Write Up - Creating the Language Model

Out of the two choices given – Hindi and Marathi, I chose Marathi as the language for building the language model and chose Marathi Stories as my domain.

The first task was to create a data set containing 1000 sentences with each sentence having more than 3 words. I chose stories as my domain for creating the data set. I found a Marathi story **Pathlag (Parts 1-7)** on the website **https://www.matrubharti.com/book/read/content/19866266/pathlag-1/** . I wrote a Python Script to scrape it and build a 1000-sentences data set. I used BeautifulSoup as the scraping tool. I also used urllib.request and langdetect modules in my script while building the data set.

The next task was tagging the words with their Part-Of-Speech tag. I manually tagged every word in these 1000 sentences. I referred the Unified Parts of Speech (POS) Standard in Indian Languages - Draft Standard –Version 1.0 Tag Set for this. While tagging the Part-Of-Speech of a word, I did it while keeping in mind the context in which the word is being used. Isolation method for Part-Of-Speech tagging would have resulted in many wrong taggings as my data set was formed from a story and hence used words repeatedly in multiple contexts.

The next task was N Gram Modeling. The task was to find out unigrams, bigrams, trigrams, 4grams and 5grams from the POS - tagged text file. I wrote a Python script to do this task. I observed the following statistical results while performing the N Gram Modeling task.

Table of Statistical Results

Number of Tokens	11879
Number of Sentences	1000
Number of Unigrams	Distinct - 4333 Overall - 11879
Number of Bigrams	Distinct - 9062 Overall - 10879
Number of Trigrams	Distinct - 9616 Overall - 9879
Number of Fourgrams	Distinct - 8842 Overall - 8879
Number of Fivegrams	Distinct - 7875 Overall - 7879

The data set was formed from a Marathi story **Pathlag**. Interestingly, it had usage of many foreign words (written in Marathi alphabet). Along with this, there were words formed using other alphabet. In order to deal with this, I used the langdetect module to make sure the data set consisted words formed using Marathi alphabet only. This says a lot about modern day novels and stories of native languages which seem to have some influence of English on them.

I had to run numerous experiments to get my code working for all cases – unigram, bigram, trigram, 4gram and 5gram. Initially, I worked to remove the newline character from the input text file before any processing. I had to appropriately manage indices of the q, tempo and done lists to form unigrams, bigrams, trigrams, 4grams and 5grams from the input text file and write them to their respective with-label, without-label and with-frequency files.

Although I had a few doubts about the assignment regarding whether the <s> and </s> tags should be included in the process or not, I got them cleared from the instructor. Hence, I did not feel the need to discuss the assignment with anyone. As a result, I did not discuss the assignment with anyone.