Assignment 3: POS Tagging

L645, Sandra Kuebler

DUE: at beginning of class on Thursday, October 9

1. **Bigram Tagging** Given the probabilities in the table below, what is the probability of the following tag sequences for the sentence `time flies like an arrow`? Note that the probabilities are shown as percentage (0-100%).

   (a) VB NNS IN DT NN

   (b) JJ VBZ VB DT NN

   If you are enrolled in B659, write a program to calculate this. Submit your code and an output.

P(time | NN) =	7.0727	P(NN | S) =	0.6823	P(IN | NNS) =	21.8302
P(time | VB) =	0.0005	P(VB | S) =	0.5294	P(VB | VBZ) =	0.7002
P(time | JJ) =	0	P(JJ | S) =	0.8033	P (VB | NNS) =	11.1406
P(flies | VBZ) =	0.4754	P(VBZ | NN) =	3.9005	P(RB | VBZ) =	15.0350
P(flies | NNS) =	0.1610	P(VBZ | VB) =	0.0566	P(RB | NNS) =	6.4721
P(like | IN) =	2.6512	P(VBZ | JJ) =	2.0934	P(DT | IN) =	31.4263
P(like | VB) =	2.8413	P(NNS | NN) =	1.6076	P(DT | VB) =	15.2649
P(like | RB) =	0.5086	P(NNS | VB) =	0.6566	P(DT | RB) =	5.3113
P(an | DT) =	1.4192	P(NNS | JJ) =	2.4383	P(NN | DT) =	38.0170
P(arrow | NN) =	0.0215	P(IN | VBZ) =	8.5862	P(E | NN) =	0.2069

   15 pts.

2. **Markov Assumption**: Some tagging decisions seem to depend upon the following word; for example, *I can/VBP tuna* vs. *I can/MD help.* Trigram taggers condition on the previous tags, $P(w_i|w_{i-2}w_{i-1})$. How is it, then, that trigram taggers can often tag these cases accurately? In other words, how do they capture cases where the tag of w_i depends upon w_{i+1}?

   15 pts.

3. **Additional Exercise: POS Annotation**:

   Annotate the data file 80days.txt with POS tags using the labels from the Penn Treebank. Make sure that the filename contains your name. The data file and the guidelines for the Penn Treebank tagging can be found as attachments to the assignment.

   15 pts.