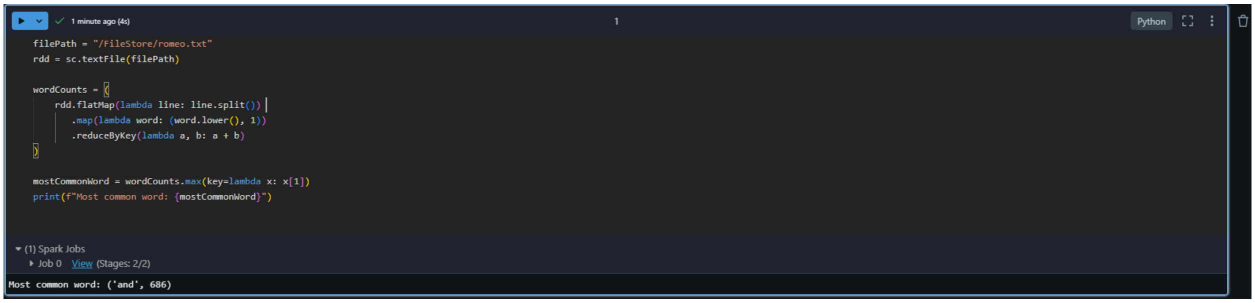
CP1370
Lab 2
John-Michael Woodrow

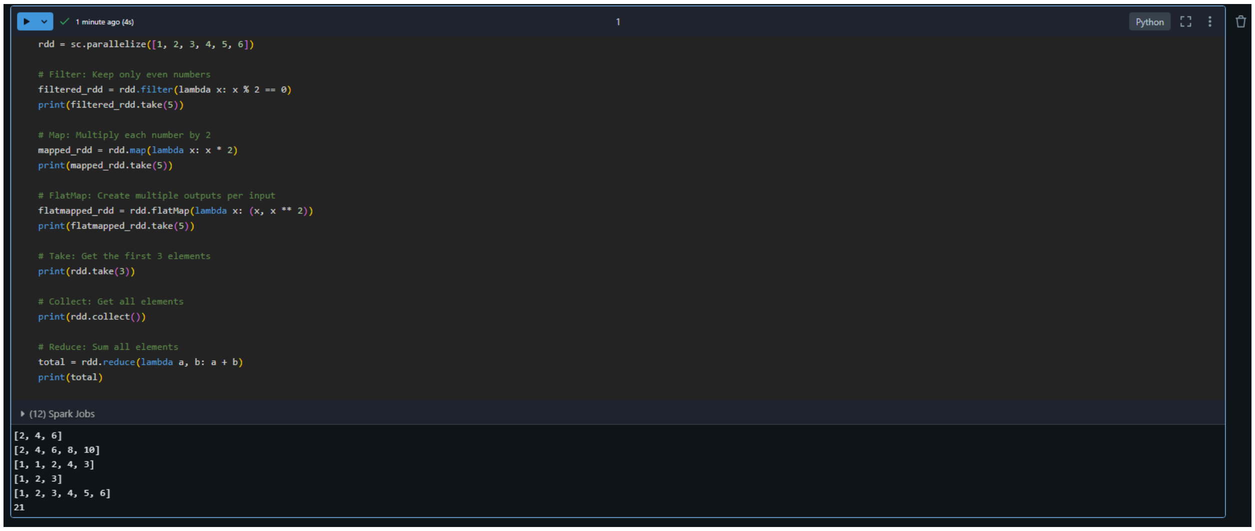
1.



```
filePath = "/FileStore/romeo.txt"
rdd = sc.textFile(filePath)

wordCounts = (
    rdd.flatMap(lambda line: line.split())
       .map(lambda word: (word.lower(), 1))
       .reduceByKey(lambda a, b: a + b)
)

mostCommonWord = wordCounts.max(key=lambda x: x[1])
print(f"Most common word: {mostCommonWord}")
```

```
Most common word: ('and', 686)
```

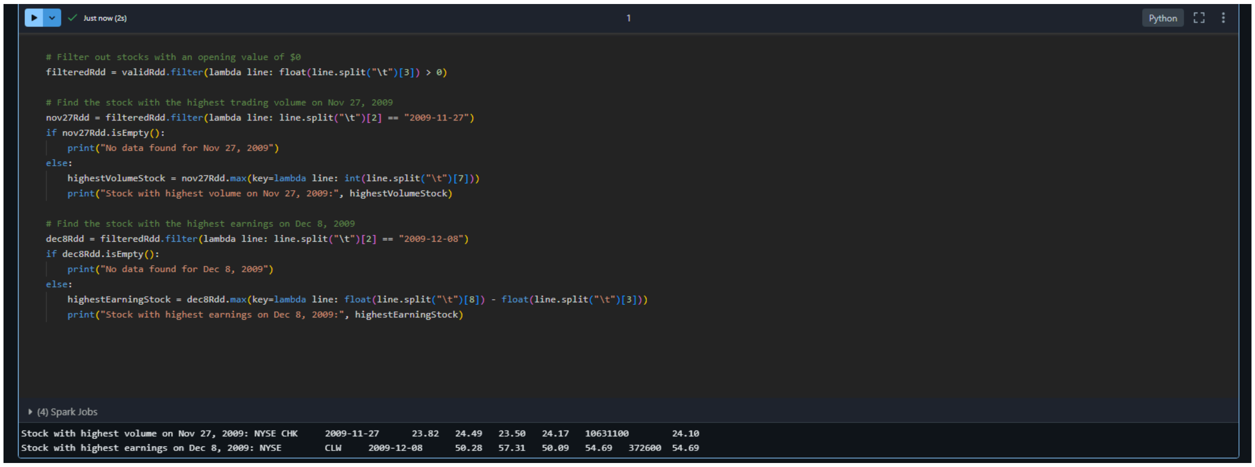
2.



```
rdd = sc.parallelize([1, 2, 3, 4, 5, 6])

# Filter: Keep only even numbers
filtered_rdd = rdd.filter(lambda x: x % 2 == 0)
print(filtered_rdd.take(5))

# Map: Multiply each number by 2
mapped_rdd = rdd.map(lambda x: x * 2)
print(mapped_rdd.take(5))

# FlatMap: Create multiple outputs per input
flatmapped_rdd = rdd.flatMap(lambda x: (x, x ** 2))
print(flatmapped_rdd.take(5))

# Take: Get the first 3 elements
print(rdd.take(3))

# Collect: Get all elements
print(rdd.collect())

# Reduce: Sum all elements
total = rdd.reduce(lambda a, b: a + b)
print(total)
```

```
[2, 4, 6]
[2, 4, 6, 8, 10]
[1, 1, 2, 4, 3]
[1, 2, 3]
[1, 2, 3, 4, 5, 6]
21
```

3.



```
# Filter out stocks with an opening value of $0
filteredRdd = validRdd.filter(lambda line: float(line.split("\t")[3]) > 0)

# Find the stock with the highest trading volume on Nov 27, 2009
nov27Rdd = filteredRdd.filter(lambda line: line.split("\t")[2] == "2009-11-27")
if nov27Rdd.isEmpty():
    print("No data found for Nov 27, 2009")
else:
    highestVolumeStock = nov27Rdd.max(key=lambda line: int(line.split("\t")[7]))
    print("Stock with highest volume on Nov 27, 2009:", highestVolumeStock)

# Find the stock with the highest earnings on Dec 8, 2009
dec8Rdd = filteredRdd.filter(lambda line: line.split("\t")[2] == "2009-12-08")
if dec8Rdd.isEmpty():
    print("No data found for Dec 8, 2009")
else:
    highestEarningStock = dec8Rdd.max(key=lambda line: float(line.split("\t")[8]) - float(line.split("\t")[3]))
    print("Stock with highest earnings on Dec 8, 2009:", highestEarningStock)
```

```
Stock with highest volume on Nov 27, 2009: NYSE CHK     2009-11-27      23.82   24.49   23.50   24.17   10631100        24.10
Stock with highest earnings on Dec 8, 2009: NYSE        CLW     2009-12-08      50.28   57.31   50.09   54.69   372600  54.69
```