CS4248 A2 by Lim En Dian John A0097082Y

Introduction

For this assignment, we are to write a program to perform part-of-speech (POS) tagging. I have decided to write it in Python and I have programmed it to take run either one of the two commands.

*java build\_tagger sents.train sents.devt model\_file*

*java run\_tagger sents.test model\_file sents.out*

In each of the commands, they require the input of the three respective filenames as arg[1], arg[2] and arg[3]. Sents.train, sents.devt, sents.out and sents.test are the data files provided for the assignment. Model\_file is the file generated after running the build\_tagger command. The tagger will reference this file when the command run\_tagger is run.

Algorithm

How my POS tagger works is first it will determine the conditional probabilities of each word type given the corpus – sents.train. This is done by first counting the occurrence of each word type. I consider same words with different tags unique word types. From here I gain the probability of a single word type occurring given the total number of word types in the corpus.

Where W = word type, T = total number of word types and C(W) = num of W

I then find the conditional probability of the word type given the previous word type.