Huong Pham (824992682)
Final Project Report
Due 12/18/2020

SENTIMENT ANALYSIS VISUALIZATION USING PYTHON PROGRAMING
-- FINAL REPORT--

Project Background

As of 12/17/2020, the U.S. marked the milestone of more than 300,000 coronavirus deaths since the beginning of the pandemic. Almost every two months we have 50,000 more deaths in the U.S. and the mortality rate is getting even higher this month. This virus is deadly, not to mention many devastating social and economic issues it has caused such as unemployment, stress on supply chains, events, business and school closure, working from home, and reduced consumer activities, etc.

Many seniors said that they have never experienced anything like this in their lives, and neither did I. The COVID-19 pandemic with its bad effects is what triggered me to do a sentiment analysis on COVID-19 tweets. Millions of Americans use Twitter to disseminate news and information on social economic events every day. This is one of the best social media to find out what people concern about and how they react to the COVID-19 pandemic.

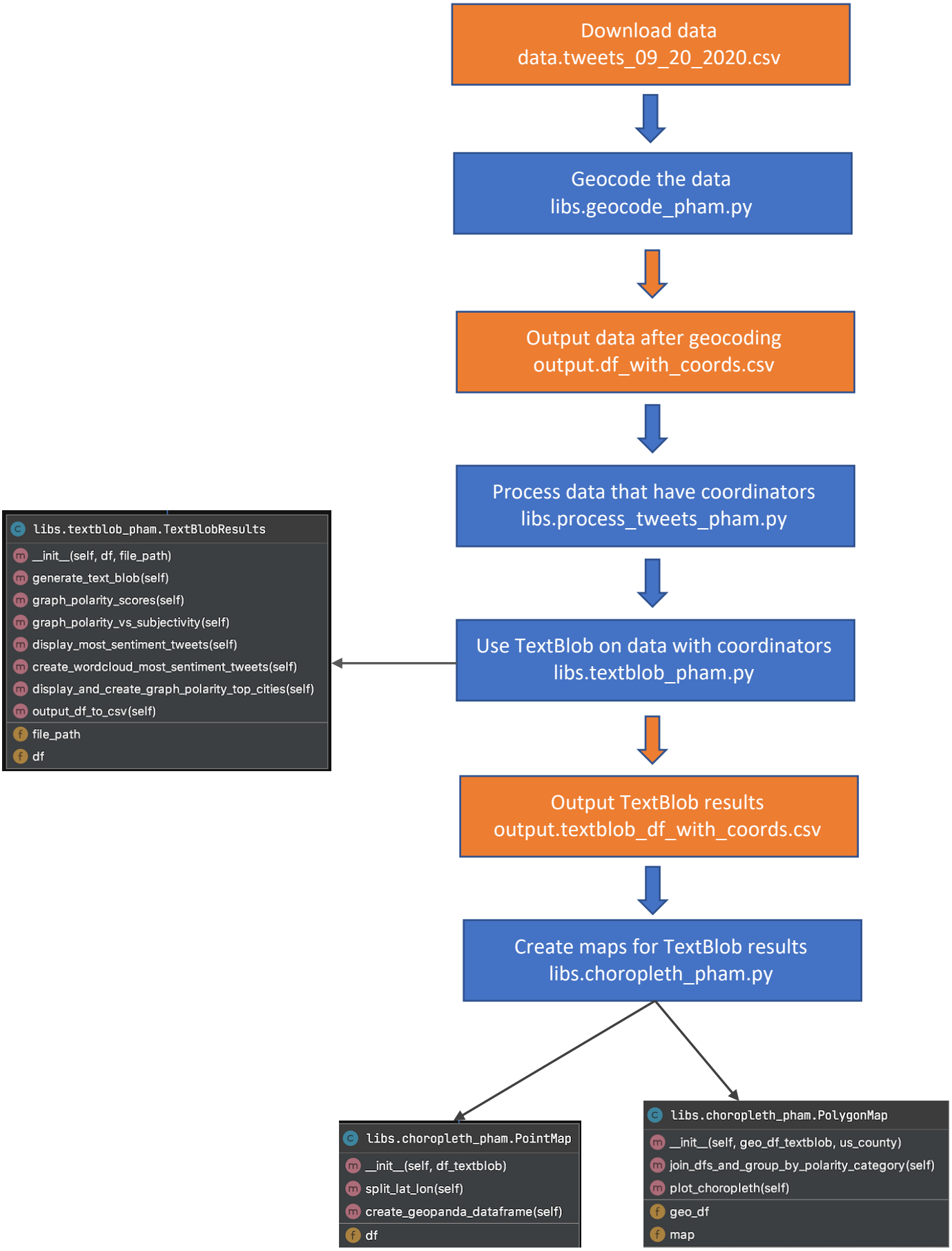
I collected tweets data from the Panacea Lab (data source). My dataset includes all the tweets that relate to COVID-19 all over the world on 09/20/2020. The data I downloaded from Panacea Lab had only one ID column at first. I used Hydrator software to extract Tweet texts and the rest of other metadata. My data originally has 204,000 tweets in English. I pared it down to tweets that have key word "CA" in the user_location column and my dataset went down to 3,600 tweets as a result.

Problem statements and GIS-related tasks

My goal in this final project is to apply the Python and GIS skills I learned from the GEOG-582 class to visualize my COVID-19 Tweets sentiment analysis.

Because of the privacy reason, The Panacea Lab do not share the coordinates of the tweets. And this is a disadvantage for me to present my analysis results. My solution is to use a geocoding service to geocode the tweet user locations and collect longitudes and latitudes for the tweets. The coordinates will then allow me to put locations of the tweet users on a map and create choropleth maps. Choropleth maps assign different colors for different range of sentiment index values, which can show us which counties have higher total number of tweets, which counties have more positive tweets, more negative tweets, or more neutral tweets than others.

Program design and code implementation


Download data
data.tweets_09_20_2020.csv
Geocode the data
libs.geocode_pham.py
Output data after geocoding
output.df_with_coords.csv
Process data that have coordinators
libs.process_tweets_pham.py
Use TextBlob on data with coordinators
libs.textblob_pham.py
Output TextBlob results
output.textblob_df_with_coords.csv
Create maps for TextBlob results
libs.choropleth_pham.py
libs.textblob_pham.TextBlobResults
__init__(self, df, file_path)
generate_text_blob(self)
graph_polarity_scores(self)
graph_polarity_vs_subjectivity(self)
display_most_sentiment_tweets(self)
create_wordcloud_most_sentiment_tweets(self)
display_and_create_graph_polarity_top_cities(self)
output_df_to_csv(self)
file_path
df
libs.choropleth_pham.PointMap
__init__(self, df_textblob)
split_lat_lon(self)
create_geopanda_dataframe(self)
df
libs.choropleth_pham.PolygonMap
__init__(self, geo_df_textblob, us_county)
join_dfs_and_group_by_polarity_category(self)
plot_choropleth(self)
geo_df
map

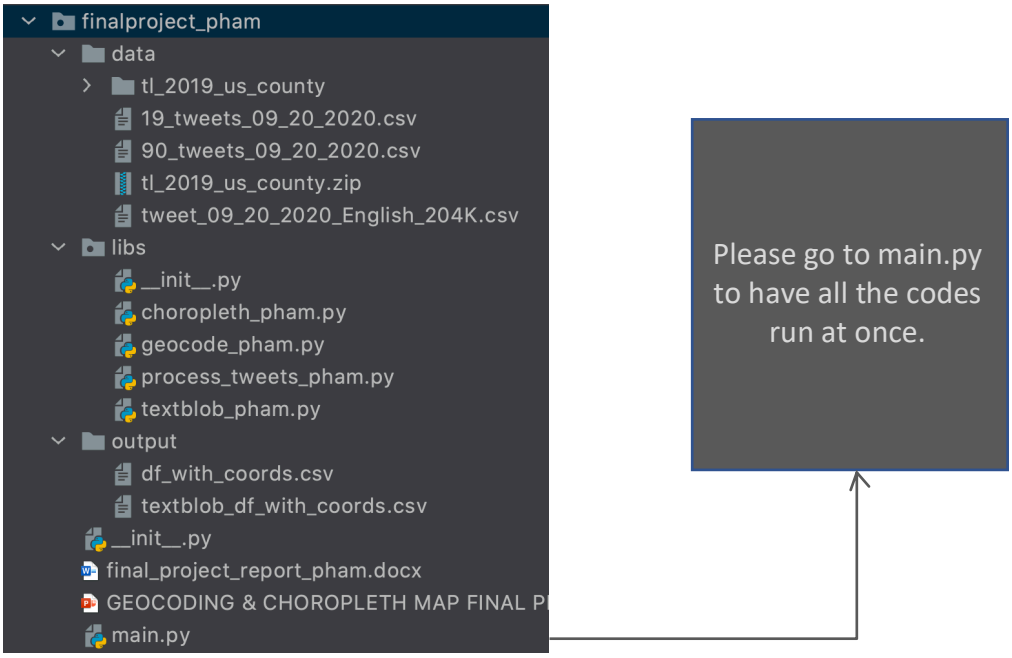
(i) Geocoding tweets by using the OpenStreetMap Nominatim service [geocode_pham.py] and output the results to a csv file [df_with_coords.csv]

(ii) Process tweet texts in the output csv file by using Regular Expressions [process_tweet_pham.py] and applying TextBlob method for sentiment analysis [textblob_pham.py]

(iii) Present TextBlob results by making Word Clouds, bar charts, and graphs to show the most common topics for the COVID-19 tweets on 09/20/2020, the most common words for the positive/negative tweets on that day, the relationship between polarity and subjectivity index, the top cities in California that have higher number of tweets and components of their polarity labels. [textblob_pham.py]

(iv) [choropleth_pham.py] Use GeoPandas library to put all user locations on a same map (Point geometry) and create Choropleth maps (polygon geometry) for CA counties' polarity label counts [choropleth_pham.py]. I downloaded a US county map from census website, then created a GeoPanda DataFrame from it, called us_county. My next task is to populate this map with colors to present the tweets numbers. In order to accomplish that, I do the following:

- Join the TextBlob GeoPanda DataFrame to us_county
- Group the new GeoPanda DataFrame by "GEOID" and polarity category columns
- Group into positive, negative, neutral labels (got three new dataframes)
- Merge three new DataFrames back to us_county
- Create a new column named 'total' to sum the total of tweets for each county
- Use the final us_county to create Choropleth maps



Major accomplishments

(i) By using the OpenStreetMap Nominatim service, I got 3,400 tweets geocoded from the total of 3,600 tweets.

```
America, you're not just poorer - you're dumber.","Oakland, CA","37.8044557, -122.2713563"
187802,Mon Sep 21 03:52:10 +0000 2020,Do not take Trump's vaccines. He does not care if you die https://t.co/oL58WromXz,"San Francisc
187881,Mon Sep 21 03:53:18 +0000 2020,"Emmys: Covid-19 'Test;' Jimmy Kimmel Mocks MAGA Rallies, Cracks Jokes About Russian Interferer
187921,Mon Sep 21 03:52:52 +0000 2020,"Not good, not good. https://t.co/XboYoX12Au","Los Angeles, CA","34.0536909, -118.242766"
187976,Mon Sep 21 03:54:02 +0000 2020,"CDC says coronavirus spreads mainly in the air, through respiratory aerosols and droplets http
187996,Mon Sep 21 03:53:48 +0000 2020,"NINE months in... CNN: Updated CDC guidance acknowledges coronavirus can spread through the ai
https://t.co/zfNDIb7LRq

via @GoogleNews","West Hollywood, CA","34.0923014, -118.3692894"
188078,Mon Sep 21 03:54:37 +0000 2020,"Now Reading: CNN: Updated CDC guidance acknowledges coronavirus can spread through the air.
https://t.co/XhfDqfjVX6

via @GoogleNews","San Jose, CA","37.3361905, -121.890583"
188081,Mon Sep 21 03:55:04 +0000 2020,"20 new cases of COVID-19 recorded in Saskatchewan push provincial total past 1,800
 https://t.co/8cDQsHbXkW","Ontario, CA","50.000678, -86.000977"
```

(ii) All the tweet user locations are put on a same map by creating a Point geometry and using GeoPandas library (Figure 1)

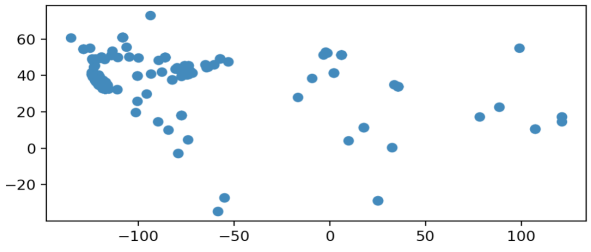


Figure 1

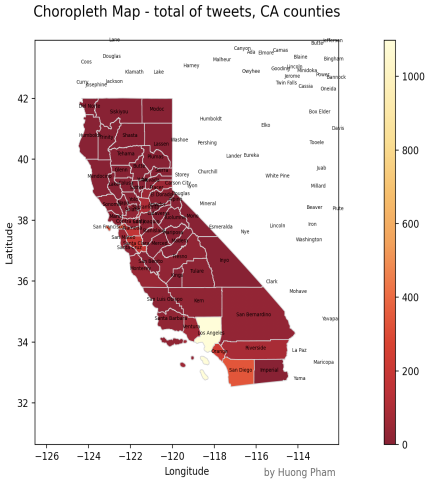


Figure 2

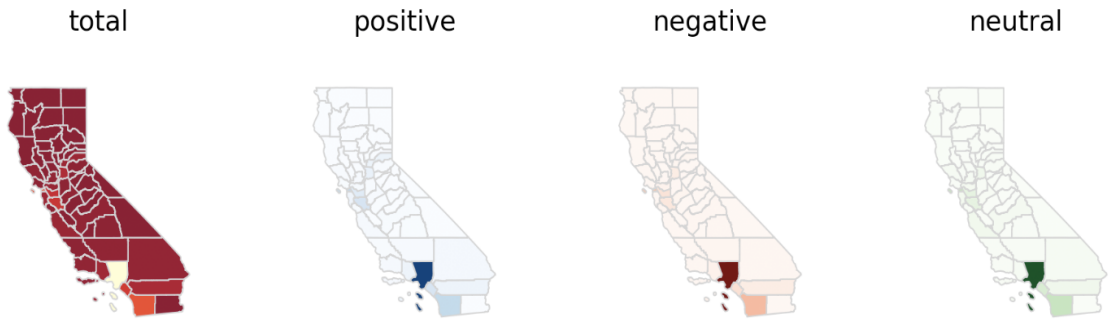


Figure 3

(i) Choropleth maps are created (Figure 2 and 3). They show that Los Angeles, San Diego, San Francisco, and Sacramento are those that have higher number of tweets.

(ii) The results match with this graph I created for the top cities in California (Figure 4)

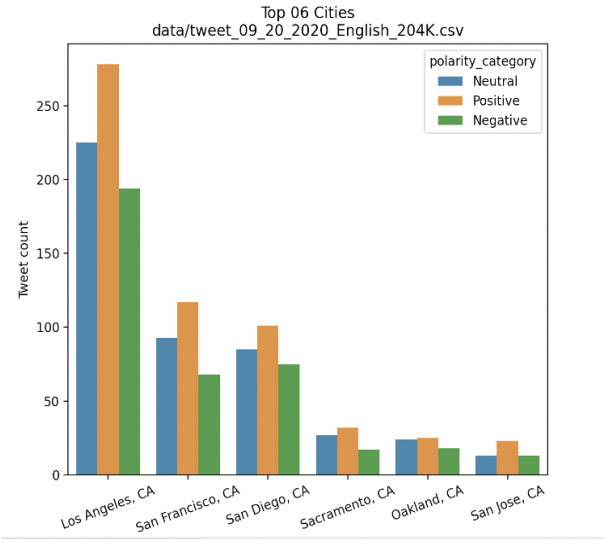


Figure 4

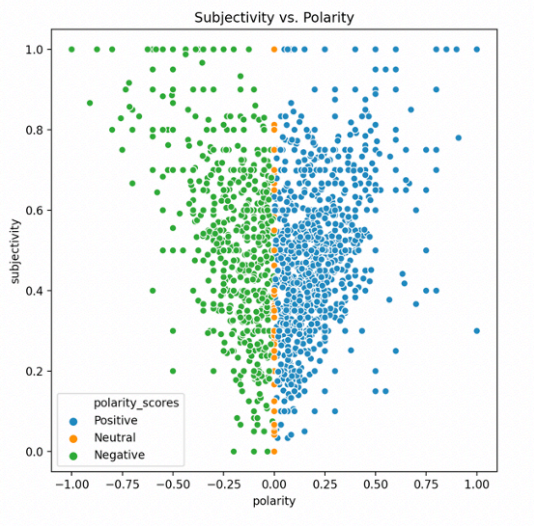


Figure 5

(iii) Figure 5 is a graph about the relationship between polarity and subjectivity index. The chart shows that when subjectivity index increases, the polarity index becomes more diverse. This makes sense because when we express our opinion, our statement often reflects some sentiment (positive or negative), and when we make a statement about a fact, our statement has little or no sentiment.

Technical challenges

(i) Some locations that have key word “CA”, “California” in the user_location column but actually are not located in California. “CA”, ”California” might be part of street or road name of the locations in other countries or other states in the U.S.

(ii) The accuracy of the TextBlob sentiment analysis method depends on the key words that the Panacea Lab used to collect tweets relating to COVID-19. It also depends on the data cleaning process which require a lot of time. The more I clean my data the higher sentiment analysis accuracy I can get.