**6CS030 Big Data**

**2018/9**

**Portfolio – Part 1**

# Worksheet Three – 10%

# Hand-out: Week 10. Due: Week 11 (See Canvas for date)

This worksheet is based on the Hadoop Workbooks 1 and 2.

1. This worksheet uses three Comma Separated Values (CSV) files generated from the *Employment Rate & Qualifications Profile of Adults* spreadsheet seen in Worksheet One.

They have undergone some cleaning to remove non-numeric fields in any fields containing figures. There is also no header row in the first line.

The files are available on hpd-srv.wlv.ac.uk in the /home/6cs030/Worksheet3 directory.

An updated version of Population.java (see Week 9’s Workbook) can also be found in the /home/6cs030/Worksheet3 directory. This has been amended to check if the figures found are numbers or floats.

Use the Linux cp command to copy the files to your own directory, for example,

cp /home/6cs030/Worksheet3/Population.java .

You need to use just one of the CSV datasets.

First take your student number and divide it by 3. Use the ***remainder*** value (modulus) to pick one of the following worksheets:

|  |  |  |
| --- | --- | --- |
| **Remainder Value** | **CSV Dataset to use** | **Java Class Name** |
| 0 | Employment\_Rate.csv | EmpRate |
| 1 | Degree-Level\_Quals.csv | DegreeQuals |
| 2 | No\_Quals.csv | NoQuals |

For example, if your student number is *1712345*, *1712345/3= 2* so you would use the *kermode.json* dataset. See the *Remainder* spreadsheet if you are not sure how to do this.

1. Examine your dataset and carry out the following tasks:

|  |  |  |
| --- | --- | --- |
| **Task no** | **Task** | **Marks** |
| a | **Java and Hadoop**  Make changes to the Population.java file to reflect the following:   * Change the Population class name to the Java Class Name shown above * Amend the Mapper and Reducer class names to include your initials * Reflect these changes in the main method | 20 |
| b | **Run the code**   * Show the steps required to run the code produced in Part a. This should include all steps (e.g., compile, creating the jar file, storing the data into the dfs). * The input and output directories produced should include your initials * Make sure you use the CSV file allocated from Part 1 * Show the contents of your output directory and some of the output produced | 30 |
| c | **Apache Spark**  Write a command to:   * Load the same CSV file from the Hadoop file system (hdfs) into Apache Spark * Show two queries that manipulate the data, one using the Data Frame and one using a SQL Query   + The columns in the SQL query should be renamed to be more meaningful | 30 |
| d | Name one advantage to using Hadoop or Spark for handling Big Data and include brief explanation of why you think this is an advantage. | 10 |
| e | Name one disadvantage to using Hadoop or Spark for handling Big Data and include brief explanation of why you think this is a disadvantage. | 10 |

See Worksheet 3 Marksheet for further marks breakdown.

Note this is an individual assessment. Any group answers will be classed as plagiarism.

You must use the dataset allocated otherwise 0 marks will be allocated.