Siddharth Hathi
CS 10
PS5
Report

In a short report, provide some example new sentences that are tagged as expected and some that aren't, discussing why. Also discuss your overall testing performance, and how it depends on the unseen-word penalty (and any other parameters you use).

Sentences that were classified incorrectly:

     The quick brown fox jumps over the lazy dog
     [DET, ADJ, ADJ, N, N, P, DET, ADJ, N]
     All that is gold does not glitter
     [DET, WH, V, N, V, *, N]
     Not all those who wander are lost
     [*, DET, DET, WH, V, V, VN]
     Gokul is a chill dude
     [PRO, V, DET, N, P]

Sentences that were classified correctly:

     henry l. bowden was listed on the petition as the mayor's attorney
     [NP, NP, NP, V, VN, P, DET, N, CNJ, DET, N, N]
     The court ruled correctly
     [DET, N, VD, ADV]

Discussion: The Markov Model is very good when the sentences being tagged are similar to the ones it was trained on. In this case, the model has been trained on "brown-train-sentences.txt," a file containing descriptions of court proceedings. When the input sentences are about courts and rulings, it has no problems. However, when they deal with new words like "Gokul," "jump," "dude," etc., the model has a much harder time. This is, in large part, a result of the high penalty for unseen words (-100).

Notice how when the penalty is reduced to -10, the model does a better job on some of the sentences:

     Gokul is a chill dude
     [PRO, V, DET, ADJ, N]

In sentences like the one above, with simple structures but new words, the model is familiar with the structuring of the sentence and it can actually classify some of the words a little better when the penalty for new words is low. However, this low penalty also decreases performance on sentences with familiar words.