feature\_importance\_ks\_new\_categories.ipynb

* Code used to rank words in order of importance via KS test scores. Resulting .csv files were combined into one Excel file called “ks\_feature\_importance\_new\_categories.xlsx”.

feature\_importance\_mlr\_new\_categories.ipynb

* Code used to rank words in order of importance via regularized multinomial logistic regression coefficients.

feature\_importance\_random\_forest\_new\_categories.ipynb

* Code used to rank words in order of importance via random forest feature importance scores.

feature\_importance\_tfidf\_new\_categories.ipynb

* Code used to rank words in order of importance via TF-IDF scores.

input\_new\_categories.xlsx

* Cleaned dataset labeled with the original categories.

ks\_feature\_importance\_new\_categories.xlsx

* Word rankings resulting from the KS test.

mlr\_feature\_importance\_new\_categories.csv

* Word rankings resulting from regularized multinomial logistic regression.

model\_output\_visualizations.ipynb

* Code used to generate confidence interval visualizations for the highest performing models.

model\_testing.ipynb

* Code used to run all selected models on the cleaned data. Filters data down to the top 300 words in word\_importance\_lists.xlsx’s “mean rank” column.

relevant\_model\_outputs.xlsx

* Shows various performance metrics of 21 statistically similar models as determined by the confidence intervals.

rf\_feature\_importance\_new\_categories.csv

* Word ranking resulting from random forest.

tfidf\_feature\_importance\_new\_categories.csv

* Word ranking resulting from TF-IDF.

word\_count\_test.ipynb

* Code used to run all models at various numbers of words ranked in word\_importance\_lists.xlsx’s “mean rank” column.

word\_count\_visualization.ipynb

* Code used to visualize the results of word\_count\_test.ipynb.

word\_count\_test\_outputs.xlsx

* Results of all models ran in word\_count\_test.ipynb.

word\_importance\_lists.xlsx

* Contains all words found in input\_new\_categories.xlsx with corresponding TF-IDF, random forest, regularized multinomial logistic regression, and KS test rankings. The “mean rank” column is used to sort words in order of importance and takes the mean of all previously mentioned rankings.

word\_importance\_lists\_code.ipynb

* Code used to generate word\_importance\_lists.xlsx.