Problem01_Final

September 26, 2019

[6]:

```
################################################################################
#
# Ulysses Carlos 09/23/2019
# Problem 1:
# From the Users.xml file, find all users which are from Georgia and output to
# screen their DisplayName only. (20 points)
#
################################################################################


import re

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



# Store the contents of the file into a rdd
lines = sc.textFile("Users.xml");



# Filter through GA or Georgia
# Regex Expression Location=[\"]\w+, GA,*

#Define function
def apply_regex(regex, line):
    x = re.search("Location=[\"]\w+, GA,*", line)
    if (x):
        return str(line)
    else:
        return str("<USER IS NOT FROM GEORGIA>")
```

```
#end

print ("Database line count: " + str(lines.count()))
pattern = "Location=[\"]\w+, GA,*"
filter_01 = lines.map(lambda line : apply_regex(pattern, str(line)))

# Now remove all Non-Georgian Locations
filter_out = filter_01.filter(lambda line : "<USER IS NOT FROM GEORGIA>" in␣
 ↪line)

filter_02 = filter_01.subtract(filter_out)

print ("Size of regex filter: " + str(filter_02.count()) + "\n")

# Define Function
def print_name(line):
    low = line.rfind("DisplayName")
    high = line.rfind("LastAccessDate")

    if low != -1 and high != -1:
        substring = line[low : high]
        return substring

#end

end = filter_02.map(lambda line : print_name(line))
print ("The List of Users from Georgia is as follows:\n")
print (end.take(end.count()))
```

```
Database line count: 66954
Size of regex filter: 139

The List of Users from Georgia is as follows:

['DisplayName="Nick Larsen" ', 'DisplayName="Michael" ', 'DisplayName="azoorob"
', 'DisplayName="ilya" ', 'DisplayName="tempusfugit" ', 'DisplayName="Peter
Woolfitt" ', 'DisplayName="YC Hu" ', 'DisplayName="Patrick Gerbes" ',
'DisplayName="Ilya Lapitan" ', 'DisplayName="Teresa Madsen" ',
'DisplayName="Brandon" ', 'DisplayName="Mr. Rooter of Savannah" ',
'DisplayName="Mr. Rooter of Southeast GA" ', 'DisplayName="Khiem Ha" ',
'DisplayName="Mac18" ', 'DisplayName="Tarun Luthra" ', 'DisplayName="Todd
Dawson" ', 'DisplayName="David F" ', 'DisplayName="PSInf" ',
'DisplayName="Chirag" ', 'DisplayName="hellofanengineer" ', 'DisplayName="Oriol
Mirosa" ', 'DisplayName="David" ', 'DisplayName="Vincent" ',
'DisplayName="cbarrick" ', 'DisplayName="Len Greski" ', 'DisplayName="PEBKAC" ',
'DisplayName="Bryce" ', 'DisplayName="donlan" ', 'DisplayName="BarclayK" ',
'DisplayName="dportman" ', 'DisplayName="Shishir Suman" ', 'DisplayName="Scott"
', 'DisplayName="JessicaRabi" ', 'DisplayName="wgreenihrcorp" ',
'DisplayName="zongyan" ', 'DisplayName="Alexandre" ', 'DisplayName="Erica Rosa"
', 'DisplayName="empoleon" ', 'DisplayName="Deepak Shenoy" ',
'DisplayName="Ragnar Lothbrok" ', 'DisplayName="Deontaé Le Pew" ',
'DisplayName="Doctorambient" ', 'DisplayName="Vidya" ', 'DisplayName="Vatsal
Srivastava" ', 'DisplayName="Dean Webb" ', 'DisplayName="wayne green" ',
'DisplayName="Floydian" ', 'DisplayName="Manikanta Reddy D" ',
'DisplayName="devarsh raghnathbhai patel" ', 'DisplayName="rmooney" ',
'DisplayName="mike_stevs" ', 'DisplayName="Gary Lai" ', 'DisplayName="Daniel" ',
'DisplayName="phos" ', 'DisplayName="Lance Ruo Zhang" ', 'DisplayName="Spencer-
Price" ', 'DisplayName="Stephen Ewing" ', 'DisplayName="addi wei" ',
'DisplayName="Alex V" ', 'DisplayName="Rob" ', 'DisplayName="Ram" ',
'DisplayName="jGaboardi" ', 'DisplayName="ps0604" ', 'DisplayName="Tiji Mathew"
', 'DisplayName="Tony Boyles" ', 'DisplayName="pkerl" ', 'DisplayName="gfritz"
', 'DisplayName="Aleksandr Blekh" ', 'DisplayName="ontek" ',
'DisplayName="Aravind R. Yarram" ', 'DisplayName="Henry Crutcher" ',
'DisplayName="Goddard" ', 'DisplayName="Matt Simpson" ', 'DisplayName="matt
biskup" ', 'DisplayName="Jason W" ', 'DisplayName="Peter Mourfield" ',
'DisplayName="Magsol" ', 'DisplayName="Bob Baxley" ', 'DisplayName="badjr" ',
'DisplayName="mplunney" ', 'DisplayName="ryan" ', 'DisplayName="pradyumnad" ',
'DisplayName="Psidom" ', 'DisplayName="jpm" ', 'DisplayName="Ahmet Cecen" ',
'DisplayName="Guy Gordon" ', 'DisplayName="C3Theo" ', 'DisplayName="niru dyogi"
', 'DisplayName="Vinitha Palani" ', 'DisplayName="Andrew" ',
'DisplayName="Aditya Gogoi" ', 'DisplayName="turtlemonvh" ', 'DisplayName="Lewis
Rodgers" ', 'DisplayName="Devendra Lattu" ', 'DisplayName="cosmosa" ',
'DisplayName="Mboolean" ', 'DisplayName="Jimd" ', 'DisplayName="Sandeep Gunda"
', 'DisplayName="Will Gao" ', 'DisplayName="Andrew King" ',
'DisplayName="rajb245" ', 'DisplayName="Sealander" ', 'DisplayName="afshin" ',
'DisplayName="Ashish Powani" ', 'DisplayName="Boris N." ', 'DisplayName="Atul
Kaushik" ', 'DisplayName="Harnoor Singh" ', 'DisplayName="Tiago Cogumbreiro" ',
'DisplayName="red_eight" ', 'DisplayName="Christoph" ', 'DisplayName="David
Hofmann" ', 'DisplayName="nburn42" ', 'DisplayName="Nick M" ',
'DisplayName="Kiran" ', 'DisplayName="Zer0k" ', 'DisplayName="Baxter" ',
'DisplayName="Rama Ananda" ', 'DisplayName="Dr. Strange" ', 'DisplayName="Shahan
M" ', 'DisplayName="rams" ', 'DisplayName="John Barbour" ',
'DisplayName="The_Flin" ', 'DisplayName="DaulPavid" ', 'DisplayName="Amandeep
Jiddewar" ', 'DisplayName="user1410665" ', 'DisplayName="Willie-G" ',
'DisplayName="Sam Washburn" ', 'DisplayName="Loisaida Sam Sandberg" ',
'DisplayName="gnorman" ', 'DisplayName="Meng Zhao" ', 'DisplayName="Myron Slaw"
', 'DisplayName="ajroot" ', 'DisplayName="zachdj" ', 'DisplayName="David Zhou"
', 'DisplayName="treysp" ', 'DisplayName="Eduardo Trunci" ',
'DisplayName="Squ1rr3lz" ', 'DisplayName="user3776637" ']
```

[ ]: