Adam Ten Hoeve

Final Project Report

COMP 4449 – Data Science Capstone
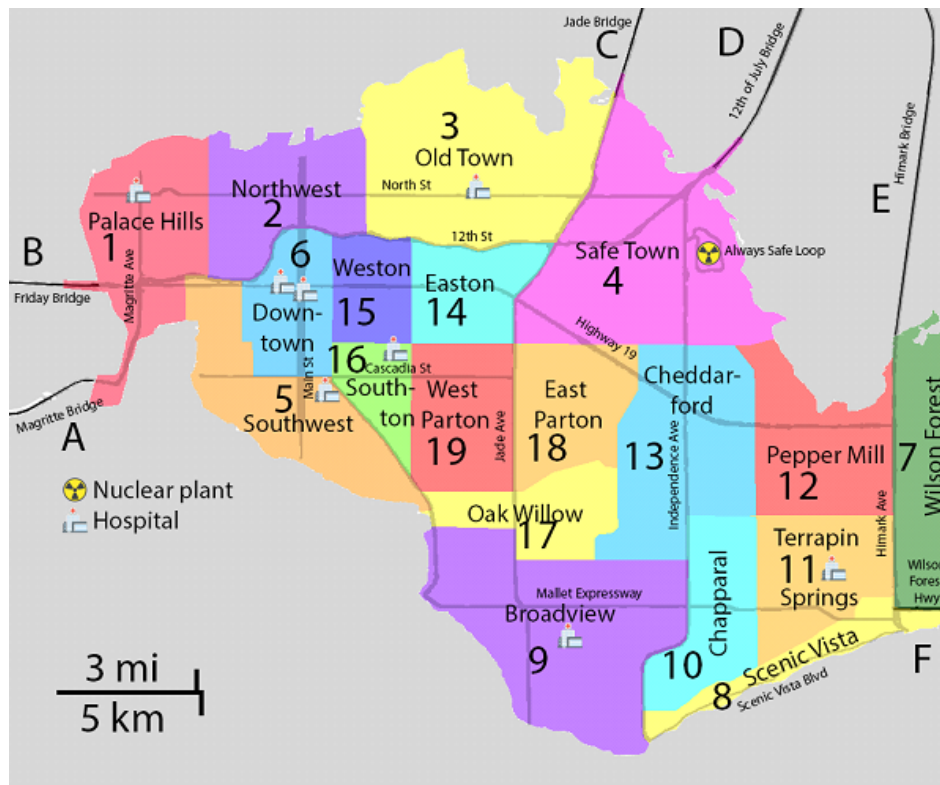
11/18/2021

Introduction

The fictional city of St. Himark has been hit by a severe earthquake and has suffered significant damages. City officials are scrambling to distribute emergency resources and response teams. However, in order to properly allocate those resources, the city officials require information on what problems are occurring and in which neighborhoods. Our task is to investigate a dataset of social media messages in order to characterize conditions across the city and recommend how emergency resources should be allocated. We must also keep in mind that priorities may change over time as the state of neighborhoods also change.

The City

Before we start working with the data, it is important for us to have an understanding of St. Himark.

St. Himark is divided into 19 neighborhoods, each with their own geography and lifestyle. The North-Western and Central areas are primarily urban city centers, the North-Eastern and Southern regions are suburban family homes, and the Western regions are mostly rural communities. An important landmark in the city is the Always Safe Nuclear Power Plant, which is the primary provider of the city's power. It is located in Safe Town. There are also 8 hospitals in the city, 4 located in the central neighborhoods, 2 along the Northern coast and 2 along the Southern coast.

Data Description and Cleaning

The provided social media data is a single table with 41,941 rows. Each row corresponds to a message posted on the platform. It contains the following features:

- account: The string identifier for each user on the social media platform.
- message: The string contents of the social media post.
- time: A datetime for when the message was sent, with granularity in minutes.

- location: The string corresponding to the neighborhood that the message was sent from.

The data spans 5 days, from 12:01am on 04/06/2020 to 11:59pm on 04/11/2020. We do not know when the earthquake occurred, so we will have to find that ourselves. There are 16,776 original messages and 25,165 reposts. Considering this, a significant portion of our investigation will be determining which messages were reposted the most, as that may give us an idea of what most citizens were concerned with. A selection of unaltered messages can be seen below:

Id	Account	Message	Location	Time
1	MabelTallR	Lots of cracks on North St going from Old Town to Northwest	Scenic Vista	2020-04-08 17:04:00
2	C2011Murphy	re: Look at that damage – this was a 4 story apartment building – help needed.	Northwest	2020-04-08 09:24:00
3	Syndicated4	Take advantage of these Ten, quaking sales!	Easton	2020-04-10 02:00:00
4	SmartFrankPlate	Oh htishis is htishe place ihtisself is well mroeade, easy htiso love, buhtis even mroeore htishan jushtis a neahtis gadgehtis.	West Parton	2020-04-10 05:22:00
5	FearsTheresa_ HaleyMouse	Ã\x83Â itÂ theyÂ itÂ itÂ theyÂ heÂ theyÂ sheÂ theyÂ anythingÂ theyÂ heÂ theyÂ someoneÂ	Southton	2020-04-08 07:53:00

An important piece of information we should gather from each message is the "intended location," meaning the location that the message is in reference to. For example, message (1) was sent from Scenic Vista but references Old Town and Northwest, so the intended location should be Old Town and Northwest. A similar problem applies to reposted messages, and which location they are referring to. For example, message (2) is a repost from Northwest saying that a building has collapsed. Does that mean that a building has collapsed in Northwest, or is the user drawing attention to the original
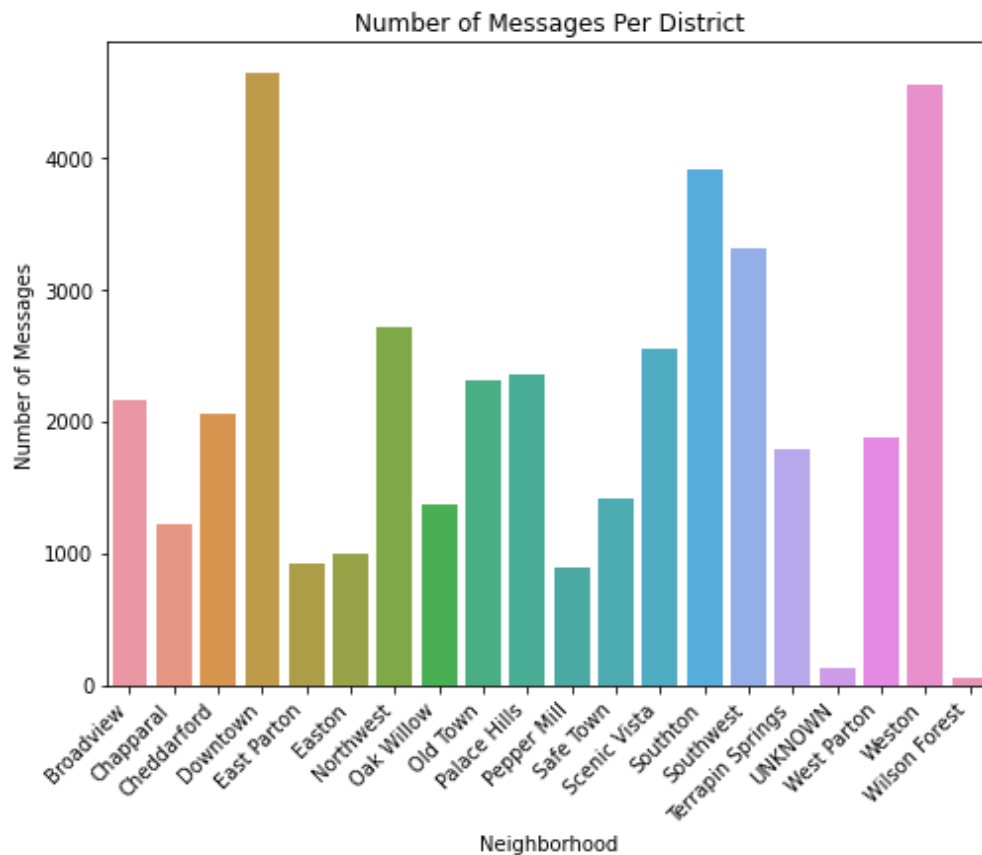
message about a building collapsing somewhere else? To address these location-based issues, the

"intended location"was determined for each message using the following rules:

1. If a message references a location, then that location is the "intended location." In this case, the location that the message was sent from is *not* the intended location.
2. If a message does not reference a location, then the location that the message was sent from is the "intended location."
3. If a message is a repost, then the "intended location" of the repost is always the same as the location of the original message.

Establishing this intended location provides us a greater understanding of the meaning of each
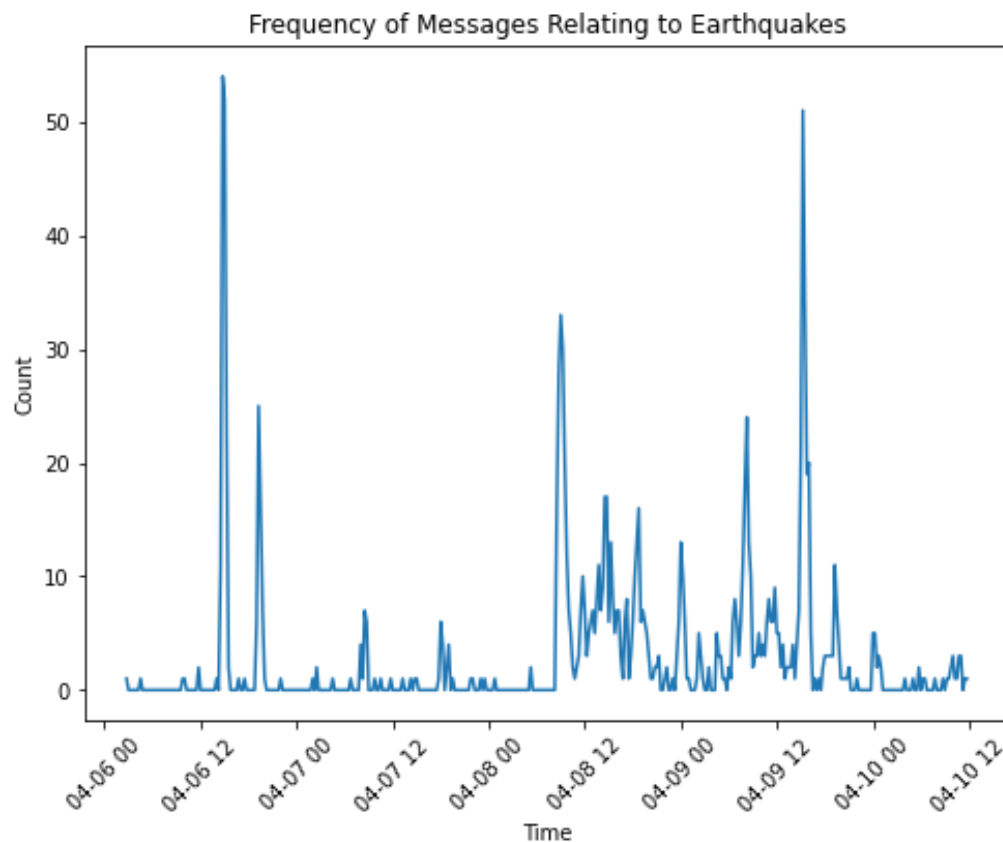
message.

Data Exploration

Before we start exploring the messages, we should get a basic overview of the data. In

particular, let's look at how many messages are sent per district.

Number of Messages Per District

There is a large amount of variation between the number of messages sent per neighborhood. The central areas like Downtown and Weston sent over 4,000 messages during the 5 day period, while most other areas only sent 1,000-3,000. We should keep this in mind when investigating the severity of issues in each location. That is, if the same number of messages mention an issue in Downtown and East Parton, then the issue may be more severe in East Parton because a greater proportion of messages concern the issue.

When Did the Earthquake Occur?

We can determine when the earthquake occured by plotting the times that the citizens mention words like "earthquake," "quaking," and "trembling." This produces the following plot.

The plot shows that there was a large amount of matches on 04/06 around 12:00pm, then a long lull until about 8:00am on 04/08. Afterwards, "earthquakes" were mentioned fairly regularly in the following days. We can not be sure when the time of the earthquake was from the plot, so we must at the most reposted messages during each of those times.

04/06 at 12:00pm	04/08 at 8:00am
`We jsutust recorded a Northwest.Old Town mild earthquake jsutust northeast of St.Himark town. Did you feel it? Probably no damage #HateDidWonder (6 reposts)` `What earthquake??? I didntididntn't feel anything! (10 reposts)` `Earthquake or not I still have to teach (7 reposts)`	`ALERT: A 6.7 earthquake just occurred off the NE shore of the town of St. Himark. This could be severe. Expect heavy damage. (6 reposts)` `My building collapsed around me! (8 reposts)` `National Earthquake Center reports a major earthquake with an epicenter near town. (7 reposts)`

The messages from 04/06 at 12:00pm mention an earthquake, but it is "mild" and many "didn't feel anything!" In contrast, the messages from 04/08 at 8:00am mention a "major earthquake," to "expect heavy damage" and even that someone's "building collapsed." From this, we can conclude that the true earthquake occurred on 04/08 at 8:00am.
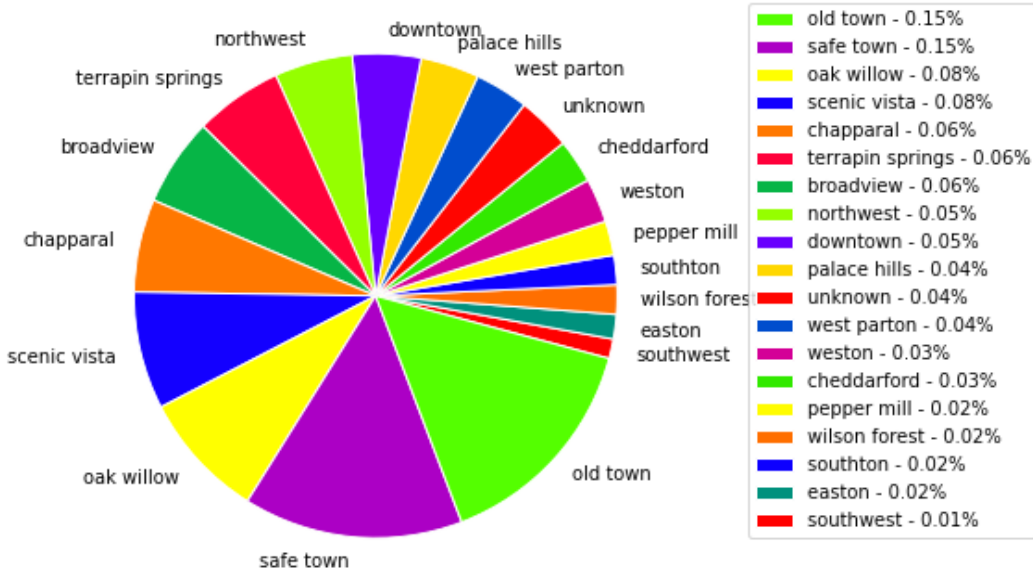
Investigating the Events

Now we should find out what areas have been damaged. To start, we can go through each issue and determine how many messages are concerned with that issue for each location. This goes off the assumption that more severe issues will have more messages and reposts about them.

I will use road issues as an example of the process taken during this investigation. Road issues are those which contain words like "rubble," "fissure" and "traffic." We want to determine the areas

that are most concerned with these issues. Also, recall that different areas produce different amounts of messages. That was taken into account during these analyses.

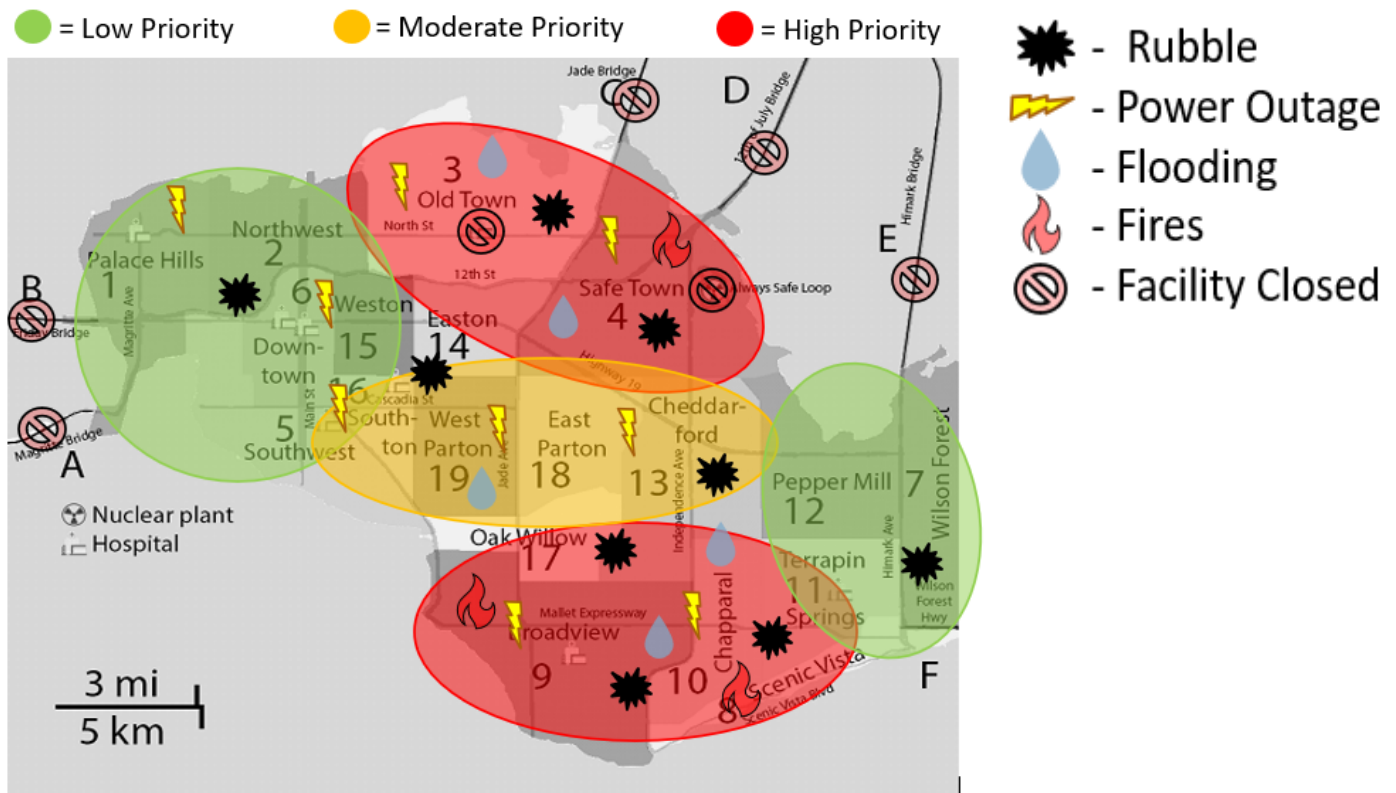Proportion of Message Pretaining to Road Issues After 5 Hours, Per District

We see that Old Town, Safe Town, Oak Willow, Scenic Vista and Chapparal sent the most messages about road issues within 5 hours of the earthquake. These areas make up the Northern and Southern territories of the city, so we have an idea of what areas were hit the hardest. Let's look at the most reposted messages from these locations to confirm that they are facing road issues.

Old Town	Safe Town	Oak Willow
– Our neighborhood has been hit hard. All the old brick buildings have collapsed or are heavily damaged. #neighborhood (9 reposts) – My building collapsed around me! (8 reposts) – Old Town hospital is heavily damaged! (2 reposts)	– Always Safe security fence is down and one their buildings has partially collapsed. You can walk right in! We need some cops out here right away (12 reposts) – Buildings are damaged, windows broken... I can smell smoke somewhere... Hope everyone is safe. (6 reposts)	– A major sinkhole has collapsed taken several cars with it. I think the watermain burst. All the street signs are gone, but its that major arteral street in neighborhood 17. I get a busy signal on 911 and people are trapped. (8 reposts)

The reposted messages confirm that these areas are suffering from heavy structural and roadway damages. This process was repeated for other issues, such as flooding, fires, power outages, trapped individuals, and so on to get an understanding of the overall conditions of St. Himark. Here are some of the key messages from that process:

- "Extensive damage on the north and south sides. Although no neighborhood has escaped damage. Severals fires throughout City"
- "Water is contaminated. Serious reactions reported in the following neighborhoods Old Town,Safe Town,Scenic Vista,Broadview,Chapparal"
- "I heard the power plant sustained damage – am I safe? Is there a radiation leak?"
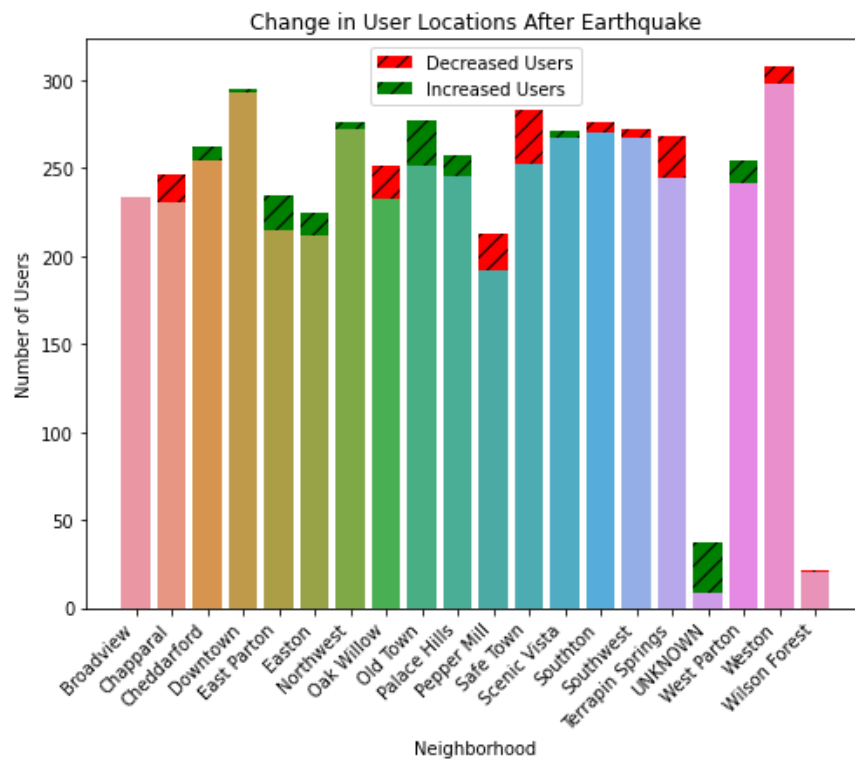- "Old Town hospital is heavily damaged!"

Through this analysis, we created the following picture of the city:

The major events that occurred immediately after the earthquake are as follows: The North-Eastern and Southern coast (the neighborhoods of Old Town, Safe Town, Scenic Vista, Broadview and Chapparal) have been hit the hardest. Their issues include collapsed buildings, flooding, fires, and

people needing rescuing. There was also damage to the sewer lines in these locations, meaning the water was contaminated and not safe to drink. The Always Safe Power Plant suffered damages and has shut down, which sent the majority of the city into a power outage. The other areas sustained some damage but were not nearly as bad as the areas listed above.

Now we must determine what happened in the days following the disaster. In particular, what changes occur in the city that may shift our priorities? Let's first look at how people moved around the city after the earthquake. We do not have population numbers but we can estimate those values, and how many people changed locations, by using the number of unique accounts in each location before and after the earthquake.
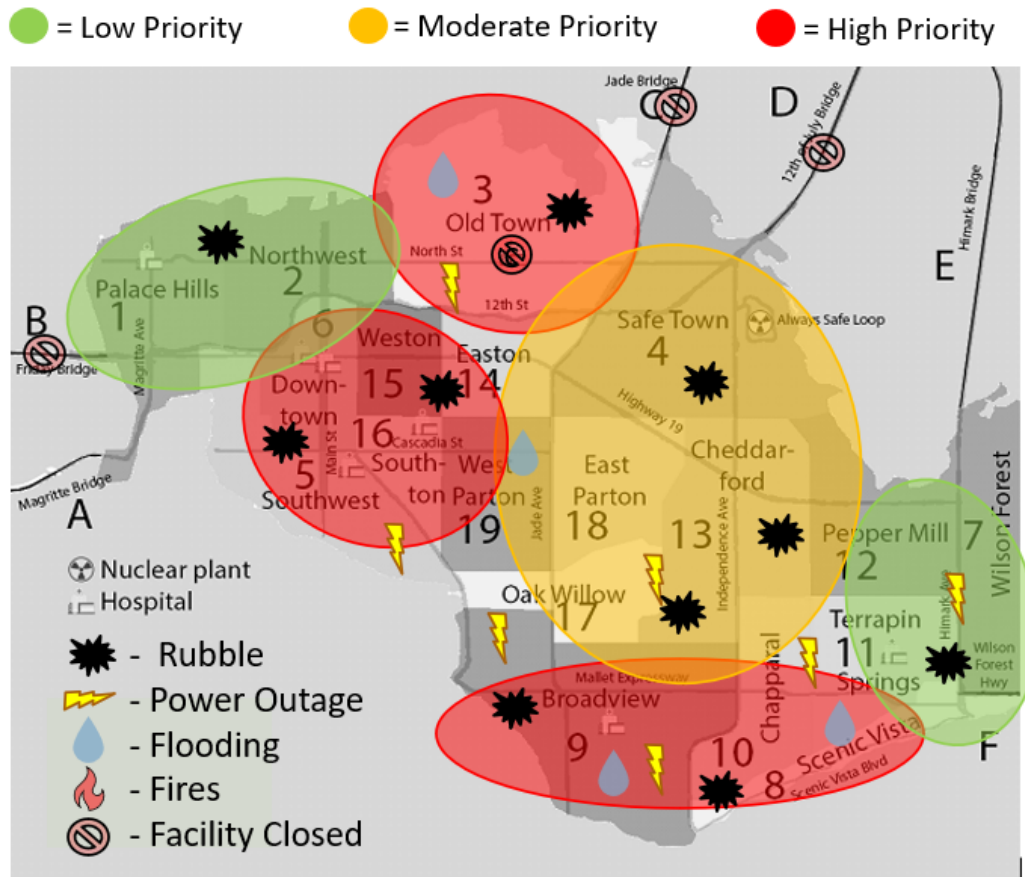

This plot shows the change in the number of unique users before and after the earthquake in each location. We can see that locations that were heavily damaged tended to have fewer users after

the earthquake, and less damaged areas had more users. From that, we can infer that people evacuated from those heavily damaged areas towards the center of the city.

At this point, we repeat the process of investigating different problems throughout the neighborhoods. This involved determining which issues were talked about, which districts talked about them, and what were each location's most reposted messages. Here are some key insights for that period:

- "We are running out of critical medical supplies, hosptials are burning through IV solutions, medicines, and much more. The roads are making it difficult to supply the hospitals and trucks can't get to our warehouse to keep us supplied. Road clearing from our warehouse in 16 to hospitals should be a priority."
- "The extent of damage to the sewer system is more than we initial thought. All St. Himark residents should boil their drinking water until further notice. We have put out a nationwide call for portable restrooms since the sewer system will take months to repair."
- "Electrical power has been restored to approximately 80% of the city. Internet service is also restored to most of the city. As we have seen in other disasters, the cell system is pretty resilient. Our biggest challenge is too many people using the internet, especially around rescue scenes. Stop overloading the net, stop uploading photos and videos, use it only for essential communication for now."

The overall state of the city looks as follows:

The Northern and Southern Coasts were still heavily damaged, and continued to need

assistance. As citizens evacuated towards the center of the city, that area became a high priority. In

particular, the roads needed to stay clear to allow access to the hospitals and allow the hospitals to

resupply as needed. The Always Safe Power Plant was restored, and power returned to the majority of

the St. Himark. Unfortunately, the extent of the sewer damage was greater than originally thought, and

all neighborhoods faced the contaminated water issue.

Conclusions and Resource Allocation

From this investigation, we constructed the following conclusions: The earthquake occurred at

8:00am on 04/08 and caused the most damage to the neighborhoods of Old Town, Safe Town, Scenic

Vista, Broadview and Chapparal. Road repair and rescue team should be sent to those areas

immediately. There is also heavy flooding along the Southern coasts, which will require sewer repair teams. The Always Safe Power Plant was also damaged, so any power and reconstruction crews should be sent there to prevent further issues.

After the initial effects of the earthquake, many citizens moved towards the hospitals in Downtown and Southton. As a result, these areas will require road teams to clear rubble, allowing the hospitals to remain accessible and supplied. The sewer system has suffered extensive damages throughout the city so all citizens will require clean drinking water and access to sanitization.