Sparks A5

February 11, 2020

[1]:
```
from pyspark import SparkContext
from pyspark.sql import SparkSession
sc = SparkContext(appName = "wordcount")
spark = SparkSession.Builder().getOrCreate()
text_file = sc.textFile("sk.txt") \
            .map( lambda x: x.replace(',',' ').replace('.',' ').replace('-','␣
 ↪').lower()) \
# line numbers
text_file.count()
```

[1]:
```
124456
```

[2]:
```
counts = text_file.flatMap(lambda line: line.split(" ")) \
             .map(lambda word: (word, 1)) \
             .reduceByKey(lambda a, b: a + b)
sorted_counts = counts.sortBy(lambda wordCounts: wordCounts[1], ascending=False)
# the #24 most used word in Shakespeares writings
# the first one is not a word
i = 0
for word, count in sorted_counts.collect()[0:25]:
    print("{} : {} : {} ".format(i, word, count))
    i += 1
```

```
0 :  : 680273
1 : the : 27572
2 : and : 26752
3 : i : 20191
4 : to : 19338
5 : of : 18135
6 : a : 14520
7 : you : 12991
8 : my : 12468
9 : that : 10964
10 : in : 10914
11 : is : 9503
12 : not : 8453
13 : for : 8215
14 : with : 7973
```

```
15 : it : 7224
16 : be : 6979
17 : me : 6962
18 : your : 6875
19 : his : 6825
20 : this : 6299
21 : but : 6272
22 : he : 6102
23 : as : 5934
24 : have : 5845
```

[ ]: