Topic Modeling Using Latent Dirichlet Allocation on Disaster Tweets

by

Virashree Hrushikesh Patel

B.S., Kansas State University, 2015

A Report

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Co-Major Professor
Cornelia Caragea

Approved by:

Co-Major Professor
Doina Caragea

Copyright

Abstract

Social media has changed the way people communicate information. It has been noted that social media platforms like Twitter are increasingly being used by people and authorities in the wake of natural disasters. The year 2017 was a historic year for the USA in terms of natural calamities and associated costs. According to NOAA (National Oceanic and Atmospheric Administration), during 2017, USA experienced 16 separate billion-dollar disaster events, including three tropical cyclones, eight severe storms, two inland floods, a crop freeze, drought, and wildfire. During natural disasters, due to the collapse of infrastructure and telecommunication, often it is hard to reach out to people in need or to determine what areas are affected. In such situations, Twitter can be a lifesaving tool for local government and search and rescue agencies. Using Twitter streaming API service, disaster-related tweets can be collected and analyzed in real-time. Although tweets received from Twitter can be sparse, noisy and ambiguous, some may contain useful information with respect to situational awareness. For example, some tweets express emotions, such as grief, anguish, or call for help, other tweets provide information specific to a region, place or person, while others simply help spread information from news or environmental agencies. To extract information useful for disaster response teams from tweets, disaster tweets need to be cleaned and classified into various categories. Topic modeling can help identify topics from the collection of such disaster tweets. Subsequently, a topic (or a set of topics) will be associated with a tweet. Thus, in this report, we will use Latent Dirichlet Allocation (LDA) to accomplish topic modeling for disaster tweets dataset.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Dr. Doina Caragea and Dr. Cornelia Caragea for giving me an opportunity to be part of their research and guiding me all the way. I express my gratitude to Dr. Doina Caragea for her time and effort spent on providing positive critique and suggestions that molded the outcome of the project and report. I thank Dr. Doina and Cornelia for introducing me to the field of Big Data Analytics and Natural Language Processing that equipped me with the knowledge to implement this project. I thank Dr. William Hsu for being part of my committee and providing valuable feedback regarding my project.

Dedication

All power is within you; you can do anything and everything - Swami Vivekananda

I dedicate this project to my parents who have always been my greatest support and strength.

Chapter 1

Introduction

During 2017, United States experienced a historic year of weather and climate disasters. According to NOAA (National Oceanic and Atmospheric Administration), in total, U.S. was impacted by 16 separate billion-dollar disaster events this year including three tropical cyclones, eight severe storms, two inland floods, a crop freeze, drought and wildfire[7]. As a consequence, 2017 had the highest number of billion-dollar disasters for a single year in U.S. history[7].

One of the first and immediate impacts of natural disaster is the collapse of communication infrastructure. Due to the widespread use of social media, more and more people are turning to social media platforms, such as Facebook and Twitter, to ask for help during natural disasters. Tweets posted during natural disasters can give valuable real-time information about the region and people affected by the disaster. Therefore, tweets can become a very powerful tool for rescue agencies and local government who execute relief operations for the victims.

Twitter allows users to send short text messages with a maximum length of 140 characters. These tweets can be collected using Twitter Streaming API service available for the public use. Tweets posted during natural disasters can be of different types. For example, some tweets express emotions, such as grief, anguish, or call for help, other tweets provide information specific to a region, place or person, while others simply help spread information

from news or environmental agencies. To extract information useful for disaster response teams from tweets, disaster tweets need to be cleaned and classified into various categories. Topic modeling can help identify topics from disaster dataset. Topic models are designed for well-designed text like newspapers, articles, blogs etc. that are mainly talks about a few selected topics and follow grammar rules. On the other side, short texts received from social media platform such as Twitter, have a limited contextual information, and they are sparse, noisy and ambiguous. Therefore, learning topics from them can be challenging[8]. In this project, I have used Latent Dirichlet Allocation to accomplish this task. Subsequently, a topic will be associated with a tweet. In this report, I will discuss data collection and cleaning process, the LDA model used to derive topics from the corpus, and finally, I will discuss some results derived from the topic model.

1.1 Data Collection

My study used Twitter data collected during Hurricane Harvey, Hurricane Irma, Hurricane Maria, and Mexico Earthquake. Twitter provides developers a streaming API to collect tweets in real-time. Using this Twitter API, I collected about 248.2 Gb of data related to these naturals disasters. The data obtained from Twitter API is in JSON format. One such tweet object example is shown in Figure 1.1.

In addition to the text content itself, a Tweet can have over 150 attributes associated with it. Several of these attributes can be missing from a tweet. Therefore, the first is to extract information that is important to us and get rid of unnecessary data from a tweet.

For the purpose of this project I have focused on the following fields:

created_at: UTC time when a tweet was created

tweet_id: Unique id associated with every tweet

text: UTF-8 text related to the tweet

The above fields were extracted from the JSON file containing the original tweets into a CSV file. The main reason to switch to the CSV format is to increase readability of data.

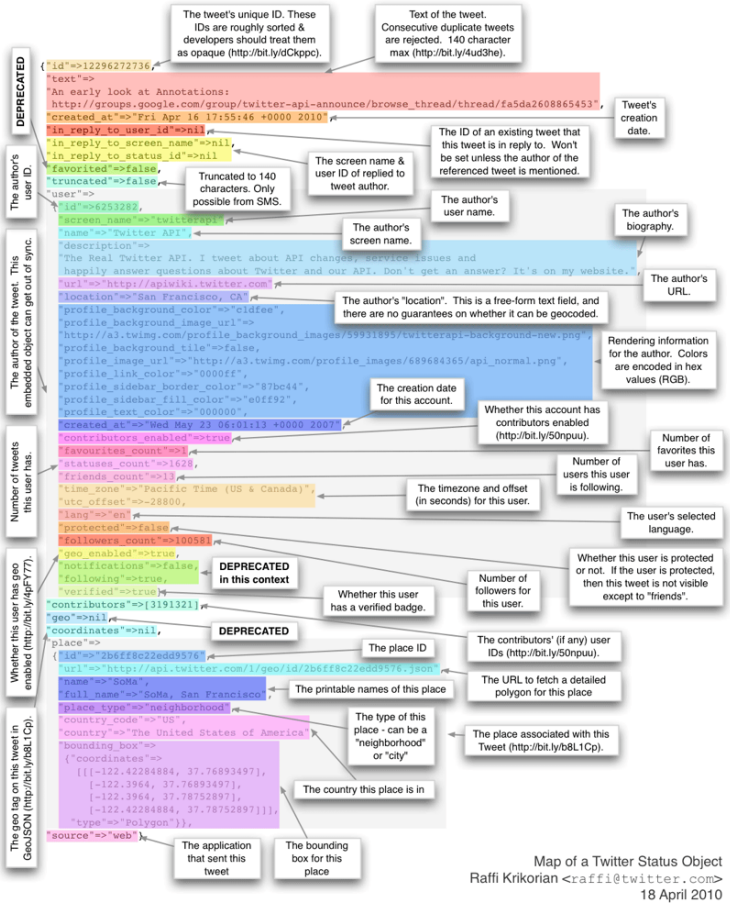

Figure 1.1: *Tweet Object*

Figure 1.2 shows the number of tweets collected per day during the course of hurricane Irma, Harvey and Maria. It is clear that all three hurricanes timeline and course of events had a lot of overlap among them. In our initial Twitter data collection process, we made an error of collecting data for all the hurricanes using a single Twitter stream with keywords related to all the hurricanes, instead of collecting data for all the hurricanes separately using separate Twitter stream. This added an overhead, later on, to classify data into different disasters.

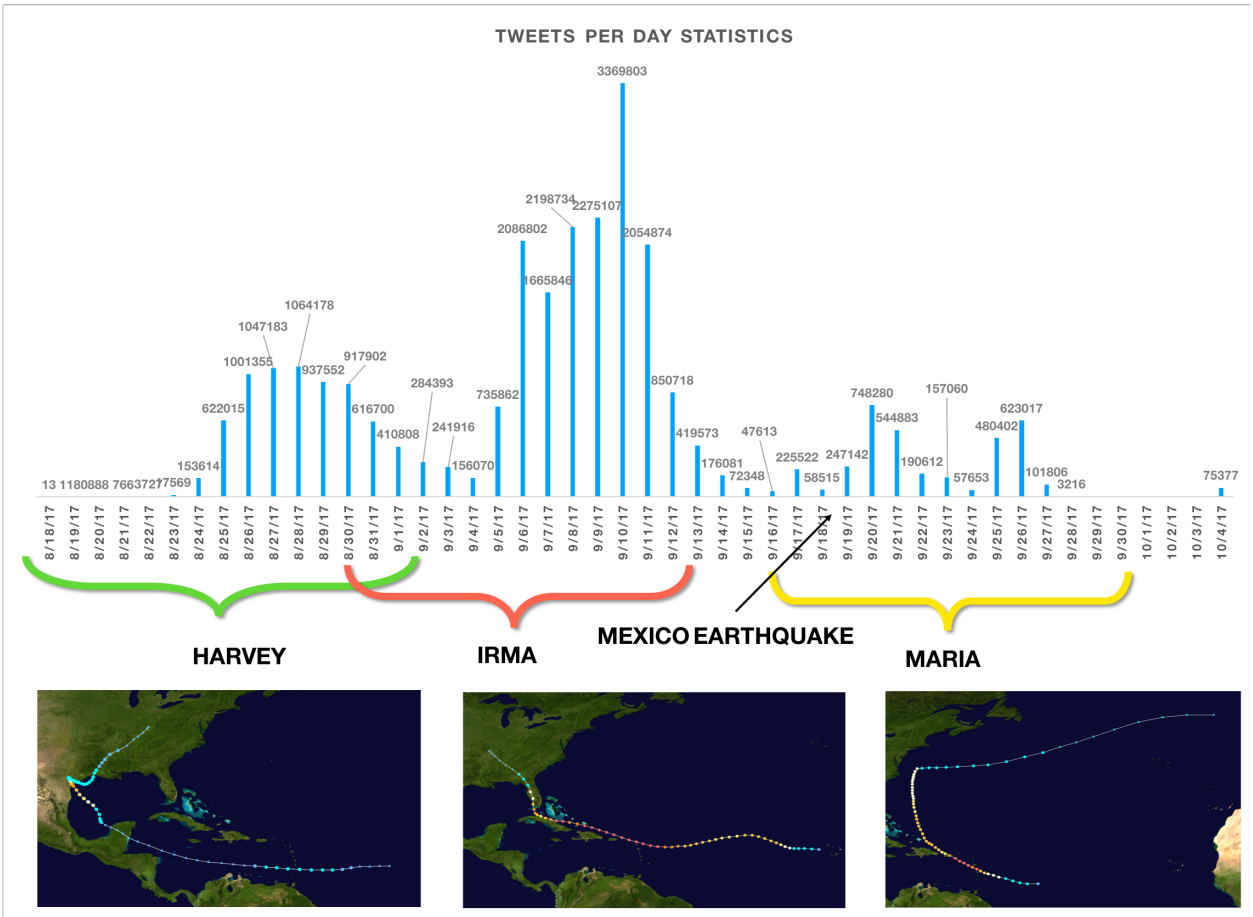

Figure 1.2: *Number of Tweets Per Day Statistics*[1]

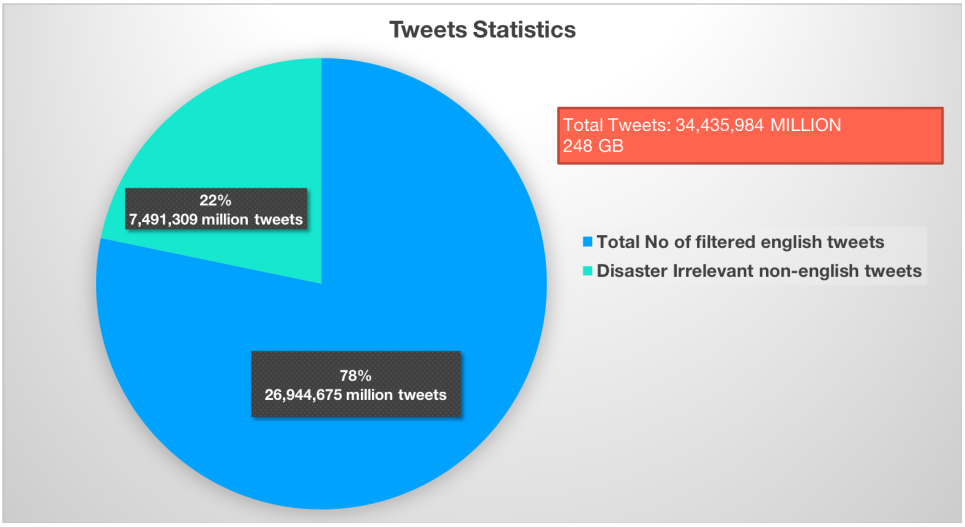

Figure 1.3: *Tweets Statistics Unused non-English disaster irrelevant tweets vs Used English tweets*

For the purpose of topic modeling task, we are treating all the tweets from all three hurricanes as a single data set. We are more interested in observing data per day or hour, rather than by disaster. Therefore, this data collection error does not make any difference to the topic model. Also, note that the Twitter data stream does not provide unique tweets due to the retweet feature of Twitter. Therefore, in our data collection, we have plenty of duplicate tweets that has same tweet text but are posted by a different. Retweeted tweets can provide useful information about the importance of a certain topic. So, in the case of topic models, we have not considered getting rid of the retweets.

Figure 1.3 shows the number of tweets per day statistics for our disaster dataset with the timeline of each hurricane and Mexico earthquake. Note that we do not have data for the start of hurricane Harvey and for the end of hurricane Maria. Also, tweets collected during the timeline of a hurricane need to be exclusive for the hurricane. For example, tweets collected during hurricane Irma can have tweets related to hurricane Harvey as well.

Chapter 2

Data Preprocessing

Before performing the actual LDA analysis on the Tweets, the disaster data-set needs to be processed to reduce potential *"noise"* which is inherently present in social media posts[9]. To ensure the accuracy of the topic model, the quality of the data is a very important part. A tweet's text is limited to 140 characters and consists of URLs, user mentions, numbers, hashtags, slang words, abbreviations, emoticons, irregular punctuation[10]. Linguistic noise present in the Twitter text makes extensive preprocessing a requirement, although on the other side we might lose some valuable information from the limited text that is available to us about a tweet.

The following paragraphs summarize the preprocessing steps performed on the disaster dataset.

2.1 Classifying English tweets

For the purpose of this project, only English tweets are used. Tweet objects lang field tells us about the machine detected tweet language. Based on that, English language tweets are separated. Note that it is possible to lose some data here as it is observed that some tweets are not detected as English tweets even though they are in English.

2.2 Classifying disaster relevant tweets

Tweets from the Twitter API are based on keywords that the user crawling data specifies. However, not all the tweets received from Twitter are relevant to a disaster of interest. For example, for a keyword like "power" or "maria", the crawled dataset might have tweets that use power and maria in a different context than "power outage" or "hurricane Maria," respectively. Therefore, it is essential that we further classify our dataset into relevant tweets and non-relevant tweets. For this purpose, a pre-trained Bernoulli Naive Bayes classification model was used. This model is designed by HongMin Li, PhD student in Computer Science at Kansas State University. The classifier was trained on using six disasters. The cleaning process of the training data involved using placeholder for disaster references, user mentions, hashtags and so on. The cleaned data then was used to to trian Bernoulli Naive Bayes that can filter out disaster irrelevant tweets from a given dataset.

For the LDA modeling, we are only focusing on the textual part of data, which is the tweet text. However, the tweet text itself can have a lot of "noise" that can deteriorate the quality of the model. It is necessary to remove the noise before the LDA model can be applied.

2.3 Removing emoticons

Use of emoticons on social media platform and text messaging is popular to express emotions in messages and they can give very useful information about emotions and sentiments related to a tweet. Although, emoticons are not standardized and thus they are hard to interpret[11]. Therefore, we remove emoticons from tweet text.

2.4 Removing user mentions

On Twitter users can tag other users using their Twitter handle. Therefore, tweet text often has user mentions starting with @ symbol. We are only interested in the disaster-related

content of a tweet. So, usernames do not contribute to disaster-related topics for our topic modeling purpose. If we only choose to remove special character @ and do not remove the username, it can cause significant errors in our model, because names are classified as nouns and topics derived from LDA will consist of nouns and adjectives. Therefore, usernames are also removed.

2.5 Removing URLs

URLs are also not used in our topic modeling approach because they contain unspecific and hardly interpretable semantic information[11]. Thus, URLs are removed from the tweet text.

2.6 Removing numbers

Numbers do not contain semantically viable information for the purpose of topic modeling[11]. In some cases, digits can give valuable information also, for example "death toll", "magnitude" of an earthquake. Regardless of numbers, we will still be able to catch these words in our topic model to describe the topics in a document sufficiently.

2.7 Removing special characters

Apart from emoticons, tweets can also contain other special characters. For example, one case can be the use of special characters like a semi-colon, parenthesis, etc. in combination with each other in place of emoticons such as ;), :D. In addition to this, sometimes tweet text contains hashtags, which start with the special character #. Hashtags can give very valuable information regarding a tweets topic. Removing special characters can help retain the hashtag information.

2.8 Convert tokens to lowercase

This is a required step to make the tokens more similar, reduce typos, and prevent words with the same spelling in the vocab. Tokens with different cases can be treated as two different tokens, For example, tweet hashtags such as HurricaneMaria, hurricanemaria, hurricaneMaria, Hurricanemaria can exist in a document at the same time. If these tokens are not converted to lowercase, they will be treated as different tokens, even though they hold the same meaning. But words that do hold special grammatical meaning due to the case cannot be differentiated anymore. However, this is not a huge issue in the English language except for nouns like a place or country name, company name, brand names etc.[11].

2.9 Removing stop words

Stop words do not carry any semantical meaning of their own (e.g. auxiliary verbs, conjunctions, and articles). Stop words are present in almost every document and would form a topic on their own because they co-occur frequently with other words[11]. I have removed the stop words using a predefined list of stop words provided by the NLTK toolkit. Apart from the stop words provided by NLTK, I have added some other stop words which are 'rt', 'aint', 'gonna', 'wanna', 'amp', 'htt', 'http', 'https'.

2.10 Tokenization

In this step, the tweet text is converted into a list of tokens by dividing them using white-space. These tokens will be later used in our model to find topics. For tokenization, I have used Gensim's library function simple_preprocess. This function, by default, ignores words that are too short or too long. By default, the minimum length for the word is set to 2 and the maximum length is set to 15. In addition to this, simple_preprocess also removed letter accents from the words. These features greatly help further cleaning our dataset.

2.11 Bi-grams, Tri-grams and Quad-grams

To construct Bi-grams, Tri-grams, and Quad-grams, I have utilized Gensim's Phrases and Phrasers library functions. Gensim's Phrases library functions to detect n-grams is inspired by the paper by Mikolov, et. al[12] "Distributed Representations of Words and Phrases and their Compositionality", published by researchers at Google. Gensim Phrases model is trained using parameters specified by the user, i.e. min_count, threshold, max_vocab_size. The default values of these parameters are 5, 10 and 40,000,000 respectively. The min_count specifies how many times a word needs to appear in the document to be considered. If for a word total collected count is lower than min_count that word will be ignored. By default, this model uses the function called original_scorer, which follows a scoring technique specified by the paper mentioned above. The score for an n-gram will determine if it will be added to the vocabulary. The model suggested by Mikolov in the paper mentioned above calculates the score for a bi-gram by taking into account co-occurrences of words participating in the bi-gram. If the score is above the threshold value, then the bi-gram will be added to the vocabulary. Same goes for tri-grams and quad-grams, where one of the words will be bi-gram and tri-gram respectively. The higher the value of min_value and threshold, the harder it will be for words to combine into bi-grams, tri-grams, and quad-grams.

I have selected min_count and threshold values 200 and 500 respectively. There are 1380 k-grams in vocabulary. Full vocabulary can be viewed under this URL under file vocab.csv. https://github.com/virashree/Topic-Modeling

2.12 Lemmatization

I have done lemmatization using spaCy. Lemmatization returns the base or dictionary form of a word, which is known as the lemma. This will help condense the corpus and increase the efficiency of the topic model without affecting the results. spaCy has a lot of capabilities, such as tokenization, sentence recognition, part of speech tagging, lemmatization, dependency parsing, named entity recognition. Using spaCy's part of speech tagging capability, I was

able to keep only nouns and adjectives along with lemmatization. I am only considering nouns and adjectives as they contain the most meaningful information about the topics.

Chapter 3

Latent Dirichlet Allocation Model

To extract the topics from the tweets, I have used the LDA (Latent Dirichlet Allocation) model. The LDA model assumes that a document is a mixture of topics (a probability distribution over topics), where a topic is a mixture of words (a probability distribution over words) in the collection. LDA model describes a simple probabilistic approach by which documents can be generated from the topics, and topics from words. To generate a document, one first chooses a topic distribution. Next, one chooses a topic form this distribution at random (latent) to form words for the document[3]. In other words, we are trying to reverse engineer documents using the topic distribution of the document, and word distribution of the topic. Due to these properties, the LDA model is classified as a generative model.

For the purpose of this project, we will consider LDA as a problem of statistical inference. Figure 3.1 illustrates the word distribution of a topic and the topic distribution of a document. Figure 4.7 shows LDA as a problem of statistical inference. In the case of statistical inference, the goal is to find the topic model that have most likely generated the document[3].

On the other hand, in the case of the generative model approach goal is to find the best set of words that can describe the document[3].

As an example of a topic model, in Figure 3.3, I have shown two topics related to money and river[3].

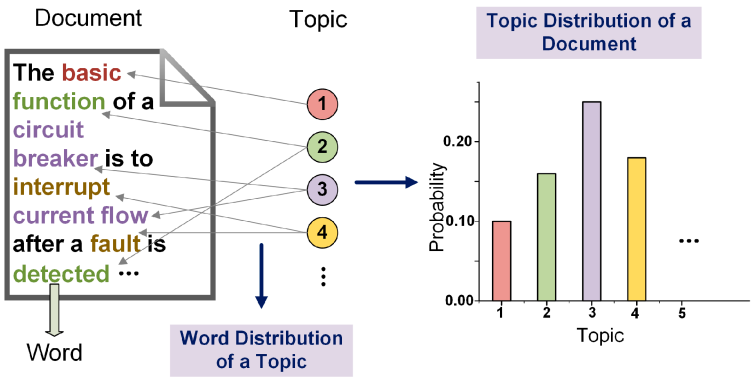

Figure 3.1: *LDA model illustration*[2]

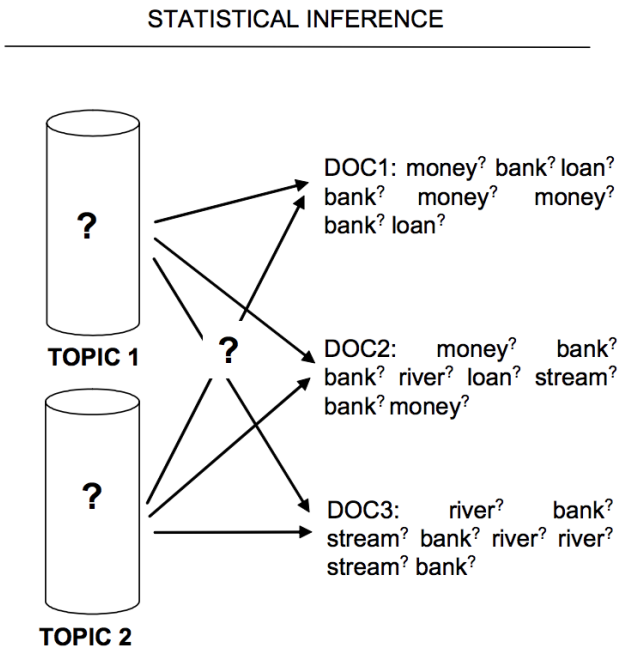

Figure 3.2: *LDA - illustration of the inference problem*[3]

Now how many ways we can generate documents using these two topics. There can be several ways. Three of such approach can be similar to what is shown in Figure 3.3 that shows three possible documents. Two documents are entirely generated using topic 1 and topic 2, respectively. While one more document can be generated using a mixture with an equal percentage of words from topic 1 and topic 2. Note that the superscript numbers on the words in a document denote which topic the word belongs to. Also, note that a word can

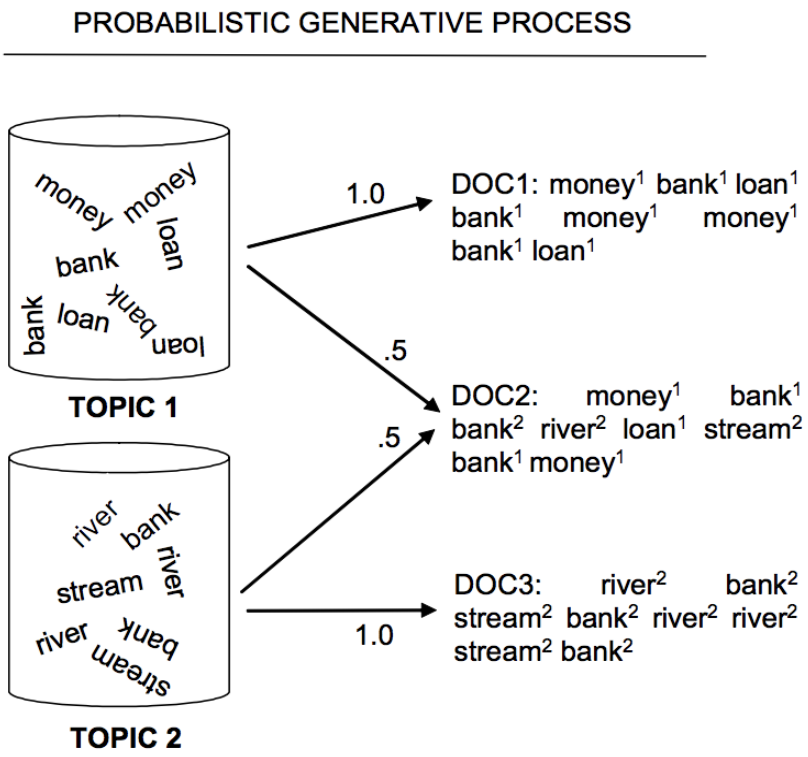

Figure 3.3: *LDA - illustration of the generative model*[3]

be part of any number of topics. For example, in Figure 3.3 notice that the word "bank" is part of both topics, although it can assume different meanings when used with other words. i.e. *river bank*, *banks money*, etc. This allows the topic model to capture the *polysemous* (the coexistence of many meanings for a word or phrase) nature of words.

The generative approach follows the *bag of words* model, which does not consider the order in which words appear in the document. However, the model can be modified to give importance to the order of words to capture some important properties of documents[3]. In our case, we have considered bi-grams, tri-grams, and quad-grams.

3.1 Parameters for LDA

I have used Gensim's ladmodel and ldamulticore model to derive topics for my disaster dataset. I have used following parameters for these LDA model[13].

corpus (Mandatory): Represents corpus in form of a stream of document vectors. It

represents a mapping of words ID in id2word dictionary and document
num_topics (Mandatory): Number of topics requested from the corpus
id2word(Mandatory): Mapping of vocabulary words and with an integer ID
passes(Optional): Number of passes through corpus
chunksize(Optional): Number of documents used in training the model. These many documents are loaded in the memory at once and model is updated. Having bigger chunksize makes the model more efficient
workers (Optional): Number of worker processes to be used (only used for ldamulticore model in Gensim)
iterations(Optional): Maximum number of iterations through corpus

I have taken two approaches to derive topics from the dataset in this paper. We will discuss each approach in chapter 4 in more detail.

I have used the following values for the parameters for the deriving topics from whole corpus for approach 1.
num_topics = 80
passes = 1
chunksize = 100000
workers = 1
iterations = 1000

I have used following parameters for deriving topics for the files divided by date for approach 2.
num_topics = 20/10/8/5
passes = 1
chunksize = 100000/50000/10000
alpha=0.001

Chapter 4

Results

We will now look closely at the results obtained from topic modeling. Data is divided according to date tweet was posted. Each tweet is considered a document. Therefore, we have a large number of documents in our corpus. From figure 3 we can tell there are over 34 million total documents in the corpus. From figure 4, shows how many tweets are there per day i.e. documents are there per day.

We will consider two different approaches to derive topics and observe results for each.

Approach 1: Whole dataset as corpus

In this approach, I have merged all documents by date into one document and applied the LDA model to this merged document. This trained model is applied to tweets separated by date and to tweets further separated by an hour for a day. This approach will give us information about how topics for different dates or different hours in a day are related to each other. We will be able to hopefully observe a hurricane timeline.

Approach 2: Document by date as a corpus

In this approach, each document by date is considered a corpus in itself and the LDA model is applied to each document and topics are derived. This approach will provide a detailed look at what topics were discussed on each day. We will be able to track unique events that

happened on that particular day using this approach.

The drawback of this approach is that vocabulary will differ from day to day. I am considering bi-grams, tri-grams, and quad-grams for LDA model as part of the vocabulary. Depending on the number of times a term appears on each day these k-grams will differ from document to document because Gensim's Phrases class derives k-grams based on term frequency in the corpus. To overcome this limitation, I have taken an innovative approach. As mentioned previously, before running the LDA model we will first create a streaming corpus using Gensim that will feed documents sequentially to LDA model. This will not load the whole corpus in memory at once and will give us an advantage in speed and efficiency. While generating the streaming corpus for the whole dataset in approach 1, I am also keeping track of cleaned tweets by day. Before merging all files for approach 1, I have separated all date files using term 'eofeofeof'. Therefore, the cleaned merged file after pre-processing steps contains 'eofeofeof' term separating the tweets, which can be again separated as date documents. By using the corpus and vocabulary created for the entire corpus in approach 1 to recreate documents by date, I am now making sure to use common vocabulary for all the documents.

4.1 Analysis

we will look closely at each approach and try to interpret our results in the following sections.

4.1.1 Dataset as corpus

For approach 1, the first step is to train the LDA model on the whole corpus and then this trained model is applied to files separated by date or hour. Gensim's get_document_topics function returns topic probability distribution for a given document in form of a topic number and topic probability. Using this function topic probability distribution is derived for each day or hour.

For approach 1, I have derived 80 topics from the whole corpus. I have listed some sam-

ple topics in 4.1 with 10 words as a limit. All 80 topics can be found under this link. https://github.com/virashree/Topic-Modeling/blob/master/Topics_Corpus_80_10Words.txt

As we discussed in chapter 3, the LDA model assumes a document is a mixture of topics. It generates the documents by first picking a topic distribution and selecting a topic and then using the topic to generate a word for that document. The numbers in front of each word in the topics in Table 4.1 shows the probability of selecting the word after the topic has been selected. In my analysis, I have chosen to observe topics till 20 words, although for the most part, we are going to be looking at the first 10 words in a topic as it gives suffice information.

One can identify the tone of a given topic from the terms that comprise it. Therefore, I have developed certain rules for topic interpretation. Each topic is tagged with following four set of rules. First, we try to identify the main tone of a topic. For example, if the overall tone of a topic is about weather information, relief efforts or politics etc. Secondly, we will tag a topic based on preparedness, during the disaster, aftermath or other in case it is about something else. We want to know if a topic is related to a particular disaster. This can be identified by place mention or direct disaster reference in a topic. For example, if the topic explicitly mentions "*hurricaneIrma*" then one can say that it is more related to hurricane Irma. But instead, there is mention let's say "*houston*" than it is more related to hurricane Harvey.

The derived 80 topics for approach 1 are tagged using following rules.

Rules 1: Determine the main tone

1 : News

2 : Emotions

3 : Urgent help people

4 : Rescue

5 : Relief

6 : Money

7 : Disaster News Damage / witness accounts / evacuation / donation / event/ fraud /

causality

8 : Weather Caution

9 : Weather News

10 : Place mentions

Rule 2: Preparedness/During/Aftermath

P: Hurricane Preparedness

D: During the disaster

A: Aftermath

O: Other

Rule 3: Disaster

M: Hurricane Maria

H: Hurricane Harvey

E: Mexico earthquake

I: Hurricane Irma

Rule 4: Place mention

Listed place mentions if there are any mentioned in the topic.

Using these rules we will first determine what is the main tone of the topic is, whether it is more related to hurricane preparedness, during the hurricane or aftermath and hopefully, it will also tell us what disaster it is in relation to. Using these set of rules I have tagged all 80 topics.

These rules will help us plot the timeline of a hurricane when we apply the LDA model trained on the whole corpus to individual days or hours. It will help us answer questions like what disaster the topic can be referring to in the given context or whether it is more related to events before the hurricane or during the hurricane etc.

Topic No	***Topics***
1	0.229*"tonight" + 0.185*"landfall" + 0.116*"central" + 0.103*"tomorrow" + 0.041*"open" + 0.040*"communication" + 0.029*"economic" + 0.028*"individual" + 0.021*"evacuee" + 0.018*"martin"
6	0.351*"storm" + 0.210*"building" + 0.089*"tropical" + 0.052*"ship" + 0.049*"atlantic" + 0.044*"hurricane" + 0.029*"season" + 0.027*"system" + 0.025*"depression" + 0.014*"nhc"
20	0.482*"relief" + 0.087*"great" + 0.063*"red" + 0.061*"hurricane" + 0.058*"donation" + 0.058*"fund" + 0.045*"cross" + 0.031*"volunteer" + 0.027*"worker" + 0.020*"urgent"
68	0.403*"trump" + 0.299*"president" + 0.089*"usns" + 0.075*"fema" + 0.025*"leader" + 0.015*"pre" + 0.012*"cut" + 0.011*"security" + 0.008*"bc" + 0.008*"fraud"
64	0.616*"earthquake" + 0.080*"day" + 0.052*"week" + 0.049*"magnitude" + 0.015*"broken" + 0.014*"break" + 0.013*"next" + 0.009*"several" + 0.009*"tragedy" + 0.007*"oaxaca"
74	0.739*"Mexico" + 0.058*"part" + 0.046*"money" + 0.023*"climate" + 0.020*"change" + 0.013*"catastrophe" + 0.013*"amazing" + 0.011*"risk" + 0.010*"healthcare" + 0.010*"view"

Table 4.1: *Sample topics derived from whole dataset with word limit of 10*

4.1.2 Analysis by Day

As described before Gensim's get_document_topics function provides topic distribution for a given document. I have used this function to get topic distribution for each day. For simplicity, I have used topics that have a probability above 0.02. For smaller files, I have picked topics with probability more than 0.035.

Date	***Topic Distribution***
9-21-2017	[(30, 0.08480384), (64, 0.072067425), (74, 0.07150304)]
9-22-2017	[(6, 0.050151225), (38, 0.05847761), (57, 0.050108533), (64, 0.1537913), (74, 0.15533683), (77, 0.08801828)]
9-23-2017	[(20, 0.054945845), (64, 0.19351476), (74, 0.118808776), (77, 0.05394344)]
9-24-2017	[(64, 0.14173648), (74, 0.085296795)]
9-25-2017	[(30, 0.16315438)]
9-26-2017	[(30, 0.17248002)]
9-27-2017	[(30, 0.15499169)]
9-28-2017	[(64, 0.18130434), (74, 0.15396737)]
10-4-2017	[(30, 0.11225968), (68, 0.053644)]

Table 4.2: *Topic distribution of topics with probability greater than 0.05 by date*

Table 4.2 shows topic distribution for some days with topics probability greater than 0.05. Topic distribution for all 43 days can be found under the following https://github.com/virashree/Topic-Modeling/blob/master/Topics_by_date_all.csv. We will try to interpret what people mostly talked about on a particular day by interpreting the topics in topic distribution. While doing this, the first precedence is given to topic with the highest probability and then one with probability less than that and so on.

For example, September 24 has topic 64 and 74 in it's topic distribution. From table 4.1 we can say that there were a lot of tweets regarding Mexico earthquake and climate change

on this day. This way I have interpreted topics for all 43 days. Results are shown in the table 4.3.

Table 4.3: Topic interpretation by Day considering, dataset as corpus

Date	Topic Interpretation	Timeline
8-19-2017	Advisories regarding tropical storm hurricane Harvey from NHC (National Hurricane Center), weather alerts regarding tropical depression	P-Harvey
8-20-2017	Advisories regarding tropical storm hurricane Harvey from NHC (National Hurricane Center), weather alerts regarding tropical depression	P-Harvey
8-21-2017	Advisories regarding tropical storm hurricane Harvey from NHC (National Hurricane Center), weather alerts regarding tropical depression, Reference to the Caribbean probably the Caribbean Sea or Island	P-Harvey
8-22-2017	Advisories regarding tropical storm hurricane Harvey from NHC (National Hurricane Center), weather alerts regarding tropical depression, Reference to Texas and some Island	P-Harvey
8-23-2017	Advisories regarding tropical storm hurricane Harvey from NHC (National Hurricane Center), weather alerts regarding tropical depression, Referring to advisories issued for the coast of Texas or some island	P-Harvey
8-24-2017	Hurricane Harvey track or forecast or strength, tropical storm warnings from NHC, reference to Texas or some island, weather news regarding wind speed or hurricane category	P-Harvey

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
8-25-2017	Attention or warnings regarding Hurricane Harvey, popular hashtag hurricaneHarvey, weather updates regarding the category of hurricane Harvey and wind speed, Reference to Texas, tweets displaying sentiments, emotions, thoughts and prayers, News regarding FEMA(Federal Emergency Management Agency) and presidency	D-Harvey
8-26-2017	Attention or warnings regarding Hurricane Harvey, popular hashtag hurricaneHarvey, weather updates regarding the category of hurricane Harvey and wind speed, Tweets regarding Texas, news about landfall	D-Harvey
8-27-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, reference to Houston, Texas area,dialysis patients related news, relief efforts by volunteers, Red Cross, relief donation and fund request, food for people or animal, weather news regarding rain, news regarding dogs and other pets, damage done by hurricane	D-Harvey

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
8-28-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, reference to Houston, Texas area, relief efforts by volunteers, Red Cross, relief donation and fund request, dialysis patients related news, News regarding damage to infrastructure, floodwater, shelter, tweets related to help, urgency, emergency ; probably helpline numbers, people in need of help or assistance, news regarding fire, from the context of real events possibly tweets regarding fire at Virgin Mary Statue, possibly news regarding effect on agriculture	D-Harvey
8-29-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, reference to Houston, Texas area, dialysis patients related news, probably tweets related to showing unity during crisis with phrases like fellow citizen, fellow Texas people etc.	A-Harvey
8-30-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, reference to Houston, Texas area, probably images of flood damage, tweets in relation to aged people, medical or health information, emergency information, news regarding fire, from the context of real events possibly tweets regarding fire at Virgin Mary Statue	A-Harvey

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
8-31-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, reference to Houston, Texas area, search & rescue missions and recovery	A-Harvey
9-1-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, reference to Texas, Possibly some tweets regarding earthquake ; although from the context of real events there wasn't any news of any major earthquake, News related president Trump, FEMA, Tweets regarding pets, hurricane relief & donation, hurricane survivors & victims, search & rescue missions and recovery, reference to Houston, Texas and some talk regarding climate change, possible use of hashtag "actonclimate"	A-Harvey
9-2-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, hurricane survivors & victims related tweets, reference to Houston, Texas, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, disaster response, and recovery	A-Harvey

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-3-2017	Tweets related to Hurricane Harvey updates, hurricane survivors & victims, popular hashtag hurricaneHarvey, reference to Houston, Texas, news related to president Trump, FEMA, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, sentiments or emotions,thoughts and prayers, from the emphasis on words prayer and president trump possibly referring to National Day of Prayer for Hurricane Harvey victims	A-Harvey
9-4-2017	Tweets related to Hurricane Harvey updates, popular hashtag hurricaneHarvey, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, hurricane survivors & victims, search & rescue missions and recovery, reference to Texas	A-Harvey
9-5-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings and alerts, use of popular hashtag hurricaneIrma, tweets related sentiments and emotions, thoughts and prayers, reference to Florida, some tweets for hurricane Harvey, and an earthquake	D-Irma, A-Harvey
9-6-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings and alerts, use of popular hashtag hurricaneIrma, reference to Puerto Rico, tweets related sentiments and emotions, thoughts and prayers, talk about an earthquake	D-Irma

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-7-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings and alerts, use of popular hashtag hurricaneIrma, reference to Florida, tweets related to hurricane Harvey, reference to Puerto Rico	D-Irma, A-Harvey
9-8-2017	Hurricane Irma related updates, reference to Florida, weather-related news such as wind speed, category, tropical storm warnings and alerts, use of popular hashtag hurricaneIrma, tweets related sentiments and emotions, thoughts and prayers, tweets related to hurricane Harvey, reference to Mexico, possibly talk about climate change	D-Irma, A-Harvey
9-9-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings and alerts, use of popular hashtag hurricaneIrma, reference to Florida, tweets related sentiments and emotions, thoughts and prayers	D-Irma
9-10-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings, alerts and advisories, use of popular hashtag hurricaneIrma, reference to Florida, Puerto Rico, Dominica, Miami, key west, tweets related sentiments and emotions, thoughts and prayers, water damage	D-Irma

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-11-2017	Hurricane Irma weather-related news such as wind speed, category, tropical storm warnings, use of popular hashtag hurricaneIrma, reference to Florida, Puerto Rico, Dominica, Miami, key west, political news related to president trump, some news related to New York	D-Irma
9-12-2017	Hurricane Irma tweets, use of popular hashtag hurricaneIrma, weather-related news such as wind speed, category, tropical storm warnings, reference to Florida, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, hurricane survivors & victims, some tweets related to hurricane Harvey	D-Irma, A-Harvey
9-13-2017	Hurricane Irma tweets, reference to Florida, hurricane Harvey related tweets, weather-related news of hurricane Irma, humanitarian aid & support, reference to Mexico, talk about climate change	A-Irma
9-14-2017	Hurricane Irma tweets, hurricane Harvey related tweets, weather-related news, reference to Florida, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, disaster response, power	A-Irma, A-Harvey

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-15-2017	Hurricane Irma tweets, hurricane Harvey related tweets, weather-related news, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, humanitarian aid & support, disaster response, power, reference to Florida, tweets related sentiments and emotions, thoughts and prayers	A-Irma, A-Harvey
9-16-2017	Hurricane Irma tweets, hurricane Harvey related tweets, weather-related news, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, reference to Florida, information regarding hurricane way humanitarian aid & support, disaster response, power, possibly alerts regarding another hurricane on the way	A-Irma, A-Harvey
9-17-2017	Hurricane Irma related tweets, tropical storm warnings, alerts and advisories, weather news such as rain, wind speed, hurricane category, Hurricane Maria forecast, track or strength, reference to Florida, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, recovery	A-Irma, A-Harvey, P-Maria
9-18-2017	Tweets regarding hurricane Irma, Weather related news such as wind speed, hurricane category, hurricane Harvey related tweets, popular hashtag hurricaneHarvey, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, search & rescue missions and recovery, humanitarian aid & support, reference to Florida	A-Irma, A-Harvey, P-Maria

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-19-2017	Reference to Puerto Rico, Hurricane Maria's forecast, strength, track, wind speed, the category of hurricane, tropical storm warnings, tweets related Hurricane Irma, Reference to Dominica, Miami, Key West	A-Irma, D-Maria
9-20-2017	Reference to Puerto Rico, Mexico, tweets related to hurricane Maria forecast, hurricane category, strength, wind speed, tropical storm warning, Mexico earthquake, earthquake magnitude, climate change, popular hashtag fuerzaMexico, tweets related sentiments and emotions, thoughts and prayers, hurricane Irma tweets	A-Irma, D-Mexico, D-Maria
9-21-2017	Reference to Puerto Rico, Mexico, tweets related hurricane Maria forecast, strength, tropical storm warning, Mexico earthquake, earthquake magnitude, climate change, popular hashtag fuerzaMexico, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, search & rescue missions and recovery, tweets related to hospital, children	D-Maria, A-Irma, A-Mexico
9-22-2017	Reference to Mexico, talk about climate change, Mexico earthquake magnitude, popular hashtag fuerzaMexico, tweets regarding people affected by hurricane or earthquake, hospital, child, tweets regarding relief efforts by volunteers, Red Cross, relief donation and fund request, search & rescue missions and recovery, reference to Puerto Rico, hurricane Maria strength and forecast	A-Mexico, D-Maria

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-23-2017	Tweets regarding Mexico earthquake and earthquake magnitude, popular hashtag fuerzaMexico, reference to Mexico, tweets about climate change, relief efforts by volunteers, Red Cross, relief donation and fund request, affected people, search & rescue missions and recovery, tweets related to victims and survivors, hurricane Maria strength and forecast	D-Maria, A-Mexico
9-24-2017	Tweets regarding Mexico earthquake and earthquake magnitude, climate change, hurricane Maria strength, forecast, tweets regarding people affected, hospital, child, tweets regarding relief efforts by volunteers, reference to Mexico, Puerto Rico, search & rescue missions and recovery, people affected by hurricane or earthquake	A-Mexico, D-Maria
9-25-2017	Reference to Puerto Rico, Mexico, tweets related to hurricane Maria, people in need, president Trump's golf habit, tweets related FEMA, humanitarian aid & support, possibly news related governor or Puerto Rico, climate change	A-Mexico, A-Maria
9-26-2017	Reference to Puerto Rico, hurricane Maria and earthquake tweets, humanitarian aid & support, people in need, president Trump's golf habit, tweets related to FEMA, hurricane way or path alert, tweets regarding Puerto Rico governor, power outage	A-Mexico, A-Maria

Continued on next page

Table 4.3 – continued from previous page

Date	Topic Interpretation	Timeline
9-27-2017	Reference to Puerto Rico, hurricane Maria and earthquake tweets, tweets related to humanitarian aid & support, people in need, president Trump's golf habit, hurricane way or path alert, tweets regarding Puerto Rico governor, power outage, relief efforts by volunteers, Red Cross, relief donation and fund request	A-Mexico, A-Maria
9-28-2017	Tweets related to Mexico earthquake, climate change, search & rescue missions and recovery, victims and survivors of the disaster, humanitarian aid & support, relief efforts by volunteers, Red Cross, relief donation and fund request, people affected	A-Mexico, A-Maria
10-4-2017	Reference to Puerto Rico, Caribbean, tweets related to hurricane Maria and Irma, president Trump, FEMA, flood water damage, disaster response, power outage, president Trump golf habit, people affected, relief supplies, Puerto Rico governer	A-Mexico, A-Maria, A-Irma

From results in table 4.3, we can try to predict hurricane timeline and see if compares to the actual timeline. Figure 4.2 shows predicted timeline of events. The actual timeline of events is shown in figure 4.1. All the time and dates are in UTC (Coordinated Universal Time).

I have used data starting from August 18 till September 28, 2017 for the predicted timeline. From the comparison of figure 4.1 and 4.2 we can see that predicted timeline does not exactly match with the actual hurricane timeline, although by comparing the most catastrophic days of each hurricane in figures 4.3 through 4.5 with 4.2 we can say that it covers most of the days that were significant during given hurricane. Also, note that the Mexico earthquake was not predicted correctly. This is the result of the data collection process for the Mexico earthquake, which was started the day after.

This method relies completely on topic distribution derived for the whole corpus and

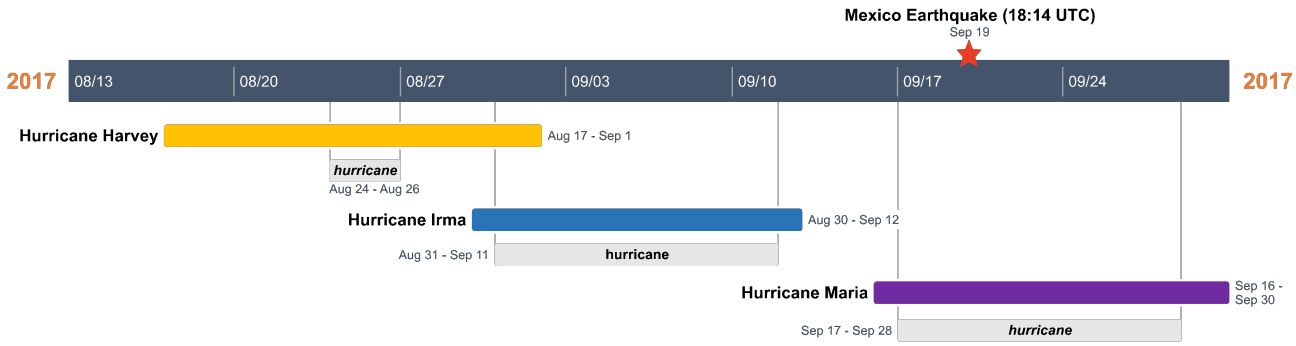

Figure 4.1: *Actual timeline of hurricane Harvey, Irma and Maria as well as Mexico Earthquake, UTC Date&Time*

gives a list of topics that can do the best job at describing a day. The corpus is a mix of four or more natural disaster events. Therefore, one topic can describe more than one events or relate to more than one events. For instance, Topic 73 is included in the topic distribution for August 18. Terms related to hurricane Harvey and Mexico earthquake hold high probability in this topic.

0.375"hurricaneHarvey" + 0.134*"Mexicocity" + 0.072*"heartbreaking" + 0.061*"Mexicoquake" + 0.046*"newsdesk" + 0.040*"unsurprising" + 0.018*"arrest" + 0.014*"beloved" + 0.014*"hurricanejose" + 0.011*"freedom"*

Therefore, this topic could have been selected if it relates to either of the events. My initial guess for Mexico earthquake occurrence was August 18, which is a day earlier than actual event happened. Although after cross-checking with Aug 18th tweets, I found no reference to the Mexico earthquake. This is a major drawback of this method, that a topic can be related to more than one events and can make it hard to interpret the actual event.

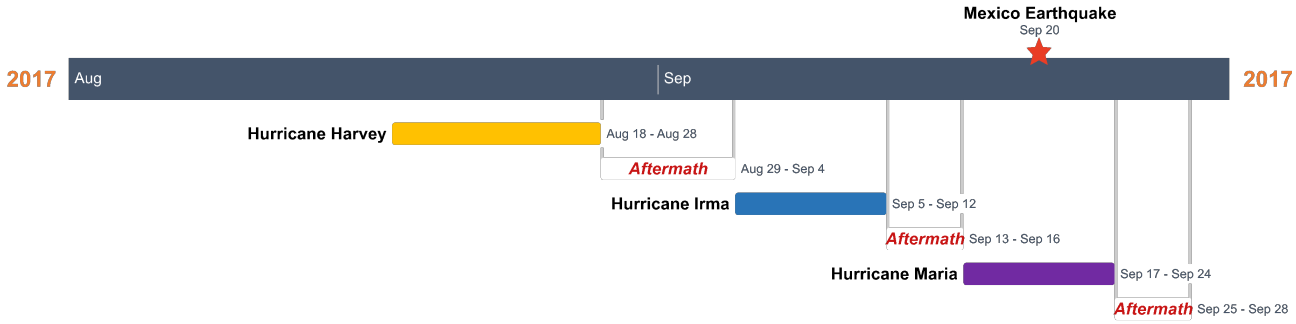

Figure 4.2: *Predicted hurricane timeline of hurricane Harvey, Irma and Maria from topic distribution per day, Data available from Aug 18 - Sep 28 2017 only, UTC Date&Time*

Another issue with this method is polysemy nature of the terms in a topic. For example, term "national" can hold multiple different meanings, such as "National Hurricane Center", "National Oceanic and Atmospheric Administration" or "national crisis" etc. Therefore, topic interpretation in this approach is ambiguous and tedious. This approach gave a very high-level view of hurricane timeline, but we do not have information such as affected areas or counties.

On the bright side, we are deriving LDA model and corpus only one time compared to the second approach where we will be deriving different LDA model and corpus for each day.

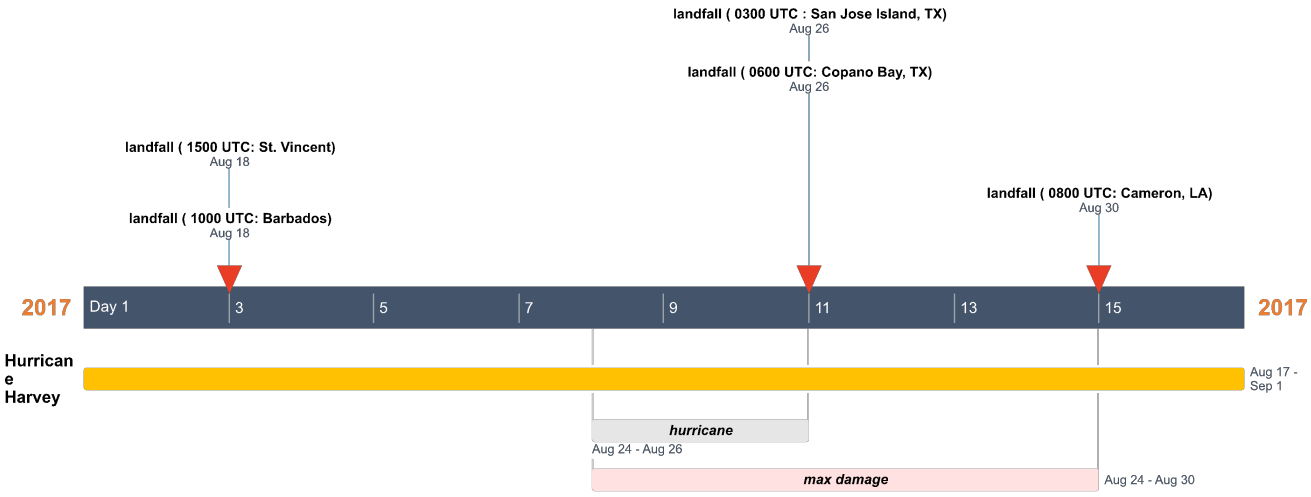

Figure 4.3: *Hurricane Harvey actual Timeline, UTC Date&Time*

The drawback, of having multiple event references to a topic may be helped by to some extent by deriving more topics from the corpus. Although, this claim can only be verified by testing the model different number of topics. Please note that I have limited myself to 80 topics due to Gensim's issue with deriving more than 80-85 topics from the corpus due to an existing bug. In my case, Gensim was listing duplicate topics for the number of topics above 80. Here is a detailed description of the problem that was faced by multiple other users. https://github.com/RaRe-Technologies/gensim/issues/217

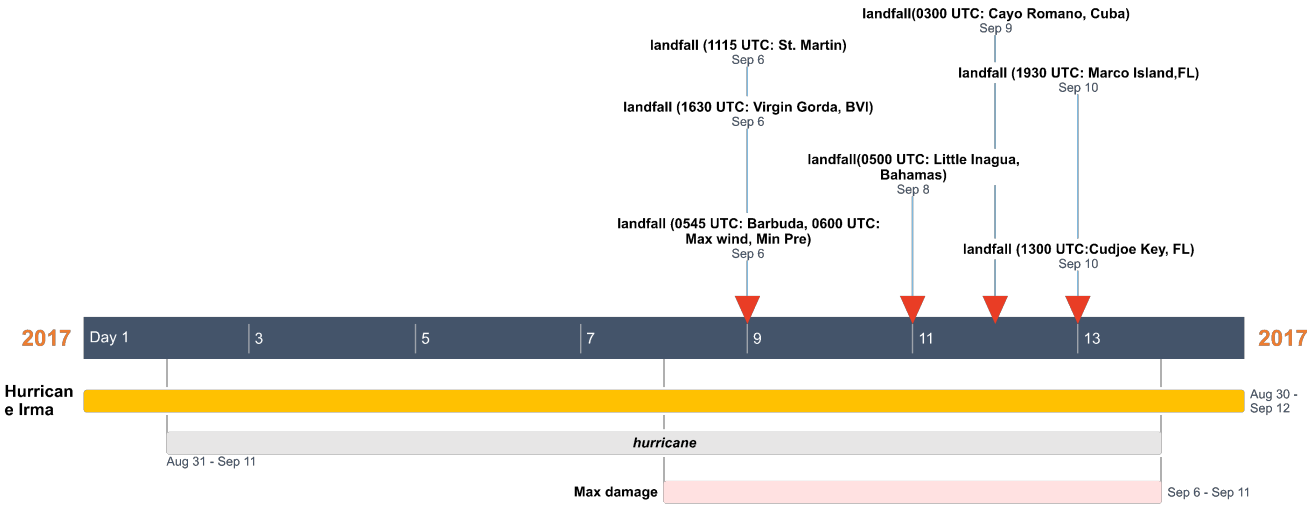

Figure 4.4: *Hurricane Irma actual Timeline, UTC Date&Time*

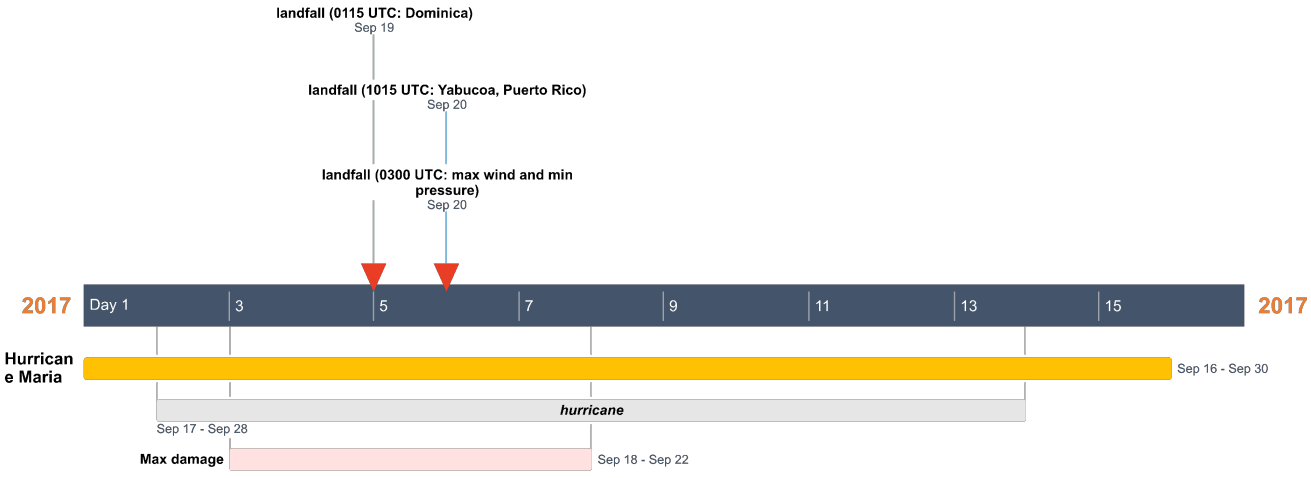

Figure 4.5: *Hurricane Maria actual Timeline, UTC Date&Time*

4.1.3 Analysis by Hour

I have done a similar analysis by changing granularity to one hour. I have only done this analysis on five days with the maximum number of topics. Those days are September 10,9,8,6 and 11. Probability distribution for all hours for these five days can be found under this link https://github.com/virashree/Topic-Modeling/blob/master/topics_by_hour_all.csv. Using this method, I want to see if I notice any major events on a particular day and predict the hurricane timeline any better.

Hour	***Popular topics***
00	Florida keys, Virgin Island , weather , emotional tweets
01	Florida keys, weather, emotional tweets
02	Florida keys, weather, emotional tweets
03	tropical storm warning, Florida keys, weather, emotional tweets
04 - 11	tropical storm warning, Florida keys, weather, emotional tweets
13	Landfall , tropical storm warning, Florida keys, weather, emotional tweets
14	Landfall , tropical storm warning, Florida keys, weather, emotional tweets
15-18	tropical storm warning, Florida keys, weather
19- 20	Landfall, tropical storm warning, Florida keys, weather, emotional tweets
21- 23	tropical storm warning, Florida keys, weather

Table 4.4: *Unique topics by hour for September 10, 2017. All data and time in UTC*

Tables 4.7,4.4,4.6, 4.8, 4.5 shows unique topics that were discussed at a particular hour on a day. Note that these tables only show the difference in topics discussed between the previous hour and the current hour. Meaning, there can be other topics mentioned during a particular hour that can be common with the previous hour. But for hourly analysis, change in topic is what is important. This is because we are interested in noticing unique events or

Hour	***Popular topics***
00 - 19	Tropical storm Irma, Florida, Puerto Rico, Dominica, Miami , Florida keys
20	Relief and recovery efforts, Donation and red cross
21	Flood warnings

Table 4.5: *Unique topics by hour for September 11, 2017. All data and time in UTC*

Hour	***Popular topics***
00 - 02	tropical storm Irma, Florida , Cuba, weather, emotional tweets
03	landfall, tropical storm Irma, Florida, weather, Cuba, emotional tweets
06- 21	tropical storm Irma, Florida , Cuba, weather, emotional tweets
21- 23	tropical storm Irma, Florida keys, weather, emotional tweets

Table 4.6: *Unique topics by hour for September 9, 2017. All data and time are in UTC*

change in the flow of what people are talking to note any significant event that may have occurred on a given day.

From the table above I have tried to speculate what events might have taken place on a given day and hour. This information can be found in Table 4.9.

Hour	***Popular topics***
00	tropical storm Irma, Florida, Harvey, weather
01	state official warning , tropical storm Irma, Florida, Harvey, weather , emotional tweets
02	tropical storm Irma, Florida, Harvey, weather , emotional tweets
03 - 04	Talk about Harveys relief and rescue and some political news, hurricane Irma , weather alerts and updates
05 - 10	hurricane Irma, hurricane Harvey and relief tweets, weather , topics from previous hour and hurricane Irma, some earthquake news in Mexico
11	tropical storm Irma, Harvey, Florida, weather news , Harvey
12- 23	tropical storm Irma, Harvey, Florida, weather news , emotional tweets

Table 4.7: *Unique topics by hour for September 8, 2017. All data and time in UTC*

Hour	***Popular topics***
00 - 01	hurricane Irma, tropical storm , Puerto Rico , witness accounts or weather hurricane news videos, Florida , weather, hurricane category , wind speed news etc. , emotional tweets
02	above topics and earthquake
03	topic same as 00
04	mention of earthquake and Mexico, climate change, all topics above , probably tweets regarding comparison of storm in past year to current year
05	all topics in 4 , mention of national (hurricane) center
06	all topics in 05 and Caribbean island , hurricane eye , radar ..leaning more towards hurricane weather related tweets. Highest number of topics that day
07	Still talking about topics in 06
08 - 11	topics similar to 07 and Florida. Florida is mentioned in the rest of the topics that day
12	Landfall, and similar topics as previous hour
13 - 24	Mention of National (hurricane) center again at 13 UTC , tweets regarding weather , Florida are popular for the rest of the hours
06- 15	Landfall mention
18, 20 -21	Landfall mention

Table 4.8: *Unique topics by hour for September 6, 2017. All data and time in UTC*

Table 4.9: Topic interpretation by Hour

Date & Time	**Topic Interpretation**
September 6	Possibly hurricane Irma has affected Puerto Rico, Mexico or some of the other Caribbean Islands. Florida has become more popular for the later hours, which might mean that hurricane Irma is slowly moving towards Florida. There are not any common topics among hours this day, which shows that this day has proved to be very catastrophic for more than one places
September 6, 1200, UTC	landfall caused by hurricane Irma possibly somewhere in the Caribbean Islands
September 6, 0600-1500, 1800, 2000-2100 UTC	landfall mentions during these hours shows that landfall has occurred in more than one places possibly in the Caribbean Islands
September 8	The pattern of topics still highly related to hurricane Harvey relief efforts rather than tropical storm warnings related to hurricane Irma might mean that hurricane Irma has not yet caused any significant damage especially on USA mainland, although mention of Florida might mean that hurricane will be heading towards Florida in the near future
September 8, 0500 - 1000 UTC	Possible earthquake news near Mexico at around 0500 UTC
September 9	Topics mentioning Cuba at the beginning of the day and later related the Florida Keys possibly mean that Hurricane Irma is moving towards the Florida Keys and away from Cuba
September 9, 0300 UTC	landfall possibly somewhere near Cuba

Continued on next page

Table 4.9 – continued from previous page

Date & Time	Topic Interpretation
September 10, 0000 UTC	From the focus of topics shifting from the Virgin Islands to the Florida Keys might mean that Hurricane Irma has already passed over the Virgin Islands and caused some significant damage
September 10, 0000 - 0200 UTC	From the combination of sentimental tweets and weather-related topics pattern for the given hours it is possible that Hurricane Irma heading towards the Florida Keys
September 10, 0300 - 2300 UTC	From the combination of sentimental tweets, tropical storm warnings and landfall and other damage related topics for all the given hours it is possible that Hurricane Irma has affected on some parts of Florida Keys and warnings has been issued for other parts of Florida Keys
September 10, 1300, UTC	Hurricane Irma possibly have made a landfall somewhere in the Florida Keys
September 10, 1900, UTC	Hurricane Irma possibly have made a landfall somewhere in the Florida Keys

Continued on next page

Table 4.9 – continued from previous page

Date & Time	Topic Interpretation
September 11	Sentimental topics are absent from the topic distribution for this day and topics related to relief, recovery, flood are introduced. There are multiple place mentions such as Florida, Puerto Rico, Dominica, Miami, Florida Keys. Although, there still tropical storm warnings related topics during most of the day. This pattern shows that hurricane Irma has decreased quite a bit in intensity, although has not completely left the land. The places mentioned are the placed that were affected during the last 2-3 days, except Miami. This might mean that hurricane Irma is still bringing flood or rain in parts of Florida and neighboring states. The topics discussed this day are very common in nature, which shows that there weren't events as catastrophic as landfall occurred this day. Although, topics related to flood and tropical storm means that there were a lot of tweets regarding either showing damage caused by flood water or warning regarding flood

Table 4.10 lists actual events that occurred during these five days. From the comparison, we can tell that we have predicted the nature of events quite accurately. The landfalls and earthquake event match well with the time and date. The hourly analysis has predicted the track of hurricane Irma more accurately than analysis by day. According to the prediction, September 6 was an event-full day. As a matter of fact, NHC issued 16 public advisories that day[4]. As suggested in the predicted hurricane path, hurricane Irma did pass over the Caribbean Islands, Cuba, the Florida Keys around September 6, September 9 and September 10 respectively. Figure 4.8 shows a comparison between the predicted and actual events with date, time and place specific.

Hour	***Popular topics***
September 6, 0600 UTC	maximum wind and minimum pressure noted for hurricane Irma
September 6, 0545 UTC	landfall on Barbuda
September 6, 1115 UTC	landfall on St. Martin
September 6, 1630 UTC	landfall on Virgin Gorda, British Virgin Islands
September 8, 0449 UTC	2017 Chiapas earthquake, Mexico
September 9, 300 UTC	Hurricane Irmas fifth landfall near Cayo Romano, Cuba
September 9, 2100 UTC	Public advisory issued regarding Irma moving away from Cuba and approaching Florida keys *"...EYE OF IRMA BEGINNING TO MOVE SLOWLY AWAY FROM THE COAST OF CUBA WHILE WEATHER IS DETERIORATING IN SOUTH FLORIDA... ...MAJOR HURRICANE FORCE WINDS EXPECTED OVER THE FLORIDA KEYS AT DAYBREAK"* [14]
September 10, 0300 UTC	Public advisory at this time and date *"...IRMA TAKING ITS TIME MOVING AWAY FROM CUBA... ...LIFE-THREATENING STORM SURGE EXPECTED IN THE FLORIDA KEYS AND THE WEST COAST OF FLORIDA..."* [14]
September 10, 1300 UTC	landfall on Cudjoe Key, Florida
September 10, 1930 UTC	landfall near Marco Island, Florida

Table 4.10: *Unique topics hour for September 11, 2017. All data and time in UTC*

Actual Events	Predicted Events	Comparison
September 6, 0545 UTC: landfall on Barbuda	September 6, 0600 -1500 UTC: landfall mentions during these hours shows that landfall has occurred in more than one places possibly in Caribbean Islands	✅ Most probably referring to same event
September 6, 1115 UTC: landfall on St. Martin	September 6, 1200 UTC: landfall caused by hurricane Irma possibly somewhere in the Caribbean Islands	✅ Mostly probably correct match
September 6, 1630 UTC: landfall on Virgin Gorda, British Virgin Islands	September 6, 1800, 2000-2100 UTC: landfall mentions during these hours shows that landfall has occurred in more than one places possibly in Caribbean Islands	✅ Most probably referring to same event
September 8, 0449 UTC: 2017 Chiapas earthquake, Mexico	September 8, 0500 - 1000 UTC: Possible earthquake news near Mexico at around 0500 UTC	✅ Compares Correctly!
September 9, 0300 UTC: Hurricane Irma's fifth landfall near Cayo Romano, Cuba	September 9, 0300 UTC: landfall possibly somewhere near Cuba	✅ Compares Correctly!
September 10, 1300 UTC: landfall on Cudjoe Key, Florida	September 10, 1300 UTC: Hurricane Irma possibly have made a landfall somewhere in Florida Keys	✅ Compares Correctly!
September 10, 1930 UTC: landfall near Marco Island, Florida	September 10, 1900 UTC: Hurricane Irma possibly have made a landfall somewhere in Florida Keys	✅ Compares Correctly!

Figure 4.6: *Comparison between actual and predicted hourly events*

4.1.4 Day as corpus

Python's library pyLDAvis provides creates interactive charts for the LDA model. It is a great way to visualize the results of a topic model. I have created these charts for all 43 days considering a day as a corpus and a tweet is treated as a document. It is obvious that topics derived for each day, in this case, will be more specific and accurate compared to approach 1. Although in this approach we will be creating a separate LDA model for each day, in our case 43 days that is 43 LDA models! We can say that approach 1 is giving an overall picture of topics discussed for each day, while approach 2 is looking more closely at each day's events.

With the help of pyLDAvis graphs, one can observe topics and how prevalent they were on a given day, what were the terms with highest term frequency on a given day. These graphs make it easy to interpret the topics. Discussing all the results here is out of the scope of this report. Figure 4.7 shows an example of pyLDAvis graph.

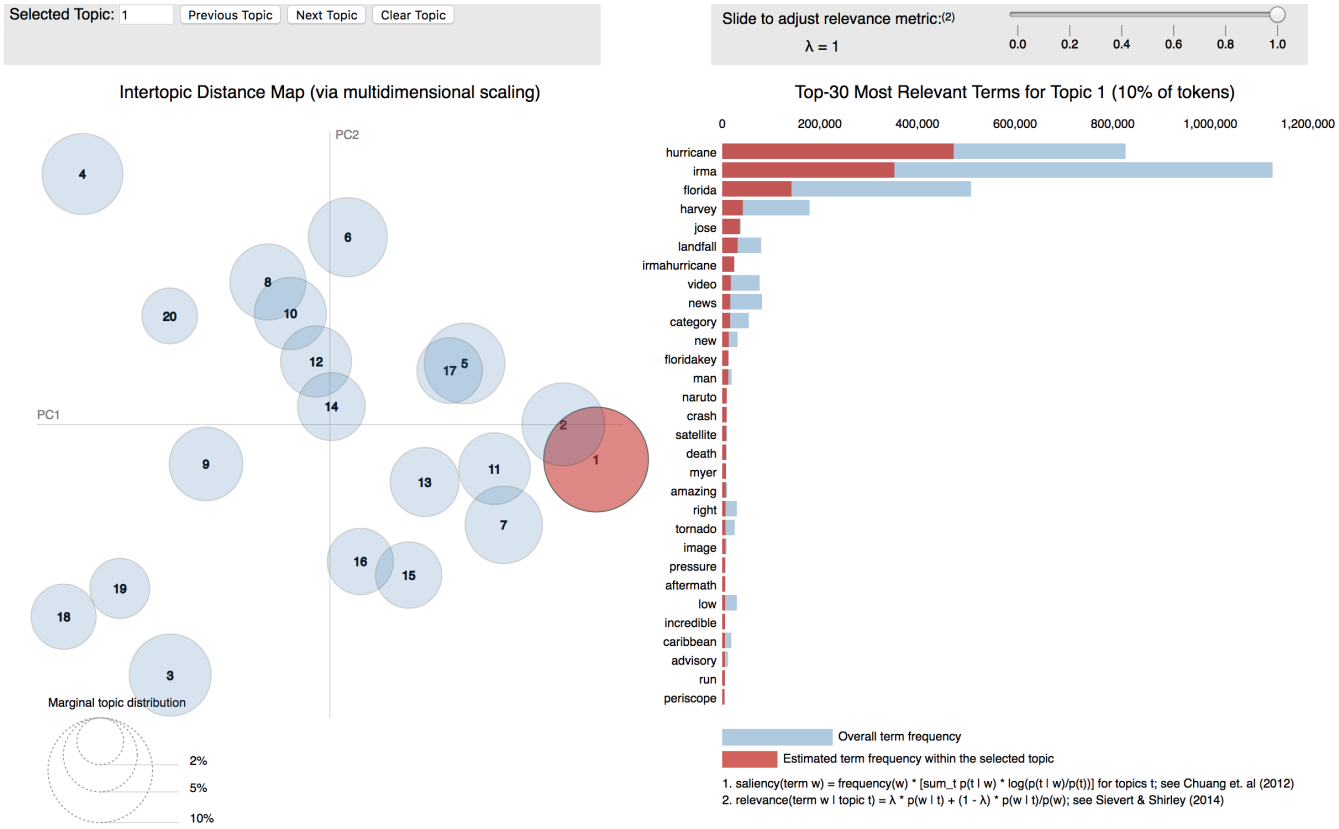

Figure 4.7: *LDAvis graph for September 10, 2017*

The interpretation for all the topics is provided under table 4.11. I have also listed places mentions that came up each day in table 4.12, and some other news that got captured are listed in table 4.13.

Table 4.11: Topic interpretation by Day, a day as corpus

Date	Topic Interpretation
8-18-2017	hurricane Harvey heading towards Barbados
8-19-2017	Harvey will impact Central America, Mexico, Caribbean next week, it has made a landfall in Barbados yesterday
	Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
8-20-2017	Tropical storm Harvey remnants related advisories. Showers and thunderstorm warnings for possibly. Very interesting mentions of terms "hunters", "plane" and "flight", which refers to NOAA's Hurricane Hunters air-crafts that are used during a hurricane to collect data regarding pressure, wind speed, direction, humidity to for meteorologists to provide forecasts regarding hurricanes. So it seems that for now, the hurricane has calmed down although they are closely watching it
8-21-2017	Same topics are the previous day only more affirmative of development of storm Harvey and now referring to Texas. So it seems that there are high chances Harvey will move towards the Gulf of Mexico and Texas
8-22-2017	There are a lot of predictions supporting that Harvey is regenerating and will affect Texas, there are also predictions related to the strength of the storm. It seems that there are warnings related to tropical storm or hurricane watch
8-23-2017	Similar topics as the previous day. Tropical depression related advisories have been issued for Texas and Louisiana
8-24-2017	It seems that Harvey has converted into a hurricane and has started affecting southern Texas. There were mandatory evacuations for some counties in Texas, There seems to be news surrounding wall and president Trump
8-25-2017	Similar topics as the previous day. Corpus Christi and Galveston seem to have been affected a lot by the storm

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
8-26-2017	It seems there was landfall this day, maybe the first landfall of hurricane Harvey in the USA and storm has moved to Houston, Texas. There are several other news related immigration, apparently, someone named Joe Arpaio. There is still talk about landfall and devastation mostly related to cities and counties in Texas
8-27-2017	It seems that storm has weakened a bit because there are tweets related to Red Cross, relief efforts, the budget is coming into the picture
8-28-2017	Similar topics as before and now more and more regarding rescue and relief. There is also more talks about climate change
8-29-2017	Similar topics as the previous day. Note that there were some topics related to agriculture. Some talks about First Lady Melania and president trump's visit
8-30-2017	similar topics as before
8-31-2017	similar topics as before, there is news related to some chemical explosion
9-1-2017	similar topics as before, lot more surrounding relief and rescue and political news
9-2-2017	similar topics as before and Some tweets regarding missing person Elijah Griffin. There seems to be Eid celebration
9-3-2017	similar topics as before and it seems that this day is a National day of Prayer. Some tweets regarding missing person Elijah Griffin
9-4-2017	similar topics as before. Some focus on DACA and political news

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
9-5-2017	Hurricane Irma is on the news now and with that storm, Jose is also being talked about. There are mentions of the Caribbean Islands. It does not seem to be in its initial stage but rather has already tuned in to a hurricane of some Category and has started the devastation. From the repeated mentions of Cuba, Bahama, Haiti, Puerto Rico, Dominica it seems that storm is around these Caribbean Islands at the moment and some mentions of Florida and Miami gives a signal that it storm is predicted to affect those areas as well. There is a mixture of topics related to Harvey relief efforts and Irma's forecast. It also seems that Irma has registered an earthquake on a seismometer
9-6-2017	Apart from hurricane Irma there are hurricane Katia and Jose are being discussed as well. It seems that there has been a lot of devastation in Barbuda, St. Martin, Barbados, Puerto Rico, Dominica, Cuba. There are lot of terms related to devastation used, ex. "devastating", "damage", "destruction", "catastrophic", "habitable". There also seems to be a landfall news mostly probably around St. Martin. Florida has been mentioned in multiple topics, although the devastation news is more related to the Caribbean Islands, which gives a clue that the storm has not reached Florida yet but is predicted to affect it. This day has a lot of mixture of topics that includes a lot of political news, relief news related to Hurricane Harvey, news related to hurricane damage, landfall, and forecast

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
9-7-2017	There are more tweets focused on hurricane track and path than hurricane damage. Florida has been mentioned in many topics. The focus towards Caribbean Islands has decreased compared to the previous day. So it seems that the hurricane has passed over them and heading towards Florida. Tweets related to hurricane Harvey's relief can be noticed on this day's topics. Apart from hurricane Irma, hurricane Jose and hurricane Katia are also most discussed topics this day. There is some mention of the earthquake. Climate change is also one of the most discussed topics on this day.
9-8-2017	Hurricane Irma has made some devastation in Cuba. There is a possible landfall as well most probably in Cuba. Apart from this, there is a lot of talk about an earthquake in Mexico, California wildfires, Hurricane Jose and hurricane Harvey's recovery efforts. Some mentions of Florida Keys.
9-9-2017	Lot of talk about Florida Keys island on this day, although tweets are still more related to hurricane's path and track. So it implies that Irma will be most definitely heading towards the Florida Keys Islands. There are time references to Saturday and Sunday, which might mean that the hurricane will pass over these areas during Saturday or Sunday. There are tweets regarding mandatory evacuation, Mexico earthquake, and hurricane Harvey relief efforts.

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
9-10-2017	Florida Keys related topics are much more apparent on this day. There is news related to landfall. There is a high possibility these landfalls are somewhere in the Florida Keys. It seems that on this day Hurricane Irma passed over Florida keys. There are mentions of places such as Georgia, Atlanta that are north of Florida Keys. This may imply that storm is heading towards these places slowly
9-11-2017	The focus from Florida keys has shifted more towards Georgia, Atlanta, Tampa, Orlando, Jacksonville. It seems that hurricane Irma is moving to North to these places
9-12-2017	There are topics related to hurricane relief now. It seems that hurricane Irma has calmed down for the most part
9-13-2017	Similar topics are last day
9-14-2017	Similar topics are last day
9-15-2017	Similar topics are last day
9-16-2017	Similar topics are last day, Some mention of Hurricane Maria
9-17-2017	Tweets related to hurricane Maria's forecast. From the mentions of Caribbean, it seems that this hurricane may impact Carribean Islands as well
9-18-2017	Still a lot of talk related to hurricane Irma and Harvey and there are some topics related to hurricane Maria but not a whole lot. There is still reference to Caribbean islands, this might mean that hurricane Maria is probably around Caribbean Islands

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
9-19-2017	There is a shift towards tweets related to Puerto Rico, Hurricane Maria. terms related to hurricane damage and intensity. This strongly suggests that hurricane Maria has started affecting Puerto Rico as well as neighboring Islands such as Bahama, Dominica, Virgin Islands, Barbuda. Apart from this there are also references to Mexico Earthquake on this day suggesting another earthquake might have taken place in Mexico
9-20-2017	Lot of topics in reference to Mexico Earthquake, hurricane Maria and Puerto Rico. From the terms such as "rubble", "magnitude", "devastating", "death", it seems that earthquake has brought a lot of damage. There are some signs of landfall due to hurricane Maria and power outage in Puerto Rico on this day. Also, tweets related to hurricane Maria eye-wall and strength showing that it has proved to be devastating for Puerto Rico so far. Hurricane Irma and Harvey's relief efforts are still being discussed quite a lot.
9-21-2017	There are tweets related to the devastation caused by the Mexico earthquake and Hurricane Maria. There are topics related to electricity and power outage in case of hurricane Maria quite a bit
9-22-2017	similar topics as the previous day. Quite a lot talk regarding devastation and electricity outage in Puerto Rico. It seems that hurricane has moved away from Puerto Rico, although has caused a huge amount of damage
9-23-2017	Tweets related to relief efforts are introduced, it seems that hurricane Maria has moved away from land and/or decreased in intensity
9-24-2017	similar topics as previous day

Continued on next page

Table 4.11 – continued from previous page

Date	Topic Interpretation
9-25-2017	similar topics as the previous day and topics related to requesting help, and reports of flood water, diseases, and power outage.
9-26-2017	similar topics as the previous day. Request for help is a prevalent topic
9-27-2017	similar topics as previous day
9-28-2017	similar topics as the previous day and there are news related to a volcanic eruption in Popocatepetl, Mexico. There is a lot of focus on this volcanic eruption and Mexico earthquake
10-4-2017	There is a lot of topics around to hurricane Maria relief, President Trump and strangely paper towels. There are also mentions of water purification kit, drinkable water and so on, which donates the tweets related to hurricane relief efforts. Although most of these are related to hurricane Maria rather than Irma or Harvey

Table 4.12: Place Mentions by Day

Date	Place Mentions
8-18-2017	Barbados
8-19-2017	Barbados, Caribbean, Mexico, Central America, Jamaica, Yucatan
8-20-2017	Caribbean, Jamaica, Central America
8-21-2017	Campeche, Texas, Caribbean, Mexico, Gulf of Mexico
8-22-2017	Texas, Caribbean, Yucatan, Mexico, Louisiana, Campeche, Houston, Gulf of Mexico, or Gulf of Texas
8-23-2017	Texas, Louisiana, Gulf of Mexico, or Gulf of Texas, Mexico

Continued on next page

Table 4.12 – continued from previous page

Date	Place Mentions
8-24-2017	Galveston, Texas, Mexico, Gulf of Mexico or Texas, Carolina, Louisiana, Houston, Corpus Christi, Aransas, Rockport, Gulf Coast
8-25-2017	Padre Island, Texas, Austin, Texas, Corpus Christi, Galveston, Houston
8-26-2017	Corpus Christi, Texas, Aransas, Houston, Rockport, Galveston, Port Aransas, Victoria, Austin, San Antonio, Louisiana
8-27-2017	Galveston, Houston, Rockport, Campbell, Austin, Corpus Christi, Greenspoint
8-28-2017	Houston, Greenspoint, Campbell, Austin
8-29-2017	Houston, Annaville
8-30-2017	Jefferson, Port Arthur, Houston, Corpus Christi
8-31-2017	Beaumont, Houston, Port Arthur, Rockport
9-1-2017	Houston, Corpus Christi, Port Arthur, Louisiana
9-2-2017	Beaumont, Houston, Corpus Christi, Lakewood, Houston, Austin
9-3-2017	Houston
9-4-2017	Houston, Louisiana
9-5-2017	Florida, Cuba, Bahama, Haiti, Caribbean, Puerto Rico, Dominica, Texas, Barbuda, Barbados, Virgin Island
9-6-2017	Barbados, Barbuda, St. Martin, Guam, Puerto Rico, Dominica, Haiti, Cuba, Virgin Island
9-7-2017	Georgia, Florida, Miami, Barbuda, Virgin Island, Puerto Rico, Bahama, Haiti, Dominica, St. Martin, Barbuda, Barbados
9-8-2017	Cuba, Florida, Barbuda, Haiti, Houston, Georgia, Miami, Florida Keys, Mexico

Continued on next page

Table 4.12 – continued from previous page

Date	Place Mentions
9-9-2017	Florida Keys, Beaumont, Miami, Mexico, Houston, Cuba, Orlando, Florida, Georgia, Barbuda, Bahama, Caribbean, Haiti, Tampa, Fort Lauderdale, Sabana-Camagey Archipelago
9-10-2017	Bahama, Florida Keys, Miami, Houston, Tampa Bay, Caribbean, Marco Island, Cudjoe Key, Brickell, Virgin Island, Fort Lauderdale, Sarasota, Naples, Atlanta
9-11-2017	Jacksonville, Georgia, Atlanta, Florida Keys, Miami, Brickell, Charleston, Keywest, Tampa, Puerto Rico, Bahama, St. Martin, Orlando, Beaumont, Houston
9-12-2017	Orlando, Houston, Virgin Islands, Carolina, Tampa, Louisiana
9-13-2017	Caribbean, Florida, Houston
9-14-2017	Florida, Texas, Charleston
9-15-2017	Charleston, Florida, Texas
9-16-2017	Florida, Texas
9-17-2017	Cuba, Florida, Caribbean, Texas, Barbuda
9-18-2017	Florida, Caribbean, Texas, Miami
9-19-2017	Puerto Rico, Virgin Islands, Dominica, Guadeloupe, Bahama, San Juan, Barbuda, Dominica, Mexico
9-20-2017	Puerto Rico, Yabucoa, Mexico, Florida, Houston, Texas, Virgin Islands, Dominica, Rio Grande, Manati, Ciales, Mexico
9-21-2017	San Juan, Puerto Rico, Mexico, Yabucoa, Rio Grande, Mexico City
9-22-2017	San Juan, Mexico, Puerto Rico
9-23-2017	Oaxaca, Mexico, Puerto Rico, Dominica, Virgin Island, Matias, Bahama
9-24-2017	St. Thomas, Puerto Rico, San Juan, Mexico

Continued on next page

Table 4.12 – continued from previous page

Date	Place Mentions
9-25-2017	Puerto Rico, Dominica, Virgin Island, Mexico
9-26-2017	San Juan, Mexico, Puerto Rico
9-27-2017	San Juan, Mexico, Puerto Rico
9-28-2017	Oaxaca
10-4-2017	Puerto Rico, San Juan

Table 4.13: Special Mentions by Day

Date	Special Mentions
8-24-2017	Kevin Doremus, NOAA, FEMA, President Trump
8-25-2017	vest Selena statue, NOAA, President Trump, Kevin Doremus
8-26-2017	Arpaio, Casey Stegall, reporter, Gorka, Gorka resign, chose Jimmy Kimmel, Joe Arpaio, North Korea ICBMS
8-27-2017	Casey Stegall, FEMA, saratoga Cimarron
8-28-2017	Cajun navy, Dallas Cowboys, blacklivesmatter, Dickinson, Roberson, USCG, National Guard, FEMA
8-29-2017	Melania, Joel Osteen, Anheuser-Busch, alligator, Cajun navy, Joel Osteen
8-30-2017	Sandra Bullock, kyle wood ct, sigma phi
8-31-2017	coach Doug Brocail, Sandra Bullock, sigma phi, Arkema, gas shortage, a chemical explosion
9-1-2017	Obama, Trump, National Guard, Redneck, delta, sigma phi, labors day weekend, gas shortage, USCG, MAGA, Jerry Jones, Joel Osteen

Continued on next page

Table 4.13 – continued from previous page

Date	Special Mentions
9-2-2017	Clinton, Joel Osteen, national day of prayer, labor day weekend, climate change, Eid, Elijah Griffin # years old, mosque, Iran, alligator, Scott Olson, NRG, ExxonMobil, a hazardous pollutant
9-3-2017	National day of prayer, MAGA, President Trump, Obama, Clinton, Tyler Perry, Joel Osteen, Elijah Griffin # years old, Scott Olson, DACA
9-4-2017	infectious disease, chemical hazard, krogercare, Ariana Grande, dreamer, DACA, Elijah Griffin # years old, Obama, Mccarthy, salvation army, Jedi wisdom professional, Vancouver, pizza hut, Kayak
9-5-2017	Alonso Guill, actress Alyssa Milano blaspheme, Ariana grande, North Korea, DACA, Dreamer, California fire, stage colon cancer, Canada
9-6-2017	Syria, Randy Muller, HHS, India, Nepal, Nigeria, Pakistan, jay cutler, Hurricane Jose, Hurricane Katia, Jet Blue
9-7-2017	Ivanka, Hurricane Jose, Hurricane Katia, Climate change, India, Nepal, Nigeria, Pakistan, Jet Blue, California, Rick Scott
9-8-2017	Rick Scott, Hurricane Jose, Hurricane Katia, Beyonce, Rush Limbaugh, Clinton, Mexico Earthquake, California wildfire, Janet Jackson, North Korea, Telethon, Susan Dell, Wasiu
9-9-2017	Rush Limbaugh, Capt Phil Blanchard, Beygood
9-10-2017	Richard Branson, Tornadoes, Hurricane Jose, Rush Limbaugh, Busch
9-11-2017	Hurricane Jose, Tornadoes, Jordan, Rush Limbaugh, Busch, Yemen
9-12-2017	HandinHand, Telethon, France, Greenpeace, NASA, climate change, expired Patanjali good, Steve Harvey, Hurricane Jose, Hillary, FEMA
9-13-2017	FEMA, HandinHand, Justin Bieber

Continued on next page

Table 4.13 – continued from previous page

Date	Special Mentions
9-14-2017	Beyonce, Rick Scott
9-16-2017	pantyfa BLM
9-19-2018	Roosevelt Skerrit, Mexico Earthquake
9-20-2017	Sergio Perez, Japan, London, New Zealand, Mexico Earthquake
9-21-2017	Canada, Salma Hayek
9-22-2017	Salma Hayek, Ben Carson
9-23-2017	Kylie Jenner, Richard Romero
9-24-2017	Kylie Jenner
9-25-2017	Despacito singer
9-26-2017	Carmen Yulin
9-28-2017	Popocatepetl volcano, tribute Santiago Flores
10-4-2017	Geraldo, Trump, Ivanka

From the table 4.13, notice that we were able to capture some other disasters that we did not collect data for. For example hurricane Jose, hurricane Katia, California wildfire, Popocatepetl volcano, a chemical explosion in Arkema factory after hurricane Harvey. some famous personalities names also came up in the topics for example Rush Limbaugh who is a host of a radio talk show and lives in Palm Beach, Florida. Apparently, he commented that hurricane Irma is fake and is made up by media. Because of that, we see a surge of tweets mentioning him on Twitter during hurricane Irma. Other names such as Salma Hayek, who made a big donation to Mexico earthquake victims, also president Trump, first lady Melania Trump, Ivanka Trump etc. I will leave it to the readers to explore the special mentions in more detail.

Notice that apart from the political news and other disaster news, we were also able to capture one SOS (stands for Save our Souls/Ship, a distress call) tweets regarding a missing person Elijah Griffin for three days in a raw. This indicates that, if we reduce the corpus

size even more and extract topics from smaller chunks of data we might be able to gather more of such topics. Note that the LDA model is designed to automatically learn about the underlying structure of a document. We will need a much more sophisticated system to be able to extract SOS tweets from the dataset. Unlike the first approach, we were able to get down to the county level information using this approach. I have listed all the place mentions in table 4.12. Using this information, I have pinpointed all the locations on a map to predict the path for all three hurricanes. Figures 4.8, 4.9, 4.10 and 4.11 show all three hurricanes predicted path based on the location information for each day. I have given areas that came up the most in topics higher importance. For example, in the case of hurricane Harvey, Corpus Christi, Beaumont and Houston came up the most during hurricane Harvey. The counties surrounded by Corpus Christi, and San Antonio and Austin came up comparatively less. Therefore, I made an assumption that Corpus Christi, Beaumont, and Houston were closer to the eye of the hurricane and experienced probably higher category hurricane than others. Note that, only the counties closer to the Gulf areas were noted the most in the topics, which implies that hurricane Harvey did not directly move towards Houston from Corpus Christi but it left the land of Texas shortly after making huge destruction in Corpus Christi and then again entered Texas and progressed towards Beaumont and Houston. We can do a similar analysis for hurricane Irma and Maria as well.

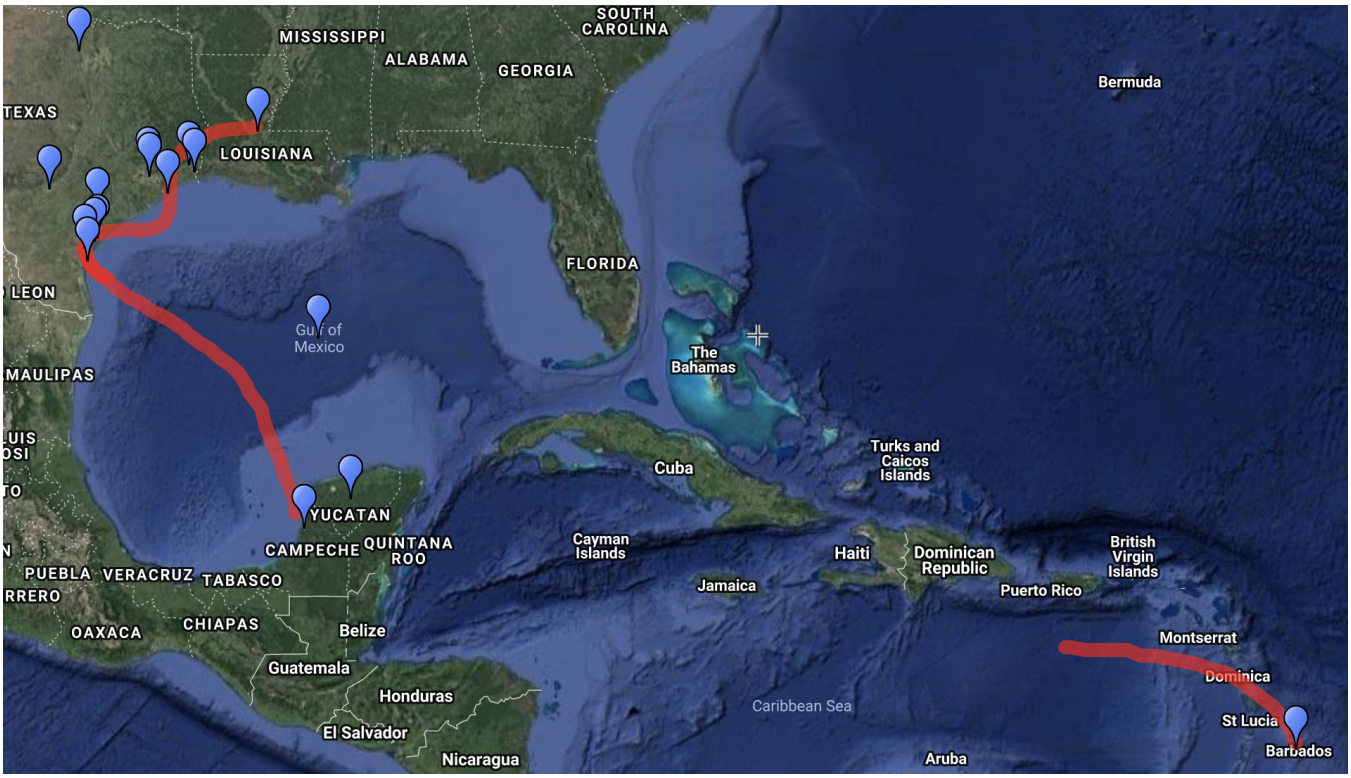

Figure 4.8: *Hurricane Harvey predicated path*

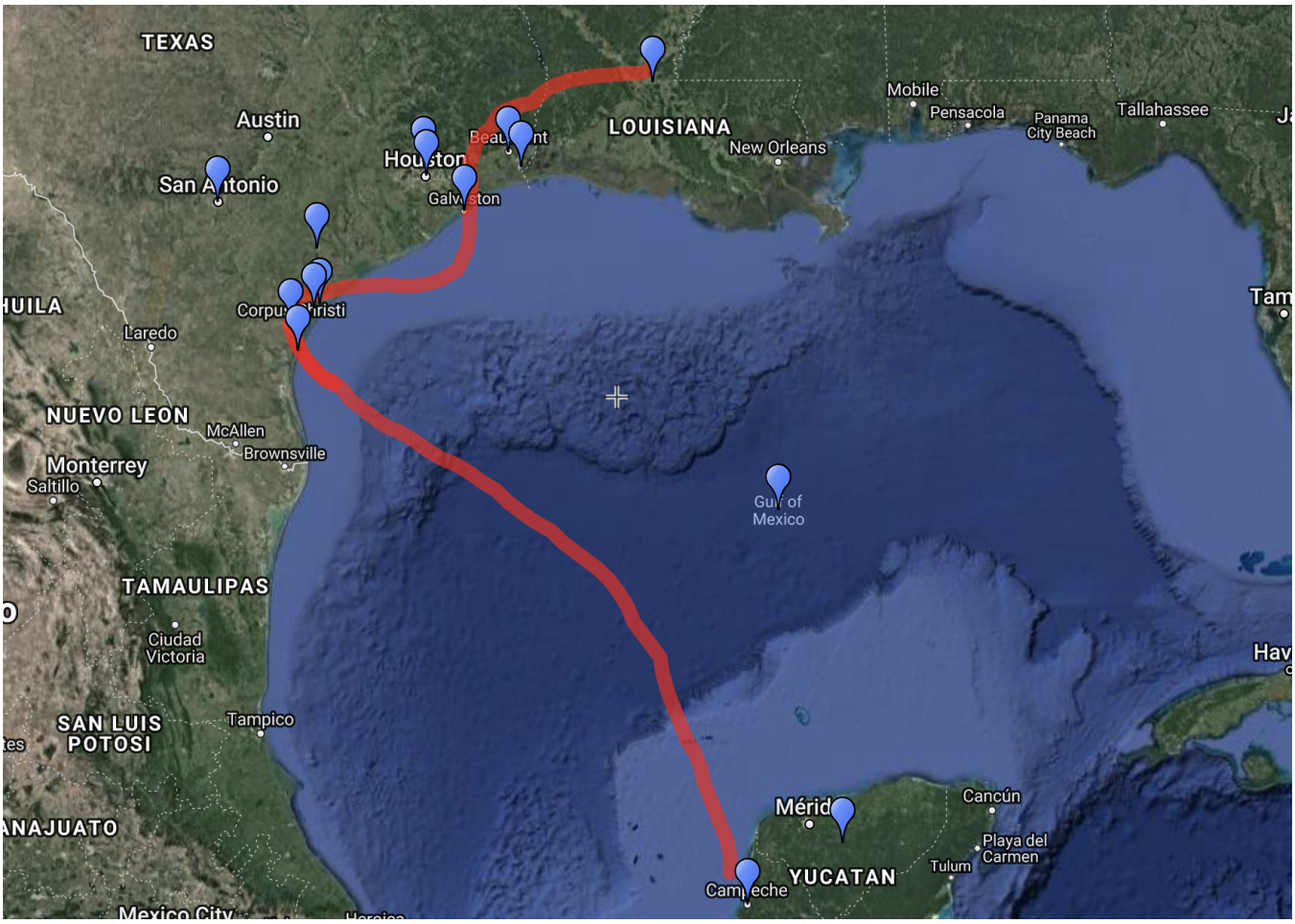

Figure 4.9: *Closer look to Hurricane Harvey's predicted path in united states*

Figure 4.10: *Hurricane Irma predicted path*

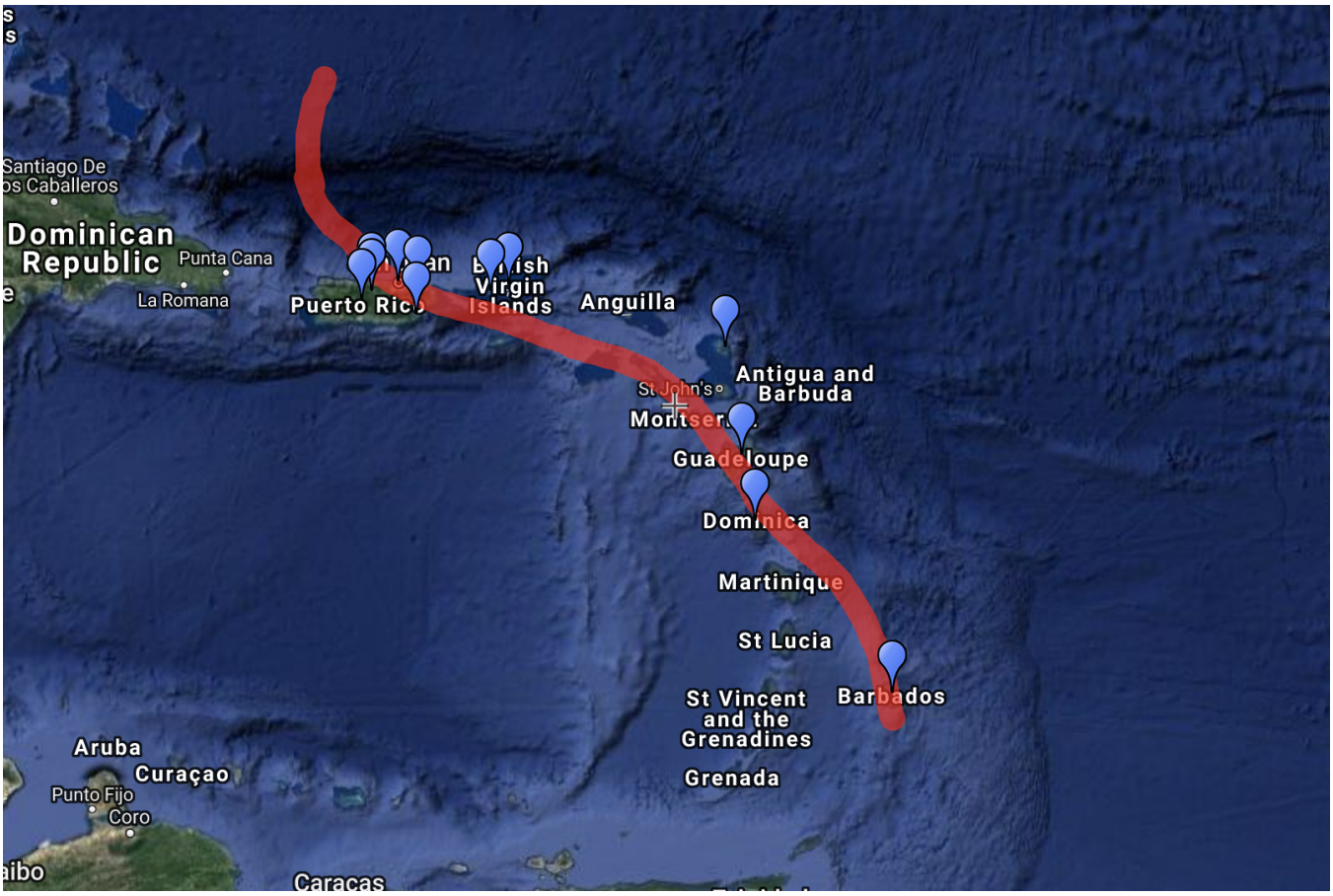

Figure 4.11: *Hurricane Maria predicted path*

The location information obtained from Twitter can be very important means for rescue and relief troops to help people during a natural disaster. This approach does give a list of places that were affected the most, although we are still long ways from getting to street level information to reach to the victim.

It will be interesting to find out how the predicted paths of hurricanes compare with the actual ones. I have gathered hurricane track data from The Weather Channel that are shown in figures 4.12 through 4.15. One can see that the predicted paths do match the actual paths of the hurricanes and the most affected areas are in fact in the direct path of the hurricane. Note that the areas that were impacted the most match with the predicted path. This supports our assumption that the more a place is mentioned in topics, the more it got damaged during a hurricane. Not only can we list the places that were damaged the most, but we can also predict what places are projected to be in the way of a hurricane. For example, in the case of hurricane Irma, Florida and the Florida Keys came up in the topics even before they were affected by the hurricane. Similarly, in case of the hurricane Harvey, Campeche appeared in the topics even before hurricane Harvey regenerated in the Gulf of Mexico. This is because tweets related to hurricane forecasts and warnings are posted prior to the arrival of a hurricane for a place.

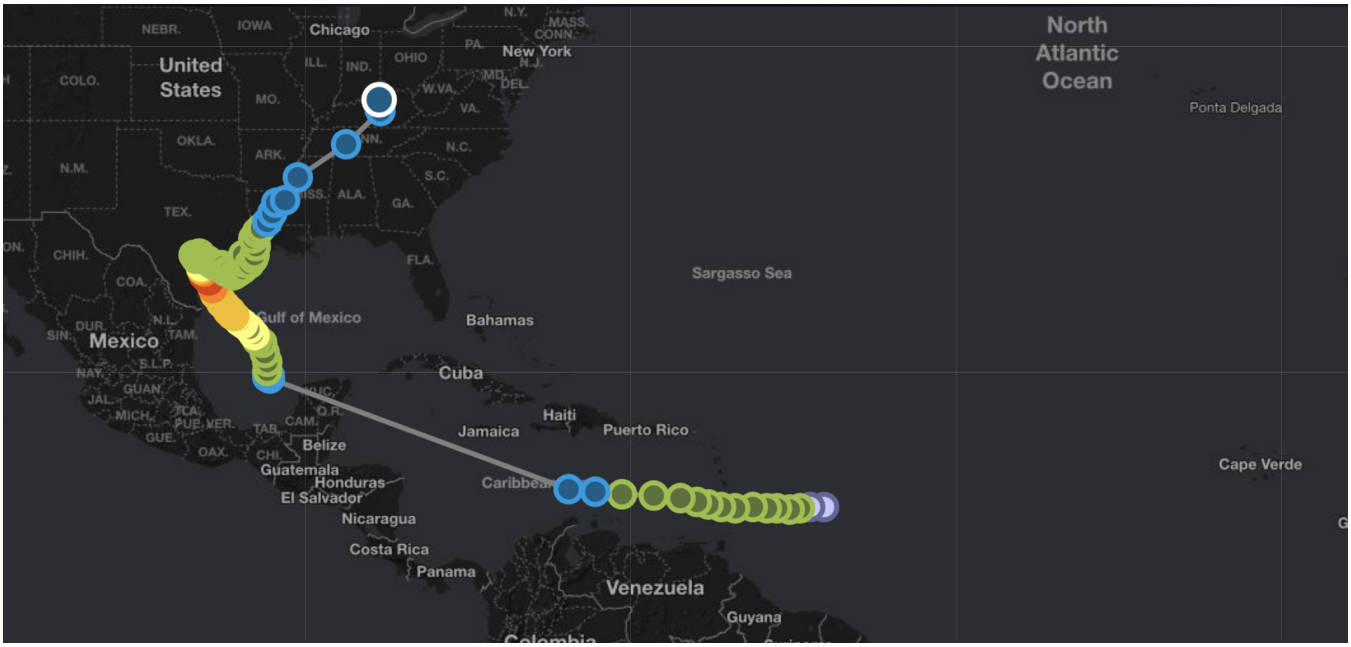

Figure 4.12: *Hurricane Harvey actual path by The Weather Channel*[4]

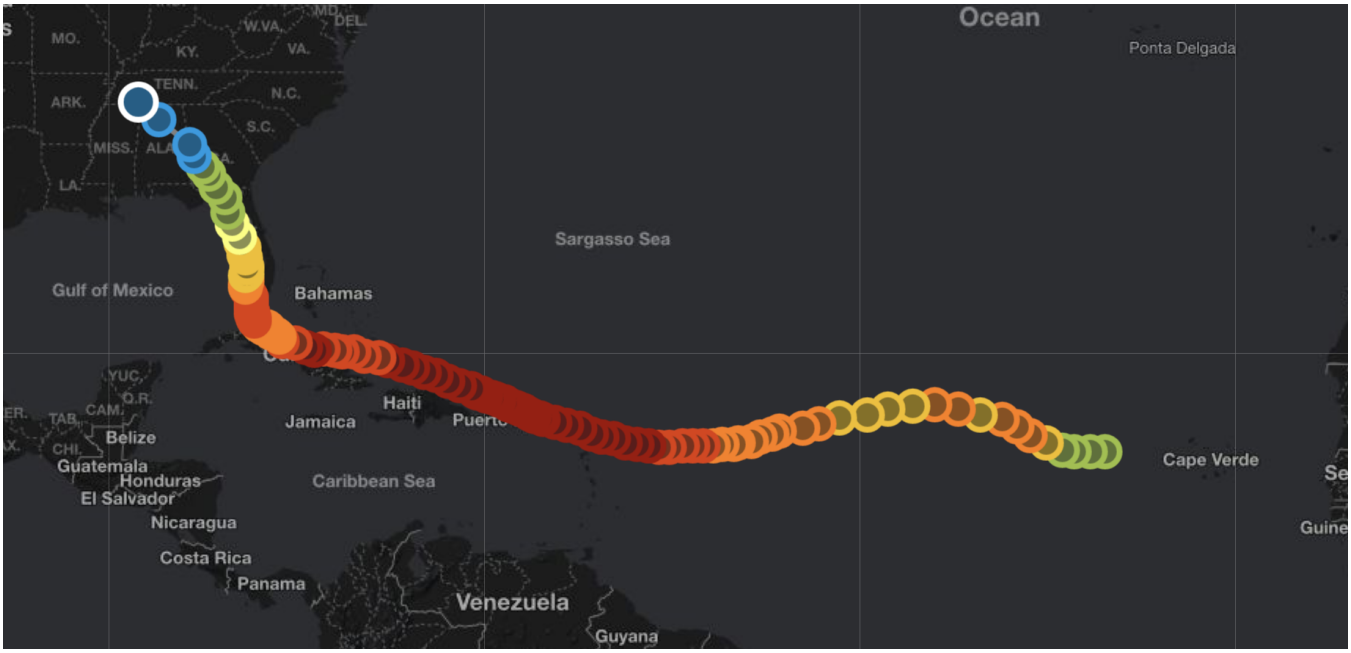

Figure 4.13: *Hurricane Irma actual path by The Weather Channel*[5]

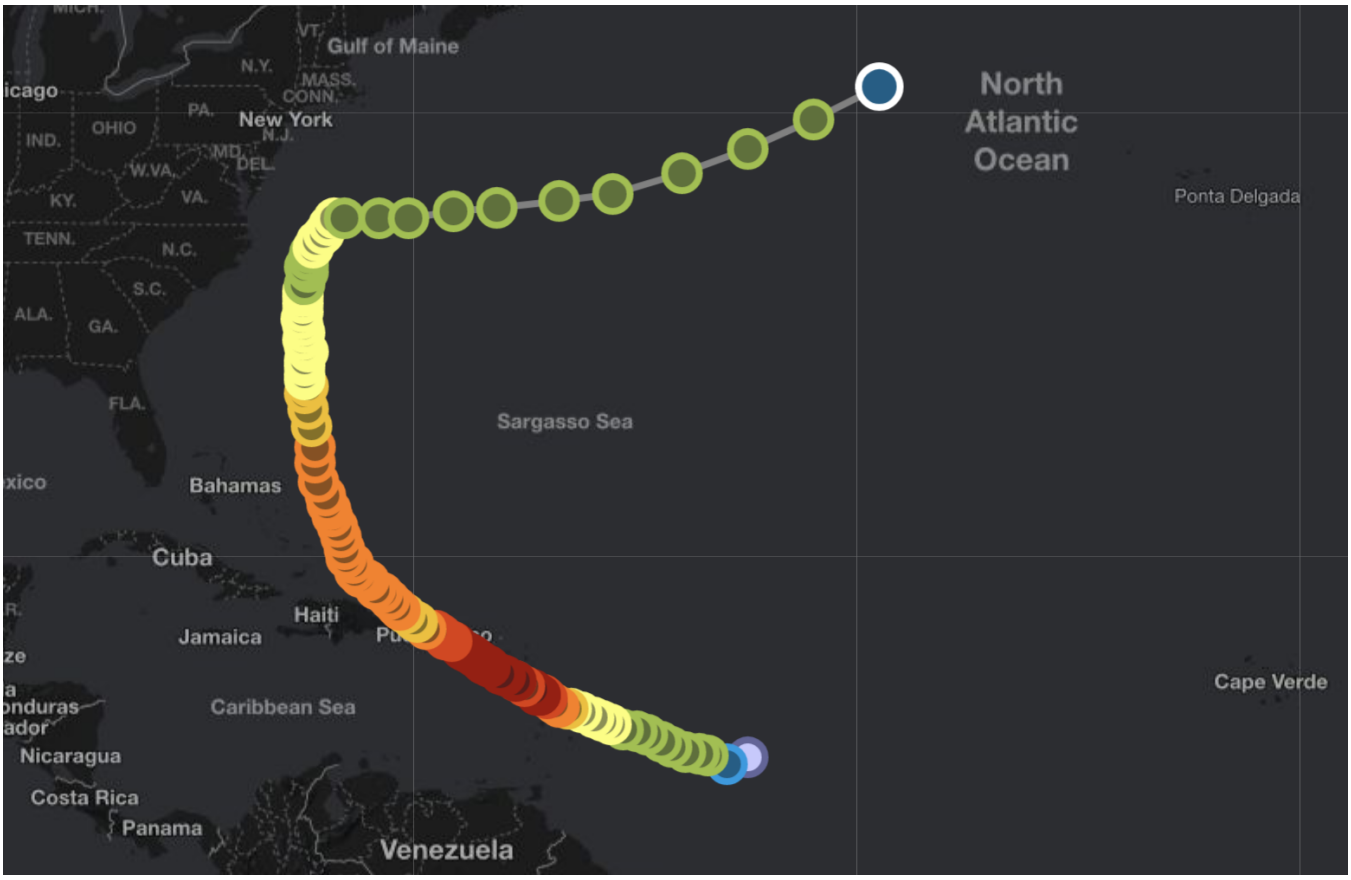

Figure 4.14: *Hurricane Maria actual path by The Weather Channel*[6]

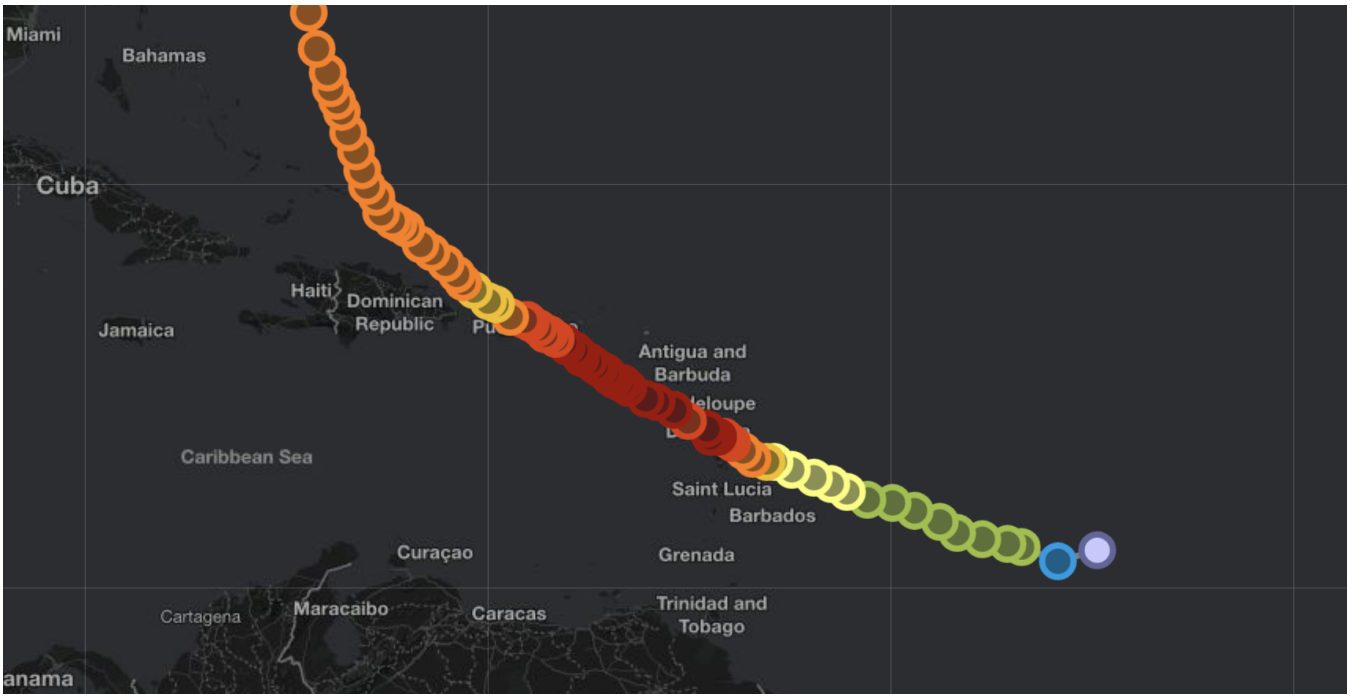

Figure 4.15: *Closer look to Hurricane Maria's actual path by The Weather Channel*[6]

Chapter 5

Conclusion

In this report, we explored Latent Dirichlet Allocation to learn topics from disaster dataset of hurricane Maria, Irma, Harvey and Mexico Earthquake. We explored two approaches, one where we considered the whole dataset as corpus and in the second approach we considered each document by day as a corpus for the LDA model. The second approach was more successful at providing a smaller picture for the disasters, where we were able to get county and town level information in relation to the hurricanes and was able to plot an approximate hurricane path. The first approach was able to provide a bigger picture of hurricane timeline covering the days that were most significant for a given hurricane. We also discovered that the first approach was ambiguous in terms of topic interpretation because of the fact that a topic can relate to several disasters and also a term can have multiple meaning given the context. Therefore, a topic can be interpreted in several ways and it can lead to wrong conclusions.

Overall, we were able to learn the underlying structure of disaster tweets better using topic modeling. We were able to look closely at day by day accounts and were able to approximately predict the path of a hurricane.

Bibliography

[1] 2017 atlantic hurricane season, 2017. URL https://commons.wikimedia.org/wiki/File:Irma_2017_track.png.

[2] Gaoyang Li, Xiaohua Wang, Aijun Yang, Mingzhe Rong, and Kang Yang. Failure prognosis of high voltage circuit breakers with temporal latent dirichlet allocation. *Energies*, 10(11):1913, 2017.

[3] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[4] Hurricane irma advisory archive, 2017. URL https://www.nhc.noaa.gov/archive/2017/IRMA.shtml.

[5] 2017 atlantic hurricane season, 2017. URL https://weather.com/storms/hurricane-central/irma-2017/AL112017.

[6] 2017 atlantic hurricane season, 2017. URL https://weather.com/storms/hurricane-central/maria-2017/AL152017.

[7] Adam B. Smith. 2017 u.s. billion-dollar weather and climate disasters: a historic year in context, January 2018. URL https://www.climate.gov/news-features/blogs/beyond-data/2017-us-billion-dollar-weather-and-climate-disasters-historic-year.

[8] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1105–1114. International World Wide Web Conferences Steering Committee, 2018.

[9] Enrico Steiger, Bernd Resch, and Alexander Zipf. Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30(9):1694–1716, 2016.

[10] Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, 2013.

[11] Bernd Resch, Florian Usländer, and Clemens Havas. Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, 45(4):362–376, 2018.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[13] Radim Rehurek. models.ldamodel latent dirichlet allocation, 2018. URL https://radimrehurek.com/gensim/models/ldamodel.html.

[14] 2017 atlantic hurricane season, 2017. URL https://weather.com/storms/hurricane-central/harvey-2017/AL092017.