Eric Rios Soderman & Wafiakmal Mitah
IDS 703
Assignment 4 : POS Tagging for Viterbi Algorithm

Introduction:

In this exercise, we created multiple functions that will operate on the segmentation of the brown corpus. From the nltk module, the tagged_sents method was utilized on the brown corpus. The output was a list of sentences with tuples of the word and its part of speech tag. Our role was to create the initial states distribution matrix, the emission matrix and the transition matrix that could work in unison with the Professor's implementation of the Viterbi algorithm.

Test Cases where it failed:

Test String	Our POS Output	NLTK (Real) POS Output
SENTENCE	10150	Of 10152
Those	DET	DET
coming	NOUN	VERB
from	ADP	ADP
other	ADJ	ADJ
denominations	NOUN	NOUN
will	VERB	VERB
welcome	VERB	VERB
the	DET	DET
opportunity	NOUN	NOUN
to	PRT	PRT
become	VERB	VERB
informed	VERB	VERB
.	.	.
SENTENCE	1051	10152
The	DET	DET

preparatory	ADJ	ADJ
class	NOUN	NOUN
Is	VERB	VERB
an	DET	DET
introductory	NOUN	ADJ
face-to-face	ADP	ADJ
group	NOUN	NOUN
In	ADP	ADP
which	DET	DET
new	ADJ	ADJ
members	NOUN	NOUN
become	VERB	VERB
acquainted	VERB	VERB
with	ADP	ADP
one	NUM	NUM
another	NOUN	DET
.	.	.
SENTENCE	10152	10152
It	PRON	PRON
provides	VERB	Verb
a	DET	DET
natural	ADJ	ADJ
transition	NOUN	NOUN
into	ADP	ADP
the	DET	DET
life	NOUN	NOUN
of	ADP	ADP

the	DET	DET
local,	ADJ	ADJ
church	NOUN	NOUN
and	CONJ	CONJ
its	DET	DET
organizations	NOUN	NOUN
.	.	.

For the correct cases, it performed well by tagging the words correctly in comparison to the truth output. We had 3 errors. We will round our numbers for these quick calculations.

The first error is “coming”, which was identified as a noun, when it should have been a verb. This word had a noun to noun transition probability of 0.2129 with coming as a noun’s probability, 0.9210. The product was 0.01556. The Noun to Verb probability was 0.1435 with coming as a verb’s probability, 0.1435. The product was 0.1414. The winning probability was Coming as a noun’s 0.01556.

The second error was “introductory”. It was an unknown word. Its probability was already defined.

The third error was “another which was tagged as a Noun when it should have been a determinant. After moving through the decision trees, “another came at a junction where it could be either a Noun or a Determinant. For the sake of brevity, we will also omit the prior probabilities that led to another, all the way from the initial state probability of “the” given that it is a determinant to “one”. As for the transitions, the probability for going from Num to Noun was 0.3779 and the one from Num to DET was 0.011. “Another”’s probability, given that it is a noun, is 0.0044, and, for the determinant, it was0.9518. The product of the pairs were 0.0042 (Num to Noun, Another - Noun) and 0.0038 (Num to DET, Another DET). This explains why the incorrect choice was chosen, given that it had a higher probability.