PA2 Report

Name(s): Emmanuel S Thomas & Sumit Gupta

Due Date: 05/01/2014

Class: CS276

Professors: Chris Manning & Pandu Nayak

**Overview**

For this assignment we build the language model using unigrams and bigrams. In the building step, only the counts were calculated leaving the probabilities calculations to be done at query time to avoid calculating the probability of each unigram and bigram. The maximum likelihood estimates was used to calculate the probabilities in the language model. Interpolated conditional probability was used as smoothing technique in the language model.

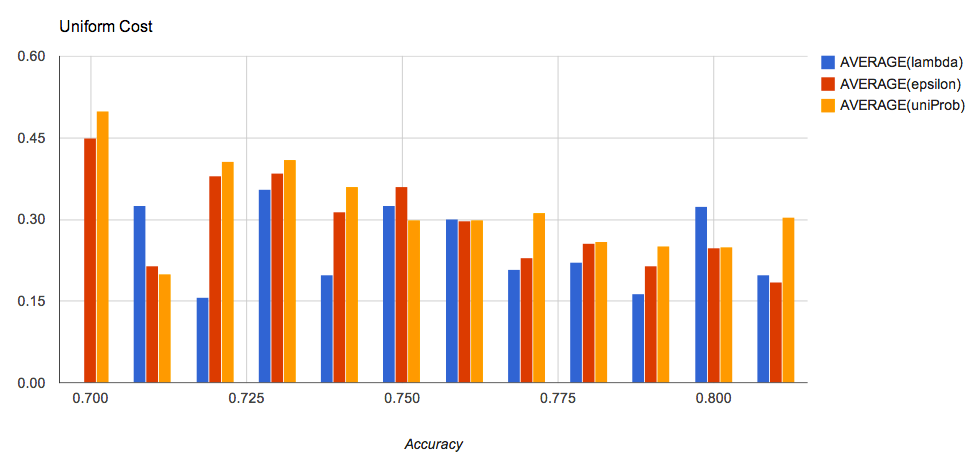
For the noisy channel model, the uniform cost probability was used first using different values for the uniform number. Then we use an empirical model in an attempt to increase the accuracy of our estimates. To generate the candidates, we only calculate the 1 edits. Pruning was then performed by taking only the candidates where all the words appear in our dictionary.

Finally to score the candidates, Naive Bayes was used.

**Expanded edit mode (Extra)** – We also tried to incorporate common spelling errors (source: aspell.net) to 1) build the language model by adding corrected words to unigram dictionary and 2) while generating candidate to do lookup in the common spelling error dictionary. It shows slight improvement (apprx. 1%).

**Note:** The code reads dictionary from **“./data/commonspellingerrors.txt”** when “**extra**” param is passed with buildmodel.sh and runcorrector.sh.

**Parameter Tuning**

As we experimented, it appears that the lower the parameters, the higher the accuracy for the uniform cost model.

However for the empirical model, the accuracy seemed to be driven by the value of epsilon. The higher the value of epsilon, the higher was the accuracy.