Book genre prediction

By Pramod Anantha
and
Shreyash Bali

Dataset : The CMU Book Summaries dataset we are using consists of 16,559 books . The Title, Author , Summary and the Genres it belongs to are given for each book.

Problem Statement : Supervised Text Classification for Multiclass Book Genre Classification using NLP

Software Used : Python , Anaconda , NLTK , Scikit-Learn, Tensorflow/Keras

Methods :

- Preprocessing : RegEx is used to remove unwanted text and create a 2D array for the genres.
  Remove Stop-Words
- Tokenizing : Here each word in the summary is tokenized based on space between words.
- Vectorizing : Scikit-Learn library is used to implement Bag of Words and TF-IDF vectorization and convert the summaries into an array representation . Google's Word2Vec and other similar Vectorization techniques which are more suitable for Book Summary will also be tried
- Feature Selection : LDA, Chi-Square test is used to select the most important features from a bigger set of features
- Dimensionality Reduction : PCA will be implemented and included if it contributes to higer accuracy
- Classification : Multinomial NB, SVM , Logistic Regressing , Random Forest , Xgboost,RNN, LSTM, GRU,
- Validation : 10-Fold Cross Validation

Current Progress: A basic pipeline of the above methods have been implemented using bag of words and Multinomial NB. We will now build on this to improve accuracy