Review Sentiment Analysis

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
- What industry/realm/domain does this apply to?
- What is the motivation behind your project? (Saying you needed to do a capstone project for flatiron is not an appropriate motivation)

This project attempts to get a sentiment of user reviews left for a product by comparing review text with its corresponding star rating. This can allow for analysis of other texts or for double checking if review texts match up with star ratings for validation. A motivation for this is seeing a large number of amazon reviews where the text is negative but the star rating is positive, indicating possible bot activity.

Data Understanding

- What data will you collect?
- Is there a plan for how to get the data (API request, direct download, etc.)?
- What are the features you'll be using in your model?

Data will be in the form of amazon reviews scraped using beautiful soup, the reviews will be hashed and tokenized to be used as features

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
- What are some of the cleaning/pre-processing challenges for this data?

Since it is text data there will be several steps in preprocessing, including removing punctuation, removing stop words, hashing and tokenizing the data and more. Some challenges will involve messy text and text that may have special characters

Modeling

- What modeling techniques are most appropriate for your problem?
- What is your target variable? (remember - we require that you answer/solve a supervised problem for the capstone, thus you will need a target)
- Is this a regression or classification problem?

This will be a classification problem, in this case the target variable is whether or not a review was positive (4 or 5 stars out of 5). Appropriate modeling techniques would be a supervised learning classifier.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?

I plan to use F1 Scores, accuracy, and AUC to determine success.

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?

Naïve Bayes and random forests