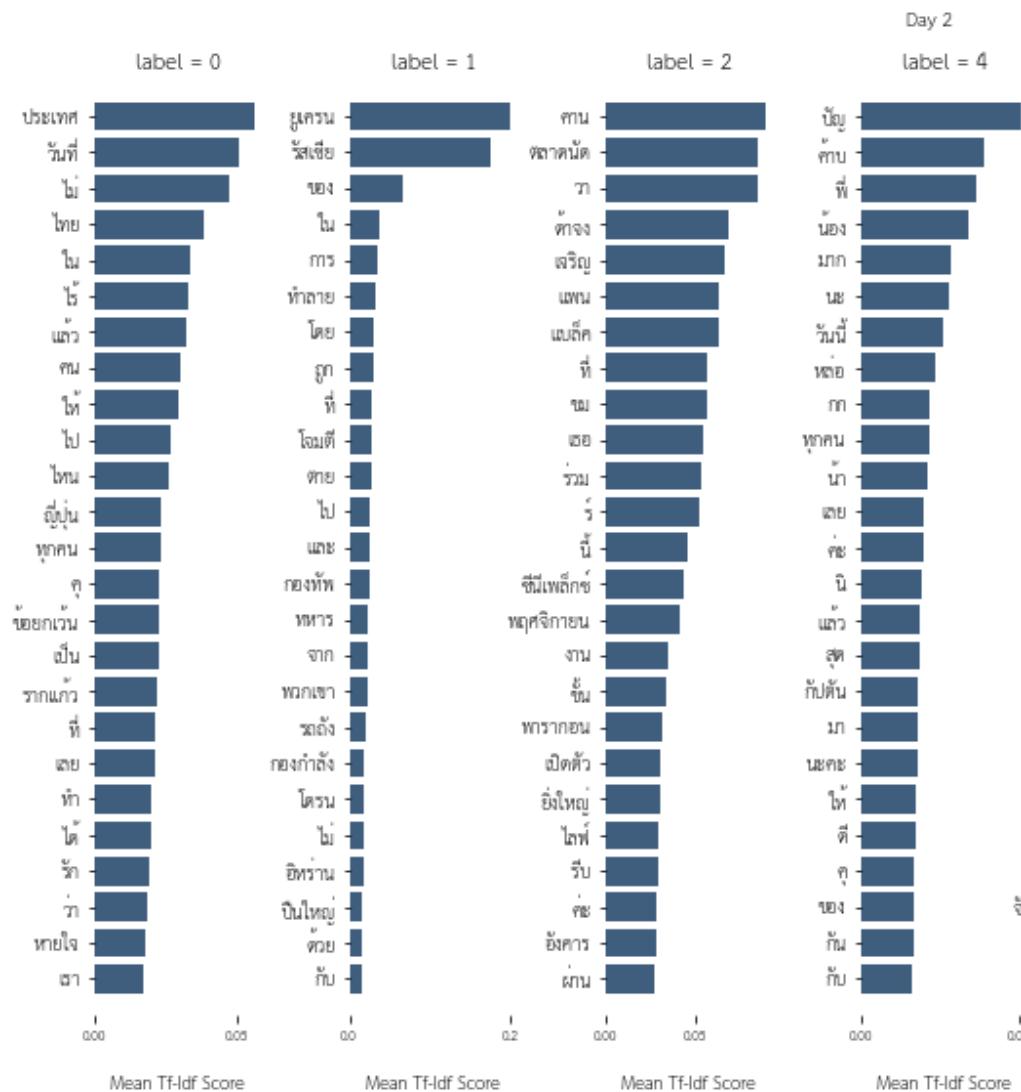
## โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.8 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window

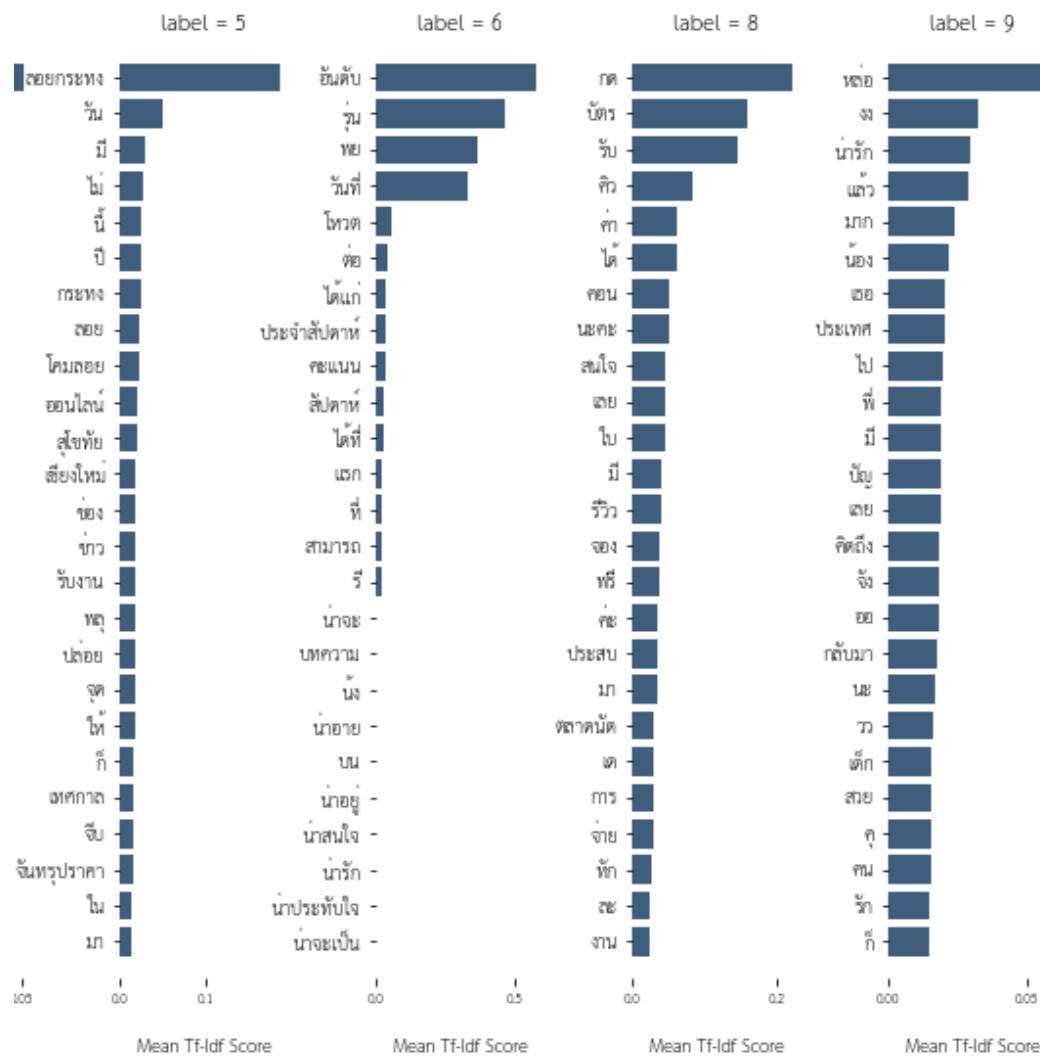
โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.7 และ ก.8 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่หนึ่ง (วันที่ 5 พฤศจิกายน 2565) พบรหัสข้อที่เห็นได้ชัด เช่น เกี่ยวกับวันลอยกระทงในหัวข้อที่ 1 เกี่ยวกับสกุลเงินกับรัฐเชียในหัวข้อที่ 2 เกี่ยวกับภาษณตร์ในหัวข้อที่ 8 และหัวข้อที่ 5 ที่มีการกล่าวถึงชื่อสมาชิกวงดนตรี BNK48 และหัวข้ออื่น ๆ เกี่ยวกับการพนัน และเงินกู้ออนไลน์ซึ่งตีความได้ยาก



รูปที่ ก.9 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window

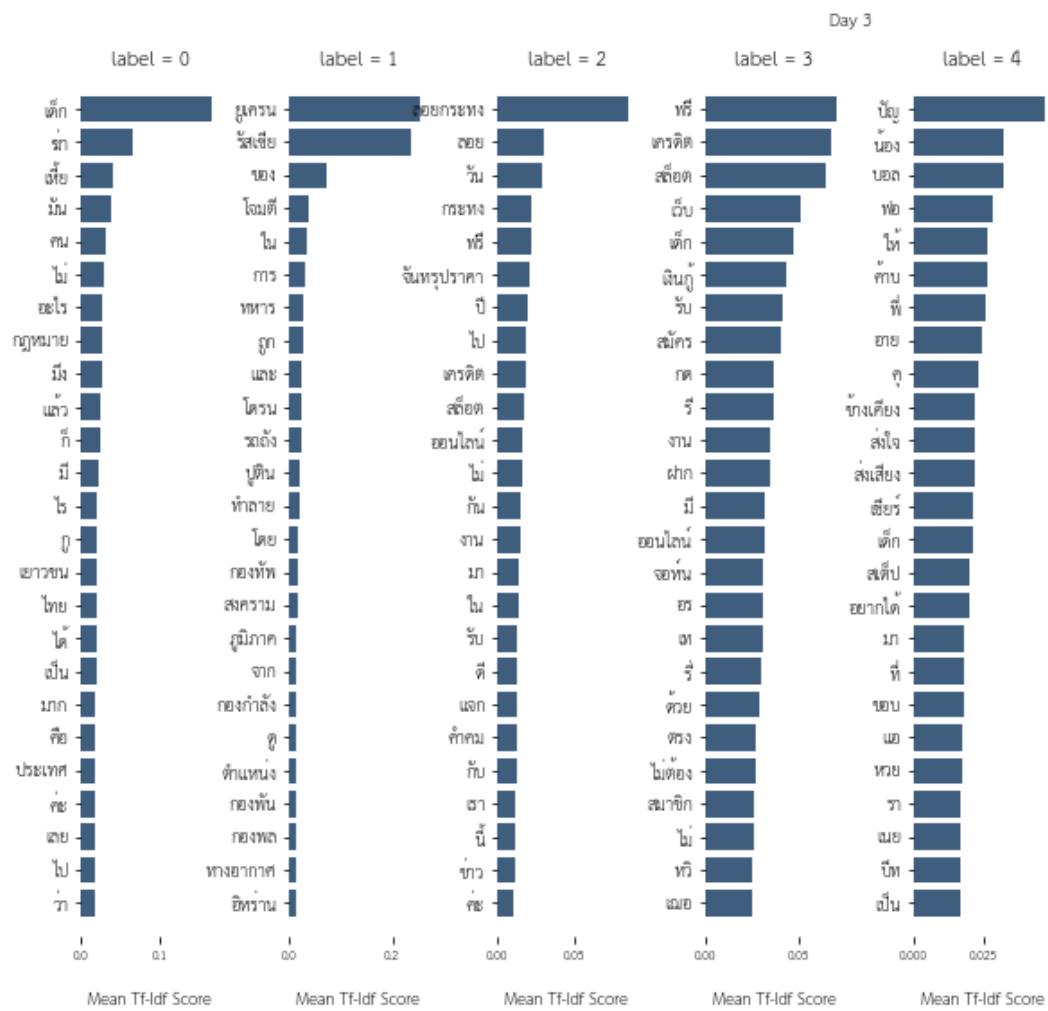
โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.10 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window

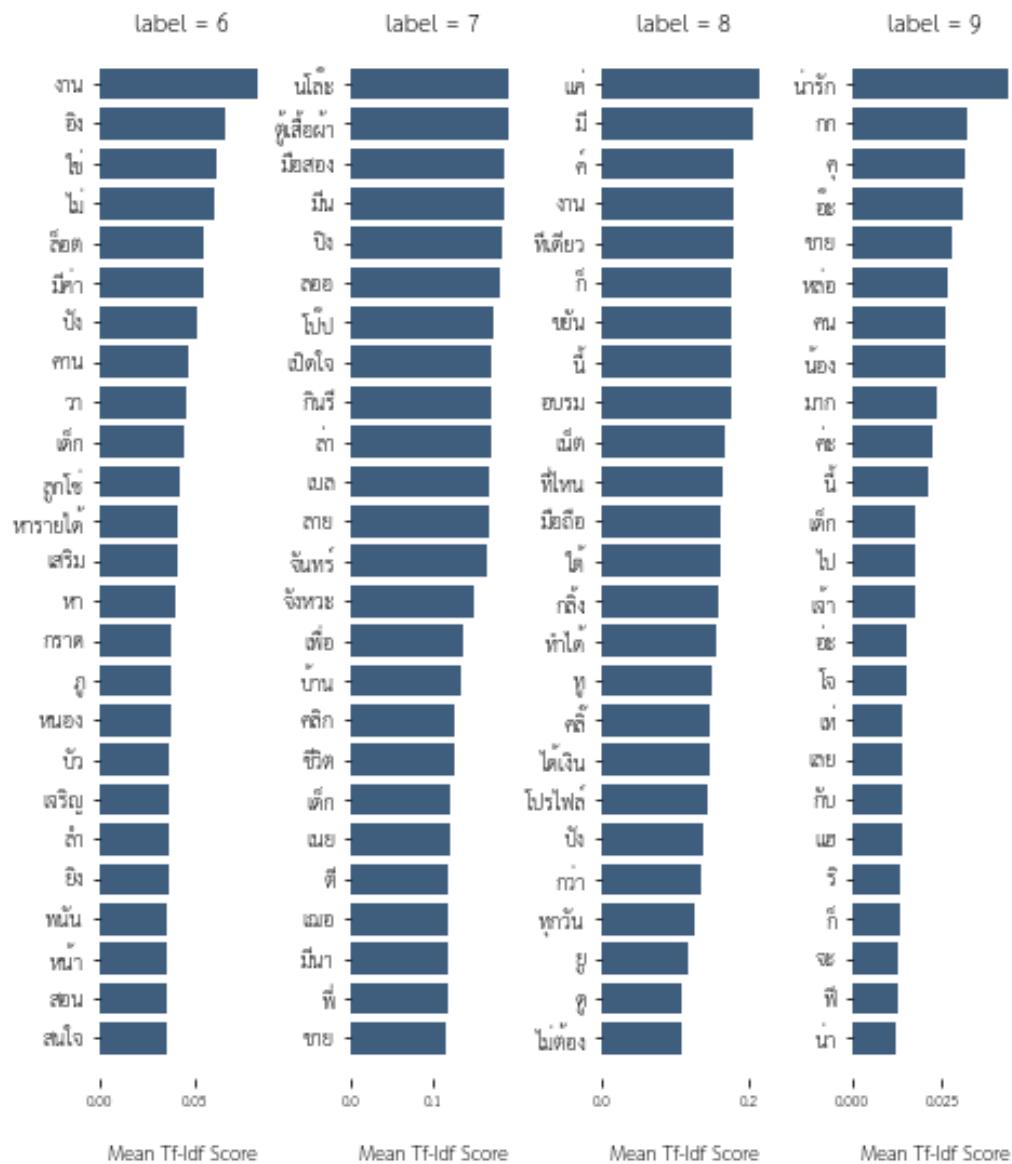
โดยไม่ใช่วลามเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.9 และ ก.10 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สอง (วันที่ 6 พฤศจิกายน 2565) พบหัวข้อที่เห็นได้ชัดคล้ายกับวันก่อนหน้า เช่น หัวข้อเกี่ยวกับวันลอยกระทง และหัวข้อเกี่ยวกับสงกรานต์ว่างญูครนกับรัสรสเซีย อีกทั้งยังมีการพบหัวข้อใหม่อีก หัวข้อที่ 6 ที่ไม่สามารถตีความได้



รูปที่ ก.11 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window

โดยไม่ใช้เวลาปืนปัจจัยประกอบ (ส่วนที่ 1)

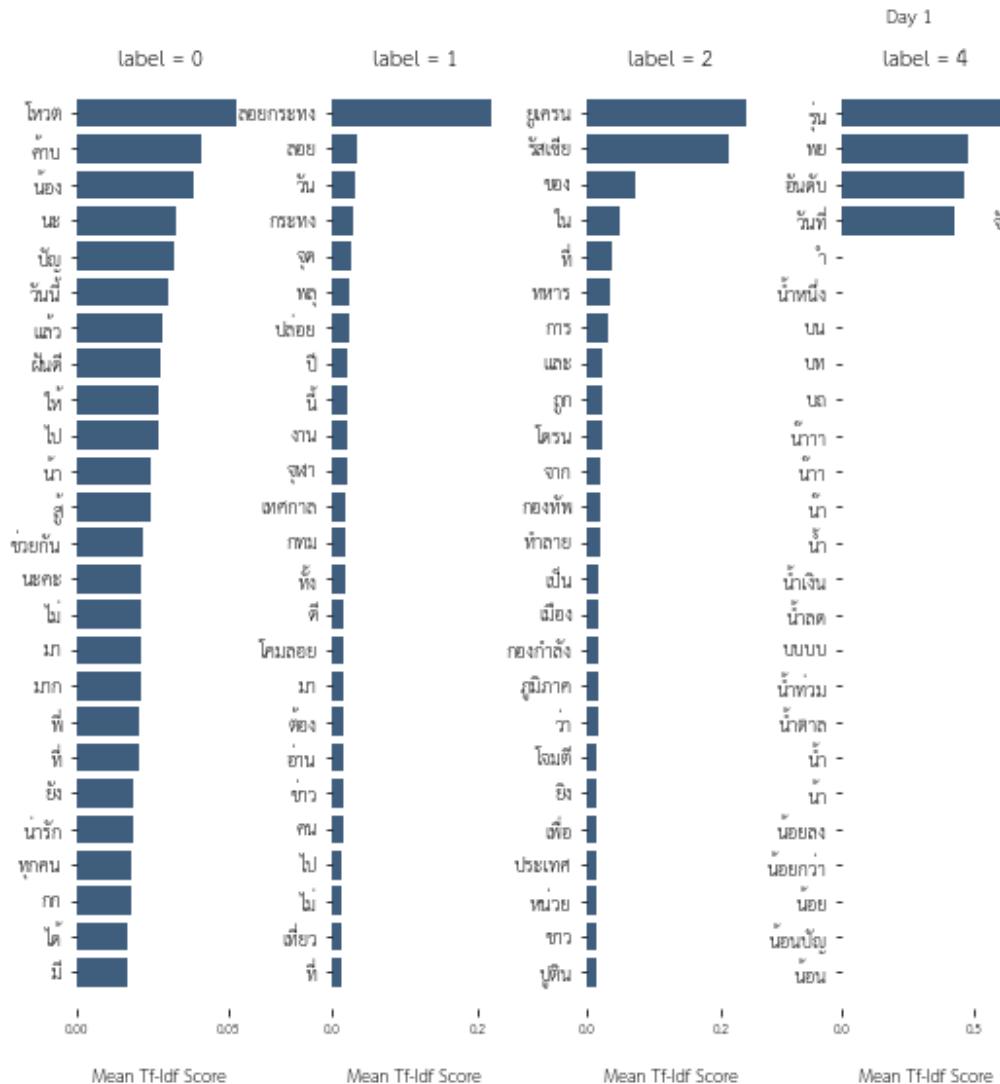


รูปที่ ก.12 ผลลัพธ์การค้นหาหัวข้อของวันที่สามจากการทดลองแบบ Expanding Window

โดยไม่ใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

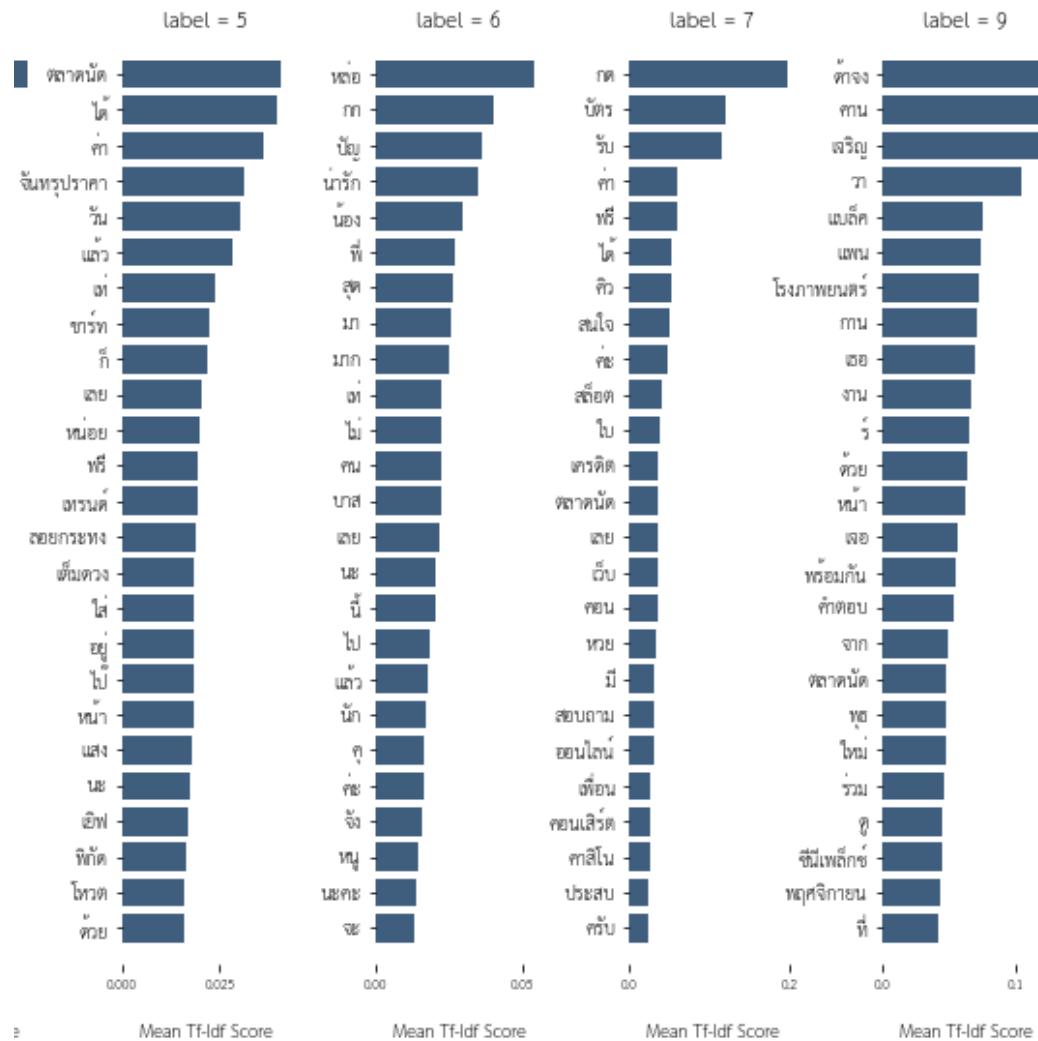
จากรูปที่ ก.11 และ ก.12 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สาม (วันที่ 7 พฤษภาคม 2565) พบหัวข้อที่มีลักษณะคล้ายกับวันก่อนหน้านี้ เช่น หัวข้อเกี่ยวกับสังคม ระหว่างยุครุนกับรัฐเซีย หัวข้อเกี่ยวกับวันลอยกระทง และหัวข้อเกี่ยวกับวง BNK48 อีกทั้งยัง มีการพบหัวข้อใหม่อีกหัวข้อที่ 0 ซึ่งมีเนื้อหารุนแรงอาจเกี่ยวกับข่าวการมาตรฐานเด็ก รวมถึงหัวข้อ 7, 8 และ 9 ซึ่งมีลักษณะเนื้อหาที่ตีความได้ยาก

ผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ



รูปที่ ก.13 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window

## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

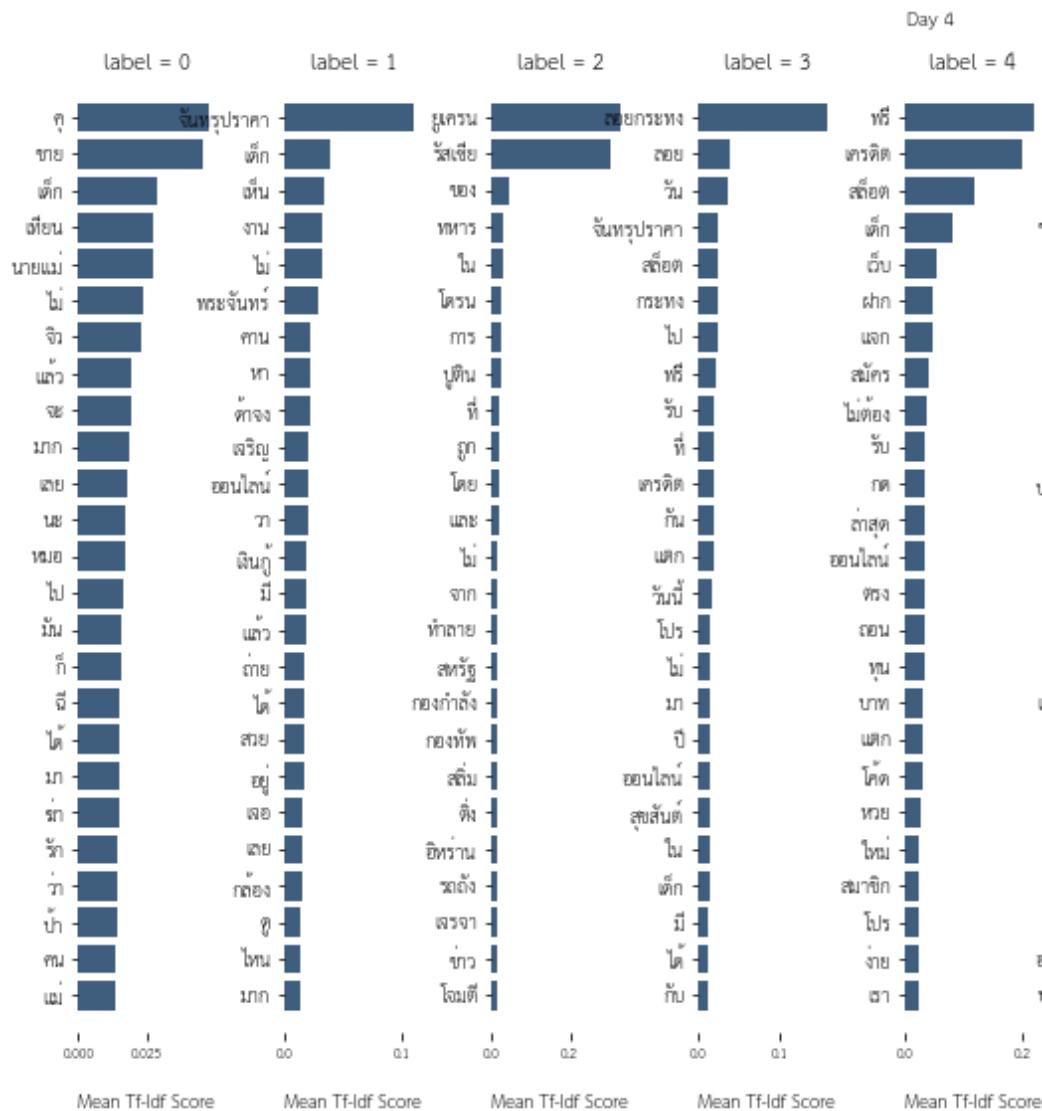


รูปที่ ก.14 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Sliding Window

## ໂຄຍໃໝ່ເວລາເປັນປັງຈີຍປະກອນ (ສ່ວນທີ 2)

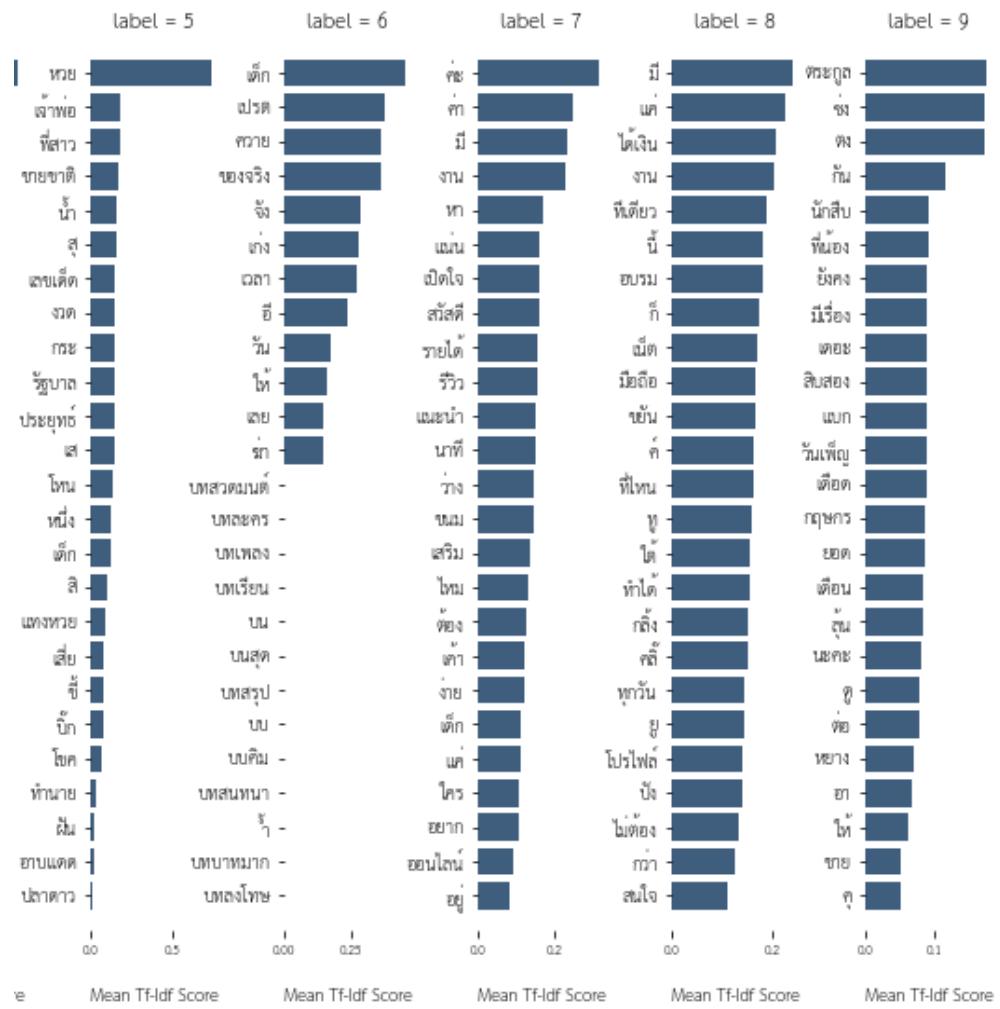
จากรูปที่ ก.13 และ ก.14 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่นึง (วันที่ 5 พฤษภาคม พ.ศ. 2565) ลักษณะเบื้องต้นของหัวข้อที่พบคือ คำที่เป็นส่วนประกอบของหัวข้อแสดงให้เห็นก่อรุ่มของเนื้อหาที่ชัดเจน เช่น หัวข้อที่ 0 อาจสามารถอนุมานได้ว่าเป็นหัวข้อที่เกี่ยวกับวง BNK48 ซึ่งมีคำที่เกี่ยวกับการ โหวตที่ในช่วงเวลาใดมีงานไว้สำหรับโหวต Center ของวงในเพลงใหม่ อีกทั้งยังมีการพูดถึง ปัญ หนึ่งในสมาชิกวงคนดัง BNK48 หัวข้อที่ 5 ที่อาจหมายถึงจันทรุปราดา เนื่องจากมีคำเป็นส่วนประกอบของหัวข้อเป็นอันดับต้น ๆ หรือแม้แต่หัวข้อที่ 9 ที่อาจอนุมานได้ว่าเป็นหัวข้อที่เกี่ยวกับภาพยนตร์ Black Panther เนื่องจากมีคำเป็น

ส่วนประกอบของหัวข้ออย่าง โรงภาพยนตร์ วากานด้วย (คำที่กล่าวถึงเมืองที่ Black Panther อาศัยอยู่ในเรื่อง) โดยวันที่หนึ่ง หัวข้อที่ทำการหายังคงมีไม่ครบ 10 หัวข้อ ซึ่งอาจเกิดจากเนื้อหาของช่วงวันดังกล่าวไม่สอดคล้องกับข้อมูลของวันที่ใช้ฝึกสอนแบบจำลอง



รูปที่ ก.15 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window

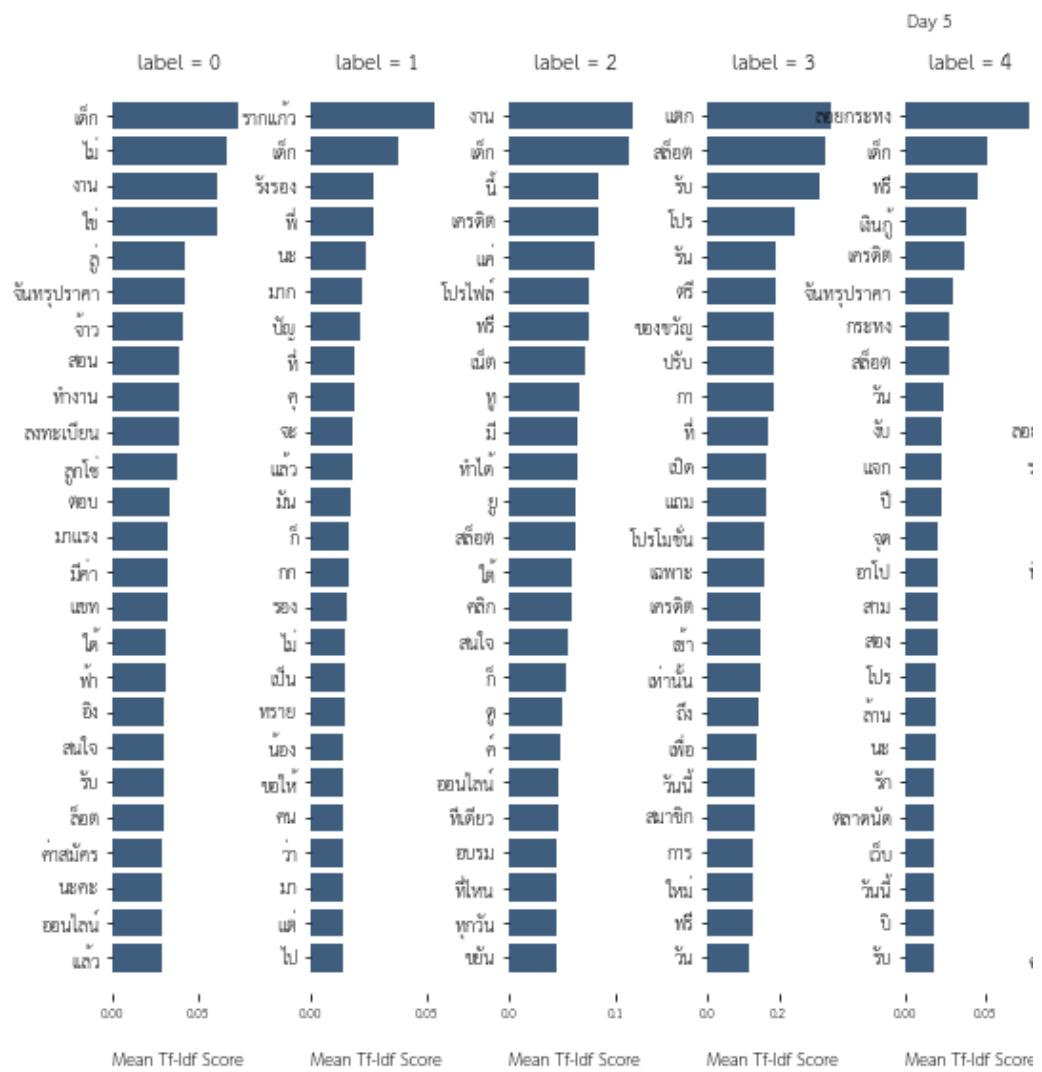
โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.16 ผลลัพธ์การค้นหาหัวข้อวันที่สี่จากการทดลองแบบ Sliding Window

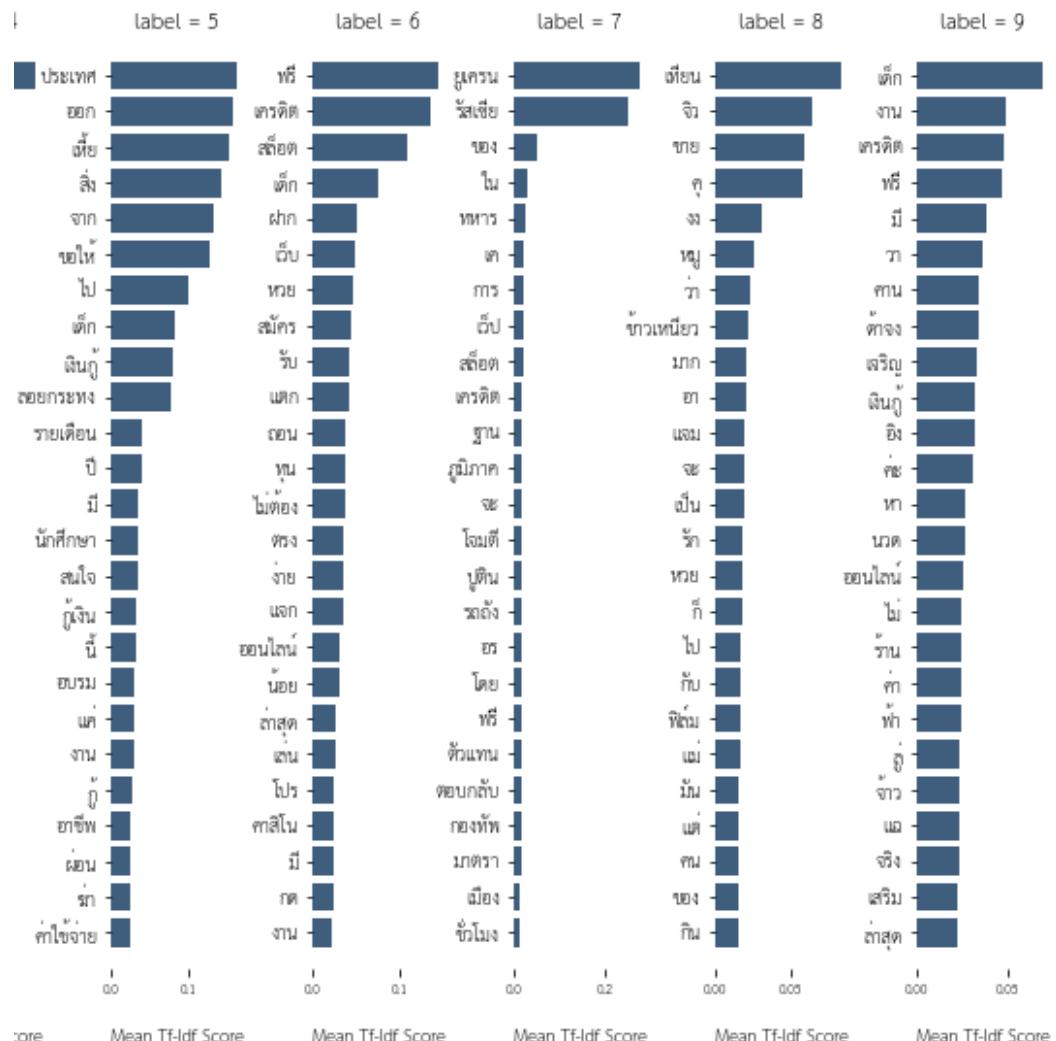
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.15 และ ก.16 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สี่ (วันที่ 8 พฤษภาคม พ.ศ. 2565) พบว่ายังคงมีการพูดถึงหัวข้อจันทรุปราสา (หัวข้อที่ 1) สองครั้ง ระหว่างผู้คนกับรัฐเชีย (หัวข้อที่ 2) และประเพณีลอยกระทง (หัวข้อที่ 3) เหมือนกับวันที่ผ่านมา อีกทั้งยังมีบางหัวข้อที่มีการกล่าวถึงในวันที่สาม แต่ไม่ถูกกล่าวถึงในวันที่สี่อย่างเรื่องของ โรมนัลโดกับวิลล่า และเช่นเดียวกับคดีที่เด็กอายุ 18 ปีมาตกรรมเด็กอายุ 13 ปี ที่ผลลัพธ์ของแบบจำลองไม่ได้แสดงผลออกมากซักเท่าไร เหมือนกับผลลัพธ์ของวันที่สาม โดยสำหรับวันที่สี่ หัวข้อที่ทำการหมายครบทั้ง 10 หัวข้อ ต่างจากวันก่อนหน้า



รูปที่ ก.17 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window

โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)

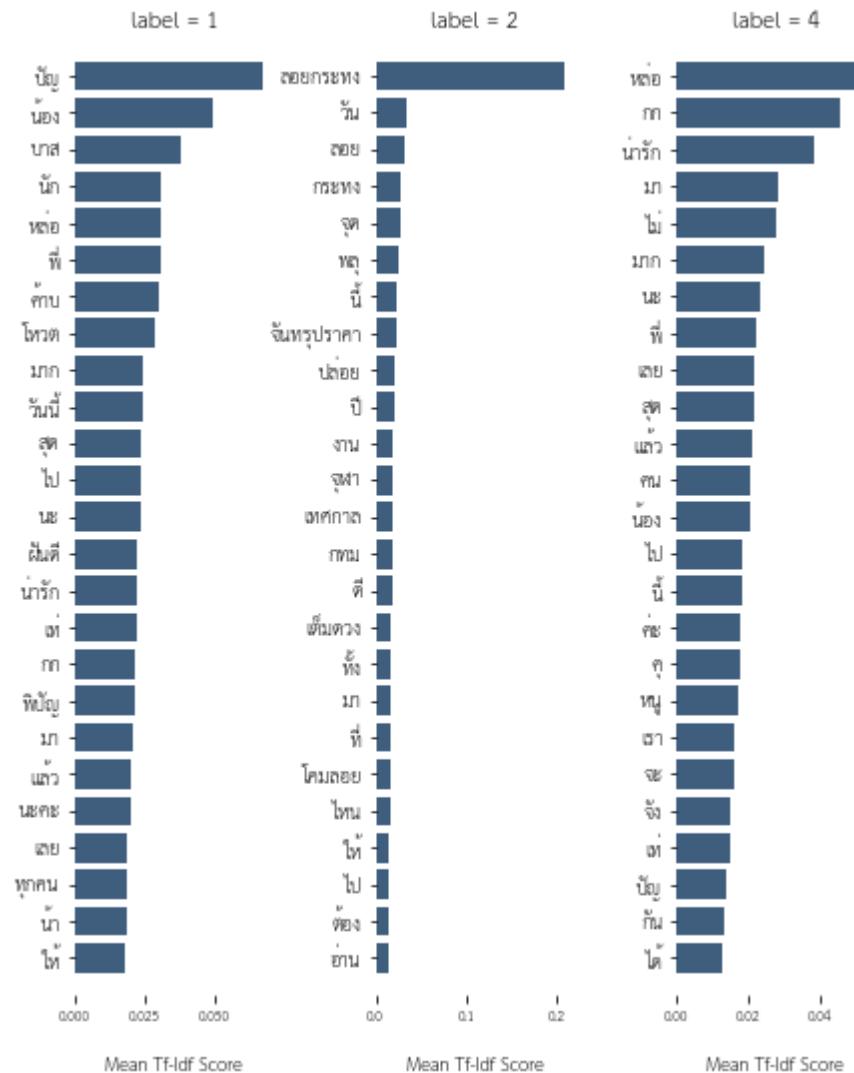


รูปที่ ก.18 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Sliding Window

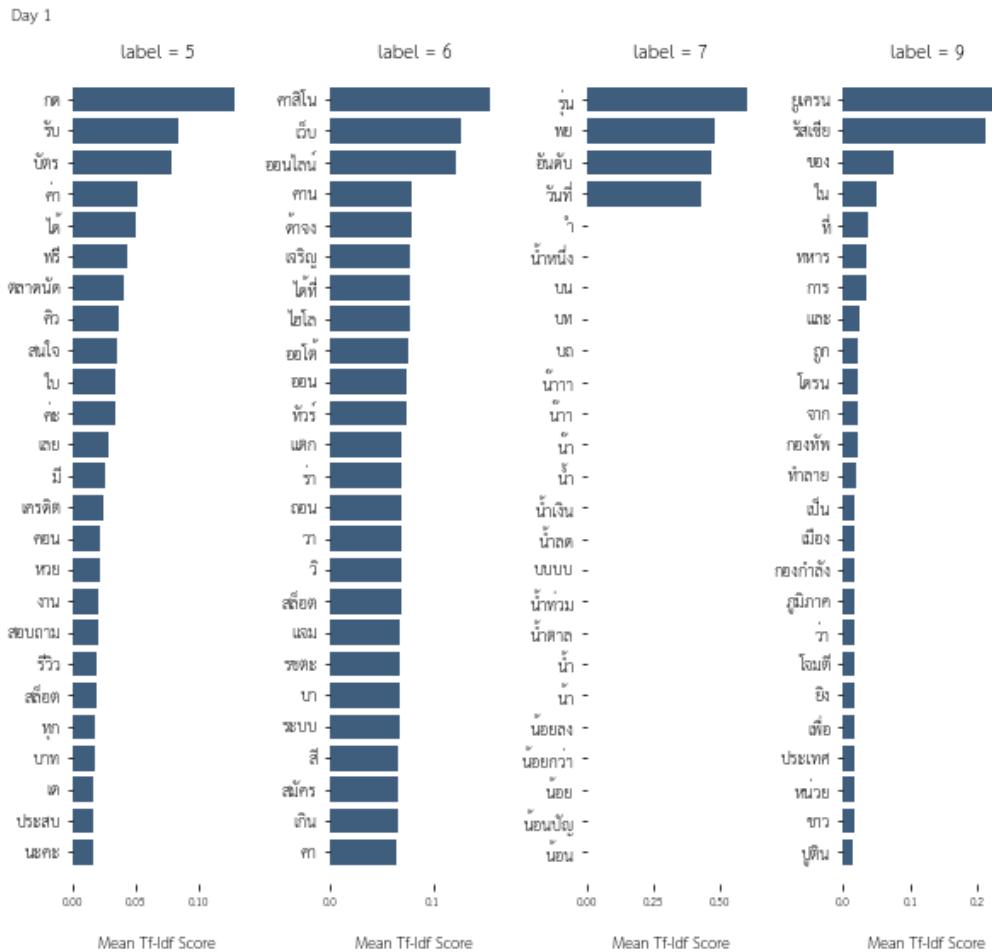
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.17 และ ก.18 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่ห้า (วันที่ 9 พฤษภาคม พ.ศ. 2565) พบว่ามีการพูดถึงหัวข้อจันทรุปราคा (หัวข้อที่ 0) ประเพณีลอยกระทง (หัวข้อที่ 4) และส่งความหว่างยุเครนกับรัสเซีย (หัวข้อที่ 7) เมื่อฉันกับวันที่ผ่านมาอีกทั้งยังพบการกล่าวถึงเรื่องใหม่อีกหลายเรื่องราวด้วยกัน เช่น อยู่ในช่วงเดียวกับวันที่มีการนายตัวเรื่องนี้ จากหัวข้อที่ 1 โดยสำหรับวันที่ห้า หัวข้อที่ทำการหมายครบทั้ง 10 หัวข้อ เช่นเดียวกับวันที่สี่ (วันที่ 8 พฤษภาคม พ.ศ. 2565)

ผลลัพธ์หัวข้อที่ได้จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ



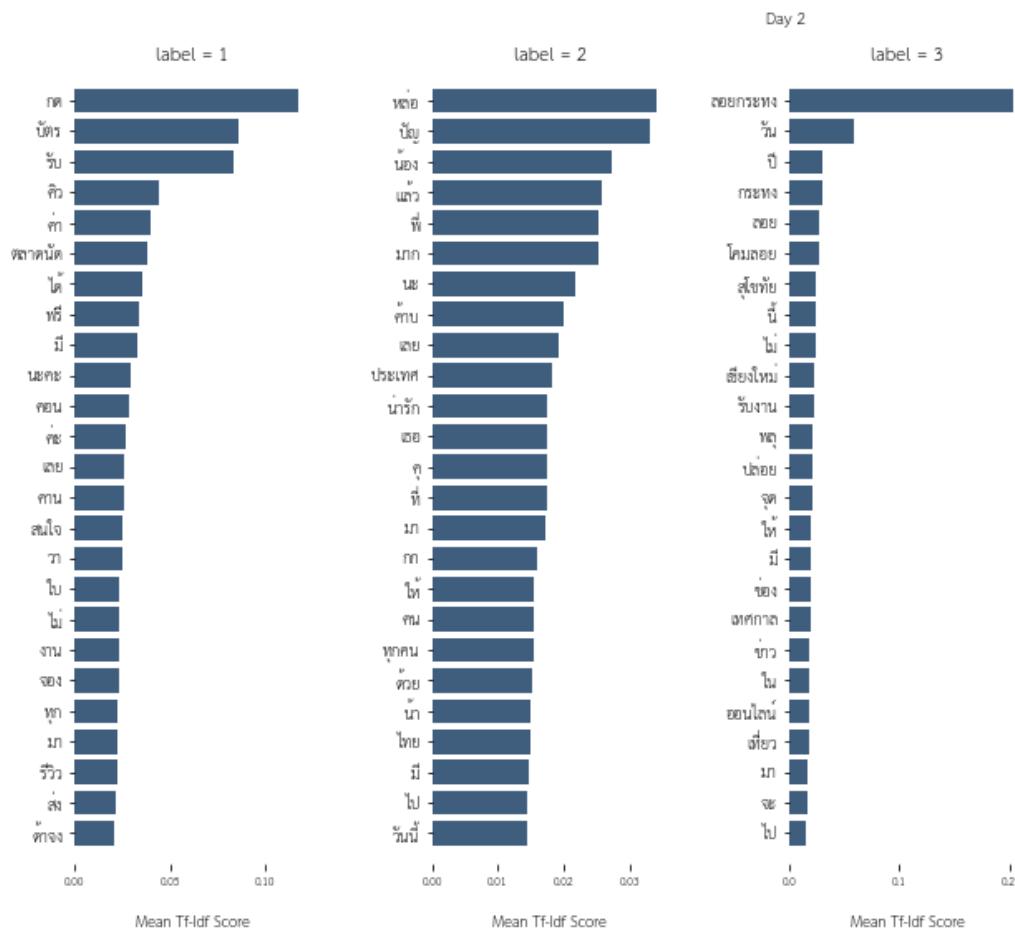
รูปที่ ก.19 ผลลัพธ์การค้นหาหัวข้อวันที่หนึ่งจากการทดลองแบบ Expanding Window โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.20 ผลลัพธ์การค้นหาหัวข้อวันที่ห้าจากการทดลองแบบ Expanding Window

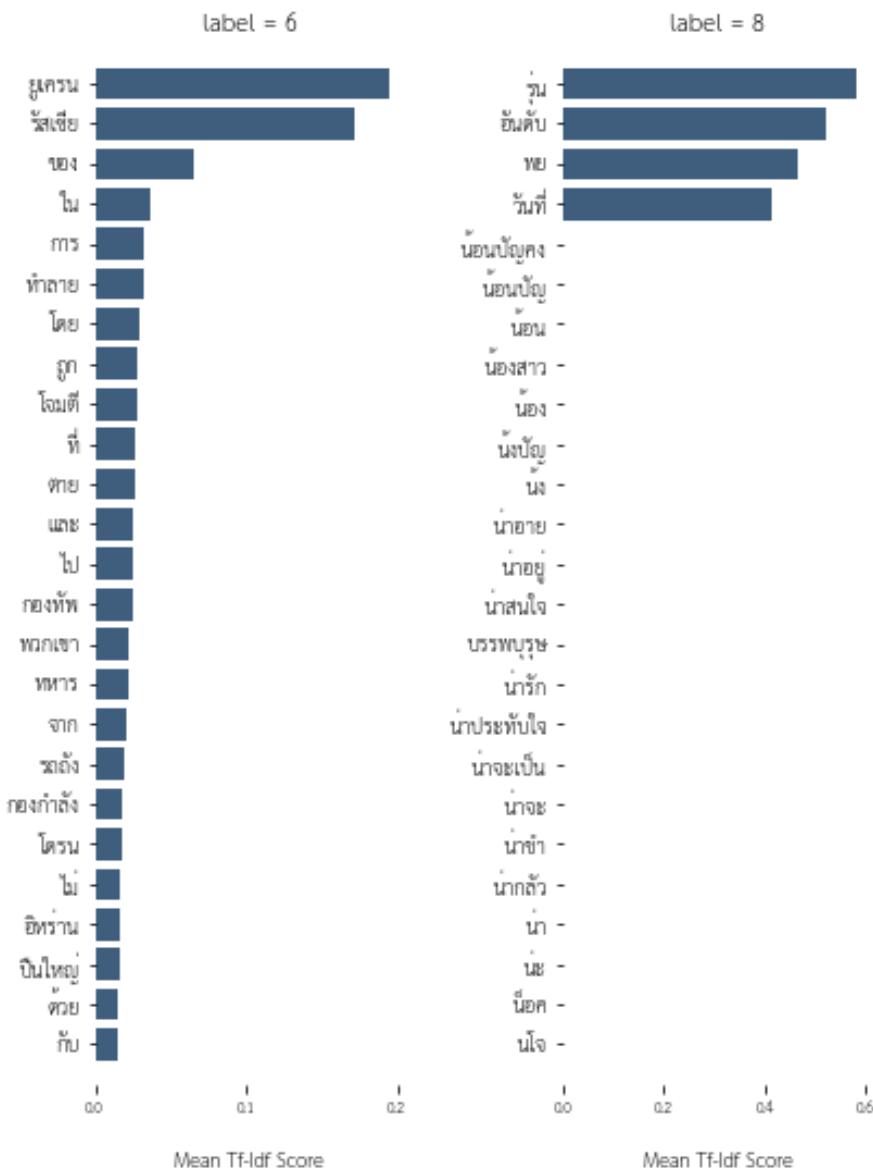
โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.19 และ ก.20 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่หนึ่ง (วันที่ 5 พฤษภาคม พ.ศ. 2565) พบรหัสข้อที่เห็นได้ชัดอย่าง หัวข้อที่ 1 ที่มีการกล่าวถึงหนึ่งในสมาชิกวงดนตรี BNK48 หัวข้อที่ 2 ที่กล่าวถึงประเพณีลอยกระทง และกล่าวถึงสังคมระหว่างยุคuren กับรัฐเชียงใหม่ในหัวข้อที่ 9 และหัวข้ออื่น ๆ ที่เกี่ยวกับการของตัวก่อนເສີຣີຕ ກາຣພັນ ແລະ ເຈິນກູ້ອອນໄລນ໌ ซึ่งตีความໄດ້ຢາກ



รูปที่ ก.21 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window

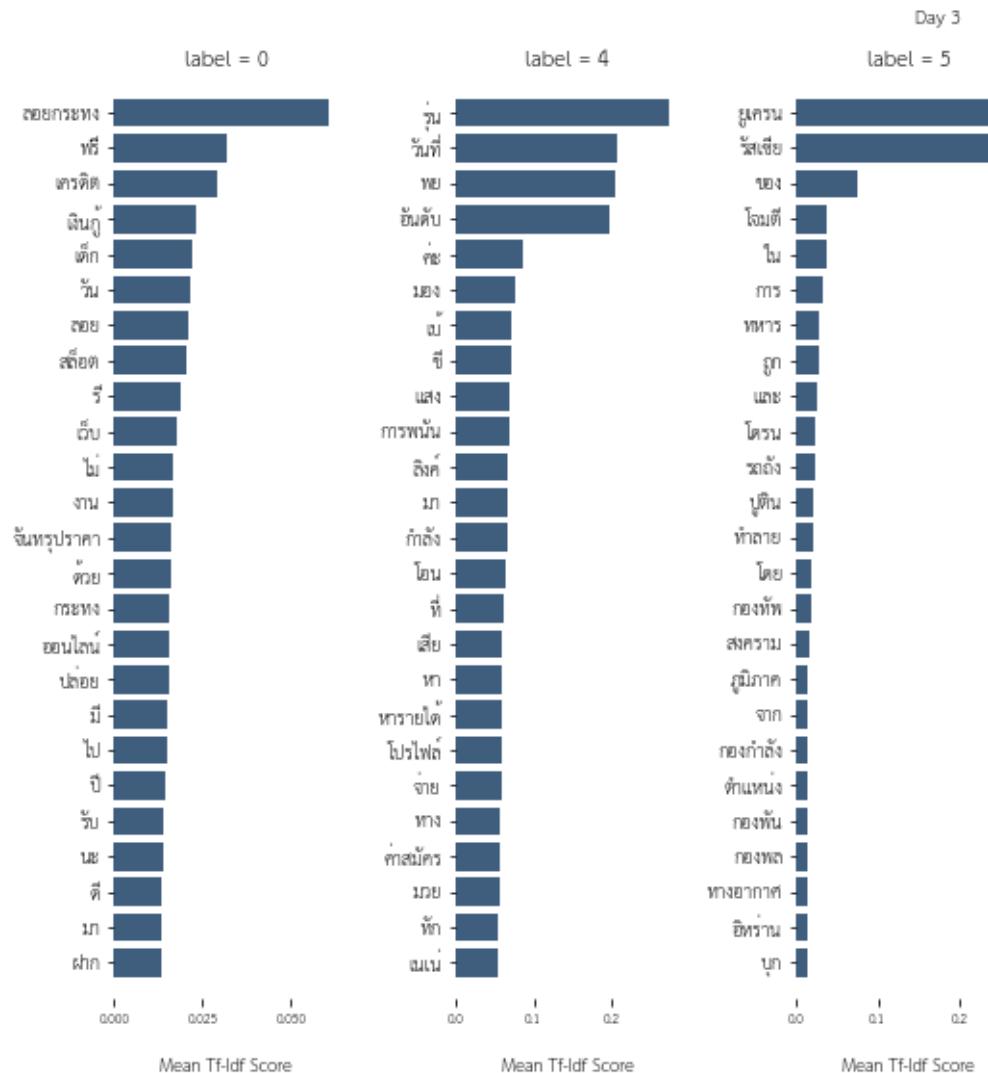
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.22 ผลลัพธ์การค้นหาหัวข้อวันที่สองจากการทดลองแบบ Expanding Window

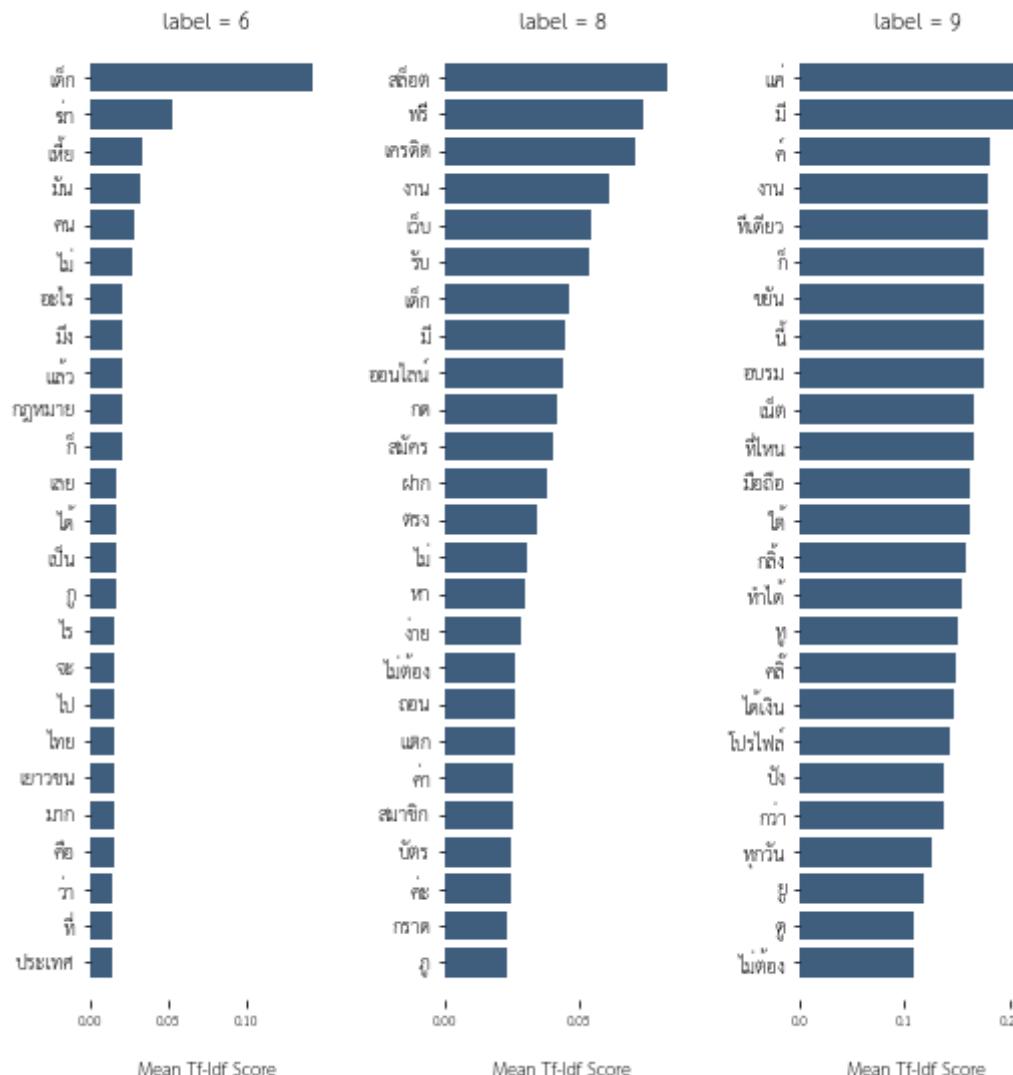
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.21 และ ก.22 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สอง (วันที่ 6 พฤษภาคม พ.ศ. 2565) พบหัวข้อที่เห็นได้ชัดคล้ายกับวันก่อนหน้าอย่าง หัวข้อที่ 1 ที่มีการกล่าวถึงการของตัวถอนเสียร์ต หัวข้อที่ 2 ที่กล่าวถึงหนึ่งในสมาชิกวงคนตรี BNK48 หัวข้อที่ 3 ที่กล่าวถึงประเพณีล้อยกระ邦 และกล่าวถึงสังคมร่วมห่วงยุเครนกับรัสเซียในหัวข้อที่ 6 อีกทั้งยังมีการพบหัวข้อใหม่อีกหัวข้อที่ 8 ที่ไม่พบในวันก่อนหน้า และไม่สามารถตีความได้



รูปที่ ก.23 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window

## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 1)



รูปที่ ก.24 ผลลัพธ์การค้นหาหัวข้อวันที่สามจากการทดลองแบบ Expanding Window

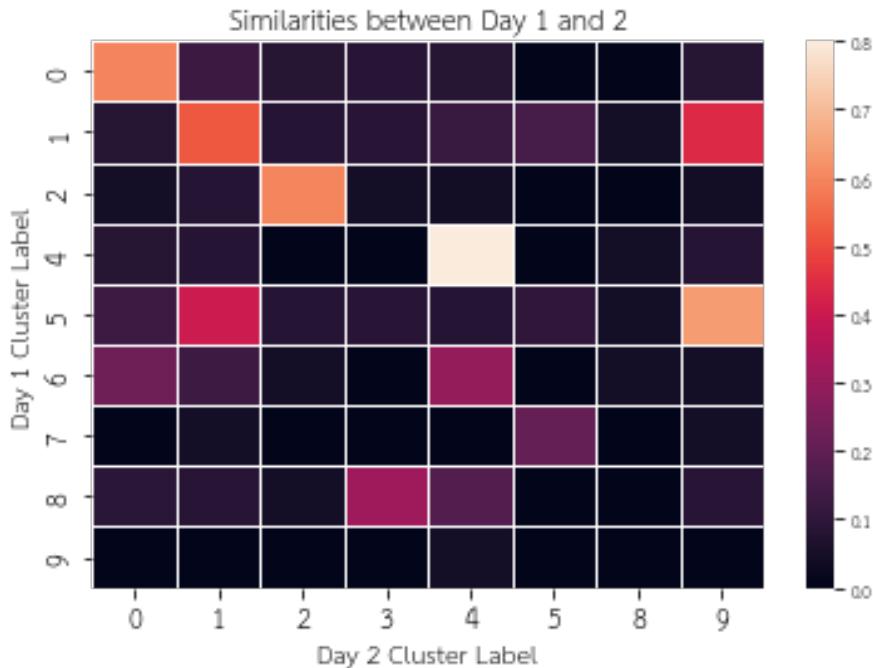
## โดยใช้เวลาเป็นปัจจัยประกอบ (ส่วนที่ 2)

จากรูปที่ ก.23 และ ก.24 เป็นผลลัพธ์ในการค้นหาหัวข้อของวันที่สาม (วันที่ 7 พฤษภาคม พ.ศ. 2565) พบหัวข้อที่เห็นได้ชัดคล้ายกับวันก่อนหน้าอย่าง หัวข้อที่ 0 ที่มีการกล่าวถึงประเพณีloykratong หัวข้อที่ 5 ที่กล่าวถึงสังคมระหว่างยุเครนกับรัสเซีย และยังมีการพบหัวข้อที่เคยแสดงในผลลัพธ์การค้นหาหัวข้อของวันที่หนึ่งอย่าง การพนัน และโ摩ฆนาออนไลน์ในหัวข้อที่ 4 และ 8 ตามลำดับ อีกทั้งยังพบผลลัพธ์ของหัวข้อที่ไม่สามารถตีความได้ในหัวข้อที่ 9

### ภาคผนวก ข.

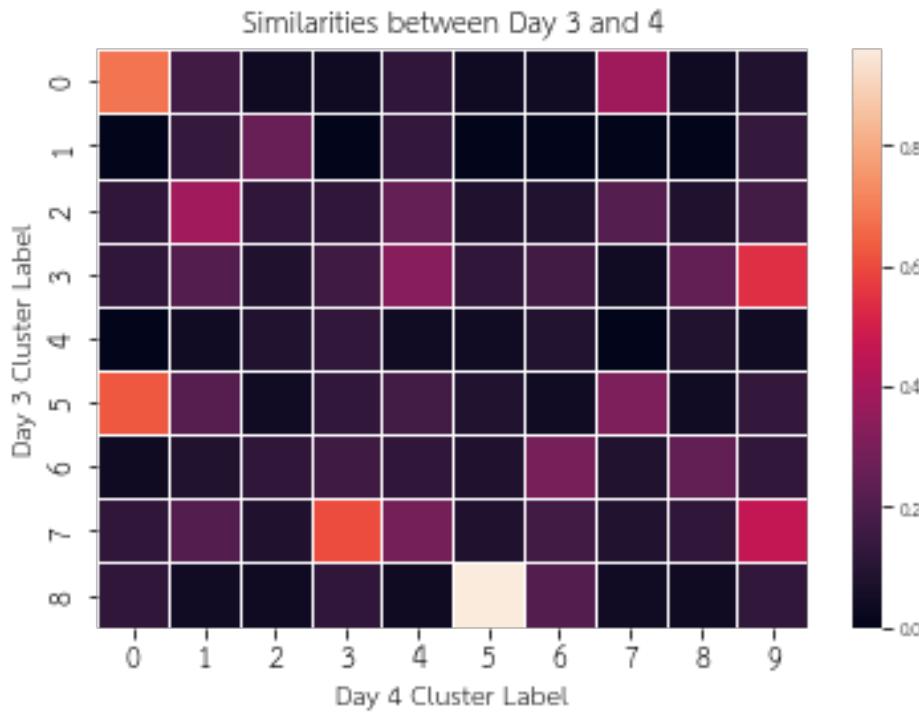
ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธีที่แตกต่างกัน

ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบไม่ใช่วลามเป็นปัจจัยประกอบ



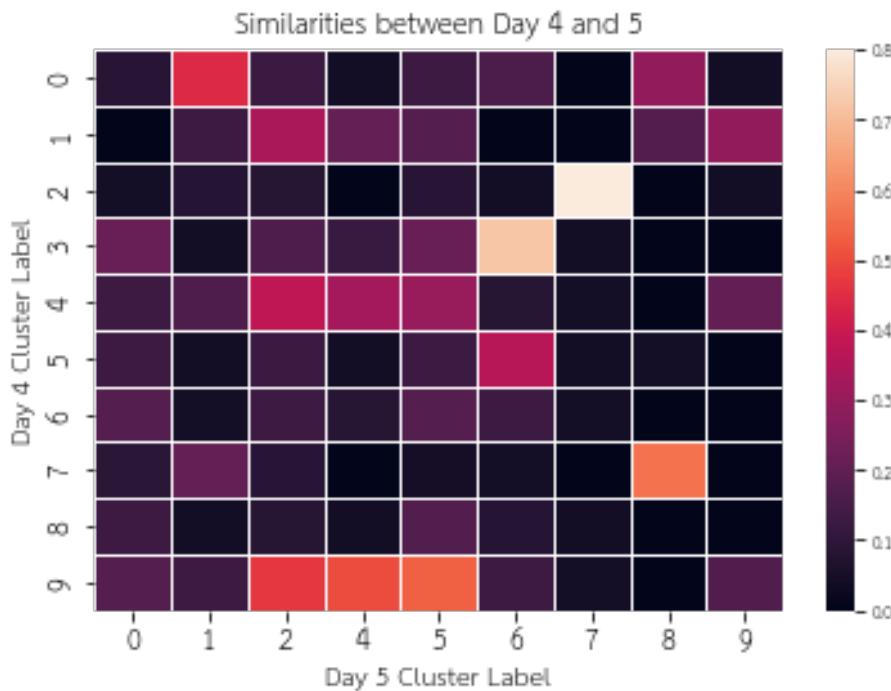
รูปที่ ๑.๑ ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช่วลามเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ๑.๑ ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่หนึ่ง (วันที่ ๕ พฤษภาคม ๒๕๖๕) และ วันที่สอง (วันที่ ๖ พฤษภาคม ๒๕๖๕) พบว่าหัวข้อที่ ๐ มีความคล้ายกับหัวที่ ๐ ของวันถัดไป หัวข้อที่ ๑ คล้ายกับหัวข้อที่ ๒ และ ๙ ของวันถัดไป หัวข้อที่ ๒ คล้ายกับหัวข้อที่ ๒ ของวันถัดไป และ หัวข้อที่ ๔ คล้ายกับหัวข้อที่ ๙ ของวันถัดไป ซึ่งเมื่อดูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นว่าเนื้อหาไม่เปลี่ยนไปเล็กน้อย แต่คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อยังคงเดิมอย่าง หัวข้อเกี่ยวกับสังคม ระหว่างยุครุนและรัตนเซีย คำว่า ยุครุน และ รัตนเซีย ยังพบบ่อย แต่มีคำอื่น ๆ ที่เปลี่ยนไปอีกทั้งยังพบหัวข้อที่มีลักษณะจางหายไปจากวันแรกสู่วันถัดไปคือหัวข้อ ๖, ๗, ๘, และ ๙ ซึ่งอาจหมายความว่าการพุดถึงหัวข้อดังกล่าวเกิดขึ้นในช่วงเวลาสั้น ๆ เท่านั้น หรืออาจเป็นหัวข้อที่ไม่สามารถตีความได้ช่นหัวข้อที่ ๗



รูปที่ ข.2 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Sliding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ข.2 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สาม (วันที่ 7 พฤศจิกายน 2565) และ วันที่สี่ (วันที่ 8 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณา มีหัวข้อที่คล้ายกับวันถัดไปคือหัวข้อที่ 0 คล้ายกับหัวข้อที่ 0 ของวันถัดไป ซึ่งมีเนื้อหาเกี่ยวกับ การมาตกรรมเด็ก หัวข้อที่ 3 คล้ายกับหัวข้อที่ 9 ของวันถัดไป ซึ่งมีเนื้อหาเกี่ยวกับเว็บพนัน ออนไลน์ หัวข้อที่ 5 คล้ายกับหัวข้อที่ 0 ของวันถัดไป ที่มีเนื้อหาเกี่ยวกับการมาตกรรมเด็ก เช่นเดียวกัน หัวข้อที่ 7 คล้ายกับหัวข้อที่ 3 ของวันถัดไป ที่มีเนื้อหาเกี่ยวกับเงินกู้ออนไลน์ และ หัวข้อที่ 8 คล้ายกับหัวข้อที่ 5 ของวันถัดไป เกี่ยวกับการทำงานออนไลน์ หัวข้อข้างต้น มีองค์ประกอบคำที่สื่อถึงหัวข้อคล้ายกัน แต่มีสติการใช้คำที่แตกต่างกันออกไปบ้าง เช่น หัวข้อ กับเนื้อหาของวันถัดไปมีหัวข้อที่สื่อถึงเรื่องเดียวกันกับวันที่พิจารณา เช่น sacrament ระหว่าง ยูเครน และ รัสเซีย ซึ่งอยู่ในหัวข้อที่ 1 ของวันที่พิจารณา และอยู่หัวข้อที่ 2 ของวันถัดไป ถึงแม้ว่าหัวข้อดังกล่าวจะมีการกล่าวถึงคำว่า ยูเครน และ รัสเซีย เป็นส่วนใหญ่ แต่คำอื่น ๆ ที่ใช้ ในการประกอบบริบทต่างกันอย่างสิ้นเชิง จึงทำให้มีความคล้ายกันน้อย

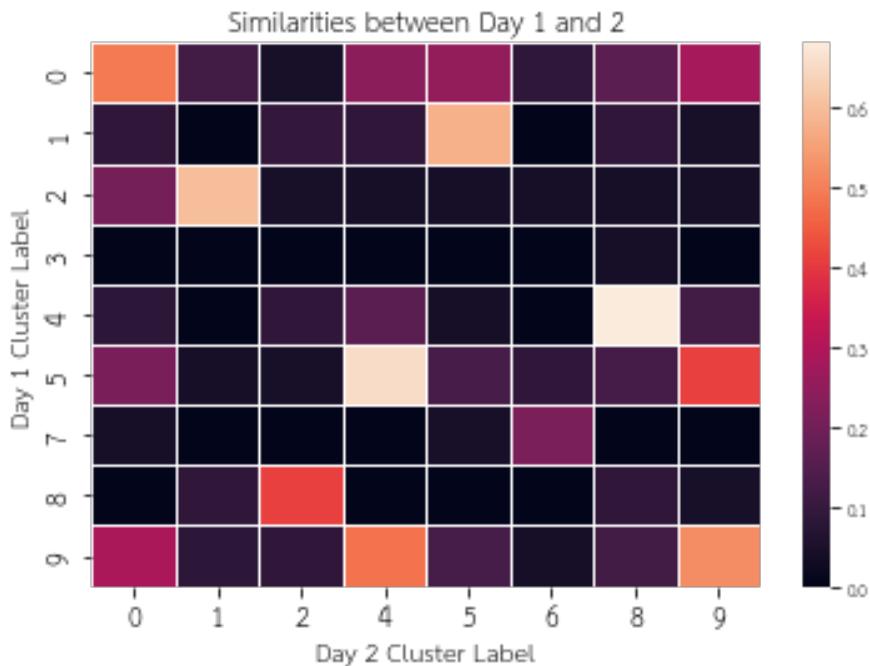


รูปที่ ข.3 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลอง

ด้วยวิธี Sliding Window แบบไม่ใช้วремาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

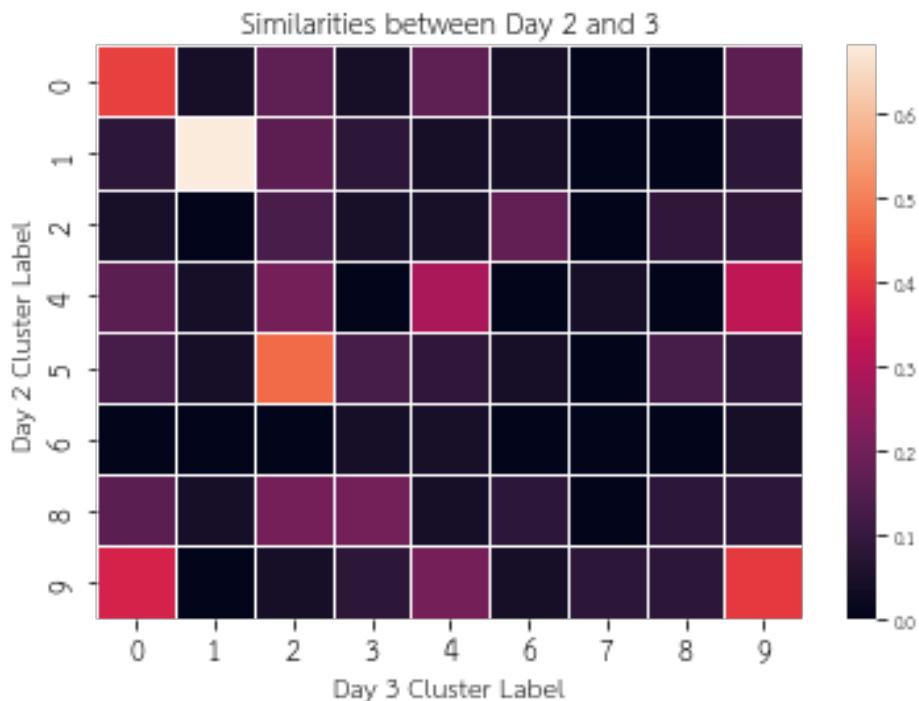
จากรูปที่ ข.3 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สี่ (วันที่ 8 พฤศจิกายน 2565) และ วันที่ห้า (วันที่ 9 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณา มีหัวข้อที่คล้ายกับวันถัดไปอย่างเห็นได้ชัดคือหัวข้อที่ 2, 3, 5, 7 และ 9 ยกตัวอย่างเช่น หัวข้อที่ 2 คล้ายกับหัวข้อ 7 ของวันถัดไป โดยมีเนื้อหาเกี่ยวกับสังคมระหว่างยุคเอนและรัฐเชีย หัวข้อที่ 7 คล้ายกับหัวข้อ 8 ของวันถัดไป โดยหัวข้อที่ 7 ของวันที่พิจารณาไม่สามารถตีความได้ แต่หัวข้อที่ 8 ของวันถัดไปสื่อถึงผลกระทบแก้วเนื่องจากมีการกล่าวถึงคำว่ารากแก้ว ส่วนหัวข้อที่ 9 คล้ายกับหัวข้อ 2, 4 และ 5 ของวันถัดไปสื่อถึงการโปรโมทการพนันออนไลน์ และหัวข้ออื่น ๆ ที่มีความคล้ายกันน้อยแต่ยังคงเนื้อหาคล้ายเดิม เช่น การมาตรฐานเด็กของหัวข้อที่ 1 ของวันถัดไป หรือเงินกู้ออนไลน์ในหัวข้อที่ 5 ของวันถัดไป เป็นต้น

ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ



รูปที่ ข.4 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

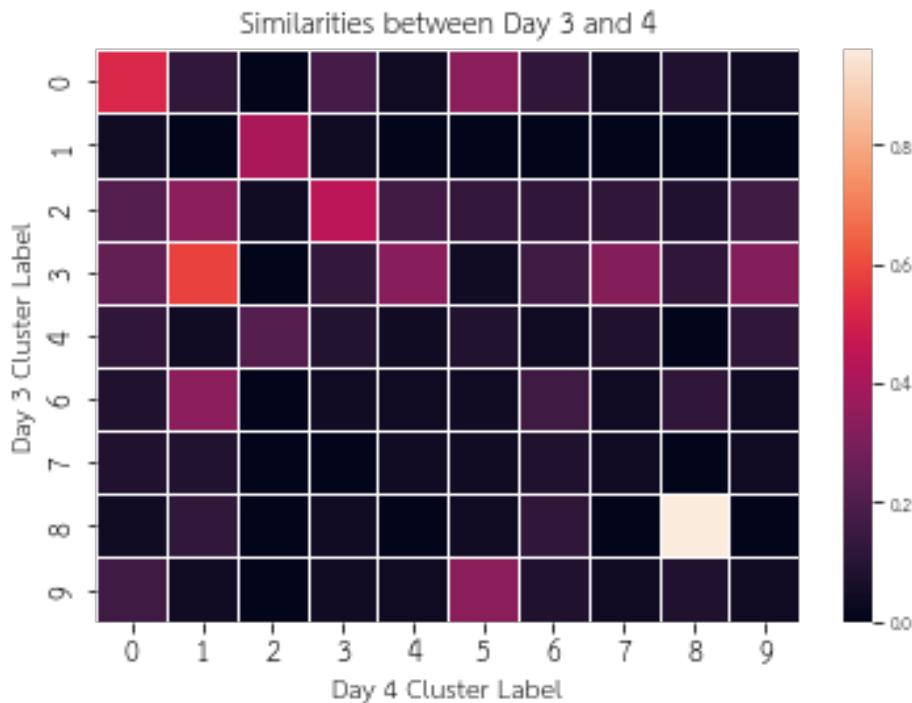
จากรูปที่ ข.4 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่หนึ่ง (วันที่ 5 พฤศจิกายน 2565) และ วันที่สอง (วันที่ 6 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณา มีหัวข้อที่คล้ายกับวันถัดไปอย่างเห็นได้ชัดคือหัวข้อที่ 0, 1, 2, 4, 5, 8, และ 9 เนื้อหาของหัวข้อ ดังกล่าวมีความคล้ายกับการทดลองแบบ Sliding Window มา กเนื่องจากการสร้างแบบจำลอง รอบแรกใช้หัวข้อมูลฝึกสอนชุดเดียวกันและทดสอบชุดเดียวกัน แต่หัวข้อในวันถัดไปจะเห็นได้ว่า เหมือนกับวันแรกมากขึ้น มีเพียงหัวข้อที่ 3 และหัวข้อที่ 7 ซึ่งอาจเกี่ยวกับการพนันออนไลน์ แต่ ไม่ปรากฏเป็นหัวข้อในวันถัดไป



รูปที่ ๑.๕ ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี

Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ๑.๕ ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สอง (วันที่ ๖ พฤศจิกายน ๒๕๖๕) และ วันที่สาม (วันที่ ๗ พฤศจิกายน ๒๕๖๕) พบว่าหัวข้อของวันที่พิจารณา มี หัวข้อที่คล้ายกับวันถัดไปอย่างเห็นได้ชัดคือหัวข้อที่ ๐, ๑, ๔, ๕, และ ๙ ซึ่งมีเนื้อหาเบื้องต้น เช่น สรุปความระหว่างยูเครนและรัสเซีย ที่มีการพูดถึงอย่างต่อเนื่องมาจนถึงวันที่ ๓ หัวข้อเกี่ยวกับ ประเพณีloyalty ในหัวข้อที่ ๕ ซึ่งคล้ายกับหัวข้อที่ ๒ ในวันถัดไป หัวข้อเกี่ยวกับวงดนตรี BNK48 สำหรับหัวข้อที่ ๔ และหัวข้อที่ ๙ ส่วนหัวข้ออื่น ๆ อาจมีการหายไป เช่น หัวข้อที่ ๖ ที่ สามารถตีความได้ยาก

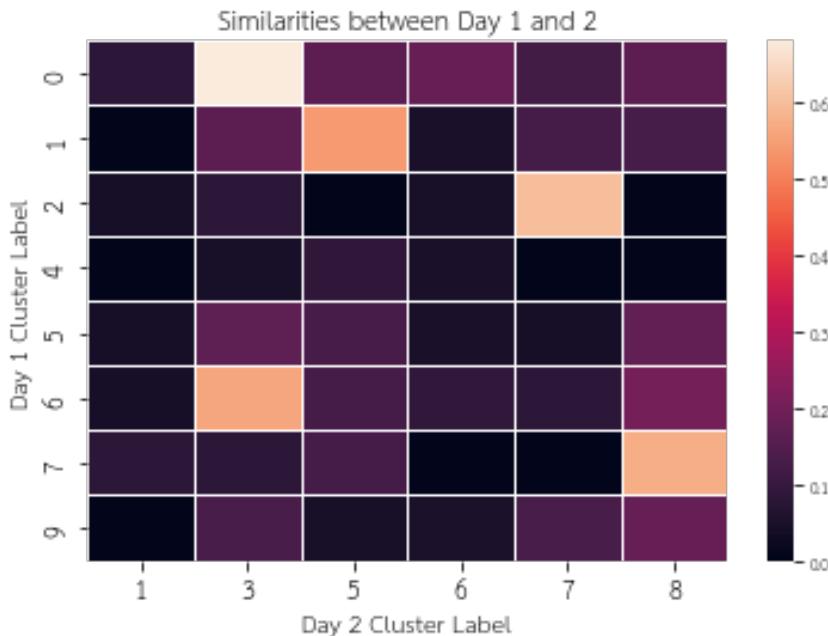


รูปที่ ข.6 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี

Expanding Window แบบไม่ใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

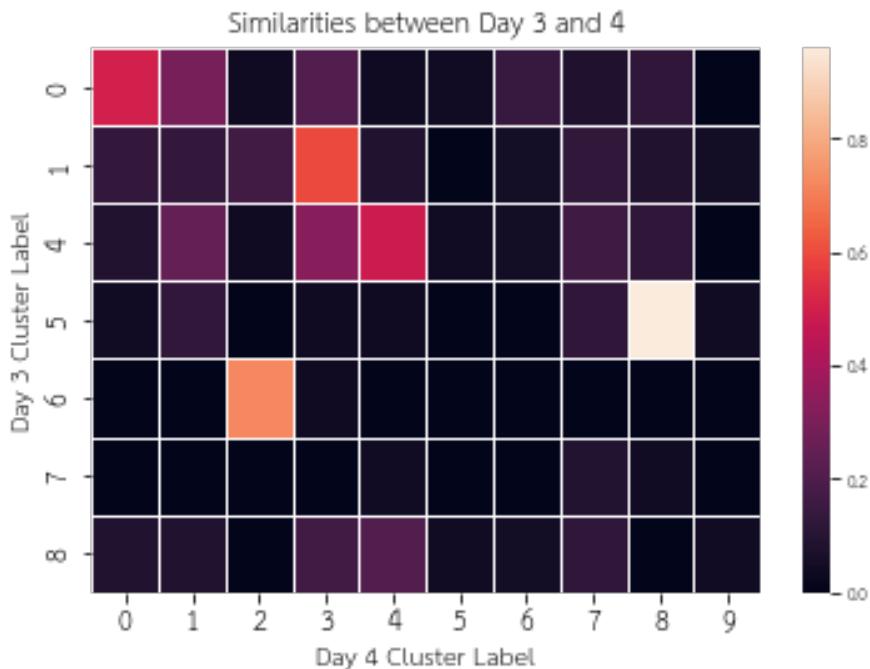
จากรูปที่ ข.6 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สาม (วันที่ 7 พฤศจิกายน 2565) และ วันที่สี่ (วันที่ 8 พฤศจิกายน 2565) พบว่าหัวข้อของวันที่พิจารณา มีหัวข้อที่คล้ายกับวันถัดไปอย่างเห็นได้ชัด อย่างหัวข้อที่ 0, 3, และ 8 มีเนื้อหาเบื้องต้น เช่น หัวข้อที่ 0 เกี่ยวกับการมาตรฐานเด็ก ซึ่งเป็นหัวข้อที่พบใหม่ และมีการเชื่อมโยงไปยังวันถัดไป หัวข้อที่ 3 เกี่ยวกับวันลอยกระทงและจันทร์ปразdroka และ หัวข้อที่ 8 ซึ่งคล้ายกับหัวข้อที่ 8 ในวันถัดไป เกี่ยวกับการทำงานออนไลน์ ส่วนหัวข้ออื่น ๆ เช่น หัวข้อที่ 1 ที่เกี่ยวกับสังคมระหว่างยุคenne และรัฐเชีย ซึ่งเชื่อมโยงไปยังหัวข้อที่ 2 แต่มีค่าความคล้ายกันน้อยซึ่งอาจเกิดจากคำอื่น ๆ ที่ถูกพูดถึงต่างกัน และหัวข้ออื่น ๆ ที่ไม่เชื่อมโยงกับวันถัดไป เช่น หัวข้อที่ 7 อาจหมายถึงการทำนาย ตลาด และหัวข้อที่ 9 ที่มีเนื้อหาเกี่ยวกับเงินกู้ออนไลน์

ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วยค่า Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ



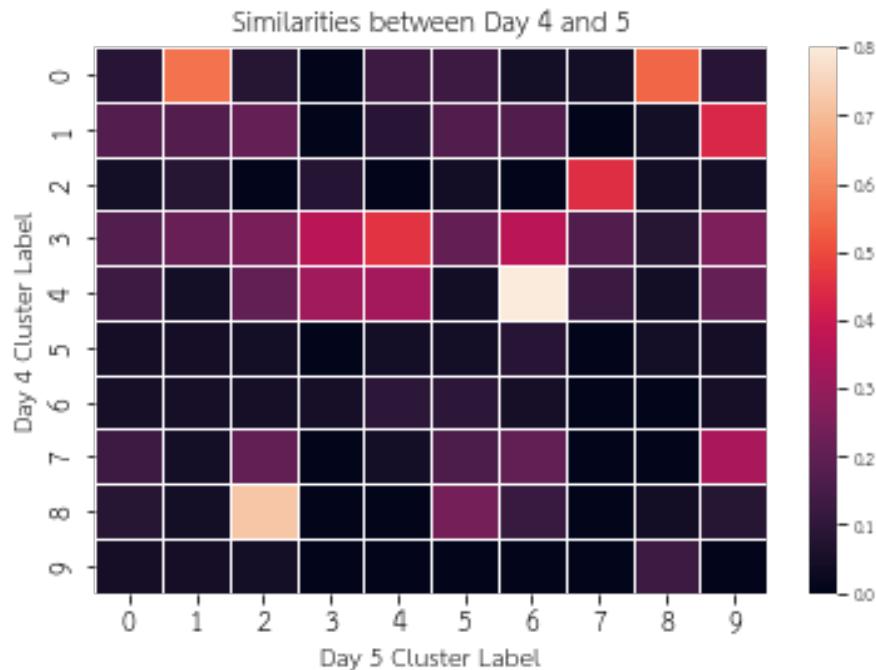
รูปที่ ข.7 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ข.7 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่หนึ่ง (วันที่ 5 พฤศจิกายน พ.ศ. 2565) และ วันที่สอง (วันที่ 6 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 0 มีความคล้ายกับหัวข้อที่ 3 ของวันถัดไป หัวข้อที่ 1 คล้ายกับหัวข้อที่ 5 ของวันถัดไป หัวข้อที่ 2 คล้ายกับหัวข้อที่ 7 ของวันถัดไป หัวข้อที่ 6 คล้ายกับหัวข้อที่ 3 ของเดือนถัดไป และหัวข้อที่ 7 คล้ายกับหัวข้อที่ 8 ของวันถัดไป ซึ่งเมื่อคุณจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นว่าเนื้อหาไม่เปลี่ยนไปเล็กน้อย แต่คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อ ยังคงเดิม เช่น หัวข้อเกี่ยวกับสิ่งแวดล้อม แต่มีคำบางคำที่เปลี่ยนไป พบหัวข้อที่มีลักษณะจากหัวข้อเดิม เช่น หัวข้อที่ 4, 5 และ 9 ซึ่งอาจหมายความว่าหัวข้อดังกล่าวเกิดขึ้นในช่วงเวลาตื้น ๆ เท่านั้น หรืออาจเป็นหัวข้อที่ไม่สามารถตีความได้ เช่น หัวข้อที่ 4



**รูปที่ ข.8** ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี Sliding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

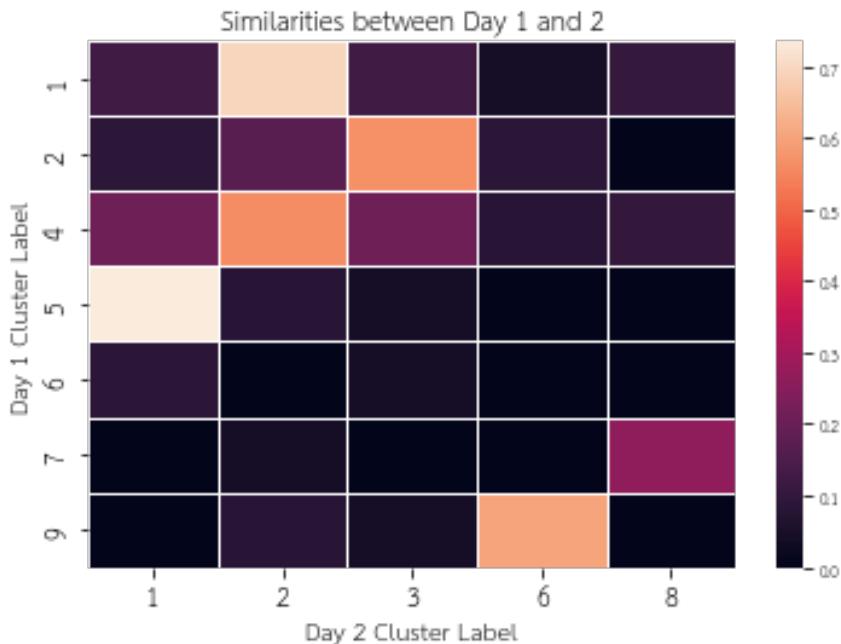
จากรูปที่ ข.8 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สาม (วันที่ 7 พฤศจิกายน พ.ศ. 2565) และ วันที่สี่ (วันที่ 8 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 5 มีความคล้ายกับหัวข้อที่ 8 ของวันถัดไปมากที่สุด โดยที่มีหัวข้อที่ 6 กับหัวข้อที่ 2 ของวันถัดไป และมีหัวข้อที่ 1 กับหัวข้อที่ 3 ของวันถัดไป ที่มีความคล้ายคลึงกันของหัวข้อรองลงมา ซึ่งเมื่อดูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นได้จากรูปที่ 4.41 และ 4.42 ว่า คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อยังคงเดิม เช่น หัวข้อเกี่ยวกับสังคม ระหว่างผู้คนกับรัฐเชีย จะเห็นว่า ผู้คน และ รัฐเชีย ยังคงเป็นคำที่ถูกกล่าวถึงบ่อย ๆ แต่ถึงหัวข้อที่ 5 และหัวข้อที่ 8 จะมีความคล้ายคลึงกันมากที่สุด แต่ก็ยากที่จะตีความหัวข้อดังกล่าวได้



รูปที่ ข.9 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สี่และวันที่ห้า ที่ทดลองด้วยวิธี Sliding Window แบบใช้วลามเป็นบล็อกจัดประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ข.9 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สี่ (วันที่ 8 พฤษภาคม พ.ศ. 2565) และ วันที่ห้า (วันที่ 9 พฤษภาคม พ.ศ. 2565) พบว่าหัวข้อที่ 4 กับหัวข้อที่ 6 ของวันถัดไป และหัวข้อที่ 8 กับหัวข้อที่ 2 ของวันถัดไปมากที่สุด โดยที่มีหัวข้อที่ 6 กับหัวข้อที่ 2 ของวันถัดไป และมีหัวข้อที่ 1 กับหัวข้อที่ 3 ของวันถัดไป ที่มีความคล้ายคลึงกันของหัวข้อรองลงมา ซึ่งเมื่อคูณกันเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นได้จากรูปที่ ก.8 ว่า คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อยังคงเดิม เช่น หัวข้อเกี่ยวกับโฆษณาหางานออนไลน์ จะเห็นว่า งาน และ ออนไลน์ ยังคงเป็นคำที่ถูกกล่าวถึงบ่อยๆ แต่หัวข้อบางหัวข้อมีความคล้ายกับหัวข้อในวันถัดไปค่อนข้างน้อยอย่าง จันทรุปรา嘉 และ ประเพณีลือยกระหง เป็นต้น

ผลลัพธ์การประเมินความเชื่อมโยงของหัวข้อด้วย Cosine Similarity จากการสร้างแบบจำลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ



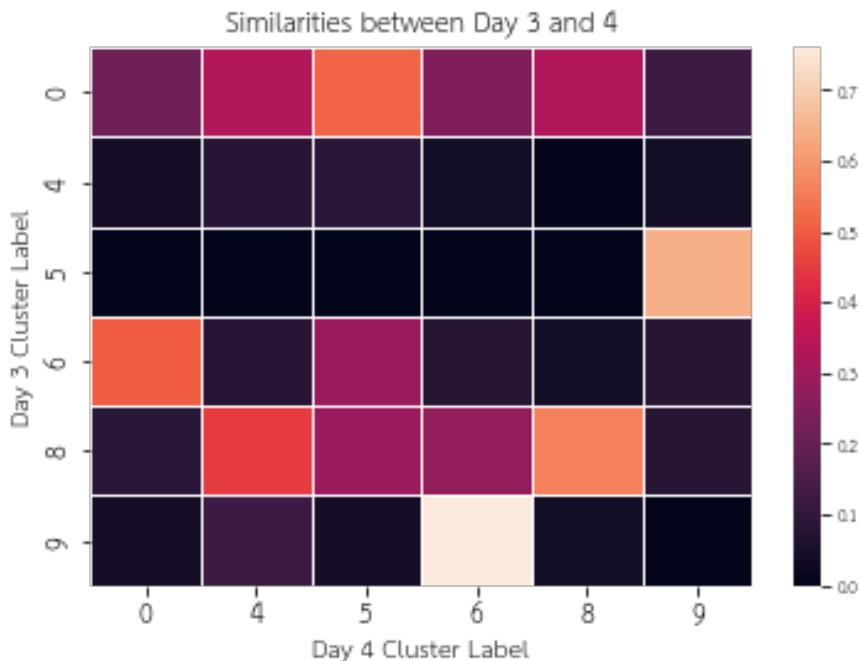
รูปที่ ช.10 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่หนึ่งและวันที่สอง ที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ช.10 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่หนึ่ง (วันที่ 5 พฤศจิกายน พ.ศ. 2565) และ วันที่สอง (วันที่ 6 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 5 กับหัวข้อที่ 1 ของวันถัดไป และหัวข้อที่ 1 กับหัวข้อที่ 2 ของวันถัดไปมากที่สุด โดยที่มีหัวข้อที่ 2 กับหัวข้อที่ 3 ของวันถัดไป หัวข้อที่ 4 กับหัวข้อที่ 2 และยังมีหัวข้อที่ 9 กับหัวข้อที่ 6 ของวันถัดไป ที่มีความคล้ายคลึงกันของหัวข้อรองลงมา ซึ่งเมื่อดูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นได้จากรูปที่ ก.10 ว่า คำที่เป็นองค์ประกอบหลักที่สืบทอด ความหมายของหัวข้อยังคงเดิม เช่น หัวข้อเกี่ยวกับการทดสอบเลิร์ต ที่ยังมีคำว่า กด รับ และ บัตร เป็นคำที่ถูกกล่าวถึงมากที่สุดในหัวข้อดังกล่าว เป็นต้น



**รูปที่ ๑.๑๑** ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สองและวันที่สาม ที่ทดลองด้วยวิธี Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ๑.๑๑ ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สอง (วันที่ ๖ พฤศจิกายน พ.ศ. ๒๕๖๕) และ วันที่สาม (วันที่ ๗ พฤศจิกายน พ.ศ. ๒๕๖๕) พบว่าหัวข้อที่ ๓ กับ หัวข้อที่ ๐ ของวันถัดไป และหัวข้อที่ ๖ กับหัวข้อที่ ๕ ของวันถัดไปมากที่สุด ซึ่งเมื่อดูจากเนื้อหา ของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นได้จากรูปที่ ๑.๑๑ ว่าคำที่เป็นองค์ประกอบ หลักที่สื่อถึงความหมายของหัวข้อยังคงเดิมเช่น หัวข้อเกี่ยวกับสังคมระหว่างยุค�훨กับ รัฐเชีย จะเห็นว่า ยุค�훨 และ รัฐเชีย และหัวข้อที่เกี่ยวกับประเพณีลอยกระทง จะเห็นว่า ลอย กระทง และ วัน เป็นคำที่ถูกกล่าวถึงบ่อยในหัวข้อดังกล่าว



รูปที่ ข.12 ผลลัพธ์ค่า Similarity ของหัวข้อระหว่างวันที่สามและวันที่สี่ ที่ทดลองด้วยวิธี

Expanding Window แบบใช้เวลาเป็นปัจจัยประกอบ ด้วยแผนภาพ Heatmap

จากรูปที่ ข.12 ผลลัพธ์การหาความคล้ายคลึงของหัวข้อระหว่างวันที่สาม (วันที่ 7 พฤศจิกายน พ.ศ. 2565) และ วันที่สี่ (วันที่ 8 พฤศจิกายน พ.ศ. 2565) พบว่าหัวข้อที่ 9 กับหัวข้อที่ 6 ของวันถัดไป และหัวข้อที่ 5 กับหัวข้อที่ 9 ของวันถัดไปมากที่สุด ซึ่งเมื่อตูจากเนื้อหาของหัวข้อที่มีความคล้ายคลึงกันในวันถัดไปจะเห็นได้จากรูปที่ ก.12 ว่า คำที่เป็นองค์ประกอบหลักที่สืบทอดความหมายของหัวข้อยังคงเดิมเช่น หัวข้อที่เกี่ยวกับสังคมระหว่างยุครุนภรัสรเซีย จะเห็นว่า ยุครุน และ รัสรเซีย เป็นคำที่ถูกกล่าวถึงบ่อยในหัวข้อดังกล่าว แต่หัวข้อที่ 9 ไม่สามารถนำมารวบรวมความหมายคำเหล่านี้ได้

## ประวัติผู้เขียน

<b>ชื่อ – นามสกุล</b>	นาย สรวิศ ยินดีอนันต์	
รหัสนักศึกษา	62070277	
วัน เดือน ปีเกิด	07 กันยายน 2543 ที่ กรุงเทพมหานคร	
ประวัติการศึกษา		
	วุฒิ ม.6 โรงเรียนบดินทรเดชา (สิงห์ สิงหนาท) ภูมิลำเนา 313/145 ซอยเคหะรัมเกล้า 64 แขวงคลองสองตันนุ่น เขตลาดกระบัง จังหวัดกรุงเทพมหานคร 10520	
เบอร์โทรศัพท์	092-598-3883	E-Mail 62070277@it.kmitl.ac.th
สาขาที่จบ	วิทยาการข้อมูล และการวิเคราะห์เชิงธุรกิจ รุ่นที่ 17	
ปีการศึกษาที่จบ	2565	

<b>ชื่อ – นามสกุล</b>	นาย อภิพล ตัวงเพียร	
รหัสนักศึกษา	62070285	
วัน เดือน ปีเกิด	17 มิถุนายน 2544 ที่ กรุงเทพมหานคร	
ประวัติการศึกษา		
	วุฒิ ม.6 โรงเรียนครุณาราชบุรี 9/51 ซอยลาดปลาเค้า 89/1 แขวงอนุสาวรีย์ เขตบางเขน กรุงเทพมหานคร 10220	
เบอร์โทรศัพท์	097-001-6987	E-Mail 62070285@it.kmitl.ac.th
สาขาที่จบ	วิทยาการข้อมูล และการวิเคราะห์เชิงธุรกิจ รุ่นที่ 17	
ปีการศึกษาที่จบ	2565	