WORKFLOW STEPS

APPROACH 1 – PCA And RFC based

1) TRAINING SET

* Created a corpus of all the fifty text files belonging to fifty different authors
* Applied transformations associated with lower casing, removal pf numbers, punctuations and whitespaces
* Removed stopwords, that are present in the smart vocabulary list, from the corpus
* Formulated a Bag of words based Sparse matrix for the training corpus
  + Number of terms 32241
  + Applied sparsity-based filter to remove columns with more than 98.5 sparsity, terms reduced to 2326
* Computed TFIDF scores for each document for the set of words
* Removed columns with sum across column values of TFIDF equal to zero
* Dimension changed to rows=2500 col/words=2310
* **Stored the set of words used in training set in a placeholder named (words\_from\_trainset)**
* **Created a ‘pseudo-word-column’ for the words that may surface in test set but are not present in training set**
  + Value for pseudo word was determined individually for each document, it was equal to the 0.2 quantile value of all the non-zero tfidf for words present in the specific row.
* Applied **PCA transformation** for the entire column sets
  + We plotted cumsum of variation\_explaination with number of components and found out that we will require at least 500 words to explain more than 60 percent of the variation
  + **We stored the feature contribution of each words (2311) in these 500 axes in another placeholder, we will use this to transform our test set**
* Now we had our predictors, we formulated the target variable using rep function 50 authors repeated 50 times
* Then we trained a random forest-based classifier model between pca predictors and target variables

2) TEST SET

* We used similar transformation as applied on the train set to create TFIDF matrix of the test set
* Then we found out the words that are common in the training and test set using intersection function, it turned out to be 1999
* We realise dthat we can only use contribution of these 1999 and 1 pseudo word feature for creating our 50 test pca predictors
* We computed mean of the tfidf scores of the words that are not present in the train set and set it as pseudo variable
* We used words present in both sets and pseudo word based variable to select pca rotation contribution, followingly computed pca scores for each of the test document

Evaluated the test set accuracy  
Accuracy score was not that high

APPROACH 2 – NAÏVE BAYES APPROACH

Workflow Steps

* For each test set, we trained 50 different naïve bayes model based on the word bag count of each repository of documents belonging to 50 individual writers
* For all trained model we computed the probability of occurrence of each word specific to that repository/folder for individual writer
* Following, we took intersection of words that occur in both the train and the test set

We had this (50 combinations of intersecting words for each test folder)

* Subsequently, we computed likelihood using the product of log of prior probability and counts of intersecting words.
  + We obtained 50 likelihood scores from all the trained models for each of the 50 documents present in the repository
  + We analysed which model gave the highest/max likelihood for each document and assigned the value of that model as predicted folder set for the document
  + Then we had a series of 50 documents with their predicted set/author
* We computed accuracy of for every test folder and stored it in a place holder variable

Finally, we plotted the computed accuracy for each test folder.

We observed a large variance in test accuracy prediction, for example documents associated with author number 21, Karl Penhaul were predicted with an accuracy of more than 0.98.

Whereas, files attributed to author 12, Graham Earnshaw was predicted with just an accuracy of 0.04

Total repository accuracy (for 50\*50 files) was 0.46.