A5-Spark+Exercise

February 8, 2020

1 A5-Spark Exercise

Ilia Ozhmegov

1.1 Import Required Libraries

[1]:
```
from pyspark import SparkConf
from pyspark import SparkContext
```

1.2 Word Count

[2]:
```
# load the text file
text_file = sc.textFile("/shared/t8.shakespeare.txt")
```

[3]:
```
counts = (text_file.flatMap(lambda line: line.lower().split())
                   .map(lambda key: (key, 1))
                   .reduceByKey(lambda i, j: i + j))
```

[4]:
```
# top 25 words
top = counts.takeOrdered(25, key=lambda x: -x[1])
[print(i+1, item, sep=": ") for i,item in enumerate(top)]
```

```
4: ('to', 18656)
5: ('of', 17958)1: ('the', 27486)
2: ('and', 25994)
3: ('i', 19540)
6: ('a', 14365)
7: ('my', 12455)
8: ('in', 10660)
9: ('that', 10474)
11: : ('you', 10597)
10('is', 9117)12: ('for', 7951)
13: ('with', 7924)
14: ('not', 7634)
15: ('your', 6862)
16: ('his', 6749)
17: ('be', 6685)
18: ('he', 5884)
```

```
19: ('but', 5881)
20: ('as', 5876)
21: ('this', 5860)
22: ('it', 5858)
23: ('have', 5675)
24: ('thou', 5138)
25: ('me', 4851)
```

[4]:
```
[None, None, None, None, None, None, None, None, None, None, None, None, None,
 None, None, None, None, None, None, None, None, None, None, None, None]
```

1.3 1.The 24th most used word in Shakespeare's writings

[5]:
```
print(top[23])
```

```
('thou', 5138)
```

1.4 2. What installation did you worked with? Where there any obstacles?

Firstly, I have tried to launch all natively (`brew install *`) or (`wget`, `tar`, `mv` , **changing path environment variables**) on my OSX machine and I have succeeded. To be more accurate I have launched `pyspark` and `spark-shell`, but when I decided to test it with a real-world task from a tutorial about `pyspark` an errror came up that meant I have the wrong JVM, ok I uninstalled jvm13 and installed jvm8 (according to the stackoverflow), but I didn't change path environment variables, because osx has diffrent pathes compared with any linux. It was the last drop. So I decided to use vagrant.

So I used vagrant, because it was extremely easy to find a box, which already has all essential and compatible software. And it was a way faster, than installing everything from a scratch on a less friendly os than linux.

```
$ vagrant init fscm/spark-jupyter
```

Then, as usual

```
$ vagrant up; vagrant ssh
```

The last command to place all created files in a shared folder. Eventually, I opened jupyter notebook in a web-broser and finished the task.

1.5 Conclusion 1: nothing but linux

1.6 Conclusion 2: why should I every time install everything from a scratch if we have Vagrant

1.7 Conclusion 3: Apache Spark is a really good technology that allows to reduce computational time significantly, because of multithreading even if you use it on a standalone machine

The file after deleting the heading has 124212 lines, it took 6.48 seconds on my machine to find top 25 of the most used words. That is amazing! I will definitely continue to learn this technology.

[ ]: