BBC News Classifier Overview

BBC News Classifier downloads and extracts the news articles dataset, splits the dataset into labeled and unlabeled datasets. It builds the text based classification models based on the labeled data and then the built models are used for predicting the news category of unlabeled data.

Corpus Description:

BBC News articles dataset downloaded from the link here. The associated publication is:
D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006. [PDF] [BibTeX].

BBC News articles are categorized as business, politics, entertainment, sports and tech articles.
The news articles are available in the raw text format.

Project Approach:

BBC News Classification process consists of following stages:

Corpus Extraction Phase:

In this phase Corpus is downloaded to the data->external folder and extracted dataset is split into labeled and unlabeled data and stored in the data->raw folder.
Below analytics are generated on the labeled data in this phase.

- Number of raw text files in the Corpus.
- Number of News categories.
- Number of paragraphs in the corpus.
- Number of sentences in the corpus.
- Number of words in the corpus.
- Number of unique words (vocabulary) in the corpus.
- The lexical diversity of the corpus.
- Number of paragraphs per document.
- Number of sentences per document.

Corpus Preprocessing and Wrangling Phase:

In this phase text contained in each of the files is converted into a list of paragraphs. Each paragraph in turn is converted into a list of sentences. Each sentence is again is converted into a list of words and their associated POS (part of speech) tag tuples.
All the text files after conversion to the nested lists format are saved to disk.

Train Test Split Phase

In this phase all the labeled data is split into Train data (67%) and Test data (33%).

Transformation Pipeline and Model Build Phase

In this phase Transformation Pipelines are created for each of the classification algorithms.
Below are the multiclass text classification algorithms used for generating models:

- MultinomialNB classification algorithm

- LogisticRegression classification algorithm
- LinearSVC classification algorithm
- Stochastic Gradient Descent Classification algorithm

The Transformation Pipeline is chained as Text Normalization process, TF-IDF Vectorization process and finally model fitting and transformation process.
In the Text Normalization Process pickle files are loaded in memory and all of the words in each file are subjected to below dimensionality reduction techniques.

- All the words are converted into lower case
- Some common English stop words and corpus specific stop words are filtered
- The words of length 1 and punctuation marks are discarded
- Remaining words are lemmatized

The output of TF-IDF vectorization process is a sparse matrix and as part of the vectorization process again dimensionality reduction is performed by instructing TF-IDF vectorization process to ignore the terms occurring in more than 50% of the documents and also the terms which are in < 10% of the documents.
After the models are successfully fitted, the models are saved to the disk in the pickle format.

Reporting and Visualization Phase

In this phase the Classification Reports, Confusion Matrices and Reports containing the Frequency Distribution of words are generated using the Training-Test data for various classification algorithms.
Classification Report consists of Precision, Recall and F1 scores for each news category is calculated.
Separate Classification Report is generated for each of the classification algorithms.
Confusion Matrix for each of the classification algorithm consists of a table with actual value counts and Predicted value counts for each of the News categories. Each row represents a predicted class and a column represents actual class.
Frequency Distribution of words report consists of a Horizontal Histogram for top 50 words.

Prediction Phase

In this phase news category for the unlabeled text files is determined using a web application.
While predicting a class for an unlabeled news article, the application loads all the classifier models into memory.
Gets the value predicted by each of the classifiers and the value predicted by majority of the classifiers is returned as category of the news article. This is called Vote classification.