Data and Preprocessing

Reuter_50_50 Data Set[1] from UCI machine learning repository was chosen for this experiment. These corpus has already been used in other authorship identification experiments. There are 2500 training texts by 50 authors and 2500 test texts by 50 authors. 50 authors of texts labeled with at least one subtopic of the class CCAT were selected that attempt to minimize the topic factor in distinguishing among the texts.

For using these text in our experiment, these texts need to be preprocessed. Each text was tokenized, lemmatized and stem. After these steps, features can be extracted from these clean texts.

Baseline

We set up a trigrams model with smoothing as the baseline of this experiment.