Final Project Proposal

Blue Cookie; Group B8

Introduction and Questions of Interest

For this project, we are going to explore a new topic. The new topic that we wish to explore is the Harry Potter franchise of books. Some questions of interest to us include the following: + How has word choice evolved over the series of the books? + How has the tone of the books changed? + Has the author's writing style also changed throughout the time writing the franchise? + What are the common themes and topics that are prevalent in each book?

The data source that we will be using is the R package `bradleyboehmke/harrypotter` which has been used in the community for various educational purposes, so we will also be using this source. This package contains the Harry Potter novels that we will be able to import into R.

Exploratory Topic

The main focus of our project will be on text analysis. However, our potential way of going beyond is to implement a model of unsupervised learning into our analysis. In particular, we are thinking of using the LDA (Latent Dirichlet allocation) to classify different topics/themes within a book. Upon brief examination, we believe that the LDA principle is somewhat similar to k-means clustering: To find patterns from a set of words that were not previously labeled. We plan to employ this model by using an already, community built package in R, which we are still working on finding. Although we do hope to understand how the model works and the math/stat behind it, our main focus is to successfully identify and utilize an available package for the sake of this project.

The Final Product

The final product that we will be creating contains multiple different pieces to display help answer our interest questions. We will create a Shiny app to display the usage of user-inputed words throughout the different books. This visual will be able to show the frequency of the words over time (although time is relative to the book). We will also create a visual that maps the sentiments/emotions of each book. We will accomplish this by using a metric that displays the average sentiment score for a grouping of words (such as chapter or paragraph). In order to determine if the author's writing style has changed over time, we will perform a Chi-Square Test of Homogeneity to determine how similar the books are to each other. The LDA information will be displayed in a table format or perhaps a cluster (to be determined when we gain better understanding).

Schedule

By April 26, 2022

- Wrangle data
- Research LDA topic, find the appropriate packages

By April 28, 2022

- Prepare visuals
- Create models

By May 3, 2022 - Status Update 2

- Start creating first visuals

By May 5, 2022

- Finish all of the visuals

By May 10, 2022

- Finish the report