Configuration Manual

Artifacts

Code

- HS_classification.ipynb
- HSCODE_Classification_FAISS.ipynb

Dataset

- DailyfiledFeb2023.xlsx (unzip file)

Requirements

Required Libraries::

- Python version 3.10.12
- Pandas version 1.5.3
- TensorFlow version 2.14.0
- CUDA version 12.0

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 525.105.17   Driver Version: 525.105.17   CUDA Version: 12.0     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  Tesla T4            Off  | 00000000:00:04.0 Off |                    0 |
| N/A   56C    P8    10W /  70W |      0MiB / 15360MiB |      0%      Default |
|                               |                      |                  N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Execution

Change the variable path to locate the datasets the file DailyfiledFeb2023.xlsx
path = '/content/drive/My Drive/Research_project/'

Execute the first file called HS_classification.ipynb. This script will execute RoBERTa classification and DistilBERT classification. It will create 3 files:

- CleanData.csv. The file contains cleaned record to be used in Roberta classification and Distilbert classification

- Once Roberta classification is executed will create an output file called robertaPred.csv. This file contains the prediction of the first 2 digits of the HS code by using Roberta classification
- Once Dilstilbert classification is executed will create an output file called distilbertPred.csv. This file contains the prediction of the first 2 digits of the HS code by using Distilbert classification

The previous generated files will be the input for the second part of the algorithm. Make sure the files are in the declared path before executing the second file.

The second file called "HSCODE_Classification_FAISS.ipynb" implements the similar search using FAISS. Execute the script, the first part of the script will prepare the datasets to later train the model.
Once the model is trained it will execute the validation section, this section is divided in two parts:

- Bert process will receive as input the file distilbertPred.csv. It will take 2 columns: first 2 digits classification and good's description to forecast the 6-digits HS code.
- RoBERTa process will received as input the file robertaPred.csv. It will use two columns: the first 2 digits and good's description to forecast the 6-digits HS code.

These scripts were executed using Colab research google platform as a Pro member to access to GPU resources. Otherwise, the classification (Roberta and Distilbert) may take a longer time to be completed.