Coursework assigned: 12 February 2021.

Coursework submission deadline: 23:59pm, 1 March 2021.

Late submission deadline (capped at 50%): 23:59pm, 2 March 2021.

Overview: The coursework aims to make you familiar with the following concepts: (i) Big Data collection, and (iii) programming using the MapReduce framework.

This coursework is formally assessed and is worth 10% of your final mark.

You will receive some feedback as part of the marking of the coursework after 4 weeks from the coursework submission deadline. You will also receive additional feedback in the form of examples solutions soon after.

Submission: Include BOTH files below:

(i) A file, **Coursework1.PDF**, containing your answers. For tasks that require writing code, write your code and comments as part of the answer. For tasks that require showing output of a program, show the output or a small part of the output.

(ii) A file, **Coursework1_code.ZIP**, containing, for each program, the code of the program (.py file) and a file containing **the entire output** of applying the program to the required dataset. Name the code and output to indicate the task it corresponds to (e.g., task2.py for the code and task2.out for the output of Task 2).

Evaluation: The maximum number of marks (out of 100) for each task is given in square brackets [] next to each question.

Plagiarism: “Plagiarism is passing off someone else’s work as your own, or submitting a piece of your own work that you have already submitted as part of a different programme, module or at a different institution. The penalties for plagiarising by the College can be severe. Uploading work to KEATS is regarded by the Department as a statement by the student concerned, confirming that the work has not been plagiarised.”

Late submission: "If you are submitting your coursework after the deadline, you must submit a Mitigating Circumstances Form (MCF), with evidence to justify why you have not submitted on time. If you do not do this or your reasons are not acceptable, your coursework may be given a mark of zero." Please speak to your personal tutor about the MCF. **Lecturers cannot provide deadline extensions.**

Task 1. Big data collection using Apache Sqoop.

Describe three features of Apache Sqoop that help import data into a distributed file system efficiently. Your description should state the feature and a brief justification of what the feature does and how it can help efficiency.

[30]

Task 2. MapReduce frequency-based computations.

Download the Adult dataset (file adult.data) from

https://archive.ics.uci.edu/ml/machine-learning-databases/adult
or download it from KEATs.

Write a program task2.py using mrjob, which outputs the 10 most frequent values in the Age attribute of the Adult dataset and their frequency. That is, the ten values in Age that appear the largest number of times in the dataset and the number of times each of these values appears.

The format of the output should be as follows:

98 "6"

88 "1"

This format says that the value 6 is the most frequent and appears 98 times followed by value 1 that appears 88 times.

In your report, provide the output file (result) from applying the program to the first 10 lines of the dataset. Also, provide your code along with comments and description about what each step does.

In your zip file, provide the code with comments (a task2.py file) and also the output file (a task2.out file) from applying the program to the entire dataset. [30]

Note: You can use redirection (e.g., python3.6 myprogram.py > myoutput.txt) to get the output. You can execute the program in local mode (i.e., without -r hadoop).

[35]

Task 3. MapReduce inverted index.

An inverted index is a data structure commonly used to map symbols into their location. Many search engines utilize this data structure to efficiently process user queries. In this task, you will implement a simple inverted index and apply it to a web dataset coming from msn.com.

Download the dataset msnbc_lines.seq from KEATS.Each record (line) in the dataset contains a first number which is the line-id, and one or more symbols (integers) that can appear one or more times in a line.

Write a program task3.py using mrjob. Given the dataset, your program must output each symbol of the dataset along with a set of line-ids indicating the lines in which the symbol appears.

For example, there is a file named *test* on KEATs which contains the following lines:

1 1 1

2 2

3 3 2 2 4 2 2 2 3 3

4 5

5 1

6 6

7 1 1

8 6

9 6 7 7 7 6 6 8 8 8 8

10 6 9 4 4 4 10 3 10 5 10 4 4 4

The first line in *test* contains 1 as line-id and then the symbol 1 two times. The second line contains 2 as line-id and then the symbol 2 once. Your program on *test* should output the following lines (the order may differ):

"5" [4, 10]

"6" [6, 8, 9, 10]

"3" [3, 10]

"4" [3, 10]

"7" [9]

"8" [9]

"9" [10]

"1" [1, 5, 7]

"10" [10]

"2" [2, 3]

In this output, the symbol “5” is associated with [4, 10] which means that it appears in lines with line-ids 4 and 10.

In your report, provide the output file (result) from applying the program to the first 10 lines of the **msnbc_lines.seq** dataset. Also, provide your code along with comments and description about what each step does.

In your zip file, provide the code with comments (a task3.py file) and also the output file (a task3.out file) from applying the program to the entire **msnbc_lines.seq** dataset.

You can execute the program in local mode (i.e., without -r hadoop).

[35]

-- END of CW1 --