{

"nbformat": 4,

"nbformat\_minor": 0,

"metadata": {

"colab": {

"name": "OCR-Tesseract-Python.ipynb",

"version": "0.3.2",

"provenance": [],

"collapsed\_sections": [

"0GHxCcNS3EhN",

"qzkn3o9H3QIc"

]

},

"kernelspec": {

"name": "python3",

"display\_name": "Python 3"

},

"accelerator": "GPU"

},

"cells": [

{

"metadata": {

"id": "Eu9Yihr34MUj",

"colab\_type": "text"

},

"cell\_type": "markdown",

"source": [

"#Introduction\n",

"\*\*The aim of this NoteBook is to be able to recognise text from an image file using the Tesseract Library in the Python Programming Language.\*\*\n",

"\n",

"\*\*Tesseract is an Open Source library for Optical Character recognition(OCR). We will be using PyTesseract to print the recognized text given an input image of any of the following formats\n",

": jpeg, png, gif, bmp, tiff, and others. \*\*\n"

]

},

{

"metadata": {

"id": "4-yKpihT25p8",

"colab\_type": "text"

},

"cell\_type": "markdown",

"source": [

"# Installation"

]

},

{

"metadata": {

"id": "Iup9C5Lon-g0",

"colab\_type": "code",

"outputId": "f1b46343-59be-4bc8-f762-2df08b207ac8",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 493

}

},

"cell\_type": "code",

"source": [

"!sudo add-apt-repository ppa:alex-p/tesseract-ocr"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

" The Tesseract OCR engine was one of the top 3 engines in the 1995\n",

" UNLV Accuracy test. Between 1995 and 2006 it had little work done on\n",

" it, but since then it has been improved extensively by Google and is\n",

" probably one of the most accurate open source OCR engines\n",

" available. It can read a wide variety of image formats and convert\n",

" them to text in over 40 languages. This package includes the command\n",

" line tool.\n",

" More info: https://launchpad.net/~alex-p/+archive/ubuntu/tesseract-ocr\n",

"Press [ENTER] to continue or Ctrl-c to cancel adding it.\n",

"\n",

"Ign:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1710/x86\_64 InRelease\n",

"Get:2 http://security.ubuntu.com/ubuntu bionic-security InRelease [83.2 kB]\n",

"Hit:3 http://archive.ubuntu.com/ubuntu bionic InRelease\n",

"Get:4 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic InRelease [15.4 kB]\n",

"Ign:5 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1604/x86\_64 InRelease\n",

"Hit:6 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1710/x86\_64 Release\n",

"Hit:7 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1604/x86\_64 Release\n",

"Get:8 http://archive.ubuntu.com/ubuntu bionic-updates InRelease [88.7 kB]\n",

"Get:9 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease [21.3 kB]\n",

"Get:11 http://archive.ubuntu.com/ubuntu bionic-backports InRelease [74.6 kB]\n",

"Get:13 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 Packages [26.4 kB]\n",

"Get:14 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [300 kB]\n",

"Get:15 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic/main amd64 Packages [27.2 kB]\n",

"Get:16 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [618 kB]\n",

"Get:17 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [135 kB]\n",

"Get:18 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [900 kB]\n",

"Fetched 2,290 kB in 2s (1,341 kB/s)\n",

"Reading package lists... Done\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "RxADPcEA39z-",

"colab\_type": "code",

"outputId": "d9cc72fc-93f9-43aa-9411-ba5184fa5fd1",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 204

}

},

"cell\_type": "code",

"source": [

"!sudo apt-get update"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"\r0% [Working]\r \rIgn:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1710/x86\_64 InRelease\n",

"\r0% [Waiting for headers] [Waiting for headers] [Waiting for headers] [Waiting f\r \rIgn:2 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1604/x86\_64 InRelease\n",

"\r \r0% [Waiting for headers] [Waiting for headers] [Waiting for headers]\r \rHit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1710/x86\_64 Release\n",

"\r0% [Waiting for headers] [Waiting for headers] [Waiting for headers]\r \rHit:4 http://archive.ubuntu.com/ubuntu bionic InRelease\n",

"\r \rHit:5 http://security.ubuntu.com/ubuntu bionic-security InRelease\n",

"\r \rHit:6 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic InRelease\n",

"\r0% [Waiting for headers] [Connecting to ppa.launchpad.net (91.189.95.83)]\r0% [Release.gpg gpgv 564 B] [Waiting for headers] [Connecting to ppa.launchpad.\r \rHit:7 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1604/x86\_64 Release\n",

"\r0% [Release.gpg gpgv 564 B] [Waiting for headers] [Connecting to ppa.launchpad.\r \r0% [Waiting for headers] [Waiting for headers]\r \rHit:9 http://archive.ubuntu.com/ubuntu bionic-updates InRelease\n",

"\r0% [Waiting for headers] [Waiting for headers]\r0% [4 InRelease gpgv 242 kB] [Waiting for headers] [Waiting for headers]\r \rHit:10 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease\n",

"Hit:11 http://archive.ubuntu.com/ubuntu bionic-backports InRelease\n",

"Reading package lists... Done\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "bOQf0ryh3\_8d",

"colab\_type": "code",

"outputId": "f35e64cf-dfa1-4c39-c7a4-1696e118b0bf",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 632

}

},

"cell\_type": "code",

"source": [

"!sudo apt install tesseract-ocr"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"Reading package lists... Done\n",

"Building dependency tree \n",

"Reading state information... Done\n",

"The following additional packages will be installed:\n",

" tesseract-ocr-eng tesseract-ocr-osd\n",

"The following NEW packages will be installed:\n",

" tesseract-ocr tesseract-ocr-eng tesseract-ocr-osd\n",

"0 upgraded, 3 newly installed, 0 to remove and 14 not upgraded.\n",

"Need to get 4,843 kB of archives.\n",

"After this operation, 15.8 MB of additional disk space will be used.\n",

"Get:1 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 tesseract-ocr-eng all 1:4.00~git30-7274cfa-1ppa1~bionic1 [1,592 kB]\n",

"Get:2 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 tesseract-ocr-osd all 1:4.00~git30-7274cfa-1ppa1~bionic1 [2,991 kB]\n",

"Get:3 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 tesseract-ocr amd64 4.0.0+git3360-e3a39c35-1ppa1~bionic1 [259 kB]\n",

"Fetched 4,843 kB in 8s (588 kB/s)\n",

"debconf: unable to initialize frontend: Dialog\n",

"debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76, <> line 3.)\n",

"debconf: falling back to frontend: Readline\n",

"debconf: unable to initialize frontend: Readline\n",

"debconf: (This frontend requires a controlling tty.)\n",

"debconf: falling back to frontend: Teletype\n",

"dpkg-preconfigure: unable to re-open stdin: \n",

"Selecting previously unselected package tesseract-ocr-eng.\n",

"(Reading database ... 110842 files and directories currently installed.)\n",

"Preparing to unpack .../tesseract-ocr-eng\_1%3a4.00~git30-7274cfa-1ppa1~bionic1\_all.deb ...\n",

"Unpacking tesseract-ocr-eng (1:4.00~git30-7274cfa-1ppa1~bionic1) ...\n",

"Selecting previously unselected package tesseract-ocr-osd.\n",

"Preparing to unpack .../tesseract-ocr-osd\_1%3a4.00~git30-7274cfa-1ppa1~bionic1\_all.deb ...\n",

"Unpacking tesseract-ocr-osd (1:4.00~git30-7274cfa-1ppa1~bionic1) ...\n",

"Selecting previously unselected package tesseract-ocr.\n",

"Preparing to unpack .../tesseract-ocr\_4.0.0+git3360-e3a39c35-1ppa1~bionic1\_amd64.deb ...\n",

"Unpacking tesseract-ocr (4.0.0+git3360-e3a39c35-1ppa1~bionic1) ...\n",

"Setting up tesseract-ocr-osd (1:4.00~git30-7274cfa-1ppa1~bionic1) ...\n",

"Setting up tesseract-ocr-eng (1:4.00~git30-7274cfa-1ppa1~bionic1) ...\n",

"Processing triggers for man-db (2.8.3-2ubuntu0.1) ...\n",

"Setting up tesseract-ocr (4.0.0+git3360-e3a39c35-1ppa1~bionic1) ...\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "qUkfdeTq4Bta",

"colab\_type": "code",

"outputId": "a81a632c-6b48-4112-fa58-21ec0bf1e1f1",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 734

}

},

"cell\_type": "code",

"source": [

"!sudo apt install libtesseract-dev"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"Reading package lists... Done\n",

"Building dependency tree \n",

"Reading state information... Done\n",

"The following additional packages will be installed:\n",

" liblept5 libleptonica-dev libtesseract4\n",

"The following NEW packages will be installed:\n",

" libleptonica-dev libtesseract-dev\n",

"The following packages will be upgraded:\n",

" liblept5 libtesseract4\n",

"2 upgraded, 2 newly installed, 0 to remove and 12 not upgraded.\n",

"Need to get 4,916 kB of archives.\n",

"After this operation, 13.7 MB of additional disk space will be used.\n",

"Get:1 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 liblept5 amd64 1.76.0-1+nmu1ppa1~bionic1 [933 kB]\n",

"Get:2 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 libleptonica-dev amd64 1.76.0-1+nmu1ppa1~bionic1 [1,321 kB]\n",

"Get:3 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 libtesseract4 amd64 4.0.0+git3360-e3a39c35-1ppa1~bionic1 [1,209 kB]\n",

"Get:4 http://ppa.launchpad.net/alex-p/tesseract-ocr/ubuntu bionic/main amd64 libtesseract-dev amd64 4.0.0+git3360-e3a39c35-1ppa1~bionic1 [1,453 kB]\n",

"Fetched 4,916 kB in 8s (611 kB/s)\n",

"debconf: unable to initialize frontend: Dialog\n",

"debconf: (No usable dialog-like program is installed, so the dialog based frontend cannot be used. at /usr/share/perl5/Debconf/FrontEnd/Dialog.pm line 76, <> line 4.)\n",

"debconf: falling back to frontend: Readline\n",

"debconf: unable to initialize frontend: Readline\n",

"debconf: (This frontend requires a controlling tty.)\n",

"debconf: falling back to frontend: Teletype\n",

"dpkg-preconfigure: unable to re-open stdin: \n",

"(Reading database ... 110889 files and directories currently installed.)\n",

"Preparing to unpack .../liblept5\_1.76.0-1+nmu1ppa1~bionic1\_amd64.deb ...\n",

"Unpacking liblept5 (1.76.0-1+nmu1ppa1~bionic1) over (1.75.3-3) ...\n",

"Selecting previously unselected package libleptonica-dev.\n",

"Preparing to unpack .../libleptonica-dev\_1.76.0-1+nmu1ppa1~bionic1\_amd64.deb ...\n",

"Unpacking libleptonica-dev (1.76.0-1+nmu1ppa1~bionic1) ...\n",

"Preparing to unpack .../libtesseract4\_4.0.0+git3360-e3a39c35-1ppa1~bionic1\_amd64.deb ...\n",

"Unpacking libtesseract4:amd64 (4.0.0+git3360-e3a39c35-1ppa1~bionic1) over (4.00~git2288-10f4998a-2) ...\n",

"Selecting previously unselected package libtesseract-dev:amd64.\n",

"Preparing to unpack .../libtesseract-dev\_4.0.0+git3360-e3a39c35-1ppa1~bionic1\_amd64.deb ...\n",

"Unpacking libtesseract-dev:amd64 (4.0.0+git3360-e3a39c35-1ppa1~bionic1) ...\n",

"Setting up liblept5 (1.76.0-1+nmu1ppa1~bionic1) ...\n",

"Processing triggers for libc-bin (2.27-3ubuntu1) ...\n",

"Setting up libleptonica-dev (1.76.0-1+nmu1ppa1~bionic1) ...\n",

"Setting up libtesseract4:amd64 (4.0.0+git3360-e3a39c35-1ppa1~bionic1) ...\n",

"Setting up libtesseract-dev:amd64 (4.0.0+git3360-e3a39c35-1ppa1~bionic1) ...\n",

"Processing triggers for libc-bin (2.27-3ubuntu1) ...\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "8gSJiO8E4DT6",

"colab\_type": "code",

"outputId": "ebab2cf1-528d-45ff-ef0f-e7e90d3bc28c",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 204

}

},

"cell\_type": "code",

"source": [

"!sudo pip install pytesseract"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"Collecting pytesseract\n",

"\u001b[?25l Downloading https://files.pythonhosted.org/packages/71/5a/d7600cad26276d991feecb27f3627ae2d0ee89aa1e3065fa4f9f1f2defbc/pytesseract-0.2.6.tar.gz (169kB)\n",

"\u001b[K 100% |████████████████████████████████| 174kB 8.1MB/s \n",

"\u001b[?25hRequirement already satisfied: Pillow in /usr/local/lib/python3.6/dist-packages (from pytesseract) (4.0.0)\n",

"Requirement already satisfied: olefile in /usr/local/lib/python3.6/dist-packages (from Pillow->pytesseract) (0.46)\n",

"Building wheels for collected packages: pytesseract\n",

" Running setup.py bdist\_wheel for pytesseract ... \u001b[?25l-\b \bdone\n",

"\u001b[?25h Stored in directory: /root/.cache/pip/wheels/d5/90/56/ab7b652592da86821293f7cadc1c554aa376a0d57ce414d0a0\n",

"Successfully built pytesseract\n",

"Installing collected packages: pytesseract\n",

"Successfully installed pytesseract-0.2.6\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "4DmjENFFpd9w",

"colab\_type": "code",

"outputId": "3222ecd7-6a6b-461a-ff74-27c37ecd22b4",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 119

}

},

"cell\_type": "code",

"source": [

"#Checking the installation.\n",

"!tesseract --version"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"tesseract 4.0.0-115-ge3a3\n",

" leptonica-1.76.0\n",

" libgif 5.1.4 : libjpeg 8d (libjpeg-turbo 1.5.2) : libpng 1.6.34 : libtiff 4.0.9 : zlib 1.2.11 : libwebp 0.6.1 : libopenjp2 2.3.0\n",

" Found AVX2\n",

" Found AVX\n",

" Found SSE\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "0GHxCcNS3EhN",

"colab\_type": "text"

},

"cell\_type": "markdown",

"source": [

"# Image Downloader"

]

},

{

"metadata": {

"id": "OUU2ixCpva\_L",

"colab\_type": "code",

"outputId": "cedc9c81-940e-44f9-a407-b906bc0ac677",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 51

}

},

"cell\_type": "code",

"source": [

"#Image Downloader(From previous video.)\n",

"import requests\n",

"\n",

"print ('Starting to Download!')\n",

"\n",

"url = 'http://zone1-af2a.kxcdn.com/wp-content/uploads/life-quotes-that-will-change-you-forever-wisdom-quotes.jpg'\n",

"r = requests.get(url)\n",

"\n",

"filename = '2.jpg'\n",

"\n",

"with open(filename, 'wb') as out\_file:\n",

" out\_file.write(r.content)\n",

"\n",

"print(\"Download complete!\")"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

"Starting to Download!\n",

"Download complete!\n"

],

"name": "stdout"

}

]

},

{

"metadata": {

"id": "dvyFYwR7zZ59",

"colab\_type": "text"

},

"cell\_type": "markdown",

"source": [

"#TESTING 5 Images\n",

"\n",

"\n",

"---\n",

"\n",

"\n",

"#IMAGE 1\n",

" \n",

" ![alt text](https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcT6z5vTpAMfE1xjxV7rrZ726lWoMARenCftjVxH4\_HPEykFfAiK)\n",

"\n",

"\n",

"---\n",

"\n",

"\n",

"#IMAGE 2\n",

"\n",

"![alt text](http://pun.me/pages/funny-quote-about-life.jpg)\n",

"\n",

"---\n",

"\n",

"#IMAGE 3\n",

"\n",

"![alt text](https://i.pinimg.com/originals/38/0b/87/380b87f345c48637cee8ff05d872fd60.png)\n",

"\n",

"---\n",

"\n",

"#IMAGE 4\n",

"![alt text](https://i.pinimg.com/originals/5a/bc/0f/5abc0feb32c88c9626afb7f8aedaea2f.jpg)\n",

"\n",

"---\n",

"\n",

"\n",

"#IMAGE 5\n",

"![alt text](http://quotesideas.com/wp-content/uploads/2015/03/nice-motivational-inspirational-quotes-thoughts-achieves-possible-great-best.jpg)\n",

"\n",

"---\n"

]

},

{

"metadata": {

"id": "qzkn3o9H3QIc",

"colab\_type": "text"

},

"cell\_type": "markdown",

"source": [

"# Final Code"

]

},

{

"metadata": {

"id": "jTrX8LtltFgX",

"colab\_type": "code",

"outputId": "f3021430-383c-4e04-fd02-7a058359c91a",

"colab": {

"base\_uri": "https://localhost:8080/",

"height": 102

}

},

"cell\_type": "code",

"source": [

"import cv2\n",

"import numpy as np\n",

"import pytesseract\n",

"from PIL import Image\n",

"\n",

"# Path of working folder on Disk\n",

"\n",

"def get\_string(img\_path):\n",

" # Read image with opencv\n",

" img = cv2.imread(img\_path)\n",

"\n",

" # Convert to gray\n",

" img = cv2.cvtColor(img, cv2.COLOR\_BGR2GRAY)\n",

"\n",

" # Apply dilation and erosion to remove some noise\n",

" kernel = np.ones((1, 1), np.uint8)\n",

" img = cv2.dilate(img, kernel, iterations=1)\n",

" img = cv2.erode(img, kernel, iterations=1)\n",

"\n",

" # Write image after removed noise\n",

" cv2.imwrite(\"removed\_noise.png\", img)\n",

"\n",

" # Apply threshold to get image with only black and white\n",

" #img = cv2.adaptiveThreshold(img, 255, cv2.ADAPTIVE\_THRESH\_GAUSSIAN\_C, cv2.THRESH\_BINARY, 31, 2)\n",

"\n",

" # Write the image after apply opencv to do some ...\n",

" cv2.imwrite(img\_path, img)\n",

"\n",

" # Recognize text with tesseract for python\n",

" result = pytesseract.image\_to\_string(Image.open(img\_path))\n",

"\n",

" # Remove template file\n",

" #os.remove(temp)\n",

"\n",

" return result\n",

"\n",

"\n",

"print ('--- Start recognize text from image ---')\n",

"print (get\_string(filename))\n",

"\n",

"print (\"------ Done -------\")"

],

"execution\_count": 0,

"outputs": [

{

"output\_type": "stream",

"text": [

" "

],

"name": "stdout"

}

]

}

]

}