Problem03_Final

September 26, 2019

[26]:

```
################################################################################
#
# Ulysses Carlos -- 09/25/2019
# Problem 03
# Using the PostHistory file, count the number of Posts that feature the words
# Spark" and Scala". Output this to the screen. (20 points)
#
################################################################################

import re

# Stop SparkContext
sc.stop()
from pyspark import SparkConf, SparkContext
conf = SparkConf().setMaster("local").setAppName("My App")
sc = SparkContext(conf = conf)


xml_file = sc.textFile("PostHistory.xml")

file_count = xml_file.count()
print ("Lines in XML file: " + str(file_count))

#Define function:
def apply_regex(regex1, line):
    test1 = re.search(regex1, line, re.IGNORECASE)
    if test1:
        return str(line)
    else:
        return str("<USER>")


#End

test_spark = xml_file.map(lambda line : apply_regex("spark", str(line)) )
```

```python

test_scala = xml_file.map(lambda line : apply_regex("scala", str(line)))

#Expensive process
filter_out = test_spark.filter(lambda line: "<USER>" in line)
test_spark = test_spark.subtract(filter_out)
print ("Lines that contain the word \"Spark\" : " + str(test_spark.count()))

filter_out = test_scala.filter(lambda line: "<USER>" in line)
test_scala = test_scala.subtract(filter_out)
print ("Lines that contain the word \"Scala\" : " + str(test_scala.count()))



end = test_spark.union(test_scala)
end = end.distinct()

print ("The number of posts that feature the words \"Spark\" or \"Scala\" is "␣
 ↪+ str(end.count()) + ".\n")
```

```
Lines in XML file: 121525
Lines that contain the word "Spark" : 1341
Lines that contain the word "Scala" : 968
The number of posts that feature the words "Spark" or "Scala" is 2098.
```

[ ]:

[ ]: