CS 690V – Visual Analytics

Final Project Report
IEEE VAST Challenge 2011

December 18th, 2017

Kriti Shrivastava, Suhas Keshavamurthy

Mini Challenge 1- Characterization of an Epidemic Spread

MC 1.1 Origin and Epidemic Spread:

1. *Preprocessing*: Remove mentions, URLs, emojis, stopwords; stemming; tokenization
2. *Extract relevant tweets*: To filter the tweets, we have used the symptoms mentioned in the challenge description as our primary keywords. For example, flu, sick, fever, pain, fatigue, cough, vomiting, diarrhea were some of the chosen keywords. We then performed the following steps:
    - *Stem the keywords*: Since the words in our tweet dataset are stemmed, we stemmed the keywords as well. The main intention behind this was to include cases like- "I have been coughing all day" or "I have a terrible cough.". Stemming would help us extract both these tweets, otherwise we would have missed the tweet which contains the word "coughing" instead of the exact keyword "cough".
    - *Add similar words to the keyword search list*: For all the words in our list of keywords we found a list of similar words using the Word2Vec model. Performing this step helped us in retaining the tweets which might not contain the exact word as our keyword but contain words with a similar meaning. For example, "I fell ill today." and "I am feeling very sick.", both these tweets are relevant. We have the word "sick" in our search list but not the word "ill" but both of these tweets should be considered. Therefore, we have added words with similar meanings to our list of keywords.
    - *Filter tweets*: As a last step, we have extracted the tweets which contain one or more of our keywords. Out of more than a million tweets, we extracted around 55k tweets by using the disease symptoms as keywords.
3. *Plot relevant tweets on the map*: We plotted the filtered tweets on the map and we could clearly see the regions with high concentration of tweets. With this we were able to outline the affected areas on the map.

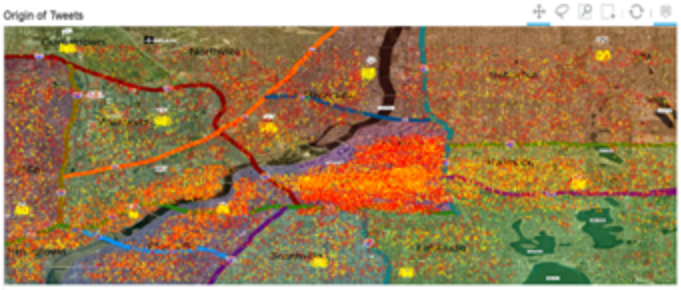

Figure 1. Plotting relevant tweets on the map. Red shows tweets by people who have tweeted about the topic before and the yellow tweets are by new people

4. *Get the trend in outbreak:* We also found the number of relevant tweets seen each hour and plotted it across time. With the help of this line graph, we observed that there is a spike in the number of cases on 18th and the number of tweets are significantly high over the next few days. Thus we were able to find the time period when the outbreak started to spread.

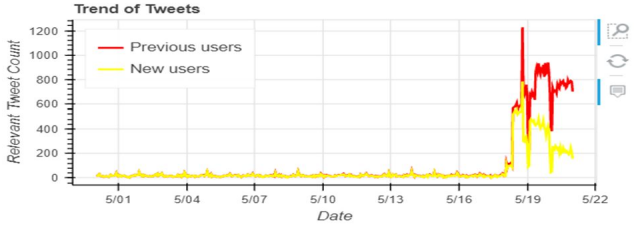

Figure 2. Trend of tweets over time

5. *Filter on outbreak start date:* Since now we have the affected area and date when the outbreak started, we try to narrow down the cause. Using the time slider, we selected the date from 05-16 to 05-19. Using the play button, we observed the rising spread of the disease in the affected regions. We see two distinct regions of disease spread. The intersection of these regions could be the possible ground zero location.

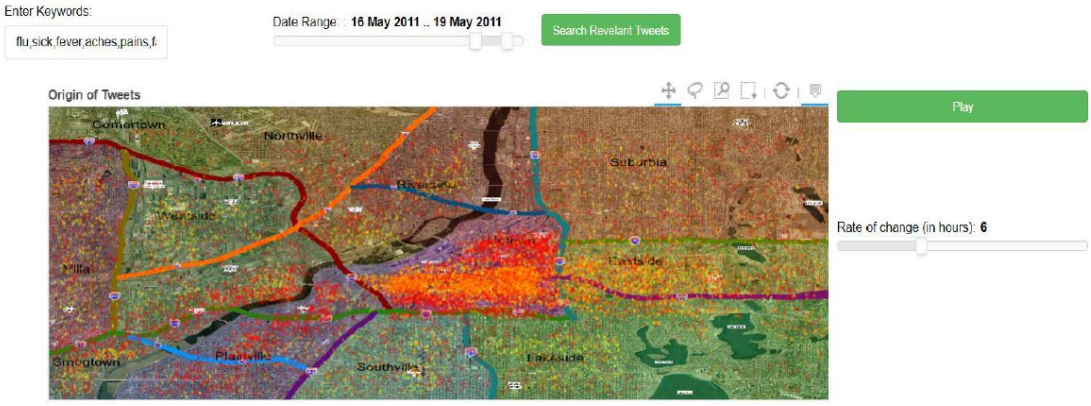

Figure 3. Filtering on date to find affected regions and approximate area of origin

6. *Select approximate area of origin:* We decided to closely observe the tweets originated from the location from step 5 around 2011-05-17. For this, we selected the approximate area of origin on the map(marked using red pen in the below screenshot) and filtered the tweets using the date.

Figure 4. Selecting approximate area of origin on the map and finding events

7. *Get possible causes of origin:* Once we had the subset of tweets with specific location and date, we tokenized and performed part of speech tagging on them. We then extracted the most frequent nouns that occur in these tweets to find suspicious events. These words are shown in a word cloud next to the map.
8. *Narrow down on the most probable cause:* Ignoring trivial words like day, it etc., there are 3 major events in the area in the given time period. We investigated each of these events further. One of them turns out to be a bomb drill in downtown. The second is a car accident on Highway 270. The third is a truck accident on highway 610. After looking at the tweets for the incidents and searching for relevant words and events for each of these three events, we inferred that the truck accident is the likely cause. We filtered the tweets further, now using the keywords like “truck”, “accident”, time and location and got a list of approximately 500 tweets. The location of the accident was narrowed down to I-610 Bridge over the VAST River.

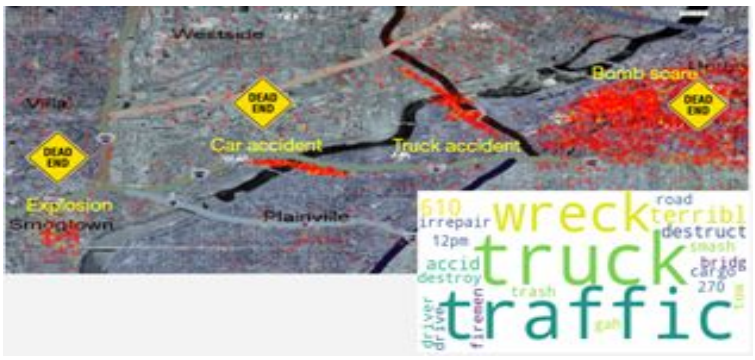

Figure 5. Finding the most probable cause and the words related to it

MC 1.2 Epidemic Spread

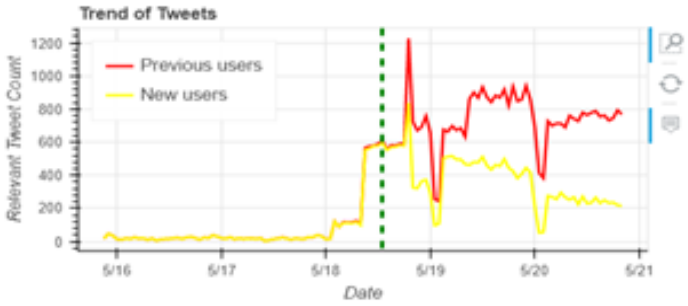

Figure 6. Timeline for epidemic

We closely observed the two distinct regions affected by the epidemic. For region 1(Downtown, Uptown and Eastside), we saw that the symptoms are mostly flu-like. We also saw that the wind direction corresponds to the direction in which the disease is spreading. For region 2(Plainsville, Smogtown) which is near the river, we saw that the symptoms are related to water-borne disease . Going back and looking at the truck accident related tweets, we found that the cargo from the truck was spilled in the river which may have caused the disease to spread in this region.

Figure 7. Spread in Region 1 and 2

From the trend graph of tweets, we observed that there was a spike in reporting of disease on 18th. There was a slump for a short time before increasing again. These two spikes can be attributed to the airborne and waterborne spread of disease. The trend graph also plots the number of new users and old users reporting in the microblogs. We observed that the trend of new users reporting of any kind of symptoms is reducing day by day. From this we concluded that the outbreak is contained and additional steps to control it is not necessary.

Learnings and observations:

1. We tried two approaches Word2Vec and Wordnet Synset for finding similar words. Word2Vec gave better results than Synset as Word2vec also considers the context of the words.
2. We started with scatterplot on the city map and the line graph and then added the functionalities of search bar, play and pause buttons to facilitate the analysis.
3. Our approach is generalizable and we used the same search bar and WordClouds to find solution to various problems for this challenge.
4. The idea of plotting the wind direction on the map was borrowed from previous challenge submissions.
5. Using the POS tagging and noun extraction approach was sufficient to reach to the truck accident as the possible cause of epidemic.
6. We did not identify the tweets where the person who posted the tweet is talking about someone else. This might have affected our analysis.

Mini Challenge 2 - Computer Networking Operations

MC 2.1 Events of Interest

To find out the events of interest, we first parsed the different network log files that were provided and converted the data in a consistent format. Next, for each of the given network log file, we counted the entries for every minute for all 3 days. This would give us an idea about the network traffic per minute. Finally, we plotted 3 heat maps that show the network traffic for Firewall, IDS and security log files. Using these heatmaps, the CNO team member would very easily be able to trace any anomalies in the company's network traffic. The team member can select the date of interest and then very quickly observe the entire computer network behavior for the chosen day at a glance.

For day 1, we saw very high rise in traffic in the Firewall heatmap. It started from 11:39 and lasted till 12:51. IDS heatmap also showed some anomalies during the same time. The IDS heatmap also showed some suspicious event from 11:15 - 11:41. On closer look at the Firewall and IDS logs for these timestamps, we figured out that there was a Denial of Service Attack and Port Scan in these time frames. For day 2, from the IDS and Firewall heatmaps, we found suspicious activities between 09:01 - 09:27 and 10:56 - 12:28. On drilling down further, these were identified as port scans. We could not find any event of interest for day 3.

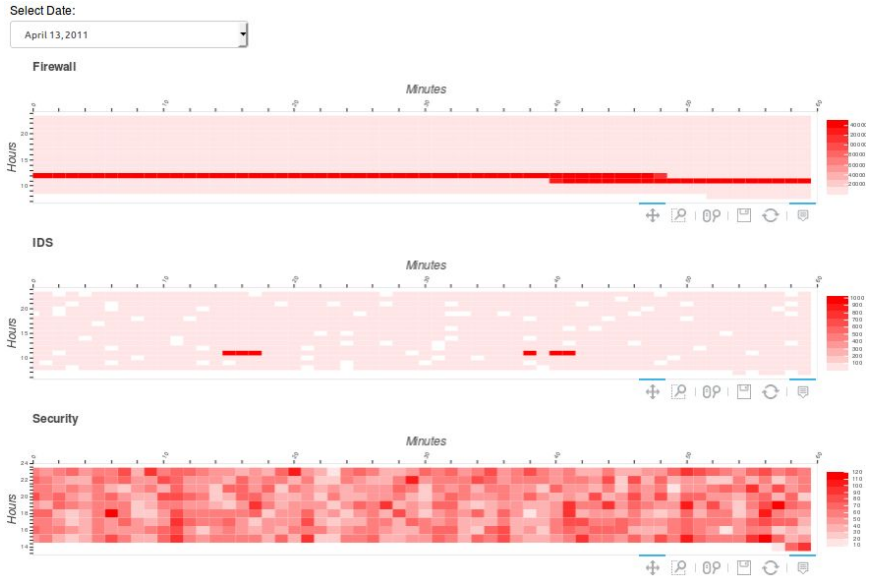

Figure 8. Day 1 - Situational awareness interface for company's computer network

MC 2.2 Timeliness

The interface shows data for the entire day at a glance and since the traffic is plotted per minute, the team member can identify the occurrence of an interesting event as early as the next minute.

Learnings and Observations:

1. Heatmaps worked well in providing quick information at a glance.
2. We chose this approach as it can be easily extended to accommodate real-time streaming data.
3. For the time duration of 3 days, we found interesting results using the Firewall and IDS logs but the same approach did not work for the given security logs. This is because the frequency of entries in the security logs does not give enough information to identify a potential security threat and additional filtering/preprocessing is required to better analyze the security log data.
4. We did not create visualizations to get to the details of what was happening. We lacked the domain knowledge to identify the key features for plotting such detailed visualizations.
5. To understand the kind of attack or the event that occurred, this interface relies on the team member's domain knowledge. The interface could only point out a possible suspicious activity, but to figure out the criticality of the event and its possible effects, a human is required in the loop. A member with domain expertise needs to manually go through the network logs entries with the timestamp alerted by the system to drill down to the actual cause.

Mini Challenge 3 - Investigation into Terrorist Activity

MC 3.1 Potential Threats:

1. Topic Modeling

The first step in the process was topic modeling. There are multiple topic modeling approaches for text like Latent Semantic Indexing, Probabilistic Latent Semantic Indexing, Latent Dirichlet Allocation and Hierarchical Dirichlet Process. We tried the following 2 approaches for this challenge,

- Latent Dirichlet Allocation (LDA)
- Hierarchical Dirichlet process (HDP)

LDA is one of the most popular topic modeling algorithms and provides fairly good results of the underlying topic distribution. The topic distribution from HDP did not yield as good results as LDA. Hence, LDA was chosen for obtaining the underlying topic distribution. The experimented with 10, 50 and 100 topics and decided to select 100 topics. It is possible to select higher number of topics, but we chose to stick with 100 as it becomes increasingly difficult to browse through all the generated topics as the number increases.

The list of topics is plotted in a datatable in Bokeh. Each row contains the topic number and the significant words which contribute to that particular topic. The user can infer the topic from these keywords. For example: in figure 9 below, topic 14 consists of 'Franc + share + stock + company + ...'. It is clear that this topic is related to economic markets. Similarly Topic 16 talks about 'car + auto + vehicle + ford + ...' which is clearly related to automobiles.

A reverse index is created which consists of the topics and the probability of each document belonging to that particular topic. This is plotted as a simple scatter plot (figure 10) of topic vs probability distribution of the news articles. Hover is enabled on the plot to show the headline of the document when user hovers over any of the points.

#	No	Topic Distribution
13	13	"rate" + "market" + "price" + "growth" + "stock" + "economy" + "month" + "week" + "report" + "analyst"
14	14	"franc" + "share" + "stock" + "company" + "french" + "france" + "belgian" + "de" + "billion" + "july"
15	15	"site" + "one" + "mr" + "family" + "news" + "time" + "may" + "stacy" + "research" + "say"
16	16	"car" + "auto" + "vehicle" + "ford" + "sale" + "motor" + "gm" + "truck" + "maker" + "model"
17	17	"mr" + "russian" + "crabb" + "rebel" + "war" + "russia" + "military" + "chechen" + "moscow" + "president"
18	18	"diamond" + "video" + "paper" + "border" + "pulp" + "photo" + "noble" + "one" + "new" + "printer"
19	19	"bond" + "million" + "yield" + "debt" + "treasury" + "peso" + "issue" + "market" + "security" + "spread"
20	20	"animal" + "chicago" + "printer" + "wildlife" + "right" + "club" + "past" + "one" + "hunting" + "official"
21	21	"yen" + "japan" + "japanese" + "company" + "tokyo" + "billion" + "share" + "penny" + "million" + "bank"
22	22	"steel" + "thailand" + "say" + "company" + "ton" + "thai" + "security" + "crescent" + "million" + "would"
23	23	"bill" + "would" + "state" + "law" + "federal" + "say" + "rule" + "house" + "new" + "congress"
24	24	"investigator" + "bomb" + "explosive" + "official" + "fbi" + "found" + "ge" + "wreckage" + "mr" + "obryan"
25	25	"mr" + "mf" + "would" + "irs" + "libya" + "one" + "say" + "vantassel" + "tax" + "codi"
26	26	"stock" + "market" + "dollar" + "index" + "dow" + "jones" + "trading" + "mark" + "rate" + "share"
27	27	"run" + "hit" + "game" + "inning" + "two" + "first" + "three" + "homer" + "season" + "yard"
28	28	"travel" + "pet" + "woman" + "one" + "politics" + "new" + "aid" + "sex" + "line" + "political"

Figure 9. List of topics generated by the LDA model

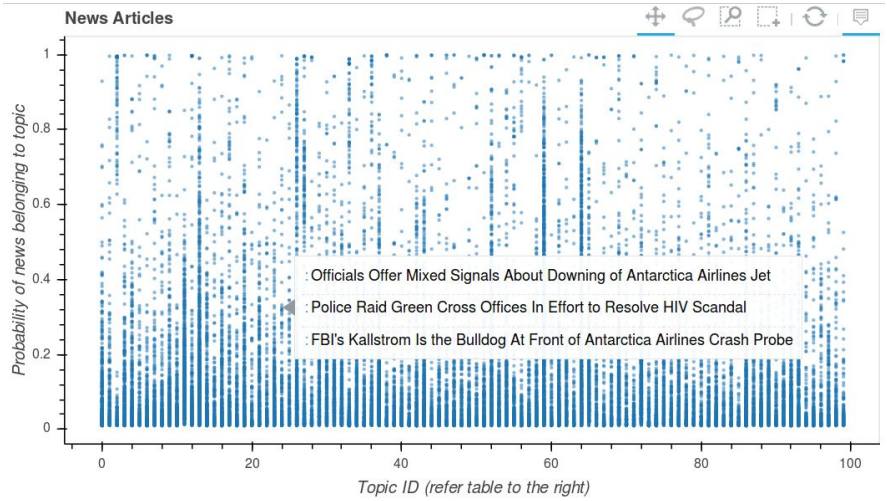

Figure 10. Distribution of news reports over topics

2. Clustering of News Reports

User can select the documents on the scatter plot using the box select or lasso tool. The selected articles are then plotted as connected graphs; the nodes represent the news articles and the length of the edges denote the similarity between the end nodes. We have used **cosine similarity** to measure the similarity between different news articles. The slider can be used to set the desired threshold for similarity. User can then investigate the clusters to find news articles of interest. User can select clusters of interest from this graph and the datatable on the right is populated with the headlines of the news articles selected.

An option is provided for the user to remove the irrelevant(not related to threats) news articles from the list. This helps to prune down the total number of news articles. For ex: There is a group of articles which talk about baseball games. This group of articles can be extracted easily and eliminated from consideration. Similarly, we can easily identify a group of news articles which are closely related to the stock market. These set of articles can also be easily eliminated from the corpus as not relevant for our investigation.

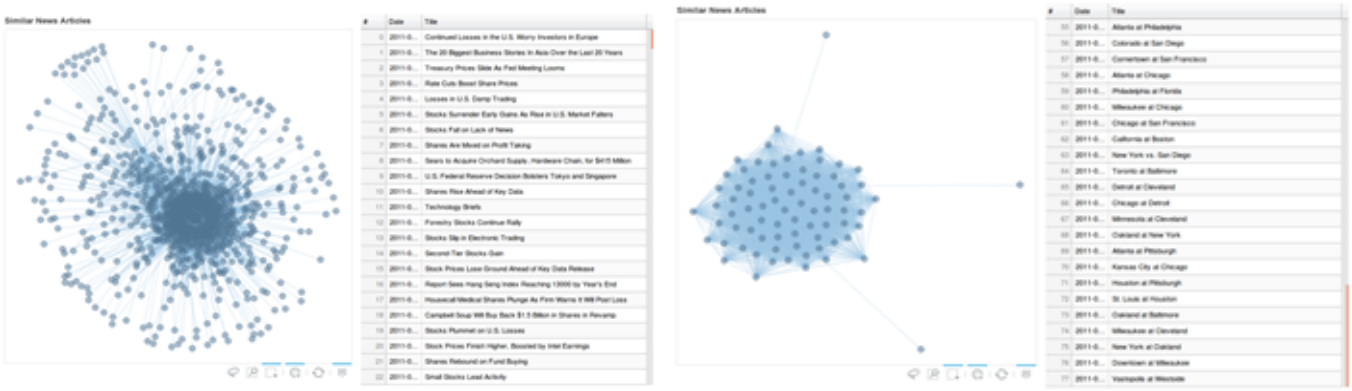

Figure 11. Selection of topic related to Baseball and Stock Market

If we find clusters that are of interest then they can be selected and added to another datatable that will contain all articles that require further analysis and investigation.
After applying the above-mentioned process on the data given, we were able to arrive at the below set of results.

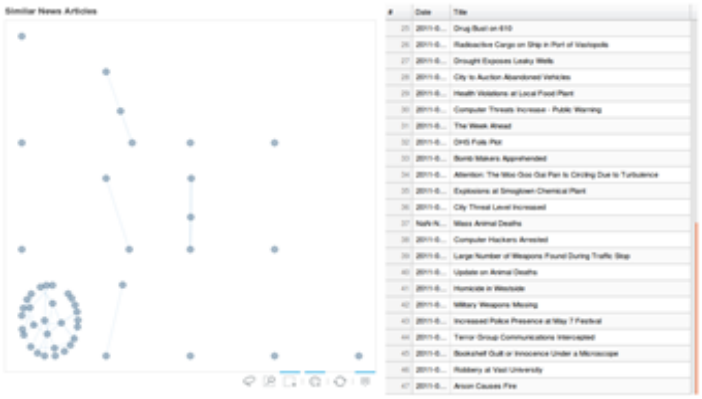

Figure 12. Selection of news reports related to Potential Threats

A total of 51 news articles (shown below) is obtained from the above analysis which is further manually inspected to identify and reason about the threats.

- The news articles are arranged chronologically.
- Some of the obvious non-relevant articles are discarded.
- Hypothesis is generated based on the remaining news reports
- The bioterror attack planned by the group ‘Paramurderers of Chaos’ stands out. Other threats include bomb threat and isolated incidents like missing military equipment, radioactive cargo, etc.

#	ID	Date	Title
41	3628	NaN-NaN-...	Mass Animal Deaths
3	315	2011-03-29	The Remaining Agriculture Subsidies
13	1078	2011-03-29	Foreclosure Auctions
30	3130	2011-03-31	City to Auction Abandoned Vehicles
7	427	2011-04-04	Selling Hunting Rights Saves Animals
15	1155	2011-04-04	Weyerhaeuser to Sell Acreage And Mills to U.S. Timberlands
16	1209	2011-04-10	Extended Training for ``Dirty Bombs'' Initiated
9	544	2011-04-11	Manufacturing Dangerous Microbes
44	3969	2011-04-14	Update on Animal Deaths
17	1450	2011-04-18	CDC Publication on Bioterrorism
42	3663	2011-04-19	Computer Hackers Arrested
14	1130	2011-04-20	Animal Deaths in City Caused by Microbes
31	3137	2011-04-20	Health Violations at Local Food Plant
25	2647	2011-04-21	Man Builds Bomb in Vastapolis
11	887	2011-04-25	Terrorism Expert Speaks Out
32	3154	2011-04-25	Computer Threats Increase - Public Warning
29	3048	2011-04-26	Drought Exposes Leaky Wells
46	4292	2011-04-26	Military Weapons Missing
50	4436	2011-04-26	Robbery at Vast University
24	2330	2011-04-27	Convicts Escape Center for the Criminally Insane
2	205	2011-05-01	Letters to the Editor Counter-terrorism: A CIA Priority
48	4427	2011-05-01	Terror Group Communications Intercepted
23	2319	2011-05-02	Overseas Terror Group Threatens Press
27	2757	2011-05-05	Drug Bust on 610
47	4318	2011-05-06	Increased Police Presence at May 7 Festival
0	179	2011-05-07	Vehicle Downtown Checked for Bomb Threat

#	ID	Date	Title
0	179	2011-05-07	Vehicle Downtown Checked for Bomb Threat
12	993	2011-05-08	Donnette Karen Hemenway DNKY In Clash of Fashion Symbols
5	332	2011-05-09	Homeland Security Talks About ``Dirty'' Bombs
18	1872	2011-05-09	Animal Activist Threatens Press
20	2039	2011-05-09	Threats to City Officials Received
40	3625	2011-05-10	City Threat Level Increased
10	571	2011-05-11	Less Logging Spurs Wildfires, Timber Trade Group Claims
28	2978	2011-05-11	Radioactive Cargo on Ship in Port of Vastopolis
33	3194	2011-05-12	The Week Ahead
36	3430	2011-05-12	Bomb Makers Apprehended
37	3575	2011-05-12	Attention: The Moo Goo Gai Pan Is Circling Due to Turbulence
45	4122	2011-05-12	Homicide in Westside
6	375	2011-05-13	Suspects Apprehended
22	2287	2011-05-13	Eight Die at Northville Airport Aviation Accident
4	321	2011-05-14	Suspicious Turkey Found with Man at Airport
19	1893	2011-05-15	Dangerous Suspect Arrested at Local Plant
21	2179	2011-05-16	Justice Approves USA Waste's Planned Acquisition of Sanifill
38	3588	2011-05-16	HEARD ON THE STREET As More Bagel Chains Go Public, Th...
43	3781	2011-05-16	Large Number of Weapons Found During Traffic Stop
51	4455	2011-05-16	Arson Causes Fire
26	2755	2011-05-17	Deadly Collision Causes Major Backup
39	3613	2011-05-17	Explosions at Smogtown Chemical Plant
49	4433	2011-05-17	Bookshelf Guilt or Innocence Under a Microscope
8	439	2011-05-18	University Professor Walks out of Class
34	3323	2011-05-18	DHS Foils Plot
1	196	2011-05-19	Flu Season Hits Hard
35	3363	2011-05-19	Pioneer Is Pressed on Discount Of Shares to Net Asset Value

Figure 13. List of new reports related to Potential Threats

We have also provided an option to save the progress made in the analysis so that the user does not have to repeat these steps. Instead she can load the previous work and pick up from where she had left.

Learnings and observations:

1. We started with topic modelling and then applied clustering to find similar documents together.
2. For clustering, we initially tried using Doc2Vec. Since the results from Doc2Vec were not good enough, we decided to explore other approaches like cosine similarity. Clustering using cosine similarity gave surprisingly good results and we decided to stick with this approach.
3. Even after getting a list of 100 topics, it was difficult to narrow down to the solution in one go. So we then decided to add the functionality of selecting the irrelevant topics and removing them. We also added the options to save the reduced list of topics and reports and then reload them later. This helped us avoid repetition of work.
4. We later learnt that the use of **LDAVis** for interactive topic model visualization is better than the scatter plot we have generated.
5. This approach is generalizable and the dashboard can be used to find news articles related to any topic like baseball, stock-markets, etc, and it is not limited to terrorist threats.

Grand Challenge

To solve the grand challenge, we gathered pieces of information for all the previous challenges. Challenge 1 gave us the information about the truck accident, epidemic symptoms and the mode of transmission of the disease. Using mini challenge 2, we figured out that the computer network of All Freight Corporation was attacked. Finally, the missing pieces of the puzzle were found using the mini-challenge 3 where we discovered various news articles which provided us the remaining information and correlation with the data from other challenges.

Challenges faced and Lessons learnt

1. Do a quick initial analysis using multiple approaches to identify most relevant path to be taken to solve the problem.
2. Spend time in understanding the input dataset and how it can be transformed to make the analysis easier.
3. Interactions are useful, especially early in the analysis. They provide faster intuition for analytical techniques to be applied.
4. Be thorough in understanding the strengths/weakness of the analytic algorithms that are applied on the data. Also, make sure to experiment and tune the parameters of the algorithm, otherwise even good approaches may give bad results.
5. Applying some of the more intensive analytic techniques require lot of time while running on a consumer grade laptop. This hampered exploration of multiple techniques to solve the challenge.
6. Some of the disadvantages of Bokeh include the loss of flexibility in creating the visualization and relative immaturity of the product. We spent a lot of time working around the tool to create some visualizations which were not supported by Bokeh.