

Final Report [VAST 2014](#)

Visual Analytics, 690V

[Sanjay Reddy Satti](#), [Aditya Agrawal](#)

When it comes to the visualization tools, we tried to make it as generic as possible. Almost all the tools will work even when the dataset is changed. Please read the report in collaboration with [video](#) (we specifically didn't include many visualizations in this report because of the space constraint. The video consists of all good visualizations. This report explains the thinking behind them and also various approaches which we tried to implement but failed).

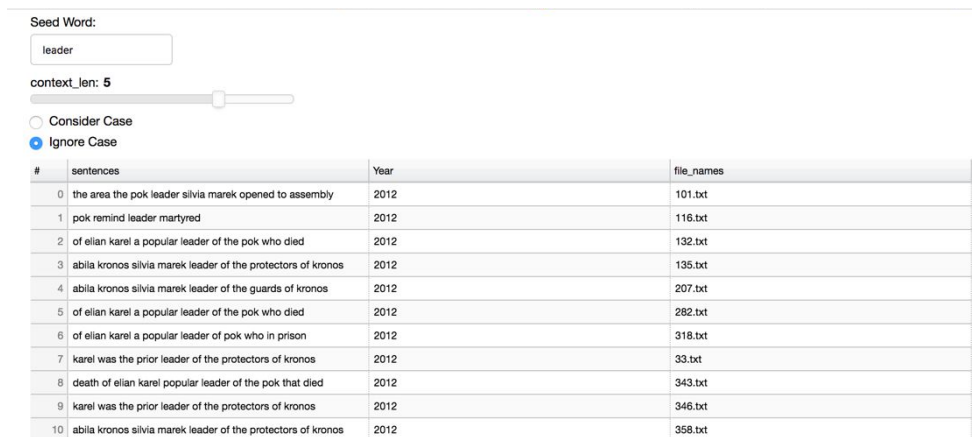
MC 1:

This deals with generating reports on text data.

Overview of Work done:

Our main focus for creating visualizations was on emails and news articles. The reasoning being that the other data files like resume and employee records were already in a structured-tabular form (querying on them directly seemed easier than creating another visualization). Focusing our NLP techniques on more free-form data like emails seemed a better approach.

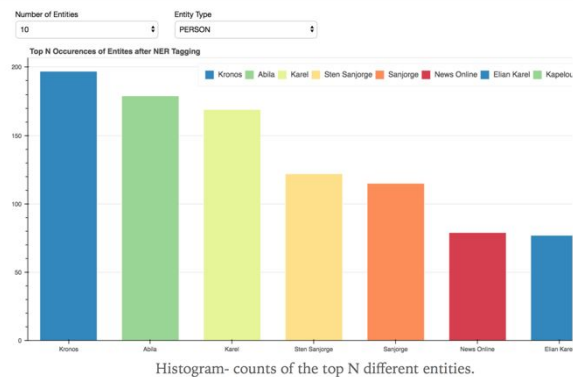
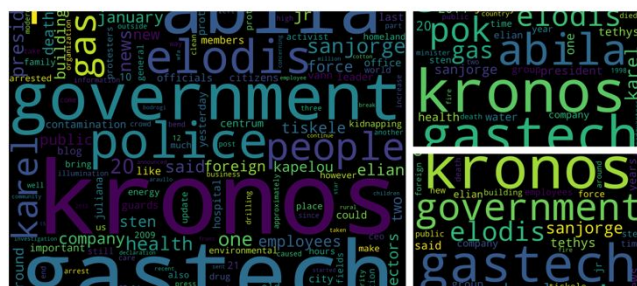
We first created a 'sentence sampler' which takes in a word as an input and returns the sentences in the emails which have them (Thus context in which the word is present is also obtained. The context length can be changed and also, other tools like 'case' of the word can be controlled).



#	sentences	Year	file_names
0	the area the pok leader silvia marek opened to assembly	2012	101.txt
1	pok remind leader martyred	2012	116.txt
2	of elian karel a popular leader of the pok who died	2012	132.txt
3	abila kronos silvia marek leader of the protectors of kronos	2012	135.txt
4	abila kronos silvia marek leader of the guards of kronos	2012	207.txt
5	of elian karel a popular leader of the pok who died	2012	282.txt
6	of elian karel a popular leader of pok who in prison	2012	318.txt
7	karel was the prior leader of the protectors of kronos	2012	33.txt
8	death of elian karel popular leader of the pok that died	2012	343.txt
9	karel was the prior leader of the protectors of kronos	2012	346.txt
10	abila kronos silvia marek leader of the protectors of kronos	2012	358.txt

We actually solved the whole of MC-1 using just this tool. But, what search phrase to search was a hurdle and therein we used Python libraries which gave us words of interest like names of people, organizations mentioned, locations specified etc. We displayed these common frequency words as interactive word clouds and histograms so that it becomes more apparent to the user.

We can control the number of words being displayed in our wordcloud- looking at 50 words for a sparser view, versus the top 500 occurrences for a detailed view. Additionally, the histogram also contains the same feature of selecting the type of entity we wish to look for and all the number of entities we wish to visualize.



Interesting Inferences/Results:

- As early as 1884, when GASTech came to 'Tethys' people were skeptical about it (One in particular: Professor 'Oskar Wertz' of University of Tethys about off-shore shale project). In fact, the government deemed that the environmental impact is unacceptable. (search: 'Tethys'). But since GASTech was not only allowed to establish there, but also thrive pushes forward the hypothesis that there are undeniable relationships between Govt. And the company.
- POK is an organization which started off as an environmental protection organization but slowly became a violent (possibly terrorist) one. (Backed up by the general tone of emails/articles of POK which become more and more radical as years progressed).
- Bodrogi was one of the leaders till 2001 and then hands over control to 2001 (search 'Bodrogi').
- Carmine Osvaldo was one of the other co-founders. (search: osvaldo). Interesting that, one of the security guards at GASTech is Hennie Osvaldo and is suspected.
- Elian Karel died under suspicious circumstances in 2009 and became a martyr for POK. This makes the organization push for more destructive (search 'Karel').
- Marek becomes the head after the passing Karel. Although in later years, his leadership is referred to as 'weak' (search: 'Marek'). Note that although leadership is considered weak, the actions suggest otherwise.
- The phrase 'drugs' and 'army' start coming up more and more frequently in the emails starting from 2011. (search: 'drugs')
- Asterian People's army mixes in with POK and makes more targeted and violent attacks. (search: 'army'). They are a paramilitary group.
- Julliana Vann has died (probably) because of cancer. Her (possible) relatives are Mandor Vann and Isia Vann. It is suspected that Mandor is the current real-force behind POK (and Marek is just a figurehead). Isia Vann is also an employee of GASTech.
- When 'Juliana' died in 'Elodis' in 1998, many reports claimed that it was because of the company. But nothing happened on that front (search: 'Elodis')
- In 2009 it was also reported that the company has not given sick people of Kronos any compensation. The Kronos people from Elodis in fact formed an alliance and pushed the government for reforms but to no avail. (search: 'Elodis')

Actual Implementation details:

- To run the sentence sampler... use: `bokeh serve --show seed_word.py` (The folder 'articles' from the the MC-1 [dataset](#) should be accessible by this. It contains all the emails).

- For the sentence sampler, we first went through all emails, extracted date using regular expressions (so that it can distinguish between both 12/18/2017 and 18th December, 2017) and stored the actual email in a dictionary (using date as key).
- Using 'nltk' library we parsed the words and put them in categories like 'Person', 'Organization' and also other categories like 'Positive Sentiment' or 'Negative'.
- For the wordcloud, we have used the wordcloud library in Python and for the entity histogram used NER features of the NLTK library. Both these visualizations can be rendered in their respective iPython notebooks. The notebooks should be place in the same folder as the data- 'nysk.xml'.
- To execute the names_places: bokeh serve --show names_places.py (should have access to the articles folder.)

Failed experiments and Possible improvements:

- Although we mentioned the people of interest as asked by the question, we still had to do lot of manual reading of the emails. The current algorithm mentions people (and the count of number of mentions gives us a heuristic of how important they are) but we still had to read through the emails to get a clearer picture. Maybe we can further automate it. Relationship recognizer ([Textual Entailment](#) and [dependency parsing](#)) is one thing which we wanted to use to combat this. Although we tried using it, we couldn't integrate into the system properly.
- We very much wanted to put up a visual timeline of the events. But it proved tougher than we expected to do so, in Bokeh. So, while going through interesting emails, we manually jotted the interesting dates.
- We have actually [done](#) a clustering of news articles for HW-6. It uses TF_IDF scores to connect the text articles and plots them on a graph (using PCA to reduce dimensions). But unfortunately, when we used the same on these emails we could not get a proper clustering. One possible reason could be that the number of emails are too small and also the text of an email is much too small to give a good score (compare that with the dataset which we used which had over 150K news articles). Also, in news articles, each article has an inherent topic associated with it (That is how news articles are structured). But here, all emails are either related to GASTech or POK. No more structuring is present and our clustering algorithm requires more.
- We found the profiles done by Tianjin university, about POK very interesting. The way they gave ratings for violence, no. Of unemployed etc. For each time period was very unique. Although qualitatively one can arrive at trends (like increasing violence) from the frequent words and tone of emails used, we could not figure out a way to quantitatively arrive at the ratings like the university submission. Also, for things like No. Of scholars and No. Of unemployed we could not find any data sources in the download zip.

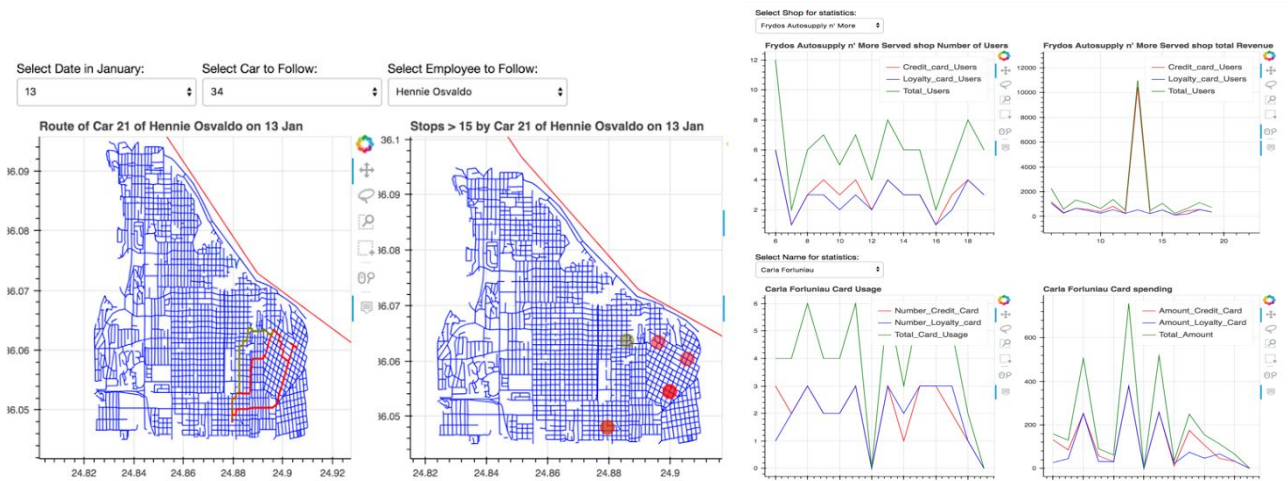
MC 2:

This deals with gps tracking of cars and credit card usage of employees.

Overview of Work done:

We first plotted the map of Abila (with roads) and gave the user the option of selecting which car/employee to follow on which day. The route is also color coded based on the time of the day. An accompanying graph just shows the places where an employee has stopped for more than 15

minutes. These might contain possible locations of interest for the police (like the residence of employees some of whom are suspected kidnappers).



Since most of them are employees, they follow a particular routine. During office days they drive to office (usually picking up coffee on the route), have lunch at a place and in the evening, go back to their homes. On weekends, they might go for shopping. (we tested these hypotheses by comparing these gps plots with the actual image of Abila. The image showed places like Coffee shop, GASTech company quarters etc).

The credit card usage also follows a particular trend. During weekends, the coffee shop revenues dip (as expected) and shopping malls rise (as expected).

Interesting Inferences/Results:

- Few of the security officers of GASTech are part of POK. Their routes are of particular importance. The suspected security officers like (Hennie Osvaldo, Isia Vann, Minke Mies, Loreto Bodrogi, Inga Ferro) have nearly identical routes. As the routes are also color coded, one can see that the time was late in the night.
- For weeks prior to January 20, the kidnappers have conducted surveillance of intended victims' houses. They have also practiced runs to safe houses from GASTech company (Southern part of Abila). Their routes remain more or less the same across various days.
- The money transactions happened using credit-cards. The shop was Frydo's. (This can be seen clearly, by the sudden spike in the spending's at the shop. That too on a weekend, a time when it is supposed to be calm). This is actually suspected to be a front for APA.
- Minke Mies (the security guard) actually stole the card of Lucas Alcazar. When the payment was done at Frydo's (by Minke), Lucas was away (The gps location shows he is somewhere else but the credit card was spent at Frydo's). It can be known that Minke is the culprit, because his car was at Frydo's.
- Two admins at GASTech, Carla Forluniau and Ruscella Mies Haber are also suspected to be co-conspirators because they also have a spike in spending on the same day where Frydo's took in lot of money.

Actual Implementation details:

- I used 'shapefile' library in Python to help with the plotting of the maps. [This](#) blog has more details. It needs to be installed for the program to be run.

- Download the MC-2 [data](#) . (Both the .py files should have access to .csv files and geospatial folder)
- To execute the gps plots: `bokeh serve --show gis_test.py`.
- To execute the credit_card plots: `bokeh serve --show card_pay.py`.

Failed experiments and Possible improvements:

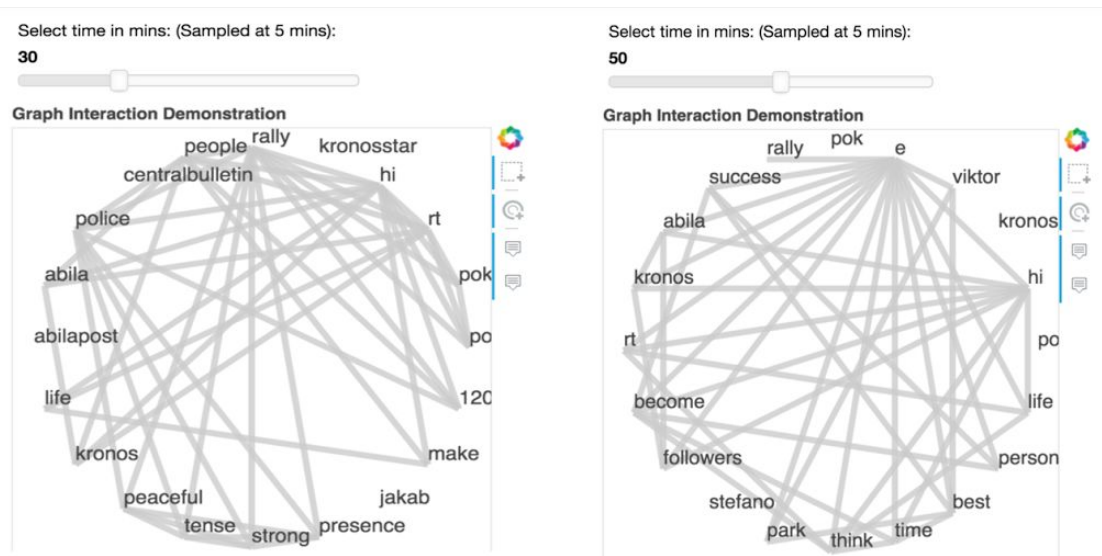
- As mentioned there are safe houses used by the kidnappers. One possible improvement was, given routes and stops across days, finding common points. (Say for example, it is known that a person stopped at a place for 5 days. In the current system, one has to manually change the day and test the hypothesis that, yes indeed the person is stopping at the same location 5 times and it could be a potential safe house). This was tough to implement, because the points of stopping are not exact. Thus, taking the intersection directly proved futile. One possible workaround could be that maybe we can include a distance threshold and if it falls within that we can combine them as the same. This automates the process of finding the safe houses (by analysing stopping locations and times across the week).
- One easier addition was adding filtering by time (Not just the date). But the program was actually crashing on a 2GB RAM machine. So, we had to leave it just at the date. (Just the date itself was taking lot of processing time because Bokeh had to draw multi-lines on the map).
- Credit Card [Fraud detection](#) is actually a huge topic in Machine Learning (specifically, Anomaly detection). Infact there was an interesting [paper](#) using Convolution Neural Networks. Agreed that it might not be needed for this dataset, but we wanted to keep it as generic as possible.
- [Central South University](#)'s visualizations served as an inspiration for our project. But after creating the tools, they have done a LOT of experiments to test their various hypotheses. Time was one resource which we didn't have enough of on our side.
- The [City of London](#)'s visualizations require a special mention. They have done 2 unique things. First, after a preliminary analysis of Abila map, they've gotten places like 'museum', 'office', 'meal' etc. And removed the background map from the analysis. Then, they modelled the routes as nodes between graph. This was a very interesting approach (We have used a similar network graph in MC-3. But we couldn't integrate into MC-2). Second innovation, was 3D visualizations. This works wonders for such a problem. Take for example an employee going to office and then coming back home on a same route. In our approach, they will overlap (although colors would be different as timestamps are different). But by using a 3D visualization, one can easily differentiate the routes. This method also very succinctly sends the user a summary of the routes. Although we looked at [3D maps](#) in bokeh, we could not integrate it with gps coordinates.
- Another failed experiment was counting the number of distinct cars passing through a road. This basically works in the reverse approach of our current method. Instead of following cars along various roads, we thought what if we fixed roads/shops and looked at distinct cars passing/stopping through/near it. As an example, say someone is conducting surveillance near an employee house by staying there at night. In this approach, when we click on the employee house, it would give us the IDs of cars near it (at various times). [One way of implementation is Euclidean distance between house coordinates and the GPS coordinates of the car.] This visualization (if made) would make narrowing suspects MUCH easier.

MC3 :

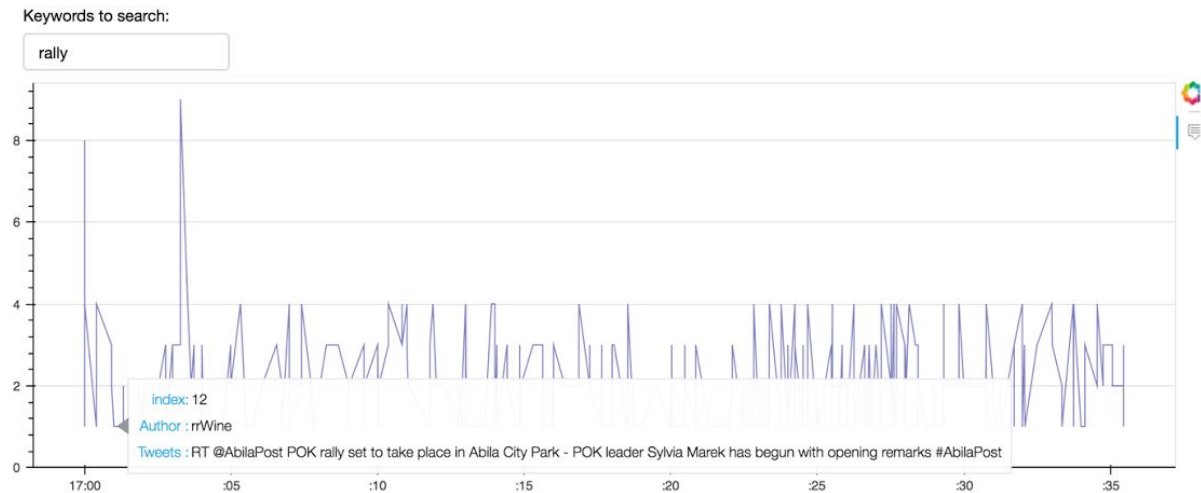
This deals with supposed streaming micro-blogging data.

Overview of Work Done:

We have first plotted a co-occurrence graph between the most frequently occurred words for all the micro-blogs sampled at 5 minutes as we just have data available for 1.5 hours / 90 minutes. The nodes in the graph show the top most words which have been posted in those 5 minutes and the edges between these nodes displays a co-occurrence relation between these words, i.e.- these 2 words occur frequently in the blog-posts for those 5 minutes. We wanted to get an idea about the events that we were happening during that 1.5 hour and wanted to try figuring that out using the most frequent words relation displayed using this network graph.



Further, we created an interactive line graph which displays the trends of occurrence for an input word for the duration of the analysis. For a given word, we can look at when it had a higher occurrence and then hovering over the line displays the micro-blog posted for that word, the no. of such tweets and the author's name.



Interesting Inferences/Results:

- We observed that some words have co-occurred alot- rally, Sylvia, lucio, peaceful, police, Abila City Park.
- The most obvious word popping was rally and we started looking up at the tweets mentioning the word rally and we realized that - leader of POK Sylvia Marek was giving an opening remark to start the POK Rally at the Abila City Park. This also covered the co-occurrence of Sylvia and rally which happened a lot during our visualization.
- Further we investigated the co-occurrence of 'Lucio' as a prominent word in our network graph and then looked at with the same in our trends graph for looking at the tweets. Looking at the tweets mentioning Lucio we learnt that special guest Dr. Newman was invited to give a talk with Prof. Stefano and Lucio Jacob, and the band Victor-E was invited to play music.
- Additionally, we can also find there were more than 1000 people attending the rally when we study the tweets mentioning the keyword 'rally' in our trends graph.

Actual Implementation Details:

- We have exploited the networkX graph integration with our Bokeh visualization interface for rendering the co-occurrence graph.
- Additionally, the folder named MC3 Data-containing all the data available from MC3 Challenge- should be in the same folder and the iPython data can handle the rest of the input of the data and one can render both visualizations in the same iPython notebook.
- All these tools are generic in nature and would work for similar data.
- We did not have access to 2 more chunks of the same and size and thus, could not solve this mini-challenge entirely and could not report on other events of interest that happened during that duration.

Failed experiments and Possible improvements:

- We wanted to visualize the tweets popping up at different times on a map using the latitude and longitude columns in the data, but for the 1.5 hour data these columns were not populated and thus we could not have plotted this data on a map of Abila. We are assuming the rest of chunks could have possibly had this data, but it was not available at the MC3 repository.

- We wanted to make the network graph more interactive (and colorful). Like for example, by hovering over the label, display more information (probably meta-info like 'sentiment' of the tweet, age and the gender of the [tweet](#) or answer other NLP tasks like Is it a sarcastic tweet or not). [In fact in our NLP project, we worked on age and gender prediction, but we developed that algorithm assuming we have much higher volume of tweets].
- The current system is actually a bit slow (because of the Network Graph in Bokeh). Although the approach we used will definitely be scalable for bigger tweet datasets also, we doubt the interaction with graph would be seamless.

Conclusion:

We have analysed the VAST 2014 challenge with simplistic, yet effectual visualizations which capture the essence of the data presented to us and have tried to approach the problem in a time-effective and productive way, which has yielded conclusive results for the majority of the challenge. Our approach was to put up the visualizations which would be simple to understand for the user, yet would be able to convey as much information as possible within a time bound manner.