NLP – Sentiment Analysis

Rain Dartt, Katie Dion, Brian Padilla, and Dharmin Shah

## Overview

Our team used the Stanford Core NLP library’s NLP parser and NLP sentiment analyzer to process and perform sentiment analysis on IMDB reviews. Stanford’s NLP library already had the functionality to perform sentiment analysis on sentences. However, we added the functionality of analyzing the sentiment of the subjects in the sentence. For example, for the sentence “I hated the movie, but the actor was great,” we can map that “the movie” received a negative sentiment, whereas, “the actor” received a positive sentiment.

## Dataset

The corpus was obtained from ai.stanford.edu/~amaas/data/sentiment. The dataset contains a set of movie reviews from IMDB. It provides 25,000 highly polarized movie reviews for training and 25,000 for testing as well. However, we performed our analysis a subset of only 10 movie reviews in raw text format from the corpus, as the bi-skip five-gram was a huge bottleneck in our algorithm.

## Details

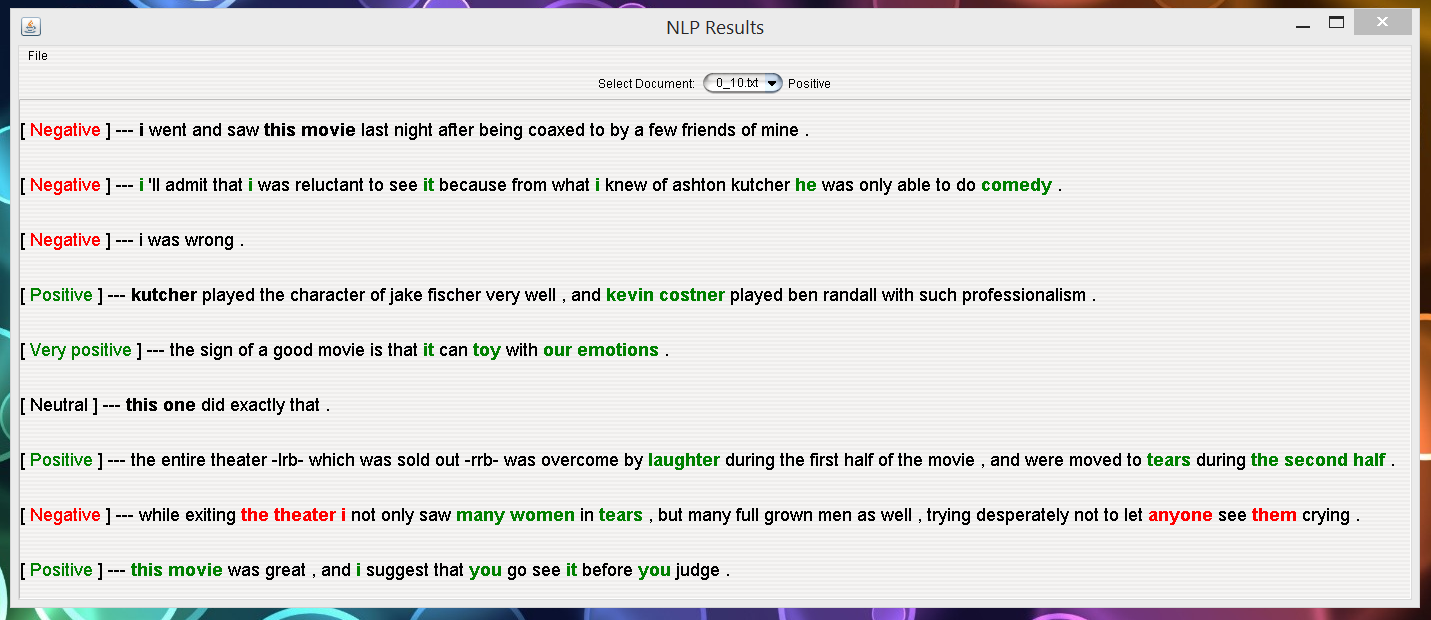
The Stanford NLP library analyzes the sentiment of entire sentences. Our contribution to the library is to analyze the sentiment of each subject within a sentence. The sentiments can range from “Very positive” to “Very negative”. We use the NLP parser to parse the sentences and extract the subjects from each parsed sentence. For example, the sentence “The movie was horrible, but the actor was great.,” is parsed as:

(ROOT (S (S (NP (DT The) (NN movie)) (VP (VBD was) (ADJP (JJ horrible)))) (, ,) (cc but) (S (NP (DT the) (NN actor)) (VP (VBD was) (ADJP (JJ great)))) (. .)))

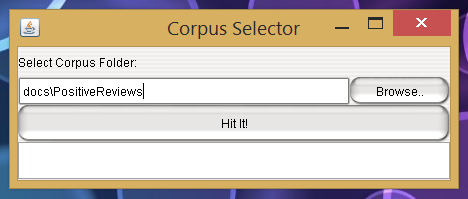
We use the words in the NP tag as the subjects. So in this case the subjects are “the movie” and “the actor.” Next, we generated bi-skip five-grams from the subjects to ignore insignificant words such as “and” when handling long or compound subjects. The bi-skip five-grams helps to produce more concise subjects of five words or less. We then calculate the sentiment of the n-grams containing the subjects by using the NLP sentiment analyzer. Finally, we display the results.

## Results

We displayed the results by showing the associated sentences within each document. Each sentence is prefixed with the appropriate sentiment. Within each sentence the subjects are bolded and colored based upon their corresponding sentiment. Positive sentiment is represented with green bolded text. Negative sentiment is represented by red bolded text. Lastly, subjects with a neutral sentiment have a bold font with no color change. We also display the overall sentiment of the document next to the dropdown menu. The document sentiment is a simple aggregation of the sentiments of each sentence.



Additionally we allowed the users to choose the directory, which contains the corpus.



## Limitations

The major limitation of our approach is the accuracies of Stanford’s NLP Parser and NLP Sentiment Analyzer. According to Stanford’s website, the NLP Sentiment Analyzer has an accuracy of 80.57% when analyzing at the sentence level. However, the performance takes a hit when applying it to entire documents, because a document is a collection of sentences related to each other. We ran the sentiment analyzer on 5000 test documents that were categorized as positive reviews, and found that the Stanford NLP sentiment analyzer has an accuracy of only 46%. However, our approach also has other limitations as well. The code takes a long time to run because of the bi-skip five gram. We look at 21 possible five-grams for every possible set of 7 words in the document. Another limitation is that sometimes the five-grams may not be enough to capture the relevant information because there might be sentences that may have lot of irrelevant words like “a”, “the” etc. which may generate irrelevant five-grams and not capture the true sentiment for the subject. The third limitation, which is also a limitation for Stanford’s library, is that the NLP Parser fails to recognize the same subjects if they are worded differently. For example, if the sentence is “The movie was great but the film was long,” the parser will give two subjects “the movie” and “the film,” even though we know that “the movie” and “the film” are referring to the same subject. This is significantly reduces the accuracy of the sentiment analysis for subjects.

## Potential/Future Uses

### In the IMDB movie reviews

Subject sentiment analysis in the IMDB can potentially be used to discover the popularity of a movie’s actors, characters, or production crew directly from the reviews. Aggregating the sentiment feedback of a subject and drawing conclusions from the cumulative reviews could achieve this. Preforming this analysis would require us to implement an aggregator function. We would also need to match similar subjects, such as an actor’s name with the name of their character.

### Other databases

Subject sentiment analysis has implications for major social networking sites or sites with a significant social review component such as Facebook, Twitter and YouTube. Facebook already has been widely discussed for its social experiments wherein the company experimented on the effects of presenting users with content based on its sentiment. Companies can use subject sentiment analysis to study feedback for their products and overall brand on social media. Another example use would involve politicians analyzing the effect of recent decisions or campaigns on their public relations by aggregating the feedback about their name or other key words on Twitter.