Indian Institute of Technology Gandhinagar

CS-613 - Natural Language Processing

——————-:Report: Assignment 1:————- Data Crawling

Instructor: Prof. Mayank Singh

Topic: ENVIRONMENT

Team No: 11

Jani Dhyey Hareshbhai	18110068
Kumar Ayush Paramhans	18110089
Shah Jay Rahul	18110154
Thakar Devanshu Nilesh	18110174

Contents

1 Question 1

Describe the subtopics assigned to each member. Why was a particular subtopic selected? What important/popular keywords will the subtopics be searched with?

The subtopics assigned to each member are as follows:

- Devanshu : "Pollution"
  - Environment is closely interlinked with pollution. With the dawn of industrialization pollution has become a global challenge. Many rivers of India are severely polluted. Air pollution reaches hazardous levels especially in north Indian cities. Pollution seems to be a widely discussed topic over the media. Pollution as a subtopic was selected because it is directly affecting the environment.
  - Keywords for search: **Pollution, India, Pollution in India**
- Jay: "Climate Change"
  - Climate change has been a highly discussed topic in recent years. It is now widely accepted that the phenomenon of climate change has affected several parts of the world leading to rising sea levels, wildfires, destruction of flora and fauna etc. There has been a huge slew of support behind several popular activists and common citizens are becoming more and more vocal about it on social media. When we scraped twitter searching for climate change, the data collected was massive. Hence we chose this topic.

    Keywords for search: **Climate change, climatechange and India**
- Dhyey: "Eco Friendly"
  - "Eco Friendly" is a word that is quite frequently used when there is a conversation about the environment going on. Here, we are assuming that the context in which the word "environment" is used, is not in terms of "work environment" or something similar, but in the context of nature and similar stuff. As it is already known, the word "Eco friendly" is used to describe something that is not harmful to the environment and the nature. It is normally associated with environment to promote environment friendly products. In many cases, the word is frequently used to market one's product over media platforms such as television or newspaper, and nowadays, even social media platforms are full of advertisements of such products and promotions. Since this word is quite closely related to the word environment, we decided to include this word as a subtopic for this assignment.

    Keywords for search: **Eco friendly and India**
- Ayush: "Floods"

- After humans' widespread exploitation of our mother nature began, one of the severe consequences of this massive destruction was floods. The word 'flood' has now become a common name in newspapers around the world. Whether it is India or Germany, no country in the world has remained untouched by this natural calamity. The primary reason attributed to the erratic time pattern of floods is climate change and deforestation. Coming to the India specific events, our country has witnessed floods every year, especially in the eastern parts of the country like Assam and Bihar. Due to the reasons mentioned above, we predicted that there might be many tweets available for this subtopic, which was later confirmed by the number of tweets that we scraped from Twitter.

  Keywords for search: **Floods in India, Floods and India**

2 Question 2

Scrap posts relevant to your topics from Twitter. Save the Tweet content along with its comment Tweets' content. Each Tweet is associated with metadata such as user info, geographical information, language tags, etc. Store the content as well as associated metadata into a CSV file.

The total number of tweets scrapped by our team is **142208**.

2.1 Link to the Drive

The link to the main folder containing the datasets along with the links to the individual CSV files for each subtopic is hyperlinked as follows :

Note: CLICK ON THE NAMES BELOW TO ACCESS THE CSV FILES.

- **Link to the main folder containing all datasets**
  - **Pollution dataset**
  - **Climate Change dataset**
  - **Eco Friendly dataset**
  - **Floods dataset**

3 Question 3

Answer the following questions based on your experience from the dataset curation.

3.1 Part (a)

What were the challenges in curating the dataset?

- The challenges involved in curating the dataset were encountered mainly in some buggy implementations in the Twint library. For example, sometimes, it stops fetching the tweets at some point, however on changing the "until" attribute for the Config class to a date previous to the date of the last tweet extracted, it again starts working fine.
- One problem encountered was in filtering with location. For example, when we tried to filter out the tweets according to the location, many times the program ran too long without fetching any tweets. One possible reason that we think, is that there are possibly fewer tweets where people are sharing the "location".
- Another thing that we tried was by applying the "near" attribute of the Config class to India, but that also yields few results, since on checking with certain profiles of the tweets fetched, it was probably the location that one mentions in their profile.
- Thus, there might be more people that share the city in that attribute, than the country. As a result, fetching by location was quite problematic. In our case, we tried to include the keyword "India" in the Search attribute to include the tweets that have India mentioned in them along with the subtopic of interest.

3.2 Part (b)

Does data curation from the different subtopics was equally easy/challenging? Argue.

- Some subtopics were highly discussed on twitter, while others not so much. In case of the latter, it was slightly annoying to scrape out larger chunks of tweets of a particular subtopic, if it was "sparse" with date (i.e. a smaller number of tweets available in a given time frame).

3.3 Part (c)

Are Geotags available for each Tweet? If not, how can we guess the geographical location of the place from where the Tweet was posted?

- No, the Geotags were not available for each tweet. However, in some cases, the CSV file generated contained a non-empty field for the "place" column, which contained coordinates in the form of a dictionary. From the coordinates, we can find out the geographical location of the user.

3.4 Part (d)

Are language tags always correct? Can you describe any four Tweets from your scraped collection, two each where tags are correct or incorrect?

- Most of the tweets that we saw had correct language tags, even when the language is not English. Like we can see for this tweet ***(link)*** and this tweet ***(link)***, which have been correctly identified as Indonesian and Italian languages respectively.
- But sometimes, twitter misinterprets the language. Also, when the tweets consisted of only hashtags, or only images/videos without any text, it marked the language as undefined.
- For example this tweet ***(link)*** has been tagged as ‘eu’, but it is actually written in a mix of English and Hindi. This tweet ***(link)*** has also been written in English and Hindi but has been tagged as undefined.

4 Question 4

What are top-10 popular hashtags and mentioned profile for each subtopic?
The top 10 popular hashtags for each subtopic are summarized in the table shown:

Top 10 hashtags			
Pollution	**Climate change**	**Eco Friendly**	**Floods**
pollution	climatechange	ecofriendly	india
environment	india	india	floods
plastic	environment	environment	dignityinfloods
india	climateaction	green	flood
waste	nature	gogreen	assamfloods
plasticwaste	climatecrisis	nature	assam
finance	globalwarming	sustainability	modiji_postponejeeneet
esg	climate	climatechange	covid19
bitcoin	sustainability	sustainable	indiaunitedtopostponejee_neet
ico	climateemergency	eco	climatechange

Table 1: Top 10 popular hashtags for each subtopic

The top 10 popular mentioned profiles for each subtopic are summarized in the table shown below:

Top 10 mentioned profiles			
Pollution	**Climate change**	**Eco Friendly**	**Floods**
pmo india	change.org india	narendra modi	narendra modi
narendra modi	narendra modi	change.org india	pmo india
moef&cc	prakash javadekar	pmo india	dr. ramesh pokhriyal nishank
prakash javadekar	moef&cc	flipkart	change.org india
central pollution control board	greta thunberg	youtube	himanta biswa sarma
ministry of health	un environment programme	ministry of health	sarbananda sonowal
arvind kejriwal	un climate change	ministry of housing and urban affairs	subramanian swamy
cnn travel	joe biden	amazon india news	amit shah
plastic finance	united nations	moef&cc	assam state disaster management authority
change.org india	licypriya kangujam	pib india	national testing agency

Table 2: Top 10 popular mentioned profiles for each subtopic

5 Question 5

List top-5 languages used for a subtopic. Can you see some strange language tags?

The top 5 languages used in each subtopic is shown as follows:

Top 5 languages			
Pollution	**Climate change**	**Eco Friendly**	**Floods**
English	English	English	English
Hindi	Hindi	Indonesian	Indonesian
Indonesian	Spanish	Tagalog	Estonian
French	Indonesian	Hindi	Tagalog
Estonian	Tagalog	Spanish	Japanese

Table 3: Top 5 languages for each subtopic

- Some of the strange language tags are the ones that are mis-classified in different languages. For example, consider this tweet ***(link)*** in the subtopic of "Eco Friendly" is classified as Tagalog language, however, it is actually Hindi, written in English script (Romanized Hindi). Another example is this tweet ***(link)***, again in the subtopic "Eco Friendly", which is mis-classified as Indonesian but is actually written in English.
- For the subtopic of "Floods", this tweet ***(link)*** is written in a mix of Romanized Hindi and English, but classified as Indonesian. Also, this ***(link)*** is another example of a tweet mis-classified as Tagalog, but written in a mix of Romanized Hindi and English.

6 Question 6

What is the monthly distribution of each subtopic? You need to prepare a line plot, X-axis: months. Y-axis: subtopic Tweet count.

The monthly distribution of each subtopic is shown in the plots below:

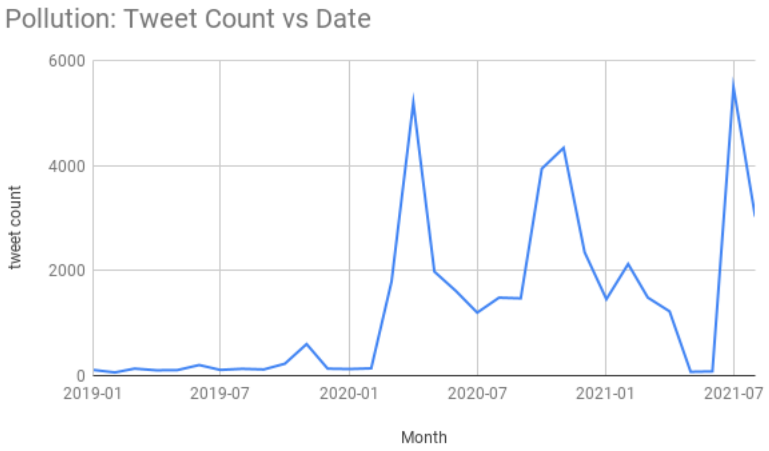

Figure 1: Plot of Tweet Count v/s Date for Pollution

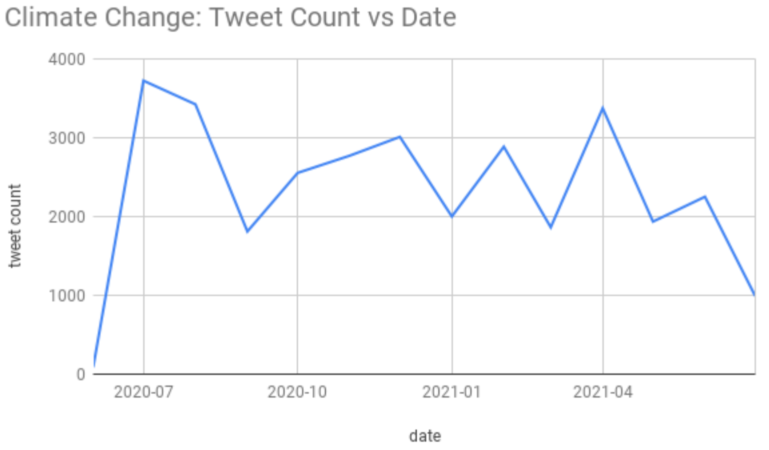

Figure 2: Plot of Tweet Count v/s Date for Climate Change

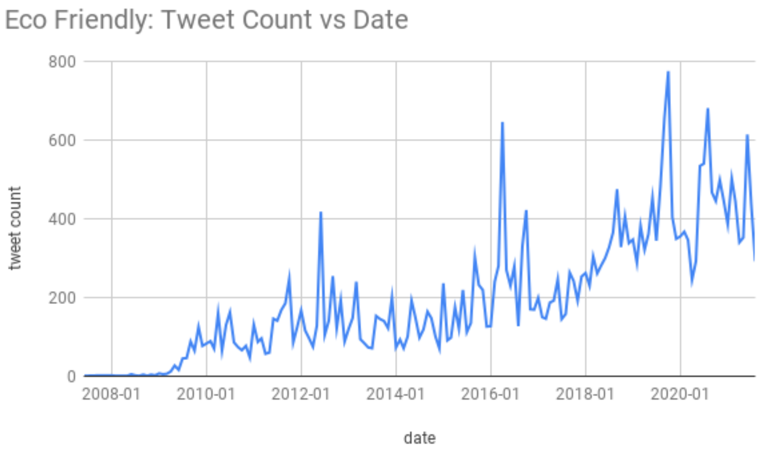

Figure 3: Plot of Tweet Count v/s Date for Eco Friendly

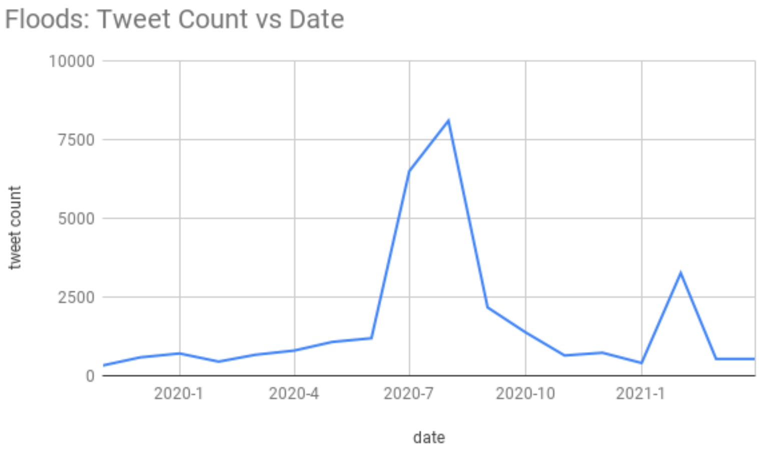

Figure 4: Plot of Tweet Count v/s Date for Floods

7 Question 7

Create a word cloud for each subtopic.

The word cloud for each subtopic are as follows:

7.1 Pollution

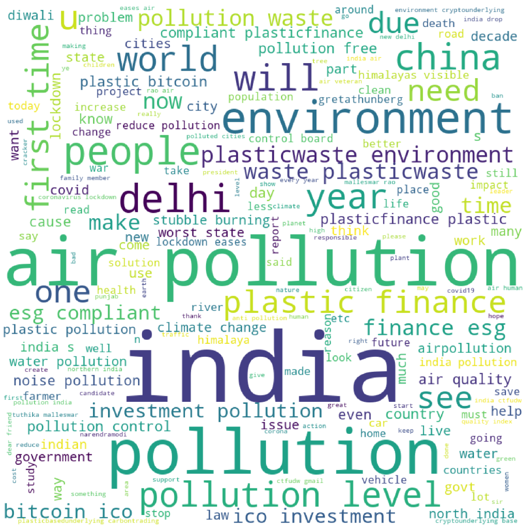

Figure 5: Plot of Word Cloud for Pollution

7.2 Climate Change

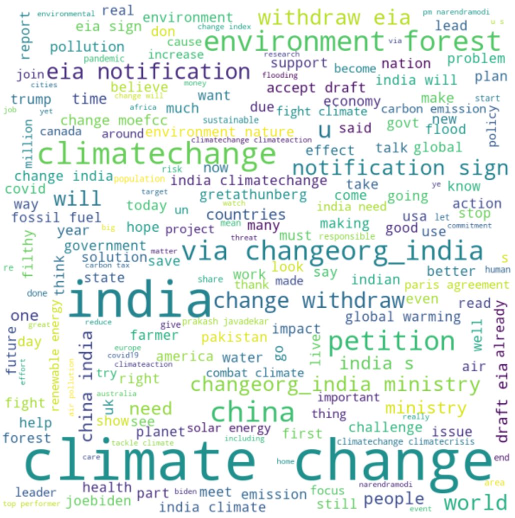

Figure 6: Plot of Word Cloud for Climate Change

7.3 Eco Friendly

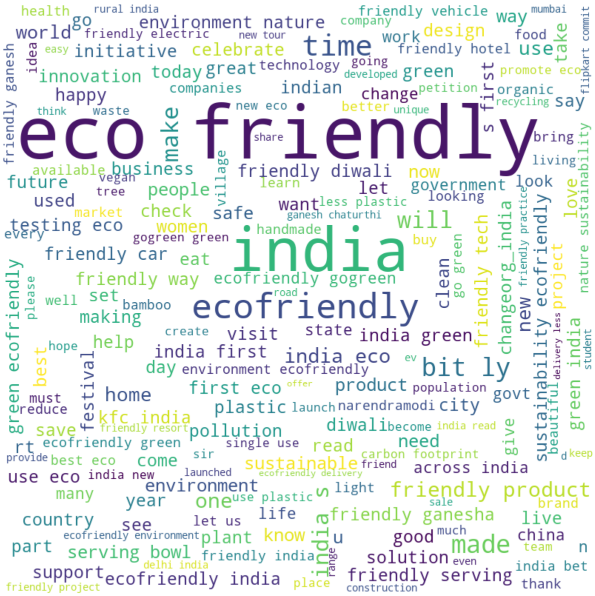

Figure 7: Plot of Word Cloud for Eco Friendly

7.4 Floods

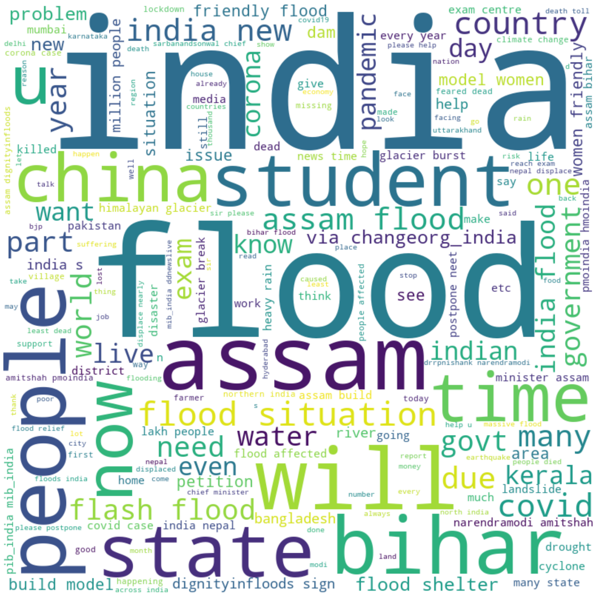

Figure 8: Plot of Word Cloud for Floods

8 Question 8

A table summarizing the total tweets, total mentions, total hashtags, total languages, etc., for each subtopic.

A summary of the different statistics from our dataset is shown in the table below:

	Subtopics			
	Pollution	**Climate change**	**Eco Friendly**	**Floods**
Total Tweets	46112	32673	33404	30019
Total Mentions	6955	6145	4866	3306
Total Hashtags	12682	11529	19235	6996
Total Languages	35	33	37	27
Total URLs	10697	11826	13912	8023

Table 4: Summary of the various subtopics

9 BIBLIOGRAPHY

References

[1] Twintproject. (n.d.). Twintproject/Twint: An advanced Twitter scraping amp; OSINT tool written in Python that doesn't use Twitter's API, allowing you to scrape a user's followers, following, tweets and more while evading most API limitations. GitHub. https://github.com/twintproject/twint.

[2] wordcloud. PyPI. (n.d.). https://pypi.org/project/wordcloud/.