



Social Listening and Road Accident Mapping

Members

Paweeckorn	Soratyathorn	65070501037
Mawin	Srichat	65070501045
Pichawat	Adulwittayakorn	65070501080
Sawit	Koseeyaumporn	65070507238

Present to

Dr. Sansiri Tarnpradab

Table of Contents

Introduction.....	1
Background.....	1
Objectives.....	1
Method.....	2
Data Acquisition.....	2
Twitter API v2 (Tweepy):.....	2
NTScrapers.....	2
Web Scraping with Selenium:.....	2
Exploratory Data Analysis (EDA).....	3
Reading and get the information about the tweets.csv.....	3
List all of the hashtags.....	4
#รถติด Hashtag data.....	4
#อุบัติเหตุ Hashtag data.....	4
#ป่าPNC Hashtag data.....	5
#ฝนตก Hashtag.....	5
#JS100 Hashtag.....	5
No hashtags and excluding from top 5 hashtags.....	6
Process for modeling.....	6
Labeling data for training model.....	7
Text Cleansing.....	7
Converts into DataFrame.....	7
Noise removal.....	7
Regular Expression (Regex).....	8
Stopwords Removal.....	8
NER Location and Road Text Extraction.....	8
Training the models.....	8
Process for Visualization.....	9
Text pre-processing.....	9
Location Extraction.....	9
Regular Expression.....	9
Name-entity recognition.....	9
Geocoding with geopy library.....	9
Result and Discussion.....	10
Results.....	10
Data Acquisition:.....	10
Exploratory Data Analysis (EDA):.....	10

Text Classification:.....	11
Visualization:.....	13
Discussion.....	15
Conclusion.....	16
Bibliography.....	17

Introduction

Background

In today's world, road accidents have a significant impact on traffic flow, causing congestion on certain routes. Route planning is especially crucial in industries like logistics, where delays due to accidents can lead to inefficiencies. To avoid such issues, traffic must be diverted to alternate routes when accidents occur or when there's a risk of congestion. By utilizing Social Listening data from sources such as JS100 Radio, we can analyze history accident reports, extract key information, and map accident locations. This enables more informed decisions for route adjustments, improving traffic flow and reducing the risk of further incidents.

Objectives

1. Collect Accident Data: Gather reports of road accidents from Thai news sources in the length of time, such as JS 100 Radio, using web scraping and social listening techniques.
2. Analyze and Process Thai Text: Perform text preprocessing and natural language processing (NLP) to clean, tokenize, and extract relevant information such as locations, accident severity, and time from the collected Thai text data.
3. Geocode Accident Locations: Convert extracted location data from the Thai texts into geographical coordinates (latitude and longitude) using the Google Maps API for accurate mapping.
4. Visualize Accident Data on Maps: displays real-time accident locations on Google Maps or Folium, helping users to avoid high-risk routes.

Datasets

Gathered by scraping tweets from JS100 using Selenium.

Method

1. Data Acquisition

In this study, the dataset was acquired from JS100 Twitter's feed through this link <https://twitter.com/js100radio>.

1.1 Twitter API v2 (Tweepy):

First, we tried to use the Twitter API v2 free tier to see if we could acquire it cleaner. But after we attempted to test the API, we found out the problem is that the Twitter API v2 has a very restrictive limit on API use. We can acquire the data for up to 100 posts and can send the request every 3 hours per request. Also, the most severe problem is they don't have the pagination feature for querying the posts. So we can only get 100 newest posts unless we pay them.

1.2 NTScraper

NTScraper is a Python library designed to scrape tweets directly from Twitter without the need for official APIs. This makes it incredibly accessible and easy to use, especially for those just starting in data science.

Key Features of NTScraper:

- No API Keys Requireds
- Real-Time Data
- CSV Conversion

But somehow, the NTScraper doesn't work with the Thailand Account. The example that is working is only @rashtrapatibhv. The President of India which I think it the government account so that it is working

1.3 Web Scraping with Selenium:

We decided to use the in-class method to gather the datasets called Web Scraping. We use Selenium tools using Chrome web driver to help with extracting data out of JS100's tweet feed through automation and compile them into a csv file to create tabular data that can be used in the next process

Tweet 1 : ไปพี่ท่านที่นี่มีอ่อนใจ กับสุภาพรื่นเรื่องใหญ่ กับ Rally..ที่คิดถึง เส้นทางกรุงเทพฯ-พัทยาแบบใหม่ เที่ยวบนเส้นทางบานาน พร้อมจัดกิจกรรมนี้ ตลอดเส้นทาง จะบินด้วยเครื่องเดิมๆ “ปีกอยู่ ม่องสกุล” และ “ไวรัส หัวเรือพิท” จังจะสวัสดี

Rally.. พี่คิดถึง ครั้งที่ #26

Tweet 2 : พานิชเพลสเดอลินท์วิลล์ เชียงใหม่ตามเดิมจริง 30 เพชรบุรีเจ้าของแจ้ง JS100 ไฟ #1888 ห้อง #5100
https://js100.com/en/site/_lost_found/view/145827 #JS100

Tweet 3 : #js100radio
เพื่อโปรดทราบในเมืองไทยนี่ อยู่บ้านทุกหลังต้องห่วงเพื่อนๆ “จำลองฯ วัน เด็ก หมกธรรมเนียมดูบ้านญาติ หอบเส้าไม่มีห้องนอน” ถนนและนา รถดินแมลงดูร้าวเก็บนาลอกหะ

Tweet 4 : #พอดิลลิสต์ นายวิรันดู ไฟ #1888 ผู้ห่วงใยเพื่อนที่ซื้อห้องนิลล์ บนบีช บ้าน 823 กม. เก็บเงินเป้าหมายห้องร้านอาหารได้ ภัยไม่ใช่เงินสุดและน้ำร้อนด้วยดี ดีดความประทับใจจ้าวลง

สถานที่ล่องราชา #JS100

Tweet 5 : @js100radio
ขอเชิญชวนมาลิว่า (ปี. 1982) บุญท้า พฤกษา 3 รถติดมากเดือนสัมภาระอยู่นี่ เนื่องจากลากมาปลูกเมืองรักษางานผลักหัวใจปี 2567 ผู้คนเข้าร่วมกิจกรรมเป็นจำนวนมาก กระหายน้ำ อ.กาญจนวนิชชัย และคนในใกล้ตัวอีก ไม่เข้าเป็นไปผลเดือนเดือนแห่งความรัก

#รถติดมาก2567 รถติด บุญท้า

Tweet 6 : กาแฟน้ำนมฟ้าเพลิดคลอกกองหงษ์ที่วัดกลางไฟใต้ และ วัดป่าเพลินหยัน แขวงมีนบุรี กทม. เวลา 20:15 น.
#ผลักหัวใจปี2567 #หัวใจแห่งประเทศไทย

(Cr.Kajonyot Saeim)

Tweet 7 : 20:15 #ชาบูหมู ศรีบูรพาพาร์ค ห้อง แยกชัยชาลิตาพัฒน์ กาแฟในหลวงรัชกาล ที่บริบูรณ์และดีเยี่ยมที่สุด 300g. สนใจกับรากยอด ชนกัน ชาวเชียงใหม่ห้าม ญี่ปุ่นขาดเจ็บ การจราจรเดือนด้วนๆ สร้างติด

Tweet 8 : แนะนำเชียงใหม่ไฟชุด23 อ.เชียงใหม่ที่ดี ล. สุขุมวิทที่ดีทั้งหมด JS100 ไฟ #1888 ห้อง #5100
https://js100.com/en/site/_lost_found/view/145828 #JS100

Tweet 9 : #ธุรกิจเดียวและยกประเทศ ที่ก็ คง พอ กู้ภัยกู้ไว้ก็ต้องดู
ชีวันนี้มีเมือง ไม่มีเมือง

@js100radio
...

Fig.1 Example of data scraped using Selenium with Chrome web driver

2. Exploratory Data Analysis (EDA)

We must examine and comprehend our data after scraping it from JS100Radio's tweet website in order to determine its purpose and method.

2.1 Reading and get the information about the tweets.csv

We gathered 926 tweets between October 26, 2024, and November 16, 2024.

```
# Load the CSV file
df = pd.read_csv('tweets.csv')

df.info()

→ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 926 entries, 0 to 925
Data columns (total 1 columns):
 #   Column   Non-Null Count  Dtype  
---  -- 
 0   Tweet     926 non-null    object 
dtypes: object(1)
memory usage: 7.4+ KB
```

Fig.2 Data of tweets.csv

	Tweet
0	ไปเพื่อยกับพี่ไม่มีอ่อน! กลั่นนาครังยิ่งใหญ่ กับ Rally..ที่คิดถึง เส้นทางกรุงเทพฯ-พัทยาแบบใหม่ เพื่อยกระดับมาตรฐาน พร้อมกิจกรรมใหม่ ๆ ตลอดเส้นทาง จนศึกด้วยคุณเสิร์ตจาก 'ปีอป ปองกูล' และ 'โรส ศิรินทิพย์' จองเลยวันนี้! ไทย————— ไทยRally..ที่คิดถึง ครั้งที่ #26
1	20:17 #อุบัติเหตุ #ถนนมอเตอร์เวย์ ช่วง ด้านเก็บเงินทางบก >ด้านเก็บเงินลาดกระบัง ที่บริเวณกม.38+600 รถปีค้อพ ไม่ทราบคุกรถี ขวาง ช่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
2	19:53 #ถนนรัชดาท่าพระ ช่วง แยกรัชดาตลาดพลู >แยกในส่วนรัชดาภิเษก6 รถบรรทุกจอดเสีย ขวางช่องทางซ้าย การจราจรชลอตัว
3	18:58 #อุบัติเหตุ #ถนนกาญจนากิจे�ก ช่วง ด้านระดับ345 >แยกบางกรวย ไทรน้อย ที่บริเวณแยกตลาดสมบัดบุรี เล็กน้อย รถปีค้อพ เสียหลัก ชนขอนทาง ขวางช่องทางขวา ของทางขวา บนทางขนาน การจราจรชลอตัว
4	18:55 #อุบัติเหตุ #ถนนมอเตอร์เวย์ (ช่วง แยกพัทยา >摹บประชาน) ที่บริเวณกม.6+200 จ.ชลบุรี รถเก่ง เสียหลักชนขอนทาง ขวางช่องทาง ขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)

Fig.3 Example of data from tweets.csv

From the raw datasets above we can totally see that we have some text that have hashtags itself and some don't have it.

2.2 List all of the hashtags

We need to identify all the hashtags in the Twitter dataset. By using regular expressions, we successfully extracted approximately 327 unique hashtags. Next, we proceeded to visualize this data using WordCloud and other graphs to better understand the frequency and patterns of these hashtags. After generating the visualization, we observed that the top five most frequent hashtags stood out prominently. To gain deeper insights, we decided to analyze each of these top five hashtags individually to understand their context and significance within the dataset.

2.2.1 #รถติด Hashtag data

We can guarantee that the #รถติด is the data where it's the traffic in those text. We will use it to plot in the future

	Tweet
1	20:17 #อุบัติเหตุ #ถนนมอเตอร์เวย์ ช่วง ด้านเก็บเงินทางบก >ด้านเก็บเงินลาดกระบัง ที่บริเวณกม.38+600 รถปีค้อพ ไม่ทราบคุกรถี ขวางช่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
4	18:55 #อุบัติเหตุ #ถนนมอเตอร์เวย์ (ช่วง แยกพัทยา >摹บประชาน) ที่บริเวณกม.6+200 จ.ชลบุรี รถเก่ง เสียหลักชนขอนทาง ขวางช่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
6	RT @Mitraphap_Road: 16 พ.ย.67 เวลา 17.50 น. #อุบัติเหตุ #ถนนมีดราฟ บุงหน้าสะรุบ (ขาเข้า กทม.) ช่วงลงมอคลังดง เจ้าหน้าที่กำลังยกเคลื่อนย้าย รถที่เกิดอุบัติเหตุ การจราจรติดขวางทาง #รถติด
10	16:58 #อุบัติเหตุ #ถนนจักรยานยนต์ ช่วง แยกบางพลัด >แยกถนนราชชั�นี ที่บริเวณเช.จักรยานยนต์56 รถจยย. เสียหลักล้ม ขวางช่องทางซ้าย มีผู้บาดเจ็บ การจราจรเคลื่อนตัวช้า #รถติด
11	16:30 สถานศูนย์ฯ เคลื่อนย้ายแล้ว อุบัติเหตุในช่วงเช้า คาดว่าจะเริ่มช้าลง #รถติด

Fig.4 Example data of #รถติด Hashtag

2.2.2 #อุบัติเหตุ Hashtag data

The #อุบัติเหตุ Hashtag data is also guarantee to be useful and can plot in the future

	Tweet
1	20:17 #อุบัติเหตุ #ถนนมอเตอร์เวย์ ช่วง ด้านเก็บเงินบางป้อ >ด่านเก็บเงินลาดกระบัง ที่บริเวณกม.38+600 รถปีค้อพ ไม่ทราบคุณรุ่น ขาวซ่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
3	18:58 #อุบัติเหตุ #ถนนกาญจนารักษ์ ช่วง ด่านชั้น345 >แยกบางกรวยไทรน้อย ที่บริเวณกม.6+200 จ.ชลบุรี รถเก่ง เสียหลักชนขอนทาง ขาวซ่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
4	18:55 #อุบัติเหตุ #ถนนมอเตอร์เวย์ (ช่วง แยกพัทยา >นามประนับ) ที่บริเวณกม.6+200 จ.ชลบุรี รถเก่ง เสียหลักชนขอนทาง ขาวซ่องทางขวา จนท.กำลังดำเนินการ การจราจรเคลื่อนตัวช้า #รถติด ไทย(Cr.มอเตอร์เวย์)
6	RT@Mitraphap_RoadIn : 16 พ.ย.67 เวลา 17.50 น. #อุบัติเหตุ #ถนนมีดราฟ 妩 หมุน้ำสรวงบุรี (นาเข้า กทม.) ช่วงลงมอคลองดง เจ้าหน้าที่กำลังยกเคลื่อนย้าย รถที่เกิดอุบัตินิดๆ การจราจรติดด้านมา #รถติด
7	17:50 #อุบัติเหตุ #ทางหลวง304 (นครราชสีมา >เชียงใหม่) ที่บริเวณทางลงเข้าศาลปูโทน ต.บุพราหมณ์ อ.นาดี จ.ปราจีนบุรี จนท.กำลังดำเนินการ การจราจรชลอตัวไทย (Cr. ข่าวด่วนทันเหตุการณ์อุบัติเหตุ ปราจีนบุรี)

Fig.5 Example data of #อุบัติเหตุ Hashtag

2.2.3 #ข่าวPNC Hashtag data

The #ข่าวPNC Hashtag data will not be used in the plot because it's news and don't related with the accident or traffic in the road

	Tweet
18	ไฟไหม้โรงพยาบาลทางหนึ่งของอินเดีย ทางการแอร์เบิดเสียงชี้วัด 10 ราย https://js100.com/en/site/news/view/145838... ไทย#ข่าวPNC #ไฟไหม้โรงพยาบาล #อินเดีย
47	เก็บกระหงชี้ปี! กทม.เผยแพร่ มีกระหงวัสดุธรรมชาติต้มอกว่าไฟฟ้า แยกทำปูอินทรีย์-ฝังกลม https://js100.com/en/site/news/view/145828... #ข่าวPNC ไทย#ล้อຍกระหง67
78	อย. จับมือ TikTok ควบคุมโฆษณาภัย กฎหมายเบี้ยรุก เริ่มแล้ววันนี้ https://js100.com/en/site/news/view/145814... #ข่าวPNC ไทย#อย.จับมือTikTok ไทย#ควบคุมโฆษณาภัยกฎหมาย
86	ตำรวจสอนสวนกลาง เตือนระวังกัน! ล้อຍกระหงดำเนินการเรื่องเบี้ยรุก เริ่มแล้ววันนี้ https://js100.com/en/site/news/view/145810... #ข่าวPNC ไทย#ล้อຍกระหงวันนี้
91	ผบ.ตร.เชิญเพื่อน เบี้ยรุกสั่งตร.ปี'68 ย้ำห้ามมีร่องเรียนไม่ได้เงินเด็ดขาด ! https://js100.com/en/site/news/view/145800... #ข่าวPNC ไทย#เบี้ยรุกสั่งตร.ปี'68

Fig.6 Example data of #ข่าวPNC Hashtag

2.2.4 #ฝนตก Hashtag

This hashtag is also not be used in the plot because raining condition doesn't affect the accident directly

	Tweet
19	15:00 กทม. #ฝนตก เขตทวีวัฒนา ตั้งลิ้นชั้น บางกอกน้อย บางพลัด ตุลิศ พritchard เคลื่อนตัวที่ศูนย์วันเด็ก ไทยความโน Hodดและพลีเคชัน #JS100 เชิญเดาร์ฟนไลต์ >> http://goo.gl/hoc9w8
24	14:00 #ฝนตก เขตหนองแขม ภาษีเจริญ ตั้งลิ้นชั้น เคลื่อนตัวที่ศูนย์วันเด็กเรียงหนึ่ง แนวโน้มคงที่ ไทยความโน Hodดและพลีเคชัน #JS100 เชิญเดาร์ฟนไลต์ >> http://goo.gl/hoc9w8
94	14:00 กทม. #ฝนตก เขตบางเขน พระรามสอง เคลื่อนตัวที่ศูนย์วันเด็กเรียงหนึ่ง / มีกลุ่มฝนเคลื่อนตัวจากชลบุรี ที่ศูนย์วันเด็ก ภายใน 3 ชั่วโมง หากไม่ สลายตัวไปความโน Hodดและพลีเคชัน #JS100 เชิญเดาร์ฟนไลต์ >> http://goo.gl/hoc9w8
102	13:00 กทม. #ฝนตก เขตสาทร ใหม่ ถนนเมือง คลองสามวา ลาดกระบัง หนองจอก เคลื่อนตัวที่ศูนย์วันเด็กเรียงหนึ่ง แนวโน้มคงที่ ไทยความโน Hodดและพลีเคชัน #JS100 เชิญเดาร์ฟนไลต์ >> http://goo.gl/hoc9w8
107	12:00 กทม. #ฝนตก เขตหลักสี่ บางเขน ถนนเมือง ดันนาภานา สายใหม่ ลาดพร้าว มีกุ่ม คลองสามวา ลาดกระบัง พระราม ถนนปานฯ บางรัก เคลื่อนตัวที่ศูนย์วันเด็กเรียงหนึ่ง แนวโน้มคงที่ ไทยความโน Hodดและพลีเคชัน #JS100 เชิญเดาร์ฟนไลต์ >> http://goo.gl/hoc9w8

Fig.7 Example data of #ฝนตก Hashtag

2.2.5 #JS100 Hashtag

This hashtag is have the various topics in the text and we can summarize it into the 3 categories including

1. News - which is not related to traffic and accident
2. Raining - this related like the #ฝนตก Hashtag
3. Traffic and Accident - have the similar keyword called “จราจร”

		Tweet
16	ผลสำรวจกิมเบ็งรัฐบาลลงดูประจําวันที่ 16 พฤษภาคม 2567 https://js100.com/en/site/post_share/view/145840... #JS100	
19	15:00 กทม. #ฝันดึก เขตที่ร่วมนา ดลึงชัน บางกอกน้อย บางพลัด ตุสิต พระนคร เคื่อนด้วยหัวใจที่ศรัทธาด้วย ใจความโน골ด์แอปพลิเคชัน #JS100 เชิญเดินทางไปที่ >> http://goo.gl/hoc9w8	
24	14:00 #ฝันดึก เขตหนองแขม ภาษีเจริญ ดลึงชัน เคื่อนด้วยหัวใจที่ศรัทธาด้วยเงินเท่านี้ แนวโน้มคงที่ไปตามโน골ด์แอปพลิเคชัน #JS100 เชิญเดินทางไปที่ >> http://goo.gl/hoc9w8	
60	พบสูบบุหรี่หลงที่นี่ริเวอร์ไซด์ อนาคตยังเจริญ 30 เขตบางขุนเทียน ผู้โดยบินเข้าของแข็ง JS100 โทร *1808 หรือ 1137 https://js100.com/en/site/lost_found/view/145823... #JS100	
62	#ผลเมืองตี นาใต้รัตน์ โพธิ์น้ำ ผู้ชั่นรอดแท็กซี่สีเขียวเหลือง ทะเบียน 1 นช 823 กทม. เก็บกระเพาในของผู้โดยสารชาวต่างชาติได้ ภายในไม่เป็นสดและบัตรล่าสุด ติดตามน่าสักดู ใจ#คนต้องการ #JS100	
66	นกหายจากชีวิตริมแม่น้ำเจ้าพระยา ฝั่งพระนคร ที่ 23 อ.พระบรมราชูปถัมภ์ จ.สุนทรปราการ ผู้โดยพิมพ์แข็ง JS100 โทร *1808 หรือ 1137 https://js100.com/en/site/lost_found/view/145823... #JS100	

Fig.8 Example data of #JS100 Hashtag

2.2.6 No hashtags and excluding from top 5 hashtags.

		Tweet
0	ไปเที่ยวภัยที่ไม่มีอ่อน! กลับมาครั้งยิ่งใหญ่ กับ Rally. ที่คิดถึง เส้นทางกรุงเทพฯ-พัทเยาแบบใหม่ ที่ยวหะเลหนาหนา พร้อมกิจกรรมนาน ๆ ตลอดเส้นทาง จนคืนด้วยถอนเสื้อจาก 'ปีอ่อน' ไป 'ปีอ่อน' และ 'โรส ศิรินทิพย์' รวมอยู่ใน! ไทย —> https://js100radio.tn/Rally.. #ที่คิดถึง ครั้งที่ #26	
2	19:53 #ถอนรัชชาท่าพระ ช่วง แยกรัชชาตลาดปลา >แยกใหญ่สระบุรี >แยกสีลม รับรถทุกรุ่นเดียวกัน ช่วยช่องทางช้าๆ การจราจรชลอดด้วย	
5	@js100radio.tn พื้นที่บังร่วง สำหรับการรับข้อมูลงานวิจิตรจักรพรรษา 2567 ตลอดทางเดินริมแม่น้ำเจ้าพระยา ฝั่งธนบุรี	
9	วันอาทิตย์ที่ 17 พ.ย.67 มีกิจกรรมวิ่งรายการ กรุงเทพมาราธอน BANGKOK MARATHON 2024 ไทยดูเริ่มนั่นและลืมสุดการแข่งขัน บริเวณถนนสบายน้ำชัย หน้าพระบรมมหาราชวัง (วัดพระแก้ว) และพระราชวังคากาโน่ แห่งหลักสี่เดือนทั่วทุกแห่งในจังหวัดเป็น...	
14	#ไอคอนสยามเก็บกระหงหันทีที่ไอคอนสยามร่วมกับสำนักสิ่งแวดล้อม และสำนักงานเขตคลองสาน ลงที่นี่ที่เพื่อเก็บกระหงหันทีที่เพื่อการรักษาน้ำเจ้าพระยาอย่างยั่งยืนควบคู่ไปกับการรีสาน ประเทศไทยที่ถูกจัดเก็บจะเข้าสู่กระบวนการแยกประเภท	
...	...	
915	พบแนวภาพลักษณะที่ซ่อนอยู่ในท้องสี 29 เขตบางกอกน้อย ผู้โดยพิมพ์แข็ง JS100 โทร *1808 หรือ 1137 https://js100.com/en/site/lost_found/view/145362...	
917	บางแสนวันนี้ บรรยายการเต็มปาก ลมหนาวมาถึงบางแสนแล้ว พัดแรงมากๆไทย(Cr.ที่นี่ชลบุรี)	
918	บรรยายกาศเช้านี้ (3 พ.ย.67) ยอดภูเรือ จ.เลย อุณหภูมิลดลงต่ำสุด 15 องศา มีลมแรงปานกลาง มีหaze เกิดจากมวลอากาศ นักท่องเที่ยวคลิก รับแสงแรกของวันและสัมผัสอากาศหนาวเย็นเป็นจันทร์ นาทีแรก(Cr.อุทุมพรพัฒนาภูเรือ)	
920	บรรยายกาศงานทำบุญกรุรุ่นสามัคคี วัดป่าพิชัยวัฒน์ คลองสัมภាន คลอง5 ต.บางป่า อ.บางป่า จ.สุนทรปราการไทย(Cr.สปอร์ตไลฟ์บางปู)	
923	บรรยายกาศเช้านี้(3 พ.ย.67) ทั้งยอดดอยอินทนนท์ และจุดชมวิวท่ามกลางป่า จ.เชียงใหม่ อุณหภูมิอยู่ 12 องศา ยังคงมีสภาพอากาศมืด หมอกลงด้วยความเย็นโดยทั่วไป(Cr.ว่าก่อจอมทอง จังหวัดเชียงใหม่ อุณหภูมิ ดอยอินทนนท์ รถสองแถวซึ่งดอยอินทนนท์)	

Fig.9 Example data from no hashtags and excluding from top 5 hashtags

This section of the dataset is particularly interesting as it encompasses various topics. However, removing this section would result in the loss of 368 rows, which is not ideal for model performance or data insights. Therefore, we will continue to investigate this section to ensure all valuable insights are captured.

In this project, we have undertaken numerous tasks to optimize the data for maximum benefit. To provide clarity, the methods will be divided into two main topics: Method 3, which includes all processes involved in preparing and refining the data for modeling, and Method 4, which focuses on the processes applied to the data to create visualizations, particularly the heatmap overlaid on the map.

3. Process for modeling

From our data, it is essential to capture all insightful information, not just the hashtags. Our goal is to classify whether the text excluded from this section is related to traffic and accident content. To achieve this, we selected text classification methods to analyze the **No Hashtags** and other data sections.

For this text classification task, we employed binary classification, where:

- **Label 1 (Accident):** Includes posts related to traffic, accidents, and rain.
 - **Label 0 (Non-Accident):** Excludes unrelated posts.

The rationale for using binary classification is as follows:

- Simpler and faster to implement, and it aligns well if your immediate goal is incident detection and mapping.
 - We can later refine the approach by introducing multi-class classification if detailed categorization becomes necessary.

3.1 Labeling data for training model

For this project, we created labeled data based on hashtags. Posts labeled as 1 (accident-related posts) included hashtags such as #รถติด and #อุบัติเหตุ. This process yielded a total of 470 rows for the 1 label. For posts labeled as 0 (non-accident-related posts), we used hashtags like #ข่าวPNC, #JS100, #ฝนตก, and otc. The result of label 0 was 456 rows.

3.2 Text Cleansing

This process involves creating structured data out of tweet data including

3.2.1 Converts into DataFrame

After obtaining data through web scraping, the collected information is often unstructured and not in the desired format. Therefore, the data must be organized into a DataFrame format consisting of one column: *Tweet*.

			Tweet	Time
0	อุบัติเหตุ ถนนเจริญสินห่วง ขาเข้า จำกัดส่วนพระราม 7 บัง麾น้ำแยกกางเพลส จุดเกิดเหตุปากของบริษัทบีทีวีทาวน์ 86/1 วันนี้ยามบ่ายมีล้วนๆในทราบครกนี่ ก็คงช่างมองทางข้าง			12:08
1	ถนนพญาไท ขาลงเบี้ยนบุณยารักษ์ขึ้นสมรภูมิ บัง麾น้ำแยกพญาไท และรายเทวี จำกัดส่วนพระราม 7 ห้ามเวลาส่วนในวันวันนี้บุราีบีชั้นฯ กรมทบกนบดินแดง ขาเข้า ห้ามแยกส่วนสะพานบ้าน้ำแยกสามเหลี่ยมดินแดง			11:51
2	สัญญาณไฟจราจรดังข้าง ถนนสุขุมวิท บริษัทบีทีวีทาวน์ 86/1 ห้ามแยกดังนี้ ให้ผู้ใช้ทางไปบีทีวีทาวน์			11:50
3	ถนนสุขุมวิท ขาเข้า บัง麾น้ำแยกอโศกมนตรี การจราจรคิดดี รถมากดีใจตัวได้ใช้ขาเข้า ห้ามแยกส่วนไม่มีที่ให้อีกด้วยหล่อ			11:47
4	ไฟฟ้าลัดวงจรที่สายไฟ ปากซอยหอทองหล่อ 25 ถนนสุขุมวิท แขวงคลองเตยเหนือ เขตวัฒนา เจ้าหน้าที่สถานีดีเด่นผลและยกย่องคล่องแยบ กำลังไปที่เกิดเหตุ			11:45
...
1090	พรุ่งนี้(4.พ.ย.67) งานแท่งองค์ห่วงห่อป่า ครอบรอบ 114 ปี วัดมงคลโคcharawas ต.คลองล้าน อ.บางปลอก จ.สุพรรณบุรีการท่องเที่ยว เวลา น. ทางรถ เวลา 09:39 น. ก้าวแรกหลังเลื่อนเข้าท่อง #รถดี ไป(Or.ที่สุดที่ประทับใจ)			07:09
1091	น. #อุบัติเหตุ ถนนพญาไท ขาดออก ช่วงวงจรส่วนทางข้ามบัง麾น้ำแยกพญาฯ เลี้ยวไป >แยกอุตุฯ จบฯ. 3 คนถูกชน มีคนเจ็บ วางชีวิตทางกลาง การจราจรคิดดีที่สุดแล้วน้า #รถดี			10:53
1092	น. #ถนนสายเอเป๊ะ ขาดออก บริเวณแยกดี หล.3027 อ.ป้อมฯ จ.อ่างทอง จุดชนวนรถดีที่ส่วนทาง ถนนสุกนันต์ ติดอยู่ได้ทางบ้านกันยังกัน สัก จนหอป่ายหัวหาน้านิการเดือนปี ห้ามเดินทาง(Cr.คุณภูมิ หันทร์สมบูรณ์ รอง ผอ.แขวงบางบอนล่าง(บางกอก))			09:40

Fig.10 Converted Tweet to DataFrame

3.2.2 Noise removal

process of correcting textual data to improve the performance of analytical algorithms. In our case, we do hashtag extraction to extract only relevant hashtags, pattern matching to find the start of the sentence, and structural filtering to only process tweets containing specific hashtags.

3.2.2.1 Regular Expression (Regex)

We used regular expressions to remove noise from the dataset text, ensuring better performance and higher-quality data for the model.

```
def clean_text(text):
    text = re.compile(r'[/(){}[\]\|@,;#+_\n]|u\.\').sub('', text)
    text = text.replace('jsradio', '')

    return text
```

Fig.11 clean_text function that uses regex for preprocessing

3.2.2.2 NER Location and Road Text Extraction

```
name="pythainlp/thainer-corpus-v2-base-model"
tokenizer = AutoTokenizer.from_pretrained(name)
model = AutoModelForTokenClassification.from_pretrained(name)
```

Fig. 12 ThaiNERv2 AutoTokenizer and pre-trained model

The model used in this task is **ThaiNERv2** from PyThaiNLP to remove the location in the training and in Text Classification because we need to generalize the data to fit in both train and predicted datasets without the bias from the location name

3.3 Training the models

To train our binary classification model, we utilized a Long Short-Term Memory (LSTM) neural network. LSTMs are a type of recurrent neural network (RNN) specifically designed to learn and capture temporal dependencies in sequential data. This makes them particularly effective for tasks where the order of the input data plays a significant role, such as time-series analysis or text classification. The model was trained using labeled data, with one class representing positive instances and the other representing negative instances. We employed appropriate preprocessing techniques to prepare the input data, followed by optimizing the LSTM model using a suitable loss function and optimizer to ensure robust classification performance.

4. Process for Visualization

We plan to focus on data visualization for purposes distinct from modeling. To achieve this, we will use the data after it has undergone cleansing and apply a fresh approach to data preprocessing.

4.1 Text pre-processing

4.1.1 Location Extraction

We aim to extract locations from each tweet and have explored various methods to achieve this effectively.

4.1.1.1 Regular Expression

We use regular expressions to extract words that start with specific keywords, such as "ถนน", "แยก", "สะพาน", "ซอย", and "ทางหลวง". After extracting these words, we create a column to store a list of the identified locations. Since each tweet may contain multiple locations, representing starting and ending points, these locations are flagged as unsuitable for transportation purposes.

4.1.1.2 Name-entity recognition

is a technique to identify and classify specific elements within unstructured text into predefined categories which in our case is hashtags.

After we have conducted a variety of methods and then we finalize that NER is the best approach for our application. So, we decided to use it.

4.1.2 Geocoding with geopy library

We plan to create data visualizations on a map, so we need to find a way to extract the latitude and longitude of each location. After some research, we discovered that the Python library `geopy.geocoder` is well-suited to meet our requirements, even when working with location names in Thai.

4.2 Data Visualization

After completing the textual data preparation and gathering all the necessary materials for data visualization, we explored various methods to plot our data on a map. The library that best met our expectations was `folium`. This library allows us to plot the heatmap directly on a map without requiring additional data processing, making it an ideal choice for our needs.

Result and Discussion

Results

1. Data Acquisition:

- The dataset was obtained from JS100 Radio's Twitter feed, covering tweets from October 26 to November 16, 2024, totaling 926 entries.
 - The scraping method utilized Selenium with Chrome WebDriver due to limitations of other methods like the Twitter API v2 and NTScrapers.

2. Exploratory Data Analysis (EDA):

- The dataset contained 327 unique hashtags, with the most frequent being traffic-related hashtags (#รถติด - 340 posts, #อุบัติเหตุ - 281 posts).
 - Word cloud visualizations highlighted key trends, and top hashtags were classified for relevance to traffic and accidents. Let's take a look at wordcloud and the bar chart below.



Fig.13 Word Cloud of all hashtags

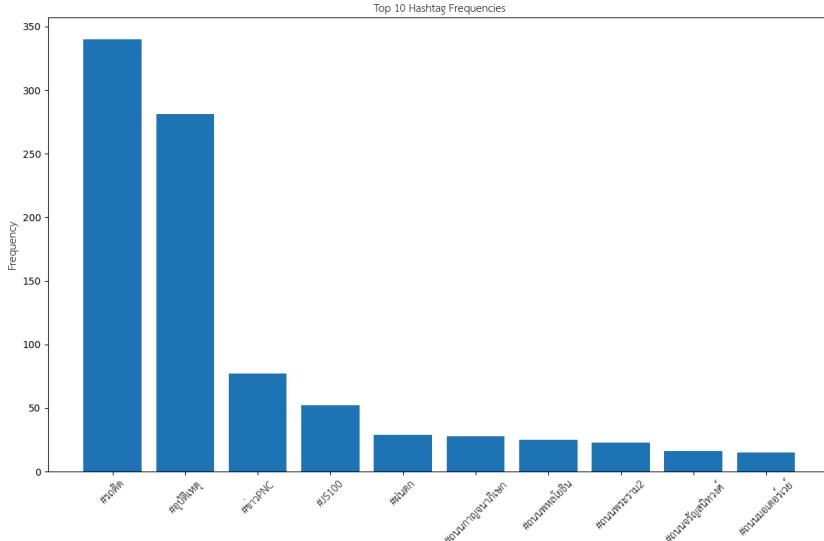


Fig.14 Top 10 hashtag frequency bar chart

From Fig.5 as a result we can assume that traffic jams have the highest frequency which can infer that user or listener of JS100 radio would typically use JS100 for monitoring traffic jam, seconded with an reported accident which are critical topic that would benefit both road-user for avoiding road that has accident occurred and would also benefit a paramedics in responding to incidents.

After identifying all the hashtags, we constructed a dataset to classify posts without hashtags for optimal performance. We assumed that posts labeled as 1 (Accident Cases) were associated with the hashtags #รถติด and #อุบัติเหตุ, while posts labeled as 0 (Other Cases) were associated with hashtags such as #แจ้งPNC, #JS100, and others. As a result, the dataset contains 470 posts labeled as 1 and 456 posts labeled as 0. Although the dataset is relatively small, it is balanced in terms of class distribution.

3. Text Classification:

- Binary classification was implemented to differentiate accident-related posts from non-accident posts.
- Small Datasets were addressed by undersampling to improve model fairness. However, oversampling and better preprocessing may be necessary for improved accuracy.
- LSTM Model

We trained an LSTM model for the binary classification task, where the labels 1 (Accident Cases) and 0 (Other Cases). To optimize the model's performance, we conducted hyperparameter fine-tuning and applied various data manipulation techniques to enhance the quality and diversity of the training dataset. The final model was trained

on a batch dataset, comprising an equal combination of samples from both label 0 and label 1. This balanced approach ensured that the model could learn effectively from both classes.

The evaluation results, as shown below, represent the best performance achieved by the model after fine-tuning and data preparation. These results demonstrate the model's ability to accurately classify the samples in the binary classification task.

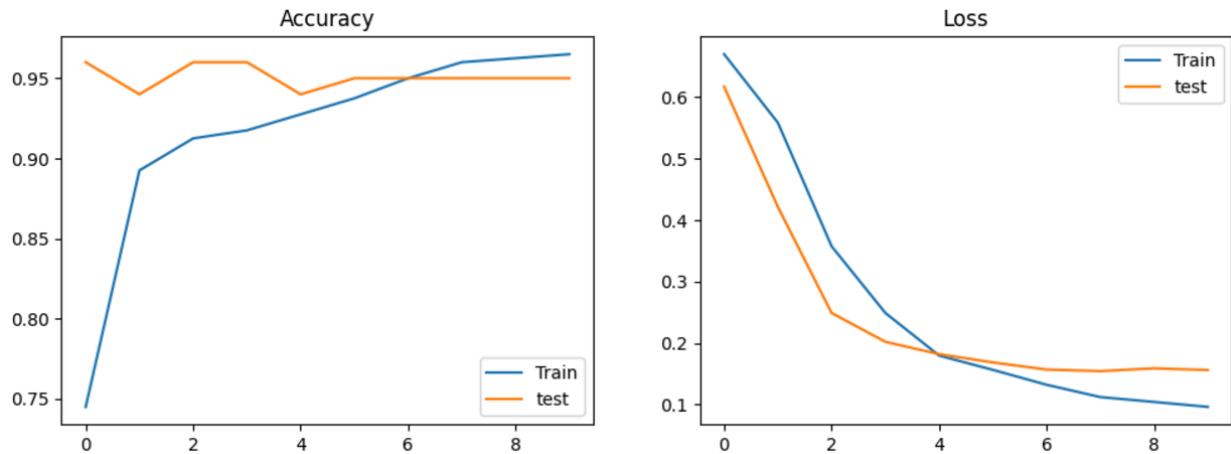


Fig.15 Result of LSTM model (batch training)

	Tweet	Prediction	Icon
82	เข้านี่ (. น. เป็นต้นไป) มีบ่วงสรวง #ขอบขามาແມ່ນ...	0	grid icon
64	: #ฝนตก เคลื่อนตัวทิศตะวันตกเฉียง...	0	bar chart icon
85	น ข้าอก จำกจุดพักรถ กม >ด่านเก็บเงิน ที่ ก...	1	
69	@jsradio\g : บรรยายกาศของหาดในหา ตอบแฉดร่มลมตกครับ	0	
90	. น. เพลิงไฟหมาภู ใกล้เดียงบลก. ภูมิชัย อชีโอด ...	0	
36	@jsradio\g ขาเข้าช่วงก่อนถึงตรงข้าม มีรถเครนจ่อ...	0	
91	: # ช่วง > ที่บริเวณก่อนถึงBTSวุฒากาศ เล็กน้อ...	1	
37	@jsradio\g เส้นโดยรอบ รถจุนมาทุกช่องทางเพื่อ...	0	
76	เวลา . น. รับแจ้งจากสายด่วน เหตุเหตุเพลิงไฟหม...	0	
60	. น. # สาย ข้าอก > ที่กม.+ - + การจราจรเคลื่...	1	

Fig.16 Prediction of Model LSTM

From **Fig. 15** and **Fig. 16**, it is evident that the model demonstrates strong performance, achieving an accuracy of approximately 95%. This high level of accuracy indicates that the model effectively distinguishes between the two classes in the binary classification task. Furthermore, we extended the evaluation by applying the trained

model to predict the values for tweets without hashtags to assess its generalization ability. The results were promising, as the predictions for non-hashtag tweets aligned well with expectations, suggesting that the model is capable of effectively handling variations in the input data. This highlights the robustness of the model in practical applications beyond the initial training dataset.

4. Visualization:

- Cleaned data was visualized using heatmaps on geographical maps with the folium library.
- Latitude and longitude extraction from tweets was achieved through geopy.geocoder, ensuring accurate location-based visualization.

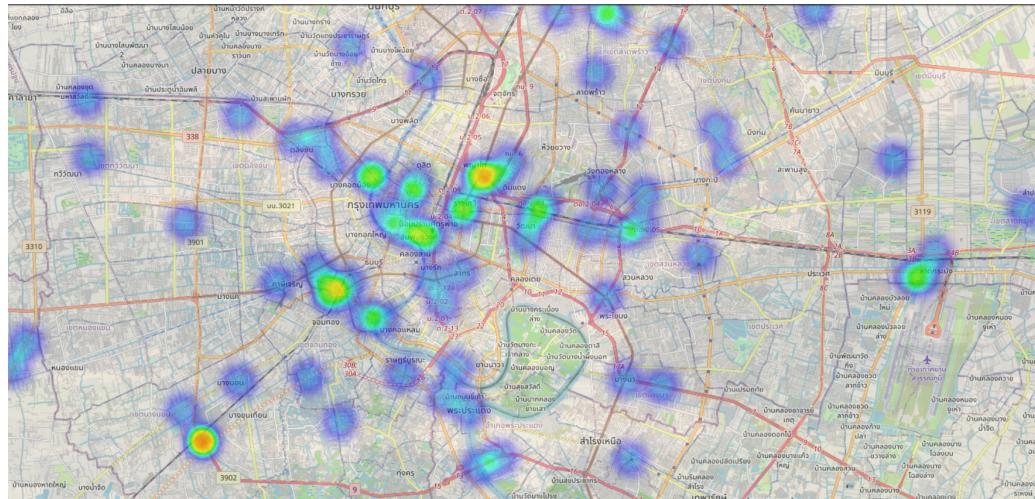


Fig.17 Closed-up visualization map

From **Fig. 17**, we can see that the heatmap was produced correctly, providing a clear visualization of the areas with the highest concentration of incidents. The heatmap highlights specific regions where incidents are most frequent, offering valuable insights for road users. Such information can help individuals identify areas that require caution while traveling or suggest alternative routes to avoid potential risks. This visualization serves as a practical tool for improving traffic safety and planning more secure travel paths.

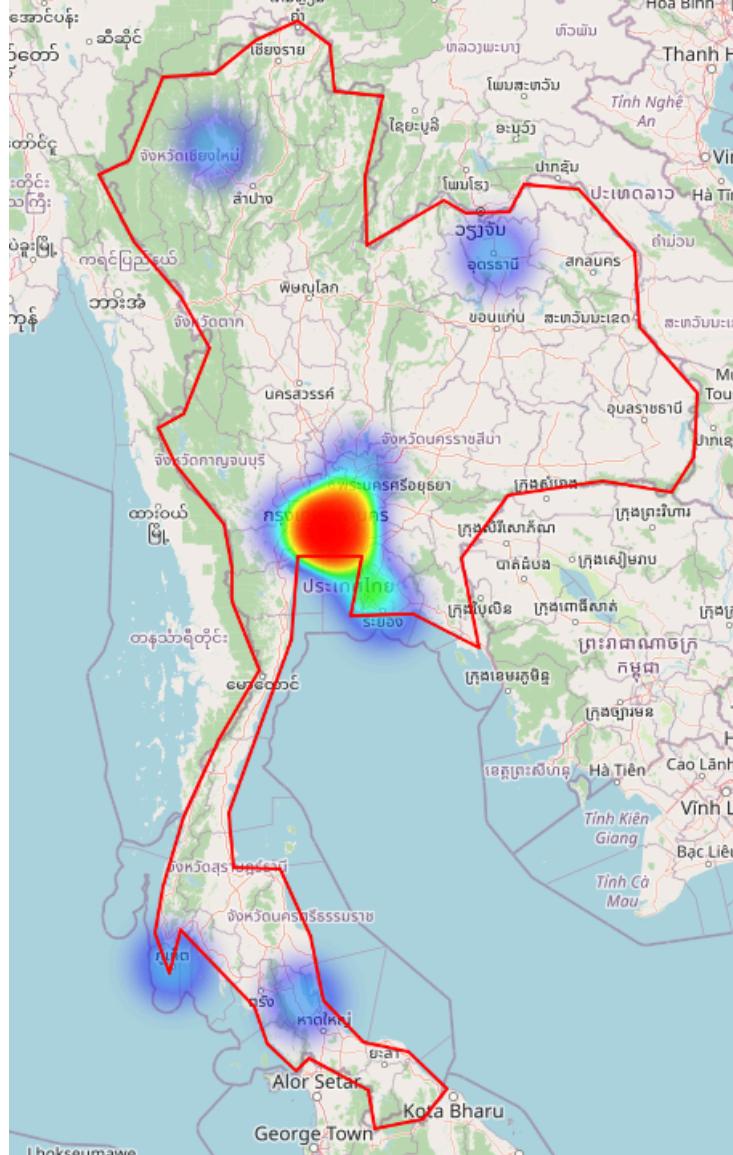


Fig.18 Visualization map zoomed-out

From this visualization, we observe that the hot spots on the map are primarily clustered in the Central region of Thailand, with some scattered spots appearing across other regions. However, upon further investigation, we identified certain errors where plotted points appeared outside the Bangkok area. These inaccuracies were found to stem from two primary sources: (1) instances where road names were identical or similar to those in Bangkok but referred to locations elsewhere, and (2) tweets that mentioned bypass roads or highways, which could span multiple regions or fall outside the intended geographical scope. This highlights the importance of refining location extraction and disambiguation techniques to improve the accuracy of spatial visualizations in future iterations.

Discussion

1. Methodology Effectiveness:
 - The choice of Selenium proved suitable for data collection under Twitter API restrictions. However, the manual effort involved in setting up and maintaining web scraping pipelines might limit scalability for future datasets.
 - The reliance on binary classification was appropriate given the dataset's focus on incident detection. However, multi-class classification could provide nuanced insights into accident types.
2. Challenges:
 - Small datasets posed significant challenges in model training, leading to potential overfitting. Undersampling addressed this partially but also limited the data's representation.
 - Noise and irrelevant content in the scraped data required extensive preprocessing to ensure quality.
3. Visualization and Practical Insights:
 - The visualization of accidents and traffic conditions on maps provides actionable insights for traffic management and urban planning.
 - The use of Python libraries like folium and geopy enabled effective data presentation without requiring additional computational overhead.
4. Future Improvements:
 - Expanding the dataset by leveraging paid API tiers or combining multiple data sources could enrich the analysis.
 - Exploring advanced techniques like synthetic data generation or transfer learning may improve classification performance for small datasets
 - Incorporating real-time visualization capabilities could enhance the utility of the system for live monitoring.

Conclusion

In this project, we successfully extracted and processed data from the JS100 Twitter feed, employing web scraping with Selenium due to the limitations of the Twitter API and other scraping methods. After obtaining a dataset of 926 entries, we conducted extensive exploratory data analysis (EDA), including hashtag extraction, frequency analysis, and the creation of labeled data for binary classification.

Our text classification focused on distinguishing traffic and accident-related posts from unrelated content. By addressing data datasets through undersampling, we evaluated the performance of LSTM models, revealing challenges such as overfitting and the need for more data and robust techniques.

For visualization, we refined the data by extracting meaningful information, such as locations using Named Entity Recognition (NER), and plotted results on an interactive map using the Folium library. This approach provided a practical method to visualize traffic and accident data in real-time, emphasizing the importance of preprocessing and data visualization in deriving insights from social media datasets.

This study highlights the potential of combining web scraping, text classification, and geospatial visualization to analyze real-time traffic and incident data for practical applications like traffic management and public safety.

Bibliography

- Daowraeng, A. (2022, January 7). การทำ Text Classification ภาษาไทยด้วย Stacking model. *Medium.* <https://medium.com/super-ai-engineer/การทำ-text-classification-ภาษาไทยด้วย-stacking-model-be12defb989e>
- Gujjar, P. K. (2024, June 30). Scraping Tweets for Real-Time Data Analysis Using ntscraper. *Medium.* <https://medium.com/@prathamsk130/scraping-tweets-for-real-time-data-analysis-using-ntscraper-a875f6d030b9>
- Kanoktipsatharporn, S. (2019, Sep 22). สอนสร้าง Word Cloud ภาษาไทย ด้วย Python ใน Jupyter Notebook / Google Colab. Bualabs. <https://www.bualabs.com/archives/2996/word-cloud-thai-language-python-matplotlib-mask/>
- pythainlp/thainer-corpus-v2 · Datasets at Hugging Face. (2023, March 23). <https://huggingface.co/datasets/pythainlp/thainer-corpus-v2>
- PyThaiNLP documentation — PyThaiNLP e034c54 documentation. (n.d.). <https://pythainlp.org/docs/5.0/index.html>