Benjamin Wiley
CSCI 5622
9/18/15

Feature Explanation

My features were tested before submission to Kaggle by splitting up the training data in an $80 - 20$ ratio, where 80% was used to train my features while 20% was used to test them. The following list of features were incorporated into my text classification model:

1. Tf-idf vectorizer: This was to account for repetition of common words, such as *the, and, but* by assigning a lower score to them, which in doing so assigned higher scores to words which occur less frequently, but are potentially better classifiers.

2. Incorporation of symbols: After trying out a character-gram classification model, my overall accuracy decreased compared to using word-grams. However, I did notice that certain symbols, paricularly the # symbol, were important predictors. Realizing that my current model was not taking this into consideration, I added a unique word classifier to account for these characters.

3. Analysis of uppercase letters: Analyzing the data, I noticed that character names were often an indication of a spoiler sentence. Because these names were uniquely uppercase, I decided to single them out and assign a unique word classifier to sentences with uppercase characters.

4. Extraction of data from Wikipedia: While very computationally intensive, I tried extracting a sentence from Wikipedia for each TV show title and episode combination which provided additional data for classifying important words occuring in other sentences.

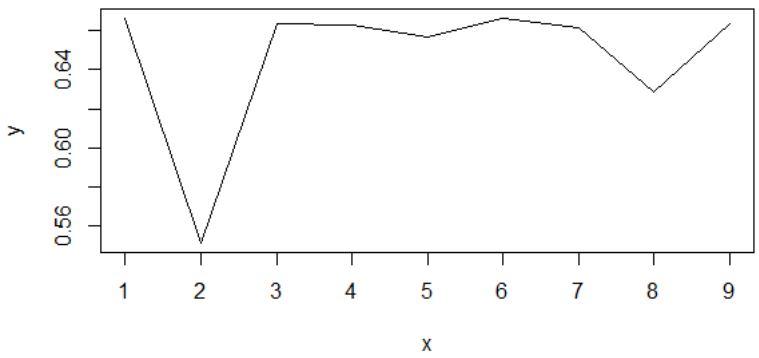
5. Stemming words: Initially, I constructed a basic classification model and discovered that certain words were repeated, only in various grammatical formats. To combine these, I striped the roots of words in sentences and analyzed these instead.

6. TV Show: Some TV shows had less spoiler sentences than others, so including the TV show title in the classification model proved effective in obtaining a better classification.

7. n-grams: These actually lowered my predictability, which I suspect was due to my stemming procedure and the tf-idf functionality. As a result, I did not include these in my final model.

8. Categorization of sentence length: I took the length of each TV show sentence and determined the quartiles of the data. Then, during classification, I assigned each a unique weight which would identify it as belonging within a certain family.

Each feature provided, on average, a marginal increase (or decrease, as in n-grams) in predictability for my classification model, and the results may be seen below, where the x-axis corresponds to the item above (note that 1 represents the base case with all features added),



As a result, my Kaggle score was 0.67344, while my ranking, listed as BenjaminWiley, was 43rd in the class.