About Dataset:

The Gutenberg English Fiction contains 1079 html files and we have to explore 996 files. There are a total of 10 genre categories. Each of the html files contains <p> tags. So, parsing of the html files will be done in order to remove these tags.

Project Plan:

1. **Data Pre-processing:**

- Each book can be split into chunks and each chunk can be analyzed further by dividing into sentences. Unwanted characters/numbers/punctuations will be removed from data.
- Stemming and Lemmatization will be performed to normalize words to their root forms.
- POS tagging to extract the content words which carries important information.
- Anaphora Resolution to resolve pronouns, verb phrases to items seen earlier or later in the document.
- Punctuation Analysis to detect the expressiveness of the book.
- Named Entity Recognition will be carried out for each sentence with part of speech tags. This helps in detection of Relations.

2. **Feature Selection:**
    - Sentence length
    - Richness of vocabulary
    - Gender Identification
    - Complexity of sentence
    - Number of characters
- Tools
    - NLTK toolkit
    - Stanford's NLP tools like "the part-of-speech (POS) tagger", "the named entity recognizer (NER)", "the parser", "sentiment analysis"
    - Semi-Markov Quotation Model

3. **Model Selection:**

Models under consideration for classification:

- Convolution Neural Networks on feature vectors to classify the documents.
- SVM by having a final feature vector for each document.
- K-Nearest Neighbor for the feature vectors.

4. **Validation:**

K-fold Cross-validation can be used for different models by making use of package sklearn.model_selection.KFold

5. **Evaluation:**

For evaluation we shall use metrics such as Confusion Matrix, Accuracy Score and other measures will be decided as the project progresses. Packages like sklearn.metrics can be used to calculate evaluation measures.

6. **Visualization:**

Python libraries such as seaborn and matplotlib can be used to visualise the features selected as well as the performance of different models.