**Using Twitter Data to analyse product shortages in the USA due to Coronavirus**

**1.0 Introduction**

Coronavirus or COVID-19 is a global pandemic, infecting approximately 4 million people with around **280,000** deaths as of 11th May 2020 (ECDC, 2020). High demand and low supply of goods have impacted product supplies, with a lower level of distribution internationally (Reuters, 2020). Through unpredictable consumption, with families stockpiling goods to ensure their families can eat, goods supply issues were further compounded (Royte, 2020). In the USA, there has been a 29% increase in purchasing of basic food supplies when compared to the same week in 2019 (22nd-29th February 2019) (The Nielsen Company, 2020). Shortages of toilet paper, bread and various staples have been widely reported, with queues of up to 800 people at supermarkets (Farzan, 2020). Consequently, Narea (2020) stated, the food supply is plentiful; however, the food may not be in the right location.

"What a crisis like the novel Coronavirus reveals about the food system…is actually its flexibility and strength under pressure."(Rubinstein, 2020). By using this web-based map, the food distribution system is provided with Twitter data from the 2nd-9th April 2020 to assess where in the USA the public mentioned Coronavirus and coffee, milk, bread, toilet paper and fries. This spatial distribution can be assessed, and heat maps created to provide visual representations of hotspots regarding particular products. This allows the food distribution system to pivot towards these areas, targeting required goods to these locations (Rubinstein, 2020). Data collection occurred from the 48 million monthly active Twitter users in the USA, assessing the public's perception of food supply, and importantly food shortages (Clement, 2020). For food distributors, this is useful to assess where in the USA the population mentions these products, which could indicate shortages of supplies, providing information with which to target supplies.

**2.0 Methods**

**2.1 Data Gathering**

This application used the Tweepy API to extract Tweets from the 2nd-9th April 2020. A Python script was written to use the Tweepy API, using the search terms outlined in Table 1, with only the tweets from the USA returned and stored to CSV files for conversion to GEOJSON using <http://www.convertcsv.com/csv-to-geojson.htm> and subsequently plotted on the web-application. Any tweets without locations were also not stored as the location is paramount to the use case for this application. Currently, the basic version of the Tweepy API only allows for Tweets from the previous seven days to be analysed and downloaded (Roesslein, 2020). These Tweets were then sanitised using Preprocessor, a module for Python that removed the twitter handles, mentions of other users, emojis and smileys.

The use cases for this application are three-fold; initially, this application will allow the food distribution network to pivot to an area of hotspots regarding certain products, focusing their distribution to these areas to increase supplies of goods. Secondly, the application will allow for flexibility and crisis planning to be completed, assessing how the public reacted to shortages of supplies. Finally, by using the display of all tweets, users can read how the public interpret these problems, if they feel the issue is with stockpiling, supply, or demand, and adjust their supply chains accordingly.

**2.2 Chosen Technologies**

Text mining, the process of extracting unstructured information from textual data, such as social media data, was chosen for this application (Surjandari et al., 2015). This text mining relied upon the Tweepy API and a Python script and followed comparable methodology to Achrekar et al., (2011) by embedding indicators, or search terms, as input parameters into the request and fetching the responses from Twitter, with the country filter of USA applied to the study, Table 1 outlines the search terms used. Only tweets with a place name included were written out to the CSV, as it was paramount that for this application locations were included.

Table 1. Products included in analysis, and the Search Terms used in the Python Script.

|  |  |  |  |
| --- | --- | --- | --- |
| **Product** | **Product Search Terms** | **Coronavirus Search Terms** | **Tweets Collected** |
| Bread | Bread | COVID, Corona, lockdown, quarantine | 13 |
| Milk | Milk | COVID, Corona, lockdown, quarantine | 10 |
| Coffee | Coffee | COVID, Corona, lockdown, quarantine | 23 |
| Toilet Paper | Toilet Paper, Toilet Roll, Loo Roll | COVID, Corona, lockdown, quarantine | 41 |
| Fries | Fries, Chips | COVID, Corona, lockdown, quarantine | 6 |

Twitter data was chosen as a medium of analysis due to its location-defined, high volume, real-time updates from the public (Vertalka, 2018; Achrekar et al., 2011). Approximately 48 million people in the USA actively use Twitter each month, providing a huge source of information for analysis (Clement, 2020). In academic research, Twitter users are referred to as 'sensors', with the search parameters referred to as 'indicators' (Vertalka, 2018; Achrekar et al., 2011). The use of the Geoweb, with geo-tagging of Tweets, a central practice is key here (Elwood and Leszczynski, 2011). By using the Geoweb and attaching location-based tweets to web-based content, spatial information can be quickly and easily disseminated (Elwood and Leszyczynski, 2011). Importantly, the use of Volunteer Generated Information (VGI) such as Twitter data, even with its inherent reliability issues, when aggregated will still reveal patterns in space, as the group will converge upon the truth (Achrekar et al., 2011; Goodchild and Li, 2012; Vertalka,2018).

Leaflet was used as it is the leading open-source JavaScript library for mapping (Agafonkin, 2019). Importantly, Leaflet is a free to use software, meaning no extra expenditure for distribution companies to freely and without limitations have access to the usage of this software. Further, plugins extend Leaflet's usage, such as the Leaflet.Heat Plugin and Leaflet.EasyPrint plugin, both vital for the usability of this application (Agafonkin, 2019). Leaflet functions well with GeoJSON data, the file type of all the Tweet data used within this application. Finally, there is a wealth of helpful documentation about Leaflet, enabling customisation of the web-map to suit the needs of the user (Agafonkin, 2019).

**2.3 Complex Methods**

The application includes the creation and usage of a Python script for the collation of twitter data, this required the creation of a Twitter Developer account, and then to plug in the various search terms attached to the products for consideration shown in Table 1. Further, a complex method for the generation and implementation of heat maps to show the spatial distribution of the tweets across the USA, using various forms was built. This method required various validation checks, including ensuring that weightings added to 100%, ensuring checked layers were weighted and ensuring the user understands what they needed to do to build these layers. From this, each of the layers included in the weightings was added to the heat map layer, and its relative weighting as a percentage divided by 100 as the heat layer plugin uses a 0-1 scale (Agafonkin,2014). Resetting the layer is also possible to allow for multiple heat map layers to be generated to assess the products and their spatial distribution in more detail. By building various validation checks and ensuring the user cannot input alpha-numeric data, the website is more secure, and the user is unable to make fatal errors that harm the websites ability to function.

Through testing of the website, the validation checks were used as visual cues for the user, aiding the user experience. For example, a red box displayed when the weightings did not sum to 100 and would only alter to green upon validation of the sum, dynamically adjusting the colours as the user inputted values. Further, submit buttons would only appear on successful validation of criteria, and the Tweet Index page displays an information window when no tweets were displayed. This level of detail and commitment to providing a positive, simple user experience was time-intensive in its implementation.

**3.0 Discussion**

**3.1 Issues and Future Developments**

Currently, the Leaflet.Heat plugin that creates the heat map contains a bug, which has been previously raised (GitHub Inc, 2020). If the user adds all five layers to the heat map creation, it can cause the website to run slowly, and a visual bug in which the map control for the layers does not display correctly. This issue does not detrimentally impact the usage of the website but is unfortunate. Hopefully, this will be rectified in future versions of the Leaflet.Heat Plugin. Further, another visual limitation was found when moving the webpage to screens of differing size and aspect ratios. I thought this had been solved with the conversion from pixel sizing to percentage-based size using CCS, but this is still not completely rectified, instead sometimes relying upon scrolling to view elements.

Without reading the tweets in detail, which is the reason for the Tweet Index webpage, the user is unable to tell whether the tweet is referring to a shortage or plentiful supply of a product. This website currently makes use of the hotspots and density of tweets to initially determine areas in which the distribution company should focus, and then allows the user to assess these tweets by using the Tweet Index webpage to do so. However, to further develop this site, perhaps more stringent text mining could occur that makes use of further search terms to hone in upon the terminology used in the tweets.

To improve the usage and representative nature of the application, upgrading to the premium Tweepy API would allow for Tweets from the past 30 days to be analysed, and allows tweets since March 2006 to be accessed (Twitter Inc, 2020a). Only tweets from the 2nd-9th April, as the height of the food shortage issues, ran from the end of March to Mid-April 2020 broadly, so the application aimed to capture tweets about these issues (Royte, 2020). However, the web map could also be converted to making use of live tweet functions, if this was deemed necessary, for subsequent analysis.

**3.2 Security and Privacy Considerations**

The use of the Tweepy API requires the user does not exceed the limitations on access to the Twitter API and that access is solely by the Developer license holder, or this could result in the Twitter Developer required being revoked (Twitter Inc, 2020b). As such, this application downloaded the data to CSV files and ensured the wait\_on\_rate\_limit variable was always set to true. However, all the keys and passwords are removed to protect the security of the application and developer account used.

Tweet data before sanitisation has various security and privacy issues, namely that users can be located with a high degree of accuracy through their geo-tagged content (Elwood and Leszycynski, 2011). Further, social media information such as handles and relationships with others provide important information to attackers, so were all sanitised (Krombholz et al., 2015). Here, only the content and location of the tweets, with the users handle and any users mentioned in the tweets remaining private, to ensure acceptable online practices were adhered to (Elwood and Leszycynski, 2011).

Website design ensured that no user alpha-numeric inputs were allowed, and no login data was required, with only buttons used for user interactivity. This removed a whole host of security considerations, such as social engineering attacks and buffer overflow attacks. Social engineering attacks occur when the user is tricked into revealing their password and login information, enabling the attacker to potentially access multiple other sites as people often the same password (Krombholz et al., 2015). Buffer overflow attacks occur when attackers add data to program processes through inputs, which is also prevented using buttons and no alpha-numeric inputs (Kuperman et al., 2005). In summary, various attacks types that could harm the site are removed by blocking alpha-numeric inputs and ensuring the user inputs only buttons clicks.

**4.0 Conclusion**

In conclusion, this web-based map allows users to see hotspots and areas in which Tweets about Coronavirus and various products occurred. By allowing the user to see the python script used to collect data, and assess each Tweets content individually, the user can understand the meaning behind each Tweet and where there is a concern regarding supply, demand or stockpiling of goods. Jointly, these two webpages allow distribution companies to build flexibility into their networks, pivot rapidly to areas of concern and ensure that the population of the USA has adequate supplies of food and goods.

**References**

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H. and Liu, B., 2011, April. Predicting flu trends using twitter data. In 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS) (pp. 702-707). IEEE.

Agafonkin, V. 2014. Leaflet.Heat. <https://github.com/Leaflet/Leaflet.heat> [22/04/2020]

Agafonkin,V. 2019. *Leaflet- an open source JavaScript library for mobile-friendly interactive maps.* <https://leafletjs.com/> [24/04/2020]

Clement, J. 2020. Countries with the most Twitter users 2020. <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/> [22/04/2020]

Elwood, S. and Leszczynski, A., 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. Geoforum, 42(1), pp.6-15.

European Centre for Disease Prevention and Control (ECDC). 2020. Situation update worldwide, as of 11th May 2020. <https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases> [11/05/2020]

Farzan, A.N. 2020. *'It's been nuts': Costco's sales spike amid coronavirus panic-buying of bottled water, hand sanitizer and toilet paper.* <https://www.washingtonpost.com/nation/2020/03/06/costco-coronavirus/> [24/04/2020]

GitHub Inc, 2020. Laggy Heatmaps #116. <https://github.com/Leaflet/Leaflet.heat/issues/116> [22/04/2020]

Krombholz, K., Hobel, H., Huber, M. and Weippl, E., 2015. Advanced social engineering attacks. Journal of Information Security and applications, 22, pp.113-122.

Kuperman, B.A., Brodley, C.E., Ozdoganoglu, H., Vijaykumar, T.N. and Jalote, A., 2005. Detection and prevention of stack buffer overflow attacks. Communications of the ACM, 48(11), pp.50-56.

Lakhani, N. 2020. 'A perfect storm': US facing hunger crisis as demand for food banks soars. <https://www.theguardian.com/environment/2020/apr/02/us-food-banks-coronavirus-demand-unemployment> [22/04/2020]

Narea, N. 2020. *The US won't run out of food during the coronavirus pandemic.* <https://www.vox.com/2020/4/18/21222028/america-food-meat-supply-chain-coronavirus> [24/04/2020]

Reuters, 2020. Explainer: How the coronavirus crisis is affecting food supply. <https://www.reuters.com/article/us-health-coronavirus-food-explainer/explainer-how-the-coronavirus-crisis-is-affecting-food-supply-idUSKBN21L0D2> [22/04/2020]

Roesslein, J. 2020. API Reference. <http://docs.tweepy.org/en/latest/api.html> [22/04/2020]

Royte, E. 2020. Food waste and food insecurity rising amid coronavirus panic. <https://www.nationalgeographic.com/science/2020/03/food-waste-insecurity-rising-amid-coronavirus-panic/> [22/04/2020].

Rubinstein, P. 2020. *Why grocery shelves won't be empty for long.* <https://www.bbc.com/worklife/article/20200401-covid-19-why-we-wont-run-out-of-food-during-coronavirus> [24/04/2020]

Surjandari, I., Naffisah, M.S. and Prawiradinata, M.I., 2015. Text mining of twitter data for public sentiment analysis of staple foods price changes. Journal of Industrial and Intelligent Information, Engineering and Technology Publishing, 3(3).

The Nielsen Company, 2020. Key consumer behavior thresholds identified as the coronavirus outbreak evolves. <https://www.nielsen.com/us/en/insights/article/2020/key-consumer-behavior-thresholds-identified-as-the-coronavirus-outbreak-evolves/> [22/04/2020]

Twitter Inc. 2020a. Search Tweets. <https://developer.twitter.com/en/docs/tweets/search/overview/premium> [22/04/2020]

Vertalka, J.J., 2018. The Augmentation, Potential, and Practicality of Twitter Data for Predicting Influenza Emergency Room Admissions. Michigan State University.