## Assignment 1

Due: 01/28/2019

Build a bigram HMM tagger. Data is packed in
http://nlp.cs.rpi.edu/course/spring19/udtb.tar.gz

1. Data format
After you untar the package, there are six files:

- en-ud-train.conllu: Training data for English POS tagging
- en-ud-dev.conllu: Development data for English POS tagging, which you can use to test and tune your system.
- en-ud-test.conllu: Blind test set for English POS tagging
- es-ud-train.conllu: Training data for Spanish POS tagging
- es-ud-dev.conllu: Development data for Spanish POS tagging, which you can use to test and tune your system.
- es-ud-test.conllu: Blind test set for Spanish POS tagging

In each file,
Column 1: token ID
Column 2: token
Column 3: normalized token
Column 4: coarse-grained POS tags
Column 5: fine-grained POS tags

2. Required assignment: English POS tagging (12pts)
(1) From the training set, learn transition and emission probabilities of an HMM based POS tagger, print them out in separate files. Use Column 4 for tag labels. (5pt)
(2) Then implement the Viterbi algorithm so that you can decode (label) an arbitrary test sentence. (5pt)
(3) Tune your tagger on the development set, e.g., you can try different out-of-vocabulary handling methods, you can try various types of features from the additional columns; any of these improvement over the baseline will get extra credits. (up to 3 extra pt)
(3) Test the final tagger on the blind test set and report accuracy. (2pt)
(4) Error analysis, and report remaining challenges and possible solutions (up to 2 extra pt)

3. Bonus assignment:
(1) Spanish POS tagging (instead of English POS tagging) (3 extra pts)
(2) Use fine-grained POS tags in Column 5 (2 extra pts)