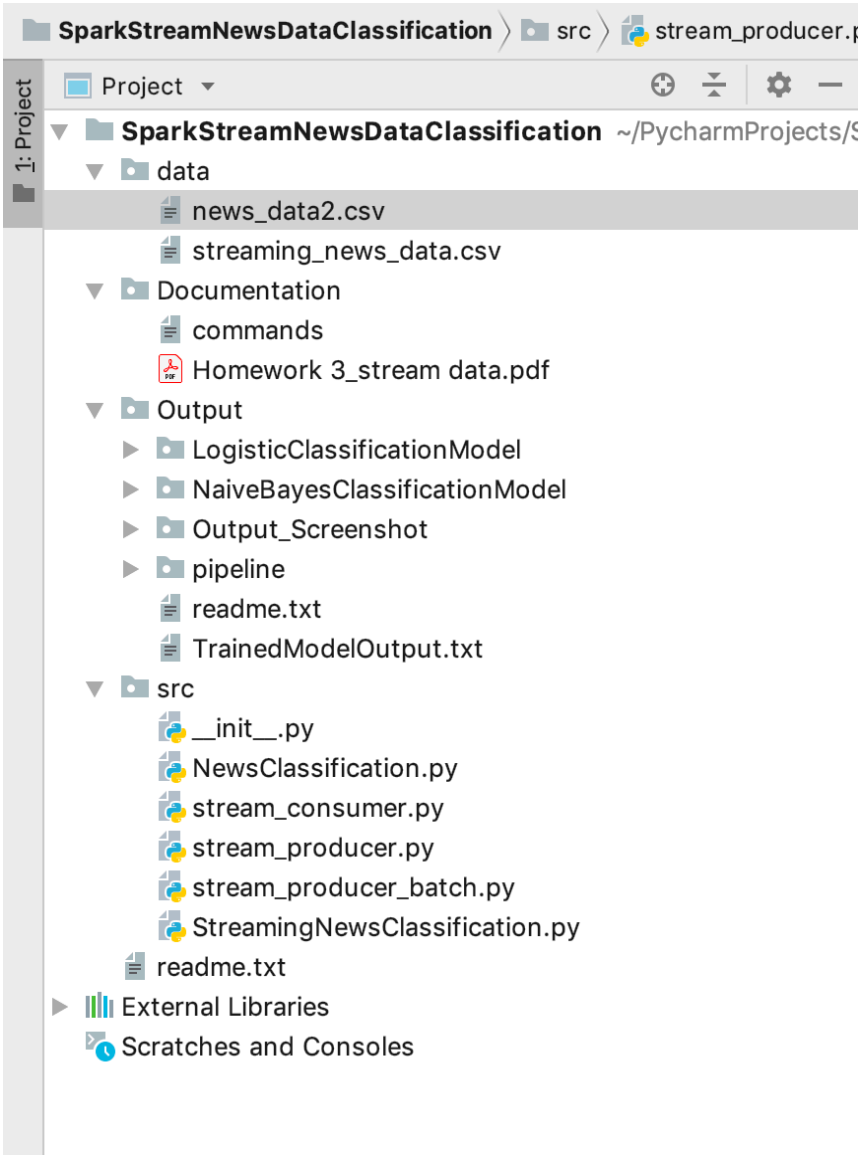
Multi News Classification
Big data Management Analytics and Management

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 :

```
-------------------------------------------
Time: 2019-04-12 20:11:53
-------------------------------------------
(2, 'Hull KRs Jimmy Keinhorst strikes in final seconds to break Hull FC heartsThe new Super League season may be only hours old but there is already a fascinating story developing at Hull FC. It would be an exaggeration to suggest there i
s any kind of pressure building this early into the season but Hulls wait for a competitive victory goes on following the most remarkable of finishes in a cauldron of emotion by the banks of the River Humber. It is now over seven months s
ince Lee Radfords side won a meaningful fixture, dating all the way back to their victory over Widnes last June. Since then Hull ended 2018 with a run of 11 consecutive league defeats, a sequence that stretched to 12 here following a fair
ly uninspiring start to 2019. Few of those losses have been as tough to take as this, though. Hull fans will take little consolation from the result or the manner of defeat of course but Jimmy Keinhorsts magnificent finish with the final
play of the game was a fitting way to end another superb spectacle for the new-look Super League in 2019, following on from Thursdays electrifying opener between St Helens and Wigan. Here Bureta Faraimos try with 13 minutes remaining look
ed as though it would be enough before Keinhorst struck in spectacular fashion. Im very happy for the club  weve had a lot to contend with during the off-season and well be a lot more coherent as the season goes on, said the Hull KR coach
, Tim Sheens. When the video referee was checking and checking I left my box, then I heard the cheering and I knew. Hull twice forged leads over the course of the evening, including a commanding 12-point advantage in the first half courte
sy of tries from Sika Manu and Matty Dawson-Jones. Rovers responded well, though, Joel Tomkins and Mitch Garbutt replying; Josh Drinkwater converted both before adding a penalty to make it 14-12 at half-time. Their heads were down when I
went in the sheds but I gave them a bit of a bollocking and told them to pick themselves up, a typically upbeat, defiant Radford insisted afterwards. If we continue to turn up with that attitude and we add some finesse down the other end,
 well be OK. Despite continuing to be stuck in this run of defeats it is hard to argue with the Hull coach. Had Rovers taken advantage of a period of prolonged pressure in the moments after half-time, the result might have been far more c
omfortable for the home fans. Instead it looked as though a litany of missed opportunities would prove to be fatal when, on 67 minutes, Faraimo charged over in robust fashion to put Hull back in the lead. Yet with seconds remaining Rovers
, though looking out on their feet, kept the ball alive long enough for Keinhorst to dive over in the corner and send three-quarters of a near sold-out stadium into raptures. It is never dull in Hull, the locals are fond of saying, and ho
w right they were in this particular instance. Hull KR Atkin; Crooks, Keinhorst, Vaivai, Hall; McGuire, Drinkwater; Masoe, Lawler, Mulhern, Tomkins, Linnett, Hauraki. Interchange Addy, Garbutt, Greenwood, Lee. Tries Tomkins, Garbutt, Kein
horst. Goals Drinkwater 3. Hull FC Shaul; Faraimo, Tuimavave, Griffin, Dawson-Jones; Washbrook, Sneyd; Taylor, Houghton, Matongo, Manu, Minichiello, Hadley. Interchange Litten, Thompson, Paea, Lane. Tries Manu, Dawson-Jones, Faraimo. Goal
s Sneyd 2. Referee B Thaler. Attendance 12,100.')

========************************************************* 2019-04-12 20:11:53 ***********************************************===========


========*********************************************************$ Raw Data From Stream $***********************************************=========
```

```
========*********************************************************** 2019-04-12 20:11:53 ***********************************************************===========


========***********************************************************$ Raw Data From Stream $***********************************************************=========
+-----+--------------------+
|label|                text|
+-----+--------------------+
|    2|Hull KRs Jimmy Ke...|
+-----+--------------------+



========***********************************************************$ Transformed Data After Running Pre Loded Pipeline $***********************************************=========
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|label|                text|               words|            filtered|         rawFeatures|            features|
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|    2|Hull KRs Jimmy Ke...|[hull, krs, jimmy...|[hull, krs, jimmy...|(3000,[2,25,36,40...|(3000,[2,25,36,40...|
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+



========***********************************************************$ Classification Using Pre Trained Logistic Classification Model $***********************************************=========
+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+
|label|prediction|Current Stream Accuracy %|Current Stream Error %|Current Stream F1 Score|Current Stream weightedRecall|Current Stream weightedPrecision|News_Category_Predicted|News_Category_InitalLabel|
+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+
|    2|       4.0|                     0.0%|               100.0%|                    0.0|                           0.0|                              0.0|                  Sport|                 Football|
+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+



========***********************************************************$ Overall Classification Metrics Logistic Classification Model $***********************************************=========
+---------------------+-----------+----------------------------+-------------------------+
|Overall Correct Count|Total Count|Overall Accuracy Percent(%)|Overall Error Percent(%)|
+---------------------+-----------+----------------------------+-------------------------+
|                    1|          8|                       12.5%|                    87.5%|
+---------------------+-----------+----------------------------+-------------------------+



========***********************************************************$ Classification Using Pre Trained Naive Bayes Classification Model $***********************************************=========


+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+
|label|prediction|Current Stream Accuracy %|Current Stream Error %|Current Stream F1 Score|Current Stream weightedRecall|Current Stream weightedPrecision|News_Category_Predicted|News_Category_InitalLabel|
+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+
|    2|      31.0|                     0.0%|               100.0%|                    0.0|                           0.0|                              0.0|             Crosswords|                 Football|
+-----+----------+-------------------------+---------------------+-----------------------+------------------------------+---------------------------------+----------------------+-------------------------+



========***********************************************************$ Overall Classification Metrics Naive Bayes Classification Model $***********************************************=========
+---------------------+-----------+----------------------------+-------------------------+
|Overall Correct Count|Total Count|Overall Accuracy Percent(%)|Overall Error Percent(%)|
+---------------------+-----------+----------------------------+-------------------------+
|                    0|          8|                        0.0%|                   100.0%|
+---------------------+-----------+----------------------------+-------------------------+



========*********************************************************** End of Single Stream ***********************************************=========
-------------------------------------------
```

Classifying Streaming News Direct Stream: Output :

```
-------------------------------------------
(15, 'Ex-Barclays chief fretted over Brown-Dalai Lama meetingThe former chief executive of Barclays worried that a 2008 meeting between Gordon Brown and the Dalai Lama would anger Chinese investors and put a multibillion rescue package at
 risk, just as the bank was finalising fundraising terms with Qatar, a court heard on Friday. Barclays was scrambling to find big investors who would commit billions of pounds to shore up the lenders balance sheet and avoid a state bailou
t  a fate that befell its high street peers Northern Rock, RBS and HBOS. Prosecutors for the Serious Fraud Office presented an email to the jury at Southwark crown court that was written by former chief executive John Varley to then-chair
man Marcus Agius. He was briefing Agius on a recent meeting with Qatari investors  codenamed Quail. The email, dated 25 May 2008, came just days after Brown held a meeting with the Tibetan spiritual leader at Lambeth Palace, the archbisho
p of Canterburys London residence. There has been a lot of work with Quail over the weekend which Id like to brief you on, Varley wrote. They [Qatari investors] are playing hardball. In addition  I cant gauge whether the Chinese  will tak
e the opportunity of playing hardball too because of their anger at the Brown-Dalai Lama meeting, he wrote. That risk influences my thinking on how we should play Quail. Conversations and emails presented in court showed Barclays bosses d
iscussing the banks need to hurry up with its fundraising. The jury was played a recorded telephone conversation between Richard Boath, the banks former European financial institutions boss, and Robert Morrice, the former CEO of Barclays
Asia. Everybody knows Barclays needs money  we need to be less cute  hurry up and pay the price and get people signed up  just fucking make it happen, Boath said. We need 4bn pretty quickly  they try and be too smart about this shit, Morr
ice said according to transcripts. Theyve got us by the balls because the price is so low, he added. The SFO alleges that four former Barclays executives  Varley, Boath, Roger Jenkins and Tom Kalaris  lied to the stock market and other in
vestors about how 322m in fees were paid to Qatar in relation to emergency fundraising of more than 11bn in 2008. Prosecutors say the executives put together two advisory services agreements in order to disguise Qatars demand for larger c
ommission payments. All four men have denied the charges. Prosecutors have not accused Qatar or Morrice of wrongdoing. The trial, which is expected to last up to six months, continues.')

========********************************************* 2019-04-12 20:12:18 ********************************************============


========**********************************************$ Raw Data From Stream $*******************************************=========
+-----+--------------------+
|label|                text|
+-----+--------------------+
|   15|Ex-Barclays chief...|
+-----+--------------------+



========**********************************************$ Transformed Data After Running Pre Loded Pipeline $*******************************************=========
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|label|                text|               words|            filtered|         rawFeatures|            features|
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+
|   15|Ex-Barclays chief...|[ex-barclays, chi...|[ex-barclays, chi...|(3000,[8,13,15,64...|(3000,[8,13,15,64...|
+-----+--------------------+--------------------+--------------------+--------------------+--------------------+



========**********************************************$ Classification Using Pre Trained Logistic Classification Model $*********************************************=========
+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+
|label|prediction|Current Stream Accuracy %|Current Stream Error %|Current Stream F1 Score|Current Stream weightedRecall|Current Stream weightedPrecision|News_Category_Predicted|News_Category_InitalLabel|
+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+
|   15|       0.0|                     0.0%|                100.0%|                    0.0|                          0.0|                             0.0|         Australia news|            Life and style|
+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+



========**********************************************$ Overall Classification Metrics Logistic Classification Model $*********************************************=========
+---------------------+-----------+----------------------------+------------------------+
|Overall Correct Count|Total Count|Overall Accuracy Percent(%)|Overall Error Percent(%)|
+---------------------+-----------+----------------------------+------------------------+
|                    2|         33|            6.0606060606061%|         93.9393939394%|
+---------------------+-----------+----------------------------+------------------------+



========**********************************************$ Classification Using Pre Trained Naive Bayes Classification Model $*******************************************=========

+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+
|label|prediction|Current Stream Accuracy %|Current Stream Error %|Current Stream F1 Score|Current Stream weightedRecall|Current Stream weightedPrecision|News_Category_Predicted|News_Category_InitalLabel|
+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+
|   15|      31.0|                     0.0%|                100.0%|                    0.0|                          0.0|                             0.0|             Crosswords|            Life and style|
+-----+----------+-------------------------+----------------------+-----------------------+-----------------------------+--------------------------------+-----------------------+-------------------------+



========**********************************************$ Overall Classification Metrics Naive Bayes Classification Model $*******************************************=========
+---------------------+-----------+----------------------------+------------------------+
|Overall Correct Count|Total Count|Overall Accuracy Percent(%)|Overall Error Percent(%)|
+---------------------+-----------+----------------------------+------------------------+
|                    0|         33|                        0.0%|                  100.0%|
+---------------------+-----------+----------------------------+------------------------+
```

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