Word clouds

A word cloud is a collection, or cluster, of words depicted in different sizes. The bigger and bolder the word appears, the more often it's mentioned within a given text and the more important it is. In this Project, you need not portray the visual simulations rather generate words of highest importance.

A. Corpus Collection : Take 10,000 sentences each from English and Indian language of your choice.

a. Use Crawling to extract the text. You can choose a particular website crawl.
b. Clean the corpus (Remove images,ads if any)
c. Remove foreign words/expressions, punctuations, symbols like currency, Abbreviations. Acronyms (WHO, UNICEF) can be retained.

B. Perform the following on the corpus.

a. Tokenization(Sentences,Tokens)
b. POS tagging
c. Remove Stopwords
d. Stemming and Lemmatization

C. Analysis using computational tools.

a. Frequency graphs for each of the above tasks. Analyse the Graphs.
b. Based on the Frequency analysis, Write your own algorithm to build the word cloud.
c. How many words do you want to include in the word cloud? Mention the reasons for your choice. Output these words in the descending order of their importance.

Bonus:

Visualisation of the word cloud.

Submissions Format:

Create a folder with RollNumber_Project1

1. Code(.py or .ipynb)
2. A PDF with name Analysis.pdf containing the following:
    a. Frequency graphs and their analysis
    b. Overview of the Algorithm used to generate word cloud.

Zip the folder and upload.