Exercise 5

Simple 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 2.1.1
- Python 2.7.12
- Jupyter Notebook

Aim
There is a file in the Github repository that contains some data about health practices (e.g. GP surgeries) in the UK.

```
~/BigData/datafiles/practices/ukpractices.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: BN1, GU27.

We are going to do this locally, NOT on EC2.

There are some hints overleaf.

Hints:

1. Create a new Jupyter Notebook as in previous exercises
2. Use the CSV reader from the SQL exercise to load the data in
3. You should know enough to do this:
    a. either as a set of Map/ReduceByKey operations.
    b. Alternatively, you can do this all in SQL if you like SQL.
4. If you like to mix and match SQL and Map/Reduce you can do that too.

    If you started with a DataFrame and then converted to an RDD, then you convert any of the resulting RDDs back to a DataFrame using rdd.toDF()
5. Ask one of us if you get stuck.