PySpark distinct().count() on a csv file

I'm new to spark and I'm trying to make a distinct().count() based on some fields of a csv file.

Csv structure(without header):

```
id,country,type
01,AU,s1
02,AU,s2
03,GR,s2
03,GR,s2
```

to load .csv I typed:

```
lines = sc.textFile("test.txt")
```

then a distinct count on `lines` returned 3 as expected:

```
lines.distinct().count()
```

But I have no idea how to make a distinct count based on lets say `id` and `country`.

python apache-spark pyspark

edited Jan 16 '15 at 15:41

elyase
14.8k 1 24 52

asked Jan 16 '15 at 15:28

dimzak
996 2 13 33

2 Answers

In this case you would select the columns you want to consider, and then count:

```
sc.textFile("test.txt")\
  .map(lambda line: (line.split(',')[0], line.split(',')[1]))\
  .distinct()\
  .count()
```

This is for clarity, you can optimize the lambda to avoid calling `line.split` two times.

answered Jan 16 '15 at 15:39

elyase
14.8k 1 24 52

The split line can be optimized as follows:

```
sc.textFile("test.txt").map(lambda line: line.split(",")[:-1]).distinct().count()
```

edited May 11 '15 at 16:37

answered May 11 '15 at 16:32

Andy
20.2k 11 63 108

rami
14 4