Proposal: Project 3
CS 4740
Jennifer Doughty jad359, Scott Allen Cambo sac355, Joseph Porter jmp392, Graham Harwood gdh56

I. Software Implementation Overview:

We will be using the python programming language with the nltk package recommended for this course. We will be employing the Hidden Markov Model already included in the nltk package in order to focus more on analysis for tagging. The hidden variables that we will be trying to uncover from the HMM are the set of numbers -2,-1,0,1,2 where a positive number represents a positive sentiment, a negative number represents a negative sentiment, and zero represents a neutral statement. This hidden variable set will match nicely with the observed variable set that is tagged in the training data. Given that the Lee et al paper had fairly good results with unigram probabilities and even better results with unigram and position probabilities, we have decided to first try to work only with unigram probabilities and if that seems to work well, then we will consider the position of a word within the text as well. For a sentence such as the one below, we might be able to predict that it is a 2 by first extracting unigram probabilities from the training set and feeding these as transition probabilities to the HMM function in NLTK along with sentences comprising an entire paragraph or an entire review. The output we would expect is that the HMM would use viterbi's algorithm to figure out that given the unigram probabilities that we calculated as features and the overall path from the beginning to the end of the sequence (be it a paragraph or a full review), that this sentence has strong positive sentiment making it a 2.

> *this is an energetic film it has lots of black comedy and incisive dialogue to recommend it to an audience appreciative of indie films*

Extension:

We will attempt to predict paragraph level sentiment by treating each paragraph as a whole sentence and implementing the same process that we are doing for the sentence level sentiment tagging. Our baseline for this process will be the mean value (rounded to the nearest sentiment integer) from the set of sentences making up the paragraph. Both the baseline and the paragraph level sentiment prediction will be compared to the gold standard paragraph sentiment analysis provided by the class during this portion of the project.

Baseline:

The baseline system that we will develop will be done by sorting all sentences from the training set into their respective categories (-2,-1,0,1,2) and counting the most frequent words used in each of these sentiment categories. We will then perform tests in which we tag a sentence accordingly if it contains more of the top 10 most frequent words than the

other sentiment categories.