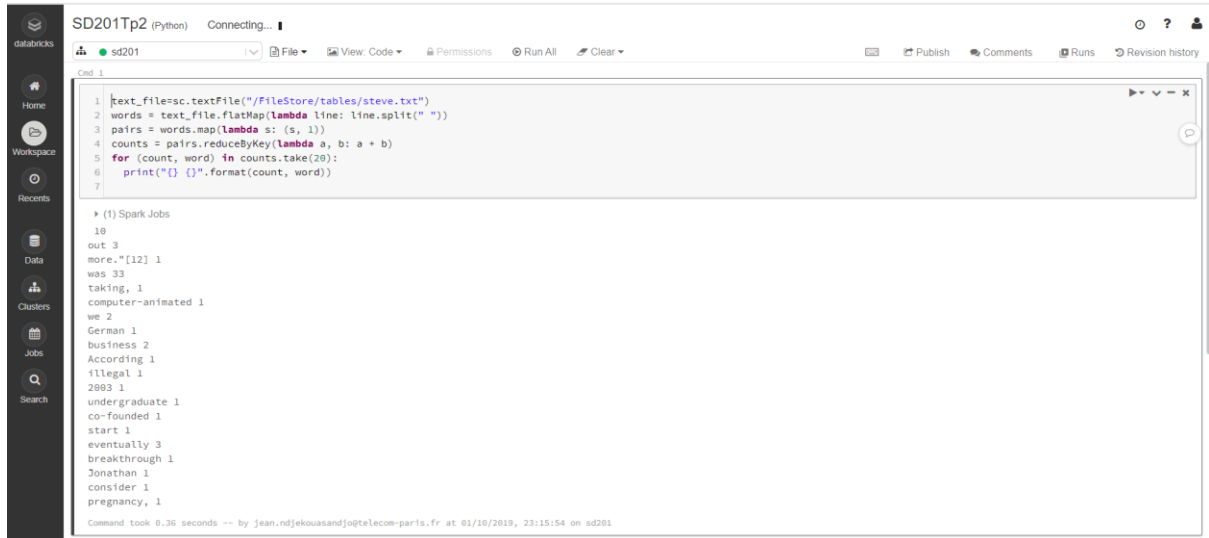


REPORT SD201 TP2

1) Question 1:



The screenshot shows the Databricks SD201Tp2 workspace. The top bar indicates the workspace is connecting. The left sidebar contains navigation icons for Home, Workspace, Recents, Data, Clusters, Jobs, and Search. The main area displays a code editor with a Python script and its output. The script reads a text file, splits it into words, and counts the frequency of each word. The output shows the top 20 words and their counts.

```
1 text_file = sc.textFile("/FileStore/tables/steve.txt")
2 words = text_file.flatMap(lambda line: line.split(" "))
3 pairs = words.map(lambda s: (s, 1))
4 counts = pairs.reduceByKey(lambda a, b: a + b)
5 for (count, word) in counts.take(20):
6     print("{} {}".format(count, word))
7
```



Output:






```
(1) Spark Jobs
10
out 3
more.[12] 1
was 33
taking, 1
computer-animated 1
we 2
German 1
business 2
According 1
illegal 1
2003 1
undergraduate 1
co-founded 1
start 1
eventually 3
breakthrough 1
Jonathan 1
consider 1
pregnancy, 1
```

Command took 0.36 seconds -- by jean.ndjekoussandjo@telecom-paris.fr at 01/10/2019, 23:15:54 on sd201

#1) modified c into sc and defined the variable text_file which stores the lines of the text
#2) modified the print format because the previous one was giving some errors
#3) modified the map from (s,2) to (s,1)

2) Question 2:

SD201Tp2 (Python) Connecting...  

sd201     

Cmd 1

```
1 |text_file=sc.textFile("/FileStore/tables/steve.txt")
2 |words = text_file.flatMap(Lambda line: line.split(" "))
3 |pairs = words.map(Lambda s: (s, 1))
4 |#counts = pairs.reduceByKey(Lambda a, b: a + b).takeOrdered(5, lambda pair: pair[1])
5 |counts = pairs.reduceByKey(Lambda a, b: a + b).sortBy(Lambda a: -a[1])
6 |counts.take(5)
7 |#for (count, word) in counts.take(10):
8 |# print("{} {}".format(count, word))
9 |
```

▶ (3) Spark Jobs

Out[34]: [('the', 66), ('and', 53), ('a', 45), ('to', 42), ('in', 41)]

Command took 1.50 seconds -- by jean.ndjekouasandjo@telecom-paris.fr at 01/10/2019, 23:54:27 on sd201

Cmd 2


1

Shift+Enter to run [shortcuts](#)

22:36 CRLF UT

3) Question 3:

SD201Tp2 (Python)  

sd201   View: Code  Permissions  Run All 

8 |counts.take(5)

▶ (3) Spark Jobs

Out[126]: [('Apple', 11), ('Jandalf', 8), ('Schieble', 8), ('Clara', 8), ('Jobs's', 8)]



Command took 1.28 seconds -- by jean.ndjekouasandjo@telecom-paris.fr at 02/10/2019, 12:52:39 on sd201






Cmd 2

1

Shift+Enter to run [shortcuts](#)

4) Question 4:

SD201Tp2 (Python)  

sd201   View: Code  Permissions  Run All 

1 |
2 |

▶ (7) Spark Jobs

Out[124]: [('United States', 8145), ('France', 7799), ('Communes of France', 5740), ('Departments of France', 5299), ('Regions of France', 4864), ('City', 3832), ('Romania', 3527), ('Category:Rivers in Romania', 2978), ('Tributary', 2799), ('England', 2277)]

Command took 11.48 seconds -- by jean.ndjekouasandjo@telecom-paris.fr at 02/10/2019, 12:50:07 on sd201

Cmd 2

1

