Installing Tesseract OCR

1. The first step is to install the Tesseract 'engine' and language training files from Git Hub.

https://github.com/tesseract-ocr/tesseract/wiki

2. Scroll down to choose the instructions for the operating system your computer is running, e.g. 'Linux', 'macOS', 'Windows'. E.g. for installation on Windows open the 'Tesseract at UB Mannheim' page.

Tesseract at UB Mannheim

The Mannheim University Library (UB Mannheim) uses Tesseract to perform OCR of historical German newspapers (Allgemeine Preußische Staatszeitung, Deutscher Reichsanzeiger). The latest results with OCR from more than 360,000 scans are available online.

Normally we run Tesseract on Debian GNU Linux, but there was also the need for a Windows version. That's why we have built a Tesseract installer for Windows.

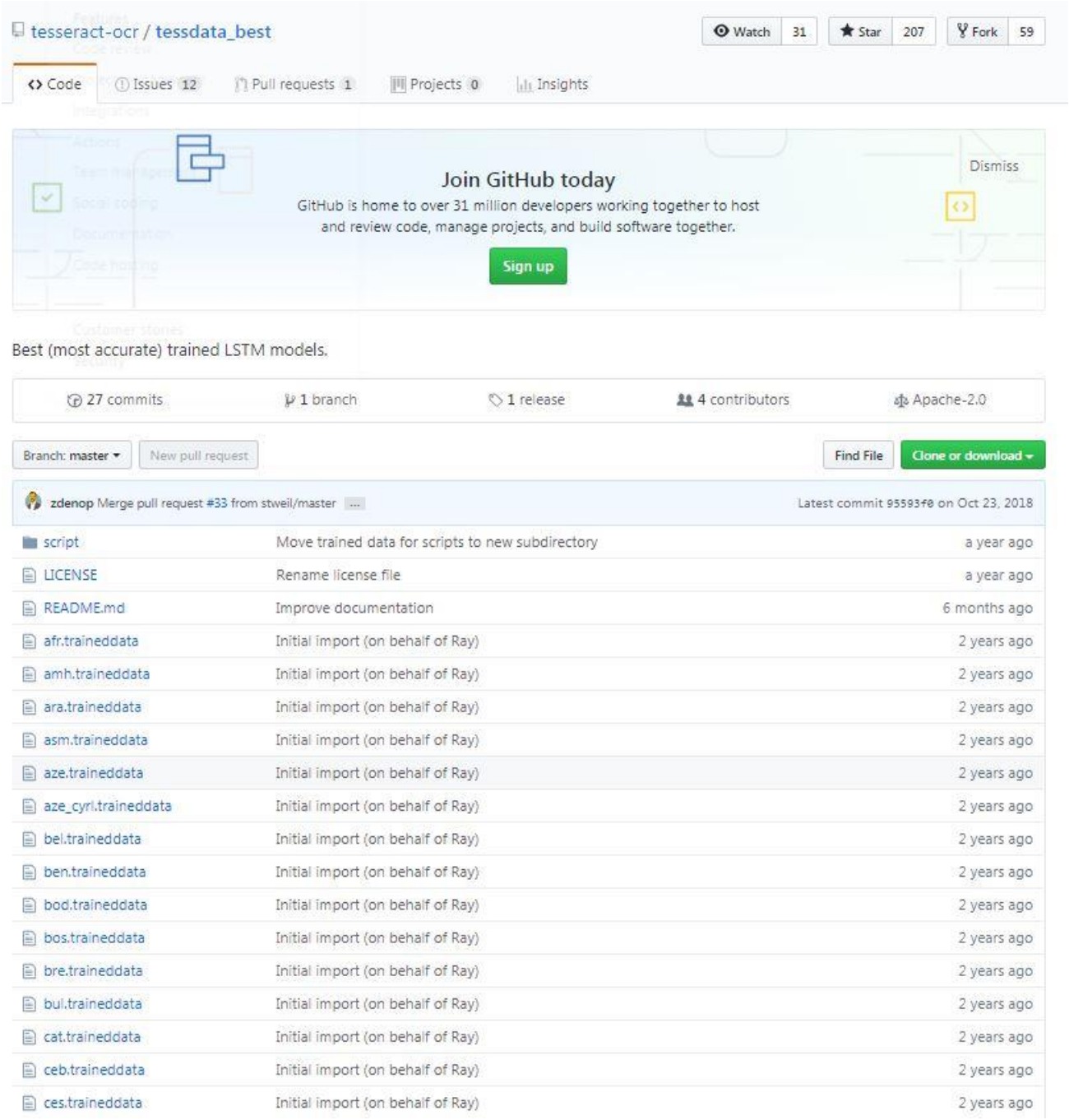
The latest installers can be downloaded here:

- tesseract-ocr-w32-setup-v4.1.0.20190314 (rc1) (32 bit) and
- tesseract-ocr-w64-setup-v4.1.0.20190314 (rc1) (64 bit) resp.
- tesseract-ocr-w64-setup-v4.1.0-bibtag19.exe a special branded #bibtag19 edition.

3. Scroll down and click the correct link for your computer depending on whether it is 32 or 64 bit. This will download the Tesseract engine and will take up about 40MB of storage space on your computer.

4. As well as the engine, you will need to install the source code. Go to https://github.com/tesseract-ocr/tesseract/releases and download the .zip file.

5. Next, go to https://github.com/tesseract-ocr/tessdata_best and select the language file(s) you need if you are working with non-English language material (see image below). For example, if the document or page you want to OCR is written in Bengali, download **'ben.traineddata'**



6. You may need to ask someone at your institution with administrator privileges to install the downloaded Tesseract application and other files you have just downloaded.

7. Once installed, the training files will be on your C drive, likely in 'C:\Program Files (x86)\Tesseract-OCR'. The folder will be called 'Tesseract-Master'. You will need to unpack the files using a programme like 7-zip.

8. Once you have done that, move the **ben.traineddata** file into the **tessdata** folder.

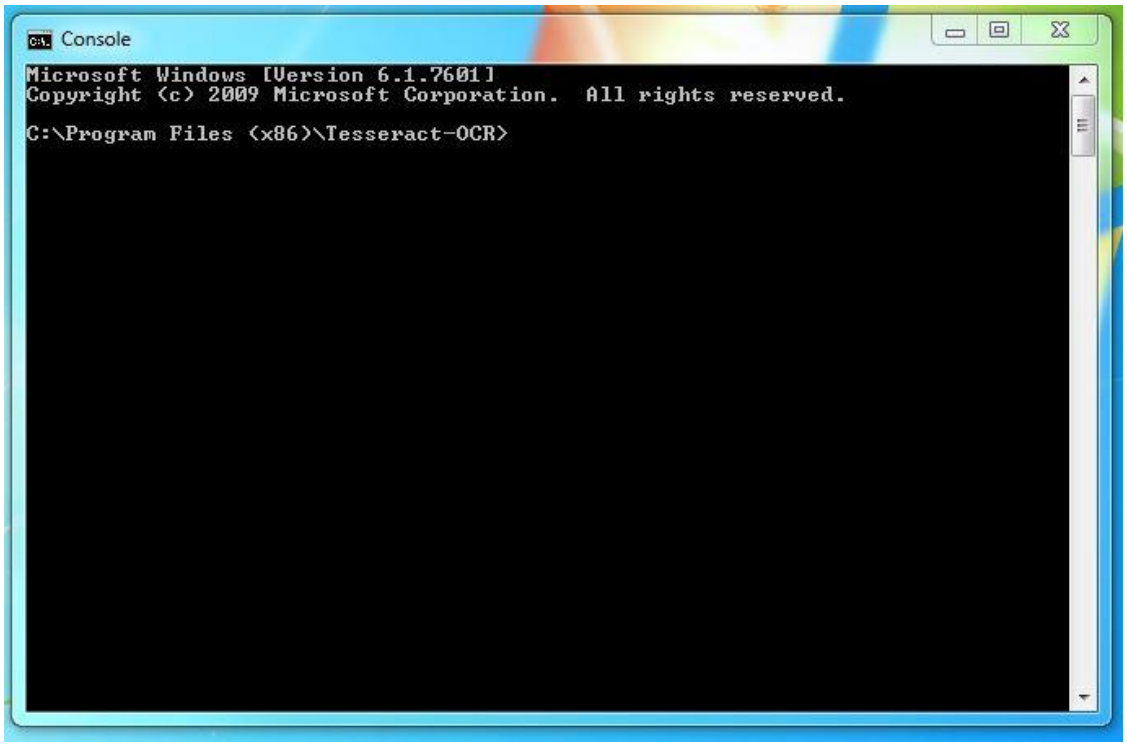
9. Move the images (TIFF, JPEG, PNG) you want to OCR into the main **tesseract-4.00.00alpha** folder.

You're now ready to OCR your documents! Scroll down to the next page to learn how to use Tesseract for OCR.

Using Tesseract Command Line for OCR of Bangla

1. Open the command prompt ‘Console’ which should be displayed on your desktop

This is where you will send write commands to OCR the images.



2. In the command prompt the folder path will show **C:\Program Files (x86)\Tesseract-OCR.** You will need to change this to point to the folder where the folder of images is you want to work with is saved. For my computer I pointed to:
C:\Program Files (x86)\Tesseract-OCR>cd “C:\Users\tderrick\Desktop\Tesseract-OCR”

Hit enter. This will give you the new source directory.

3. The next step is to write the command to OCR your desired image. Because you performing OCR on a language other than English you need to specify the language you are working with. The command is:

>tesseract filename.tif out –l ben

which makes the whole command…

C:\Users\tderrick\Desktop\Tesseract-OCR>tesseract nameoffile.tif out –l ben
(note: the character after '–' is a lower case 'L' rather than upper case I).

4. Great! You have just turned an image into OCR text. Check your folder of images. You should see both your original image file and a **txt** file (the OCR output). Open both to compare how accurate the .txt file is. Open the .txt with Notepad or Microsoft Word.

5. Next, try applying OCR to the whole folder of images. The command for a folder of .tif images is:

>for %i in (*.tif) do tesseract %i %i –l ben

The process is quite slow so be prepared to wait a few minutes if you are converting even just a few files into .txt files.