How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Table of Contents

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Overview

Tesseract is an open-source text recognition (OCR) Engine, available under the Apache 2.0 license.

Major version 5 is the current stable version and started with release 5.0.0 on November 30, 2021.

Newer minor versions and bugfix versions are available from GitHub.

The latest source code is available from the main branch on GitHub. Open issues can be found in the issue tracker, and planning documentation.

Tesseract can be used directly via the command line, or (for programmers) by using an API to extract printed text from images. It supports a wide variety of languages. Tesseract doesn't have a built-in GUI, but there are several available from the 3rdParty page. External tools, wrappers, and training projects for Tesseract are listed under AddOns.

Tesseract can be used in your project, under the terms of the Apache License 2.0. It has a fully featured API and can be compiled for a variety of targets including Android and the iPhone. See the 3rdParty and AddOns pages for samples of what has been done with it.

In this tutorial, we'll be showing you how to install Tesseract OCR for Windows. On a Mac, this is fairly straightforward, but on Windows, it's a little more complicated because we need to download the .exe from UB Mannheim, then

Step 1: Install Tesseract OCR .exe Installer from UB Mannheim

The first step to install Tesseract OCR for Windows is to download the .exe installer that corresponds to your machine's operating system (related: how to tell if you have Windows 64-bit or 32-bit). Chances are, if you're running any version of Windows later than Windows XP, you're likely running Windows 64-bit on your machine.

Links to the .exe installer for both 32-bit and 64-bit operating systems are provided below:

Download the 64-bit Tesseract OCR for Windows .exe installer from UB Mannheim here:

https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-w64-setup-v5.0.0-alpha.20210506.exe

Download the 32-bit Tesseract OCR for Windows .exe installer from UB Mannheim here:

https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-w32-setup-v5.0.0-alpha.20210506.exe
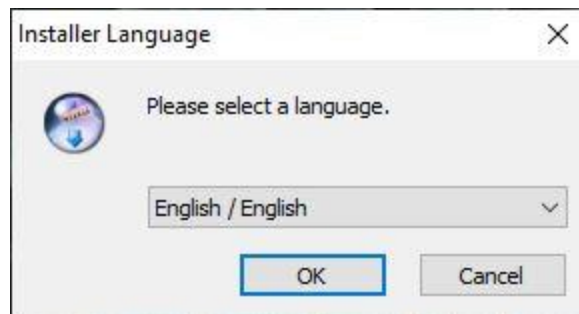
Step 2: Configure Installation

Next, we'll need to configure the Tesseract installation. If you're feeling confident and only want to run Tesseract OCR for Windows with the default language of English, running through the installation screens with all the default options selected should work.

If you would like to extract text from multiple languages or would like a more detailed explanation, I have provided screenshots of each step and their significance below.

Installer Language

This is just the language for the dialog boxes and helps info. If we want to run Tesseract OCR for Windows on multiple languages, we'll have an opportunity to install them later.
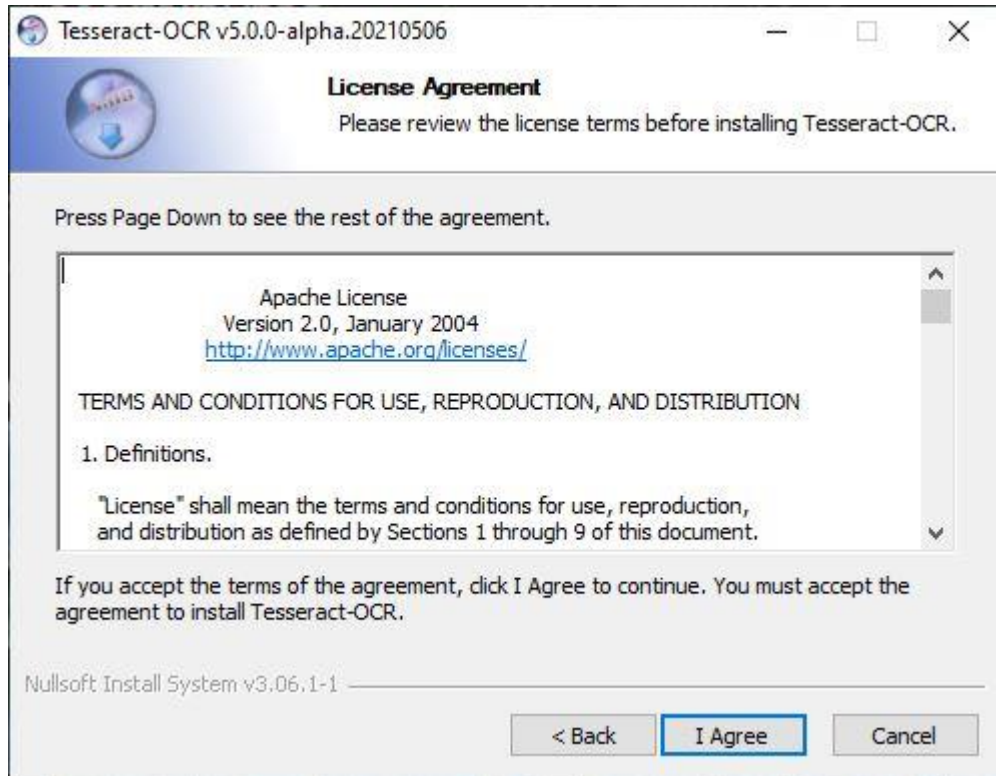
Installer language for Tesseract OCR for Windows

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

License Agreement

Tesseract OCR is licensed under the open-source Apache License 2.0, so you are entitled to redistribute Tesseract and modified versions of Tesseract without concern for royalties. You'll need to accept the license agreement to continue the installation.

Agree to the Apache License 2.0 to continue with the installation.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Tesseract OCR Setup

The setup screen recommends that all other applications are closed before continuing with the installation. It's not necessary to close all of your other applications to install Tesseract OCR for Windows, but if you don't, you may need to restart your computer for the installation to be completed.
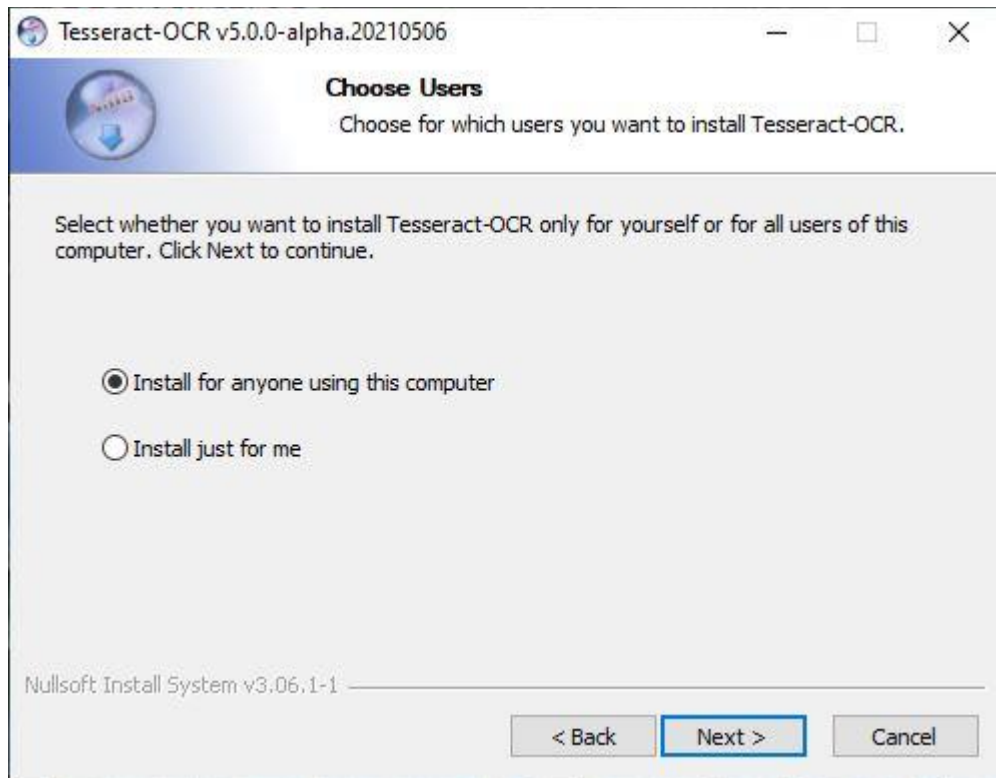

The Tesseract OCR for Windows installation screen.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Choose Users

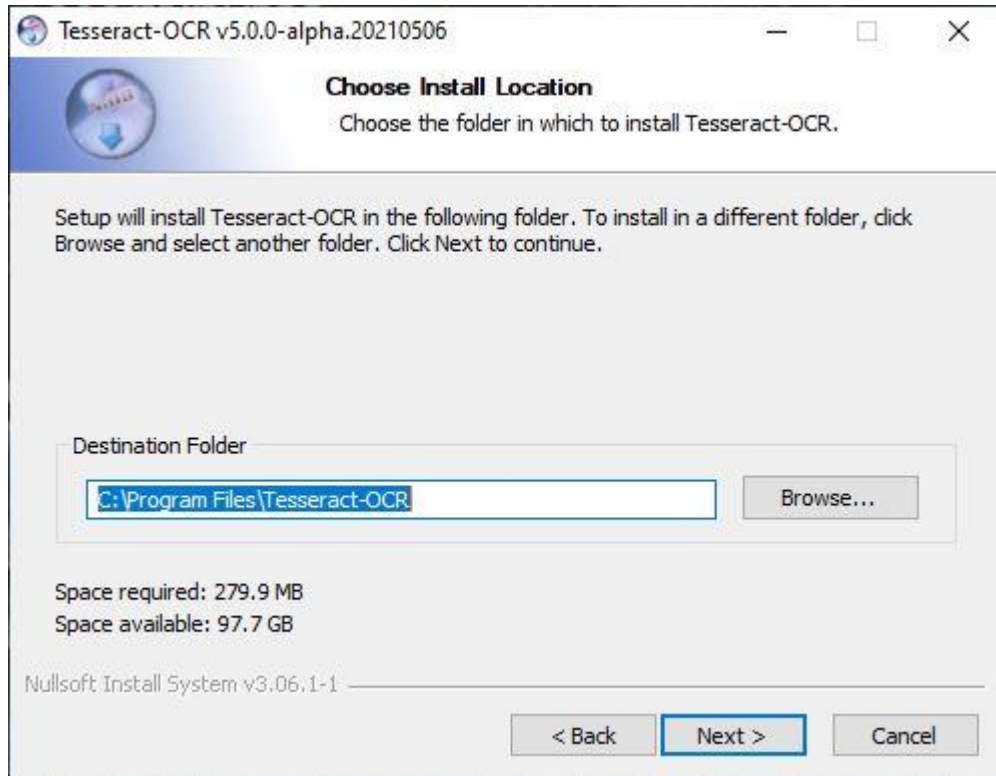Here we will choose which users on our machine we will install Tesseract for.

Choose the users whom we want to install Tesseract for.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Choose Install Location

Next, we'll choose the installation location. Before proceeding to the next step, make sure to copy the install location to a .txt file. We will need to add the installation location to our machine's environment variables once the installation is complete.
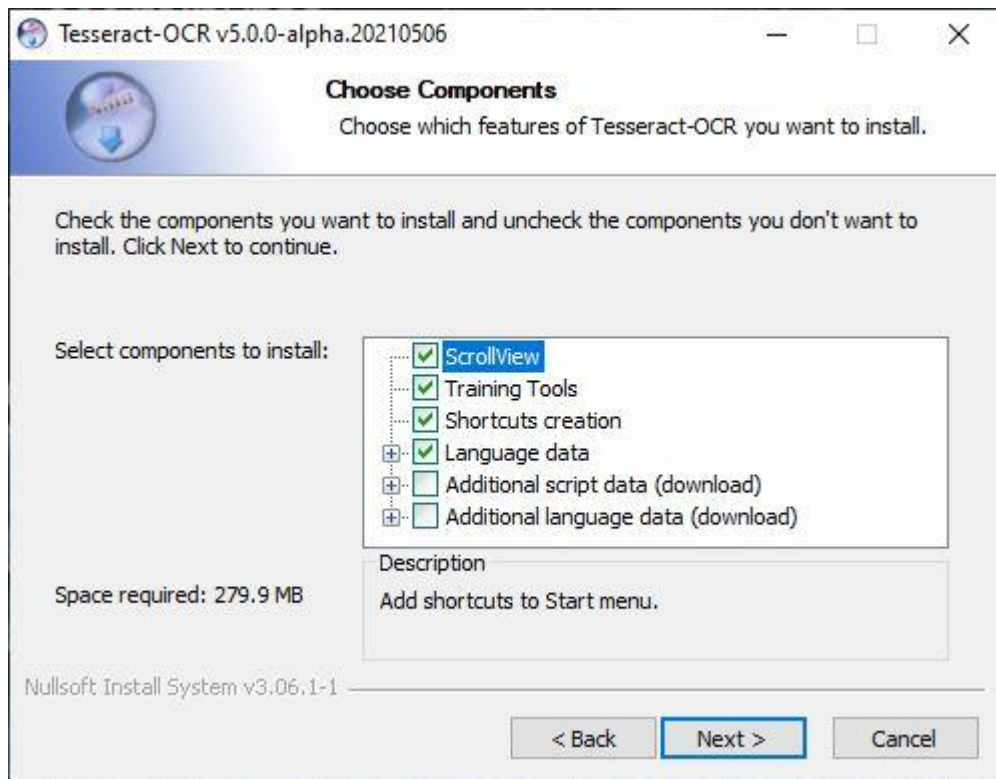

Choose the installation location.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps
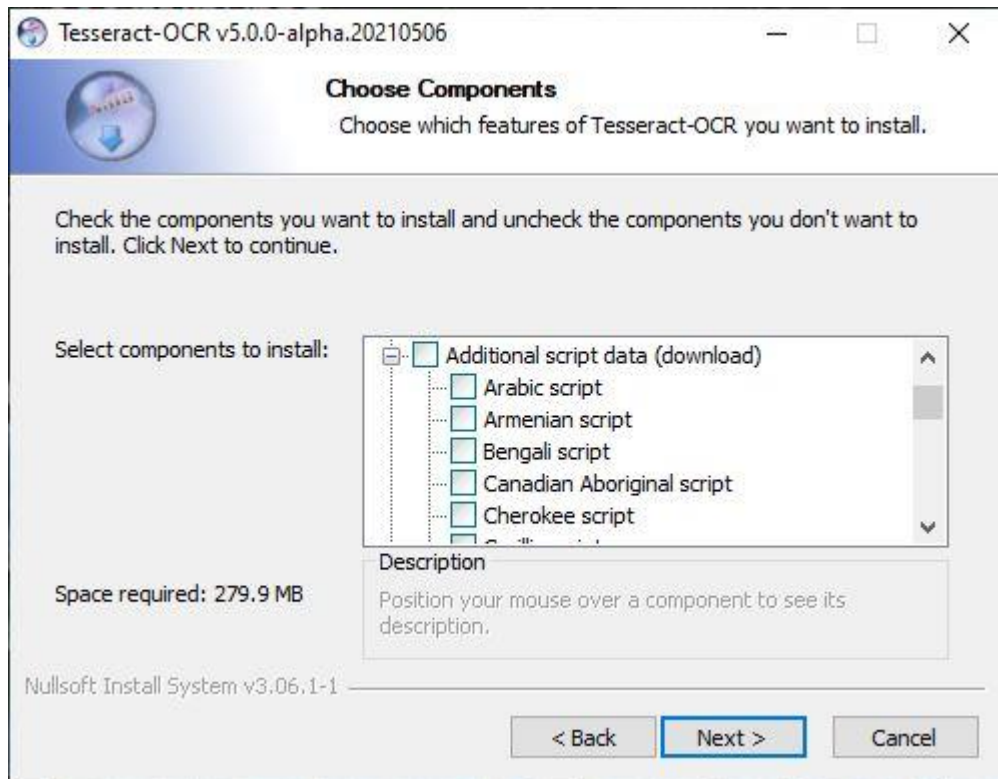
Choose Components

By default, the ScrollView, Training Tools, Shortcuts creation, and Language data are all selected. Unless you have a specific reason not to install these, we will want to keep all of these selected.

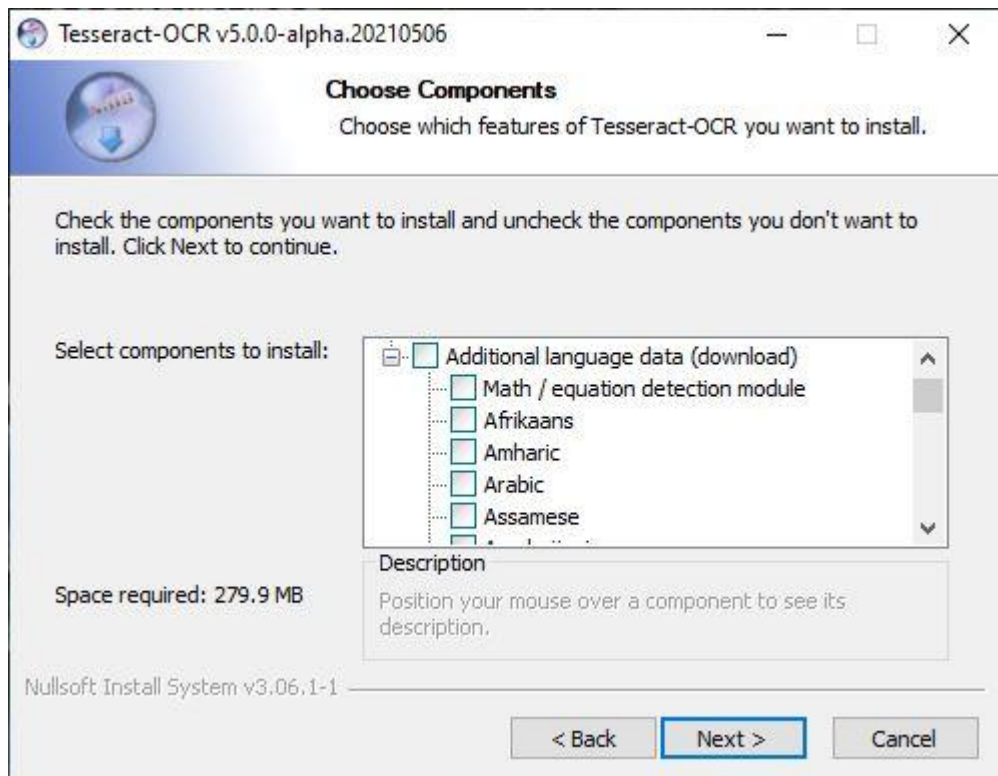Default Tesseract OCR for Windows installation components.

If we scroll down some and expand the 'Additional script data', we will see that we have the option to download and install additional script data. This can help improve the accuracy of text extraction from certain scripted languages. It's up to you if you want to install these.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Optional script installation components

If we go down one box further, we'll see that we have the option to download and install additional language data. If you plan on extracting text in multiple languages, look for the languages you plan to work with and check the box next to them.
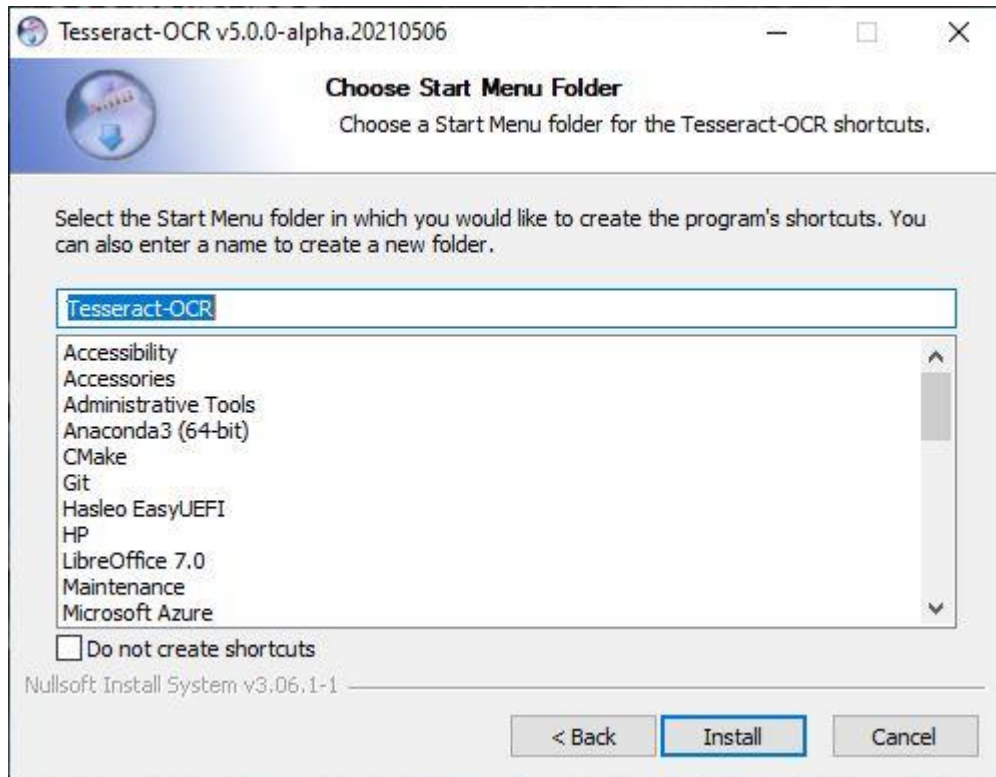
How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Add additional languages to use with Tesseract OCR for Windows.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Choose the Start Menu Folder

In the last step of the installation, we'll be asked to choose the start menu folder for Tesseract OCR for Windows shortcuts. I've left mine set to the default name 'Tesseract-OCR'.

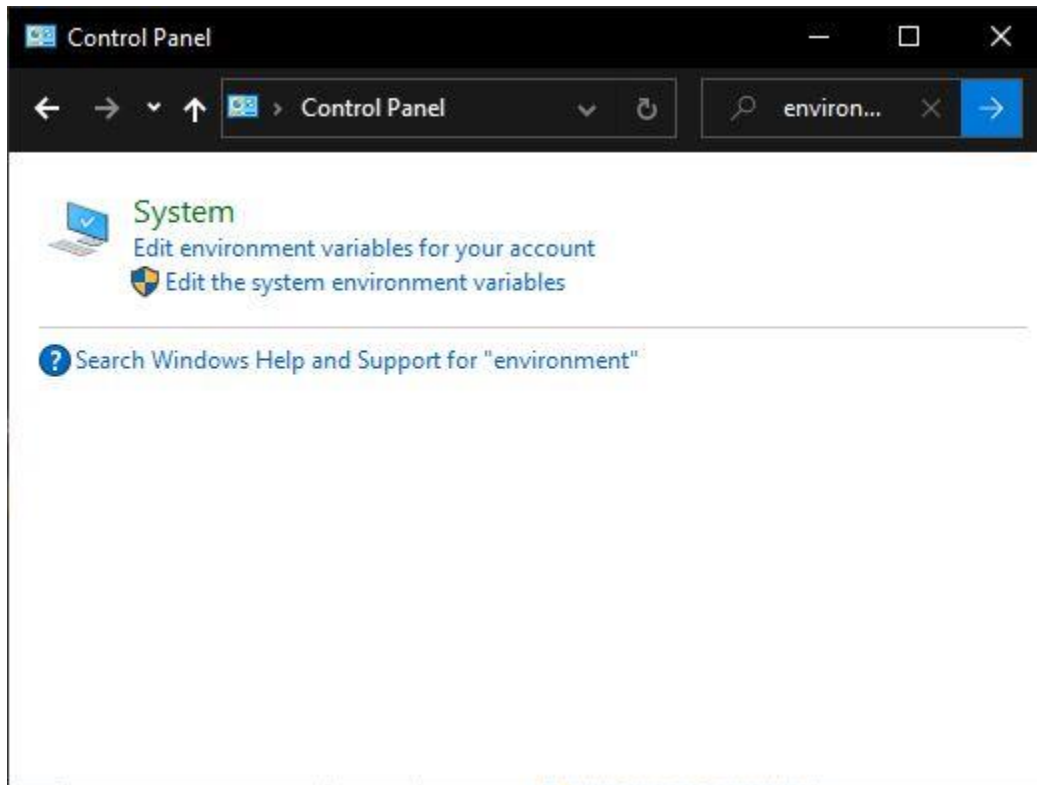Choose the start menu folder for the Tesseract OCR for Windows shortcuts.

After we click install, Tesseract OCR for Windows will begin installing. Our next step is to add the installation path to our machine's environment variables.

Step 3: Add Installation Path to Environment Variables

Control Panel

To add the installation location to our environment variables, go to the Start menu and search for 'environment variables. You should see a result to edit the system environment variables. If you don't, you can always use the following steps: Start menu > Control Panel > Edit the system environment variables.
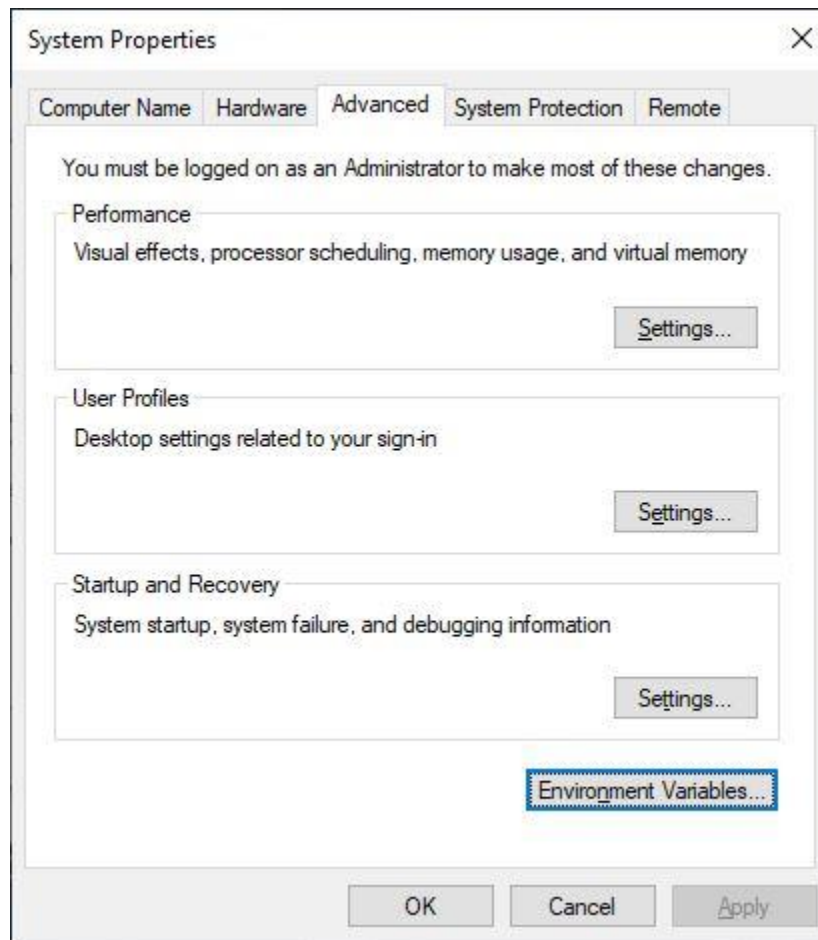
Searching for 'environment variables

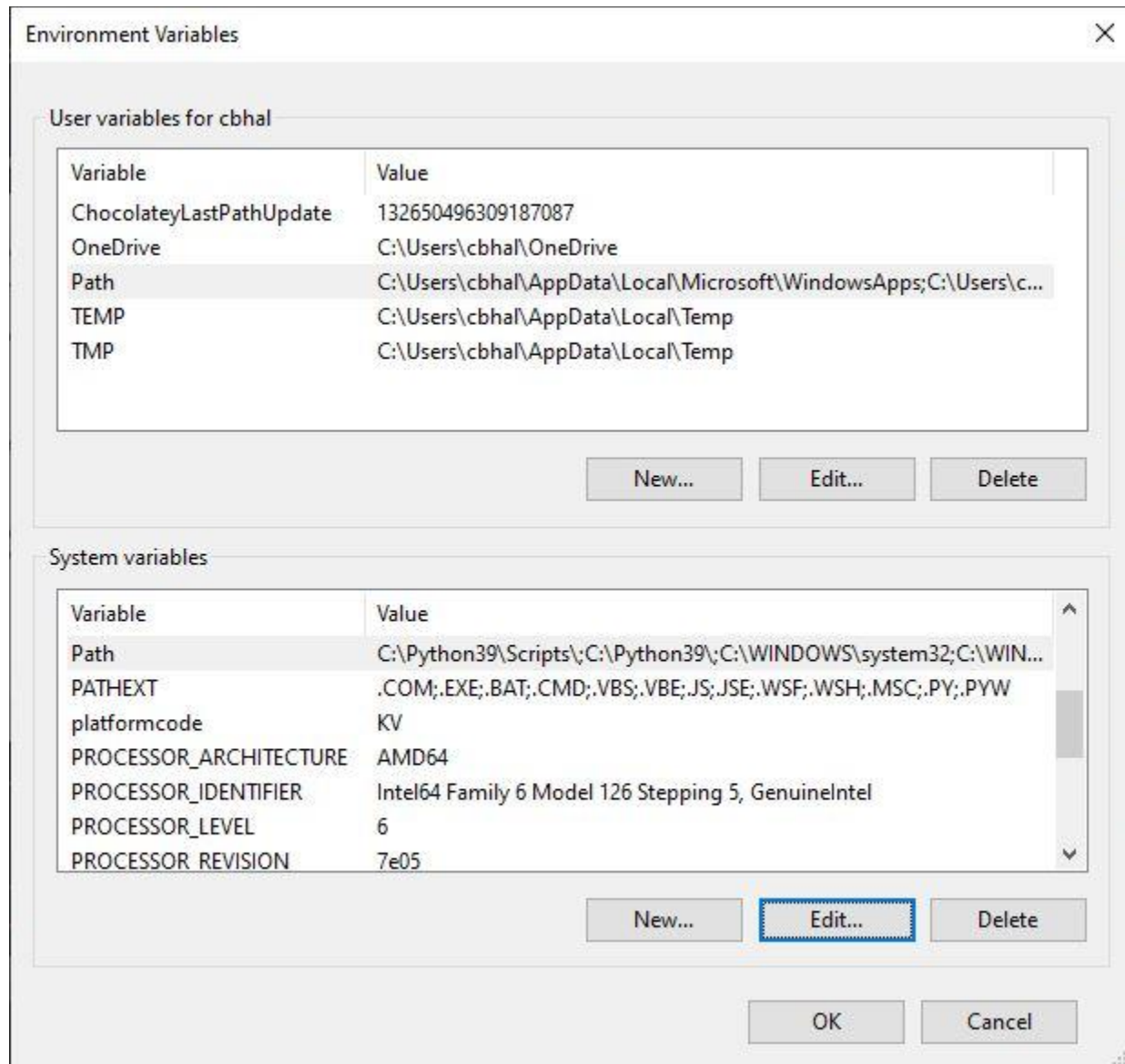# How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

System Properties

When presented with the 'System Properties dialog box, we'll want to make sure the 'Advanced' tab is clicked, then click the 'Environment Variables…' button towards the bottom right of the screen.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps
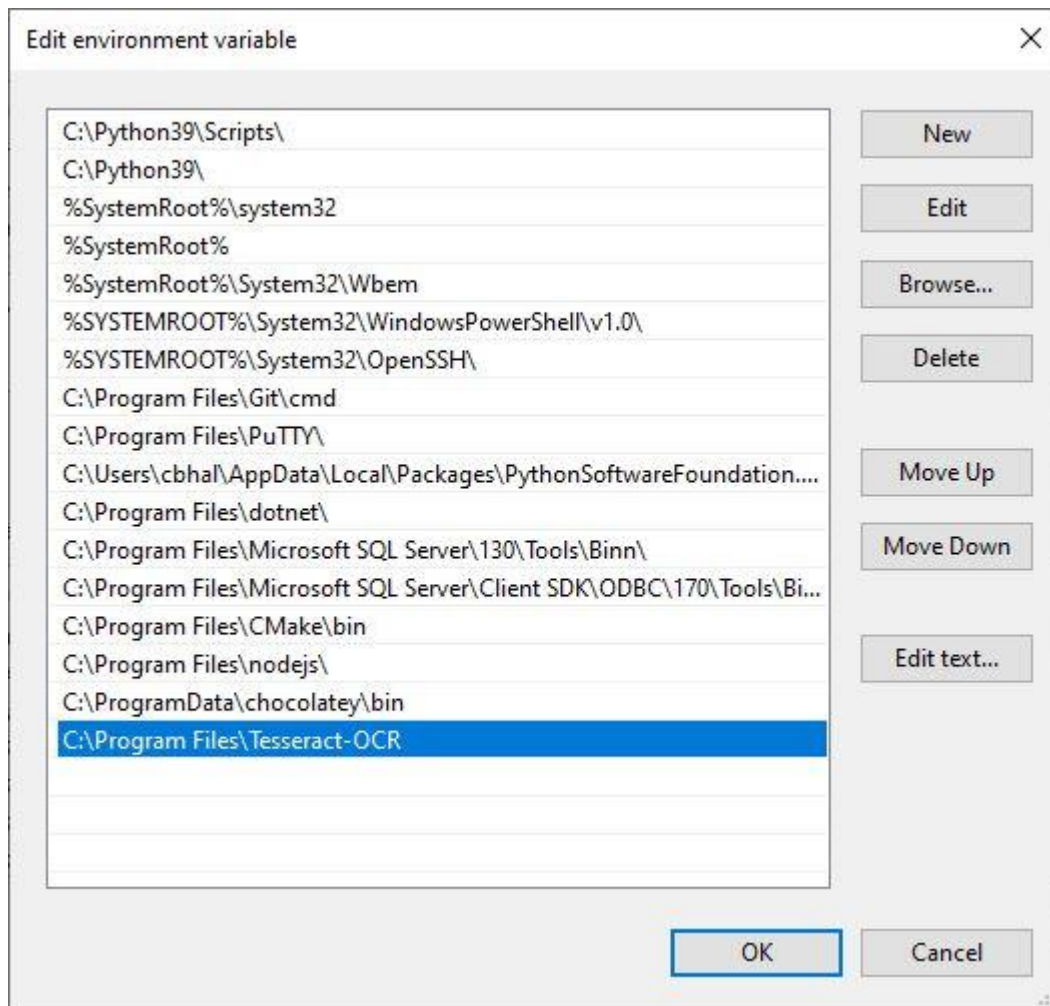
Environment Variables

Under system variables, we will click the 'Edit…' button.

When presented with the Environment Variable window, click the 'New' button, and paste it into your Tesseract OCR installation path that we copied earlier in Step 2. Once you've done this, click the 'OK' button.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Add Tesseract OCR for Windows Installation Directory to Environment Variables
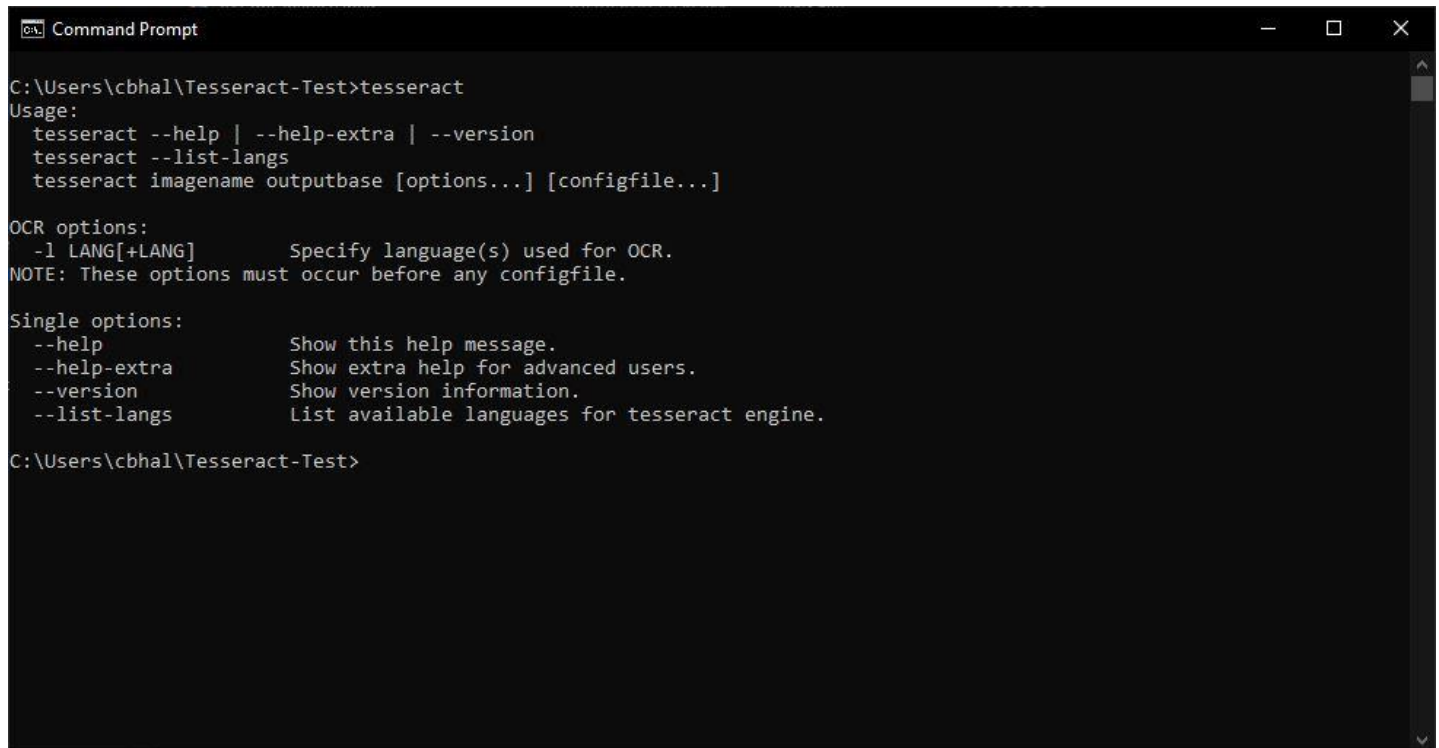

That's it! Now that we've run the .exe installer and added the Tesseract OCR for Windows install location to our environment variables, we can test that our installation is working by running Tesseract on a test image.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Step 4: Run Tesseract OCR for Windows on a Test Image

Check for Successful Installation

To test that Tesseract OCR for Windows was installed successfully, open the command prompt on your machine, then run the tesseract command. You should see an output with a quick explanation of Tesseract's usage options.


```
C:\Users\cbhal\Tesseract-Test>tesseract
Usage:
  tesseract --help | --help-extra | --version
  tesseract --list-langs
  tesseract imagename outputbase [options...] [configfile...]

OCR options:
  -l LANG[+LANG]        Specify language(s) used for OCR.
NOTE: These options must occur before any configfile.

Single options:
  --help                Show this help message.
  --help-extra          Show extra help for advanced users.
  --version             Show version information.
  --list-langs          List available languages for tesseract engine.

C:\Users\cbhal\Tesseract-Test>
```

Checking successful installation of Tesseract OCR for Windows

Congratulations! You've successfully installed Tesseract OCR for Windows on your machine. This is where the real fun begins.

Extract Text from the document using Tesseract OCR for Windows

Now that we know Tesseract was installed successfully, let's run Tesseract's OCR on a test document to check for accuracy. Download the sample invoice image here, and move the image into the same directory your command prompt is in. We can then run Tesseract on the image with the following command:

tesseract 200-dpi-invoice.jpeg 200-dpi-invoice

In this command, we're running Tesseract, supplying the sample image 200-dpi-invoice.jpeg to Tesseract, and telling Tesseract to put the extracted text in a new file called 200-dpi-invoice.txt.

Now if we look inside our current directory, we'll see that a new .txt file was generated which includes the contents of our image. If we compare the text in the sample image to the text in our newly generated .txt file, we can see that Tesseract's text extraction was incredibly accurate with only a few visible errors.

How to Install and Run Tesseract OCR for Windows in 4 Easy Steps

Accurately extracted text from Tesseract OCR for Windows.

Conclusion

In this tutorial, we:

Downloaded the .exe Tesseract OCR for Windows installer from UB Mannheim.

Ran the installer with our configurations.

Added the installation location to our environment variables

Tested Tesseract OCR for Windows by running Tesseract's OCR on a sample image in our command line.

I hope you enjoyed this tutorial and found it helpful.