**Introduction-**

This project addresses the problem of Genre Identification from the given subset of Gutenberg Corpus. The data subset contains around 1000 “19 Century English Fiction books. Given data is labeled with additional details about author name and book names etc. This is supervised classification problem where we aim to predict the genre of any given Fiction book.

**Why It’s a problem?**

Since Books are lengthy in nature and there doesn’t exists any informative abstract/preface of the book to help identify the genre of the book, one has to read the complete book or take expert’s help in order to label it while categorizing. Manual labeling is a time consuming task which we aim to solve by training a classification model which can learn from this dataset about what features make a book fictional and then correctly predicts the exact genre of the book.

In the given dataset for there are issues like having more books from Literacy genre and very less from analogies hence class imbalance.

**Project Planning-**

We shall Share the task of each phase such as Data loading/pre-processing/EDA/Feature Selection/Model Selection and Model Evaluations equally as we progress, based on understanding and also shall documents the progress step by step.

**Problems to be addressed-**

1. Correctly predict the genre of a unknown Fiction book
2. Handle the book data corpus, cleaning and pre-processing the data in order to get suitable features
3. Find features which makes a book fiction book by trying to determine the overall sentiment of a book, plot, characters etc.
4. Find the sentiment of the a book from text
5. Address class imbalance problem
6. Find if there is underlying pattern/correlation between genre and Title of the book
7. Find if there is underlying pattern//correlation between Genre and the Authors
8. Find out which classification model works best on large books set
9. Compare performance of various models
10. While trying to choose a classification model try to get a model with most accurate and faster predictions.

**Tools**

Tools are being used as of now- *Python 3*

Libraries- *SkLearn, NLTK* etc, *Git* for Code sharing/Versioning