CS690V PROJECT: VAST CHALLENGE 2011

By: Divyesh Harit

1. PROBLEM DESCRIPTION:

Vast Challenge 2011 provides us with a fictional city called Vastopolis. Within this city, some sort of mysterious disease seems to be moving through the population. There's a hint of potential foul play with some terrorist groups being involved.

Mini Challenge 1: We are given ~1 million geo-coded "messages" in a CSV file. We have to try and look for signs when the epidemic started, how it spread etc.

Mini Challenge 2: We are provided with timestamped network logs and we have to create a visual dashboard to report network intrusions.

Mini Challenge 3: We are given ~4500 text articles and we have to determine if there was a potential terrorist action.

2. WHY I CHOSE THIS CHALLENGE:

I was new to dealing with text data, and wanted to try something new and challenge myself. The 2011 challenge genuinely sounded interesting. I thought it would be fun to try and find the epidemic cause, all the while learning more about text data.

This is why I spent most of my efforts on Mini Challenge 1, and a little bit on Mini Challenge 3 as I wanted to experiment with text more.

3. MINI CHALLENGE 1:

a. Checking message frequency:

The first thought I had was this: Since we are given messages of the residents of this city, when the epidemic started out, surely a lot of people starting messaging about this; panicking, complaining about symptoms, etc. So, the first thing I did was checking for the frequency of messages on each day.

Now, around 97% of the messages were in the month of May 2011, while others were in April. I discarded the April messages and focused on the May ones. I grouped the messages by date, and threw up a bar graph. A bar graph was really intuitive to me as it would easily display the highs and lows of the number of messages. So, this was the result:

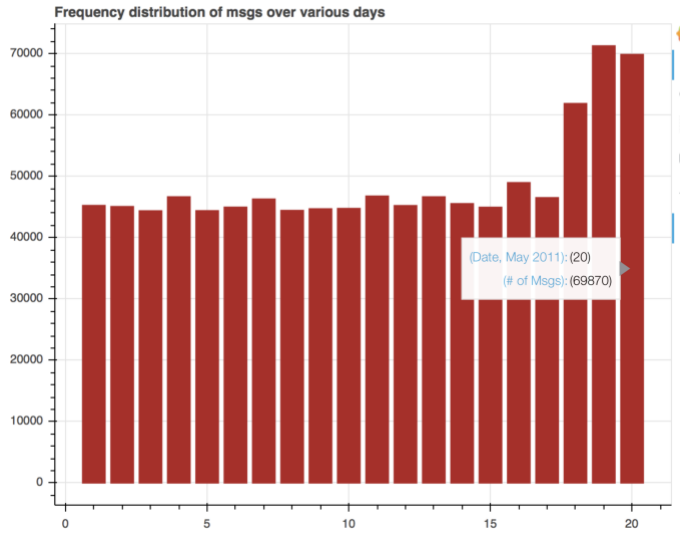


Fig. 1: Frequency distribution of messages per day

As we can see, the number of messages are pretty much constant for dates 1st-17th May, hovering around ~45000. However, it really shoots from 18th till 20th, reaching as high as ~70000. This tells us that there's a good chance that the epidemic started on the 18th. Next logical step would be to look into the messages over these last 3 days.

b. Visualizing popular words for most busy days:

So, that's what I do here. I wanted to visualize the most popular words over these 3 days, which would confirm our suspicion of timeline regarding the outbreak. My first instinct here was that a word cloud would be ideal for this purpose.

However, Bokeh does not have an inherent word cloud. I thought about doing it in Bokeh, somehow using scatter plots with the images of words instead of circle, and make their size according to their frequency. But as I was doing the project alone, it seemed an uphill task for frankly one of the trivial aspects of the project, especially with the timeline in mind.

So, I took the help of an external library called wordcloud, and used it to display the following:

Fig 2.1: May 17th

Fig 2.2: May 18th

Fig 2.3: May 19th

Fig 2.4: May 20th

Here, we see a pattern appear. Till 17th May, the most popular words were regular words used in everyday life. But on 18th, symptoms and health related words such as "sweats", "headache", "cold", "fatigue" etc. start to pop up. On 19th, we see more relevant words such as "fever", "flu", "runny nose", "chills", "stomach ache", "sick" etc. appear. On 20th, we see "flu", "chills", "pneumonia", "diarrhea", "fever" appear.

This confirms our suspicion: The epidemic started on 18th May. Now, what do we do with this piece of information? I thought, we could do two things:

One would be to cluster messages from these 3-4 days by location and superimpose it on top of the map. This, combined with the fact that there was an increase in messages, could help us identify which regions have higher concentration, and thus narrow down the epidemic spread.

The other thing we can do is to use the word cloud. We can use these symptoms, and cluster them per day by location, and then superimpose on map. This would help us identify which symptoms are spreading in which regions, thus potentially help us identify the cause/spread pattern of epidemic.

c. Cluster messages by location on map:

I try the first option here. I took the messages for 17th, 18th, 19th and 20th May. I cluster them by location, plot it, and superimpose it over the image of map of Vastopolis. The result is something like this:

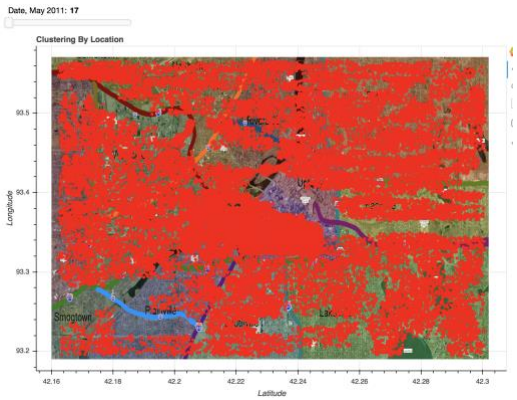


Fig 3.1: Msgs by location, 17th May

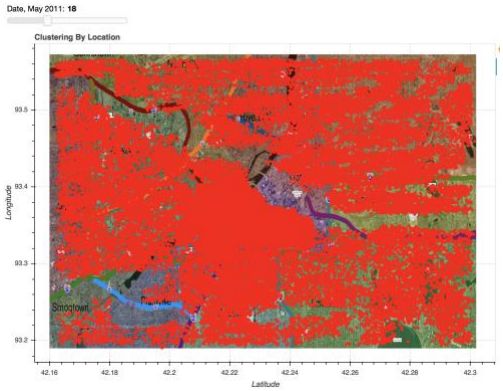


Fig 3.2: Msgs by location, 18th May

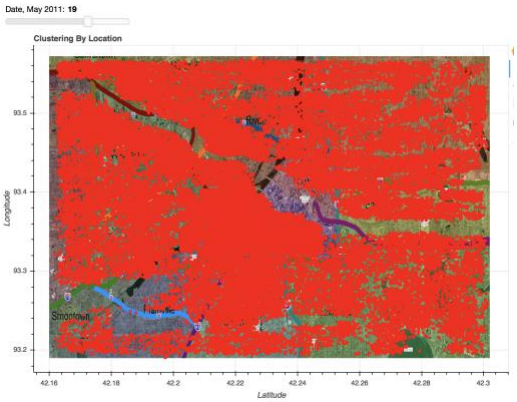


Fig 3.3: Msgs by location, 19th May

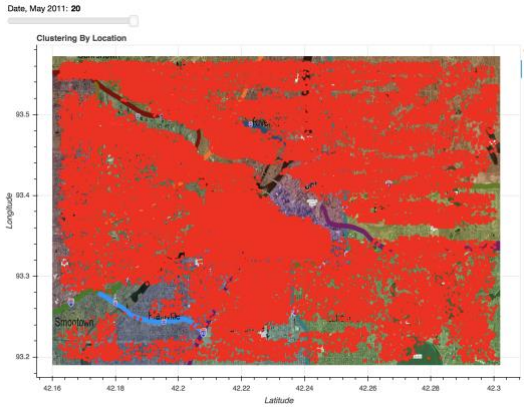


Fig 3.4: Msgs by location, 20th May

Here, as expected, we see a pattern. On 17th, the messages are scattered throughout, with only an above average concentration in the central (downtown and uptown) areas. On 18th, we see an increased concentration in these areas. This trend maintains and frequency of messages in these areas increases over 19th and 20th as well. There's a pattern alongside the river as well (this map is inverted).
This very likely points toward either start of the epidemic or its spread in these areas. A logical next step appears to be looking at the symptoms in these areas to find out which disease it is.

d. Cluster messages by symptoms on map:

So, this is what I do here. I create a dictionary of symptoms/health related words, taking cue from the wordclouds obtained earlier:
{"cold", "headache", "fever", "flu", "chills", "diarrhea", "pneumonia"}
I initially had 10+ symptoms. But as I'm plotting each symptom by different color, it became really unintuitive and hard to make sense, so I trimmed it down. I take this dictionary, and look in messages between 17th-20th May for each symptom. If a message text contains a symptom, I take that message location and plot it with a particular color per symptom, and superimpose it on the map. The result is the set of graphs in Figure 4.
Here, on 17th, we see very few people talking about any of the symptoms. On 18th, in the downtown and uptown areas, we see an increased concentration of messages talking about chills and headache, with headache and cold also spread all over the map.

On 19th, we see a very heavy concentration of chills and fever in the central regions, while 20th has pneumonia in the central regions with diarrhea along the river (current map is reversed).

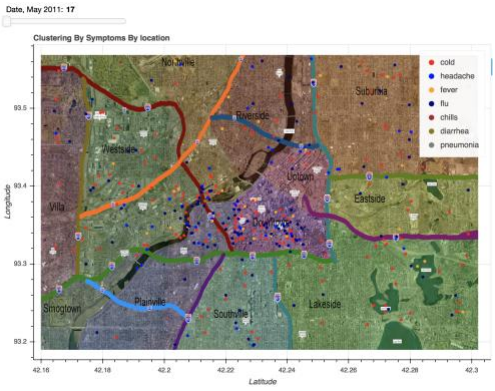


Fig 4.1: Msgs by symptoms, 17th May

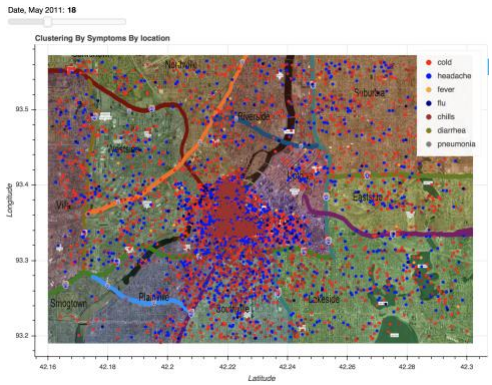


Fig 4.2: Msgs by symptoms, 18th May

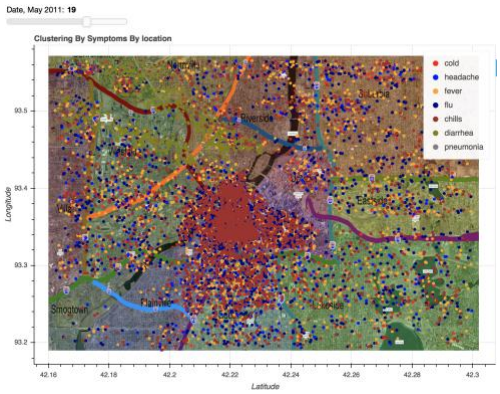


Fig 4.3: Msgs by symptoms, 19th May

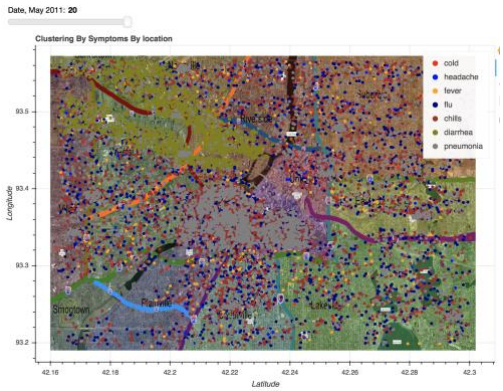


Fig 4.4: Msgs by symptoms, 20th May

e. My MC1 conclusion:

Based on the results obtained, I conclude that the epidemic started on the 18th. I believe the symptom pattern of diarrhea obtained on 20th tells us the epidemic started along the river and it's a water-based disease. I could not get far along to find out *what caused* the outbreak.

4. MINI CHALLENGE 3:

a. Looking for terror-related words

For dabbling a little bit in MC3, since we're required to look for evidence of terror activity, I thought about looking for terror-related words. I created a dictionary of terror-related words:
{"terror", "terrorist", "terrorists", "terrorism", "threat", "bomb", "scare", "explosion", "torture", "violence", "shoot", "shot", "panic", "gun", "missile", "horror", "scary", "attack"}
I then look within the text articles for these keywords. I increment the respective counter if I come across a particular word. I threw up a bar chart, and the result is Fig. 5.
Well, it's apparent from the figure that there definitely is occurrence of these words. But there's no telling so far what *context* they appear in. Just because an article talks about *some* terrorist activity, doesn't mean there's an actual terrorist activity happening in Vastopolis.
A logical next step, I thought, would be looking at most frequent topics being talked about in these articles, see if there's a terrorist-related topic, and then go from there. That didn't work for me! More about it in section 6.

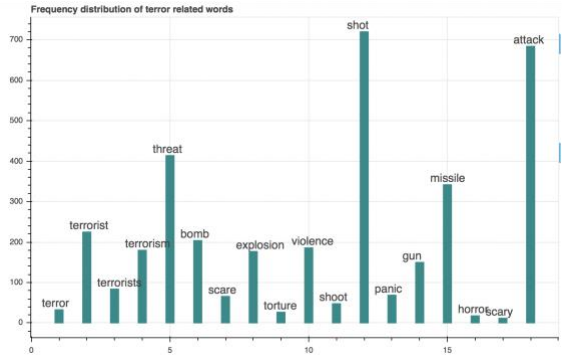


Fig 5: Terror-words frequency

5. WHAT WORKED:

I felt I succeeded quite a bit in MC1. I was able to identify the days on which the epidemic started, identify the symptoms, and identify a potential pattern in the spread and its location.

My approach of using basic graphs and interactions to learn more about the data seemed to work here. I thought this was a more intuitive approach than making elaborate graphs and not learning much about the trends.

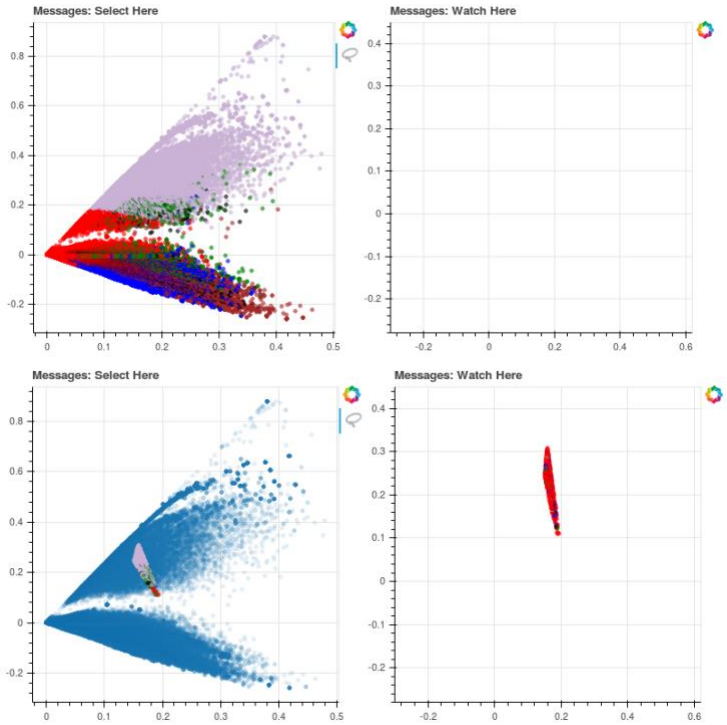
In MC3, I was able to confirm the talk of terrorist activity, although it couldn't be correlated to actual terrorist activity happening in Vastopolis.

Also, going through others' presentations and picking up points here and there helped!

6. WHAT DIDN'T WORK:

a. MC1:

I tried clustering by just the message content. I took the messages, reduced their dimensionality by PCA, and plotted on the map:



Yeah, it didn't work. I thought it would give me some insight about a pattern or structure within the messages, but that definitely did not happen.

b. MC3:

I tried LDA topic modeling on the text articles. None of the top 20 topics were terrorist-related. This was unexpected!

```
Topic 0: exchange tuesday political tax derryberry going rates late chicago policy
Topic 1: group index expected higher use early products board director states
Topic 2: world national general unit city results set current costs large
Topic 3: shares codi york net chairman funds hong economy yen francs
Topic 4: mr state make lost issues bonds communications value dow employees
Topic 5: market like trading far month fund white game losses family
Topic 6: billion investors fell industry games thursday end including line pay
Topic 7: year law lower increase volume th european offer told today
Topic 8: sales week house news work deal major high number trade
Topic 9: company analysts chief officials financial right firm department point july
Topic 10: share companies money friday international securities kong small better used
Topic 11: new time cents day months campaign income power buy strong
Topic 12: bank federal growth home profit investment technology network rate ended
Topic 13: said million quarter plans long help local hit japan press
Topic 14: government corp american way good public country want services ms
Topic 15: stock stocks big second past issue ago reported team little
Topic 16: president years rose report economic revenue wednesday union come need
Topic 17: says people business price earlier plan markets foreign monday internet
Topic 18: say party according convention left sell university vice loss job
Topic 19: prices china executive earnings recent service march close school problems
```

7. WHAT I WOULD CHANGE/WHAT WAS MISSING:

Group work! I started out in a group but the schedules didn't work and we worked individually. Twice the amount of work I did would have definitely solved at least MC1, maybe even MC3.

I didn't look at any submissions to the challenge. Looking at the solutions would have definitely helped, but I wanted to try it on my own – like I was presented the VAST challenge as if it was released this semester. So, while I would have benefited from looking at previous solutions, I don't think I would have, if given the opportunity again.

Also, regardless of working individually or in group, I genuinely believe more time to work on the problem (I realize it's unrealistic as this is a semester project) would have helped.