Literature Analysis

Divya Iyer, John Berry, Tijana Cosic, Brennan Ayres

**Motivation**

Every author has a different writing style. A sort of ‘literary signature,’ each author’s writing style comprises everything from the words used to the length of their sentences. It’s easy for us to state that two authors who typically write books of different genres will have different signatures. However, our project will allow us to enhance this understanding of authors’ signatures by comparing specific elements of their writing style and prose. This project could be interesting to future recruiters because it has the potential to be applied to various other types of publications in order to perform valuable author sentiment and signature analysis. Additionally, it will have the ability to process large amounts of information in order to discover similarities and differences between the content of two pieces of literature.

**Goals and Objectives**

The goal of our project is to use text mining and data visualization to compare a pair of ‘documents’ and show the similarities and differences between them. In order to achieve this output, we will provide our program with bodies of text to read. From this data, our program will glean information about authors based on their writing style and content.

* Utilize text analysis concepts (including mining, normalization, tokenization, etc.) to produce insightful data visualizations that compare and contrast two different pieces of text.
* Compare two works by the same author – what do they do differently? If they write books in different genres, how do they compare?
* Compare works by different authors in the same genre, different authors across different genres, and so on.
* Apply our software to other types of publications (blogs, poetry, scientific research articles, etc.)

**Downstream Goals**

* Analyze a body of text and determine its genre
* Determine author bias in blog posts, journal entries, articles, etc.

**Data Sources**

Kaggle Classic Literature Dataset

* <https://www.kaggle.com/mylesoneill/classic-literature-in-ascii?select=AUTHORS>

Google Books API

* <https://developers.google.com/books/docs/v1/using>

Reedsy Short Story Database

* ​​<https://blog.reedsy.com/short-stories/>

Books for Text Mining Training

(Source of 18k .txt files which could be used to train an AI. We might want to start smaller with something like the kaggle dataset.)

* <https://github.com/soskek/bookcorpus/issues/27>

Reddit post containing ~200k books. Might be too much for us to process and train AI on, and we probably don’t have enough GPU to realistically use this much material.

* <https://www.reddit.com/r/MachineLearning/comments/ji7y06/p_dataset_of_196640_books_in_plain_text_for/>

Database for downloading eBooks and journal articles

* <http://libgen.li>

Northeastern DRS potential datasets

* <https://library.northeastern.edu/services/digital-repository>

**Platform Architecture**

Our goal for this project is to output a database populated with various information including author sentiment/tone, complexity/sophistication, and prose/genre, all in order to establish an author’s unique signature.