Andrey Shprengel
Kaggle Username: ansh2964
CSCI 5622
Homework 3

In order to begin analyzing new features and how well they were working I split the training set into a development training set and a development test set. The training set contained 13,000 examples or ~90%, the rest of the data was used for development testing.

With initial testing I found that without any modification using just the word count as features the accuracy hovered around 62%.

My first intuition was to use 2 grams to see if we could extract more meaning from the sentence. To do so I changed the n gram range to include 2 grams. This seemed to increase the accuracy to about 64%.

I then noticed that wont words repeated even in the top 10 features for each class. For example the words[kill, kills, killed] and [die, dies]. Because this may be wasting our features I added a preprocessor that would stem the words. To dd so I used the the porter stemmer in NLTK python toolkit. This Increased the accuracy the about 65%

Knowing very little about tv(haven't had one since 8th grade) I wasn't sure what other sort of features to look at but realized that tropes are certain rhetorical devices that may be used. I figured that certain types of episodes may be more likely to cause viewers to discuss what happens such as if it caught them by surprise, because of this I added the trope field as a feature. This increased to accuracy to 66-67%

There were a few other things that I tried that did not seem to help increase the accuracy, for one i tried adding the page field but that did not seem as useful as the trope. The other was that I noticed that often particular characters names showed up in the top features for each label. In order to avoid this i tried to leave out all proper nouns using NLTK pos_tag however this seemed to decrease accuracy