EAI6980(Text-CNN and ORG System) set up — Windows Version
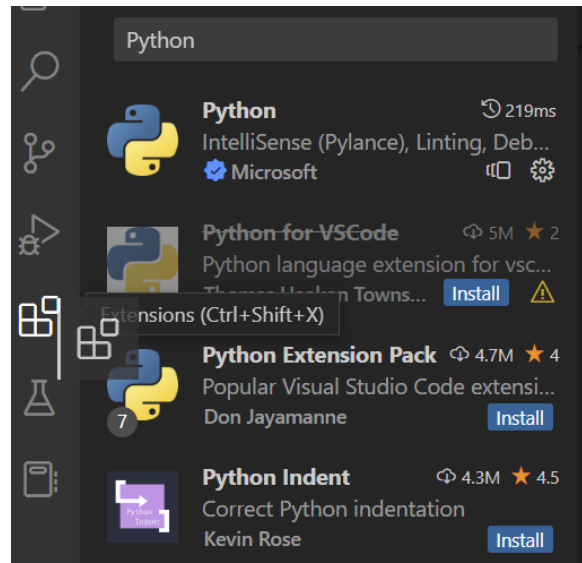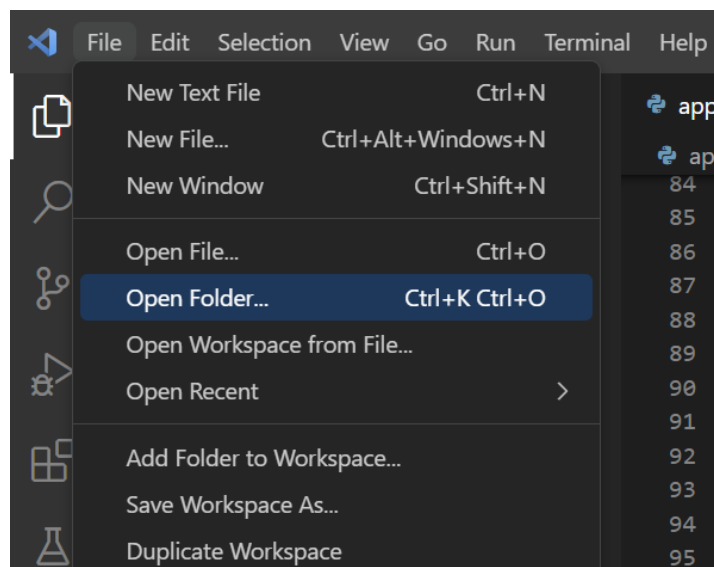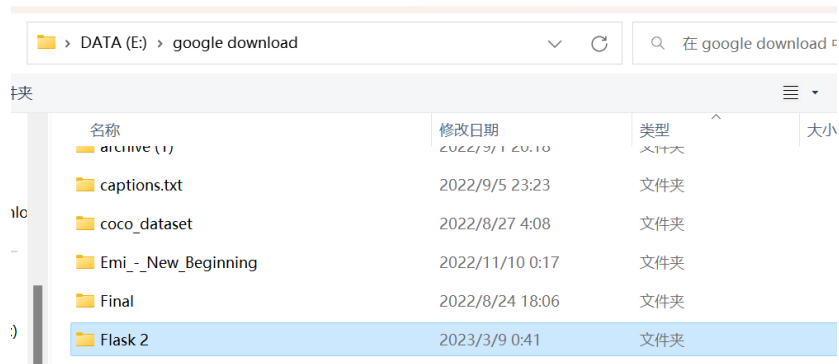
1. Make sure you have downloaded the visual studio code: https://code.visualstudio.com/

2. Search for "Python" in Extensions on the left menu, and download Python (linting, debugging, code navigation, code formatting, refactoring, variable explorer, test explorer)
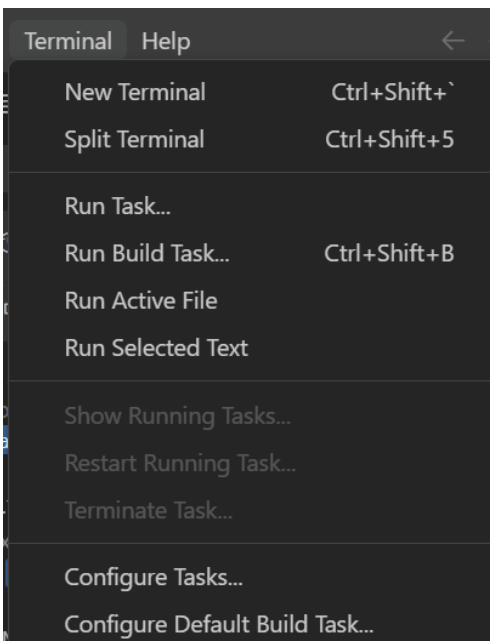

3. In the "File" option in the upper left corner, click "Open Folder" and select the folder named "Flask2".
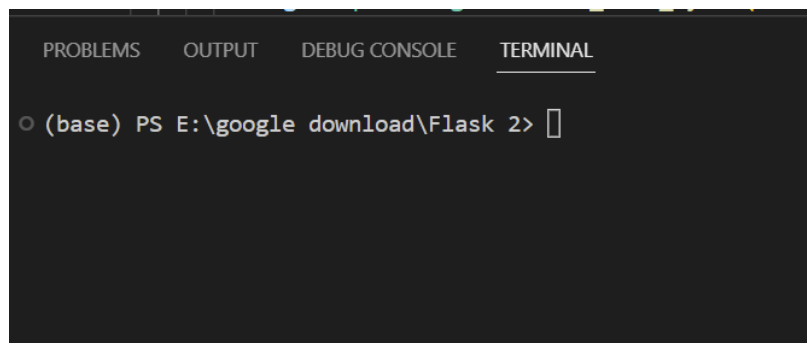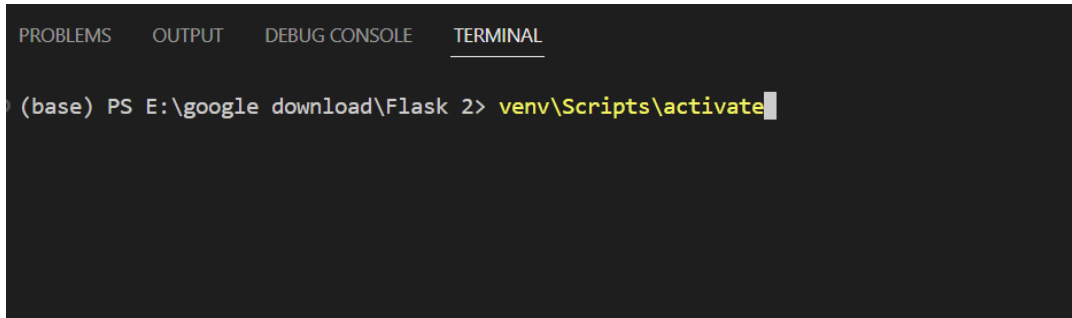
4. In the "Terminal" option in the upper left corner, click "new Terminal."

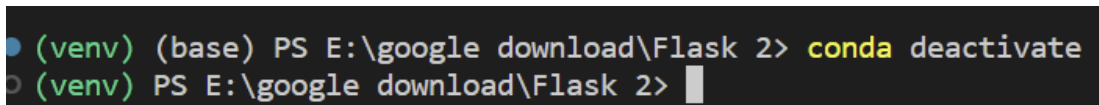You will see a "base" based environment to create.

5. Enter the following code in Terminal to create a virtual environment, **venv\Scripts\activate** & **conda deactivate**

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL

(base) PS E:\google download\Flask 2> venv\Scripts\activate
```
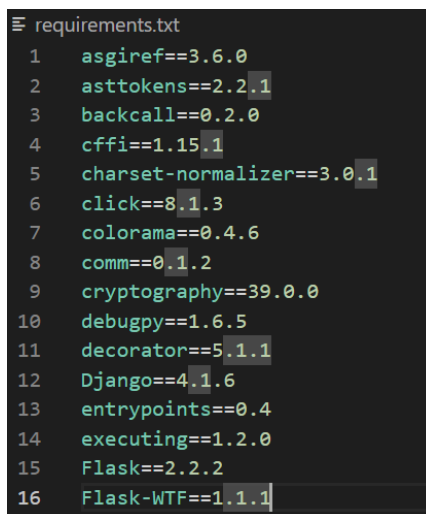
Finally you will see the "base" environment converted to the "venv" virtual environment

```
(venv) (base) PS E:\google download\Flask 2> conda deactivate
(venv) PS E:\google download\Flask 2>
```
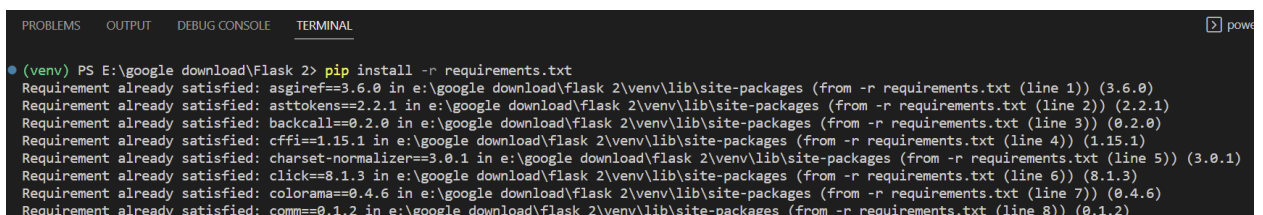
1. (Python)Package set up

By freeze > requirements.txt, we have gotten all the packages we use in this application.

```
requirements.txt
1    asgiref==3.6.0
2    asttokens==2.2.1
3    backcall==0.2.0
4    cffi==1.15.1
5    charset-normalizer==3.0.1
6    click==8.1.3
7    colorama==0.4.6
8    comm==0.1.2
9    cryptography==39.0.0
10   debugpy==1.6.5
11   decorator==5.1.1
12   Django==4.1.6
13   entrypoints==0.4
14   executing==1.2.0
15   Flask==2.2.2
16   Flask-WTF==1.1.1
```

The only thing we have to do is installing all the packages that the project depends on

through **pip install -r requirements.txt**

```
PROBLEMS    OUTPUT    DEBUG CONSOLE    TERMINAL                                                                        powe

(venv) PS E:\google download\Flask 2> pip install -r requirements.txt
Requirement already satisfied: asgiref==3.6.0 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 1)) (3.6.0)
Requirement already satisfied: asttokens==2.2.1 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 2)) (2.2.1)
Requirement already satisfied: backcall==0.2.0 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 3)) (0.2.0)
Requirement already satisfied: cffi==1.15.1 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 4)) (1.15.1)
Requirement already satisfied: charset-normalizer==3.0.1 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 5)) (3.0.1)
Requirement already satisfied: click==8.1.3 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 6)) (8.1.3)
Requirement already satisfied: colorama==0.4.6 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 7)) (0.4.6)
Requirement already satisfied: comm==0.1.2 in e:\google download\flask 2\venv\lib\site-packages (from -r requirements.txt (line 8)) (0.1.2)
```

The purpose of doing this is twofold:

1. Through requirements.txt, sponsors and future users can quickly get started with this function without using google or chatpgt to find out how to download a certain package.

2. Some functions may be updated or deleted after the package undergoes a version update, resulting in unusable functions. By keeping the package at the same version, the guarantee function can be used normally.

2. **(Path install) Poppler and tesseract ocr set up**

   Poppler installation guide video::

   https://www.youtube.com/watch?v=IDu46GjahDs

   Poppler is an open-source software library for rendering Portable Document Format (PDF) documents. It is written in C++ and provides various tools and utilities for working with PDF files.

   A. First of all, let's go to poppler's official website to download the latest package:

      https://blog.alivate.com.au/poppler-windows/

   - **PDFToText** – Extract all the text from PDF document. I suggest you use the -Layout option for getting the content in the right order.
   - **PDFToHTML** – Which I use with the -xml option to get an XML file listing all of the text segments' text, position and size, very handy for processing in C#
   - **PDFToCairo** – For exporting to images types, including SVG!
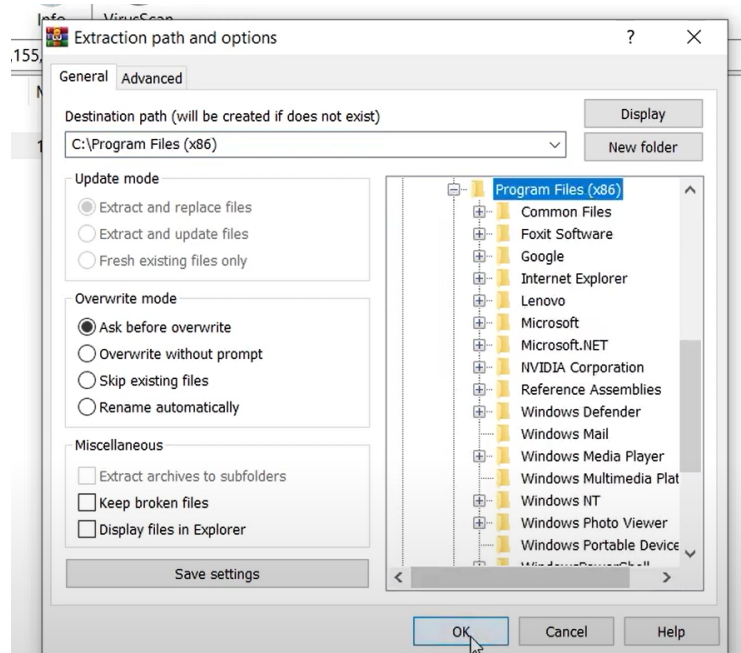   - Many more smaller utilities

   Download

   Latest binary : poppler-0.68.0_x86

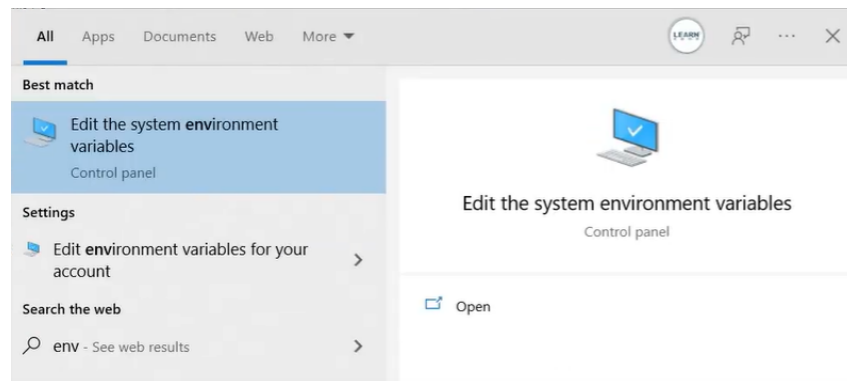   Older binaries:
   poppler-0.67.0_x86
   poppler-0.58.0_x86

B. Then unzip Latest binary: poppler-0.68.0_x86 to the C drive (you can choose any folder), but you must pay attention to the unified path. Suppose we set up in Program File(x86).
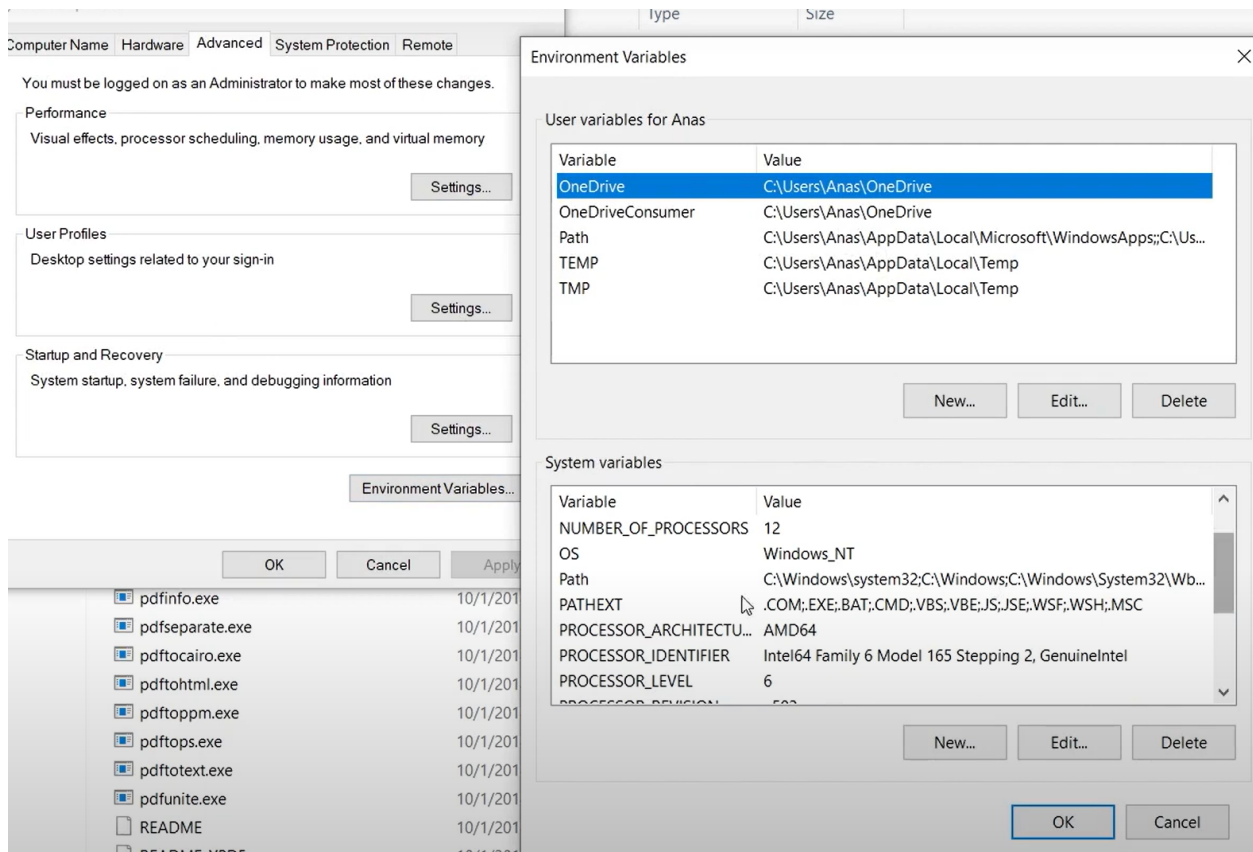

C. Then we click **"bin"** in the poppler folder of Program File (x86), pay attention to the key point to enter the **"bin"** folder, and then copy the path address of the folder.

D. Then we click on "Edit the system environment variables" in the computer,

Select "Path" in "**System variable**" in "Environment Variables."

Select "New" and enter the path address of poppler inside.
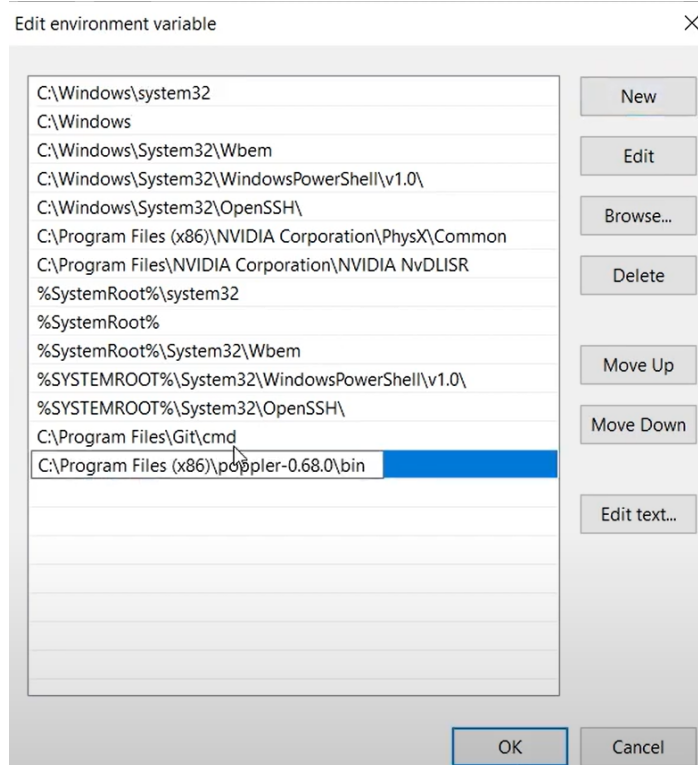
E. Finally, open the CMD control panel, enter **pdftoppm -h** to activate poppler.

Congratulations! The setting of poppler is complete.

Tesseract ocr installation guide video::

https://www.youtube.com/watch?v=Rb93uLXiTwA

Tesseract OCR is an open-source optical character recognition (OCR) engine developed by Google. OCR is the process of converting scanned images or printed text into machine-readable text that can be edited, searched, or analyzed by computer software. Tesseract OCR can recognize over 100 languages and is widely used in various applications such as document digitization, data entry, and text-to-speech conversion.

A. First of all, let's go to Tesseract ocr's official website to download the latest package: https://github.com/UB-Mannheim/tesseract/wiki

The latest installers can be downloaded here:

- tesseract-ocr-w32-setup-5.3.0.20221222.exe (32 bit) and
- tesseract-ocr-w64-setup-5.3.0.20221222.exe (64 bit) resp.

Please choose to download the 32bit or 64bit version based on your windows system.

B. Note that you can choose to install other language translators in Tesseract OCR, if necessary, please choose by yourself.
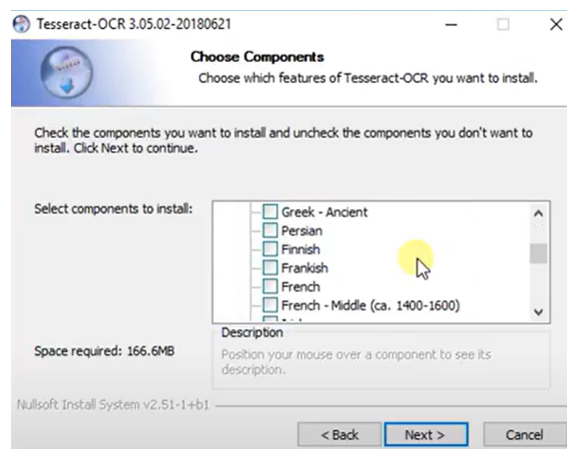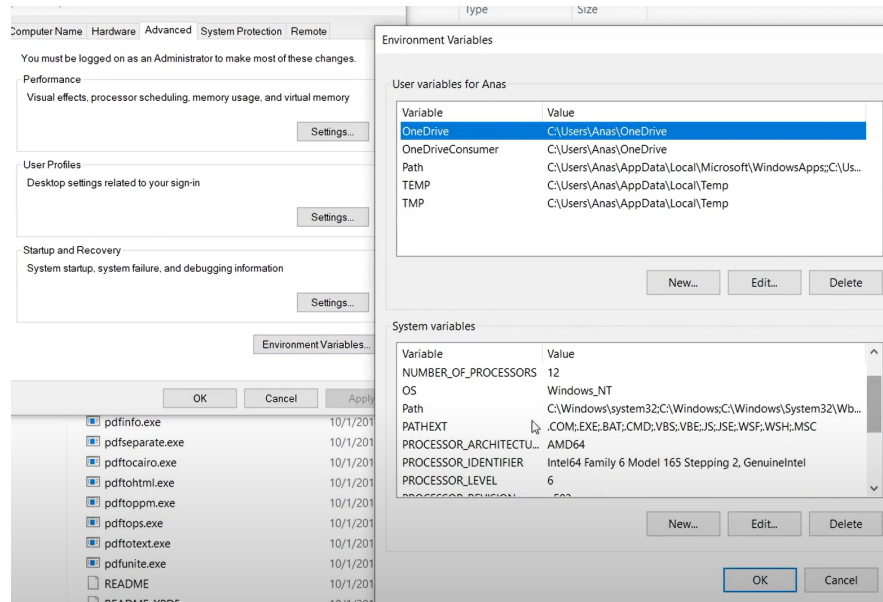
C. Please pay attention to the path address of your installation because we also need to set
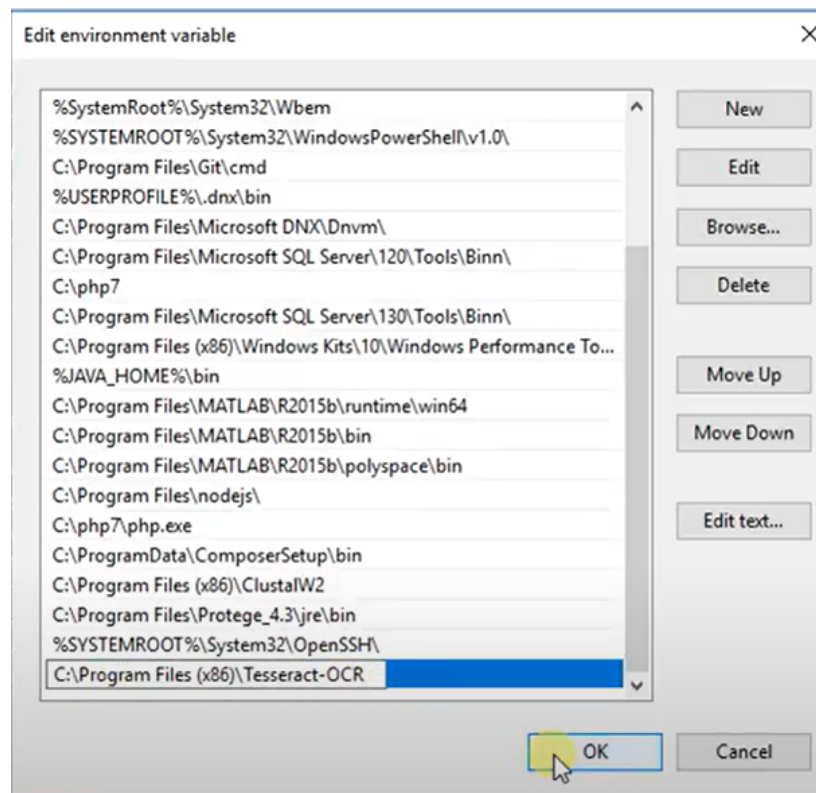   it in the environment variable.
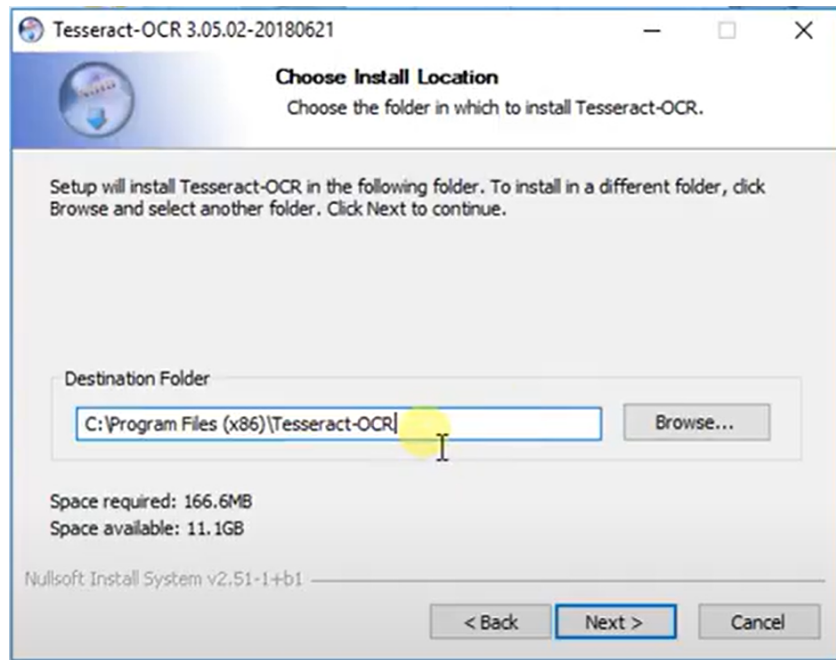
D. Then we click on "Edit the system environment variables" in the computer,
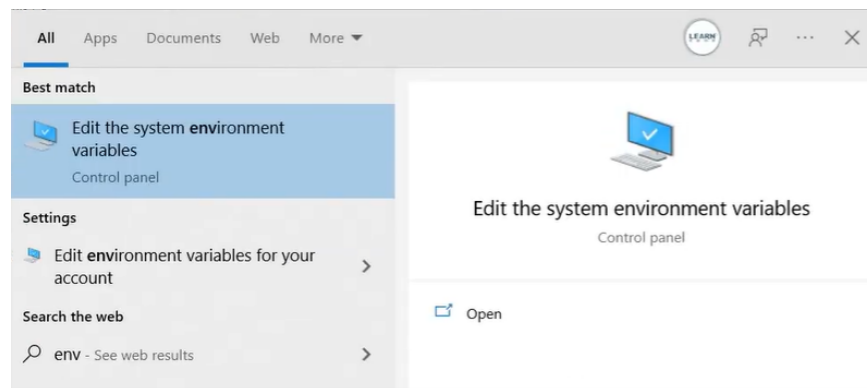
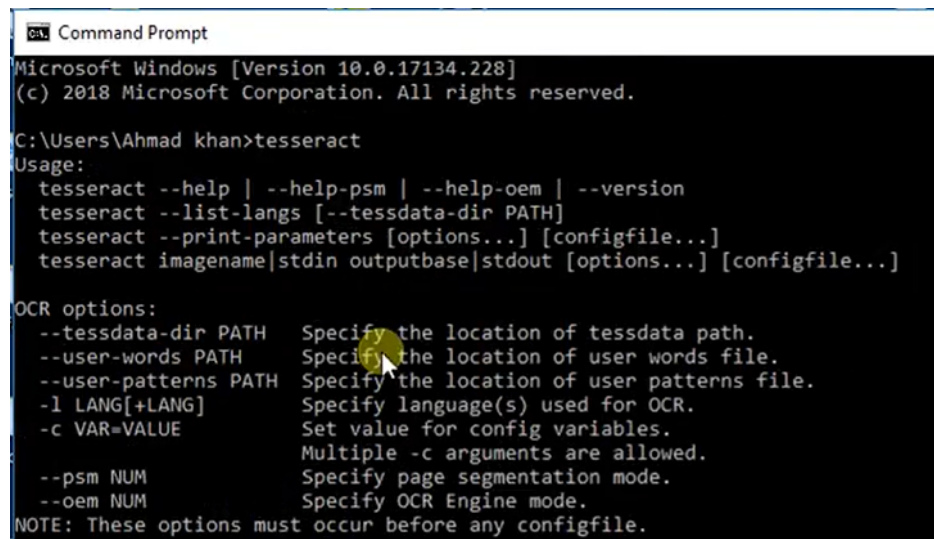Select "Path" in "**System variable**" in "Environment Variables."

Select "New" and enter the path address of Tesseract ocr inside.

E. Finally, open the CMD control panel, and enter "tesseract" to activate tesseract ocr.

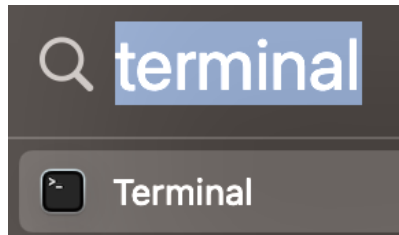Congratulations! The setting of tesseract ocr is complete.

Final work:

In the last step, don't forget to replace your own Tesseract ocr environment on lines 110 and 111 of app.py.

Finally, click the "F5" key to enter the debug interface.

Step-by-Step Guide to Setting up a Flask App on a Mac Machine

1. Open terminal.

2. Install Homebrew on the Mac machine. Homebrew provides an easy way to install packages.You will be asked to enter your password. Please type it in and enter.

For Apple Intel Mac machines:
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
For Apple Silicon Mac machines:
This can be done through Terminal with the command
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"
Then paste and enter:
eval "$(homebrew/bin/brew shellenv)"
brew update --force --quiet
chmod -R go-w "$(brew --prefix)/share/zsh"

3..Check to see if Python is installed with the command "python3 --version". If not then install Python 3, this can be done through Homebrew w
ith the command "brew install python3".

4. Download and unzip the Flask application, make sure to cd to the directory containing the unzipped application with command "cd FlaskNEW"

5. Create a virtual environment for the Flask app with command "python3 -m venv venv" and activate it with ". venv/bin/activate".

6. Install dependencies with command "pip install -r requirements.txt".

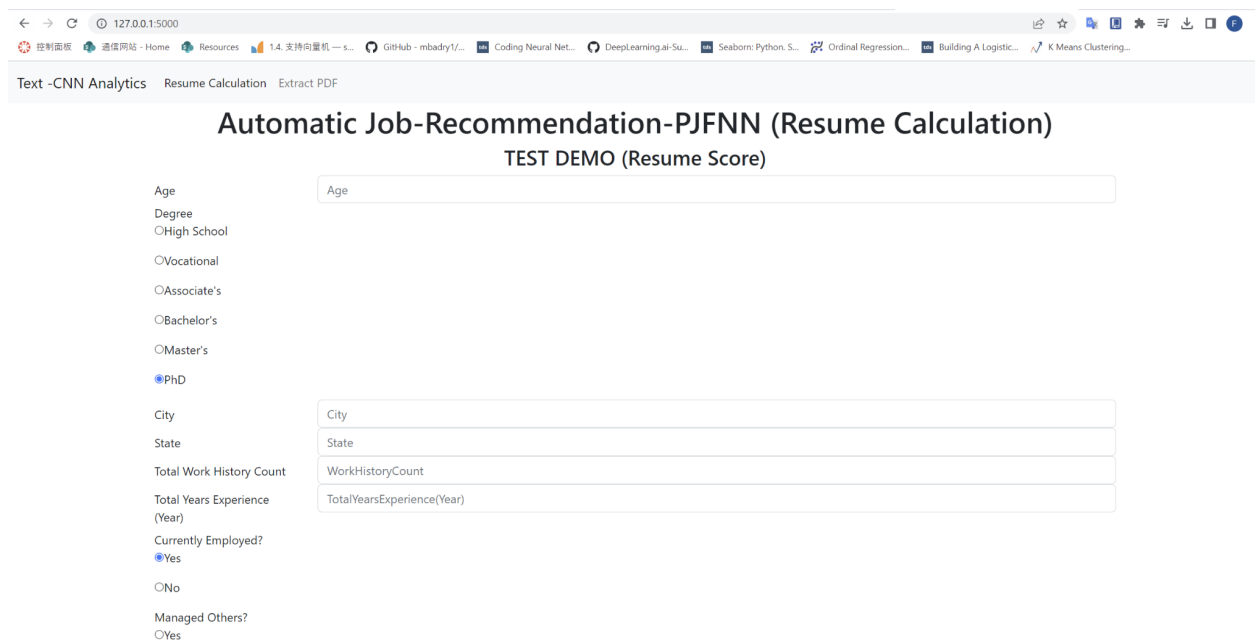7. Install packages such as poppler and tesseract used for document and image processing with command "brew install poppler tesseract". This will likely take awhile to download as both packages are powerful tools for working with documents and images in a variety of ways.

```
$ brew install poppler tesseract
```

8. Run the application with the command "flask run".

9. Open a web browser and enter the address "http://127.0.0.1:5000" to view the application.

Team menbers (contact)：

1. Jianming Ma: ma.jianm@northeastern.edu

2. Harshita Atray: atray.h@northeastern.edu