Indian Institute of Technology Gandhinagar

CS-613 - Natural Language Processing

————————-:Report: Assignment 2:————-
Processing and Understanding Data

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 Sub-topic specific tasks

Each team member should select at least 100 random tweets (each Tweet should have at least five tokens after removing hashtags and mentions) from his/her curated corpus. Paste these random tweets in a spreadsheet and annotate the following for each tweet (create separate columns to answer the following questions in your spreadsheet).

The total number of random tweets selected in each subtopic from the curated corpus are as follows:

- Pollution: 128 tweets
- Climate Change: 129 tweets
- Eco Friendly: 127 tweets
- Floods: 130 tweets

1.1 Question 1

For each selected tweet, mark if the Twitter assigned language tag is correct. If not, what is the correct language tag (see list of tags here)? In case the text mixes multiple languages, assign a combined tag by separating them using a hyphen. For example, if your Tweet text mixes Hindi (either in Devanagari or Roman) and English tokens, your first answer will be 'No' and the second answer would be 'Hi-En'.

A snapshot of the spreadsheet with classification of the language tags for each subtopic is as shown below:

Tweet	Language		Correct (Yes) / Incorrect (No)	Correct Language tag	Main Language	Embedded Language
Air pollution, the silent killer was the fourth leadir	en		Yes			
Pollution in Ganga: NGT directs UP Jal Nigam tc	en		Yes			
I want to go out and run but pollution makes me	en		Yes			
WASTE WATCH: There's no better time than NC	en		Yes			
Pollution levels have plummeted so much during	en		Yes			
Dear , can you please keep an eye on today's F	en		No	En-Hi	En	Hi

Figure 1: Pollution

Tweet	Language		Correct(Yes)/Incorrect(No)	Correct lang tag	Main Languange	Embedded Language
The talk was part of a Faculty Develc	en		Yes			
Speaking at the 15th Summit I said,	en		Yes			
Vijay Sherawat from Youth For Clima	en		Yes			
India is filthy, says Trump while talki	en		Yes			
is still a peripheral issue in litigation:	en		Yes			
Greta ke umar ke bachhon ko log se	hi		No	Hi-En	Hindi	English

Figure 2: Climate Change

Tweet	Language		Correct (Yes) / Incorrect(No)	Correct Language Tag	Main Language	Embedded Language
the only deadly solution for all the pesky ! For more	en		YES			
SAVE EARTH WITH LOCAL! Earthen pots are on	en		YES			
Now in India protests are done only if the project ar	en		YES			
India's eco-friendly rechargeable LED study lamp b	en		YES			
1.5 lakh passengers use domestic flights in india d	en		YES			
Environment pe Gyan dena world envirc	tl		NO	HI	HI	EN

Figure 3: Eco Friendly

Tweet	Language		Correct (Yes) / Incorrect (No)	Correct Language Tag	Main Language	Embedded Language
During numerous floods in she o	en		Yes			
SIR,PREVIOUSLY I wrote ABOU	en		Yes			
Assam floods affect over 16 lakh p	en		Yes			
I am just waiting when Anushka cc	en		Yes			
JUST IN: China expresses condol	en		Yes			
I think in logo nai pesa acche se k	ht		No	Hi-En	Hi	En

Figure 4: Floods

1.2 Question 2

In case the Tweets are mixed ones. What is the main language (whose grammar is followed) and what is the embedded language/s (whose few tokens are embedded in the main language). For example, in the Tweet "items ko cart me daal ke app band kar dena is not funny", the main language is 'Hi' and embedded language is 'En'.

As shown in the figures 1-4, some of the incorrectly classified tweets have mixed correct language tags where **Hi** (Hindi) is the main language and **En** (English) is the embedded language. While, some tweets have **En** (English) as the main language and **Hi** (Hindi) as the embedded language.

The links to the CSV files of each subtopic containing the random tweets are as follows:

- **Pollution**: Link
- **Climate Change**: Link
- **Eco Friendly**: Link
- **Floods**: Link

The link to the main folder containing the files is : Link

2 Topic specific tasks

Now, the team has to combine the annotations from all teams members and answer the following questions:

2.1 Question 1

What is the percentage of tags that were incorrect in question 1 in Section 1? What could be some of the possible reasons for this error? Any suggestions to improve the language identification?

After combining the tweets from each of the four subtopics, the following statistics are obtained:

- Total tweets: 514
- Incorrectly classified tweets: 13
- Percentage of incorrect tags: 2.53%

Most of the language tags incorrectly determined were those having a small phrase in a language other than English. One of the reason might be the inability of twitter to assign multi-lingual language tag. Other errors were in those tweet where the complete tweet was in Hindi language but the tweet was written in Roman script. This can be because the twitter may not be trained to detect Hindi language from scripts other than Devanagari.

One of the possible suggestions to improve language identification could be that if we can train the twitter language classifier on large Romanized Hindi dataset and classify the language with a new tag. Similar steps can be followed in the case of false language detection in tweets having two different scripts embedded (example Roman and Devanagari) in it.

2.2 Question 2

Do you think your data curation method (in assignment 1) was biased to some specific Indian languages? Why and why not?

Among Indian languages, we found the Hindi **Hi** was the dominant language among other Indian languages. The reason for this might be that the topic of environment is not confined to a particular region of the country, and also it is more discussed in the urban parts of the country where the primary means of communication is either English or Hindi rather than regional language.

2.3 Question 3

Can you give some examples of neologisms from your data?

Some of the neologisms obtained among the 514 tweets are:

- **Godi media** (which is typically used as a slang for media biased towards BJP government).
- **Pappu** (is used as a satire towards a particular politician of the Congress party).
- **Tamashas** (is a Hindi word used here as a slang to signify showmanship).
- **Bhumafia** (it means a group of criminals that operate in land dealings).
- **Climate Smart** (used to describe cities taking appropriate measures against climate change).
- **Blue gold** (refers to water and highlights its importance).

3 Statistical Analysis of Entire Curated Data

The next task is to conduct a statistical analysis of the entire curated data (not just the annotated data). The following are the tasks:

3.1 Question 1

Create a frequency list of tokens present in the dataset in the decreasing order. What are the top-20 words? What is the percentage of stop-words in the top-20 rank list?

After creating a frequency list of tokens present in the dataset, the top-20 words in the decreasing order of frequency are as follows:

Word token	Frequency
in	130439
the	123875
india	112074
to	91937
of	90205
and	80052
is	54256
a	46935
pollution	45812
for	39056
s	32272
climate	28892
on	28250
are	27549
friendly	27101
change	26708
eco	26275
amp	25407
it	23069
we	21150

Table 1: Top 20 words in the frequency list of tokens

Out of the top 20 words in the frequency list, 13 of them are stop-words. Hence, the percentage of unique stop words in the top-20 list is **65%**

3.2 Question 2

Does the data follow heap's law? Show by plotting $|V|$ vs N. What are the values of curve fitting parameters K and β?

Heap's Law:

- Let $|V|$ be the size of the vocabulary and N be the number of tokens in the dataset, then according to Heap's law:

$$|V| = KN^{\beta} \tag{1}$$

Yes, the data follows Heap's law. The plot of $|V|$ vs N is shown below:

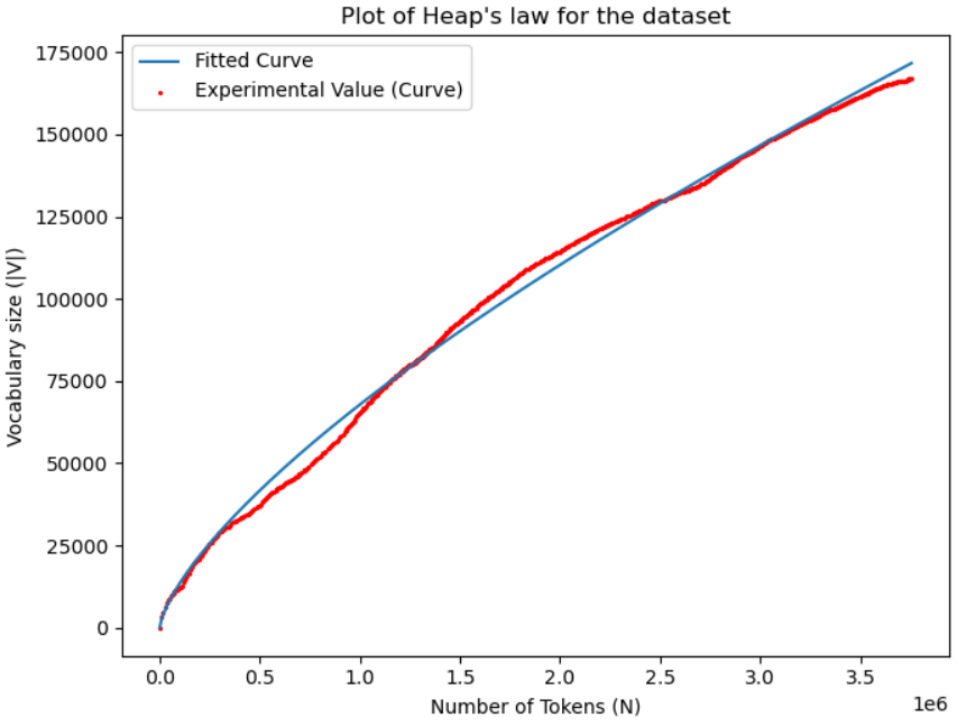

Figure 5: Plot of Heap's Law

The values of the curve fitting parameters are:

- K: 4.194
- β: 0.701