Big Data Processing
COSC 2637/2633
Assignment 2

Assessment Type	– Individual assignment. – Submit online via Canvas → Assignment 2. – Marks awarded for meeting requirements as closely as possible. – Clarifications/updates may be made via announcements or relevant discussion forums.
Due Date	23:59, 26 September
Marks	40

Overview

Write advanced MapReduce programs which give your chance to develop in-depth understanding of principles when solving complex problems on Hadoop execution platform, and analyze solutions by applying the knowledge learned in this course to achieve the optimal outcome.

Learning Outcomes

The key course learning outcomes are:

CLO 1: model and implement efficient big data solutions for various application areas using appropriately selected algorithms and data structures.
CLO 2: analyse methods and algorithms, to compare and evaluate them with respect to time and space requirements and make appropriate design choices when solving real-world problems.
CLO 3: motivate and explain trade-offs in big data processing technique design and analysis in written and oral form.
CLO 4: explain the Big Data Fundamentals, including the evolution of Big Data, the characteristics of Big Data and the challenges introduced.
CLO 6: apply the novel architectures and platforms introduced for Big data, i.e. Hadoop, MapReduce and Spark.

Assessment details

Task 1 – Count word co-occurrence frequency (10 marks)

Write a MapReduce program that uses pairs approach and outputs the frequency of word pairs.

- Given "(a, b)" and word pair "(b, a)", they are considered as different word pairs,
- Do not output count the pair of same words, e.g., "(a, a)",
- The words are considered co-occurred if they are in the same line and the number of words between them <=3.

Task 2 – Count word pair relative frequency (10 marks)

Write a MapReduce program that uses pairs approach and outputs the relative frequency of word pairs.

- Given "(a, b)" and word pair "(b, a)", they are considered as different word pairs,
- Do not output count the pair of same words, e.g., "(a, a)",
- The words are considered co-occurred if they are in the same line and the number of words between them <=3.

Task 3 – Implement PAM algorithm with a MapReduce Program (20 marks)

The most common realization of k-medoid clustering is the Partitioning Around Medoids (PAM) algorithm which is described below:

step 1.	Initialize: randomly select k of the n data points as the medoids
step 2.	Assignment step: Associate each data point to the closest medoid.
step 3.	Update step: For each medoid m and each data point o associated to m , swap m and o, and compute the total cost of the configuration (that is, the average dissimilarity of o to all the data points associated to m). Select the medoid o with the lowest cost of the configuration.
step 4.	Repeat alternating steps 2 and 3 until there is no change in the assignments.

(a) Your program must correct implement PAM. In your code, provide detailed comments to specify where each step is implemented. For example

```
//Step 2 start.
…
Block of code;
…
//Step 2 end.
```

Run your PAM MapReduce program to cluster a point dataset NYTaxiLC1000 [1] (with 1000 points in longitude and latitude from line 1 to line 1000) where $1 \leq k \leq 6$. Note the initial medoids are always points at line 100, 200, 300, 400, 500 and 600 (i.e., $k = 1$, the initial medoid is point at line 100; $k = 2$, the initial medoids are points at line 100, 200; and so on for k=3, 4, 5 and 6).

(b) Visualize the clustering results. The points belonging to the same cluster are with the same color. The medoid of each cluster is highlighted.

(c) Analyse what is the best setting of k ($3 \leq k \leq 6$) and explain why.

Submission

Your assignment should follow the requirement below and submit via Canvas > Assignment 2.

Assessment declaration: when you submit work electronically, you agree to the assessment declaration: https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration

Format Requirement

Failure to follow the requirements incur penalty

(a) Submit a zip file including three Maven projects for Task 1, 2, 3. Each project must include the standalone jar file. The zip file should be named as sxxxxx_BDP_A2.zip (replace sxxxxx by your student ID). (1 mark)

(b) You need include a "README" file in the zip file. In the README, specify how to run each project using the standalone jar in Hadoop. (1 mark)

(c) Besides Maven project, answer Task 3 (a)(b)(c) in a PDF file which is included in the zip file. (1 mark)

(d) Paths of input file and output file should not be hard-coded. (1 mark)

Functional Requirement

Failure to follow the requirements incur penalty

(a) For Task 1 and 2, Use "StringTokenizer(String str)" when constructing a string tokenizer for the specified string. The tokenizer uses the default delimiter set, which is " \t\n\r\f": *the space character, the tab character, the newline character, the carriage-return character, and the form-feed character*. Delimiter characters themselves will not be treated as tokens. (2x1 mark)

(b) Your MapReduce programs must be well written, using good coding style and including appropriate

[1] NYTaxiLC dataset used in this assignment is extracted from NYC taxi pickup and drop-off points from https://registry.opendata.aws/nyc-tlc-trip-records-pds/

comments. (4x1 marks)

Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarized, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods,
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites.

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviors, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to
https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity

Marking Guide

Task 1 Implementation Correctness	0-1 marks cannot operate, no output, >1 major logic error in code	2-4- marks 1 major logic error in code	5-7 marks >1 minor logic error in code		8-9 marks There are non-logic errors or 1 minor logic error in code	10 marks output correct and no code error
Task-2 Implementation Correctness	0-1 marks cannot operate, no output, >1 major logic error in code	2-4 marks 1 major logic error in code	5-7 marks >1 minor logic error in code		8-9 marks There are non-logic errors or 1 minor logic error in code	10 marks output correct and no code error
Task-3 (a) Implementation Correctness	0-1 marks cannot operate, no output, >1 major logic error in code	2-4 marks 1 major logic error in code	5-7 marks >1 minor logic error in code the and the implementation of each step is not clearly specified in code	8-9marks >1 minor logic error in code the and the implementation of each step is clearly specified in code	10-11 marks output correct but there are non-logic errors or 1 minor logic error in code and the implementation of each step is clearly specified in code	12 marks output correct and no code error and the implementation of each step is clearly specified in code.
Task-3 (b) visualization	0 marks Task-3 (a) <= 4 marks OR No visualization of clustering is provided	1 mark Task-3 (a) >=5 marks OR visualization of clustering has obvious errors for some setting of k	2 marks Task-3 (a) >=8 marks AND visualization of clustering has minor errors for some setting of k		3 marks Task-3 (a) >=10 marks AND visualization of clustering is correct in general for different settings of k	4 marks Task-3 (a) = 12 marks AND visualization of clustering is correct for different settings of k
Task-3 (c) Analysis	0 marks Task-3 (a) <= 4 marks OR no analysis is provided.	1 mark Task-3 (a) >=5 marks OR explain how to select but analysis is improper	2 marks Task-3 (a) >=8 marks AND best setting of k is provided with analysis, but based on test results not reliable		3 marks Task-3 (a) >=10 AND best setting of k is provided with analysis but based on solid test results	4 marks Task-3 (a) = 12 marks and best setting of k is provided with analysis based on solid test results
Functional requirement	Penalty on failure to follow functional requirements detailed in specification					
Format requirement	Penalty on failure to follow format requirements detailed in specification					