Decoding the Curious Case of Missing Employees

Joyce Woon Shi Hui
Singapore Management University
joyce.woon.2020@mitb.smu.edu.sg

Lang Shuang
Singapore Management University
shuang.lang.2020@mitb.smu.edu.sg

Vertika Poddar
Singapore Management University
vertikap.2020@mitb.smus.edu.sg

ABSTRACT

This project was developed using R and Shiny packages with text data from multiple sources. The aim is to help identify conflicts and correlations in large volume of text data, detect patterns of group of people, recognize the most popular topics and locations to find out the high-risk areas to focus on. Hence, the local police department can discover the development of events and find suspicious people in the events. This project was developed in response to the VAST Challenge 2021.

1. INTRODUCTION

The VAST Challenge 2021 describes a hypothetical scenario where some of the employees of a fictitious organization, GAStech have gone missing on 23 Jan 2014 (Thursday). It is suspected that Protectors of Kronos (POK), an environmental activist group might be related to this incident. Dataset given in this case contains a collection of microblog records and text transcripts of emergency calls by local public service departments from the days surrounding the disappearance. This challenge requires identification of suspicious activities and behaviour hidden in data. In addition, it is also required to evaluate the changing levels of risk to the public. Our team has developed a web-based visual analytics system to analyse microblog and call centre messages exchanged on the event evening. It also provides interactive methods to analyse the popular locations and the connection among the employees of GAStech.

2. DATASET

This study uses selected datasets given by VAST 2021 Mini-Challenge 1, 2 & 3. The datasets provide information containing 4,063 records of microblog and call centre data, credit and debit card transaction location and timestamp (cc_data.csv) and car assignment data which contains 44 vehicle ids with assigned employee names and employment type and title (car_assignemnts.csv). In addition, email headers dataset was used which contains the subjects of the emails exchanged between the employees of the company. As the data volume is large, we will therefore focus on the transcripts from the call centre, employment relations as well as timestamp and location of the card transactions and the email headers in the following three aspects: a) Word Cloud b) Network Graph c) Location Bar Chart

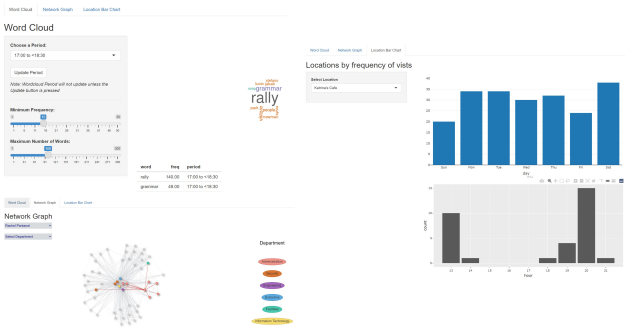

Figure 1: Web interface for our VAST Challenge solution. The interface is split into 3 parts: word cloud, network graph and location bar chart

3. METHODOLOGY DESISGN FRAMEWORK

To reveal the change of risk level and the riskiest location where the event took place, we need to investigate important information using files from different sources. The interactive web-based application was developed using R and Shiny packages for interfacing with a backend server for the visualization frontend. It serves as a quick overview tool of the original data which allows us to focus only on related period and locations. Given the nature of our main dataset, frequency analysis would be appropriate for our study. The visualization is split up into three tabs as seen in Figure 1: a word cloud, a network graph and a location bar chart.

3.1 Word Cloud

The word cloud tab displays the word cloud from call center messages and microblog messages, not including spam data which was removed during data pre-processing, filtered by the selected period and frequency of appearance. To find out top words used in the messages exchanged in the evening on 23 Jan 2014 (Thursday) from 5:30pm to 23:34pm, four options were used to allow users to narrow down the period: 17:00 to <18:30, 18:30 to <20:00, 20:00 to 21:34, and All

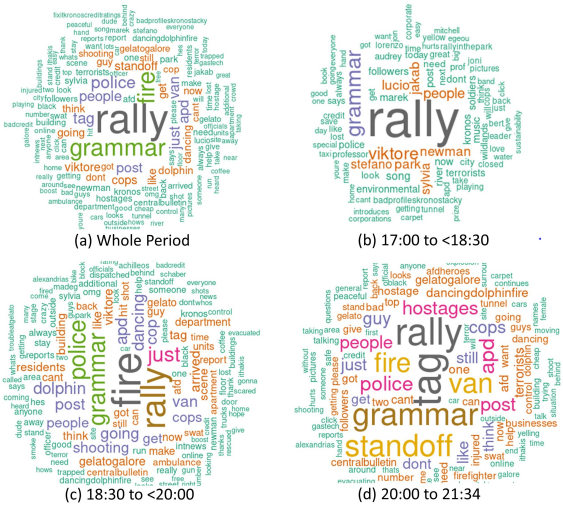

Figure 2: Word cloud for 23 Jan 2014

Periods. In addition, users can specify the minimum frequency of words and maximum number of words appeared which allow users to find out if risk is happening and what kind of risk is happening.

3.2 Network Graph

The network graph provides the relations of email communications between GAStech employees. By selecting the name of the employee from the dropdown list, his/her email communications with other employees are shown. By selecting the department, all the employees of the department get highlighted. We can zoom in to know the names of the employees.

3.3 Location Bar Chart

The location bar chart tab displays the location by frequency of visits and the day and time of the visits when zoom in. It accumulates the total number of transactions made by visitors between 6 Jan 2014 to 19 Jan 2014.

4. FINDINGS AND DISCUSSIONS

We were asked to distinguish meaningful events of interest in the Challenge. For Word Cloud the main action of discovery was to select the period to find out the frequently used words in both the microblog and the call center messages. In each case the Shiny app we developed helped immensely in determining events. For each case the main action of discovery was looking at their occurrence in both the microblog and the call center messages.

By observing the word cloud of the whole period, most frequently mentioned word is "rally" as seen in Figure 2 (a). The riskiest period was from 18:30 to 20:00 with the highest frequency words being "Fire." See Figure 2 (c). In the last period (20:00 to 21:34), we found that even though there was the word "Fire" in large frequencies, it is less frequent than before. There is an increasing trend of mentioning "Police" and "Cop" after 20:00 which indicate the risky event already took place in Figure 2 (d). Less risky words are higher in frequency now, including the word "Standoff" which shows that a response to the risk has occurred.

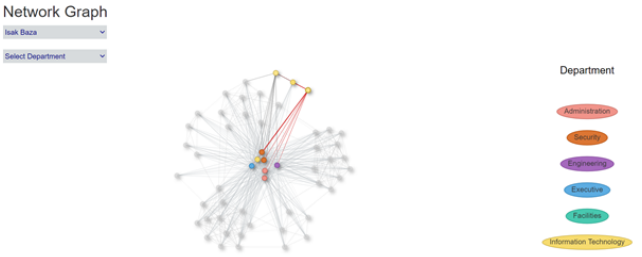

Figure 3: Network diagram of GAStech employees

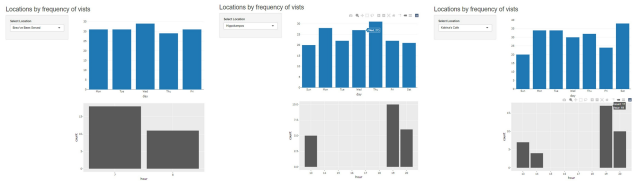

Figure 4: Location bar chart of Brew've Been Served, Hippokampos and Katrina's Café

This implies that the riskiest period is 18:30 to 20:00 followed by 20:00 to 21:34 and then 17:00 to 18:30 on the event evening on 23 Jan 2014 To better visualize the events, we also use network diagram to show the relations of GAStech employees and their department type and titles (as shown in Figure 3). After we select the name, the dot representing the employee is highlighted along with those employees with whom emails were exchanged. We can easily mouse over the corresponding bubbles, and then all the related bubbles will be highlighted. This graphs helps us to get the name of the employees with whom maximum email communications was done.

Brew've Been Served is the most popular shop among GAStech employees especially during weekday 7-8 am where the employees go for morning coffee. Katrina's Café and Hippokampos are the top two most visited shops for lunch and dinner (Figure 4).

5. CONCLUSION & FUTURE WORK

None of the above visualizations on its own was sufficient to answer the questions in VAST Challenge. Instead, we used information from combinations of the Shiny apps to gain a better understanding of the data. We did this by using zoom in and out the location and frequency diagram to visually organize our findings from the various visualizations together. There were some improvements that we could work on in the future including: Linking up Word Cloud and Location Popularity data, for example, Location Popularity on a certain date (23 January 2014) when the risk occurred to find out exactly how many people were in the risks areas and of course, how many people were affected by the incident at the various periods of time. Another improvement would be to possibly add a frequency element to the network graph in order to find which users are the closest together. If we can map these users to a time element, we can possibly find out which people are involved in the incident and provide a more focused response.

6. REFERENCES

M. Whiting et al., "VAST challenge 2014: The Kronos incident," 2014 IEEE Conference on Visual Analytics Science and Technology (VAST), 2014, pp. 295-300, doi: 10.1109/VAST.2014.7042536.

, Ü. den, & , N. (2019, June 21). Interactive network visualization with r. STATWORX. https://www.statworx.com/ch/blog/interactive-network-visualization-with-r/.

---

references: - id: 1 title: VAST challenge 2014: The Kronos incident author: - family: Whiting et al. given: M type: article-journal DOI: 10.1109/VAST.2014.7042536. page: 295-300 publisher: IEEE issued: year: 2014 month: 9 - id: 2 title: Interactive network visualization with r. STATWORX. author: - family: Autor given: Üden - family: Junker given: Niklas URL: 'https://www.statworx.com/ch/blog/interactive-network-visualization-with-r/' issued:article year: 2019 month: 7 ...