## Captcha Recognition using pytesser & tesseract

Reference: <http://www.debasish.in/2012/01/bypass-captcha-using-python-and.html>

\*\* *-- I have done all operation in “download” Directory.*

**1. Install Pillow**

Pillow is a fork of PIL (Python Image Processing Library) that has support for opening, manipulating, and saving many different image file formats.

reference : <http://pillow.readthedocs.org/en/latest/installation.html>

**A.** sudo apt-get install python-dev python-setuptools

**B.** sudo apt-get install libtiff5-dev libjpeg8-dev zlib1g-dev \

libfreetype6-dev liblcms2-dev libwebp-dev tcl8.6-dev tk8.6-dev python-tk

**C.** sudo pip install pillow

**2. Install tesseract :**

Tesseract is probably the most accurate open source OCR engine available. Combined with the [Leptonica Image Processing Library](http://leptonica.com/) it can read a wide variety of image formats and convert them to text in over 60 languages.

reference : <https://code.google.com/p/tesseract-ocr/wiki/Compiling>

**A.** sudo apt-get install autoconf automake libtool

sudo apt-get install libpng12-dev

sudo apt-get install libjpeg62-dev

sudo apt-get install libtiff4-dev

sudo apt-get install zlib1g-dev

sudo apt-get install libicu-dev

sudo apt-get install libpango1.0-dev

sudo apt-get install libcairo2-dev

sudo apt-get install libleptonica-dev

**B.** wget <https://tesseract-ocr.googlecode.com/files/tesseract-ocr-3.02.02.tar.gz>

**C.** tar -xzvf tesseract-ocr-3.02.02.tar.gz

**D.** cd tesseract-ocr-3.02.02

./autogen.sh

./configure

make

sudo make install

sudo ldconfig

**E.** wget <http://tesseract-ocr.googlecode.com/files/tesseract-ocr-3.01.eng.tar.gz>

tar -xzvf tesseract-ocr-3.01.eng.tar.gz

sudo cp -R tesseract-ocr/tessdata /usr/local/share

export TESSDATA\_PREFIX=/home/ajay/Downloads/tesseract-ocr/

**3. Use PyTesser :**

PyTesser is an Optical Character Recognition module for Python. It takes as input an image or image file and outputs a string.

PyTesser uses the [Tesseract OCR engine](http://code.google.com/p/tesseract-ocr/), converting images to an accepted format and calling the Tesseract executable as an external script.

PyTesser can read all image types supported by the Python Imaging Library(Pillow), including jpeg, png, gif, bmp, tiff, and others, whereas tesseract-ocr by default only supports tiff and bmp.

Reference : <https://code.google.com/p/pytesser/>

**A.** wget <https://pytesser.googlecode.com/files/pytesser_v0.0.1.zip>

mkdir pytesser

unzip pytesser\_v0.0.1.zip -d pytesser  
  
**B.** Copy the crack.py file in pytesser folder & run

mkdir test

sudo python crack.py

**Library Version with which I have tested :**

ajay-HP-ProBook-4445s:pytesser# python -V

Python 2.7.6

ajay-HP-ProBook-4445s:pytesser# pip freeze | grep -E '(Pillow|PIL)'

Pillow==2.3.0

**Test accuracy :**

Tesseract sometimes can not understand 0 & 8 correctly.

**Future scope :**

1. Extend python script to read all the order numbers that need to be tracked.

Fetch captcha.gif from Indian post website.

Recognise the captcha number.

Make a call to website with captcha text & order number. Get the status of the order.

Update the status of order in DB & alert the CSR if criteria for order status is not met.

2. Train tesseract.

3. Understand OCR in detail to remove small error rate in recognising 0 & 8.