Progress Completed So Far:

1. Joined training and test set to get a more complete dataset
2. Implemented bag of words model
3. Removed stopwords from bag of words model
4. Modified bag of words model to remove unknowns
5. Added “text\_” to words to distinguish them from labels in dataset
6. Implemented an algorithm to remove words from dataset that are correlated (> 0.5 correlation)
7. Organized functions to make them easier to use
8. Started working on a function to partition the data into training and test data
9. Refactored code, added comments to make it easier to read

Work left:

1. Run feature selection on the entire dataset (not just the bag of words)
2. Implement rfe feature selection
3. Implement sentiment model (current dataset includes this, but would be good practice)
4. Create visualizations of the word frequency distribution
5. Create visualizations of the word frequency distribution by subreddit
6. Create visualizations of the word frequency distribution by label
7. Create visualizations of the sentiment distribution
8. Create visualizations of the sentiment distribution by subreddit
9. Create visualizations of the sentiment distribution by label
10. Create decision tree model to run on dataset
11. Run decision tree model on the dataset
12. Tune decision tree model for better performance/easier to interpret
13. Create visualizations of the decision tree model

Challenges I Have Run Into

1. Defining a good cutoff for rare words (experimenting and researching online tutorials for good definitions)
2. The number of words, even with removal of stopwords is currently over 5000 (feature selection will help with this but I’m concerned about lowered performance
3. Slow processing speeds due to the high number of features and complexity of the dataset (once again, feature selection)
4. Defining a good algorithm for feature selection (currently removing correlated features, will also apply rfe to full dataset

Resolved Challenges

1. Getting the removal of unknown words to work properly
2. Dealing with words that occur in both the dataset and the features