Data 620 Proposal

Tony Mei and Lin Li

Proposal : Analysis of Highly polar movie reviews

Hypothesis: We want to find what words are used most often in movie ratings

Data Sources: By the Stanford AI Laboratory, this text classification dataset contains a set of 25,000 highly polar movie reviews, with an additional 25,000 reviews for training. The dataset is useful for sentiment analysis experiments. It also includes unlabeled data which can be used for further training or testing.

Guideline of approach:

Proposal

Data Aquisition

Data Preparation

Data Exploration

Sentiment Analysis

Conclusion

Review

Final Presentation and Video

Concerns:

1. First time sentiment analysis on movie ratings and seeing how it pans out
2. Collaborating with each other to finish the project before deadline

Work Cited:

http://ai.stanford.edu/~amaas/data/sentiment/