Determining Genre of Classical Literature with Machine Learning

Tony Martinez
Department of Computer Science
Brigham Young University
martinez@cs.byu.edu

Abstract

Text classification is a popular problem within machine learning. The ability for a program to understand classifications and distinctions between texts can be useful for a wide array of real world applications. It was because of this that we decided to focus on teaching our learning model to classify entire books worth of text as either fiction or nonfiction. Using indicators like word frequency and word count allowed us to approach the problem. We trained on a variety of different models. Initially we found that due to possible imbalance in our data set, MLP tended to perform poorly, while Decision Trees and Ensembles of Random forests performed much better. [Additional results here]

1 Introduction

Classification of literary works is a different beast than normal text classification. A big reason for this can be attributed to length, books are just much longer than most other text mediums. Additionally, the classifications of literary works are more nuanced than other sources such as a newspaper article. This has lead to a wide array of different results in previous models for literary classification. A large part of the background information that contributed to this project comes from things that we learned throughout the semester. However there were additional avenues beyond class materials that we explored in an effort to improve and refine our model.

One such avenue was the paper *Genre Identification and the Compositional Effect of Genre in Literature* authored by Joseph Worsham and Jugal Kalita. This study was of interest to us in part because they used the same data set that we pulled our data set from. Their study focused on specific genre classification such as romance or adventure stories. Although our learner is focused more broadly on just fiction versus nonfiction, it was useful to read the work of those who approached a similar topic. One thing that they talked about at length was how using just the frequency of words is usually inadequate for the purposes of genre classification. This lead to us including additional features beyond a bag of words to our model.

Another model that we looked at was a model based off predicting genre only by the title of the book. THis model was created by a github user by the name Akshay Bhatia. Similarly to the paper by Worsham and Kalita, Bhatias model focuses on classifying works into several genres like adventure and romance. However the ability to get a initial guess just off the title of a book was useful to us when we began refining our model and trying different approaches to increase our prediction accuracy.

2 Methods

As we stated earlier, we focused on the classification of literary texts as either fiction or nonfiction. In this situation we defined nonfiction as anything that purports to be nonfictions (autobiographies, biographies, textbooks, etc.). Although works like biographies and autobiographies tend to come with a lot of bias and possible embellishments, we considered them nonfiction for our purposes. We decided on this because the author viewed the works as nonfiction and therefore likely wrote them using the conventions of the genre.

2.1 Data Source

The source for our data set was the web API of Project Gutenberg. This is a project aimed at making a selected count of literary classics available as free ebooks. This allowed us to quickly access hundreds of literary works to use in our model. Using a parser, we downloaded and parsed 963 instances. We then labeled each instance as either fiction or nonfiction based on wikipedia and goodreads.

There were a few hurdles that we had to address when it came to our data source. The first was that Project Gutenberg would have duplicates of some books. We didnt quite know why this was the case, but ultimately we decided that having a few duplicates in our data set wouldnt hurt our models ability to work. This is because the average scenario would be that we would have duplicates of fiction works at the same rate as those of nonfiction works. This meant that our overall data would set would have the same ratio. Another problem we had was that many books would be in different languages. We decided that since we were using english words, we should disregard these entries. As a result, we did not include any works that were written in different languages.

The last major hurdle was that the API that we pulled these books from did not have their genre listed in any obvious place. This meant that for our entire data set we had to hand label each instance as either fiction or nonfiction. This was the main reason that our data set ended up being smaller than we would have initially liked.

2.2 Data Set

2.3 Selected Models

3 Initial Results

4 Data and Feature Improvements

5 Final Results