Homework 3 [Due date 26 Oct 2015]

Build a HMM based part-of-speech tagger. A corpus for this task containing a training set (entrain.txt) and a test set (entest.txt) is provided to you. From the labeled training set, calculate the transition and the observation probabilities of the HMM tagger (slide pos-tagging.ppt, page 78,106) where states are the part-of-speech tags and the observations are the words. Then implement the Viterbi algorithm (slide pos-tagging.ppt, page 98-102) discussed in the class to find the most probable state sequence (decoding) for an arbitrary test sentence. Run your algorithm on the test set and report its error rate, where the error rate is defined as follows,

$$error\_rate = \frac{\#words \ \ in\ test\ set\ for\ which\ predicted\ label\ matches\ the\ hand\ tagged\ label}{\#total\ words\ in\ test\ set}$$

Corpus (was taken from here)

The file format is quite simple. Each line has a single word/tag pair separated by the / character. (In the raw file, only the word appears.) Punctuation marks count as words. The special word ### is used for sentence boundaries, and is always tagged with ###. Test set contains some unknown words that do not appear in the training set. Handle these unknown words by assuming that the probability of observing these words from is same for all the states. The following table describes all the tags used in the corpus.

C	Coordinating conjunction **or** Cardinal number
D	Determiner
E	Existential *there*
F	Foreign word
I	Preposition or subordinating conjunction
J	Adjective
L	List item marker (*a., b., c., . . .*) (rare)
M	Modal (*could, would, must, can, might . . .*)
N	Noun
P	Pronoun **or** Possessive ending (*'s*) **or** Predeterminer
R	Adverb **or** Particle
S	Symbol, mathematical (rare)
T	The word *to*
U	Interjection (rare)
V	Verb
W	*wh*-word (question word)
###	Boundary between sentences
,	Comma
.	Period
:	Colon, semicolon, or dash
-	Parenthesis
'	Quotation mark
$	Currency symbol