**DOCUMENTATION**

* **EXTRACTING DATA FROM NEWS ON AIR PDF**

**Downloading and installation tesseract on ubuntu os---**

**For 18.04**

sudo apt install tesseract-ocr

sudo apt install libtesseract-dev

**For 16.04**

**These commands should be typed in the terminal**

**Install requirements**--

sudo apt-get install -y build-essential &&\

sudo apt-get install -y python3-dev &&\

sudo apt-get install -y python3 &&\

sudo apt-get install -y python3-pip &&\

sudo apt-get install -y python3-setuptools &&\

sudo pip install -y cython

**Upgrade g++5**

sudo add-apt-repository ppa:ubuntu-toolchain-r/test &&\

sudo apt-get update &&\

sudo apt-get upgrade &&\

sudo update-alternatives --install /usr/bin/gcc gcc /usr/bin/gcc-5 1

**Install dependency package**

sudo apt-get install -y autoconf automake libtool &&\

sudo apt-get install -y autoconf-archive &&\

sudo apt-get install -y pkg-config &&\

sudo apt-get install -y libpng12-dev &&\

sudo apt-get install -y libjpeg8-dev &&\

sudo apt-get install -y libtiff5-dev &&\

sudo apt-get install -y zlib1g-dev &&\

sudo apt-get install -y libicu-dev &&\

sudo apt-get install -y libpango1.0-dev &&\

sudo apt-get install -y libcairo2-dev

**Install Leptonica 1.74**

Download the source code from Leptonica

<http://www.leptonica.com/source/leptonica-1.76.0.tar.gz>

Open terminal in downloaded folder and type the commands below:

sudo tar xf leptonica-1.76.0.tar.gz &&\

cd leptonica-1.76.0 &&\

If you have extracted manually then open a terminal in the extracted folder named leptonica-1.76.0

sudo ./configure &&\

sudo make &&\

sudo make install

**Install Tesseract 4.0**

Checkout code from Github

<https://launchpad.net/~alex-p/+archive/ubuntu/tesseract-ocr>

Compiling it and install it.

Extract the file and open terminal inside the extracted folder and run the below commands

sudo sh autogen.sh &&\

./configure &&\

LDFLAGS="-L/usr/local/lib" CFLAGS="-I/usr/local/include" make &&\

sudo make install &&\

sudo make install -langs &&\

sudo ldconfig

Source-- <https://github.com/tesseract-ocr/tesseract/wiki>

<https://bingrao.github.io/blog/post/2017/07/16/Install-Tesseract-4.0-in-ubuntun-16.04.html>

Downloading tesserect language files:

From the link download san-deva-float.traineddata san-siddhanta-float.traineddata

<https://github.com/Shreeshrii/tessdata_shreetest>

For hindi fonts

<https://github.com/tesseract-ocr/tessdata/raw/4.00/hin.traineddata>

For sanskrit fonts

<https://github.com/tesseract-ocr/tessdata/raw/4.00/san.traineddata>

Go inside computer

/usr/share/tesseract-ocr/4.00/tessdata and open a terminal

or open a terminal with ctrl+alt+t and type

cd /usr/share/tesseract-ocr/4.00/tessdata/(or /usr/local/share/tesseract-ocr/4.00/tessdata/)

And remove the default traineddata

Open terminal in downloaded folder and type the commands below:

sudo rm -r hin.traineddata

sudo rm -r san.traineddata

Open a terminal in the downloads location and type these commands to copy the traineddatas to the tessdata folder

sudo cp hin.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san-deva-float.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san-siddhanta-float.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

**Downloading and installation on centos7---**

sudo yum-config-manager --add-repo https://download.opensuse.org/repositories/home:/Alexander\_Pozdnyakov/CentOS\_7/

sudo yum update

sudo yum install tesseract

If you get public key error replace

sudo yum install tesseract

with

sudo yum install tesseract –nogpgcheck

Source--- <https://github.com/tesseract-ocr/tesseract/wiki>

Go to /usr/share/tesseract-ocr/4.00/tessdata and open a terminal

or open a terminal with ctrl+alt+t and type

cd /usr/share/tesseract-ocr/4.00/tessdata/

And remove the default traineddata

for this type this command on terminal opened in that place

sudo rm -r hin.traineddata

sudo rm -r san.traineddata

Open a terminal in the downloads location and type these commands to copy the traineddatas to the tessdata folder

sudo cp hin.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san-deva-float.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

sudo cp san-siddhanta-float.traineddata /usr/share/tesseract-ocr/4.00/tessdata/

**Convert to text(Same for both os)**

Now for save all the pdfs and the on.sh and off.sh inside a folder named test

now run command

chmod +x on.sh (making the script files executable)

chmod +x off.sh

now for converting them to text files(replace siba with your user name)

sudo ./on.sh /home/siba/Documents/test/ /home/siba/Documents/test/on/ hin+san+san-deva-float+san-siddhanta-float

sudo ./off.sh /home/siba/Documents/test/ /home/siba/Documents/test/off/ hin+san+san-deva-float+san-siddhanta-float

on.sh---File converted using alpha on option

off.sh---File converted using alpha off option

/home/siba/Documents/test/ ----location where the pdf files are stored

(/home/siba/Documents/test/on/)/(/home/siba/Documents/test/off/)--- location where the result text files will be

--- The language you are using for ocr you can replace with 1 lang like hin or san or add multiple by using + sign like hin+san (Here we have 4 languages for ocr which has proved to give good result).

---- put language.traineddata files into /*usr/*share/tesserect-ocr/4.00/tessdata and also in /usr/local/share/tessdata/ folder

----open file by typing ‘nano ~/.bashrc’ and and type

**export TESSDATA\_PREFIX='**/*usr/*share/tesserect-ocr/4.00/tessdata**'** inside .bashrc then run ‘source ~/.bashrc’