HOME ASSIGNMENT 4

Roll No: 208W1A1299

Name: MOHAMMAD RIZWANULLAH

Aim: To do analysis using Apache spark in data bricks.

Execution:

1)Open databricks and create a new cluster with latest configuration version.

2) Now create a Notebook to work with spark.

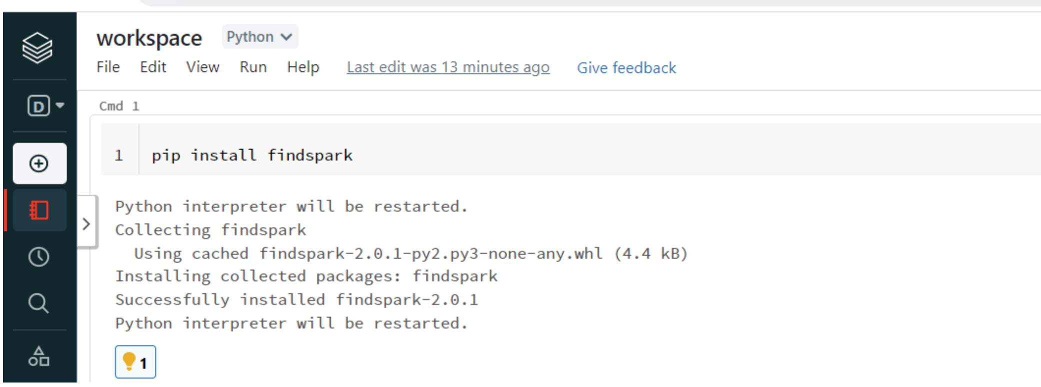
3) Import csv file or any file for using it and do analysis using pyspark.

4) Now write code to import dataset and create some RDD to read input in spark.

5) Run the cell.

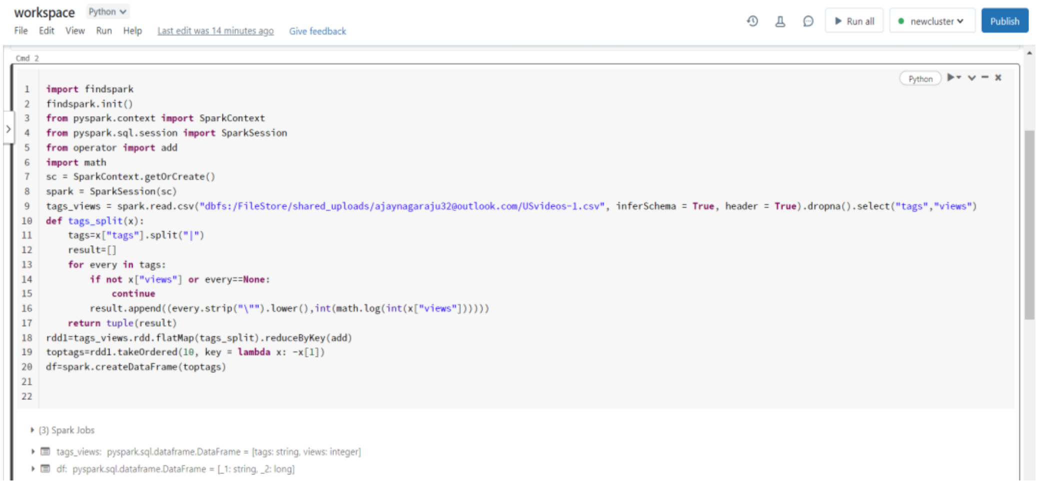
Below are the screenshots of execution:

Twitter followers count and protected data analysis.



```
pip install findspark
```

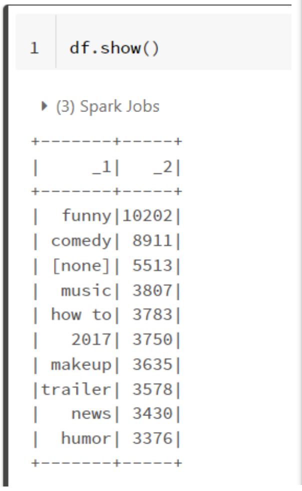
```
Python interpreter will be restarted.
Collecting findspark
  Using cached findspark-2.0.1-py2.py3-none-any.whl (4.4 kB)
Installing collected packages: findspark
Successfully installed findspark-2.0.1
Python interpreter will be restarted.
```



```
import findspark
findspark.init()
from pyspark.context import SparkContext
from pyspark.sql.session import SparkSession
from operator import add
import math
sc = SparkContext.getOrCreate()
spark = SparkSession(sc)
tags_views = spark.read.csv("dbfs:/FileStore/shared_uploads/ajaynagaraju32@outlook.com/USvideos-1.csv", inferSchema = True, header = True).dropna().select("tags","views")
def tags_split(x):
    tags=x["tags"].split("|")
    result=[]
    for every in tags:
        if not x["views"] or every==None:
            continue
        result.append((every.strip("\"").lower(),int(math.log(int(x["views"])))))
    return tuple(result)
rdd1=tags_views.rdd.flatMap(tags_split).reduceByKey(add)
toptags=rdd1.takeOrdered(10, key = lambda x: -x[1])
df=spark.createDataFrame(toptags)


```

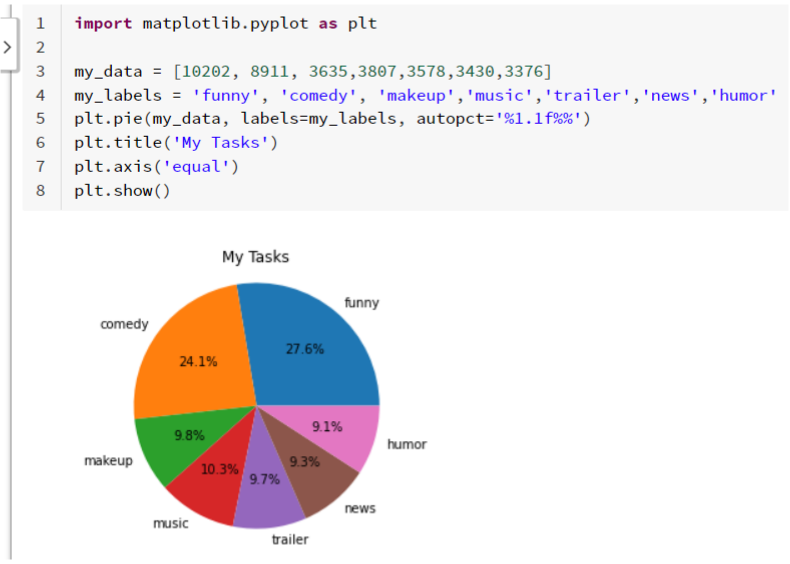
```
(3) Spark Jobs
tags_views:  pyspark.sql.dataframe.DataFrame = [tags: string, views: integer]
df:  pyspark.sql.dataframe.DataFrame = [_1: string, _2: long]
```



```
df.show()
```

▸ (3) Spark Jobs

```
+-------+-----+
|     _1|   _2|
+-------+-----+
|  funny|10202|
| comedy| 8911|
| [none]| 5513|
|  music| 3807|
| how to| 3783|
|   2017| 3750|
| makeup| 3635|
|trailer| 3578|
|   news| 3430|
|  humor| 3376|
+-------+-----+
```



```
import matplotlib.pyplot as plt

my_data = [10202, 8911, 3635,3807,3578,3430,3376]
my_labels = 'funny', 'comedy', 'makeup','music','trailer','news','humor'
plt.pie(my_data, labels=my_labels, autopct='%1.1f%%')
plt.title('My Tasks')
plt.axis('equal')
plt.show()
```