SCALABLE MACHINE LEARNING

Assignment 1

RAMEEZ HUSSAIN
180128228

Question 1

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 slots.

I have decided to filter out the values to contain the month of July 1995 only. I had to take into consideration each day in the month and the requests for the four-hour slots of each day in July 1995. Upon looking at the dataset I realised that there were 28 days in July so therefore, I only considered 28 days not 31 days to get accurate results. These results are presented below in Table 1.

Time Slots	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
Average	64407.4	84317.4	61735.4	59339.4	56608.25	72290.5

Table 1: Average number of requests on each of the 4-hour slots

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.

These results obtained in A can be seen in Figure 1 below. We can notice that the number of requests is at its highest during the time of 04:00:00-07:59:59. Another interesting is that the number of requests is at its lowest during the times of 16:00:00-19:59:59, a possibility for this is that people either will be working or come back tired from work. **NOTE:** All figures are saved as a .png file by using plt.savefig .

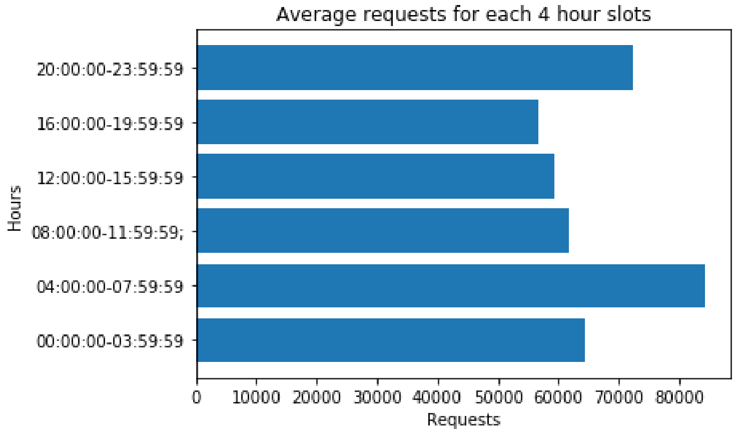

Figure 1: Average number of requests on each of the 4-hour slots

C. Find out the top 20 most requested .html files (pages). Report the file name and number of requests for each of these 20 files (pages).

For this question, I had to filter out our initial data frame and then check if it contains the extension .html in the path column. Furthermore, I also had to count the total requests for each file and show the top 20 results. Please note that I was getting duplicate Images.html and Movies.html files but in different paths but by adjusting the regular expressions I managed to overcome this problem. These results are presented below in Table 2.

```
+--------------------+------+
|path                |_count|
+--------------------+------+
|ksc.html            |40081 |
|missions.html       |24881 |
|images.html         |24505 |
|liftoff.html        |21987 |
|mission-sts-71.html|16717 |
|mission-sts-70.html|16116 |
|apollo.html         |14500 |
|apollo-13.html      |14339 |
|movies.html         |12529 |
|history.html        |11859 |
|countdown.html      |8566  |
|stsref-toc.html     |7510  |
|winvn.html          |6994  |
|mission-sts-69.html|6973  |
|apollo-13-info.html|5791  |
|lc39a.html          |5258  |
|apollo-11.html      |5002  |
|tour.html           |4318  |
|fr.html             |4212  |
|atlantis.html       |3637  |
+--------------------+------+
```

Table 2: Top 20 .html files and the number of requests

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.

The results obtained in C can be seen in Figure 2 below. We can notice that the most popular in terms of .html requests is ksc.html with 40081 requests. Another observation is that the lowest number of .html requests is atlantis.html with 3637 requests.

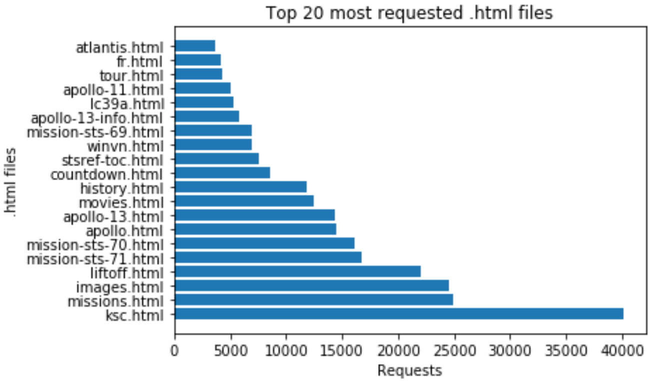

Figure 2: Top 20 .html files and the number of requests

Question 2

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 in Lab 3 notebook with “drop” as the coldStartStartegy, 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, the mean & the std in one Table for the three ALSs in the report. Visualise the mean and std of RMSE and MAE for each of the three versions of ALS in one single figure.

The three models that I have selected have the following parameters:

1. Model 1: maxIter=10, regParam =0.1, coldStartStartegy=”drop”, as stated in Lab 3
2. Model 2: maxIter=5, regParam =0.1, coldStartStartegy=”drop”
3. Model 3: maxIter=1, regParam =0.1, coldStartStartegy=”drop”

After I tested the ALS algorithm on my three models, for the three splits, the results are listed below in Table 3.

Split	**RMSE Model 1**	**MAE Model 1**	**RMSE Model 2**	**MAE Model 2**	**RMSE Model 3**	**MAE model 3**
1	0.8067635270095144	0.6235658345884876	0.8132929967688093	0.6298471796741532	3.4901884887769636	3.3075820900723443
2	0.8069389744830621	0.6238901582003696	0.8142679112978647	0.6309127938752682	3.4824660357644843	3.299126961944328
3	0.806999846907764	0.6239931449028059	0.8141497680472022	0.6309332190660907	3.4791602266292454	3.2961541187345222
Mean	0.8069007828001135	0.6238163792305543	0.8139035587046254	0.630564397538504	3.4839382503902314	0.004621053518257195
std	0.00010018558466505403	0.00018208245105071628	0.0004344182815520116	0.0005072181620125216	3.300954390250398	0.004841091673365308

Table 3: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for three versions of ALS

Please note that due to the Mean and std being so low it is not visible in Figure 3 but paying close attention you will be able to see it.

Find the Mean and Standard Deviation of the Root Mean Square Error and Mean Absolute Error for the three versions of ALS below in Figure 3 below.

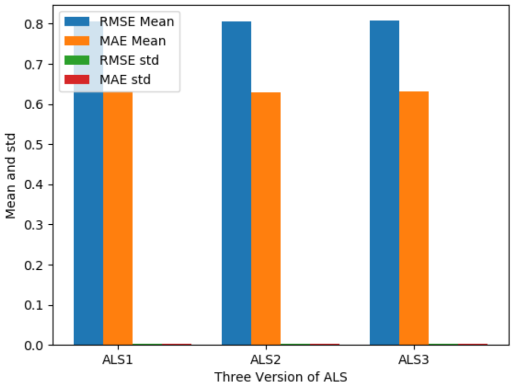

Figure 3: Mean and Standard Deviation for three versions of ALS

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

One observation that was made was that of the Root Mean Square Error is much higher compared against the Mean Absolute Error. A possible reason for this is that the errors are squared before they are averaged, so therefore the Root Mean Square Error is much higher compared against the Mean Absolute error. Another interesting observation is that model 2 and model 3 performed worse than model 1. This happens because we have decreased the number of iterations.