COM6012 Assignment 1 - Deadline: 4:00PM, Wed 11 March 2020

Assignment Brief

How and what to submit

A. Create a .zip file containing the following:

1) **AS1_report.pdf**: A report in PDF **containing answers to ALL questions**. The report should be concise. You may include appendices/references for additional information but marking will focus on the main body of the report.
2) **Code, script, and output files:** All files used to generate the answers for individual questions above, **except the data**. These files should be named properly starting with the question number: e.g. your python code as **Q1_code.py** (**one each question**), your script for HPC as **Q1_script**.sh, and your output files on HPC such as **Q1_output**.txt or **Q1_fig**B.jpg. If you develop your answers in Jupyter Notebook, you **MUST** have these files (code, script, output, images etc specified above in **bold**) prepared and submitted after you finalising your code and output. The results should be generated from the HPC, **not your local machine**.

B. Upload your .zip file to MOLE before the deadline above. Name your .zip file as **USERNAME_STUDENTID_AS1.zip**, where USERNAME is your username such as **abc18de**, and STUDENTID is your student ID such as 19xxxxxxx.

C. **NO DATA UPLOAD**: Please do not upload the data files used. We have a copy already. Instead, please use **relative file path in your code (data files under folder 'Data')**, as in the lab notebook so that we can run your code smoothly.

D. **Code and output**. 1) Use **PySpark** as covered in the lecture and lab sessions to complete the tasks; 2) **Submit your PySpark job to HPC** with **qsub** to obtain the output.

Assessment Criteria (Scope: Session 1-4; Total marks: 20)

1. Being able to use PySpark to analyse big data to answer questions.
2. Being able to perform log mining tasks on large log files.
3. Being able to perform movie recommendation with scalable collaborative filtering.
4. Being able to use scalable k-means to analyse big data.

Late submissions: We follow Department's guidelines about late submissions, i.e., a deduction of 5% of the mark each working day the work is late after the deadline, but **NO late submission will be marked one week after the deadline** because we will release a solution by then.

Use of unfair means: "*Any form of unfair means is treated as a serious academic offence and action may be taken under the Discipline Regulations.*" (from the MSc Handbook). Please refer to the handbook or consult your tutor on what constitutes Unfair Means if not sure.

Question 1. Log Mining [10 marks]

Please finish critical & essential tasks in Lab 1 and Lab 2 before solving this question.

Data: the NASA access log July 1995 data (click to download). Please read the dataset description to understand the data and complete the following four tasks.

A. Find out the **average** number of requests on each four hours of a day of July 1995 by the local time (i.e. 00:00:00-03:59:59; 04:00:00-07:59:59; 08:00:00-11:59:59; 12:00:00-15:59:59; 16:00:00-19:59:59; 20:00:00-23:59:59). The average is taken over the days in July. You need to report **SIX** numbers, one for each of these four-hour slot. [3 marks]

B. Visualise the results in A above as a figure (e.g. bar graph or pie chart) and discuss at least two observations (e.g., any trend, contrast, something expected, unexpected or interesting), with 1 to 3 sentences for each observation. To plot on HPC, you need to activate your environment and install matplotlib via **conda install -c conda-forge matplotlib** [2 marks]

C. Find out the top 20 most requested "**.html"** files (i.e. rows containing "**xxx**.html", where **xxx** is the filename, .html is the file extension). Report the file name and number of requests for each of these 20 files (pages). [3 marks]

D. Visualise the results in C above as a figure (e.g. bar graph or pie chart) and discuss at least two observations (e.g., anything interesting), with 1 to 3 sentences for each observation. [2 marks]

Question 2. Movie Recommendation [10 marks]

Please finish essential tasks in Lab 3 before solving A/ B, and in Lab 4 before solving C/D.

Data: the MovieLens 25M Dataset (click to go to the download page). Please read the dataset description to understand the data & complete the following tasks.

A. Perform a **three-fold cross validation** of ALS-based recommendation on the rating data **ratings.csv**. Study **three** versions of ALS: one with the ALS setting used in Lab 3 notebook, and another **two different settings decided by you**. For each split, compute the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for the three ALSs. Then compute the mean and standard deviation (std) of RMSE and MAE over the three splits. Put these RMSE and MAE results for each of the three splits as well as the mean & the std in one **Table** for the three ALSs in the report (2 x 5 x 3 = 30 numbers in total). **Visualise** the mean and std of RMSE and MAE for each of the three versions of ALS in one single **figure**. [4 marks]

B. Discuss at least two observations on results in A, with 1 to 3 sentences for each observation. [1 mark]

C. After ALS, each movie is modelled with some factors. Use *k*-means with ***k=25*** to cluster the movie factors (**hint**: see **itemFactors** in ALS API, id=movieid in this problem) learned with the ALS setting in Lab 3 notebook in A for each of the three splits. For each of the three splits, use **genome-scores.csv** to find the top **three** tags for each of the **top three** largest clusters (i.e., **9 tags** in total for each split), find out the names of these top tags using **genome-tags.csv**, and count the number of movies having each of these three tags in each cluster (one number per tag per cluster). For each cluster and each split, report the top three tags **and** the respective number of movies having that tag in one table (3 cluster x 3 split x 3 tag = 27). **Hint**: For each cluster, sum up tag scores for all movies in it; find the largest three scores and their indexes; go to genome-tags to find their names. You can use any information provided by the dataset to answer the question. [4 marks]

D. Discuss at least two observations on results in C, with 1 to 3 sentences for each observation. [1 mark]

Summary of clarification from my answers to questions in discussion board

Files to submit: 1) the PDF report must contain ALL answers to questions asked. If an answer (including table/figure) is NOT found in the PDF report, you will LOSE the respective mark. 2) All the files and outputs used to provide answers in the PDF report must be included for record and verification purpose.

Q1A: The average is taken all available days in July. You decide how to deal with missing data (if any) and explain in your report.

Q1C: Here we count all ".html" files in the log. The filename refers to the filename only so files with the same filename but under different paths are considered as the same ".html" files (i.e. the same by name not necessarily by content).

Q2A: The table needs you to report 2 metrics x 5 (3 splits + mean +std) x 3 ALS settings = 30 numbers.

Q2C: 1) You need to provide 3 clusters x 3 splits x 3 tags = 27 tags and 27 numbers (one number for each tag). 2) You can use any information provided by the dataset to answer the question.