Programming Assignment 3 (**due Tuesday 15 March 2022**)

Write a python program called tagger.py which will take as input a training file containing part of speech tagged text, and a file containing text to be part of speech tagged. Your program should implement the "most likely tag" baseline.

For each word in the training data, assign it the POS tag that maximizes P(tag|word). Assume that any word found in the test data but not in training data (ie an unknown word) is an NN, and find the accuracy of your most likely tagger on a given test file. **Record that accuracy in your overview comments.**

The input for this assignment is found in the files section of the web site (PA3.zip). The training data is pos-train.txt, and the text to be tagged is post-test.txt. There is also a gold standard (manually tagged) version of the test file found in post-test-key.txt that you will use to evaluate your tagged output. **THEN add at least 5 rules to your tagger and see how those rules affect your accuracy. Make certain to also include the rules you add and the resulting accuracy in the overview comment as well.**

Here's an example of how your tagger.py program should be run from the command line. **Note that your program output should go to STDOUT, so the file named used below could be anything.** This program will learn the most likely tag tagger from the train data, and then tag the test file based on that model.

python tagger.py pos-train.txt pos-test.txt > pos-test-with-tags.txt

Note that your tagger should not modify pos-test.txt in any way, and that the output of the program should make certain to handle each tagged item in the test data. You will note that in both the training and test data phrases are enclosed in brackets [ ] - those indicate phrasal boundaries, and you may ignore these since we don't use them in POS tagging.

You should also write a utility program called scorer.py which will take as input your pos tagged output and compare it with the gold standard "key" data which I have placed in the Files section of our group (pos-test-key.txt). Your scorer program should report the overall accuracy of your tagging, and provide a confusion matrix . Again, this program should write output to STDOUT.

The scorer program should be run as follows:

python scorer.py pos-test-with-tags.txt pos-test-key.txt > pos-tagging-report.txt

Note that if your accuracy is unusually low (less than the most likely tag baseline) that is a sign there is a significant problem in your tagger, and you should work to resolve that before submission.

Please do not modify any of the files found in PA3.zip. If there is some unusual situation in that text please ask in lecture or via the discussion list. Note that there are a small number of "ambiguous" tags, where two tags are joined with an | symbol (e.g. broker-dealer/NN|JJ). In these cases, only use the first part of speech tag and ignore the rest.

Please submit a soft copy of your program source code (tagger.py and scorer.py) to the Canvas.