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

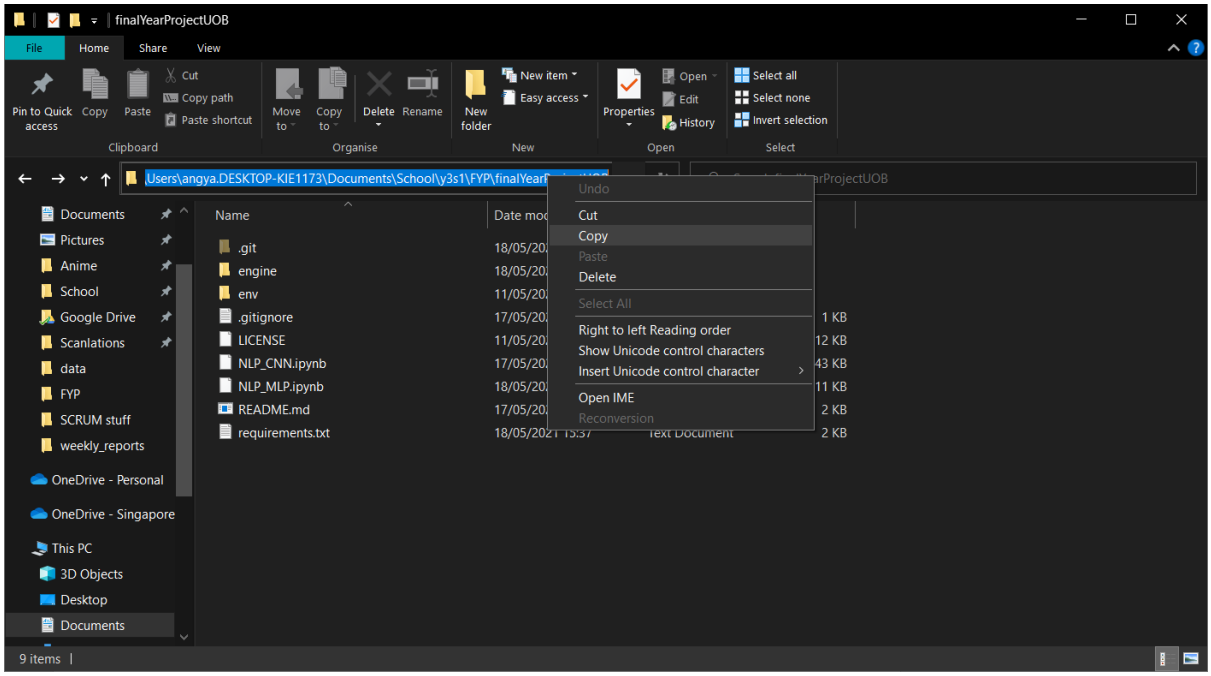
1. Download and unzip project zip file: "*fyp_code_vX*", where *vX* is the version number.

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

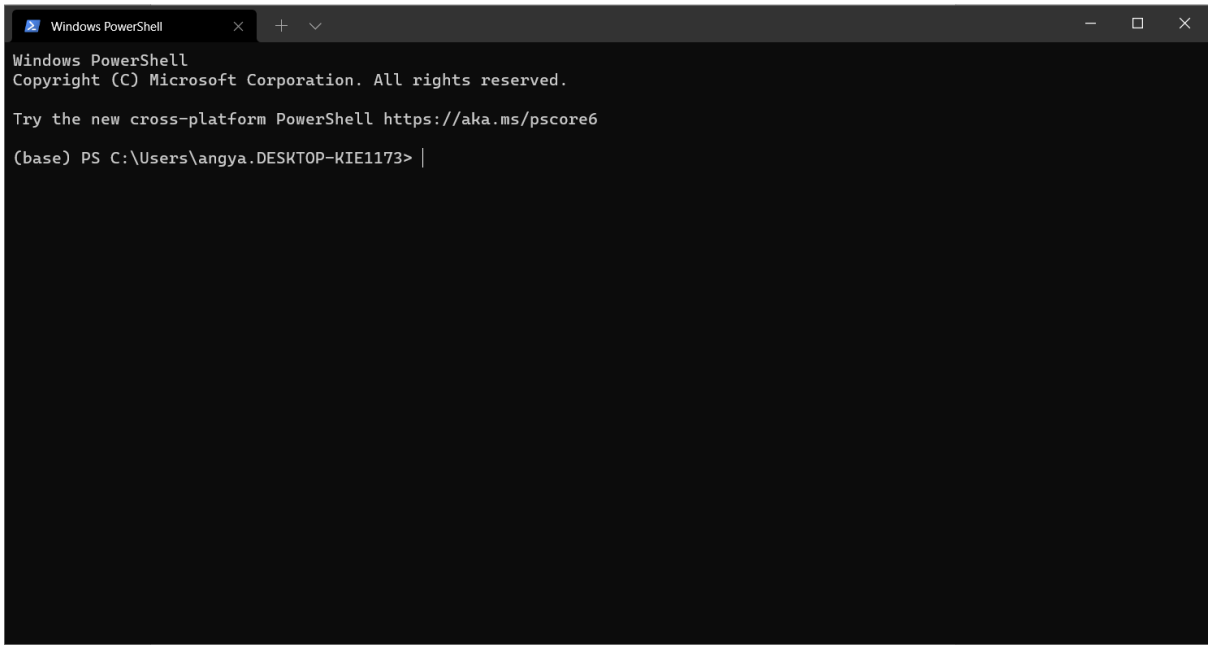
A. Engine
   - __pycache__
   - files
   - weights
   - app.py
   - predict.py
   - preprocess.py
   - sector_master_definition.xlsx

B. .gitignore
C. LICENSE
D. NLP_CNN.ipynb
E. NLP_MLP.ipynb
F. README.md
G. requirements.txt

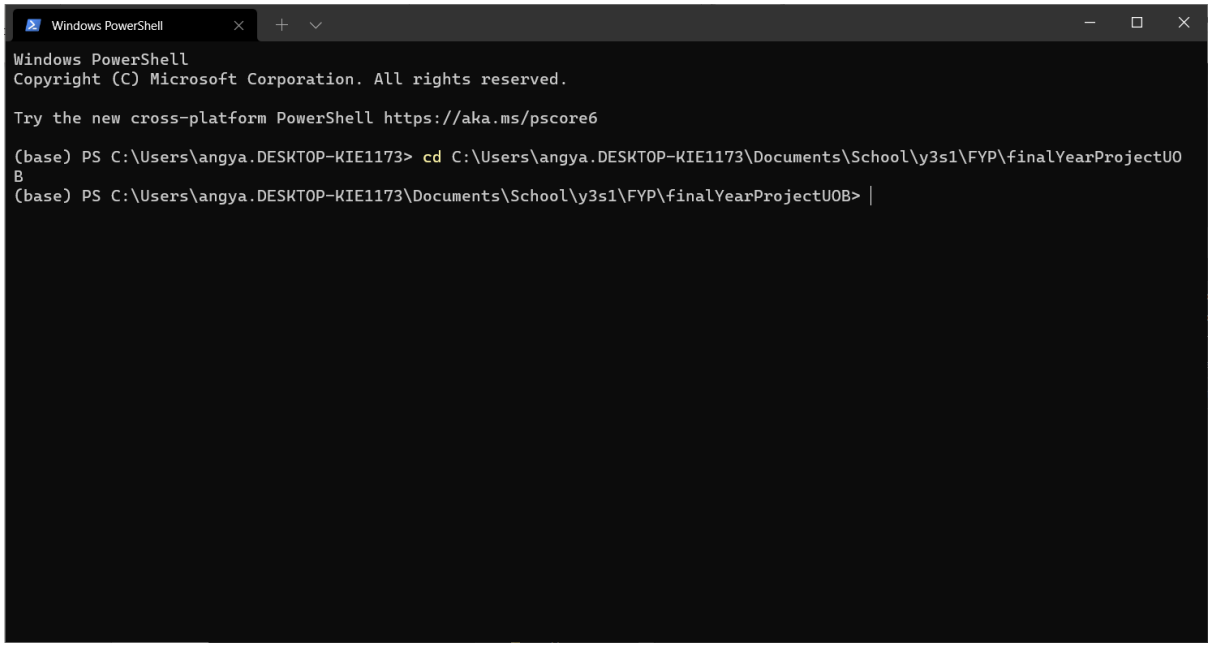
2. Navigate into unzipped project folder on your file explorer.
3. Copy path to project folder as such:



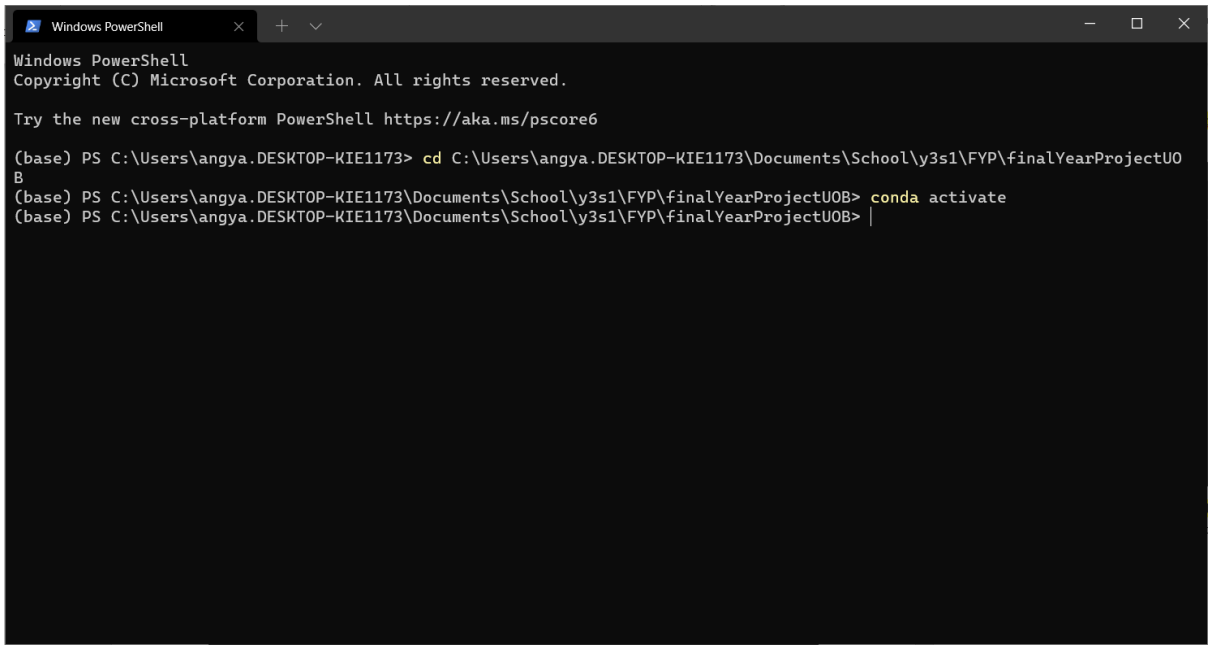
4. Open Anaconda Prompt or Powershell.



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

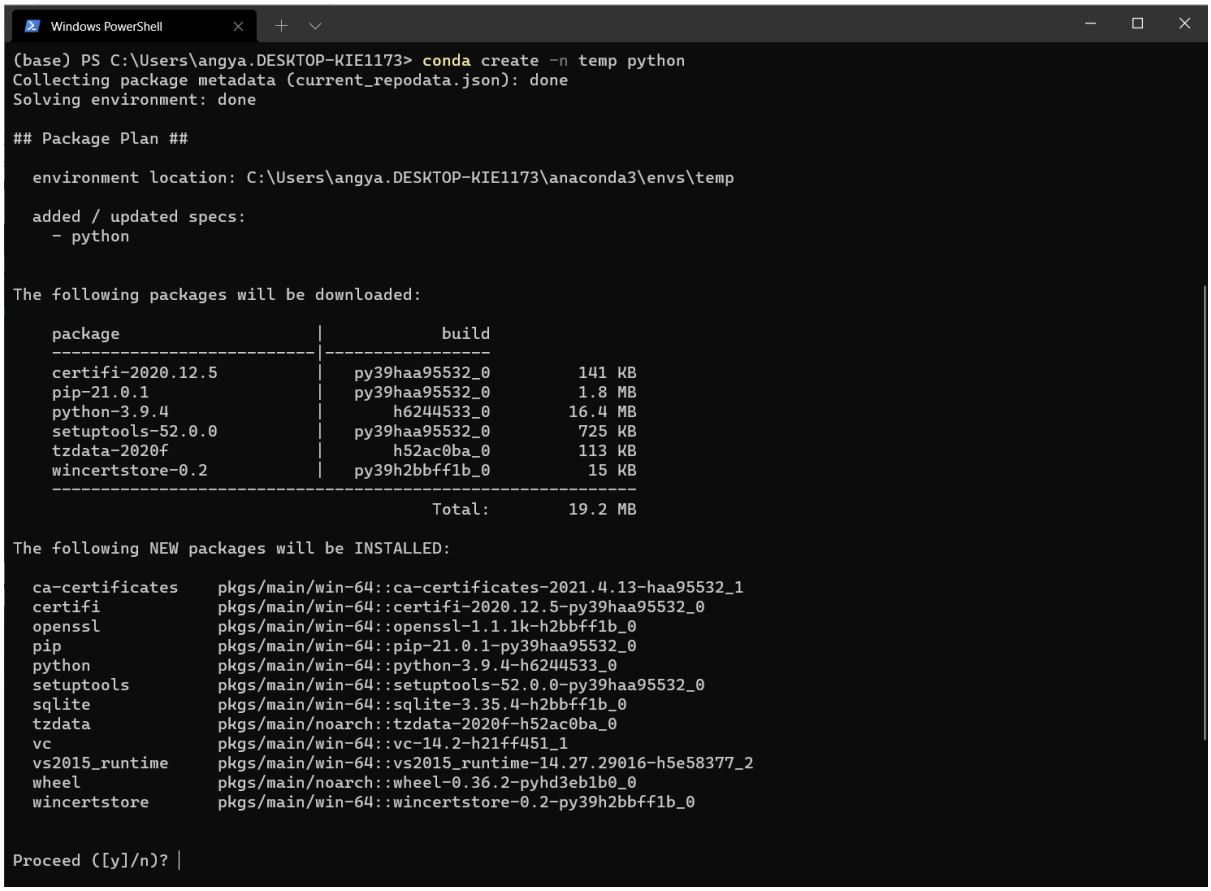


6. Run `conda activate` to ensure that the Anaconda environment is enabled. (There should be a (base) visible if it is enabled)

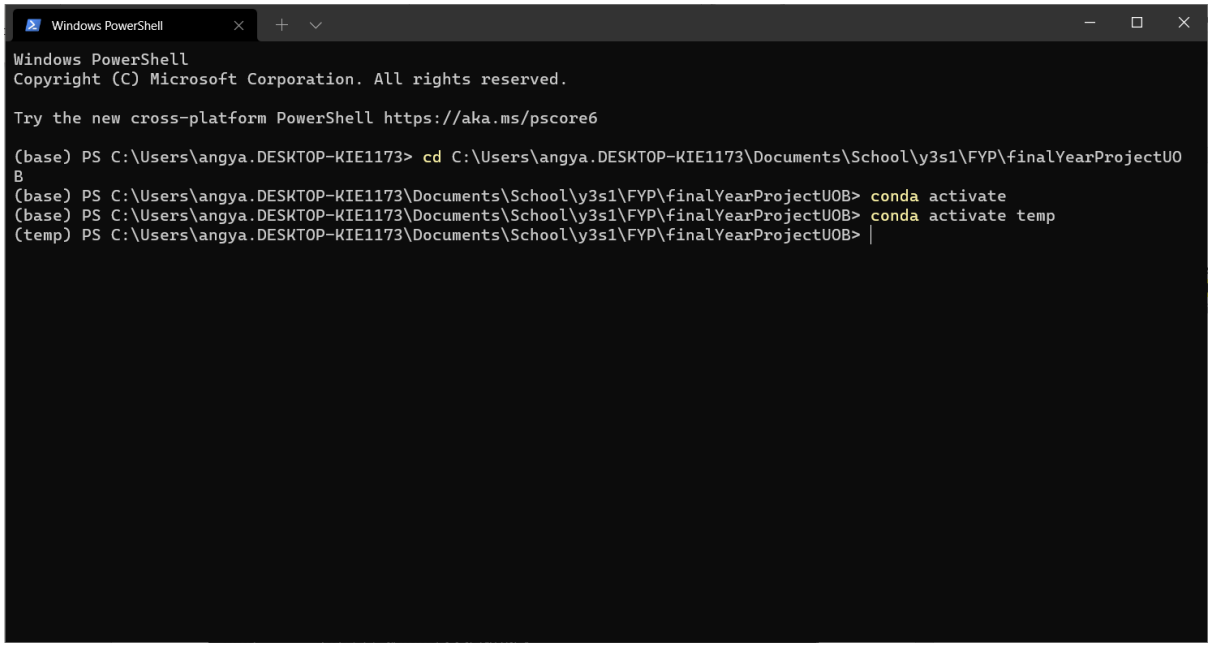


DISCLAIMER: PLEASE RUN STEP 7 TO 10 IF NVIDIA CUDNN PACKAGE IS NOT INSTALLED IN YOUR DEVICE BEFORE. ONLY RUN IF YOUR DEVICE HAS AN NVIDIA GPU.

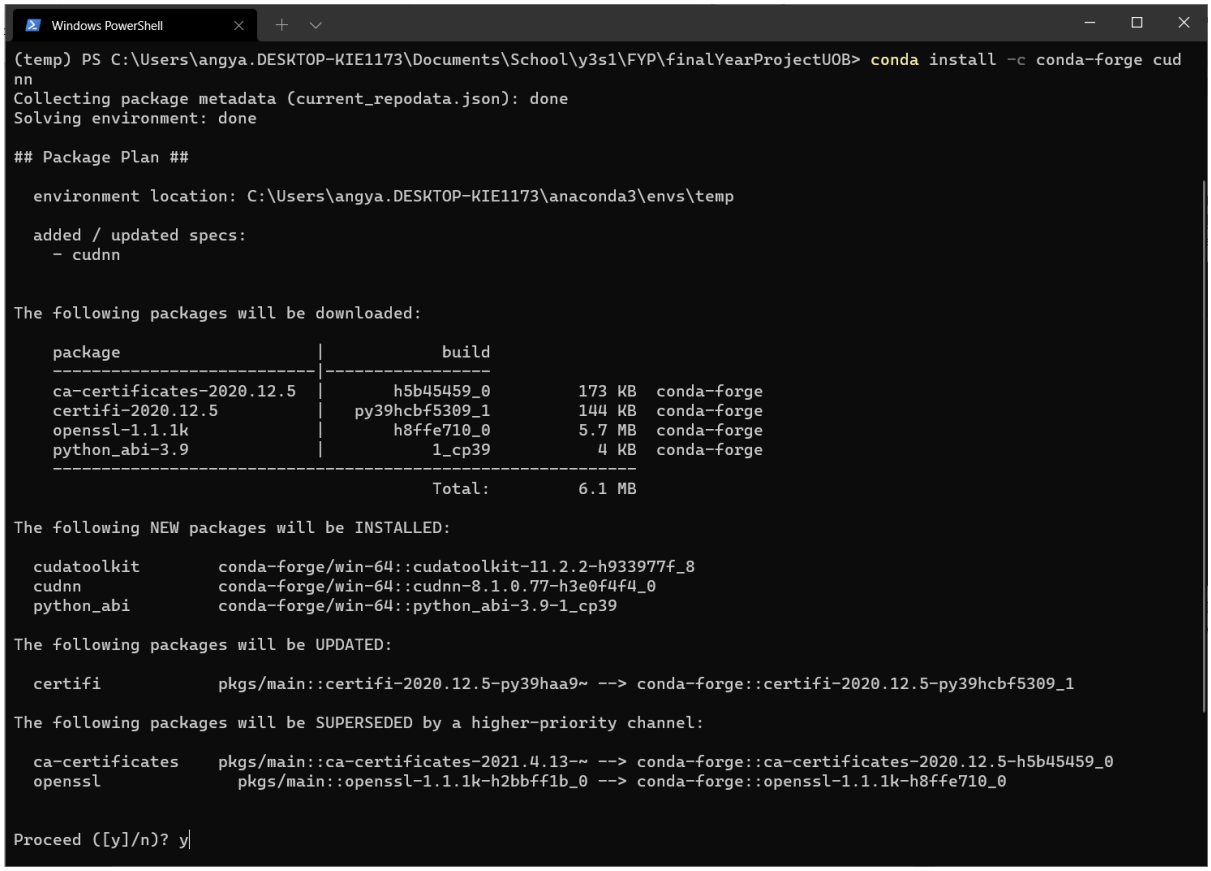
7. OPTIONAL: Run `conda create -n temp python` to create a fresh conda environment, so that your default anaconda environment is not dirtied with the libraries required to run the code. Press `y` then `Enter` when prompted.



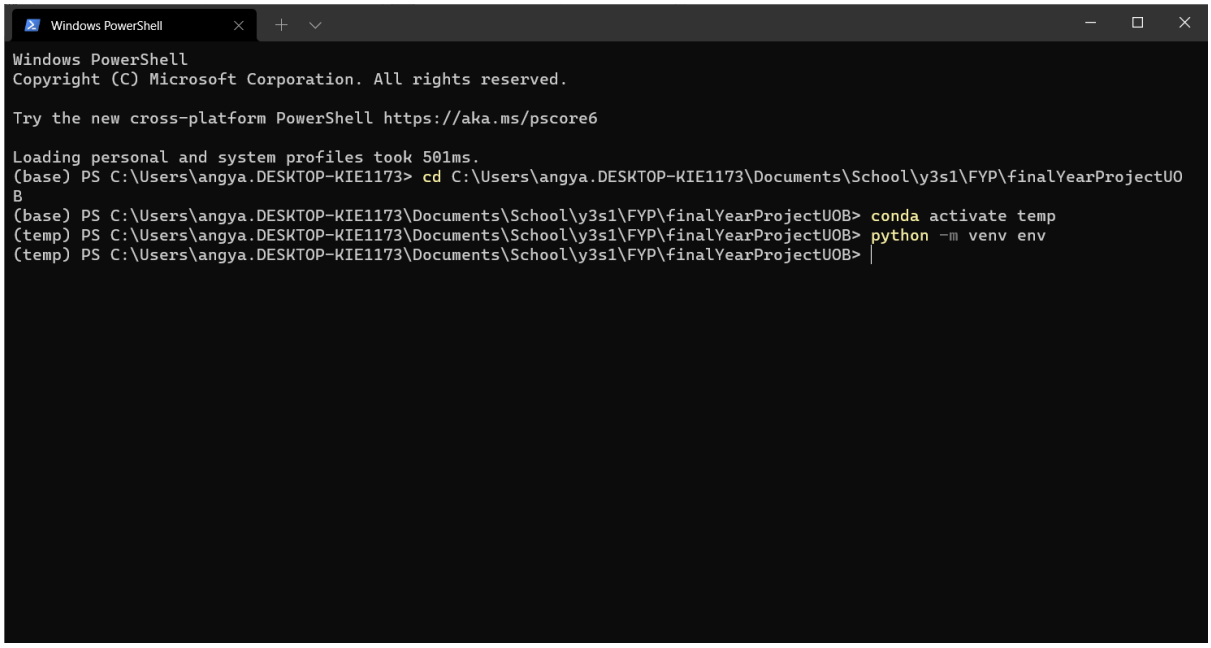
8. OPTIONAL: Wait until the setup is complete.
9. OPTIONAL: Run `conda activate temp` to activate the created anaconda environment. (There should be a (temp) visible if it is successfully enabled)



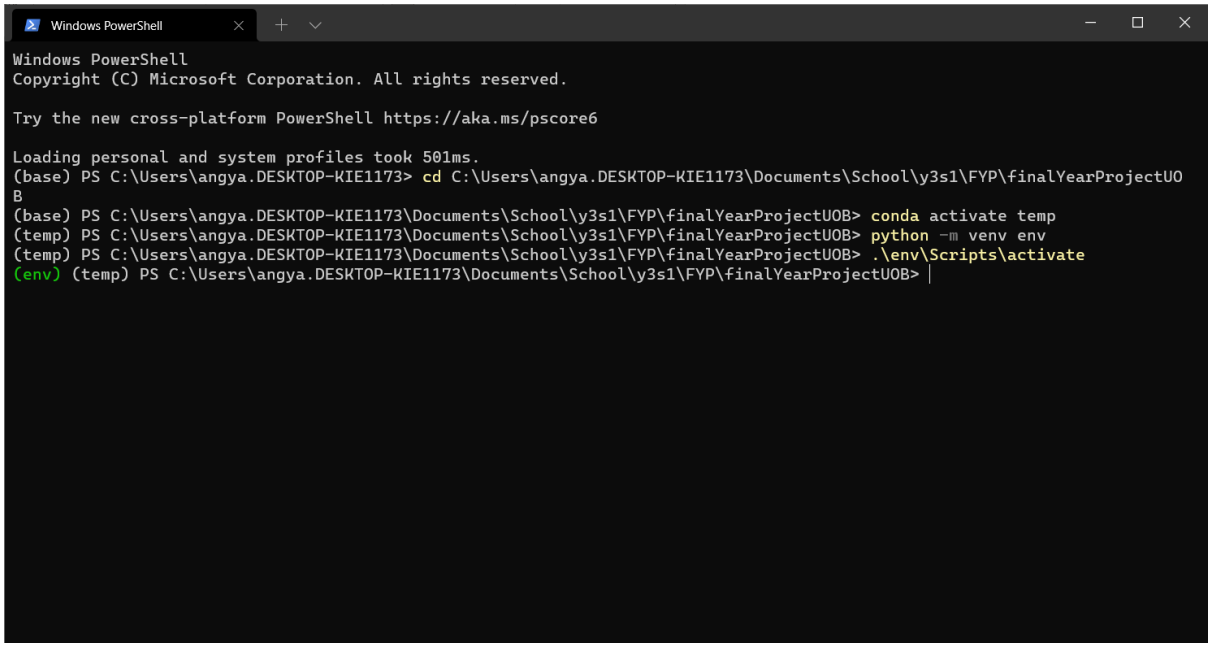
10. OPTIONAL: Run `conda install -c conda-forge cudnn` to install libraries required for GPU usage. Press `y` then `Enter` when prompted. Wait for installation to complete.



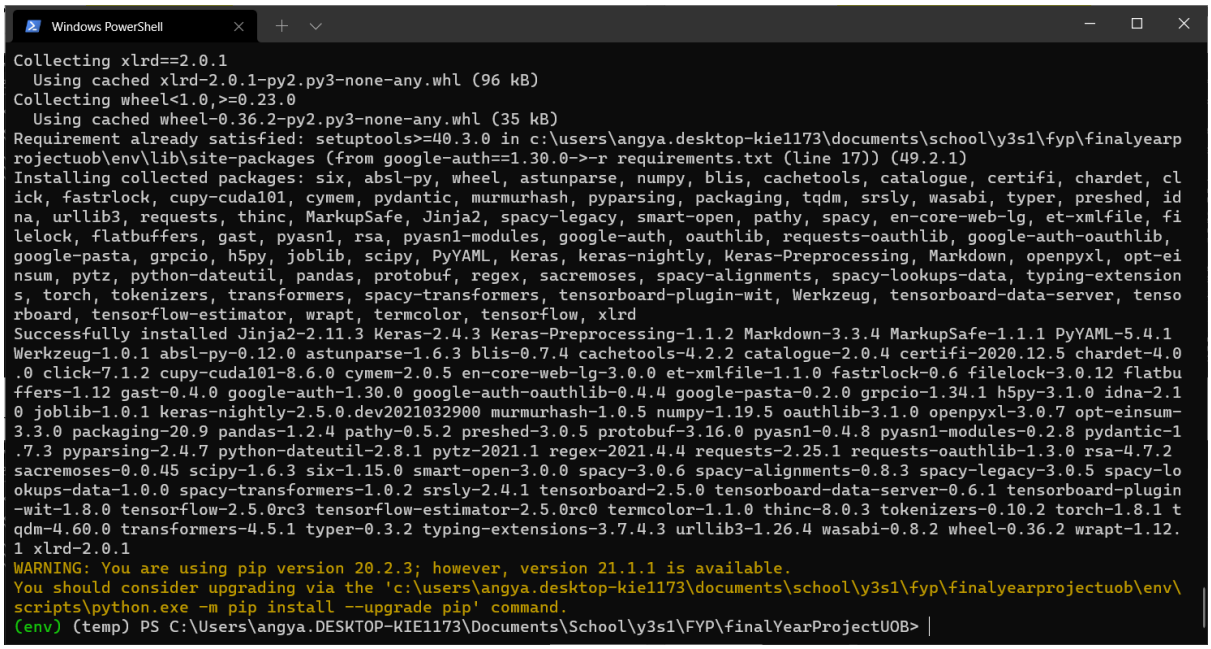
11. Run `python -m venv env` to make a new python virtual environment. This is to ensure that the installed python libraries do not dirty your python environment, and that there are no conflicts with existing packages that can cause the Engine to fail.



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

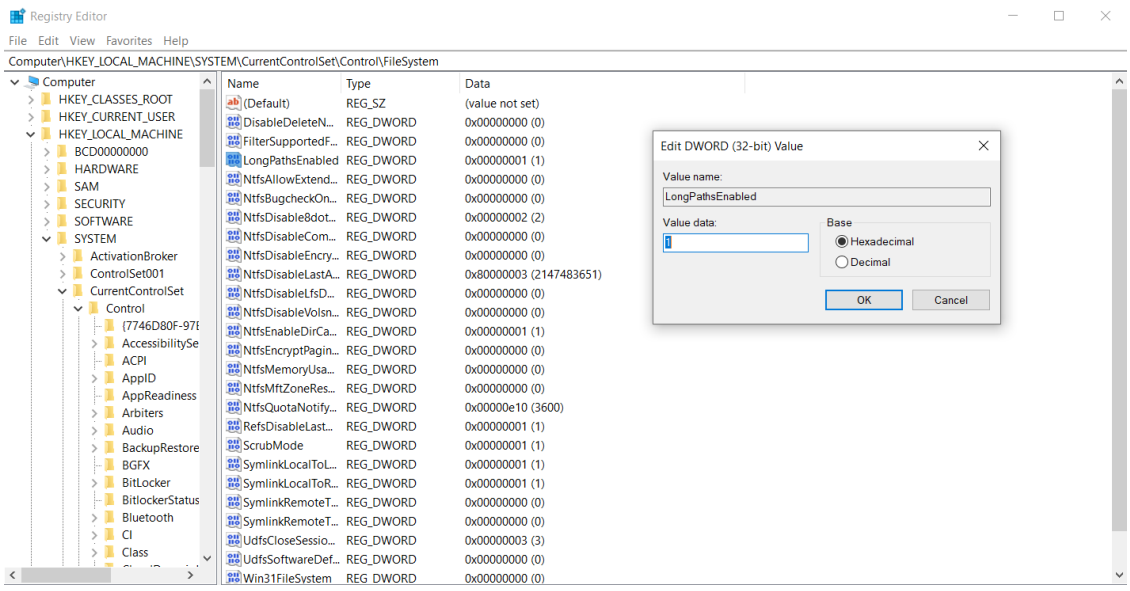


13. Run `pip install -r requirements.txt` to install required libraries. Wait for installation to finish.

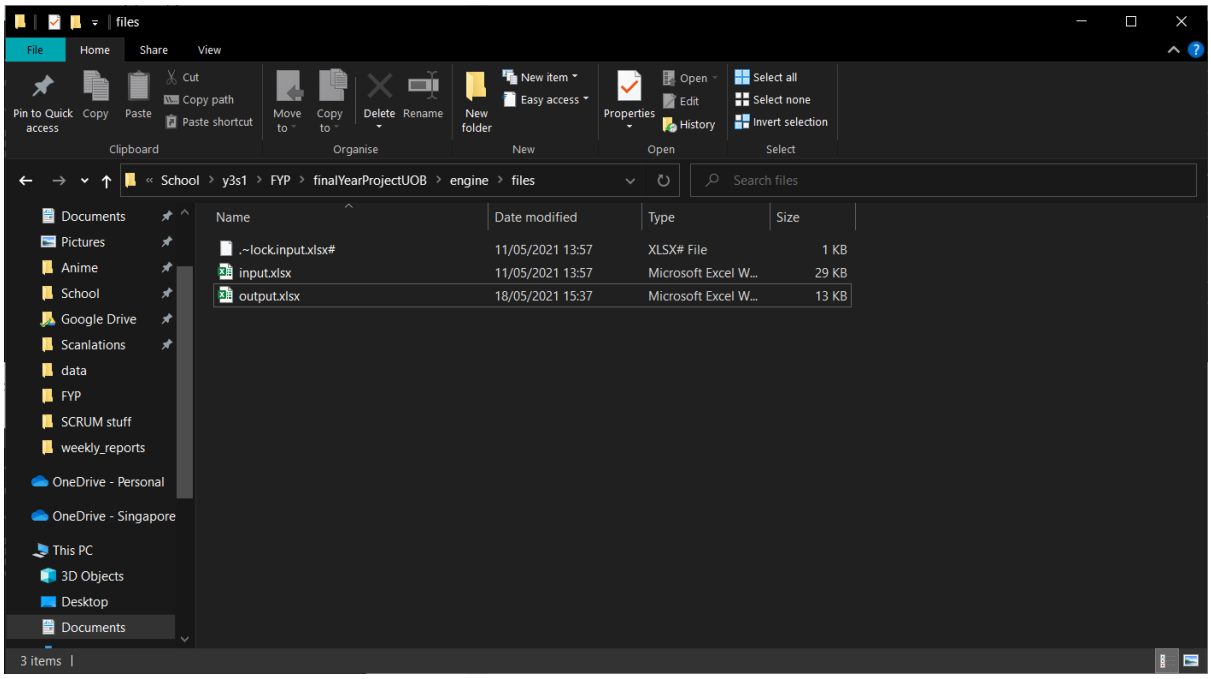


14. OPTIONAL: If you encounter an OS error saying that the package could not be installed as there is no such file or directory when running the above command, it is most likely due to the file path being too long. To solve this, search up Registry Editor in your search bar and copy paste this path into 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**

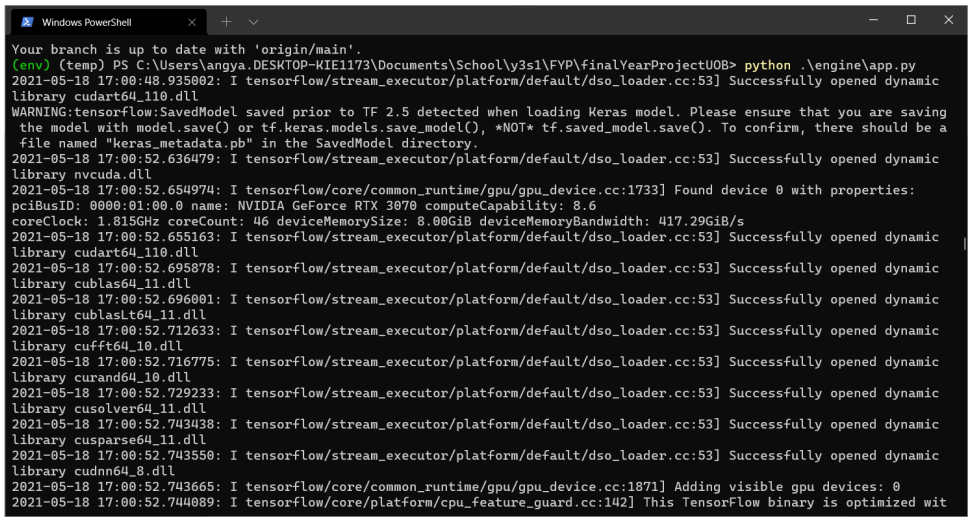


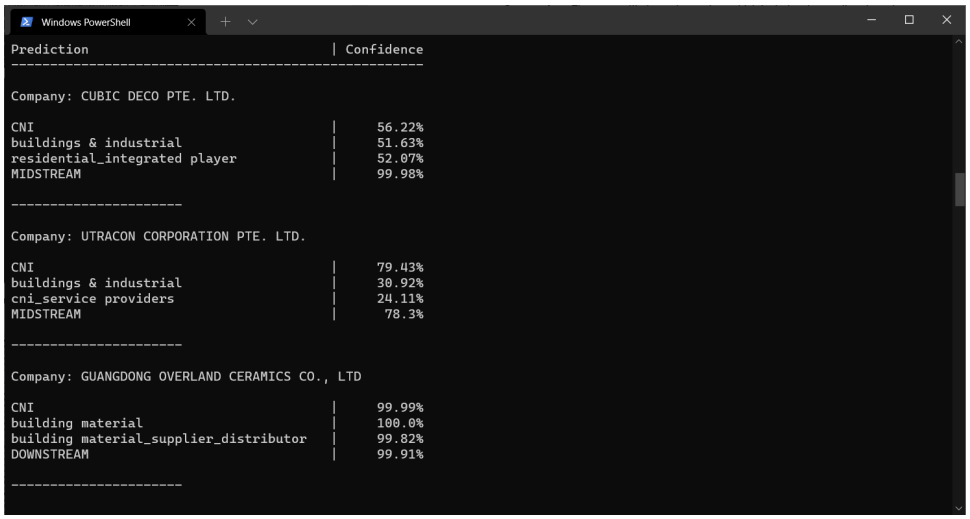
15. Ensure that there is an “input.xlsx” in the folder `/engine/files` in the project folder. An “output.xlsx” is optional.



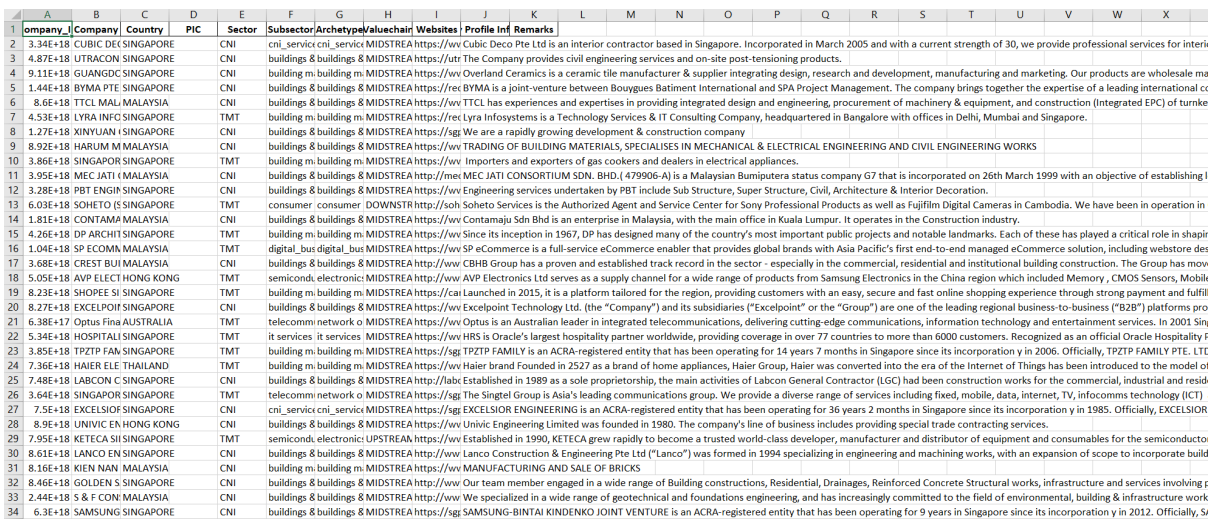
16. OPTIONAL: If you want to create a custom input file, do follow format of the already existing “input.xlsx” file, and make sure that the custom file is in `/engine/files`. (If the file contains invalid entries, it will be automatically removed and will not show up in the “output.xlsx” file)

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.





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

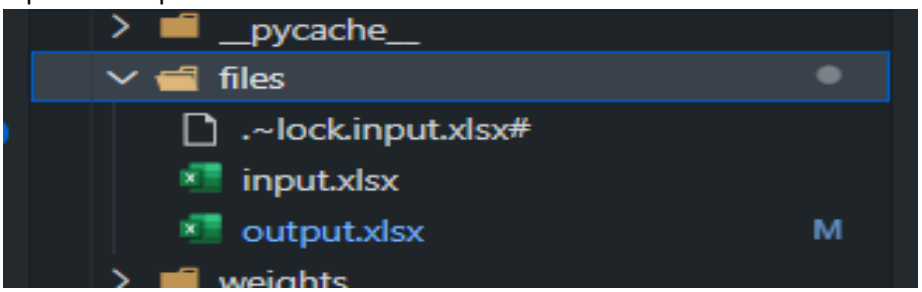


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	ompany_I	Company	Country	PIC	Sector	Subsector	Archetype	Valuechair	Websites	Profile Inf	Remarks													
2	3.34E+18	CUBIC DE	SINGAPORE		CNI	cni_servic	cni_servic	MIDSTREA	https://w	Cubic Deco Pte Ltd is an interior contractor based in Singapore. Incorporated 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 &	MIDSTREA	https://utr	The Company provides civil engineering services and on-site post-tensioning products.														
4	9.11E+18	GUANGDC	SINGAPORE		CNI	building m	building m	MIDSTREA	https://w	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 &	MIDSTREA	https://re	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 MAL	MALAYSIA		CNI	buildings &	buildings &	MIDSTREA	https://w	TTCL has experiences and expertises in providing integrated design and engineering, procurement of machinery & equipment, and construction (Integrated EPC) of turnke														
7	4.53E+18	LYRA INFO	SINGAPORE		TMT	building m	building m	MIDSTREA	https://re	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 &	MIDSTREA	https://sg	We are a rapidly growing development & construction company														
9	8.92E+18	HARUM M	MALAYSIA		CNI	buildings &	buildings &	MIDSTREA	https://w	TRADING OF BUILDING MATERIALS, SPECIALISES IN MECHANICAL & ELECTRICAL ENGINEERING AND CIVIL ENGINEERING WORKS														
10	3.86E+18	SINGAPOR	SINGAPORE		TMT	building m	building m	MIDSTREA	https://w	Importers and exporters of gas cookers and dealers in electrical appliances.														
11	3.95E+18	MEC JATI (	MALAYSIA		CNI	buildings &	buildings &	MIDSTREA	http://me	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 l														
12	3.28E+18	PBT ENGI	SINGAPORE		CNI	buildings &	buildings &	MIDSTREA	https://w	Engineering services undertaken by PBT include Sub Structure, Super Structure, Civil, Architecture & Interior Decoration.														
13	6.03E+18	SOHETO (S	SINGAPORE		TMT	consumer	consumer	DOWNSTR	http://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 operation in														
14	1.81E+18	CONTAMA	MALAYSIA		CNI	buildings &	buildings &	MIDSTREA	https://w	Contamaju 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	MIDSTREA	https://w	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 role in shapir														
16	1.04E+18	SP ECOMM	MALAYSIA		TMT	digital_bus	digital_bus	MIDSTREA	https://w	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 &	MIDSTREA	http://w	CBHB Group has a proven and established track record in the sector - especially in the commercial, residential and institutional building construction. The Group has mov														
18	5.05E+18	AVP ELECT	HONG KONG		TMT	semicondu	electronic	MIDSTREA	http://w	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, Mobile														
19	8.23E+18	SHOPEE SI	SINGAPORE		TMT	building m	building m	MIDSTREA	https://ca	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 fulfill														
20	8.27E+18	EXCELPOI	SINGAPORE		CNI	buildings &	buildings &	MIDSTREA	https://w	Excelpoint Technology Ltd. (the “Company”) and its subsidiaries (“Excelpoint” 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 o	MIDSTREA	https://w	Optus is an Australian leader in integrated telecommunications, delivering cutting-edge communications, information technology and entertainment services. In 2001 Sin														
22	5.34E+18	HOSPITALI	SINGAPORE		TMT	it services	it services	MIDSTREA	https://w	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 FAN	SINGAPORE		TMT	building m	building m	MIDSTREA	https://sg	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 FAMILY PTE. LTD														
24	7.36E+18	HAIER ELE	THAILAND		TMT	building m	building m	MIDSTREA	https://w	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 C	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, industrial and reside														
26	3.64E+18	SINGAPOR	SINGAPORE		TMT	telecomm	network o	MIDSTREA	https://sg	The Singtel Group is Asia's leading communications group. We provide a diverse range of services including fixed, mobile, data, internet, TV, infocomms technology (ICT)														
27	7.5E+18	EXCELSIOF	SINGAPORE		CNI	cni_servic	cni_servic	MIDSTREA	https://sg	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, EXCELSIOR														
28	8.9E+18	UNIVIC EN	HONG KONG		CNI	buildings &	buildings &	MIDSTREA	https://w	Univic Engineering Limited was founded in 1980. The company's line of business includes providing special trade contracting services.														
29	7.95E+18	KETECA SI	SINGAPORE		TMT	semicondu	electronic	UPSTREAM	https://w	Established in 1990, KETECA grew rapidly to become a trusted world-class developer, manufacturer and distributor of equipment and consumables for the semiconducto														
30	8.61E+18	LANCO EN	SINGAPORE		CNI	buildings &	buildings &	MIDSTREA	http://w	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 NAN	MALAYSIA		CNI	building m	building m	MIDSTREA	https://w	MANUFACTURING AND SALE OF BRICKS														
32	8.46E+18	GOLDEN S	SINGAPORE		CNI	buildings &	buildings &	MIDSTREA	http://ww	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 &	MIDSTREA	http://ww	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 &	MIDSTREA	https://sg	SAMSUNG-BINTAI KINDENKO JOINT VENTURE is an ACRA-registered entity that has been operating for 9 years in Singapore since its incorporation y in 2012. Officially, SA														

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

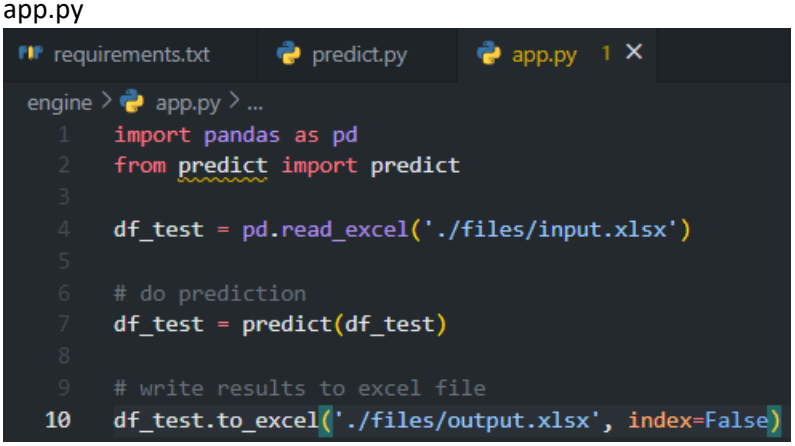
In-depth Descriptions

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

```python
# import necessary libraries
from tensorflow.keras.models import load_model
import pandas as pd
import numpy as np
from preprocess import preprocess

# load models
sector_model = load_model('engine/weights/model_1')
subsector_model = load_model('engine/weights/model_2')
archetype_model = load_model('engine/weights/model_3')
valuechain_model = load_model('engine/weights/model_4')

models = [sector_model, subsector_model, archetype_model, valuechain_model]

# import sector master definition file
df_keywords = pd.read_excel('engine/sector_master_definition.xlsx')

# preprocess dataset
df_keywords.drop(['Explanations', 'Notes'], axis=1, inplace=True)
df_keywords['Value Chain'].fillna(' ', inplace=True)
df_keywords['Sector Keywords'].fillna('[]', inplace=True)
df_keywords['Sector Keywords'] = df_keywords['Sector Keywords'].str.upper()
df_keywords.dropna(axis=0, how='any', inplace=True)

# read and split classification tags
sectors = sorted(list(df_keywords['Sector'].str.upper().unique()))
subsectors = sorted(list(df_keywords['Subsector'].unique()))
archetypes = sorted(list(df_keywords['Archetype'].unique()))
valuechains = sorted(list(df_keywords['Value Chain'].str.upper().unique()))

class_counts = [len(sectors), len(subsectors), len(archetypes), len(valuechai
classes = [sectors, subsectors, archetypes, valuechains]

# build keyword master list
keywords = []
for index, item in df_keywords['Sector Keywords'].iteritems():
    keywords += eval(item)

keywords = sorted(list(set(keywords)))

# function for processing prediction results
def __process_results(result):
    temp = []

    for r in result:
        temp.append((np.argmax(r), r[np.argmax(r)]))

    return temp
```

```python
# === MAIN PREDICT FUNCTION === #
def predict(df):
    df = preprocess(df, keywords)

    X_pred = np.array(list(df['BoW_vectors']))

    # do prediction
    results = []
    for model in models:
        results.append(model.predict(X_pred))

    # process output into rows
    processed = []
    for result in results:
        processed.append(__process_results(result))

    processed = np.array(processed)

    # print results in human readable form
    print('Prediction' + ' '*31 + '| Confidence')
    print('-'*53 + '\n')

    for index, row in enumerate(processed.swapaxes(0, 1)):
        print('Company:', df['Company'].iloc[index], '\n')

        for i in range(len(row)):
            print(f'{classes[i][int(row[i][0])]: <40.40} | {row[i][1] * 100: >9.4}%')

        print('\n' + '-'*22 + '\n')

    # add results to df

    processed_tags = []
    for i, result in enumerate(processed):
        temp = []
        for j, _ in result:
            temp.append(classes[i][int(j)])

        processed_tags.append(temp)

    df['Sector'] = processed_tags[0]
    df['Subsector'] = processed_tags[1]
    df['Archetype'] = processed_tags[2]
    df['Valuechain'] = processed_tags[3]

    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

```python
# 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"(?<=[{al}{q}])\.(?=[{au}{q}])".format(
            al=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
```

```python
        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: