COM6012 Assignment 1 - Deadline: 11:00AM, Friday 12 March 2021

Assignment Brief

Additional explanations/clarifications (mainly from the discussion board) are shown in italic fonts.

How and what to submit

A. Create a .zip file containing the following:

1) **AS1_report.pdf**: A report in PDF **containing answers (*including all figures and tables*) to ALL questions** at the root of the zipped folder (*like readme.txt in the lab solutions*). If an answer to a question is not found in this PDF file, you will lose the respective mark. The report should be concise but you may include appendices/references for additional information.
2) **Code, script, and output files:** All files used to generate the answers for individual questions in the report, **except the data,** should be organised following the file structure used in this module on GitHub. These files should be named properly starting with the question number (separate files for the two questions): **for example**, your python code as **Q1_code.py and Q2_code.py** under folder “Code”, your HPC script as **Q1_script.sh and Q2_script.sh** under folder “HPC”, and your output files on HPC as **Q1_output.txt and Q2_output.txt** (and Q1_figB2_uk.jpg, etc.) under folder “Output”. The results must be generated from the HPC, **not your own computer**.

B. Upload your .zip file to MOLE before the deadline. Name your .zip file as **USERNAME_STUDENTID_AS1.zip**, where USERNAME is your username such as **abc19de**, and STUDENTID is your student ID such as 19xxxxxxx.

C. **NO DATA UPLOAD**: Please do not upload the data files used. Instead, use the **relative file path** in your code, assuming data files downloaded (and unzipped if needed) under folder ‘Data’, as in the lab.

D. Use **PySpark 3.0.1** as in the Labs to complete the tasks and submit your PySpark job to HPC with **qsub** to obtain the output.

Assessment Criteria (Scope: Session 1-4; Total marks: 30)

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 the 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**.

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.

Note: To plot and save figures on HPC, see Lab 3 solution but you need to activate your environment and install matplotlib via **conda install -y matplotlib** first. When using it in your code, you should do the following before using pyplot:

```
import matplotlib
matplotlib.use('Agg') # Must be before importing matplotlib.pyplot or pylab!
import matplotlib.pyplot as plt
```

Question 1. Log Mining and Analysis [15 marks]

You need to finish Lab 1 and Lab 2 before solving this question.

Data: Use **wget** to download the NASA access log July 1995 data (using the hyperlink) to the "Data" folder. The data description is the same as in Lab 2 Task 4 Question 1 so please review it to understand the data before completing the four tasks below.

A. Find out the **total** number of requests for 1) all hosts from Japanese universities ending with "**.ac.jp**", 2) all hosts from UK universities ending with "**.ac.uk**", and 3) all hosts from US universities ending with "**.edu**". Report these three numbers and visualise them using one bar graph. [3 marks]

B. 1) For each of the three countries in Question A (Japan, UK, and US), find the top 9 most frequent universities according to the host domain, e.g. for host "pc021133.**shef.ac.uk**", the university is "shef.ac.uk" (no need to convert to text like University of Sheffield, though welcome if you figure out how). 9 for Japan, 9 for the UK and 9 for the US. *For US Universities, look at what is in front of .edu (after then "next" dot, from back to front), e.g. eecs.**mit.edu**, (mit.edu is the university). For Japanese (UK) Universities: what is right in front of .ac.jp (.ac.uk) but after then "next" dot, from back to front, e.g. math.**kyoto-u.ac.jp** (dcs.**shef.ac.uk**).* [3 marks]

2) For each country, produce a pie chart visualising the percentage (with respect to the total) of requests by each of the top 9 most frequent universities and the rest, i.e. a pie with 10 pieces, 9 universities plus the rest. The university host domain such as "shef.ac.uk" (*or simply the part for the university such as "shef"*) should be labeled on the pie chart. Three pie charts need to be produced [3 marks].

C. For the most frequent university from each of the three countries, produce a heatmap plot with day as the x-axis (*the range of x-axis should cover the range of days available in the log file. If there are 31 days, it runs from 1st to 31st. If it starts from 5th and ends on 25th, it runs from 5th to 25th*), the hour of visit as the y-axis (0 to 23, as recorded on the server), and the number of visits indicated by the colour. **Three** x-y heatmap plots need to be produced with the day and hour clearly labelled. [3 marks]

D. Discuss two most interesting observations from A to C above, each with three sentences: 1) what is the observation? 2) what are the possible causes of the observation? 3) how useful is this observation to NASA? [2 marks]

E. Your report must be clearly written and your code must be well documented so that it is clear what each step is doing. [1 mark]

Question 2. Movie Recommendation and Analysis [15 marks]

You need to finish Lab 3 and Lab 4 before solving this question.

Data: Use **wget** to download the MovieLens ml-latest Dataset to the “Data” folder and unzip there. Please read the dataset description to understand the data before completing the following tasks.

A. Time-split Recommendation

1) Perform time-split recommendation using ALS-based matrix factorisation on the rating data **ratings.csv**, where **all** data need to be sorted by the timestamp and **splitting needs to be performed according to the sorted timestamp**. **Earlier time should be used for training and later time should be used for testing**. Consider three such splits with three training data sizes: 50%, 65%, and 80%. [2 marks]

2) For each of the three splits above, study **three** versions (*settings*) of ALS using your student number as the seed as the following [2 marks]

- one with the ALS **setting** used in Lab 3 except the seed
- Based on results (see the next step 3 below) from the first ALS setting, choose another **two different ALS settings that can potentially improve the results.** Provide at least a **one-sentence justification to explain** why you think the chosen setting can potentially improve the results. [*This is to imagine a real scenario. You need to think about how the performance might be improved, provide a justification, and then make changes. This implies that failing to improve the results is acceptable but we expect you provide a good justification when you make changes aiming to improve the results and such justification is sound.*]

3) For each split and each version of ALS, compute three metrics: the Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE). Put these RMSE, MSE and MAE results for each of the three splits in one **Table** for the three ALSs in the report. You need to report 3 metrics x 3 splits x 3 ALS settings = 27 numbers. [3 marks]

B. User Analysis

1) After ALS, each user is modelled with some factors. For each of the three time-splits, use *k*-means in pyspark with ***k=20*** to cluster the user factors learned with the ALS setting in Lab 3 but with your student number as the seed (as above in A.2 first bullet), and find the top three largest user clusters. Report the size of each cluster (number of users) in one **Table**, in total 3 splits x 3 clusters = 9 numbers. [2 marks]

2) For each of the three splits, for all users in the largest user cluster, use **movies.csv** to find the top five genres (*each movie may have multiple genres, separated by* ‘|’, *top refers to the number of appearances*) among movies with **ratings 4 and above (including 4)** in the training set and in the test set. Report these 3 splits x 5 genres x 2 sets = 30 genres in one **Table**. [3 marks]

3 splits: the three splits as in Q2 A1

5 genres: the top five genres for movies with ratings >=4, among moveis rated by those users in the largest user cluster in Q2 B1

2 sets: training set and test set

For the training set of each split: report 5 genres as above.

For the test set of each split: report 5 genres as above.

C. Discuss two most interesting observations from A & B above, each with three sentences: 1) what is the observation? 2) what are the possible causes of the observation? 3) how useful is this observation to a movie website such as **Netflix**? [2 marks]

D. Your report must be clearly written and your code must be well documented so that it is clear what each step is doing. [1 mark]