INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT AKURDI, PUNE

Documentation On

“InfoStream Data lake for Sentiment Analysis”

PG-DBDA September 2023

Submitted by-

Group No. 07

Roll No.	Name:
239517	Harsh Yadav
239549	Vishal Yerne

Mrs. Priyanka Bhor
Project Guide

Mr. Rohit Puranik
Centre Coordinator

ABSTRACT

Predicting stock market sentiment has been a topic of interest among both analysts and researchers for a long time. Stock market analyses are hard to predict because of their high volatile nature which depends on diverse political and economic factors, change of leadership, investor sentiment, and many other factors. This project is about taking non quantifiable data such as financial news articles about a company and predicting its future market trend with news sentiment analysis. Assuming that news articles have an impact on the stock market, this is an attempt to study the relationship between news and market trends.

Several sentiment analysis studies have been attempted using Natural language processing. But, till date, most of the attempts are confined to the amount or type of data being processed in the sentiment analysis. Under most of the available solutions today, scalability is a key issue. Most of the solutions either fail while processing big data, or they have limitations for the data processing time. In our project, we attempted to improve the performance and achieve the linear scalability of the news data lake by gathering live news data on a daily basis and analyzing it using john snow lab's pre-trained model using the big data's scalable toolstack

The dataset we have gathered includes daily stock news of companies for a week, along with more than 2000 financial news articles per day related to these companies. Taking in consideration of dataset and the scalability of our project, we have use big data technologies such as hadoop, spark, delta lake, airflow as an invaluable resource for prediction models and performing sentiment analysis for a given news data lake in real time.

Acknowledgement

I take this occasion to thank God, almighty for blessing us with his grace and taking our endeavor to a successful culmination. I extend my sincere and heartfelt thanks to our esteemed guide, Dr. Shantanu Pathak for providing me with the right guidance and advice at the crucial juncture sand for showing me the right way. I extend my sincere thanks to our respected **Centre Co- Ordinator Mr. Rohit Puranik**, for allowing us to use the facilities available. I would like to thank the other faculty members also, at this occasion. Last but not the least, I would like to thank my friends and family for the support and encouragement they have given me during the course of our work.

Harsh Yadav 230941225017
Vishal Yerne 230941225048

TABLE OF CONTENTS

1. INTRODUCTION AND OVERVIEW OF PROJECT

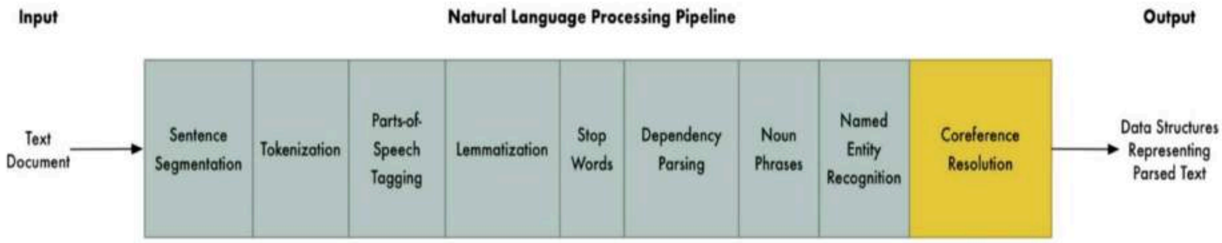
The news data lake for sentiment analysis is platform where we are trying to apply sentiment analysis using nlp on news data lake, which will further give the overall sentiment for the domain specific market, whether the sentiment is positive, negative or neutral. News has always been an important source of information to build perception of market investments.

As the volumes of news and news sources are increasing rapidly, it's becoming impossible for an investor or even a group of investors to find out relevant news from the big chunk of news available. But, it's important to make a rational choice of investing timely in order to make maximum profit out of the investment plan. And having this limitation, computation comes into the place which automatically extracts news from all possible news sources, filters and aggregates relevant ones, analyzes them in order to give them real time sentiments to know whether stock will go high or down. To discover the future trend of a stock by considering news articles about a company as prime information and tries to classify news as positive, negative or neutral. If the news sentiment is positive, there are more chances that the stock price will go up and if the news sentiment is negative, then stock price may go down. This project is to check the impact of news articles on stock prices.

SENTIMENT ANALYSIS

Sentiment analysis can be defined as analyzing the positive or negative sentiment of the customer in text. The contextual analysis of identifying information helps businesses understand their customers' social sentiment by monitoring online conversations. Sentiment analysis has become a powerful tool to monitor and understand online conversations.

NATURAL LANGUAGE PROCESSING (NLP): Natural language processing is the development of human language applications and services. With the NLP, developers can organize and structure their expertise to carry out tasks such as automated summary, translation, and recognition of entities, relationship removal, feelings analysis, speech recognition, and the segmentation of the topic. NLP used for text analyses to allow computers to understand how people communicate. This interaction of human-computers enables real-world applications like automatic resuming text, sentiment analysis, extraction of subject matter, identified entities, speaking components, extraction of relationships, stemming, etc. NLP (Natural Language Processing) also used for the mining of documents, computer translations, and automatic answering queries.



Natural language processing is a branch of artificial intelligence concerned with teaching computers to read and derive meaning from language. Since language is so complex, computers have to be taken through a series of steps before they can comprehend text. The following is a quick explanation of the steps that appear in a typical NLP pipeline.

1. Sentence Segmentation
   The text document is segmented into individual sentences.
2. Tokenization
   Once the document is broken into sentences, we further split the sentences into individual words. Each word is called a token, hence the name tokenization.
3. Parts-of-Speech-Tagging
   We input each token as well as a few words around it into a pre trained part-of-speech classification model to receive the part-of-speech for the token as an output.
4. Lemmatization
   Words often appear in different forms while referring to the same object/action. To prevent the computer from thinking of different forms of a word as different words, we perform lemmatization, the process of grouping together various inflections of a word to analyze them as a single item, identified by the word's lemma (how the word appears in the dictionary).
5. Stop Words
   Extremely common words such as "and", "the" and "a" don't provide any value, so we identify them as stop words to exclude them from any analysis performed on the text.
6. Dependency Parsing
   Assign a syntactic structure to sentences and make sense of how the words in the sentence relate to each other by feeding the words to a dependency parser.
7. Noun Phrases
   Grouping the noun phrases in a sentence together can help simplify sentences for cases when we don't care about adjectives.
8. Named Entity Recognition
   A Named Entity Recognition model can tag objects such as people's names,
   company names, and geographic locations.

9. Coreference Resolution

Since NLP models analyze individual sentences, they become confused by pronouns referring to nouns from other sentences. To solve this problem, we employ coreference resolution which tracks pronouns across sentences to avoid confusion.

Objective of the project :

1. To find overall sentiment of the market so we can have quite a good idea regarding how the market is performing.
2. Also to find out domain specific sentiment of the market, how pharma, automobile, banking, energy, finance etc are performing in their respective domains.
3. And assisting the end user with the simple insights of the market using positive, negative and neutral keywords as per the performance of the market which helps in financial investment planning.

2.DATA COLLECTION AND DATA DESCRIPTION

For the project, the data was primarily obtained from RSS feeds of various news website. RSS stands for Really Simple Syndication, and it is a simple, standardized content distribution method that helps to stay up-to-date with newscasts, blogs, and websites. We have taken several rss feeds of various news sites containing all news from different domains. All this data was in xml format. This data from there has been converted to parquet format using spark code and stored on hadoop distributed file system.

BEAUTIFUL SOUP

It is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work**.**

The quality and amount of data has a great impact on the concerned work and below is the various sources of it.

Data collection process

There are four popular ways of collecting data:

Internal Data

Companies collect some exclusive data by its actions or through collaborations with other workers which is generally the most applicable data.

Searching Online

If we require a labeled set of a huge number of videos, a webpage is there to provide that. The collection is really surprising to everyone. Available datasets permit us in prototyping prior investment in exclusive data.

API's

API's permit us for accessing programmatic datasets gathered by other businesses. Anything can be obtained from Twitter feeds of climate change to monetary data.

Web Scraping

Web scraping and crawling is an influential means that must be used sensibly. A total new world opens up which ensures to obey terms of services.

PARQUET FORMAT

Parquet is an open source file format built to handle flat columnar storage data formats. Parquet operates well with complex data in large volumes. It is known for its both performant data compression and its ability to handle a wide variety of encoding types. Parquet format have some advantages such as

1)Fast queries that can fetch specific column values without reading full row data

2)Highly efficient column-wise compression

3)High compatibility with OLAP

DATA FRAME BUILDING

There are several ways to approach data structure building. Primarily we consider the headlines of the news. There is the possibility to compile a collection of data by human effort according to specific conditions, such as gathering economic news titles filtered by a given company name from the collection built from a start time (which is the oldest possible economic news titles) until reaching a certain limit. There is the possibility of approaching the analysis using data from a previous archive collection of data, but the main goal is to use the most up-to-date data as possible. There is also the possibility of using human effort in the case of data collection from the stock market values of companies, but today many economic portals and other libraries and frameworks are available to fully automate the process. In this case, automation plays a more important role than in the previous economic news title data collecting. The error factor can be significantly reduced when compiling companies' economic data. In addition, the source and the values of the stock data are easier to manage this way than in the economic news title data collecting.

DATA DESCRIPTION

Banking: Banking and investment have been one of the major application areas of Data Science and it has allowed banks to be active in the competition. Using Data Science, banks are able to control the concerned resources effectively; additionally, users are able to make smarter decisions by analyzing sentiment.

Finance: The financial sector is a section of the economy made up of firms and institutions that provide financial services to commercial and retail customers. This sector comprises a broad range of industries including banks, investment companies, insurance companies, and real estate firms.

Auto: Automobile industry in India is huge and The contribution of the automobile sector to the overall GDP of India stands at 7.1 per cent and 49 per cent of the manufacturing GDP, with an annual turnover of Rs 7.5 lakh crores and export of Rs 3.5 lakh crores.so its an important sector to keep eye on. Therefore news of auto industry from various sources have been taken into consideration

Pharma: The Indian pharmaceutical industry is the world's 3rd largest by volume and 14th largest in terms of value. Total Annual Turnover of Pharmaceuticals was Rs. 2,89,998 crore for the year 2019-2020.This industry has also seen boom in the pandemic period and people are intended to invest in the pharma sector as well. For this purpose we have provided sentiment for pharma industry also.

Energy: The energy sector includes corporations that primarily are in the business of producing or supplying energy such as fossil fuels or renewables. The energy sector has been an important driver of industrial growth over the past century, providing fuel to power the rest of the economy. Various private players are also emerging in this field and it becomes a crucial sector to be added in our project.

3.TECHNOLOGY USED

Big Data Technologies:

1.HDFS: The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. HDFS employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters. Hadoop itself is an open source distributed processing framework that manages data processing and storage for big data applications. HDFS is a key part of the many Hadoop ecosystem technologies. It provides a reliable means for managing pools of big data and supporting related big data analytics applications.

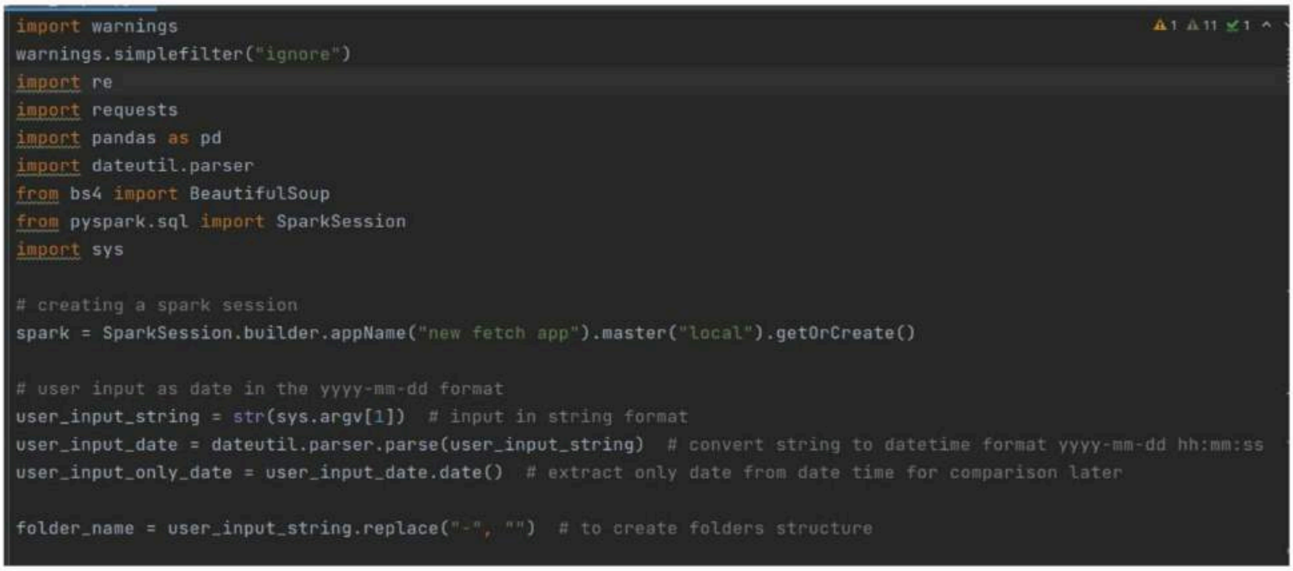
2. PySpark: Apache Spark is written in Scala programming language. To support Python with Spark, Apache Spark community released a tool, PySpark. Using PySpark, we can work with RDDs in the Python programming language also. It is because of a library called Py4j that they are able to achieve this. This is an introductory tutorial, which covers the basics of Data-Driven Documents and explains how to deal with its various components and sub-components.

3. Spark NLP: Spark NLP is an open-source natural language processing library, built on top of Apache Spark and Spark ML. It provides an easy API to integrate with ML Pipelines and it is commercially supported by John Snow Labs. Spark NLP's annotators utilize rule-based algorithms, machine learning and some of them Tensorflow running under the hood to power specific deep learning implementations. The library covers many common NLP tasks, including tokenization, stemming, lemmatization, part of speech tagging, sentiment analysis, spell checking, named entity recognition, and more. The full list of annotators, pipelines, and concepts is described in the online reference. All of them are included as open-source and can be used by training models with our data. It also provides pre-trained pipelines and models.

4. Airflow: Apache Airflow is used for the scheduling and *orchestration of data pipelines* or workflows. Orchestration of data pipelines refers to the sequencing, coordination, scheduling, and managing complex data pipelines from diverse sources. These data pipelines deliver data sets that are ready for consumption either by business intelligence applications and data science, machine learning models that support big data applications.

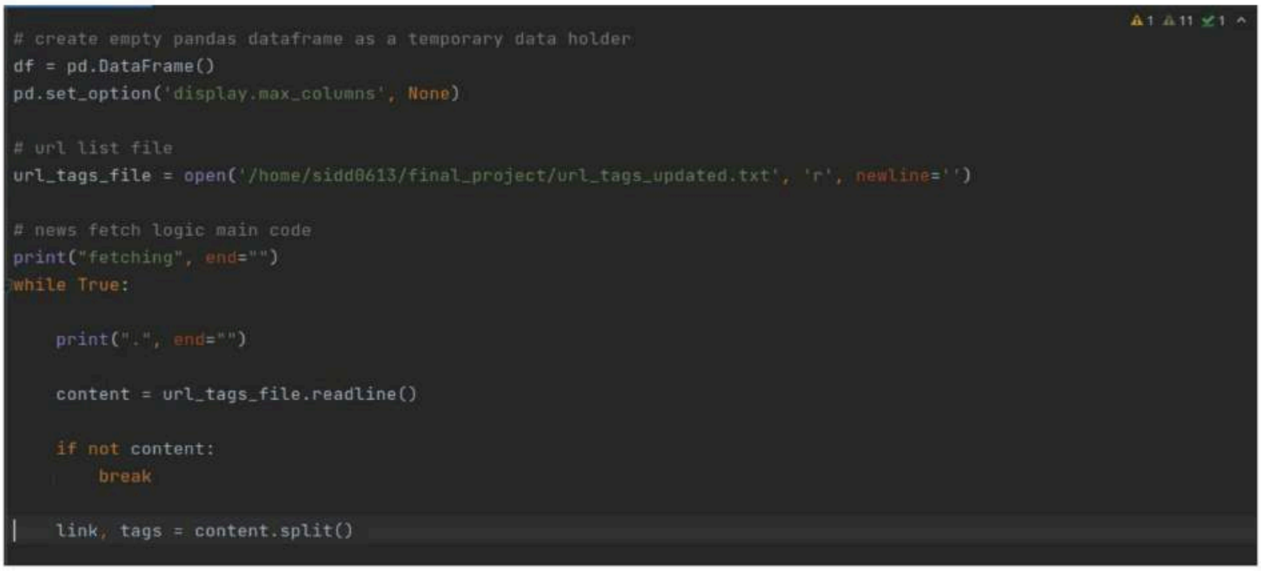
4.PROCEDURE

Importing the required modules and Creating spark session

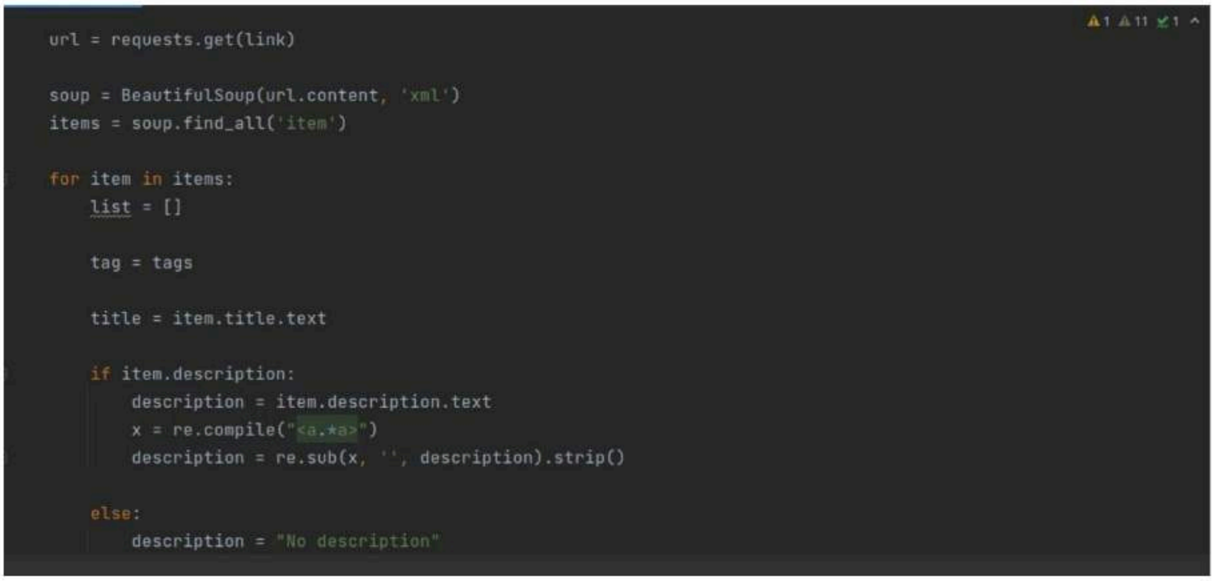


```python
import warnings
warnings.simplefilter("ignore")
import re
import requests
import pandas as pd
import dateutil.parser
from bs4 import BeautifulSoup
from pyspark.sql import SparkSession
import sys

# creating a spark session
spark = SparkSession.builder.appName("new fetch app").master("local").getOrCreate()

# user input as date in the yyyy-mm-dd format
user_input_string = str(sys.argv[1])  # input in string format
user_input_date = dateutil.parser.parse(user_input_string)  # convert string to datetime format yyyy-mm-dd hh:mm:ss
user_input_only_date = user_input_date.date()  # extract only date from date time for comparison later

folder_name = user_input_string.replace("-", "")  # to create folders structure
```

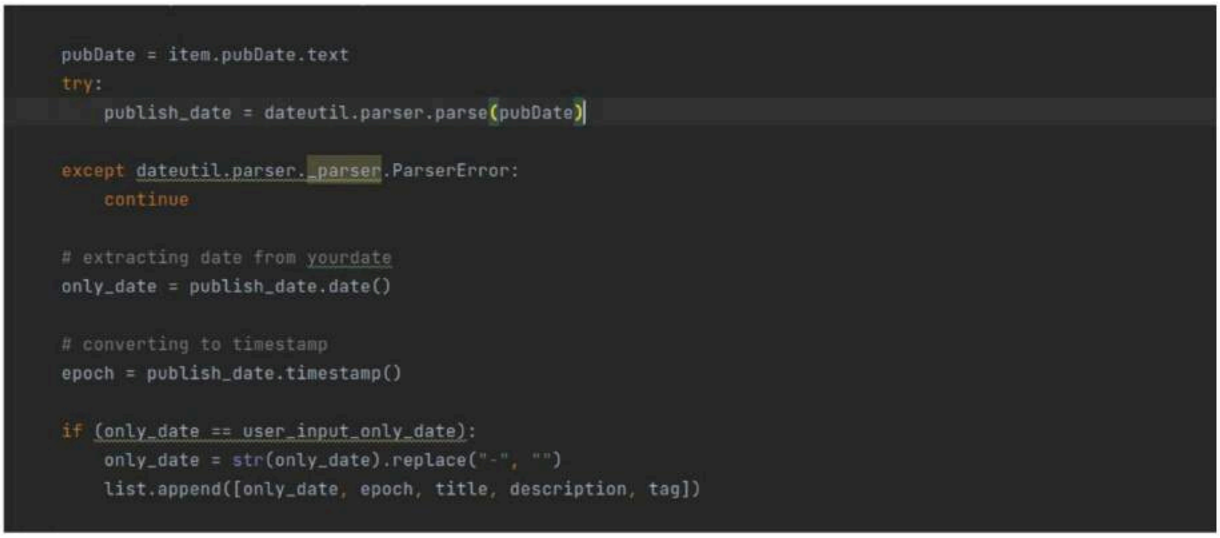
Create spark DataFrame from pandas DataFrame



```python
# create empty pandas dataframe as a temporary data holder
df = pd.DataFrame()
pd.set_option('display.max_columns', None)

# url list file
url_tags_file = open('/home/sidd0613/final_project/url_tags_updated.txt', 'r', newline='')

# news fetch logic main code
print("fetching", end="")
while True:

    print(".", end="")

    content = url_tags_file.readline()

    if not content:
        break

    link, tags = content.split()
```

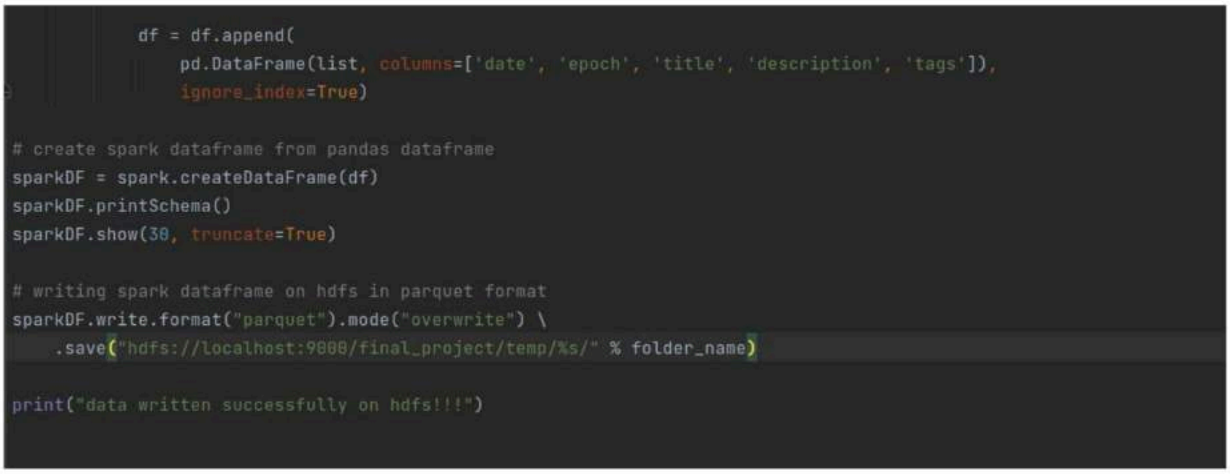


```
url = requests.get(link)

soup = BeautifulSoup(url.content, 'xml')
items = soup.find_all('item')

for item in items:
    list = []

    tag = tags

    title = item.title.text

    if item.description:
        description = item.description.text
        x = re.compile("<a.*a>")
        description = re.sub(x, '', description).strip()

    else:
        description = "No description"
```



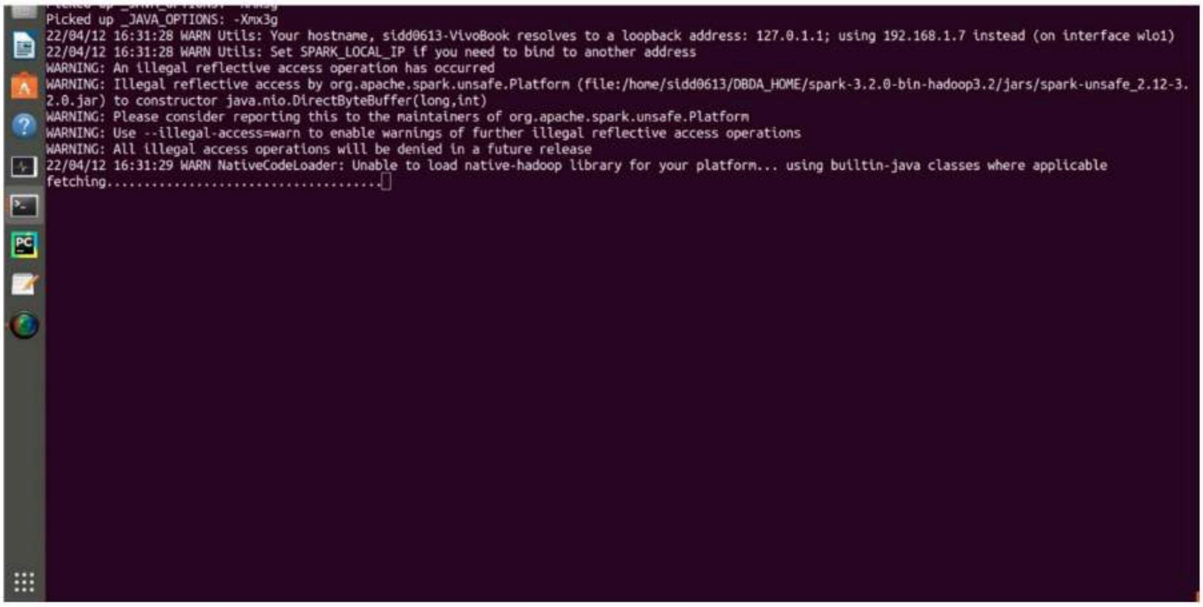
```
pubDate = item.pubDate.text
try:
    publish_date = dateutil.parser.parse(pubDate)

except dateutil.parser._parser.ParserError:
    continue

# extracting date from yourdate
only_date = publish_date.date()

# converting to timestamp
epoch = publish_date.timestamp()

if (only_date == user_input_only_date):
    only_date = str(only_date).replace("-", "")
    list.append([only_date, epoch, title, description, tag])
```

Storing data on HDFS

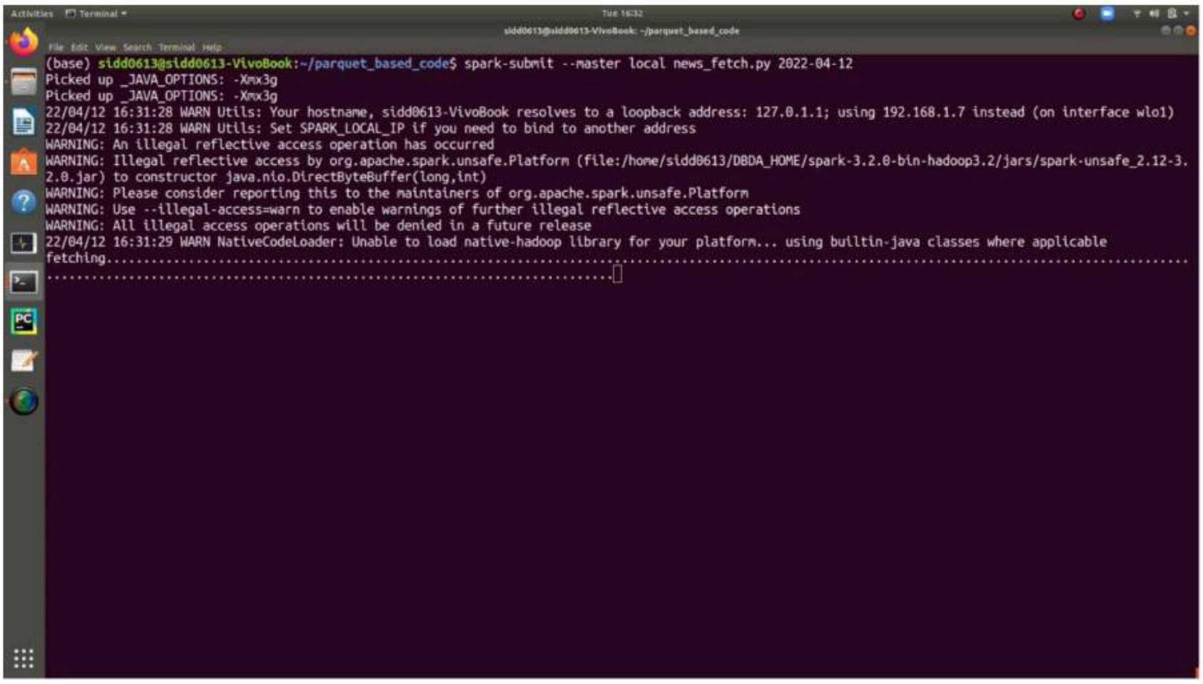


```
            df = df.append(
                pd.DataFrame(list, columns=['date', 'epoch', 'title', 'description', 'tags']),
                ignore_index=True)

# create spark dataframe from pandas dataframe
sparkDF = spark.createDataFrame(df)
sparkDF.printSchema()
sparkDF.show(30, truncate=True)

# writing spark dataframe on hdfs in parquet format
sparkDF.write.format("parquet").mode("overwrite") \
    .save("hdfs://localhost:9000/final_project/temp/%s/" % folder_name)

print("data written successfully on hdfs!!!")
```

OUTPUT:

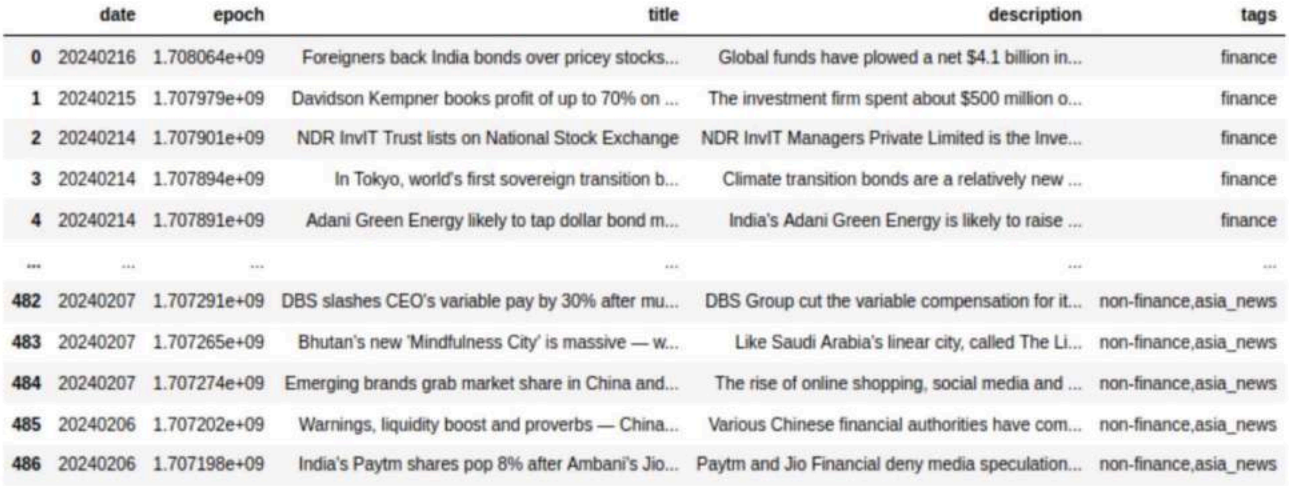


```
Picked up _JAVA_OPTIONS: -Xmx3g
22/04/12 16:31:28 WARN Utils: Your hostname, sidd0613-VivoBook resolves to a loopback address: 127.0.1.1; using 192.168.1.7 instead (on interface wlo1)
22/04/12 16:31:28 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/sidd0613/DBDA_HOME/spark-3.2.0-bin-hadoop3.2/jars/spark-unsafe_2.12-3.
2.0.jar) to constructor java.nio.DirectByteBuffer(long,int)
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
22/04/12 16:31:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
fetching....................................
```



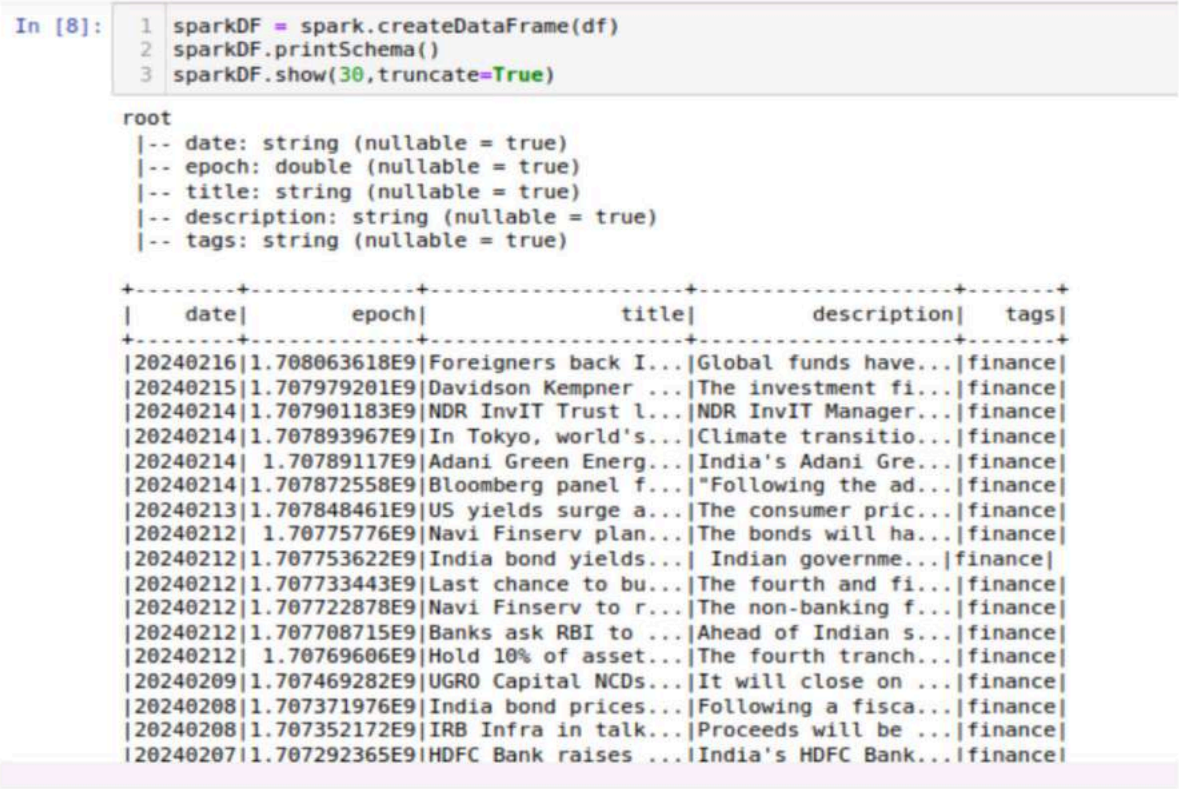
Activities
Terminal
Tue 16:32
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

```
(base) sidd0613@sidd0613-VivoBook:~/parquet_based_code$ spark-submit --master local news_fetch.py 2022-04-12
Picked up _JAVA_OPTIONS: -Xmx3g
Picked up _JAVA_OPTIONS: -Xmx3g
22/04/12 16:31:28 WARN Utils: Your hostname, sidd0613-VivoBook resolves to a loopback address: 127.0.1.1; using 192.168.1.7 instead (on interface wlo1)
22/04/12 16:31:28 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/sidd0613/DBDA_HOME/spark-3.2.0-bin-hadoop3.2/jars/spark-unsafe_2.12-3.
2.0.jar) to constructor java.nio.DirectByteBuffer(long,int)
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
22/04/12 16:31:29 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
fetching.....................................................................................................................................
........................................................................
```



	date	epoch	title	description	tags
0	20240216	1.708064e+09	Foreigners back India bonds over pricey stocks...	Global funds have plowed a net $4.1 billion in...	finance
1	20240215	1.707979e+09	Davidson Kempner books profit of up to 70% on ...	The investment firm spent about $500 million o...	finance
2	20240214	1.707901e+09	NDR InvIT Trust lists on National Stock Exchange	NDR InvIT Managers Private Limited is the Inve...	finance
3	20240214	1.707894e+09	In Tokyo, world's first sovereign transition b...	Climate transition bonds are a relatively new ...	finance
4	20240214	1.707891e+09	Adani Green Energy likely to tap dollar bond m...	India's Adani Green Energy is likely to raise ...	finance
...	...	...	...	...	...
482	20240207	1.707291e+09	DBS slashes CEO's variable pay by 30% after mu...	DBS Group cut the variable compensation for it...	non-finance,asia_news
483	20240207	1.707265e+09	Bhutan's new 'Mindfulness City' is massive — w...	Like Saudi Arabia's linear city, called The Li...	non-finance,asia_news
484	20240207	1.707274e+09	Emerging brands grab market share in China and...	The rise of online shopping, social media and ...	non-finance,asia_news
485	20240206	1.707202e+09	Warnings, liquidity boost and proverbs — China...	Various Chinese financial authorities have com...	non-finance,asia_news
486	20240206	1.707198e+09	India's Paytm shares pop 8% after Ambani's Jio...	Paytm and Jio Financial deny media speculation...	non-finance,asia_news

5. Model Building

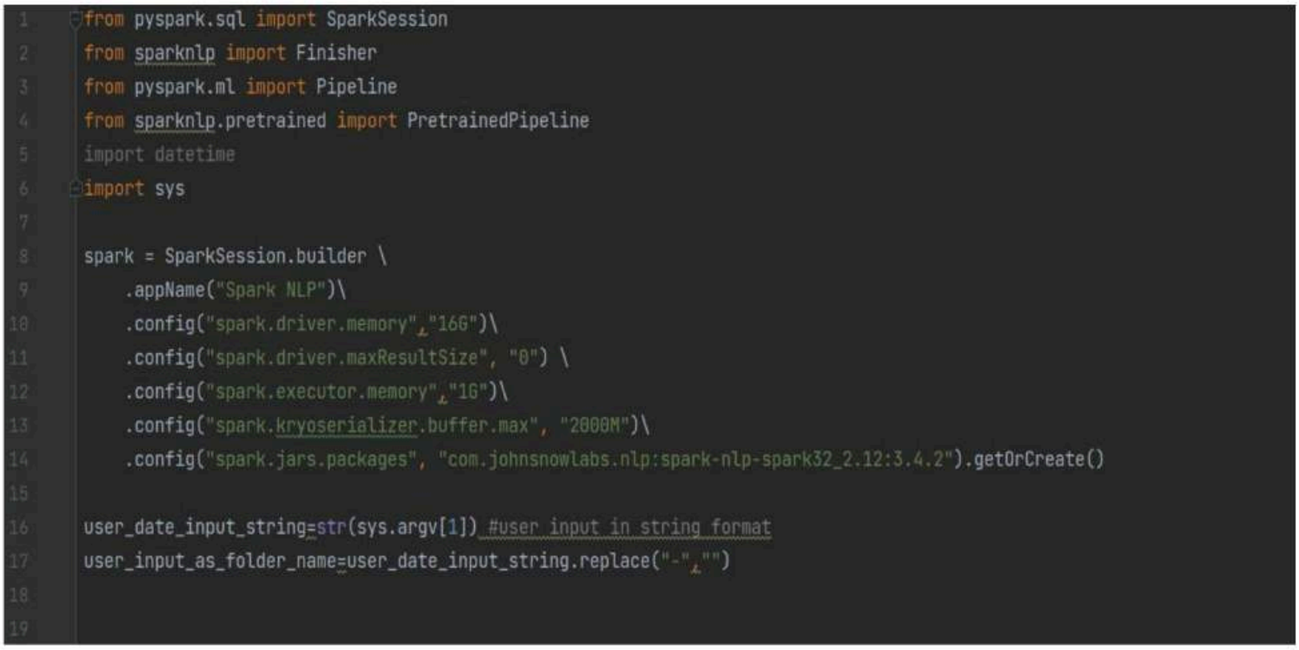


In [8]:

```
sparkDF = spark.createDataFrame(df)
sparkDF.printSchema()
sparkDF.show(30,truncate=True)
```

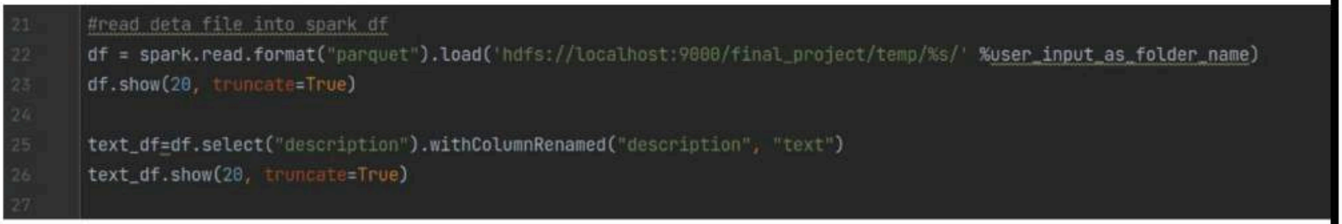
```
root
 |-- date: string (nullable = true)
 |-- epoch: double (nullable = true)
 |-- title: string (nullable = true)
 |-- description: string (nullable = true)
 |-- tags: string (nullable = true)

+--------+------------+--------------------+--------------------+-------+
|    date|       epoch|               title|         description|   tags|
+--------+------------+--------------------+--------------------+-------+
|20240216|1.708063618E9|Foreigners back I...|Global funds have...|finance|
|20240215|1.707979201E9|Davidson Kempner ...|The investment fi...|finance|
|20240214|1.707901183E9|NDR InvIT Trust l...|NDR InvIT Manager...|finance|
|20240214|1.707893967E9|In Tokyo, world's...|Climate transitio...|finance|
|20240214| 1.70789117E9|Adani Green Energ...|India's Adani Gre...|finance|
|20240214|1.707872558E9|Bloomberg panel f...|"Following the ad...|finance|
|20240213|1.707848461E9|US yields surge a...|The consumer pric...|finance|
|20240212| 1.70775776E9|Navi Finserv plan...|The bonds will ha...|finance|
|20240212|1.707753622E9|India bond yields...| Indian governme...|finance|
|20240212|1.707733443E9|Last chance to bu...|The fourth and fi...|finance|
|20240212|1.707722878E9|Navi Finserv to r...|The non-banking f...|finance|
|20240212|1.707708715E9|Banks ask RBI to ...|Ahead of Indian s...|finance|
|20240212| 1.70769606E9|Hold 10% of asset...|The fourth tranch...|finance|
|20240209|1.707469282E9|UGRO Capital NCDs...|It will close on ...|finance|
|20240208|1.707371976E9|India bond prices...|Following a fisca...|finance|
|20240208|1.707352172E9|IRB Infra in talk...|Proceeds will be ...|finance|
|20240207|1.707292365E9|HDFC Bank raises ...|India's HDFC Bank...|finance|
```

Taking user input for NLP pipeline



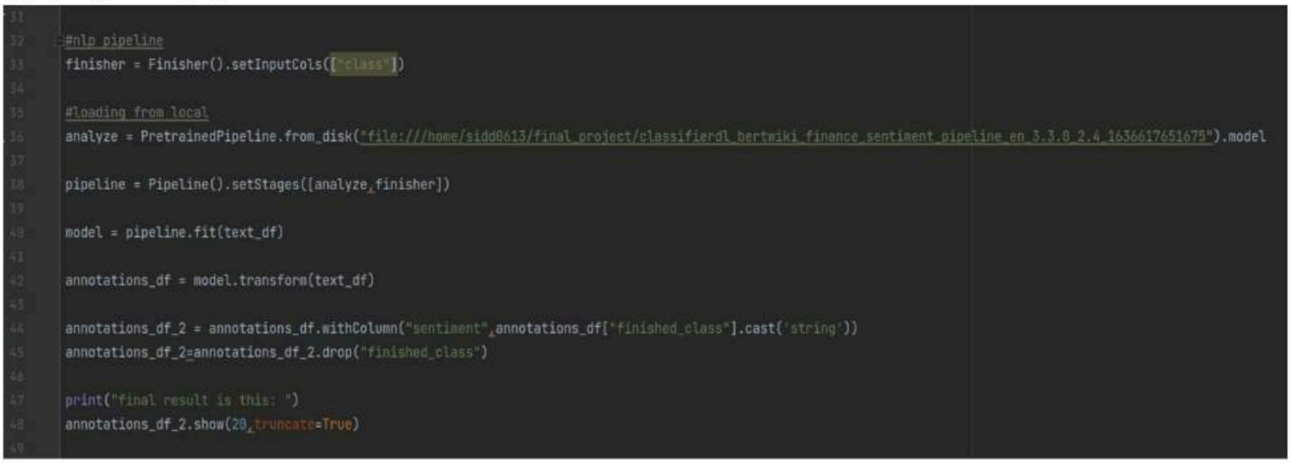
```
from pyspark.sql import SparkSession
from sparknlp import Finisher
from pyspark.ml import Pipeline
from sparknlp.pretrained import PretrainedPipeline
import datetime
import sys

spark = SparkSession.builder \
    .appName("Spark NLP")\
    .config("spark.driver.memory","16G")\
    .config("spark.driver.maxResultSize", "0") \
    .config("spark.executor.memory","1G")\
    .config("spark.kryoserializer.buffer.max", "2000M")\
    .config("spark.jars.packages", "com.johnsnowlabs.nlp:spark-nlp-spark32_2.12:3.4.2").getOrCreate()

user_date_input_string=str(sys.argv[1]) #user input in string format
user_input_as_folder_name=user_date_input_string.replace("-","")


```

Read data file into spark df



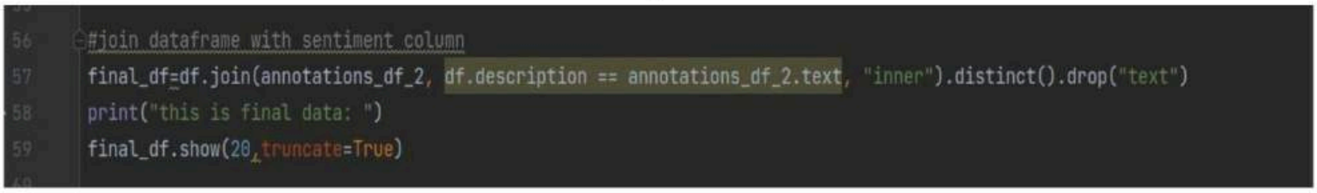
```
#read deta file into spark df
df = spark.read.format("parquet").load('hdfs://localhost:9000/final_project/temp/%s/' %user_input_as_folder_name)
df.show(20, truncate=True)

text_df=df.select("description").withColumnRenamed("description", "text")
text_df.show(20, truncate=True)
```

Applying NLP pipeline

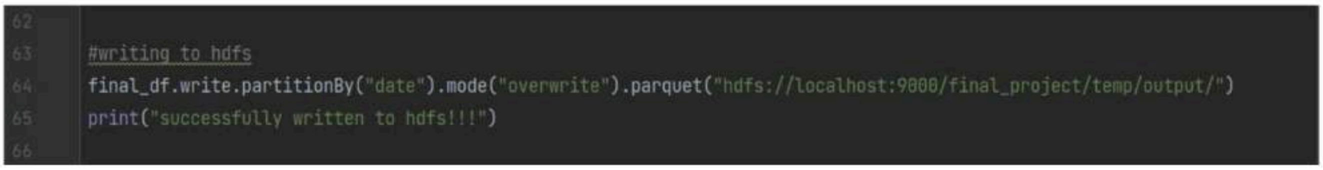


```

#nlp pipeline
finisher = Finisher().setInputCols(["class"])

#loading from local
analyze = PretrainedPipeline.from_disk("file:///home/sidd0613/final_project/classifierdl_bertwiki_finance_sentiment_pipeline_en_3.3.0_2.4_1636617651675").model

pipeline = Pipeline().setStages([analyze,finisher])

model = pipeline.fit(text_df)

annotations_df = model.transform(text_df)

annotations_df_2 = annotations_df.withColumn("sentiment",annotations_df["finished_class"].cast('string'))
annotations_df_2=annotations_df_2.drop("finished_class")

print("final result is this: ")
annotations_df_2.show(20,truncate=True)
```

Join dataframe with sentiment column and writing back to hdfs

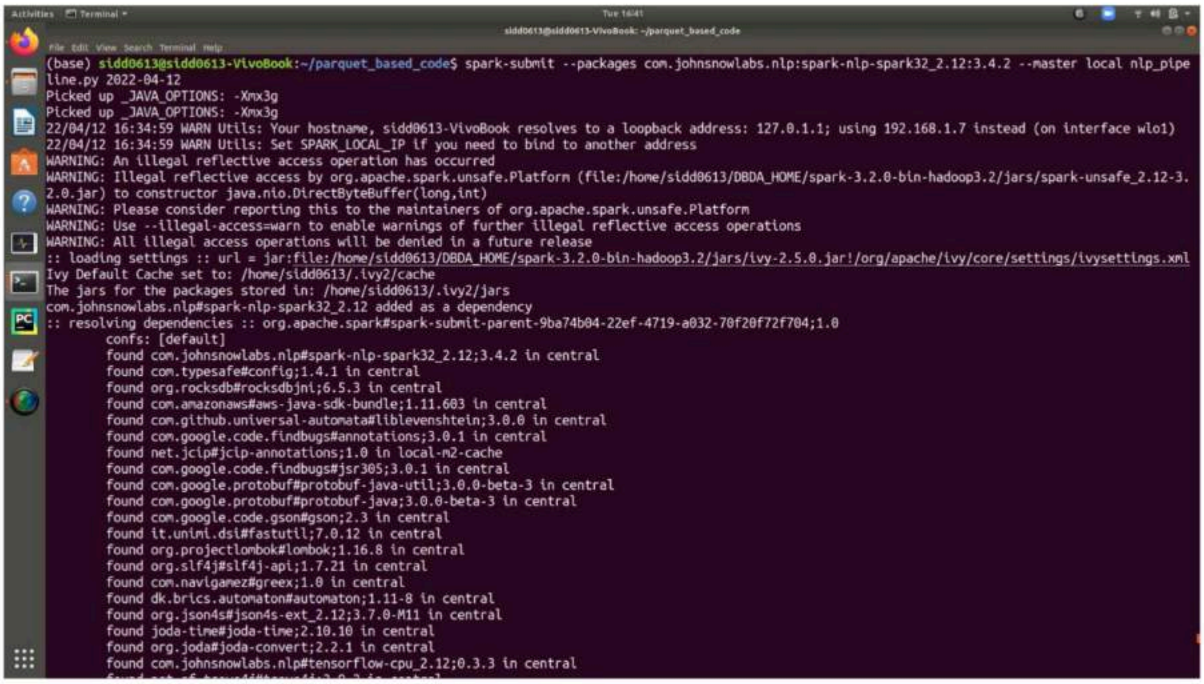


```
#join dataframe with sentiment column
final_df=df.join(annotations_df_2, df.description == annotations_df_2.text, "inner").distinct().drop("text")
print("this is final data: ")
final_df.show(20,truncate=True)
```



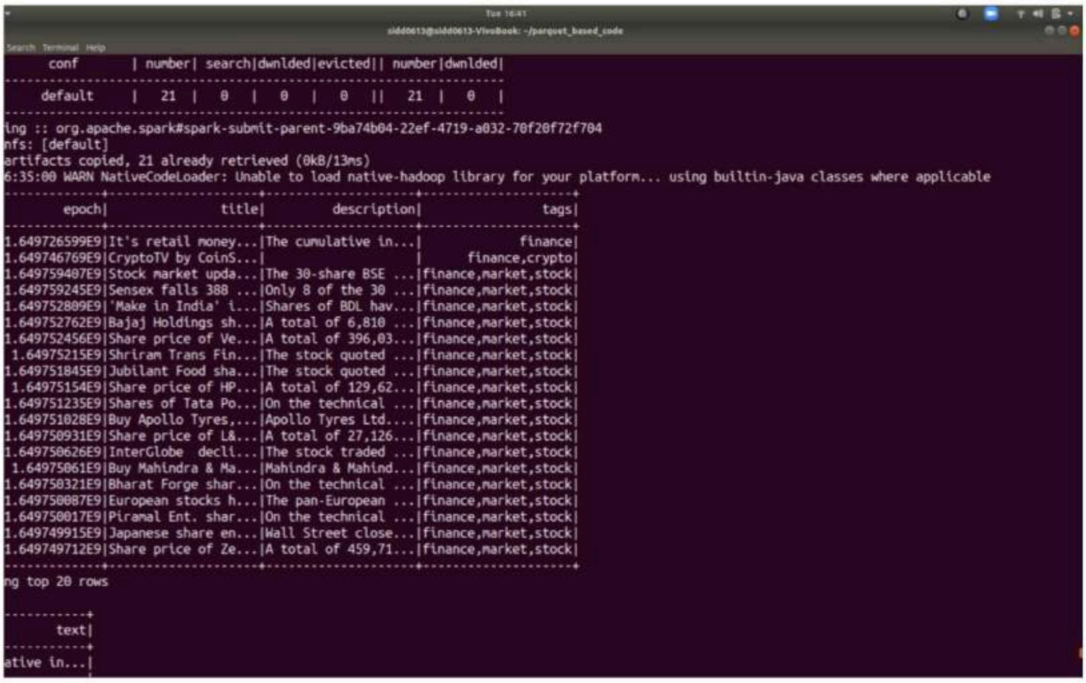
```

#writing to hdfs
final_df.write.partitionBy("date").mode("overwrite").parquet("hdfs://localhost:9000/final_project/temp/output/")
print("successfully written to hdfs!!!")
```

OUTPUT:

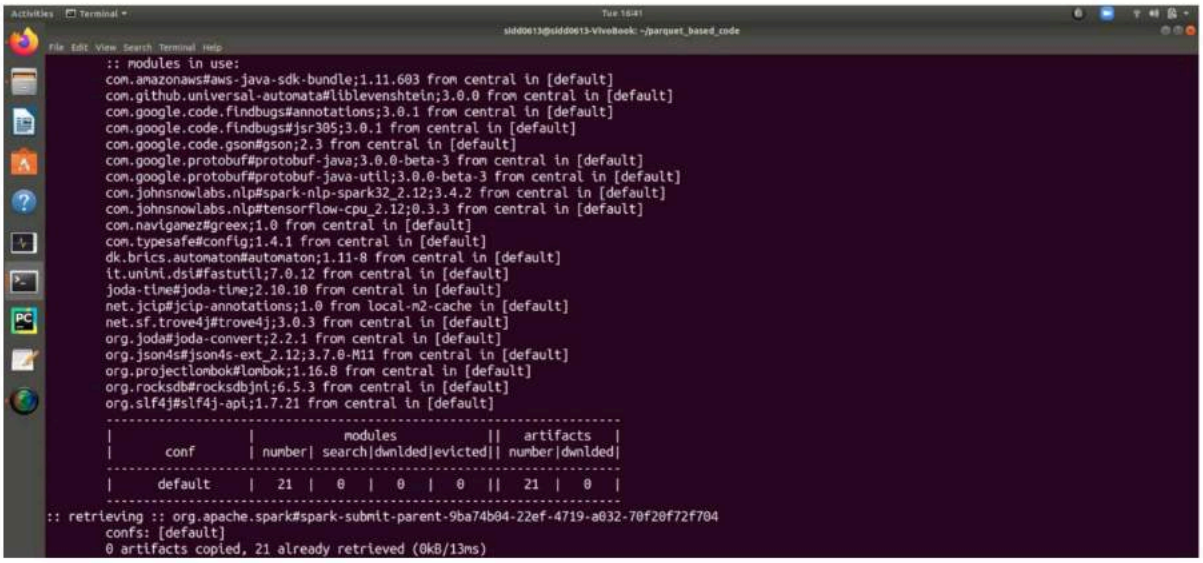


Activities
Terminal
Tue 16:41
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

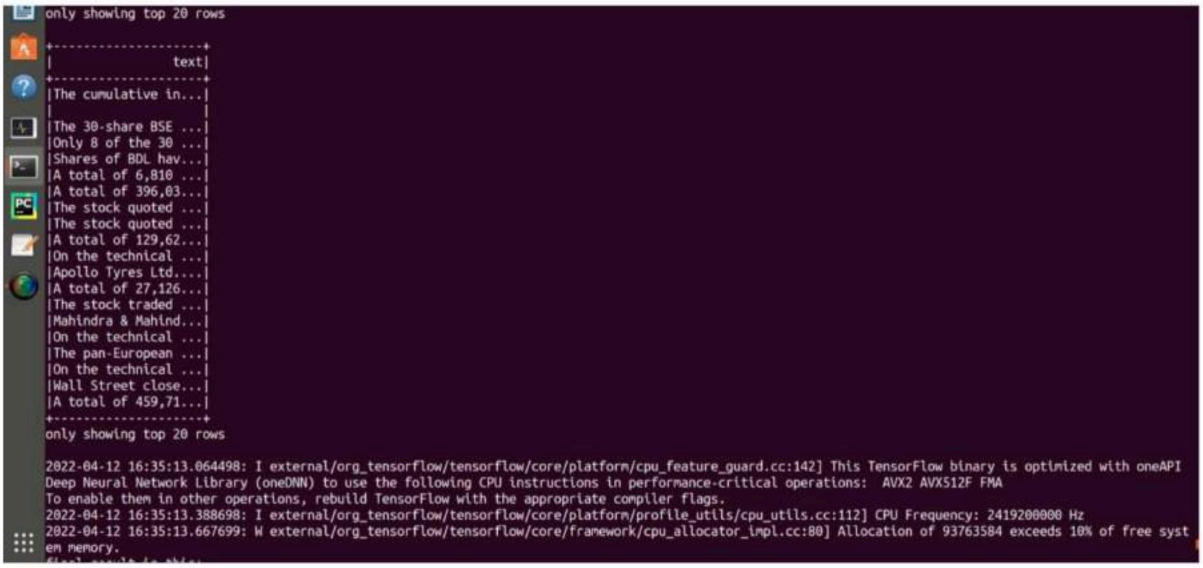
```
(base) sidd0613@sidd0613-VivoBook:~/parquet_based_code$ spark-submit --packages com.johnsnowlabs.nlp:spark-nlp-spark32_2.12:3.4.2 --master local nlp_pipe
line.py 2022-04-12
Picked up _JAVA_OPTIONS: -Xmx3g
Picked up _JAVA_OPTIONS: -Xmx3g
22/04/12 16:34:59 WARN Utils: Your hostname, sidd0613-VivoBook resolves to a loopback address: 127.0.1.1; using 192.168.1.7 instead (on interface wlo1)
22/04/12 16:34:59 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/sidd0613/DBDA_HOME/spark-3.2.0-bin-hadoop3.2/jars/spark-unsafe_2.12-3.
2.0.jar) to constructor java.nio.DirectByteBuffer(long,int)
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
:: loading settings :: url = jar:file:/home/sidd0613/DBDA_HOME/spark-3.2.0-bin-hadoop3.2/jars/ivy-2.5.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: /home/sidd0613/.ivy2/cache
The jars for the packages stored in: /home/sidd0613/.ivy2/jars
com.johnsnowlabs.nlp#spark-nlp-spark32_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-9ba74b04-22ef-4719-a032-70f20f72f704;1.0
        confs: [default]
        found com.johnsnowlabs.nlp#spark-nlp-spark32_2.12;3.4.2 in central
        found com.typesafe#config;1.4.1 in central
        found org.rocksdb#rocksdbjni;6.5.3 in central
        found com.amazonaws#aws-java-sdk-bundle;1.11.603 in central
        found com.github.universal-automata#liblevenshtein;3.0.0 in central
        found com.google.code.findbugs#annotations;3.0.1 in central
        found net.jcip#jcip-annotations;1.0 in local-m2-cache
        found com.google.code.findbugs#jsr305;3.0.1 in central
        found com.google.protobuf#protobuf-java-util;3.0.0-beta-3 in central
        found com.google.protobuf#protobuf-java;3.0.0-beta-3 in central
        found com.google.code.gson#gson;2.3 in central
        found it.unimi.dsi#fastutil;7.0.12 in central
        found org.projectlombok#lombok;1.16.8 in central
        found org.slf4j#slf4j-api;1.7.21 in central
        found com.navigamez#greex;1.0 in central
        found dk.brics.automaton#automaton;1.11-8 in central
        found org.json4s#json4s-ext_2.12;3.7.0-M11 in central
        found joda-time#joda-time;2.10.10 in central
        found org.joda#joda-convert;2.2.1 in central
        found com.johnsnowlabs.nlp#tensorflow-cpu_2.12;0.3.3 in central
```



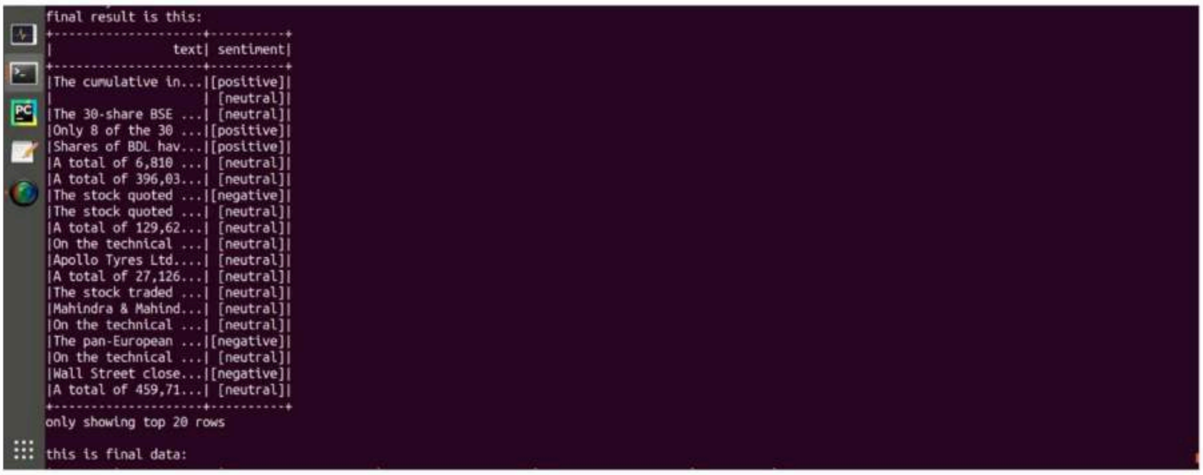
```
     conf       | number| search|dwnlded|evicted|| number|dwnlded|
-----------------------------------------------------------------
    default     |   21  |   0   |   0   |   0   ||   21  |   0   |
-----------------------------------------------------------------
ing :: org.apache.spark#spark-submit-parent-9ba74b04-22ef-4719-a032-70f20f72f704
nfs: [default]
artifacts copied, 21 already retrieved (0kB/13ms)
6:35:00 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
-------------+--------------------+--------------------+--------------------+
        epoch|               title|         description|                tags|
-------------+--------------------+--------------------+--------------------+
1.649726599E9|It's retail money...|The cumulative in...|             finance|
1.649746769E9|CryptoTV by CoinS...|                    |      finance,crypto|
1.649759407E9|Stock market upda...|The 30-share BSE ...|finance,market,stock|
1.649759245E9|Sensex falls 388 ...|Only 8 of the 30 ...|finance,market,stock|
1.649752809E9|'Make in India' i...|Shares of BDL hav...|finance,market,stock|
1.649752762E9|Bajaj Holdings sh...|A total of 6,810 ...|finance,market,stock|
1.649752456E9|Share price of Ve...|A total of 396,03...|finance,market,stock|
 1.64975215E9|Shriram Trans Fin...|The stock quoted ...|finance,market,stock|
1.649751845E9|Jubilant Food sha...|The stock quoted ...|finance,market,stock|
 1.64975154E9|Share price of HP...|A total of 129,62...|finance,market,stock|
1.649751235E9|Shares of Tata Po...|On the technical ...|finance,market,stock|
1.649751028E9|Buy Apollo Tyres,...|Apollo Tyres Ltd....|finance,market,stock|
1.649750931E9|Share price of L&...|A total of 27,126...|finance,market,stock|
1.649750626E9|InterGlobe  decli...|The stock traded ...|finance,market,stock|
 1.64975061E9|Buy Mahindra & Ma...|Mahindra & Mahind...|finance,market,stock|
1.649750321E9|Bharat Forge shar...|On the technical ...|finance,market,stock|
1.649750087E9|European stocks h...|The pan-European ...|finance,market,stock|
1.649750017E9|Piramal Ent. shar...|On the technical ...|finance,market,stock|
1.649749915E9|Japanese share en...|Wall Street close...|finance,market,stock|
1.649749712E9|Share price of Ze...|A total of 459,71...|finance,market,stock|
-------------+--------------------+--------------------+--------------------+
ng top 20 rows

-----------+
       text|
-----------+
ative in...|
```


Activities
Terminal
Tue 16:41
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

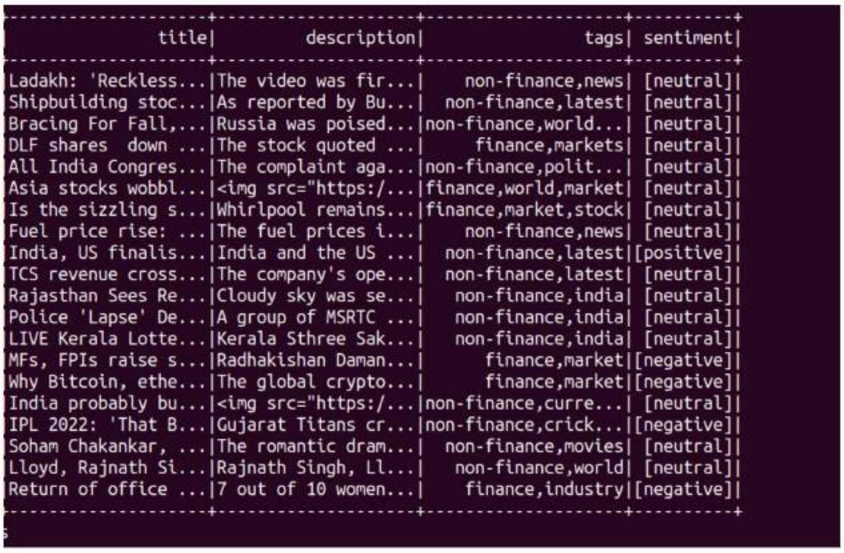
```
        :: modules in use:
        com.amazonaws#aws-java-sdk-bundle;1.11.603 from central in [default]
        com.github.universal-automata#liblevenshtein;3.0.0 from central in [default]
        com.google.code.findbugs#annotations;3.0.1 from central in [default]
        com.google.code.findbugs#jsr305;3.0.1 from central in [default]
        com.google.code.gson#gson;2.3 from central in [default]
        com.google.protobuf#protobuf-java;3.0.0-beta-3 from central in [default]
        com.google.protobuf#protobuf-java-util;3.0.0-beta-3 from central in [default]
        com.johnsnowlabs.nlp#spark-nlp-spark32_2.12;3.4.2 from central in [default]
        com.johnsnowlabs.nlp#tensorflow-cpu_2.12;0.3.3 from central in [default]
        com.navigamez#greex;1.0 from central in [default]
        com.typesafe#config;1.4.1 from central in [default]
        dk.brics.automaton#automaton;1.11-8 from central in [default]
        it.unimi.dsi#fastutil;7.0.12 from central in [default]
        joda-time#joda-time;2.10.10 from central in [default]
        net.jcip#jcip-annotations;1.0 from local-m2-cache in [default]
        net.sf.trove4j#trove4j;3.0.3 from central in [default]
        org.joda#joda-convert;2.2.1 from central in [default]
        org.json4s#json4s-ext_2.12;3.7.0-M11 from central in [default]
        org.projectlombok#lombok;1.16.8 from central in [default]
        org.rocksdb#rocksdbjni;6.5.3 from central in [default]
        org.slf4j#slf4j-api;1.7.21 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   21  |   0   |   0   |   0   ||   21  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-9ba74b04-22ef-4719-a032-70f20f72f704
        confs: [default]
        0 artifacts copied, 21 already retrieved (0kB/13ms)
```



```
only showing top 20 rows

+--------------------+
|                text|
+--------------------+
|The cumulative in...|
|                    |
|The 30-share BSE ...|
|Only 8 of the 30 ...|
|Shares of BDL hav...|
|A total of 6,810 ...|
|A total of 396,03...|
|The stock quoted ...|
|The stock quoted ...|
|A total of 129,62...|
|On the technical ...|
|Apollo Tyres Ltd....|
|A total of 27,126...|
|The stock traded ...|
|Mahindra & Mahind...|
|On the technical ...|
|The pan-European ...|
|On the technical ...|
|Wall Street close...|
|A total of 459,71...|
+--------------------+
only showing top 20 rows

2022-04-12 16:35:13.064498: I external/org_tensorflow/tensorflow/core/platform/cpu_feature_guard.cc:142] This TensorFlow binary is optimized with oneAPI
Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations:  AVX2 AVX512F FMA
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
2022-04-12 16:35:13.388698: I external/org_tensorflow/tensorflow/core/platform/profile_utils/cpu_utils.cc:112] CPU Frequency: 2419200000 Hz
2022-04-12 16:35:13.667699: W external/org_tensorflow/tensorflow/core/framework/cpu_allocator_impl.cc:80] Allocation of 93763584 exceeds 10% of free syst
em memory.
final result is this:
```

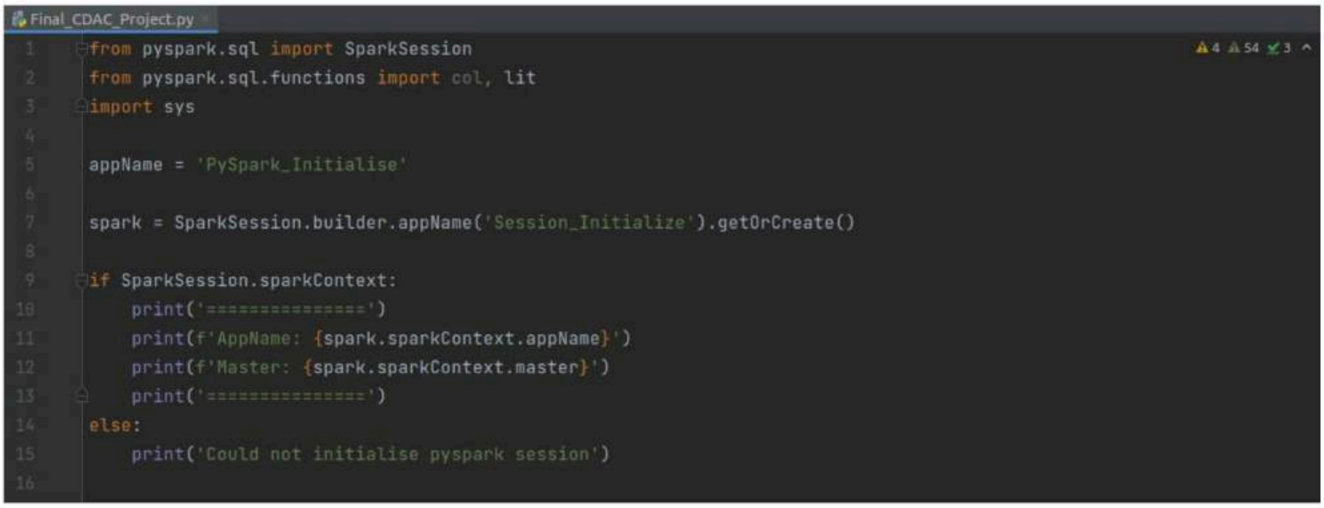


```
final result is this:
+--------------------+----------+
|                text| sentiment|
+--------------------+----------+
|The cumulative in...|[positive]|
|                    | [neutral]|
|The 30-share BSE ...| [neutral]|
|Only 8 of the 30 ...|[positive]|
|Shares of BDL hav...|[positive]|
|A total of 6,810 ...| [neutral]|
|A total of 396,03...| [neutral]|
|The stock quoted ...|[negative]|
|The stock quoted ...| [neutral]|
|A total of 129,62...| [neutral]|
|On the technical ...| [neutral]|
|Apollo Tyres Ltd....| [neutral]|
|A total of 27,126...| [neutral]|
|The stock traded ...| [neutral]|
|Mahindra & Mahind...| [neutral]|
|On the technical ...| [neutral]|
|The pan-European ...|[negative]|
|On the technical ...| [neutral]|
|Wall Street close...|[negative]|
|A total of 459,71...| [neutral]|
+--------------------+----------+
only showing top 20 rows

this is final data:
```

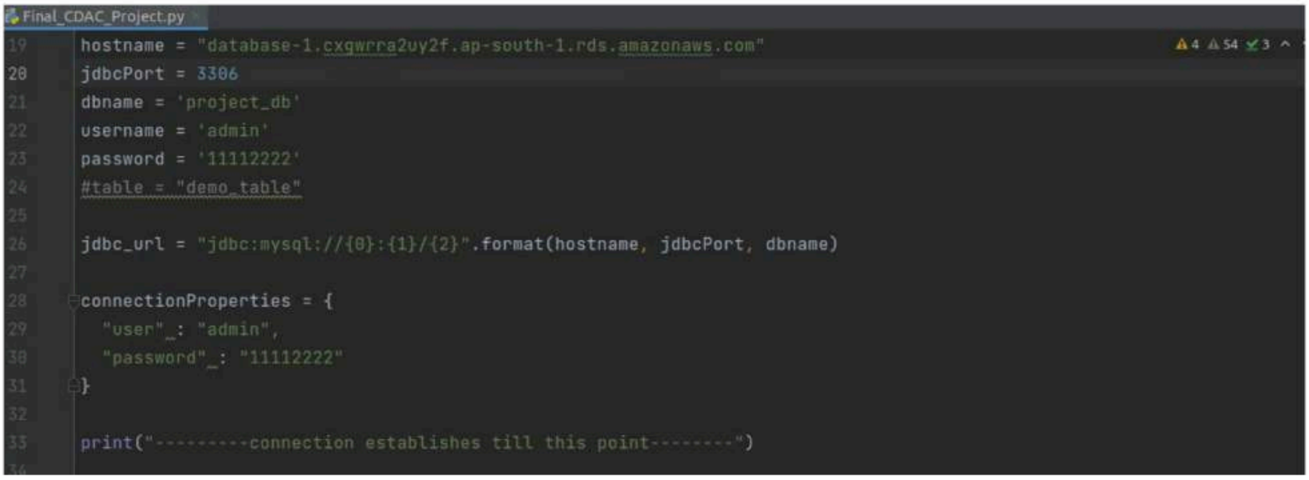


```
--------------------+--------------------+--------------------+----------+
               title|         description|                tags| sentiment|
--------------------+--------------------+--------------------+----------+
Ladakh: 'Reckless...|The video was fir...|   non-finance,news| [neutral]|
Shipbuilding stoc...|As reported by Bu...| non-finance,latest| [neutral]|
Bracing For Fall,...|Russia was poised...|non-finance,world...| [neutral]|
DLF shares  down ...|The stock quoted ...|    finance,markets| [neutral]|
All India Congres...|The complaint aga...|non-finance,polit...| [neutral]|
Asia stocks wobbl...|<img src="https:/...|finance,world,market| [neutral]|
Is the sizzling s...|Whirlpool remains...|finance,market,stock| [neutral]|
Fuel price rise: ...|The fuel prices i...|   non-finance,news| [neutral]|
India, US finalis...|India and the US ...| non-finance,latest|[positive]|
TCS revenue cross...|The company's ope...| non-finance,latest| [neutral]|
Rajasthan Sees Re...|Cloudy sky was se...|  non-finance,india| [neutral]|
Police 'Lapse' De...|A group of MSRTC ...|  non-finance,india| [neutral]|
LIVE Kerala Lotte...|Kerala Sthree Sak...|  non-finance,india| [neutral]|
MFs, FPIs raise s...|Radhakishan Daman...|     finance,market|[negative]|
Why Bitcoin, ethe...|The global crypto...|     finance,market|[negative]|
India probably bu...|<img src="https:/...|non-finance,curre...| [neutral]|
IPL 2022: 'That B...|Gujarat Titans cr...|non-finance,crick...|[negative]|
Soham Chakankar, ...|The romantic dram...| non-finance,movies| [neutral]|
Lloyd, Rajnath Si...|Rajnath Singh, Ll...|  non-finance,world| [neutral]|
Return of office ...|7 out of 10 women...|   finance,industry|[negative]|
--------------------+--------------------+--------------------+----------+
s
```

Create spark session and Establish

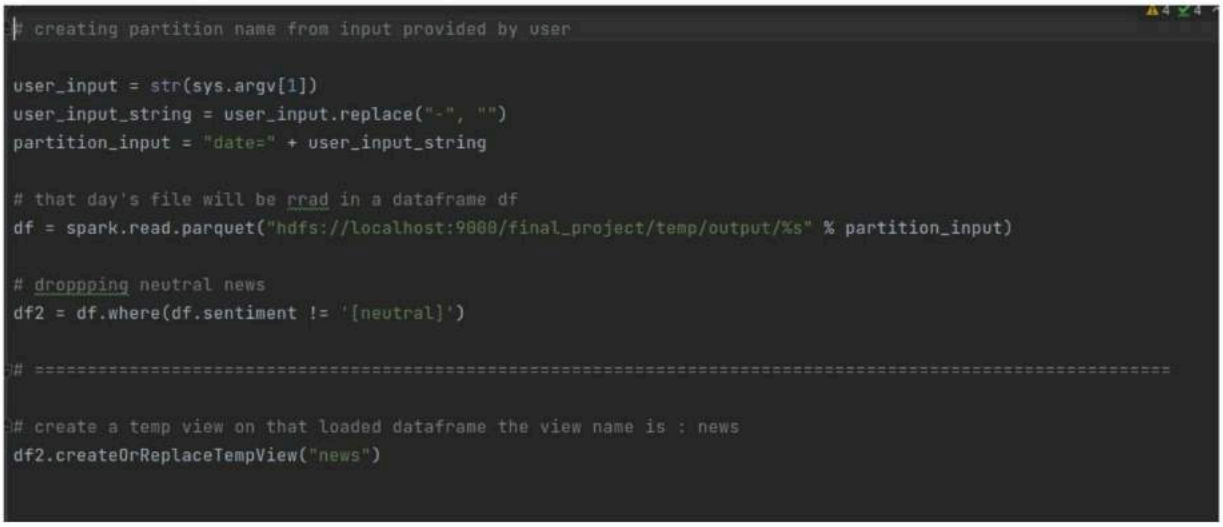


```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, lit
import sys

appName = 'PySpark_Initialise'

spark = SparkSession.builder.appName('Session_Initialize').getOrCreate()

if SparkSession.sparkContext:
    print('===============')
    print(f'AppName: {spark.sparkContext.appName}')
    print(f'Master: {spark.sparkContext.master}')
    print('===============')
else:
    print('Could not initialise pyspark session')
```



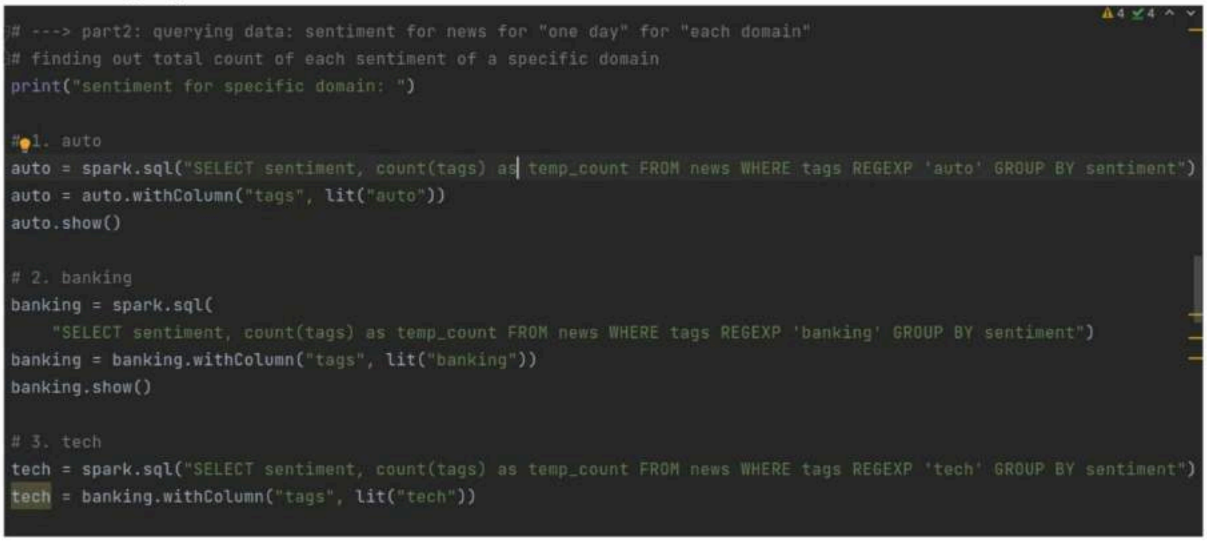
```python
hostname = "database-1.cxgwrra2uy2f.ap-south-1.rds.amazonaws.com"
jdbcPort = 3306
dbname = 'project_db'
username = 'admin'
password = '11112222'
#table = "demo_table"

jdbc_url = "jdbc:mysql://{0}:{1}/{2}".format(hostname, jdbcPort, dbname)

connectionProperties = {
  "user" : "admin",
  "password" : "11112222"
}

print("---------connection establishes till this point--------")
```

Taking input for performing queries and reading parquet format

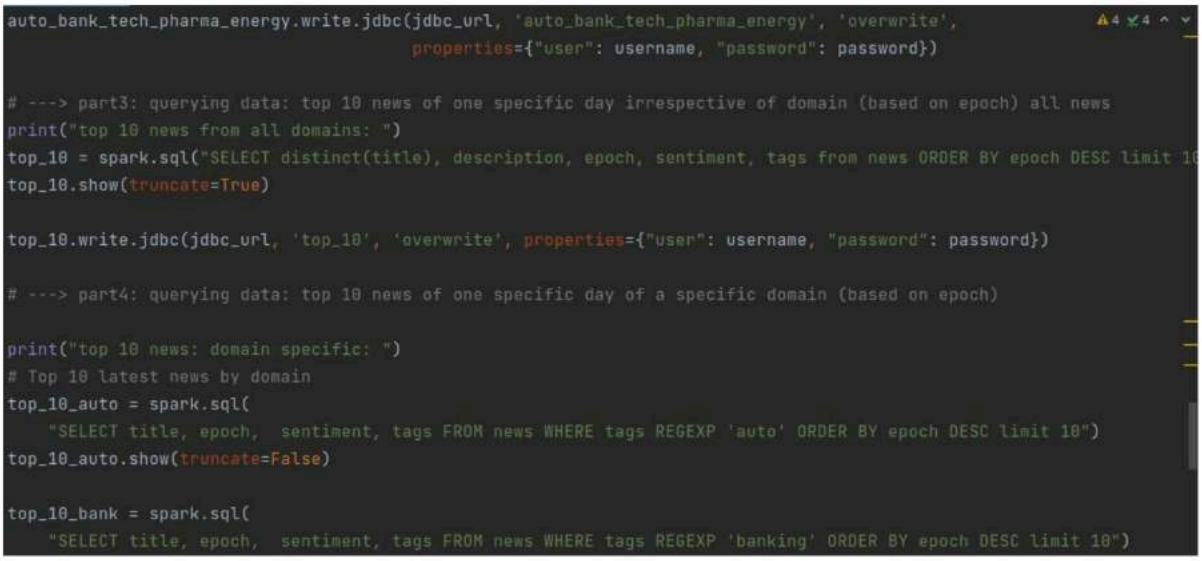


```
# creating partition name from input provided by user

user_input = str(sys.argv[1])
user_input_string = user_input.replace("-", "")
partition_input = "date=" + user_input_string

# that day's file will be rrad in a dataframe df
df = spark.read.parquet("hdfs://localhost:9000/final_project/temp/output/%s" % partition_input)

# droppping neutral news
df2 = df.where(df.sentiment != '[neutral]')

# ===========================================================================================================

# create a temp view on that loaded dataframe the view name is : news
df2.createOrReplaceTempView("news")
```

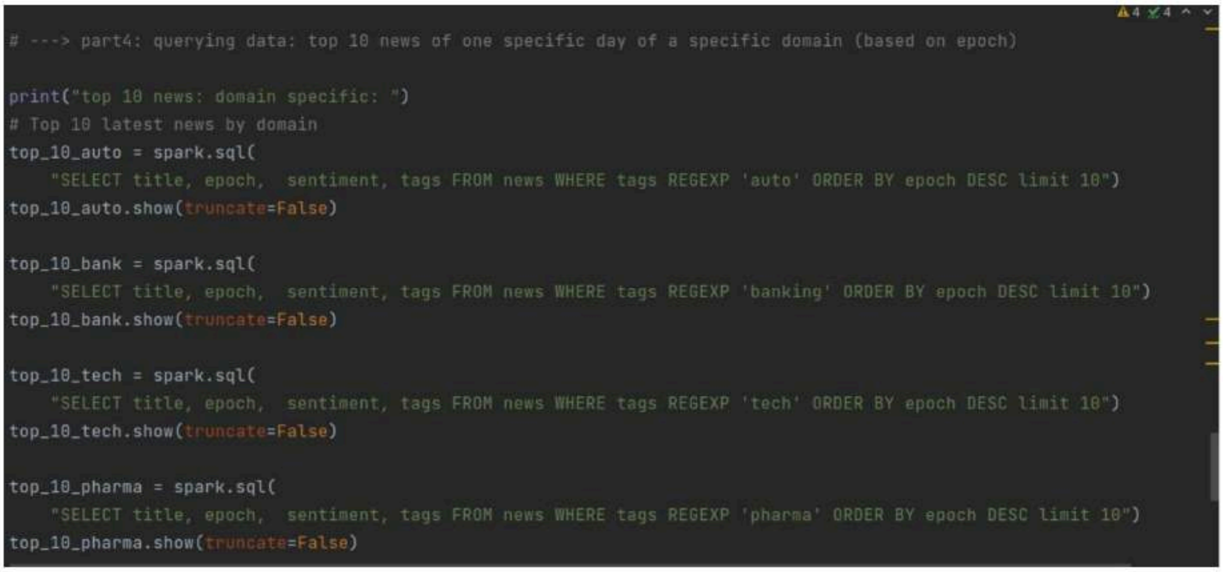
Performing queries on the available data



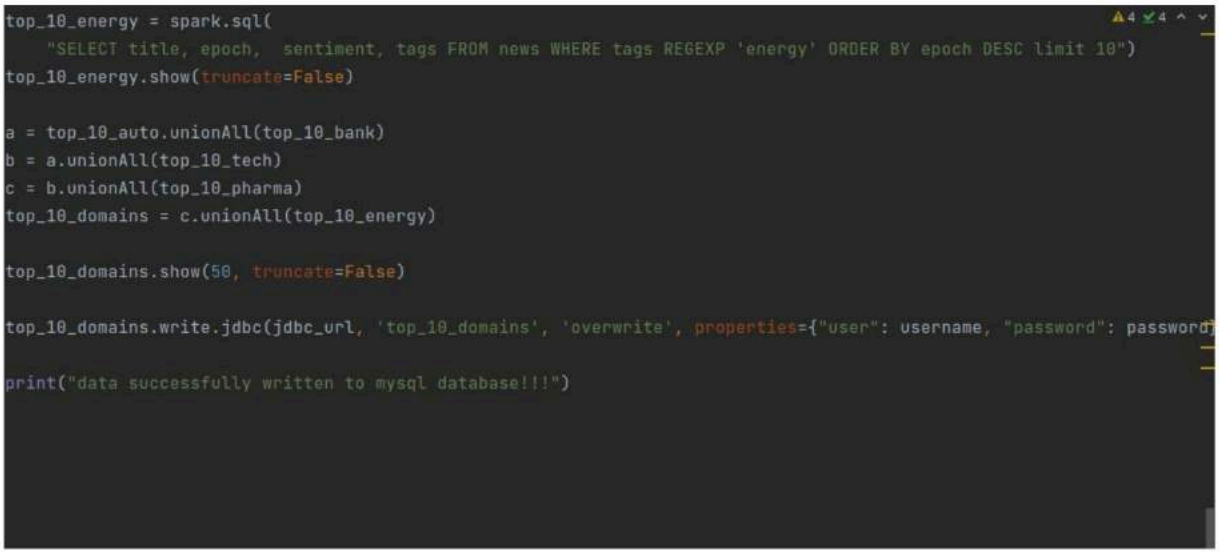
```
# ---> part2: querying data: sentiment for news for "one day" for "each domain"
# finding out total count of each sentiment of a specific domain
print("sentiment for specific domain: ")

#1. auto
auto = spark.sql("SELECT sentiment, count(tags) as temp_count FROM news WHERE tags REGEXP 'auto' GROUP BY sentiment")
auto = auto.withColumn("tags", lit("auto"))
auto.show()

# 2. banking
banking = spark.sql(
    "SELECT sentiment, count(tags) as temp_count FROM news WHERE tags REGEXP 'banking' GROUP BY sentiment")
banking = banking.withColumn("tags", lit("banking"))
banking.show()

# 3. tech
tech = spark.sql("SELECT sentiment, count(tags) as temp_count FROM news WHERE tags REGEXP 'tech' GROUP BY sentiment")
tech = banking.withColumn("tags", lit("tech"))
```



```
auto_bank_tech_pharma_energy.write.jdbc(jdbc_url, 'auto_bank_tech_pharma_energy', 'overwrite',
                                        properties={"user": username, "password": password})

# ---> part3: querying data: top 10 news of one specific day irrespective of domain (based on epoch) all news
print("top 10 news from all domains: ")
top_10 = spark.sql("SELECT distinct(title), description, epoch, sentiment, tags from news ORDER BY epoch DESC limit 1
top_10.show(truncate=True)

top_10.write.jdbc(jdbc_url, 'top_10', 'overwrite', properties={"user": username, "password": password})

# ---> part4: querying data: top 10 news of one specific day of a specific domain (based on epoch)

print("top 10 news: domain specific: ")
# Top 10 latest news by domain
top_10_auto = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'auto' ORDER BY epoch DESC limit 10")
top_10_auto.show(truncate=False)

top_10_bank = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'banking' ORDER BY epoch DESC limit 10")
```

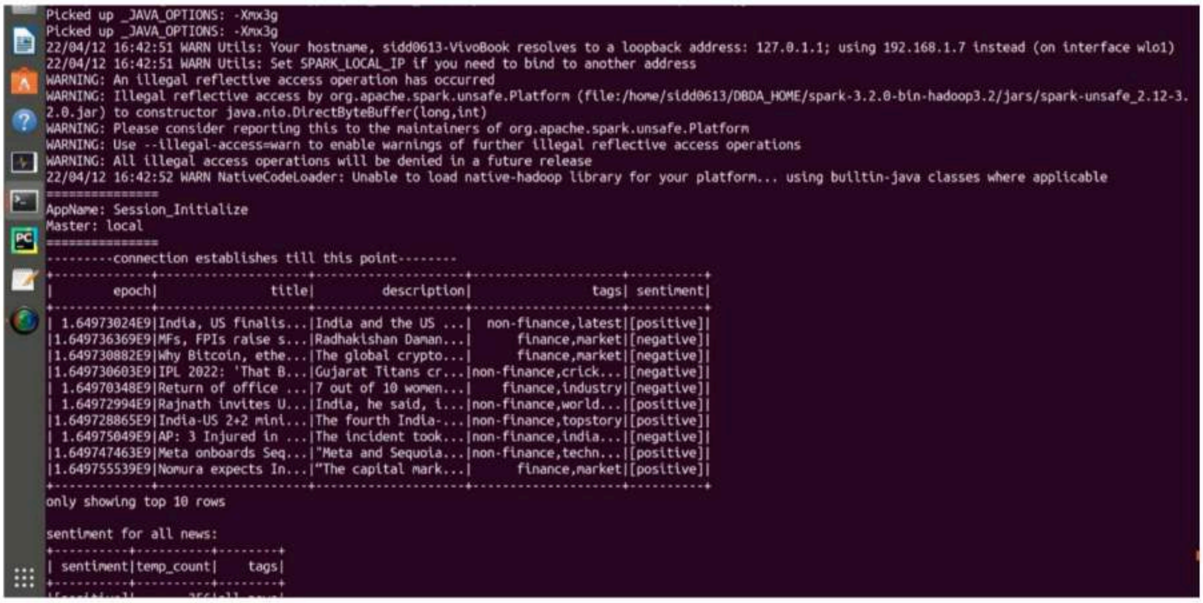


```
# ---> part4: querying data: top 10 news of one specific day of a specific domain (based on epoch)

print("top 10 news: domain specific: ")
# Top 10 latest news by domain
top_10_auto = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'auto' ORDER BY epoch DESC limit 10")
top_10_auto.show(truncate=False)

top_10_bank = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'banking' ORDER BY epoch DESC limit 10")
top_10_bank.show(truncate=False)

top_10_tech = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'tech' ORDER BY epoch DESC limit 10")
top_10_tech.show(truncate=False)

top_10_pharma = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'pharma' ORDER BY epoch DESC limit 10")
top_10_pharma.show(truncate=False)
```

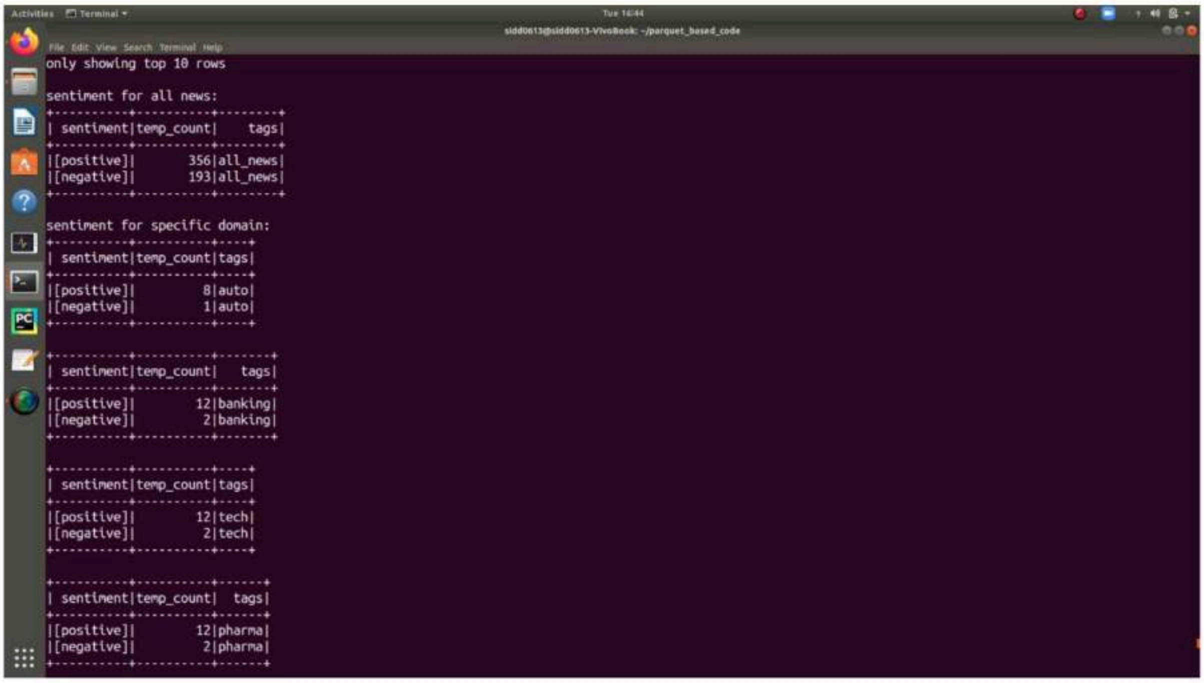


```
top_10_energy = spark.sql(
    "SELECT title, epoch,  sentiment, tags FROM news WHERE tags REGEXP 'energy' ORDER BY epoch DESC limit 10")
top_10_energy.show(truncate=False)

a = top_10_auto.unionAll(top_10_bank)
b = a.unionAll(top_10_tech)
c = b.unionAll(top_10_pharma)
top_10_domains = c.unionAll(top_10_energy)

top_10_domains.show(50, truncate=False)

top_10_domains.write.jdbc(jdbc_url, 'top_10_domains', 'overwrite', properties={"user": username, "password": password

print("data successfully written to mysql database!!!")
```

6. OUTPUT

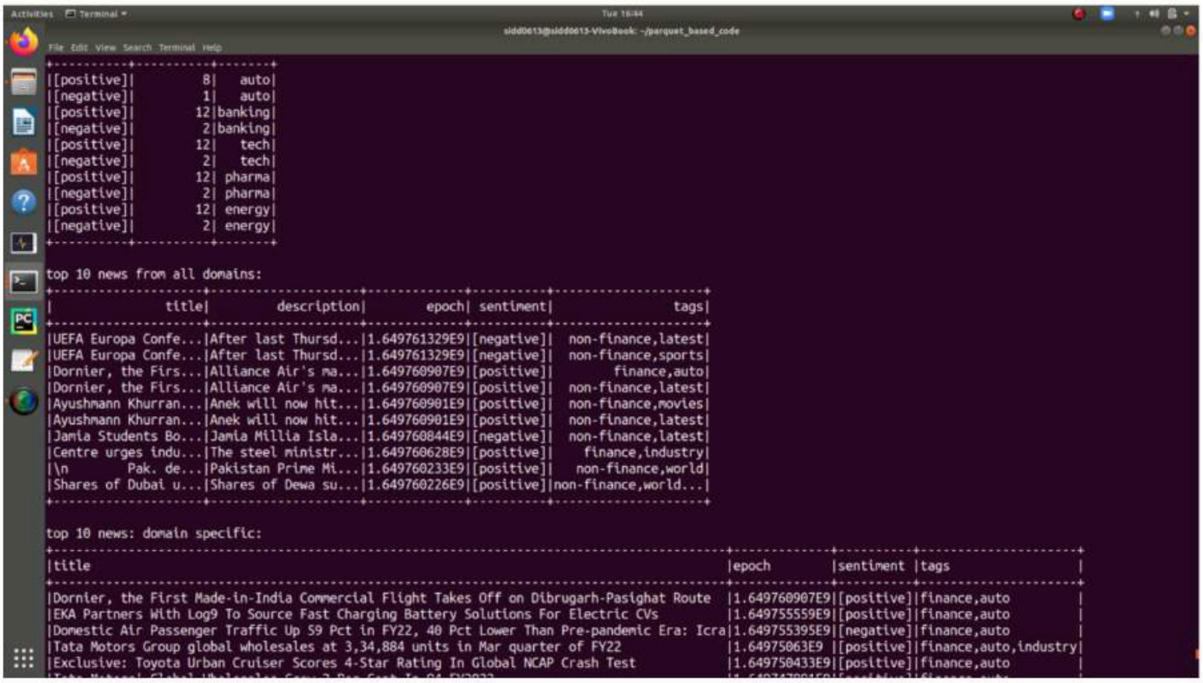


```
Picked up _JAVA_OPTIONS: -Xmx3g
Picked up _JAVA_OPTIONS: -Xmx3g
22/04/12 16:42:51 WARN Utils: Your hostname, sidd0613-VivoBook resolves to a loopback address: 127.0.1.1; using 192.168.1.7 instead (on interface wlo1)
22/04/12 16:42:51 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/home/sidd0613/DBDA_HOME/spark-3.2.0-bin-hadoop3.2/jars/spark-unsafe_2.12-3.
2.0.jar) to constructor java.nio.DirectByteBuffer(long,int)
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
22/04/12 16:42:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
===============
AppName: Session_Initialize
Master: local
===============
---------connection establishes till this point--------
+------------+--------------------+--------------------+--------------------+----------+
|       epoch|               title|         description|                tags| sentiment|
+------------+--------------------+--------------------+--------------------+----------+
| 1.64973024E9|India, US finalis...|India and the US ...|  non-finance,latest|[positive]|
|1.649736369E9|MFs, FPIs raise s...|Radhakishan Daman...|      finance,market|[negative]|
|1.649730882E9|Why Bitcoin, ethe...|The global crypto...|      finance,market|[negative]|
|1.649730603E9|IPL 2022: 'That B...|Gujarat Titans cr...|non-finance,crick...|[negative]|
| 1.64970348E9|Return of office ...|7 out of 10 women...|    finance,industry|[negative]|
| 1.64972994E9|Rajnath invites U...|India, he said, i...|non-finance,world...|[positive]|
|1.649728865E9|India-US 2+2 mini...|The fourth India-...|non-finance,topstory|[positive]|
| 1.64975049E9|AP: 3 Injured in ...|The incident took...|non-finance,india...|[negative]|
|1.649747463E9|Meta onboards Seq...|"Meta and Sequoia...|non-finance,techn...|[positive]|
|1.649755539E9|Nomura expects In...|"The capital mark...|      finance,market|[positive]|
+------------+--------------------+--------------------+--------------------+----------+
only showing top 10 rows

sentiment for all news:
+----------+----------+--------+
| sentiment|temp_count|    tags|
+----------+----------+--------+
```

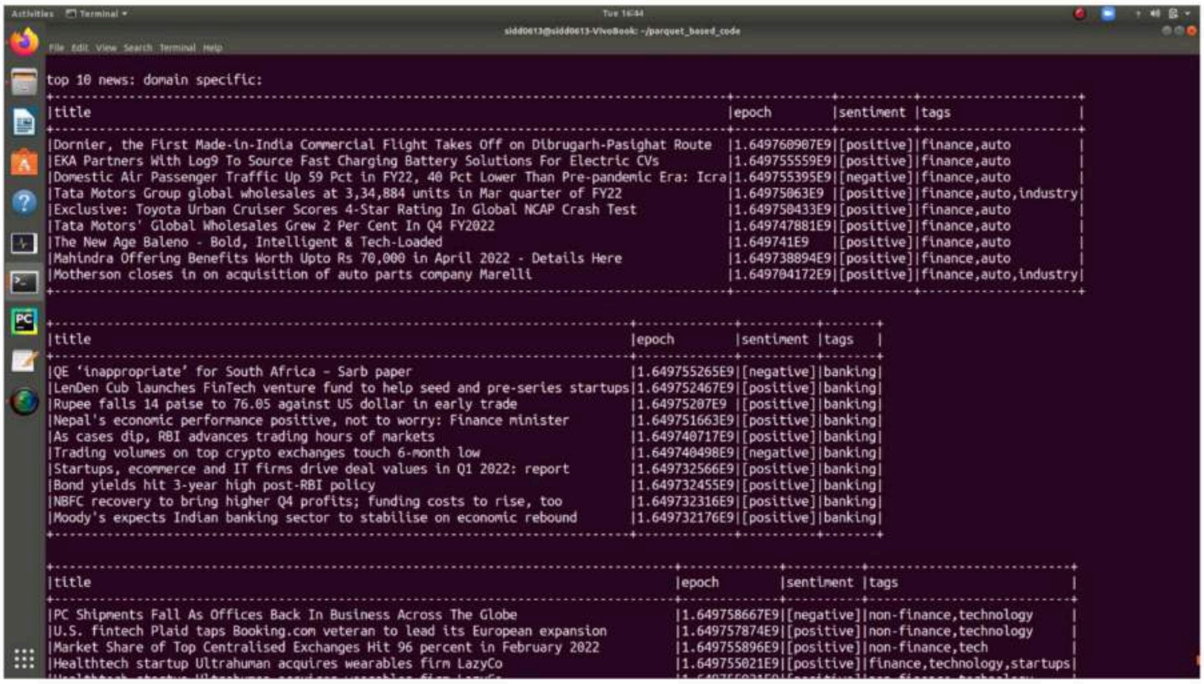


Activities
Terminal
Tue 16:44
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

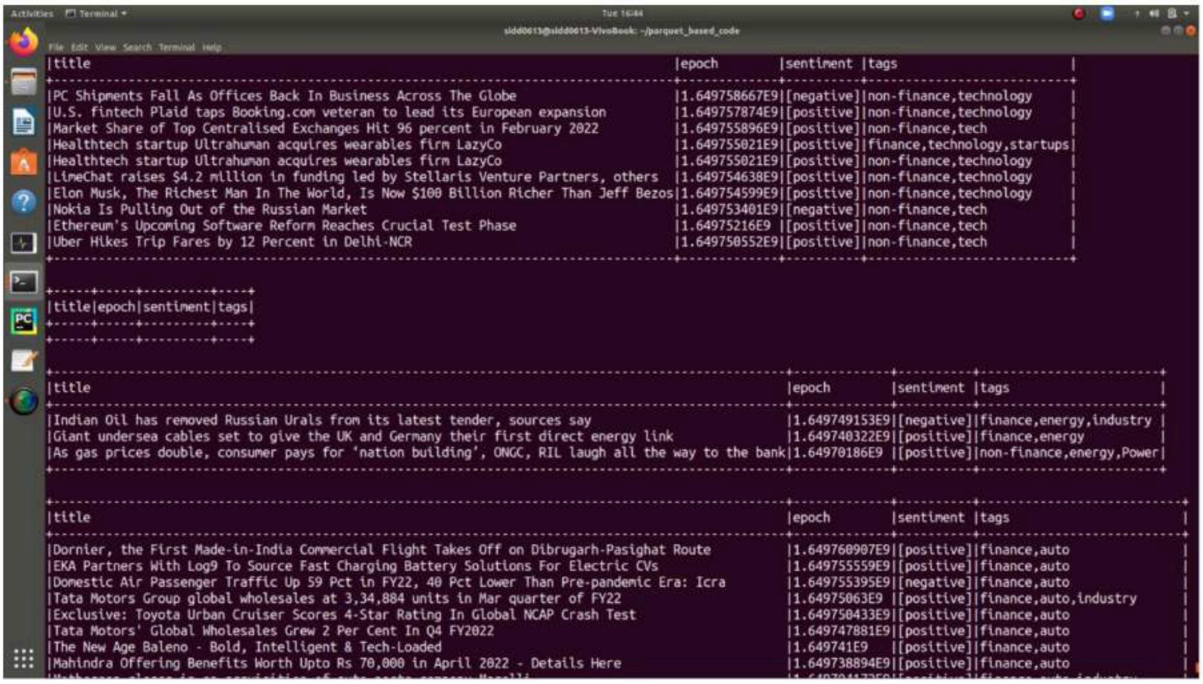
```
only showing top 10 rows

sentiment for all news:
+----------+----------+--------+
| sentiment|temp_count|    tags|
+----------+----------+--------+
|[positive]|       356|all_news|
|[negative]|       193|all_news|
+----------+----------+--------+

sentiment for specific domain:
+----------+----------+----+
| sentiment|temp_count|tags|
+----------+----------+----+
|[positive]|         8|auto|
|[negative]|         1|auto|
+----------+----------+----+

+----------+----------+-------+
| sentiment|temp_count|   tags|
+----------+----------+-------+
|[positive]|        12|banking|
|[negative]|         2|banking|
+----------+----------+-------+

+----------+----------+----+
| sentiment|temp_count|tags|
+----------+----------+----+
|[positive]|        12|tech|
|[negative]|         2|tech|
+----------+----------+----+

+----------+----------+------+
| sentiment|temp_count|  tags|
+----------+----------+------+
|[positive]|        12|pharma|
|[negative]|         2|pharma|
+----------+----------+------+
```


Activities
Terminal
Tue 16:44
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

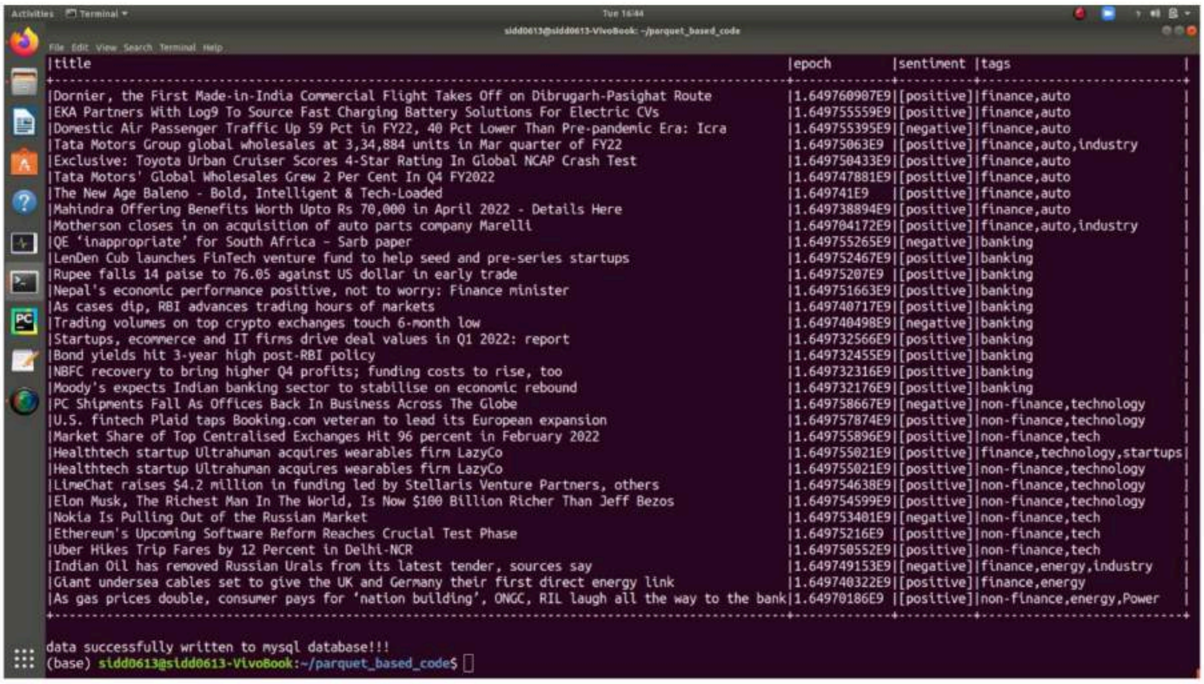
```
+----------+---------+-------+
|[positive]|        8|   auto|
|[negative]|        1|   auto|
|[positive]|       12|banking|
|[negative]|        2|banking|
|[positive]|       12|   tech|
|[negative]|        2|   tech|
|[positive]|       12| pharma|
|[negative]|        2| pharma|
|[positive]|       12| energy|
|[negative]|        2| energy|
+----------+---------+-------+

top 10 news from all domains:
+--------------------+--------------------+-------------+----------+--------------------+
|               title|         description|        epoch| sentiment|                tags|
+--------------------+--------------------+-------------+----------+--------------------+
|UEFA Europa Confe...|After last Thursd...|1.649761329E9|[negative]|  non-finance,latest|
|UEFA Europa Confe...|After last Thursd...|1.649761329E9|[negative]|  non-finance,sports|
|Dornier, the Firs...|Alliance Air's ma...|1.649760907E9|[positive]|        finance,auto|
|Dornier, the Firs...|Alliance Air's ma...|1.649760907E9|[positive]|  non-finance,latest|
|Ayushmann Khurran...|Anek will now hit...|1.649760901E9|[positive]|  non-finance,movies|
|Ayushmann Khurran...|Anek will now hit...|1.649760901E9|[positive]|  non-finance,latest|
|Jamia Students Bo...|Jamia Millia Isla...|1.649760844E9|[negative]|  non-finance,latest|
|Centre urges indu...|The steel ministr...|1.649760628E9|[positive]|    finance,industry|
|\n        Pak. de...|Pakistan Prime Mi...|1.649760233E9|[positive]|   non-finance,world|
|Shares of Dubai u...|Shares of Dewa su...|1.649760226E9|[positive]|non-finance,world...|
+--------------------+--------------------+-------------+----------+--------------------+

top 10 news: domain specific:
+------------------------------------------------------------------------------------------+-------------+----------+---------------------+
|title                                                                                     |epoch        |sentiment |tags                 |
+------------------------------------------------------------------------------------------+-------------+----------+---------------------+
|Dornier, the First Made-in-India Commercial Flight Takes Off on Dibrugarh-Pasighat Route  |1.649760907E9|[positive]|finance,auto         |
|EKA Partners With Log9 To Source Fast Charging Battery Solutions For Electric CVs         |1.649755559E9|[positive]|finance,auto         |
|Domestic Air Passenger Traffic Up 59 Pct in FY22, 40 Pct Lower Than Pre-pandemic Era: Icra|1.649755395E9|[negative]|finance,auto         |
|Tata Motors Group global wholesales at 3,34,884 units in Mar quarter of FY22              |1.64975063E9 |[positive]|finance,auto,industry|
|Exclusive: Toyota Urban Cruiser Scores 4-Star Rating In Global NCAP Crash Test            |1.649750433E9|[positive]|finance,auto         |
```


Activities
Terminal
Tue 16:44
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

```
top 10 news: domain specific:
+------------------------------------------------------------------------------------------+-------------+----------+---------------------+
|title                                                                                     |epoch        |sentiment |tags                 |
+------------------------------------------------------------------------------------------+-------------+----------+---------------------+
|Dornier, the First Made-in-India Commercial Flight Takes Off on Dibrugarh-Pasighat Route  |1.649760907E9|[positive]|finance,auto         |
|EKA Partners With Log9 To Source Fast Charging Battery Solutions For Electric CVs         |1.649755559E9|[positive]|finance,auto         |
|Domestic Air Passenger Traffic Up 59 Pct in FY22, 40 Pct Lower Than Pre-pandemic Era: Icra|1.649755395E9|[negative]|finance,auto         |
|Tata Motors Group global wholesales at 3,34,884 units in Mar quarter of FY22              |1.64975063E9 |[positive]|finance,auto,industry|
|Exclusive: Toyota Urban Cruiser Scores 4-Star Rating In Global NCAP Crash Test            |1.649750433E9|[positive]|finance,auto         |
|Tata Motors' Global Wholesales Grew 2 Per Cent In Q4 FY2022                               |1.649747881E9|[positive]|finance,auto         |
|The New Age Baleno - Bold, Intelligent & Tech-Loaded                                      |1.649741E9   |[positive]|finance,auto         |
|Mahindra Offering Benefits Worth Upto Rs 70,000 in April 2022 - Details Here              |1.649738894E9|[positive]|finance,auto         |
|Motherson closes in on acquisition of auto parts company Marelli                          |1.649704172E9|[positive]|finance,auto,industry|
+------------------------------------------------------------------------------------------+-------------+----------+---------------------+

+-----------------------------------------------------------------------------+-------------+----------+-------+
|title                                                                        |epoch        |sentiment |tags   |
+-----------------------------------------------------------------------------+-------------+----------+-------+
|QE ‘inappropriate’ for South Africa - Sarb paper                             |1.649755265E9|[negative]|banking|
|LenDen Cub launches FinTech venture fund to help seed and pre-series startups|1.649752467E9|[positive]|banking|
|Rupee falls 14 paise to 76.05 against US dollar in early trade               |1.64975207E9 |[positive]|banking|
|Nepal's economic performance positive, not to worry: Finance minister        |1.649751663E9|[positive]|banking|
|As cases dip, RBI advances trading hours of markets                          |1.649740717E9|[positive]|banking|
|Trading volumes on top crypto exchanges touch 6-month low                    |1.649740498E9|[negative]|banking|
|Startups, ecommerce and IT firms drive deal values in Q1 2022: report        |1.649732566E9|[positive]|banking|
|Bond yields hit 3-year high post-RBI policy                                  |1.649732455E9|[positive]|banking|
|NBFC recovery to bring higher Q4 profits; funding costs to rise, too         |1.649732316E9|[positive]|banking|
|Moody's expects Indian banking sector to stabilise on economic rebound       |1.649732176E9|[positive]|banking|
+-----------------------------------------------------------------------------+-------------+----------+-------+

+-----------------------------------------------------------------------------------+-------------+----------+---------------------------+
|title                                                                              |epoch        |sentiment |tags                       |
+-----------------------------------------------------------------------------------+-------------+----------+---------------------------+
|PC Shipments Fall As Offices Back In Business Across The Globe                     |1.649758667E9|[negative]|non-finance,technology     |
|U.S. fintech Plaid taps Booking.com veteran to lead its European expansion         |1.649757874E9|[positive]|non-finance,technology     |
|Market Share of Top Centralised Exchanges Hit 96 percent in February 2022          |1.649755896E9|[positive]|non-finance,tech           |
|Healthtech startup Ultrahuman acquires wearables firm LazyCo                       |1.649755021E9|[positive]|finance,technology,startups|
```

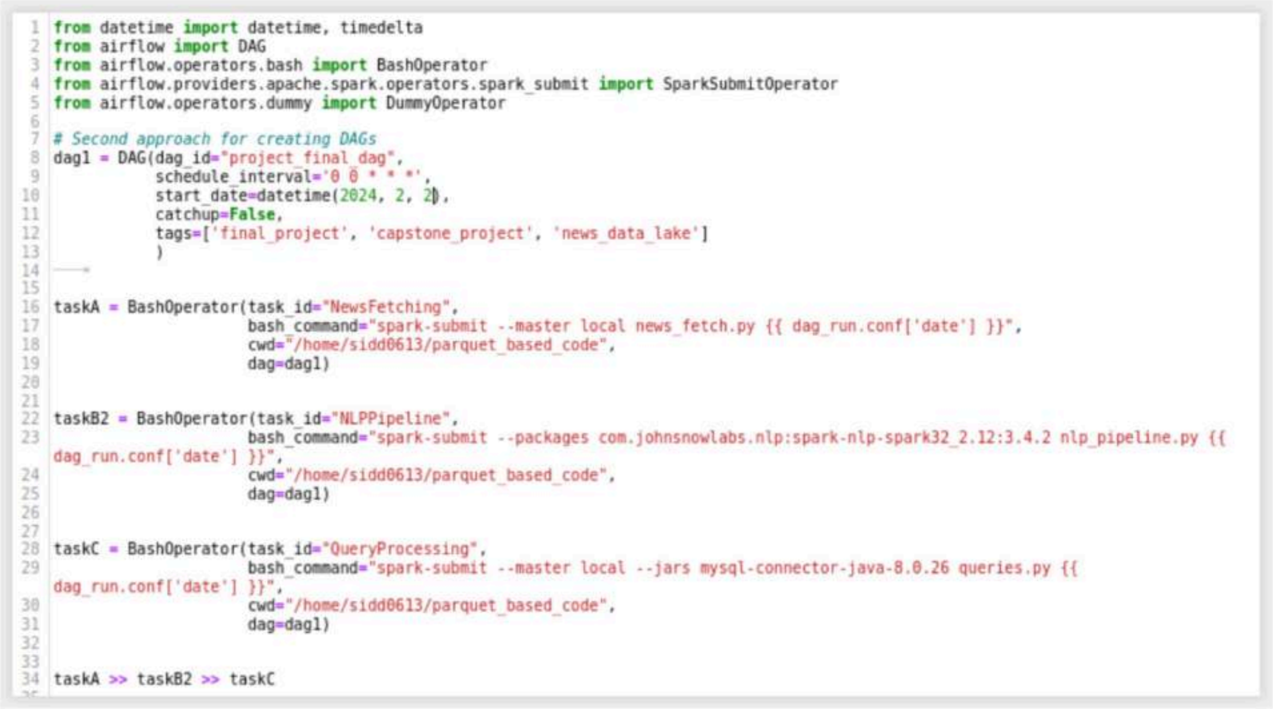

Activities
Terminal ▾
Tue 16:44
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

```
|title                                                                              |epoch        |sentiment |tags                       |
+-----------------------------------------------------------------------------------+-------------+----------+---------------------------+
|PC Shipments Fall As Offices Back In Business Across The Globe                     |1.649758667E9|[negative]|non-finance,technology     |
|U.S. fintech Plaid taps Booking.com veteran to lead its European expansion         |1.649757874E9|[positive]|non-finance,technology     |
|Market Share of Top Centralised Exchanges Hit 96 percent in February 2022          |1.649755896E9|[positive]|non-finance,tech           |
|Healthtech startup Ultrahuman acquires wearables firm LazyCo                       |1.649755021E9|[positive]|finance,technology,startups|
|Healthtech startup Ultrahuman acquires wearables firm LazyCo                       |1.649755021E9|[positive]|non-finance,technology     |
|LimeChat raises $4.2 million in funding led by Stellaris Venture Partners, others  |1.649754638E9|[positive]|non-finance,technology     |
|Elon Musk, The Richest Man In The World, Is Now $100 Billion Richer Than Jeff Bezos|1.649754599E9|[positive]|non-finance,technology     |
|Nokia Is Pulling Out of the Russian Market                                         |1.649753401E9|[negative]|non-finance,tech           |
|Ethereum's Upcoming Software Reform Reaches Crucial Test Phase                     |1.64975216E9 |[positive]|non-finance,tech           |
|Uber Hikes Trip Fares by 12 Percent in Delhi-NCR                                   |1.649750552E9|[positive]|non-finance,tech           |
+-----------------------------------------------------------------------------------+-------------+----------+---------------------------+


+-----+-----+---------+----+
|title|epoch|sentiment|tags|
+-----+-----+---------+----+
+-----+-----+---------+----+


+--------------------------------------------------------------------------------------------------+-------------+----------+------------------------+
|title                                                                                             |epoch        |sentiment |tags                    |
+--------------------------------------------------------------------------------------------------+-------------+----------+------------------------+
|Indian Oil has removed Russian Urals from its latest tender, sources say                          |1.649749153E9|[negative]|finance,energy,industry |
|Giant undersea cables set to give the UK and Germany their first direct energy link               |1.649740322E9|[positive]|finance,energy          |
|As gas prices double, consumer pays for ‘nation building’, ONGC, RIL laugh all the way to the bank|1.64970186E9 |[positive]|non-finance,energy,Power|
+--------------------------------------------------------------------------------------------------+-------------+----------+------------------------+

+--------------------------------------------------------------------------------------------------+-------------+----------+---------------------------+
|title                                                                                             |epoch        |sentiment |tags                       |
+--------------------------------------------------------------------------------------------------+-------------+----------+---------------------------+
|Dornier, the First Made-in-India Commercial Flight Takes Off on Dibrugarh-Pasighat Route          |1.649760907E9|[positive]|finance,auto               |
|EKA Partners With Log9 To Source Fast Charging Battery Solutions For Electric CVs                 |1.649755559E9|[positive]|finance,auto               |
|Domestic Air Passenger Traffic Up 59 Pct in FY22, 40 Pct Lower Than Pre-pandemic Era: Icra        |1.649755395E9|[negative]|finance,auto               |
|Tata Motors Group global wholesales at 3,34,884 units in Mar quarter of FY22                      |1.64975063E9 |[positive]|finance,auto,industry      |
|Exclusive: Toyota Urban Cruiser Scores 4-Star Rating In Global NCAP Crash Test                    |1.649750433E9|[positive]|finance,auto               |
|Tata Motors' Global Wholesales Grew 2 Per Cent In Q4 FY2022                                       |1.649747881E9|[positive]|finance,auto               |
|The New Age Baleno - Bold, Intelligent & Tech-Loaded                                              |1.649741E9   |[positive]|finance,auto               |
|Mahindra Offering Benefits Worth Upto Rs 70,000 in April 2022 - Details Here                      |1.649738894E9|[positive]|finance,auto               |
```


Activities
Terminal ▾
Tue 16:44
sidd0613@sidd0613-VivoBook: ~/parquet_based_code
File Edit View Search Terminal Help

```
|title                                                                                             |epoch        |sentiment |tags                       |
+--------------------------------------------------------------------------------------------------+-------------+----------+---------------------------+
|Dornier, the First Made-in-India Commercial Flight Takes Off on Dibrugarh-Pasighat Route          |1.649760907E9|[positive]|finance,auto               |
|EKA Partners With Log9 To Source Fast Charging Battery Solutions For Electric CVs                 |1.649755559E9|[positive]|finance,auto               |
|Domestic Air Passenger Traffic Up 59 Pct in FY22, 40 Pct Lower Than Pre-pandemic Era: Icra        |1.649755395E9|[negative]|finance,auto               |
|Tata Motors Group global wholesales at 3,34,884 units in Mar quarter of FY22                      |1.64975063E9 |[positive]|finance,auto,industry      |
|Exclusive: Toyota Urban Cruiser Scores 4-Star Rating In Global NCAP Crash Test                    |1.649750433E9|[positive]|finance,auto               |
|Tata Motors' Global Wholesales Grew 2 Per Cent In Q4 FY2022                                       |1.649747881E9|[positive]|finance,auto               |
|The New Age Baleno - Bold, Intelligent & Tech-Loaded                                              |1.649741E9   |[positive]|finance,auto               |
|Mahindra Offering Benefits Worth Upto Rs 70,000 in April 2022 - Details Here                      |1.649738894E9|[positive]|finance,auto               |
|Motherson closes in on acquisition of auto parts company Marelli                                  |1.649704172E9|[positive]|finance,auto,industry      |
|QE ‘inappropriate’ for South Africa - Sarb paper                                                  |1.649755265E9|[negative]|banking                    |
|LenDen Cub launches FinTech venture fund to help seed and pre-series startups                     |1.649752467E9|[positive]|banking                    |
|Rupee falls 14 paise to 76.05 against US dollar in early trade                                    |1.64975207E9 |[positive]|banking                    |
|Nepal's economic performance positive, not to worry: Finance minister                             |1.649751663E9|[positive]|banking                    |
|As cases dip, RBI advances trading hours of markets                                               |1.649740717E9|[positive]|banking                    |
|Trading volumes on top crypto exchanges touch 6-month low                                         |1.649740498E9|[negative]|banking                    |
|Startups, ecommerce and IT firms drive deal values in Q1 2022: report                             |1.649732566E9|[positive]|banking                    |
|Bond yields hit 3-year high post-RBI policy                                                       |1.649732455E9|[positive]|banking                    |
|NBFC recovery to bring higher Q4 profits; funding costs to rise, too                              |1.649732316E9|[positive]|banking                    |
|Moody's expects Indian banking sector to stabilise on economic rebound                            |1.649732176E9|[positive]|banking                    |
|PC Shipments Fall As Offices Back In Business Across The Globe                                    |1.649758667E9|[negative]|non-finance,technology     |
|U.S. fintech Plaid taps Booking.com veteran to lead its European expansion                        |1.649757874E9|[positive]|non-finance,technology     |
|Market Share of Top Centralised Exchanges Hit 96 percent in February 2022                         |1.649755896E9|[positive]|non-finance,tech           |
|Healthtech startup Ultrahuman acquires wearables firm LazyCo                                      |1.649755021E9|[positive]|finance,technology,startups|
|Healthtech startup Ultrahuman acquires wearables firm LazyCo                                      |1.649755021E9|[positive]|non-finance,technology     |
|LimeChat raises $4.2 million in funding led by Stellaris Venture Partners, others                 |1.649754638E9|[positive]|non-finance,technology     |
|Elon Musk, The Richest Man In The World, Is Now $100 Billion Richer Than Jeff Bezos               |1.649754599E9|[positive]|non-finance,technology     |
|Nokia Is Pulling Out of the Russian Market                                                        |1.649753401E9|[negative]|non-finance,tech           |
|Ethereum's Upcoming Software Reform Reaches Crucial Test Phase                                    |1.64975216E9 |[positive]|non-finance,tech           |
|Uber Hikes Trip Fares by 12 Percent in Delhi-NCR                                                  |1.649750552E9|[positive]|non-finance,tech           |
|Indian Oil has removed Russian Urals from its latest tender, sources say                          |1.649749153E9|[negative]|finance,energy,industry    |
|Giant undersea cables set to give the UK and Germany their first direct energy link               |1.649740322E9|[positive]|finance,energy             |
|As gas prices double, consumer pays for ‘nation building’, ONGC, RIL laugh all the way to the bank|1.64970186E9 |[positive]|non-finance,energy,Power   |
+--------------------------------------------------------------------------------------------------+-------------+----------+---------------------------+

data successfully written to mysql database!!!
(base) sidd0613@sidd0613-VivoBook:~/parquet_based_code$
```

DATA INTEGRATION



```
from datetime import datetime, timedelta
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.providers.apache.spark.operators.spark_submit import SparkSubmitOperator
from airflow.operators.dummy import DummyOperator

# Second approach for creating DAGs
dag1 = DAG(dag_id="project_final_dag",
           schedule_interval='0 0 * * *',
           start_date=datetime(2024, 2, 2),
           catchup=False,
           tags=['final_project', 'capstone_project', 'news_data_lake']
           )


taskA = BashOperator(task_id="NewsFetching",
                     bash_command="spark-submit --master local news_fetch.py {{ dag_run.conf['date'] }}",
                     cwd="/home/sidd0613/parquet_based_code",
                     dag=dag1)


taskB2 = BashOperator(task_id="NLPPipeline",
                     bash_command="spark-submit --packages com.johnsnowlabs.nlp:spark-nlp-spark32_2.12:3.4.2 nlp_pipeline.py {{
dag_run.conf['date'] }}",
                     cwd="/home/sidd0613/parquet_based_code",
                     dag=dag1)


taskC = BashOperator(task_id="QueryProcessing",
                     bash_command="spark-submit --master local --jars mysql-connector-java-8.0.26 queries.py {{
dag_run.conf['date'] }}",
                     cwd="/home/sidd0613/parquet_based_code",
                     dag=dag1)


taskA >> taskB2 >> taskC
```

7. CONCLUSION

In conclusion, we provide a sentiment analysis on the basis of the news data lake, whether the market is performing well or not on the basis of positive and negative sentiments. First of all the news which are fetched from various sources are given to nlp pipeline and afterwards the nlp pipeline provides the sentiment for it. Now we have all data and sentiment of it so we have performed some queries on top of available data ,which can help user to find domain specific, overall and latest sentiment of market. Later which can be shown with the help of powerBi/tableau dashboard.

8. FUTURE SCOPE

In the future scope of this project we can include following features :

1) Project can be scalable for future
2) Delta format can be used to have some added features such as time travel
3) More news sources such as rss feeds, api's can be added
4) End to end project can be created on cloud

9. References

1. https://www.johnsnowlabs.com/
2. https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining
3. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
4. https://docs.astronomer.io/learn/intro-to-airflow