CS 4740 - Project 3 Part 1

Alex Maass (am838), Chris Heelan (cjh276), Casey Mak (cm522), Joe Cardali (jmc475), Kevin Lin (kl524)

Model:
We will implement a Hidden Markov Model for this project. Each author will have its own model.

Algorithmic Details
Observed variables: Sentences in the review (delimited by <>'s)

Hidden variables: The sentiment scores of sentences.

Transition probabilities: The probability of transitioning from one score to another. In theory, when you have a score of -2, your next sentence should have a higher probability of also being negative. This is learned by counting the number of times one score follows any other score in the training data. For the first sentence of a paragraph (or review), we will assume the previous score is zero, or pick the most common score.

Emission probabilities: The probability that a score could emit a certain sentence. For each author and for each sentiment score (-2 to 2), train an n-gram model. After training is complete, the emission probability of a test sentence will be inversely proportional to the perplexity of that sentence on that model. One possibility is to use Naive Bayes and simply multiply the probability of each n-gram in the sentence. One factor to determine experimentally is the removal of irrelevant stopwords in sentences. In the previous project, POS filtering and stemming worked well, but we will also try using NLTK's built in stopword list.

For example, in the sentence:

```
the nicolas cage performance caught the nuances of his character in a bewitching manner
as in his crazed ems role he catches the futility and desperation he feels gnawing at him
from inside and displays an engaging chemistry with the three drivers and with patricia
arquette <2>
```

There are n-grams that reveal the positive sentiment, such as "bewitching manner" and "engaging chemistry." These phrases generally do not include stopwords like "the," or "a."

Given transition probabilities between states, emission probabilities for each state, the current state, and the observed sentence, we can determine the sentiment scores for a sequence of observed sentences using the Viterbi algorithm.

Extension:
We think that a simple n-gram model using Naive Bayes to determine the probability of a sentence appearing given a score will work well, and therefore we will analyze its performance compared to the Hidden Markov Model.

We are also interested in taking into account the overall review score and using it as a base score from which each Markov Model will start at. Initially, the starting sentiment is going to be neutral (zero) for our main algorithm, but as an extension, we will use the paragraph level sentiment score and use it as a starting sentiment for the HMM to aid in determining the sentence level sentiments.