

Data Extractor Utility (Setup & Usage Guide)

Setup: (Page 1 : Page 8)

Usage: (Page 9)

Notes: (Page 10)

To use the utility, Python, Tesseract, and some packages must be installed properly. Please follow the guide step by step to get them installed successfully. Skipping any step may raise errors in the future.

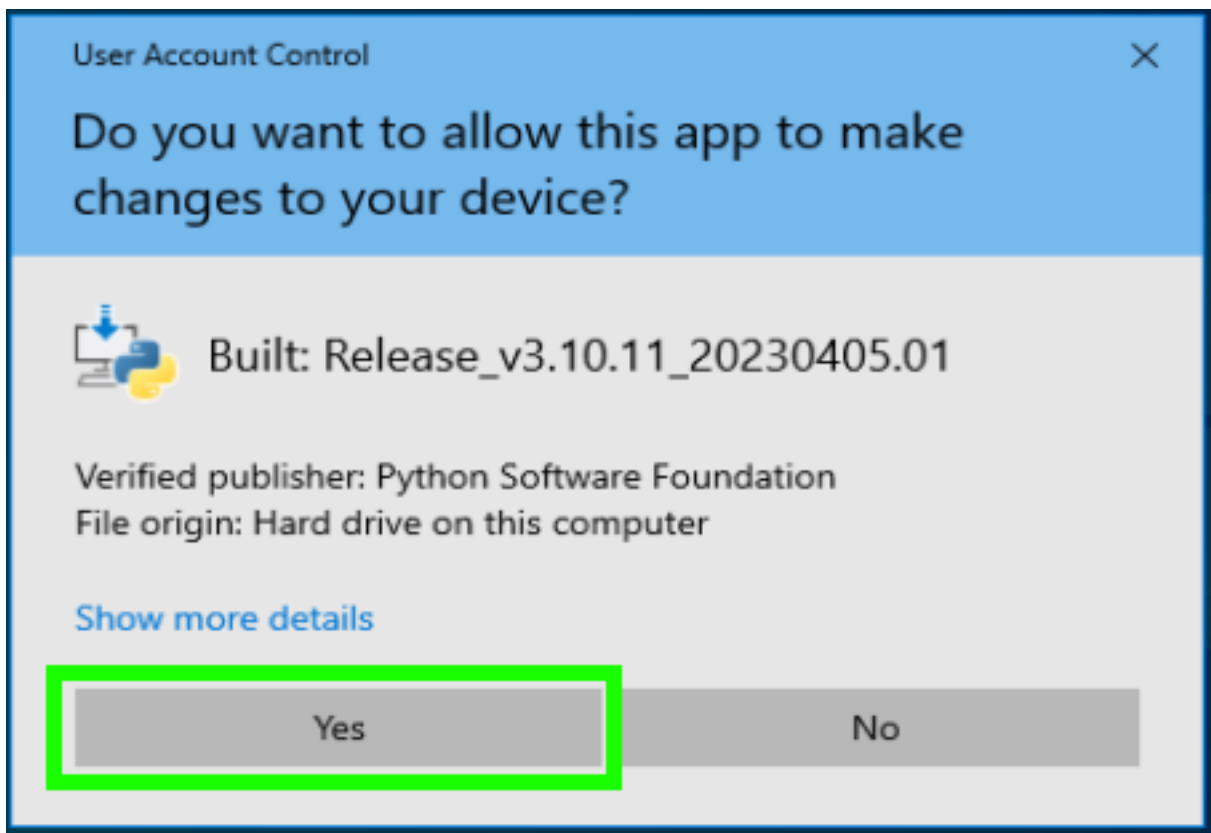
Setup:

1) Please use the following link to install Python 3.9 installer for Windows:

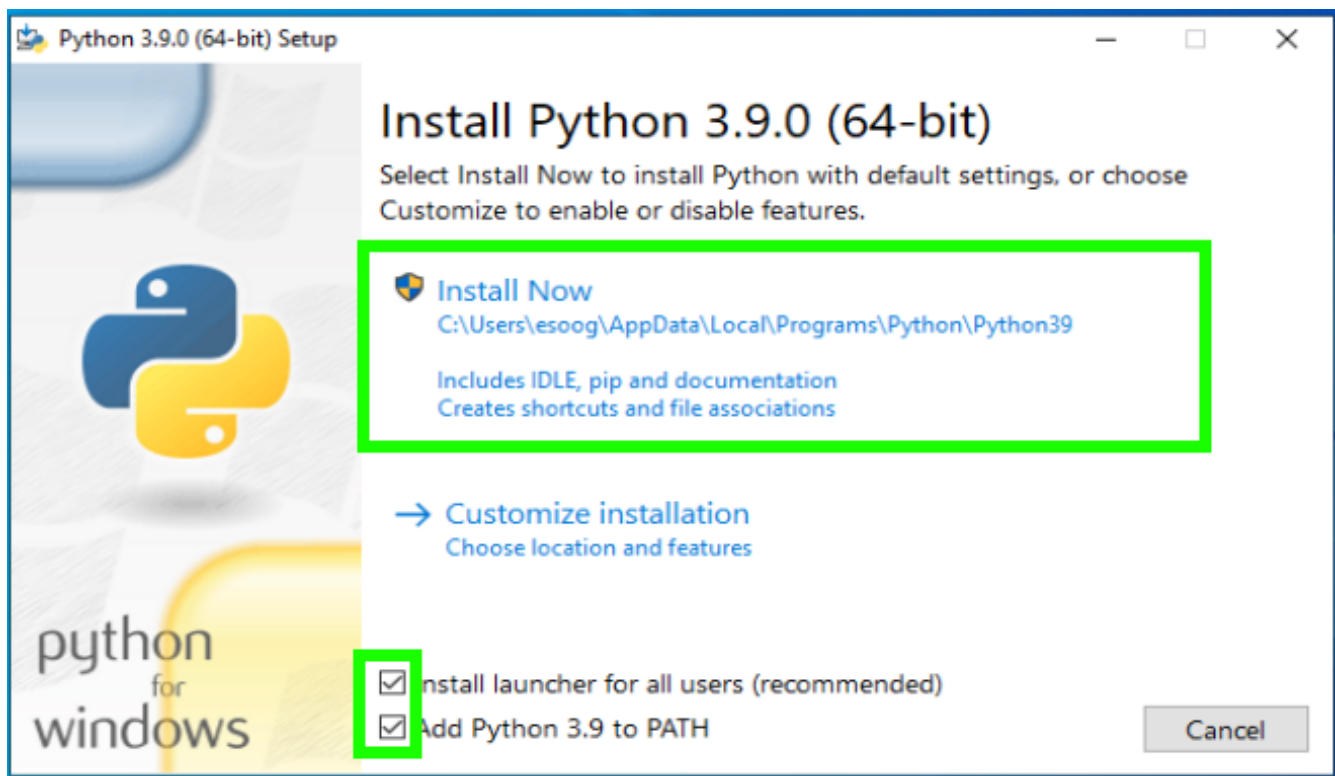
(Windows - 64 bit): <https://www.python.org/ftp/python/3.9.0/python-3.9.0-amd64.exe>

(Windows - 32 bit): <https://www.python.org/ftp/python/3.9.0/python-3.9.0.exe>

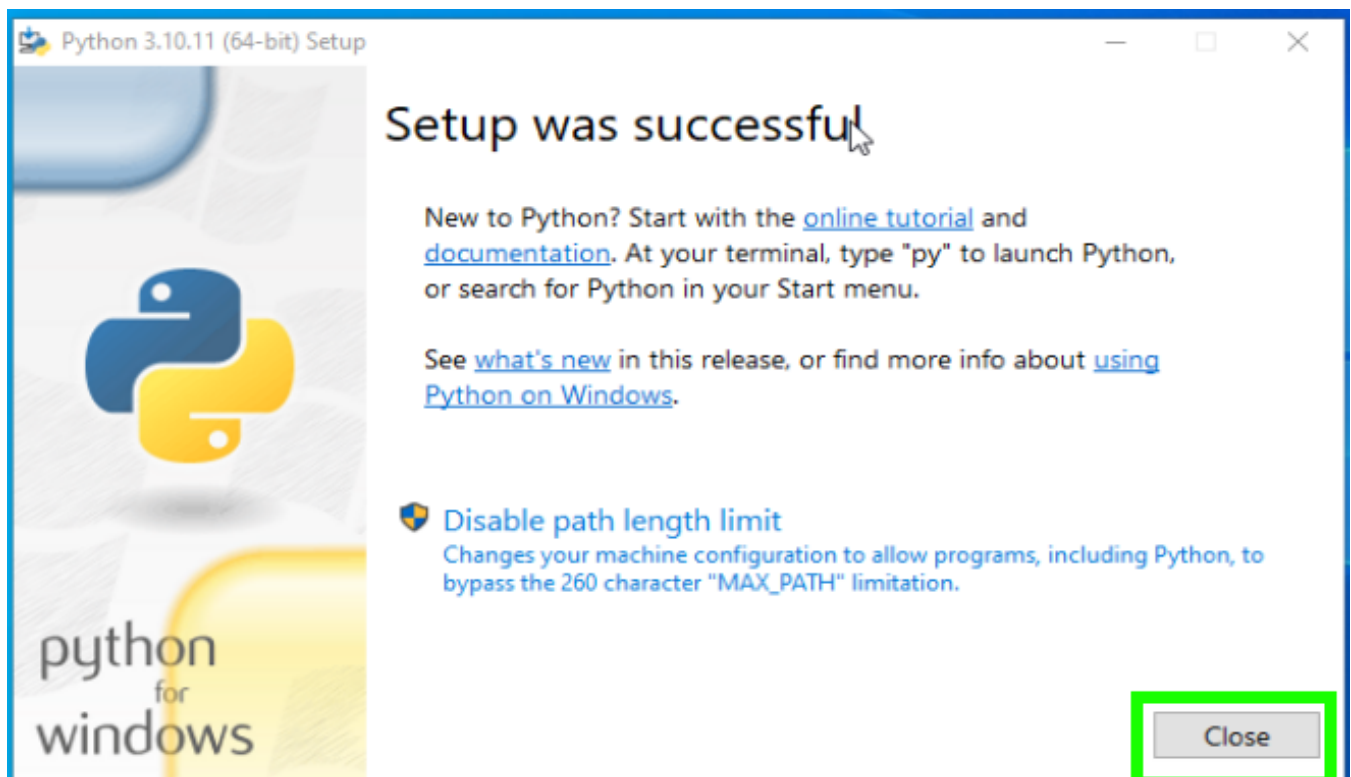
2) Run your downloaded Python installer file and allow it to make changes to your computer. (Always accept permissions when asked at any step.)



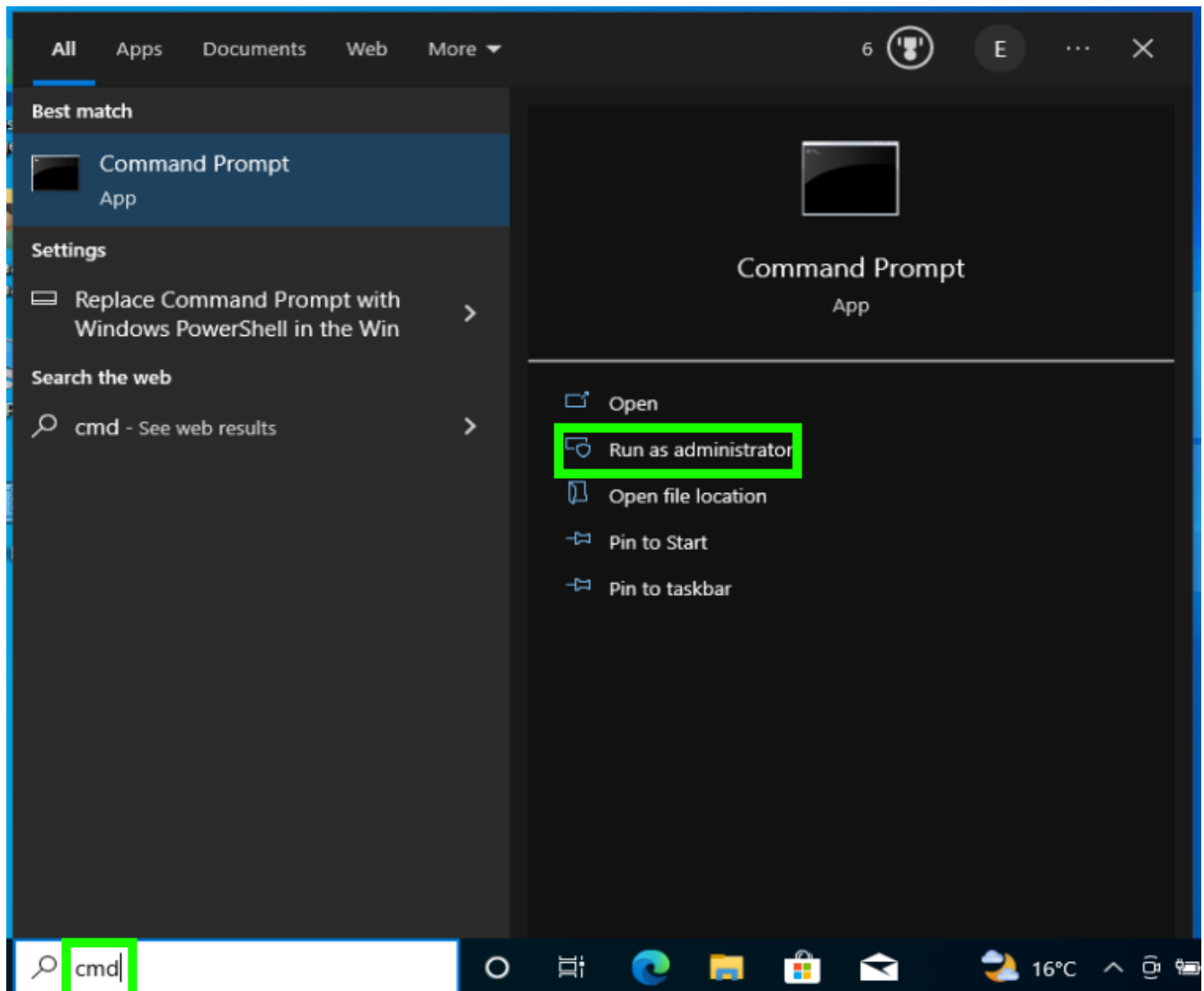
3) Check the boxes “Install for all users” and “Add path”. Then, click “Install Now”.



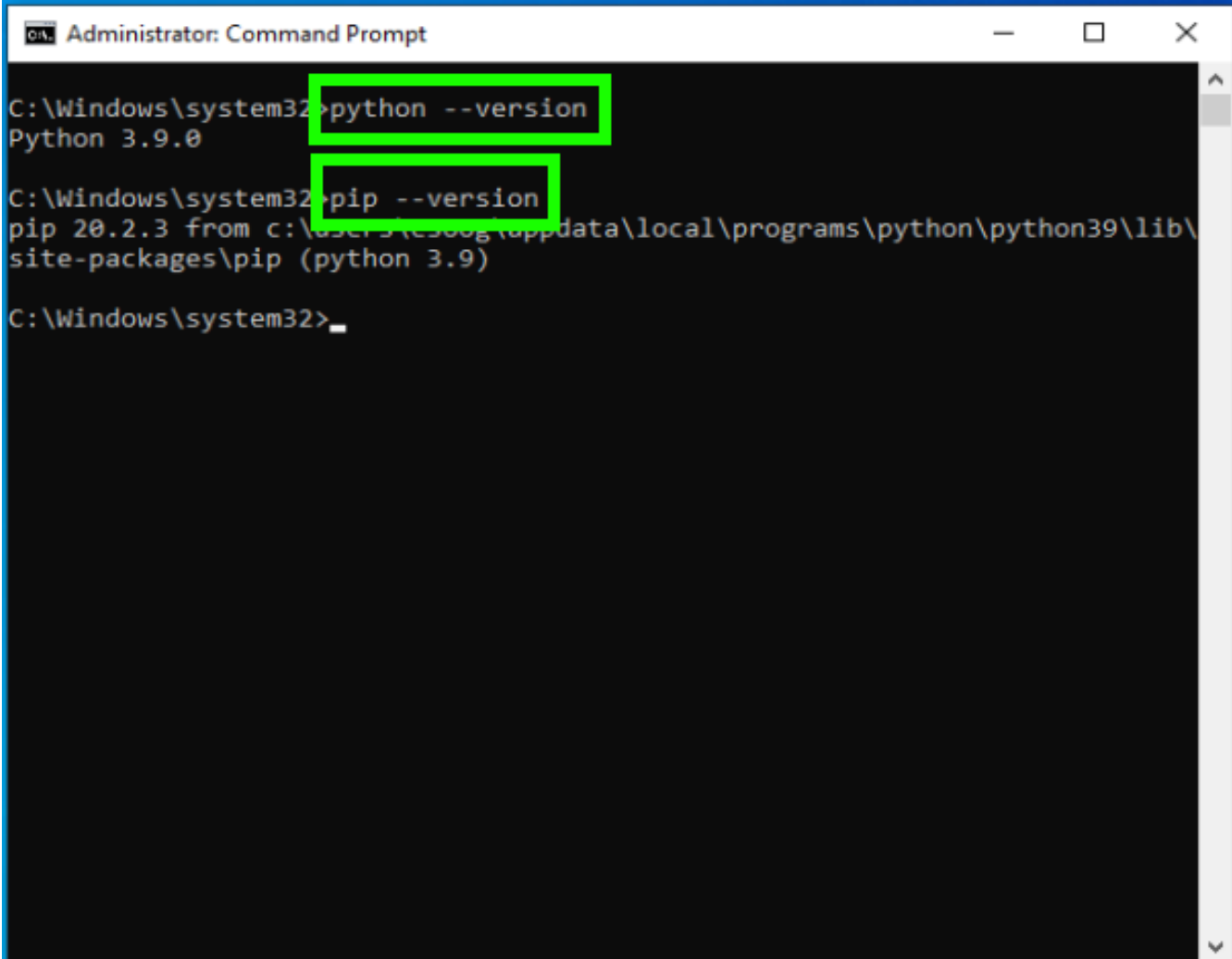
4) Once it is installed, click “Close”.



5) In your Windows Taskbar, type “cmd” to see the command prompt. Then, open it as an administrator.



6) Type “python --version” and press enter. If Python was installed properly, you should see the python version number. Then, type “pip --version” and press enter. You should be able to see the pip version number too. The “--” are 2 hyphens without spaces.



```
Administrator: Command Prompt
C:\Windows\system32>python --version
Python 3.9.0

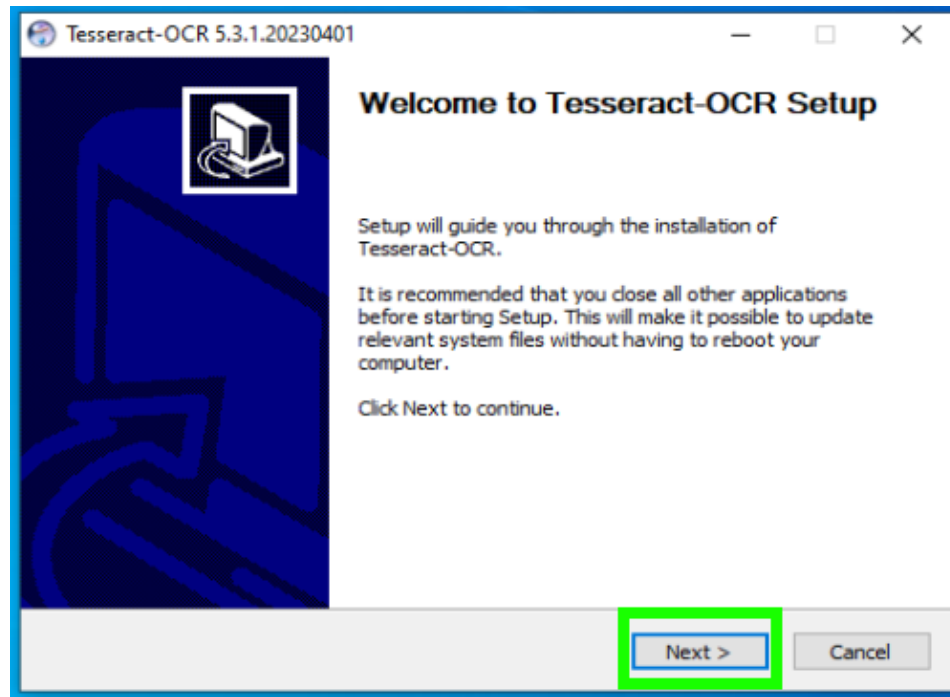
C:\Windows\system32>pip --version
pip 20.2.3 from c:\users\c300g\appdata\local\programs\python\python39\lib\
site-packages\pip (python 3.9)

C:\Windows\system32>_
```

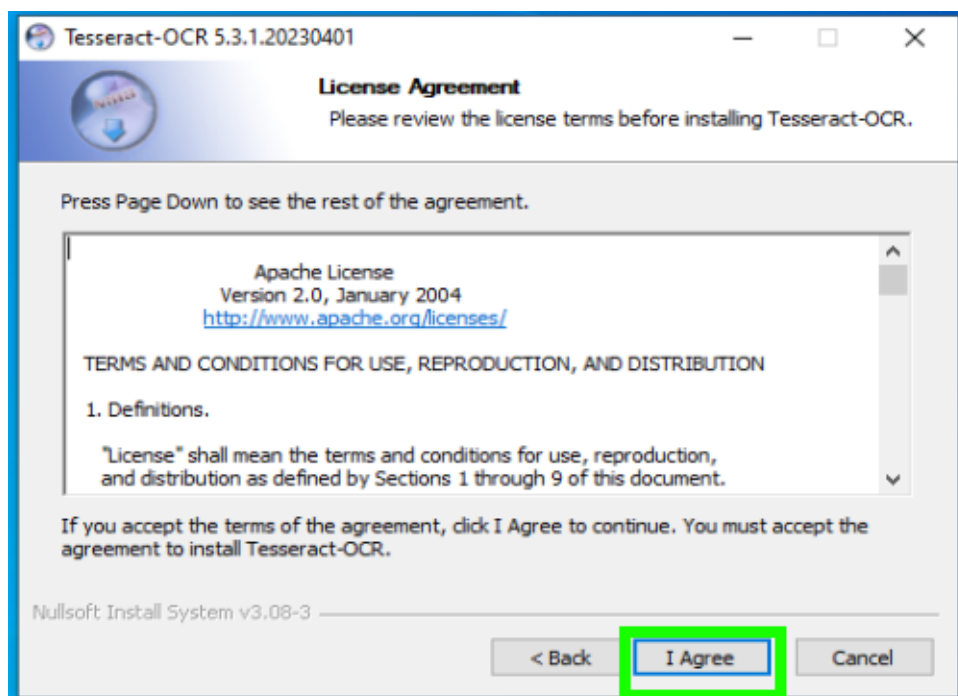
7) Now, download Tesseract using the following link:

<https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-w64-setup-5.3.1.20230401.exe>

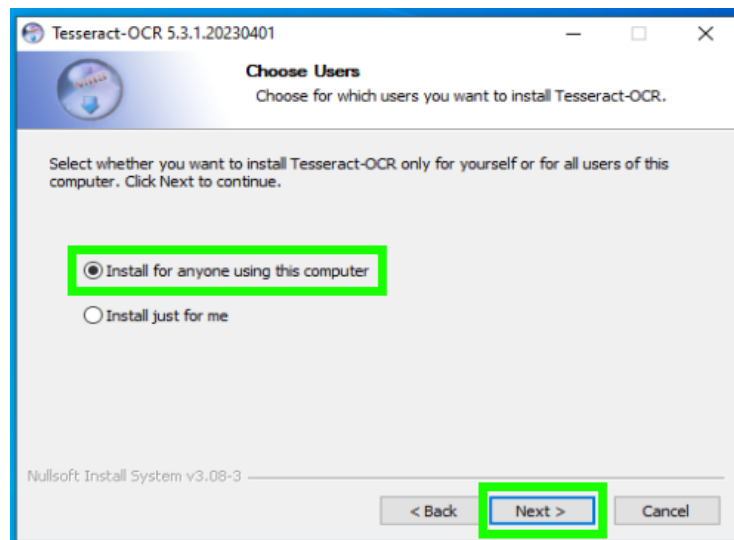
8) Run the file as an administrator to start. Then, click “Next”.



9) Click “I Agree”.

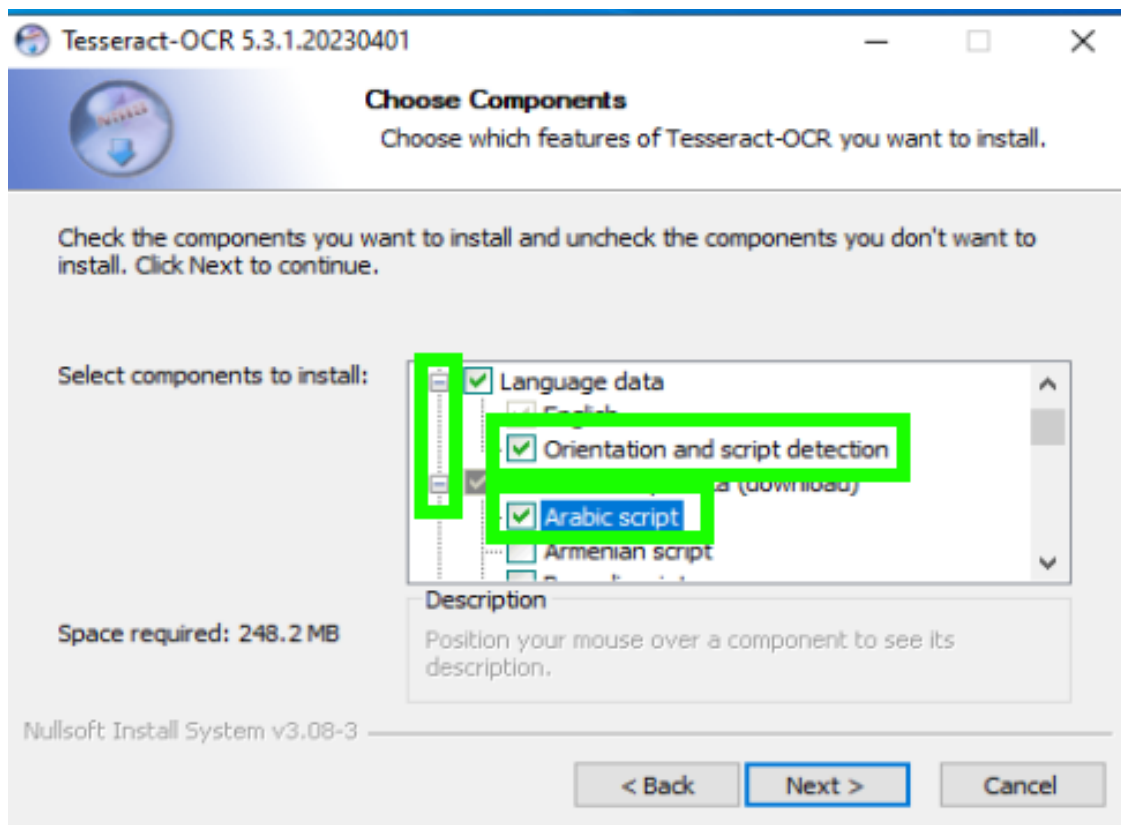


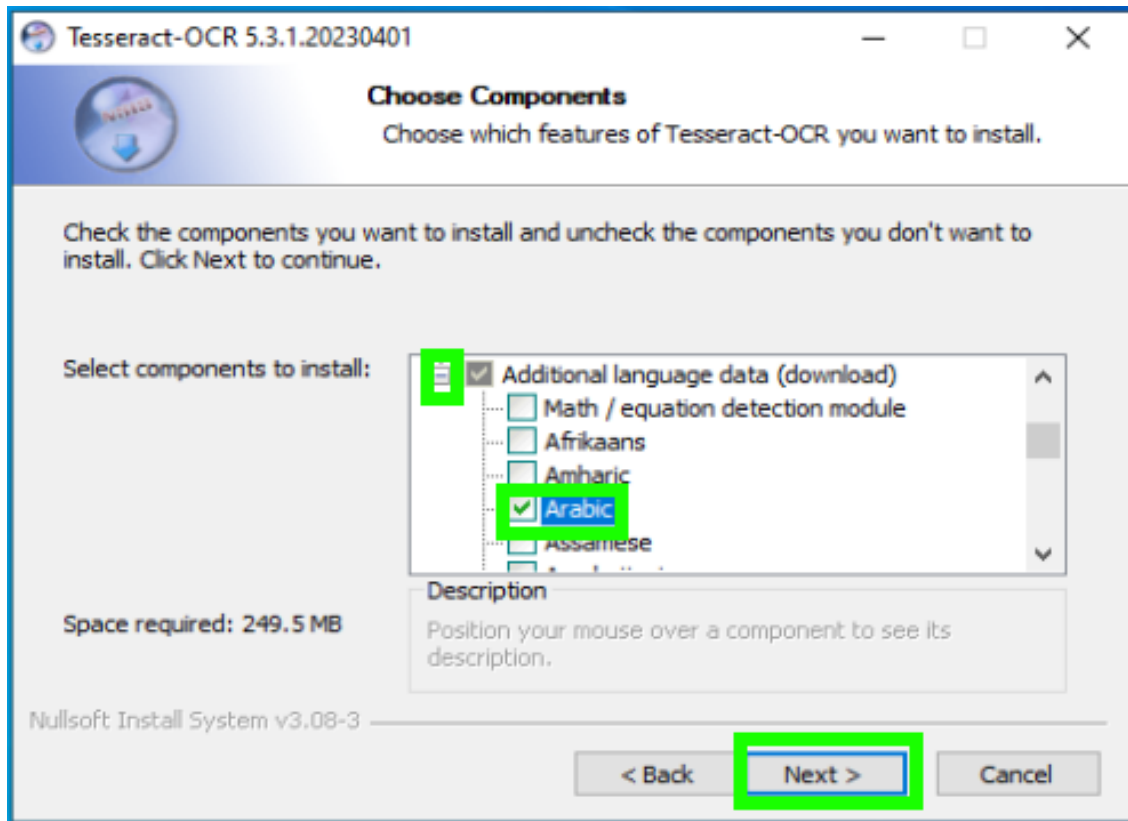
10) Choose “Install for anyone using this computer”. Then click “Next”.



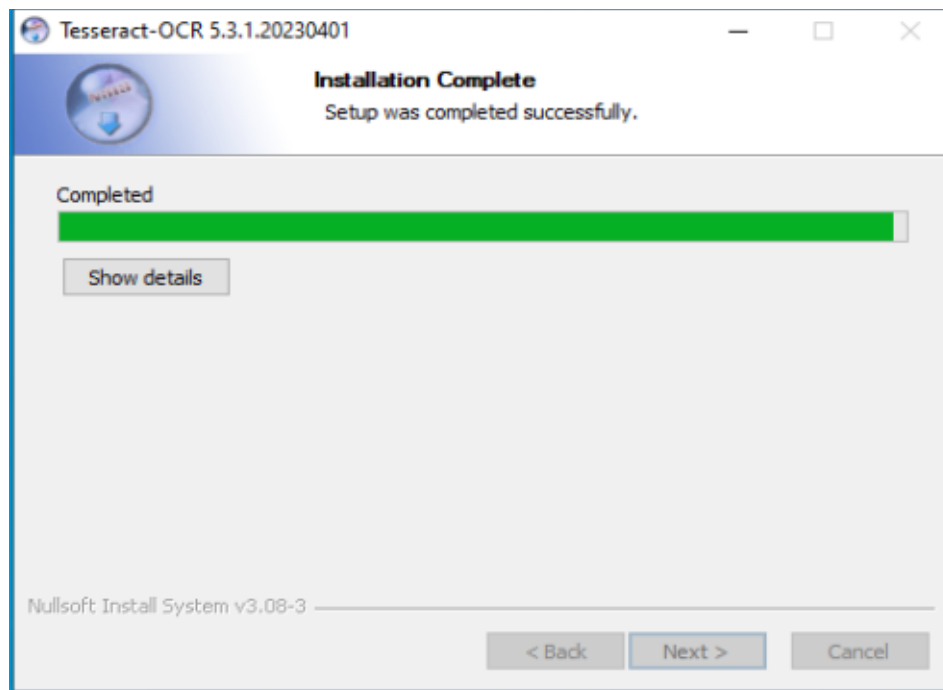
11) Click on the choice collections “+” to unpack them.

- Make sure “Orientation and script detection” is selected.
- Select “Arabic script” from “Additional script data (download)”.
- Scroll down and select “Arabic” from “Additional language data (download)”.

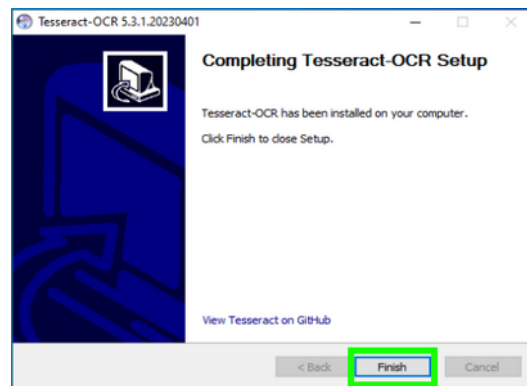




12) Don't change the installation directory if asked. Wait until the installation is completed.



13) Tesseract is now installed. Click “Finish”.



14) Now, let's install the packages required by the utility. Open the “Command Prompt” as administrator using your taskbar as in step no.5.

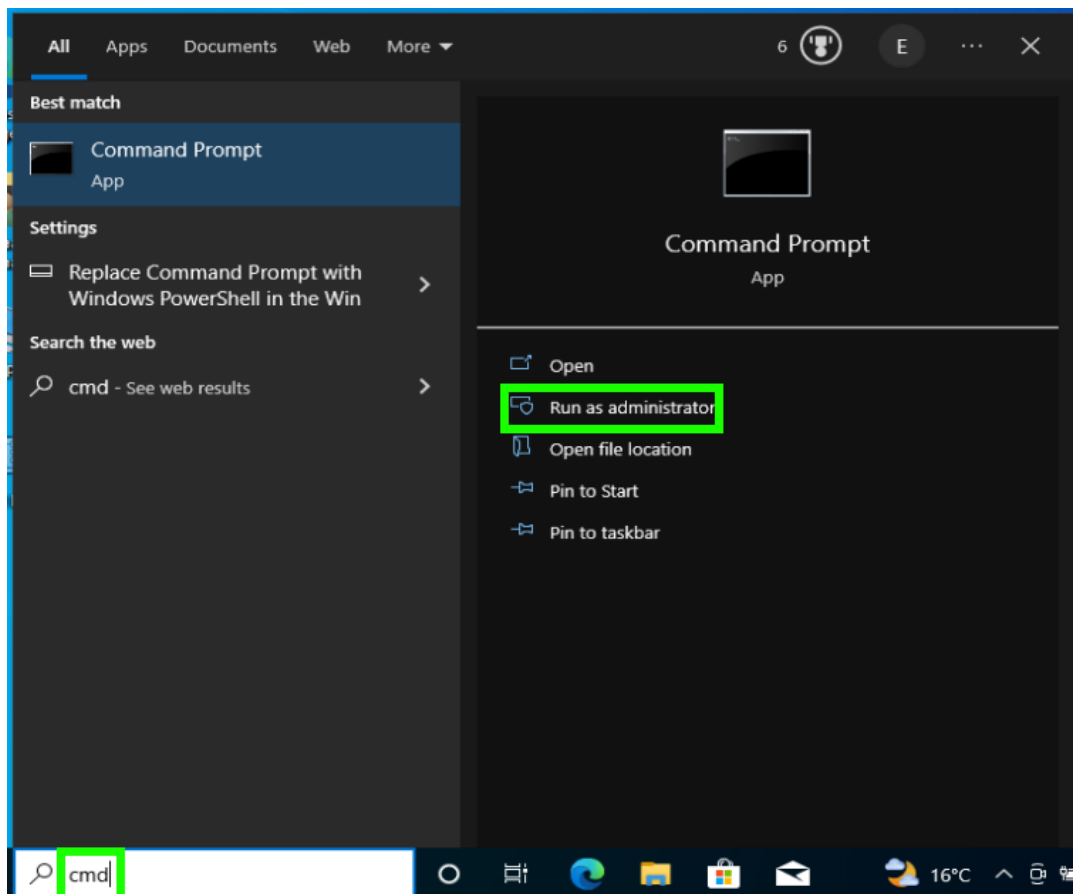
15) Find the path where you downloaded the “requirements.txt” files your received. In this example, the file path is “C:\Work\requirements.txt”. So, we will write “pip install -r C:\Work\requirements.txt” and press enter. Your utility packages will be installed.
Note: If your path includes **spaces**, write it between **double quotations** “C:\My Work\requirements.txt”

A screenshot of an 'Administrator: Command Prompt' window. The title bar is blue. The command prompt shows the command `C:\Windows\system32\cmd.exe pip install -r C:\Work\requirements.txt` being executed. The output shows the installation progress for several packages: click==8.0.3, Pillow==9.5.0, PyMuPDF==1.20.2, pytesseract==0.3.10, colorama; platform_system == "Windows", and packaging>=21.3. Each package has a progress bar and download information. At the bottom, it says 'Installing collected packages: colorama, click, Pillow, PyMuPDF, packaging, pytesseract' and 'Successfully installed Pillow-9.5.0 PyMuPDF-1.20.2 click-8.0.3 colorama-0.4.6 packaging-23.0 pytesseract-0.3.10'. There is a yellow warning message: 'WARNING: You are using pip version 20.2.3; however, version 23.0.1 is available. You should consider upgrading via the 'c:\users\esoog\appdata\local\programs\python\python39\python.exe -m pip install --upgrade pip' command.'

If no errors occurred during the previous steps, we can run the utility now. The steps above are done once only.

Usage:

1) Type “cmd” in your taskbar to open your Command Prompt.



2) Write your command in form of “python <utility file path> <PDFs input folder path> <output folder path>”. Add only one space after each part. If your path includes **spaces**, write it between **double quotations**.

A screenshot of a Windows Command Prompt window. The title bar reads 'C:\Users\esoog\ - Command Prompt'. The command prompt shows the following text:

```
C:\Users\esoog>python C:\Work\extractor.py "C:\Work\my pdf files" "C:\Work\output folder"
```

 The command and its arguments are highlighted with a green box. Below the command, the output of the script is displayed:

```
Validating your files...
Please don't use the CSV file until the end of the process.
Running the OCR...
Your files are being processed... 100%
Your data is ready in C:\Work\output folder
```

 The prompt then returns to

```
C:\Users\esoog>
```

Your files should now be ready in your output folder.

Notes:

- 1) Remember to use the double quotations for paths that include spaces.
- 2) Don't move the CSV file at all during the process.
- 3) If you already have a CSV file that is an output of an earlier process in the output folder, move it to avoid overwriting the file.
- 4) If you get an error message, read it carefully to know the reason. For example:
 - If you write an input path that doesn't exist, you'll get the error message "Please use a valid input directory path."
 - If you pass the same path as input folder and output folder, you'll will get a warning "Warning: You are using the same path as the input and output folder."
 - If you pass an input folder that doesn't contain PDF files, you'll get "No PDF files in the input directory."
 - If the CSV file cannot be written, you'll get "Cannot write files to the output directory."
 - If processing a file is failed, you'll get the message "Reading or writing <file_name> has failed." but the rest of the files will be processed.
 - If you get "Reading or writing <file_name> has failed." for all files when you run the utility for the first time, please contact me.