PROJECT PLAN

Problem Definition: Genre identification on a subset of the Gutenberg Corpus.

guten_genre	Count of guten_genre
Literary	794
Detective and Mystery	111
Sea and Adventure	36
Love and Romance	18
Western Stories	18
Ghost and Horror	6
Humorous and Wit and Satir	6
Christmas Stories	5
Allegories	2

Dataset Characteristics:

- Ground Truth : {<book_Content>, <genre_label>}
- Observation : High class imbalance problem observed
  83 non-classified instances

Initial Project Strategies:
We first were posed with a dilemma of either using a Discriminative approach vs a Generative approach. Then the preliminary Machine Learning Pipeline meant, employing the three basic blocks i.e. Data Pre-Processing and Model building, Model Selection, and Model Evaluation. Further,

Dataset Pre- Processing: This block focusses on improving the 'relevancy' of the features with respect to the genre.
Feature Reduction

- Stop words recognized and removed for English Language using nltk library.
- Lemmatization, tokenization and stemming using the nltk library.
- Feature Selection and Extraction using sklearn library and PCA.
- Word Embeddings using Wor2Vec.
- Feature Scaling using CBOW and TFIDF from sklearn.
- Feature map using English Sentiment Analysis from nltk library.

Model Building: This sub-block emphasizes on building the actual models from some of the approaches mentioned above. The models currently planned for build which will be passed on to Model Selection.

- A model built on Matrix Factorization methods, SVD and dimensionality reduction
- A neural net built on Word Embeddings.
- A custom training loop based on the correlation of sentiments from book_ data to genre(clustering).
- An SVM using scaled features.

Model Selection & Model Evaluation: Model Selection and Model Evaluation have basic objectives of improving classification and reducing generalization error respectively. All the above-mentioned classifier settings has some inconsistencies. To mitigate further errors and perform some viable validation, we plan on using a Train-Test-Split setting. We plan on also using cross validation.

Models Limitations/weak points:

- The MF Model may suffer from calculating the principal component because of sparseness.
- It might not be possible to learn all the possible examples with sentiments on a scale [+pos, -ve, neu].

Evaluation Limitations and Proposals:
Traditionally, classification problems can be most efficiently solved using either supervised or unsupervised learning methods when the class division is similar. However, in our specific scenario we are forced to deal with the class imbalance problem that is described in the above table. For instance, it would be highly erroneous when accuracy is used as an evaluation metric in Model Evaluation. We plan on using an alternate evaluation criterion/procedure which would tax the most frequent genre, in our case 'Literary'. We plan on using these Evaluation metrics which can be imported from the sklearn library,

- F1 measure
- Precision
- Recall
- Accuracy

***please note that we have only mentioned the experiments that would like to perform, and the final model would be an experiment with the best performance measure.**