W205 summer 2017 section 2 – Lab04 - K Iwasaki

SUBMISSION 1:

Create an RDD with tuples were the there is a key and a value. But in contrast to the example above the key is removed from the value portion of the key-value tuple. Submit the code and a print out of the first tuple.

```
>>> narcoticsCrimeTuples2 = narcoticsCrimes.map(lambda x:(x.split(",")[0], x.split(",")[1:]))
>>> narcoticsCrimeTuples2.first()
17/06/20 08:26:46 INFO spark.SparkContext: Starting job: runJob at PythonRDD.scala:361
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Got job 14 (runJob at PythonRDD.scala:361) with 1 output partitions
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Final stage: ResultStage 14(runJob at PythonRDD.scala:361)
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Parents of final stage: List()
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Missing parents: List()
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Submitting ResultStage 14 (PythonRDD[16] at RDD at PythonRDD.scala:43)
, which has no missing parents
17/06/20 08:26:46 INFO storage.MemoryStore: ensureFreeSpace(6864) called with curMem=275188, maxMem=556038881
17/06/20 08:26:46 INFO storage.MemoryStore: Block broadcast_15 stored as values in memory (estimated size 6.7 KB, fre
e 530.0 MB)
17/06/20 08:26:46 INFO storage.MemoryStore: ensureFreeSpace(4352) called with curMem=282052, maxMem=556038881
17/06/20 08:26:46 INFO storage.MemoryStore: Block broadcast_15_piece0 stored as bytes in memory (estimated size 4.3 K
B, free 530.0 MB)
17/06/20 08:26:46 INFO storage.BlockManagerInfo: Added broadcast_15_piece0 in memory on localhost:38528 (size: 4.3 KB
, free: 530.2 MB)
17/06/20 08:26:46 INFO spark.SparkContext: Created broadcast 15 from broadcast at DAGScheduler.scala:861
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 14 (PythonRDD[16] at RDD a
t PythonRDD.scala:43)
17/06/20 08:26:46 INFO scheduler.TaskSchedulerImpl: Adding task set 14.0 with 1 tasks
17/06/20 08:26:46 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 14.0 (TID 254, localhost, PROCESS_LOCAL,
2218 bytes)
17/06/20 08:26:46 INFO executor.Executor: Running task 0.0 in stage 14.0 (TID 254)
17/06/20 08:26:46 INFO rdd.HadoopRDD: Input split: file:/home/w205/w205-summer-17-labs-exercises/data/Crimes_-_2001_t
o_present_data/Crimes_-_2001_to_present.csv:0+33554432
17/06/20 08:26:46 INFO python.PythonRDD: Times: total = 13, boot = 5, init = 7, finish = 1
17/06/20 08:26:46 INFO executor.Executor: Finished task 0.0 in stage 14.0 (TID 254). 2497 bytes result sent to driver
17/06/20 08:26:46 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 14.0 (TID 254) in 21 ms on localhost (1/1
)
17/06/20 08:26:46 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
17/06/20 08:26:46 INFO scheduler.DAGScheduler: ResultStage 14 (runJob at PythonRDD.scala:361) finished in 0.021 s
17/06/20 08:26:46 INFO scheduler.DAGScheduler: Job 14 finished: runJob at PythonRDD.scala:361, took 0.030512 s
(u'10184515', [u'HY372204', u'08/06/2015 11:55:00 PM', u'033XX W DIVERSEY AVE', u'2027', u'NARCOTICS', u'POSS: CRACK'
, u'STREET', u'true', u'false', u'1412', u'014', u'35', u'22', u'18', u'1153440', u'1918377', u'2015', u'08/13/2015 1
2:57:42 PM', u'41.931870591', u'-87.711546895', u'"(41.931870591', u' -87.711546895)"'])
```

SUBMISSION 2: *Submit the the code for executing the above query as a Spark SQL python call. Also submit the number of rows in the result*

```
>>> results = sqlContext.sql('select count(*) from Web_Session_Log where REFERERURL= "http://www.ebay.com"')
>>> results.show()
```

```
+-----+
|  _c0|
+-----+
|3943|
+-----+
```