Big Data Solutions and Architecture

PROG8450-23F-Sec2-Big Data Integration and Storage

Section – 1

ETL Pipeline with Spark Streaming

JEMIMA (8890669)

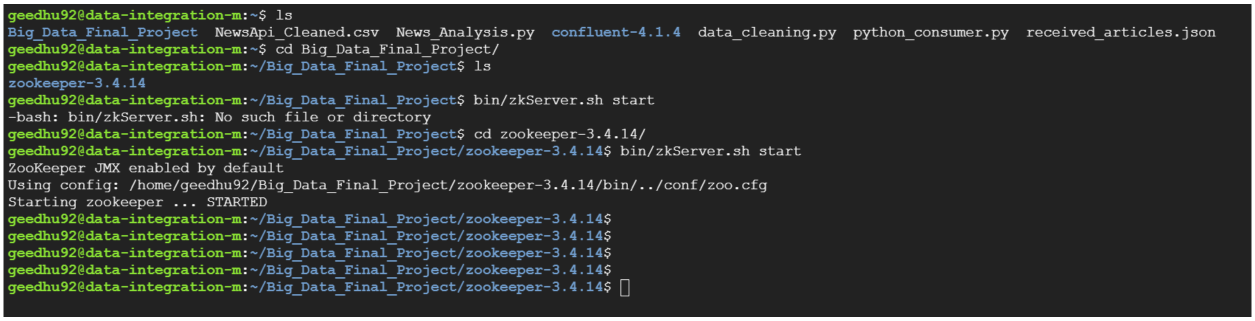
DELAXSHANA (8932185)

GEEDHU (8899510)

Streaming Section

Step 1: Start Zookeeper from the zookeeper directory

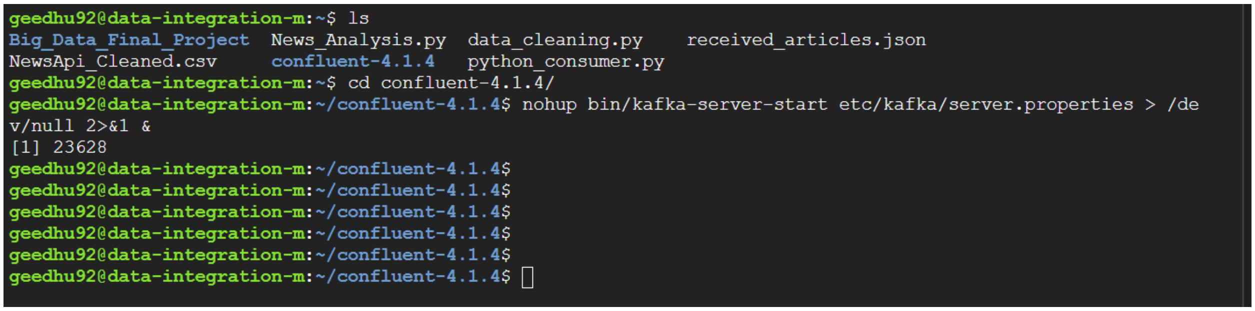
bin/zkServer.sh start



```
geedhu92@data-integration-m:~$ ls
Big_Data_Final_Project  NewsApi_Cleaned.csv  News_Analysis.py  confluent-4.1.4  data_cleaning.py  python_consumer.py  received_articles.json
geedhu92@data-integration-m:~$ cd Big_Data_Final_Project/
geedhu92@data-integration-m:~/Big_Data_Final_Project$ ls
zookeeper-3.4.14
geedhu92@data-integration-m:~/Big_Data_Final_Project$ bin/zkServer.sh start
-bash: bin/zkServer.sh: No such file or directory
geedhu92@data-integration-m:~/Big_Data_Final_Project$ cd zookeeper-3.4.14/
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$ bin/zkServer.sh start
ZooKeeper JMX enabled by default
Using config: /home/geedhu92/Big_Data_Final_Project/zookeeper-3.4.14/bin/../conf/zoo.cfg
Starting zookeeper ... STARTED
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$
geedhu92@data-integration-m:~/Big_Data_Final_Project/zookeeper-3.4.14$
```

Step 2: Start Kafka from Kafka directory that is from Confluent directory.

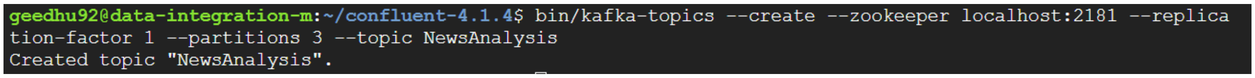
nohup bin/kafka-server-start etc/kafka/server.properties > /dev/null 2>&1 &



```
geedhu92@data-integration-m:~$ ls
Big_Data_Final_Project  News_Analysis.py  data_cleaning.py    received_articles.json
NewsApi_Cleaned.csv     confluent-4.1.4   python_consumer.py
geedhu92@data-integration-m:~$ cd confluent-4.1.4/
geedhu92@data-integration-m:~/confluent-4.1.4$ nohup bin/kafka-server-start etc/kafka/server.properties > /de
v/null 2>&1 &
[1] 23628
geedhu92@data-integration-m:~/confluent-4.1.4$
geedhu92@data-integration-m:~/confluent-4.1.4$
geedhu92@data-integration-m:~/confluent-4.1.4$
geedhu92@data-integration-m:~/confluent-4.1.4$
geedhu92@data-integration-m:~/confluent-4.1.4$
geedhu92@data-integration-m:~/confluent-4.1.4$
```

Step 3: Create the topic “NewsAnalysis” in the confluent directory.

bin/kafka-topics --create --zookeeper localhost:2181 --replication-factor 1 --partitions 3 --topic NewsAnalysis

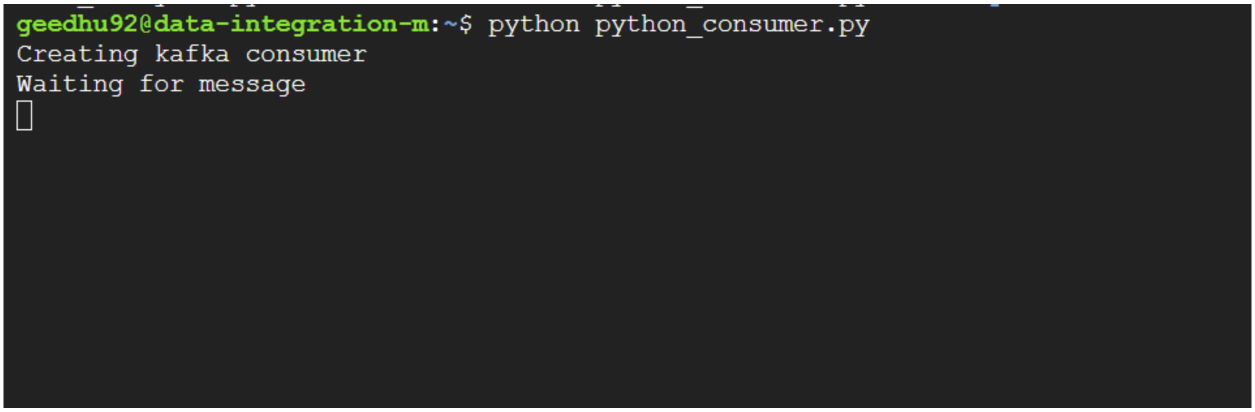


```
geedhu92@data-integration-m:~/confluent-4.1.4$ bin/kafka-topics --create --zookeeper localhost:2181 --replica
tion-factor 1 --partitions 3 --topic NewsAnalysis
Created topic "NewsAnalysis".
```

Step 4: Create a consumer to listen to the topic created that is “NewsAnalysis”.

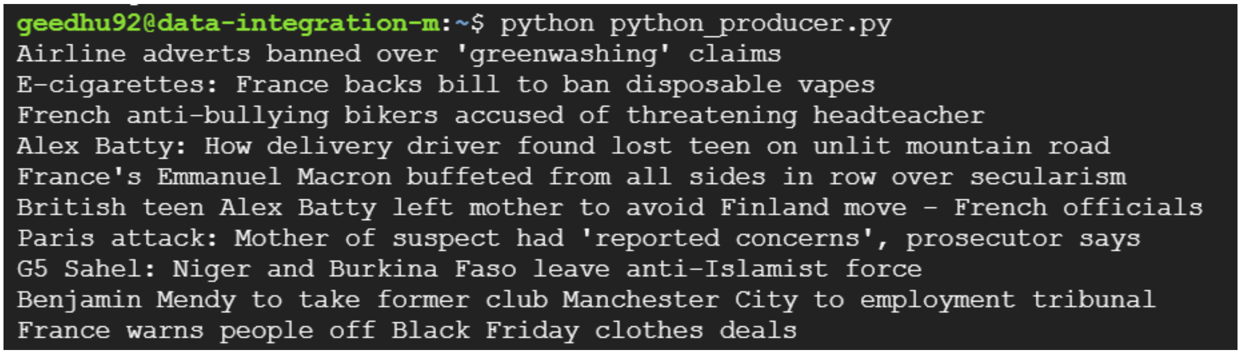
General command: bin/kafka-console-consumer --bootstrap-server localhost:9092 --topic NewsAnalysis

Here we are running a python code to collect data from the topic.



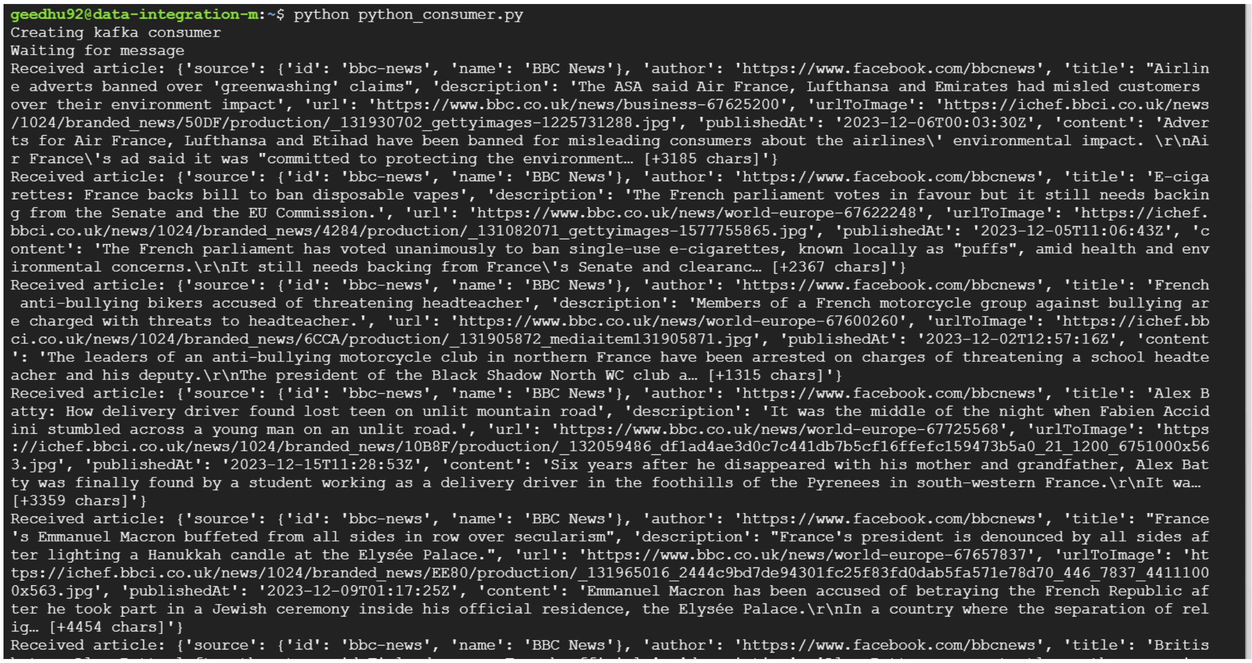
```
geedhu92@data-integration-m:~$ python python_consumer.py
Creating kafka consumer
Waiting for message
```

Step 5: Run the Producer code.



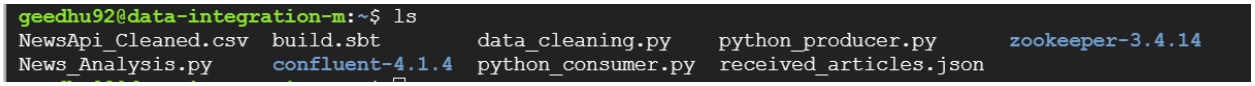
```
geedhu92@data-integration-m:~$ python python_producer.py
Airline adverts banned over 'greenwashing' claims
E-cigarettes: France backs bill to ban disposable vapes
French anti-bullying bikers accused of threatening headteacher
Alex Batty: How delivery driver found lost teen on unlit mountain road
France's Emmanuel Macron buffeted from all sides in row over secularism
British teen Alex Batty left mother to avoid Finland move - French officials
Paris attack: Mother of suspect had 'reported concerns', prosecutor says
G5 Sahel: Niger and Burkina Faso leave anti-Islamist force
Benjamin Mendy to take former club Manchester City to employment tribunal
France warns people off Black Friday clothes deals
```

Step 6: Data collected by the Consumer.



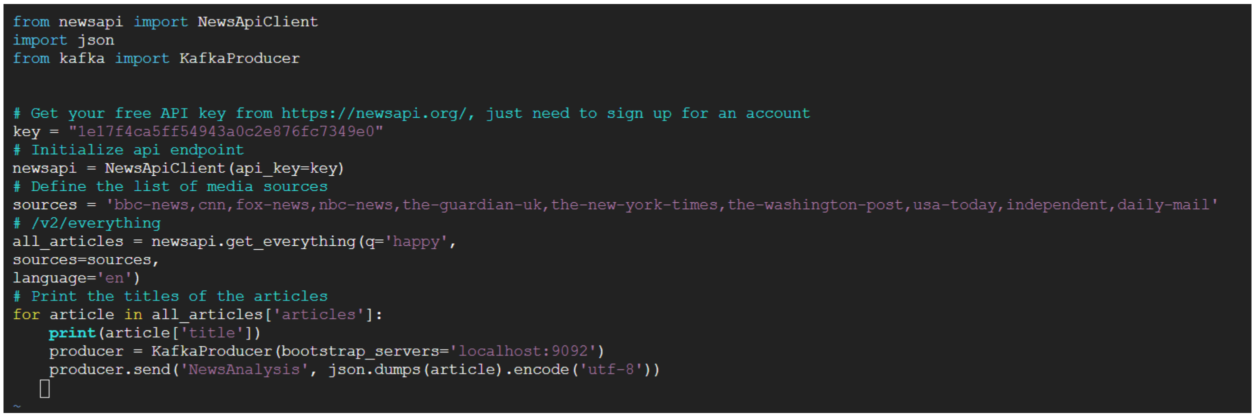
```
geedhu92@data-integration-m:~$ python python_consumer.py
Creating kafka consumer
Waiting for message
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': "Airlin
e adverts banned over 'greenwashing' claims", 'description': 'The ASA said Air France, Lufthansa and Emirates had misled customers
over their environment impact', 'url': 'https://www.bbc.co.uk/news/business-67625200', 'urlToImage': 'https://ichef.bbci.co.uk/news
/1024/branded_news/50DF/production/_131930702_gettyimages-1225731288.jpg', 'publishedAt': '2023-12-06T00:03:30Z', 'content': 'Adver
ts for Air France, Lufthansa and Etihad have been banned for misleading consumers about the airlines\' environmental impact. \r\nAi
r France\'s ad said it was "committed to protecting the environment… [+3185 chars]'}
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': 'E-ciga
rettes: France backs bill to ban disposable vapes', 'description': 'The French parliament votes in favour but it still needs backin
g from the Senate and the EU Commission.', 'url': 'https://www.bbc.co.uk/news/world-europe-67622248', 'urlToImage': 'https://ichef.
bbci.co.uk/news/1024/branded_news/4284/production/_131082071_gettyimages-1577755865.jpg', 'publishedAt': '2023-12-05T11:06:43Z', 'c
ontent': 'The French parliament has voted unanimously to ban single-use e-cigarettes, known locally as "puffs", amid health and env
ironmental concerns.\r\nIt still needs backing from France\'s Senate and clearanc… [+2367 chars]'}
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': 'French
 anti-bullying bikers accused of threatening headteacher', 'description': 'Members of a French motorcycle group against bullying ar
e charged with threats to headteacher.', 'url': 'https://www.bbc.co.uk/news/world-europe-67600260', 'urlToImage': 'https://ichef.bb
ci.co.uk/news/1024/branded_news/6CCA/production/_131905872_mediaitem131905871.jpg', 'publishedAt': '2023-12-02T12:57:16Z', 'content
': 'The leaders of an anti-bullying motorcycle club in northern France have been arrested on charges of threatening a school headte
acher and his deputy.\r\nThe president of the Black Shadow North WC club a… [+1315 chars]'}
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': 'Alex B
atty: How delivery driver found lost teen on unlit mountain road', 'description': 'It was the middle of the night when Fabien Accid
ini stumbled across a young man on an unlit road.', 'url': 'https://www.bbc.co.uk/news/world-europe-67725568', 'urlToImage': 'https
://ichef.bbci.co.uk/news/1024/branded_news/10B8F/production/_132059486_df1ad4ae3d0c7c441db7b5cf16ffefc159473b5a0_21_1200_6751000x56
3.jpg', 'publishedAt': '2023-12-15T11:28:53Z', 'content': 'Six years after he disappeared with his mother and grandfather, Alex Bat
ty was finally found by a student working as a delivery driver in the foothills of the Pyrenees in south-western France.\r\nIt wa…
[+3359 chars]'}
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': "France
's Emmanuel Macron buffeted from all sides in row over secularism", 'description': "France's president is denounced by all sides af
ter lighting a Hanukkah candle at the Elysée Palace.", 'url': 'https://www.bbc.co.uk/news/world-europe-67657837', 'urlToImage': 'ht
tps://ichef.bbci.co.uk/news/1024/branded_news/EE80/production/_131965016_2444c9bd7de94301fc25f83fd0dab5fa571e78d70_446_7837_4411100
0x563.jpg', 'publishedAt': '2023-12-09T01:17:25Z', 'content': 'Emmanuel Macron has been accused of betraying the French Republic af
ter he took part in a Jewish ceremony inside his official residence, the Elysée Palace.\r\nIn a country where the separation of rel
ig… [+4454 chars]'}
Received article: {'source': {'id': 'bbc-news', 'name': 'BBC News'}, 'author': 'https://www.facebook.com/bbcnews', 'title': 'Britis
```

Step 7: Data collected by the consumer is stored in the local directory.



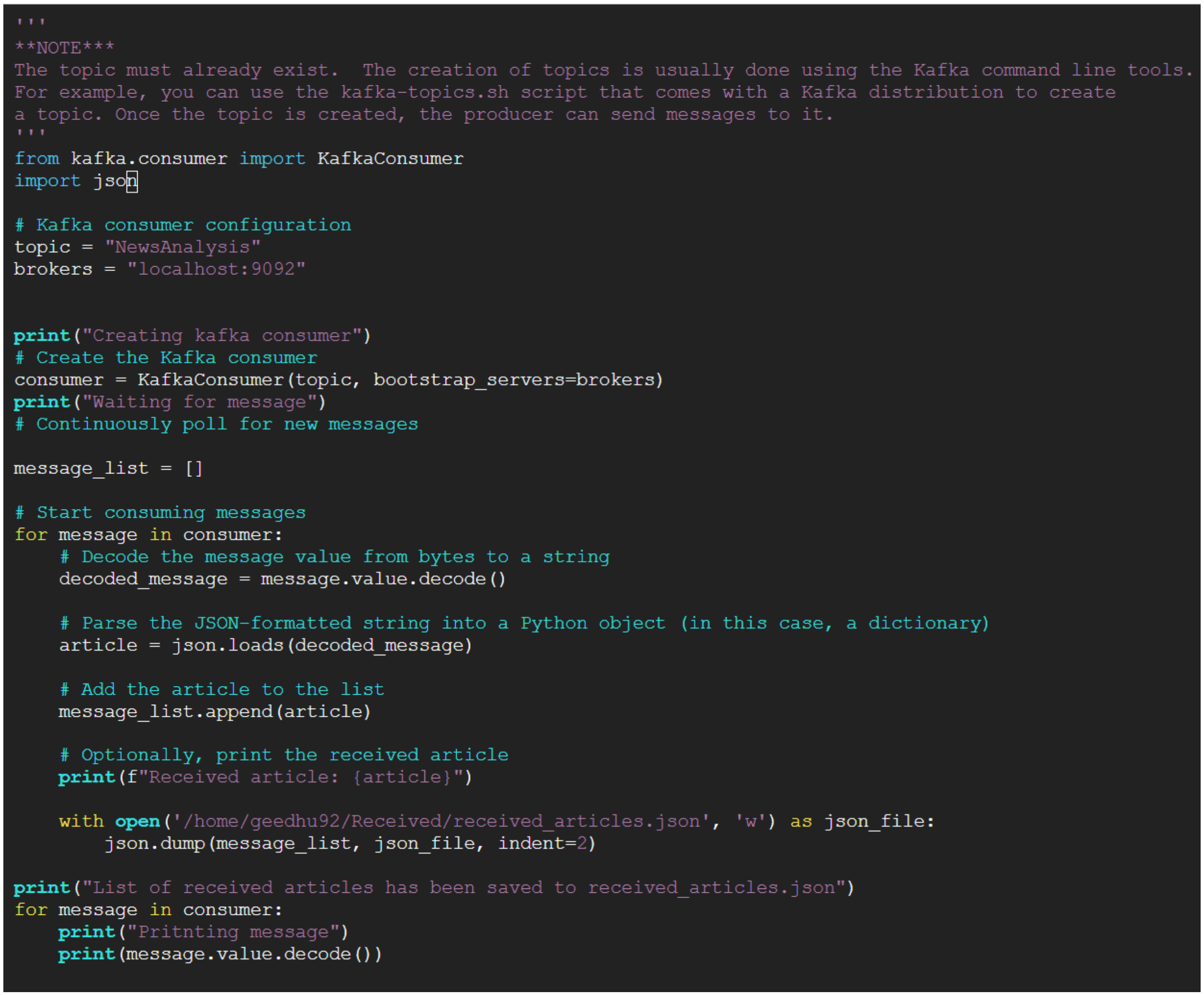
```
geedhu92@data-integration-m:~$ ls
NewsApi_Cleaned.csv  build.sbt        data_cleaning.py    python_producer.py      zookeeper-3.4.14
News_Analysis.py     confluent-4.1.4  python_consumer.py  received_articles.json
```

Producer Code



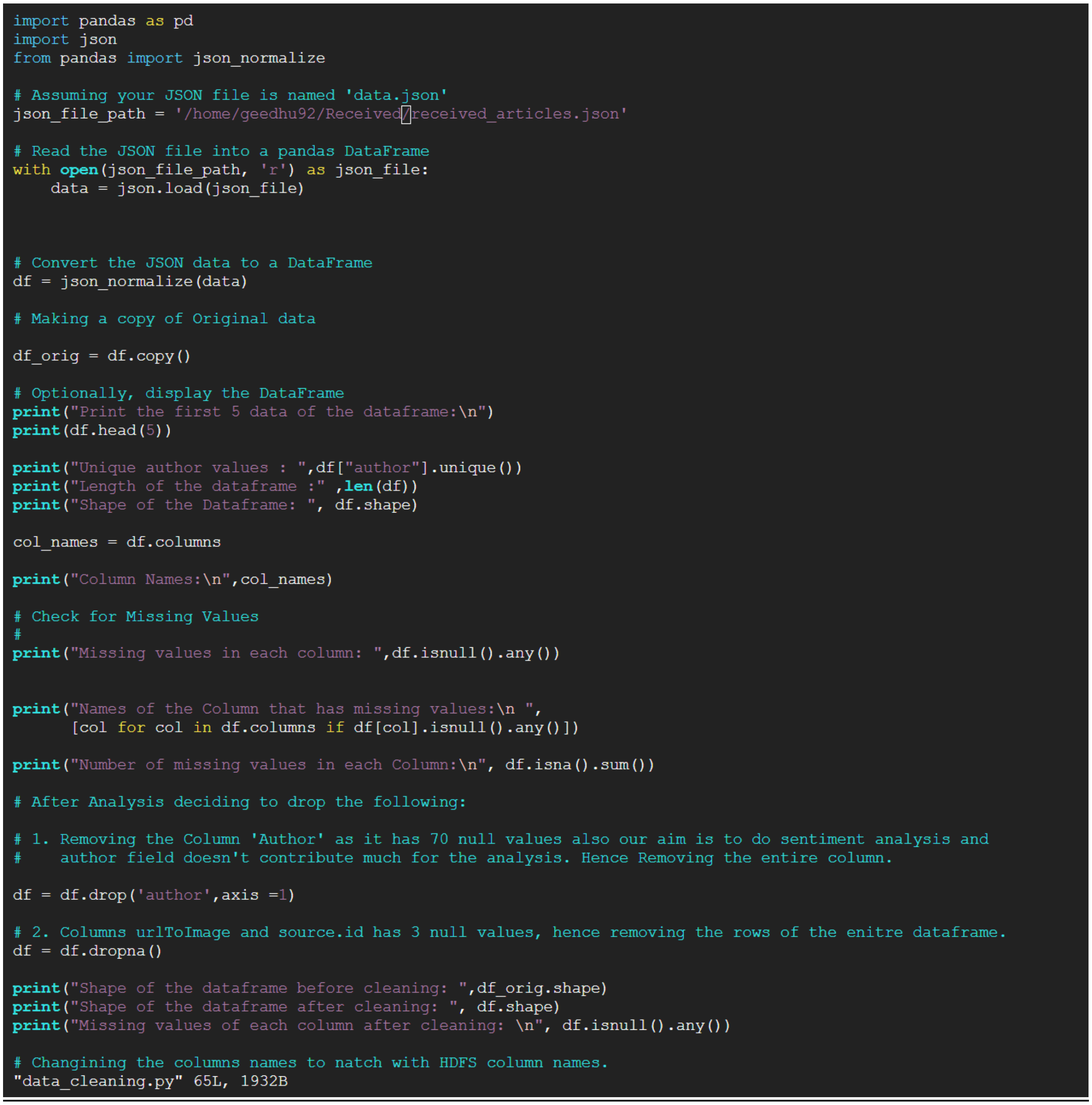
```python
from newsapi import NewsApiClient
import json
from kafka import KafkaProducer


# Get your free API key from https://newsapi.org/, just need to sign up for an account
key = "1e17f4ca5ff54943a0c2e876fc7349e0"
# Initialize api endpoint
newsapi = NewsApiClient(api_key=key)
# Define the list of media sources
sources = 'bbc-news,cnn,fox-news,nbc-news,the-guardian-uk,the-new-york-times,the-washington-post,usa-today,independent,daily-mail'
# /v2/everything
all_articles = newsapi.get_everything(q='happy',
sources=sources,
language='en')
# Print the titles of the articles
for article in all_articles['articles']:
    print(article['title'])
    producer = KafkaProducer(bootstrap_servers='localhost:9092')
    producer.send('NewsAnalysis', json.dumps(article).encode('utf-8'))

~
```

Consumer Code

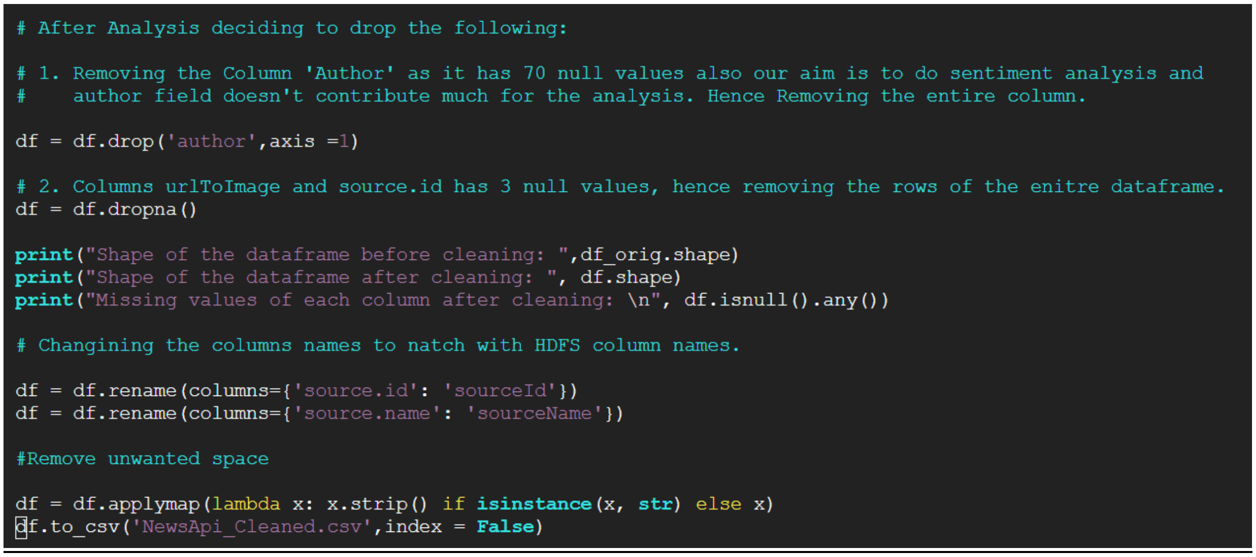


```
'''
**NOTE***
The topic must already exist.  The creation of topics is usually done using the Kafka command line tools.
For example, you can use the kafka-topics.sh script that comes with a Kafka distribution to create
a topic. Once the topic is created, the producer can send messages to it.
'''
from kafka.consumer import KafkaConsumer
import json

# Kafka consumer configuration
topic = "NewsAnalysis"
brokers = "localhost:9092"


print("Creating kafka consumer")
# Create the Kafka consumer
consumer = KafkaConsumer(topic, bootstrap_servers=brokers)
print("Waiting for message")
# Continuously poll for new messages

message_list = []

# Start consuming messages
for message in consumer:
    # Decode the message value from bytes to a string
    decoded_message = message.value.decode()

    # Parse the JSON-formatted string into a Python object (in this case, a dictionary)
    article = json.loads(decoded_message)

    # Add the article to the list
    message_list.append(article)

    # Optionally, print the received article
    print(f"Received article: {article}")

    with open('/home/geedhu92/Received/received_articles.json', 'w') as json_file:
        json.dump(message_list, json_file, indent=2)

print("List of received articles has been saved to received_articles.json")
for message in consumer:
    print("Pritnting message")
    print(message.value.decode())
```

Data Cleaning Code



```
import pandas as pd
import json
from pandas import json_normalize

# Assuming your JSON file is named 'data.json'
json_file_path = '/home/geedhu92/Received/received_articles.json'

# Read the JSON file into a pandas DataFrame
with open(json_file_path, 'r') as json_file:
    data = json.load(json_file)



# Convert the JSON data to a DataFrame
df = json_normalize(data)

# Making a copy of Original data

df_orig = df.copy()

# Optionally, display the DataFrame
print("Print the first 5 data of the dataframe:\n")
print(df.head(5))

print("Unique author values : ",df["author"].unique())
print("Length of the dataframe :" ,len(df))
print("Shape of the Dataframe: ", df.shape)

col_names = df.columns

print("Column Names:\n",col_names)

# Check for Missing Values
#
print("Missing values in each column: ",df.isnull().any())


print("Names of the Column that has missing values:\n ",
      [col for col in df.columns if df[col].isnull().any()])

print("Number of missing values in each Column:\n", df.isna().sum())

# After Analysis deciding to drop the following:

# 1. Removing the Column 'Author' as it has 70 null values also our aim is to do sentiment analysis and
#    author field doesn't contribute much for the analysis. Hence Removing the entire column.

df = df.drop('author',axis =1)

# 2. Columns urlToImage and source.id has 3 null values, hence removing the rows of the enitre dataframe.
df = df.dropna()

print("Shape of the dataframe before cleaning: ",df_orig.shape)
print("Shape of the dataframe after cleaning: ", df.shape)
print("Missing values of each column after cleaning: \n", df.isnull().any())

# Changining the columns names to natch with HDFS column names.
"data_cleaning.py" 65L, 1932B
```

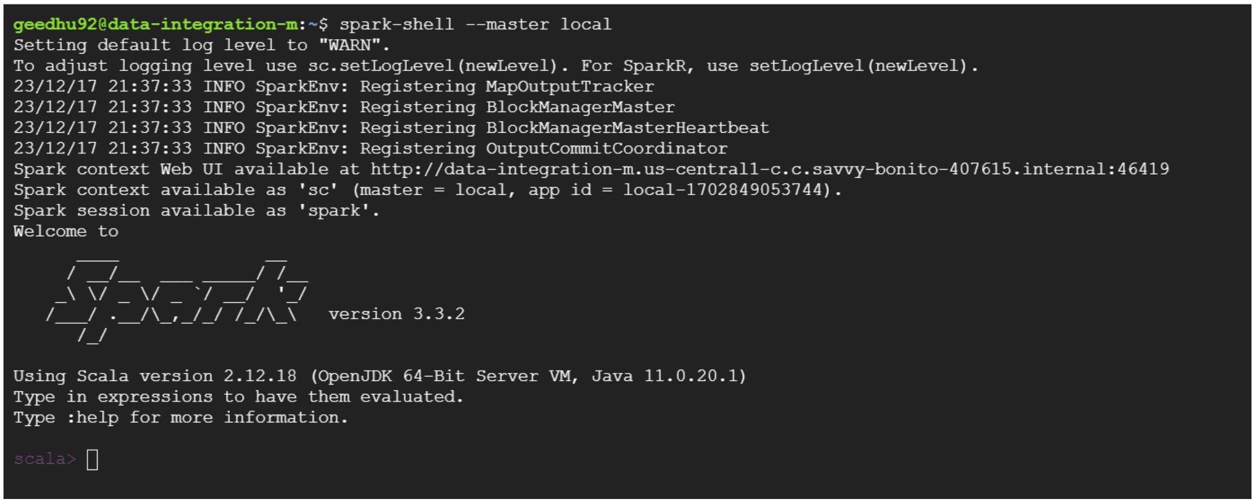


```
# After Analysis deciding to drop the following:

# 1. Removing the Column 'Author' as it has 70 null values also our aim is to do sentiment analysis and
#    author field doesn't contribute much for the analysis. Hence Removing the entire column.

df = df.drop('author',axis =1)

# 2. Columns urlToImage and source.id has 3 null values, hence removing the rows of the enitre dataframe.
df = df.dropna()

print("Shape of the dataframe before cleaning: ",df_orig.shape)
print("Shape of the dataframe after cleaning: ", df.shape)
print("Missing values of each column after cleaning: \n", df.isnull().any())

# Changinging the columns names to natch with HDFS column names.

df = df.rename(columns={'source.id': 'sourceId'})
df = df.rename(columns={'source.name': 'sourceName'})

#Remove unwanted space

df = df.applymap(lambda x: x.strip() if isinstance(x, str) else x)
df.to_csv('NewsApi_Cleaned.csv',index = False)
```

Analysis

Start SPARK.

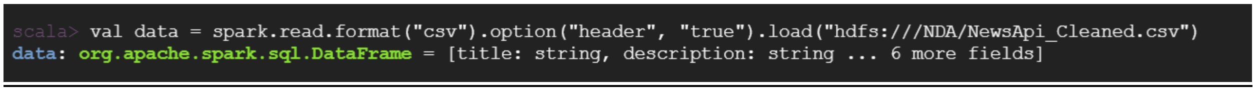
spark-shell --master local



```
geedhu92@data-integration-m:~$ spark-shell --master local
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/12/17 21:37:33 INFO SparkEnv: Registering MapOutputTracker
23/12/17 21:37:33 INFO SparkEnv: Registering BlockManagerMaster
23/12/17 21:37:33 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
23/12/17 21:37:33 INFO SparkEnv: Registering OutputCommitCoordinator
Spark context Web UI available at http://data-integration-m.us-central1-c.c.savvy-bonito-407615.internal:46419
Spark context available as 'sc' (master = local, app id = local-1702849053744).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.3.2
      /_/

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.20.1)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

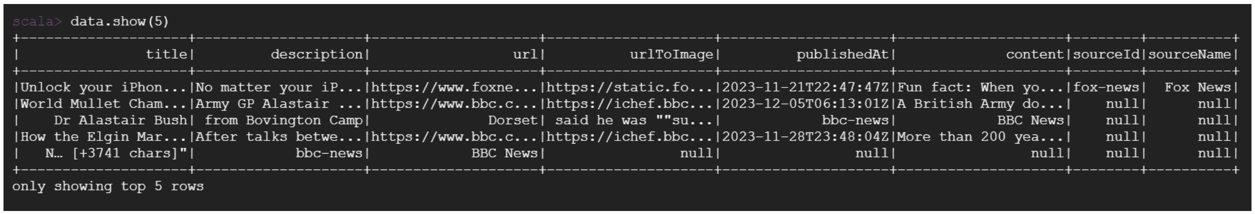
Loading the data in HDFS to Scala

val data = spark.read.format("csv").option("header", "true").load("hdfs:///NDA/NewsApi_Cleaned.csv")



```
scala> val data = spark.read.format("csv").option("header", "true").load("hdfs:///NDA/NewsApi_Cleaned.csv")
data: org.apache.spark.sql.DataFrame = [title: string, description: string ... 6 more fields]
```

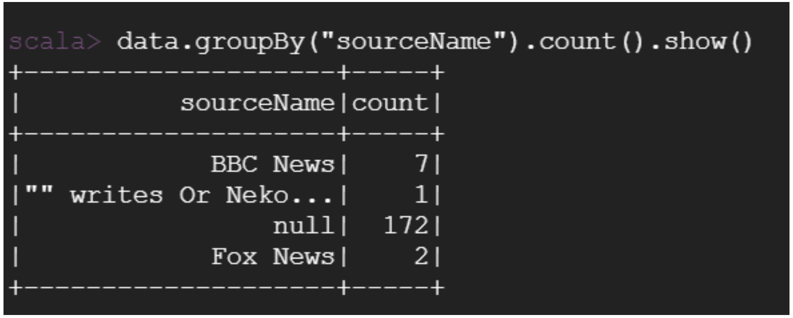
First 5 rows



```
scala> data.show(5)
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------+----------+
|               title|         description|                 url|          urlToImage|         publishedAt|             content|sourceId|sourceName|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------+----------+
|Unlock your iPhon...|No matter your iP...|https://www.foxne...|https://static.fo...|2023-11-21T22:47:47Z|Fun fact: When yo...|fox-news|  Fox News|
|World Mullet Cham...|Army GP Alastair ...|https://www.bbc.c...|https://ichef.bbc...|2023-12-05T06:13:01Z|A British Army do...|    null|      null|
|    Dr Alastair Bush| from Bovington Camp|              Dorset| said he was ""su...|            bbc-news|            BBC News|    null|      null|
|How the Elgin Mar...|After talks betwe...|https://www.bbc.c...|https://ichef.bbc...|2023-11-28T23:48:04Z|More than 200 yea...|    null|      null|
|   N… [+3741 chars]"|            bbc-news|            BBC News|                null|                null|                null|    null|      null|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------+----------+
only showing top 5 rows
```

Sources Available and their Count

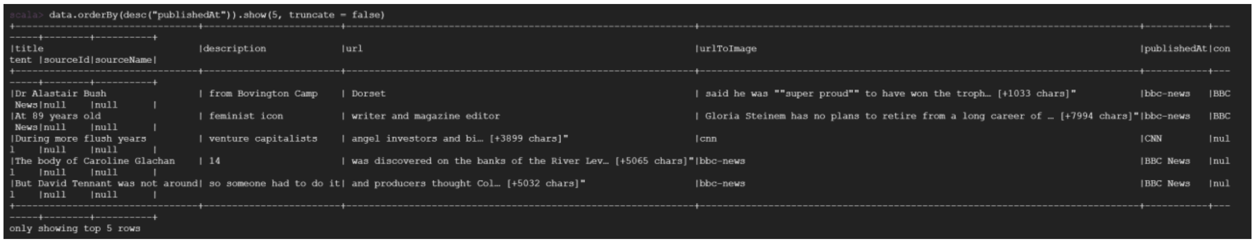
data.groupBy("sourceName").count().show()



```
scala> data.groupBy("sourceName").count().show()
+--------------------+-----+
|          sourceName|count|
+--------------------+-----+
|            BBC News|    7|
|"" writes Or Neko...|    1|
|                null|  172|
|            Fox News|    2|
+--------------------+-----+
```

Most Recent Articles:

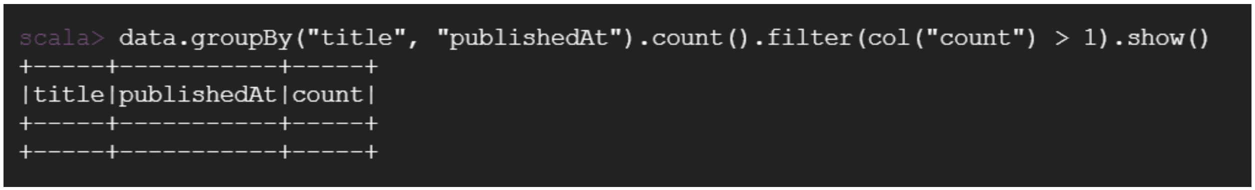
data.orderBy(desc("publishedAt")).show(5, truncate = false)



```
scala> data.orderBy(desc("publishedAt")).show(5, truncate = false)
+--------------------------------+------------------------+-------------------------------------------------------------+------------------------------------------------------------------------------+-----------+---
-----+--------+----------+
|title                           |description             |url                                                          |urlToImage                                                                    |publishedAt|con
tent |sourceId|sourceName|
+--------------------------------+------------------------+-------------------------------------------------------------+------------------------------------------------------------------------------+-----------+---
-----+--------+----------+
|Dr Alastair Bush                | from Bovington Camp    | Dorset                                                      | said he was ""super proud"" to have won the troph… [+1033 chars]"           |bbc-news   |BBC
 News|null    |null      |
|At 89 years old                 | feminist icon          | writer and magazine editor                                  | Gloria Steinem has no plans to retire from a long career of … [+7994 chars]"|bbc-news   |BBC
 News|null    |null      |
|During more flush years         | venture capitalists    | angel investors and bi… [+3899 chars]"                      |cnn                                                                           |CNN        |nul
l    |null    |null      |
|The body of Caroline Glachan    | 14                     | was discovered on the banks of the River Lev… [+5065 chars]"|bbc-news                                                                      |BBC News   |nul
l    |null    |null      |
|But David Tennant was not around| so someone had to do it| and producers thought Col… [+5032 chars]"                    |bbc-news                                                                      |BBC News   |nul
l    |null    |null      |
+--------------------------------+------------------------+-------------------------------------------------------------+------------------------------------------------------------------------------+-----------+---
-----+--------+----------+
only showing top 5 rows
```

Check for Duplicate Articles

data.groupBy("title", "publishedAt").count().filter(col("count") > 1).show()



```
scala> data.groupBy("title", "publishedAt").count().filter(col("count") > 1).show()
+-----+-----------+-----+
|title|publishedAt|count|
+-----+-----------+-----+
+-----+-----------+-----+
```

Machine Learning Algorithm

Step 1: Import necessary packages.

import org.apache.spark.ml.Pipeline

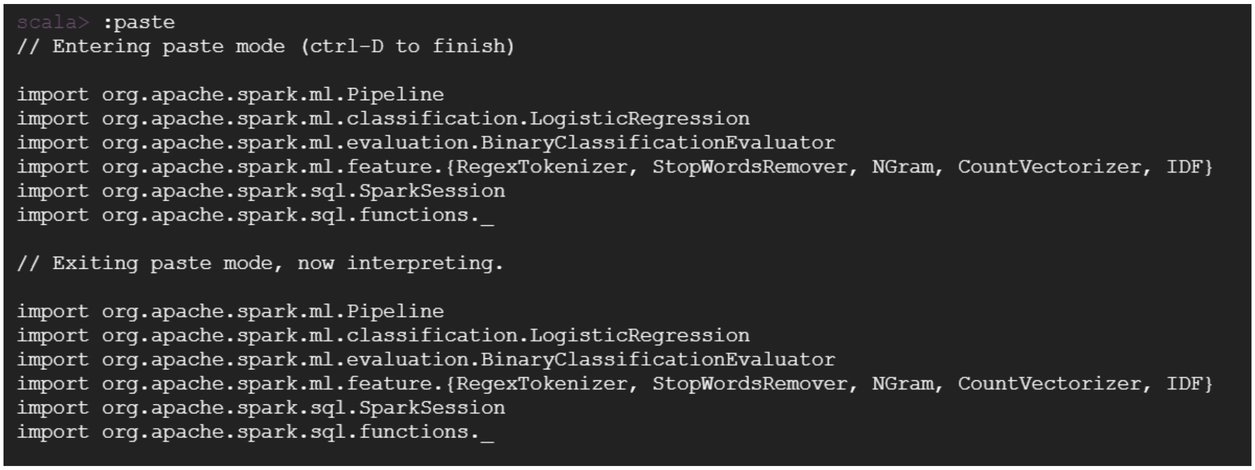
import org.apache.spark.ml.classification.LogisticRegression

import org.apache.spark.ml.evaluation.BinaryClassificationEvaluator

import org.apache.spark.ml.feature.{RegexTokenizer, StopWordsRemover, NGram, CountVectorizer, IDF}

import org.apache.spark.sql.SparkSession

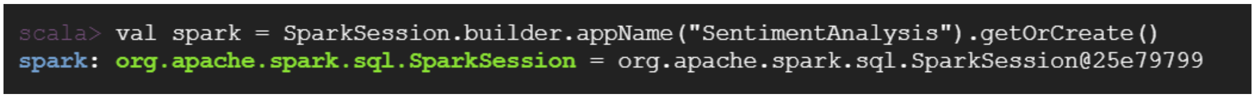
import org.apache.spark.sql.functions._



```
scala> :paste
// Entering paste mode (ctrl-D to finish)

import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.evaluation.BinaryClassificationEvaluator
import org.apache.spark.ml.feature.{RegexTokenizer, StopWordsRemover, NGram, CountVectorizer, IDF}
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

// Exiting paste mode, now interpreting.

import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.classification.LogisticRegression
import org.apache.spark.ml.evaluation.BinaryClassificationEvaluator
import org.apache.spark.ml.feature.{RegexTokenizer, StopWordsRemover, NGram, CountVectorizer, IDF}
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
```

Step 2: Creating Spark Session.

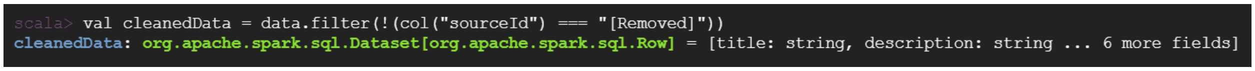
val spark = SparkSession.builder.appName("SentimentAnalysis").getOrCreate()



```
scala> val spark = SparkSession.builder.appName("SentimentAnalysis").getOrCreate()
spark: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@25e79799
```

Step 3: Removing all the articles which are “[Removed]”.

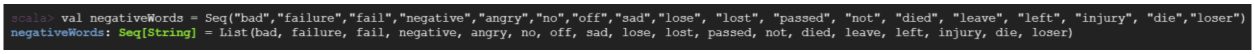
val cleanedData = data.filter(!(col("sourceId") === "[Removed]"))



```
scala> val cleanedData = data.filter(!(col("sourceId") === "[Removed]"))
cleanedData: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [title: string, description: string ... 6 more fields]
```

Step 4: Initiailizing all the negative keywords

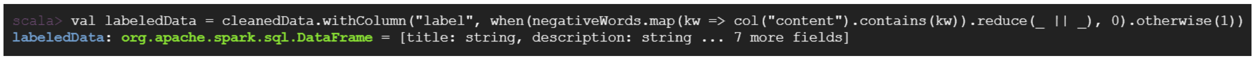
val negativeWords = Seq("bad","failure","fail","negative","angry","no","off","sad","lose", "lost", "passed", "not", "died", "leave", "left", "injury", "die","loser")



```
scala> val negativeWords = Seq("bad","failure","fail","negative","angry","no","off","sad","lose", "lost", "passed", "not", "died", "leave", "left", "injury", "die","loser")
negativeWords: Seq[String] = List(bad, failure, fail, negative, angry, no, off, sad, lose, lost, passed, not, died, leave, left, injury, die, loser)
```

Step 5: Mapping the data to whether positive or negative.

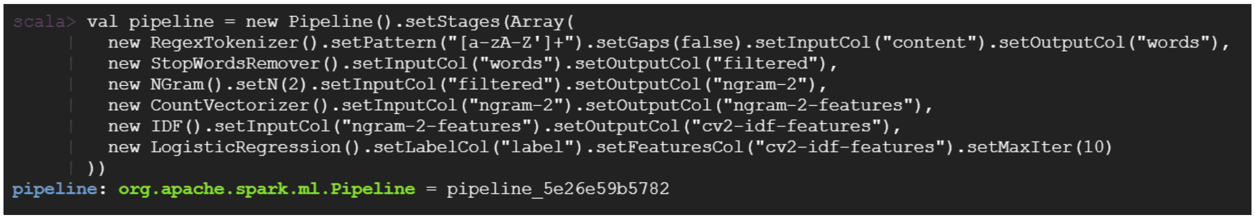
```
val labeledData = cleanedData.withColumn("label", when(negativeWords.map(kw =>
col("content").contains(kw)).reduce(_ || _), 0).otherwise(1))
```



```
scala> val labeledData = cleanedData.withColumn("label", when(negativeWords.map(kw => col("content").contains(kw)).reduce(_ || _), 0).otherwise(1))
labeledData: org.apache.spark.sql.DataFrame = [title: string, description: string ... 7 more fields]
```

Step 6: Preprocessing the data and creating data pipeline.

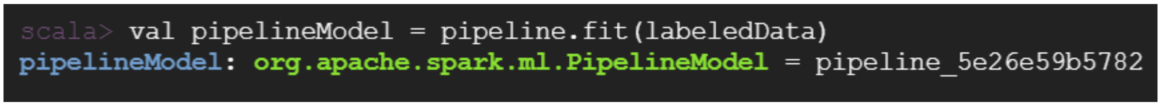
```
val pipeline = new Pipeline().setStages(Array(
  new RegexTokenizer().setPattern("[a-zA-Z']+").setGaps(false).setInputCol("content").setOutputCol("words"),
  new StopWordsRemover().setInputCol("words").setOutputCol("filtered"),
  new NGram().setN(2).setInputCol("filtered").setOutputCol("ngram-2"),
  new CountVectorizer().setInputCol("ngram-2").setOutputCol("ngram-2-features"),
  new IDF().setInputCol("ngram-2-features").setOutputCol("cv2-idf-features"),
  new LogisticRegression().setLabelCol("label").setFeaturesCol("cv2-idf-features").setMaxIter(10)
))
```



```
scala> val pipeline = new Pipeline().setStages(Array(
     |   new RegexTokenizer().setPattern("[a-zA-Z']+").setGaps(false).setInputCol("content").setOutputCol("words"),
     |   new StopWordsRemover().setInputCol("words").setOutputCol("filtered"),
     |   new NGram().setN(2).setInputCol("filtered").setOutputCol("ngram-2"),
     |   new CountVectorizer().setInputCol("ngram-2").setOutputCol("ngram-2-features"),
     |   new IDF().setInputCol("ngram-2-features").setOutputCol("cv2-idf-features"),
     |   new LogisticRegression().setLabelCol("label").setFeaturesCol("cv2-idf-features").setMaxIter(10)
     | ))
pipeline: org.apache.spark.ml.Pipeline = pipeline_5e26e59b5782
```

Step 7: Fitting the data to the Pipeline

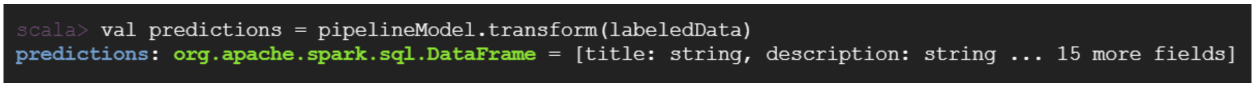
```
val pipelineModel = pipeline.fit(labeledData)
```



```
scala> val pipelineModel = pipeline.fit(labeledData)
pipelineModel: org.apache.spark.ml.PipelineModel = pipeline_5e26e59b5782
```

Step 8: Predictions

```
val predictions = pipelineModel.transform(labeledData)
```

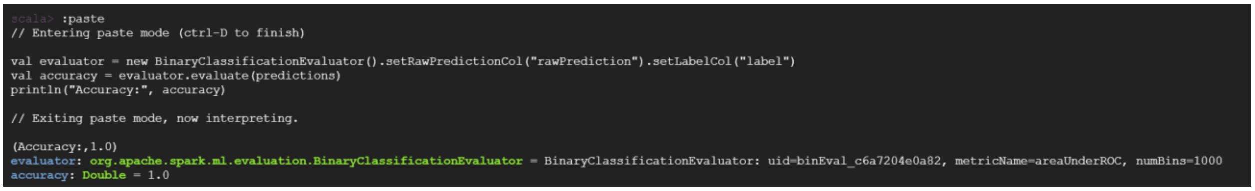


```
scala> val predictions = pipelineModel.transform(labeledData)
predictions: org.apache.spark.sql.DataFrame = [title: string, description: string ... 15 more fields]
```

Step 9: Evaluating the Accuracy of the Model

```
val evaluator = new
BinaryClassificationEvaluator().setRawPredictionCol("rawPrediction").setLabelCol("label")

val accuracy = evaluator.evaluate(predictions)
```

println("Accuracy:", accuracy)



```
scala> :paste
// Entering paste mode (ctrl-D to finish)

val evaluator = new BinaryClassificationEvaluator().setRawPredictionCol("rawPrediction").setLabelCol("label")
val accuracy = evaluator.evaluate(predictions)
println("Accuracy:", accuracy)

// Exiting paste mode, now interpreting.

(Accuracy:,1.0)
evaluator: org.apache.spark.ml.evaluation.BinaryClassificationEvaluator = BinaryClassificationEvaluator: uid=binEval_c6a7204e0a82, metricName=areaUnderROC, numBins=1000
accuracy: Double = 1.0
```