

Phase 1 Engine Documentation

The purpose of this documentation is to demonstrate the steps a user should take to successfully use the Engine, as well as interpret the outputs of the Engine. This documentation is for any potential user of the Engine. **The engine uses the Multilabel MLP.**

Instructions

1. Download and unzip project zip file: “fyp_code_vX”, where vX is the version number.

1.1. Please ensure the folder has the following content before proceeding:

A. Engine

- __pycache__
- files
- weights
- app.py
- predict.py
- preprocess.py
- sector_master_definition.xlsx

B. .gitignore

C. LICENSE

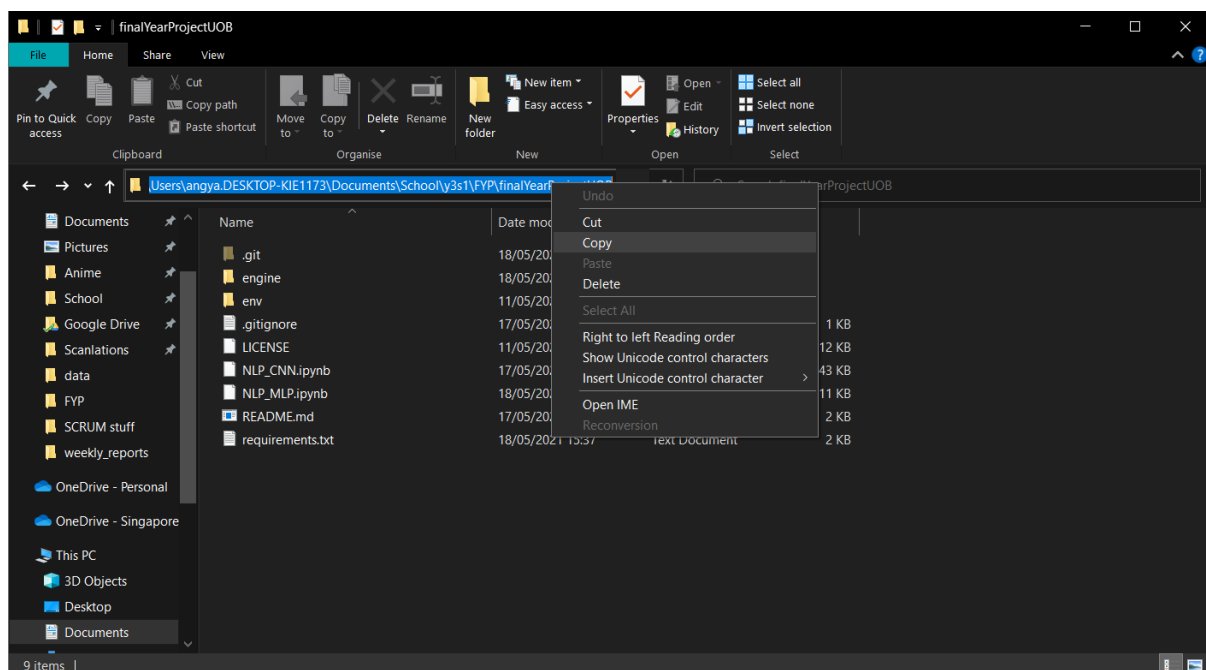
D. NLP_CNN.ipynb

E. NLP_MLP.ipynb

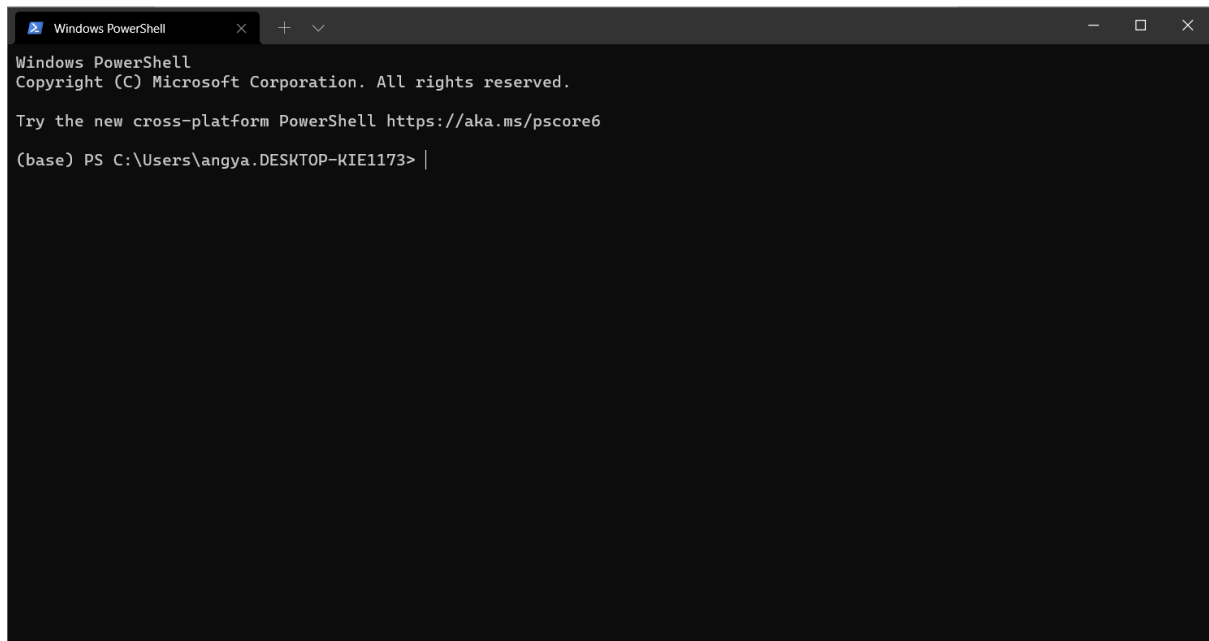
F. README.md

G. requirements.txt

2. Navigate into unzipped project folder on your file explorer.
3. Copy path to project folder as such:



4. Open Anaconda Prompt or Powershell.



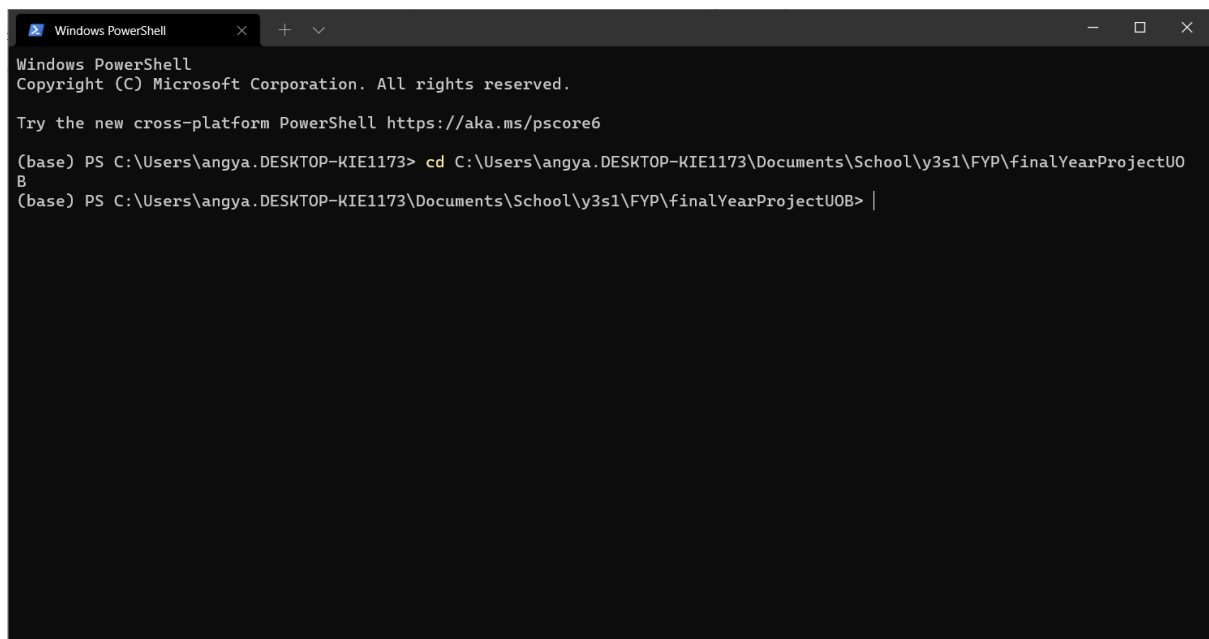
A screenshot of a Windows PowerShell terminal window. The title bar shows 'Windows PowerShell' with standard window controls. The terminal text includes the Windows PowerShell logo, copyright information for Microsoft Corporation, a link to the new cross-platform PowerShell, and the current prompt: (base) PS C:\Users\angya.DESKTOP-KIE1173> |

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

(base) PS C:\Users\angya.DESKTOP-KIE1173> |
```

5. Run `cd <project_path>`, where <project_path> is replaced with the copied path from the previous step.



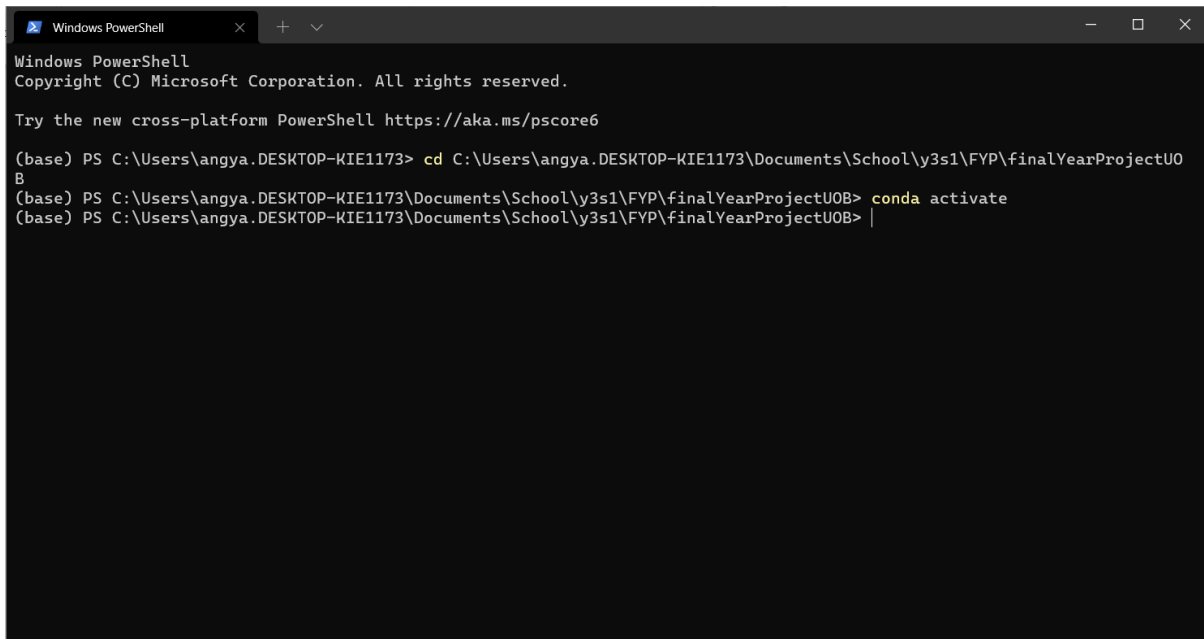
A screenshot of a Windows PowerShell terminal window, similar to the one above. It shows the same initial text, followed by the execution of the 'cd' command with a specific path. The prompt now reflects the new directory: (base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectUOB> |

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

(base) PS C:\Users\angya.DESKTOP-KIE1173> cd C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectUOB
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectUOB> |
```

6. Run **conda activate** to ensure that the Anaconda environment is enabled. (There should be a (base) visible if it is enabled)



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

(base) PS C:\Users\angya.DESKTOP-KIE1173> cd C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0
B
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda activate
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> |
```

DISCLAIMER: PLEASE RUN STEP 7 TO 10 IF NVIDIA CUDNN PACKAGE IS NOT INSTALLED IN YOUR DEVICE BEFORE. ONLY RUN IF YOUR DEVICE HAS AN NVIDIA GPU.

7. **OPTIONAL:** Run `conda create -n temp python` to create a fresh conda environment, so that your default anaconda environment is not dirtied with the libraries required to run the code. Press **y** then **Enter** when prompted.

```
Windows PowerShell
(base) PS C:\Users\angya.DESKTOP-KIE1173> conda create -n temp python
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\angya.DESKTOP-KIE1173\anaconda3\envs\temp

added / updated specs:
- python

The following packages will be downloaded:

package | build | size
-----|-----|-----
certifi-2020.12.5 | py39h... | 141 KB
pip-21.0.1 | py39h... | 1.8 MB
python-3.9.4 | h6244533_0 | 16.4 MB
setuptools-52.0.0 | py39h... | 725 KB
tzdata-2020f | h52ac0ba_0 | 113 KB
wincertstore-0.2 | py39h2bbff1b_0 | 15 KB
-----|-----|-----
Total: | | 19.2 MB

The following NEW packages will be INSTALLED:

ca-certificates | pkgs/main/win-64::ca-certificates-2021.4.13-haa95532_1
certifi | pkgs/main/win-64::certifi-2020.12.5-py39h...
openssl | pkgs/main/win-64::openssl-1.1.1k-h2bbff1b_0
pip | pkgs/main/win-64::pip-21.0.1-py39h...
python | pkgs/main/win-64::python-3.9.4-h6244533_0
setuptools | pkgs/main/win-64::setuptools-52.0.0-py39h...
sqlite | pkgs/main/win-64::sqlite-3.35.4-h2bbff1b_0
tzdata | pkgs/main/noarch::tzdata-2020f-h52ac0ba_0
vc | pkgs/main/win-64::vc-14.2-h21ff451_1
vs2015_runtime | pkgs/main/win-64::vs2015_runtime-14.27.29016-h5e58377_2
wheel | pkgs/main/noarch::wheel-0.36.2-pyhd3eb1b0_0
wincertstore | pkgs/main/win-64::wincertstore-0.2-py39h2bbff1b_0

Proceed ([y]/n)? |
```

8. **OPTIONAL:** Wait until the setup is complete.
9. **OPTIONAL:** Run `conda activate temp` to activate the created anaconda environment.
(There should be a (temp) visible if it is successfully enabled)

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

(base) PS C:\Users\angya.DESKTOP-KIE1173> cd C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda activate
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda activate temp
(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> |
```

10. **OPTIONAL:** Run `conda install -c conda-forge cudnn` to install libraries required for GPU usage. Press `y` then `Enter` when prompted. Wait for installation to complete.

```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda install -c conda-forge cudnn
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: C:\Users\angya.DESKTOP-KIE1173\anaconda3\envs\temp

  added / updated specs:
    - cudnn

The following packages will be downloaded:



| package                   | build          | size   | channel     |
|---------------------------|----------------|--------|-------------|
| ca-certificates-2020.12.5 | h5b45459_0     | 173 KB | conda-forge |
| certifi-2020.12.5         | py39hcbf5309_1 | 144 KB | conda-forge |
| openssl-1.1.1k            | h8ffe710_0     | 5.7 MB | conda-forge |
| python_abi-3.9            | 1_cp39         | 4 KB   | conda-forge |
| Total:                    |                | 6.1 MB |             |



The following NEW packages will be INSTALLED:

  cudatoolkit      conda-forge/win-64::cudatoolkit-11.2.2-h933977f_8
  cudnn            conda-forge/win-64::cudnn-8.1.0.77-h3e0f4f4_0
  python_abi       conda-forge/win-64::python_abi-3.9-1_cp39

The following packages will be UPDATED:

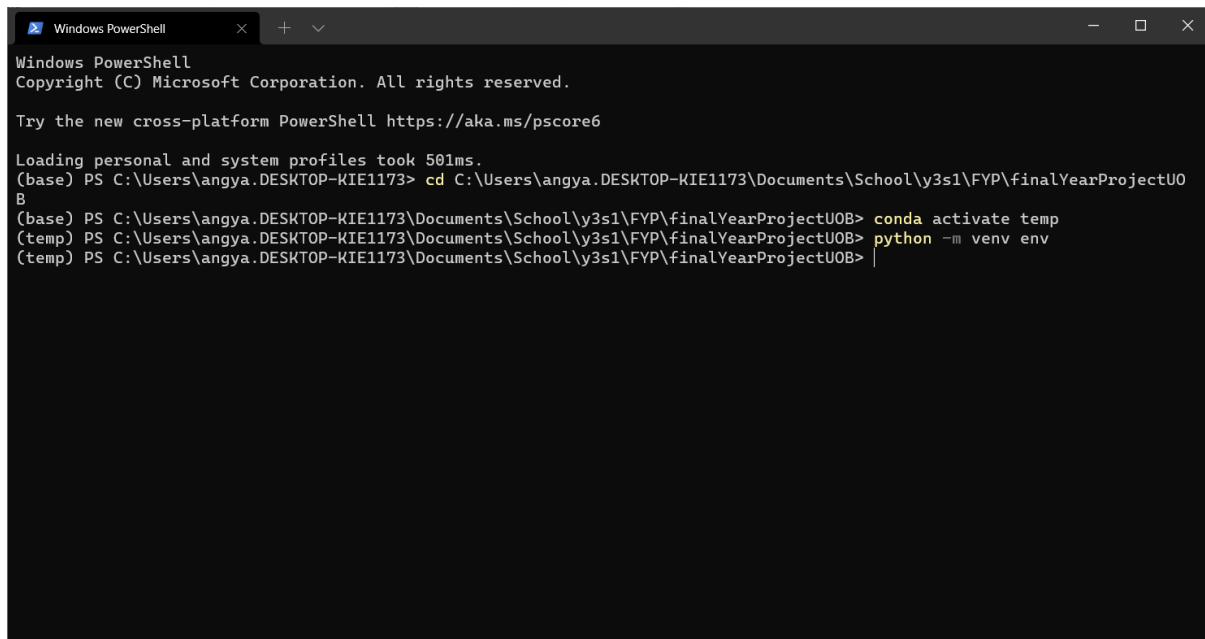
  certifi          pkgs/main::certifi-2020.12.5-py39haa9~ --> conda-forge::certifi-2020.12.5-py39hcbf5309_1

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates  pkgs/main::ca-certificates-2021.4.13~ --> conda-forge::ca-certificates-2020.12.5-h5b45459_0
  openssl          pkgs/main::openssl-1.1.1k-h2bbff1b_0 --> conda-forge::openssl-1.1.1k-h8ffe710_0

Proceed ([y]/n)? y
```

11. Run `python -m venv env` to make a new python virtual environment. This is to ensure that the installed python libraries do not dirty your python environment, and that there are no conflicts with existing packages that can cause the Engine to fail.

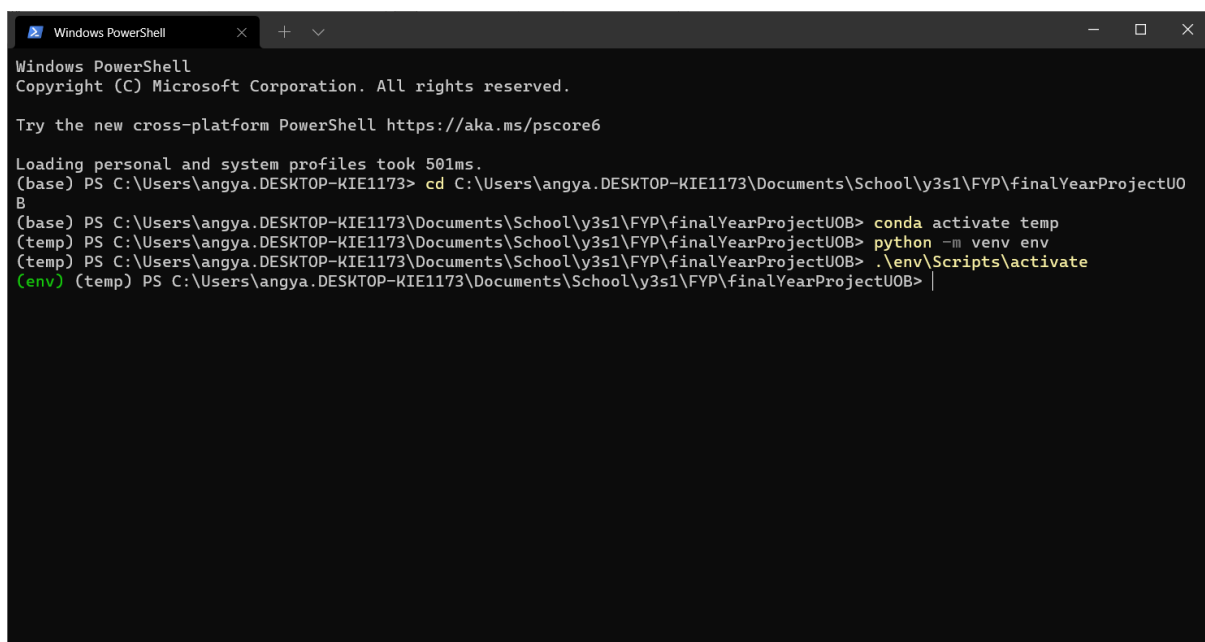


```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

Loading personal and system profiles took 501ms.
(base) PS C:\Users\angya.DESKTOP-KIE1173> cd C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda activate temp
(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> python -m venv env
(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> |
```

12. Run `.\env\Scripts\activate` to activate the created python virtual environment. (There should be a (env) visible.)



```
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/pscore6

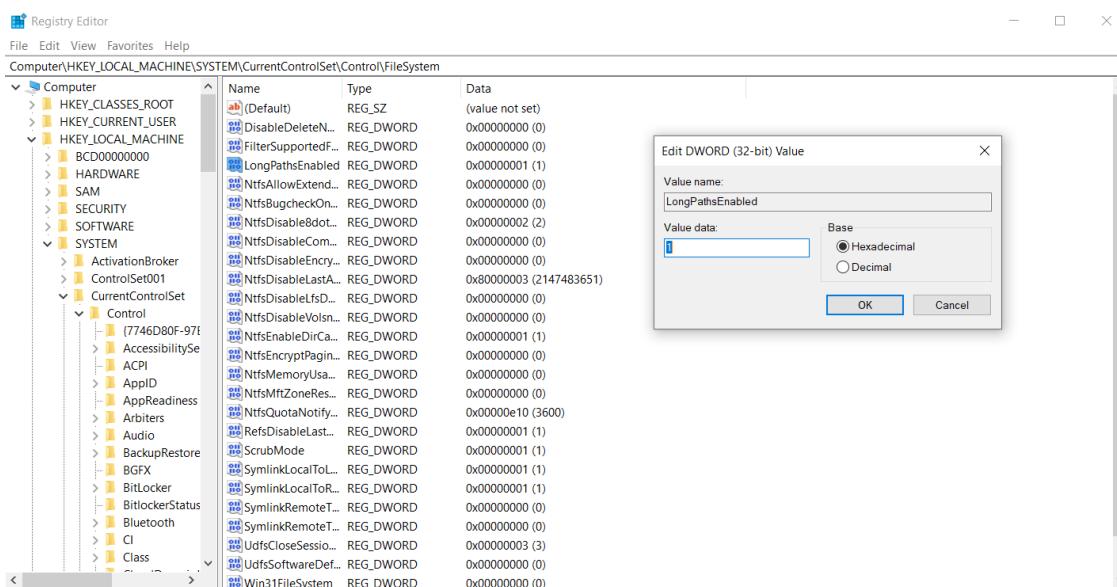
Loading personal and system profiles took 501ms.
(base) PS C:\Users\angya.DESKTOP-KIE1173> cd C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B
(base) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> conda activate temp
(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> python -m venv env
(temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> .\env\Scripts\activate
(env) (temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> |
```

13. Run `pip install -r requirements.txt` to install required libraries. Wait for installation to finish.

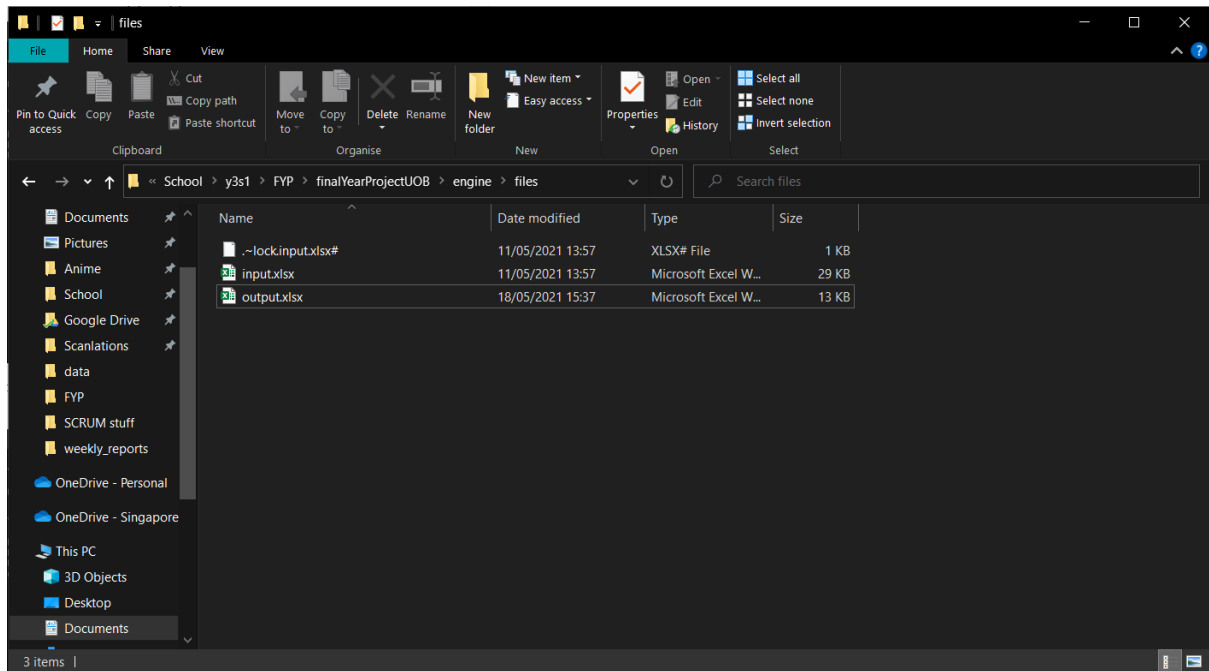
```
Windows PowerShell
Collecting xlrd==2.0.1
  Using cached xlrd-2.0.1-py2.py3-none-any.whl (96 kB)
Collecting wheel<1.0,>=0.23.0
  Using cached wheel-0.36.2-py2.py3-none-any.whl (35 kB)
Requirement already satisfied: setuptools>=40.3.0 in c:\users\angya.desktop-kie1173\documents\schoo\y3s1\fyf\finalyearprojectuob\env\lib\site-packages (from google-auth==1.30.0->-r requirements.txt (line 17)) (49.2.1)
Installing collected packages: six, absl-py, wheel, astunparse, numpy, blis, cachetools, catalogue, certifi, chardet, cl
  ick, fastlock, cupy-cuda101, cymem, pydantic, murmurhash, pyparsing, packaging, tqdm, srsly, wasabi, typer, preshed, id
  na, urllib3, requests, thinc, MarkupSafe, Jinja2, spacy-legacy, smart-open, pathy, spacy, en-core-web-lg, et-xmlfile, fi
  lellock, flatbuffers, gast, pyasn1, rsa, pyasn1-modules, google-auth, oauthlib, requests-oauthlib, google-auth-oauthlib,
  google-pasta, grpcio, h5py, joblib, scipy, PyYAML, Keras, keras-nightly, Keras-Preprocessing, Markdown, openpyxl, opt-ei
  nsum, pytz, python-dateutil, pandas, protobuf, regex, sacremoses, spacy-alignments, spacy-lookups-data, typing-extension
  s, torch, tokenizers, transformers, spacy-transformers, tensorboard-plugin-wit, Werkzeug, tensorboard-data-server, tenso
  rboard, tensorflow-estimator, wrapt, termcolor, tensorflow, xlrd
Successfully installed Jinja2-2.11.3 Keras-2.4.3 Keras-Preprocessing-1.1.2 Markdown-3.3.4 MarkupSafe-1.1.1 PyYAML-5.4.1
Werkzeug-1.0.1 absl-py-0.12.0 astunparse-1.6.3 blis-0.7.4 cachetools-4.2.2 catalogue-2.0.4 certifi-2020.12.5 chardet-4.0
.0 click-7.1.2 cupy-cuda101-8.6.0 cymem-2.0.5 en-core-web-lg-3.0.0 et-xmlfile-1.1.0 fastlock-0.6 filelock-3.0.12 flatbu
ffers-1.12 gast-0.4.0 google-auth-1.30.0 google-auth-oauthlib-0.4.4 google-pasta-0.2.0 grpcio-1.34.1 h5py-3.1.0 idna-2.1
0 joblib-1.0.1 keras-nightly-2.5.0.dev2021032900 murmurhash-1.0.5 numpy-1.19.5 oauthlib-3.1.0 openpyxl-3.0.7 opt-einsum-
3.3.0 packaging-20.9 pandas-1.2.4 pathy-0.5.2 preshed-3.0.5 protobuf-3.16.0 pyasn1-0.4.8 pyasn1-modules-0.2.8 pydantic-1
.7.3 pyparsing-2.4.7 python-dateutil-2.8.1 pytz-2021.1 regex-2021.4.4 requests-2.25.1 requests-oauthlib-1.3.0 rsa-4.7.2
sacremoses-0.0.45 scipy-1.6.3 six-1.15.0 smart-open-3.0.0 spacy-3.0.6 spacy-alignments-0.8.3 spacy-legacy-3.0.5 spacy-lo
okups-data-1.0.0 spacy-transformers-1.0.2 srsly-2.4.1 tensorboard-2.5.0 tensorboard-data-server-0.6.1 tensorboard-plugin
-wit-1.8.0 tensorflow-2.5.0rc3 tensorflow-estimator-2.5.0rc0 termcolor-1.1.0 thinc-8.0.3 tokenizers-0.10.2 torch-1.8.1 t
qdm-4.60.0 transformers-4.5.1 typer-0.3.2 typing-extensions-3.7.4.3 urllib3-1.26.4 wasabi-0.8.2 wheel-0.36.2 wrapt-1.12.
1 xlrd-2.0.1
WARNING: You are using pip version 20.2.3; however, version 21.1.1 is available.
You should consider upgrading via the 'c:\users\angya.desktop-kie1173\documents\schoo\y3s1\fyf\finalyearprojectuob\env\
scripts\python.exe -m pip install --upgrade pip' command.
(env) (temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\Schoo\y3s1\FYF\finalYearProjectUOB> |
```

14. OPTIONAL: If you encounter an OS error saying that the package could not be installed as there is no such file or directory when running the above command, it is most likely due to the file path being too long. To solve this, search up Registry Editor in your search bar and copy paste this path into the search bar of the Registry Editor found at the top: `“Computer\HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\FileSystem”`. Then, right click on “LongPathsEnabled” and change the “Value data” from 0 to 1. After this, you must close the Powershell or Anaconda Prompt, delete the env folder and restart from the very beginning.

**** Please do note if you are worried that this step might affect your file system in the future, you can change it back to 0 after running the engine****



15. Ensure that there is an “input.xlsx” in the folder /engine/files in the project folder. An “output.xlsx” is optional.



16. OPTIONAL: If you want 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 /engine/files. (If the file contains invalid entries, it will be automatically removed and will not show up in the “output.xlsx” file)

Company ID	Company	Country	SEC	Subsector	Architecture	Valuechain	Description	Customer Acquisition Strategy	Remarks
3.336E+18	CUBIC DECO PTE. LTD.	SINGAPORE					Cubic Deco Pte Ltd is an interior contractor based in Singapore, incorporated in March 2005 and with a current strength of 80, we provide professional services for interior fit out and construction works for residential, offices and retail spaces. We also offer a complete custom-made carpentry and joinery works right here at our outlet.		
4.880E+18	UTRACON CORPORATION PTE. LTD.	SINGAPORE					The Company provides civil engineering services and on-site post-tensioning products. Overland Ceramics is a ceramic tile manufacturer & supplier integrating design, research and development, manufacturing and marketing. Our products are wholesale market tile & stone tile.		
9.1054E+18	GUANGDONG OVERLAND CERAMICS CO., LTD.	SINGAPORE					OWMA is a joint-venture between Bouygues Batiment International and SFA Project Management. The company brings together the expertise of a leading international construction group (Bouygues Construction) and the local knowledge of SFA Strategic Holdings focusing on the Myanmar market. In April 2013, OWMA was awarded by Thaksin Estate Development the design and build contract for the 1st phase of Star City (Zone B), a prominent residential estate featuring 4,980 apartments, car parks and commercial facilities located along the Raga Road in the suburbs of Yangon. In December 2014, OWMA was also awarded the 3rd phase of Star City (Zone C). The development will be composed of residential units, including 900 residential units, as well as a 28,000sqm car park, located around the periphery of the site over two levels, providing L&T car park spaces. With more than 1,500 employees, OWMA will soon be PTC's key experience and expertise in providing integrated design and engineering, procurement of machinery & equipment, and construction (integrated EPC) of turnkey projects for industrial and process plants, mainly in energy, petrochemical, chemical and power industries.		
1.4434E+18	BTMA PTE LTD	SINGAPORE					Lure Infocystems is a Technology Services & IT Consulting Company, headquartered in Singapore with offices in Delhi, Mumbai and Singapore.		
8.6047E+18	PTCL MALAYSIA SDN. BHD.	MALAYSIA					We are a rapidly growing development & construction company		
4.5329E+18	DATA INFORMATION PTE. LTD.	SINGAPORE					TRADING OF BUILDING MATERIALS, SPECIALISES IN MECHANICAL & ELECTRICAL, ENGINEERING AND CIVIL ENGINEERING WORKS		
1.2859E+18	XINYUAN CONSTRUCTION PTE. LTD.	SINGAPORE					Importers and exporters of gas coolers and dealers in electrical appliances.		
9.9161E+18	MARUM MAUNDUNA SDN. BHD.	MALAYSIA					MET-INT CONSTRUCTION SDN. BHD. (MET-INT) is a Malaysian Bumiputera status company that is incorporated on 26th March 1999 with an objective of establishing leading engineering services and products.		
8.8596E+18	SINGAPORE RADIO & INDUSTRY PTE LTD	SINGAPORE					Engineering services undertaken by PRET include Sub Structure, Super Structure, Civil, Architecture & Interior Decoration.		
3.949E+18	MCC JATI CONSORTIUM SDN. BHD.	MALAYSIA					Soheto Service is the Authorized Agent and Service Center for Sony Professional Products as well as Fujifilm Digital Camera in Cambodia. We have been in operation in Cambodia since 1996, and we have a specialized sales and service team to cater to all your needs for broadcasting as well as photography needs.		
3.2762E+18	PRET ENGINEERING PTE LTD	SINGAPORE					Contemporio Sdn Bhd is an enterprise in Malaysia, with the main office in Kuala Lumpur. It operates in the Construction industry.		
6.0317E+18	SONETO (SINGAPORE) PTE. LTD.	SINGAPORE					Since its inception in 2007, OP has designed many of the country's most important public projects and notable landmarks. Each of these has played a critical role in shaping Singapore's civil urban landscape and development over a long history of the city in the formation of a continuous urban fabric. These sites serve as social and cultural anchors for Singapore and as public nodes of human density that have contributed greatly to the city's success. OP has also been involved with urban projects of immense scale and increasing diversity of function, some of the first's works are the		
1.814E+18	CONTAMAU SDN BHD	MALAYSIA					AP Commerce is a full-service eCommerce enabler that provides global brands with Asia Pacific's first end-to-end managed eCommerce solution, including website design and implementation, marketplace management, cross-border commerce, omnichannel order management, global delivery & fulfillment, digital marketing, customer care and Merchant of Record services.		
4.2543E+18	OP ARCHITECTS PTE LTD	SINGAPORE							
1.0402E+18	SP ECOMMERCE (MALAYSIA) SDN. BHD.	MALAYSIA							

17. To run the Engine, simply run **cd engine**, then **python app.py** in Powershell or Anaconda Prompt. Below is a partial snapshot of the terminal output.

```
Windows PowerShell
Your branch is up to date with 'origin/main'.
(env) (temp) PS C:\Users\angya.DESKTOP-KIE1173\Documents\School\y3s1\FYP\finalYearProjectU0B> python .\engine\app.py
2021-05-18 17:00:48.935002: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cudart64_110.dll
WARNING:tensorflow:SavedModel saved prior to TF 2.5 detected when loading Keras model. Please ensure that you are saving
the model with model.save() or tf.keras.models.save_model(), *NOT* tf.saved_model.save(). To confirm, there should be a
file named "keras_metadata.pb" in the SavedModel directory.
2021-05-18 17:00:52.636479: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library nvcuda.dll
2021-05-18 17:00:52.654974: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1733] Found device 0 with properties:
pciBusID: 0000:01:00.0 name: NVIDIA GeForce RTX 3070 computeCapability: 8.6
coreClock: 1.815GHz coreCount: 46 deviceMemorySize: 8.00GiB deviceMemoryBandwidth: 417.29GiB/s
2021-05-18 17:00:52.655163: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cudart64_110.dll
2021-05-18 17:00:52.695878: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cublas64_11.dll
2021-05-18 17:00:52.696001: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cublasLt64_11.dll
2021-05-18 17:00:52.712633: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cufft64_10.dll
2021-05-18 17:00:52.716775: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library curand64_10.dll
2021-05-18 17:00:52.729233: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cusolver64_11.dll
2021-05-18 17:00:52.743438: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cusparse64_11.dll
2021-05-18 17:00:52.743550: I tensorflow/stream_executor/platform/default/dso_loader.cc:53] Successfully opened dynamic
library cudnn64_8.dll
2021-05-18 17:00:52.743665: I tensorflow/core/common_runtime/gpu/gpu_device.cc:1871] Adding visible gpu devices: 0
2021-05-18 17:00:52.744089: I tensorflow/core/platform/cpu_feature_guard.cc:142] This TensorFlow binary is optimized wit
```

Prediction	Confidence
Company: CUBIC DECO PTE. LTD.	
CNI	56.22%
buildings & industrial	51.63%
residential_integrated_player	52.07%
MIDSTREAM	99.98%
Company: UTRACON CORPORATION PTE. LTD.	
CNI	79.43%
buildings & industrial	30.92%
cni_service_providers	24.11%
MIDSTREAM	78.3%
Company: GUANGDONG OVERLAND CERAMICS CO., LTD	
CNI	99.99%
building material	100.0%
building material_supplier_distributor	99.82%
DOWNSTREAM	99.91%

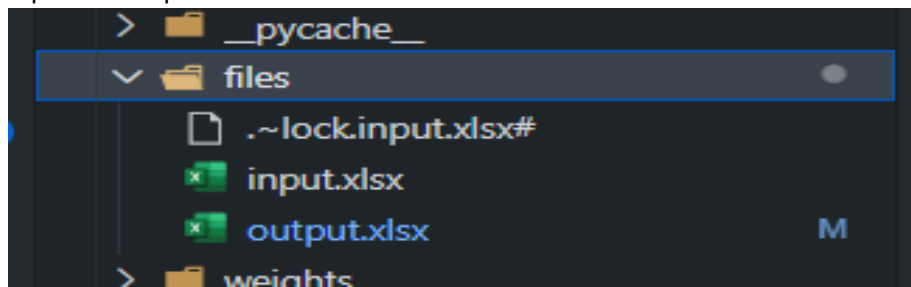
18. The output Excel file will be placed in the folder **/engine/files** and named **“output.xlsx”**.
Once opened, it will look like this:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	Company	Country	PIC	Sector	Subsector	Archetype	Valuechain	Websites	Profile	Inf	Remarks													
2	3.34E+18	CUBIC DECO SINGAPORE	CNI	cni_service	cni_service	MIDSTREAM	https://www.cubicdeco.com.sg/	https://www.cubicdeco.com.sg/	Cubic Deco Pte Ltd is an interior contractor based in Singapore. Incorporated in March 2005 and with a current strength of 30, we provide professional services for interi															
3	4.87E+18	UTRACON SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.utracor.com.sg/	https://www.utracor.com.sg/	The Company provides civil engineering services and on-site post-tensioning products.															
4	9.11E+18	GUANGDONG SINGAPORE	CNI	building m	building m	MIDSTREAM	https://www.guangdongceramics.com.sg/	https://www.guangdongceramics.com.sg/	Overland Ceramics is a ceramic tile manufacturer & supplier integrating design, research and development, manufacturing and marketing. Our products are wholesale ma															
5	1.44E+18	BYMA PTE SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.byma.com.sg/	https://www.byma.com.sg/	BYMA is a joint-venture between Bouygues Batiment International and SPA Project Management. The company brings together the expertise of a leading international co															
6	8.6E+18	TTCL MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.ttcl.com.my/	https://www.ttcl.com.my/	TTCL has experiences and expertise in providing integrated design and engineering, procurement of machinery & equipment, and construction (Integrated EPC) of turnke															
7	4.53E+18	LYRA INFO SINGAPORE	TMT	buildings m	building m	MIDSTREAM	https://www.lyra.com.sg/	https://www.lyra.com.sg/	Lyra Infosystems is a Technology Services & IT Consulting Company, headquartered in Bangalore with offices in Delhi, Mumbai and Singapore.															
8	1.27E+18	XINYUAN SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.xinyuan.com.sg/	https://www.xinyuan.com.sg/	We are a rapidly growing development & construction company															
9	8.92E+18	HARUM M MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.harum.com.sg/	https://www.harum.com.sg/	TRADING OF BUILDING MATERIALS, SPECIALISES IN MECHANICAL & ELECTRICAL ENGINEERING AND CIVIL ENGINEERING WORKS															
10	3.86E+18	SINGAPORE SINGAPORE	TMT	building m	building m	MIDSTREAM	https://www.hrs.com.sg/	https://www.hrs.com.sg/	Importers and exporters of gas cookers and dealers in electrical appliances.															
11	3.95E+18	MEC JATI MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.mecjati.com.my/	https://www.mecjati.com.my/	MEC JATI CONSORTIUM SDN. BHD. (479906-A) is a Malaysian Bumiputera status company G7 that is incorporated on 26th March 1999 with an objective of establishing I															
12	3.28E+18	PBT ENG'G SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.pbt.com.sg/	https://www.pbt.com.sg/	Engineering services undertaken by PBT include Sub Structure, Super Structure, Civil, Architecture & Interior Decoration.															
13	6.03E+18	SOHETO SINGAPORE	TMT	consumer	consumer	DOWNSTR	http://www.soheto.com.sg/	https://www.soheto.com.sg/	Soheto Services is the Authorized Agent and Service Center for Sony Professional Products as well as Fujifilm Digital Cameras in Cambodia. We have been in operation in															
14	1.81E+18	CONTAMAM MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.contamam.com.my/	https://www.contamam.com.my/	Contamaj Sdn Bhd is an enterprise in Malaysia, with the main office in Kuala Lumpur. It operates in the Construction industry.															
15	4.26E+18	DP ARCHIT SINGAPORE	TMT	building m	building m	MIDSTREAM	https://www.dparchitect.com.sg/	https://www.dparchitect.com.sg/	DP has designed many of the country's most important public projects and notable landmarks. Each of these has played a critical role in shapin															
16	1.04E+18	SP ECOMA MALAYSIA	TMT	digital	bus digital	MIDSTREAM	https://www.spcommerce.com.my/	https://www.spcommerce.com.my/	SP eCommerce is a full-service eCommerce enabler that provides global brands with Asia Pacific's first end-to-end managed eCommerce solution, including webstore des															
17	3.68E+18	CREST BUI MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.crestbui.com.my/	https://www.crestbui.com.my/	CBHG Group has a proven and established track record in the sector - especially in the commercial, residential and institutional building construction. The Group has now															
18	5.05E+18	AVP ELECT HONG KONG	TMT	semicond	electronic	MIDSTREAM	https://www.avp.com.hk/	https://www.avp.com.hk/	AVP Electronics Ltd serves as a supply channel for a wide range of products from Samsung Electronics in the China region which included Memory, CMOS Sensors, Mobik															
19	8.23E+18	SHOPEE SINGAPORE	TMT	building m	building m	MIDSTREAM	https://www.shopee.com.sg/	https://www.shopee.com.sg/	Launched in 2015, it is a platform tailored for the region, providing customers with an easy, secure and fast online shopping experience through strong payment and fulfil															
20	8.27E+18	EXCELPOT SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.excelpt.com.sg/	https://www.excelpt.com.sg/	Excelpt Technology Ltd. (the "Company") and its subsidiaries ("Excelpt" or the "Group") are one of the leading regional business-to-business ("B2B") platforms pro															
21	6.38E+17	Optus Fina AUSTRALIA	TMT	telecomm network	telecomm network	MIDSTREAM	https://www.optus.com.au/	https://www.optus.com.au/	Optus is an Australian leader in integrated telecommunications, delivering cutting-edge communications, information technology and entertainment services. In 2001 Sin															
22	5.34E+18	HOSPITAL SINGAPORE	TMT	services	services	MIDSTREAM	https://www.hrs.com.sg/	https://www.hrs.com.sg/	HRS is Oracle's largest hospitality partner worldwide, providing coverage in over 77 countries to more than 6000 customers. Recognized as an official Oracle Hospitality F															
23	3.85E+18	TPZTP FAM SINGAPORE	TMT	building m	building m	MIDSTREAM	https://www.tpztptfamily.com.sg/	https://www.tpztptfamily.com.sg/	TPZTP FAMILY is an ACR-registered entity that has been operating for 14 years 7 months in Singapore since its incorporation in 2006. Officially, TPZTP FAMILY PTE. LTC															
24	7.36E+18	HAIER ELE THAILAND	TMT	building m	building m	MIDSTREAM	https://www.haier.com/th/	https://www.haier.com/th/	Haier brand Founded in 2527 as a brand of home appliances, Haier Group, Haier was converted into the era of the Internet of Things has been introduced to the model of															
25	7.48E+18	LABCON SINGAPORE	TMT	buildings	buildings	MIDSTREAM	https://www.labcon.com.sg/	https://www.labcon.com.sg/	Established in 1989 as a sole proprietorship, the main activities of Labcon General Contractor (LGC) had been construction works for the commercial, industrial and resid															
26	3.64E+18	SINGAPORE SINGAPORE	TMT	telecomm network	telecomm network	MIDSTREAM	https://www.singtel.com.sg/	https://www.singtel.com.sg/	The Singtel Group is Asia's leading communications group. We provide a diverse range of services including fixed, mobile, data, internet, TV, info-comms technology (ICT)															
27	7.5E+18	EXCELSIOR SINGAPORE	CNI	cni_service	cni_service	MIDSTREAM	https://www.excelshior.com.sg/	https://www.excelshior.com.sg/	EXCELSIOR ENGINEERING is an ACR-registered entity that has been operating for 36 years 2 months in Singapore since its incorporation in 1985. Officially, EXCELSIOR															
28	8.9E+18	UNIVIC EN HONG KONG	CNI	buildings	buildings	MIDSTREAM	https://www.univice.com.hk/	https://www.univice.com.hk/	Univice Engineering Limited was founded in 1980. The company's line of business includes providing special trade contracting services.															
29	7.95E+18	UPSTREA SINGAPORE	TMT	semicond	electronic	MIDSTREAM	https://www.upstream.com.sg/	https://www.upstream.com.sg/	Established in 1990, KETEC grew rapidly to become a trusted world-class developer, manufacturer and distributor of equipment and consumables for the semiconductor															
30	8.61E+18	LANCO EN SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.lanco.com.sg/	https://www.lanco.com.sg/	Lanco Construction & Engineering Pte Ltd ("Lanco") was formed in 1994 specializing in engineering and machining works, with an expansion of scope to incorporate build															
31	8.16E+18	KIEN NAM MALAYSIA	CNI	building m	building m	MIDSTREAM	https://www.kiennam.com.my/	https://www.kiennam.com.my/	MANUFACTURING AND SALE OF BRICKS															
32	8.46E+18	GOLDEN S SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.goldens.com.sg/	https://www.goldens.com.sg/	Our team member engaged in a wide range of Building constructions, Residential, Drainages, Reinforced Concrete Structural works, infrastructure and services involving p															
33	2.44E+18	S & F CON MALAYSIA	CNI	buildings	buildings	MIDSTREAM	https://www.sandf.com.my/	https://www.sandf.com.my/	We specialized in a wide range of geotechnical and foundations engineering, and has increasingly committed to the field of environmental, building & infrastructure work															
34	6.3E+18	SAMSUNG SINGAPORE	CNI	buildings	buildings	MIDSTREAM	https://www.samsung.com.sg/	https://www.samsung.com.sg/	SAMSUNG-BINTAI KIDENKO JOINT VENTURE is an ACR-registered entity that has been operating for 9 years in Singapore since its incorporation in 2012. Officially, S															

19. To run the Engine with another Excel file, simply repeat steps 14 – 18.

In-depth Descriptions

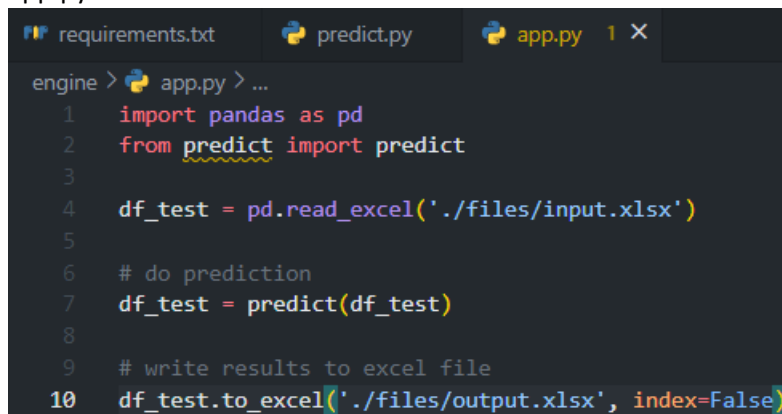
1. Input files: input.xlsx



input.xlsx – The input file for the dataset for the engine to predict results.

Output.xlsx – The output file from the engine, which includes the predicted results.
The engine reads the input.xlsx file by feeding the data into app.py:

2. app.py



Here, we read the dataset from the excel file and then populate it into a pandas dataframe. We then feed the data to the predict function found in predict.py. Finally when the prediction is done, we write the result to an output excel file.

1. predict.py

```
1 # import necessary libraries
2 from tensorflow.keras.models import load_model
3 import pandas as pd
4 import numpy as np
5 from preprocess import preprocess
6
7 # load models
8 sector_model = load_model('engine/weights/model_1')
9 subsector_model = load_model('engine/weights/model_2')
10 archetype_model = load_model('engine/weights/model_3')
11 valuechain_model = load_model('engine/weights/model_4')
12
13 models = [sector_model, subsector_model, archetype_model, valuechain_model]
14
15 # import sector master definition file
16 df_keywords = pd.read_excel('engine/sector_master_definition.xlsx')
17
18 # preprocess dataset
19 df_keywords.drop(['Explanations', 'Notes'], axis=1, inplace=True)
20 df_keywords['Value Chain'].fillna(' ', inplace=True)
21 df_keywords['Sector Keywords'].fillna('[]', inplace=True)
22 df_keywords['Sector Keywords'] = df_keywords['Sector Keywords'].str.upper()
23 df_keywords.dropna(axis=0, how='any', inplace=True)
24
25 # read and split classification tags
26 sectors = sorted(list(df_keywords['Sector'].str.upper().unique()))
27 subsectors = sorted(list(df_keywords['Subsector'].unique()))
28 archetypes = sorted(list(df_keywords['Archetype'].unique()))
29 valuechains = sorted(list(df_keywords['Value Chain'].str.upper().unique()))
30
31 class_counts = [len(sectors), len(subsectors), len(archetypes), len(valuechains)]
32 classes = [sectors, subsectors, archetypes, valuechains]
33
34 # build keyword master list
35 keywords = []
36 for index, item in df_keywords['Sector Keywords'].iteritems():
37     keywords += eval(item)
38
39 keywords = sorted(list(set(keywords)))
40
41 # function for processing prediction results
42 def __process_results(result):
43     temp = []
44
45     for r in result:
46         temp.append((np.argmax(r), r[np.argmax(r)]))
47
48     return temp
49
50
51 # === MAIN PREDICT FUNCTION === #
52 def predict(df):
53     df = preprocess(df, keywords)
54
55     X_pred = np.array(list(df['Bow_vectors']))
56
57     # do prediction
58     results = []
59     for model in models:
60         results.append(model.predict(X_pred))
61
62     # process output into rows
63     processed = []
64     for result in results:
65         processed.append(__process_results(result))
66
67     processed = np.array(processed)
68
69     # print results in human readable form
70     print('Prediction' + ' '*31 + '|' Confidence')
71     print('-'*53 + '\n')
72
73     for index, row in enumerate(processed.swapaxes(0, 1)):
74         print('Company:', df['Company'].iloc[index], '\n')
75
76         for i in range(len(row)):
77             print(f'{classes[i][int(row[i][0]): <40.40]} | {row[i][1] * 100: >9.4}%')
78
79         print('\n' + '-'*22 + '\n')
80
81     # add results to df
82
83     processed_tags = []
84     for i, result in enumerate(processed):
85         temp = []
86         for j, _ in result:
87             temp.append(classes[i][int(j)])
88         processed_tags.append(temp)
89
90     df['Sector'] = processed_tags[0]
91     df['Subsector'] = processed_tags[1]
92     df['Archetype'] = processed_tags[2]
93     df['Valuechain'] = processed_tags[3]
94
95     return df.drop(['Bow_vectors', 'processed'], axis=1)
```

At predict.py, what we did first was to load the pre-trained model weights that were done in google collab and also gather the sector master definition file which was already found in the master directory of the repository, and then build our master list keywords.

__process_results(): A function to preprocess the predicted results, getting the argmax, or the highest percentage recorded score.

predict(): A function whereby we can preprocess our raw data first by calling another function called preprocess() coded in another file. We then use the returned result from that function to format the result and print out the confidence score based on each company.

predict(): A function whereby we can preprocess our raw data first by calling another function called preprocess() coded in another file. We then use the returned result from that function to format the result and print out the confidence score based on each company.

2. preprocess.py

```
# import libraries
import spacy
import pandas as pd
from spacy.language import Language
from spacy.tokens import Doc
from spacy.lang.char_classes import ALPHA, ALPHA_LOWER, ALPHA_UPPER, CONCAT_QUOTES, LIST_ELLIPSES, LIST_ICONS
from spacy.util import compile_infix_regex

# set up spacy
spacy.prefer_gpu()

# init nlp
nlp = spacy.load('en_core_web_lg')

# ----- passthrough ----- #
# keywords masterlist
#

# declare custom properties
Doc.set_extension('processed', default=True, force=True)
Doc.set_extension('word_bag', default=True, force=True)

# Modify tokenizer infix patterns
infixes = (
    LIST_ELLIPSES
    + LIST_ICONS
    + [
        r"(?<=[0-9])[+\-\*^](?=[0-9-])",
        r"(?<=[{a1}{q}])\. (?=[{au}{q}])".format(
            a1=ALPHA_LOWER, au=ALPHA_UPPER, q=CONCAT_QUOTES
        ),
        r"(?<=[{a}]), (?=[{a}])".format(a=ALPHA),
        r"(?<=[{a}0-9])([:<>=]/) (?=[{a}])".format(a=ALPHA),
    ]
)

infix_re = compile_infix_regex(infixes)
nlp.tokenizer.infix_finditer = infix_re.finditer

dictionary = dict.fromkeys(keywords, 0)
for word in company:
    if word in keywords:
        dictionary[word] += 1

# append to dataframe
bow_vectors.append(list(dictionary.values()))

# print(f'{sum(dictionary.values()):>3}/{len(dictionary.values()):<3} |', dictionary.values())

df['Bow_vectors'] = bow_vectors

return df
```

At preprocess.py, where the data will be fed in from the predict.py after getting called, we will feed the data into spacy to do preprocessing, which includes tokenization, lemmenization, word cleaning, removing stop words, and punctuations.

3. Results

Once the entire process has finished running, users can see the following results:

Prediction	Confidence
CNI	99.99%
cni_service providers	99.96%
cni_service providers	99.18%
MIDSTREAM	100.0%
CNI	69.4%
buildings & industrial	52.23%
buildings & industrial_contractor	44.03%
MIDSTREAM	64.41%
CNI	100.0%
building material	100.0%
building material_manufacturer	100.0%
MIDSTREAM	100.0%
CNI	100.0%
buildings & industrial	100.0%
buildings & industrial_contractor	100.0%
MIDSTREAM	100.0%

And the output in the excel file:

Company ID	Company	Country	PIC	Sector	Subsector/Archetype/Aluechair	Websites	Profile Inf	Remarks
1	3.34E+18 CUBIC DEC	SINGAPORE	CNI	cni_service cni_service	MIDSTREA	https://wv Cubic Deco Pte Ltd is an interior contractor based in Singapore. Incorporated in March 2005 and with a current strength of 30, we provide professional service		
2	4.87E+18 UTRACON	SINGAPORE	CNI	buildings & buildings & MIDSTREA	https://utr The Company provides civil engineering services and on-site post-tensioning products.			
3	9.11E+18 GUANGDC	SINGAPORE	CNI	building m building m	MIDSTREA	https://wv Overland Ceramics is a ceramic tile manufacturer & supplier integrating design, research and development, manufacturing and marketing. Our products are w		
4	1.44E+18 BYMA PTE	SINGAPORE	CNI	buildings & buildings & MIDSTREA	https://rec BYMA is a joint-venture between Bouygues Batiment International and SPA Project Management. The company brings together the expertise of a leading int			
5	8.6E+18 TTCL MAL	MALAYSIA	CNI	buildings & buildings & MIDSTREA	https://wv TTCL has experiences and expertises in providing integrated design and engineering, procurement of machinery & equipment, and construction (Integrated EP			
6	4.53E+18 LYRA INFO	SINGAPORE	TMT	building m building m	MIDSTREA	https://rec Lyra Infosystems is a Technology Services & IT Consulting Company, headquartered in Bangalore with offices in Delhi, Mumbai and Singapore.		
7	1.27E+18 NINYUAN	SINGAPORE	CNI	buildings & buildings & MIDSTREA	https://sgf We are a rapidly growing development & construction company			
8	8.92E+18 HARUM M	MALAYSIA	CNI	buildings & buildings & MIDSTREA	https://wv TRADING OF BUILDING MATERIALS, SPECIALISES IN MECHANICAL & ELECTRICAL ENGINEERING AND CIVIL ENGINEERING WORKS			
9	3.86E+18 SINGAPOR	SINGAPORE	TMT	building m building m	MIDSTREA	https://wv Importers and exporters of gas cookers and dealers in electrical appliances.		
10	3.95E+18 MEC JATI	MALAYSIA	CNI	buildings & buildings & MIDSTREA	http://mei MEC JATI CONSORTIUM SDN. BHD. (479906-A) is a Malaysian Bumiputera status company G7 that is incorporated on 26th March 1999 with an objective of e			
11	3.28E+18 PBT ENGI	SINGAPORE	CNI	buildings & buildings & MIDSTREA	https://wv Engineering services undertaken by PBT include Sub Structure, Super Structure, Civil, Architecture & Interior Decoration.			
12	6.03E+18 SOHETO	SINGAPORE	TMT	consumer consumer DOWNSTRE	https://soh Soheto Services is the Authorized Agent and Service Center for Sony Professional Products as well as Fujifilm Digital Cameras in Cambodia. We have been in c			
13	1.81E+18 CONTAMPA	MALAYSIA	CNI	buildings & buildings & MIDSTREA	https://wv Contamaju Sdn Bhd is an enterprise in Malaysia, with the main office in Kuala Lumpur. It operates in the Construction industry.			
14	4.26E+18 DP ARCHIT	SINGAPORE	TMT	building m building m	MIDSTREA	https://wv Since its inception in 1967, DP has designed many of the country's most important public projects and notable landmarks. Each of these has played a critical r		
15	1.06E+18 SP ECOM	MALAYSIA	TMT	digital_bui digital_bui	MIDSTREA	https://wv SP Commerce is a full-service eCommerce enabler that provides global brands with Asia Pacific's first end-to-end managed eCommerce solution, including v		
16	3.68E+18 CREST BUI	MALAYSIA	CNI	buildings & buildings & MIDSTREA	http://wv CBHB Group has a proven and established track record in the sector - especially in the commercial, residential and institutional building construction. The Gro			
17	5.05E+18 AVP ELECT	HONG KONG	TMT	semicond; electronic; MIDSTREA	http://wv AVP Electronics Ltd serves as a supply channel for a wide range of products from Samsung Electronics in the China region which included Memory, CMOS Sen			
18	8.23E+18 SHORPEE	SINGAPORE	TMT	building m building m	MIDSTREA	https://cai Launched in 2015, it is a platform tailored for the region, providing customers with an easy, secure and fast online shopping experience through strong payme		
19	8.27E+18 EXCELPOT	SINGAPORE	CNI	buildings & buildings & MIDSTREA	https://sgf The Singtel Group is Aia's leading communications group. We provide a diverse range of services including fixed, mobile, data, internet, TV, infocomms techn			
20	6.38E+17 Optus Fina	AUSTRALIA	TMT	telecommi network o	MIDSTREA	https://wv Optus is an Australian leader in integrated telecommunications, delivering cutting-edge communications, information technology and entertainment services		
21	5.34E+18 HOSPITALI	SINGAPORE	TMT	it services it services	MIDSTREA	https://wv HRS is Oracle's largest hospitality partner worldwide, providing coverage in over 77 countries to more than 6000 customers. Recognized as an official Oracle		
22	3.85E+18 TPZTP FAM	SINGAPORE	TMT	building m building m	MIDSTREA	https://sgf TPZTP FAMILY is an ACRA-registered entity that has been operating for 14 years 7 months in Singapore since its incorporation y in 2006. Officially, TPZTP FAM		
23	7.36E+18 HAIER ELE	THAILAND	TMT	building m building m	MIDSTREA	https://wv Haier brand Founded in 2527 as a brand of home appliances, Haier Group, Haier was converted into the era of the Internet of Things has been introduced to		
24	7.48E+18 LABCON	SINGAPORE	CNI	buildings & buildings & MIDSTREA	http://labc Established in 1989 as a sole proprietorship, the main activities of Labcon General Contractor (LGC) had been construction works for the commercial, industr			
25	3.64E+18 SINGAPOR	SINGAPORE	TMT	telecommi network o	MIDSTREA	https://sgf The Singtel Group is Aia's leading communications group. We provide a diverse range of services including fixed, mobile, data, internet, TV, infocomms techn		
26	7.5E+18 EXCELSIOR	SINGAPORE	CNI	cni_service cni_service	MIDSTREA	https://sgf EXCELSIOR ENGINEERING is an ACRA-registered entity that has been operating for 36 years 2 months in Singapore since its incorporation y in 1985. Officially		
27	8.9E+18 UNIVIC EN	HONG KONG	CNI	buildings & buildings & MIDSTREA	https://wv Univic Engineering Limited was founded in 1980. The company's line of business includes providing special trade contracting services.			
28	7.95E+18 KETECA	SINGAPORE	TMT	semicond; electronic; UPSTREA	https://wv Established in 1990, KETECA grew rapidly to become a trusted world-class developer, manufacturer and distributor of equipment and consumables for the se			