### Cmpe493 Assignment 2 Report

(i) Data preprocessing steps:

* Loaded the data from the files and unescaped the HTML.
* Tokenized the text by splitting it into words using a regex pattern.
* Loaded stopwords from a file and removed them from the tokenized text.
* Extracted articles from the data, along with their topics and LEWISSPLIT values.
* Identified the top 10 topics based on their frequency in the dataset.
* Filtered the articles to include only those with at least one of the top 10 topics.
* Split the articles into training, and test sets based on their LEWISSPLIT values.
* Split the train set into half to train and development sets.

Vocabulary size of the trained data is 23939. This value does not contain the number of test data.

(ii)

Top 10 classes and the number of documents in the training, test, and dev sets:

**earn**: Training - 1251, Test - 1087, Dev - 1626

**acq**: Training - 781, Test - 719, Dev - 869

**money-fx**: Training - 336, Test - 179, Dev - 202

**crude**: Training - 150, Test - 189, Dev - 239

**grain**: Training - 198, Test - 149, Dev - 235

**trade**: Training - 209, Test - 118, Dev - 160

**interest**: Training - 177, Test - 131, Dev – 170

**wheat**: Training - 88, Test - 71, Dev - 124

**ship**: Training - 99, Test - 89, Dev - 98

**corn**: Training - 78, Test - 56, Dev - 104

* Total number of documents in the training set: 3031
* Total number of documents in the dev set: 3459
* Total number of documents in the test set: 2545

(Before splitting data into training and dev, there were 6490 train documents)

* Number of documents with more than one of the top 10 classes in the training set: 298
* Number of documents with more than one of the top 10 classes in the dev set: 322
* Number of documents with more than one of the top 10 classes in the test set: 214

(iii) Parameter tuning:  
Development set is used to perform parameter tuning. The development set was selected by splitting the original training set into two equal halves. The first half was used as the development set, and the second half was used as the new training set. Different alpha values (1, 0.5, 2) and model types ('multinomial', 'bernoulli') are used to find the best parameters. At the end, models are trained with the merged data of train and development sets with the best alpha value as described in the assignment. Corresponding values of test data are predicted by using the models that are trained by the merged data.

(iv) Evaluation results on the test set and randomization test results:

The evaluation results on the test set are printed in the code for both 'multinomial' and 'bernoulli' models, including macro and micro precision, recall, and F-score. The randomization test results are also printed as the P-value.

Multinomial Naive Bayes Test Results with the Best Alpha Value 1:

* Macro Precision: 0.7005
* Macro Recall: 0.7696
* Macro F-score: 0.7334
* Micro Precision: 0.9277
* Micro Recall: 0.9277
* Micro F-score: 0.9277

Bernoulli Naive Bayes Test Results with the Best Alpha Value 2:

* Macro Precision: 0.0730
* Macro Recall: 0.2216
* Macro F-score: 0.1098
* Micro Precision: 0.0244
* Micro Recall: 0.0244
* Micro F-score: 0.0244

Randomization test results yield to the 0 value of p\_value. This seemed okay hence these two models are significantly different from each other as can be seen above.

Text

Description automatically generated