MINI PROJECT

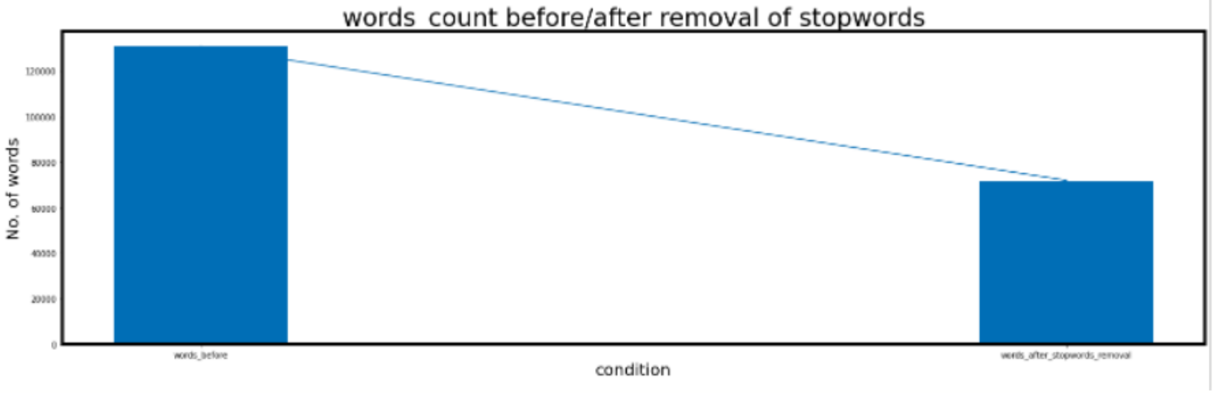
Computational linguistics

NAME:- Suyash Sethia
Roll No:- 2021114010

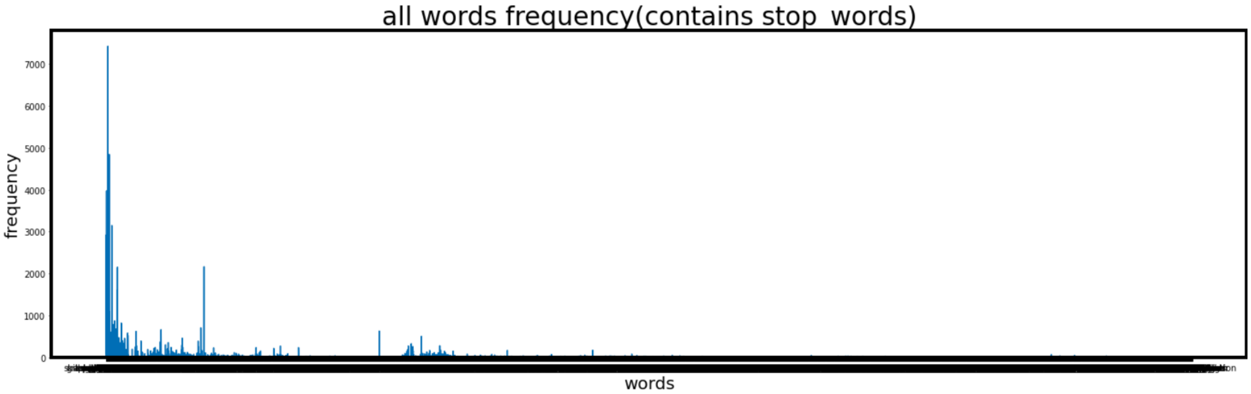
PART 1 :- English Text

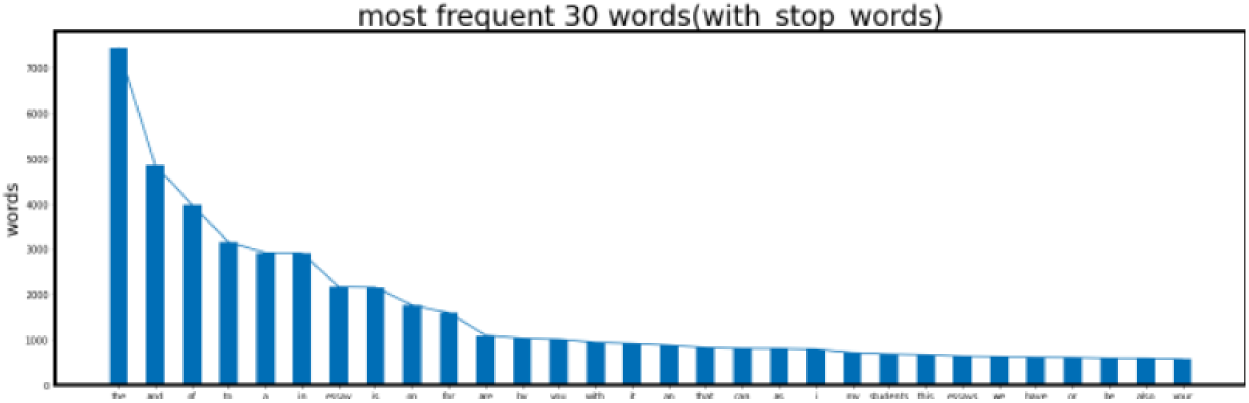
Steps

1. First download necessary libraries
2. Then extract text from all the links and print the raw text
3. Cleaning of corpus is done by removing punctuations and convert whole text into lower case
4. Tokenization of words and sentences
5. POS tagging is done on all words
6. Words without stop words are stored and printed
7. Lemmatization is done on all words
8. Stemming is done on all words
9. Calculation of frequency of all words
10. Calculation of frequency of all words after removal of stop words



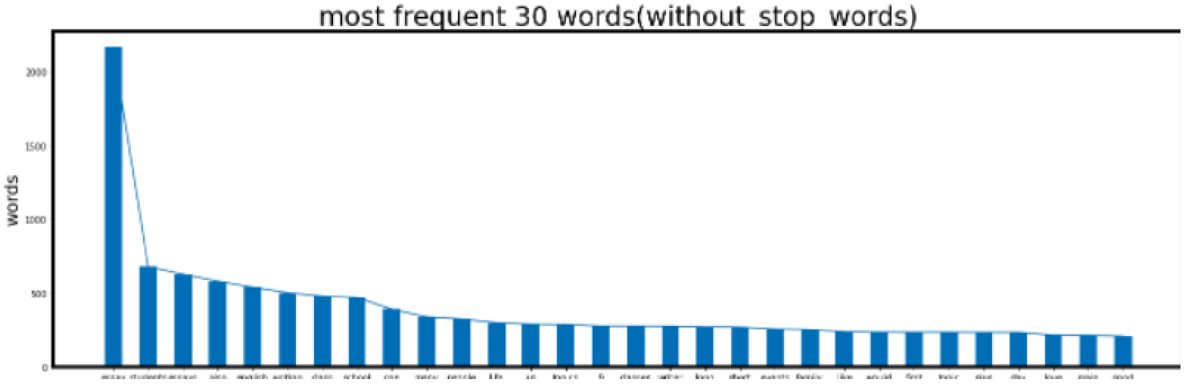
- Graph between word count before and after removal of stop words shows that almost 50% of the words were stop words





11. Graph of frequency of words shows that most frequent words are the stop words like the ,and ,of ,etc

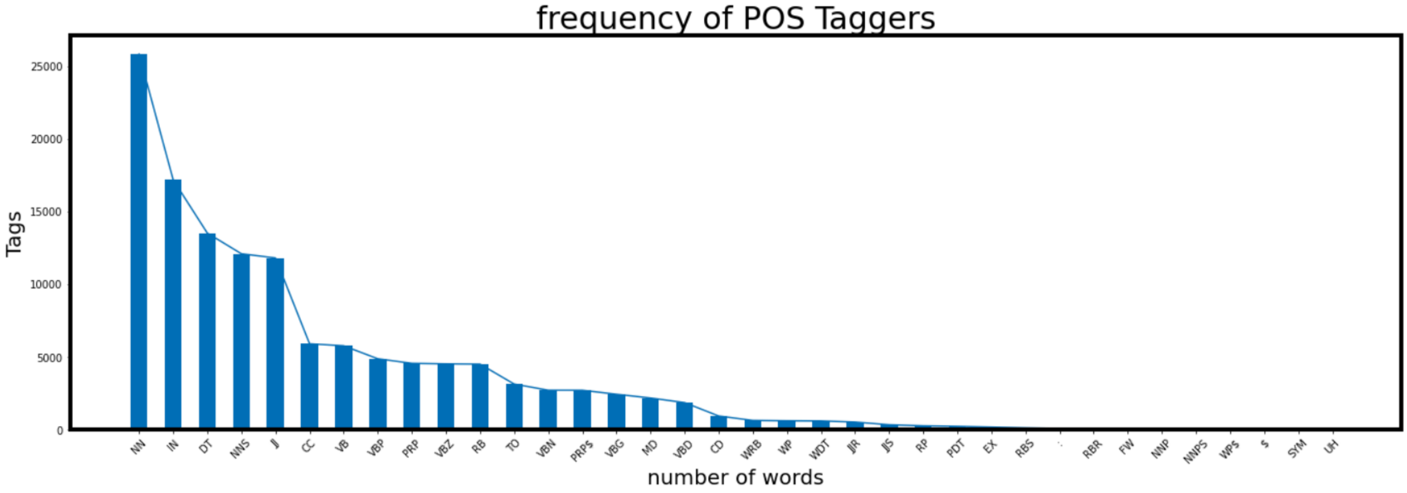
12.



13. Graph of frequency of words shows that most frequent words are essay,students ,english . as the corpus was related to essays and

studies and the website used is basically has essays on many topics so word “essay” has exceptionally more frequency

14.



15. Graph is made showing number of words for every POS tag

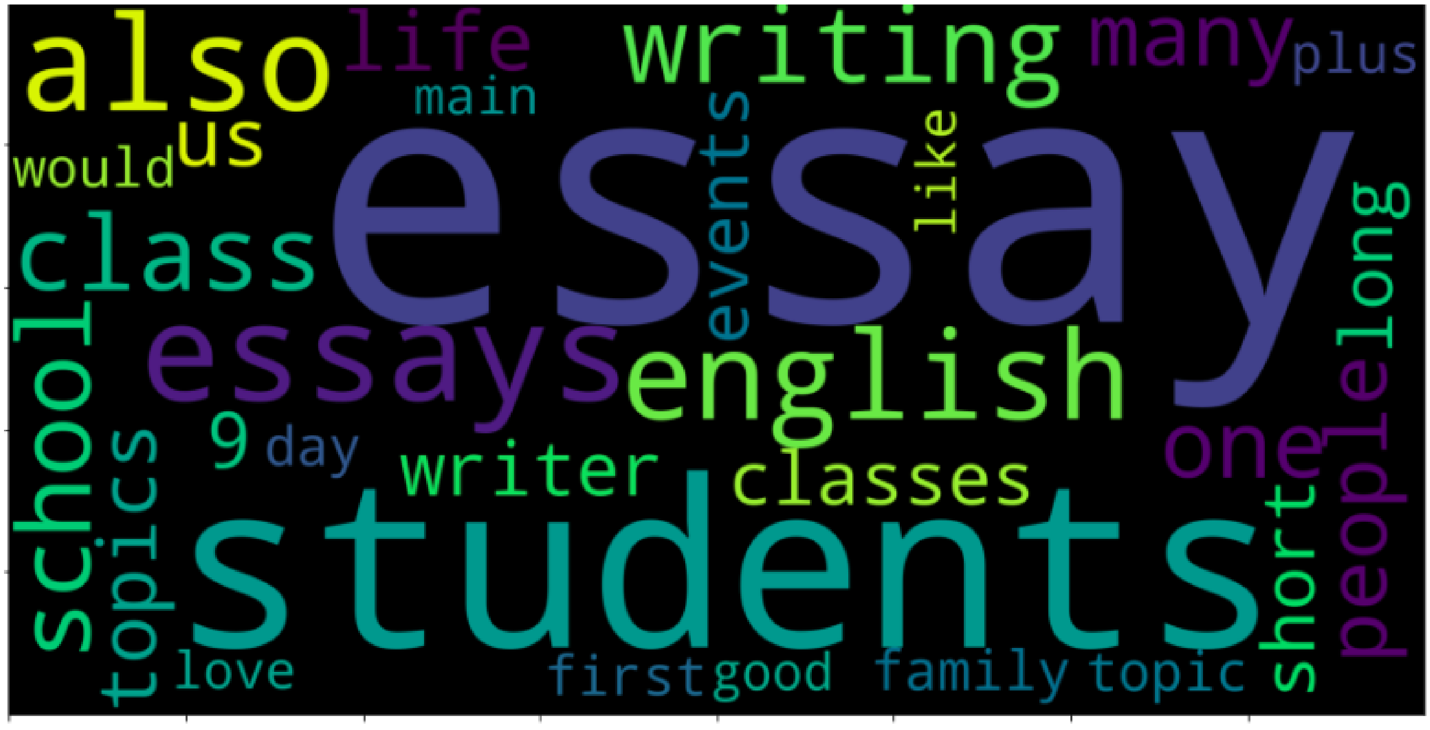
It shows the maximum number of words are nouns

16. 12. Then word cloud visualization is done using the first 30 maximum frequency words after the

Removal of stop words because after the first 30 words the frequency declines.

And these 30 words describe the topics covered in the Corpus and can be used to represent the whole corpus . These words do not include stops words because stop words do not give much information related to the corpus .

\


essay
students
essays
english
also
writing
class
school
life
many
plus
main
us
would
events
like
long
people
one
short
topics
9
day
writer
classes
love
first
good
family
topic

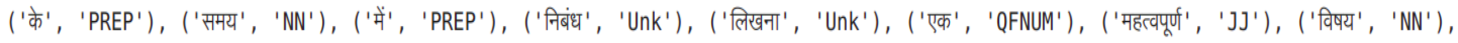
PART 2: -HINDI TEXT

Steps

1. First download necessary libraries
2. Then extract text from all the links and print the raw text(extraction is taking approximately 10 minutes . it may be because devnagiri scripts does not have normal ascii values so they are scored differently)
3. Cleaning of corpus is done by removing punctuations and foriegn words and numbers
4. Tokenization of words and sentences
5. Words without stop words are stored and printed
6. POS tagging is done on all words

   Pos tagger is taking long time to process (aprox 12 minutes for whole corpus)

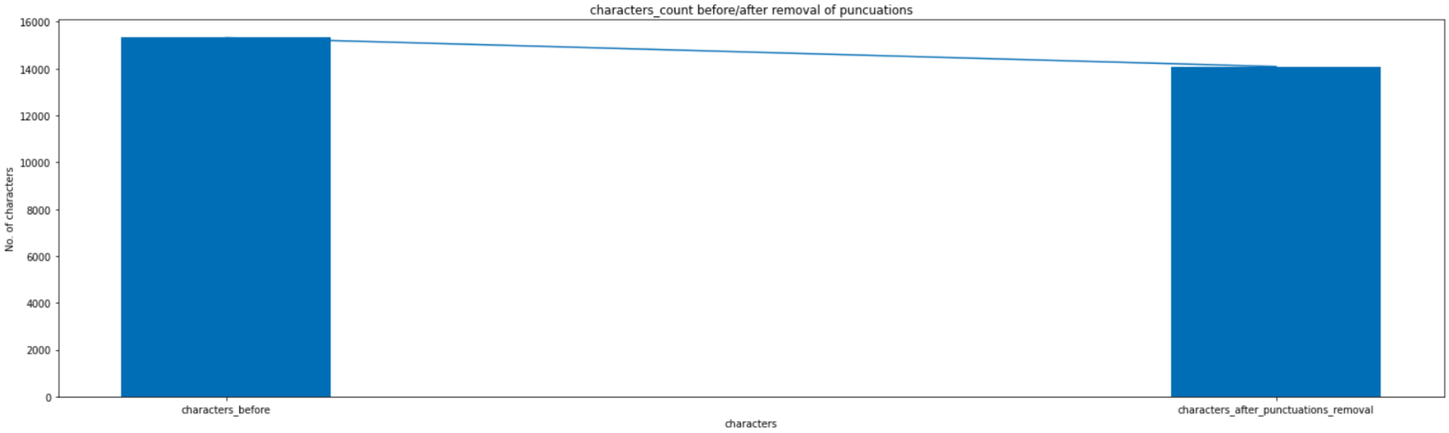
   Part of output :



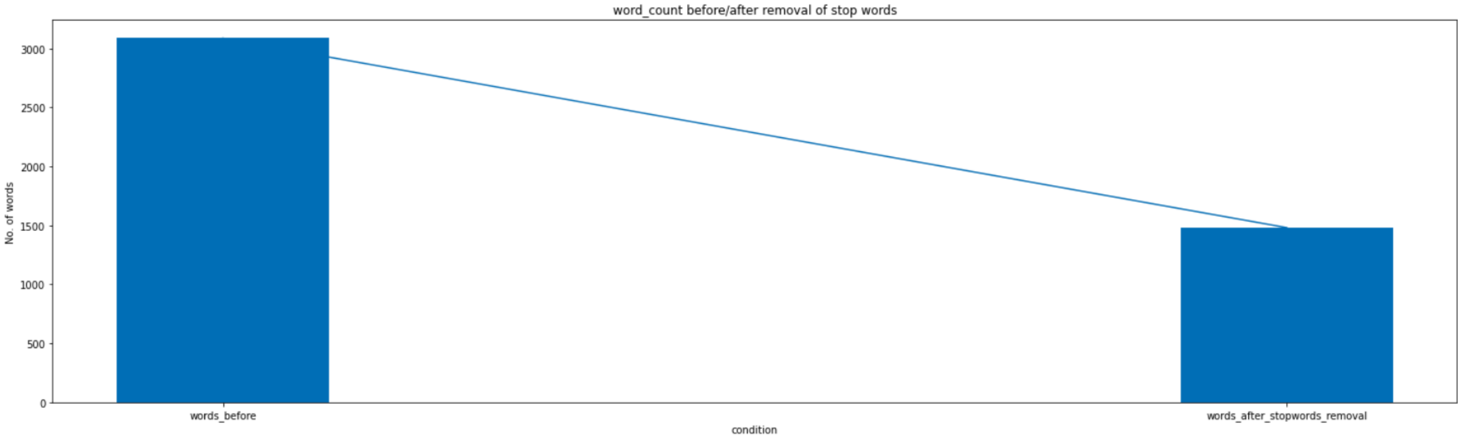
```
('के', 'PREP'), ('समय', 'NN'), ('में', 'PREP'), ('निबंध', 'Unk'), ('लिखना', 'Unk'), ('एक', 'QFNUM'), ('महत्वपूर्ण', 'JJ'), ('विषय', 'NN'),
```

7. Lemmatization is done on all words
8. Stemming is done on all words
9. Calculation of frequency of all words
10. Calculation of frequency of all words removing stop words
11. printing most frequent 30 words in romanized form (as matplot do not support devnagiri script)
12. printing most frequent 30 words(removing stop sords) in romanized form (as matplot do not support devnagiri script)

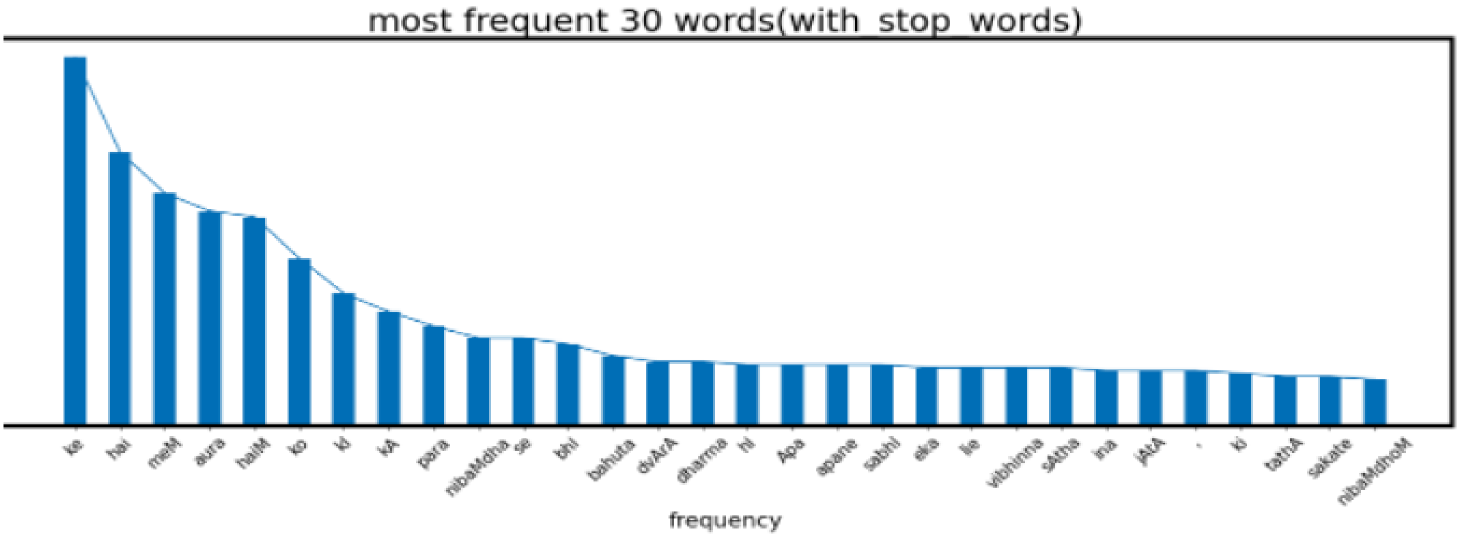
13. This graph shows that after removing punctuations their is a significant change in the number of characters in the corpus



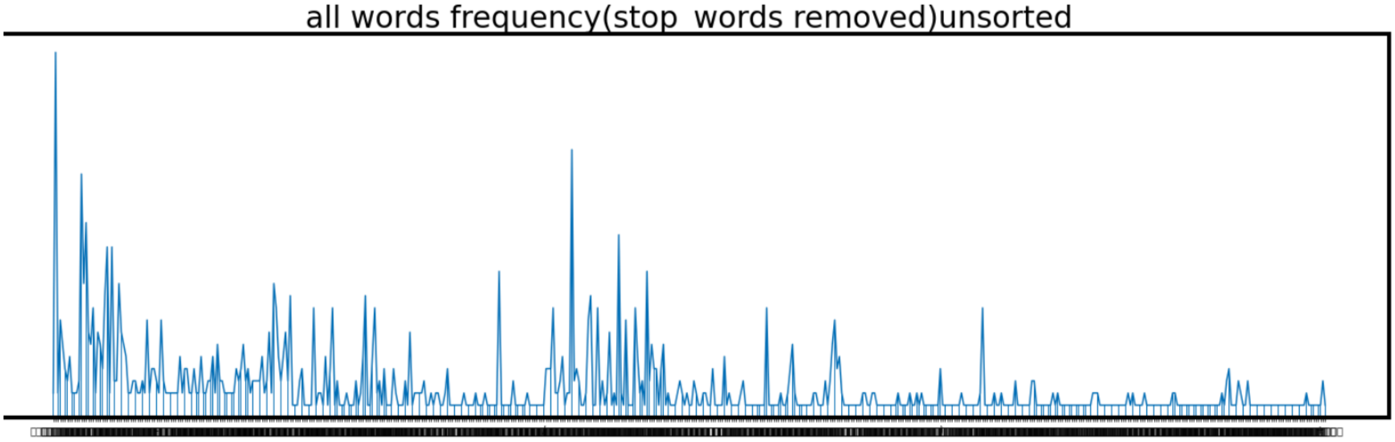
14. This graph which is depicting word count before and after removing stop words shows that more than 50% words are stopwords in corpus



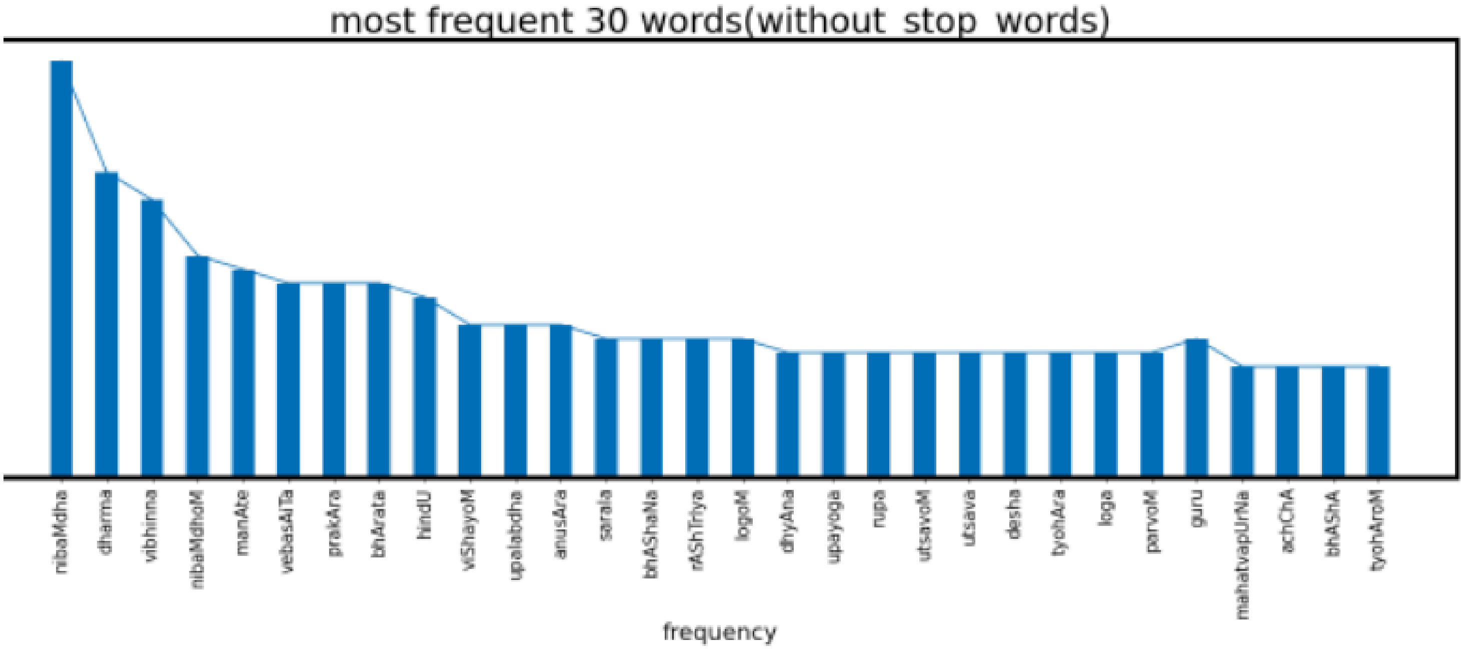
15. This graph shows words frequency of all words and most frequency is shown by stopwords like ke,hai,etc



```
set_text(s, 0, flags=flags)
```

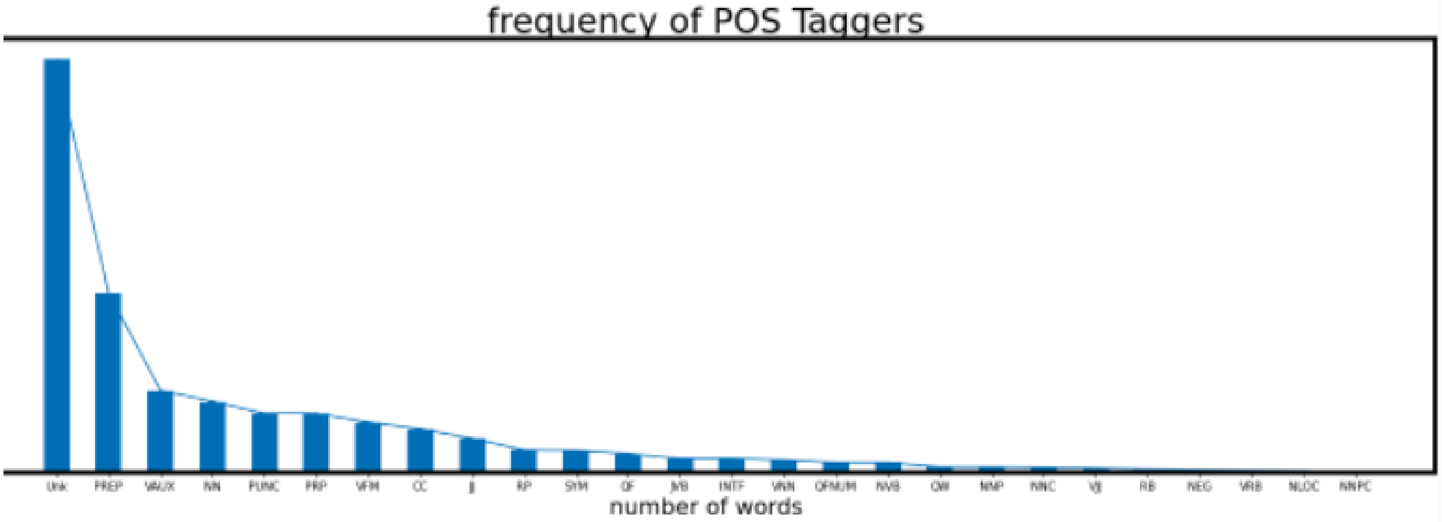


16. Graph of frequency of words shows that most frequent words are not too much in number compared to others as the stop words are removed.

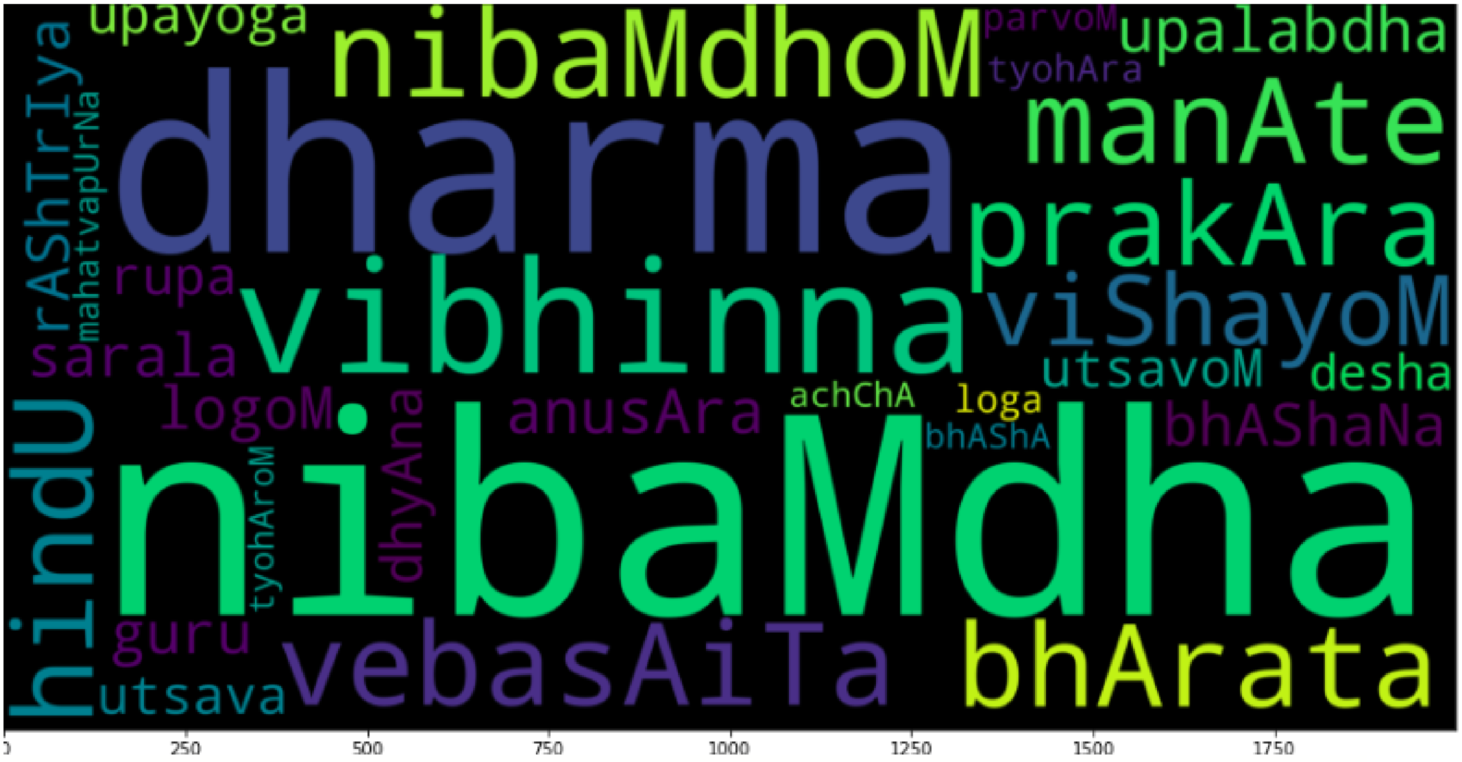


17. Graph is made showing number of words for every POS tag

It shows the maximum number of words are nouns



18. Then word cloud visualization is done using the first 30 maximum frequency words after the
Removal of stop words because after the first 30 words the frequency declines.
And these 30 words describe the topics covered in the Corpus and can be used to represent the whole corpus . These words do not include stops words because stop words do not give much information related to the corpus .


upayoga
nibaMdhoM
parvoM
upalabdha
tyohAra
dharma
manAte
rAShTrIya
mahatvapUrNa
prakAra
rupa
vibhinna
viShayoM
sarala
utsavoM
desha
logoM
anusAra
achChA
loga
bhAShaNa
bhAShA
hindU
nibaMdha
dhyAna
tyohAroM
guru
utsava
vebasAiTa
bhArata