Homework #3 - Language Modeling, and HMMs
Due at 11:59pm on Wednesday, November 11

1. [15] Read Fernando Pereira’s 2000 paper, “Formal Grammar and Information Theory: Together Again?”. Write a 3-4 paragraph response which minimally 1. identifies the major claims of the paper, 2. assess the approaches used to address these claims and 3. assesses the conclusions drawn. Any additional discussion on the relationship between statistical estimation and formal grammar is very appropriate in this response.

2. [5] Language Modeling: Examining the training portion of the Reuters Corpus as delivered as part of NLTK, identify the percentage of uni-grams that are unique (i.e. appear exactly one time).

3. [5] Language Modeling: Now identify the percentage of unique N-grams for N in {2, 3, 4, 5, 6}.

4. [5] Language Modeling: Calculate the out-of-vocabulary (OOV) rate of the testing portion of the Reuters Corpus, with respect to the training portion.

5. [20] Language Modeling: Train a normal bigram language model on the train portion of the Reuters Corpus. Calculate the perplexity of the training data. Calculate the perplexity of the testing data (ignore the perplexity contribution from OOV tokens.)

6. [40] Language Modeling: Open-Ended. Use whatever methods (smoothing, backoff) you like to train a language model on the training data to generate the lowest perplexity possible on the testing data. Note: you should not tune the parameters of your model on the testing data. Describe your technique, and report the perplexity.

7. [10] Calculate the perplexity and OOV rate of the news portion of the Brown Corpus (as delivered as part of NLTK) against the best model trained in Question 6.

Note: This page (http://www.nltk.org/book/ch02.html) will provide some useful information for accessing the Brown Corpus and Reuters Corpus. If you do not want to write the other questions in python, you can use nltk only to download the files, and then read them into a program written in another language.