How to use this project

This guide will help you understand how to use this set of script that we have developed for months.

## ****Preparation****

### First of all, we need to install Python 3.9.13

from <https://www.python.org/downloads/release/python-3913/>

表格

描述已自动生成

Scroll down to the bottom of the page and select a good version for your system😊

Important:

Check “Add Python 3.9 to PATH” when open the installer!!

### Install an IDE

You can run Python code directly through terminal, but I would recommend using an IDE instead. I’d prefer VS Code (<https://code.visualstudio.com/Download>) to run our project.

Please follow this link (<https://code.visualstudio.com/docs/languages/python>) to Section Run Python code to check if python can run correctly.

*If you encounter any problem during this step, don’t hesitate to email Yuhan about your questions.*

### File Structures

Here is a list of files in the repository:

文本

描述已自动生成

There are 4 python script we are going to run.

1. read\_article\_from\_xls\_training\_set.py
2. train\_regression\_model.py
3. read\_mhtml\_to\_object.py
4. score\_object\_using\_trained\_model.py

### Open this project in VS Code

Once you opened VS Code, you will see a window like this:

电脑屏幕的截图

描述已自动生成

Click “File – Open Folder” to select the folder that contains the entire project.

If the project loaded successfully, the window is like this:

电脑的屏幕截图

描述已自动生成

### Install dependencies

In terminal, enter the command below:

*pip install -r requirements.txt*

文本

描述已自动生成

## Train a machine learning model using scored articles (Training Set)

### Pre-process spreadsheet

Open read\_article\_from\_xls\_training\_set.py. To run this script, we need a spreadsheet that contains all scored data, and put it in folder “data”.

Please make sure the file only contains one spreadsheet.

文本

描述已自动生成

After excel file is placed in data folder, modify the path in Line 23 to the filename you are using for training. Then, make sure anything inside a square bracket is corresponding to an exact column name in the spreadsheet.

图形用户界面, 文本, 应用程序

描述已自动生成

Now it’s time to run the script! You can either hit F5 on keyboard or click “Run – Start Debugging”.

图形用户界面, 应用程序

描述已自动生成

This script will take some time to complete, because it will extract features for each article by our tagging\_core. *Script that extracts features always takes time.*

### Train the model

文本

描述已自动生成

These json files are what we’ve got from the previous step. We are going to use them to train our model.

Open train\_regression\_model.py and run. It will train a regression model and save it as “AdaBoostRegressor.joblib” in the folder. You will see a score for this model printed in the terminal. You can run this multiple times to get a score you like, and then proceed to the next step.

We are going to use this model to evaluate everything we downloaded.

## Evaluate all articles

### Read articles from mhtml files

Open read\_mhtml\_to\_object.py and run. It will prompt a window for you to select a folder. You can select any folder that contains mhtml files, or even a folder contains a folder that have mhtml files in it. This script will recursively check mhtml files in the select folder, and generate a json file for each mhtml in the same folder.

We will use these json files later.

Example:

Double click on the folder

电脑萤幕截图

描述已自动生成

Make sure folder name is in the box, then confirm

电脑萤幕的截图

描述已自动生成

This is just an example. You can select each country separately for divided tasks.

This step takes a lot of time because of feature extraction. If the script terminated by accident, just run again with the same folder. It will automatically continue the work😊

Also, we can split the task by folders (or countries, because we have separate folders for countries) for everyone in the group.

### Evaluate

Open score\_object\_using\_trained\_model.py and run.

Select a same folder as 3.1 (or any folder with processed article json files). It will go through everything and save results to [mhtml\_filename].csv. These CSV files will in the same folder as target mhtml file.

表格

描述已自动生成

文本

描述已自动生成

Mhtml -> json -> csv

Now you can use scores in csv files to do anything you like😊