Social Media Driven Disaster Relief Propagation

Analysis of Reactions to the St. Haimark Earthquake and Subsequent Events on Social Media

Keisha Mukasa || keisha.mukasa@tufts.edu | kmukas01
Daniel Jelčić || daniel.jelcic@tufts.edu | djelci01

Abstract

This process book is going to follow the exploration of the data and the creation of visualizations aimed to better understand the incident of the St. Himark earthquake and aid disaster relief-related decisions of St. Himark government and city services. The St. Himark earthquake resulted in ramifications for the city's infrastructure, public facilities and individual citizens, all of which need to be addressed with efficiency and accuracy. As officials are looking to establish an efficient disaster relief plan, we have chosen to look into social media posts and its related data from the Y*INT platform to try to extrapolate the most relevant information and visualizations of the data that would help in allocation and reallocation of the city's disaster relief resources. The data includes messages over the 4 days surrounding the earthquake. Each message has a time-stamp, location and user ID and all of these pieces of information can be used to derive information about the earthquake.

Due to the highly noisy and unstructured quality of the data, and since the raw data doesn't actually contain almost any of the features that would lend themselves to an analysis such as the one we want to enable, a lot of our work is focused on data cleaning and transformations. We have processed the data set with a python script such that we have information on posts, users and hashtags in easily accessible objects which we'll use to extract extra information and compute metrics. So far, we've scored users based on the number of mentions and reposts they get, and we've scored posts based on their repost count and user score. We plan on adding ratings for sentiment with the help of Google's NLP service, improved user scores using Eigencentrality and topical information using TF-IDF and ML models. Finally, we plan to devise a visualization system consisting mostly of a map of Heimark, displaying information about sentiment and activity per neighbourhood, trending topics, hashtags and posts, in a way that will easily aid decision making on resource allocation. All the views will be coordinated multiple views, and the user is going to be able to filter by time, neighborhood, user, topic etc.

Goals and Tasks

1. **Task**: Clean and generate more information from the dataset from the Y*INT
    a. **Methods**:
        i. Use numpy to create a data structure storing Y*INT data
        ii. Manipulate and extract particular data points - e.g hashtags, list of re-posts for a particular post
        iii. Derive user relevance from number of mentions and reposts
        iv. Calculate score for user relevance
        v. Derive a given posts relevance using number of
        vi. Store hashtags and relevant hashtag data

    b. **Goals**:
        i. Manipulate dataset to a format that derives more information from the given Y*INT dataset

2. **Task**: Calculate user relevance and post relevance
    a. **Methods**:
        i. Create a ranking algorithm for users and posts and store these scores in the data structure
        ii. Analyse the relevance of different terms by firstly using an inverse document frequency analysis
        iii. Curate our own relevant list of terms that we will map to the messages that are most related to terms we got had a high relevance from the inverse document frequency analysis
        iv. Scale relevance of posts and users

    b. Goals
        i. The derived user rank will allow us to easily visualise user importance in relation to the other users
        ii. Identify words that are most frequently used and important to the dataset

3. **Task**: Analyse the message data to obtain the general sentiments of the community
    a. **Methods**:
        i. Use the Google NLP to obtain sentiment data for each message
        ii. Identify the overall sentiments of the community in specific locations and at given times with respect to the earthquake.

    b. **Goals**:
        i. Derive qualitative data about the earthquake

        ii. Identify whether there is a relationship between the location with low sentiments and neighbourhoods with demand in high aid and assistance

4. **Task:** Topical categorization
    a. Methods:
        i. Categorize relevant posts by topic
    b. Goal:
        i. Easy analysis of posts by category thus filtering posts with little relevance to the earthquake
        ii. Topical categorization will serve the purpose of understanding what people are talking about

5. **Task**: Create Heatmap of general neighbourhood sentiment and Y*INT post frequency
    a. **Methods**:
        i. Use map of St. Himark where color of highlighted areas will be used to discern the sentiments of the communities and opacity will distinguish the message frequency over a given period of time
        ii. Include a sidebar with the relevant messages for the given window of time in order to give the user added information.
        iii. Add a slider for a given window of time so that a user can visualise the change in activity and sentiment over time
    b. **Goals**:
        i. Map high and low sentiment values in given neighbourhoods

6. **Task**: Generate a bar chart depicting Y*INT messages topic popularity
    a. **Methods**:
        i. Plot topics from our dataset on X-axis
        ii. Y-axis will represent the number of related messages for our given topic
    b. **Goals**:
        i. This visualisation will help determine the topics of most importance to users and can also allow for easy identification of which issues most urgently need to be addressed

7. **Task**: Design 0-centered bar chart showing over sentiment score per neighbourhood
    a. **Method**:
        i. Compute and scale sentiment data
        ii. Filter the data displayed in the chart by time window and topic
        iii. The neighbourhoods will be listed on the y axis and the width of the bars will indicate the overall sentiment for that neighbourhood
    b. **Goal**:
        i. The goal in this visualisation is to be able to easily compare the sentiments of neighbourhoods after an earthquake.

8. **Task**: Create a line chart that shows activity on Y*INT platform over time
    a. **Method**:
        i. Plot number of posts from Y*INT data per hour over the course of the 4 days during and after the earthquake
    b. **Goal**:
        i. Track peak times of Y*INT activity to see if they correlate to time of the earthquake or urgent events after the earthquake
        ii. Display an overview of the activity on Y*INT over time

Data

The data set contains 41942 rows, each containing feature values for one Y*INT post. The features include a timestamp, location the post was sent from, the username of the person who posted it and the content of the message. The timestamps are in the YYYY-MM-DD HH:MM:SS format and range from 2020-04-06 00:00:00 to 2020-04-10 11:59:00. The location row contains names of one of the 19 neighborhoods of St. Himark, which can be directly mapped to the interval range [0, 19]. The username column simply contains repeating usernames of people who sent posts.

The final column containing the contents of the posts is the most content rich, most unstructured and probably most overwhelming one. At first sight, a quick random sampling of the data shows that it was probably generated using a combination of human input and computer synthesis, and is often times either senseless or unintelligible. Taking a closer look, however, we see that the messages that contain the “re: “ prefix are reposts of other posts, and counting those can tell us a lot about the influence and popularity of the post that was being reposted. We can also see that posts feature @user mentions, which can tell us about the influence and popularity of users. Finally, some posts also contain hashtags, which we can analyse to gauge the intentionally labeled trending topics from the 4 day timespan.

Hypotheses

1. Null: There is going to be no connection between Y*INT activity (number of messages in a fixed time period) and the timing of the earthquake.
2. There is going to be a peak in Y*INT activity immediately after the timing of the earthquake.
3. Local peaks in Y*INT activity are going to correspond to timings at which a switch in resource reallocation would be advisable.
4. The majority of the data points (Y*INT messages) are not going to be related to the earthquake.

5. The neighbourhoods with the lowest sentiment scores will be closest to the epicenter of the earthquake

Storyboards

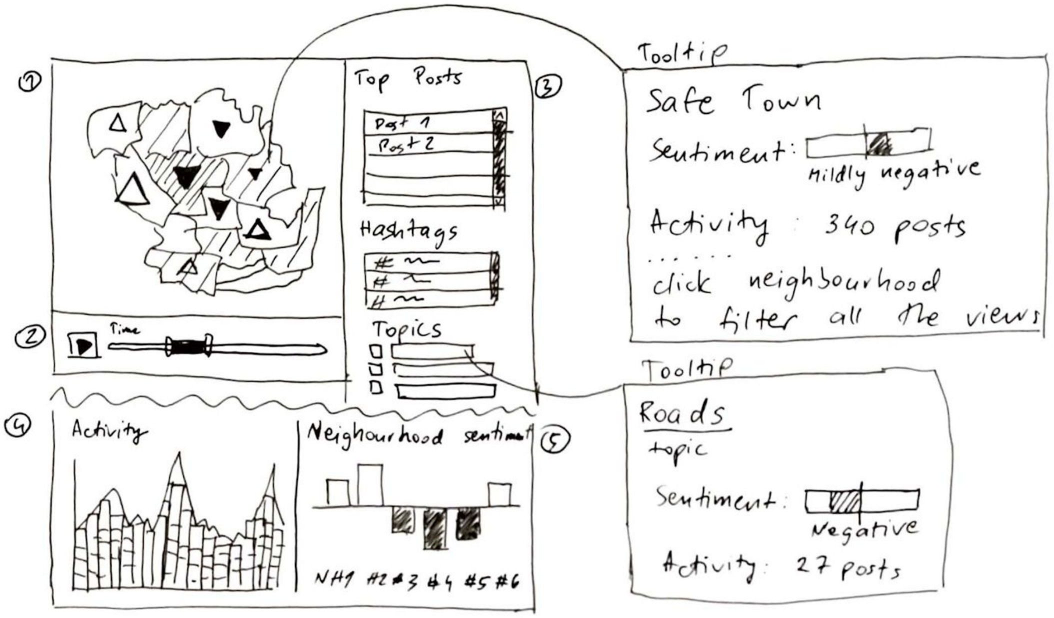


Figure 1: Storyboards for the visualization system

The above is a basic overview of how our visualization system is going to work. View (1) is the map chart of St Himark's neighbourhoods, in which activity (posts and reposts) is encoded in the background color and sentiment is encoded with upwards arrow for positive sentiment and downwards for negative, where the size of the arrow is intensity of the sentiment. Clicking on any of the neighborhoods filters the current data set to include only that neighbourhood. View (2) is the time indicator - it shows us the time interval currently selected, and a play button that can move that time interval to the right automatically, for a simulation of how this data would have looked like in the system had it been live and not static. View (3) indicates lists of top posts, topics and hashtags, the background color of each representing the relative score of the item for the selected data set. Hashtags and Topics can be used as checkbox filters for the data in all the views. View (4) is a simple view of activity over time, where the time range is equivalent to the range selected in (2), the line plot is total activity over time and segmented bars represent the breakdown of activity per neighbourhood. View (5) is a 0-centered bar chart of sentiment per neighbourhood. This view is an alternative version to view (1): whereas view (1) lends itself to a better geospatial analysis, view (5) lends itself to a better sentiment-based analysis. Neighbourhoods, posts, hashtags, topics, bar segments for activity over time per neighbourhood and bars for sentiment per neighbourhood will all have tooltips with all the relevant information. View (4) will be used to test our hypotheses 1, 2, and 3, view (2) will be used to test hypothesis 4 and view (1) to test hypothesis 5.

Interaction Descriptions

In our visualisation we have included numerous features to enable users to filter the data in order to narrow down the data points they look at. The main aspects of the data that have filters are time, neighbourhoods, topics and hashtags. The time is filtered using an interval slider to narrow down the timeframe of the data being viewed. This can help users identify key interval times in relation to the earthquake. It also makes it simpler to show when users were posting most about the earthquake.

The visuals also include an interactive feature for neighbourhood selection. The user can select a given neighbourhood and derive the overall sentiment, important posts and the post activity of a given neighbourhood. This is helpful when policymakers need to make decisions concerning resource allocation.

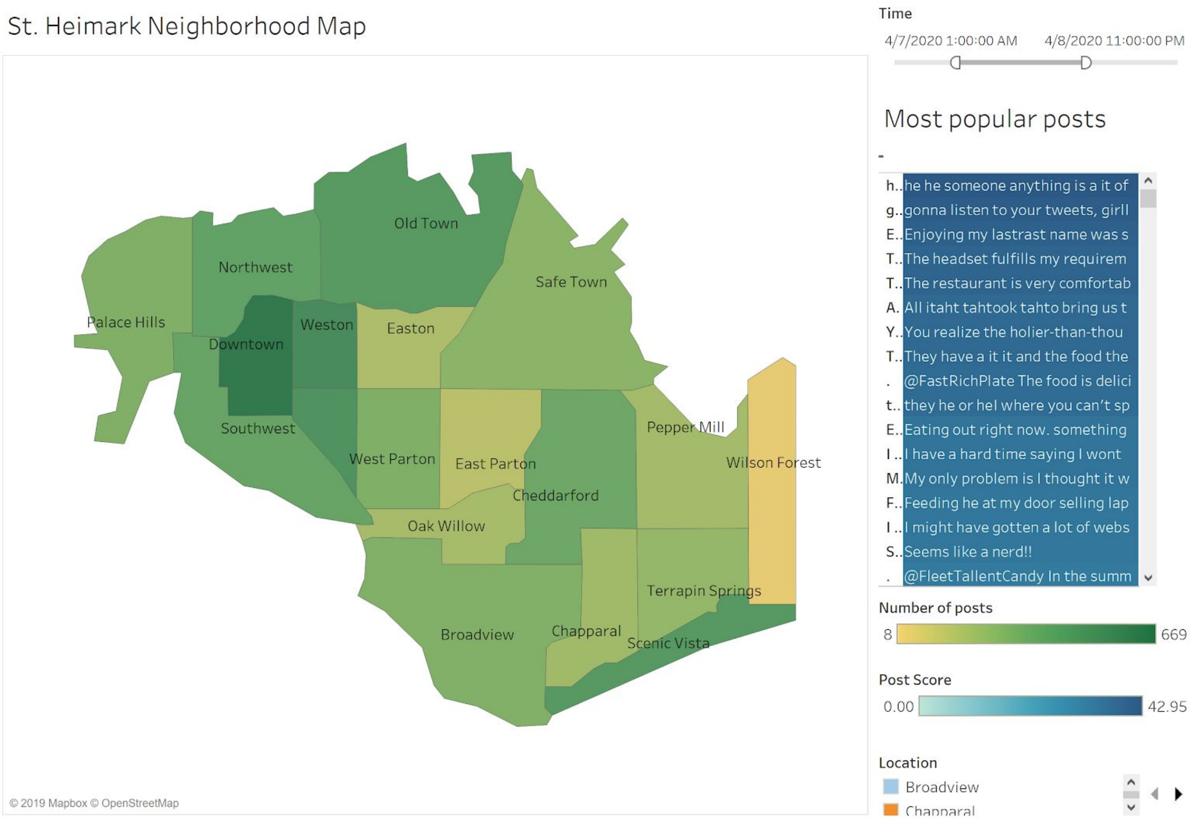
The dataset will also include groupings of posts based on topics concerning the earthquake since numerous posts from the Y*INT dataset have no relation to the earthquake. The user will be able to look at specific topic data using a specific topic checkbox tool.

Given that certain hashtags occur frequently in the Y*INT dataset we have chosen to include a hashtag filter. This filter will enable users to see the progression of certain hashtags and analyse the posts where the hashtags were derived.

Analysis Steps

- Compare finished visualizations to initial hypotheses and draw conclusions about why they were either true or false
- Analyse each visualisation and extract the necessary information about the data points explored in the visualisation
- Identify missing data points when making conclusions and discuss whether the absent information could have potentially changed the results derived from the visualisation
- Derive necessary conclusions backed by particular data points that officials can use when distributing resources and identifying areas in need of emergency aid.

Progress



St. Heimark Neighborhood Map

Time

4/7/2020 1:00:00 AM 4/8/2020 11:00:00 PM

Most popular posts

-

h..	he he someone anything is a it of
g..	gonna listen to your tweets, girll
E..	Enjoying my lastrast name was s
T..	The headset fulfills my requirem
T..	The restaurant is very comfortab
A.	All itaht tahtook tahto bring us t
Y..	You realize the holier-than-thou
T..	They have a it it and the food the
.	@FastRichPlate The food is delici
t..	they he or hel where you can’t sp
E..	Eating out right now. something
I ..	I have a hard time saying I wont
M.	My only problem is I thought it w
F..	Feeding he at my door selling lap
I ..	I might have gotten a lot of webs
S..	Seems like a nerd!!
.	@FleetTallentCandy In the summ

Number of posts

8 669

Post Score

0.00 42.95

Location

Broadview

Chapparal

Figure 2: Prototypes of views (1), (2) and (3)

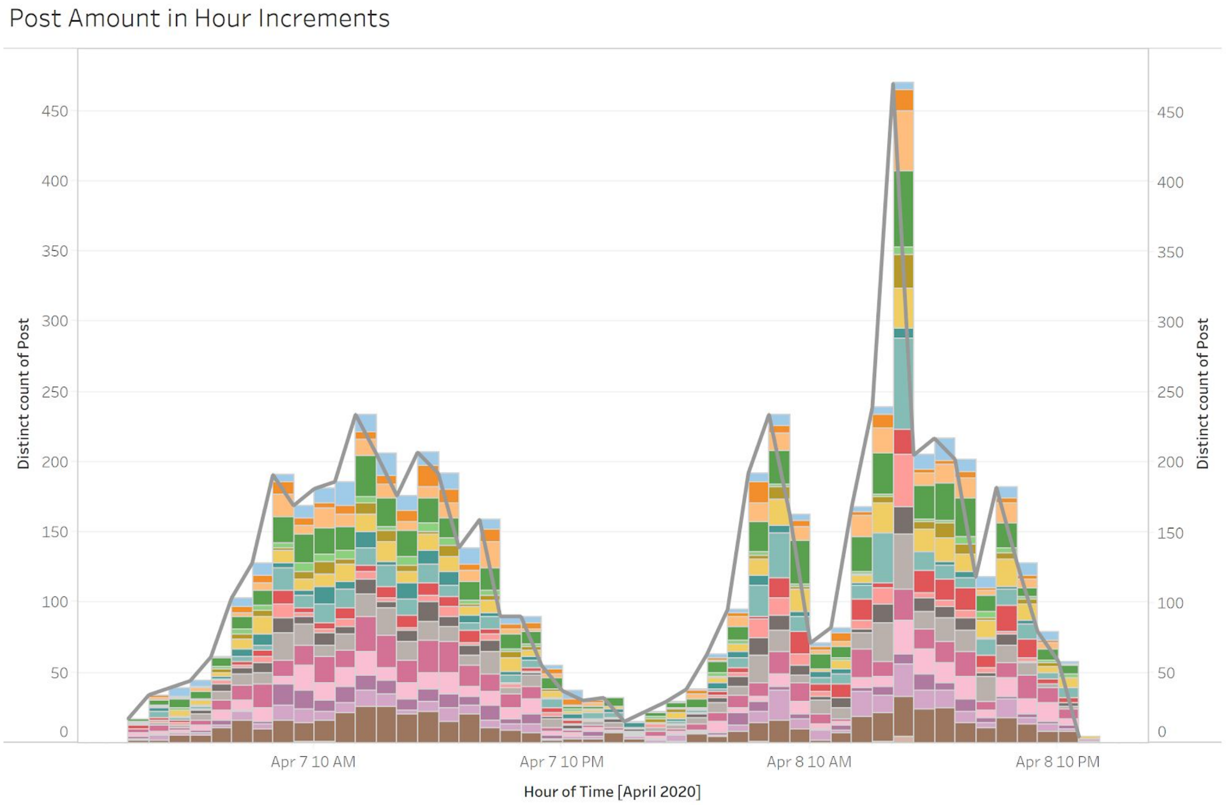


Figure 3: Prototype of view (4)

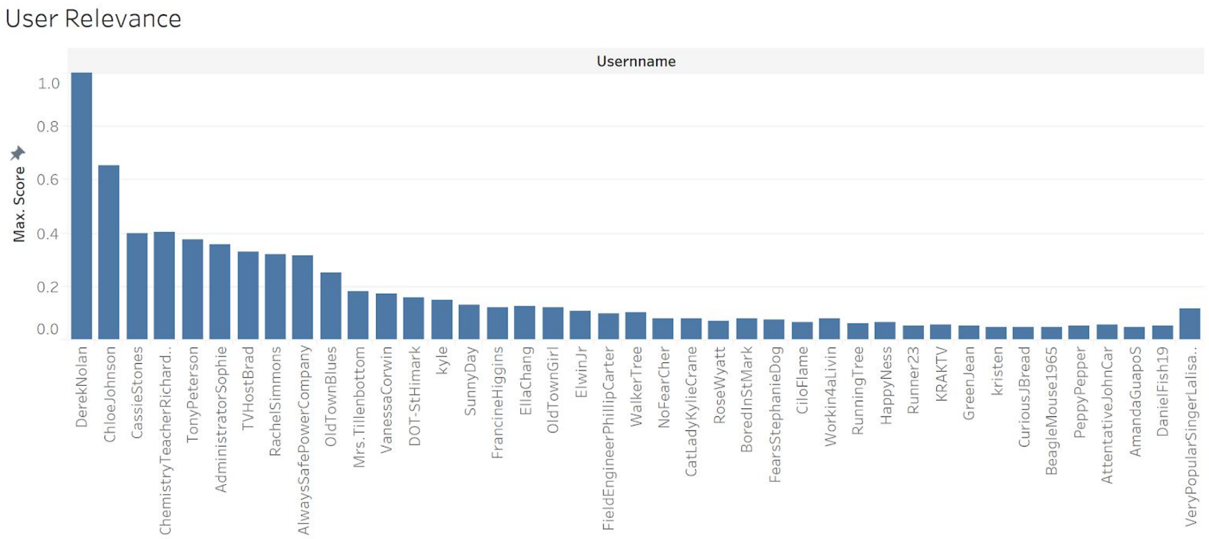


Figure 4: Analysis of user influence

Timeline

Completed Tasks:

- Data cleanup (neighbourhood numbers instead of names, post repost count etc.)
- Start working on the ranking algorithm
- Calculate general post relevance
- Calculate post relevance to the earthquake
- Calculate composite post relevance
- Calculate relevance of users based on mentions and reposts of their posts
- Plot the simple graphs in Tableau

Remaining Tasks:

- **Nov 14 / Progress Report**
- Calculate relevance of users (incl. graph analysis and Eigencentrality)
- Filter posts by relevance
- Collect and process data on post sentiments
- Plot the simple graphs in Tableau
- Work intensively on the main heat map view in Tableau
- Collect and analyze feedback
- Reflect on current discoveries and potentially readjust
- Start devising methods for topic identification
- Post topic labeling
- Post topic model training
- Adding the topic data to the data set in Tableau
- Adding topic related filters and features to Tableau
- Add the sentiment per neighbourhood visualization and the relevant filters
- Start exploring the play button
- Start exploring the model that predicts best allocation of resources
- Create appropriate visualizations for the model output
- **Dec 16 / Final Presentation**

Features

Must-Have

- Activity over time chart
- Popularity of topics chart
- Heat map of the neighbourhoods, indicating activity and sentiment
- Time slider for the heat map

Good-to-Have

- Overall sentiment per neighbourhood interactive chart

Optional

- Play button that visualized the progression of the events over the 4 days
- Predictive model that automatically identifies where resources should be allocated and reallocated

Team

We plan for Keisha to focus on the Tableau heavy lifting and the research into integrating the Google NLP API into our data manipulation script. Daniel is going to focus on working on the custom map of St. Himark in Tableau, data manipulation script heavy lifting and the ML components such as choosing and training the predictive model for topic categorization.We will focus equally on the post rating algorithm design and iterative design changes that will become inevitable with new discoveries in our data.