Name : Md Maminur Islam

Assignment number : 3

Due Date : March 01, 2018

Problem statement:
Build a baseline statistical tagger.
(i) [10 points] Use the assignment#2's hash of hashes to train a baseline lexicalized statistical tagger on the entire BROWN corpus.

(ii) [20 points] Use the baseline lexicalized statistical tagger to tag all the words in the SnapshotBROWN.pos.all.txt file. Evaluate and report the performance of this baseline tagger on the Snapshot file.

(iii) [20 points] add few rules to handle unknown words for the tagger in (ii). The rules can be morphological, contextual, or of other nature. Use 25 new sentences to evaluate this tagger (the (ii) tagger + unknown word rules). You can pick 25 sentences from a news article from the web and report the performance on those.

Summary:
I have implemented the solution in python. There are three classes SSBParser and POS and unknown_word_handler. SSBparser parses the 'SnapshotBROWN.pos.all.txt' and 'BROWN.pos.all' and generates the outfput files 'BROWN-clean.pos.txt' and 'SnapshotBROWN-clean.pos.txt' respectively. POS generates hash of hash and calculates performance. Unknown_word_handler tags unknown words and some rules are there. Please run the main.py to see the output for each problem.

Problem 1(i) Solution:
Please run main.py and the clean files will be generated in the project directory. Hash of hash is generated which is not printed as the hash is too large.

Problem 1(ii) Solution:
While you run main.py and the performance will be printed

Problem 1(iii) Solution:
While you run main.py and performance will be printed for test news from web