Project 2

Part A

This part is due at 8.20 pm on the day that it is handed out. Submissions must be made in .ipynb format, no other format is accepted. Please ensure that all output is displayed in the notebook before submission. Late submissions will be accepted up to midnight on the day is due but will be subject to a 10% penalty. No submissions will be accepted after the day on which it is due.

In this part of the project, you will need to preprocess the UTK face dataset in preparation for model building in Part B. The UTK face dataset UTKFace | Large Scale Face Dataset (susanqq.github.io) contains over 20,000 images of people ranging in age from 0 to 116. The images are annotated with Gender, Race and Age information.

As explained in the Project 2 Specification document the overall objective is construct a hash filter to search this dataset efficiently. However, in this part of the project your sole focus will be:

a) Parse the strings containing the age, gender and race information and extract each of these features. The dataset for this purpose is the utk filenames dataset and is available from Kaggle.
b) Produce meaningful visualizations on the Gender and Race features
c) Discretize the data on the Age feature and visualize it. Out of the two options for discretization: equal-width and equal-frequency, which do you consider to be a better choice and why?
d) Optional: This part does not carry any marks and is to be done only if time permits adapt the simple CNN classifier at: image_classification_from_scratch - Colaboratory (google.com) to classifying the UTK face dataset on the Age feature only using the discretization you proposed in c) above.