INF 551 – Spring 2016
Homework #5 (100 points)
Due: 11:59pm on 4/29/2016 to Blackboard

1.  [Spark, 50 points] Consider again the tables used in Homework 3 shown below.
   - Company(cname, stockPrice, country)
   - Person(name, phoneNumber, city)
   - Product(name, price, category, maker)
   - Purchase(buyer, seller, store, product)

   The data for the tables are provided to you in tab-delimited files as company.txt, person.txt, product.txt, and purchase.txt.

   Write Spark program in Python to answer the following questions.
   a.  Find names of people who bought cell phones.
   b.  Find names of people who bought US products but did not buy Chinese products.
   c.  Find the number of products sold for each product category and buyer city, e.g., cell phone, los angeles, 5; cell phone, san gabriel, 3; etc.
   d.  Find the average price of products for each product category.
   e.  Find the number of different stores in purchase table. For this part, you are required to use mapPartitions, but do not use Spark’s distinct and counting functions including: distinct(), count(), and countByKey().

   For each part, submit a Python script for the Spark job. Name your script for part (a) as “part-a.py”, part (b) as “part-b.py”, and so on.

   The program should be executed as follows.
   $ path-to-/bin/spark-submit part-a.py
   $ path-to-/bin/spark-submit part-b.py
   ...

   The output from each program should be printed to stdout.

   Note: You can NOT use Spark SQL in this homework.

I attached part-a.py, part-b.py, part-c.py, part-d.py, part-e.py
Development Environment: PySpark, Mac OS

Note that company.txt, person.txt, product.txt, purchase.txt and part.py are in the same directory!
Simple snapshot

Last login: Sat Apr 30 20:18:05 on ttys001
sin-yoengmin-ui-MacBook-Pro:~ Youngmin$ cd desktop/spark-1.6.1-bin-hadoop2.6
sin-yoengmin-ui-MacBook-Pro:spark-1.6.1-bin-hadoop2.6 Youngmin$ bin/spark-submit
part-a.py
[u'steve', u'mary', u'bill', u'mark']
sin-yoengmin-ui-MacBook-Pro:spark-1.6.1-bin-hadoop2.6 Youngmin$ bin/spark-submit
part-b.py
[u'mark']

part-c.py
defaultdict(<type 'int'>, {('alhambra', 'cell phone'): 2, ('los angeles', 'cell phone'): 3, ('alhambra', 'software'): 2, ('alhambra', 'laptop'): 1, ('los angeles', 'laptop'): 3, ('los angeles', 'software'): 8})

part-d.py
[(u'cell phone', 559.9075), (u'laptop', 850.2), (u'software', 64000.0)]

part-e.py
[3]

python      R          sbin       README.md

part-a.py  part-b.py  part-c.py  part-d.py

part-e.py  company.txt  person.txt  product.txt

purchase.txt