CL3.101 Computational Linguistics 1
Assignment 3 - Part 1

Course Instructor: Parameswari Krishnamurthy

Deadline: 11th March 2024, 5:00 PM

Instructions

- Your assignment must be implemented in Python.
- Do NOT use any standard library.
- Make sure the submitted assignment is your original work. Do not copy any part of the assignment from your friends. Do not refer any AI systems to generate the code.
- No deadline extension will be possible. Please start early in order to finish it on time.
- Make sure to follow the submission format properly. You will be penalised for not following the naming and submission format.

1 Part of Speech Tagging

This is the Part 1 for this assignment. Further Task(s) will be released later.

1.1 Task 1: Annotate Sentences

You have to annotate ∼ 500 tokens in English language, and ∼ 500 tokens for some other language, preferably an Indian language (In case you are not familiar with any Indian language, you can use a language other than English) for part of speech tagging.
The tagset to be used for POS tagging is the BIS Tagset. [Refer to the POS Tagging Folder uploaded on Moodle for the BIS POS tagset documentation]. Specify the source of the data you have chosen for annotation. Do not use data from any existing POS dataset, try to collect data from sources like news articles.
Store the annotations in a text file in the format: **word "\t" tag** i.e the word and tag seperated by a tabspace, after tokenizing the sentence into different words.
For Example:

```
India NNP
, PUNC
Australia NNP
and CC
England NNP
are VAUX
the DT
Big JJ
Three CD
in IN
Cricket NN
. PUNC
```

Also provide a frequency distribution graph for the tags for both the languages. Make sure the tags are as evenly distributed as possible.

2 Submission Format

Submit a single zip file named **RollNo_FirstName_POS_Annotations.zip** in which you will submit three files:

- English_Annotations.txt
  Containing annotations for English
- (Your Language)_Annotations.txt
  Containing annotations for your language
- Freq_Distribution graph
  Two such files, one for each language
- ReadME.md
  Containing sources for the data.

Kindly follow this submission format properly.