Project Number: 23

Project Title: IMDB Movie Ratings Sentiment Analysis

Anonymous Reviewer Team ID: 22

Summary of Project

This group's project has to do with sentiment analysis of IMDb movie comments. The original data set contains 40,000 comments from IMDb and a "positive" or "negative" sentiment associated with the comment. After much pre-processing of the data, they apply an unsupervised method (Latent Dirichlet Allocation) and two supervised learning methods (naive Bayes and logistic regression).

Strength

Overall, I think this is a very intriguing data set and problem. The report has a very clear description for the pre-processing of the data and this leads very clearly into the methods for the paper.

Weakness

I would have like to seen more clear description of what variables are most indicative of a "positive" or "negative" sentiment. The paper hints at such when talking about key words in the LDA model.

To Bayes

For the Latent Dirichlet Allocation, `gensim` package was used to get a topic model for the two classes. It is not clear how the `gensim` package applies LDA to the data. The same goes for the Naive Bayes model. That model was applied through `scikit-learn` package. I think implementing these methods with a data set this large would have been quite difficult. For the logistic regression, the `rstanarm` package was used. This method had the most interesting Bayes features to it. The table showing the 90% confidence intervals and posterior means show the uncertainty associated with each coefficient. I would have liked to see a comparison to a frequentist logistic regression model.

Ratings

Writing Quality: 5

Technical Quality: 4

Overall rating: 4