Phase 2 MLP Multilabel Model Training Documentation

ST1506: DSDA FINAL YEAR PROJECT

COMPANY PROFILING TEXT MINING

August 2021

Synopsis

The purpose of this documentation is to demonstrate the steps a user should take to successfully run the ipython file for model training.

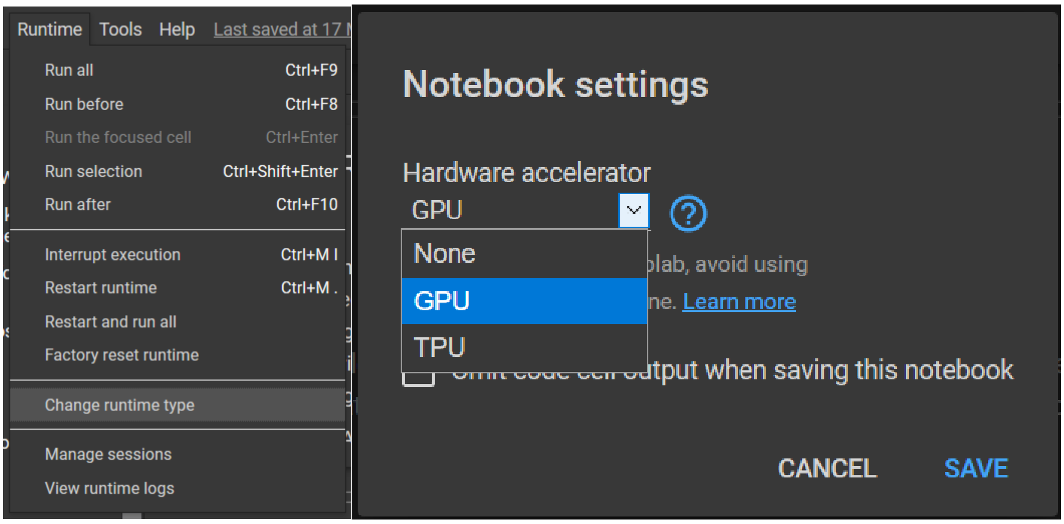
This documentation is for data scientists who want to see what the process of training the model is, and possibly to use their own data to train the model.

Table of Contents

Running The Notebook - GPU 4

Running The Notebook - GPU

1. Upload “NLP_MLP_Multilabel.ipynb” to Google Collaboratory (i.e. Google Colab).
2. Under the Runtime tab, change runtime type to “GPU” and click “SAVE”. This will allow for faster running times when running the notebook so that time can be saved.



Notebook settings

Hardware accelerator
GPU

lab, avoid using
ne. Learn more

□ tput when saving this notebook

Figure 1: Changing Runtime to GPU

3. Connect to Colab runtime by clicking on the connect button found on the top right bar if not already converted. You should see the following if the connection is successful.

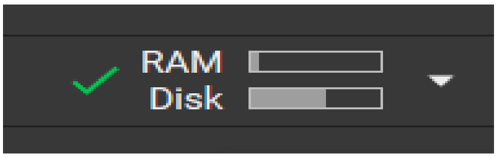
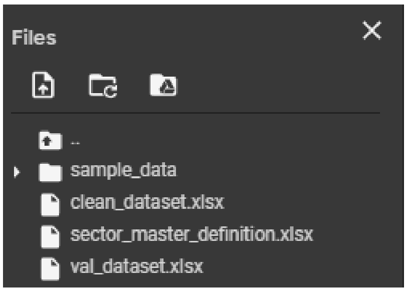


Figure 2:Ensuring runtime is connected

4. On the left tab, click on the icon of a folder and upload “clean_dataset.xlsx”, “sector_master_definition.xlsx” and “val_dataset.xlsx”. Once uploaded you should see that the files are in the temporary working directory in Google Colab.

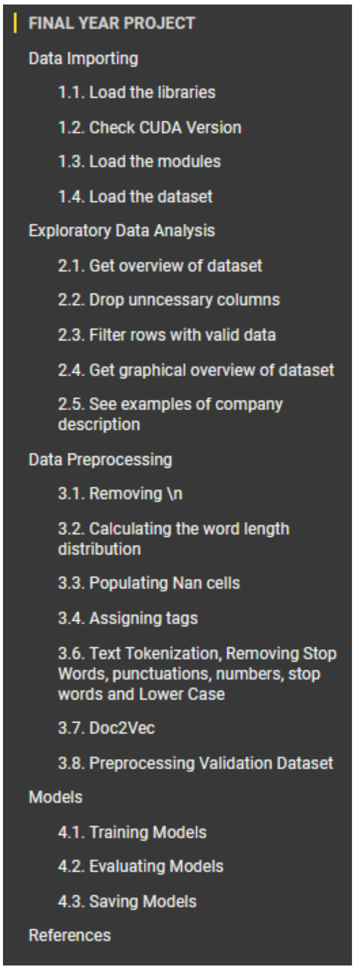


Files

..
sample_data
clean_dataset.xlsx
sector_master_definition.xlsx
val_dataset.xlsx

Figure 3: Uploaded excel data files

5. Now, the notebook is ready to run. Run the notebook cells one at a time starting from the start of Section 1, labelled as “Data Importing” until Section 4.2, labelled as “Evaluating Models”



FINAL YEAR PROJECT

Data Importing

1.1. Load the libraries

1.2. Check CUDA Version

1.3. Load the modules

1.4. Load the dataset

Exploratory Data Analysis

2.1. Get overview of dataset

2.2. Drop unncessary columns

2.3. Filter rows with valid data

2.4. Get graphical overview of dataset

2.5. See examples of company description

Data Preprocessing

3.1. Removing \n

3.2. Calculating the word length distribution

3.3. Populating Nan cells

3.4. Assigning tags

3.6. Text Tokenization, Removing Stop Words, punctuations, numbers, stop words and Lower Case

3.7. Doc2Vec

3.8. Preprocessing Validation Dataset

Models

4.1. Training Models

4.2. Evaluating Models

4.3. Saving Models

References

Figure 4: List of Contents to run

6. To save the model, run *Section 4.3: Saving Model* of the notebook.