Classifying Reddit

Problem Statement

This project aims to train a classifier to predict which subreddit a given post’s title came from.

Methodology

- Data Collection
  - Using Reddit’s API, post titles were collected from the subreddits r/WritingPrompts and r/ShowerThoughts and collected into a combined dataframe.
- Train-Test-Split
- NLP: CountVectorizer / Tfidf
- Classification Modeling
- Parameter Tuning

NLP: CountVectorizer / Tfidf Vectorizer

- Both prepare the titles for classification modeling
- CountVectorizer: word count vectors
- TfidfVectorizer: word frequency vectors, inversely weighted relative to document frequency

CountVectorizer performed marginally better on this data.

Modeling

- DummyClassifier
  - Train Score: ~51% accuracy
  - Test Score: ~51% accuracy

Modeling (Default Parameters)

Classifiers Fit	Test Set Accuracy
Multinomial Naive Bayes	73%
Logistic Regression	**81.76% (best)**
Decision Tree	75%
Random Forest	75%
Support Vector	50%
K Nearest Neighbor	53%

Modeling (Tuned Parameters)

○ Classifiers Fit	Test Set Accuracy
■ Multinomial Naive Bayes	74% (+1)
■ **Logistic Regression**	**81.76% (same)**
■ Decision Tree (gs)	77% (+2)
■ Random Forest (gs)	79% (+4)
■ **Support Vector**	**81.96% (+32)**
■ K Nearest Neighbor	54% (+1)

Conclusion

Best Classifiers

Logistic Regression (default parameters)

- Train Score: ~96%
- Test Score: 81.76%

Support Vector Machine (kernel = ‘linear’, C=0.1)

- Train Score: ~90%
- Test Score: 81.96%