Embedding + (LR, RF, XGBoost) - READMe

-----------------------

Description: This folder contains a Google Colab notebook "Embedding + (Log Regression + Random Forest + XGBoost).ipynb", with which several, in the name indicated, experiments were conducted.

-----------------------

Prerequisites:

- Logged in with a Google account, with access to Google Colab, Google Colab's CPUs and GPUs (subscription needed to have prioritised access GPU's for a significant longer time) (for this notebook, you only need CPUs)

- Granting permission to run each of our Google Colab notebooks and python files.

- Mount drive at /content/drive/ [This is done when running the notebook, see below]

- The folders/files regarding Preprocessed training and validation datasets are at the described location.

-----------------------

How to run:

Before running, the user can specify

- which preprocessed datasets they want to work with

- which embedding they want to work with

In the first section, there is a variable

PREPROCESSING\_CHOICE defined by PREPROCESSING\_OPTIONS[INDEX], where INDEX is an integer [0;8]. The user can choose INDEX from that range, which corresponds to the preprocessed dataset.

In the first section, there is also a variable "embedding\_choice", an integer from [1;5], where each integer encodes an embedding (as described in the comments)

After selecting INDEX and embedding\_choice:

Run all sections, top to bottom, up to (and exclusivly) the section "Choose and train model".

In that section, the user has to decide which experiment they want to conduct.

It is highly recommended only to run one of the 3 subsections, from top to bottom, due to potentially memory limit. Furthermore, running multiple experiments consecutively without restarting runtime could result in wrong values.

-----------------------

Output:

In the directory /content/drive/CIL22022, the trained models are stored as file (if chosen to save the mode) and the performance of the chosen model is logged.