Exercise 6b

Unstructured Spark exercise

Prior Knowledge
Unix Command Line Shell
Simple Python
Spark Python
Simple SQL syntax

Learning Objectives
Pulling together your skills from previous exercises

Software Requirements
(see separate document for installation of these)

- Apache Spark 2.0.0
- Python 2.7.12
- Nano text editor or other text editor

Aim
There is a file on your VM that contains some data about health practices (e.g. GP surgeries) in the UK:
`~/datafiles/practices/ukpractices2015.csv`

The CSV file has a header line with titles of each column.

The aim is simple:

I'd like you to calculate the number of practices per postcode prefix for the data. The postcode prefix I define as the first few characters of the postcode up to the space.

Please tell me the number of surgeries for the postcode areas: OX1, SW11.

There are some hints overleaf.

Hints:

1. Create a directory to hold your code and files

2. Use the same databricks CSV reader to load the data in
    a. If you are using Spark 2.0.0 (on the local VM) then you no longer need to specify the package on the pyspark command line to use CSV reading… it is built in.

3. You should know enough to do this as a set of Map/ReduceByKey operations. You could also look at countByKey

4. Alternatively, you can do this all in SQL if you like SQL.

5. If you like to mix and match SQL and Map/Reduce you can do that too. I've shown you how to do DataFrame → RDD. The following page shows you how to do RDD→ DataFrame:

    https://spark.apache.org/docs/latest/sql-programming-guide.html#interoperating-with-rdds

6. Ask me or David if you get stuck.