Phase 2 Engine 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 use the Engine. This documentation is for any potential user of the Engine.

This engine uses the BERT-uncased model.

Technical Terms

 OPTIONAL: These are usually denoted beside the steps bulleted points. When labelled beside a numbered list, users may omit that step depending on the conditions/scenarios further explained with an amber warning box.

Table of Contents

Running The Engine – GPU	. 4
Checking folder items	. 4
Activating Conda Environment	. 5
Preparing libraries and dataset	. 7
Execution Of Engine	.9
Appendix A	10
Unable to install packages due to an EnvironmentError	10

Running The Engine – GPU

Checking folder items

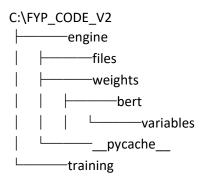


Note: The model weights were trained with GPU. Therefore, we advise users to use anaconda, which has the necessary CUDA files.



Note: This guide is based on Windows OS. Not all steps may be the same on other operating systems.

- 1. Download and unzip the project zip file "fyp_code_vX" in your C Drive (C:), where vX is the version number.
- 2. Please ensure that the folder structure is as follows or contains:



Alternatively, you may wish to replicate the following commands to do a quick check in your powershell. Assuming that "fyp_code_vX" is located in your C Drive (C:),

Activating Conda Environment



Note: Before further reading, please ensure that Anaconda has been installed in your device, and that your device has a recognised NVIDIA GPU.

 Firstly, in the windows search bar, search for "Anaconda Powershell Prompt (anaconda3)".

Select "Run as Administrator".

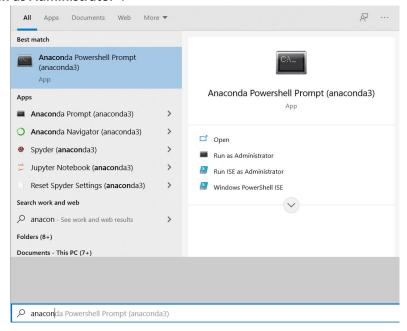


Figure 1: Search for Anaconda Powershell Prompt (anaconda3)

- 2. Once done, assuming you have saved the folder earlier on in the root directory of the C Drive (C:), cd to the folder such that the directory is **C:\FYP_CODE_VX.**
- 3. To ensure that the current working directory is clean, run

conda create -n fypEngine python -y

to create a fresh conda environment, so that your default anaconda environment is not contaminated with the libraries required to run the code.

Figure 2: Snippet of the creation of the environment

4. Once done, run **conda activate fypEngine** to activate the environment you have just created. Notice that once done execution, the bracketed text content beside the final command line will change to "fypEngine".

```
(base) PS C:\fyp_code_v2> conda activate fypEngine
(fypEngine) PS C:\fyp_code_v2> _
```

Figure 3: Notice the bracketed text changes from base to fypEngine upon activation of the environment



Note: The following steps are the methods to install the CUDNN libraries. If you have the libraries included, please omit the bulleted points marked as OPTIONAL.

5. OPTIONAL: Run conda install -c conda-forge cudnn -y to install libraries required for GPU usage. Wait for installation to complete.

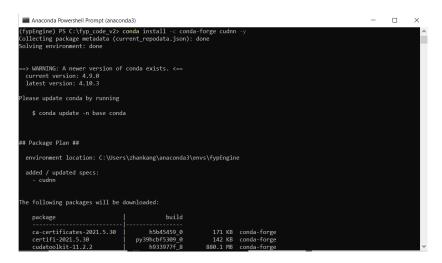


Figure 4: Snippet of installing the CUDNN packages

6. Once done, run the following command to ensure that the latest conda version is being used.

conda update -n base conda -y

7. Once done, run the following command to make a new python virtual environment. This is to ensure that the install python libraries do not dirty your python environment, and that there are no conflict with existing packages that may cause the engine to fail.

```
python -m venv env
.\env\Scripts\activate
```

You should now see that a green bracketed "env" should be beside the bracketed "fypEngine" also.

```
(fypEngine) PS C:\fyp_code_v2> python -m venv env
(fypEngine) PS C:\fyp_code_v2> .\env\Scripts\activate
(env) (fypEngine) PS C:\fyp_code_v2>
```

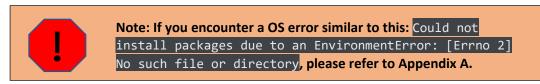
Figure 5: A new green env icon has been added

Preparing libraries and dataset

1. In order to install the required libraries that is necessary to run the engine, we will have to install the libraries declared in the requirements.txt file. To do so, run the following command and wait for it to complete.

pip install -r requirements.txt

Figure 6: Snippet of installing libraries defined in requirements.txt



2. Once done install the required libraries, ensure that there is an "input.xlsx" file in the folder fyp_code_vX/engine/files in the project folder.

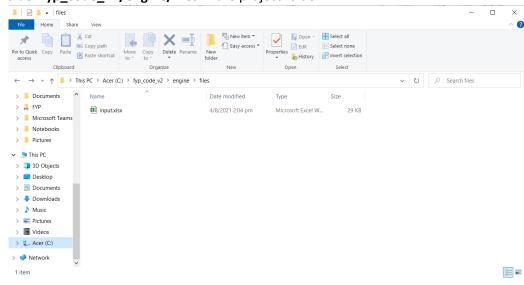


Figure 7: Example of where the input.xlsx should be at

Please also ensure that there is a "sector_master_definintion.xlsx" file in the folder fyp_code_vX/engine in the project folder.

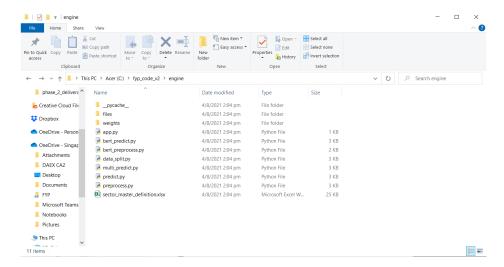


Figure 8: Example of where the sector_master_definition.xlsx should be at



Note: The following step number 3 is optional. This is valid for users who wish to create a custom input file.

3. OPTIONAL: If the user wishes to create a custom input file, do follow format of the already existing "input.xlsx" file, and make sure that the custom file is in the directory fyp_code_vX/engine/files.

Do note that if the file contains invalid entries, it will be automatically removed and will not show up in the "output.xlsx" file.



Figure 9: Snippet example of input.xlsx

Execution Of Engine

To run the Engine, simply run the following code in powershell terminal setup previously
 cd engine
 python app.py

The code may take long to run as the model is predicting.

2. Once done, the program will end in the terminal. Below is an example of what should be reflected on the terminal once the program has run finish.

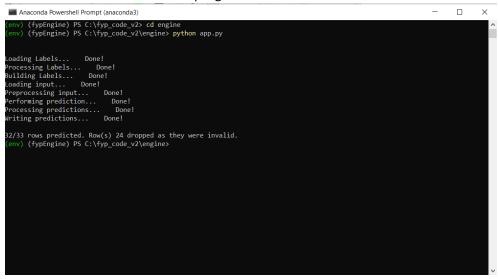


Figure 10: Screenshot of the execution of the program once completed.

The result of the program will be an output file labelled as "output.xlsx". The file can be found in the directory fyp_code_vX/engine/files. The output file will include the filled values for the respective columns: Sector, Subsector, Archetype and Valuechain.



Figure 11: Snippet of how it looks like in output.xlsx



Note: The following step number 3 is optional. This is valid for users who wish to predict results based on another separate excel file.

3. OPTIONAL: To do another prediction with another excel file, simply repeat from step 2 of the previous section: Preparing libraries and dataset.

Appendix A

Unable to install packages due to an EnvironmentError

If users encounter an OS error stating that the package could not be installed as there is
no such file or directory when running the command in point 1 of <u>Preparing libraries</u>
<u>and dataset</u>, it is most likely due to the file path being too long. To solve this, search up
the Registry Editor in the search bar. When promoted to allow this app to make changes
to your device, click Yes.

Copy paste the following path into the search bar of the Registry Editor found at the top and replace with the word "Computer". Once done, press enter.

Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\FileSystem

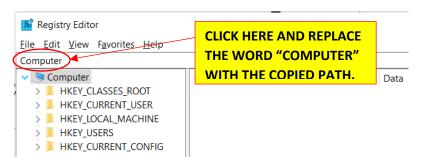


Figure 12: Snippet of where the search bar is at

2. Once done, scroll to find the row name labelled as "LongPathsEnabled". Right click it and click Modify.

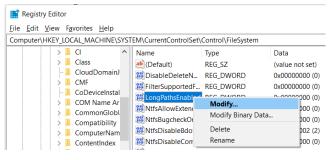


Figure 13: Snippet of selecting the row name and clicking on Modify

3. Then change the value of **Value data** from 0 to 1. Click Ok.

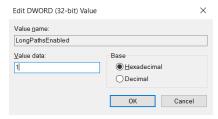


Figure 14: Changing LongPaths binary value from 0 to 1

4. Once completed, run the following command to deactivate the env and delete the environment folder with all its contents.

First run this command to get the file path of where the env folder is saved at python -c "import os; print(os.environ['VIRTUAL_ENV'])"

Once done, copy the output file path.

```
Anaconda Powershell Prompt (anaconda3)

(env) (fypEngine) PS C:\fyp_code_v2\engine> python -c "import os; print(os.environ['VIRTUAL_ENV'])"

C:\fyp_code_v2\engine\env
(env) (fypEngine) PS C:\fyp_code_v2\engine> _
```

Figure 15: Example of the output

In the powershell terminal, type deactivate.

Followed by this this following command, whereby **FILE_PATH>** is the output file path as copied previously.

```
Remove-Item '<FILE_PATH>'
```

And when prompted if children and recurse parameters are not specified, type **Y** and click enter. The env folder is now deleted.

```
(env) (fypEngine) PS C:\fyp_code_v2\engine> deactivate
(fypEngine) PS C:\fyp_code_v2\engine> rm -rf C:\fyp_code_v2\engine\env
Remove-Item : A parameter cannot be found that matches parameter name 'rf'.
At line:1 chan:4
+ rm -rf C:\fyp_code_v2\engine\env
+ CategoryInfo : InvalidArgument: (:) [Remove-Item], ParameterBindingException
+ FullyQualifiedErrorId : NamedParameterNotFound, Microsoft.PowerShell.Commands.RemoveItemCommand

(fypEngine) PS C:\fyp_code_v2\engine> Remove-Item 'C:\fyp_code_v2\engine\env'

Confirm
The item at C:\fyp_code_v2\engine\env has children and the Recurse parameter was not specified. If you continue, all children will be removed with the item. Are you sure you want to continue?

[Y] Yes [A] Yes to All [N] No [L] No to All [S] Suspend [?] Help (default is "Y"): Y
(fypEngine) PS C:\fyp_code_v2\engine> _
```

Figure 16: Example of how step 4 is executed

5. Once completed, close the Anaconda Powershell Prompt and restart from the section **Activating Conda Environment.**



Note: If users are worried that step 3 of this appendix may affect the file system in future, they can change the binary back to 0 after running the engine.