Text Analytics Final Project Part 1: Proposal

Rachel Rosenberg

November 2019

For my project, I want to do a news article classification. Therefore, I will choose Option 2: Comparison of multi-label or multi-class text classification approaches on one or more text classification datasets. For this I plan to use the very clean and well-curated Reuters-21578 dataset, of news articles that ran in 1987. This dataset was curated and compiled from 1990 to 1996, and is widely used for text categorization. I am particularly interested in this topic because I am big into reading and following the news as a hobby; I enjoy the subject matter and frequently choose news-based topics for class projects in MSiA.

This project will involve 6-class classification; I will clean and preprocess the data with my standard set of functions that remove numbers, convert to lowercase, tokenize, stem, and find n-grams. I will then find TF-IDF vectors for each segment and start to build models; I will start with the same models used in Homework 3, and will run the same experiments, but then will find optimal parameters and make tweaks to the structure of the models if it is necessary to improve the accuracy of my models.

The dataset is found at http://www.daviddlewis.com/resources/testcollections/reuters21578/ and is also available through UC Irvine at http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html.