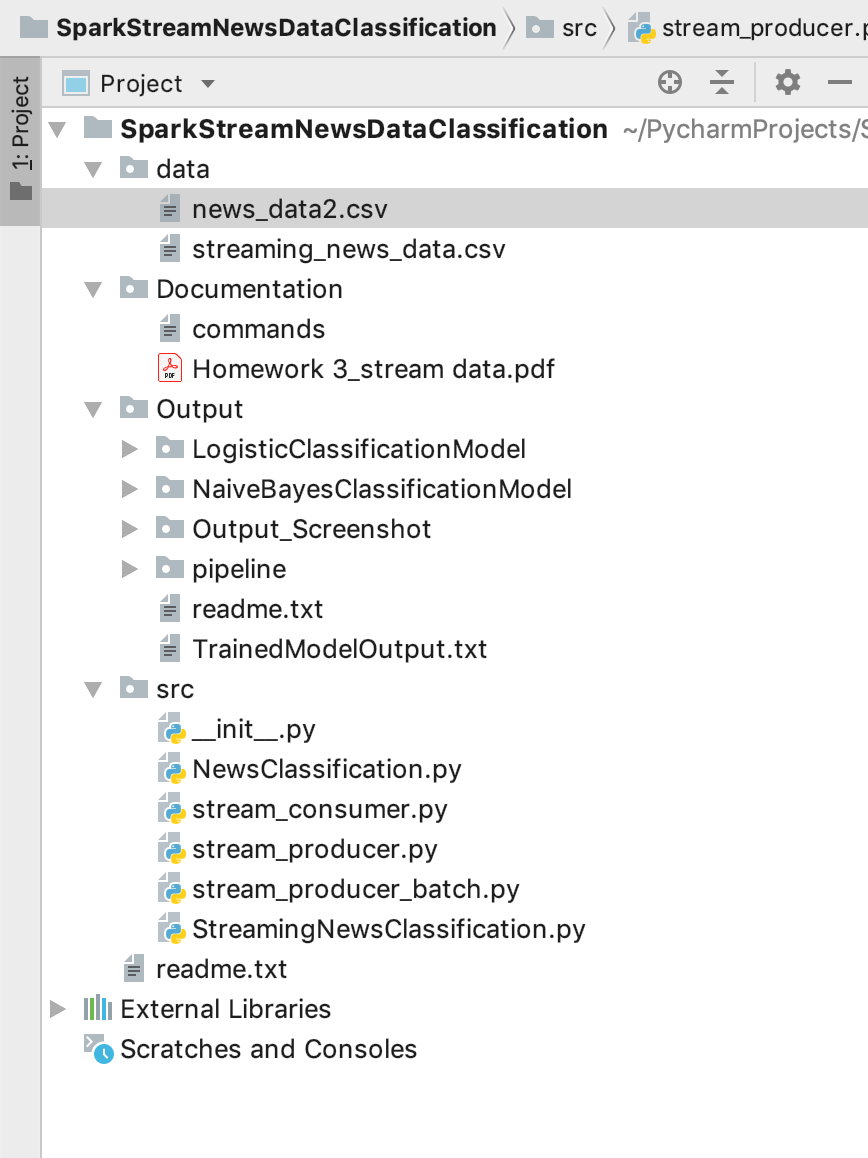
**Multi News Classification**

**Big data Management Analytics and Management**

The Program trains a Pipeline model with Tokenizer, stop word remover, Labialize, TF-IDF, vectorizer and two classifiers i.e. Logistic Regression and Naïve Bayes. Then it compares result of both classifier(Logistic and Naïve Bayes) on spark streaming data.

**Project Structure :**

**Dataset :**

The Consumer News Dataset file is present data folder with name: news\_data2.csv

**How to Execute and Run Project :**

1) Execute the following command on your command prompt to run the stream\_producer script:  
  
**python3 stream\_producer.py API-key fromDate toDate**  
  
**Ex: python3 stream\_producer.py 405cb3e5-b364-4df8-9f4a-905210534c1d 2019-01-3 2019-03-24**  
2) Start Zookeper and Kafka :

**zookeeper-server-start /usr/local/etc/kafka/zookeeper.properties & kafka-server-start /usr/local/etc/kafka/server.properties**  
  
3) Start the Consumer for kafka producer :  
Run: **python3 stream\_consumer.py to consume the kafka topic data and save to news\_data.csv file**  
  
4) Train and Test the Classifier Logistic and Naive Bayes on the consumed data with pipilene to perform data analysis and pre-processing on set of 30000 rows of news articles.  
Run: **python3 StreamingNewsClassification.py**  
  
-> This will Train both the classifiers and save the model in output folder along with pipeline  
  
5) Create Kafka Direct Stream Topic "guardian2stream":  
  
**kafka-topics --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic guardian2stream**  
  
6) Get the Spark Streaming jar from :  
https://search.maven.org/artifact/org.apache.spark/spark-streaming-kafka-0-8-assembly\_2.11/2.4.1/jar  
Search : org.apache.spark:spark-streaming-kafka-0-8-assembly\_2.11:2.4.1 and Download  
  
7) Start and Run the kafka Batch Streaming to producer stream data on newly created Topic "guardian2stream":  
  
Run: **python3 stream\_producer\_batch.py**  
8) Start the Streaming Data Classifier in spark using below command:  
**spark-submit --jars /Users/harshverma/Downloads/spark-streaming-kafka-0-8-assembly\_2.11-2.4.1.jar ~/PycharmProjects/SparkStreamNewsDataClassification/src/StreamingNewsClassification.py**  
9) Check Batch Streaming Classification output, Multiclassification Metrics, Performance of both classifiers in console

**Model Train and Metrics Evaluation Output :**

**Pipeline Output:**

/Users/harshverma/anaconda3/bin/python /Users/harshverma/PycharmProjects/SparkStreamNewsDataClassification/src/NewsClassification.py  
WARNING: An illegal reflective access operation has occurred  
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/Users/harshverma/anaconda3/lib/python3.6/site-packages/pyspark/jars/spark-unsafe\_2.11-2.4.1.jar) to method java.nio.Bits.unaligned()  
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform  
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations  
WARNING: All illegal access operations will be denied in a future release  
19/04/12 19:20:32 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
+---+-----------------+--------------------+  
|\_c0| \_c1| \_c2|  
+---+-----------------+--------------------+  
| 19| b'Coming of age|b'Coming of age: ...|  
| 13| b'Billie Eilish|b'Billie Eilish: ...|  
| 18| b'Experience|b'Experience: I r...|  
| 26|b'Madeline Miller|b'Madeline Miller...|  
| 23| b'Joaqu|b'Joaqu\xc3\xadn ...|  
+---+-----------------+--------------------+  
only showing top 5 rows  
  
root  
 |-- \_c0: integer (nullable = true)  
 |-- \_c1: string (nullable = true)  
 |-- \_c2: string (nullable = true)  
  
+-------+------------------+-------------------+--------------------+  
|summary| index| heading| text|  
+-------+------------------+-------------------+--------------------+  
| count| 28441| 28441| 28441|  
| mean|11.506100348089026| null| null|  
| stddev| 7.870803620593797| null| null|  
| min| 0|"b""'County lines'"|"b""'County lines...|  
| max| 33| b'\xe2|b'\xe2\x80\x98The...|  
+-------+------------------+-------------------+--------------------+  
  
+-----+-----+  
|index|count|  
+-----+-----+  
| 8| 2969|  
| 4| 2956|  
| 1| 2277|  
| 12| 2116|  
| 7| 1838|  
| 19| 1265|  
| 2| 1263|  
| 17| 1125|  
| 13| 1016|  
| 10| 992|  
| 16| 856|  
| 18| 848|  
| 21| 846|  
| 23| 845|  
| 5| 716|  
| 26| 709|  
| 24| 708|  
| 6| 598|  
| 15| 572|  
| 3| 443|  
+-----+-----+  
only showing top 20 rows  
  
  
+--------------------+--------------------+  
| text| features|  
+--------------------+--------------------+  
|b'Coming of age: ...|(3000,[8,13,19,22...|  
|b'Billie Eilish: ...|(3000,[3,36,37,75...|  
|b'Experience: I r...|(3000,[2,22,24,25...|  
|b'Madeline Miller...|(3000,[13,29,33,7...|  
|b'Joaqu\xc3\xadn ...|(3000,[3,21,57,80...|  
|b'Edwyn Collins: ...|(3000,[13,16,23,2...|  
|"b""Back Brexit d...|(3000,[433,1103,1...|  
|b'Prisoner by Jas...|(3000,[0,6,23,24,...|  
|b'What to see thi...|(3000,[3,6,13,21,...|  
|b'Toxins in the a...|(3000,[6,8,23,28,...|  
|b'The Matthew Her...|(3000,[25,117,120...|  
|b'House prices in...|(3000,[8,14,25,28...|  
|b'Football transf...|(3000,[5,25,28,41...|  
|b'Nathan Chen v Y...|(3000,[0,6,13,17,...|  
|b'Authorities at ...|(3000,[5,9,35,59,...|  
|b'Far-right terro...|(3000,[1,28,34,41...|  
|"b""'I'd like to ...|(3000,[330,378,39...|  
|b'Another year of...|(3000,[8,15,24,63...|  
|b'The Other Ameri...|(3000,[3,14,18,21...|  
|b'Brexit: Theresa...|(3000,[18,22,23,2...|  
+--------------------+--------------------+  
only showing top 20 rows  
  
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----+  
|index| heading| text| words| filtered| rawFeatures| features|label|  
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----+  
| 19| b'Coming of age|b'Coming of age: ...|[b'coming, of, ag...|[b'coming, age:, ...|(3000,[8,13,19,22...|(3000,[8,13,19,22...| 5.0|  
| 13| b'Billie Eilish|b'Billie Eilish: ...|[b'billie, eilish...|[b'billie, eilish...|(3000,[3,36,37,75...|(3000,[3,36,37,75...| 8.0|  
| 18| b'Experience|b'Experience: I r...|[b'experience:, i...|[b'experience:, r...|(3000,[2,22,24,25...|(3000,[2,22,24,25...| 11.0|  
| 26| b'Madeline Miller|b'Madeline Miller...|[b'madeline, mill...|[b'madeline, mill...|(3000,[13,29,33,7...|(3000,[13,29,33,7...| 15.0|  
| 23| b'Joaqu|b'Joaqu\xc3\xadn ...|[b'joaqu\xc3\xadn...|[b'joaqu\xc3\xadn...|(3000,[3,21,57,80...|(3000,[3,21,57,80...| 13.0|  
| 13| b'Edwyn Collins|b'Edwyn Collins: ...|[b'edwyn, collins...|[b'edwyn, collins...|(3000,[13,16,23,2...|(3000,[13,16,23,2...| 8.0|  
| 4|"b""Back Brexit d...|"b""Back Brexit d...|["b""back, brexit...|["b""back, brexit...|(3000,[433,1103,1...|(3000,[433,1103,1...| 1.0|  
| 26|b'Prisoner by Jas...|b'Prisoner by Jas...|[b'prisoner, by, ...|[b'prisoner, jaso...|(3000,[0,6,23,24,...|(3000,[0,6,23,24,...| 15.0|  
| 0|b'What to see thi...|b'What to see thi...|[b'what, to, see,...|[b'what, see, wee...|(3000,[3,6,13,21,...|(3000,[3,6,13,21,...| 25.0|  
| 12| b'Toxins in the air|b'Toxins in the a...|[b'toxins, in, th...|[b'toxins, air,, ...|(3000,[6,8,23,28,...|(3000,[6,8,23,28,...| 3.0|  
| 13|b'The Matthew Her...|b'The Matthew Her...|[b'the, matthew, ...|[b'the, matthew, ...|(3000,[25,117,120...|(3000,[25,117,120...| 8.0|  
| 20|b'House prices in...|b'House prices in...|[b'house, prices,...|[b'house, prices,...|(3000,[8,14,25,28...|(3000,[8,14,25,28...| 21.0|  
| 1|b'Football transf...|b'Football transf...|[b'football, tran...|[b'football, tran...|(3000,[5,25,28,41...|(3000,[5,25,28,41...| 2.0|  
| 7|b'Nathan Chen v Y...|b'Nathan Chen v Y...|[b'nathan, chen, ...|[b'nathan, chen,   
year,...|(3000,[8,15,24,63...|(3000,[8,15,24,63...| 3.0|  
| 26|b'The Other Ameri...|b'The Other Ameri...|[b'the, other, am...|[b'the, americans...|(3000,[3,14,18,21...|(3000,[3,14,18,21...| 15.0|  
| 4| b'Brexit|b'Brexit: Theresa...|[b'brexit:, there...|[b'brexit:, there...|(3000,[18,22,23,2...|(3000,[18,22,23,2...| 1.0|  
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+-----+  
only showing top 20 rows

**Dataset Count:**

Training Dataset Count: 22757  
Test Dataset Count: 5684

**Logistic Classification Output:**

19/04/12 19:21:09 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS  
19/04/12 19:21:09 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS  
+------------------------------+-----+------------------------------+-----+----------+  
| text|index| probability|label|prediction|  
+------------------------------+-----+------------------------------+-----+----------+  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
|b'City of love? Christian r...| 8|[0.9483385908666192,0.00581...| 0.0| 0.0|  
+------------------------------+-----+------------------------------+-----+----------+  
only showing top 10 rows

**Test Error for Logistic Regression :3.6985719078893364%  
Test Accuracy for Logistic Regression :96.30142809211067%  
Test weightedRecall for Logistic Regression :0.9623504574243491  
Test weightedPrecision for Logistic Regression :0.9691435850273702  
Test f1 score for Logistic Regression :0.9630142809211066**

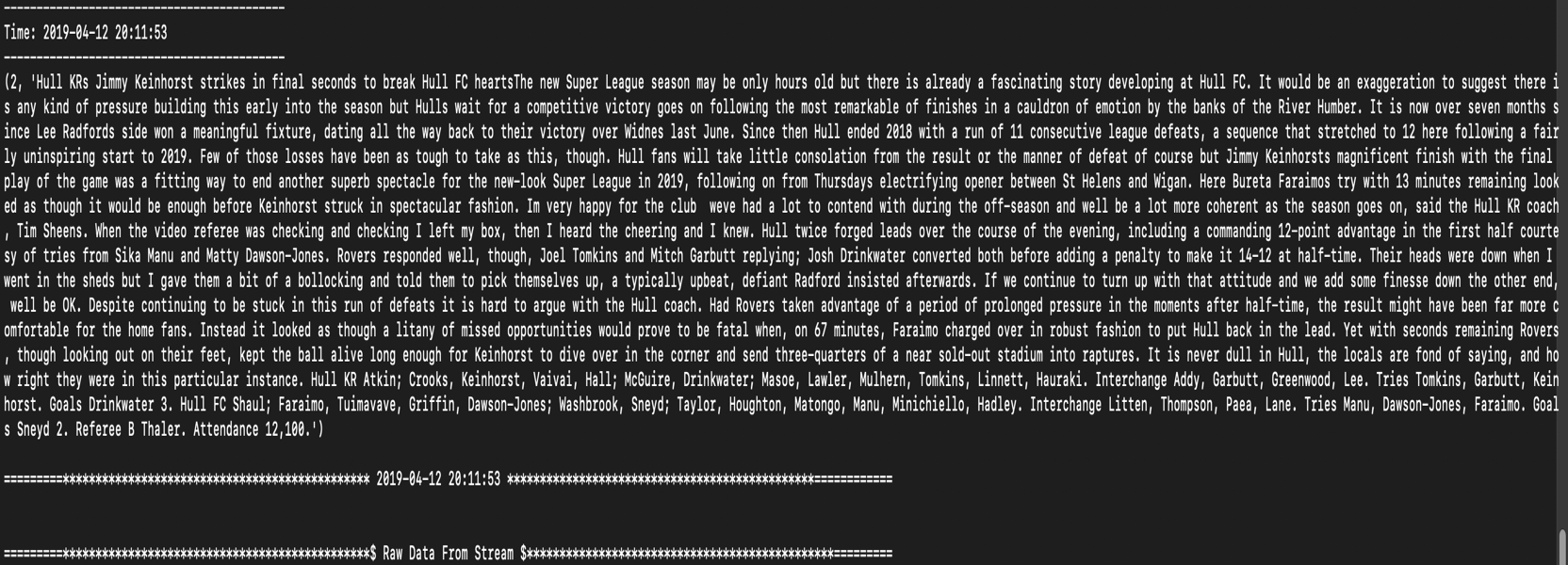
19/04/12 19:22:46 WARN TaskSetManager: Stage 112 contains a task of very large size (821 KB). The maximum recommended task size is 100 KB.  
Logistic Classification Model Successfully trained and saved in project Output directory

**Naïve Bayes Classification Output:**

+------------------------------+-----+------------------------------+-----+----------+  
| text|index| probability|label|prediction|  
+------------------------------+-----+------------------------------+-----+----------+  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
|"b""Foreign Office admits i...| 8|[1.0,8.12709876200264E-39,1...| 0.0| 0.0|  
+------------------------------+-----+------------------------------+-----+----------+  
only showing top 10 rows  
  
**Test Error for Naive Bayes :1.2560765014749453%  
Test Accuracy for Naive Bayes :98.74392349852505%  
[Stage 148:======================================> (25 + 10) / 35]Test weightedRecall for Naive Bayes :0.9811752287121748  
Test weightedPrecision for Naive Bayes :0.9945375621234787  
Test f1 score for Naive Bayes :0.9874392349852505**

19/04/12 19:24:40 WARN TaskSetManager: Stage 151 contains a task of very large size (821 KB). The maximum recommended task size is 100 KB.  
Naive Bayes Model Successfully trained and saved in project Output directory  
  
Process finished with exit code 0

**Classification and Evaluation Output on Streaming Data :**

****

**Classifying Streaming News Direct Stream: Output :**

**Conclusion:**

The news classification is done using the spark streams by creating a streaming direct stream in spark. For a batch of news articles, it classifies the news type.

**Accuracy on Streaming Test Data: 15-20% due to less training data.**

**Accuracy on Train/Test Data with 80-20 split:->**

**Logistic Classification Model: 96.3**

**Naïve Bayes Classification Model: 98.7**

At run Time Both shows overall performance and Individual Spark Stream Performance.

### References:

* <https://spark.apache.org/docs/latest/ml-decision-tree.html>
* <https://spark.apache.org/docs/2.2.0/mllib-naive-bayes.html>
* <https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35>
* <https://blog.insightdatascience.com/spark-pipelines-elegant-yet-powerful-7be93afcdd42>
* <https://scalac.io/scala-spark-ml-machine-learning-introduction/>
* <https://towardsdatascience.com/multi-class-text-classification-with-pyspark-7d78d022ed35>