COMPSCI690V Final Report

Qi Pan, Bhavarth Pandya

December 18th, 2017

1 Running Instructions

Our visualizations can be run from our submission folder as the visualization using bokeh's serve command `bokeh serve --show X.py` where X is the python file name based on the visualization we are running. Below we outline how to run each visualization, and briefly describe what the visualizations show. We will also give our analysis for how they help towards solving 2014 VAST mini-challenge 1.

Note: There may be some library warnings in the command line when running our visualizations but rest assured they will not affect anything so please ignore them.

2 Dependencies

Unzip our submission folder and use it as a working directory from your command line (ie. cd to it). Please make sure the following things are in the working directory (should come from our submission zip):

1. `FP_emails_network.py`
2. `FP_news_timeline.py`
3. `FP_car_track.py`
4. `FP_credit_card.py`
5. `gps.pickle`
6. `VASTChal2014MC2-20140430`
7. `mich_twit_sent_data.txt`
8. `rt-polarity.neg`
9. `rt-polarity.pos`
10. `poi.csv`
11. `MC1 Data(folder)`

Please make sure you are using the Anaconda distribution with Python 3.6 and please install the following NLP packages by typing the following commands:

1. `conda config --add channels conda-forge`
2. `conda install spacy`
3. `python -m spacy download en`
4. `conda install nltk`
5. `conda install -c pelson pyshp`

Just as a precaution please also update the following packages, we will be using them as well (we are using bokeh 0.12.10):

1. `conda update pandas`
2. `conda update bokeh`
3. `conda update numpy`
4. `conda update matplotlib`

3 Approach

Our main approach was to not look at any of the 2014 VAST-solutions and try to come up with novel visualization/analysis approaches. We did this to prevent letting any of the past solutions bias our approach and to ensure that we were truly trying to innovate instead of putting a spin on past ideas. We decided that no matter how honest we were, once we saw some solution there's no way they could not at least subconsciously influence our tools.

4 Visual-Analytical Tools

4.1 Email Network w/ NLTK Naive Bayes Classifier

Please run `bokeh serve --show FP_emails_network.py` from our submission directory.

4.1.1 Concept

We started brainstorming for this tool with the idea of showing who the most important GAStech employees were in terms of communication, as well as try to identify unusual emailing/coordination patterns. If there were suspects within GAStech, maybe they coordinated through company emails. To help in noticing unusual communication patterns, we also wanted to run NLP sentiment analysis on the subject headers of the emails and display that somehow.

4.1.2 Data Setup

Behind the scenes, a Naive Bayes Classifier (NBC) was trained on a movie review data set combined with a twitter data set. Using this model, we predicted the sentiment of emails based on their subject headers as positive or negative on all internal emails between GASTech employees from January 6th 2014 to January 17th 2014.

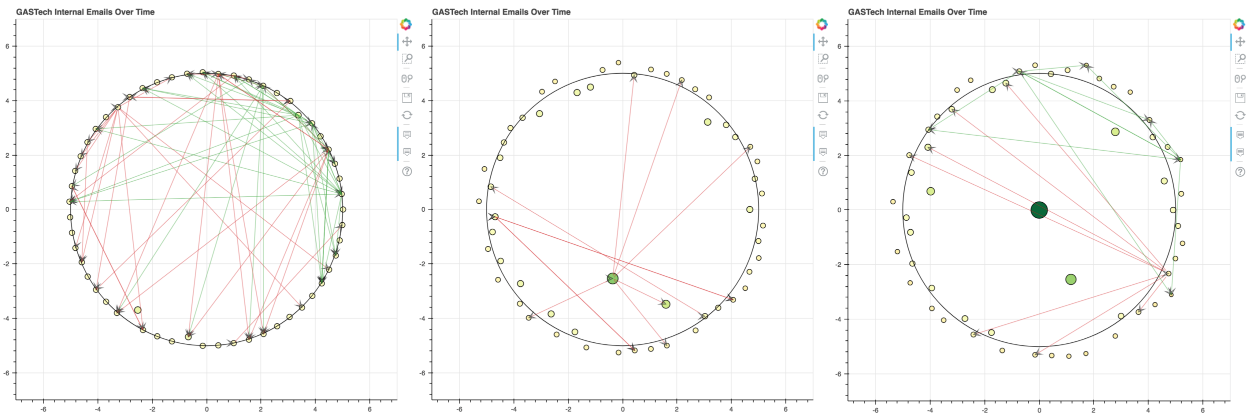


Figure 1: Email Network animation from time 0 to the end

4.1.3 Tool Description

A network showing each company member as a dot in a circle is displayed. By default, the slider allows you to "scroll" through each email sent in chronological order throughout the company. When a new slider value is selected, arrows originating from the sender of the email will propagate outwards to those the email was sent to. The arrows will be colored red (negative sentiment) or green (positive sentiment) based on the sentiment prediction of our NBC model. To automatically play through the emails, click the play button, you can click the button again to pause at any time (see fig. 1 for snapshots of the network animating through time). Another element captured in the model is the "importance" of users. Each time a user sends an email, their node will move closer to the center of the circle, and each time a user receives an email they will move away from the center (a reference circle of the starting positions is provided as well). We use this to infer who are the "leaders"/"commanders" based on who is issuing the most emails and who are the "followers" based on those who never send but just receive. One can also change the granularity of time and view the emails on a per hour or per day manner by selecting the correct radio button (although per day is way too messy to infer anything).

4.1.4 Analysis

The network animation helped us see who the primary "commanders" were, these employees were later connected to relevant news articles later (in connection with our other tools). Also, once we identified some suspect people via our other tools, we were able to track them on the email network and see that they collaborated often with suspicious email headers. One of the employees we identified as a communication leader was Ingrid Barranco who appears near the center at the end of the animation, she is highly likely to have been one of the kidnapped employees when relating this insight with our other like the timeline (where she was a popular subject of many articles due to her leadership in GAStech).

Looking to the future, we are happy with our tools ability to track suspect people, and identify primary voices within a communication network. It also generalizes to any dataset with a from, to, text format. Admittedly, the NLP we applied could have been improved to take into account the granularity of positive/negative sentiment instead of a binary choice.

4.2 News Article Entity Extraction w/ Spacy

Please run `bokeh serve --show FP_news_timeline.py` from our submission directory.

4.2.1 Concept

Given a collection of 845 news articles (a lot!) and the goalof this viz was to find contextual information about entities involved in these news articles without reading all the articles. With a word cloud approach it would have been difficult to realize a chronological order, thus we came up with the 'chronological entity pyramids'. Where we could stack up extracted entities (of a given type) from all the news articles on a chronological time-line. We also added some sentiment analysis in the mix by coloring the word pyramid accordingly. To add further context to entities, we setup an option to view the full article if need be by clicking on the article.

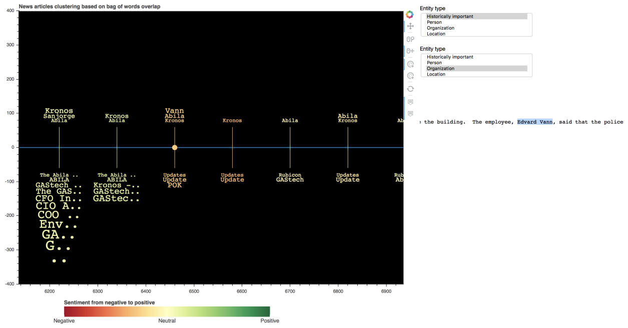
4.2.2 Data Setup

The news articles range over a 25 year period totaling to 845 articles. Many of these articles were translated from different languages and thus contain erroneous words and noisy data which presents as a challenge for us in sentiment analysis and entity recognition.This part uses the NER tagger of spaCy package to extract entities from news articles pertaining to kronos, gastech and pok. Spacy's sentiment analysis tool is also used to get average sentiment score for each article. The sentiment analysis tool is trained on spacy's own english corpus. The spaCy tools are based on a pre trained english language corpus and they require proper context to tag the entity types which sometimes is not fully available due to the noisy language data from the articles. Thus the methods can sometimes give incorrect results.

4.2.3 Tool Description

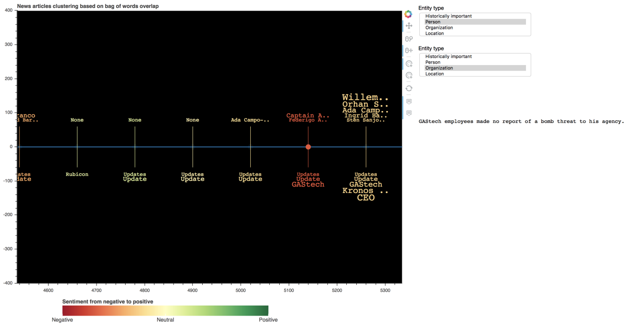
The visualization presents two timelines, top one for 20th and 21st January 2014 and bottom one for all the older years.The timelines present entitity lists in two directions up and down. The following list displays the properties and interactions involved with the timeline plots.

- The types of entities you want to display can be selected from the multi-select widgets adjoining the plots. The 'Historically important' category displays entities mentioned in the historic documents and having ties to the POK.
- Clicking on the circles representing the articles (make sure TapTool from the tool list is enabled) will display the full articles next to the plot in a pretext widget box.
- The entities shown in the plot are clipped after some letters to avoid text overlap. Hovering over the entities will display the full name of the entity and the sentiment score for the entity.
- Make sure you enable the 'WheelPan' tool from the tool list adjoining the plot. That will enable you to scroll through the timeline.
- For the top plot (articles for the 20th and 21st January 2014), if you scroll further ahead you will see a vertical blue line which marks the separation between the articles dated 20th and 21st



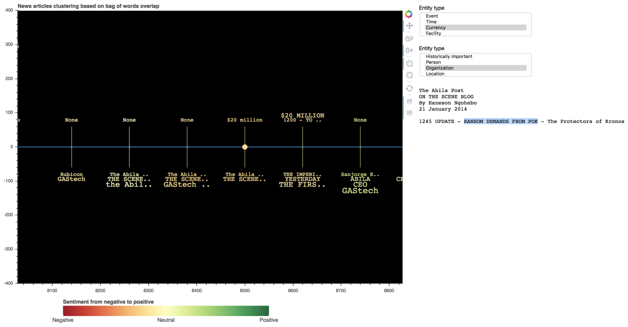
: the building. The employee, Edvard Vann, said that the police

Figure 2: Suspicious entities



GAStech employees made no report of a bomb threat to his agency.

Figure 3: Negative sentiment article



The Abila Post
ON THE SCENE BLOG
By Haneson Ngohebo
21 January 2014

1245 UPDATE - RANSOM DEMANDS FROM POK - The Protectors of Kronos

Figure 4: Article reporting the ransom demand from POK

- For the bottom time line plot, the articles are on the time line are in chronological order of their dates. Hovering over the articles in the bottom timeline displays the date of the article.

4.2.4 Analysis

The tool allowed us to visualize the entities in the articles which had high negative sentiment. We found that the POK was actually responsible for the disappearance of the GASTech company employees. We found a negative sentiment article describing a diversion created by the perpetrators in the form of false fire alarm which served as an opportunity to move the important people of GASTech to 'secure' location but instead they went missing, suggesting involvement of some GASTech employees. The entities under tag 'CURRENCY' gave us articles containing information about ransom demands by POK and a strong statement of purpose by a POK spokesperson Maha Salo. We also cross referenced the entities found from the articles with the historic documents and the employee records to find out mole/s inside GASTech, we present these entities as 'Historically Important' category on the timeline. The moles we found were Loreto Bodrogi, Isia Vann and Hennie Osvaldo, all three of them had ties to the POK families and were security employees at GASTech. We noticed that the name Vann appeared frequently with the POK and it also appeared in the list of entities from the historic documents. Further the visualization helped us to determine a timeline of events prior to the disappearance of the employees which we documented as a part of Homework 7 report.

4.3 GAStech Car GPS Tracker Animated Map

Please run `bokeh serve --show FP_car_track.py` from our submission directory.

4.3.1 Concept

In this tool, we wanted to observe two main things. What were the movement patterns of all employee cars leading up to the disappearances? Given a location, who was near that location the most on average over time? We felt the only way to see movement patterns was to obviously animate the cars on a map. We also needed some way for the user to specify a location of interest on the map and analytically compute who has been in close proximity with that location over time.

4.3.2 Data Setup

The gps data was given at a very granular level of time, so the first thing we did was sample from it at a frequency of minutes. This is enough to give a smooth animation of the cars movements. The map was constructed from shapefiles that describted Abila and Kronos using Pyshp. Besides this, no other prepossessing was needed.

4.3.3 Tool Description

The tool allows the user to play, pause, skip around the time before the disappearances and see by the minute, where each GAStech car was at that time. Locations are labelled based on their label in the shapefiles, to avoid clutter we only labelled "shapes" that had 3 vertices or greater. On top of this, we analytically calculate the time-cumulative proximity of each car to a user specified point (specified with lat-long sliders) and display the top 10 closest at the given time of the animation. This metric is defined for each car at time T as:

$$\text{cumulative distance} = \frac{\sum_{t=1}^{T} \text{euclidean\_dist}(\text{car\_location}[t], \text{user\_specified\_location})}{T} \tag{1}$$

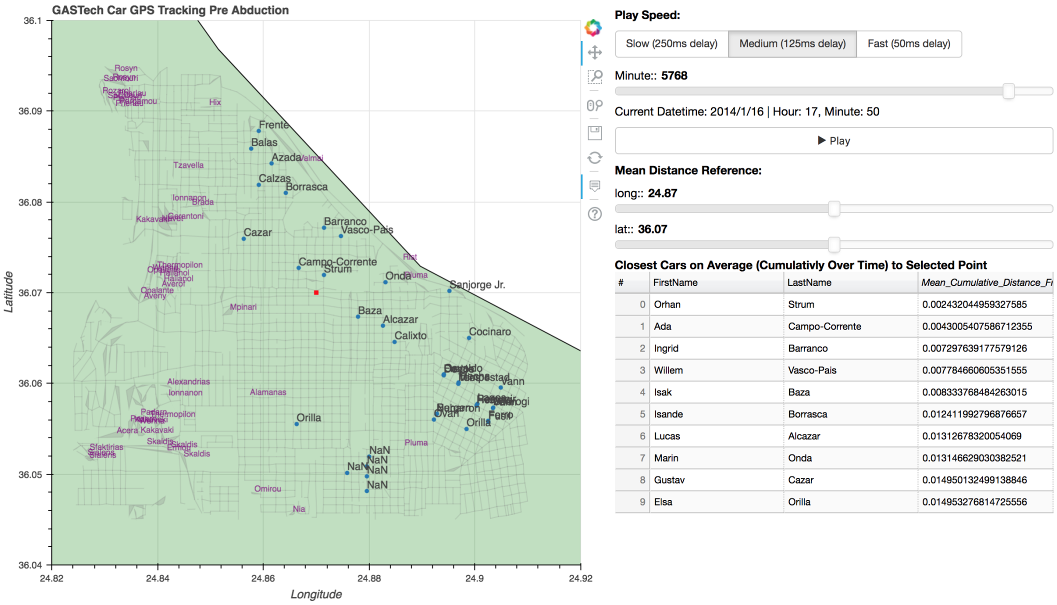
The full dashboard of this tool can be seen in fig. 5.

4.3.4 Analysis

We used this tool to follow the suspects from our email network and timeline tools. We were able to find a common meeting spot for a subset of the suspects on the north-east edge of town and verified that they met there more than any other employee analytically using our distance metric. One of these suspects was Isia Vann, someone we identified earlier as a possible supporter of POK (the group suspected of abducting GAStech employees).

This tool worked especially well for identifying who was in close proximity of a specific spot over time as the distance metric is specific, and ranked. There is no dispute or ambiguity as to who was actually closer to a location up to that point in time. This tool also generalizes to any kind of gps data in form of (datetime, lat, long) and given any shapefiles, can generate the map too.

Even though we could see the general patterns of car movements, we found it difficult to keep track of the movement patterns of cars without some kind of cumulative trail. For future improvements of this tool, we would consider adding a trail behind the cars as well as a filter for displaying only certain select cars at a time to avoid cluttering the screen with too many cars being displayed at once.



Minute:: **5768**

Current Datetime: 2014/1/16 | Hour: 17, Minute: 50

long:: **24.87**

lat:: **36.07**

Closest Cars on Average (Cumulativly Over Time) to Selected Point

#	FirstName	LastName	*Mean_Cumulative_Distance_Fi*
0	Orhan	Strum	0.002432044959327585
1	Ada	Campo-Corrente	0.0043005407586712355
2	Ingrid	Barranco	0.007297639177579126
3	Willem	Vasco-Pais	0.007784660605351555
4	Isak	Baza	0.008333768484263015
5	Isande	Borrasca	0.012411992796876657
6	Lucas	Alcazar	0.01312678320054069
7	Marin	Onda	0.013146629030382521
8	Gustav	Cazar	0.014950132499138846
9	Elsa	Orilla	0.014953276814725556

Figure 5: GAStech Car GPS animation, calculates and sorts the average distance of each car to a user specified point on the map of Abila, Kronos

4.4 Credit Card History Dashboard

Please run `bokeh serve --show FP_credit_card.py` from our submission directory

4.4.1 Concept

The goal of this visualization was to visualize the credit card usage patterns of the employees of GASTech. The questions asked here are who frequents a given place, who visits specific places at odd hours (anomalies) and who all are at the same place at the same time on the same day (detect groups of people meeting up at a place). To target the second question we create a scatter plot to visualize the credit card usage on different days. For the first question we plot a histogram of who frequents a place and how many times over the given two week period and for the third question we use a heat map to see what two people use their credit cards in the same hour on the same say at the same place.

4.4.2 Data Setup

We have credit card usage data for 56 people over two weeks. The data is quite clean thus easier to deal with. It includes names, places, amount and time stamp. The amount data hasnt been used in the visualization.

4.4.3 Tool Description

The visualization contains a scatter plot for the credit cards usage data of day vs. hour. This shows the daily patterns of credit card usage by the employees at the selected location. We have also colored the suspects (from the previous visualizations and the names appearing in the historical

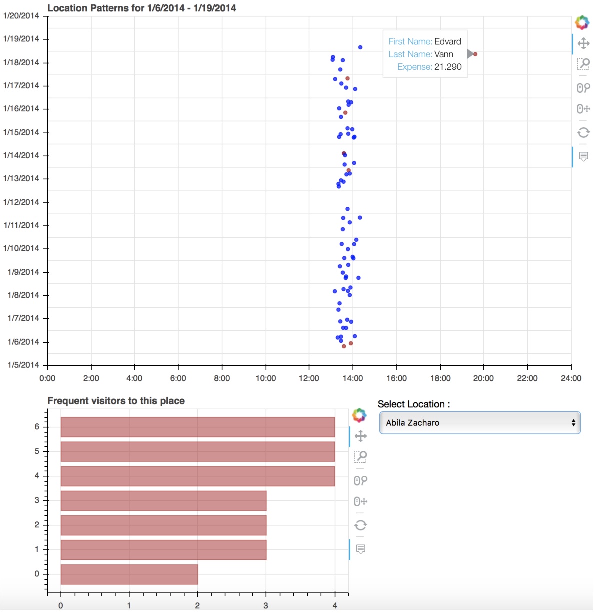


Figure 6: SSE with different k

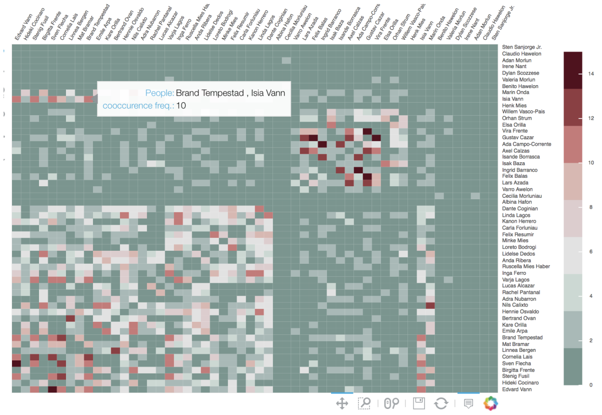


Figure 7: SSE with different k

documents) as red in the scatter plots to track their movements. The location can be selected by the selector drop down. It also changes the histogram of frequent visitors at the selected place. Hovering over the scatter plots shows the full name of the person and the amount spent as well as the timestamp. Hovering over the histogram bars gives the full names of that employee. The heatmap is colored according to the frequency of co occurrence of two people. Hovering over the heatmap blocks gives the number of times the two people were together at the same place at the same time.

4.4.4 Analysis

This tool allowed us to pin point people who visited places at odd hours, like Edvard Vann who visited the Abila Zacharo late in the evenings had ties to the Vann family of POK. We flagged the credit card activities of the people from a compiled suspicion list (people having ties to POK according to the historic documents and suspicious employees of GASTech identified from the news timeline) by coloring them as red in the scatter plot. The sleeper cell of POK that we identified in the Homework 7 report namely Vann, Bodrogi and Osvaldo are seen meeting together over 6-10 times in 14 days. This fact is also supported by the GPS tracking data which puts them together in the south east corner (Abila zacharo possibly).

5 References

- https://www.twilio.com/blog/2017/09/sentiment-analysis-python-messy-data-nltk.html
- https://www.kaggle.com/c/si650winter11/data
- http://www.cs.cornell.edu/people/pabo/movie-review-data/