CSE 587

DATA INTENSIVE COMPUTING

LAB 3

SIDDHARTH SELVARAJ

#50247317

VIGNESHWARAN VASANTHAKUMAR

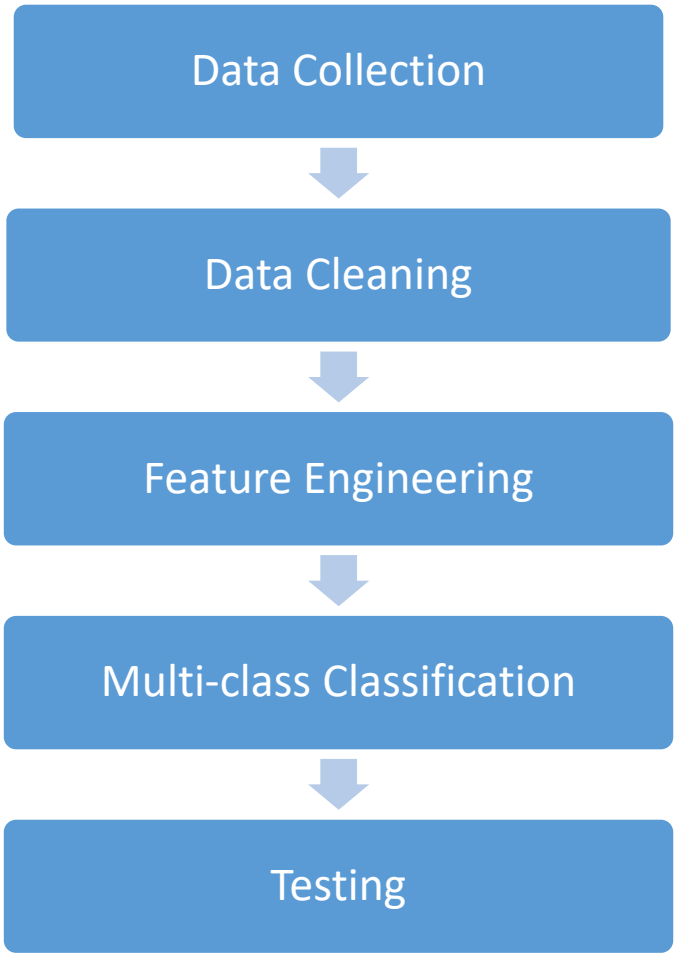
#50248708

LAB 3: DATA ANALYTICS PIPELINE USING APACHE SPARK

In this project, multi-class classification is used to classify the articles into the respective categories they belong to. Logistic Regression and NaiveBayes classifiers are used for predicting the categories of the articles and their accuracies are studied.

The environment chosen for the project is PySpark in VM.

The following is the block diagram of the steps involved in the project:



1. Data Collection:

New York Times articles are used as the input data. Data is collected with the NewYorkTimes API using the nytimesarticle package in python. Data is collected for the following categories: Business, Politics, Sports and Medical. The collected data are stored in separate directories. The data is then processed in the next step.

2.Cleaning the data:

The data that has been collected is cleaned in the following steps:

Tokenization:

Data is initially split into words(tokenized) using the regexTokenizer() which tokenizes the data with regular expressions.

Removing stop words:

After the tokenization of the data, the stop words are removed using the StopWordsRemover() which removes all the stop words in the data.

3.Feature Engineering:

After the data has been cleaned, the features are extracted characterizing each category. The categories are converted into labels from 0 to 3 using the StringIndexer(). The features are extracted using the functions HashingTF() and IDF().

HashingTF:

HashingTF is a Transformer which takes sets of terms and converts those sets into fixed-length feature vectors.

IDF:

IDF is an Estimator which is fit on a dataset and produces an IDFModel. The IDFModel takes feature vectors and scales each column. Intuitively, it down-weights columns which appear frequently in a corpus.

HashingTF is used to hash the sentence into a feature vector and IDF is used to rescale the feature vectors.

A pipeline is constructed using the above functions which cleans the data and extracts the features.

pipeline = Pipeline(stages=[regexTokenizer, stopwordsRemover, hashingTF, idf, label_stringIdx])

4.Multi-class Classification:

The classification algorithms used in this project are Logistic Regression and Naïve Bayes.
The dataset is split into training data (80%) and test data (20%). The package pyspark.ml is used for the multi-class classification.

Logistic Regression:

Logistic Regression is carried out using the function,
LogisticRegression(maxIter=20, regParam=0.3, elasticNetParam=0).

Naïve Bayes classification:

The Naïve Bayes classification is carried out using the function,
NaiveBayes(smoothing=1).

5.Testing:

Unknown set of articles are used as the test set and it is processed into the pipeline which cleans and extract features for the test data.
This is then used in the classification algorithm for predicting the categories.

6.Result

The accuracy for test data set using Logistic Regression is 66.53%.

```
18/05/11 16:29:02 INFO TaskSchedulerImpl: Removed TaskSet 76.0, whose tasks have all completed, from pool
18/05/11 16:29:02 INFO DAGScheduler: ResultStage 76 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.365 s
18/05/11 16:29:02 INFO DAGScheduler: Job 43 finished: collectAsMap at MulticlassMetrics.scala:53, took 43.475108 s
-------Accuracy of test data using logistic_regression-----: 66.5336339614%
18/05/11 16:29:02 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 16:29:02 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 16:29:02 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 16:29:02 INFO MemoryStore: MemoryStore cleared
```

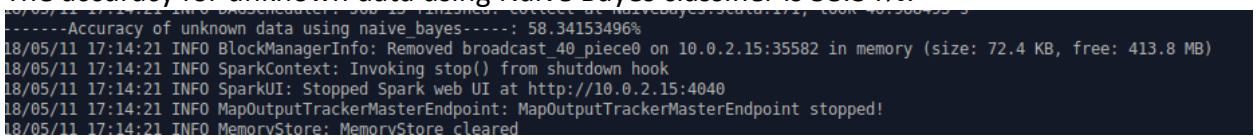
The accuracy for unknown data set using Logistic Regression is 57.47%.

```
-------Accuracy of unknown data using logistic_regression-----: 57.47841685%
18/05/11 17:05:08 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 17:05:09 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 17:05:09 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 17:05:09 INFO MemoryStore: MemoryStore cleared
```

The accuracy for test data set using Naïve Bayes classifier is 66.81%.

```
18/05/11 16:47:47 INFO TaskSchedulerImpl: Removed TaskSet 23.0, whose tasks have all completed, from pool
18/05/11 16:47:47 INFO DAGScheduler: ResultStage 23 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.478 s
18/05/11 16:47:47 INFO DAGScheduler: Job 17 finished: collectAsMap at MulticlassMetrics.scala:53, took 46.736485 s
-------Accuracy of test data using naive_bayes-----: 66.8069939376%
18/05/11 16:47:47 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 16:47:47 INFO BlockManagerInfo: Removed broadcast_55_piece0 on 10.0.2.15:41367 in memory (size: 1997.0 B, free: 413.0 MB)
18/05/11 16:47:47 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 16:47:47 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 16:47:47 INFO MemoryStore: MemoryStore cleared
```

The accuracy for unknown data using Naïve Bayes classifier is 58.34%.



Reference:

1) https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35