lab3 Python

lab3cluster

Cmd 1

Python

```
## Krista Miller

#first make directory to store files.
dbutils.fs.mkdirs("/FileStore/tables/lab3short")

#move or copy the files to that location with the mv or cp command.  write python code with a loop that creates the filePaths for the given files
fList =dbutils.fs.ls("/FileStore/tables/")

for i in range(0, len(fList)):
    if "shortLab" in fList[i].name:
        dbutils.fs.cp("/FileStore/tables/"+fList[i].name, "/FileStore/tables/lab3short/"+fList[i].name)

#confirm the files are in correct directory:
display(dbutils.fs.ls('/FileStore/tables/lab3short'))
```

(3) Spark Jobs

	path	name	size
1	dbfs:/FileStore/tables/lab3short/shortLab3data0.txt	shortLab3data0.txt	577
2	dbfs:/FileStore/tables/lab3short/shortLab3data1.txt	shortLab3data1.txt	527

Showing all 2 rows.

Cmd 2

```
sc = spark.sparkContext

#can pass in a directory name rather than a specific file name:
file = sc.textFile("/FileStore/tables/lab3short/")

#pair each target URL as the first item in the pair with the URL of a page which links it as a second element:
urls = file.map(lambda line: (line.split(" ")[0], line.split(" ")[1:]))

#mapValues to remove duplicates from value list and sort it:
def f(x): return sorted(list(set(x)))
urls2 = urls.mapValues(f)

#flatMapValues to return pair/value tuples of (target, reference):
def g(x): return x
urls3 = urls2.flatMapValues(g)

#map to flip target and references:
rddInverted = urls3.map(lambda x: (x[1], x[0]))

#group and sort by key:
final = rddInverted.groupByKey().mapValues(list).sortByKey(ascending = True)
```

```
#sort by value:
finalSorted = final.mapValues(f)

#output for small file:
print(finalSorted.collect())
```

(3) Spark Jobs

```
[('www.example1.com', ['www.example14.com', 'www.example16.com', 'www.example3.com', 'www.example4.com', 'www.example7.com', 'www.example9.com']),
('www.example10.com', ['www.example20.com']), ('www.example11.com', ['www.example8.com']), ('www.example12.com', ['www.example19.com', 'www.example4.com']), ('www.example13.com', ['www.example5.com']), ('www.example14.com', ['www.example10.com', 'www.example11.com', 'www.example3.com', 'www.example7.com']), ('www.example15.com', ['www.example1.com', 'www.example18.com']), ('www.example16.com', ['www.example12.com']), ('www.example18.com', ['www.example15.com', 'www.example2.com', 'www.example4.com', 'www.example7.com']), ('www.example19.com', ['www.example2.com', 'www.example5.com']), ('www.example2.com', ['www.example3.com']), ('www.example20.com', ['www.example14.com', 'www.example17.com', 'www.example5.com']), ('www.example4.com', ['www.example16.com', 'www.example17.com', 'www.example6.com']), ('www.example5.com', ['www.example1.com', 'www.example20.com', 'www.example3.com']), ('www.example6.com', ['www.example12.com', 'www.example13.com']), ('www.example7.com', ['www.example12.com', 'www.example16.com']), ('www.example8.com', ['www.example12.com', 'www.example4.com'])]
```

Command took 0.88 seconds -- by krista.miller445@du.edu at 4/13/2022, 11:57:19 AM on lab3cluster

Cmd 3

```
#Now running for large file:

#first make directory to store files.
dbutils.fs.mkdirs("/FileStore/tables/lab3full")

#move or copy the files to that location with the mv or cp command.  write python code with a loop that creates the filePaths for the given files
fList =dbutils.fs.ls("/FileStore/tables/")

for i in range(0, len(fList)):
    if "fullLab" in fList[i].name:
        dbutils.fs.cp("/FileStore/tables/"+fList[i].name, "/FileStore/tables/lab3full/"+fList[i].name)

#confirm the files are in correct directory:
display(dbutils.fs.ls('/FileStore/tables/lab3full'))
```

▸ (3) Spark Jobs

	path	name	size
1	dbfs:/FileStore/tables/lab3full/fullLab3data0.txt	fullLab3data0.txt	3347768
2	dbfs:/FileStore/tables/lab3full/fullLab3data1.txt	fullLab3data1.txt	3405715
3	dbfs:/FileStore/tables/lab3full/fullLab3data2.txt	fullLab3data2.txt	3418525
4	dbfs:/FileStore/tables/lab3full/fullLab3data3.txt	fullLab3data3.txt	3386848

Cmd 4

Python

```
sc = spark.sparkContext

#can pass in a directory name rather than a specific file name:
file = sc.textFile("/FileStore/tables/lab3full/")

#pair each target URL as the first item in the pair with the URL of a page which links it as a second element:
urls = file.map(lambda line: (line.split(" ")[0], line.split(" ")[1:]))

#mapValues to remove duplicates from value list and sort it:
def f(x): return sorted(list(set(x)))
urls2 = urls.mapValues(f)

#flatMapValues to return pair/value tuples of (target, reference):
def g(x): return x
urls3 = urls2.flatMapValues(g)

#map to flip target and references:
rddInverted = urls3.map(lambda x: (x[1], x[0]))

#group and sort by key:
final = rddInverted.groupByKey().mapValues(list).sortByKey(ascending = True)
```

```
#sort by value:
finalSorted = final.mapValues(f)

#output for large file:
print(finalSorted.take(10))
print(finalSorted.count())
```

▸ (4) Spark Jobs

```
[('www.example1.com', ['www.example10444.com', 'www.example33108.com', 'www.example38177.com', 'www.example55699.com', 'www.example57376.com', 'www.example72549.com', 'www.example82521.com', 'www.example94619.com']), ('www.example10.com', ['www.example31480.com', 'www.example32441.com', 'www.example35833.com', 'www.example37322.com', 'www.example39583.com', 'www.example59772.com', 'www.example7313.com', 'www.example94259.com']), ('www.example100.com', ['www.example15270.com', 'www.example30972.com', 'www.example39623.com', 'www.example42465.com', 'www.example8219.com', 'www.example88000.com']), ('www.example1000.com', ['www.example21434.com', 'www.example4238.com', 'www.example81432.com', 'www.example9523.com', 'www.example96395.com']), ('www.example10000.com', ['www.example23061.com', 'www.example33274.com', 'www.example44484.com']), ('www.example100000.com', ['www.example40090.com', 'www.example40104.com', 'www.example48175.com', 'www.example58537.com', 'www.example65426.com', 'www.example69557.com']), ('www.example10001.com', ['www.example56344.com', 'www.example9856.com']), ('www.example10002.com', ['www.example13258.com', 'www.example15859.com', 'www.example16921.com', 'www.example18482.com', 'www.example34053.com', 'www.example48273.com', 'www.example60617.com']), ('www.example10003.com', ['www.example24039.com', 'www.example7180.com', 'www.example72155.com']), ('www.example10004.com', ['www.example11907.com', 'www.example30952.com', 'www.example50633.com', 'www.example64570.com'])]
99584
```

Command took 9.58 seconds -- by krista.miller445@du.edu at 4/13/2022, 11:58:18 AM on lab3cluster