Lily, Megan, Sophia, Francisco

Professor Ying Lin

IST 736 Text Mining

9 June 2024

Final Project Proposal: Gutenberg Books

**Data Mining Problem**

The problem consists in text mining an online library of ebooks to classify them into fiction and non-fiction categories based on their titles and subcategories.

**The Dataset -** [**https://www.kaggle.com/datasets/mateibejan/15000-gutenberg-books**](https://www.kaggle.com/datasets/mateibejan/15000-gutenberg-books)

For our final project, we have decided to utilize a large dataset from Kaggle containing a corpus of books scraped from the Gutenberg website. The dataset includes 15,000 book texts, their titles, and authors. In addition, the metadata comprises the categories or genres in which each book falls, stored in volunteer-curated “bookshelves”. The dataset contains links to download the book texts as well as the audiobook files however we will not be utilizing this data during our execution of the project. The languages of these books are extensive, so we will be focusing on English-language titles for our analysis.

**Initial Strategies for Preprocessing, Transformation, and Analysis**

‘Bookshelves’ contains 118 different categories, with an average of 23 cases per category. Therefore, the initial step in preprocessing will involve grouping these categories into fiction and non-fiction based on the text of the Bookshelf label (ex: Maps and Cartography). Once the classification categories are decided, we will proceed to clean the text and extract features using techniques like tokenization and lemmatization. We will test several classification algorithms and fine-tune them to find the best model for predicting the selected categories. Some metrics for evaluation we will use include the F1-score and the area under the curve (AUC).