Case Study

Chennai Flood

March 2017

Problem

- In November–December 2015, the annual northeast monsoon generated heavy rainfall in the Coromandel Coast region of the South Indian States of Tamil Nadu and Andhra Pradesh, the union territory of Pondicherry with the city of Chennai particularly hard-hit
- The ‘black swan’ caused an enormous destruction
- Around 500 people lost their lives and over 18 lakh people were displaced
- Damages and losses ranging from ₹50,000 crore to ₹100,000 crore
- The costliest natural disasters of the year 2015
  - The flooding has been attributed to the *El Niño* phenomenon
- Rescue and rehabilitation efforts spanned well over 2 weeks

Social Media for the Problem

- Coordination of **relief efforts** over social media platforms such as Twitter, Facebook and WhatsApp
  - Volunteers, NGOs and other rescue parties
    - Sent out alerts, requested and shared useful information related to the flooding on social media
  - Others used social media
    - To check up on the loved ones, share information, express opinions and sending out requests for help

Social Media for the Problem

- Media content included
  - Needy
  - Referee for the needy
  - Helpers

Outline

- Problem Definition
- Dataset collection
- Dataset and Analysis
- Mining
- Limitations
- Challenges in Real Time Analysis of Tweets
- Applications and Scope of Future work

Problem Definition

- Event Detection
  - To analyze the crisis communication on Twitter happened during Chennai Floods
  - To discover patterns and themes of communication
    - How the twitter used to share information
    - How it shaped response to the crisis

Problem Objectives

- Analysis of social media interactions
  - Understand the different subject of interactions
  - Information sharing
  - Grouping geographies into risk and the actions
    - Rescue
    - Food
    - Supplies
    - Ambulance calls

Problems to achieve this…

- Where to search for data
- Data really helpful
  - Preprocessing
    - Do it from Scratch
    - Use Tools
- Which method/technique to use

Dataset Collection - Twitter

- Worst days of the crisis
  - First two days of flood
  - 6000 tweets
  - With hashtag #ChennaiFloods

Tweet Characteristics

- A social media message posted on Twitter.com
- Restricted to 140 characters
  - Mostly contains text
  - Possible to embed URLs/pictures/videos/GIFs
- Contains components called hashtags
  - Words that capture the subject of the tweet
  - Prefixed by the '#' character
  - Also convey emotion (#sarcasm) or an event (#IndiaVsPakistan) or popular catchphrase in pop culture (#ICantBreathe)

Tweet Characteristics

- Users
  - Usernames or handles of those who post are recognized by the '@' symbol
  - A user can direct a message to another user by adding the handle, with the '@' symbol
- A retweet is a tweet by a user X that has been shared by user Y to all of Y's followers
  - A way of measuring how popular is the tweet

Tweet Dataset Collection

- Source
  - Twitter
  - GitHub

Tweet Dataset Collection

- Source
  - Twitter has an official API called Oauth
    - A token-based authentication system that indexes tweets that match a given search string and writes the output to a file
    - Free service and convenient to perform a quick and efficient extraction of tweets
    - A crucial limitation: Retrieve tweets only from the previous week

Tweet Dataset Collection

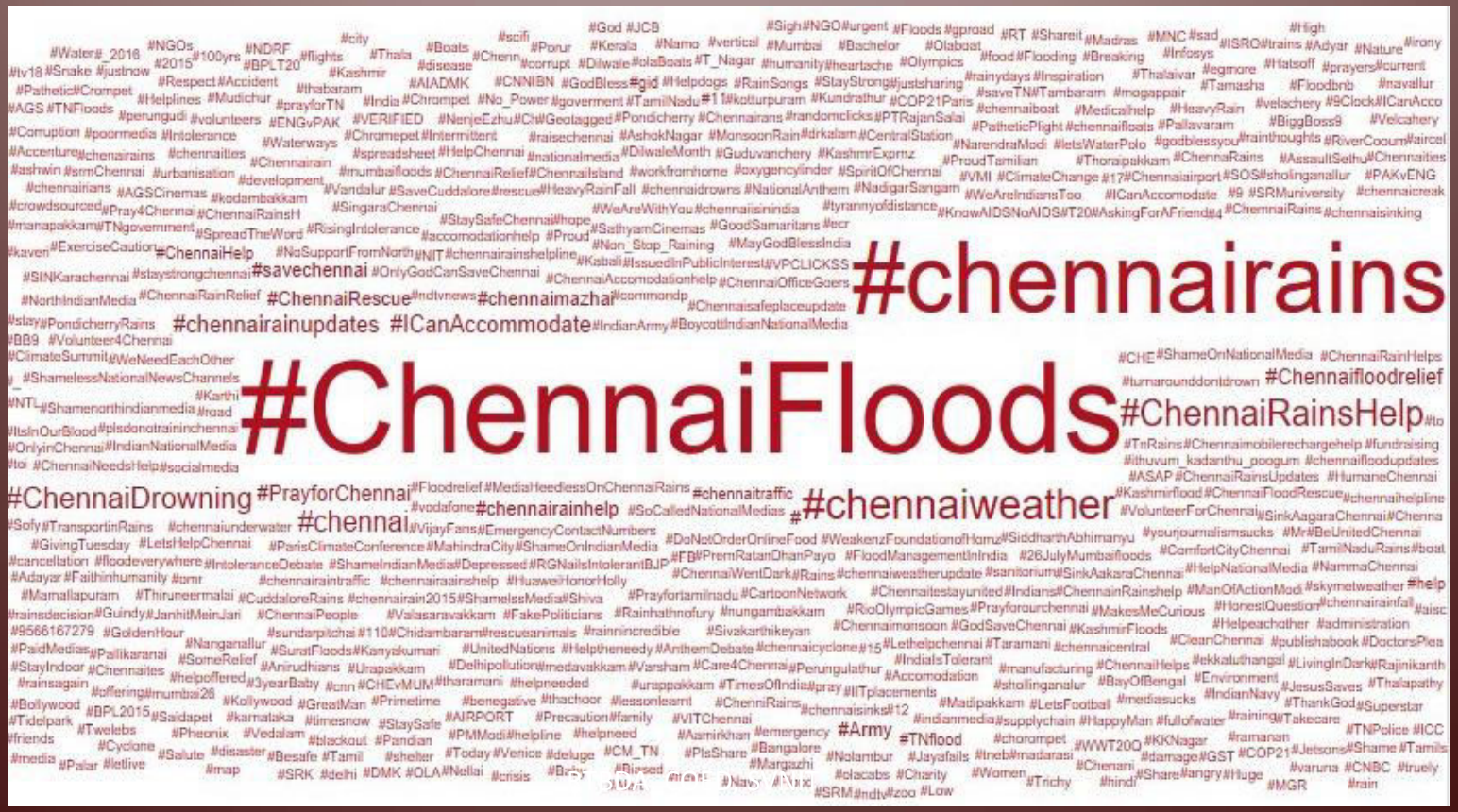
- Source
  - GitHub
    - A repository on the code sharing portal
    - Utilize to search for and to extract tweets
    - Python language was used (in an Ubuntu Linux environment) to perform the extraction
    - The extracted tweets were written to a comma-separated value file

Data Preparation & Exploration

- Tweets
  - Contains hashtags and users who tweeted about this event
  - To separate these
    - Ready tool of R using regular expressions

Data Preparation & Exploration

- Wordcloud shows the frequency of hashtags used in the tweets



Data Preparation & Exploration

- From the hashtags used, the following themes are evident, other than the disaster itself:
  - Sympathy (#PrayForChennai)
  - Requests for help (#savechennai, #ChennaiRainsHelp)
  - Information on further weather forecasts (#chennaiweather)
  - Information on specific areas in Chennai (#airport, #Chromepet)
  - Cautionary messages (#ExerciseCaution)
- Various hashtags for the same topic are observable
- This would make it challenging to separate all tweets on the subject

Data Preparation & Exploration

- Tweets were parsed into a corpus for text analysis
- Steps to clean the corpus to prepare for analysis
  - **Removing numbers:** TweetIDs are number generated by Twitter to identify each tweet Numbers as such of no use, hence they are discarded
  - **Removing URLs & links:** Many tweets contained links to webpages and videos elsewhere on the Internet, remove with regular expressions
  - **Removing stopwords:** Stopwords are words in English that are commonly used in every sentence, but have no analytical significance. Examples are ‘is’, ‘but’, ‘shall’, ‘by’ etc. Removed by matching the corpus with the stopwords list in the tm package of R. Expletives were also removed.

Data Preparation & Exploration

- Steps to clean the corpus and to prepare for further analysis
  - **Removing non-English words:** The corpus generated after performing the last 3 steps were broken into their constituent words and all English words and words less than 4 characters long were removed
    - What remained was the list of non-English words, words that mentioned areas in Chennai, words that are actually Tamil words written in English and misspellings of normal English words.
    - The non-English words for the names of localities were used to form a wordcloud
  - **Stemming words:** *'the process of reducing inflected/derived words to their word stem, base or root form—generally a written word form'*
  - Stemming is sometimes derivationally related-forms of a word to a common base form
  - Many methods exist to stem words in a corpus

Data Preparation & Exploration

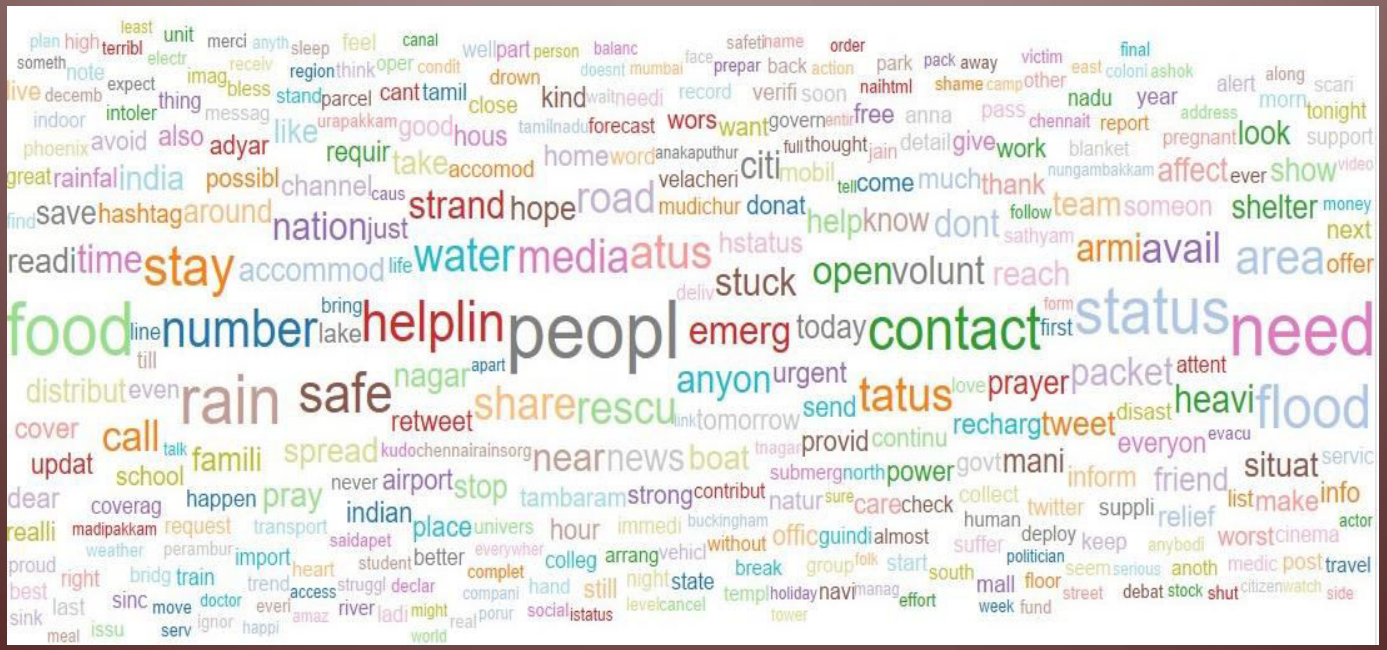
- **Stemming**
  - **Suffix-dropping Algorithms**
    - The last parts of all the words get truncated
    - For example, words like ‘programming’, ‘programmer’, ’programmed’, ‘programmable’ can all be stemmed to the root ‘program’
    - Whereas, ‘rescuing’, ‘rescue’, ‘rescued’ are stemmed to form ‘rescu’, which is not a word or a root
    - This method was chosen for this study for simplicity
  - **Lemmatisation Algorithms**
    - Each word is the determination of the *lemma* for a word in the corpus
    - This is done with the understanding of the context, part of speech and the lexicon for the language
    - For example, ‘better’ is related to ‘good’, ‘running’ is related to ‘walk’ and so on

Data Preparation & Exploration

- **n-gram Analysis**
  - Each word is broken into a part of its whole by ‘n’ characters, and the one that makes most sense is retained
  - For example, for n=1 (unigram), the letters ‘f’, ’l’, ’o’, ’o’, ’d’ are individually parsed from ‘flood’
  - For a higher n (say n=5), ‘flood’ is retained from ‘flooding’, although at n=4, ‘ding’ can also be construed as a word
- **Removing punctuation**
  - Punctuation marks make no impact to the analysis of text and are hence removed
- **Stripping whitespace**
  - Words that have extra whitespaces at the beginning, middle or end are subjected to a regular expression that removes the whitespace and retains only the words themselves
- **Checking for impure characters**
  - A check on the corpus after the modifications made thus far revealed that some URLs were left behind, due to the removal of whitespaces, numbers and punctuations
  - Regular expressions were used to remove them

Data Preparation & Exploration

- Word Frequencies and Associations after cleaning



Data Preparation & Exploration

- Word Frequencies and Associations
  - After the necessary cleaning, another wordcloud was plotted to understand the most frequently used terms

Data Preparation & Exploration

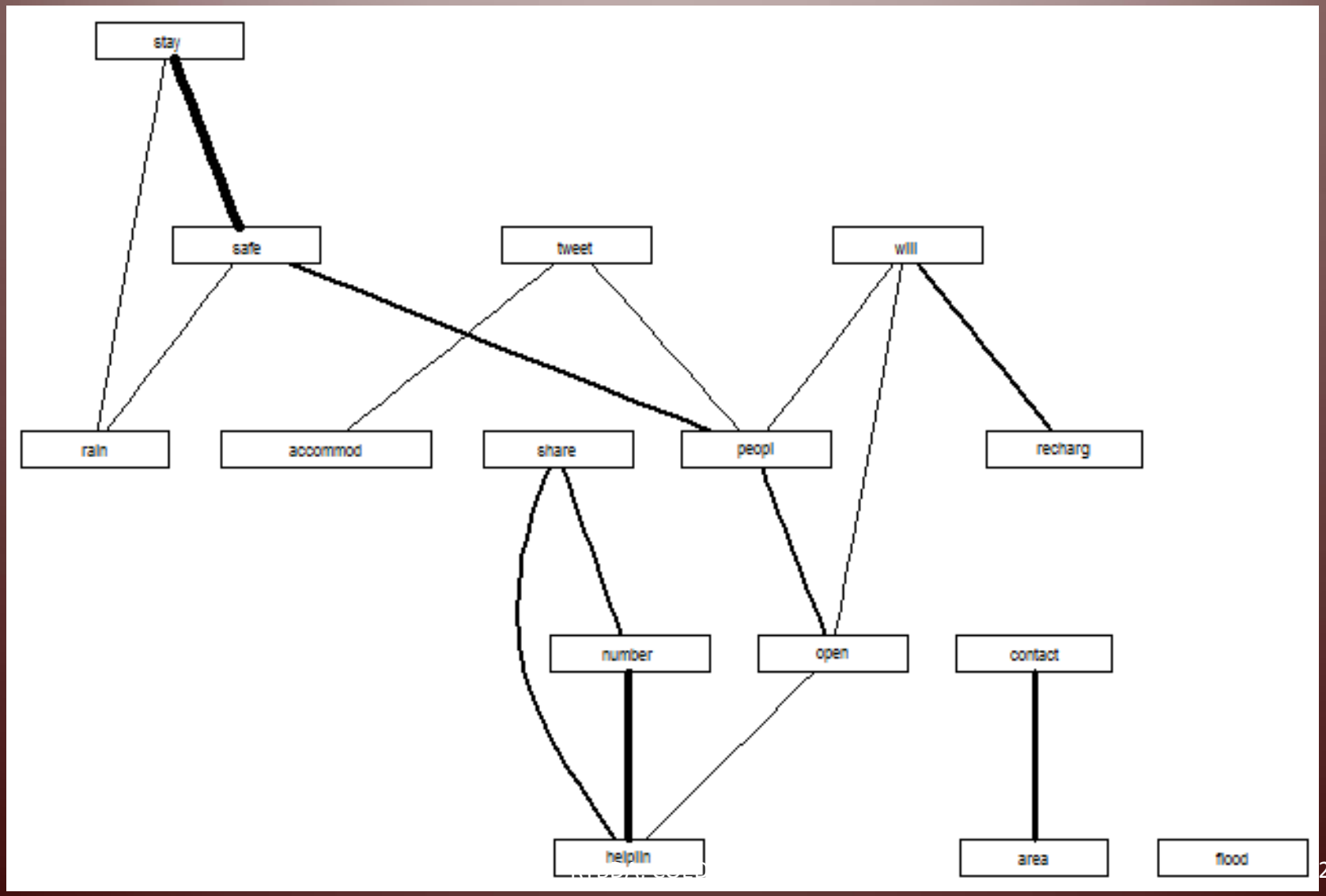
- Word Frequencies and Associations Observations
  - ‘people’, ‘stay’, ’safe’, ‘food’ etc. -
    - Few simple words, repeat more often than others
    - Indicate **immediate reactions and responses to the crisis**
  - ‘street’, ‘nagar’, ‘mudichur’, ‘road’ etc. - Some infrequent terms
    - Provide information about the situation in a few areas
  - ‘pray’, ‘hope’ and ‘proud’
    - Messages that mostly convey sympathy or hope
  - ‘channel’, ‘media’ etc.
    - Tweets about news reports covering the crisis
  - ‘help’, ‘please’, ‘food’ etc.
    - Tweets that requested for volunteers to participate in rescue and rehabilitation efforts

Data Preparation & Exploration

- A few words are associated strongly with the most frequent words than others
- The Table describes the most common word associations

Word	Associated with	%
help	hand	26%
number	helpline	44%
	emergency	33%
stay	safe	59%
	indoor	32%
	strong	31%
news	channel	50%
open	phoenix	33%
	cinema	29%
	mall	27%
	royapetah	27%
	door	25%
media	nation	43%
	social	32%
packet	food	35%
	deliv	28%
	aadambakkam	27%
	biscuit	27%
	thiruvallur	27%
food	packet	35%
	distribut	26%

Data Preparation & Exploration



Clustering and Topic Modeling

- Challenge is to make meaningful interpretations from the huge volumes of data that need to be processed
  - From a crisis like the Chennai floods strikes, as large number of similar tweets get generated
- One solution is to cluster similar tweets together after performing the necessary Exploratory Data Analysis operations that it becomes easier to manage the flow of information
- 1 Hierarchical Clustering

Clustering and Topic Modeling

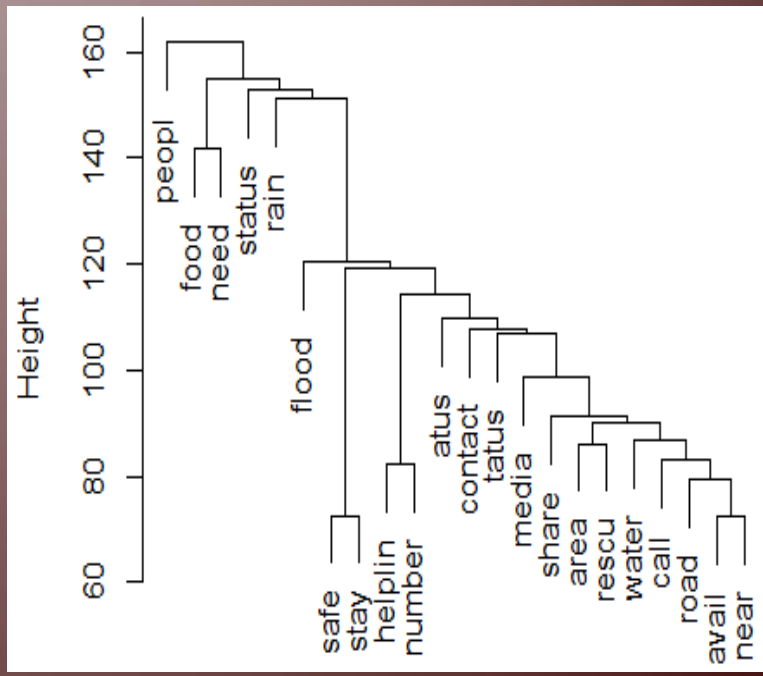
- Hierarchical Clustering
  - Attempts to build different levels of clusters
  - Two types:
    - **Agglomerative:** where we start out with each document in its own cluster. The algorithm iteratively merges documents or clusters that are closest to each other until the entire corpus forms a single cluster. Each merge happens at a different (increasing) distance.
    - **Divisive:** where we start out with the entire set of documents in a single cluster. At each step the algorithm splits the cluster recursively until each document is in its own cluster. This is basically the inverse of an agglomerative strategy.
  - The results of hierarchical clustering are presented in a dendrogram

Clustering and Topic Modeling

- Hierarchical Clustering
  - Use the R function, hclust()-agglomerative method
  - The following steps explain hierarchical clustering in simple terms:
    - Assign each document to its own (single member) cluster
    - Find the pair of clusters that are closest to each other and merge them, leaving us with one less cluster
    - Compute distances between the new cluster and each of the old clusters
    - Repeat steps 2 and 3 until you have a single cluster containing all documents
    - To perform this operation, Convert the corpus into a matrix with each tweet (or 'document') given an ID. Remove corpus, extremely sparse rows, i.e. rows with elements that are part of less than 2% of the entire corpus. Here used **Ward's method for hierarchical clustering**

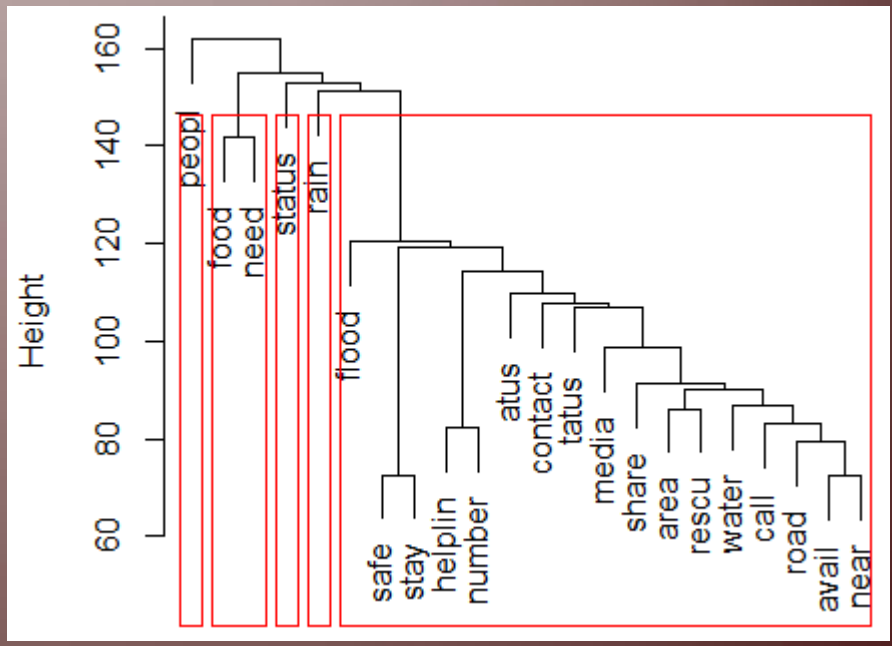
Clustering and Topic Modeling

- Hierarchical Clustering
  - The dendrogram output interpreted as:
    - Farther the nodes, greater is the dissimilar than the closer the node
    - The height of each node in the plot is proportional to the value of the intergroup dissimilarity



Clustering and Topic Modeling

- Interpretation of distinct clusters of tweets from the dendrogram:
  - Talk about general information about affected individuals, areas and news about the crisis
  - Talk about food, supplies and rescue efforts
  - Describe the weather, forecasts of rain and further developments
  - Caution people on risky areas and share information on relief efforts

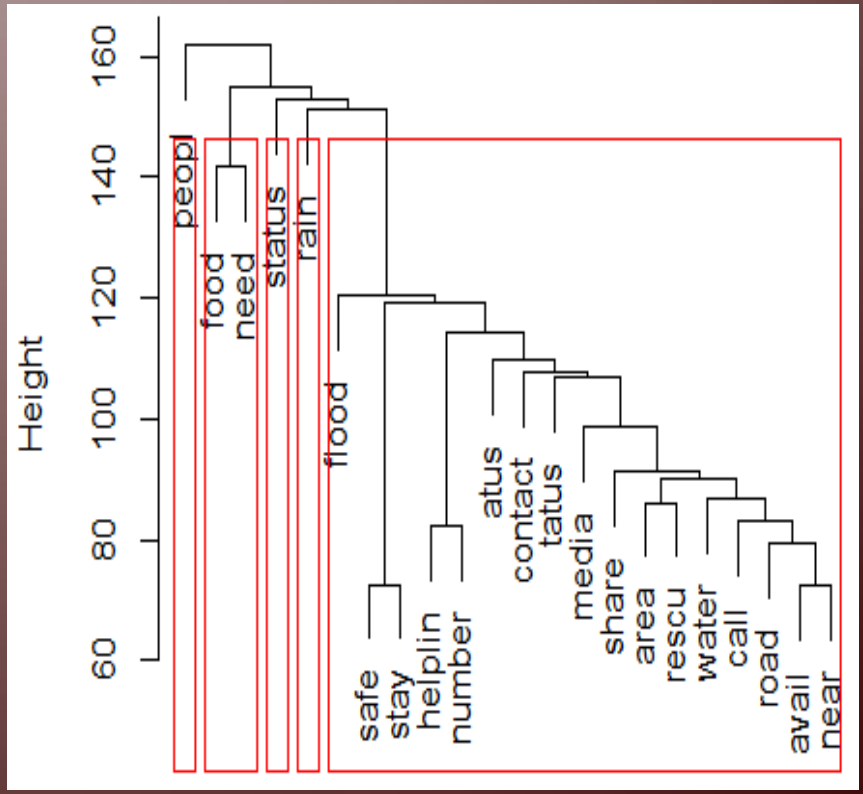


Clustering and Topic Modeling

- Interpretation of distinct clusters of tweets from the dendrogram:

  -- Seen that there is a significant similarity between clusters of tweets; this is expected as the terms used across tweets are more or less similar

  -- But, No locality or specific names are mentioned as the clustering was performed on a matrix that did not contain such rarely-occurring terms



Clustering and Topic Modeling

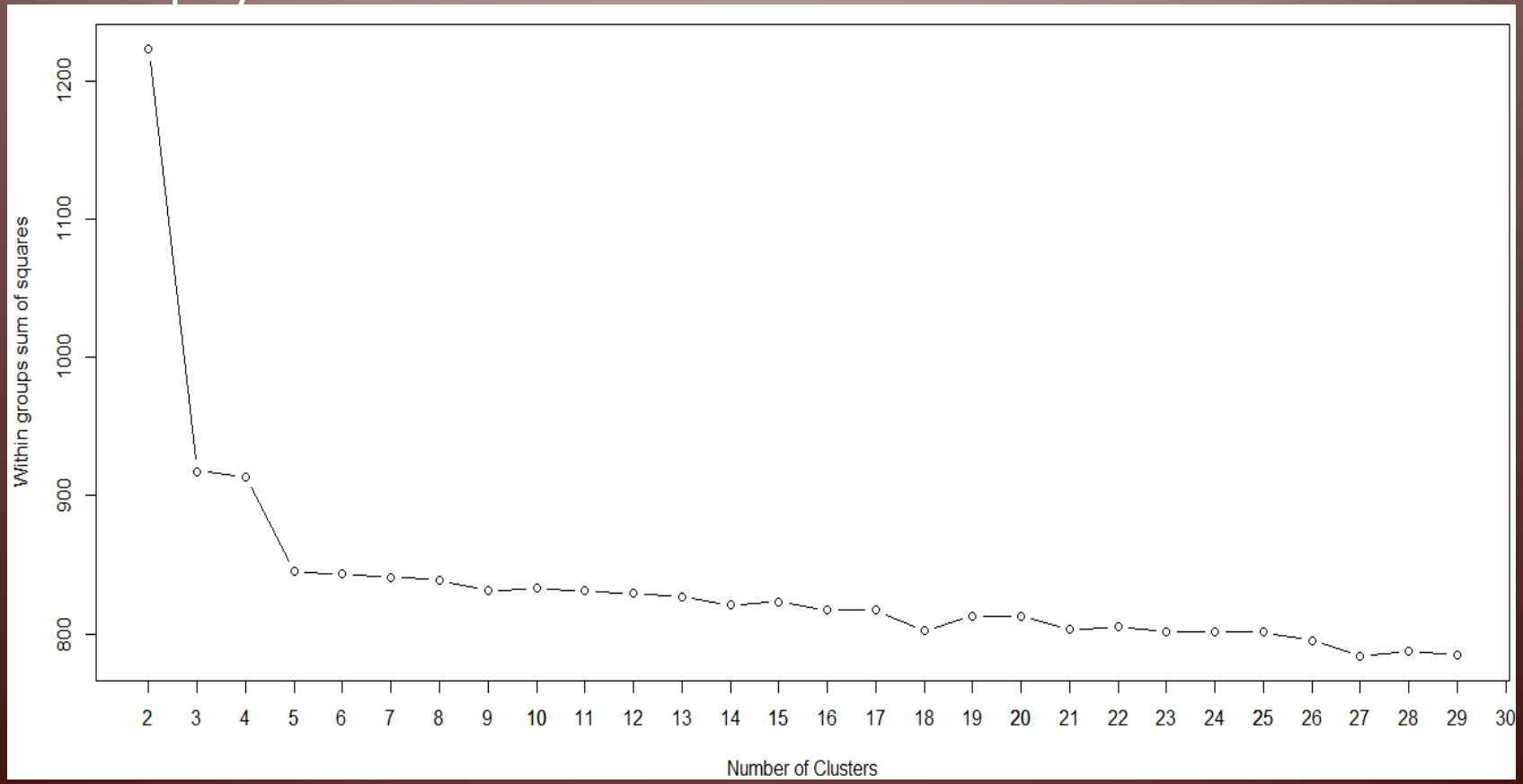
- 2. K-Means Clustering
  - As opposed hierarchical clustering, where one does not arrive at the number of clusters until after the dendrogram, in K-means, the number of clusters is decided beforehand
  - The algorithm then generates k document clusters in a way that ensures the within-cluster distances from each cluster member to the centroid (or geometric mean) of the cluster is minimized
  - A simplified description of the algorithm is as follows:
    - Assign the documents randomly to k bins
    - Compute the location of the centroid of each
    - Compute the distance between each document and each centroid
    - Assign each document to the bin corresponding to the centroid closest to
    - Stop if no document is moved to a new bin, else go to step

Clustering and Topic Modeling

- 2. K-Means Clustering
- Choosing k
  - The most significant factor of employing k-means clustering is choosing the no. of clusters, 'k'. The '*elbow method'*, wherein the SUM of Squared Error (SSE, the sum of the squared distance between each member of the cluster and its centroid) decreases abruptly at that value that is theoretically the optimal value of k, is widely applied to arrive at k
  - When k is plotted against the SSE, observed that the error decreases as k gets larger; this is because when the number of clusters increases, they become smaller, and hence the distortion is also smaller

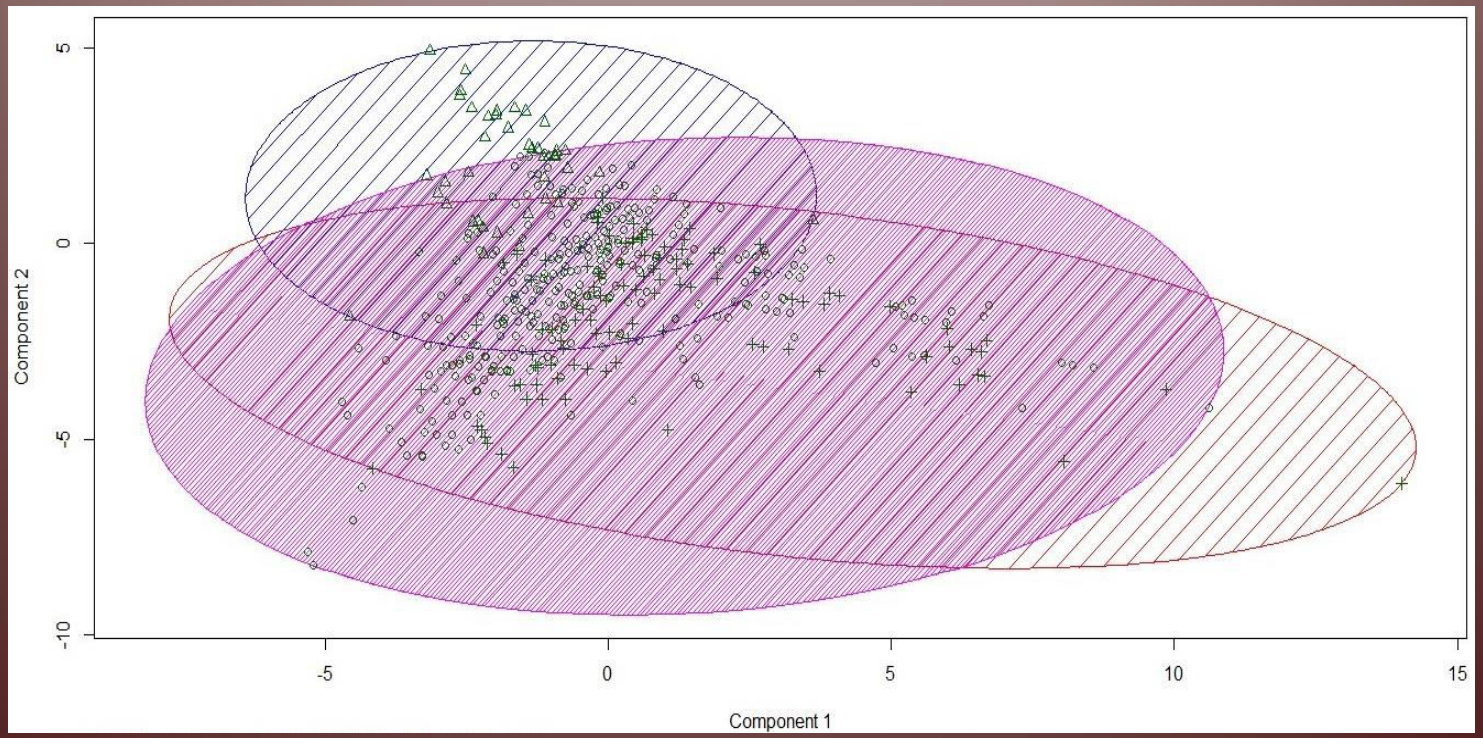
Clustering and Topic Modeling

- Here, the optimal value for k=3, where the SSE decreases abruptly



Clustering and Topic Modeling

- ClusPlot(m3)- With k=3
- These two components explain 15.01% of the point variability



Clustering and Topic Modeling

- The plot clearly shows that there is only marginal dissimilarity with a corpus at 98% sparsity
- Which can be seen from the top 10 words in each of the three clusters
  - Cluster 1: rain need status flood helplin number contact stay safe status
  - Cluster 2: food need contact avail peopl near water area call status
  - Cluster 3: peopl safe stay flood rain need near road contact media

Clustering and Topic Modeling

- With respect to clustering, subject matter and corpus knowledge is the best way to understand cluster themes. With the insights gleaned thus far, it is reasonable to assume the following:
- Cluster 1: rain need status flood helplin number contact stay safe status
- Cluster 2: food need contact avail peopl near water area call status
- Cluster 3: peopl safe stay flood rain need near road contact media

  - Cluster 1 contains news updates and cautionary
  - Cluster 2 contains messages about requests for help and volunteers
  - Cluster 3 contains messages about area-specific updates and some cautionary

Clustering and Topic Modeling

- 3. Topic Modeling
  - Another technique to deduce the themes of Twit text
  - A type of statistical model for discovering the abstract “topics” that occur in a collection of documents (tweets in this case)
  - Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: in this case, ‘help’ is quite common to almost every tweet
  - Base: A document typically concerns **multiple topics in different proportions**; thus, in a document that is 10% about subject A and 90% about subject B, there would probably be about 9 times more words about ‘B’ than words about ‘A’
  - Various algorithms are there, but the most common algorithm in use is Latent Dirichlet Allocation (LDA)

Clustering and Topic Modeling

3 Topic Modeling

- Latent Dirichlet Allocation (LDA)
  - A statistical model
  - Allows the possibility of a document to arise from a combination of topics
  - Uses probabilities-word/topic and word/documents
  - For example, the Tweet may be classified as 90% information & 10% sympathy/hope

Clustering and Topic Modeling

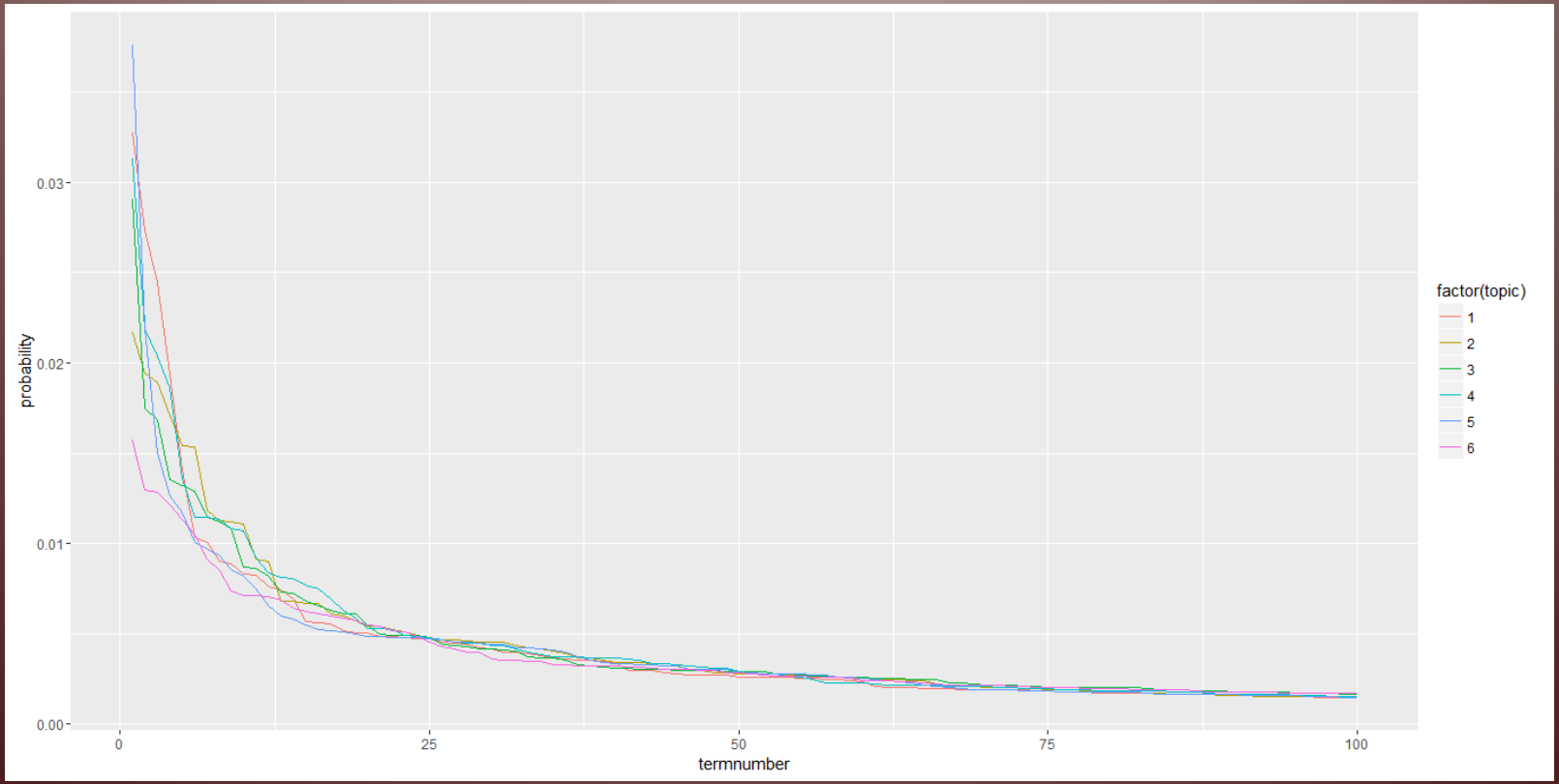
- 3 Topic Modeling
  - Performance of LDA with an objective to discover 6 topics
  - The output gives the following set of words for each topic

Topic 1	food, helplin, flood, contact, status, number	General Information
Topic 2	peopl, need, food, rain, packet, flood	Requests
Topic 3	contact, need, rain, stay, flood, number	General Information
Topic 4	rain, safe, peopl, stay, need, share	Caution
Topic 5	media, status, nation, volunt, area, share	Volunteers & relief efforts
Topic 6	peopl, food, status, rain, need, tatus	News about crisis

- Topics 1 & 3 are quite similar; this is in agreement with the results of the K-means exercise

Clustering and Topic Modeling

- 3 Topic Modeling



Summary

- The general sentiment across the tweets render the tweets quite similar and thus crucial information like the **worst-hit areas can be identified by analyzing tweets and performing basic text analytics**
- Power of social media can be harnessed to great effect in times of crisis as the practice of **creating hashtags specific to individual crises to index tweets easily**
- Facebook launches ‘Mark Safe’ feature to those who have listed a crisis-hit location as their place of residence
- The government agencies and other relief agencies would do well to develop analytics capabilities focused on mining Twitter for real-time, tangible updates **to take meaningful action**

Limitations of this study

- Considers only 6000 tweets of the whole set of tweets that would have been sent on the subject
- Did not consider captions of pictures, news reports, and other social media reports which could have generated additional insights

Challenges to Real-Time Analysis of Tweets

- If the tweets contain no hashtags. In the dataset generated for this analysis, 1399 tweets (22%) had no hashtags. These tweets may also be highly relevant to the crisis but may be missed due to the lack of hashtags
- Twitter has a 140 character-limit on all tweets (not including pictures and videos). This leads users to **truncate or shorten words to forms that is easily interpretable to humans, but is challenging for a machine**. Eg: 'afcted ppl' is easily understandable to mean 'affected people' for a human, but not for a program. One way to solve this problem is to maintain a dictionary of such terms and match them in real-time

Applications & Scope for Further Work

- This is an **active field**. The power of social media will continue to be researched for new applications
- One area is **quashing rumors**. During the Chennai floods, quite a number of false ‘news reports’ and ‘alerts’ circulated on Facebook, Twitter and the mobile messaging application WhatsApp. Machine learning can be employed to check the veracity of social media by comparing contents from actual news reports

Applications & Scope for Further Work

- Every civic authority would do well to **develop a framework and system to manage crises also through social media**. This covers all disasters, both natural (earthquakes, floods, hurricanes) and man-made (terror strikes, shootouts)
- Media outlets and government agencies can work together in planning for incidents that are expected **by creating distinct identifiers and hashtags for each scenario** and making the public aware of them
- Disasters may strike at any time. While it may not be possible to prevent them, it is prudent to be prepared for unfortunate eventualities. Having a dedicated social network analysis platform to analyze information in real-time will definitely aid in this endeavor

References

- *The Capstone Project titled* ***Tapping Social Media Exchanges on Twitter: A Case-Study of the 2015 Chennai Floods***