A Visual Investigation of the Kronos Incident

Connie XIA Yi Jing
Singapore Management University
connie.xia.2020@mitb.smu.edu.sg

Nikitha BANDA
Singapore Management University
nikithab.2020@mitb.smu.edu.sg

TAN Kar Yee
Singapore Management University
karyee.tan.2020@mitb.smu.edu.sg

ABSTRACT

This is the abstract.

It consists of two paragraphs.

1. INTRODUCTION

The fictitious Kronos Incident saw the disappearance of several employees from the Tethys-based GASTech in January 2014 after a successful initial public offering (IPO) of the company. Given that GASTech has not been very environmentally friendly in its operations of a natural gas production site in the island country of Kronos, it was suspected that a Kronos-based organisation (POK) is involved in the disappearance of the employees, as a form of retaliation. In order to have a better idea on what exactly transpired to lead to the vanishing of the GASTech employees, we will be applying visual analytical techniques on the datasets provided.

This study will be handling visualisations on newspaper articles, employee records and emails, call center reports and microblog tweets before structuring them into an interactive web application. Users can then investigate the application and understand more about GASTech's reputation. Furthermore, one can navigate around the app to find out how certain events unfolded on the incident day itself.

2. MOTIVATION AND OBJECTIVES

The motivation behind this study is to look into analytical techniques to visualise large chunks of text data effectively using R Studio. By doing so, we are able to better understand the relationships among people and organisations of importance, as well as see how multiple events of high consequences unfolded in Abila on the incident day.

This interactive Shiny app aims to provide information on:

1. Media portrayal of GASTech over the years
2. Relationships among GASTech, POK, the APA and Government
3. Meaningful event reports during the incident day
4. Risks identified during the incident day and their corresponding locations

3. REVIEW & CRITICS OF PAST WORKS

This study is based on the VAST Challenge 2021, which in turn is adapted from a similar VAST Challenge in 2014. Literature review is conducted on the previous VAST Challenge 2014 submissions to look at the analytical techniques used to solve the challenge back then, even though the exact questions were slightly different. While useful, some of the techniques adopted have certain areas that can be further improved.

3.1 Text Visualisations

A study conducted by Peking University (2014) on Mini Challenge 1 presented their text analysis in a form of a timeline to showcase different events occurring between January 20 – 21. Articles in the form of text boxes were layered over the timeline for comparison. While it showcased all the news reporting of different events occurring over the two-day time period, it might be difficult for a user to interpret the main concepts of those articles. Hence, a better alternative might be to utilise a word cloud function to pull out key words of the articles for view and interpretation. In addition, interactive comparisons of different newsgroups can also be performed, giving the user flexibility to choose the newsgroups they are interested in to view and evaluate.

While word clouds are generally useful in identifying topic content for a broad overview as shown in the study performed by Tianjing University (2014) on Mini Challenge 3, their results might be less consistent and harder to make sense of due to the presence of spam data. Hence, to be able to distinguish important events from typical chatter, TF-IDF would be a better statistical tool to use.

3.2 Network Graphs

Network graphs are a good visualisation tool to establish the relationships between different parties of interest. By and large, network graphs would be densely populated with nodes and edges if there are numerous parties involved. Yet, this brings about an issue of overcrowding and overlaps of texts, making the entire visualisation looks cluttered, as seen in Fig. x.

One way to overcome this issue of cluttering will be to divide the network graph into sub graphs. When the graph is divided, the density of the visualisation will be reduced, with the readability enhanced.

3.3 Geospatial maps

Static geospatial maps tends to show many different points of interests, which might overload the user with content. Hence, to enhance the use of geospatial maps, we intend to include in the interactivity function so as to allow users to click and explore different points as desired.

4. DESIGN FRAMEWORK

This application makes use of the open-source R language to conduct visual analysis. The application design considerations are as follows:

- Utilise standard R packages to create reproducible text and visual analysis
- Utilise the embedded Shiny Web Application in R to translate the codes into a webpage for users' ease of understanding
- Provides interactivity functions for users to navigate through the app to discover trends and insights

The design of the application will consists of five major tabs for navigation at the top panel.

i) Introduction - Describes the main purpose of our application
ii) History of GASTech - Sub tabs of Text Analysis and Network Graph to respectively discover insights from the newspaper articles and employee relationships
iii) Message Stream Exploration - Explores tweets from different users on the incident day itself to sieve out notable keywords from otherwise spam information
iv) Risk Level Timeline tab - Further split into Call Centre Reports and Microblog Messages to detect how the public risk levels changed over time on the incident day
v) Message Stream Geomap - Shows locations where those key messages appeared

Coupled with the interactivity aspect to choose different options from dropdown boxes to sliders for exploration, the combination of these views gives the user the flexibility and autonomy to navigate through the app to find out more information regarding the Kronos Incident.

To faciliate users in their exploration around the application, we have also provided a user guide for their reference.

4.1 Data Used

The datasets used in this study is taken from the VAST Challenge 2021 Mini Challenges 1 and 3. They are collated in the following tables as shown below.

#HOW TO INSERT TABLE GAHHHH#

These datasets are then loaded into R and further data preparation is conducted to clean these raw data mainly with the tidyverse package.

4.2 Analytic Techniques used in Shiny App

A variety of standard R packages were used to conduct text and visual analysis on the datasets to draw useful insights. The following techniques are used for analysis.

4.2.1 Text Analysis

Our application will build several text analysis outputs to break down large chunks of text data from newspaper articles and tweets into understandable visualisation for users to view:

Comparison Cloud

Comparison clouds are used to visualise the similarity and differences of important words used by different newsgroups. Wordclouds are able to extract out keyword metadata and the frequency of the appearance of a particular word in the articles is determined by the size of the word in the wordcloud. This type of visualisation is useful in the quick pickup of prominent terms to determine the significance of those words. In turn, based on the keywords extracted out, we are also able to get a sense of the sentiments and attitude of these newsgroups regarding their articles about GASTech.

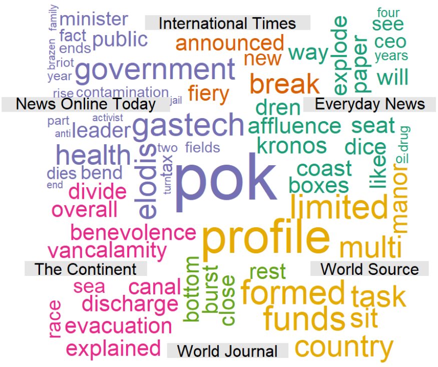
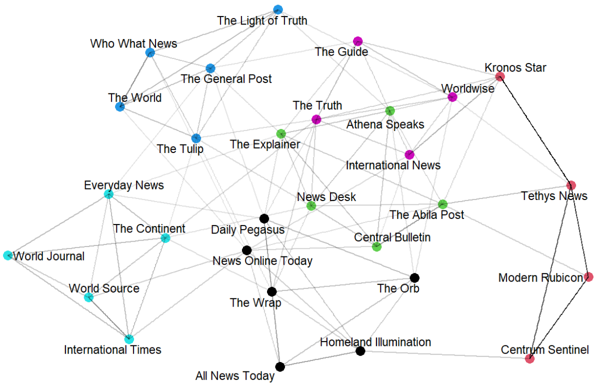


Fig. x above shows the comparison cloud between six newsgroups based on the words taken from the published news articles' titles. Our application will have the option for choosing between viewing keywords from "Content of Articles" or "Titles of Articles." Moreover, users can choose the newsgroups they would have to like to compare from a dropdown list of all the newsgroups available.

The packages used to build this visualisation are **tidytext**, **tm** and **wordcloud**. **Tidytext** is used to convert text into a format that is visualisable with the use of 'unnext_tokens' function. **tm** has built in functions such as 'removeStopwords' to help in the removal of unimportant words. Lastly, the comparison cloud will be developed using the **wordcloud** package.

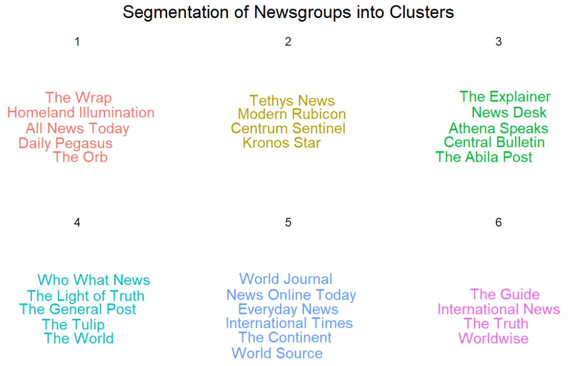
Textnet, Cluster Analysis and Text Plot

The Textnet approach is used to represent relationships between words. The usage of nodes and links is able to explicit show the relationships between each neighbouring text [cite]. In this case, the first node set is the words found in news articles and the second node set is the newsgroups themselves. That way, a network can be created where newsgroups are connected by their use of the same words, as shown in Fig. X.

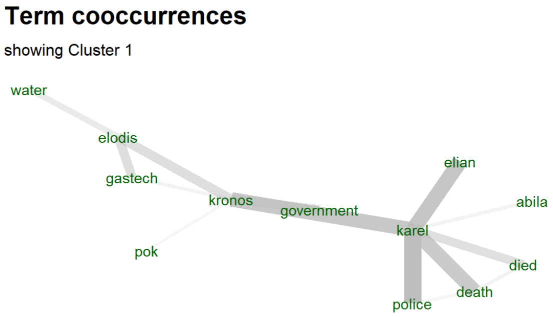


The node color corresponds to the text communities, with the same colour indicating a strong relationship between its components. In this way, clusters are formed to segment the newsgroups with similar characteristics in terms of the types of words used in the news articles.

To delve into the cluster segments of different newsgroup, we visualised each segmentation as follows:

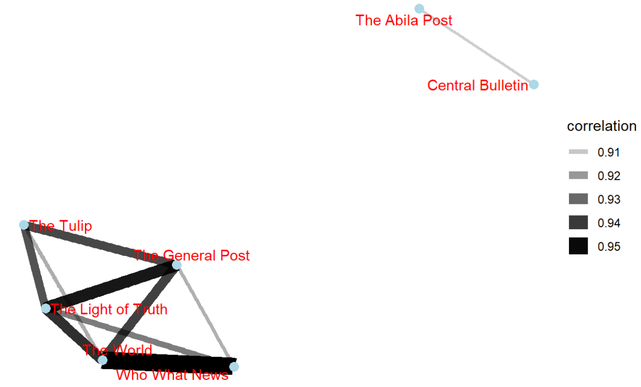


Upon visualising the components in each cluster, text plot visualisations are conducted to pull out the word cooccurences between word-pairs, so as to determine the context and content of each clusters.



The main R packages used are **textnet**, **ggwordcloud**, **tidytext**, **udpipe** and **textplot**. **Textnet** is currently the only R package available to implement text network techniques in R. To display the word cloud by cluster, **ggwordcloud** was utilised. **Tidytext** helps to convert text into a format that is visualisable with the use of 'unnest_tokens' function. The **udpipe** package provides language-agnostic tokenisation, tagging, lemmatisation and dependency parsing of raw text, which is an essential part in natural language processing (NLP). Lastly, to plot data as a text plot, we will be needing the **textplot** package.

Correlation Graphs Correlation graphs are plotted to determine the correlation between different newsgroups. From this, we are able to determine which newsgroups might be highly related in terms of their reports of certain events over the years. The correlation values are obtained using the widely used Pearson method.

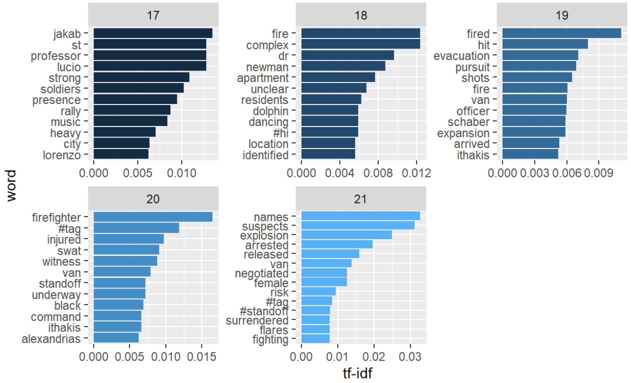


The R packages used are **widyr** and **ggraph**. **Widyr** is able to cast a tidy dataset into a wide matrix, performs an operation such as computing the correlation on it, and then re-tidies the result. 'pairwise_cor' function is found in this package. **ggraph** is then used to plot the relationship between different newsgroups based on their correlation values.

Term Frequency-Inverse Document Frequency (TF-IDF)

As mentioned in the **Review & Critics of Past Works** section, word cloud alone might not be very useful to visualise a collection of microblog message due to its consistency issue. As such, TF-IDF approach is more helpful when trying to pick out specific key events/topics of relevance. This approach aims to measure the importance of a word is to a document in a collection/corpus of documents, by including the inverse document frequency (IDF) to balance the term frequency (TF) used in wordclouds. Hence, we will view each hour of tweets as a document and all 5 hours of the dataset as the corpus, to determine the key events that dominate each hour. With that, TF-IDF can be a powerful heuristic to sieve out the most important events occurring during each period from the chatter spread throughout the dataset.

We plotted out both unigrams and bigrams to look at both the singular word and word-pairs respectively.



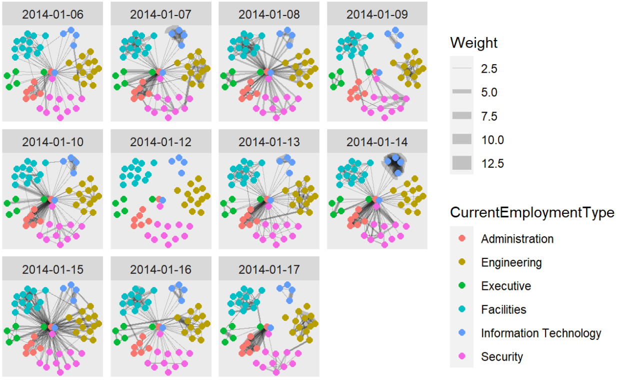
The relevant R package used is **tidytext**, where the 'bind_tf_idf' function helps to compute and bind the term frequency, inverse document frequency and td-idf of a tidy text dataset to the dataset.

4.2.2 *Network Graphs*

Network graphs are informative visualisations to mainly show the relationship between different entities. In this study, network graphs are created to visualise the different official and unofficial relationships of GASTech employees and the overall distribution flow of emails.

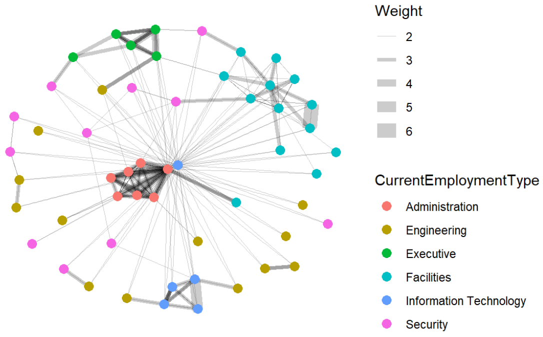
Email Flow

The email distribution flow from one person to another is visualised in terms of a network graph. The thickness of the edge implies the number of email passed between two nodes. The nodes are grouped by the employees' **Current Employment Type**, and users also have the option to view select the email distribution flow on a daily basis or on a weekly basis.



Relationships among GASTech Employees

The email relationships of the employees are split into work-related and non-work related, in attempt to identify potential suspicious activities that might be transpiring between different employees. There are four different views of the network graphs, with the nodes sorted by "Citizenship," "Current Employment Type," "Gender" and "Current Employment Title."



The main R packages needed to create network graphs are **tidygraph** and **ggraph**. Under **tidygraph**, the 'tbl_graph' function is used to convert data to an object to display network belons to this package. **ggraph** is used to plot the network object.

4.2.3 *Bar Charts*

Separate bar charts are plotted to look into the frequency of the microblog and call center reports coming in over a fixed time period of between 1700H to 2200H. Through that, we aim to discover the risk levels indicated by these data to draw out a basic timeline of major events that occurred during that evening.

A data table is also linked to the bar chart to show the exact messages that are received tagged to both time and location.

plotly package is used in this visualisation to be able to create interactive web-based graphs.

4.2.4 *Geospatial Mapping*

Interactive geospatial mapping visualisation is performed so as to track the locations where those microblogs tweets and call center reports originate from. The interactive view mode allow layers to be removed or added, as well as show the exact messages by clicking on the points.

The R package used is **tmap** to generate thematic maps with high flexibility and interactivity,

5. APPLICATION INSIGHTS

From our Shiny application, different insights can be drawn to learn more about the Kronos Incident.

5.1 Correlated News Groups

5.2 Frequently Occurring Phrases

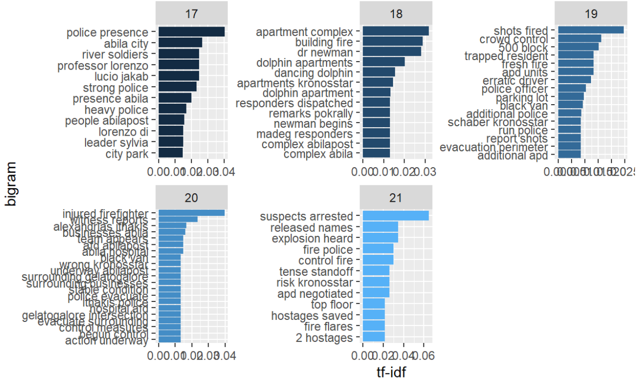
5.3 Content of Newsgroups

5.4 Context of Clusters

5.5 Email Flow and Distribution

5.6 Sensitive Emails Identified

5.7 Microblog Tweets Analysis



From this bigram plotted, we are able to extract certain important information as follows:

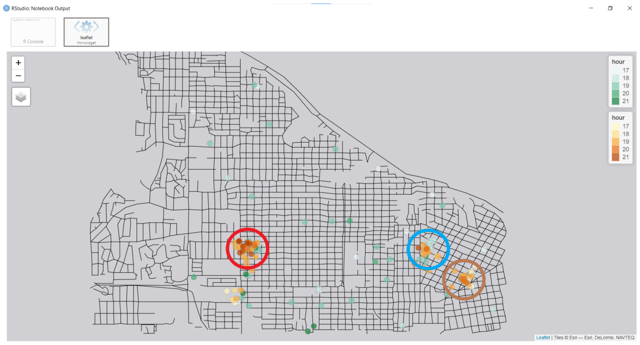
- 1800H+: The fire was likely related to a building, the apartment complex of Dancing Dolphon
- 1900H+: Shots were fired, with a trapped resident and black van being other points of interest.
- 2000H+: A firefighter was injured. A phrase "alexandrias ithakis" appears (this is the cross-street where the standoff happened). Several businesses appear to be affected as well.
- 2100H+: Besides the arrests already spotted in the wordcloud, an alarming event of explosions heard also appears.

5.8 Risk Levels From Call Center Reports

The basic timeline of major event identified over the course of the evening with its respective risk ratings (on a scale of 1-5, 5 being highest) are as follows:

1. The first big event was a fire detected at about 1830H (exact time: 1842H). An ambulance and fire truck was dispatched to the fire. Crowd control was also requested. The risk appears to have elevated, but appeared under control (risk level 3).
2. The next big event happened from 1915H: what appeared as a vehicle accident (exact time: 1919H) turned out to be a rogue black van running amok in the crowd. It was pursued by police units. Risk level would have gone up to 4, due to the uncontrollable nature of the van.
3. From 1930H, at least one officer was down because of the event (exact time: 1941H). A dire emergency was called, and additional support was requested. Risk level would have been at 5 (given that the situation was termed "dire").
4. Nearing the end after the black van and fire incident seemed to have died down, from 2045H there were several reports of crime scene investigations, with continued suspicious reports. Risk level would be at 3 (moderate).

5.9 Map of Key Messages with Geolocation



From Fig. X, we are able observe that there are three locations with high density of message origination, suggesting that major events are occurring around those location clusters,

The red cluster appears to be the place with highest risk and potential consequences, due to the high number of negative tweets and several serious call center reports spread over time. Both types of messages mention the standoff and police casualties.

The blue cluster also has many call center reports, but lesser negative tweets. Closer inspection by clicking on the points show that these tweets are mainly about the fire which appeared to have been under control, and the call center reports are mostly about dispatch requests.

The brown cluster seems to be about the explosion that occurred late. Closer inspection shows that many of such messages are sent by a spam account footfingers.

6. CONCLUSION AND FUTURE WORK

In conclusion, this study

Further developments to the application can be implemented such that the application is able to take out other text data such as text corpus

References

[1] Fenner, M. 2012. One-click science marketing. *Nature Materials*. 11, 4 (Mar. 2012), 261–263.

[2] Meier, R. 2012. *Professinal Android 4 Application Development*. John Wiley & Sons, Inc.