Computational Linguistics

Summer 2021, IIIT Hyderabad

Project 1 – Wordcloud Generation

Algorithm

The first section of the code scrapes 136 Wikipedia pages for sentences. The pages are all related to the culture of India and Kerala.

The raw data is written to a file `raw.txt` in the same directory, and the user is prompted to enter a filename to read the new raw data. If the user wishes to use the same data, `raw.txt` should be given as input. The file is then read.

The raw data is then

- transliterated into WX notation
- tokenised into words and then sentences
- cleaned (rid of punctuation, abbreviations, symbols and foreign words)
- rid of stopwords
- stemmed

The NLTK library is used for tokenisation and graphing. Cleaning is done by simply filtering out tokens that contain extraneous symbols. Removing stopwords is done similarly. Stemming is done by a set of regexes and is extremely rudimentary. The wxconv library is used for transliteration. As there are four symbols that it does not transliterate, these are taken care of individually. The BeautifulSoup library is used for scraping.

Thereafter, the frequency graphs are created with the top 20 tokens at each step (four graphs). Then the wordcloud is generated with the top 50 words that are not stopwords (not stems). Note that a new wordcloud will be generated every time the code is run.

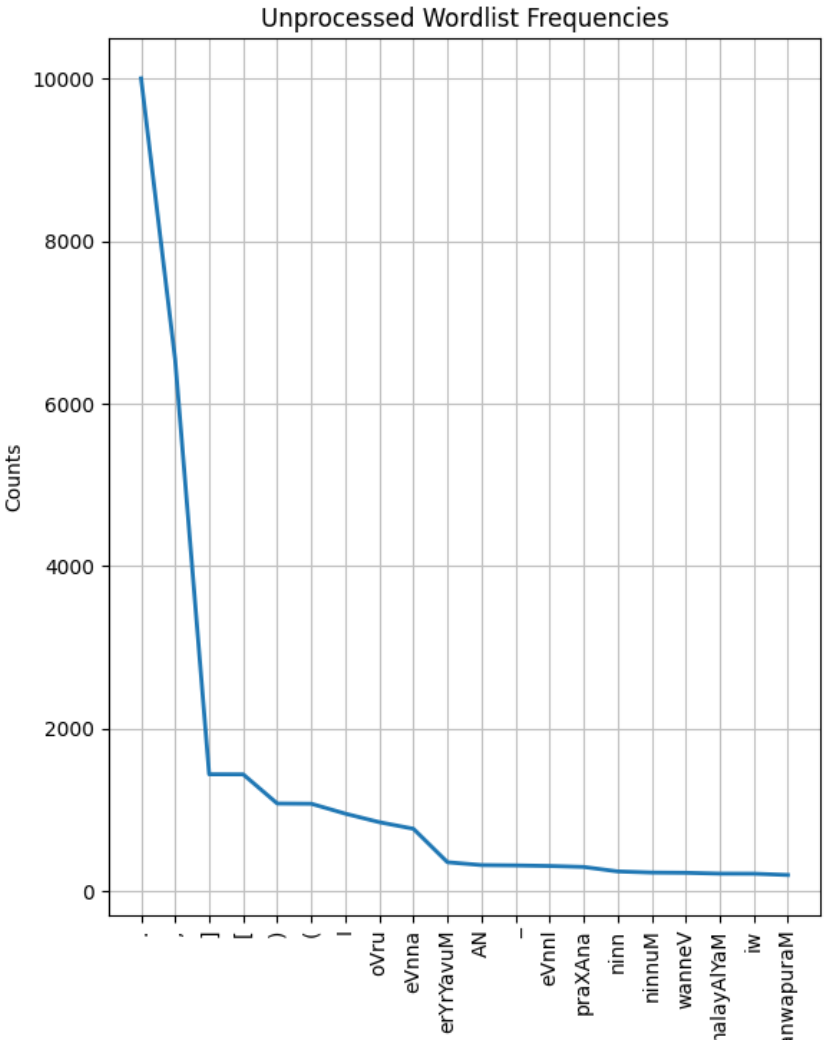
These numbers were chosen taking into a account a reasonable size and shape of the graphs. More than 50 words do not fit in a 2000 x 3000 px image without getting very cluttered, and 50 words is sufficient to give an idea of the contents of the text. Furthermore, a graph with more than 20 words gets too big, and 20 words are enough to estimate the correctness of the method and modify it accordingly.

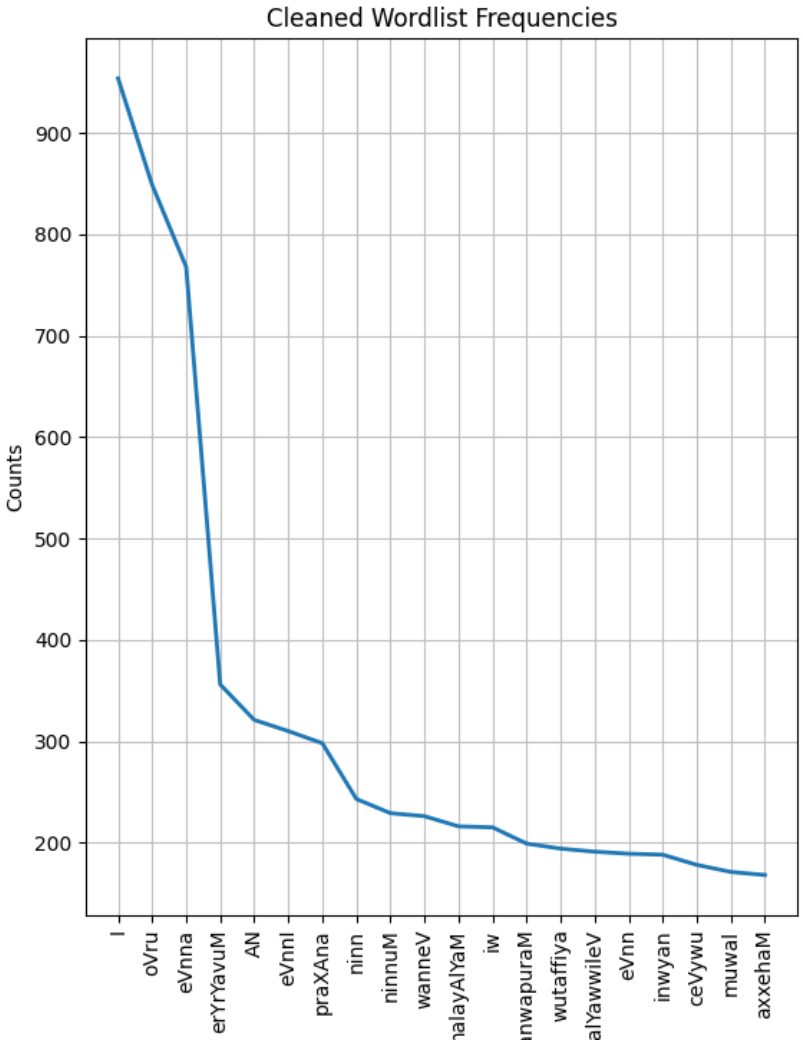
The above steps are all done in parallel on Malayalam text as well as transliterated text, since stemming and graphing can only be done on transliterated text, but the scraped data is in the Malayalam script.

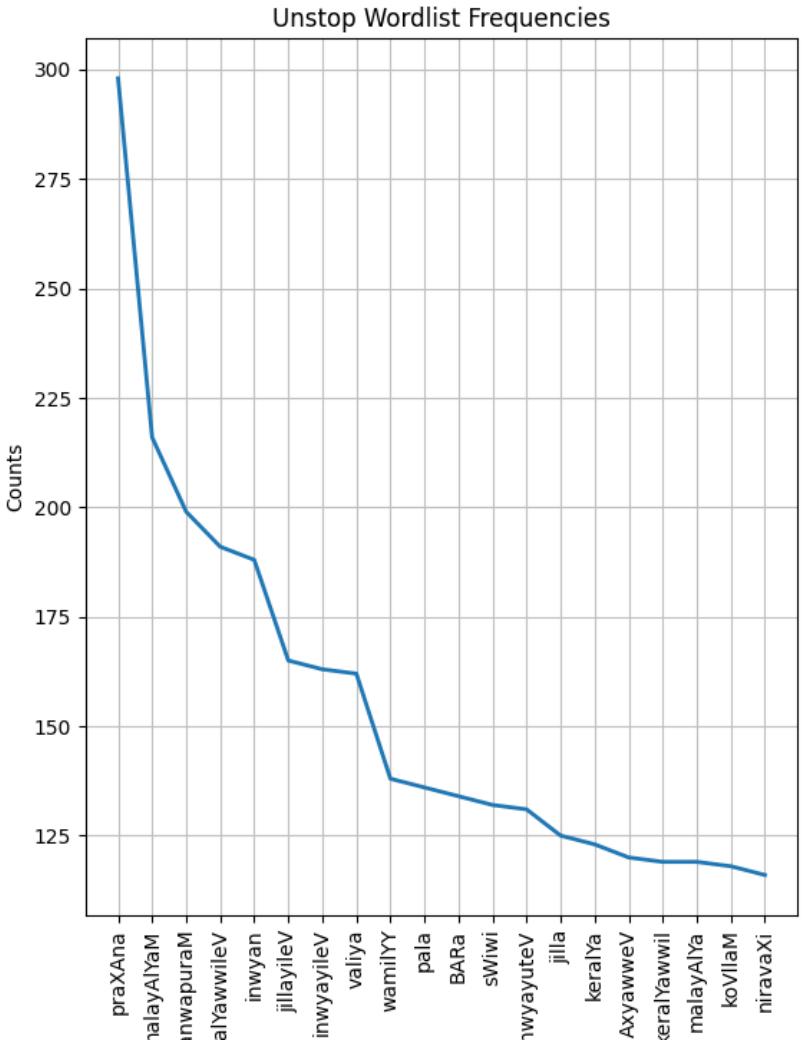
Analysis

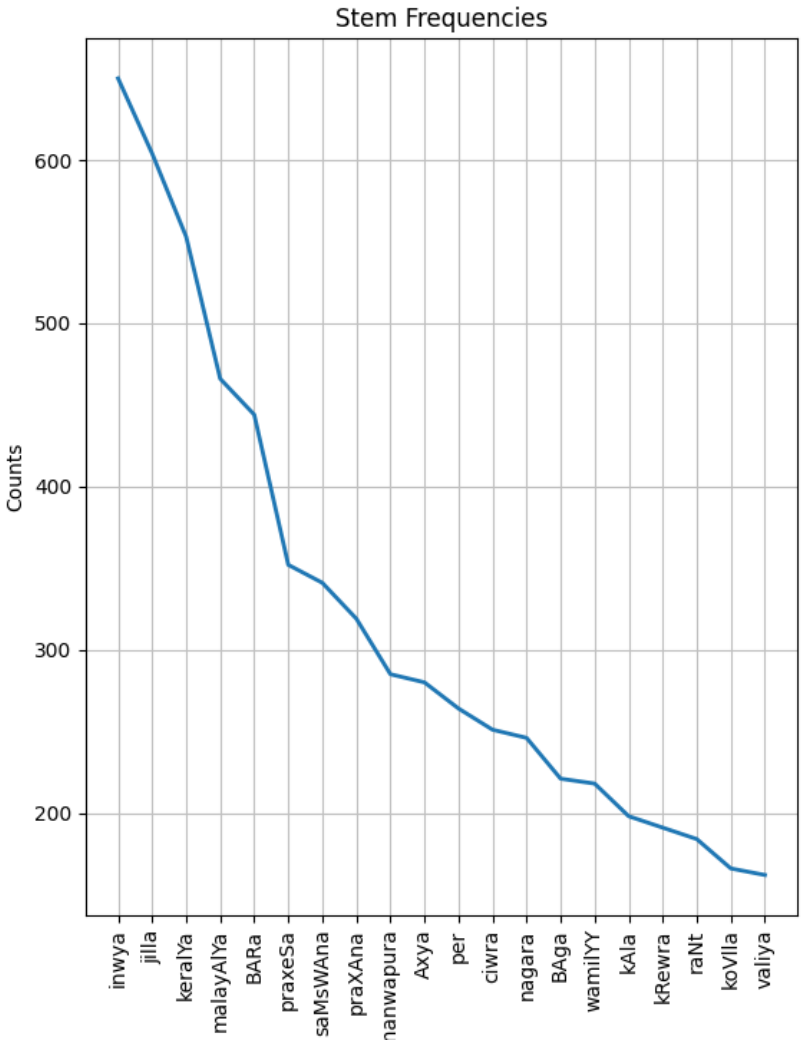
- The forms of all graphs of tokens were roughly similar, displaying a steep drop from the beginning and a flattening towards the end.

- The most common tokens in the unprocessed graph were punctuation symbols, justifying the need to remove them from the text by cleaning.
- The top 20 tokens in the cleaned data were mainly stopwords, with a few exceptions. These, too, convey no idea as to the overall meaning of the corpus and are removed.
- The most frequent words with stopwords removed are fairly representative of the text, but there are repetitions. In the top 20, for instance, we see both `malayAlYaM` and the corresponding adjective `malayAlYa`; and both `keralYam` and `keralYawwile` ("in Kerala", adjective phrase). This justifies the need for stemming.
- The stemmer is about as accurate as in English, although extremely rudimentary. This could be because Malayalam words tend to be slightly more regular in their morphology.


Unprocessed Wordlist Frequencies
Counts
10000
8000
6000
4000
2000
0
oVru
eVnna
praXAna
ninn
ninnuM
wanneV
iw


Cleaned Wordlist Frequencies
Counts
900
800
700
600
500
400
300
200
l
oVru
eVnna
erYrYavuM
AN
eVnnI
praXAna
ninn
ninnuM
wanneV
nalayAlYaM
iw
inwapuraM
wutaffiya
alYawwileV
eVnn
inwyan
ceVywu
muwal
axxehaM


Unstop Wordlist Frequencies
Counts
300
275
250
225
200
175
150
125
praXAna
nalayAlYaM
ınwapuraM
alYawwileV
inwyan
jillayileV
inwyayileV
valiya
wamilYY
pala
BARa
sWiwi
ɯwyayuteV
jilla
keralYa
AxyawweV
ɑeralYawwil
malayAlYa
koVllaM
niravaXi


Stem Frequencies
Counts
600
500
400
300
200
inwya
jilla
keralYa
malayAlYa
BARa
praxeSa
saMsWAna
praXAna
anwapura
Axya
per
ciwra
nagara
BAga
wamilYY
kAla
kRewra
raNt
koVlla
valiya