“Ebola Covid คู่หูทำลายล้าง”

รายวิชา SC312105
Basics of Data Engineering

คณะผู้จัดทำ

1). นาย สิทธิพงศ์ บับพาน รหัสนักศึกษา 613020029-1

2). นาย มีชัย หนูพิศ รหัสนักศึกษา 613020594-0

ความเป็นมาและความสำคัญ

- เนื่องจากในปัจจุบันโรคCovid 19 กำลังเป็นปัญหาใหญ่สำหรับโลกของเรา กลุ่มของพวกเราจึง จะนำข้อมูลการพูดถึงเรื่องไวรัส Covid19 มาเปรียบเทียบกับโรค Ebola ที่เคยเกิดขึ้นในอดีตและยังมีอยู่ในปัจจุบัน กลุ่มของพวกเราจึงอยากทราบว่าประชากรในโลก Twitter และ Reddit พูดถึงโรค Covid 19 และ Ebola มีความเหมือน และแตกต่างกันอย่างไร

วัตถุประสงค์

- เพื่อดึงข้อมูลจาก Twitter กับ Reddit
- เพื่อนำข้อมูลมาทำความสะอาด
- เพื่อวิเคราะห์ข้อมูล

ขอบเขตการศึกษา

- ดึงข้อมูลเกี่ยวกับโรค Covid19 และ Ebola

  จาก Twitter และ Reddit มาวิเคราะห์

วิธีการดำเนินโครงงาน

โปรแกรมที่ใช้

- Anaconda
  - Jupyter Notebook

ภาษาที่ใช้

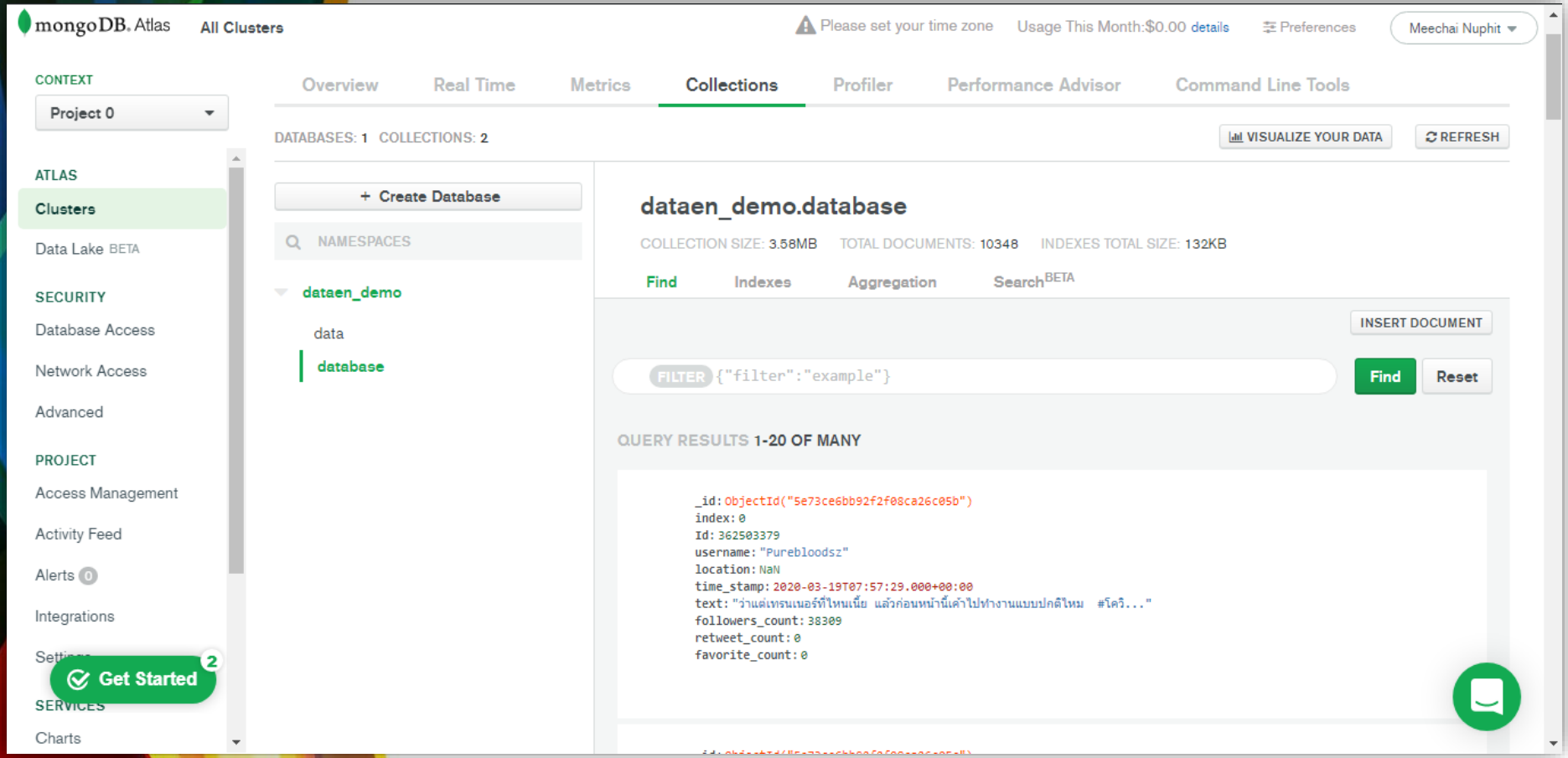
- Python 3

ผลดำเนินโครงงาน

ระบบฐานข้อมูลที่ใช้

MongoDB

MongoDB เป็น open-source document database โดยเป็นฐานข้อมูลแบบ NoSQL คือไม่มี relation (ความสัมพันธ์) ของตารางแบบ SQL ทั่วๆไป แต่จะเก็บข้อมูลเป็นแบบ JSON (JavaScript Object Notation) แทน


mongoDB. Atlas
All Clusters
Please set your time zone
Usage This Month:$0.00 details
Preferences
Meechai Nuphit
CONTEXT
Project 0
ATLAS
Clusters
Data Lake BETA
SECURITY
Database Access
Network Access
Advanced
PROJECT
Access Management
Activity Feed
Alerts 0
Integrations
Settings
Get Started 2
SERVICES
Charts
Overview
Real Time
Metrics
Collections
Profiler
Performance Advisor
Command Line Tools
DATABASES: 1 COLLECTIONS: 2
VISUALIZE YOUR DATA
REFRESH
+ Create Database
NAMESPACES
dataen_demo
data
database
dataen_demo.database
COLLECTION SIZE: 3.58MB
TOTAL DOCUMENTS: 10348
INDEXES TOTAL SIZE: 132KB
Find
Indexes
Aggregation
Search BETA
INSERT DOCUMENT
FILTER {"filter":"example"}
Find
Reset
QUERY RESULTS 1-20 OF MANY
_id: ObjectId("5e73ce6bb92f2f08ca26c05b")
index: 0
Id: 362503379
username: "Purebloodsz"
location: NaN
time_stamp: 2020-03-19T07:57:29.000+00:00
text: "ว่าแต่เทรนเนอร์ที่ไหนเนี่ย แล้วก่อนหน้านี้เค้าไปทำงานแบบปกติไหม #โควิ..."
followers_count: 38309
retweet_count: 0
favorite_count: 0

การ Query Data จากเว็บ Twitter

การ Query Data จากเว็บ Twitter

Query ข้อมูลโรคต่างๆ

import libary ที่ต้องใช้ ¶

In [77]:
```
pip install tweepy
```

```
Requirement already satisfied: tweepy in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (3.8.0)
Requirement already satisfied: PySocks>=1.5.7 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from tweepy) (1.7.0)
Requirement already satisfied: requests-oauthlib>=0.7.0 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from tweepy) (1.3.0)
Requirement already satisfied: requests>=2.11.1 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from tweepy) (2.22.0)
Requirement already satisfied: six>=1.10.0 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from tweepy) (1.12.0)
Requirement already satisfied: oauthlib>=3.0.0 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests-oauthlib>=0.7.0->t
weepy) (3.1.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests>=2.11.1->tweep
y) (2019.6.16)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests>=2.11.1->t
weepy) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from req
uests>=2.11.1->tweepy) (1.24.2)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests>=2.11.1->tweepy)
(2.8)
Note: you may need to restart the kernel to use updated packages.
```

In [114]:
```
import os
import tweepy as tw
import pandas as pd
```

In [115]:
```
#Keys and Access Tokens
consumer_key       = 'wXeF8FTaYBt59eIThBU33DPpJ'
consumer_secret    = '6BqCMOG9Okd5igSbSDfSN3A6pqjgGg8WCpmZLObOV6yM5XGBRY'
access_token       = '987188176962727936-c4q0igSeNcHBKvKQzWRd38XZ7FzInPx'
access_token_secret= 'xfhi5WPaALj1LvrRMYtWPcGm3K0eZnmOBkMqJ8OmK3etE'
```

In [116]:
```
auth = tw.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tw.API(auth, wait_on_rate_limit=True)
```

การ Query Data จากเว็บ Twitter

สร้าง function เพื่อค้นหา # และเก็บเป็น Dataframe

In [82]:
```
def search_tweets(language, hashtag, date, number):

    # Define the search term and the date_since date as variables
    search_words = hashtag
    date_since = date
    new_search = search_words + " -filter:retweets" # Do not get retweet of tweets


    # Collect tweets
    tweets = tw.Cursor(api.search, q=new_search, lang=language, since=date_since  ).items(number)

    users_locs = [[tweet.user.id_str, tweet.user.screen_name, tweet.user.location,
                   tweet.created_at, tweet.text, tweet.user.followers_count,
                   tweet.retweet_count, tweet.favorite_count] for tweet in tweets]


    # To Dataframe
    tweet_text = pd.DataFrame(data=users_locs, columns=[ 'Id', 'username', 'location','time_stamp',
                                                         'text', 'followers_count', 'retweet_count',
                                                         'favorite_count'])


    # return Dataframe
    return tweet_text
```

Covid-19

In [5]:
```
# Call function search_tweets
covid19_en = search_tweets('en', '#COVID19', '2020-01-01', 2000)
```

In [23]:
```
# Export Dataframe to CSV file
covid19_en.to_excel(r'D:\my_data\covid19_en.xlsxv', encoding='utf-8-sig', index = False)
```

In [57]:
```
covid19_en.shape
```

Out[57]: (2000, 8)

การ Query Data จากเว็บ Twitter

```
In [12]: covid19_th = search_tweets('th', '#โควิด19', '2020-01-01', 2000)

In [21]: # Export Dataframe to CSV file
         covid19_th.to_excel(r'D:\my_data\covid19_th.xlsx', encoding='utf-8-sig', index = False)

In [56]: covid19_th.shape
Out[56]: (2000, 8)

In [129]: covid19_th.head()
```

Out[129]:

	Id	username	location	time_stamp	text	followers_count	retweet_count	favorite_count
0	362503379	Purebloodsz		2020-03-19 07:57:29	ว่าแต่เทรนเนอร์ที่ไหนเนี้ย แล้วก่อนหน้านี้เค้า...	38309	0	0
1	856823519107317760	proxumer	Bangkok, Thailand	2020-03-19 07:57:25	COVID -19 ไข้หวัด หรือภูมิแพ้ 😷 แต่เราสามารถ #...	5781	0	0
2	723891243340587010	whathcn		2020-03-19 07:57:05	ขาย 🎗🎗🎗🎗🎗 พึ่งสมัครวันนี้ค่ะ 19/3/63\nแพ็กเกจ ...	0	0	0
3	1116699085166145537	sorkorlao		2020-03-19 07:57:03	คนไทยต้องช่วยกัน \n\nอย่าหวังพึ่งพาลุง\n\n#โคว...	39917	0	0
4	2479547882	BrightTodayTh	Bangkok	2020-03-19 07:56:56	สาธารณสุข แถลงสถานการณ์ โรคโควิด-19 วันที่ 19 ...	41058	1	0

การ Query Data จากเว็บ Twitter

Ebola

In [22]:
```
#Ebola
ebola_en = search_tweets('en', ' #Ebola ', '2014-01-01', 2000)
```

In [127]:
```
# Export Dataframe to CSV file
ebola_en.to_excel(r'D:\my_data\ebola_en.xlsx', encoding='utf-8-sig', index = False)
```

In [54]:
```
ebola_en.shape
```

Out[54]: (2000, 8)

In [111]:
```
ebola_en.head()
```

Out[111]:

	Id	username	location	time_stamp	text	followers_count	retweet_count	favorite_count
0	1024331753941622784	MDMEDICINE1	United Arab Emirates	2020-03-19 08:55:08	@WHO #pandemic expert @DrMikeRyan: “what we le...	47	1	1
1	853736260397084674	BeeNewsDaily	United States	2020-03-19 08:55:07	#KungFlu, #KungFuTea, and #Ebola are not raci...	4161	1	1
2	1024331753941622784	MDMEDICINE1	United Arab Emirates	2020-03-19 08:45:31	@WHO #pandemic expert @DrMikeRyan: “What we le...	47	0	0
3	3402412635	ADFmagazine	Africa	2020-03-19 08:45:08	#SierraLeone President Bio engages members of ...	374	0	0
4	18754412	MacJordaN	ACC/YFB/YEG/SFO/NBO/YYZ	2020-03-19 08:43:18	Update from #Liberia LR: \n\nThe country is ap...	16928	0	1

In [ ]:
```

```

การ Query Data จากเว็บ Reddit

การ Query Data จากเว็บ Reddit

In [2]:
```
pip install praw
```

```
Collecting praw
  Downloading https://files.pythonhosted.org/packages/25/c0/b9714b4fb164368843b41482a3cac11938021871adf99bf5aaa3980b0182/praw-6.5.1-py3-none-any.whl (134kB)
Collecting prawcore<2.0,>=1.0.1 (from praw)
  Downloading https://files.pythonhosted.org/packages/76/b5/ce6282dea45cba6f08a30e25d18e0f3d33277e2c9fcbda75644b8dc0089b/prawcore-1.0.1-py2.py3-none-any.whl
Collecting websocket-client>=0.54.0 (from praw)
  Downloading https://files.pythonhosted.org/packages/4c/5f/f61b420143ed1c8dc69f9eaec5ff1ac36109d52c80de49d66e0c36c3dfdf/websocket_client-0.57.0-py2.py3-none-any.whl (200kB)
Collecting update-checker>=0.16 (from praw)
  Downloading https://files.pythonhosted.org/packages/17/c9/ab11855af164d03be0ff4fddd4c46a5bd44799a9ecc1770e01a669c21168/update_checker-0.16-py2.py3-none-any.whl
Requirement already satisfied: requests<3.0,>=2.6.0 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from prawcore<2.0,>=1.0.1->praw) (2.22.0)
Requirement already satisfied: six in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from websocket-client>=0.54.0->praw) (1.12.0)
Requirement already satisfied: idna<2.9,>=2.5 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests<3.0,>=2.6.0->prawcore<2.0,>=1.0.1->praw) (2.8)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests<3.0,>=2.6.0->prawcore<2.0,>=1.0.1->praw) (2019.6.16)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests<3.0,>=2.6.0->prawcore<2.0,>=1.0.1->praw) (3.0.4)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in c:\users\meech\appdata\local\continuum\anaconda3\lib\site-packages (from requests<3.0,>=2.6.0->prawcore<2.0,>=1.0.1->praw) (1.24.2)
Installing collected packages: prawcore, websocket-client, update-checker, praw
Successfully installed praw-6.5.1 prawcore-1.0.1 update-checker-0.16 websocket-client-0.57.0
Note: you may need to restart the kernel to use updated packages.
```

In [2]:
```
import praw
import pandas as pd
import datetime as dt
```

In [3]:
```
# ใส่ข้อมูลเพื่อเข้า API ของ Reddit
reddit = praw.Reddit(client_id='ZkVmj1TgPUl3sQ',
                     client_secret='malppNk9bF_fK04KMtO16sjKQvM',
                     user_agent='dataen_demo',
                     username='nineo_sadboy',
                     password='catcatogc5678')
```

การ Query Data จากเว็บ Reddit

สร้าง fuction เพื่อเอาไว้ เปลี่ยนเวลา แบบ Unix time ให้เป็นวันเวลา แบบมาตรฐาน

- unix time คือมาตรฐานเวลาที่ถูกนับเริ่มมาตั้งแต่วันที่ 1 มกราคม ค.ศ.1970 โดยถูกนับเพิ่มเรื่อยมาทุกๆวินาที ดังนั้นเวลา unix time ก็คือจำนวนวินาทีที่นับมาตั้งแต่ 1 มกราคม ค.ศ.1970 นั่นเอง และยังคงนับต่อไป
- ดูเวลา Unix time ในปัจจุบันได้ที่ : http://service.meewebfree.com/php/unixtime

In [4]:
```
# function เปลี่ยนแปลง Unix time
def change_date(created):
    return dt.datetime.fromtimestamp(created)
```

สร้าง function เพื่อค้นหา กระทู้

In [5]:
```
def search_reddit(topics, number):  #input topics(type string): หัวข้อกระทู้ที่จะค้นหา and number(integer type ) จำนวนข้อมูลที่ต้องการ

    subreddit = reddit.subreddit(topics)  #กำหนดหัวข้อกระทู้
    top_subreddit = subreddit.top(limit=number)   #ดึง top ของกระทู้นั้นๆ และ กำหนด limitที่ต้องการ


    # สร้าง Data dict เอาไว้เก็บค่าจากกระทู้ที่ค้นหา key : ชื่อ column, value : list
    topics_dict = { "title":[], "score":[], "id":[], "url":[],"comms_num": [], "created": [],"body":[]}

    # ใช้ loop for เพื่อทำการดึงข้อมูล ไปเก็บใน Data dict ที่สร้างไว้ข้างบน
    for submission in top_subreddit:
        topics_dict["title"].append(submission.title)
        topics_dict["score"].append(submission.score)
        topics_dict["id"].append(submission.id)
        topics_dict["url"].append(submission.url)
        topics_dict["comms_num"].append(submission.num_comments)
        topics_dict["created"].append(submission.created)
        topics_dict["body"].append(submission.selftext)
```

การ Query Data จากเว็บ Reddit

```
# นำ Data dict มาสร้างเป็น DataFrame
topics_data = pd.DataFrame(topics_dict)


'''
    ทำการดึงเอา Unix time ในแต่ละ row ของ column created มาแปลงค่าให้เป็น วันเวลา แบบปกติ
โดย apply จาก function change_date ที่สร้างไว้ด้านบน

'''
new_time = topics_data.created.apply(change_date)

# ทำการเพิ่ม column timestamp
topics_data = topics_data.assign(timestamp = new_time)

topics_data = topics_data.drop(['created'], axis=1)



#ส่ง Dataframe
return topics_data
```

การ Query Data จากเว็บ Reddit

Covid19

In [65]:
```
reddit_covid19 = search_reddit("COVID19", 2000)
```

In [66]:
```
reddit_covid19.to_excel(r'D:\my_data\reddit_covid19.xlsx', encoding='utf-8-sig', index = False)
```

In [67]:
```
reddit_covid19.head()
```

Out[67]:

	title	score	id	url	comms_num	body	timestamp
0	Please consider downloading BOINC or folding@h...	2391	fd29vj	https://www.reddit.com/r/COVID19/comments/fd29...	1047	Hello all.\n\nI believe this has been posted b...	2020-03-04 12:56:00
1	CDC recommends cancelling or postponing all pu...	2184	fjbv0q	https://www.cdc.gov/coronavirus/2019-ncov/comm...	444		2020-03-16 15:35:33
2	“We were able to ascertain that patients who h...	1998	fkizd0	https://www.mediterranee-infection.com/wp-cont...	442		2020-03-18 18:33:00
3	Data from SARS outbreak showed that mask weari...	1795	ffy8av	https://www.cochranelibrary.com/cdsr/doi/10.10...	476		2020-03-10 08:26:30
4	Relationship between the ABO Blood Group and t...	1734	fjzjpc	https://www.medrxiv.org/content/10.1101/2020.0...	407		2020-03-17 20:05:45

การ Query Data จากเว็บ Reddit

ebola

In [6]:
```
reddit_ebola = search_reddit("Ebola", 2000)
```

In [7]:
```
reddit_ebola.to_excel(r'D:\my_data\reddit_ebola.xlsx', encoding='utf-8-sig', index = False)
```

In [8]:
```
reddit_ebola.head()
```

Out[8]:

	title	score	id	url	comms_num	body	timestamp
0	I met 25-year-old Moses Massaquoi today. He's ...	1957	2jg459	http://i.imgur.com/kLhaux5.jpg	129		2014-10-17 10:43:08
1	Ebola survivor Christopher, 43, washes blue pa...	1738	2jii6m	http://i.imgur.com/eEhQPJu.jpg	91		2014-10-18 04:12:32
2	Yesterday I showed you the ebola kit on our IC...	1296	2jeilp	http://imgur.com/a/ZQWR4	493		2014-10-17 00:05:01
3	outside an Ebola hospital, guy pulls a mattres...	997	2j7j4s	http://i.imgur.com/19857Nm.jpg	251		2014-10-15 02:59:30
4	NYC Doctor tests positive for ebola	746	2k5fkp	http://www.nytimes.com/2014/10/24/nyregion/cra...	822		2014-10-24 15:30:47

In [9]:
```
reddit_ebola.shape
```

Out[9]: (1000, 7)

Type *Markdown* and LaTeX: α^2

การ Query Data จากเว็บ Reddit

```
In [10]: reddit_ebola.to_excel(r'D:\my_data\ebola.xlsx', index = False)
```

```
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=newssearch&cd=1&cad=rja&uact=8&ved=0CB0QqQIoADAA&url=http%3A%2F%2Ffox13now.com%2F2014%2F10%2F03%2Fprimary-childrens-hospital-confirms-patient-does-not-have-ebola%2F&ei=hD0vVOOcNLTGsQTZtICoDw&usg=AFQjCNG9pBIdes91lXiHDnz7PEKijOzitA&sig2=vDgd7Gd6C1oYcWodTC9a5g' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'http://www.reddit.com/r/news/comments/2jfwyp/another_ebola_case_inova_hospital_in_northern?sort=new

Apparently%20this%20guy%20was%20an%20inmate%20and%20had%20come%20from%20W%20Africa%20within%20the%20past%2021%20days.%20Let's%20see%20if%20reddit%20is%20correct%20and%20beats%20the%20msm%20to%20the%20punch%20again.' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'http://stacks.cdc.gov/view/cdc/24900

Makes sense, I guess, since it was government-funded. The spreadsheet seems limited in a lot of ways though. (Lack of very many variables, and I wonder how they are doing the PDE's.) This is their toolbox?

When I offer 3 initial contacts for a population of 350,000,000, I see the S-curve maxing out at 8,000. Raising that initial number to ten maxes out the S-curve at 20,000. Take all with a teaspoon of salt. I haven't looked at their techniques yet.

See also:

http://www.cdc.gov/media/releases/2014/s0923-ebola-model-Factsheet.html

http://www.cdc.gov/mmwr/preview/mmwrhtml/su6303a1.htm

For an interesting alternative model to show you all the varieties of ways to model this outbreak, see, e.g., http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000984' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
```

การ Cleaning Data จากเว็บ Twitter

การ Cleaning Data จากเว็บ Twitter

Cleaning data

In [1]:
```
import pandas as pd
```

In [127]:
```
# open file
df_covid_th = pd.read_excel('D:/my_data/covid19_th.xlsx')
df_covid_en = pd.read_excel('D:/my_data/covid19_en.xlsx')
df_ebola_en = pd.read_excel('D:/my_data/ebola_en.xlsx')
```

เช็คข้อมูล

In [25]:
```
df_covid_th.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          1275 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

In [26]:
```
df_covid_en.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          1639 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

In [27]:
```
df_ebola_en.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          1587 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

จะเห็นได้ว่า column location มี missing value

- ส่วน column location ของ df_covid_th ทำการสมมติค่าให้เป็น Thailand ทั้งหมด เพราะทำการสมมติให้เป็นการ tweet จากประเทศไทยทั้งหมด
- ส่วน column location ของ df_covid_en , df_ebola_en เราจะทำการลบ row ที่มีค่า NAN ทิ้ง เพราะไม่สามารถสมมติได้ว่าจะให้เป็นเมืองหรือประเทศใด

In [128]:
```
df_covid_th.location = 'Thailand'
df_covid_en = df_covid_en.dropna()
df_ebola_en = df_ebola_en.dropna()
```

ทำการตรวจสอบข้อมูลอีกรอบ

In [40]:
```
df_covid_th.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          2000 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

การ Cleaning Data จากเว็บ Twitter

In [27]:
```
df_ebola_en.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          1587 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

จะเห็นได้ว่า column location มี missing value

- ส่วน column location ของ df_covid_th ทำการสมมติค่าให้เป็น Thailand ทั้งหมด เพราะทำการสมมติให้เป็นการ tweet จากประเทศไทยทั้งหมด
- ส่วน column location ของ df_covid_en , df_ebola_en เราจะทำการลบ row ที่มีค่า NAN ทิ้ง เพราะไม่สามารถสมมติได้ว่าจะให้เป็นเมืองหรือประเทศใด

In [128]:
```
df_covid_th.location = 'Thailand'
df_covid_en = df_covid_en.dropna()
df_ebola_en = df_ebola_en.dropna()
```

ทำการตรวจสอบข้อมูลอีกรอบ

In [40]:
```
df_covid_th.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 8 columns):
Id                2000 non-null int64
username          2000 non-null object
location          2000 non-null object
time_stamp        2000 non-null datetime64[ns]
text              2000 non-null object
followers_count   2000 non-null int64
retweet_count     2000 non-null int64
favorite_count    2000 non-null int64
dtypes: datetime64[ns](1), int64(4), object(3)
memory usage: 125.1+ KB
```

การ Cleaning Data จากเว็บ Twitter

In [44]:
```
df_covid_en.head()
```

Out[44]:

	Id	username	location	time_stamp	text	followers_count	retweet_count	favorite_count
1	2965655845	Fukkard	Hyderabad, India	2020-03-19 07:36:11	A word of caution from Mega Star Chiranjeevi g...	92184	0	0
2	31264880	AFNtelevision	March ARB, California	2020-03-19 07:36:11	"...because every red light eventually turns g...	4908	0	0
3	1182257542455539968	nukelda1	London, England	2020-03-19 07:36:09	Day 3 in quarantine #coronavirus #londonlockdo...	7	0	0
4	4732405876	sindaniv	Bungoma, Kenya	2020-03-19 07:36:09	When you experience high body temperatures, fe...	1453	0	0
6	28541686	DrAlexShehata	GVA | Cairo | NYC | Montclair	2020-03-19 07:36:07	Here a few ideas/observations I had about #COV...	758	0	0

In [45]:
```
df_ebola_en.head()
```

Out[45]:

	Id	username	location	time_stamp	text	followers_count	retweet_count	favorite_count
0	1024331753941619968	MDMEDICINE1	United Arab Emirates	2020-03-19 08:55:08	@WHO #pandemic expert @DrMikeRyan: "what we le...	47	1	1
1	853736260397084032	BeeNewsDaily	United States	2020-03-19 08:55:07	#KungFlu, #KungFuTea, and #Ebola are not raci...	4161	1	1
2	1024331753941619968	MDMEDICINE1	United Arab Emirates	2020-03-19 08:45:31	@WHO #pandemic expert @DrMikeRyan: "What we le...	47	0	0
3	3402412635	ADFmagazine	Africa	2020-03-19 08:45:08	#SierraLeone President Bio engages members of ...	374	0	0
4	18754412	MacJordaN	ACC/YFB/YEG/SFO/NBO/YYZ	2020-03-19 08:43:18	Update from #Liberia LR: \n\nThe country is ap...	16928	0	1

การ Cleaning Data จากเว็บ Twitter

สร้างฟังก์ชัน ลบอักขระ ไอคอน สติ๊กเกอร์

In [129]:
```
import re
import sys
import string
```

In [130]:
```
# function ลบ icon
def remove_emoji(string):
    emoji_pattern = re.compile("["
                           u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        u"\U00002500-\U00002BEF"  # chinese char
        u"\U00002702-\U000027B0"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U00010000-\U0010ffff"
        u"\u2640-\u2642"
        u"\u2600-\u2B55"
        u"\u200d"
        u"\u23cf"
        u"\u23e9"
        u"\u231a"
        u"\ufe0f"  # dingbats
        u"\u3030"
                           "]+", flags=re.UNICODE)
    return emoji_pattern.sub(r'', string)
```

In [131]:
```
# สร้าง function เพื่อลบอักขระต่างๆ
def remove_punct(text):
    text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub('[0-9]+', '', text)
    return text
```

In [132]:
```
# ลบ URL
df_covid_th.text = df_covid_th.text.replace(r'http\S+', '', regex=True).replace(r'www\S+', '', regex=True)
df_covid_en.text = df_covid_en.text.replace(r'http\S+', '', regex=True).replace(r'www\S+', '', regex=True)
df_ebola_en.text = df_ebola_en.text.replace(r'http\S+', '', regex=True).replace(r'www\S+', '', regex=True)
```

In [133]:
```
#ลบ icon
df_covid_th.text = df_covid_th.text.apply(lambda x : remove_emoji(x))
df_covid_en.text = df_covid_en.text.apply(lambda x : remove_emoji(x))
df_ebola_en.text = df_ebola_en.text.apply(lambda x : remove_emoji(x))
```

การ Cleaning Data จากเว็บ Twitter

In [134]:
```
#ลบ อักขระ
df_covid_th.text = df_covid_th.text.apply(lambda x : remove_punct(x))
df_covid_en.text = df_covid_en.text.apply(lambda x : remove_punct(x))
df_ebola_en.text = df_ebola_en.text.apply(lambda x : remove_punct(x))
```

In [135]:
```
# ลบเครื่องหมาย เพิ่มเติมจากฟังก์ชัน
df_covid_th.text = df_covid_th.text.str.replace('\n','').str.replace('\n\n','').str.replace('\n\n\n','').str.replace('B','').str.replace('LR ','')
df_covid_en.text = df_covid_en.text.str.replace('\n','').str.replace('\n\n','').str.replace('\n\n\n','').str.replace('B','').str.replace('LR ','')
df_ebola_en.text = df_ebola_en.text.str.replace('\n','').str.replace('\n\n','').str.replace('\n\n\n','').str.replace('B','').str.replace('LR ','')
```

In [148]:
```
# ลบเครื่องหมายคำพูด
df_covid_th.text = df_covid_th.text.str.replace("'",'').str.replace('"','')
df_covid_en.text = df_covid_en.text.str.replace("'",'').str.replace('"','')
df_ebola_en.text = df_ebola_en.text.str.replace("'",'').str.replace('"','')
```

- ลองตรวจสอบข้อมูลอีกรอบ

In [140]:
```
df_covid_th.text.head()
```

Out[140]:
```
0    ว่าแต่เทรนเนอร์ที่ไหนเนี้ย แล้วก่อนหน้านี้เค้า...
1    COVID  ไข้หวัด หรือภูมิแพ้  แต่เราสามารถ เช็กอ...
2    ขาย  พึงสมัครวันนี้ค่ะ แพ็กเกจ  เดือนค่ะ ขาย...
3    คนไทยต้องช่วยกัน อย่าหวังพึ่งพาลุงโควิท โควิด ...
4    สาธารณสุข แถลงสถานการณ์ โรคโควิด วันที่  มิค พ...
Name: text, dtype: object
```

In [141]:
```
df_covid_th.text.head()
```

Out[141]:
```
0    ว่าแต่เทรนเนอร์ที่ไหนเนี้ย แล้วก่อนหน้านี้เค้า...
1    COVID  ไข้หวัด หรือภูมิแพ้  แต่เราสามารถ เช็กอ...
2    ขาย  พึงสมัครวันนี้ค่ะ แพ็กเกจ  เดือนค่ะ ขาย...
3    คนไทยต้องช่วยกัน อย่าหวังพึ่งพาลุงโควิท โควิด ...
4    สาธารณสุข แถลงสถานการณ์ โรคโควิด วันที่  มิค พ...
Name: text, dtype: object
```

ทำการ export ลงเครื่อง

In [151]:
```
df_covid_th.to_excel(r'D:\my_data\clean\twitter_covid_th_cleaned.xlsx', encoding='utf-8-sig', index = False)
df_covid_en.to_excel(r'D:\my_data\clean\twitter_covid_en_cleaned.xlsx', encoding='utf-8-sig', index = False)
df_ebola_en.to_excel(r'D:\my_data\clean\twitter_ebola_en_cleaned.xlsx', encoding='utf-8-sig', index = False)
```

การ Cleaning Data จากเว็บ Reddit

การ Cleaning Data จากเว็บ Reddit

ตัด column body

In [82]:
```
df_covid = df_covid.drop(columns=['body'], axis=1)
df_ebola = df_ebola.drop(columns=['body'], axis=1)
```

In [83]:
```
df_covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 836 entries, 0 to 835
Data columns (total 6 columns):
title        836 non-null object
score        836 non-null int64
id           836 non-null object
url          836 non-null object
comms_num    836 non-null int64
timestamp    836 non-null datetime64[ns]
dtypes: datetime64[ns](1), int64(2), object(3)
memory usage: 39.3+ KB
```

In [84]:
```
df_ebola.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 6 columns):
title        1000 non-null object
score        1000 non-null int64
id           1000 non-null object
url          1000 non-null object
comms_num    1000 non-null int64
timestamp    1000 non-null datetime64[ns]
dtypes: datetime64[ns](1), int64(2), object(3)
memory usage: 47.0+ KB
```

Cleaning Data

In [1]:
```
# import libary
import pandas as pd
```

In [26]:
```
# path to open file
path_covid = 'D:/my_data/reddit_covid19.xlsx'
path_ebola = 'D:/my_data/reddit_ebola.xlsx'
```

In [81]:
```
# open file
df_covid = pd.read_excel(path_covid)
df_ebola = pd.read_excel(path_ebola)
```

ทำการตรวจสอบค่า missing value

In [28]:
```
df_covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 836 entries, 0 to 835
Data columns (total 7 columns):
title        836 non-null object
score        836 non-null int64
id           836 non-null object
url          836 non-null object
comms_num    836 non-null int64
body         278 non-null object
timestamp    836 non-null datetime64[ns]
dtypes: datetime64[ns](1), int64(2), object(4)
memory usage: 45.8+ KB
```

In [29]:
```
df_covid.shape
```

Out[29]: (836, 7)

```
df_ebola.info()
```

In [33]:
```
# ลองแสดงค่าในตาราง
df_covid.head()
```

Out[33]:

	title	score	id	url	comms_num	timestamp
0	Please consider downloading BOINC or folding@h...	2391	fd29vj	https://www.reddit.com/r/COVID19/comments/fd29...	1047	2020-03-04 12:56:00
1	CDC recommends cancelling or postponing all pu...	2184	fjbv0q	https://www.cdc.gov/coronavirus/2019-ncov/comm...	444	2020-03-16 15:35:00
2	"We were able to ascertain that patients who h...	1998	fkizd0	https://www.mediterranee-infection.com/wp-cont...	442	2020-03-18 18:33:00
3	Data from SARS outbreak showed that mask weari...	1795	ffy8av	https://www.cochranelibrary.com/cdsr/doi/10.10...	476	2020-03-10 08:26:00
4	Relationship between the ABO Blood Group and t...	1734	fjzjpc	https://www.medrxiv.org/content/10.1101/2020.0...	407	2020-03-17 20:05:00

In [34]:
```
df_ebola.head()
```

Out[34]:

	title	score	id	url	comms_num	timestamp
0	I met 25-year-old Moses Massaquoi today. He's ...	1956	2jg459	http://i.imgur.com/kLhaux5.jpg	129	2014-10-17 10:43:00
1	Ebola survivor Christopher, 43, washes blue pa...	1738	2jii6m	http://i.imgur.com/eEhQPJu.jpg	91	2014-10-18 04:12:00
2	Yesterday I showed you the ebola kit on our IC...	1293	2jeilp	http://imgur.com/a/ZQWR4	493	2014-10-17 00:05:00
3	outside an Ebola hospital, guy pulls a mattres...	995	2j7j4s	http://i.imgur.com/19857Nm.jpg	251	2014-10-15 02:59:00
4	NYC Doctor tests positive for ebola	747	2k5fkp	http://www.nytimes.com/2014/10/24/nyregion/cra...	822	2014-10-24 15:30:00

ทำการลบ อักขระต่างๆ ออกจาก column title

In [89]:
```
import string
import re
```

In [91]:
```
string.punctuation
```

Out[91]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [92]:
```
# สร้าง function เพื่อลบอักขระต่างๆ
def remove_punct(text):
    text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub('[0-9]+', '', text)
    return text
```

In [105]:
```
# call function remove_punct
df_covid.title = df_covid.title.apply(lambda x : remove_punct(x))
df_ebola.title = df_ebola.title.apply(lambda x : remove_punct(x))
```

In [119]:
```
# ลบเครื่องหมายคำพูด
df_covid.title = df_covid.title.str.replace("'",'').str.replace('"','')
df_ebola.title = df_ebola.title.str.replace("'",'').str.replace('"','')
```

In [122]:
```
df_covid.title.head()
```

Out[122]:
```
0    Please consider downloading BOINC or foldingho...
1    CDC recommends cancelling or postponing all pu...
2    We were able to ascertain that patients who ha...
3    Data from SARS outbreak showed that mask weari...
4    Relationship between the ABO Blood Group and t...
Name: title, dtype: object
```

In [112]:
```
df_ebola.title.head()
```

Out[112]:
```
0    I met yearold Moses Massaquoi today Hes a hyge...
1    Ebola survivor Christopher  washes blue paint ...
2    Yesterday I showed you the ebola kit on our IC...
3    outside an Ebola hospital guy pulls a mattress...
4                 NYC Doctor tests positive for ebola
Name: title, dtype: object
```

In [116]:
```
# ดูจำนวน Row, column
df_covid.shape
```

Out[116]:
```
(836, 6)
```

In [117]:
```
# ดูจำนวน Row, column
df_ebola.shape
```

Out[117]: (1000, 6)

ทำการ export ลงเครื่อง

In [123]:
```
df_covid.to_excel(r'D:\my_data\clean\reddit_covid_th_cleaned.xlsx', encoding='utf-8-sig', index = False)
df_ebola.to_excel(r'D:\my_data\clean\reddit_covid_en_cleaned.xlsx', encoding='utf-8-sig', index = False)
```

```
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'https://jamanetwork.com/journals/jama/fullarticle/2762028%20This%20tied%20to%20other%20initial%20research%20is%20of%20concern.%20This%20article%20on%20Children%20https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa198/5766430%20who%20were%20hospitalized%20is%20also%20revealing.%20The%20extremely%20mild%20case%20presentation%20in%20this%20limited%20set%20of%20cases%20and%20the%20implied%20population%20of%20children%20NOT%20hospitalized%20needs%20further%20study%20including%20a%20better%20understanding%20of%20seroprevalence%20in%20children%20utilizing%20serologic%20data%20and/or%20case%20specific%20information%20on%20adult%20cases%20in%20relation%20to%20their%20contact%20with%20children%20where%20other%20potential%20exposures%20can%20be%20excluded.%20%20This%20may%20or%20may%20not%20be%20practical.' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'https://www.researchgate.net/profile/Jianqing_Wang/publication/339243337_ACE2_Expression_in_Kidney_and_Testis_May_Cause_Kidney_and_Testis_Damage_After_2019-nCoV_Infection/links/5e4bb6be299bf1cdb933e804/ACE2-Expression-in-Kidney-and-Testis-May-Cause-Kidney-and-Testis-Damage-After-2019-nCoV-Infection.pdf' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'https://www.nature.com/articles/s41591-020-0820-9?utm_source=fbk_nnc&utm_medium=social&utm_campaign=naturenews&utm_medium=social&utm_content=organic&utm_source=facebook&utm_campaign=NatureNews_&sf231597135=1&fbclid=IwAR1sK_7p7J1Djx8ZkEt3k4ARraMmC_2tDLyhTdfVBwmQANa6j9cr19qfCoM' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
C:\Users\meech\AppData\Local\Continuum\anaconda3\lib\site-packages\xlsxwriter\worksheet.py:915: UserWarning: Ignoring URL 'https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=newssearch&cd=1&cad=rja&uact=8&ved=0CB0QqQIoADAA&url=http%3A%2F%2Ffox13now.com%2F2014%2F10%2F03%2Fprimary-childrens-hospital-confirms-patient-does-not-have-ebola%2F&ei=hD0vVOOcNLTGsQTZtICoDw&usg=AFQjCNG9pBIdes91lXiHDnz7PEKijOzitA&sig2=vDgd7Gd6C1oYcWodTC9a5g' with link or location/anchor > 255 characters since it exceeds Excel's limit for URLS
  force_unicode(url))
```

สรุปผลการดำเนินโครงงาน

สรุปผลการดำเนินโครงงาน

ทำการวิเคราะห์ข้อมูล

In [1]:
```
import pandas as pd
import numpy as np
import re
import string
```

In [2]:
```
reddit_covid_en  = pd.read_excel('D:/my_data/clean/reddit_covid_th_cleaned.xlsx')
reddit_ebola_en  = pd.read_excel('D:/my_data/clean/reddit_covid_en_cleaned.xlsx')
twitter_covid_th = pd.read_excel('D:/my_data/clean/twitter_covid_th_cleaned.xlsx')
twitter_covid_en = pd.read_excel('D:/my_data/clean/twitter_covid_en_cleaned.xlsx')
twitter_ebola_en = pd.read_excel('D:/my_data/clean/twitter_ebola_en_cleaned.xlsx')
```

ลองแสดงข้อมูล

In [3]:
```
reddit_covid_en.head()
```

Out[3]:

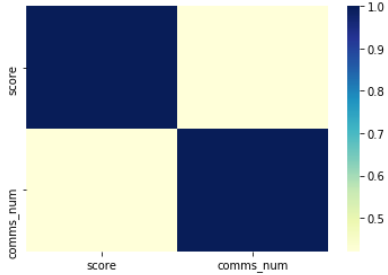
	title	score	id	url	comms_num	timestamp
0	Please consider downloading BOINC or foldingho...	2391	fd29vj	https://www.reddit.com/r/COVID19/comments/fd29...	1047	2020-03-04 12:56:00
1	CDC recommends cancelling or postponing all pu...	2184	fjbv0q	https://www.cdc.gov/coronavirus/2019-ncov/comm...	444	2020-03-16 15:35:00
2	We were able to ascertain that patients who ha...	1998	fkizd0	https://www.mediterranee-infection.com/wp-cont...	442	2020-03-18 18:33:00
3	Data from SARS outbreak showed that mask weari...	1795	ffy8av	https://www.cochranelibrary.com/cdsr/doi/10.10...	476	2020-03-10 08:26:00
4	Relationship between the ABO Blood Group and t...	1734	fjzjpc	https://www.medrxiv.org/content/10.1101/2020.0...	407	2020-03-17 20:05:00

ดูความสัมพันธ์ของแต่ละ column มนแต่ละตาราง

ทำการ plot Heatmap เพื่อดูความสัมพันธ์ของข้อมูลในตารางต่างๆ

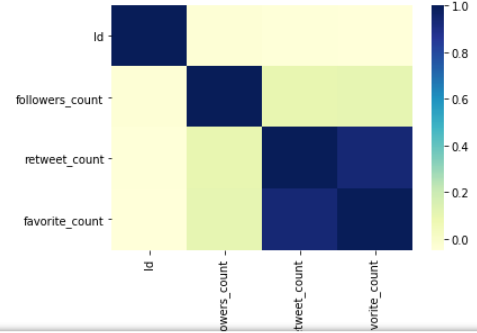
```
In [168]: sns.heatmap(reddit_covid_en.corr(), cmap = 'YlGnBu')
```

```
Out[168]: <matplotlib.axes._subplots.AxesSubplot at 0x1ebbf1fc470>
```




```
In [172]: sns.heatmap(twitter_ebola_en.corr(), cmap = 'YlGnBu')
```

```
Out[172]: <matplotlib.axes._subplots.AxesSubplot at 0x1ebc0ca3400>
```



จะเห็นได้ว่าข้อมูลจาก Twitter ทั้งสามตาราง ในแต่ละ column มีความสัมพันธ์ค่อนข้างมาก เช่น column favorite_count กับ retweet_count

สร้างกราฟจากข้อมูล

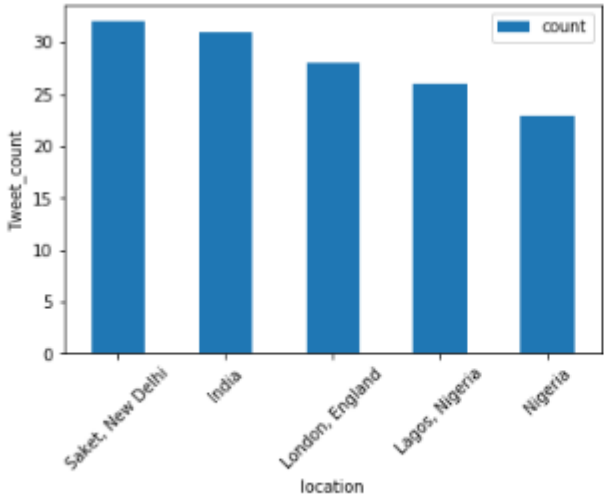
In [8]:
```
import matplotlib.pyplot as plt
```

5 location ที่มีการ Tweets ที่ติด #covid มากที่สุด จากข้อมูล Twitter ภาษาอังกฤษ

In [137]:
```
df = twitter_covid_en.groupby('location').username.agg(['count']).sort_values(by='count', ascending=False).head()
```

In [139]:
```
df.plot.bar(rot=45).set_ylabel('Tweet_count')
```

Out[139]:
```
Text(0, 0.5, 'Tweet_count')
```



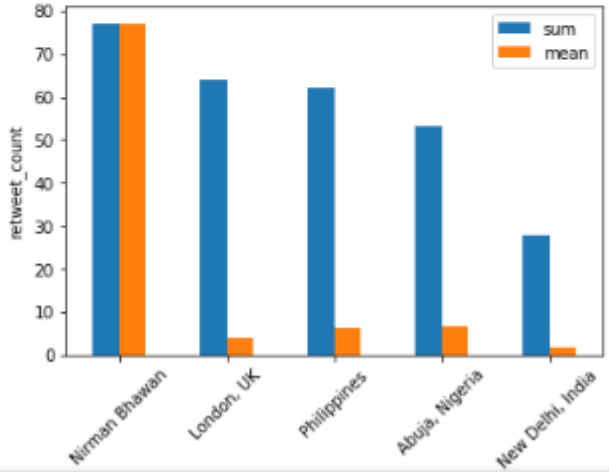
5 location ที่มีการ Re Tweets ที่ติด #covid มากที่สุด จากข้อมูล Twitter ภาษาอังกฤษ

In [111]:
```
df1 = twitter_covid_en.groupby('location').retweet_count.agg(['sum','mean'])
```

- เรียงตามผลรวม

In [164]:
```
df1.sort_values(by='sum', ascending=False).head(5).plot.bar(rot=45).set_ylabel('retweet_count')
```

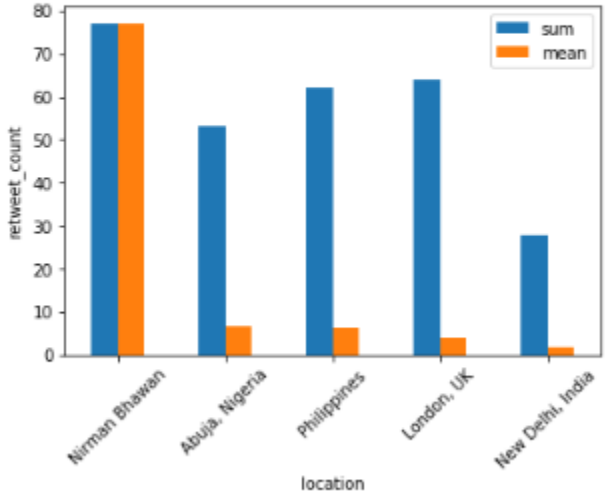
Out[164]: Text(0, 0.5, 'retweet_count')



- เรียงตามค่าเฉลี่ย

```
In [165]: df1.sort_values(by='mean', ascending=False).head(5).plot.bar(rot=45).set_ylabel('retweet_count')
```

```
Out[165]: Text(0, 0.5, 'retweet_count')
```

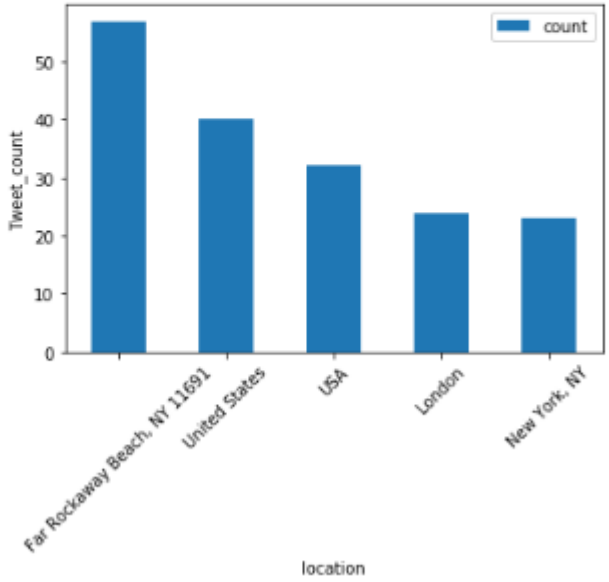


5 location แรก ที่มีการ Tweets ที่ติด #ebola มากที่สุด จากชุดข้อมูล Twitter ภาษาอังกฤษ

In [143]:
```
df2 = twitter_ebola_en.groupby('location').username.agg(['count']).sort_values(by = 'count', ascending=False).head(5)
```

In [147]:
```
df2.plot.bar(rot=45).set_ylabel('Tweet_count')
```

Out[147]:
```
Text(0, 0.5, 'Tweet_count')
```



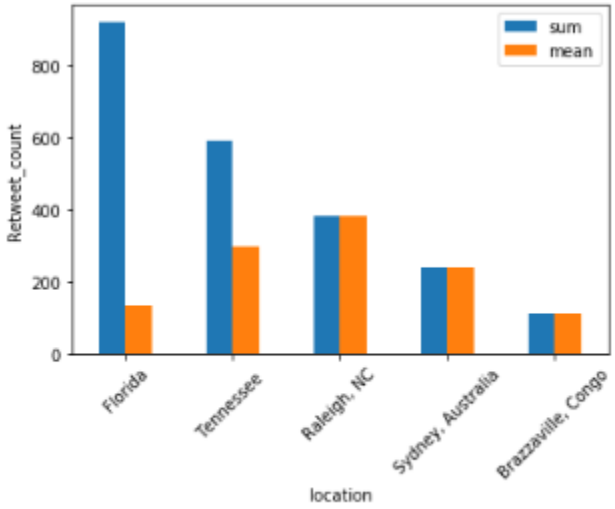
5 location แรก ที่มีการ Re Tweets ที่ติด #ebola มากที่สุด จากข้อมูล Twitter ภาษาอังกฤษ

In [162]:
```
df3 = twitter_ebola_en.groupby('location').retweet_count.agg(['sum', 'mean'])
```

- เรียงตามผลรวม

In [159]:
```
df3.sort_values(by='sum', ascending=False).head(5).plot.bar(rot=45).set_ylabel('Retweet_count')
```

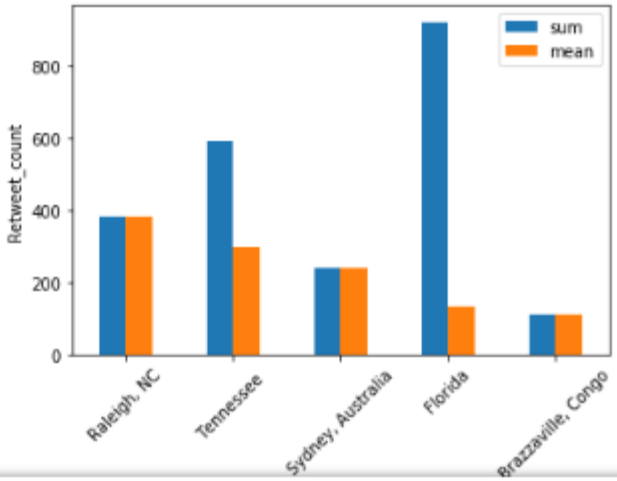
Out[159]:
```
Text(0, 0.5, 'Retweet_count')
```



- เรียงตามค่าเฉลี่ย

```
In [158]: df3.sort_values(by='mean', ascending=False).head(5).plot.bar(rot=45).set_ylabel('Retweet_count')
```

```
Out[158]: Text(0, 0.5, 'Retweet_count')
```



สรุปผลที่ได้จากกราฟเกี่ยวกับโรค Covid

ภาพแสดงจำนวนผู้ที่ติดเชื้อและเสียชีวิตจาก Covid-19 ในแต่ละประเทศ

ข้อมูลจากเว็บไซต์ : https://kku.world/7-w4d

รายงานสถานการณ์จากทั่วโลก จำนวนผู้ติดเชื้อ ไวรัสโคโรนาสายพันธุ์ใหม่ 2019 (Novel Coronavirus;2019-nCoV) หรือ COVID-19 ประจำวันที่ 21 มีนาคม 2563 ดังนี้

ประเทศ / ภูมิภาค Country / Territory	จำนวนผู้ติดเชื้อไวรัส Confirmed	จำนวนผู้ติดเชื้อรายใหม่ New Cases	ผู้เสียชีวิต Deaths	จำนวนผู้เสียชีวิตเพิ่ม New Deaths	รักษาหาย Recoveries
จีน Mainland China	81,008	+41	3,255	+7	71,740
อิตาลีItaly	47,021	+5,986	4,032	+627	5,129
สเปนSpain	21,510	+3,433	1,093	+262	1,588
เยอรมนี Germany	19,848	+4,528	68	+24	180
อิหร่าน Iran	19,644	+1,237	1,433	+149	6,745

ประเทศ / ภูมิภาค Country / Territory	จำนวนผู้ติดเชื้อไวรัส Confirmed	จำนวนผู้ติดเชื้อรายใหม่ New Cases	ผู้เสียชีวิต Deaths	จำนวนผู้เสียชีวิตเพิ่ม New Deaths	รักษาหาย Recoveries
จีน Mainland China	81,008	+41	3,255	+7	71,740
อิตาลีItaly	47,021	+5,986	4,032	+627	5,129
สเปนSpain	21,510	+3,433	1,093	+262	1,588
เยอรมณี Germany	19,848	+4,528	68	+24	180
อิหร่าน Iran	19,644	+1,237	1,433	+149	6,745
สหรัฐอเมริกา United States	19,429	+5,640	257	+50	147

- จากภาพแสดงจำนวนผู้ที่ติเชื้อและเสียชีวิตจาก Covid-19 ในแต่ละประเทศ จะเห็นได้ว่าประเทศที่มีผู้ติดเชื้อและผู้เสียชีวิตมากที่สุด ไม่มีประเทศใดเลยที่มีรายชื่อในกราฟ 5 location แรกที่มีการ Tweet หรือ Retweet #covid

สรุปผลที่ได้จากกราฟเกี่ยวกับโรค **Ebola**

ภาพแสดงจำนวนผู้ที่ติดเชื้ออีโบล่าและจำนวนผู้เสียชีวิต ในประเทศคองโก จากองค์การอนามัยโลก (world health organization)

- จากเว็ปไซต์ : https://www.who.int/emergencies/diseases/ebola/drc-2019

- เมื่อทำการนำกราฟที่แสดงผล 5 location แรกที่มีการ Tweet หรือ Retweet #ebola มาเปรียบเทียบกับข้อมูลจำนวนผู้เสียชีวิตจากโรคอีโบล่าในคองโก จะเห็นได้ว่าผู้คนในคองโกมีการพูดถึง โรค Ebola ผ่าน Twitter ซึ่งก็เป็นไปตามสถานการณ์การระบาดของโรคอีโบล่าในประเทศ ซึ่งแสดงให้เห็นถึงความสนใจของผู้คนในคองโก เกี่ยวกับปัญหาโรคอีโบล่า ที่เกิดขึ้นประเทศของตน

จากข้อมูลที่แสดงด้านบนทั้งหมด เราอาจจะสรุปได้ว่า

1. ประเทศที่มีการ Tweet หรือ Retweet เกี่ยวกับโรค Covid มากที่สุด คือประเทศที่มีจำนวนผู้ที่ติดเชื้อหรือเสียชีวิตน้อยกว่าประเทศที่ไม่ได้ทำการ Tweet หรือ Retweet

2. อาจจะสรุปได้อีกอย่างว่าผู้คนในประเทศดองโก มีความสนใจและใส่ใจเกี่ยวกับปัญหาโรคอีโบล่าในประเทศของตนเป็นอย่างมาก

3. การที่ผู้คนออกมา Tweet หรือ Retweet นั้น อาจจะแสดงให้เห็นถึงความใส่ใจเกี่ยวกับข่าวสารหรือปัญหาที่กำลังเกิดขึ้นในสังคมโลกหรือประเทศของตนเอง

กราฟแท่งแสดงจำนวนคำว่า ebola, covid ใน Reddit, Twitter ¶

- โดยหารด้วยจำนวน row

In [153]:
```
# set ให้เป็น Lower เพื่อง่ายในการค้นหา

# Reddit
reddit_covid_en.title = reddit_covid_en.title .str.lower()
reddit_ebola_en.title = reddit_ebola_en.title .str.lower()

# Twitter
twitter_covid_th.text =  twitter_covid_th.text .str.lower()
twitter_covid_en.text =  twitter_covid_en.text .str.lower()
twitter_ebola_en.text =  twitter_ebola_en.text .str.lower()
```

count covid, ebola

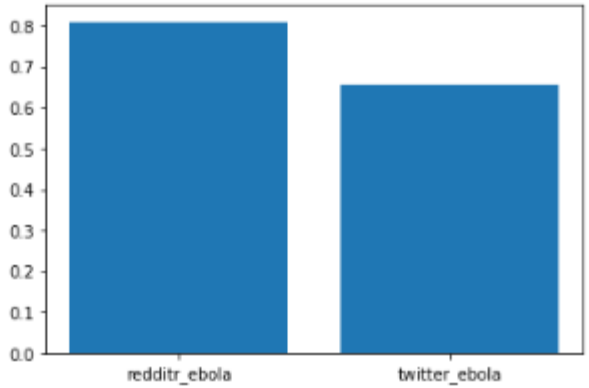
In [10]:
```
# count covid
count_reddit_covid     = reddit_covid_en.title.str.count('covid').sum()
count_twitter_covid_en = twitter_covid_en.text.str.count('covid').sum()
count_twitter_covid_th = twitter_covid_th.text.str.count('covid').sum()


# count ebola
count_reddit_ebola  = reddit_ebola_en.title.str.count('ebola').sum()
count_twitter_ebola = twitter_ebola_en.text.str.count('ebola').sum()
```

In [11]:
```
# จำนวน column, row
m,n = reddit_ebola_en.shape
j,k   = twitter_ebola_en.shape
```

In [161]:
```
plt.bar(['redditr_ebola', 'twitter_ebola'],  [count_reddit_ebola/m , count_twitter_ebola/j])
```

Out[161]: <BarContainer object of 2 artists>

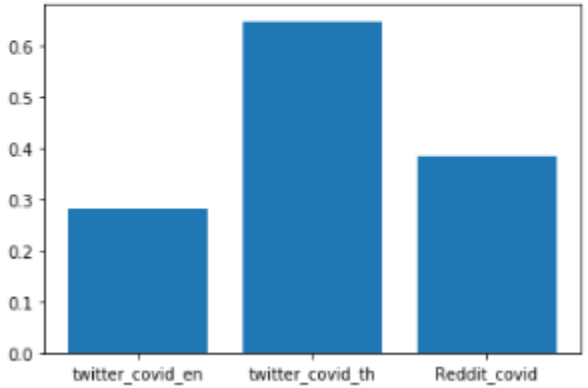


- จะเห็นได้ว่าใน Reddit มีการพูดถึง Ebola มากกว่า

In [13]:
```
a,b = reddit_covid_en.shape
c,d = twitter_covid_en.shape
e,f = twitter_covid_th.shape
```

In [14]:
```
plt.bar( ['twitter_covid_en', 'twitter_covid_th', 'Reddit_covid'],  [count_twitter_covid_th/e , count_twitter_covid_en/c, count_
```

Out[14]: <BarContainer object of 3 artists>



- จะเห็นได้ว่า ใน Twitter มีการพูดถึง Covid มากที่สุด

Thai text preprocessing using Python

- เราจะทำการวิเคราะห์ข้อความที่ถูกทวิตจาก Twitter และ Reddit ว่าคำใดถูกใช้มากที่สุด

1. ทำการดึงข้อความ text จากตาราง

In [15]:
```
# เก็บข้อความไว้ใน List
# Reddit
list_reddit_covid_en  = [i for i in reddit_covid_en.title]
list_reddit_ebola_en  = [i for i in reddit_ebola_en.title]

#Twitter
list_twitter_covid_th = [i for i in twitter_covid_th.text]
list_twitter_covid_en = [i for i in twitter_covid_en.text]
list_twitter_ebola_en = [i for i in twitter_ebola_en.text]
```

In [16]:
```
# นำคำใน List แต่ละคำมาต่อกันเป็น sring

#Reddit
text_reddit_covid_en  = ''.join(item for item in list_reddit_covid_en)
text_reddit_ebola_en  = ''.join(item for item in list_reddit_ebola_en)


#Twitter
text_twitter_covid_th = ''.join(item for item in list_twitter_covid_th)
text_twitter_covid_en = ''.join(item for item in list_twitter_covid_en)
text_twitter_ebola_en = ''.join(item for item in list_twitter_ebola_en)
```

In [17]:
```
# ลองแสดงข้อมูล
text_twitter_covid_en
```

Out[17]:
```
'a word of caution from mega star chiranjeevi garu stay safe covid covidindia because every red light eventually turns greenenc
ouraging words from mcconaughey covid coronavirus  day  in quarantine coronavirus londonlockdown covid quarantinewhen you exper
ience high body temperatures fever headache runny nose coughing and sneezing then difficult in br… here a few ideasobservations
i had about covid inhospital feel free to add your ideas n medtwitter… a pandemic disease has spreadis nepal out of this worldc
ovidstaysafequarantinelifewhat parents of deaf children need to know before their childs school closes  please share and help g
et this in… covidthis is the time to our homeopaths friends and researchers to come forwardhomeopathy is also efficient in… ret
```

ทำการนับคำซ้ำ และแสดงผลแบบภาพ

- import libary ที่ต้องใช้

In [74]:
```
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from PIL import Image
```

In [19]:
```
import nltk
```

In [20]:
```
nltk.download()
```

```
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
```

Out[20]:
```
True
```

In [21]:
```
from nltk.corpus import stopwords
set(stopwords.words('english'))
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

In [22]:
```
# Importing necessary library
import pandas as pd
import numpy as np
import nltk
import os
import nltk.corpus
```

In [23]:
```
# importing word_tokenize from nltk
from nltk.tokenize import word_tokenize
# ทำการตัดคำ

# Reddit
token_reddit_covid_en = word_tokenize(text_reddit_covid_en)
token_reddit_ebola_en = word_tokenize(text_reddit_ebola_en)


# Twitter
token_twitter_covid_th = word_tokenize(text_twitter_covid_th)
token_twitter_covid_en = word_tokenize(text_twitter_covid_en)
token_twitter_ebola_en = word_tokenize(text_twitter_ebola_en)
```

In [24]:
```
# finding the frequency distinct in the tokens
# Importing FreqDist library from nltk and passing token into FreqDist
from nltk.probability import FreqDist

#m ทำการนับคำซ้ำ
# Reddit
fdist_reddit_covid_en = FreqDist(token_reddit_covid_en)
fdist_reddit_ebola_en = FreqDist(token_reddit_ebola_en)


#Twitter
fdist_twitter_covid_th = FreqDist(token_twitter_covid_th)
fdist_twitter_covid_en = FreqDist(token_twitter_covid_en)
fdist_twitter_ebola_en = FreqDist(token_twitter_ebola_en)
```

In [81]:
```
# To find the frequency of top  words

#Reddit
fdist_reddit_covid_en1 = fdist_reddit_covid_en.most_common(20)
fdist_reddit_ebola_en1 = fdist_reddit_ebola_en.most_common(20)

# Twitter
fdist_twitter_covid_th1 = fdist_twitter_covid_th.most_common(100)
fdist_twitter_covid_en1 = fdist_twitter_covid_en.most_common(20)
fdist_twitter_ebola_en1 = fdist_twitter_ebola_en.most_common(20)
```

คำที่ถูกใช้งานมากที่สุดใน Reddit หัวข้อ Covid

In [26]:
```
fdist_reddit_covid_en1
```

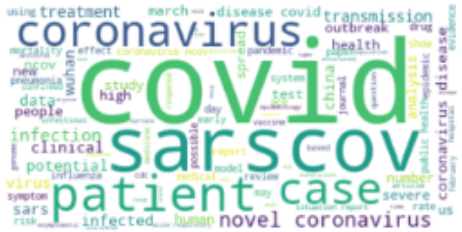
Out[26]:
```
[('of', 534),
 ('the', 377),
 ('and', 283),
 ('in', 263),
 ('covid', 230),
 ('for', 193),
 ('to', 189),
 ('a', 179),
 ('coronavirus', 168),
 ('on', 109),
 ('with', 92),
 ('from', 76),
 ('is', 75),
 ('sarscov', 67),
 ('novel', 65),
 ('patients', 57),
 ('disease', 53),
 ('are', 51),
 ('that', 50),
 ('cases', 47)]
```

คำที่ถูกใช้งานมากที่สุดใน **Reddit** หัวข้อ **Covid**

In [26]:
```
fdist_reddit_covid_en1
```

Out[26]:
```
[('of', 534),
 ('the', 377),
 ('and', 283),
 ('in', 263),
 ('covid', 230),
 ('for', 193),
 ('to', 189),
 ('a', 179),
 ('coronavirus', 168),
 ('on', 109),
 ('with', 92),
 ('from', 76),
 ('is', 75),
 ('sarscov', 67),
 ('novel', 65),
 ('patients', 57),
 ('disease', 53),
 ('are', 51),
 ('that', 50),
 ('cases', 47)]
```

In [38]:
```
# lower max_font_size, change the maximum number of word and lighten the background:
text = text_reddit_covid_en
wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white").generate(text)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

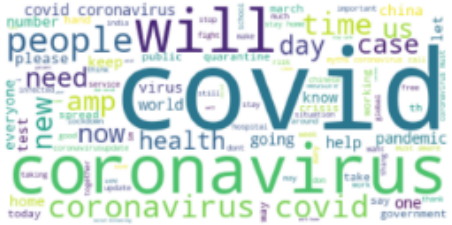


คำที่ถูกใช้งานมากที่สุดใน Twitter หัวข้อ Covid

In [27]:
```
fdist_twitter_covid_en1
```

Out[27]:
```
[('the', 933),
 ('covid', 774),
 ('to', 770),
 ('of', 519),
 ('in', 434),
 ('a', 419),
 ('and', 404),
 ('is', 368),
 ('for', 334),
 ('coronavirus', 299),
 ('are', 259),
 ('you', 234),
 ('this', 210),
 ('on', 186),
 ('we', 178),
 ('’', 167),
 ('i', 162),
 ('from', 161),
 ('that', 156),
 ('be', 155)]
```

In [80]:
```
# Lower max_font_size, change the maximum number of word and lighten the background:
text = text_twitter_covid_en
wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white").generate(text)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



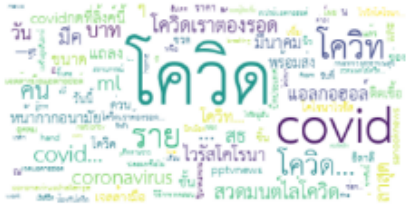
คำที่ถูกใช้งานมากที่สุดใน Twitter (ภาษาไทย) หัวข้อ Covid

In [28]:
```
fdist_twitter_covid_th1
```

Out[28]:
```
[('โควิด', 906),
 ('covid', 329),
 ('ราย', 121),
 ('โควิท', 115),
 ('โควิด…', 83),
 ('คน', 61),
 ('…', 59),
 ('covid…', 52),
 ('บาท', 41),
 ('ml', 41),
 ('มีค', 39),
 ('coronavirus', 39),
 ('สวดมนต์ไล่โควิด', 38),
 ('โควิดเราต้องรอด', 37),
 ('สธ', 35),
 ('ล่าสุด', 33),
 ('วัน', 32),
 ('แอลกอฮอล์', 30),
 ('ไวรัสโคโรนา', 28),
 ('มีนาคม', 26)]
```

In [82]:
```
# เทียบเป็น Data dict
word_dict = dict((x,y) for (x,y) in fdist_twitter_covid_th1)
```

In [83]:
```
wordcloud = WordCloud(font_path='D:/2-2562/Data Engineering/Project/font/THSarabunNew.ttf',
                      background_color ='white',max_words=100).fit_words(word_dict)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```



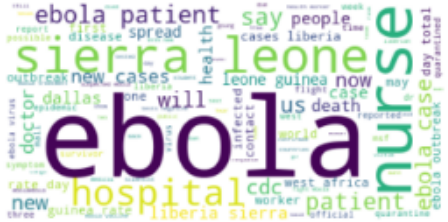
จากกราฟทั้ง 3 กราฟด้านบน แสดงให้เห็นว่า #covid ในไทยกับต่างประเทศมีการพูดถึงที่มีความแตกต่างกันเล็กน้อยเช่น ในต่างประเทศจะมีคำว่า covid, coronavirus, patients, people แต่ในไทยจะเห็นได้ว่ามี สวดมนต์ไล่โควิด, แอลกอฮอล์, หน้ากากอนามัย, โควิดเราต้องรอด นั่นอาจจะแสดงให้เห็นถึงวิกฤติการขาดแคลนหน้ากากอนามัย แอลกอฮอล์ล้างมือในไทย และการรับมือเกี่ยวกับโรคโควิดของรัฐบาลไทยที่มีนโยบายให้ประชาชน "สวดมนต์ไล่โควิด"

คำที่ถูกใช้งานมากที่สุดใน Reddit หัวข้อ ebola

In [29]:
```
fdist_reddit_ebola_en1
```

Out[29]:
```
[('ebola', 572),
 ('in', 387),
 ('the', 370),
 ('to', 367),
 ('of', 325),
 ('for', 153),
 ('and', 147),
 ('a', 142),
 ('is', 130),
 ('from', 90),
 ('cases', 88),
 ('with', 86),
 ('on', 85),
 ('new', 80),
 ('who', 79),
 ('sierra', 79),
 ('at', 77),
 ('leone', 77),
 ('liberia', 70),
 ('patient', 68)]
```

In [79]:
```
# Lower max_font_size, change the maximum number of word and lighten the background:
text = text_reddit_ebola_en
wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white").generate(text)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```

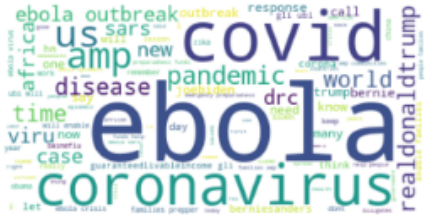


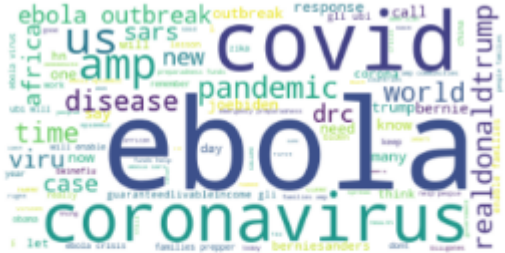
ค่าที่ถูกใช้งานมากที่สุดใน Twitter หัวข้อ ebola

In [30]:
```
fdist_twitter_ebola_en1
```

Out[30]:
```
[('the', 1012),
 ('ebola', 882),
 ('to', 571),
 ('in', 459),
 ('of', 425),
 ('a', 405),
 ('and', 404),
 ('is', 397),
 ('for', 250),
 ('it', 223),
 ('covid', 213),
 ('’', 212),
 ('coronavirus', 208),
 ('we', 207),
 ('that', 203),
 ('this', 203),
 ('i', 197),
 ('from', 179),
 ('you', 161),
 ('are', 150)]
```

In [84]:
```
# lower max_font_size, change the maximum number of word and lighten the background:
text = text_twitter_ebola_en
wordcloud = WordCloud(max_font_size=100, max_words=100, background_color="white").generate(text)
plt.figure()
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show()
```





เนื่องจาก ข้อมูลโรคอิโบล่าที่ดึงจาก Reddit นั้นเป็นข้อมูลตั้งแต่ปี 2014 แต่ข้อมูลโรคอิโบล่าที่ดึงจาก Twitter เป็นข้อมูลในปี 2020 จะเห็นได้ว่าคำที่ถูกใช้มากที่สุดนั้นแตกต่างกัน เช่น ใน #ebola ในปี2020 จะมีการพูดถึงโรค covid, coronavirus แต่ในปี 2014 จะพูดถึงแต่โรค ebola

Thank You

ขอบคุณครับ