

Lab - 10

1. Using RDD & flatmap count how many times each word appears in a file & write out a list of words whose count is strictly greater than the using spark

word count.py

```
> from pyspark import SparkContext
> sc = SparkContext("file1.txt")
counts = text_file.flatMap(lambda line:
    line.split()) \
    .map(lambda word: (word, 1)) \
    .reduceByKey(lambda a, b: a + b)
```



```
Output = counts.collect()
for (word, count) in output:
    print(f"{word} : {count}")
```

Output:

hello : 2

bmsee : 3

how : 1

are : 1

you : 1

file1.txt

hello bmsee,

bmsee how are you

bmsee


```
1 from pyspark import SparkContext
2
3 sc = SparkContext("local", "SimpleWordCount")
4 text_file = sc.textFile("file1.txt")
5 counts = text_file.flatMap(lambda line: line.split()) \
6                        .map(lambda word: (word, 1)) \
7                        .reduceByKey(lambda a, b: a + b)
8 output = counts.collect()
9
10 for (word, count) in output:
11     print(f"{word}: {count}")
12
```

hello: 2

bmsce: 3

how: 1

are: 1

you: 1