DESCRIPTION OF COURSEWORK

Course Code	SWE404
Course Name	Big Data Analytics
Lecturer	Lu Yang
Academic Session	2020/04
Assessment Title	Project

A. Introduction/ Situation/ Background Information

This project evaluates the students' understanding of Lecture 5-11, which are mainly about Spark, MLlib and machine learning algorithm.

B. Course Learning Outcomes (CLO) covered

At the end of this assessment, students are able to:

CLO 1 Describe the concept and understand broad knowledge of Big Data.

CLO 2 Demonstrate appropriate knowledge of Hadoop, Spark, Storm and Map Reduce framework and apply them to build a VM-based environment.

CLO 3 Integrate knowledge and understanding of the basic principles, techniques and methodologies of organizing and searching Big Data, and apply them to create value with business insight.

CLO 4 Identify the awareness of the wide applicability of Big Data for real-world practical purposes.

CLO 5 Demonstrate the need to continually follow Big Data development trends.

C. University Policy on Academic Misconduct

1. Academic misconduct is a serious offense in Xiamen University Malaysia. It can be defined as any of the following:
   i. **Plagiarism** is submitting or presenting someone else's work, words, ideas, data or information as your own intentionally or unintentionally. This includes incorporating

published and unpublished material, whether in manuscript, printed or electronic form into your work without acknowledging the source (the person and the work).

ii. **Collusion** is two or more people collaborating on a piece of work (in part or whole) which is intended to be wholly individual and passed it off as own individual work.

iii. **Cheating** is an act of dishonesty or fraud in order to gain an unfair advantage in an assessment. This includes using or attempting to use, or assisting another to use materials that are prohibited or inappropriate, commissioning work from a third party, falsifying data, or breaching any examination rules.

2. All the assessment submitted must be the outcome of the student. Any form of academic misconduct is a serious offense which will be penalised by being given a zero mark for the entire assessment in question or part of the assessment in question. If there is more than one guilty party as in the case of collusion, both you and your collusion partner(s) will be subjected to the same penalty.

D. Instruction to Students

This is an **individual** project. Each student should submit:

1. A zip file containing:
    a. A Jupyter Notebook file as your report, named as "Project_YourStudentID.ipynb". All the codes, analysis, figures should be included in a single file with nice and neat format. Use Markdown cell with different headers for different sections of your report.
    b. The Python source code of all your implementation. Just export all of your codes from Jupyter Notebook in a .py file, named as "Project_YourStudentID.py". This is for plagiarism detection.
    c. A text file named as "Video_link.txt". It contains the OneDrive link of a 10 minutes recorded video of your project presentation. The presentation should use a PowerPoint slides.

The deadline of Project is **18:00, 20th July**. Overdue penalty will be given to the project that is submitted after the deadline.

* Your codes will be sent to a **Plagiarism** detection system for duplication checking. Please write your codes independently. (Modify your code if you copy some fragment from the Internet because your classmates may copy the same fragment.)

E. Evaluation Breakdown

No.	Component Title	Percentage (%)
1.	Project presentation	40
2.	Project report	60
	TOTAL	**100**

F. Task(s)

You are required to analyze the Movie Dataset using PySpark. The dataset can be downloaded on Moodle. These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset consists of the following files:

movies_metadata.csv: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

keywords.csv: Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.

credits.csv: Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

links.csv: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset. You may obtain more information by using a crawler program with the TMDB and IMDB IDs:

- imdbId is an identifier for movies used by http://www.imdb.com. E.g., the movie Toy Story has the link http://www.imdb.com/title/tt0114709/.

- tmdbId is an identifier for movies used by https://www.themoviedb.org. E.g., the movie Toy Story has the link https://www.themoviedb.org/movie/862.

ratings.csv: This file contains 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Questions and tasks:

1. Build regression models to predict movie revenue and vote averages based on a certain metric. (40 marks)
2. Analyze that what movies tend to get higher vote averages on TMDB. Try to use more figures with data visualization methods to illustrate your analysis. (20 marks)
3. Use collaborative filtering to build a movie recommendation system with two functions:
    a. Suggest top N movies similar to a given movie title (20 marks).
    b. Predict user rating for the movies they have not rated for. You may use a test set to test your prediction accuracy, in which the test ratings can be regarded as not rated during training (20 marks).

Show all steps including data preprocessing, modeling, testing, evaluations with concise explanation in Markdown cell. You may also try different models and compare them in different ways with discussion. If your personal computer is not powerful enough to handle this project, you may try to use some public computation resources like Google Colab.

APPENDIX 1

MARKING RUBRICS

No.	Criteria	0 (Below Expectation)	1 ~ 2 (Beginning)	3 ~ 4 (Developing)	5 (Exemplary)	Marks
1	Presentation	No presentation is submitted	The presentation is not well organized.	The presentation is completed with required content.	The presentation is well organized with clear structure, good logic and solid content.	/40
2	Report writing	No report or the report is of no logic or no organized.	Report is organized, yet without much details.	Report is organized, and with enough details.	Report is well written without grammar and spelling mistakes. good presentation format. Easy to read.	/50
3	Report format	No markdown cell is used to write down analysis.	Markdown cells are used but the contents are poorly represented.	Markdown cells are used with proper design.	Analysis and discussion are represented by markdown cells in organized and neat manner.	/10
					total	/100