Twitter Data Analysis By University

Glory Scheel

Big Data Platforms

December 11,2020

Presentation Overview

Executive Summary

Methodology and Data Overview

Conclusions and Recommendations

Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing.

Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing.

Lorem Ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing.

Executive Summary

Beginning with a 2 Terabyte file of twitter user data, I used Google Cloud Platform to filter and clean this data and grab tweets only relevant to the four schools: University of Chicago, Yale University, Florida State College and University of Florida.

This data was exported to csv files and then continued to be processed on my local machine. Upon completing some further data cleaning to achieve a more powerful dataset analysis began.

In order to identify the most influential users a weighted “influence score” was calculated for each user University of Florida to show the users with the highest influence score among the universities in this analysis.

To explore tweet and user location I created a density map. This map did not show too noticable of differences among universities, locations were slightly different and slightly more populated where the universities were located.

Users who tweeted about University of Chicago generally showed a uniform distribution of tweets let out at every time frame which was not the case for the other universities. They also showed to have the highest amount of median followers among the universities and the least similar tweets to one another.

Data Description

In this presentation we will be discussing the use of a dataset that was pulled from Twitter resulting in a 2 TB file which was condensed by filtering through the data and only pulling out tweets that had text that related to the four chosen universities: University of Chicago, Yale, Florida State University and University of Florida. Before filtering through the data all redundant and irrelevant columns were removed to make for easier processing. This dataset was then exported as four separate csv's, by university, and further processed on my local machine.

After Further cleaning the data consisted of columns:

1. "Id": Unique ID number for each user
2. "Created_at": When the tweet was created
3. "Quote_count": The amount of quoted tweets the user has done
4. "Favorite_count": The amount of favorites on the tweet
5. "Favorited": Whether the tweet has been favorited or not
6. "Reply_count": How many replies the tweet received
7. "Retweet_count": How many retweets the tweet received
8. "Retweeted": Whether or not the tweet was retweeted
9. "Location": The location of the tweet
10. "Followers": How many followers the user has
11. "Total_tweets": How many total tweets the user has conducted

Methodology

Determining Similarity Among Tweets

To measure similarity between tweets, within each university, I first needed to address the large datasets being used. Because of the size of these datasets I subset each of the university datasets and ran a simhash function I created on each dataset.

Distinguishing Uchicago

To understand any differences in the data among universities that may exist I grouped the data by university and produced a visualization. The visualization showed the median values for each of the universities. (Figure 2)

Location

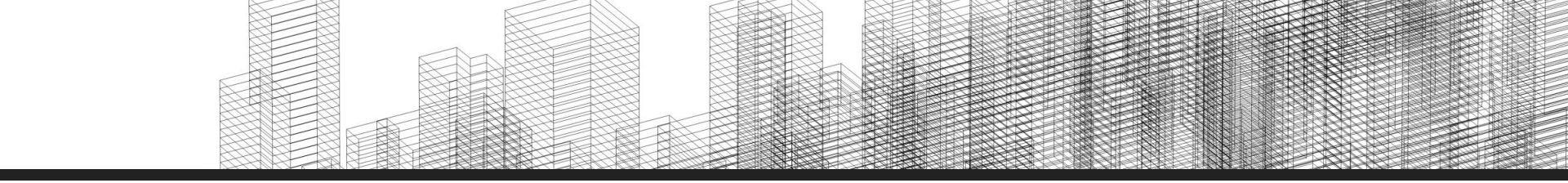
I used the variable location and a package called geopy to extract latitude and longitude coordinates for tweets when the location was valid. I then used these latitude and longitudes with a package called basemap to map the tweet densities on a map of the globe. (Figure 1)

Identifying Influential Users

In order to identify the most influential twitter users for each university I created a “influence score”. This score used a weighted linear combination of the amount of retweets, total tweets, percent of tweets that are about that university and followers. I then identified the top 5 influential twitter users for each university

Tweet Timelines

I created a histogram of the density of the tweets by time for each university. The x-axis consisted of when the tweet was created and the y-axis was how many tweets were done at that time.



Conclusions and Recommendations

While the location variable did not show very much of a difference between the different universities, besides the increased concentration of tweets surrounding universities, it seems the other variables had more complex information to share.

The similarity measures showed Uchicago had the most disimilar tweet texts, by a large amount.

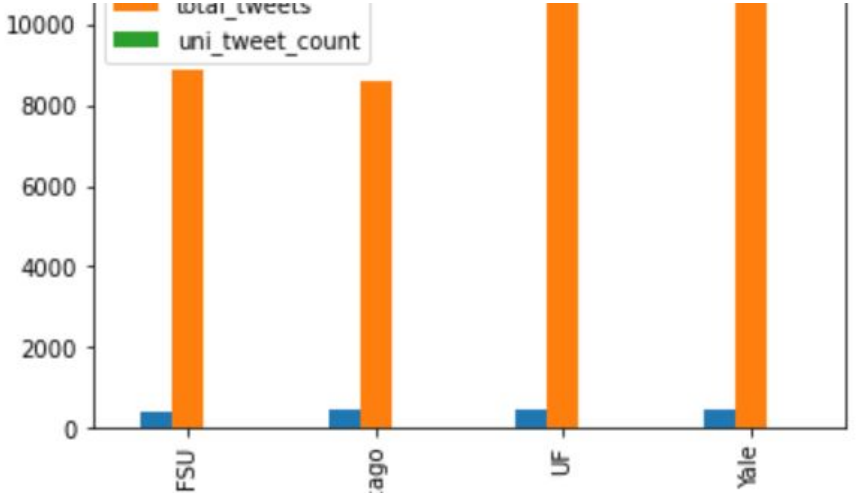
Uchicago also showed the most uniformly distributed tweet densities throughout time. Schools like Yale and University of Florida showed big peaks and valleys, this could be due to sport off and on seasons since these are very sports oriented schools.

The Twitter users that had the highest influence scores were University of Florida users. This could be partially due to their large volume of total tweets per each user.

Uchicago showed to have the highest median followers and lowest median amount of total tweets per follower. This is a good attribute because it means they have a good ratio of followers to total tweets however it could suggest that if users who tweeted about Uchicago tweeted more than Uchicago could have higher influencing users.

University of Florida Density of Tweets Map

Figure 1



Median Total Tweets, Followers, and tweets about University per University

Figure 2

Appendix