Entry Name: **"LCEE-MC3"**
VAST Challenge 2019
<u>**Mini-Challenge 3**</u>

Team Members:

Ana Larissa Dias, Federal University of Pará, `larissa.engcomp@gmail.com`
Cassio Batista, Federal University of Pará, `cassio.batista.13@gmail.com`
Edwin Rueda, Federal University of Pará, `ejrueda95g@gmail.com`
Erick Campos, Federal University of Pará, `erick.c.modesto@gmail.com`

Student Team: YES

Tools Used:

- Python (v3.5.3)

- Bokeh (v1.2.0)

- NLTK (v3.4.1)

- TextBlob (v0.15.3)

- NumPy (v1.16.0)

- pandas (v0.24.2)

- Matplotlib (v3.0.2)

- word_cloud (v1.5.0)

- scikit-learn (v0.20.2)

- Node.js (v10.16.0)

- LibreOffice Calc (v5.2.7.2)

- grep (v2.27)

Approximately how many hours were spent working on this submission in total? 168 hours

May we post your submission in the Visual Analytics Benchmark Repository after VAST Challenge 2019 is complete? YES

Video: `https://youtu.be/vNlLfvNMXN8`

Questions

The City has been using Y*INT to communicate with its citizens, even post-earthquake. However, City officials needs additional information to determine the best way to allocate emergency resources across all neighborhoods of St. Himark. Your 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.

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. 1000 words, 12 images.

In order to discern between reliable and unreliable messages, we have considered a bar chart of frequency of users that tweeted the most with a hover tool to show the most frequent words spoken by each user. Figure 1.1 shows the unfiltered distribution, in which the red bars represent users that are probably sellers since their messages include too many words like "*deal*", "*offer*", "*opportunity*", "*chances*", etc. All tweets from those users have therefore been discarded.
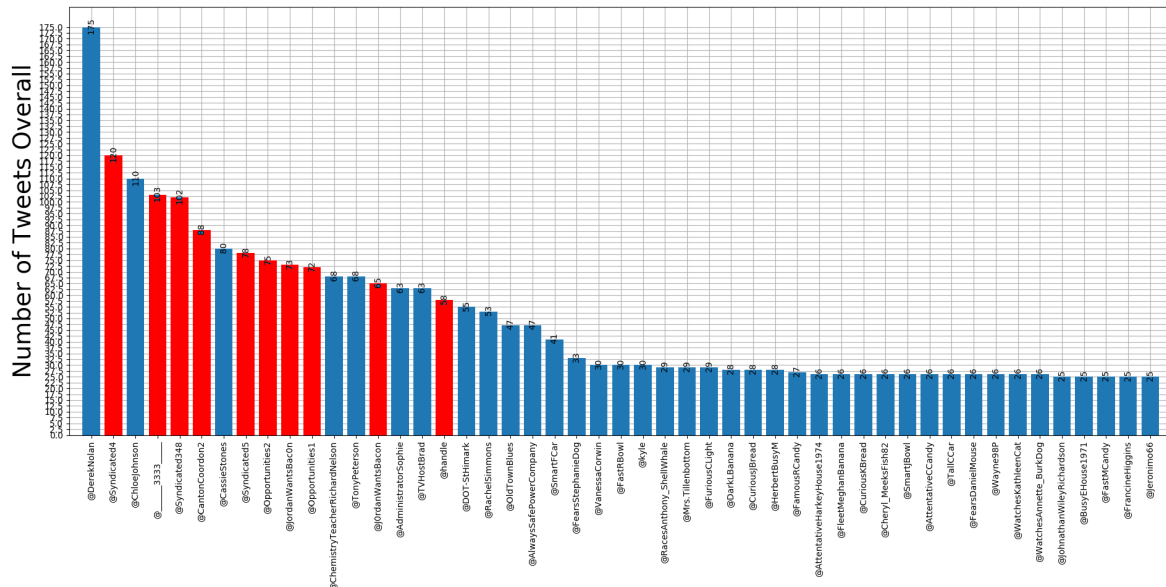

Figure 1.1: Most frequent tweeters (usernames). Red bars represent sellers.

We have then counted the number of tweets per a small time interval and presented in a histogram-like plot of Figure 1.2 that shows four peaks, where things possibly might have changed across the city.


Figure 1.2: Overall frequency of tweets over time considering all locations.

By analysing the heatmap in Figure 1.3a, there appears to have occurred three earthquakes, but the first one seems to have started at 2 PM of the first day (April 6th). The heatmap is divided by a time interval of one hour and by neighbourhood location. It was generated by considering a list of keywords that were tweeted and are similar in meaning (synonyms) and are directly

2

related to the word "earthquake". The list is: "*shake*", "*shudder*", "*vibrate*", "*wobble*", "*tremor*", "*tremble*", "*quaver*", "*quiver*", "*hazard*", "*disaster*", "*destruction*", and "*rubble*".
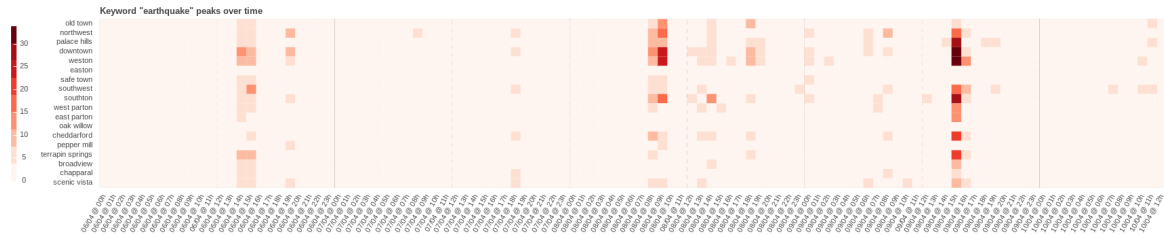
(a) Heatmap considering cluster of similar keywords (synonyms) for the word "earthquake".
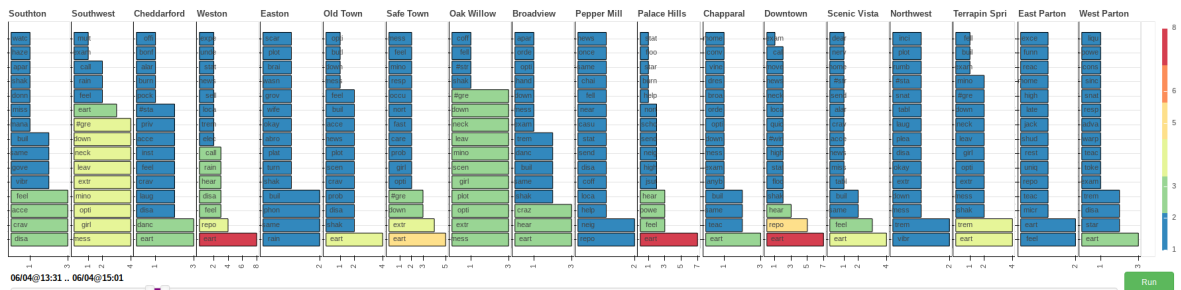
(b) Bar chart with blue-to-red colormap considering frequency of words from 1:00 PM to 3:00 PM of April 6th.

Figure 1.3: Conditions at the start of the earthquake.

To ensure the heatmap is providing a reliable information, a horizontal bar chart, shown in Figure 1.3b, was used from 1:30 PM to 3:00 PM. It counts isolated words and discards words such as adverbs, pronouns, adjectives, articles and some nouns and verbs that were considered to be useless such as "anyone", "make", "know", "food", "hate", etc. Then the Porter Stemmer from the `nltk` package was used to clip the words by its invariant parts (word root), and that root was further reduced to 4-chars only. A heat-like colormap from blue to red was also included to enhance frequency distinction.

The bar chart shows some interesting other words such as "*feel*", "*hear*", "*report*" apart from the keywords aforementioned, in which "*earthquake*" is the most frequent one in accordance with the red color.

Figure 1.4 shows a dynamically-colored SVG map of the city, where the neighbourhood inner colors range from blue to red in a heatmap fashion according to the average mean of the colors of the five bigger bars of the horizontal bar chart of Figure 1.3b for each location. It appears most reports come from the northwest and southeast of the city.
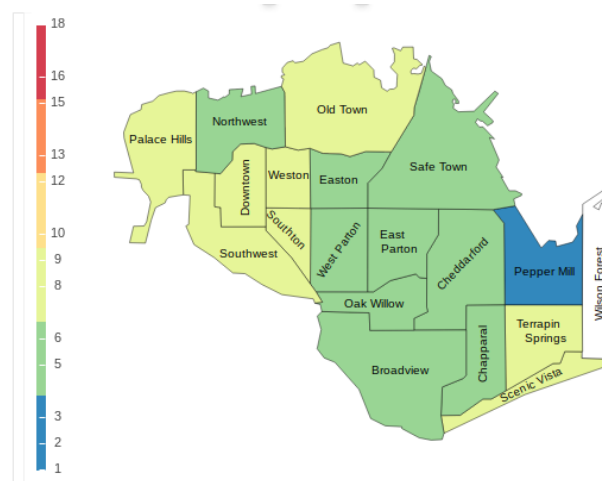

Figure 1.4: St. Himark's map 5h after the first earthquake. Lighter shades of green represent a higher frequency of messages.

3

Figure 1.5 shows multiple heatmaps, one per keyword, in a five-hour time interval from 2:00 PM to 6:59 PM. The charts at the top show three blank graphs for the keywords "building" (Figure 1.5a), "medical" (Figure 1.5b), and "road" (Figure 1.5c), which means these resources do not appear to be requested by any neighbourhood. On the other hand, the 3 graphs at the bottom show the number of mentions for keywords related to "sewer and water" (Figure 1.5d), "power" (Figure 1.5e), and "rain" (Figure 1.5f).

(a) Building   (b) Medical   (c) Roads and Bridges

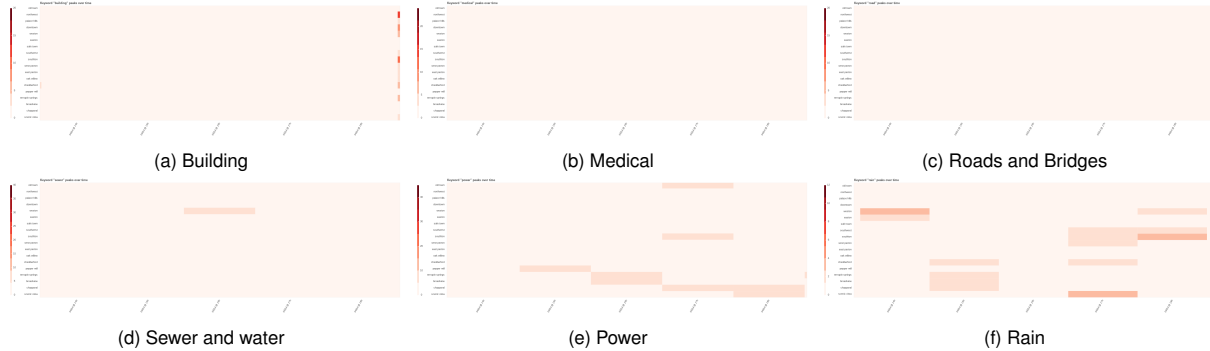(d) Sewer and water   (e) Power   (f) Rain

Figure 1.5: Conditions after 5h of the first earthquake

Suggestions for crew allocation is detailed as follows:

- *Sewer and water:* A crew must be sent only to Weston between 4:00 PM and 4:59 PM.

- *Power*: Issues have occurred in Pepper Mill, **Terrapin Springs**, **Broadview**, Chapparal, **Southton**, **Old Town** and Scenic Vista, but we'll consider only the locations highlighted in bold because they have hospitals.

    - A crew must be sent to Terrapin Springs between 3:00 PM and 3:59 PM. Broadview also has a power demand at this time interval but the tweet frequency is much lower considering the five-hour period.
    - Two crews must be sent to Old Town and Southton between 4:00 PM and 4:59 PM.
    - Lastly, the crew from Terrapin Springs can be reallocated to Chapparal between 5:00 PM and 5:59 PM. Although Chapparal does not have hospitals, it has been nearly two hours with electrical issues.

- *Rescue, sewer and water*: A crew must be sent to Southton only because there have been small issues in Weston, Southton and Downtown, and therefore Southton is geographically in the middle of such neighbourhoods.

Looking at the useful-words colormap over the SVG map of St. Hirmak, shown in Figure 1.6, it can be inferred right at the outset that the neighbourhoods that are in most need are Downtown, Southton, Old Town and Weston.
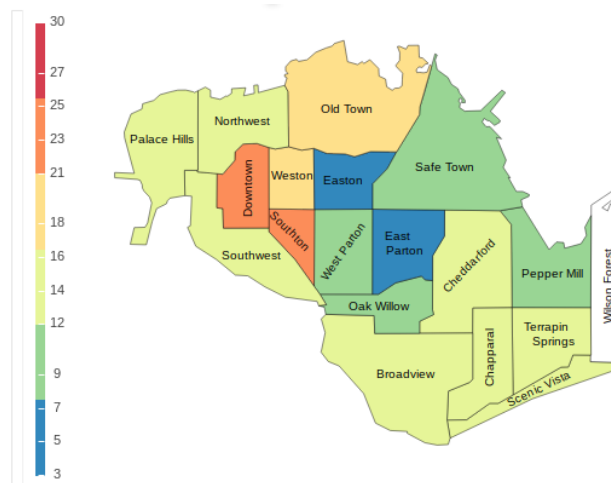

Figure 1.6: St. Himark's map 30h after the first earthquake. Yellow and orange shades represent a higher frequency of messages.

4

(a) Medical   (b) Sewer and water   (c) Rain   (d) Roads and Bridges
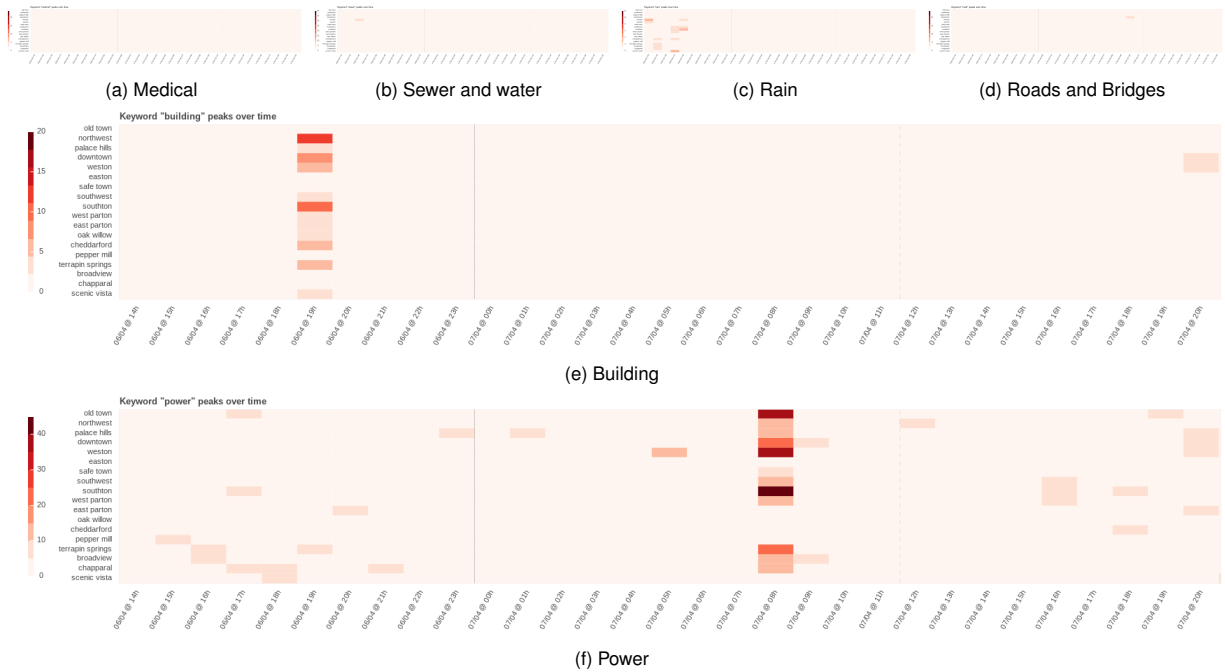

(e) Building

(f) Power

Figure 1.7: Conditions after 30h of the first earthquake

By looking at the heatmap of Figure 1.7 for each keyword, it can be seen that there have been no occurrences for medical, and the ones related to sewer/water and rain have already been attended within the first five hours. With respect to roads and bridges in particular, there have been 3 occurrences from 10:00 AM to 11:00 AM of April 8th at Downtown, but this neighbourhood is under resurfacing maintenance, which implies a road crew is already working there.

Suggestions for crew allocation is detailed as follows:

- *Building*: On April 7th from 7:00 PM to 7:59 PM there have been multiple casualties on almost all locations so we would prioritize the dark-red-colored ones according to the heatmap of Figure 1.7e: Northwest, Southton, Downtown, and Weston. All four have a high density of buildings and people, apart from being geographically close to each other, which can be an advantage for an eventual reallocation of crews in the following hours. Terrapin Springs and Cheddarford also have some less-intense occurrences, but they must be ignored due to the absence of high buildings.

- *Power*: According to Figure 1.7f on April 7th there have been sporadic, less-intense occurrences that could be solved by sending small units to individual locations, but from 8:00 AM to 9:00 AM an energy disaster appear to have affected almost all neighbourhoods. Again we would prioritize regions where the keywords were mentioned the most: Southton, Old Town, and Weston. The other can be later attended in the following hours.

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.

As already mentioned, there have been 3 independent earthquakes, as could be seen in Figure 1.3a: Apr 6th 2:00 PM, Apr 8th 7:00 AM and Apr 9th 3:00 PM, approximately. Those two last earthquakes are emphasized in Figure 2.1.
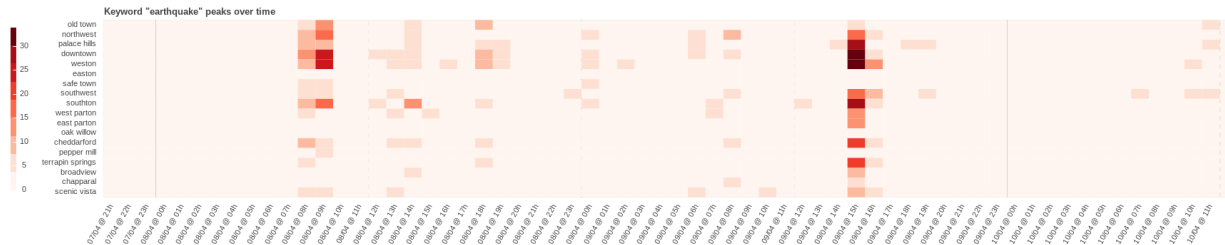

Figure 2.1: Heatmap for the two last earthquakes using keywords related to the word "earthquake".

Figure 2.2 show the root of the most-mentioned words in a 18h-time interval considering the starting point of the last two earthquakes. Power has been the most mentioned issue.
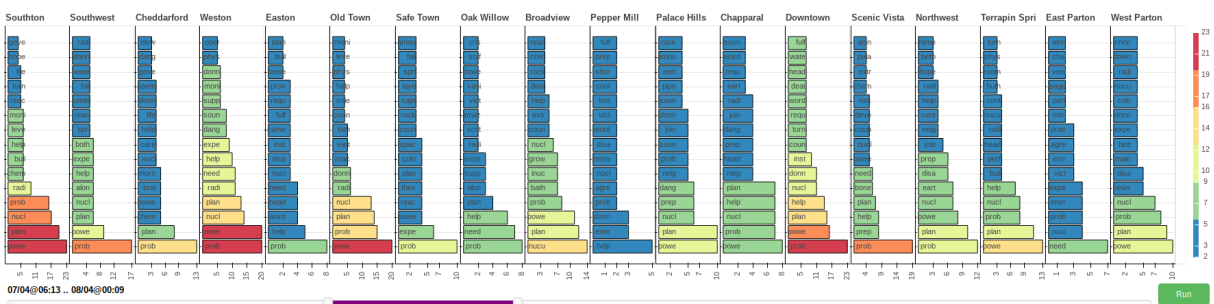
(a) Bar chart for the second earthquake in a 18h time interval from Apr 7th at 6:00 PM to Apr 8th at 12:00 AM. Issues with the nuclear power plant appear to have occurred over almost all neighbourhoods.
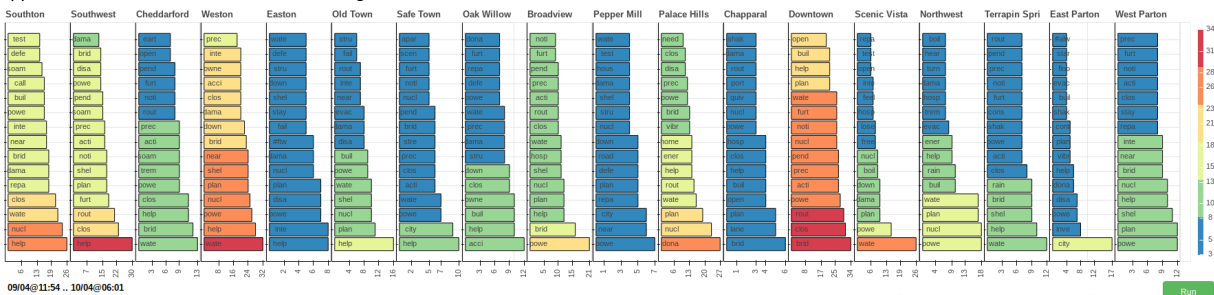

(b) Bar chart for the third earthquake in a 18h time interval from Apr 9th at 12:00 PM to Apr 10th at 6:00 AM. Apart from the problems with power, water and bridges issues also start to appear. Mentions to "shelter" and "hospital" can also be seen.

Figure 2.2: Bar charts for a 18-hour time interval with respect to the two last earthquakes.

Figure 2.3 shows the color distribution related to frequency of tweets by location at the second and third earthquakes over the SVG map. In general, the city's northwest region seem to have been the most affected by both earthquakes.

By looking at the heatmap of Figure 2.4 for each keyword, it can be seen that the damaged inflicted by the second earthquake was greater than both first and second ones. Suggestions for re-allocation of city resources are detailed as follows:

- *Medical:* At least 12 hours before 9:00 AM of April 8th, there was no occurrences regarding medical keywords.
    - On April 8th from 9:00 PM to 10:59 PM, the second earthquake time, there have been casualties on multiple locations but we would prioritize the darker colored ones according to the heatmap of Figure 2.4a: Northwest,

6
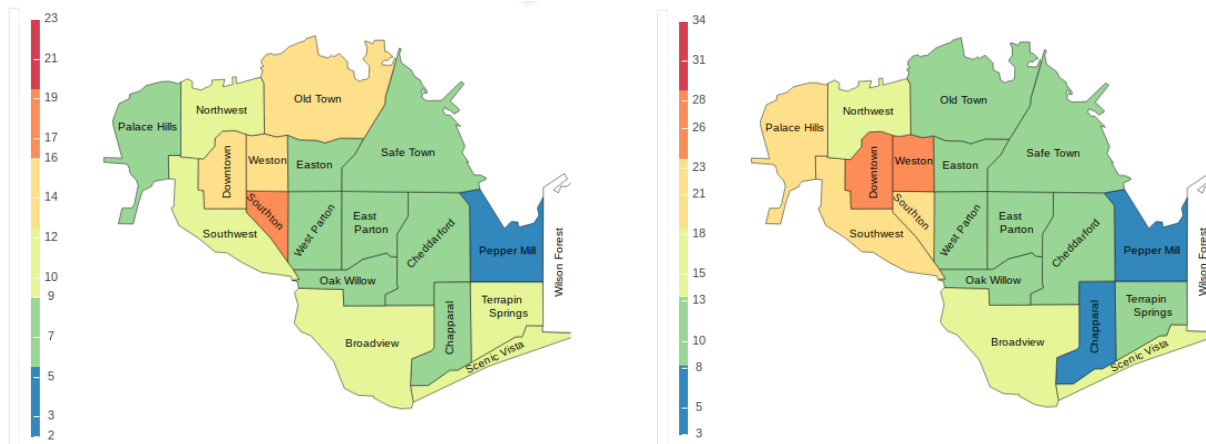
(a) Map for the second earthquake. Yellow and orange shades show locations where the frequency of tweets is higher. Southton, Downtown, Weston and Old Town appear to be, in that order, the neighbourhoods that most need help.

(b) Map for the third earthquake. Yellow and orange shades show locations where the frequency of tweets is higher. Downtown, Weston, Pallace Hills, Southwest, and Southton appear to be, in that order, the neighbourhoods that most need help.

Figure 2.3: St. Himark's SVG maps with blue-to-red colormap for the second and third earthquakes.

Southton and Downtown. First, a rescue crew must be sent to Northwest and Southton between 9:00 AM and 9:59 AM. Then, from 9:00 AM to 9:59 AM, for being geographically close a crew must be reallocated to Downtown.

– Again, between 12:00 PM and 13:00 PM there have been no medical occurrences but from 13:00 PM to 16:00 PM, the number of occurrences increased significantly and affected almost all neighbourhoods except Safetown and Terrapin Springs. Accordingly, we would prioritize the dark-red-colored ones, Downtown and Palace Hills, sending rescue team from 14:00 PM to 15:00 PM for both neighbourhoods.

– On following hours after that peak in occurrences, there have been only sporadic requests that could be solved by sending small units to individual locations.

- *Sewer and water:* According to the Figure 2.4b before 13:00 PM on April 8th there have been less-intense occurrences that could be solved by sending small units to individual locations.

– From 13:00 PM to 14:00 PM there have been water and sewer requests from all neighbourhoods, but Oak Willow and Easton. Since there was an ocurrance overload we would again prioritize the darker colored areas. Thus, a water and sewer crew must be sent to Weston and Downtown between 13:00 PM and 14:00 PM.

– On following hours the conditions changed back to infrequent occurrences.

- *Rain*: Major issues have occurred in Downton and Weston according to the coloration of those neighbourhoods shown in Figure 2.4c

– Between 19:00 PM and 19:59 PM a rescue crew must be sent to Downtown and Weston. Northwest also has a rain demand at this time interval but the frequency of the request is much lower than the aforementioned neighbourhoods.

– After this time interval there is a conditions change and so between 12:00 PM and 13:00 PM on April 9th the resources can be relocated from the already attended neighbourhoods to the remaining, such as Northwest, Southton and Broadview.

- *Roads and Bridges*: For at least 11 hours the number of mentions related to roads and bridges ceased. Between 9:00 AM and 13:00 PM on April 8th the number of mentios start to increase again mainly at Downtown. Since delays are expected at this neighbourhood because some roads are under resurfacing a crew will not be sent there.

- *Building*: Between 14:00 PM and 18:00 PM the neighbourhood with the largest number of mentions related to building changed threee times. Thus, at 14:00 PM crew must be sent to Weston, then at 15:00 PM a crew must be sent to Southwest and lastly at 18:00 PM a crew could be reallocated from any of the neighborhoods already attended to Downtown.

- *Power*: According to the Figure 2.4f there have been at least three times that conditions regarding power keyword changed considerably.
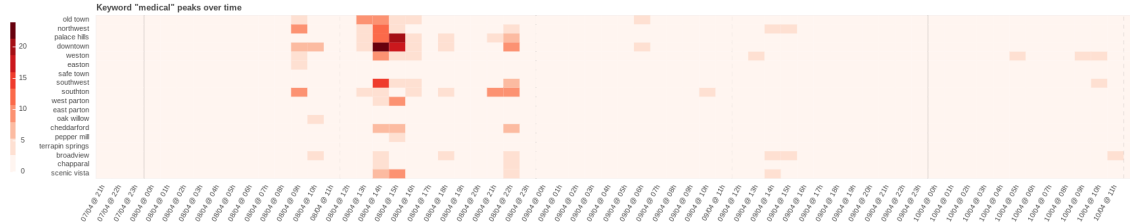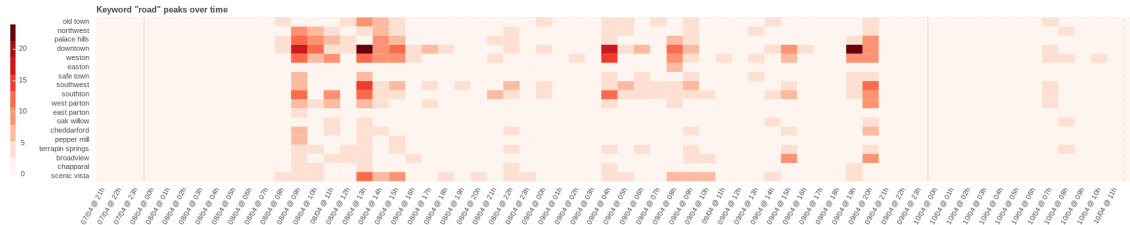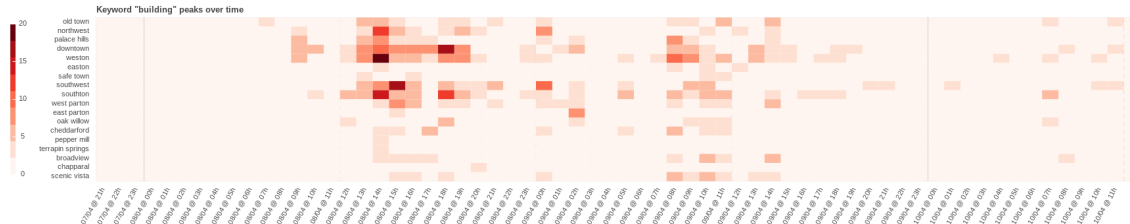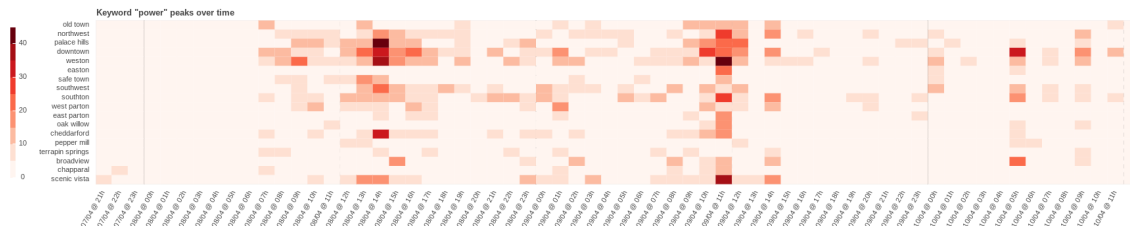
(a) Medical
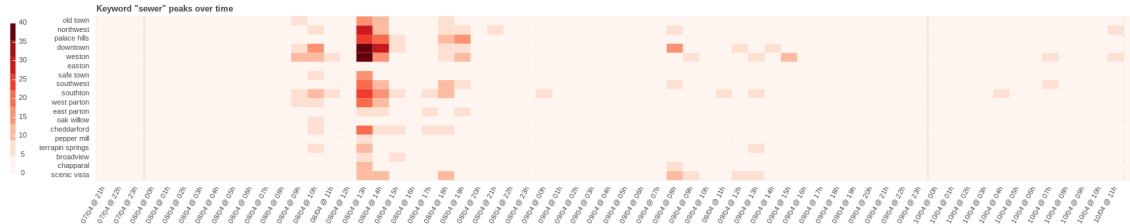
(b) Sewer and water
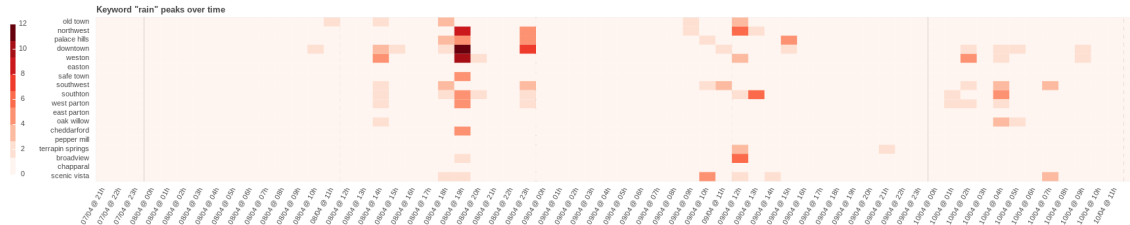
(c) Rain

(d) Roads and Bridges

(e) Building

(f) Power

Figure 2.4: Heatmap of conditions for the two last earthquakes

– Between 07:00 AM and 14:00 PM on April 8th the number of mentions for keywords related to power progressively increased mainly at Palace Hills, Weston, Downtown and Cheddarford. Since Palace Hills and Downtown have hospitals, they will be prioritize over the others and a power crew must be sent to both of them.

– Between 09:00 AM and 11:00 AM on April 9th the conditions change and the power crew must be now alocated to Scenic View and Weston, which are dark-red-colored neighbourhoods.

– Between 04:00 AM and 05:00 AM on April 10th the mentions related to power had a slight increase after having stopped for some hours. A power crew must be sent to Downtown at this time interval.

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 earthquake has produced some already expected situations such as chaotic traffic, despair among the population, etc., which has also generated some kind of uncertainty at the data. For example, the keyword "*bridge*" has been mentioned multiple times, which suggests a structural problem that could've ended in collapse if analysing the isolated word. However, by analysing the whole context by means of a bar chart of retweets frequency, as shown in Figure 3.1, we can see that most mentions refer to closing bridges instead of reporting physical problems at the bridge.
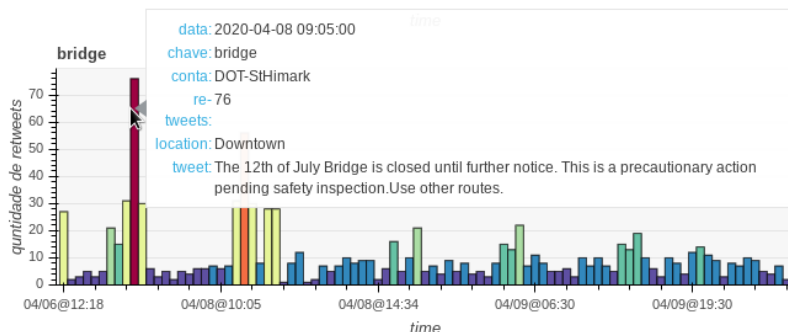

Figure 3.1: Bridge closed

Another point is with respect to fake news. At 8:00 AM of Apr 8th, the tweet shown in Figure 3.2a appeared to warn users about an eventual tsunami, and a lot of retweets have followed to spread such "news". Nevertheless, as reported on Apr 9th at 9:00 AM in the tweet shown in Figure 3.2b, the city's unique geography prevents such disaster to happen, and according to the greater number of retweets we can infer this is the real truth about existence of upcoming tsunamis in St. Himark.

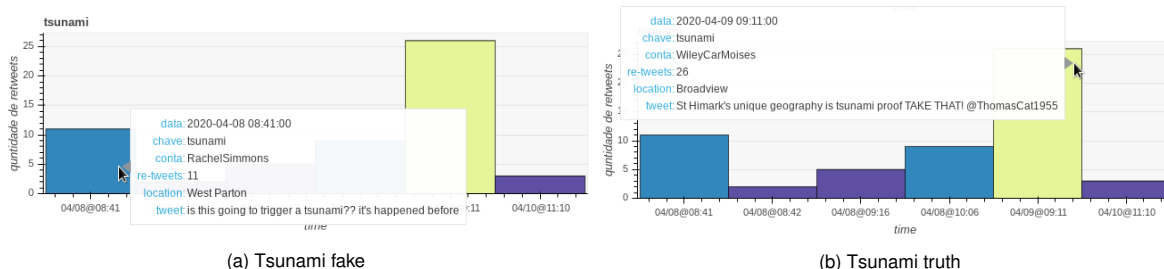(a) Tsunami fake

(b) Tsunami truth

Figure 3.2: Tsunami fake news

A third discovery made through the retweets frequency bar chart is that there was a circus show going on in town at the same week. So after the disaster that endend with buildings falling down on the floor, the elephants from the circus started to be used to help lifting heavy blocks of rocks in an attempt to search for people or corpses. This information can be seen at Figure 3.3.
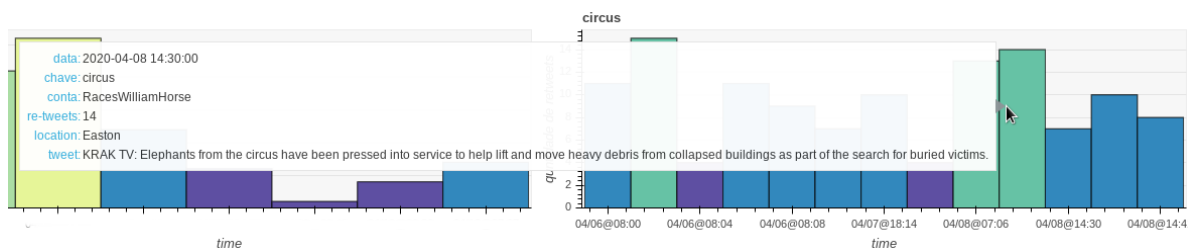

Figure 3.3: Circus elephants

10

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 were analyzed as static collection in order to facilitate interpretation of the problems. Using a stream analysis many decision-making for the allocation of resources might be misleading. Figure 4.1 shows an example for the keyword *"rain"*.
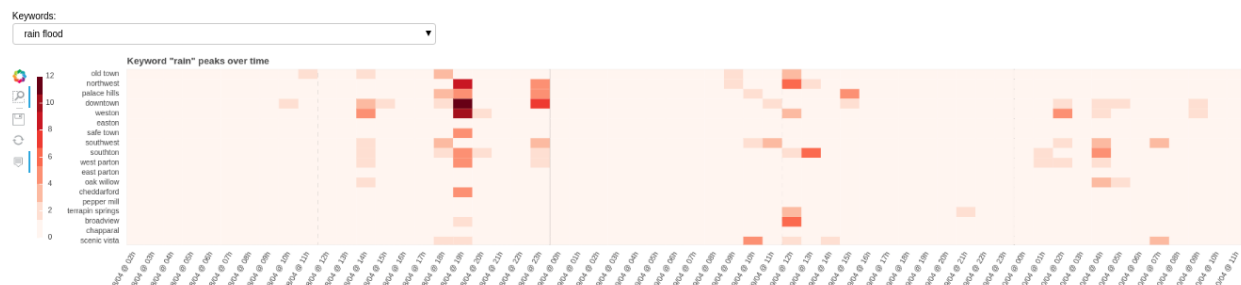

Figure 4.1: Heatmap for the "*rain*".

At 6:00 PM, five neighborhoods tweeted about rain or flood. If for each locality a rescue team was sent, at 7:00 PM where almost all localities reported rain or flood occurrences, so it would have been more difficult to efficiently serve all those affected neighborhoods.