README

June 5, 2021

1 Introduction

The topic of our mini project is building a text generator model, which is one of the Natural Language Processing fields. To do so, we used a dataset that contains different blog posts by different people, which we preprocessed, separated according to age groups, and later used to built individual neural networks. As multiple files are involved in each step, this notebook is intended to lay out the structure of the folder system of the project in hope of helping our readers guide through the files more conveniently.

1.1 Note:

As some of the file sizes involved are relatively large, they are not included as part of the submission file that we sent via email. Instead, we have put them in this Google Drive link, which has the same folder structure as the submission and all the related files that we used to run the notebooks. Therefore, if one is interested, please find/download them using this link.

2 Structure of folder & files

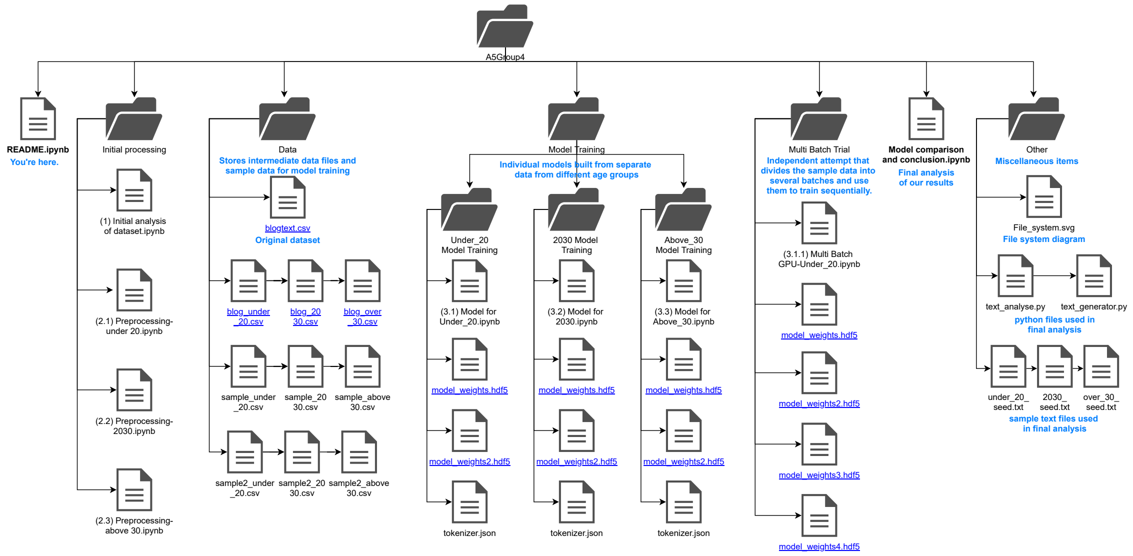
Below is a diagram of the file system of this project; a corresponding Google Drive link is provided if it is not included in the submission file due to its size. (All the .pdf files are not shown here.)

2.1 Note:

In case the display is too small for reading in this notebook, please open **File_system.svg** inside the `Other` folder manually.

[ ]:

```
from IPython.core.display import SVG
SVG(filename='Other/File_system.svg')
```



3 Recommended reading order

1. Subsequent to this notebook, one can start by looking at the `Initial processing` folder, which contains notebooks for analysis and preprocessing of the dataset:
   - **(1) Initial analysis of dataset.ipynb** provides an overview of the data we use and some data splitting.
   - The other three notebooks starting with (2.x) prefix are further data preprocessing for each age group.

Note: As the three (2.x) notebooks have almost the same structure, only *(2.1) Preprocessing-under20.ipynb* will contain detailed explnations of the steps involved.

2. The `Data` folder is simply output from the notebooks of the `Initial processing` folder and therefore is not of importance here.

3. The `Model Training` folder contains training noteboks and some related output files for each age group's data.

Note: As the three training notebooks have almost the same structure, only *(3.1) Model for Under_20.ipynb* will contain detailed explnations of the steps involved.

4. The `Multi Batch Trial` is intended to record one of our attempts that divided a large sample data file into different groups and used them to train the model sequentially. It's worth noting that this attempt was only applied on the sample data of the under-20 age group (please refer to the corresponding notebook for more detailed reasons). Moreover, the attempt is independent from the ones inside the `Model Training` and can be ignored if one is only interested in the main part of the project.

5. Finally, the **Model comparison and conclusion.ipynb** loads the previously trained models from the .hdf5 files within the folder `Model Training` and shows predictions of each model, from which some analysis and conclusions will be drawn.