DA228 – Big Data tech and App
Assignment-3

The map function: for each t in R, produce key-value pair (t, R), and for each t in S, produce key-value pair (t, S).

The reduce function: for each key t, if the associated value list is [R] but not in [S}, then produce (t, t), otherwise, produce nothing.

Contents of file R.txt :

```
1
1
2
3

5
3
```

File s.txt:

```
2
3
4
```

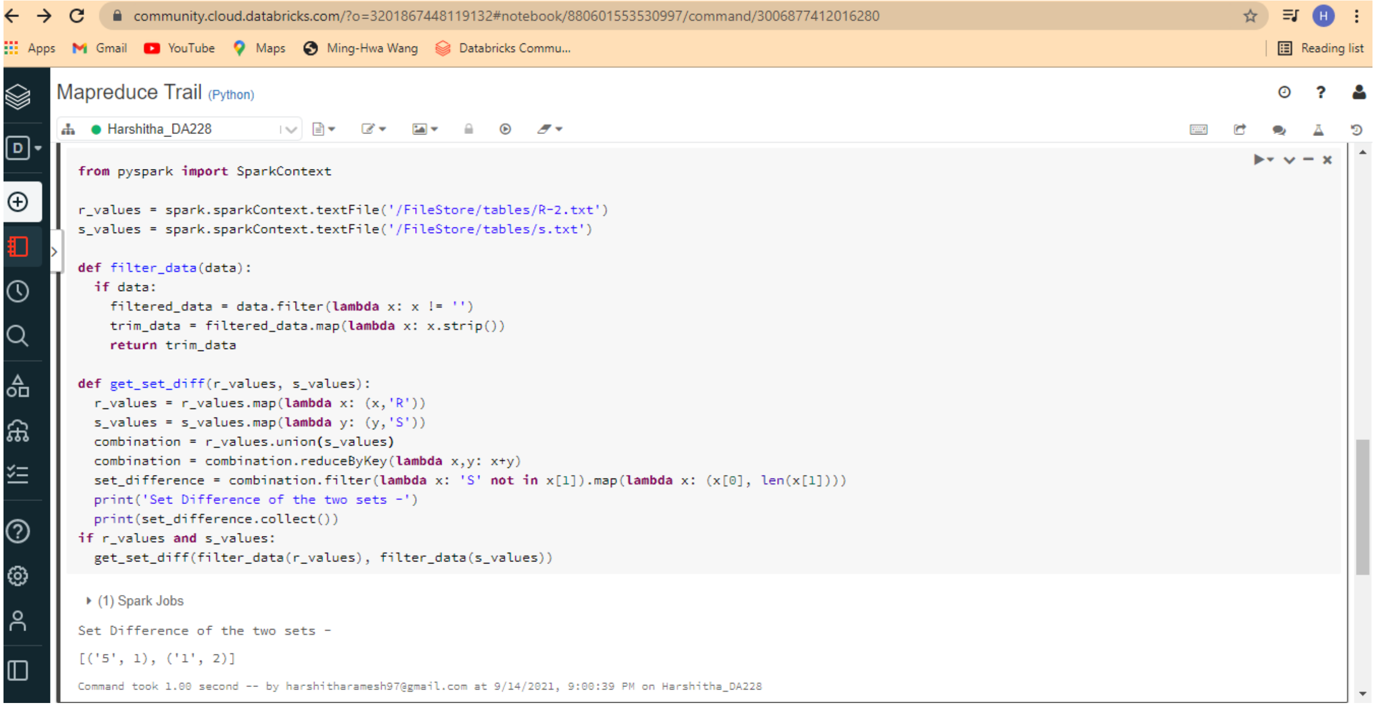
1. Code

```
from pyspark import SparkContext

r_values = spark.sparkContext.textFile('/FileStore/tables/R-2.txt')
s_values = spark.sparkContext.textFile('/FileStore/tables/s.txt')

def filter_data(data):
  if data:
    filtered_data = data.filter(lambda x: x != '')
    trim_data = filtered_data.map(lambda x: x.strip())
    return trim_data
```

DA228 – Big Data tech and App
Assignment-3

```
def get_set_diff(r_values, s_values):
  r_values = r_values.map(lambda x: (x,'R'))
  s_values = s_values.map(lambda y: (y,'S'))
  combination = r_values.union(s_values)
  combination = combination.reduceByKey(lambda x,y: x+y)
  set_difference = combination.filter(lambda x: 'S' not in x[1]).map(lambda x: (x[0], len(x[1])))
  print('Set Difference of the two sets -')
  print(set_difference.collect())
if r_values and s_values:
  get_set_diff(filter_data(r_values), filter_data(s_values))
```

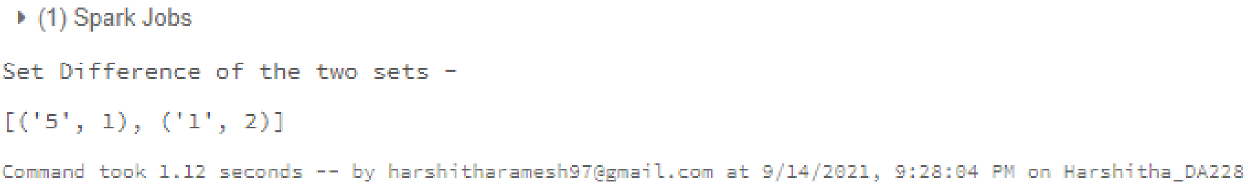
2. Execution screenshot



```
from pyspark import SparkContext

r_values = spark.sparkContext.textFile('/FileStore/tables/R-2.txt')
s_values = spark.sparkContext.textFile('/FileStore/tables/s.txt')

def filter_data(data):
  if data:
    filtered_data = data.filter(lambda x: x != '')
    trim_data = filtered_data.map(lambda x: x.strip())
    return trim_data

def get_set_diff(r_values, s_values):
  r_values = r_values.map(lambda x: (x,'R'))
  s_values = s_values.map(lambda y: (y,'S'))
  combination = r_values.union(s_values)
  combination = combination.reduceByKey(lambda x,y: x+y)
  set_difference = combination.filter(lambda x: 'S' not in x[1]).map(lambda x: (x[0], len(x[1])))
  print('Set Difference of the two sets -')
  print(set_difference.collect())
if r_values and s_values:
  get_set_diff(filter_data(r_values), filter_data(s_values))
```

```
(1) Spark Jobs

Set Difference of the two sets -

[('5', 1), ('1', 2)]

Command took 1.00 second -- by harshitharamesh97@gmail.com at 9/14/2021, 9:00:39 PM on Harshitha_DA228
```

3. Output of Execution



```
Set Difference of the two sets -

[('5', 1), ('1', 2)]

Command took 1.12 seconds -- by harshitharamesh97@gmail.com at 9/14/2021, 9:28:04 PM on Harshitha_DA228
```