COSC2671 | Social Media and Network Analytics

Assignment 2

Analyzing global trends for climate change based on tweets

10th Oct, 2019

Vishwa Gandhi – s3714805, Jigar Mangukiya - s3715807, Vikas Virani – 3715555

Group - 06**Master of Data Science, RMIT University**

Contents

1. Introduction

- The reports is written to follow study conducted as part of assignment 2, Social Media & Network analysis. The study aims to track, analyze and visualize the posts, events & trends related to "Climate Change" using tweets, their timings and their geographical location.

➢ 1.1 Climate Change

- Climate change is substantial shift in climate patterns across world noticed due to human activities from late 50s to present time. Climate change is mainly attributed to elevated Carbon dioxide levels in environment and other green-house gases. Some of the major reasons of these elevated levels of green-house gases are increase in usage of fossil fuels, deforestation, drop in plankton levels in the oceans due to large oil spills, wars and large wild fires. Consequences of climate change are observed in form of global warming, polar ice melting & rising ocean levels, unpredictable weather events & reduced crop yields due to damage to crops due to worsening weather.
- Climate Change has been a major concern for later part of this decade for governments, scientists and in general all residents of our planet.
- Lately there has been significant increase in attention towards climate change, in general, by both, government bodies & people.

➢ 1.2 Climate Change & Social Media

- Various International agencies like UN Climate Action Summit & United Nations Framework Convention on Climate Change are actively pressuring governments to take action against worsening climate situation.
- People are also forming large groups and communities to spread awareness, protest and gather support to remedy climate change. There have been significant uproar in general awareness towards climate change in population and protests, rallies, strikes have increased in numbers, intensity & size.
- Celebrities also time and again make this point at world stage, like Leonardo DiCaprio making climate change a part of his Oscar acceptance speech.
- Due to the ease of availability and increasing popularity of social media networks and their large user base, social media sites like Facebook & Twitter, people prefer expressing their views on the ongoing activities around them using these platforms. These posts describe people's orientation, trends, and events about climate change.
- Utilizing the advances in the NLP & Data analytics, we have conducted this study, to gather, process, analyze and visualize different aspects of views, events, awareness & activities about climate change globally.

➢ 1.3 Problem Formation

- Study aims to Use tweets acquired from twitter to,

➢ Major goal – Track awareness & activity level of Climate Change related activities across world

➢ Approach -

1. Analyze popular hashtags & keywords associated with Climate Change
2. Track events related to climate change like strikes or conferences

3. Analyze public sentiment orientation towards these events or trends
4. Find out which locations, states or countries are more active and aware about climate change based on tweets
5. Track geographical footprint of the prominent influencers for climate activism and deduce potential growth area

- The study can potentially help stakeholders at multiple levels in identifying key trends and events happening across world regarding climate change, measure public interest and find out active locations where climate change events are happening.
- Analyzing large number of tweets can help give some idea about how climate change is being tackled globally, active groups, agencies & communities that are trying to tackle climate change and awareness about climate change in different geographical locations globally.

2. Data Collection

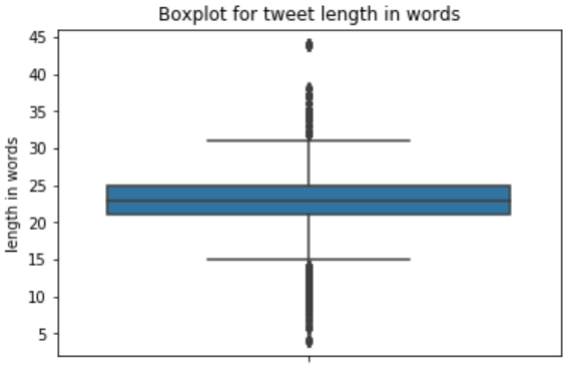
➢ 2.1 API Description

- To collect the tweets from twitter, we've used Twitter's REST API service to and through developer portal & accounts. We have chosen to use REST API and not streaming API is due to the searching capabilities and relatively small number of tweets we are utilizing for this study.
- Streaming API is aimed at providing tracking capabilities over a large period for collecting millions of tweets, however, due to limited scope of this study and restricted computational power, we have chosen to analyze and track only recent events and not past events.
- Using a twitter developer account, we've acquired necessary authentication information & used it to fetch tweets.
- We have used "#climatechange" to mine the tweets of past 7 days for majority of our study. However, we have used some other event specific hashtags to mine tweets particular to that event after identifying these events from results of analysis of "#climatechange".

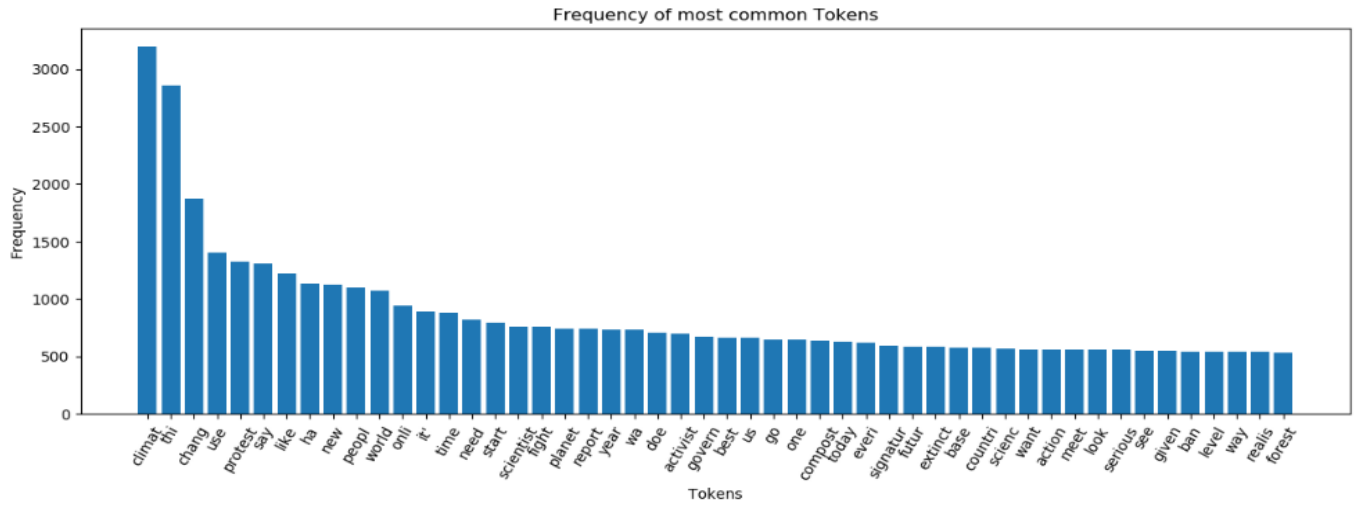
➢ 2.2 Data Description

- We have retrieved 25000 tweets from past week to perform analysis as mentioned in section 1.3. We've used a language filter to mine tweets in only English language to help simplify the analysis. Retrieved tweets are stored in a file in json format. For 25000 tweets (documents), the size of corpus is roughly 130 MB and to downloading entire corpus took around 40 minutes (largely due to waiting on limit feature of API).

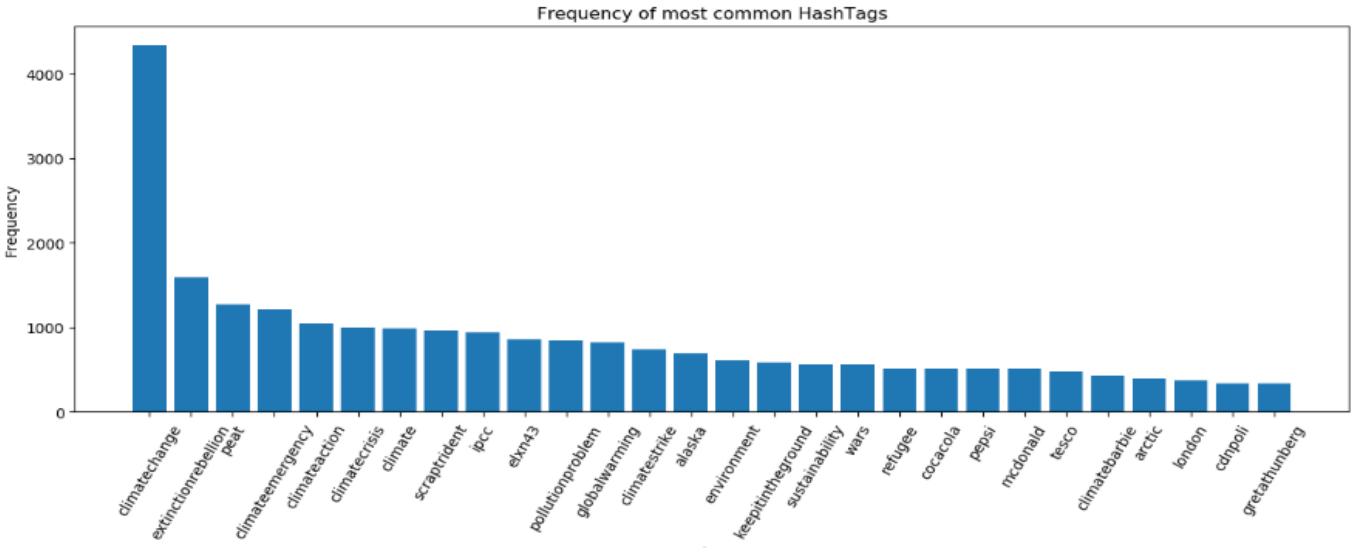
➢ 2.3 Data Exploration



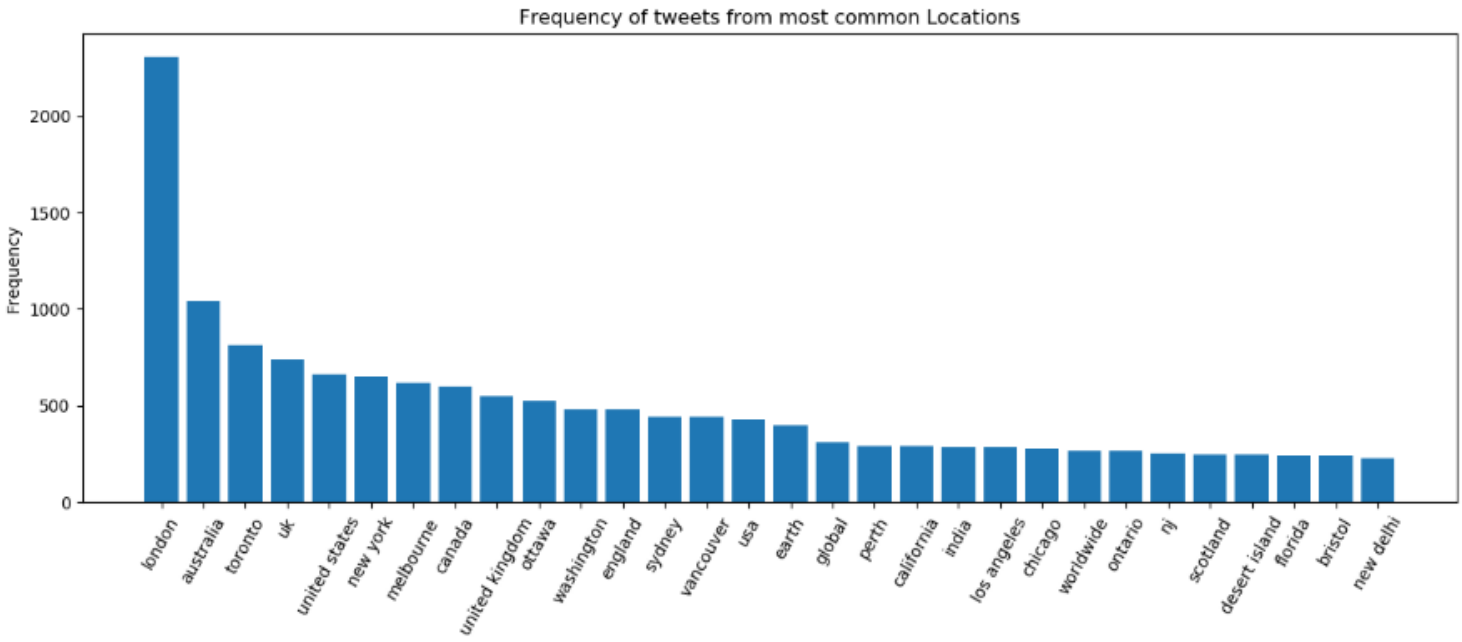
- Tweets are generally small 2-3 sentence long post expressing views of a person on a topic. That makes it considerably small document in NLP terms, as per the boxplot, average tweet length is approximately 23 words. Which will be further reduced by filtering the noise.
- There are some outliers with very low or comparatively higher word count of nearly 0 or 40 +.
- To analyze further, we extracted 50 words with highest frequency to have an intuition of most occurring words. Some of the noticeable words are protest, climate, strike, activist, ban, and scientist. Which might help us when further investigate event or trend detection.



- Following bar plot shows the most mentioned hashtags other than the hashtag we used to search for our data.

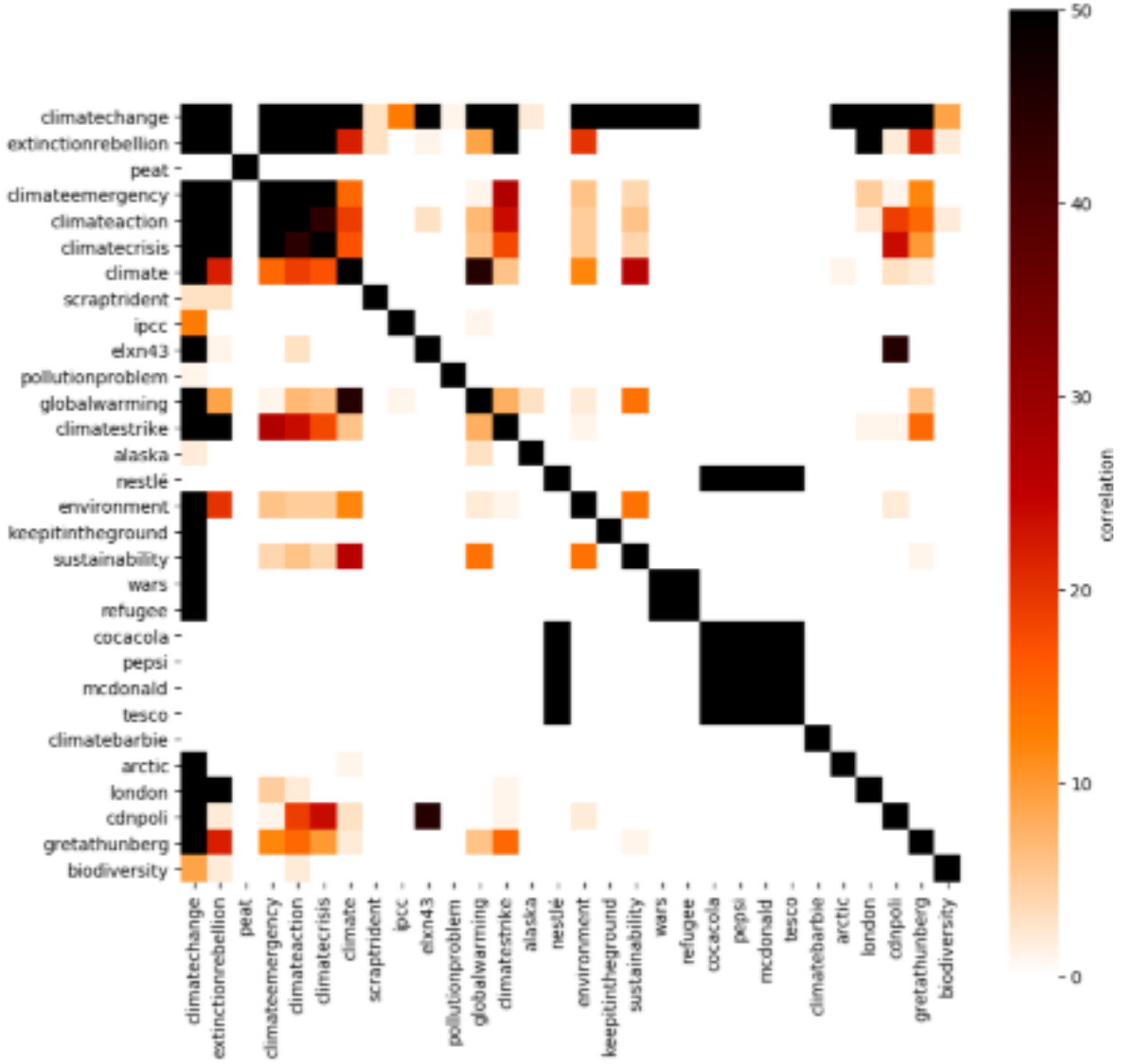


- Most of the hashtags are interesting for our study which shows significance of the hash tags in tweet analysis. We will analyze hashtags like Extinctionrebellion, PEAT, scraptrident, ipcc & climate strike in more detail in further sections.
- Apart from the hashtags and keywords, geographical locations are also important part of our analysis in this study. Knowing how many tweets originated about a topic from a particular location can tell a lot about certain events, activity or trending item in a context of a geographical location. In a broader scope, it can indicate relative awareness about the event or trend at particular location and can be compared to other locations.
- Following graph shows city, state or country level geographical affinity of tweet origins.



- The plot indicates high level of activity in UK in general followed by Australia, Canada, USA. There is however, some noise in the data, like earth, global etc. Also locations like London, UK, United Kingdom & England can be mapped to a single entity "United Kingdom". However, these are shown separately in the plot above. However, we will be taking care of this things to get a clearer picture of using maps to visualize similar data.
- For now, the plot revels that Countries like UK, Canada, Australia, USA have higher tweet count regarding climate change. This can help climate change action planners in targeting areas that are not so active like India & China, which comes near the end of the list or doesn't occur in the list.
- One interesting study is to see which hashtags are mentioned simultaneously. That can shade some light on events or topics that are related to climate change and to each other.
- To analyze this, we selected top 30 hashtags with highest frequency and created vectors for each tweet. We created concurrence matrix indicating how many times the hashtags appear together in the same tweet. By visualizing this using a heatmap, we checked which things are related to each other.
- The following heatmap indicates certain things like,
    - Hashtag "ExtinctionRebellion" occurs with many other hashtags like "ClimateStrike", "ClimateAction" and "GlobalWarming". This points towards involvement of Climate Activist Group "Extinction Rebellion", being active in multiple aspects related to climate change. It also occurs with Hashtag "London", which indicates towards the massive rally organized in London to protest against climate change and government schemes in past week by the group.
    - Hashtag "Nestle", co-occurs with tags like "CocaCola", "Pepsi", "Tesco" & "McDonalds", which provides evidence of association of these food product manufacturers in some sort of activities regarding climate change. A quick search on internet reveled latest backlash these organizations faced due to excessive usage of water resources and usage of plastic which supposedly ends up in ocean, which is believed to have hurt the local ecosystem around their manufacturing plants and oceans globally.

- There are many other similar and significant associations that can be made based on this heatmap or an extended version of similar heatmap covering more hashtags or keywords. However, due to restrictions of reporting, we have stuck with top 30 hashtags.



3. Pre-processing and Data cleaning

- Next step of the process is to clean and preprocess our data. Tweets are expressed in high level natural language which means it will incorporate complexity of human language into our data. TO rectify this complexity, we need to perform sequence of cleaning and pre-processing to transform our text into a more analysis friendly corpus.

➢ 3.1 Cleaning Tasks

- I've implemented a method which reads the data from the dump file created in 2nd section. Method processes the data tweet at a time, performing following cleaning tasks for each tweet.

No.	Cleaning Task	Description
1	Case – Folding	Convert upper case letters to lower case letters
2	Tokenize	Form a list of tokens from a single tweet using delimiters
3	Whitespace stripping	Trim the unnecessary space from either side of the words
4	Stop word filtering	Filter language specific stop words from the text, words which don't have a specific meaning associated with them
5	Stemming	Utilizing stemmer provided in NLTK package (porter stemmer) to convert words to a standard format
6	Miscellaneous filtering	- Remove URL links, Numbers, Tweet Handles, Punctuations etc.

- First step in the preprocessing phase, case-folding aims at removing analytically meaningless polymorphic property induced by multiple cases of English alphabet by converting entire text to single case, here, lowercase.
- Next step breaks down tweets into atomic words by using delimiters like space. This helps analysis by giving atomic tokens for tweets which are easier to interpret and analyze than entire sentences or groups of sentences.
- Like every language, English also has its rules of grammar, which in general uses certain words like 'the', 'what', 'were' and 'is' etc, to form syntactically and semantically meaningful sentences. However, for analyzing large corpus of texts, these words hardly convey any inferable meaning. These stop words are generally removed in these kind of analysis to reduce processing & semantic overhead.
- NLTK python package for natural language processing provides porter stemmer. Stemming is process of removing grammatically suffixes to converts words to a standard form. These stemmed words might not be real words. For example words like analyzed, analysis and analyzing are all converted to some word like analyz. This maps multiple words with similar contextual meaning to a single word and thus reducing complexity induced by grammatical variants of a single stem.
- Other preprocessing tasks include the removing URLs, numbers and other unnecessary forms of text which are not directly useful to our analysis.
- Emoticons & hashtags might seem the helpful in analyzing sentiment of a single tweet, but emoticons are often inaccurate and hashtags are mostly repetitive and tend to overpower other less recurring terms which convey more meaning. Hence, after trying out both approaches, I've removed them as they had little effect on the analysis.

Below example describes how preprocessing affects the volume of information to be processed and reduces language complexity by reducing number or words.

Raw Tweet –

We conduct research on public climatechange knowledge, attitudes, policy support & behavior. Posted events are not endorsements.

Processed Tweet -

[conduct, research, public, climatechange, knowledge, attitud, policy …]

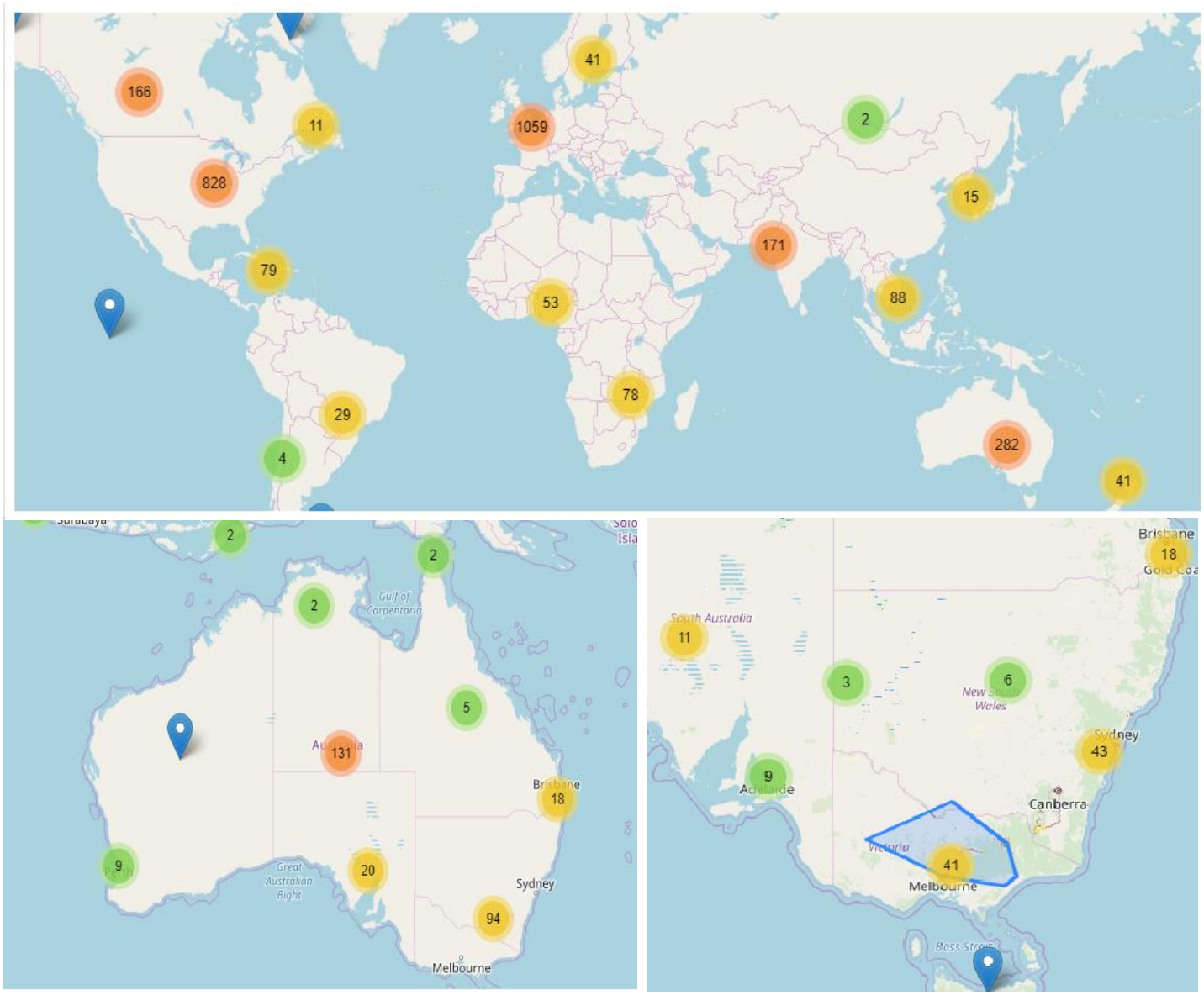
- Processed tweets has lot less words and lots of words are changed to standard form. This reduces the amount of information to be processed when performing further analysis and also results in improved accuracy of analysis method by removing noise
- Following table shows the reduction in data after pre-processing and cleaning. You can observe significant decrease in number of tokens after cleaning.

Total tokens	Tokens after preprocessing	Total Hashtags
567642	534793	30631

4. Analysis Approach

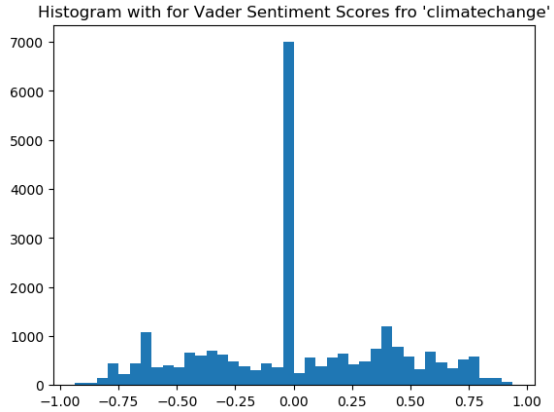
➢ 4.1 Trend Analysis & Event Detection

- Twitter provides an effective way of detecting unplanned events by analyzing trends over a specific time. By temporally analyzing the tweets, we can deduce when some particular topic was trending. That can allow the tracking of event duration and occurrence time.
- Twitter users use some particular keywords of hashtags to attach their posts to some particular topic. Monitoring this hashtags allows for tracking beginning and ending of the event. Hashtag frequencies can be monitored over long duration to track sudden bursts in the popularity of events.
- Calculating bursty-ness in the hashtag frequency is the simplest way of detecting an unplanned event. To calculate bursty-ness, we have created ratio of frequency at current time quantum to frequency at previous time quantum. There is a problem of artificially inflated ratios, where previous frequency is near 0 and new frequency is larger than that. That causes the ratio to be significantly larger. However, for current analysis, this measures are sufficient to detect past events as we are not using streaming API to detect live events.
- Apart from detecting the beginning time and duration of the event, we can also detect geographical location of events using tweets. This is important information from many perspectives. For example, journalists can find the locations of interesting events unfolding near them. Police can monitor illegal activities, confrontations around them and take action. Emergency services, targeted marketing campaigns, monitoring large gatherings are just some of the potential use cases of the tracking locations of events through social media posts.
- Tweets are often geo-tagged, meaning, they associate location of tweet origin in the tweet data itself. By analyzing large amount of tweets by their origin can help us find exact location of where an event is trending. Clustering the hashtags based on their locations allows for a quick overview of popularity of some event or topic in different parts of globe.
- We are using trend analysis to identify past events related to climate change, as well as clustering based on longitude and latitude to identify activities and awareness of events related to climate change around the world.
- We will look at global trends for hashtag climatechange as well as geographic distribution of the hashtag mentions to analyze when certain climate related events have occurred in past couple of days and at which locations climatechange hashtag is more frequent, i.e. more attention and awareness is oriented towards climate change.
- The following hybrid bubble map is generated as per frequency of mentions of keyword “climatechange” in tweet text or hashtags. The plotting is done using “IPyleaflet” package. The package works by loading interactive world maps and dropping markers using (longitude, latitude) coordinates. We have used the location information provided with tweets, which in some cases are noisy, to find out coordinates using “Geopy” package. We buffered coordinates of all the regions appearing in our dataset so as to speed up things in subsequent runs. “Geopy” has APIs to find out coordinates based either on (city/state, country) or (country) queries. As tweets’ geo-tags are often noisy, we have replaced noise location with null coordinates and they are not plotted.
- Interactive maps implicitly clusters nearby markers and forms their clusters and also displays their Centroidal Voronoi Tessellation (CVT) as can be seen in the last image. Readjusting zoom level will run clustering algorithm again so as to allow get finer or coarser view of frequency distribution.



- The images show that regions UK, USA, Canada, Australia & India have high frequency in decreasing order. In Australia, East coast, south east and central Australia shows high frequency of tweets relating to climatechange. Specifically, Melbourne & Sydney have higher tweet count than other cities.
- Absence of any data from China can be attributed to governmental restrictions of social media sites and not to less awareness about climate change. However, other major countries like Russia, Middle-eastern countries & North African Countries can be classified in to less active on social media on climatechange drives, however, this only refers to this particular keyword and there can potentially be other keywords as well. It still doesn't fade the impression of lack of awareness in this parts of globe completely.
- After analyzing where part of trend, we move on to analyze what part of events and trends. It is natural that events or trends will receive both positive as well as negative sentiment orientation from people. We analyze global sentiment towards climate change and check what topics are

being described in tweets regarding climate change. We use Vader sentiment analysis model to associate sentiment orientation to our corpus and further utilized LDA based topic modelling to find out the topics that are being discussed with this hashtag.



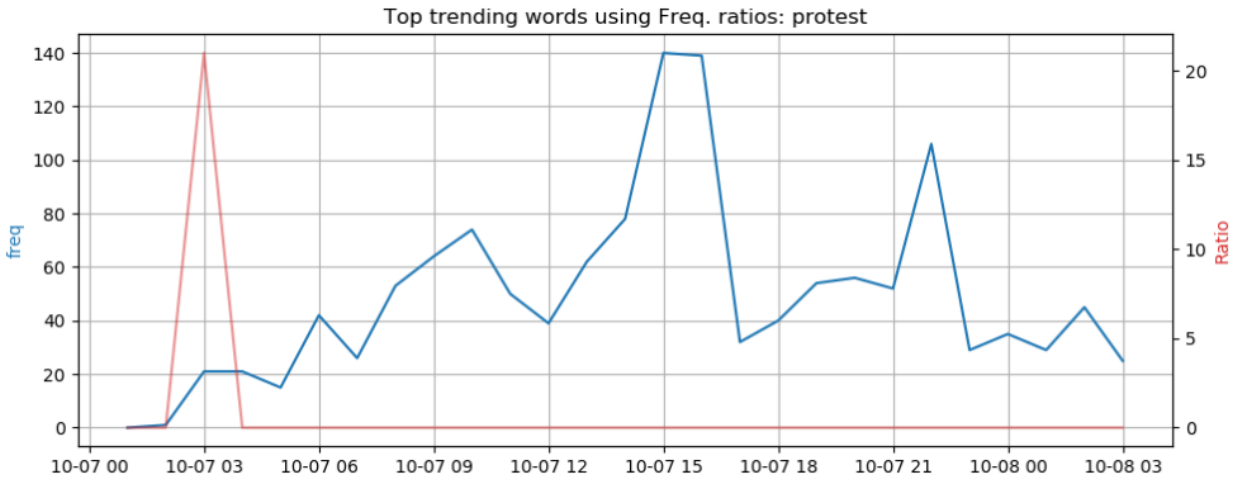
	Vader Analysis
positive count	9609
neutral count	7257
negative count	8134

- From the above histogram and table, it seems the sentiment orientation of entire dataset is almost balanced, with roughly equal number of tweets in either category, with slightly more tone towards positive sentiment. The spike in the middle is due to binning including 0 in that category which forms neutral sentiment.

- This was a generalized view of climate change based activities around the world. Following sections take a more detailed approach towards particular trending events and topics related to climate change, like climate strikes and protests, deforestation, trending climate change activist group extinction rebellion and other significant trends and events. We will try to identify trends from historical data, their location, people's sentiment towards that trend and topics that are being discussed in that trend.
- We use burstyness based trend analysis and modelled frequency-ratio for identified trending words.

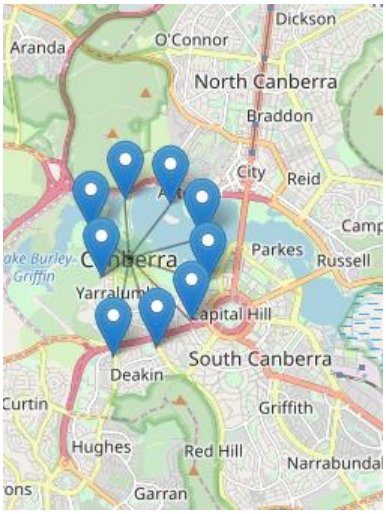
➢ 4.2 Trending – Protests, planet-wide trend.

- Based on the Frequency-Ration based trend analysis, word protest was found to be trending on 7th October, 2019. Protest and rallies were held in multiple locations around globe as per the news. It was pretty obvious that people will be talking about protest on twitter.



- The plot above shows the time based plotting of frequency counts. Multiple bursts can be observed in the plot at various time points. There is large burst in the center which is the reading of AEST 3:00 pm. As per the local information, there was a large protest held in Melbourne CBD. Similar protest were held in UK as well as US East Coast cities.
- The below figure incorporates the hybrid bubble map generated by using frequency count of the word protest occurring in tweet or hashtags in our climatechange tweet dataset. As it can be inferred from the figure, UK, USA, Australia & Canada have high frequency of tweets regarding protests.
- We used Google to verify the trend and geographical distributions and found sufficient evidence about the credibility of geo-tagging. Most of the times, news articles were published after the tweets were trending. It shows the power and suitability of twitter and crowd-sourced information in general in indicating and relaying local information on global social media platforms.
- The bottom left plot is zoomed in image of Canberra. There were roughly 9 markers. The image shows how "IPyleaflet" package forms clusters of multiple markers by taking a centroid and merging it according to zoom level dynamically.

Global tweet frequency for 'protest'



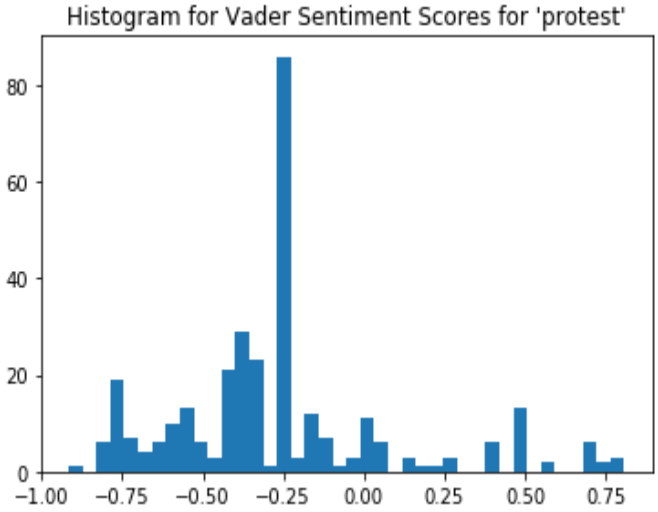
'A duty to disobey': who are Extinction Rebellion?
The Sydney Morning Herald - 7 Oct 2019
It seeks to compel governments to act on **climate change** and .... turned out to **protests** in Melbourne and Sydney and, in **October**, activists were ...

London police arrest 135 **climate change protesters**
The Sunday Times Sri Lanka - 7 Oct 2019
LONDON (Reuters) - **London's** Metropolitan Police have arrested 135 people taking part in **protests** organized by the **climate change protest** ...

Climate Change Protests: With Fake Blood, Extinction ...
The New York Times - 7 Oct 2019
By disrupting several landmarks in the heart of **New York's** financial ... book about **climate change**, marched with her **7**-week-old daughter, Ada, ...

Extinction Rebellion **protesters** pour fake blood over **New** ...
The Guardian - 7 Oct 2019

- We further analyzed and checked the public sentiment orientation towards protests. We used Vader Sentiment Analysis on a set of filtered tweets that had word protest in it.
- The histogram reveals that sentiment orientation towards the in the tweets mentioning word protest is skewed towards negative sentiment. That makes sense as people can either be reflecting upon the bad effects of climate change via protest or can be reflecting upon the fact that how protest have hindred the life of nearby places causing traffic and ruckus.
- We further filtered the tweets according to their sentiment orientation and performed topic modelling. We used LDA topic modelling and word clouds to visualize the topics that were mined.



Positive tweets

Negative tweets

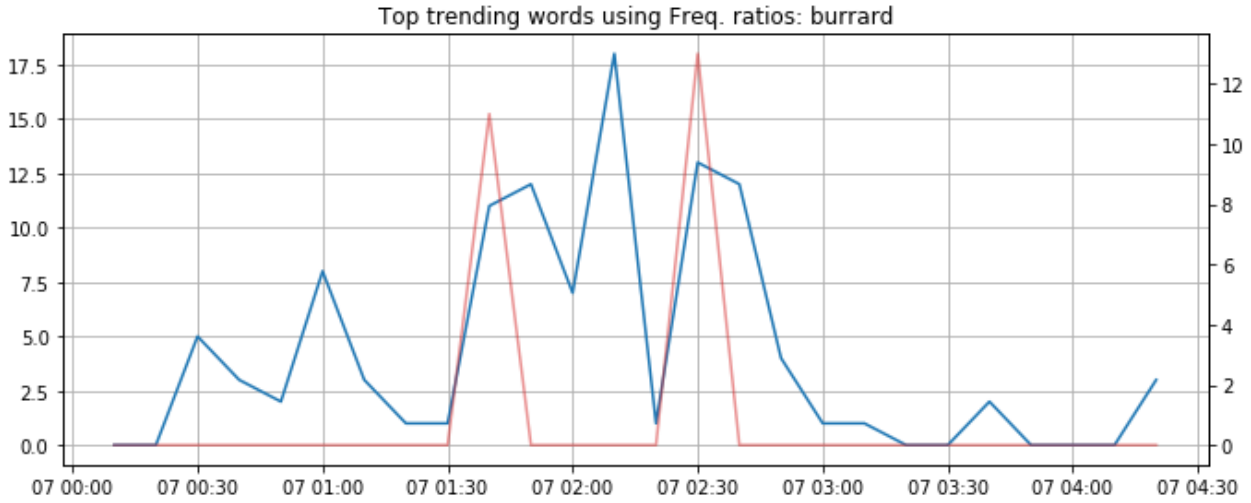
- For tweets with positive sentiment orientation, topics were revolving around Pope Francis who, recently displayed keen concern regarding climate change issue and had a major gathering on 10 october, which was very well recived by public and hailed by most of the people. It is understandable that public sentiment will be positive about it.

Pope Francis considers dropping celibacy requirements for ...
USA TODAY - 7 Oct 2019
Pope Francis considers dropping celibacy requirements for some ..... of rainforests and local cultures to **climate change**, migration and clean ...
First synod talks look at **climate**, priests, inculturation, Vatican ...
Catholic News Service - 8 Oct 2019

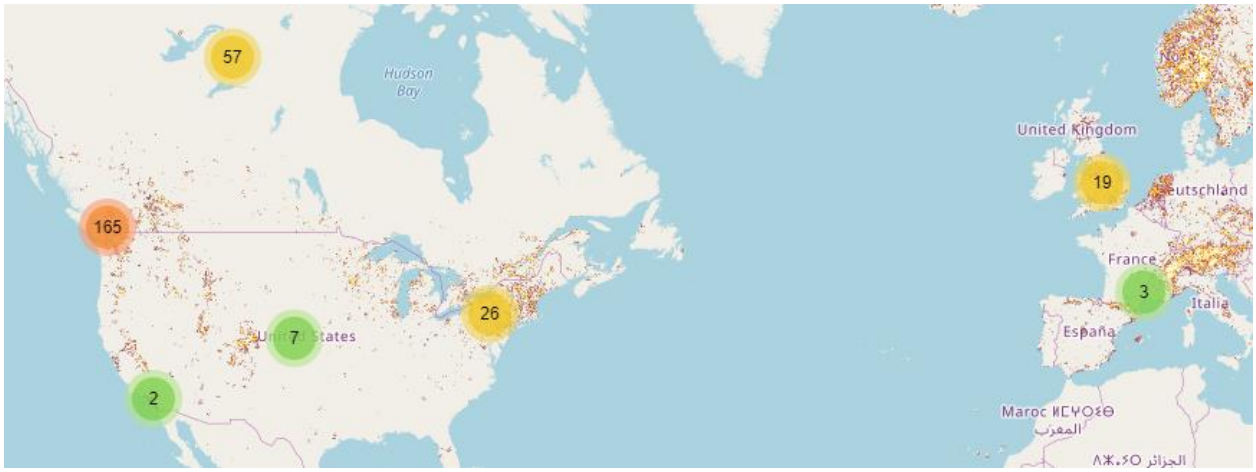
- Simillarly for tweets mentioning word protest and having negative sentiment revolves around topics like extinction, missile program – Trident & Defence, extinction rebellion, protest, inaction ect. Which falls well within the scope of negative aspect of protest and climate change.

➢ 4.3 Event Detection – Burrard Bridge Arrests, Vancouver, Canada

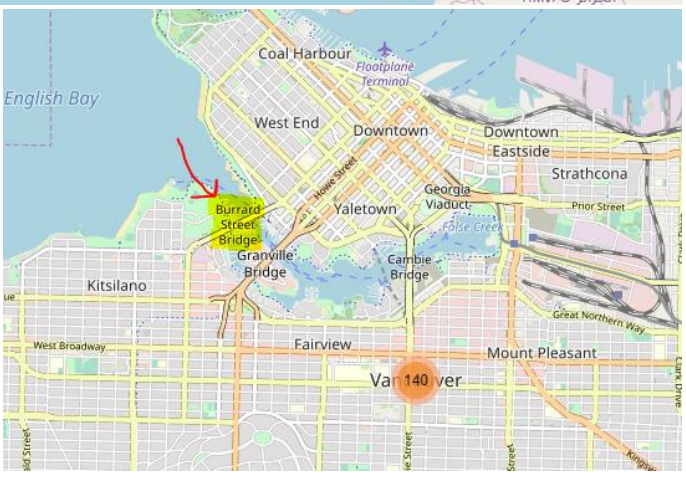
- One of the interesting event that was detecting was associated with term “Burrard”. Which can be observed in the plot below. 7th October, 2019, the observed frequency of term Burrard is displayed by the following plot. The frequency is quite bursty and peaks around 2:00 am AEST.



- We analyzed geo-tagging of the term “Burrard” on all tweets and frequency distribution across globe is shown in the hybrid bubble map below. Significant number of tweets seemed to originate from Vancouver, Canada, as well as from New York, USA & London, UK.



- Exploring Vancouver region in detail, we found that centroids were located near North Vancouver & Vancouver Downtown. Considering noise and coarse accuracy of geo-tagging, area of tweets origin (centroid), following image showed the actual location of event.



- After crunching through news sources, we confirmed that there was a major incident

in form of climate change protest by activist group “Extinction Rebellion”.

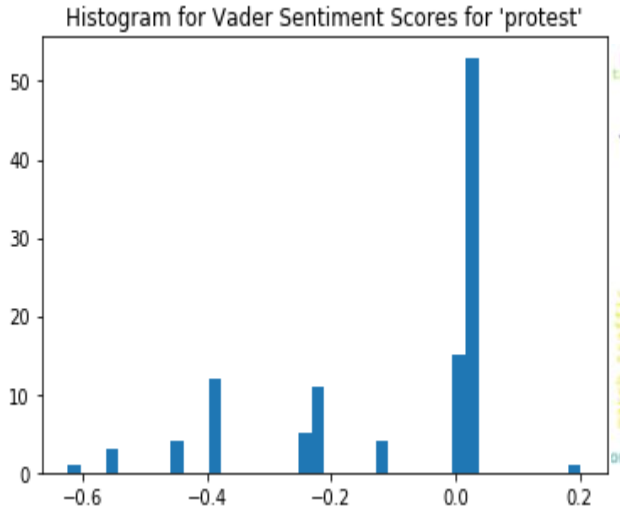
- Some of the news cutouts looked something like this. Apparently, large number of climate activists gathered on Burrard Bridge to protest against climate change resulting on blockage. Authorities took action by arresting large number of protestors which resulted in increased traction over social media.

Vancouver's **Burrard** Bridge shut down due to climate activists
Toronto Star - 7 Oct 2019
VANCOUVER—For 22-year-old Vancouverite Edison Huang, the wake-up-call came in the form of the October 2018 UN Intergovernmental ...

Climate change protesters shut down **Burrard** Street Bridge
CTV News - 7 Oct 2019
VANCOUVER - Protesters shut down the **Burrard** Street Bridge to traffic Monday for an "all day" protest calling for action on climate change.

Vancouver's **Burrard** Bridge reopens after day-long climate ...
Globalnews.ca - 7 Oct 2019

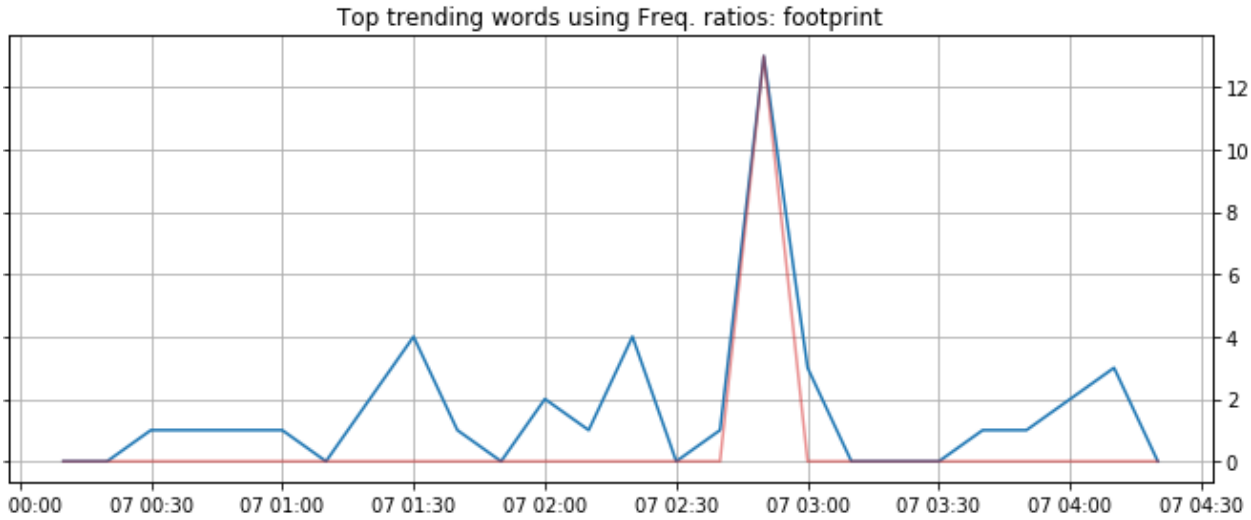
- We further analyzed the sentiment associated with blockage and topics that were being discussed by tweet authors using Vader sentiment analysis model & LDA model for topic modelling.
- Sentiments expressed in tweets were largely negative and there was lot of backlash as the blockage resulted in heavy traffic and commute came to a standstill.



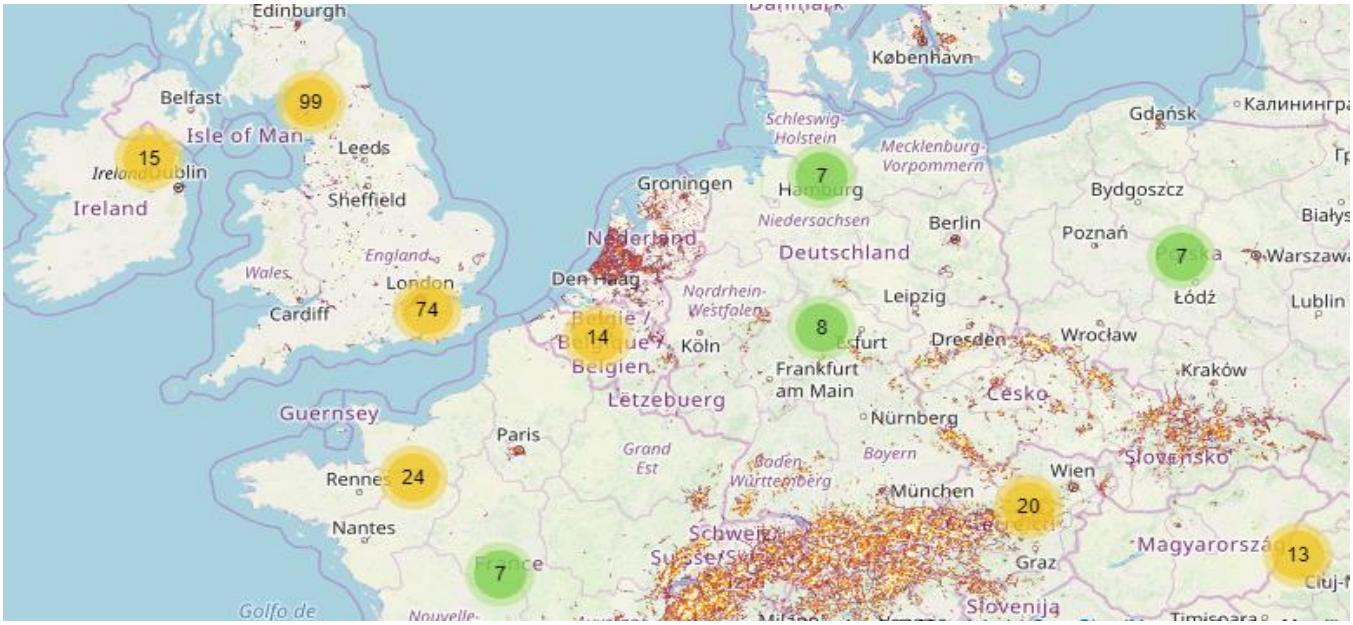
- Topic Modelling reveals the “Vehicle Traffic”, “Blocking”, “Extinction Rebellion”, “Blocking Vancouver” etc keywords that reflect upon the major points being discussed about the event on social media.
- This study reveals the prowess of social media posts in revealing information about event detection, geographical affinity & sentiment orientation measuring of events.

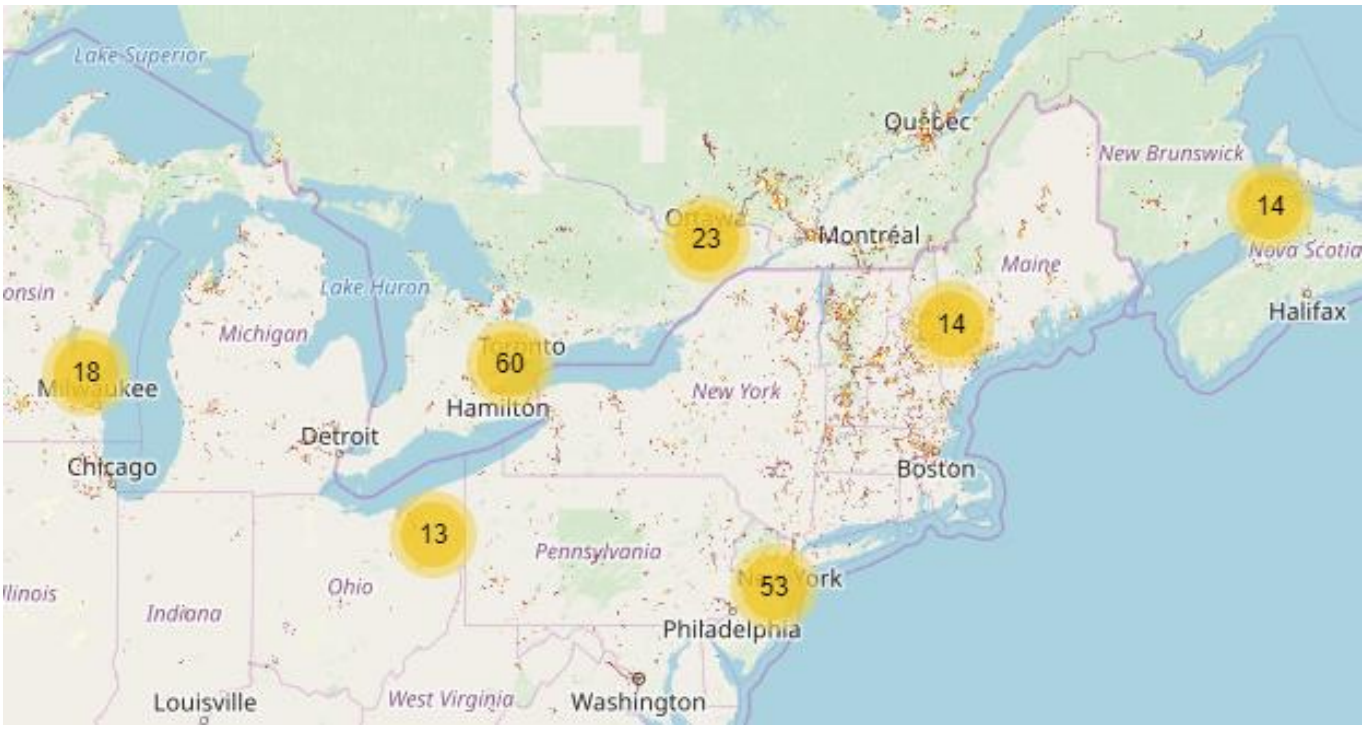
➢ 4.4 Trend Detection – Carbon Footprint Discussions, European Union (EU)

- One of the trend that sprung out of our trend model was term “Footprint” which was trending for 2:30 am to 3:30 am (AEST) with a large spike in frequency ratio. Bit more exploration revealed the association of word with term “Carbon Footprint”. Which in general refers to Carbon dioxide and other greenhouse gases emissions by fossil fuel burning.

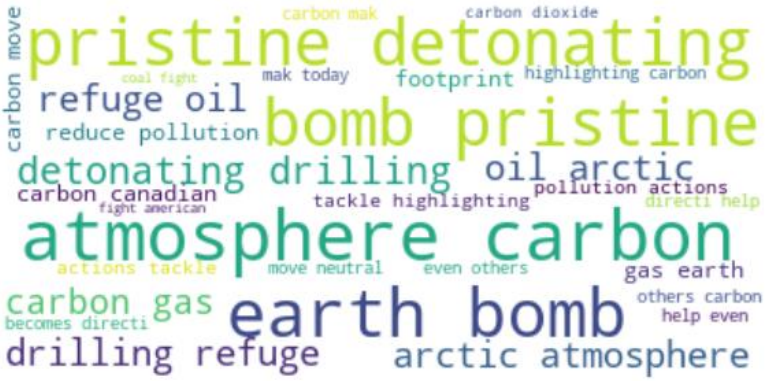


- We tracking down the locations of origin of tweets containing terms “footprint”, which lead us to 2 main regions from where the tweets were being made. Countries in European Union and US – Canadian cities. UK, France, Ireland & Germany have significant number of tweets in EU, while Toronto, Ottawa in Canada & New York in US have significant amount of tweets concerning carbon footprint.
- We modeled topics to understand what are the topics associated with carbon footprint tweets. Topic modelling revealed that the topics being discussed are revolving around arctic oil drilling, preserving pristine landscapes, carbon masks. Which are some of very hot topics in the arena of climate change tackling.





- This brief study helps in analyzing which countries are actively discussing the issue of carbon emissions and usage of oil drilling. North American states & EU countries are the major contributors to the topic on social media.



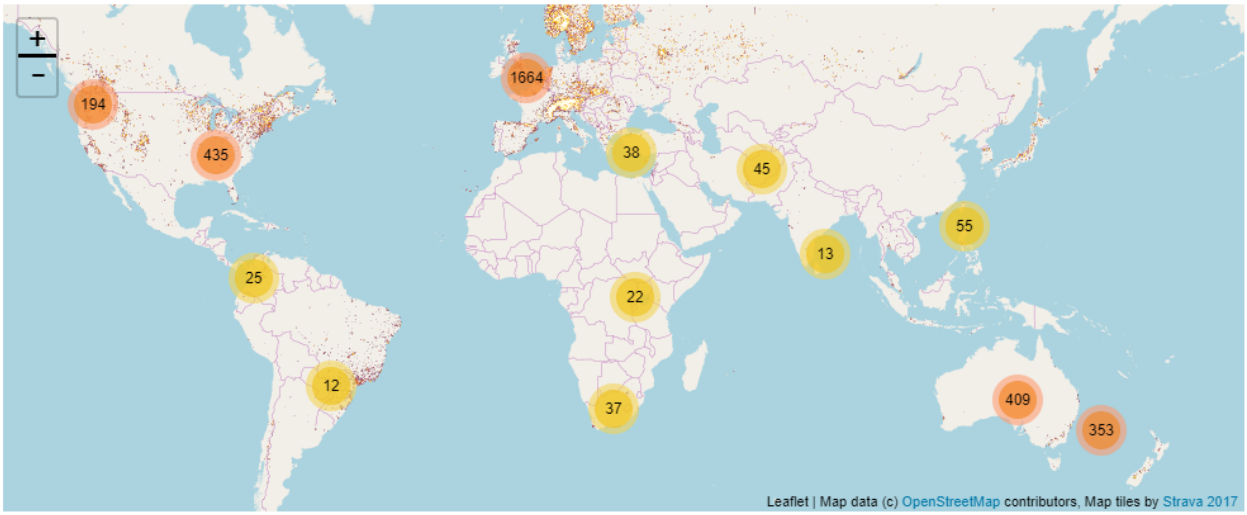
- Following news items support the intuitional interpretations of our analysis. However, we could not find any specific event like a conference or international summit happening on that duration which caused the burst in tweets mentioning carbon footprint.

French and Riviera News Monday **7th October** 2019
Riviera Radio - 6 Oct 2019
French and Riviera News Monday **7th October** 2019 ... that a move to glass would have a "dreadful effect" on the **carbon footprint** of packaging.

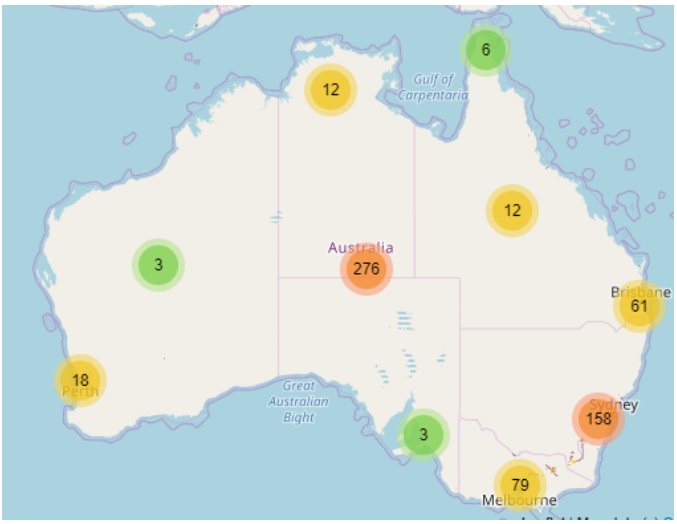
The madness of Extinction Rebellion
Spiked - 7 Oct 2019
Extinction Rebellion says mankind is doomed if we do not cut **carbon emissions** to Net Zero by 2025. That's six years' time. Think about it: they ...

➢ 4.5 Groups, Communities & their Footprint

- Actions against climate change is mostly driven by some central personalities, groups and communities spread out across the world. Some prominent examples would be
- "Al Gore" – Climate activist with highest number of followers
- "Greta Thunberg" – Young climate activist who gave a well-received speech at 2019 UN Climate Action Summit
- "Extinction Rebellion (ER)" – Highly active climate action group which organized many events across the world to protest against climate change
- In our dataset, we observed high number of tweets from "Extinction Rebellion" group in our dataset. They have been actively organizing protests in UK, Germany, France, USA, Canada & Australia. We plotted tweet frequency mentioning the group to analyze which location has high number of mentions translating to higher activity in that region.
- Following hybrid bubble plot confirms the news sources about the active regions of ER group. They are highly active in all the above mentioned countries and also have some footprint in other regions like Hong Kong, India, Pakistan, African countries and Latin American countries. However, in the later mentioned regions, activity levels are very primitive.



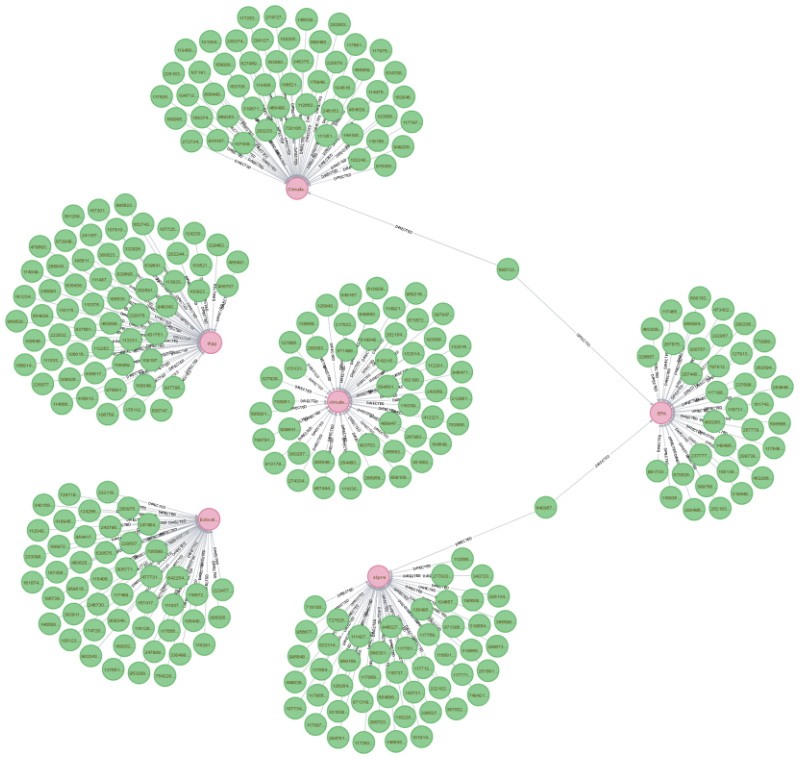
- In Australia, ER is has some foothold in all the major cities, with most active regions being Sydney, Melbourne & Brisbane. Notice that 276 tweets shown in the northern territory is due selection of improper cluster centroid by the mapping module and noise in the dataset.



- Similarly, Influence level for other prominent climate activist accounts can also be tracked to particular geographical regions. This becomes an important source of information for planning and strategizing action plans to get maximum traction and public support.

➢ 4.6 Network graph modelling & community detection

- To demonstrate scope of network graph modelling and interesting insights that can be gained from this novel approach based on graph theory, we created a sample network graph of twitter groups (users) and followers and partitioned them into communities using community detection algorithm.
- Network graphs are powerful and handy tools to analyze real world social networks as they are intuitively similar.
- Communities are smaller subset of graphs, members of which are strongly connected to each other and to form our network graph, we selected 11 prominent & influential accounts of people and groups to mine the user data. Essentially, we selected mix of people like Al Gore & Greta Thunberg and highly active groups like Extinction Rebellion & Climate Group. We fetched 1000 followers for each 11 of them.
- By allocating users to communities, we can profile them and utilize profiling information to plan targeted marketing or targeted membership drives to increase the memberships for groups.
- Each of these 11 central accounts have followers count in between 10000 & 2.3 million. Selecting a smaller subset from them, we get mostly unique users and very small fraction of users will be following multiple of these central accounts.
- Now, as the 11 accounts that we chose as seed accounts are the source of other data, they implicitly form communities and they automatically become the most influential members of that community. That makes this exercise trivial in nature. However, to demonstrate applicability of the graph algorithms and essentially explain how it can be used to analyze and understand the network is the main goal of this activity.
- After running the data acquisition in python, we merged all the followers in a set, to remove duplicates from them. Each of these follower will become a node in our graph, identified by twitter user_id. We will associate an attribute with these follower node “type”: ”user”. Similarly, the 11 seed accounts will be typed as “group” and will form a node each. We further added an edge to one of the seed account, every time it appeared in a follow list. This resulted in edge list, which contained directed edges from user to one or more seed account which it follow. Some users will follow multiple accounts, while some will follow only a single account.
- We initially used Neo4J to visualize our graph. Neo4J is an Open source graph database equipped to handle network graphs relations. However, due to high spec requirement and low visualization

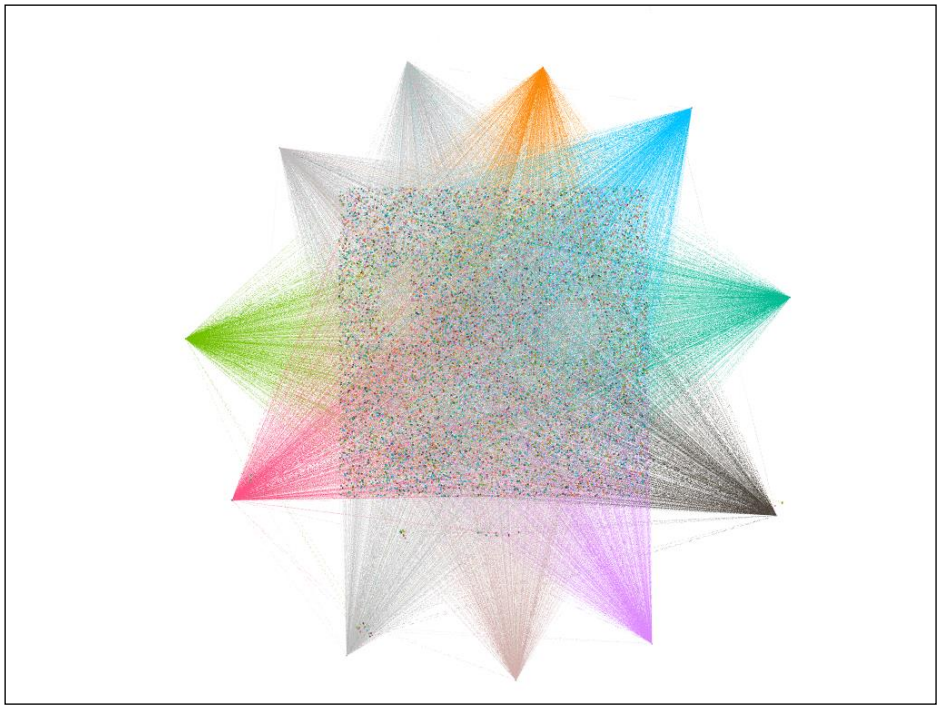


prowess, we could only visualize part of the network at a time.

- Image shows the network graph of around 300 nodes with 6 seed accounts from original 11 accounts and rest of the followers. Only 2 followers, in this minor subset of entire graph, follow more than one group nodes. Pink nodes are seed accounts & green are their respective followers connected via a directed edge.
- To perform community detection, we moved our data to “gephi”. Which is a really powerful tool to visualize and process graphs. We created a networkX graph in python notebook and exported it to GEXF format, which was later imported in gephi.
- The graph we were modelling had following statistics.

```
Name:
Type: DiGraph
Number of nodes: 10708
Number of edges: 11000
Average in degree:   1.0273
Average out degree:   1.0273
```

- Using this comparatively small/medium sized graph we first ran modularity algorithm to decide optimal number of communities in the graph. This however is a redundant step as we already know that there precisely 11 central components that connect majority of graph models.
- After running modularity, we partition the nodes based on modularity and color code them for easy viewing.
- Following image shows the end graph. Each color Portion shows a community and as stated earlier there are 11 of them.



- This entire process is bit redundant for this scenario as we already knew about community and influencers. However, this exercise shows the effectiveness and power of network graphs in analyzing social media networks.

5. Conclusions & Summary

- We concluded analysis phase by making following contextual observations/conclusions.
- Some countries are more aware and active about climate change related activities like protests, rallies or scientific conferences and meet-ups. Some of these countries are Australia, UK, Canada and Denmark. These countries have multiple active groups and individual with high frequency of tweets regarding topics – Climate Change, Carbon Footprint, Protests, and Strikes etc.
- Information regarding the awareness and activity levels can be used in 2 major ways. A) Using the active regions to hold stronger front on dealing with climate change. B) Targeting zones with lower activity levels by first identifying these regions and developing strategies to find influencers using network modelling and achieving the momentum to increase the activities regarding climate change.
- Some of the less active regions on this fronts are Russia, Japan, China, Middle Eastern Countries, India & central & eastern European countries. These regions have substantial population and a successful climate change awareness drive can bring in real impact by occupying large masses of crowds in the fight against climate change.
- Some countries have issues which are particular to their own interests, like Arctic Drilling in Russia & USA and Carbon Tax in Canada and Scandinavia. Hence, these regions see higher tweet frequency regarding these topics than other countries who are not directly involved with these activities. Hence these topics can be said to have a geographical affinity. We have linked drilling for fossil fuel issues with USA & Canada while carbon footprints are more affined to USA and EU.
- Movements against the climate change are often organized by groups. These groups have large number of followers and hence have high centrality in social networks and also acts as major influencers in that area. As our findings in section 4, groups like “Extinction Rebellion” work at global level and have large network of follower across the world. However, their effect is somewhat limited in highly populated areas like India & China, which have offer great potential to these groups to gather more supporters. They can leverage information regarding their user-base to expand their network in other places by running targeted campaigns.
- In general, Twitter and other social media have great potential to increase the effectiveness of the strategies to fight climate change by supporting movements with insights derived from analyzing large open source data like tweets and posts.

Limitations

- Entire analysis was performed on hypothesis validation model, where we used past data to perform analysis and gain insights about past events and trends, which was due to limited scope of the assignment, which was to demonstrate usability of tweet datasets to support data driven decision making. However, in practice, using live data acquired from streaming api would give much more valuable insights.
- Tweets are only in one language which can induce a bias towards English speaking countries and overlook other countries. In a real application, equal focus should be given to all major languages.
- Limitations in hardware and processing power has kept graph exploration techniques beyond the scope of this assignment, however, network graphs can deliver really high values insights about communities and influencers.

6. References

[1] D. Sarkar, "Text Wrangling & Pre-processing: A Practitioner's Guide to NLP", *Kdnuggets.com*, 2019. [Online]. Available: https://www.kdnuggets.com/2018/08/practitioners-guide-processing-understanding-text-2.html. [Accessed: 21- Aug- 2019].

[2] M. Kelecheva, "Using LDA Topic Models as a Classification Model Input. [Online]. Available: https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28 [Accessed: 22- Aug- 2019].

[3] P. Barhate, "Latent Dirichlet Allocation for Beginners: A high level intuition, *Medium.com* [Online]. Available: https://medium.com/@pratikbarhate/latent-dirichlet-allocation-for-beginners-a-high-level-intuition-23f8a5cbad71. [Accessed: 26- Aug- 2019].

[4] J. Chan, " COSC2671 | Social Media and Network Analytics, *Class Notes (Internal)* [Online] [Private]. [Accessed: 26- Aug- 2019].

[5] J. Chan, " COSC2671 | Social Media and Network Analytics, *Lab Notes (Internal)* [Online] [Private]. [Accessed: 26- Aug- 2019].

[6] R. McCreadie, C. Macdonald and I. Ounis, "MapReduce indexing strategies: Studying scalability and efficiency", Glasgow, 2011.

[7] "ipyleaflet: Interactive maps in the Jupyter notebook — ipyleaflet documentation" , *Ipyleaflet.readthedocs.io*, 2019. [Online]. Available: https://ipyleaflet.readthedocs.io/en/latest/. [Accessed: 11- Oct- 2019].

[8] "Welcome to GeoPy's documentation! — GeoPy 1.20.0 documentation", *Geopy.readthedocs.io*, 2019. [Online]. Available: https://geopy.readthedocs.io/en/stable/. [Accessed: 11- Oct- 2019].

[9] T. Rathod and M. Barot, "Trend Analysis on Twitter for Predicting Public Opinion on Ongoing Events", International Journal of Computer Applications, 2018.

[10] S. Corlay, "Interactive GIS in Jupyter with ipyleaflet", *Medium*, 2019. [Online]. Available: https://blog.jupyter.org/interactive-gis-in-jupyter-with-ipyleaflet-52f9657fa7a. [Accessed: 11- Oct- 2019].

[11] "Climate Change in Australia", *Climatechangeinaustralia.gov.au*, 2019. [Online]. Available: https://www.climatechangeinaustralia.gov.au/en/. [Accessed: 11- Oct- 2019].

[12] "Tweepy Documentation — tweepy 3.8.0 documentation", *Docs.tweepy.org*, 2019. [Online]. Available: http://docs.tweepy.org/en/latest/. [Accessed: 11- Oct- 2019].

[13] "Gephi Tutorials", *Seinecle.github.io*, 2019. [Online]. Available: https://seinecle.github.io/gephi-tutorials/. [Accessed: 13- Oct- 2019].

[14] "Overview of NetworkX — NetworkX 2.3 documentation", *Networkx.github.io*, 2019. [Online]. Available: https://networkx.github.io/documentation/stable/. [Accessed: 13- Oct- 2019].

[15] "The Neo4j Operations Manual v3.5", *Neo4j.com*, 2019. [Online]. Available: https://neo4j.com/docs/operations-manual/current/. [Accessed: 13- Oct- 2019].

[16] J. Ladd, J. Otis, C. Warren and S. Weingart, "Exploring and Analyzing Network Data with Python", *Programminghistorian.org*, 2019. [Online]. Available: https://programminghistorian.org/en/lessons/exploring-and-analyzing-network-data-with-python. [Accessed: 13- Oct- 2019].