CSE 4309/5361 - *Artificial Intelligence II*

Homework 4- Spring 2013

Due Date: May 2, 2013

Note: Problems marked with * are required only for students enrolled in CSE 5361. They will be graded for students enrolled in CSE 4309 for extra credit.

N-Gram Models

N-Gram models provide a means of probabilistically modeling language and using this model to infer correct sentence completions and evaluate the likelihood of a particular sentence being actually the one spoken.
For this assignment you are provided a small dataset of sentences over a very small vocabulary, all related to grid navigation. The data format of the dataset (given as a text file) is such that it first provides a list of the words in the vocabulary and then a list of sentences.

1. Using this dataset you are to build N-gram models.
   a) Write a program that builds a unigram model for the dataset.
   b) Write a program that builds a bigram model for the dataset.
   c) Write a program that builds a trigram model for this dataset.
2. Evaluate the three models by determining how well they predict a missing word in a sentence. For this, randomly pick a set of at least 30 sentences in which you remove a single word and then determine whether each of the three models correctly completes the sentence when choosing the maximum likelihood sentence completion (i.e. the word that has the highest likelihood given the respective n-gram model). Briefly discuss the result.

3.* Build a probabilistic model by combining the three models formed in part 1. (either as a weighted sum of the models or by selecting the one with the best support) and use it for the same completion task of part 2.