**Documentation**

The aim of this file is to describe what each file does and where it is used, how to run the files and which files/paths are required to be modified.

The files must be run in **Jupyter Notebook** and **Python 3.8.**

**Note:** All the files require various installations of libraries. Most of the installations are in the form (!pip install numpy) or similar and have been commented out (remove the comment and perform the installation). However, there is a chance that I may have missed something, which would lead to an error **explaining what must be installed**.

**Note-2:** The **InsertDirtyData.ipynb** must be run before the **TweetsPython.ipynb** file. Since the first inserts errors in the dataset.

**Handwriting Test Images** – folder containing images with my handwriting that were used for the experiments in the *Handwritten Data Recognition.ipynb* file.

**Letters** – contains the ready test and training alphabet letters, meaning that there is no need to rerun the code.

**RealWorldDocumentsCollections** – contains the email dataset and others used in the *TextRecognition.ipynb* file.

**Segmented\_img** – it is a folder location where the handwritten image is segmented to its character images (if there are any images currently, please delete them).

**Tesseract-OCR** – it is the installation of Tesseract which is used in the *TextRecognition.ipynb*, it can be used if you cannot install it.

**A\_Z-Handwritten-Data** – dataset file with the csv representations of the letter images.

**Cleaned\_English\_Tweets** – dataset output from the *TweetsPython.ipynb* file, which shows the cleaned tweets dataset.

**Contractions –** file containing all of the contractions used in the *TweetsPython.ipynb* file.

**DirtyTweetsDataset** – the output file from the *InsertDirtyData.ipynb,* which is a dataset with manually inserted dirty data.

**Email\_Output** – output dataset file from the *TextRecognition.ipynb* file.

**Functions** – this file contains all of the functions used in the *Handwritten Data Recognition.ipynb*

**InsertDirtyData** – script that inserts dirty data into the sample 1000 data in *Tweets1000* file*.*

**My\_model.pth** – this is a ready CNN model with 3 epochs trained, could be used if you don’t want to waste 2-3 hours on training.

**SlangWordMeaning** – text file containing all slang words and their meanings, used in the *TweetsPython.ipynb* file.

**Text\_After\_NLP** – file containing the prepared tweets dataset output from the *TweetsPython.ipynb file.*

**TextRecognition** – script that performs text recognition through Tesseract-OCR.

**Tweets1000** – file containing 1000 sample Tweets data.

**TweetsPython** – script performing the NLP and cleaning tasks on the tweets dataset.

**TweetsWorkbookCSV** – full tweets dataset.