PROJECT-1

BY:YASH AGRAWAL

ROLL NUMBER: 2020114005

COMPUTATIONAL-LINGUISTICS-1

CONTENTS:

Corpus Collection

- For both English and Hindi, we are using BeautifulSoup to crawl through the URLs of wikipedia that are inputted in the crawl.py and it extracts all the data present in the <p> tag of html and the output is redirected to text_with_stop.txt for further use.
- removfor.py is used to remove foreign words in both English and Hindi. In english we are using the nltk library and for Hindi we are using the inltk library.
- punctuation.py is used to remove punctuations from our corpus. We use a basic python code to print everything except the punctuations back into the file from where we took input

WORKING ON THE CORPUS

1) **Tokenization:**
    - For Hindi we are using inltk library for word tokenization.
    - For English we are using nltk for word tokenization and punkt for sentence tokenization.

2) **POS Tagging:**
    - For Hindi, we are using the stanza library of Stanford NLP. Though it's a bit slow as compared to the nltk library, it completes the POS tagging of the entire dataset while nltk can only be used for a small data set consisting of only around 2000-10000 words.
    - For English, we are using nltk to do POS tagging.

3) **Remove Stopwords:**
    - For Hindi, we used the spacy library to remove the stopwords
    - For English, we used the NLTK library to remove stopwords and then store it in a file named nostop.txt

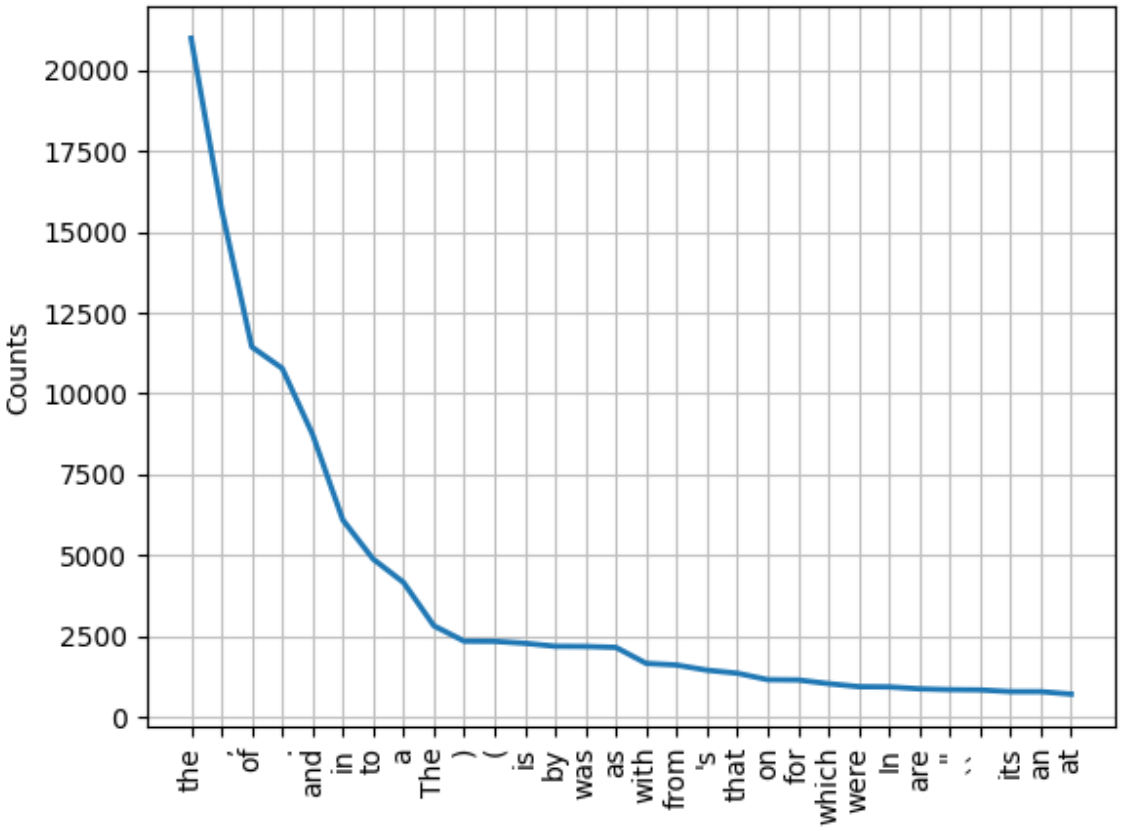
4) **Stemming:**
    - For English, we used PorterStemmer of NLTK to stem through the data that is provided to us from nostop.txt
    - For Hindi, we are using spacy library to stem through the words.

5) **Lemmatization:**
    - For Hindi, we are using stanza library by Stanford NLP to lemmatize the data. Though it's very slow, it does the job.
    - For English, we are using WordNetLemmatizer library of nltk to lemmatize through the data.

ANALYSIS OF THE ABOVE TASKS AND WORDCLOUD

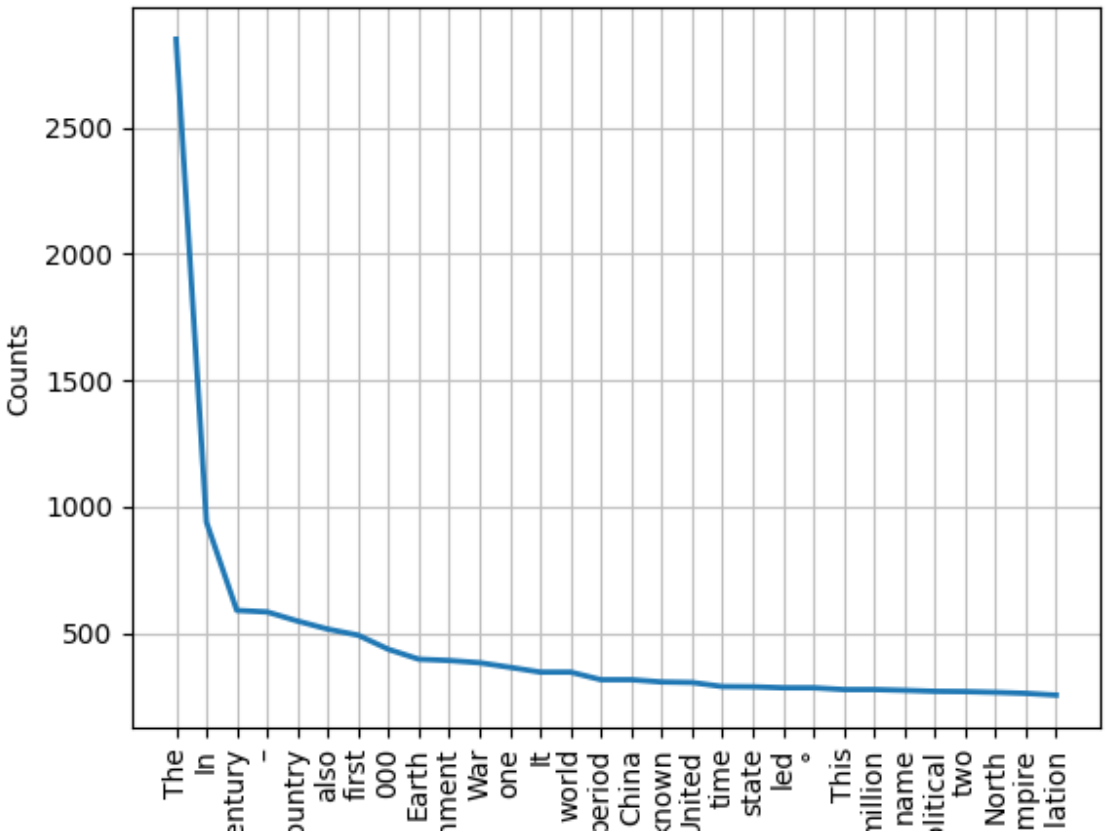
1) ENGLISH:

- Frequency graph of the data set without removing stopwords



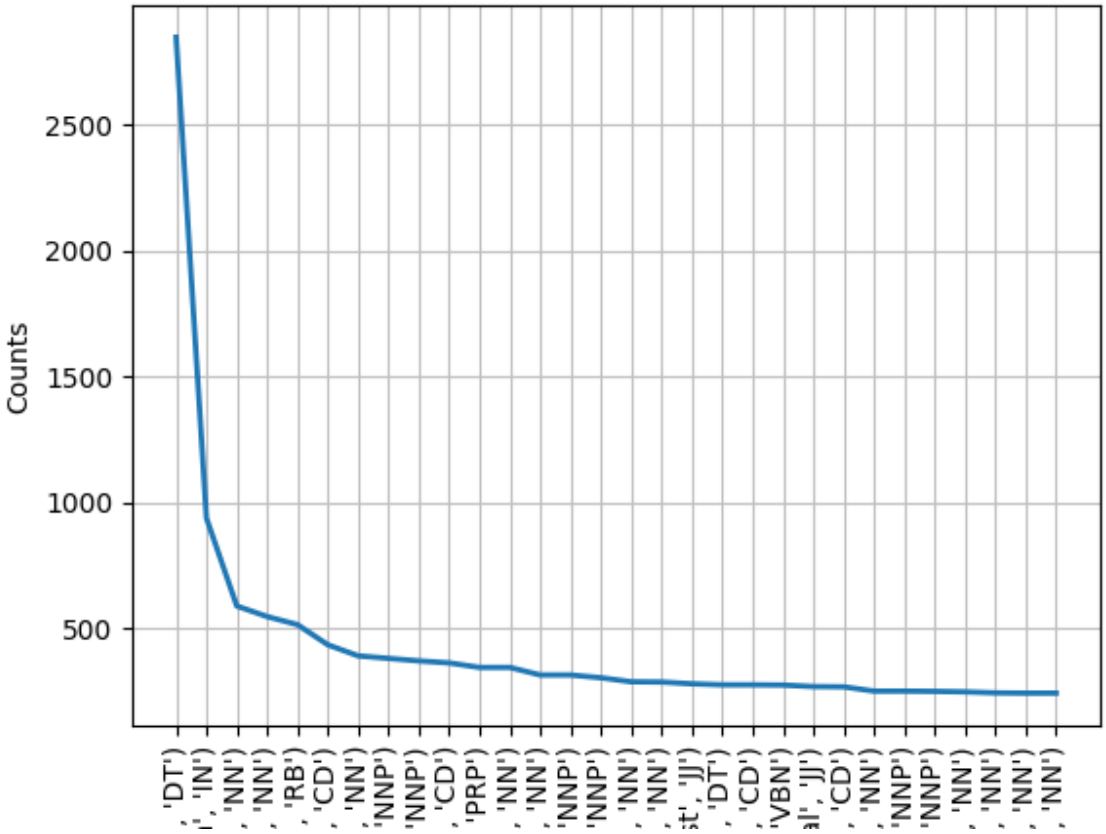
→ “the” is the most common stopword that we can see in the corpus and it occurs more than 20000 times in the 10K sentence corpus we have.
→ as this corpus has not been cleaned and is in it’s most original state, we can see that among punctuations, ‘,’ is the most common with >15000 occurrences, followed by ‘.’ and the with ‘(‘ and ‘)’

- Frequency graph of the data after removing stopwords

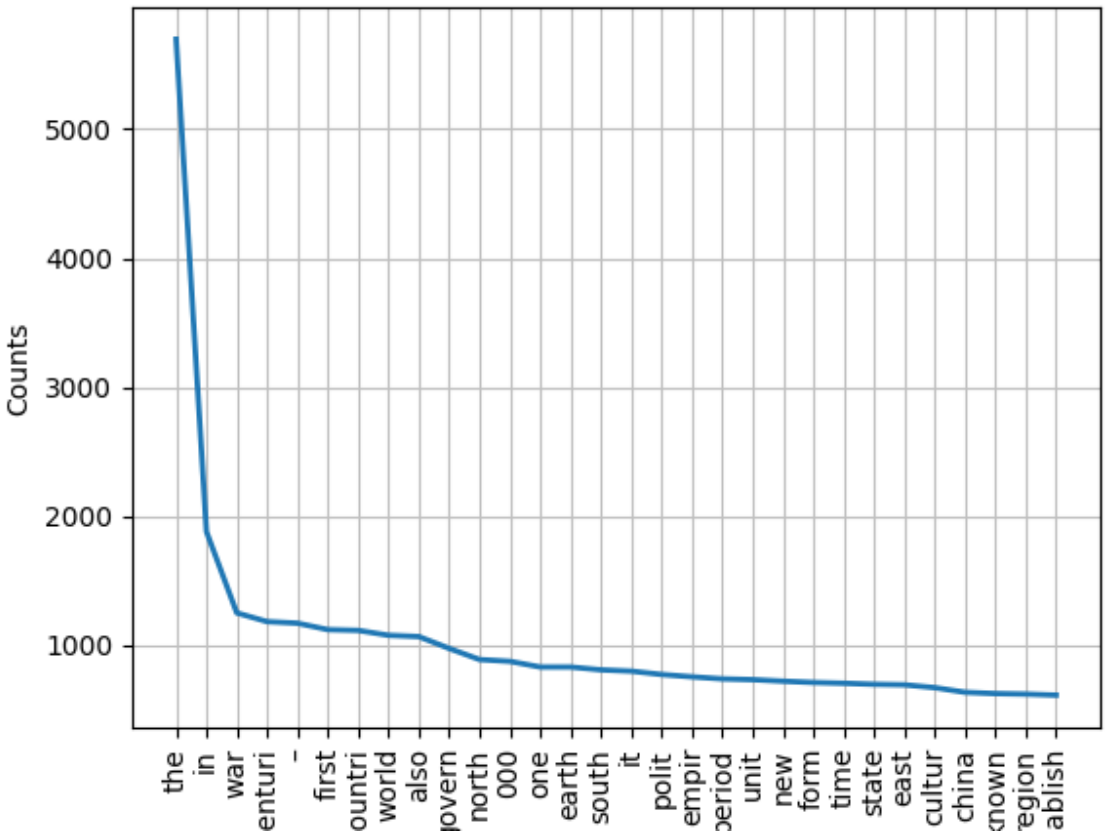


→ This graph was made after the removal of stopwords and the punctuations and we can see that “The” is the most common (probably because ‘the’ is in stopwords but “The” isn’t) with more than 2500 occurrences.
→ As numbers were not removes from the data set, we can see the ‘000’ is the most common number here with almost 500 occurrences

- POS frequency graph

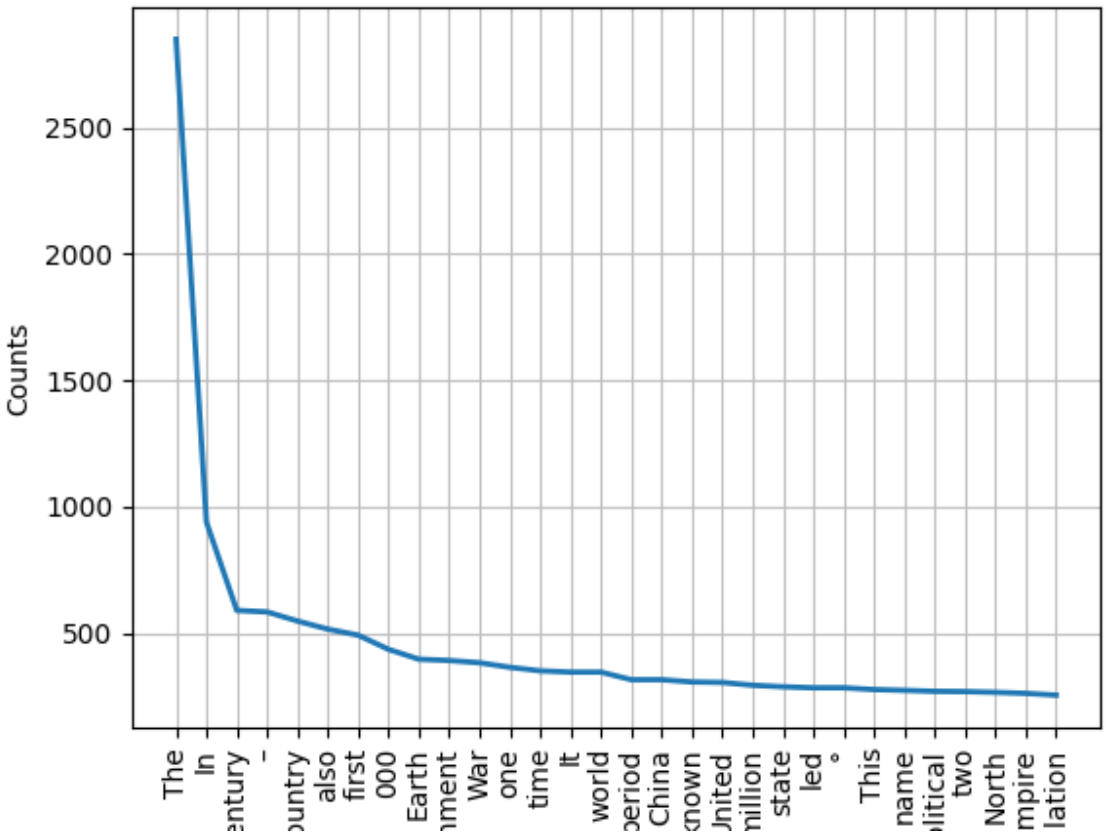


- Stemming frequency graph:



→ This is the frequency graph depicting the number of occurrences of a stemmed word in the corpus. ‘the’ still comes at the 1st position with >5000 cases followed br ‘in’ which has a little less than 2000 cases.

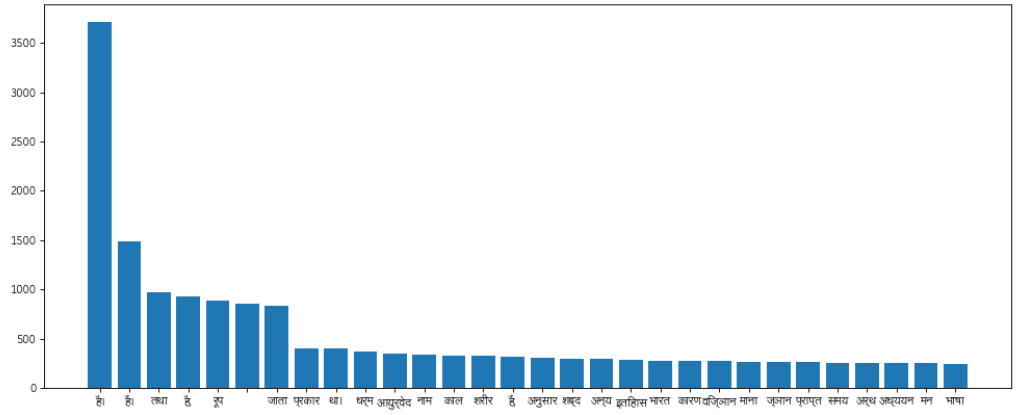
- Lemmatization frequency graph



→ This is the graph of the lemmatized data frequency. At first glance it's awfully similar to the frequency graph of the corpus without stop words, but there are some differences such as the presence of 'time' after 'one' which is not present in the previous one.

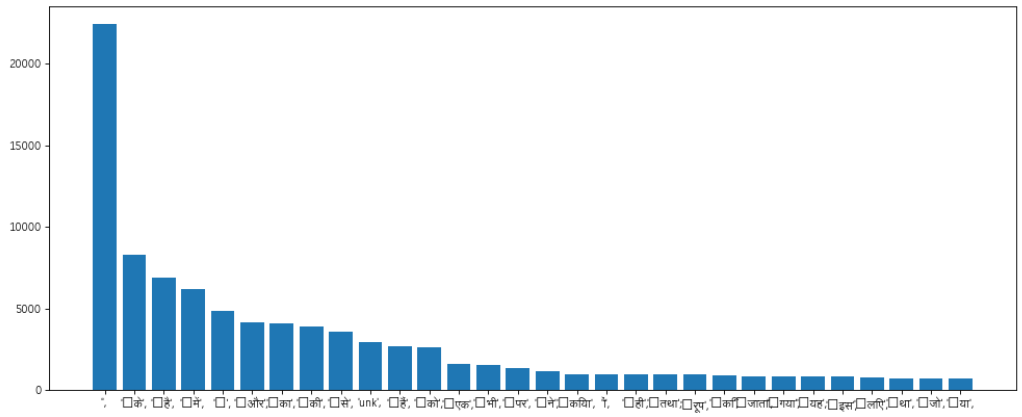
2) Hindi:

- Frequency graph of the data set without removing stopwords



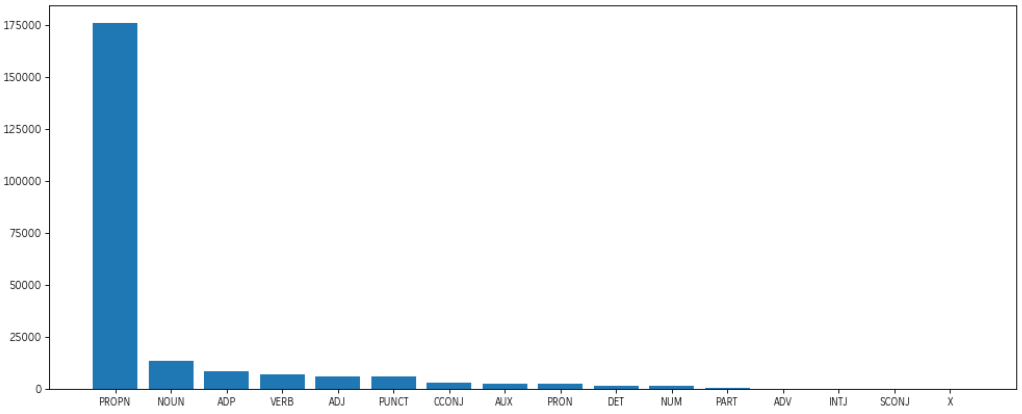
→ This is the graph of the original data when no cleaning has been done. We can see that ‘ है।’ is the most common word with > 3500 occurrences followed by ‘हैं।’ which has ~1500 cases. This shows the importance of ‘है’and ‘हैं’ in Hindi language.

- Frequency graph of the data after removing stopwords



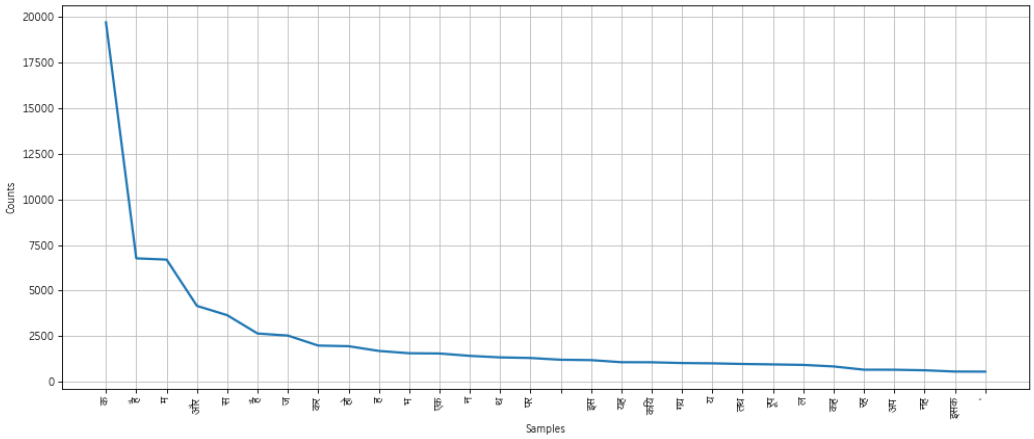
→ Here, we have removed all the foreign words and cleaned the corpus. As space could not be cleared, the highest occurrence is the occurrence of “ “, followed by ‘_के’.

- POS frequency graph



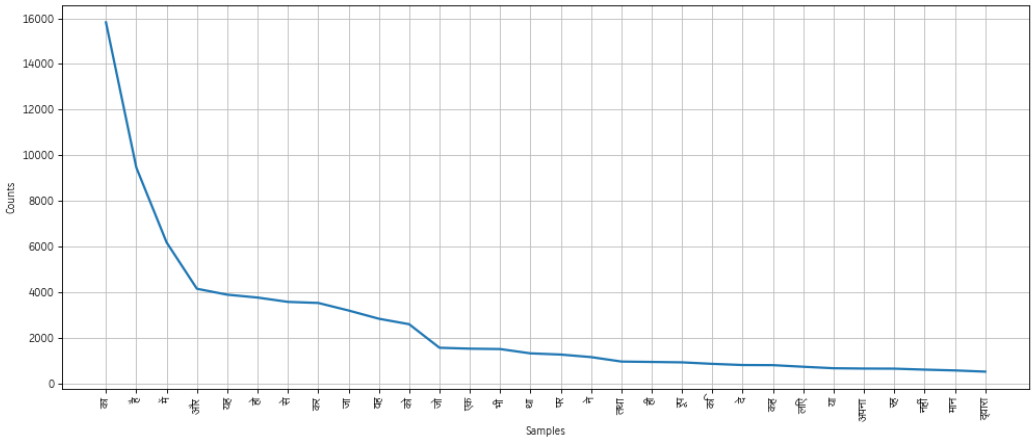
→ Here we can see that the majority of words in our corpus are Proper Nouns with more than 175000 cases, this is followed by Nouns and there are very less occurrence of Adjectives in wikipedia documents

- Stemming frequency graph:



→ This is the graph for stemming and as we can see, ‘क’’is the most common stemmed word in our corpus from wikipedia which shows that even though stemming is faster than lemmatization, the probability of getting errors also increases.

- Lemmatization frequency graph



→ Here, as we can see that the highest occurrence of lemmatized word is 'का’

which is the correct lemmatization because on stemming we were getting ‘क‘ instead of ‘का‘.

3) WORD-CLOUDS

- **ENGLISH:**

→ Here, we have used 75 most common words to make the wordCloud using the WordCloud library of python. We have taken 75 most common words because any more than this causes the wordcloud to look congested and it’s difficult to look through the data in it. Moreover, using only these many words we can see the most important and the most frequent words used in wikipedia.

- **HINDI:**

→ Here, we used “Lohit-Devnagiri” font to write the wordCloud and used 50 words to be printed in it because Hindi words also have मात्रा in it and anymore text would be difficult to manage and analyse through the data-set. Moreover, using only these many words we can see the most important and the most frequent words used in wikipedia.