EPL660: Information Retrieval and Search Engines – Lab 10

University of Cyprus
Department of
Computer Science

Παύλος Αντωνίου

Γραφείο: Β109, ΘΕΕ01

Task 1: Sum of sales per country

- Write a program on Apache Spark to calculate the sum of sales (prices) per country using the dataset SalesJan2009.csv:
- Hint: define a function (to be given as input to map) that splits each column and returns a tuple with the requested information

Transaction date	Product	Price	Payment Type	Name	City	State	Country	Account Created	Last Login	Latitude	Longitude
01-02-2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	01-02-2009 6:00	01-02-2009 6:08	51.5	-1.1166667
01-02-2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United States	01-02-2009 4:42	01-02-2009 7:49	39.195	-94.68194
01-02-2009 13:08	Product1	1200	Mastercard	Federica e Andrea	Astoria	OR	United States	01-01-2009 16:21	01-03-2009 12:32	46.18806	-123.83
01-03-2009 14:44	Product1	1200	Visa	Gouya	Echuca	Victoria	Australia	9/25/05 21:13	01-03-2009 14:22	-36.1333333	144.75
01-04-2009 12:56	Product2	3600	Visa	Gerd W	Cahaba Heights	AL	United States	11/15/08 15:47	01-04-2009 12:45	33.52056	-86.8025
01-04-2009 13:19	Product1	1200	Visa	LAURENCE	Mickleton	NJ	United States	9/24/08 15:19	01-04-2009 13:04	39.79	-75.23806

Task 1: Results

```
[('United Kingdom', 144000), ('United States', 738300), ('Australia', 64800),
('Israel', 1200), ('France', 53100), ('Netherlands', 44700), ('Ireland', 69900),
('Canada', 124800), ('India', 2400), ('South Africa', 12300), ('Finland', 2400),
('Switzerland', 76800), ('Denmark', 18000), ('Belgium', 12000), ('Sweden', 22800),
('Norway', 21600), ('Luxembourg', 1200), ('Italy', 37800), ('Germany', 42000),
('Moldova', 1200), ('Spain', 16800), ('United Arab Emirates', 12000), ('Bahrain',
1200), ('Turkey', 7200), ('Kuwait', 1200), ('Malta', 4800), ('Hungary', 3600),
('Austria', 10800), ('Jersey', 1200), ('Malaysia', 1200), ('Iceland', 1200),
('South Korea', 1200), ('Brazil', 12300), ('New Zealand', 7200), ('Russia', 3600),
('Monaco', 2400), ('Hong Kong', 1200), ('Thailand', 4800), ('Bulgaria', 1200),
('Latvia', 1200), ('Poland', 2400), ('Philippines', 2400), ('Argentina', 1200),
('The Bahamas', 2400), ('Japan', 2400), ('Czech Republic', 6000), ('Cayman Isls',
1200), ('Ukraine', 1200), ('Dominican Republic', 1200), ('China', 1200),
('Greece', 1200), ('Costa Rica', 1200), ('Bermuda', 1200), ('Romania', 1200),
('Guatemala', 1200), ('Mauritius', 3600)]
```

Task 2: Spark Streaming

- Use the Spark Streaming library to modify the program of Task 1 in order to be able to process the same type of data (SalesJan2009.csv) coming from a socket
- Test your .py program using netcat (nc) tool
  - See slides 42-44 from Lab9.pdf
- Send the 15 first lines of SalesJan2009.csv files through the nc tool and capture the results printed by your .py program

```
ubuntu@lab-0:/usr/local/spark/bin$ nc -lk 5556
1/2/09 6:17,Product1,1200,Mastercard,carolina,Basildon,England,United Kingdom,1/2/09 6:00,1/2/09 6:08,51.5,-1.1166667
1/2/09 4:53,Product1,1200,Visa,Betina,Parkville,MO,United States,1/2/09 4:42,1/2/09 7:49,39.195,-94.68194
1/2/09 13:08,Product1,1200,Mastercard,Federica e Andrea,Astoria,OR,United States,1/1/09 16:21,1/3/09 12:32,46.18806,-123.83
1/3/09 14:44,Product1,1200,Visa,Gouya,Echuca,Victoria,Australia,9/25/05 21:13,1/3/09 14:22,-36.1333333,144.75
1/4/09 12:56,Product2,3600,Visa,Gerd W,Cahaba Heights,AL,United States,11/15/08 15:47,1/4/09 12:45,33.52056,-86.8025
1/4/09 13:19,Product1,1200,Visa,LAURENCE,Mickleton,NJ,United States,9/24/08 15:19,1/4/09 13:04,39.79,-75.23806
1/4/09 20:11,Product1,1200,Mastercard,Fleur,Peoria,IL,United States,1/3/09 9:38,1/4/09 19:45,40.69361,-89.58889
1/2/09 20:09,Product1,1200,Mastercard,adam,Martin,TN,United States,1/2/09 17:43,1/4/09 20:01,36.34333,-88.85028
1/4/09 13:17,Product1,1200,Mastercard,Renee Elisabeth,Tel Aviv,Tel Aviv,Israel,1/4/09 13:03,1/4/09 22:10,32.0666667,34.7666667
1/4/09 14:11,Product1,1200,Visa,Aidan,Chatou,Ile-de-France,France,6/3/08 4:22,1/5/09 1:17,48.8833333,2.15
1/5/09 2:42,Product1,1200,Diners,Stacy,New York,NY,United States,1/5/09 2:23,1/5/09 4:59,40.71417,-74.00639
1/5/09 5:39,Product1,1200,Amex,Heidi,Eindhoven,Noord-Brabant,Netherlands,1/5/09 4:55,1/5/09 8:15,51.45,5.4666667
1/2/09 9:16,Product1,1200,Mastercard,Sean,Shavano Park,TX,United States,1/2/09 8:32,1/5/09 9:05,29.42389,-98.49333
1/5/09 10:08,Product1,1200,Visa,Georgia,Eagle,ID,United States,11/11/08 15:53,1/5/09 10:05,43.69556,-116.35306
1/2/09 14:18,Product1,1200,Visa,Richard,Riverside,NJ,United States,12/9/08 12:07,1/5/09 11:01,40.03222,-74.95778
```

```
-------------------------------------------
Time: 2020-11-20 14:53:10
-------------------------------------------
('United States', 14400)
('Australia', 1200)
('France', 1200)
('Netherlands', 1200)
('United Kingdom', 1200)
('Israel', 1200)
```

Submission

- Save the results of both Tasks to a document (.docx) file
  - For Task1, provide a screenshot of the program output
  - For Task2, provide 2 screenshots, similar to those shown in slide 4
  - You can use either the terminal or Spyder IDE for running your programs
- Zip the 2 .py files and the .docx file
- Submit the zip file to Moodle by the 3rd of December @ 15.00