READ ME

The zip file contains
1) brown.txt
2) split_corpus.py
3) HMM_1.py
4) HMM_2.py
5) unigram_accuracy.py
6) bigram_accuracy.py
7) uni_test_output.txt
8) bi_test_output.txt
9) test.txt

“split_corpus.py” divides the brown corpus into 2 parts namely, “brown-train.txt” and “brown-test.txt”. The brown-test.txt contains first 500 sentences of the brown corpus and the brown-train contains the remaining sentences.

HMM_1.py & HMM_2.py are HMM taggers using two different methods to tag unknown words. Running the 2 files produces 2 text files , output_1.txt & output_2.txt containing the tagged sentences for the sentences in brown-test.txt . They also output confusion matrix and accuracy on the terminal. The confusion matrix is printed only for common tags such as nouns, adjectives etc, and has its entries as percentages.

HMM_1.py calculates the emmision probabilities of unknown words from the corpus i.e, it calculates the probability of occurance of each tag in the corpus and uses this as emmision probabilities whenever an unknown is encountered.

HMM_2.py takes the transition probabilities as emmision probabilities whenever an unknown word is encountered.The intution behind this is that the probabilities for a tag depends on the tag of previous word .So it calculates the probabilities of all the tags given the tag of the previours word of an unknown word which are nothing but the transition probabilities.

The “unigram_accuracy.py” and “bigram_accuracy.py” are 2 files which output the accuracy of unigram and bigram models on brown-test.txt.This is for comparison of HMM based taggers with unigram and bigram tagges. uni_test_output.txt & bi_test_output.txt are the test files required for the python programs to run.

“test.txt” contains some troublesome/challenging sentences used for analysis of the performance of the tagger.
1st 4 sentences – Past participle problem
next 2 sentences -VB vs NN (for word ‘race’)
next 13 sentences - In vs RB vs RP
next 3 sentences – Unknown words
next 2 sentences – HMM_1 vs HMM_2

For running the model on “test.txt”, delete the previous output files(output_1 & output_2) and change
line no. 161 in HMM_1.py
line no. 137 in HMM_2.py