Anatomy of a Terror Plot

IEEE VAST Challenge 2011

Investigation into Terrorist Activity

Latent Dirichlet Allocation (LDA)

- Topic modeling of given text corpus
- Generate LDA model for user-defined number of topics (in this case, 100)
- From the list generated, select the relevant topics

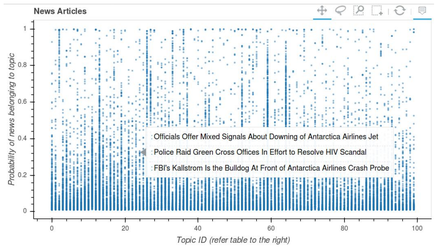


Figure 1. Distribution of News Reports across topics

Figure 2. List of topics generated by the LDA model

Clustering Similar News Reports

- Select news reports related to the topic of interest (example, potential terrorist threats)
- Cosine similarity to identify reports similar to the selected news reports and generate clusters
- Manipulate the similarity threshold to get better results (less/more connected clusters)

Figure 3. Clustering of selected news reports

- Browse through the list of news reports in a cluster to find the relationship between them
- Remove the irrelevant clusters
- Narrowed down to around 50 reports from more than 4k news reports
- Apply Named Entity Recognition to identify the important people, locations and organizations
- Identify threats which are related to each other and generate hypothesis

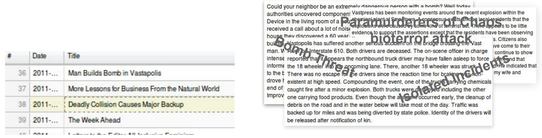


Figure 4. Creating a hypothesis using news reports

Computer Networking Operations

Network Traffic Tracking

- Parse different types of network log files and convert them into a consistent format
- Calculate frequency of entries reported in the log files per minute
- Select date and view traffic across the network at a glance
- Quickly detect anomalies
- Analyze the log entries with the corresponding timestamp to find out what is happening
- Suspicious activities like Denial of Service attack, Port Scan detected

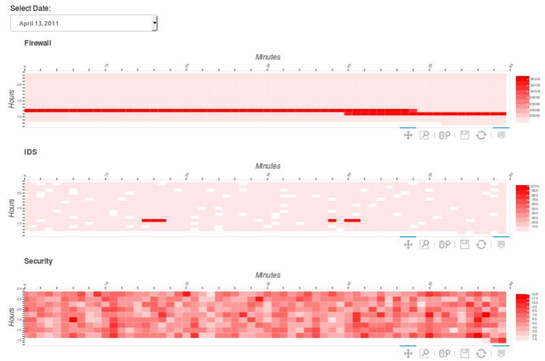
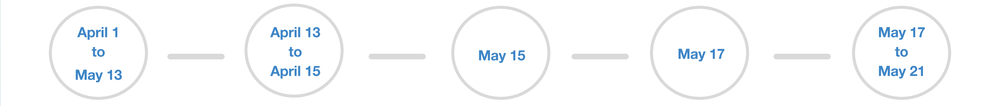


Figure 5. Situational awareness interface for corporate computer network

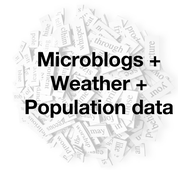
Timeline



Characterization of an Epidemic Spread

Data Preprocessing

- Remove emoticons, URLs and reserved words like 'RT'
- Remove stopwords and punctuations
- Stemming
- …..



Word2Vec

- Generate word embedding model
- Find words similar to the user-specified search terms
- Filter microblogs relevant to the search context
- Identified 55k relevant microblogs out of more than 1 million microblogs using disease symptoms as search terms
- Find out when the epidemic started
- Analyze symptoms in the affected region to find the mode of disease transmission

Figure 6. Trend of microblogs over time

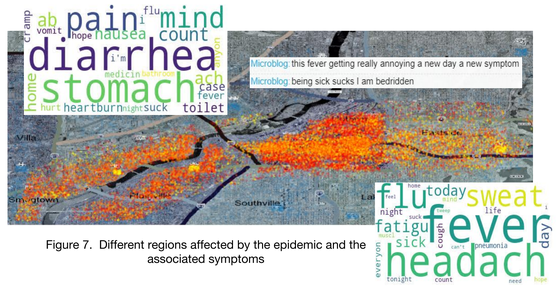


Figure 7. Different regions affected by the epidemic and the associated symptoms

Part-of-Speech Tagging & TD-IDF

- Perform PoS tagging and extract the most frequent nouns that occur in the microblogs
- Select area on the map to observe what people are talking about in that region
- Narrow down the ground zero location

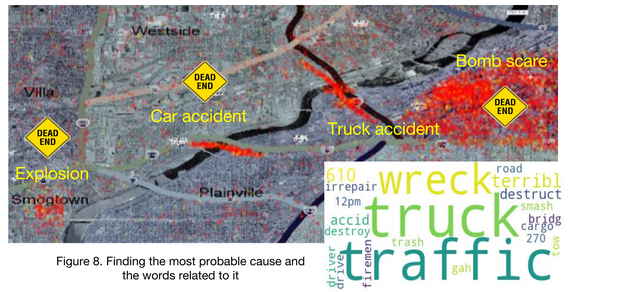


Figure 8. Finding the most probable cause and the words related to it

UMASS AMHERST

CS 690V Visual Analytics, Fall 2017

Kriti Shrivastava
Suhas Keshavamurthy