# COM6012 Assignment Part 2 - Deadline: 15:00 Thursday May 125, 2022

**New Deadline: 15:00 Thursday May 12, 2022**

## Assignment Brief

**Please, carefully read the assignment brief before starting to complete the assignment**

**Check out the FAQ with important clarifications/tips (last update: 9:13pm, 06 May - Q2A clarified in Q9/A9)**
**Correction of typos (last update: 3pm 12 April): Q2A: three splits → FIVE splits**

**How and what to submit**

A. Create a **folder Part2** containing the following:

1) **ASPart2_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. 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 in the report above, **except the data,** should be included. 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**, your HPC script as **Q1_script.sh and Q2_script.sh**, and your output files on HPC as **Q1_output.txt and Q2_output.txt** (and Q1_figB2.jpg, etc.). The results must be generated from the HPC, **not your local machine**.

B. Upload a .zip file to Blackboard before the deadline including your submissions for Part 1 in a **folder Part1** and your submissions for Part 2 (with all files in A above) in a **folder Part2**.

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. **Code and output**. 1) Use **PySpark 3.2.1** 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, 2, 7, and 8; 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., "If you submit work to be marked after the deadline you will incur a deduction of 5% of the mark each working day the work is late after the deadline, up to a maximum of 5 working days" but **NO late submission will be marked after the maximum of 5 working days** because we will release a solution by then. Please see this link.

**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 carefully read this link on what constitutes Unfair Means if not sure.

**Note:** To plot and save figures on HPC, see Lab 7 solution. 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 ftp://ita.ee.lbl.gov/traces/NASA_access_log_Jul95.gz) 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 maximum number and minimum number of requests on each of the seven days in a week (i.e., Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday) during July 1995. You need to report 14 numbers, one max number and one min number for each day of the week. Hint: see pyspark sql API related to data format. [4 marks]

B. Visualise the 14 numbers in A above in ONE figure to help gain insights from them [2 marks]

C. Find out the 12 most requested and 12 least requested .mpg videos. Report the video file name, e.g. *abc.mpg*, and the total number of requests for each of these 24 videos during July 1995. [4 marks]

D. Visualise the 24 total request numbers in C as ONE figure to help gain insights from them [2 marks]

E. Discuss two most interesting observations from A to D 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]

F. 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 7 and Lab 8 before solving this question.

**Data**: Use **wget** to download the MovieLens 25M Dataset to the "Data" folder and unzip there. Please read the dataset description to understand the data before completing the following tasks.

A. Perform a **five-fold cross validation** of ALS-based recommendation on the rating data **ratings.csv** with **two** versions of ALS to compare: one with the ALS setting used in Lab 7 notebook, and **another different setting decided by you *with a brief explanation of why***. For each split,

find the top 10% users **in the training set** who have rated the most movies, calling them as **HotUsers**, and the bottom 10% users **in the training set** who have rated the least movies (but rated at least one movie), calling them **CoolUsers**. Compute the Root Mean Square Error (RMSE) on the test set for the HotUsers and CoolUsers separately, for each of the **FIVE** splits and each ALS version. Put these RMSE results in one **Table** in the report (2 versions x 5 splits x 2 user groups = 20 numbers in total). Visualise these 20 numbers in ONE single **figure**. [6 marks]

B. After ALS, each movie is modelled with some factors. Use k-means with **k=10** to cluster the movie factors (hint: see itemFactors in ALS API) learned with the ALS setting in Lab 7 notebook in A for each of the five splits. Note that each movie is associated with several tags. For each of the five splits, find the **top tag** (with the most movies) **and bottom tag** (with the least movies, if there are ties, randomly pick one from them) for the top two largest clusters (i.e., **4 tags** in total for each split). For each cluster and each split, report the two tags (one top one bottom) in one table (so 2 clusters x 5 splits x 2 tags = 20 tags to report in total). You can use any information provided by the dataset to answer the question. [6 marks]

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]

**The END of Assignment**

---

**FAQs**

Q1: **How to deal with “Error: spark-submit: command not found”**

A1: If you are sure that it works before, you’ve installed/loaded modules properly, and you cannot recall what changes you made in between, an ultimate solution is to reinstall pyspark either as a new environment (e.g. myspark2, to use in all future runs) or by removing the previous installations via “conda remove --name myspark --all” or (brute-force) deleting your myspark environment stored at "/home/abc1de/.conda/envs/myspark" and reinstall, following Lab 1. See Blackboard General Forum in Discussion Board: Thread: Error: spark-submit: command not found

Q2: **How to reset your environment if you found that you’ve messed it up and encountered seemingly unrecoverable errors?**

A2:

login ShARC

qrshx

resetenv

rm ~/.conda

logout fully & then back in again

Start over with Lab 1 again to install everything

Q3: What is a split in Q2A?

A3: In five-fold cross validation, we do five splits. Each split has four folds as training and one fold as test. Each fold will be used as a test set only once.

Q4: What are the max and min number of requests asked in Q1A?

A4: Let us take Monday as an example. There are more than one Mondays in a month. Find the number of requests (regardless of the host) on all Mondays and report the max and min.

Q5: Where should I find the hot/cool users?

A5: Do the split first and then find the hot/cool users in **the training set.**

Q6: Q2B seems confusing

A: I have revised Question 2B of Assignment Part 2. Please read the current version.

Q7: HotUsers, CoolUsers in Q2A

A7: The question only asks you to consider HotUsers and CoolUsers to report the MSE, i.e. after the training (and testing). Therefore, you do not consider HotUsers or CoolUsers when you do the splitting and training. Instead of the usual case where you report the MSE for the whole test set, here I ask you to report the results only for a subset of the test set.

Q8: Q1C, how about seemingly the same videos under different paths?

A8: Please assume that if the video file names are the same, they are the same videos. In practice, this is not always the case but here we make such an assumption to make things easier to manage.

Q9: **Q2A** Steps

A9: Some have difficulties understanding what Q2A asks for. Each full stop means a step as follows. Finish one step before going to the next.

Step 1: Perform a five-fold cross validation of ALS-based recommendation on the rating data ratings.csv with two versions of ALS to compare: one with the ALS setting used in Lab 7 notebook, and another different setting decided by you with a brief explanation of why.

Step 2: For each split, find the top 10% users in the training set who have rated the most movies, calling them as HotUsers, and the bottom 10% users in the training set who have rated the least movies (but rated at least one movie), calling them CoolUsers.

Step 3: Compute the Root Mean Square Error (RMSE) on the test set for the HotUsers and CoolUsers separately, for each of the FIVE splits and each ALS version.

Step 4: Put these RMSE results in one Table in the report (2 versions x 5 splits x 2 user groups = 20 numbers in total).

Step 5: Visualise these 20 numbers in ONE single figure.

Q:

A:

Q:

A: