VAST 2019 Mini Challenge 3: Voice from the People
Group 3

1. INTRODUCTION

Data Visualization can be used to study different trends and patterns in the outlying data and give solutions for it. One such aspect is the usage of social media during any disaster. In this report, we analyse the social media data of St. Himark as presented in the VAST 2019 Mini Challenge 3.
We analyze the human impacts, particularly panic, during the unprecedented situations and how to combat them with the help of visual analytics on the social media data.

Implementing the MC3 "Voice from the people", we used the data for city St.Himark from the 04/06/2020 to 04/12/2020 to extract useful information of the damage caused by the earthquake and the distribution of resources which are portrayed using visualization techniques.

The visualizations provide answers to all the questions asked by the challenge and sentimental information of the people across various time ranges.

2. VISUALIZATION DESIGN

Our Visualization dashboard contains 5 visualizations (4+1 innovative): each one of them covers various aspects and attributes of the data using different marks, channels and interactions. All the visualizations except the emotional line chart (that is connected to the Choropleth map) are connected to the area chart and change over the selection of time interval.

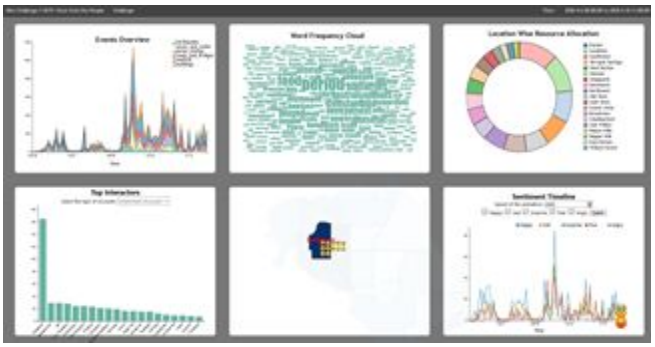

Figure 1: Data Visualization Dashboard. (Left to right on the top): 1. Events overview, 2. Word frequency graph, 3. Location wise resource allocation. (Left to right on the bottom) 4. Top interactors, 5. St. Himark map, 6. Sentiment timeline

1. Area chart (Events overview): This chart represents time on x axis and number of messages. Each stacked area illustrates an event or a resource changing over time. Hovering on the legend highlights the category. This also supports the selection of time range, which is also used to affect other visualizations. Marks for this chart are line, area and channels used are x-position, y-position and color hue.

2. Word Cloud(Word frequency cloud): This chart represents the significance of each word in the given time period obtained from the area chart. Marks for the word cloud are words and the channel is the size of the word.

3. Pie-in-Donut chart (Location wise resource allocation): This chart contains two charts donut chart and a pie chart which appears after the interaction on the donut chart. Donut chart represents the number of messages related to resources for each location. Mark in the donut chart is the arc of donut and channels in the donut chart are color and angle of arc. when an arc is selected in the donut chart pie chart appears showing distribution of messages related to different resources in the selected region. Mark for the pie chart is arc and channels are color and angle of arc.

4. Bar Chart (Top Interactors): This chart focuses on the number of messages sent by the important accounts or official accounts in the given time period. Mark of bar chart is bar and channels are x position and length of bar.

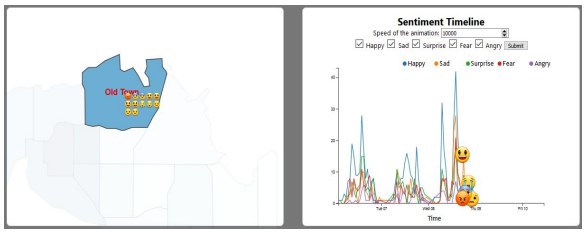

Figure 2: Innovative Design

5. **Innovative Design (St. Himark and Sentiment Timeline)**
Choropleth map: This map shows the number of messages sent from the region and emotional state of the region. Marks for this map are area and emojis and channels are color luminance and number of emojis. When a region is clicked on the map it zooms out showing the emotional state of the region using emojis and produces an emotional line chart.
Progression emotional chart: It illustrates the variation of emotional state of the people in given location obtained from map and given time period obtained from area chart. Mark for this chart is line and channels are x-position, y-position and emoji.

3. DESCRIPTION OF VAST MINI CHALLENGE

"Seismic and survey data are useful for capturing the objective damage that the earthquake has caused St. Himark. However, this data has limitations. First, official surveys are time consuming and

do not stay current in a rapidly changing situation. Second, they don't establish how citizens are reacting to the current crisis. Third, they are often insufficiently granular, providing little insight into differences between neighborhoods. In other words, the seismic and survey data do not provide an up-to-date view of the structural and humanitarian impact caused by the earthquake on a neighborhood-by-neighborhood basis. The City has concluded that this knowledge is necessary to determine where to allocate emergency resources.

City Officials have identified a subset of Y*INT, a community-based social media platform, as a potential source for revealing the current state of St. Himark's neighborhoods and people. Knowing that you are skilled in visual analytics, the City has asked you to analyze Y*INT messages in order to determine the appropriate actions it must take in order to assist the community in this disaster."

4. DATASET DESCRIPTION

The mini challenge has provided one csv file for the dataset which contains several fields dating from 04/06/2020 to 04/12/2020. The following fields have been included:

1. Time: date and time when the message was posted
2. Location: The St. Himark neighborhood from where the message was posted
3. Account : user handle of the person who posted the message
4. Message: text from the message

Apart from the csv, the challenge provided details regarding the important resources of the city. The utilities of the city included water and sewer, roadways, streets, gas, garbage, power, and a nuclear plant.
The neighbourhood cities include Palace hill, Safetown, Old Town, Northwest, etc. Similarly, there are 8 hospitals in the city along with emergency services like Radiation monitoring.

4.1 Data Abstraction

This level of abstraction is responsible for defining what type of data is shown and how it can be used to perform the given tasks. For a given nested block, the data abstraction level must be defined using several dataset types and attribute types. These types will generate domain specific problems into generic visualization problems.

For the given MC3 Dataset,

1. Cardinality: Number of Items- 41942.
2. Attributes:
   - Time : Ordinal/cyclic
   - Location : Categorical
   - Account : Categorical
   - Message : Categorical
3. Dataset types: relational table. The St. Himark dataset has 4 attributes and 41942 data items
4. Transformation of data: derived attributes like categorizing the accounts into normal user accounts, official accounts and advertising accounts will be used to importance of a message and use in task abstraction and implementation of idioms. In several cases, the data would be then modified to extract features and perform the required tasks.

4.2 Task Abstraction

This level combined with the data abstraction is used to map the data according to the task into a more real-world terminology. This describes us what tasks the user wants to perform and what kind of resources are available to the user domain. Some of {action,target} pairs are mentioned below:

- Density of messages and resource allocation:
  {discover→lookup→summarize,distribution}
- Usage of utilities and maintenance for different disaster-affected areas.
  {present→lookup→summarize,trends/distribution }
- Sentiment analysis of the people:
  {derive→explore→compare,trends/features }
- Compare different towns in terms of safety:
  {annotate→lookup→summarize,shape}

4.3 Pre-Processing of data:

The first step in pre-processing of the data was to clean the data, the Y*INT app data has a lot of advertising and meaningless conversations between accounts. We created a stop words list of our own and used it along with the nltk stop words which are to be removed from the messages. Then extracted name entities and noun phrases identified by using nlp module of spacy library. We filtered the top 600 of these words and manually coded them to categories such as "Important", "Unimportant" and "May be Required". We removed around 3000 messages which had no important category word in it.

We identified a few important user accounts, official accounts and advertising accounts. We then removed the messages from advertising accounts and categorized the official accounts and advertising accounts into csv.

The emotional state of people went through many changes because of the earthquake to analyze that, we have used text2emotion [3] library which will provide probabilities of a message for each emotion such as happy, sad, angry, surprise and fear so that we can categorize the message into one of them.

5. TASK SOLUTIONS

The City has been using Y*INT to communicate with its citizens, even post-earthquake. However, City officials need additional information to determine the best way to allocate emergency resources across all neighborhoods of St. Himark. Our task, using your visual analytics on the community Y*INT data is to determine the types of problems that are occurring across the St. Himark. Then, advise the City on how to prioritize the distribution of resources. Keep in mind that not all sources on Y*INT are reliable, and that priorities may change over time as the state of neighborhoods also changes.

5.1 Task 1

"Using visual analytics, characterize conditions across the city and recommend how resources should be allocated at 5 hours and 30 hours after the earthquake. Include evidence from the data to support these recommendations. Consider how to

allocate resources such as road crews, sewer repair crews, power, and rescue teams. Limit your response to 1000 words and 12 images."

From the "Events Overview" visualization we can observe that there are three peak points in the graph for the earthquake event:

- Earthquake 1: April 6th at around 1:00 pm - 2:00 pm.
- Earthquake 2: April 8th at around 7:00 am - 8:00 am
- Earthquake 3: April 9th at around 2:00 pm - 3:00 pm

To further support these claims we can see that the words: "earthquake", "wobbling","quaking", "shaking", "safety" and "quivering" can be found to be used most frequently during those times in the word frequency cloud.

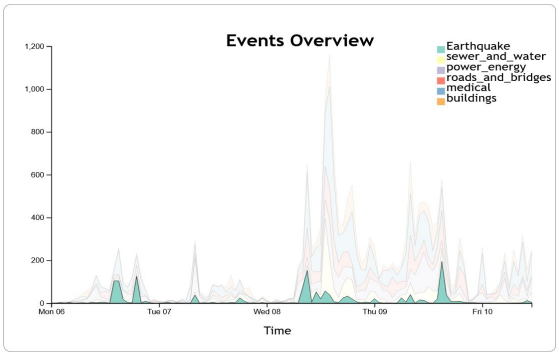

Figure 3: Earthquake peaks

5.1.1 Earthquake 1

5 hours after Earthquake 1 the main problem was with food as we can see that the "word frequency cloud (*Figure 4*) mentions words like "served", "blandest" and "bite" which tell us that something was wrong with the food and the words "buildings" and "nuclear" hint towards something going wrong with the nuclear plant, supported by seeing that the only interactions during that time are from the official account of AlwaysSafePowerComapany..We can also notice that the most resources that need to be allocated are to Weston, Scenic Vista and Downtown requiring a majority of medical resources whereas Downtown also reports bridge issues as seen in *Figure 4.*

Similarly, we deduce that after 30 hours there are still resources needed for medical and building damages as we find that using the word frequency cloud and the resources viz.

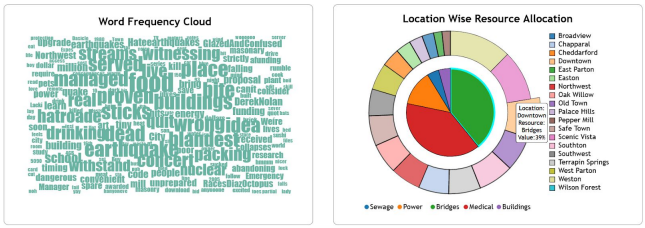

Figure 4: Earthquake 1 stats (a) Word frequency cloud (b) Location wise Recurrence Allocation

5.1.2 Earthquake 2

5 hours after Earthquake 2 the main problem was with "power", "water" and "fatalities" as these can be seen in the "word frequency graph" and supported by other words like "hospital", "collapsed", "nuclear", "damaged", "collapsed" and "rescue". With these we can understand that the water pipes might have been damaged or collapsed due to the earthquake, the nuclear power plant has had a problem or an incident that involved fatalities which are being rescued & rushed to the hospital. The nuclear incident can be backed up by the official account having maximum interactions during that time and by the fact that Safe Town represents the majority of people in the state of surprise, fear and sadness. The fatalities can be confirmed as the resource allocation viz (*Figure 5*) shows that almost all of the areas report need of medical resources and resources for buildings or bridges.

Similarly, after 30 hours we see that the people are evacuating and there are fatalities along with problems with the city bridge which has collapsed. We can analyze this using the word frequency cloud and the emotional levels of the people situated in the locations where bridges are there along with DOT being the top interactor during that time.

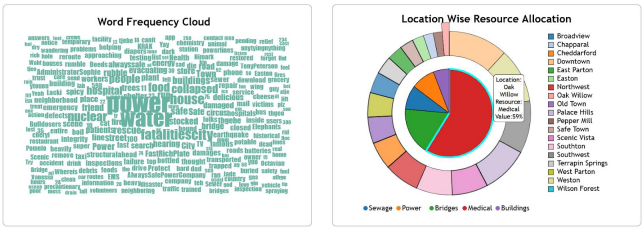

Figure 5. Earthquake 2 stats (a) Word frequency cloud (b) Location wise Recurrence Allocation

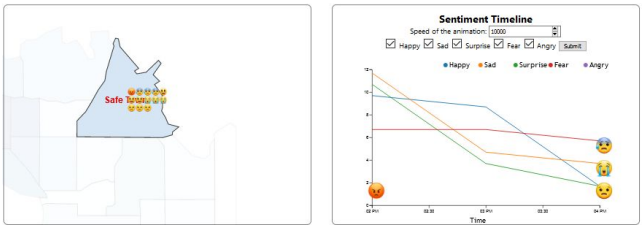

Figure 6: Earthquake 2 stats (a) St. Himark area (Safe Town) (b) Sentiment Timeline

5.1.3 Earthquake 3

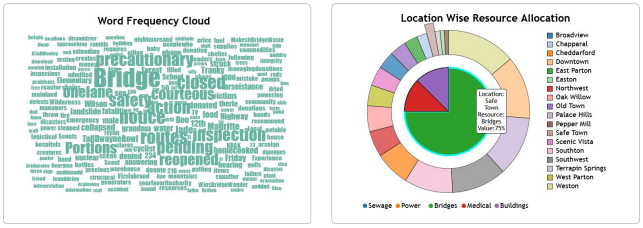

Figure 7: Earthquake 3 stats (a) Word frequency cloud (b) Location wise Recurrence Allocation

5 hours after Earthquake 3 the main problem was with "bridges" as it can be seen in the "word frequency graph" by other supportive words like ""safety", "closed", "inspection", "reopened", "onelane" and "routes". We can assume here with the occurrence of these words that the bridges and routes might be closed and reopened for inspection after the earthquake to see the consequences of them. We can confirm this by looking at the resource allocation graph where all of the locations are mainly talking about bridge resource allocation and also the sentimental

timeline for Old Town, Safe Town, Palace Hills and Pepper Mill where the sentiment of fear and sadness occupies a majority.

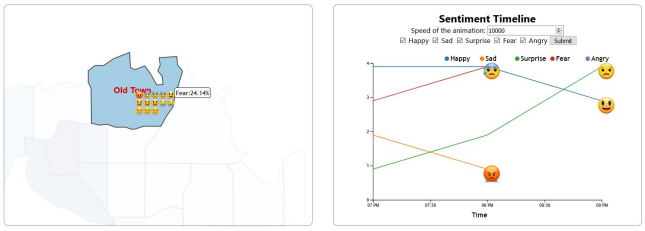

Figure 6: Earthquake 3 stats (a) St. Himark area (Safe Town) (b) Sentiment Timeline

Similarly, after 30 hours we can see that there is still talk about the nuclear plant and food as we can see that the most used word is food from all the visualization.

5.2 Task 2

"Identify at least 3 times when conditions change in a way that warrants a re-allocation of city resources. What were the conditions before and after the inflection point? What locations were affected? Which resources are involved? Limit your response to 1000 words and 10 images."

For the task 2 solution, the visualizations from the task 1 can be extended given that task 2 requires us to compare the conditions before and after the inflection point. Following were the observations:

5.2.1 Occurrence 1

The onset of Earthquake 1 is the first time when we see massive change in the resources; before and after the quake. The evidence for the same can be observed in Fig. 8 below. During 12:00 pm - 4:00pm, the Events overview chart shows that the mention of "Buildings" and "Medical" had fluctuated. There was a surge in medical emergencies around 1:30pm specially in the area of Downtown and Weston. This theory can be proved by the Location wise Resource Allocation chart as the selection of Downtown gives a breakup of the resources wherein "Medical" and "Bridges" have been mentioned most before the Earthquake 1 and tones down a little bit 2 hours later. This shows that the residents were looking for the bridge path to cross cities or report any damages to the town bridge.

The Word Frequency cloud tells us the story that "Water" was mentioned the most that depicts scarcity of water or wreckage of sewer pipes. However, there's a decline in need for all these resources after a few hours.

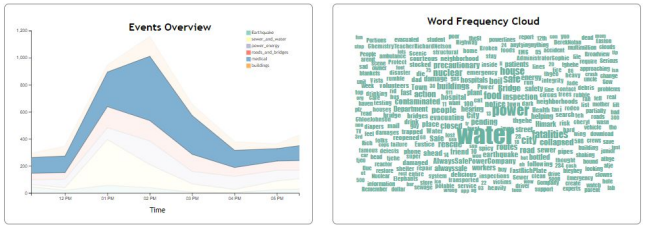

Figure 8: Occurrence 1 stats (a) Event Overview (b) Word frequency graph

5.2.2 Occurrence 2

April 9th at around 7:00am - 9:00am marked the second occurrence when there was a steep fluctuation between the resources as per the residents. If observed in the Events overview, there is a high demand of "Medical" services as well as "Power" systems. The Word Cloud has the highest frequency words as "Collapsed","Power","repair",etc. This marks the reconstruction of the buildings and restoring of the power supply after Earthquake 2 as the AllSafePowerCompany seems to be an active account as well.

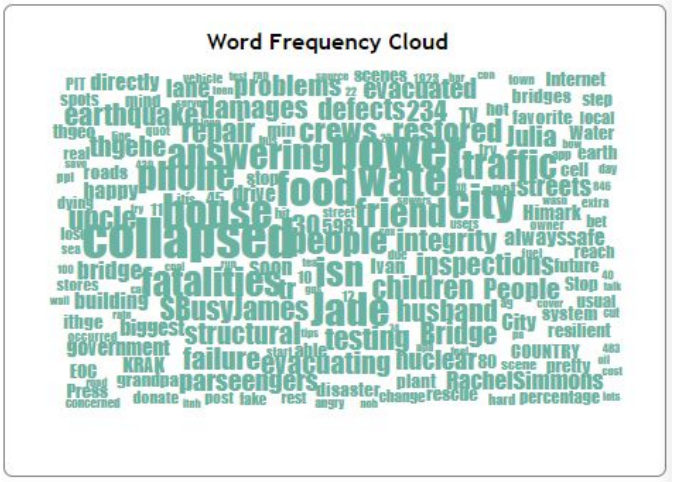

Figure 9: Occurrence 2 word frequency cloud

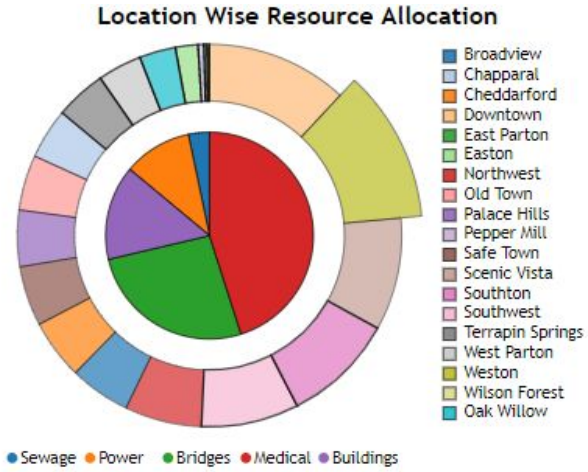

Figure 10: Occurrence 2 location wise resource allocation

Again, Downtown and Weston require the most resource allocation during this period while Wilson Forest seems to be the least affected. Using this information, resources like Transportation can be utilized to carry around citizens for evacuation or shelter purposes from the damaged areas to the safer areas. The Sentiment timeline proves that the residents of Downtown were initially sad and angry because of the scenario. Meanwhile, around 10am-11 am, there was a decline in the resources but emotion of fear increased around as the city prepared for Earthquake 3.

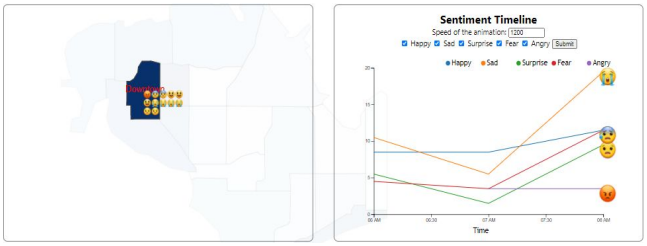

Figure 11: Occurrence 2 (a) Downtown (b) Sentiment Timeline

5.2.3 Occurrence 3

This occurrence marks the onset of Earthquake 3: April 9th at around 2:00 pm - 3:00 pm. There was a lot of commotion between the timings of 2pm and 5pm, i.e. before and after the earthquake. Around 3pm (earthquake onset), the frequencies of the following words increased:"quaking","shaking","closed","precautionary", "routes","bridges","nuclear","transportation". This inferred that the Earthquake trembling began around 2:30pm. The messages portray a problem with the power plant and the residents needed transportation. During the same time, the Top Interactors chart had "DOT-StHimark" as the most active official accounts with 6 conversations as the residents were in need of residents to reach a safer place.

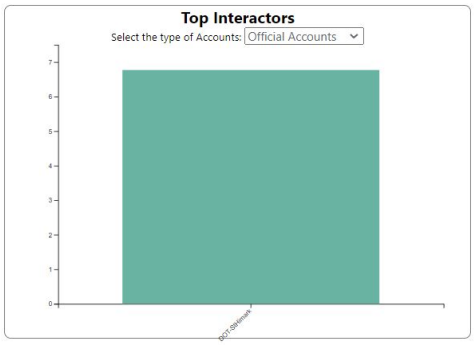

Figure 12: Department of Transportation was the most active account with 6 messages

The cities of Weston and Downstown were again most affected in terms of the resources as the Location Wise resource allocation chart visualizes. Weston had 32% of the messages related to medical, which implies a dire need of medical support in the town. The sentiment timeline also shows how the mood of the town residents changed before and during the earthquake (Major mood: Fear and Surprised).

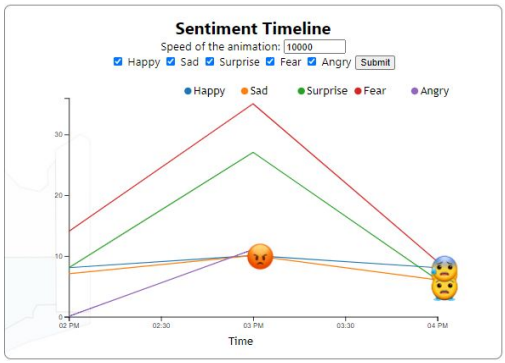

Figure 13: Sentiment analysis of Downtown (Fear: 35%,Surprise:25%)

Around 5pm, when the shaking had subsided a lot, the majority mood of the residents lightened and changed to Happy. This trend could be seen in all the towns. Similarly towns like Safetown and Chaparral saw an increase in the use of resources like "bridges" as the Location-wise chart and Word Cloud shows. This could mean the transportation of the residents from severely affected towns towards these safer towns by the Department of Transportation for St. Himark.

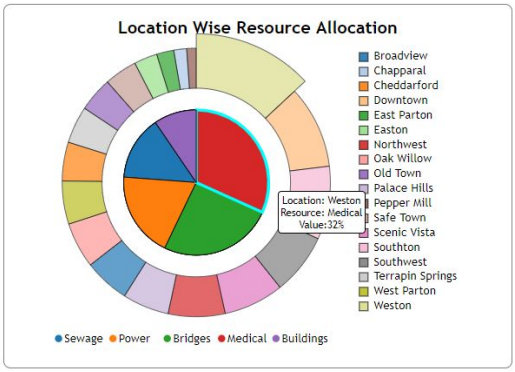

Figure 14: Location wise resource allocation

5.3 Task 3

Take the pulse of the community. How has the earthquake affected life in St. Himark? What is the community experiencing outside the realm of the first two questions? Show decision makers summary information and relevant/characteristic examples. Limit your response to 800 words and 8 images

The city was devastated because of the earthquake. But the earthquake wasn't the only disaster that affected the people as it was followed by fire eruptions and power outage. Before the earthquake, the residents enjoyed a happy and healthy lifestyle as the word cloud below depicts for the date of 7th april before the quake started. The terms like "wonderful" "songs" "geeks" portray that they were involved in their own tasks and were unknown of the situation. Similarly, towns had normal sewer problems and nothing major.

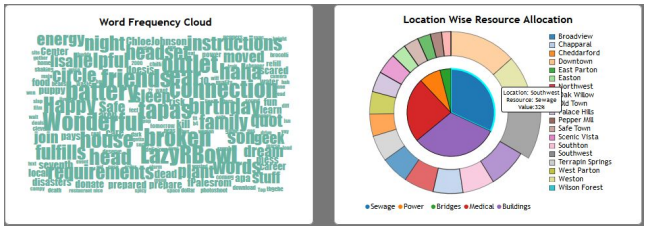

Figure 15: *Task 3 stats (a) Word frequency cloud (b) Location wise Recurrence Allocation*

After the onset of the earthquake, there was proof of fire and power outage which was something that the residents did not deal with regularly. The transportation company DOT was more involved in the process as the residents were evacuating to better places like Terrapin springs.

Post earthquake, the rumble app had a lot of interactions amongst the residents which showed the jolly mood of the residents returning as the town started healing. the mention of the terms such as "food", "reduced", "repair" infers that the cafes and

restaurants were back to normal business and the reconstruction of the buildings was in process.

Besides these, over the happy and jolly lifestyle that the people of St. Himark enjoyed before (*Figure 16)*, this was completely contrasted by the sentiments that the people had to go through during and even after the disaster where people were sad, had to go through trauma, anguish, were scared and even had fatalities during the disaster. All of this combined, we can see how the emotional state of the people drastically changed and affected them in more ways than one. But by the end of it, everything again went back to normal and people were trying to be happier, as we can see that the happiness of the people in the sentiment timeline (*Figure 16*) increases when all of this is over and people start living a healthy, happy life again.

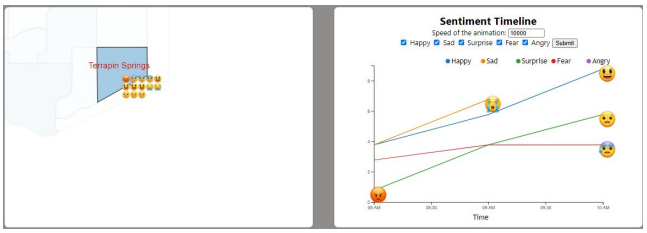

Figure 16:After disaster sentiments of the people of Terrapin Springs (a) Downtown (b) Sentiment Timeline

5.4 Task 4

The data for this challenge can be analyzed either as a static collection or as a dynamic stream of data, as it would occur in a real emergency. Describe how you analyzed the data - as a static collection or a stream. How do you think this choice affected your analysis? Limit your response to 200 words and 3 images.

The data for this challenge is analyzed statically. Though we did incorporate the stream aspect of the data into our computations and visualizations as they are in form of a stream where the data is considered to be taken under different time frames/windows as per the interactions of the user. When a particular section of the data is selected in the "Events Overview" vis, all of the visualizations change according to that time frame/window and the acquired data. Even on clicking on different locations of the St. Himark map, we can see different results for the Sentiment Timeline vis hence considering the stream of data in visualizations but the overall analysis of the data still remains to be static.

It was because of this choice that we see interactive visualizations which change with respect to a window or any other present interactions available to them as we consider a stream of data but get all the details of the events preemptively because of the static analysis of the data that we do initially.

6. LESSONS LEARNED

The following are the lessons we have learned while completing the tasks of this mini challenge.

- We understood how visual analytics can help us analyze accurately and help in making decisions at uncertain times. Here, in this challenge, our visualization helps to analyze how resources were needed at different locations after an earthquake has occurred.
- We have learned to develop functionalities in a D3 visualization, which are easy and intuitive to use while providing the necessary features that help satisfying the requirements [1].
- Visualization is mainly used to clarify the relations and reveal the hidden patterns in the data rather than focusing on representing data as a real world entity. While building a visualization we learnt how to keep them more comprehensible and simple instead of embellishing it and overwhelming the user.
- Though there's not much light thrown on data cleaning it is the most crucial step to be followed towards data visualization. Cleaning the data made a significant difference in extracting the patterns, hence we understand how important it is.

7. FUTURE ENHANCEMENTS

Our system deals with the interaction among data in a given dataset and how these interactions can be used to infer a set of tasks to solve them and provide an enhanced solution. The similar pattern can be used to perform disaster management techniques for any given geographical region in the world. Usage of geojson maps help us to analyse any part of the world.

Similarly, the given system worked on a small dataset of messages for a particular. Given the current form of static dataset, we can inculcate the same system for dynamic data stream and even for real time interactions around the globe. With more knowledge about the libraries, more complex visualizations can be delivered as well to further enhance the efficacy of the model.

8. TEAM MEMBERS

Name	Student ID
Akanksha Magod	1217135543
Akshay Kumar Gunda	1217179379
Anish Oswal	1217914776
Bhavya Talluri	1217178937
Pranay Singhal	1217208395
Sai Srujan Jaligama	1217179392

9. REFERENCES

[1] M. Bostock, V. Ogievetsky and J. Heer, "D³ Data-Driven Documents," in IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2301-2309, Dec. 2011, doi: 10.1109/TVCG.2011.185.

[2] J. Heer and M. Bostock, "Declarative Language Design for Interactive Visualization," in IEEE Transactions on Visualization and Computer Graphics, vol. 16, no. 6, pp. 1149-1156, Nov.-Dec. 2010, doi: 10.1109/TVCG.2010.144.

[3] Text2Emotion, is a python package which will help to extract emotions from the content. (https://pypi.org/project/text2emotion/)