UNIVERSITY OF CALIFORNIA
UC RIVERSIDE

”Spatial and Temporal Analysis on Twitter data using ArcGIS”

Report By

Kamalika Poddar (SID :- 862002289)

Pratheek Chindodi Rajashekar (SID :- 862002345)

Saturday 2nd June, 2018

1 Table of Contents

2 Contribution by Team Members

Kamalika Poddar: Data Collection and Data Processing
Pratheek Chindodi Rajashekar: Data Visualization

3 Introduction

Social media is currently playing an important role in the process of information diffusion. Exploring the pattern of message propogation on social network helps us better prepare for natural disasters or human crises. Twitter is among the fastest-growing microblogging and online social networking services. Messages posted on Twitter (tweets) have been reporting everything from daily life stories to the latest local and global news and events. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable information, enabling users and organizations to acquire actionable knowledge. Sentiments in tweets can be leveraged to understand variation in people's emotion towards varied subjects, during various times of the day and at different locations. This project analyzes sentiment variation during the entire span of the week. Furthur on we explore different locations using twitter data over entire USA.

4 Platform and Software used

This project uses Python libraries to filter the streaming data and json file to store the data. Here is the list of Software and Python Libraries used for this project.

4.1 Python Packages

- **Tweepy**: - Tweepy is open-sourced, hosted on GitHub and enables Python to communicate with Twitter platform and use its API. The current version of tweepy used is 3.3.0, which fixes various bugs and offers better functionality than the previous version. This library provides ability to crawl tweets generated all over the world.
- **Argparse**: - This module makes it easy to write user-friendly command-line interfaces.The program defines what arguments it requires, and argparse will figure out how to parse those out of sys.argv. The argparse module also automatically generates help and usage messages and issues errors when users give the program invalid arguments.
- **JSON**: - This is a lightweight data interchange format inspired by JavaScript object literal syntax. It was derived from JavaScript, but as of 2017 many programming languages include code to generate and parse JSON-format data.

- **Sys**: - This module provides access to some variables used or maintained by the interpreter and to functions that interact strongly with the computer.Built-in file objects representing standard input, output, and error are included in the sys module and are called stdin, stdout, and stderr.
- **Matplotlib**:- Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

4.2 Editors

- **Vim**:- Vim is designed for use both from a command-line interface and as a standalone application in a graphical user interface. Vim is free and open source software and is released under a license that includes some charityware clauses, encouraging users who enjoy the software to consider donating to children in Uganda.
- **Sublime Text** :- Sublime Text is a proprietary cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and functions can be added by users with plugins, typically community-built and maintained under free-software licenses.
- **Pycharm**:- PyCharm is an integrated development environment (IDE) used in computer programming, specifically for the Python language.It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django.

4.3 Tools Used

- **Open Refine**:-OpenRefine, formerly called Google Refine, is a standalone open source desktop application for data cleanup and transformation to other formats, the activity known as data wrangling. It is similar to spreadsheet applications (and can work with spreadsheet file formats); however, it behaves more like a database.
- **ArcGIS**:- ArcGIS is a geographic information system (GIS) for working with maps and geographic information. It is used for creating and using maps, compiling geographic data, analyzing mapped information, sharing and discovering geographic information, using maps and geographic information in a range of applications, and managing geographic information in a database.

5 Project Description

This project performs Data Ingestion, Aggregation and Analysis that allows the user to explore the spatial and temporal relationships present in twitter data. The project considers every single tweet as an expression of positive, neutral and negative sentiment. A collection of past tweets and their associated sentiments resides in a JSON file and is used for the study of sentiment across various places.

The project provides data visualization in the form of heat map, bar graph and pie charts. This visualization can be used to quicky visualize the current mode of each tweet. A more quantitative representation of each sentiment can be obtained with this project. A person can also dive deeper into the sentiment trends of any particular city by defining the bounding co-ordinates.

6 Related Work

There has been lot work done in this area, but below are just few of the related work we found most important of all.

- **Spatial and Temporal Sentiment Analysis of Twitter data**:- This article focuses on spatio-temporal variation of georeferenced Tweets' sentiment polarity, with a view to understanding how opinions evolve on Twitter over space and time and across communities of users. The results were identified as the highest percentage of positive Tweets occurred in the social science area, while science and engineering and dormitory areas had the highest percentage of negative postings. The number of negative Tweets increases in the library and science and engineering areas as the end of the semester approaches, reaching a peak around an exam period, while the percentage of negative Tweets drops at the end of the semester in the entertainment and sport and dormitory area. This study provided some insights into understanding students and staff 's sentiment variation on Twitter, which could be useful for university teaching and learning management.

- **Spatial and temporal analysis of Twitter: a tale of two countries**:- People share information with their peers using social media services (e.g. sharing their latest news over Facebook or Twitter) in order to inform the peers about their current situation. This has become a huge part of our social life. During crises this behaviour becomes even more acute because it allows people to reassure their peers (followers and friends) of their well being expeditiously. Of late, social media services have been also used for another purpose during crises: that of informing oneself over the current evolution of the crises. However obtaining relevant information from social media can be a difficult challenge as the bar for posting information, good or bad, is very low. Filtering the flow of messages such that only relevant information is remaining is critical in times of crises. To aid in this, we propose a spatial-temporal model that collects the data from Twitter. The data is further processed to evaluate the density of tweets surrounding the area. We also evaluate the possibility of shared user accounts by determining the physical distance and velocity between messages originating from the same user account.

- **Atmospheres**:- In this article, the team analysed sentiments for San Francisco and developed a WebApp showcasing the current sentiment of San Francisco using Javascript packages like jQuery, Bootstrap, Angular.js, MapBox, Leaflet.js. They mainly focused on the Visualization of data and the app is a current tweet streaming application and can be visualized from anywhere in the entire world.

- **Spatial, temporal, and content analysis of Twitter for wildfire hazards**: - Online networking information are progressively being utilized for improving situational mindfulness and helping calamity administration. They have analyzed the out of control fire related Twitter exercises regarding their ascribes appropriate to space, time, substance, and system, in order to pick up experiences into the helpfulness of online networking information in uncovering situational mindfulness. Discoveries demonstrate that online networking information can describe the fierce blaze crosswise over space and after some time, and subsequently are pertinent to give helpful data on catastrophe circumstances. Second, individuals have solid land mindfulness amid rapidly spreading fire risks and are occupied with imparting situational refreshes identified with out of control fire harm (e.g., regulation rate and consumed sections of land), out of control fire reaction (e.g., departure), and thankfulness to firefighters. Third, news media and nearby specialists are assessment pioneers and assume an overwhelming part in the out of control fire retweet arrange.

- **Spatial and Temporal Analysis of Tornado Fatalities in the United States: 1880–2005**:- A dataset of executioner tornadoes is aggregated and analyzed spatially keeping in mind the end goal to survey district particular vulnerabilities in the United States from 1880 to 2005. Results uncover that most tornado fatalities happen in the lower– Arkansas, Tennessee, and lower– Mississippi River valleys of the southeastern United States—a locale outside of conventional "tornado rear way." Analysis of factors including tornado recurrence, arrive cover, manufactured home thickness, populace thickness, and nighttime tornado probabilities exhibits that the relative most extreme of fatalities in the Deep South and least in the Great Plains might be because of the one of a kind juxtaposition of both physical and social vulnerabilities. The spatial dissemination of these executioner tornadoes recommends that the over the national normal manufactured home thickness in the Southeast might be a key purpose behind the casualty most extreme found around there. A statistic examination of fatalities amid the last piece of the database record delineates that the moderately aged and elderly are at a substantially more serious hazard than are more youthful individuals amid these occasions. Information issues found amid this examination uncover the requirement for a coordinated push to get basic data about how and where all setbacks happen amid future tornado and perilous climate occasions. These new, upgraded information, joined with aftereffects of spatially express investigations investigating the human humanism and brain research of these dangerous occasions, could be used to enhance future cautioning spread and relief procedures.

7 Data Collection

First we generated the Consumer Key, Consumer Secret key, Access Token and Access Token Secret from the Twitter API. The entire data was collected with the help of a python library "Tweepy" and stored it in a jSON file. Each row in a json file consisted of each tweet and the column represents different features like: 'id', 'id_str', 'created_at', 'text', 'retweet count', 'truncated tweet', 'timestamp','placename', 'place-country', all the information of users and co-ordinated from where the tweet generated.

We collected tweets for 1 hour each day from Monday through Friday during evening. From the huge json file of approximately 25GB each, we collected the geotagged tweets. The collected data was stored in three files each containing positive, negative and neutral keywords. There were 21 keywords used to collect data and the list of the keywords are represented in Table1.

Positive	Negative	Neutral
happy	shame	baffled
enjoy	doubt	authoritative
cheerful	envy	clinical
great	grief	detached
love	fear	nostalgic
enjoying	sadness	objective
challenge	frustration	restrained
learning	guilt	sentimental
curious	disgusted	candid
prefer	failure	frank
demand	afraid	preoccupied
advice	hate	unequivocal
trusting	pain	probing
unique	sick	nonchalant
like	overwhelmed	callous
easy	problem	consoling
nice	stressed	didactic
good	boring	direct
helpful	botthered	impartial
pretty	weird	unambiguous
fun	greedy	understated

8 Data Processing

In this project we tried to concetrate most of our findings for positive and negative and discarded the neutral tweets for better results. In this section we convert the crawled data into different formats for further processing.

- From all the features generated in the original jSON file, the geocoded data with the user information and geo co-ordinates was obtanied using the ***'argparser'*** python module and stored in a json file. Here the co-ordinates was obtained as a bounding box. ArcGIS is compatible with file formats such as CSV, GeoJSON, so the json file needed to be converted to a .csv file for further processing. OpenRefine was used to convert the jSON file to CSV file. While converting the file, lot of time got wasted and even after converting due to the file size, it became difficult to incorporate the layer in ArcGIS.
- For this reason we echoed the geotagged tweets from the tweets using Mac Terminal with the help of the below command.

```
ucrwpa-1-5-10-25-18-13:~ kamalika$ echo '{"type":"FeatureCollection","features":'"$(cat filename_of_coll
ected_tweets.json | jq 'select(.geo)| {type: "Feature", geometry:{type:"Point", coordinates:[.geo.coordi
nates[1],.geo.coordinates[0]]},properties:{tweet_body:.body,handle:.actor.displayName}}' | jq -s .)"'}'
> filename_of_geotagged_tweets.json
```

  The GeoJson file obtained contained the attributes as co-ordinates of a place and the text that it is being created from. It was easier to map the data in ArcGIS and hence the analysis was faster, even though we had to map each element.

9 Data Visualization

- Using each CSV file, a layer was added to the ArcMAP to analyse the tweets based on the location. For small files it was easier to upload the data then when we tried to upload large files it became difficult. For this reason we stopped our analysis and folowed a different approach.
- We used our second technique to upload the file and tweets mapped to the location perfectly. We added each layer in the project for Positive and negative tweets. Then after assigning location symbols and color code of *'green'* and *'red'* respectively, we started visualy analysing the data. We performed two types of analysis listed below:-

  Spatial: Here we used the 'Find Nearest' function within the 'Use proximity' function in AcrGIS, to find points in the layer which are nearer to Park, University and Hospital. This analysis was performed on mostly negative tweets to analyze the pattern of negative tweets origination from different location during each day of the week.

Temporal: We used the analysis feature in ArcGIS online to first Summarize the 'Aggregate' of all the points present in the boundary. To map the boundary we used an inbuilt Feature Layer of ArcGIS, called 'United States Country Boundary 2016'. With the help of 'United States State Boundary 2016', we could also find the number of tweets within the state.

10 Experiments

We conducted our experiment on 2.9 GHz Intel Cor i7 processor with ArcGIS Online Desktop version of ArcGIS. In the total of 402 geotagged tweets generated and mapped, 332 were positive tweets and 70 were negative.

11 Results

- Figure 1 represents the location of different tweets in Unites States within 60 sec.
- Figure 2 represents Positive, Negative and Neutral tweets generated from Los Angeles.
- Figure 3 represents all the positive tweets generated for five days.
- Figure 4 represnts all the negative tweets generated for five days.
- Figure 5 represents total positive and negative tweets for five days.

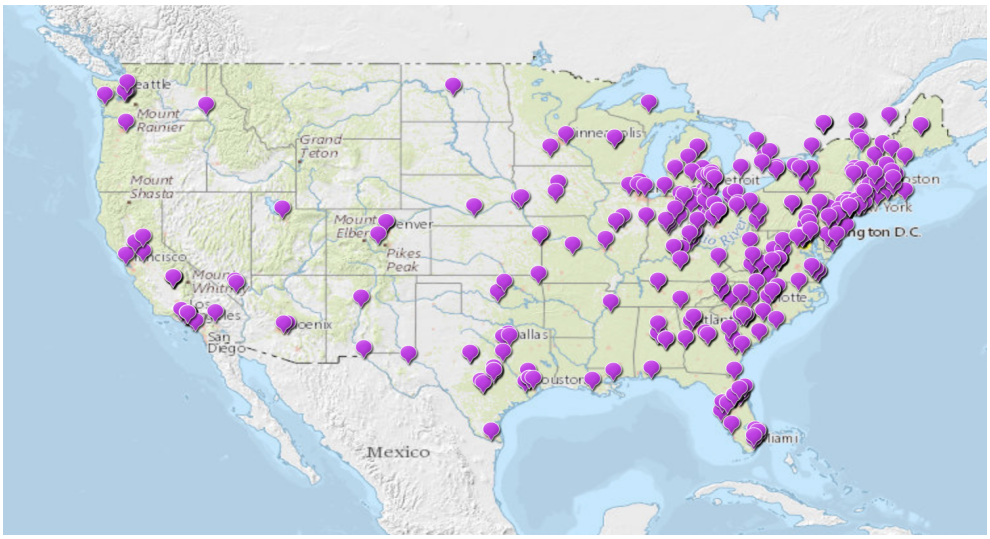

Figure 1: Tweets generated in different location in United States in 60 seconds.

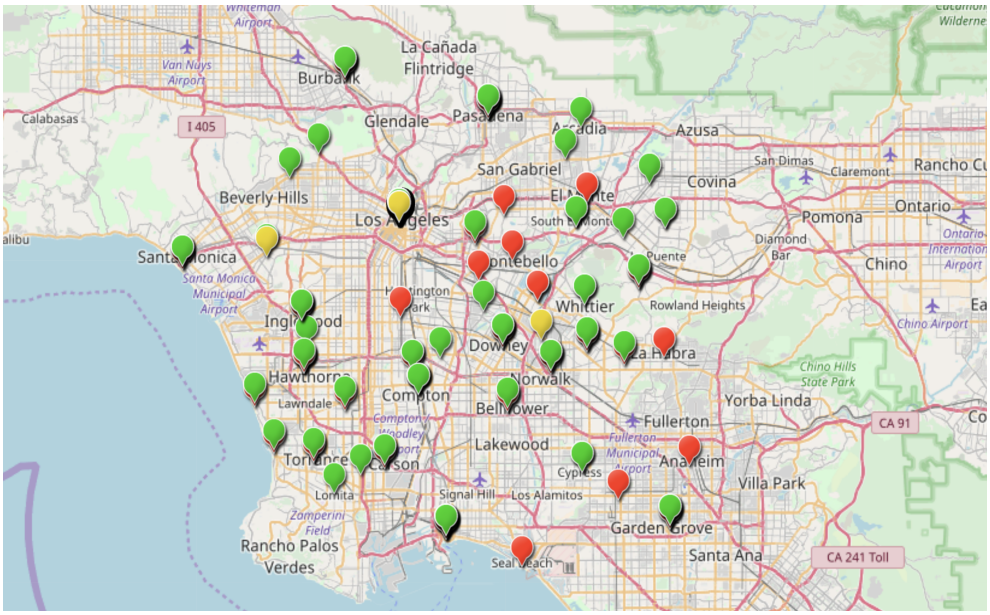

Figure 2: Positive, Negative and Neutral Tweets generated in LosAngeles for Monday.

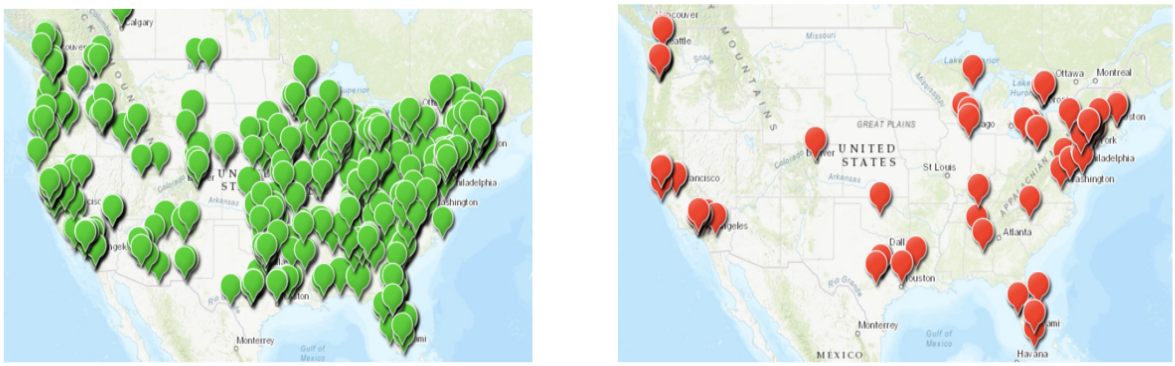

Fig a: Positive

Fig b: Negative

Figure 3: Total Positive and Negative tweets genrated for the five days in United States.

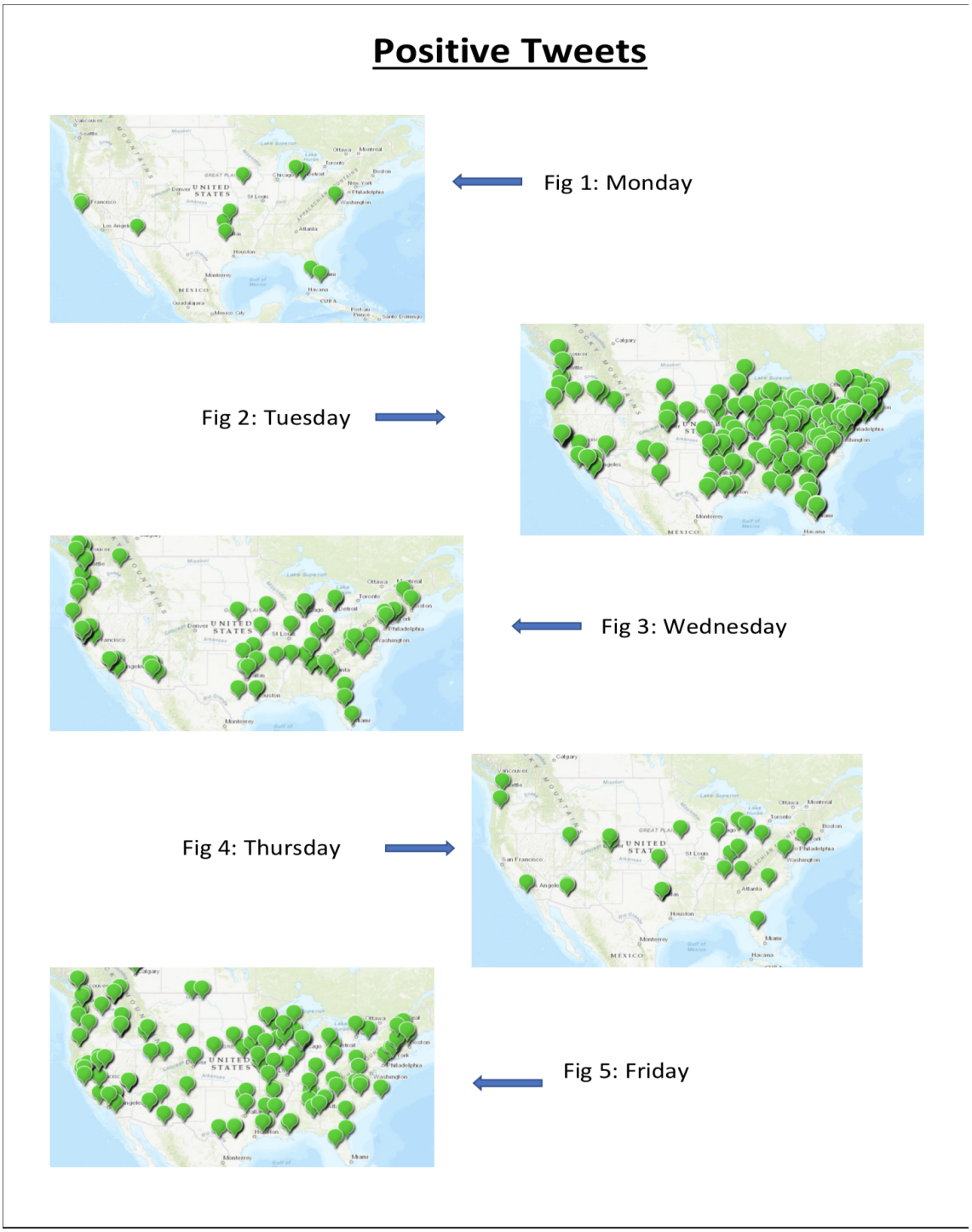

Figure 4: Positive Tweets during five days in a week.

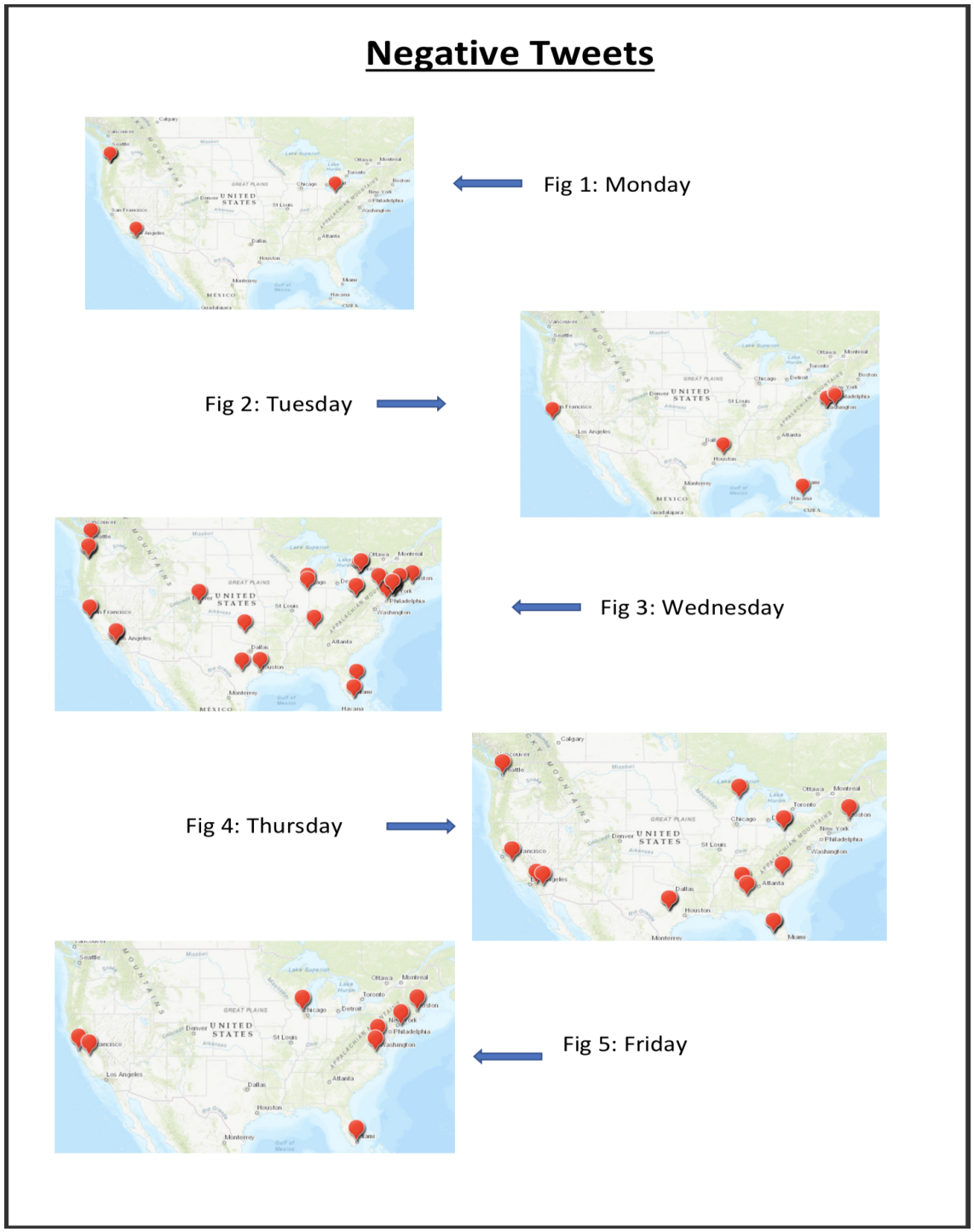

Figure 5: Negative Tweets during five days in a week.

11.1 Spatial Analysis

- Figure 6 represents table of data collected over five days from different location like Park,University and Hospital.
- Figure 7 represents the line graph variation of park, university and hospital for different days in a week.
- Figure 8 represents a bar graph showing the dependency of Negative tweets at different places.
- Figure 9 shows the Weekly Weather Report for the city of NewYork.

	Park	University	Hospital	Total
Monday	-	1	4	5
Tuesday	3	-	5	8
Wednesday	8	8	18	34
Thursday	4	4	7	15
Friday	1	2	5	8
Total	16	15	39	

Figure 6: Table representing negative tweets for Park, University and Hospital for five days.

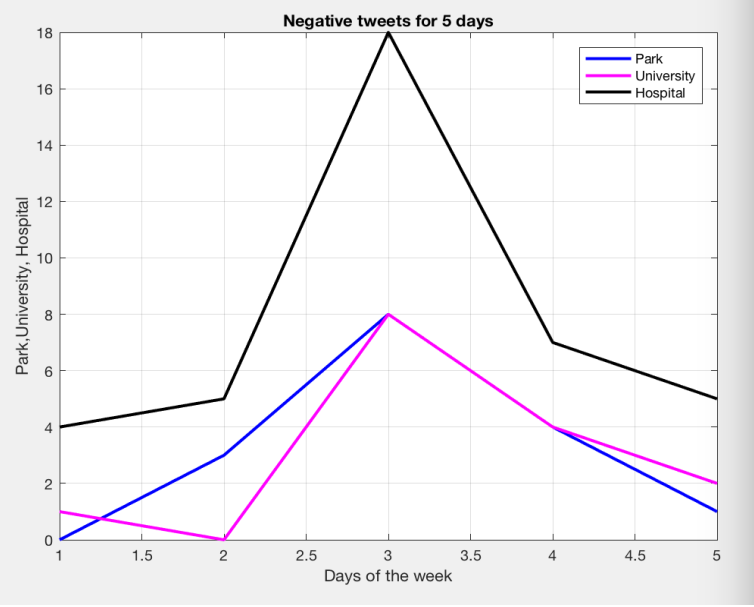

Figure 7: Line graph representation of Negative Tweets at different locations.

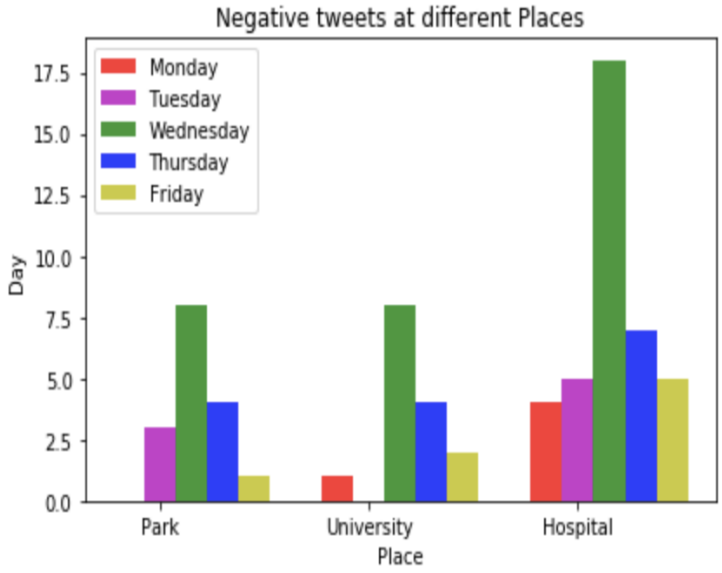

Figure 8: Line graph representation of Negative Tweets at different locations.

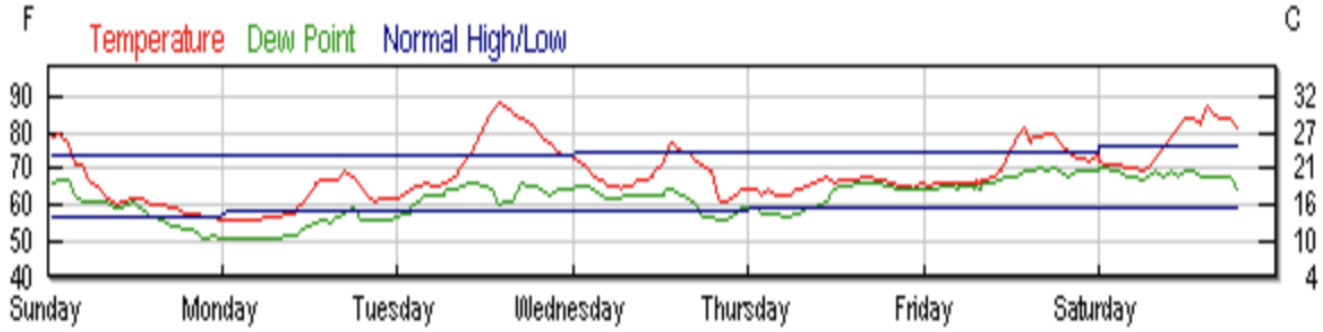

Figure 9: Weekly weather report for the city of New York

11.2 Temporal Analysis

- Figure 10 represents the table for the count of number of positive and negative tweet for five days.
- Figure 11 Line graph representing positve and negative tweets for five days.
- Figure 12 compares the different tweets and represents in the form of bars.
- Figure 13 shows comparison between positive and negative tweets during five days.
- Figure 14 compares the tweets with the help of pie charts and represnt in the form of percentages for better analysis.

	Positive	Negative
Monday	12	5
Tuesday	143	8
Wednesday	79	34
Thursday	26	15
Friday	72	8
Total	332	70

Figure 10: Table representing Positive and Negative Tweets during five days in a week.

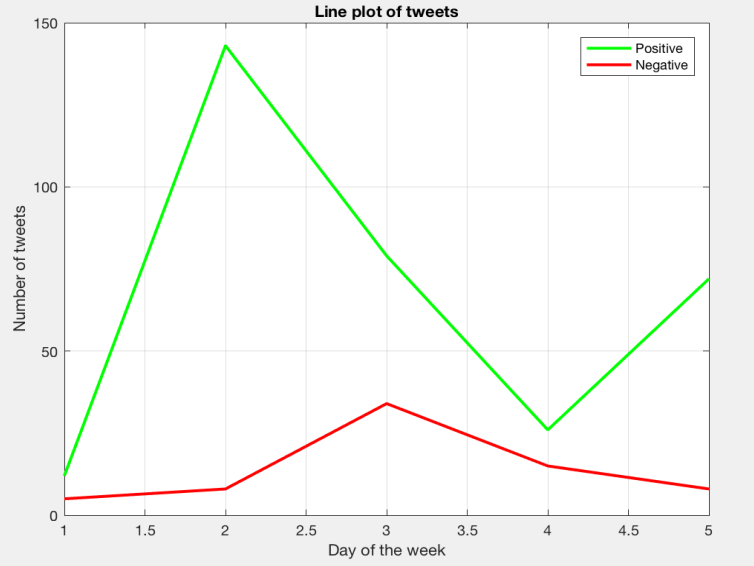

Figure 11: Positive and Negative Tweets generated for five days.

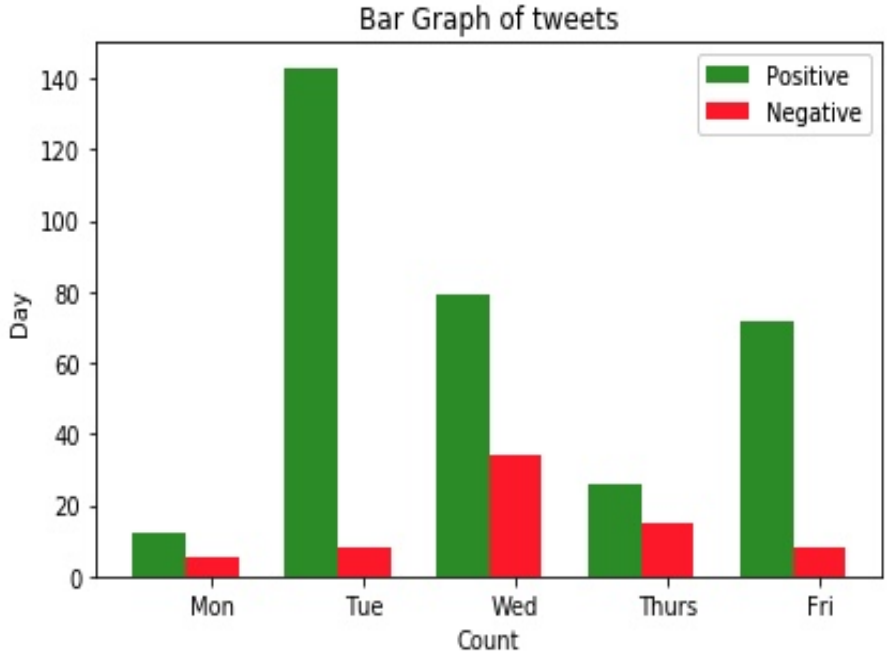

Figure 12: Bar graph representing positive and negative tweets vs days of week.

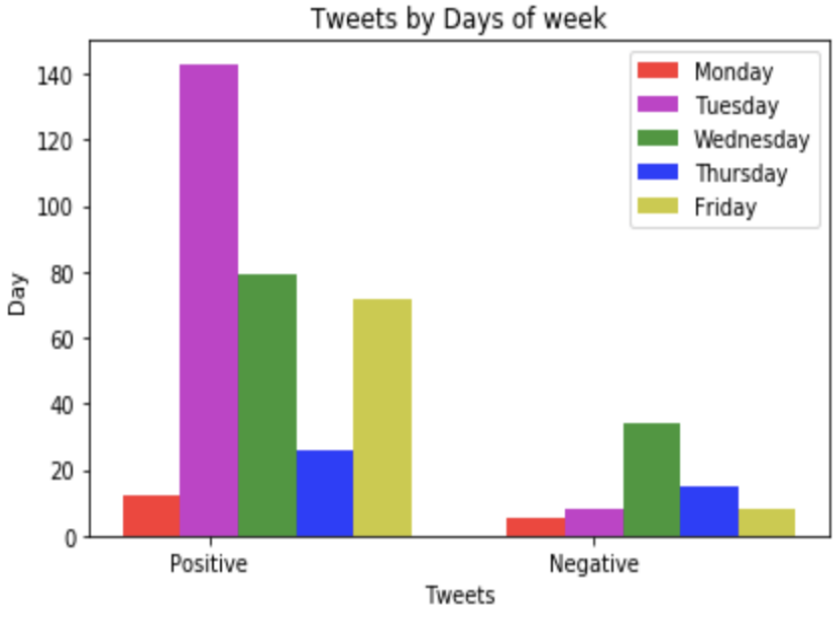

Figure 13: Positive and negative Tweets during five days in a week.

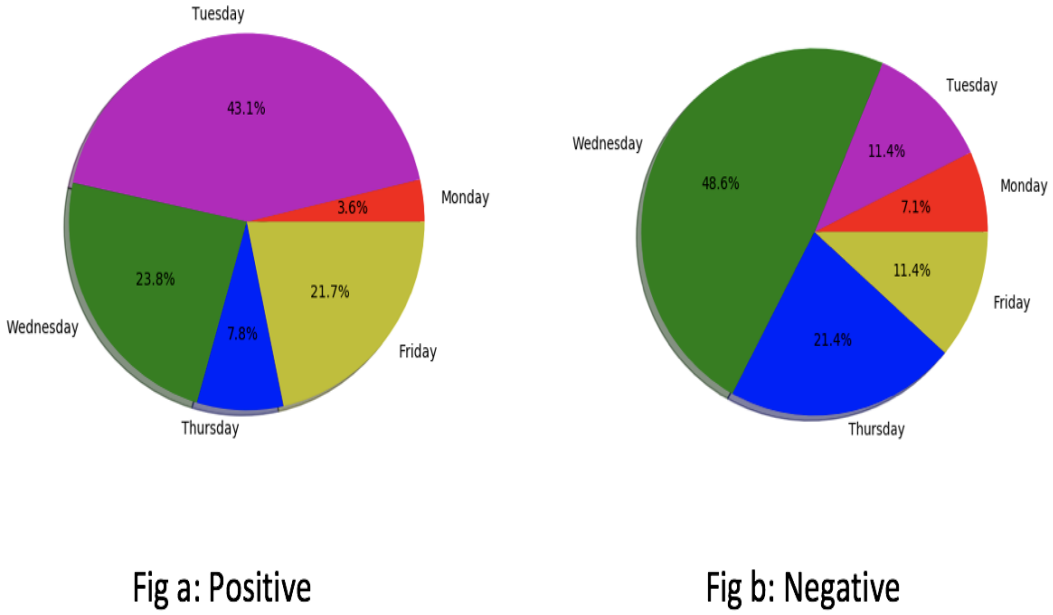

Fig a: Positive

Fig b: Negative

Figure 14: Percentage Representation of Positive, Negative Tweets generated in United States for five days.

12 Challenges Faced

- Geo co-ordinates are not enabled for all the twitter users, therefore we needed to extract the tweets with 'geo-enabled=true', which could be mapped. This in return reduced the number of tweets generated.
- Due to change in IP address there was problem in collecting the data and every time and error was raised with the change.
- When we tried to crawl the data for large amount of time, due to huge size of the file, sublime text used to stop responding. Therefore we started crawling for one hour instead.
- Data being huge, it was difficult to convert and upload the file in ArcGIS.
- Mapping was difficult as, when all layers were mapped together, each layer was individual and not all of the layers would converge together as a single layer.

13 Conclusion

After several iterations of the code and analysing each and every detail we reached to these final conclusions which are listed below.

- Performing Spatial Analysis on negative tweets resulted that number of negative tweets generated are mostly from the areas where there is a nearby hospital. And negative tweets generated from park and university almost had the equal variation throughout the week.
- The highest amount of negative tweets generated was on Wednesday with the vicinity of a hospital. This might be a due to the temperature variation in the climate, that might have made people fall sick. According to the weekly weather report by the city of New York, it was reported that there was a rise in the temperature between the period, Tuesday and Wednesday, which concludes our findings.
- The weather report also aligns with our findings that the least amount of negative tweets happened to be on Monday and highest amount of positive tweets turned out be on Tuesday.
- The total number of tweets generated for the week, Monday[05/27/2018] to Friday [05/1/2018] were mostly 'POSITIVE'. We also observed a phenomenon, that the generation of most positive tweets is directly co-related with the occurence of positive and negative events within a given time frame. Given the fact that Monday[05/27/2018] was 'Memorial day', the most amount of tweets was generated on that day. The representaion shows it on Tuesday due to different time zones in United states.
- Most of the tweets were generated from San Francisco, Los Angeles, New York, , Florida and Boston. Therefore we can conclude that the number of tweets generated is proportional to the population of the city.

14 References

1. Name: Xuebin Wei, Source: Github (Data-Mining-on-Social-Media), Link: https://github.com/xbwei/Visualizing-SocialMediaData
2. Name: Tejal Patted, Source: Github (Spatial-temporal-tweet-analysis), Link: https://github.com/TejalPatted/Spatial-temporal-tweet-analysis
3. Name: MikaelL Brunila, Source: Google (Scraping, extracting and mapping geodata from Twitter), Link:http://www.mikaelbrunila.fi/2017/03/27/scraping-extracting-mapping-geodata-twitter/
4. Name: Marco Bonzanini, Source: Google (Mining Twitter Data with Python), Link:https://marcobonzanini.com/2015/03/23/mining-twitter-data-with-python-part-4-rugby-and-term-co-occurrences/

5. Name: Pablobarbera, Source: Github (subset-geolocated-tweet), Link:https://github.com/pablobarbera/pytwools

6. Source: Quora, Link: https://www.quora.com/What-is-the-longitude-and-latitude-of-a-bounding-box-around-the-continental-United-States

7. Source: Youtube, Link: https://www.youtube.com/watch?v=ySxeoo2bB4U&list=PLH utrxqbP1Bwc7Q0yTWnAlod_tbHFbUBJ&index=18

8. Source: ArcGis Online, Link:https://kamalika.maps.arcgis.com/home/web map/viewer.html?webmap=3a80d802e13f400fae567668e49f9771

9. Source: OpenRefine

10. Source: Github, Link: https://github.com/amyxzhang/boundingbox-cities/blob/master/ boundbox.txt

11. Link:http://www.fallriverschools.org/Tone%20and%20Mood%20words%20%28unedited %29.pdf

12. https://www.pythoncentral.io/introduction-to-tweepy-twitter-for-python/

13. https://stackoverflow.com/questions/10214827/find-which-version-of-package-is-installed-with-pip

14. https://en.wikipedia.org/wiki/JSON

15. https://espace.curtin.edu.au/handle/20.500.11937/36187

16. http://eshlefest.github.io/pdfs/atmospheres_docs.pdf

17. https://tex.stackexchange.com/questions/128185/how-to-give-floats-figures-titles

18. https://www.wunderground.com/history/airport/KNYC/2018/5/27/WeeklyHistory.html?req_city=New20York&req_state=NY&req_statename=New%20York&reqdb.zip=10001&reqdb.magic=8&reqdb.wmo=99999

19. https://github.com/pubnub/tweet-emotion

20. http://support.gnip.com/articles/visualizing-twitter-geo-data.html