CS 4624: Multimedia, Hypertext, and Information Access

Virginia Tech, Blacksburg, VA 24061
Spring 2020

Twitter Disaster Behavior

Final Report

Kayley Bogemann, Shane Burchard, Jessie Butler, Austin Spencer, Taylor Thackaberry

Client: Ziqian (Alice) Song

Professor: Edward Fox

May 5, 2020

Table of Contents

Table of Figures

Table of Tables

Executive Summary/Abstract

The purpose of this research is to identify patterns in behavior immediately following natural disasters by analyzing Twitter data. The data being analyzed consists of the geolocation, the contents of the tweet, the number of retweets, and the date/time it was tweeted. With this analysis, we hope to better predict how people will respond to natural disasters. By providing this information to emergency personnel and law enforcers, the research aims to improve the response time to these events.

In order to gain this information, tweets relating to Hurricane Maria and the recent Puerto Rico earthquake were collected. All tweets pertaining to Hurricane Maria collected were created between September 15, 2017 through October 14, 2017. Similarly, all tweets pertaining to the earthquakes in Puerto Rico were created between January 7, 2020 and February 6. These tweets were analyzed for their content, number of retweets, and the location associated with the author of the tweet. We considered key words in topics relating to preparation, response, impact, and recovery, and counted the occurrence of these words. This data was then graphed using Python and Matplotlib. Additionally, using a Twitter crawler, we extracted a large dataset of tweets by users that used geotags. These geotags show location changes among the users before, during, and after each natural disaster. Finally, after performing these analyses, we developed easy to understand visuals and compiled these figures into a well-organized poster.

Using these figures and graphs, we compared the two datasets in order to identify any significant differences in behavior and response. The main differences we noticed stemmed from two key quality of the different disasters. First, Hurricane Maria's arrival was predicted, but the earthquakes were not, resulting in different patterns of activity spikes. Hurricane Maria was also a single, isolated event, while the initial earthquake was followed by aftershocks that also affected the area. Thus, the Hurricane Maria dataset experienced the highest amount of tweet activity at the beginning of the event and the Puerto Rico earthquake dataset experienced peaks in tweet activity throughout the entire period, usually corresponding to aftershock occurrences. This report will delve into these differences, as well as other important trends we identified.

Introduction

Since December 31, 2019, Puerto Rico has experienced multiple earthquakes exceeding a 5.0 magnitude. Earthquakes are unpredictable, and it is essential to have as much information as possible to assist people affected by natural disasters such as these. To better understand how people respond to events such as earthquakes, our team considers information posted on social media, specifically Twitter. Our client collected 332,927 tweets written between December 31, 2019 and January 15, 2020 that reference these Puerto Rico earthquakes. Our team used these tweets to determine information about individuals in Puerto Rico during these events. By using geotags, identifying timestamps, and searching for relevant keywords, we were able to create graphs and heatmaps to display how often the public talks about certain topics, and where in the world people are tweeting from.

Our project has been further expanded to include analysis on tweets written during Hurricane Maria. Because hurricanes are able to be predicted when compared to earthquakes, we can compare the public's response to the two events.

Requirements & Constraints

The purpose of this project was to identify common behavioral patterns during natural disasters through Twitter data analysis. Additionally, after these patterns were identified, graphs and visuals need to be created to communicate them. Thus, the two main requirements of this project are:

1. Analyze Twitter data to identify novel or significant behavioral patterns in response to the Puerto Rico earthquakes and Hurricane Maria; and
2. Create and display meaningful graphs, visuals, and other figures to effectively communicate these findings.

This analysis should also explore similarities and differences between the behavioral patterns of the two natural disaster events. The charts and figures must be visually pleasing, easy to read and understand, and well organized. All analysis needs to be done with the overarching goal in mind and with the intent to benefit those interested in natural disaster response and recovery behavior.

Design

The main points of analysis we chose to explore were categorical analysis, topic analysis, top tweet analysis, and location tracking analysis. Additionally, a major aspect of the categorical analysis and the top tweet analysis we chose to include was exploring tweet frequency over time. All analysis decisions were made based off of both our client's requests and the kind of information we received from the Twitter API.

We chose to perform categorical analysis and topic analysis because identifying trends in discussion topics could help us pinpoint particular failures in infrastructure and emergency response. The categories and topics we explored included response, impact, and recovery. Top tweet analysis also facilitated this goal because it showed which tweets received the most attention, in the form of likes and retweets, over the course of the natural disaster. Finally, location tracking analysis was chosen because we were interested in determining if there were any location change patterns. To do this, we decided to generate heat maps of Puerto Rico's twitter activity before, during, and after the earthquakes. These patterns are important because they can better help researchers to understand and predict how people may move around during and after natural disasters.

Implementation

I. Graphing Tweets based on Day

The dataset of tweets already includes only tweets discussing the disasters in question, so creating a graph depicting the popularity of these topics was simply a matter of graphing the number of tweets in the dataset over time. The parser stepped through each tweet in the dataset, parsing the time it was created. Then each tweet was placed into an array with blocks for every hour. For each tweet, a +1 was added to the corresponding block. Since this graph was concerned with how much users were talking about the incidents, retweets were also counted, albeit on the day they were retweeted. At the end of the dataset, the array that counted the number of tweets every hour was then graphed against time in a line graph to show the hourly progression of tweets. The parser ran through both the dataset of Earthquake tweets and Hurricane tweets. Then we could compare the graphs side by side. To compare the timeline of these different events, the x-axis reflects the time since the initial disaster struck. The y-axis was also rescaled because the different datasets had different volumes of tweets, and we wanted to compare the relative spikes in data. The earthquakes only impacted residents in Puerto Rico, while Hurricane Maria also threatened Twitter users in Florida and along the American East Coast [1]. Since Hurricane Maria's impact was more widely discussed, but we are more focused on comparing the patterns in discussion rather than the overall quantity of tweets, the Hurricane Maria y-axis was rescaled so that it would fit on the same graph as the earthquake tweet data. This graph is shown in Figure 1.

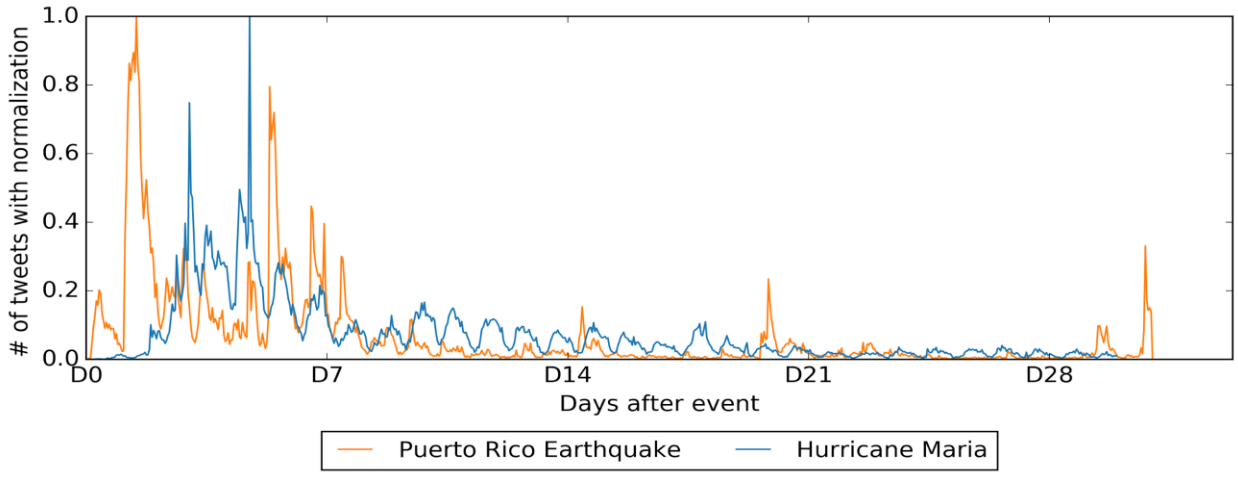

Figure 1: The number of tweets on each topic by day.

After this graph was created, the distinct tweet behavior of the different types of disasters became evident. Hurricane Maria was a natural disaster that could be predicted, with its path of movement predicted by FEMA, so individuals living in areas that could possibly be affected could prepare in advance. Despite all technological attempts to do so, earthquakes still cannot be predicted, so those affected were unable to prepare. Also, while the effects of a hurricane's flooding and damage are long-lasting, a hurricane is only one single event. Earthquakes are often followed by multiple aftershocks. The data reflected this, as seen in Figure 1, with the hurricane tweets reaching a peak during and immediately after the initial landfall, and then petering out as rebuilding efforts were underway (or the public lost interest in the story). The earthquake tweets, however, continue spiking throughout the time span of the dataset, the largest being during aftershocks of the highest magnitude.

II. Graphing Tweets Based on Word Frequency

We were particularly interested to learn how frequently people discussed specific topics about infrastructure and recovery from the events. We created a parser to search the contents of each tweet for key words of interest. Key terms are broken up into different categories, each containing words and phrases that pertain to that category. When performing the actual comparison, a variable stores the number of times that each chosen word/phrase appears in a tweet and increments the number accordingly. Once the final count was retrieved, we stored that value for that day within an array. This array stored the word count for the specific key terms being searched for pertaining to each day. Once all of the data was collected, it was then graphed based on the data within the array. This represents the total period of time in which data was collected and the number of times a particular phrase or keyword was tweeted. In Figures 2 and 3, we selected the subcategories that had the most dynamic trends over the course of the two disasters, to highlight how the responses and discussion of these disasters changed over the course of the incidents. For example, immediately after the January 7 earthquake, fundraising efforts for earthquake relief in Puerto Rico began. American charity watchdog CharityWatch released an article on January 8 advising prospective donors of the best way to choose charities so that their money could reach the people who needed it most [2]. Residents of Puerto Rico such as Orlando Crum-Echevarria, as well as the Happy NPO (a 501(c)3 nonprofit) created online fundraisers on January 8 using websites such as purecharity.com and GoFundMe to support the reconstruction of destroyed buildings [3][4]. Twitter data reflected a sharp uptick in discussion about fundraising and donation on this day as people donated to these and other relief efforts. Fundraising drops off after a couple of days, but in the case of the earthquake, repeat events trigger public interest again, resulting in multiple spikes.

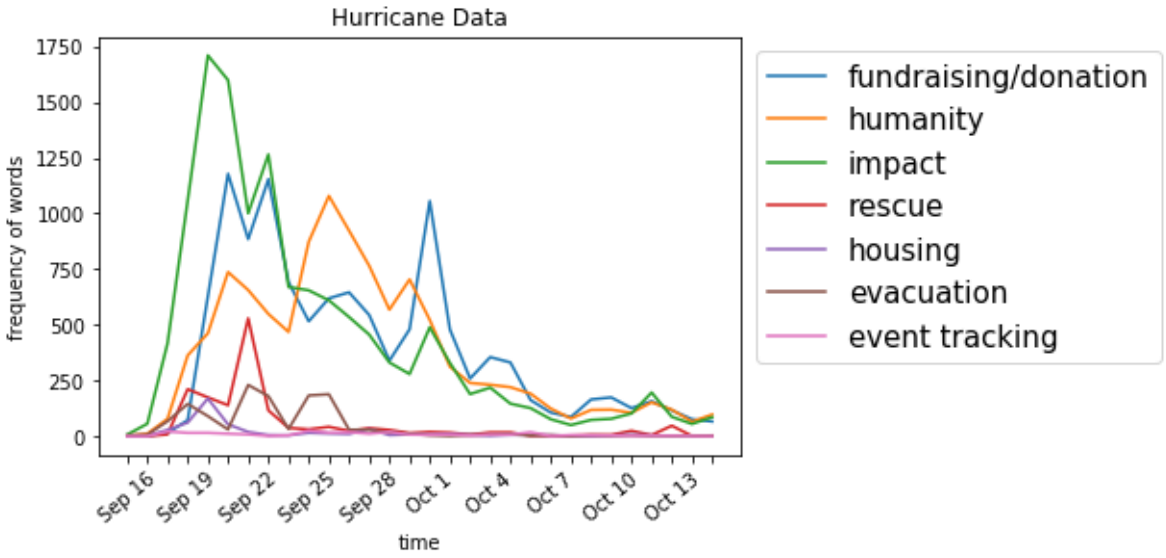

Figure 2: Word Frequency analysis for tweets about Hurricane Maria.

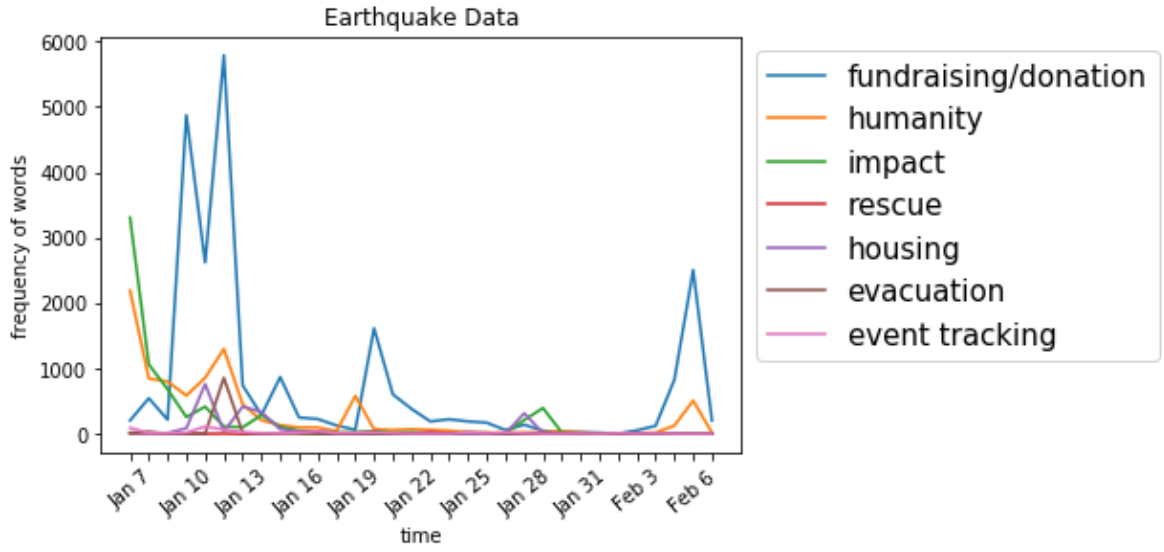

Figure 3: Word Frequency analysis for tweets about the 2020 earthquakes.

III. Selecting Top Tweets to Feature

A major feature of Twitter is the ability to "retweet" or share a tweet from a different Twitter account to one's own followers. Retweets are often seen as a measure of a tweet's popularity, as well as an indication of how many people have seen the tweet. When a tweet is created, it is visible to everyone who follows that account. However, if a follower retweets a tweet, that tweet

will also be shared with the followers of the other account, greatly increasing the messages' exposure.

We decided to find some of the most retweeted-tweets each day to see what people were talking about on the days where there were some of the biggest spikes. We created another parser to pass through each dataset, sorting the tweets by day. Then, the parser visited each day in the array and sorted the tweets by the number of retweets, returning the users and the text of the top three tweets that garnered the most retweets that day. On our poster for ISCRAM 2020 (which will be discussed in more detail in a later section), some of these notable top tweets for peak days (featured in Figure 1) were included to give context about the events transpiring during those statistically-notable days.

In addition to the top tweets per day, we also conducted analysis on the top tweets per category. Similar to the previous parser, this parser put the tweets into categories based on specific words mentioned in the tweets (for a more detailed explanation, see the section on Graphing Tweets Based on Word Frequency). Then, these tweets were sorted in their respective categories in terms of the number of retweets. The parser then returned the top ten tweets in each category, for a total of forty tweets. While these tweets were not featured on the ISCRAM poster in the interest of space, they provided some insight into the types of content people most engaged with during these crises.

The reason this approach of utilizing multiple parsers was taken was primarily due to the size and structure of our dataset. Since the data file updated multiple times throughout the project and the file type itself was changed from a CSV to a JSON format, it became easier to create multiple parsers to handle the different file types as well as tackle the issue of updating data files. This allowed us to continue to view older data while keeping track of changes within the code. Additionally, it never occurred to us to load the data into a database and generate the data through SQL statements. This approach may have streamlined analysis. The different parsers that were created for this project were a topic analysis and a top tweets per day, each taken in a JSON file of tweets pertaining to either Hurricane Maria or the Puerto Rico earthquake. The data produced for the topic analysis yielded graphs that represented the occurrence of phrases and words that were preselected by our client. On the other hand, the top tweets parser generated the top tweets per day.

After we created a Python script that would retrieve the most popular tweets per day during each crisis, we started to find some patterns about which tweets people engaged with the most. A majority of the top tweets retrieved by our parser were either American politicians or news sources. Since the American government is technically within its ever-growing election cycle, the weak response by the current (starting in 2016) administration garnered considerable backlash from former president Barack Obama as well as democratic candidates for 2020 such as

Bernie Sanders. These political figures have a considerable following, so the number of retweets they received was naturally high. News sources also were the first to report earthquakes and aftershocks. They have a considerable following both inside and outside of Puerto Rico, so they were also popular with their followers. These tweets were less focused on small infrastructure failures or personal struggles, and more on the big picture of relief in Puerto Rico as a whole. The popularity of these tweets suggests that people are unhappy with the way that political figures in power are handling the crisis, and that sympathy for Puerto Rico can be found on the mainland.

IV. Topic Analysis on Tweets

Using the same categories as before, we also identified the most common words and phrases tweeted by users relating to each category. Each category was split into two subsections of common words and phrases: those relating to the Puerto Rico earthquakes and those relating to Hurricane Maria. To do this, we used Python code in a Jupyter Notebook that takes in a .csv file of tweets and outputs a set of .dict, .lda, and .mm files. Each of these files are created for three types of analyses: single-word topics, two-word topics, and multi-word topics. Single-word analysis identifies the most common words used in the tweet set (e.g., earthquake, damage, power). Two-word analysis identifies the most common pairs of words used (e.g., “power plant”, “damage inside”, “without power”). Phrase analysis identifies the most common phrases of three or more words (e.g., “reports show damage”, “thousands (sic) people homeless”).

Those output files are fed into the topic_vis_2015election-tweets.py file, which creates a visual mapping of the most common topics. The most common “topic groups” are identified (usually 3-5, shown by the five circles in Figure 4). These are groups of words that typically go together in tweets. The Intertopic Distance Map describes the relationship between topic groups. If there is any overlap between two circles, it means that the two topic groups share some similar words and phrases. Figure 4 shows topic group #1 highlighted. Under the Top-30 Most Relevant Terms, we see that “power” is the most common word in the topic group, and the red highlight suggests that this topic group accounts for most occurrences of the word “power”.

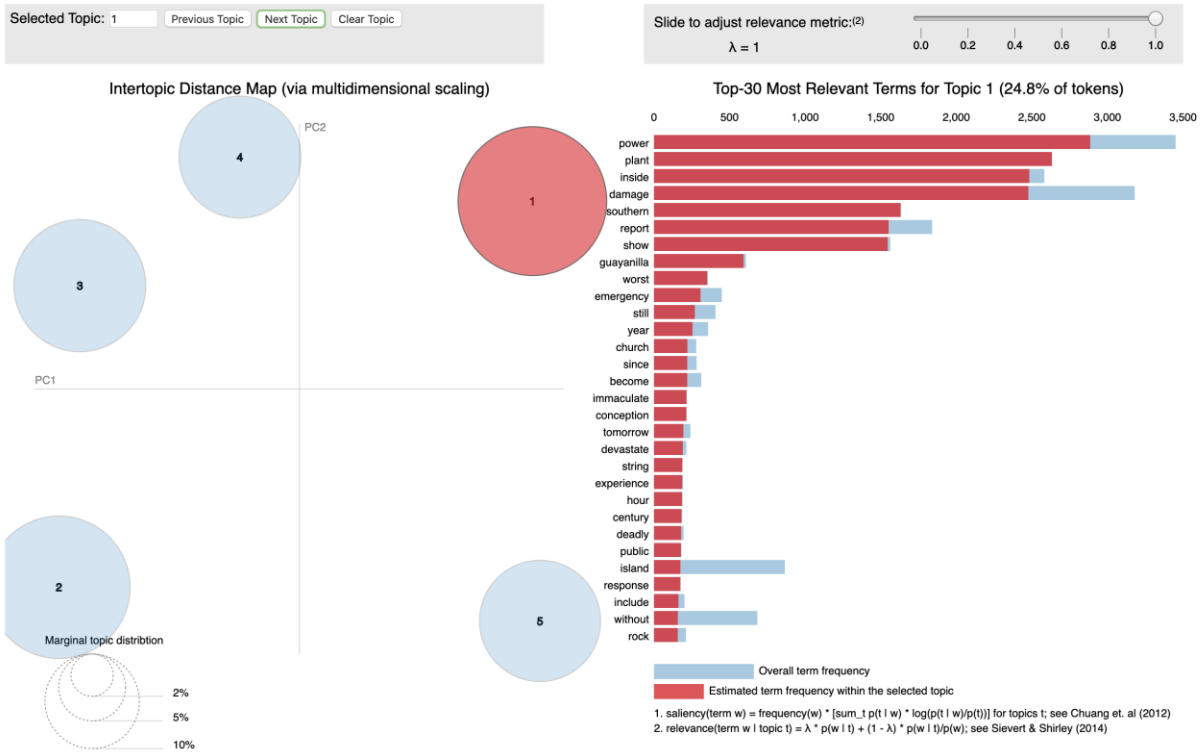

Figure 4: Single-word topic analysis for Twitter accounts run by organizations.

After analyzing these visualizations, we identified the most common topics of discussion on Twitter. The table below shows these topic categories, as well as their occurrences in one-word, two-word, and phrase analysis. The main topics of discussion included electricity (the earthquake damaged the Costa Sur Power Plant, cutting off power to residents), homelessness, a destroyed church (the Immaculate Conception Church in Guayanilla), and general earthquake severity (aftershocks, health emergency, federal aid).

	Subject	One-Word	Two-Word	Phrase
Topic 1	Power, Electricity	power, plant, costa, sur, damage, disaster, repair	power plant, costa sur, severe damage, rico electricity, large crack	costa sur power plant damage, power outage affects, serious threat
Topic 2	Homelessness	home, homeless, many, people, leave, destroyed, thousands	thousands homeless, home destroyed, sleep car	thousands residents sleep car, puerto rico homeless

Topic 3	Destroyed Church	immaculate, conception, church, damage, devastate, inside, deadly	church guayanilla, building destroyed, damage inside	earthquake devastation church guayanilla, immaculate conception church damaged inside
Topic 4	Earthquake Severity	magnitude, hit, tremor, center, response, natural, disaster, aftershock	magnitude earthquake, 6.4 magnitude, health emergency, earthquake relief,	report show earthquake damage, magnitude tremor island, withhold aid, earthquake aftershock

Table 1: Most common one-word, two-word, and phrase topics.

V. Tweet Geotag Mapping

A separate dataset was given for tweet geotags, including over 2,200,000 tweets. However, the dataset included tweets without geotags. First, the geotagging code narrowed the dataset down to only tweets with geotags. The final dataset included 206,000 tweets from 14,000 unique users. The code performs two types of analyses on the dataset: general activity heatmaps, and user movement tracking.

General Activity Heatmaps: We selected two major events, a 6.4 magnitude earthquake on January 7, and a 5.9 magnitude aftershock on January 11. The code extracted tweets for each user before, the day of, and after the events. Each tweet contained a geotag location, which was fed into a geocoder that converts addresses to geographic coordinates. Using Tableau, a data visualization software, we used the coordinates to create three heatmaps for each event that show the density of Twitter activity across Puerto Rico. We only included users that were tweeting often enough that they had tweets in each of the before, during, and after time frames. This reduced our user sample from 14,000 to about 4,000. When looking at Figure 5, we see that there is not much noticeable movement on a large scale. However, we can see a slight increase of user density in urban areas. The capital, San Juan (located in the upper right corner), especially saw an increase in Twitter users after the day of the earthquake. This could be for a number of reasons. Perhaps people from more rural areas need to enter the city to buy things like supplies in reaction to the earthquake, or people whose homes were damaged from the earthquake are moving into the city to stay temporarily.

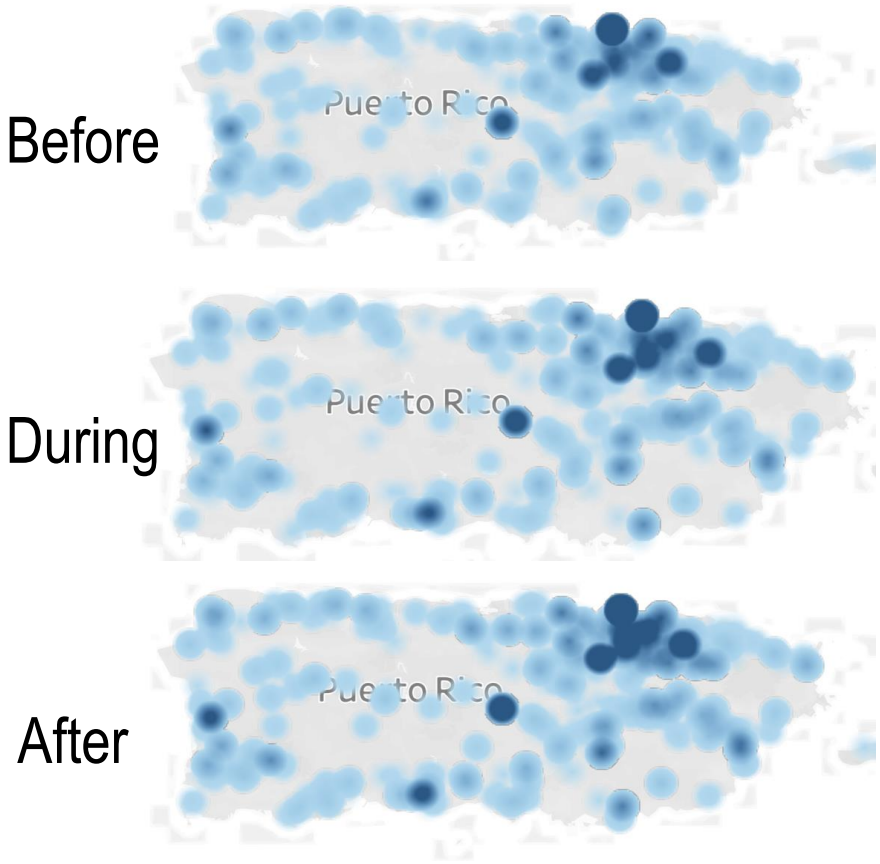

Figure 5: Twitter user heatmaps for before, during, and after the January 7 earthquake.

User Movement Tracking: The code iterated through each user's set of tweets during the timeframe (January 1 to January 15). The code compared coordinates for a single user during that time frame. If any changes in coordinates were greater than 0.25 degrees (about 17 miles), that user was marked, and their coordinates were mapped as a path over time.

Testing/Evaluation/Assessment

Testing was performed using a Jupyter Notebook as well as Eclipse's Python IDE. Jupyter Notebook allowed the testing of different sections, while providing the capability of rapid deployment of certain sections of code. Eclipse's IDE has a better built-in debugger for ensuring accuracy during development.

Since the sheer size of the dataset prevented us from simply "eyeballing" the data to see if our results were correct, we often created sanity benchmarks in our code. For example, in the World Frequency analysis code, we would explicitly check to make sure that the date ranges were within our expected range. We also first tested it on a smaller set of about 1000 tweets instead of the full dataset. That way, we could determine if the numbers we were receiving seemed reasonable for the amount of data given. During our development phase, our client provided us with a sample set of data. During our initial code creation, we discovered that this dataset was incomplete, and entire days of data were missing from the dataset. When we received the more complete dataset from our client, however, this problem was resolved. When we started receiving results, we would sanity check the data again, checking the news for the date ranges listed if there was a particularly large spike in discussion to see if such a spike was logical. In the earthquake data, it made sense that tweets would drastically increase on the days that aftershocks occurred, while the hurricane data leveled off slowly after the event.

Since we conducted multiple types of analysis on the same data, seeing the same trends occur in different analyses also served as proof that our results were consistent with the data and not a result of errors in our code.

After conducting many different types of analysis, we worked with our client to discuss which ones would be the most impactful to our audience and made sure to highlight those on the poster to take advantage of our limited space.

ISCRAM Poster

The main deliverable for this project was a poster created for the ISCRAM 2020 conference. The completed poster, as well as enhanced images of the graphs we created for the poster, can be found in the **Appendix.** Although this conference was ultimately cancelled due to COVID-19 concerns, our team completed the poster, and our client submitted a document write-up outlining our work to the ISCRAM website. The poster will be published in 2020 and presented by our instructor, Dr. Edward Fox, at ISCRAM 2021.

User Manual

The graphs and analysis supplied by our research are a culmination of tweets posted during and related to the natural disasters of Hurricane Maria in 2017 and the Puerto Rico earthquakes of 2020. Tweets were by people who live in Puerto Rico, as well as those located outside of the area. The graphs created reflect the contents of these tweets. Users can test hypotheses regarding reactions and human behavior in response to these events. Ideally, these tweets could be used to find patterns in assistance or disaster relief needed, but since this information is drawn from a corpus of natural language (posted online and not generated for this purpose), results are not always guaranteed. However, we hope that some of the illustrations of this data can tell a story of how these disasters affected this area, as well as help inform you as to which actions you can take to best benefit the people of Puerto Rico.

The **Word Frequency Analysis** graphs aim to determine at which times during these events topics such as preparation, impact, or response were mentioned in tweets. If a certain topic's line is increasing or reaches a peak, this means that people were most interested in discussing that particular topic. Two of our notable word frequency analysis graphs are featured on the top row of our poster, as illustrated as Figures 7 and 8 in the appendix. The graph on the left (Figure 7) highlights the frequency at which people used words relating to impact, response, and recovery in their tweets during the earthquake. The graph on the right (Figure 8) shows the frequency of the same words, but during Hurricane Maria. If people are talking a lot about impact, for example, this may mean they are trying to draw attention to something that has been destroyed and needs repair. On the other hand, if they are talking (or not talking) about preparing for a disaster, then you might be able to determine how prepared they were for the event.

Similar to word frequency analysis, **Topic Analysis** searches for meaning and content patterns within tweets. Instead of searching for particular phrases of interest, however, Topic Analysis aims to find recurring phrases in tweets to see more specifically about what users are discussing. In addition to word frequency analysis, topic analysis assists leaders in government or social services so that they can better determine what services are needed, and where.

The **Geotag Analysis** includes location data for Twitter users who choose to share their location when they tweet. From these graphs, hot spots may indicate areas of high population or high concern, since there are more tweets coming from those areas than others. The movement of these users may also give clues as to how people travel (or do not) during emergencies, which could be important information for first responders.

All these types of analysis are included for both the earthquake disasters and for Hurricane Maria to show how people react differently to two different types of disasters. Some disasters can be predicted, but others cannot, so impact and human reactions to these events differ immensely.

Hopefully, by considering the information illustrated in this data, better methods and structures can be developed to prepare for, respond to, and minimize the damaging effects of disasters of this scale in the future. This could include, but is not limited to, improved infrastructure, redesigned evacuation and response plans, and improved disaster relief measures.

Developer's Manual

I. File Inventory

Datasets		
File Name	**Extension**	**Description**
PR_Earthquake	.csv	Original tweet set used for word frequency and topic analysis. Each row contains tweet text, favorite and retweet counts, user information such as followers, whether or not the account is verified, and whether or not the tweet was an original or a retweet. Does not contain geotags.
PR_Earthquake_Location	.csv	Original set of 2.2 million tweets used for location tracking. Contains all of the details in PR_Earthquake while also including geotags for some users. Of these 2.2 million tweets, 206,000 of them include geotags, from 13,806 users. We performed location tracking on just these 13,806 users.
maria_tweets	.json	Original tweet set used for analysis of tweets during Hurricane Maria.

Table 2: Descriptions for the original tweet datasets used in all of the projects

Graphing Tweets Based on Word Frequency		
File Name	**Extension**	**Description**
parse_tweets_earthquake	.ipynb	Takes in a set of tweets and produces graphs that show the frequency of particular word sets over the time of the earthquake
parse_tweets_hurricane	.ipynb	Takes in a set of tweets and produces graphs that show the frequency of particular word sets over the time of Hurricane Maria
Earthquake-new-topics	.png	Images of the graphs produced from the parsing

		program. Words are split up into categories: Preparation, Response, Impact, and Recovery, in a file provided to us. These words are found significant by the NSF.
Hurricane-new-topics	.png	Images of the graphs produced from the parsing program. Words are split up into categories: Preparation, Response, Impact, and Recovery

Table 3: Descriptions for files used in graphing tweets based on word frequency

Location Tracking		
File Name	**Extension**	**Description**
extract-geotags	.py	Takes in .csv of tweets, removes rows without geotags, changes date format to be sortable, creates list of users using geotags.
tweet-geocoder	.py	Takes in a .csv of geotags representing physical locations (i.e., Caguas, Puerto Rico), and outputs a .csv of coordinates.
PR_Earthquake_Location	.csv	Original dataset of tweets for location tracking
EQL_sample_***	.csv	Smaller subsets of original dataset

Table 4: Descriptions for files used in location tracking

Topic Analysis		
File Name	**Extension**	**Description**
2015ElectionGenerateTopic	.py/.ipynb	Takes in a .csv of tweets, outputs .dict, .lda, and .mm files that will be used to visualize the most common topics in the tweets
topic_vis_2015Election	.py/ipynb	Takes in .dict, .lda, and .mm files from 2015ElectionGenerateTopic and outputs visualizations of the most common topics in three

		categories: single-word topics, two-word topics, and multi-word topics
PR_Earthquake	.csv	The tweet dataset used for topic analysis

Table 5: Descriptions for files used in topic analysis

II. Required Software

All components of the Twitter Disaster Behavior project were completed using Python. Some code was written and run in an IDE such as PyCharm, while some were completed using Jupyter Notebook.

Word Frequency: Python, Matplotlib, NumPy

NumPy and Matplotlib were used for creating the graphs that show word frequency over time.

Location Tracking: Python, NumPy, Pandas, GeoPy, GetOldTweets, Twarc, Tableau

NumPy and Pandas were used for parsing through the tweet samples and narrowing down the geotagged users. The GeoPy package was used for converting Twitter geotags to coordinates. GetOldTweets (got) and Twarc are packages used for extracting larger sets of tweets from users. Tableau Data Visualization is used for creating heatmaps.

Topic Analysis: Python, Gensim, NLTK, pyLDAvis

Gensim, NLTK, and pyLDAvis are used for the Topic Generation code that identifies the most common single-word, two-word, and multi-word topics. pyLDAvis is used for creating the visuals of the topic analysis.

Lessons Learned

I. Timeline

Because the work on our project resulted from recent findings, it was difficult to construct a full timeline from the beginning of the project. Our team essentially had smaller timelines outlining our tasks and goals for the next one or two weeks. We set goals for the week at each meeting with our client. While it was difficult to predict what our overall workload would be for the whole semester, we found this style of work to be effective for a project like this. We made sure the work we took on each week was manageable. The final timeline for this project can be viewed in Appendix B.

II. Problems

Early on, our team did not encounter many problems in terms of cooperating on the project. However, as the semester went on, sometimes our schedules would not align. This was difficult, especially during the week before poster submissions for the ISCRAM conference were due. Aligning schedules has become increasingly difficult as the semester continues due to adjustments made by the university in response to the COVID-19 pandemic. As we no longer can meet in person, our meetings are conducted online. Because of this, all of our group members' internet quality impacts the quality of our meetings. There was no easy solution to this, as we are all in distant locations, but we made sure to keep all group members updated if they were unable to join a meeting due to technical issues.

III. Solutions

To aid with communication and organization, our team communicates via GroupMe and organizes all work in a shared Google Drive. We have one drive dedicated to the documents and presentations relating to the course, and another that is organized by our client to share code and resources used for the project. The shared drive for the course consists of folders for each assignment, such as the presentations, reports, or in-class assignments. This helps us organize our work so we can find relevant documents quickly. The client's shared drive has organized folders for the data we use in our code, some sample code, resources and drafts of our poster, and references to previous semesters' work.

As for the latter part of the semester, Zoom conference rooms have greatly assisted our team in staying in touch. We currently create a separate Zoom room after our class meetings so we can

make sure we are on track with our objectives. We have also set up weekly client meetings via Zoom to replace our weekly in-person meeting.

IV. Future Work

The types of analysis we conducted during this project could be expanded to other types of disasters that impact an area, not just earthquakes and hurricanes. The difference between public response to these disasters was significant and adding more data about other types of disasters in Puerto Rico could expand our understanding of how the island is prepared--or not prepared--to handle these events. Also, by comparing different disasters that impact the same area, our geotagging research can find repeat “trouble spots” in an area that are infrastructurally unsound and unprepared for repeat disasters. Recognizing these spots can help people in political office divert funds to improve these areas structurally so they do not continue to fail in the event of disaster.

Our code is not just limited to analyzing disasters in Puerto Rico, however; we tried to make the code modular, so that future researchers could modify it easily to analyze datasets of tweets about disasters in other parts around the globe. One global disaster at the forefront of minds recently is the COVID-19 pandemic, a disease not limited by country boundaries or economic strata. Analysis of this disaster in comparison to natural disasters such as earthquakes might produce interesting similarities, as both have strong effects on public health and the economy.

Acknowledgements

Client: Ziqian (Alice) Song (ziqian@vt.edu)

Professor: Edward Fox (fox@vt.edu)

"Collaborative Research: Coordinated, Behaviorally-Aware Recovery for Transportation and Power Disruptions" (NSF CMMI-1638207)

“Global Event and Trend Archive Research (GETAR)” (NSF IIS-1619028)

References

[1] R. Pasch, A. Penny, R. Berg, “National Hurricane Center Tropical Cyclone Report: Hurricane Maria”. *National Oceanic and Atmospheric Administration, National Weather Service,* Feb. 14, 2019. [Online]. Available: https://www.nhc.noaa.gov/data/tcr/AL152017_Maria.pdf. [Accessed: Apr 21, 2020].

[2] “Puerto Rico Earthquake Relief,” CharityWatch, Jan. 08, 2020. [Online]. Available: https://www.charitywatch.org/charity-donating-articles/puerto-rico-earthquake-relief. [Accessed: Apr. 21, 2020].

[3] “Fundraiser by Orlando Crum-Echevarria: Puerto Rico Earthquake Relief,” GoFundMe, Jan 08, 2020. [Online]. Available: https://www.gofundme.com/f/qah39f-puerto-rico-earthquake-relief. [Accessed: Apr 21, 2020].

[4] “Puerto Rico Earthquake Relief Funds”. The Happy NPO, Pure Charity, Jan. 08, 2020. [Online]. Available: https://www.purecharity.com/puerto-rico-earthquake-relief-funds. [Accessed: Apr. 21, 2020].

Appendix

Appendix A: Poster and Graphs

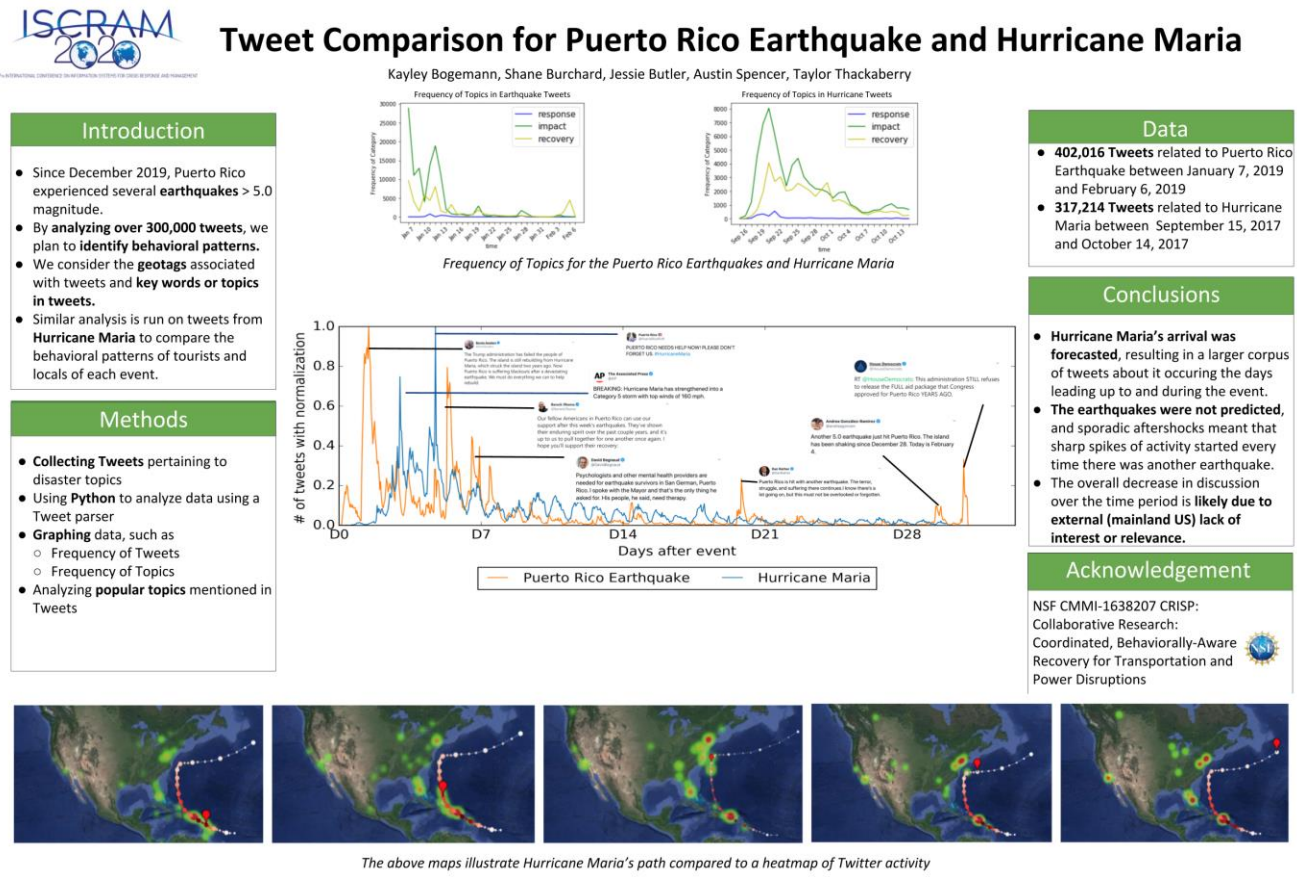

Figure 6: The poster created for the ISCRAM 2020 conference

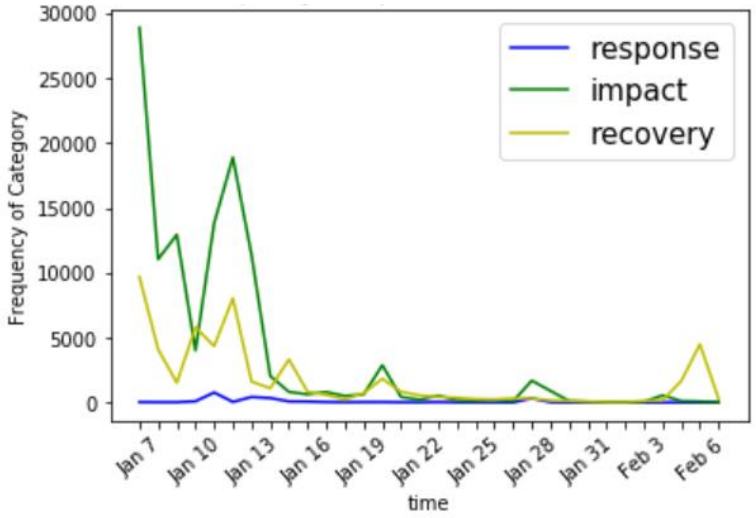

Figure 7: Frequency of topics in earthquake tweets

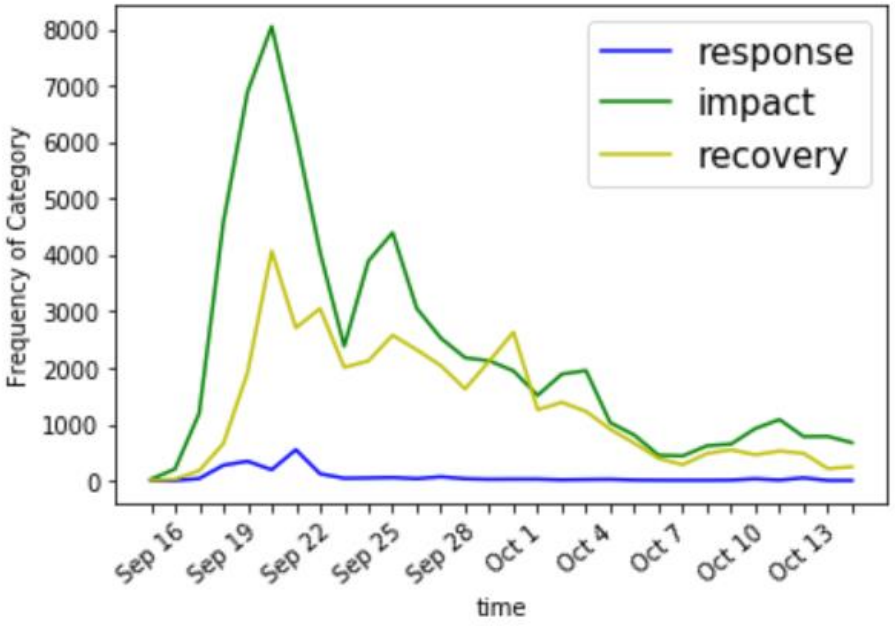

Figure 8: Frequency of topics in hurricane tweets

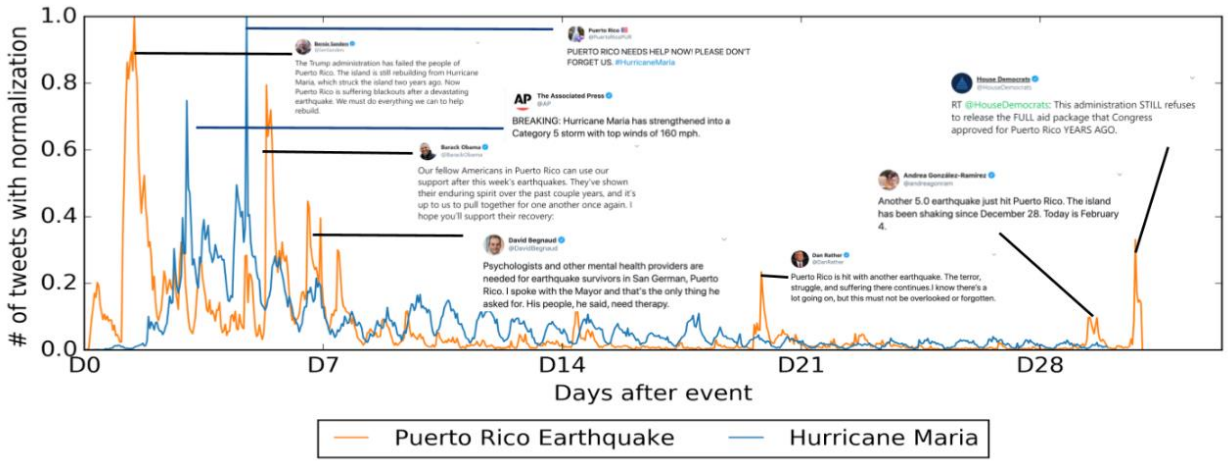

Figure 9: Graph showing number of tweets over time, emphasizing top tweets written during select peaks in data

Figure 10: Maps that illustrate Hurricane Maria's path compared to a heatmap of Twitter activity

Appendix B: Timeline

Date	**Description**
January 21, 2020	First group meeting and assignment of team roles
January 29, 2020	Began word frequency analysis on an initial, smaller set of data
February 4, 2020	Created smaller dataset of tweets containing geotags
February 29, 2020	First group presentation
March 2, 2020	Completed word frequency analysis for final datasets of earthquake and hurricane twitter data

March 4, 2020	Completed draft of ISCRAM poster
March 30, 2020	Created map of geotag location changes during and after earthquake and hurricane
April 5, 2020	Second group presentation
April 6, 2020	Conducted topic analysis on all tweets related to earthquake
April 20, 2020	Completed analysis on topics discussed by individuals vs. organizations
April 26, 2020	Initial submission to VTechWorks
May 4, 2020	Final presentation

Table 6: Final timeline describing work completed for this project