FinalReport.Rmd

Book Detector

Abstract: [FIX THIS TO ADD CONCLUSIONS]

The main question that we hope to answer is: can we create an algorithm that can successfully identify what book series a paragraph is from? We will be using 2 book series: Harry Potter and Percy Jackson. In order to answer this question, we are creating a predictive model that will compare the words in the given paragraph to the common words in each book series. To do this, we used TF-IDF and Cosine similarity.

Introduction:

The purpose of this analysis is to be able to see if there is a recognizable difference in the words used in the Harry Potter vs Percy Jackson books. We are training a predictive model to classify a given text paragraph as from either Harry Potter or Percy Jackson. This analysis would benefit someone who is trying to create a copyright detector where they are trying to identify which source a specific chunk of text is from. The different sources could be the text corpuses, and a similar predictive model can be used to classify the copyrighted text.

Data:

The data that we are using are txt files of the Harry Potter and Percy Jackson books that we found online. There are no specific variables or observations, but rather we have a large dataset of text. To clean the data, we converted all of the letters to lowercase and removed the unwanted symbols, extra spaces, line breaks, special characters, and numbers.

Visualization:

These visualizations show that although some of the most frequent words are similar among the book series (for example, “said”, “like”, back”, “one”, “know”), there are also words that are unique to each series, mainly names of characters.

Analysis: [FIX THIS TO ADD RESULTS]

We plan to use TF-IDF to to calculate the “importance” of each word in the 2 series. Once we have that, we will use Cosine similarity text classification machine learning algorithm to classify the paragraph of text into either the Harry Potter or Percy Jackson series. At first, we were planning to use Naive Bayes, but then switched to Cosine similarity because it is more ideal for comparing the content similarity in text regardless of length (because the total number of words in the Harry Potter series is much larger than in the Percy Jackson series).

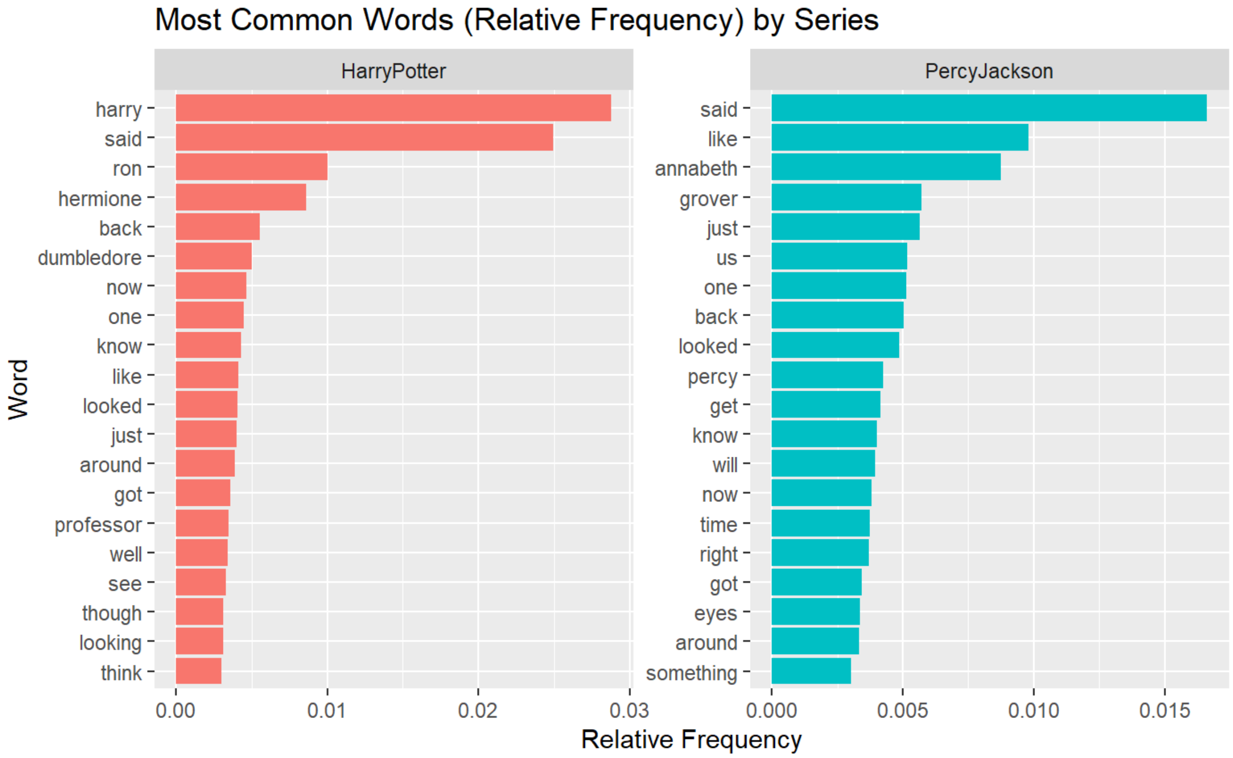

Figure 1: Word Frequency Histograms

Conclusions: [DO THIS]

Team Contributions:

Angelina How: - Created the Github Repository and all the project files, and shared them with Sara and Bethany - Added the book files into the directory - Edited the code to remove the stop words in the data - Typed up the README.RMD - Typed up the Final Report - Created the visualizations - Created the TF-IDF Cosine Similarity text classifier - Added additional necessary packages to the 00_requirements.R file

Bethany Galias: - Found the Percy Jackson books online - Added most of the required packages to the 00_requirments.R and described them - Cleaned, prepared, and saved the data of the Harry Potter books

Sara Munkhbayar: - Found the Harry Potter books online - Cleaned, prepared, and saved the data of the Percy Jackson books - Wrote the code to remove stop words in the data