Final Project Proposal

Marcel Colvin 912033961

[mmmcolvin@ucdavis.edu](mailto:mmmcolvin@ucdavis.edu)

For the final project, I would like to do an analysis of the [*New York Times Comments*](https://www.kaggle.com/aashita/nyt-comments) data set on Kaggle. For this analysis, I would like to use two other data sets in conjunction with the *New York Times Comments*, the [*Toxic Comments Dataset*](https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge)and the [*Meyer Briggs Personality Type Dataset*](https://www.kaggle.com/datasnaek/mbti-type). Using these datasets, I would like to create an algorithm, using machine learning techniques such as SVM, logistic regression, or a clustering algorithm, to see whether comments in the *New York Times Comments* are toxic and the personality types of posters in the *New York Times Comments*. I hope to in the end find out what Meyer Briggs personality type posts the most toxic comments in the *New York Times Comments*.

Firstly I hope to train either an SVM or logistic regression model on the toxic comments data set, which can classify comments as either toxic or non-toxic. I will start with a binary classification, but if the model works better with more in-depth classification, I will use that instead. I will train this model, then run it on the test data to make sure it has a decent classification rate. Then I hope to train a logistic regression model to predict the personality type using text after breaking the data into training and testing data. I will then apply these models on the *New York Times Comments* data set and hopefully gain some insight into the commenters on the New York Times’ website.

The main difficulty of my work will be cleaning the text data using natural language processing. I will have to remove URLs, stop words, and many other problem words in order to make my predictors the best they can be. The next one will be applying an SVM model to either data set if necessary. While I understand what the model does, I have never applied it in code. I hope to optimize and experiment with many different ML models to get the best results, so I know this will take a lot of research in the field.