Technical Overview of BBC News Classifier

BBC News classifier consists of 2 components:

Batch Component:

It is developed using Python modules, packages and libraries.

Corpus Extraction Phase:

- Used **requests** library to download the zip file from the given url and used other modules to extract the corpus.

Corpus Preprocessing and Wrangling Phase:

- Used **multiprocessing** library for parallel preprocessing of the extracted raw text files.
- Used **nltk** library for tokenization of paragraphs, sentences and words in the files.
- Used **pickle** library for saving the preprocessed documents.

Train-Test Split Phase:

- Used **scikit-learn** library for splitting the labeled into Train and Test data

Transformation Pipeline and Model build Phase:

- Used **WordNetLemmatizer** from **nltk** library for lemmatization.
- Used **scikit-learn** library for building various classification models.

Reporting and Visualization Phase:

- Used yellowbrick library to build classification reports, confusion matrices and frequency distribution reports.

Online Component:

It is a Python flask web application and the front-end part of it is developed using JQuery, JavaScript, JSON, CSS and HTML.

Batch and online components are deployed using bdist_wheel generated from the source code.

Tools

The development tools used for building the BBC News classifier are

Python IDLE Editor,

Jupyter Notebooks,

Python CookieCutter template for data science and Cookie-cutter template is downloaded using **Git-Bash**.