EPL448: Data Mining on the Web – Lab 11

University of Cyprus
Department of
Computer Science

Παύλος Αντωνίου

Γραφείο: B109, 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() transformation – can be a regular function not a lambda) that splits each row and returns a python tuple with the requested information

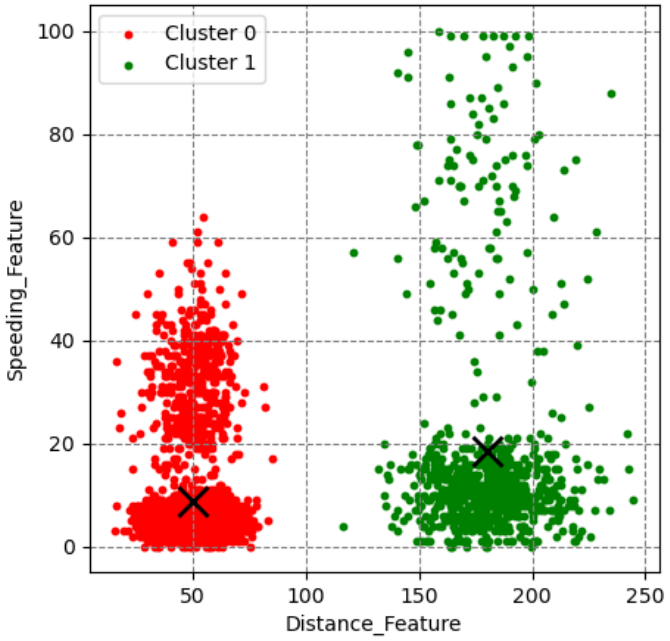
Transaction date	Product	Price	Payment Type	Name	City	State	Country	Account Created	Last Login	Latitude	Longitude
01-02-2009 6:17	Product1	1200	Mastercard	carolina	Basildon	England	United Kingdom	01-02-2009 6:00	01-02-2009 6:08	51.5	-1.1166667
01-02-2009 4:53	Product1	1200	Visa	Betina	Parkville	MO	United 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	-36.133333	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

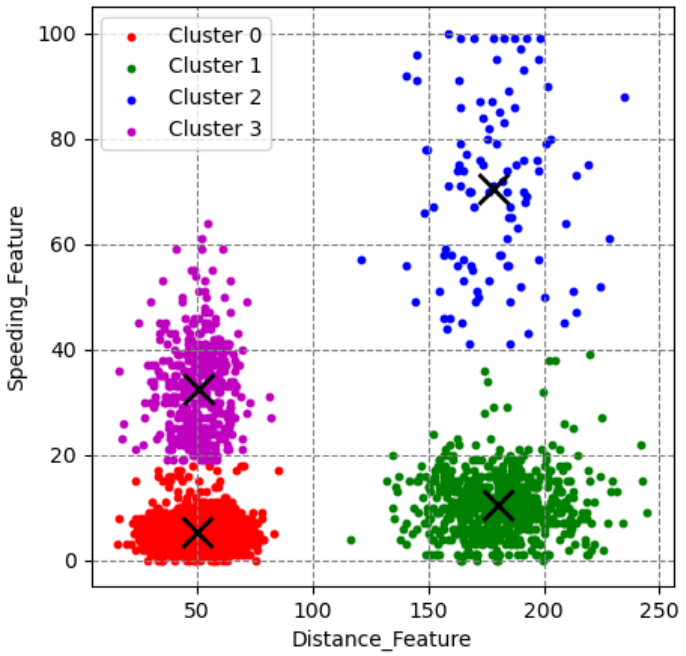
```
[('United Kingdom', 144000), ('United States', 738300), ('Australia', 64800),
('Israel', 1200), ('France', 53100), ('Netherlands', 44700), ('Ireland', 69900),
('Canada', 124800), ('India', 2400), ('South Africa', 12300), ('Finland', 2400),
('Switzerland', 76800), ('Denmark', 18000), ('Belgium', 12000), ('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: K-means on fleet data

- Modify the given `kmeans-fleet.py` to cluster fleet data (presented in Lab5) for k = 2 and k=4
  - Replace `None` with appropriate commands
  - Results for each k are shown below



k=2



k=4

Submission

- Save the results of both Tasks to a single document (.docx) file
  - For Task1, provide a screenshot of the program output showing the results of slide 3
  - For Task2, provide 2 screenshots, similar to those shown in slide 4 for k=2 and k=4
  - You can use either the terminal or Spyder IDE for running your programs
- Zip the 2 .py files and the 1 .docx file
- Submit the zip file to Moodle by Wednesday 17th of April @ 09.00 am