COM6012 Scalable Machine Learning **Lida Petropoulou**
Assignment 1 Registration No: 180128240

Question 1

In order to answer this question, the data was loaded as a dataframe with six columns. The columns are: host, timestamp, path, code, bytes and times. The last column is the same as timestamp, with the difference that it doesn't have the timezone. It was created to help with convertion of the timestamp. Afterwards, the data was cleaned, which means that the null values, which were only present at the bytes column were convert to zero. However, we noticed that the dataframe had one row which was useless, and it was removed.

A. Find out the average 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.

For this question, a new dataframe was created from the initial one where the column times (which contains strings), was converted to a timestamp with a standard format. Then a dataframe called weekdays, was created, which contains 2 columns: one with the date (without the hour and minutes) and one with the day of the week that it corresponds. From this dataframe, we can create another one, which contains every date with the number of requests done. Finally, we count the average of each day, and we sort it based on the day. The results can be seen in Table 1.

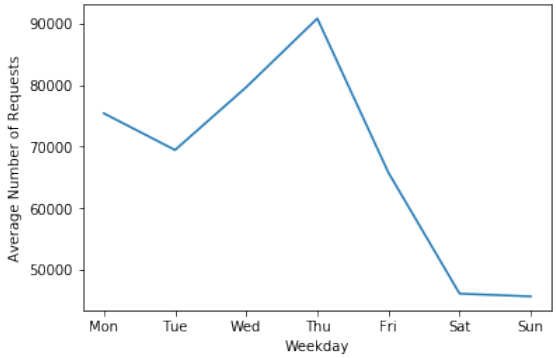
Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Average	75421.0	69460.0	79674.5	90859.0	65771.5	46095.0	45647.5

Table 1: Average number of requests on each of the seven days in a week

B. Visualise the results in A above as a figure (e.g., bar graph or pie chart) and discuss your observations (e.g., any trend, contrast, something expected, unexpected or interesting) using 1 to 3 sentences.

The results in A can be shown in the following figure (1). We notice that the number of requests during the weekends is much lower than during the rest of the days of the week. That was expected, since people don't work during the weekend.

Figure 1



C. Find out the top 20 most requested .gif images. Report the file name and number of requests for each of these 20 images

For this question, we just have to filter our initial dataframe, and check if it contains the extension .gif or .GIF in the path column. Subsequently, we can count the total requests for each file, and show the top 20. The results are presented in the Table 2.

File Path	Requests
/images/NASA-logosmall.gif	111087
/images/KSC-logosmall.gif	89530
/images/MOSAIC-logosmall.gif	60300
/images/USA-logosmall.gif	59845
/images/WORLD-logosmall.gif	59325
/images/ksclogo-medium.gif	58616
/images/launch-logo.gif	40841
/images/ksclogosmall.gif	33555
/history/apollo/images/apollo-logo1.gif	31052
/shuttle/countdown/count.gif	22189
/shuttle/countdown/count70.gif	20921
/images/launchmedium.gif	20788
/shuttle/missions/sts-71/sts-71-patch-small.gif	19832
/shuttle/missions/sts-70/sts-70-patch-small.gif	18135
/history/apollo/images/footprint-logo.gif	16168
/history/apollo/images/footprint-small.gif	13912
/history/apollo/apollo-13/apollo-13-patch-small.gif	13004
/shuttle/countdown/video/livevideo.gif	10143
/history/apollo/images/apollo-small.gif	10089
/shuttle/countdown/video/livevideo2.gif	9816

Table 2: Average number of requests on each of the seven days in a week

D. Visualise the results in C above as a figure (e.g., bar graph or pie chart) and discuss your observations (e.g., anything interesting) using 1 to 3 sentences.

The results in C can be seen in the following figure (2). We notice that the most popular gifs are the ones that are logos.

Figure 2

Question 2

In order to answer this question, we have to load the file ratings.csv with the command `spark.read.csv`, and change the strings to either integers or doubles.

A. Perform a five-fold cross validation of ALS-based recommendation on the rating data ratings.csv. Study two versions of ALS: one with the ALS setting in Lab 3 notebook with drop as the coldStartStartegy, and another different setting decided by you. For each of the five splits, report the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for the two versions of ALS. Then compute and report the mean and standard deviation (std) of RMSE and MAE.

The two models selected have the following parameters:

Model 1: `maxIter=10, regParam=0.1, coldStartStrategy="drop"`, as in Lab 3
Model 2: `maxIter=5, regParam=0.1, coldStartStrategy="drop"`

After running the ALS algorithm for our two models, for the five splits, the results were the following (Table 3):

	RMSE Model1	MAE Model1	RMSE Model2	MAE Model2
1	0.8054755856425707	0.6249901276813115	0.8253819162338106	0.6489031741714635
2	0.8050334387627959	0.6245698170044005	0.825218808972137	0.6488484710951736
3	0.8105421582542981	0.6328820886181125	0.8246893919423764	0.6483965330927223
4	0.8115185456129608	0.6334942963959526	0.8259099684198886	0.6493120844599589
5	0.8115634772279994	0.633614613882222	0.8260052014914236	0.6494646147972614
mean	0.8088266411	0.629910188716	0.825441057412	0.648984975523
std	0.00294270547812	0.00419826996047	0.00048083866356	0.000376468724792

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

B. Briefly discuss your observations on results in A with 3 to 5 sentences

We notice that the second model performs worse that the first one (by comparing the mean values of both evaluators). This probably happens because we have decreased the iterations to their half value. However we can see, that the value of the standard deviation is much lower for the second model. This means that the second model produces simila results, even if the splits are not the same.

C. After ALS, each movie is modelled with some factors. Use k-means with k=20 to cluster the movie factors learned with the ALS setting in Lab 3 notebook in A for each of the five splits. For each of the five split, use genome-scores.csv to find the top five tags for each of the top three largest clusters (i.e., 15 tags in total for each split) and report the names of these top tags using genome-tags.csv.

After using k-means with k=20 to cluster the movie factors, we managed to acquire the top tags for each split for the top three largest clusters. The results are presented below.

D. Briefly discuss your observations on results in C with 3 to 5 sentences

From the results, we can see that the clusters have a lot of tags in common, with the tag original being almost always the top tag. However, this means that the clustering wasn't done well enough. Also, we

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	story telling	mentor
good	good soundtrack	great ending
story	social commentary	dialogue
great ending	great ending	good soundtrack

Model 2

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	great ending	great ending
good	good soundtrack	storytelling
great	dialogue	mentor
great ending	mentor	good soundtrack

Split 1

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	great ending	criterion
good	good soundtrack	talky
great ending	brutality	drama
story	dialogue	great ending

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	drama	mentor
great ending	mentor	good
dialogue	good soundtrack	destiny
good soundtrack	criterion	great ending

Split 2

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	great ending	mentor
great ending	dialogue	drama
dialogue	good soundtrack	good soundtrack
good soundtrack	mentor	great ending

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	criterion	storytelling
great ending	talky	good soundtrack
dialogue	enigmatic	mentor
story	melancholic	great ending

Split 3

Model 1

Cluster 1	Cluster 2	Cluster 3
original	criterion	original
great ending	original	good soundtrack
dialogue	talky	storytelling
mentor	imdb top 250	weird
good soundtrack	runaway	social commentary

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	great ending	mentor
great ending	good soundtrack	good
dialogue	dialogue	story
good soundtrack	mentor	great ending

Split 4

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	criterion
mentor	good soundtrack	original
great ending	storytelling	talky
dialogue	mentor	enigmatic
good	great ending	golden palm

Model 1

Cluster 1	Cluster 2	Cluster 3
original	original	original
mentor	good soundtrack	mentor
dialogue	storytelling	good
great ending	great ending	story
good soundtrack	mentor	great ending

Split 5

notice that the two models produce very similar results, but in some splits model 2 gives us diverse tags, which might mean that it did a better clustering.