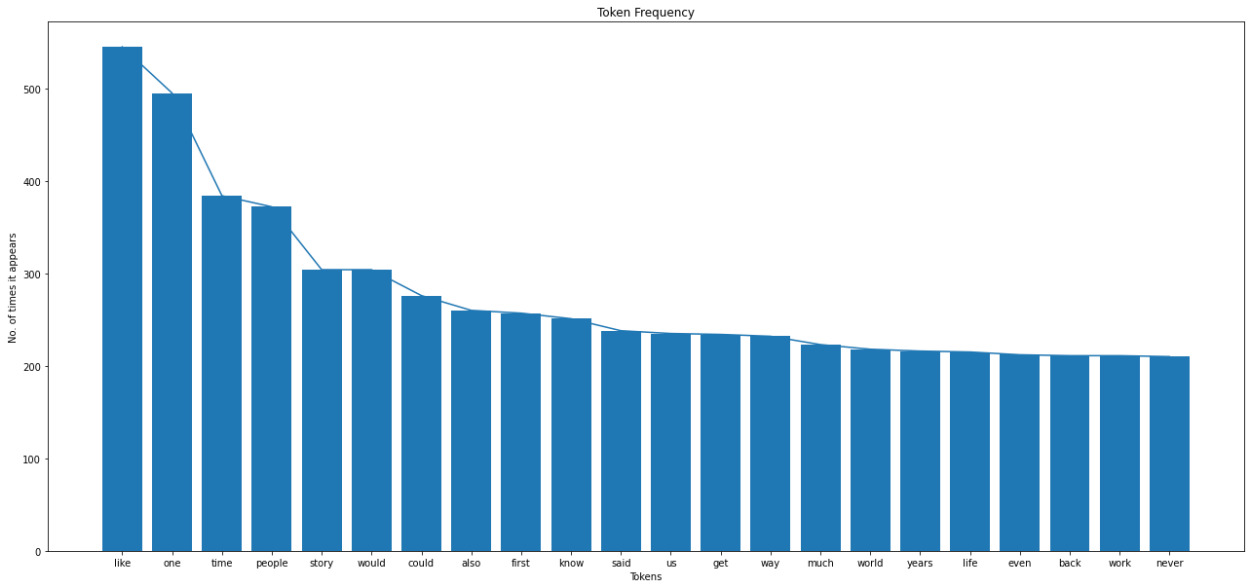
Computational Linguistics-1

Mini Project - Analysis

English corpus

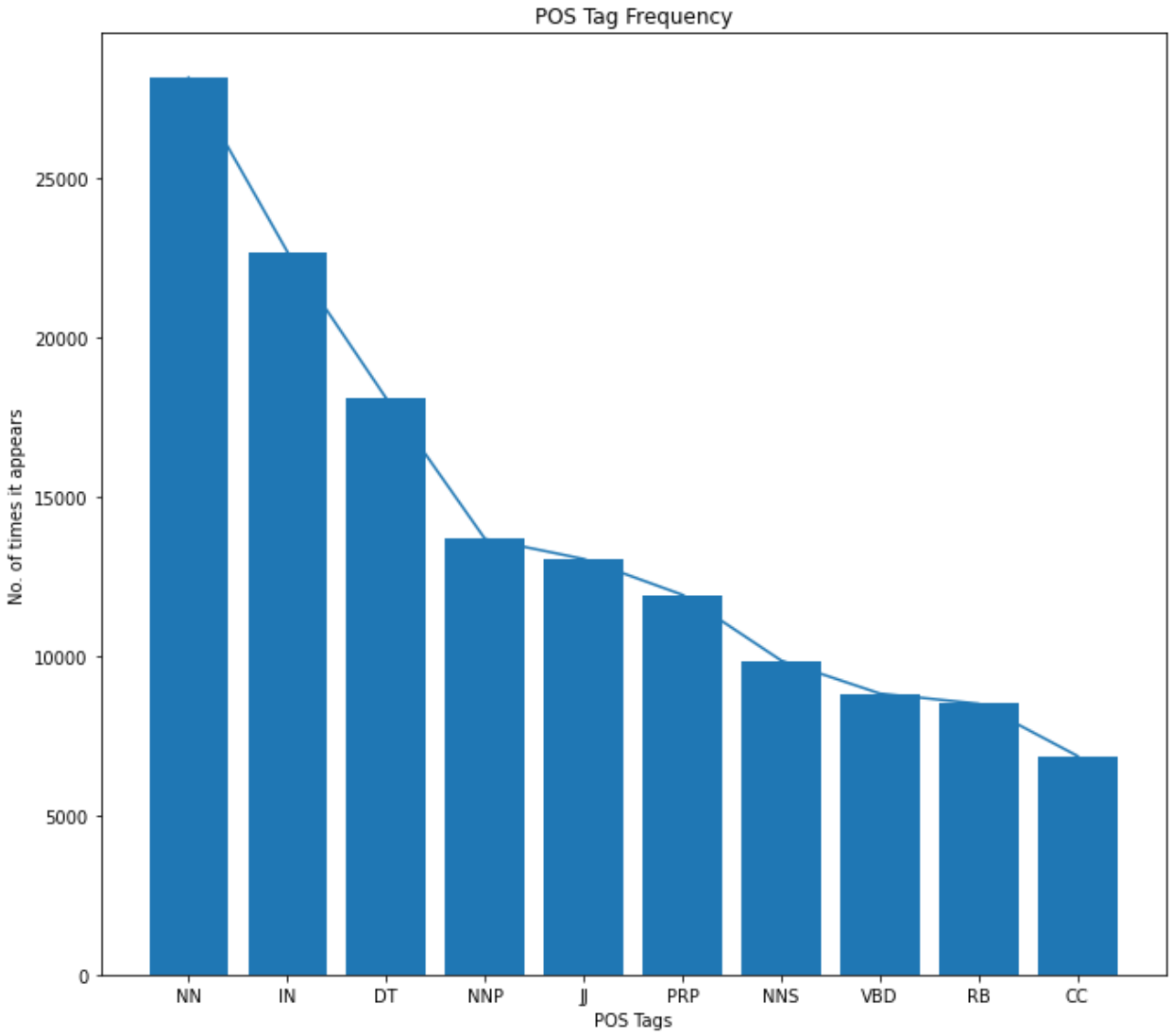
I scraped the website https://longreads.com/ for the corpus, so I expect the data to be more story/article like. I used the NLTK library for processing.

Token frequency



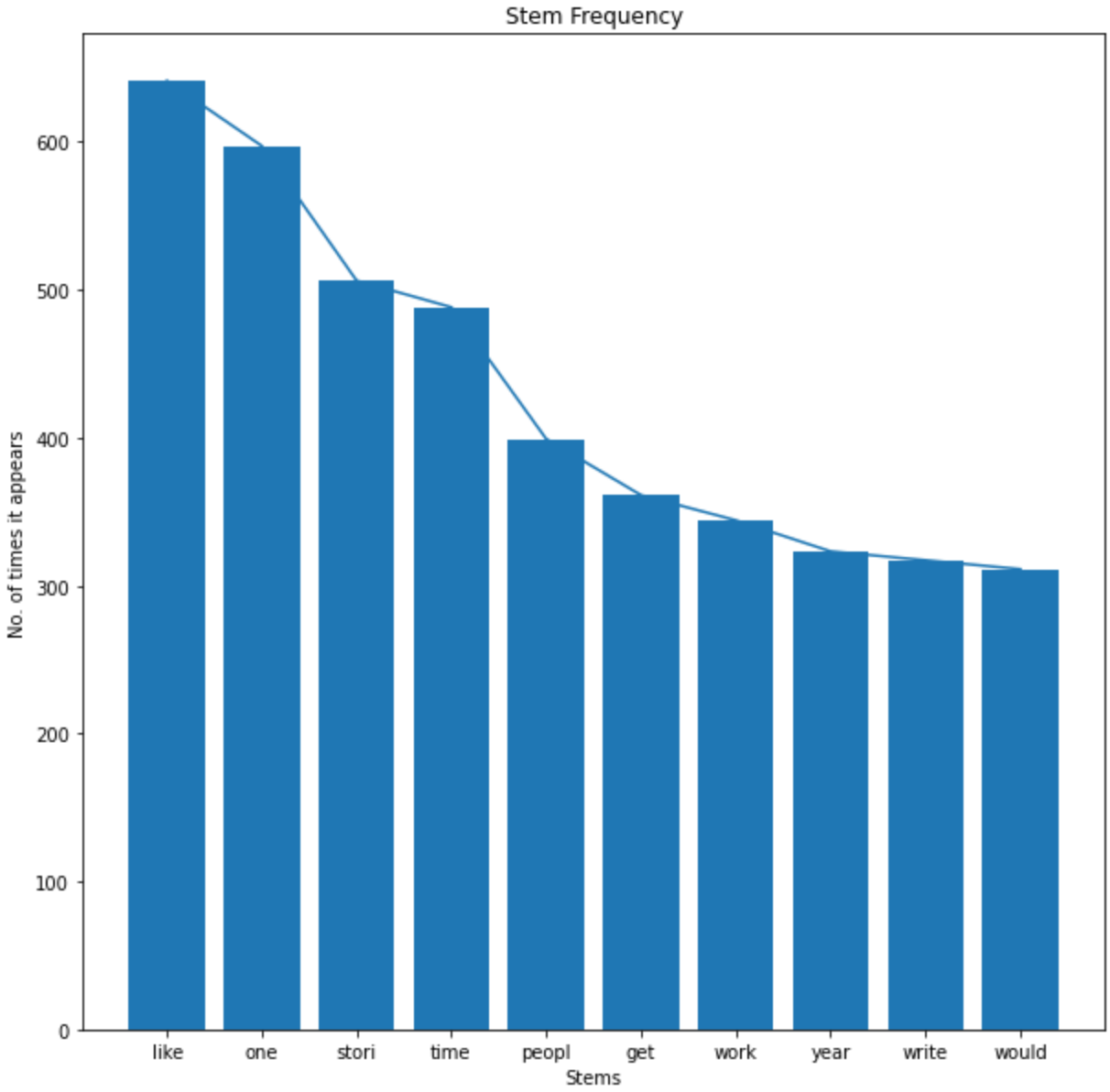
As we can see in the above graph, words like ‘story’, ‘people’, ‘said’, ‘life’, etc. are amongst the most frequently occurring ones. This fits with our expectations of the frequent words being words you would read in a story. I’ve taken 22 words because the gap between words #22 and #23 is big enough that we need not focus on the subsequent words

Parts-of-Speech Tag frequency

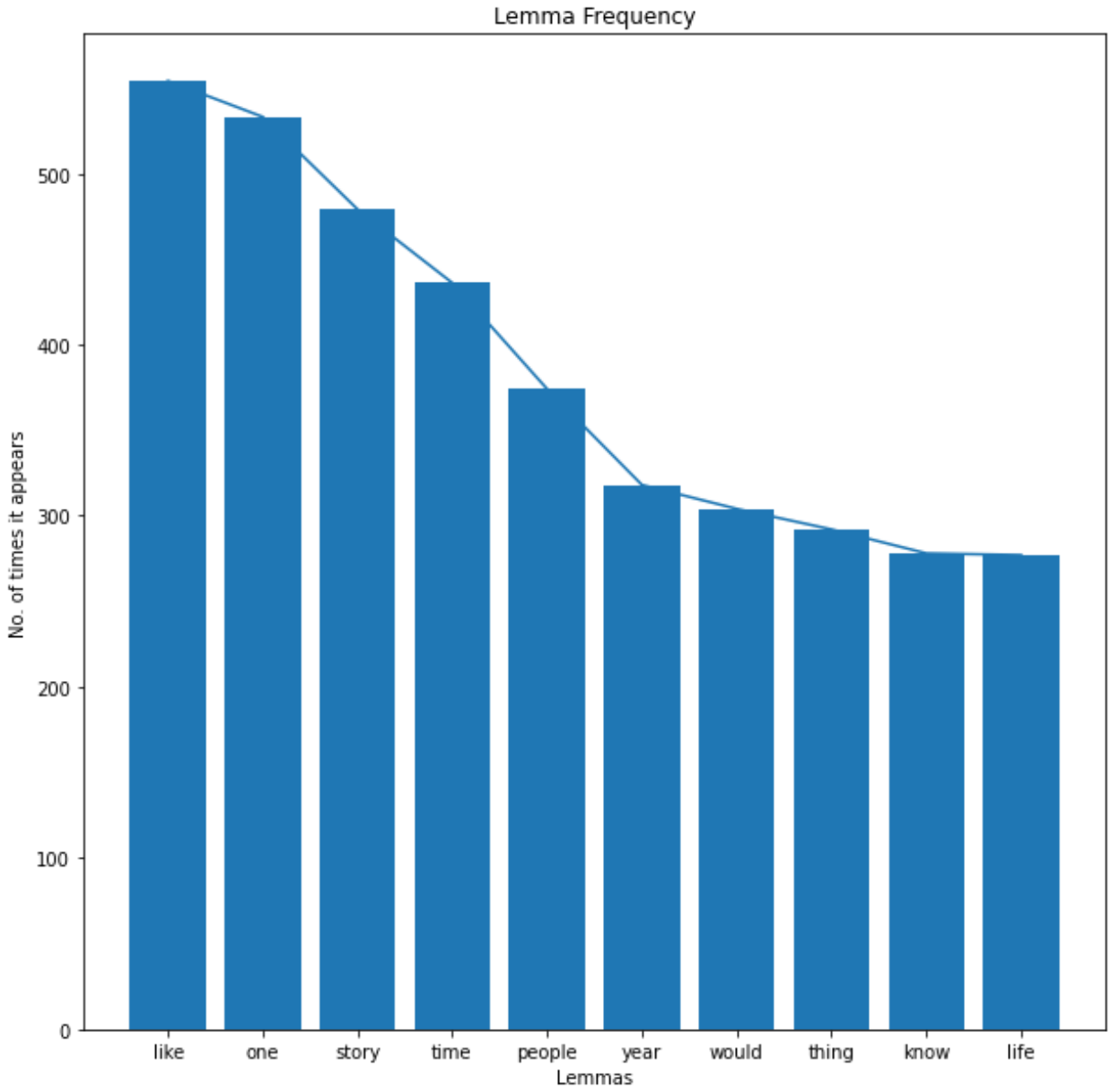


As expected, NN(Nouns), IN(Prepositions), and DT(Determiners) dominate the corpus.

Stem frequency



Lemma frequency



Word Cloud

Algorithm

My algorithm uses the ratios of the difference between the frequencies and the font sizes. Using that, you solve for fontsize using

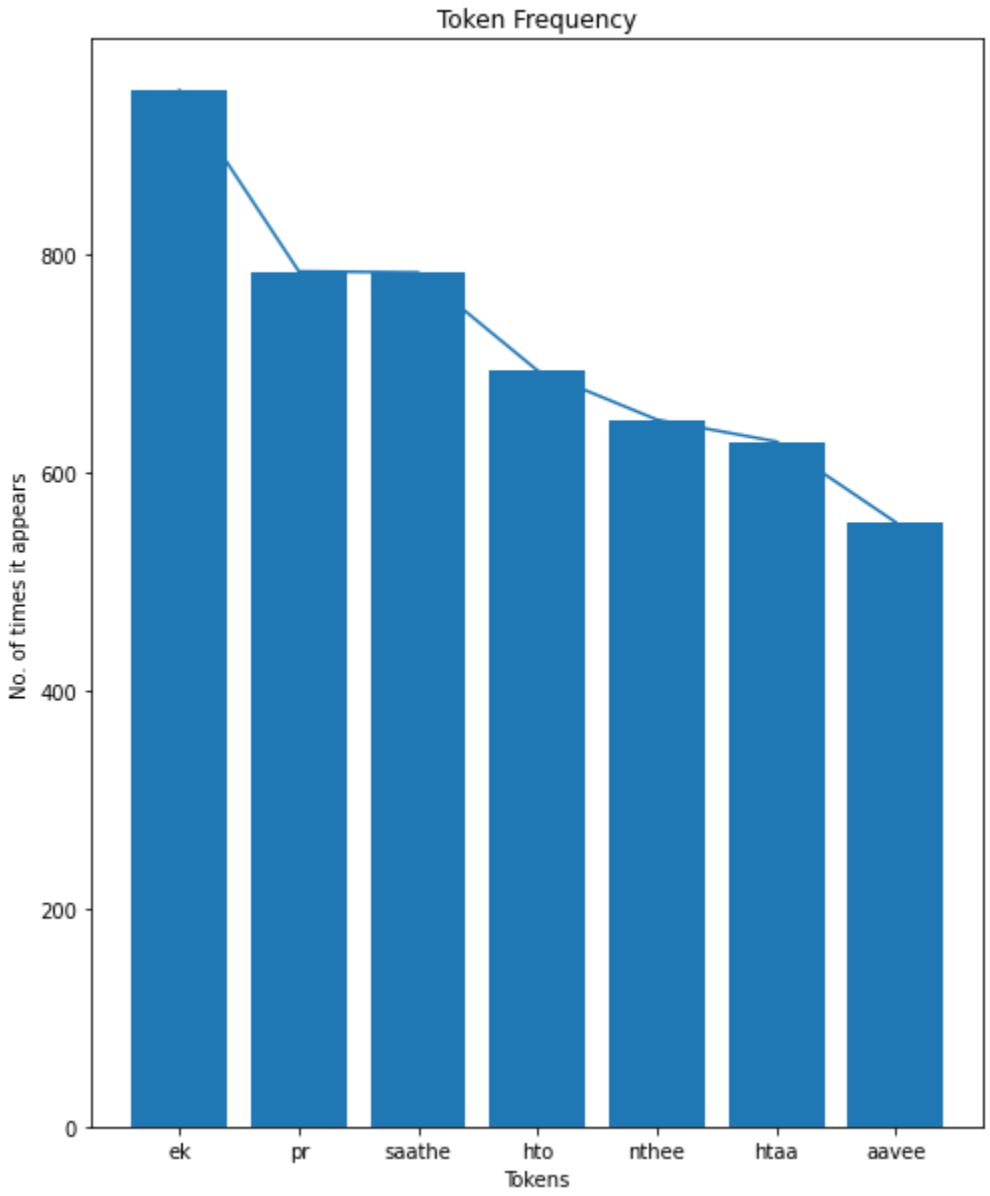
$(maxfreq - freq)/(minfreq - freq) = (maxfontsize - fontsize)/(minfontsize - fontsize)$

However, I have not used this algorithm for the word cloud visualization. I have used the WordCloud python library for the purpose.

Gujarati corpus

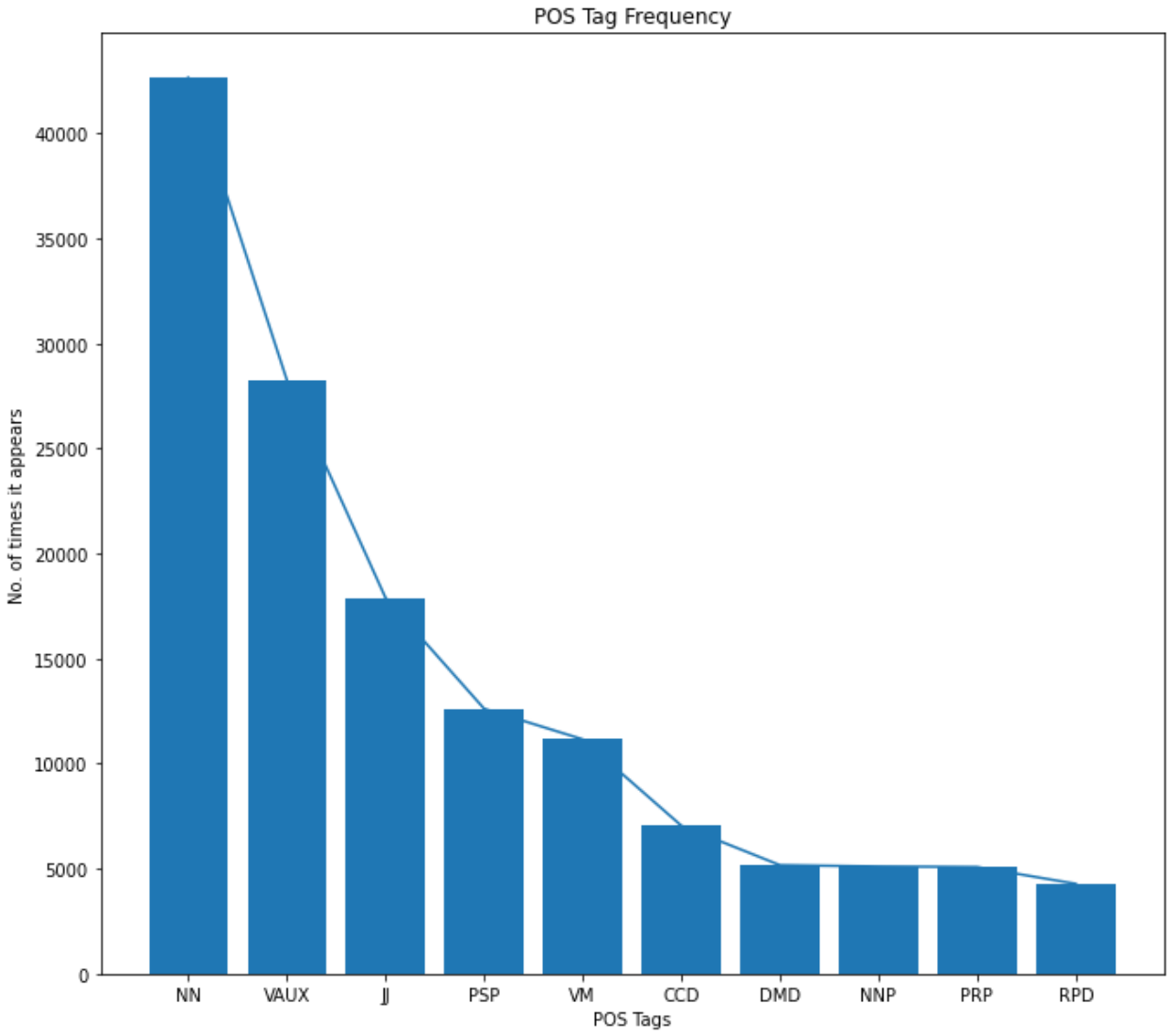
I scraped the website https://navgujaratsamay.com/ for the corpus. I used this GitHub library for processing: https://github.com/Rutvik-Trivedi/Gujarati-NLP-Toolkit

Token Frequency



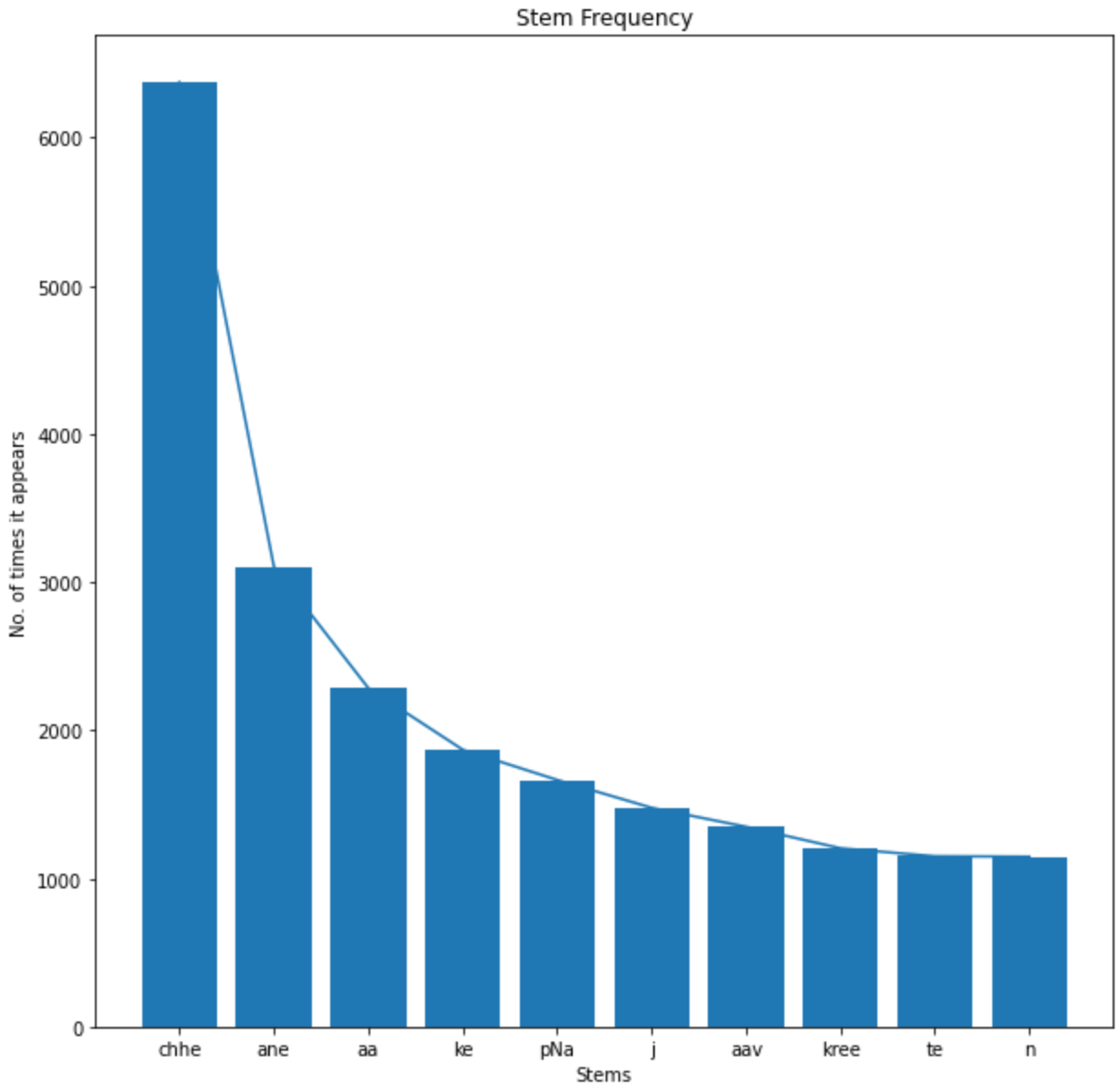
The most frequently used words(’ek’,’saathe’,’hto’,’htaa’,’aavee’) are of the reporting kind, which makes sense considering I scraped a news website.

Parts-of-Speech Tagger



One thing to note is that news articles generally contain a lot of nouns, which can be seen in the graph.

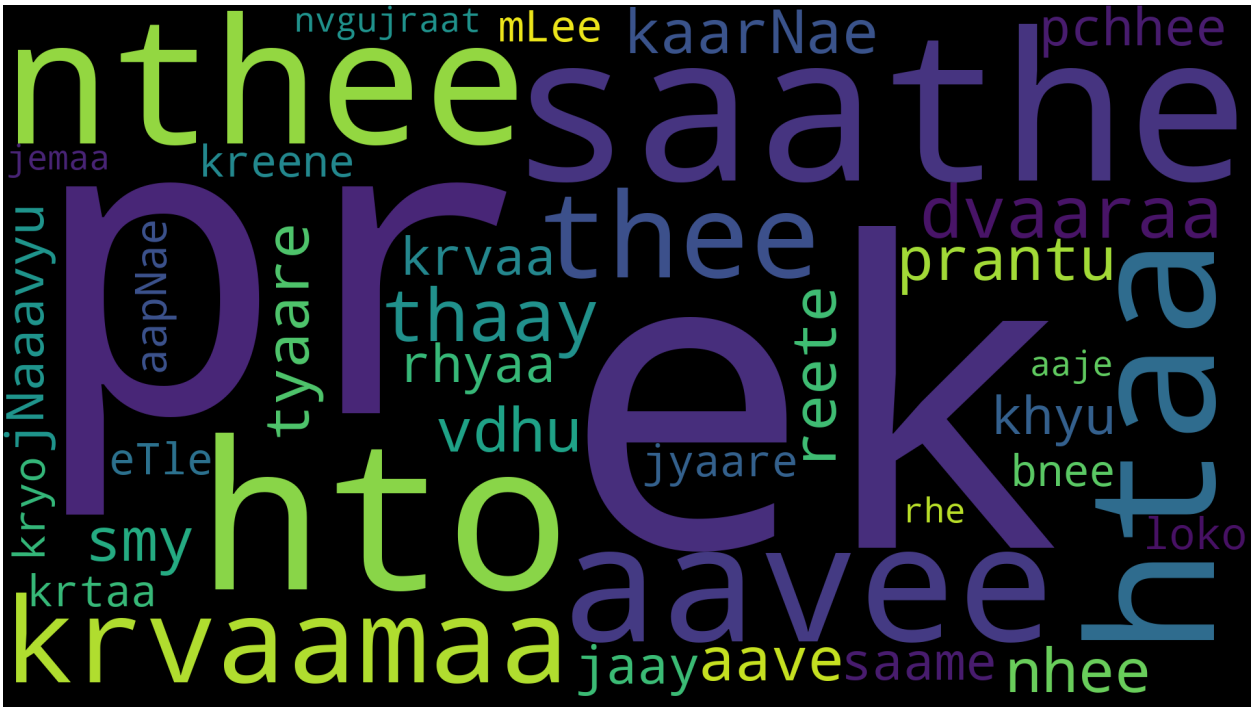
Stem frequency



This graph is quite close to a graph which follows Zipf's law.

> 📌 I couldn't find a Gujarati lemmatizer anywhere, so I haven't performed that operation in this report.

Word Cloud



- Most of the output is printed in the Jupyter notebook, and the corpus and its cleaned and processed forms are stored in text files
- The word clouds have been attached separately
- Files pertaining to Gujarati have names beginning with ‘g’
- File names are self-explanatory

📚 Aryan Chandramania
2021114004