PRINCIPLES OF SOCIAL MEDIA AND DATA MINING (CIS 600)

TERM PROJECT PROPOSAL

Name of Project

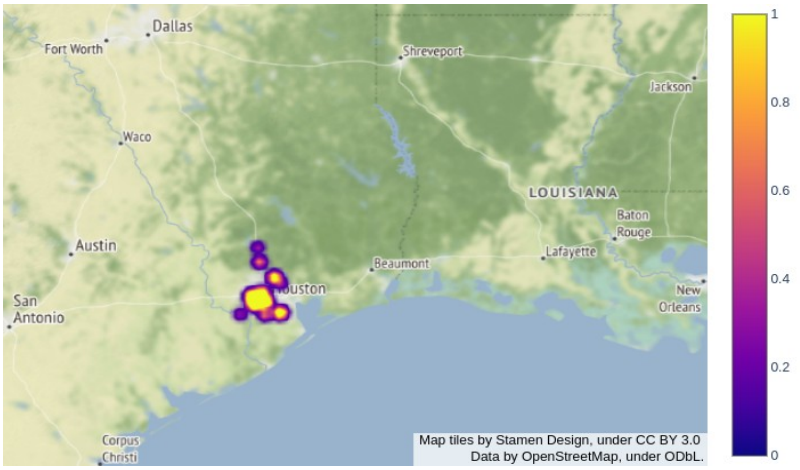
"Tweets patterns in the USA"

Group members

1. Danylo Honcharov (SUID 567804880)
2. Shwetha Koushik (SUID 317264697)
3. Yuchao Xu (SUID 625378449)
4. Vindhya Ravi Prakash (SUID 777803438)

Project Idea and features

This project mainly focuses on the analysis of the tweets based on the geolocation and performing analysis on the various features that are extracted from the tweets. Some of the patterns that are being identified in this project are; if influential people on Twitter tend to post lengthy tweets or not, Finding the trending topics in a particular region based on the hashtags, Analyzing if a particular tweet is positive or negative with regards to a particular entity (Sentiment Analysis) and recognizing trending topics in a particular region using entities. Basically, the main idea of this project is to discover interesting patterns between tweets/authors of the tweets and their locations, considering entities/hashtags. For this project, we're mainly going to use already collected dataset with geotagged tweets from the TheFollowHashtag.com. The results are represented in the form of plots and patterns. For example, tweets with entity "Memorial Hermann" can be visualized in the following way:



Significance of the idea

The primary objective of this project is to discover, understand and analyze important patterns about users with respect to different regions in the US. This is further visualized to draw vital conclusions. These conclusions can help advertising agencies to determine what kind of products can be best sold in which regions. Further sentiment analysis is applied in order to find out the impact of the tweets in a specific region. This might help in deducing how to improve that region. For example: Improving the economic growth or the technological capabilities.

Dataset description

The dataset that is being used in this project is taken from TheFollowHashtag.com, which contains the geotagged tweets. This twitter dataset has a database of 200,000 USA geolocated Tweets with a size of 47 Mb and the retweets are excluded. For each tweet there are some additional user information (For example, number of followers, languages, number of retweets of tweet and so on). These datasets are collected during the 48 hours timeframe from the Twitter Streaming API. The dataset can be downloaded in the form of 4 excel files using the link mentioned below.

http://followthehashtag.com/datasets/free-twitter-dataset-usa-200000-free-usa-tweets/

Implementation details

This project will be implemented using python language and the following python packages.

Name of package	Description
Pandas	Package for tabular data processing.
Flair	State-of-the-art Named Entity Recognition library.
TextBlob	Package for sentiment/objectivity analysis of the sentences.
Plotly	Package for the maps plotting (heatmaps of the USA, etc)
Matplotlib	Package for the plotting histograms and other plots.
reverse_geocoder	Package for the fast-offline reverse geocoding.
Jupyter notebook	IDE for the Python, suited for convenient and interactive data analysis.

Individual tasks for each team member

Team member	Task Description	Allotted time
Danylo Honcharov	Discovering patterns for the entire country and for every region in the country. For example, one of the most frequently discussed tweet topics around US is related to hiring.	Nov 2nd to Nov 16th
Danylo Honcharov	Determining the distinctive words of a region	Nov 16th to Nov 23rd
Shwetha Koushik	Discovering most popular hashtag/entity in a region. Finding out the hashtag used by people in a region and determining the most frequently used hashtag for that region. For example, #ILoveNY could be the most popular hashtag in New York while the most popular hashtag in california could be #travel	Nov 2nd to Nov 13th
Shwetha Koushik	Discovering the relationship between the number of followers and friends for a region. For example, determining if the tweets that are generated by users from popular regions have a greater number of followers.	Nov 16th to Nov 23rd
Yuchao Xu	Finding out the length of the tweet for a region. Discovering patterns from that: Relationship between the length of tweets and their number of followers, their retweets. For example, the people who have more followers have a longer tweet when	Nov 2nd to Nov 16th

	compared to those with a smaller number of followers.	
Vindhya Ravi Prakash	Identifying the language of the tweet with respect to a region. For example, identifying the popular language used in a region. This helps in identifying the distribution of people across the US.	Nov 2nd to Nov 16th
Vindhya Ravi Prakash & Yuchao Xu	Performing sentiment Analysis to determine if a tweet is positive or negative. Also, recognizing which region has major positive/ negative tweets.	Nov 16th to Nov 26th
All the team members	Documentation, presentation and representing the output using pie chart, heat maps and tables.	Always