2014 VAST Challenge

Background

In the roughly twenty years that Tethys-based GAStech has been operating a natural gas production site in the island country of Kronos, it has produced remarkable profits and developed strong relationships with the government of Kronos. However, GAStech has not been as successful in demonstrating environmental stewardship.In January, 2014, the leaders of GAStech are celebrating their new-found fortune as a result of the initial public offering of their very successful company. In the midst of this celebration, several employees of GAStech go missing. An organization known as the Protectors of Kronos (POK) is suspected in the disappearance, but things may not be what they seem.

Mini-Challenge 1

(1) Question description

This min-challenge focuses on the disappearance at GASTech. Specifically, we finished the first subtask which aims at detecting the leaders, extended networks of POK and connections between the POK and GAStech.

(2) Search Tool

Essentially, this task tries to find out the relations between different kinds of entities such as Person and Organizations. Hence, we build an interactive search tool for exploring different kinds of relations.

Figure 1. Search Tool for Detecting Entities and Relations

This tool can be used to search with keyword and entity of interest. Note the candidate list is generated with Named Entity Recognition tool provided by Spacy.

1. For detecting potential leaders, we can just search with keyword ‘leader’. All the sentences that contain this keyword will be demonstrated on the right side. Keyword itself is highlighted with red color, entities in the retrieved sentences are highlighted with green.
2. For detecting extended network of POK, we can search with selecting candidate entities listed. The targeted entity is highlighted with yellow while other entities occur in the same sentence are highlighted with light green. Note these highlights reveals key information of relations between different entities.

Figure 1 shows a concrete example of searching with keyword ‘leader’ and targeted entity ‘Henk Bodrogi’. Firstly, we confirm that Henk was one of the leaders in early years. Also, we can easily observe other potential leaders such as Karel which is highlighted in the last sentence. Following this strategy, we can trace back the relation network underlying numerous articles.

Mini-Challenge 2

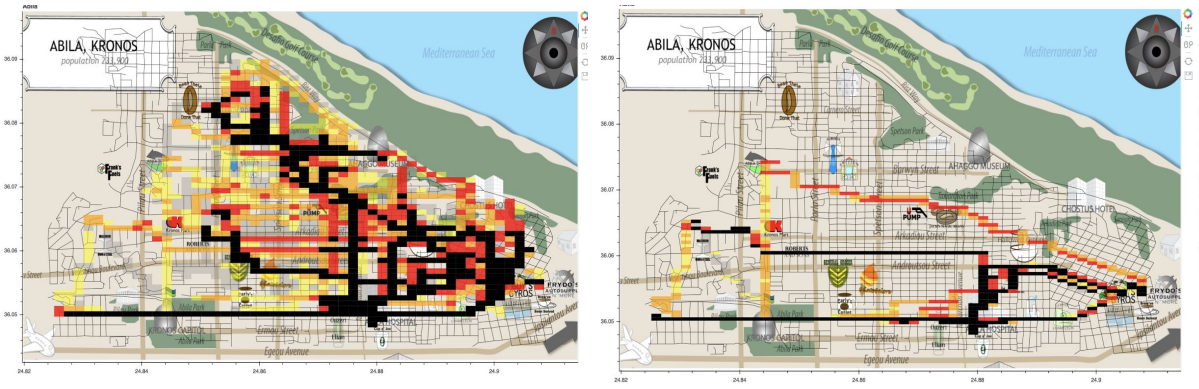
(1) Question description

Abila GAStech company provides employees with company cars for the personal and business use while provides those who do not have company cars with trucks for only business use but no personal use. Both company cars and trucks are equipped with the geospatial tracking software and thus can be tracked periodically as long as they are moving. However, the tracking data is only available for the two weeks leading up to the disappearance but not for the day that GAStech employees went missing. Our goal is to make sense of the tracking data and analyze movements to get the common daily routine and identify suspicious behavior.

(2) Timestamp movements

In order to make a good sense of GASTech employees’ movements, we set the given tourist map as background and draw multiple lines on it according to the geo information from ABILA.shp file. Then we read the gps information which includes GASTech employee’s locations in timestamp over the last 14 days. To analyze the frequency of employees’ movement, we calculate the occurrence of each grid that 35 GASTech employees appear and represent them with five color regions, in which the black ones represent the most frequency places, and the next four are red ones with fewer frequency, orange ones, yellow ones and grey ones with just little appearance.

Employees in GASTech can be classified into two types---general staffs and truck drivers. Figure 2 shows the frequency of movements plots separately. According to the pattern in the plot, the city center is GAStech while the most travelled routes are the one that leads to the airport zone and the one that leads to the port of Abila specially Ipsilantou Avenue. Furthermore, general staffs and truck drivers have different daily routines and seems to have very little contact outside of GAStech.



(a) general staffs movements (b) truck drivers movements

Figure 2. Frequency of employees‘ movements.

(3) Time period control

The function of the time period control is powerful, in which we can select the movements of employees in just one day or multiple days while can also choose different day time periods by hours, minutes and even seconds. For example, we can choose the date from 01/08/2014 to 01/10/2014 with day time period from 13:34:29 to 18:05:31. As a result, we can analyze the daily routine and activities effectively.

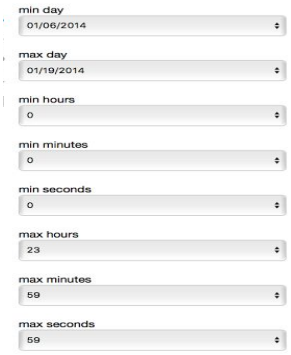
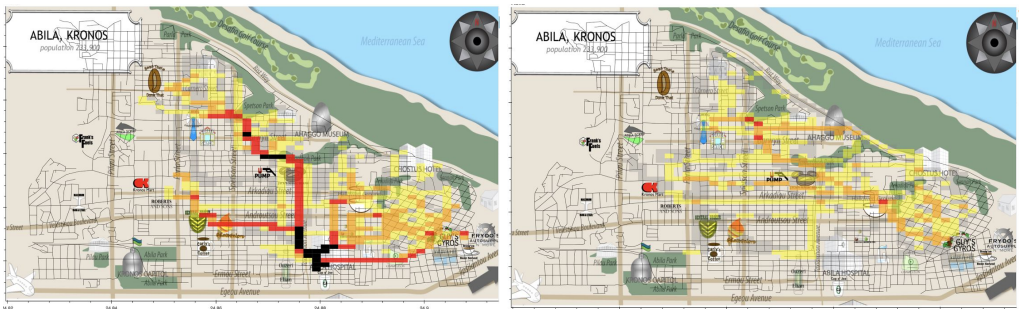


Figure 3. Time period control

(4) Common daily routine for general staffs

The daily routine can also be classified into two types---weekdays and weekends as shown in figure 4. There are little activities for general staffs on weekends as 4(a) while complicated activities on weekdays as 4(b) and thus we need to analyze them in detail.

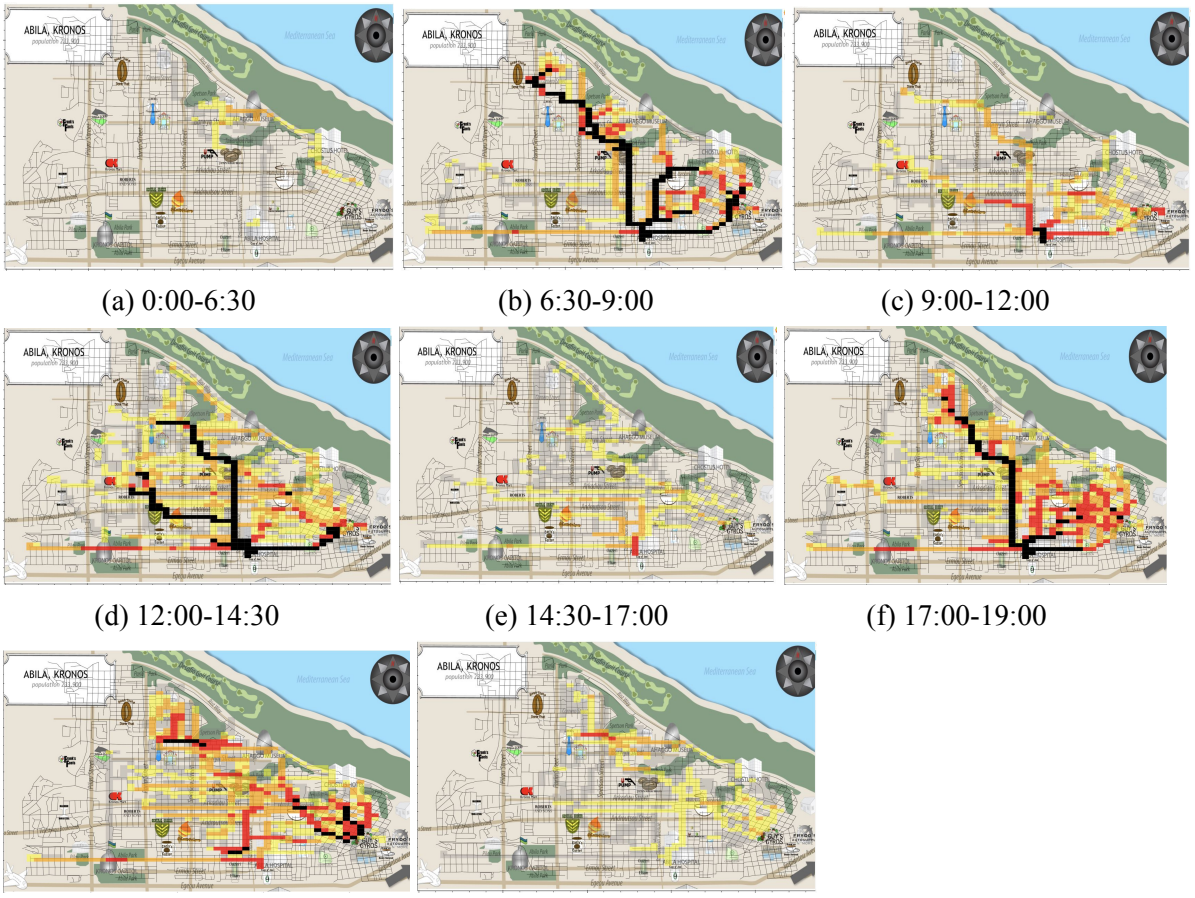


(a) weeksdays (b)weekends

Figure 4. common daily routine for general staffs

Control the day time periods by hours and we can get the common daily routine for general staffs on weekdays. According to figure, a general staff sleep in home during until 6:30 as 5(a) and start their morning by first stopping by a coffee shop during 6:30-7:30 and arriving at GAStech before 9:00am as 5(b).Then they work at GAStech till 12:00 as 5(c) and leave for lunch. Lunchtime is usually between 12:00 and 14:30, which includes the time to travel to and from the restaurant and to eat as 5(d). In the afternoon, they work from 14:30 to 17:00 as 5(e). Then they go back home or go to

restaurant for dinner as 5(f). During the evening, they will visit shops or other restaurants. In the end, most of them go back home before 21:00 as 5(h).



(a) 0:00-6:30 (b) 6:30-9:00 (c) 9:00-12:00

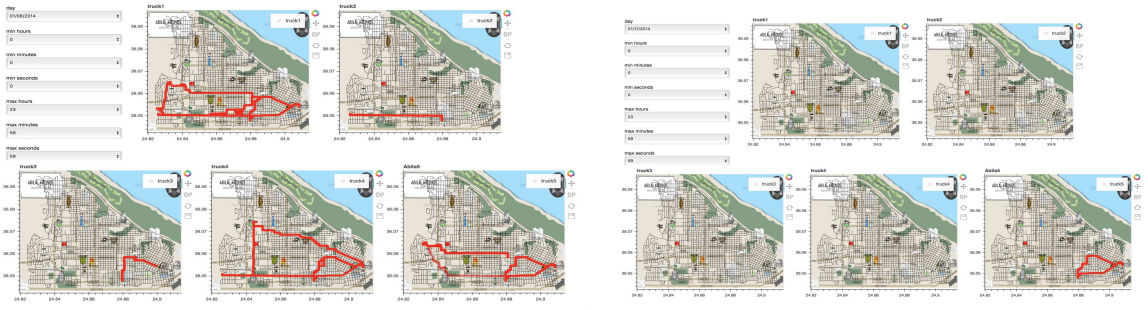
(d) 12:00-14:30 (e) 14:30-17:00 (f) 17:00-19:00

(g) 19:00-21:00 (h) 21:00-23:59:59

Figure 5. common daily routine for general staffs on weekdays.

(5) Common daily routine for truck drivers

In this section, we need to analyze the common daily routine for truck drivers one by one due to the limited datasets. We found that there is no record for truck drivers on weekends and several important suspicious behaviors. For example, lots of activities on 01/07/2014 while only one truck on 01/17/2014.



(a) lots of activities on 01/07/2014 (b) only one truck on 01/17/2014

Figure 6. Common daily routine for truck driver.

Mini-Challenge 3

(1) Question description

In MC3, we were asked to analyze streaming data to identify major events or suspicious activities which might help find the missing employees of GAStech company. The main task we have to solve is: "how to identify major events or suspicious activities in streaming data? ". We develop three tools to help us observe and investigate.

- Hashtag Analysis
- Sentiment Analysis
- Tweet Clustering

Note: Since we only got limited dataset (first segment), our results were based on the first 1.5 hour of the streaming data.

(2) Hashtag Analysis

Hashtags often provide rich information of the tweets, such as members, locations, or organizations. To detect major events, we want to first observe the distribution of hashtags over time to find out if there is any specific hashtag that appears frequently within a short period of time. Thus, we draw time-series plots to show the distribution of hashtag count interactively. We can select the time period we want to observe or any specific hashtag in the tweets.

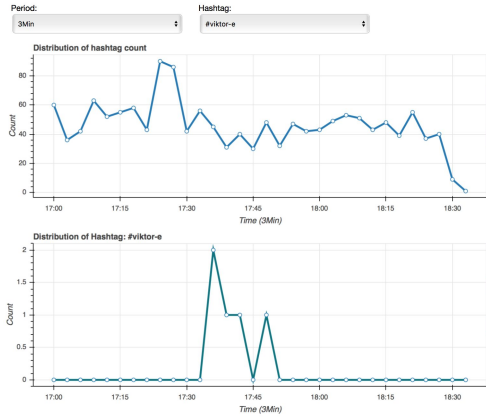


Figure 7. Distribution of hashtags.

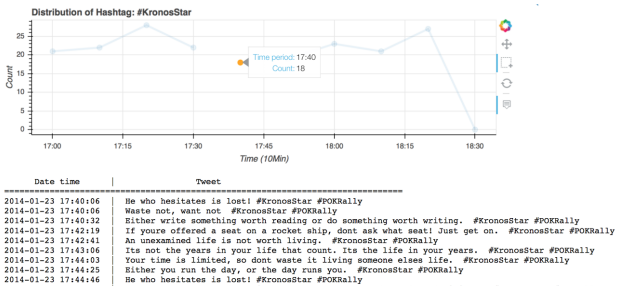


Figure 8. Show the tweets which contains the hashtag.

Given that we are dealing with streaming data and we were asked to send the results within limited time (MC3.1), we want to develop efficient tools that can deal with time-dependent data. Instead of looking into all the tweets to detect suspicious events, we want to use hashtags as an indexing tool to search for tweets that we are interested in. Thus, we embed the function of searching into our tool. By selecting a period of time on the time-series plot, the tweets which contain the specific hashtag will show interactively in our pretext area.

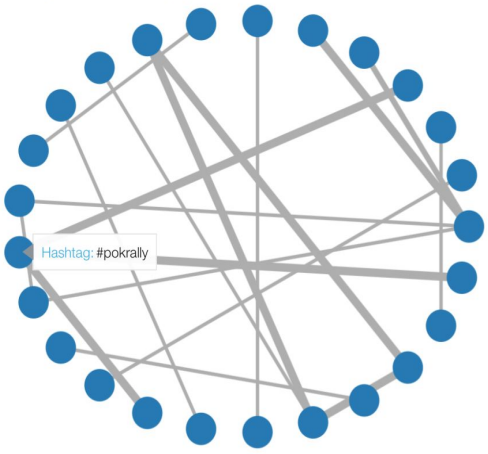


Figure 9. The weighted co-occurrence plot of hashtags.

Moreover, to better understand the relations between hashtags, we draw a weighted, co-occurrence plot to visualize their connections in the tweets. The weight is based on the count of the occurrence of a single hashtag pair. Then, we run PageRank algorithm to find the most influential hashtags and learn their relations. The plot help us understand which hashtags are correlated and provide us the target keywords to investigate.

(3) Sentiment Analysis

Secondly, we perform sentiment analysis on the tweets and classify them into positive/negative ones. We show the distribution of the sentiments to detect the occurrence of suspicious events. The purpose of this approach is that if any major event such as terrorism or bomb explosion happened, we expect the number of negative tweets will increase dramatically within a short period of time. However, in this streaming dataset, we did not observe what we are expecting.

Figure 10. The distribution of positive/negative tweets.

Figure 11. Show the selected tweets.

(4) Tweet Clustering

In the third part, we run clustering algorithm to cluster tweets into groups. The basic strategy to detect events in tweets is to understand what's happening in a period of time from tweets. Hence, we develop an interactive visualization that consists of selecting time window and clustering algorithms to visualize these tweets. Other visualization techniques like hover tool, slider to select number of clusters are also employed here.

We first convert each tweet to a Bag-of-Words representation. Then, we apply K-Means clustering algorithm to group tweets into different event. To visualize the high-dimensional vectors, t-sne is used to project the data to 2D space so that they can be easily plotted.

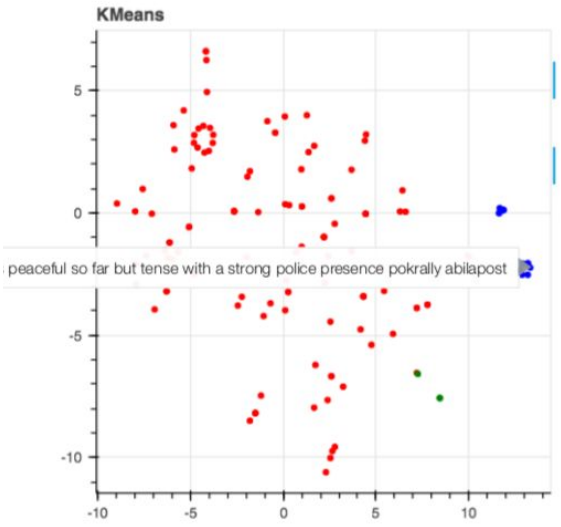


Figure 12. The projected vectors in 2D space.

Figure 13. A real-time word cloud plot.

A slider is used for interactively tuning the number of clusters. Hover tool is also employed such that we can directly see the tweet that each plot represents. Word cloud is also added to capture some keywords. Note that every component above only shows the tweets that were sent in a specific time range, which is determined by the time value we select. The purpose of this design is that we can easily switch between different time windows since events often happen in a particular time period.

We found that from 17:00, POK rally took place at Abila City Park. It could be easily detected through the word cloud since these words are frequently used since that time.

The clustering helps us focus on the tweets that talk about pok rally. And we can use the zoom tool to look into each cluster in detail. For example, as shown in the image above, we found an event that this rally leads to a heavy police presence. From the negative tweets on the right, we can also find that people are also worried about potential security risks.

As we push forward the time window, we can also easily detect the events that happen in the rally, such as introducing guests and music show. The rally remains peaceful till the end without unexpected events occurring.

Conclusion

- In mini-challenge 1, we successfully detect the five leaders of POK, Henk Bodrogi, Elian Karel, Carmine Osvaldo, Silvia Marek and Mandor Vann. Also, extended networks of POK such as APA, WFA, SOW are detected.

- In mini-challenge 2, we make sense of the tracking data and analyze the movements. As a result, we get the common daily routine both for employees, among which we divide them into weekdays and weekend and analyze routine of weekdays in detail, and truck drivers among which we analyze them one by one and identify several suspicious behaviors.

- In mini-challenge 3, we proposed three approaches to tackle the problem of major events identification in streaming data. We first analyzed the distribution of hashtags over time to find the important hashtags, and viewed them as keywords for efficient searching of suspicious activities. These hashtags provide us the meta information of the events. Then, we analyzed the sentiments of the tweets and applied K-means algorithm to cluster them into groups. These groups provide us a sketch of where, when and how the events happened during the first 1.5 hours.