NAMSEL OCR

OPTICAL CHARACTER RECOGNITION

FOR

-བོད་ཡིག། TIBETAN -

LINUX

Installation

Pre Requirements:

*Linux machine

*Decent Internet Connection

*Basic Knowledge of Terminal

*Good Quality Pecha (300 dpi min)

Installation Requirements:

1. apt-get update && apt-upgrade

-Takes linux machine to latest updates

2. Download namsel OCR from github or though terminal

git clone https://github.com/thubtenrigzin/namsel-ocr.git

3. apt-get install scantailor

-need for namsel OCR so install separately

4. git clone https://github.com/opencv/opencv.git

-need for Namsel OCR so install separately

Resources:

* http://digitaltibetan.org/index.php/Tibetan_OCR
* https://github.com/thubtenrigzin/namsel-ocr
* Linux repositories apt-get

5. git clone https://github.com/opencv/opencv_contrib.git

-need for Namsel OCR so install separately

6. apt-get install pagerecognizer

-need for Namsel OCR so install separately

7. bash ubuntu_install.sh

-this will update your file structure

8. apt-get install cmake

-need for Namsel OCR so install separately

9. apt-get install libgtk2.0-dev pig-config libavcodec-dev libavformat-dev libswscale-dev

-need for Namsel OCR so install separately

- Now you have everything you need -

How To Run

1. Go to namsel-ocr folder
2. copy your tiff image file in myfolder
3. python namsel.py preprocess ./myfolder

-This will process all the image

-Remove unwanted lines

-make higher contrast

-Analyse the Image

FLIES LOCATION:

* namsel-ocr /home/namsel-ocr
* myfolder /home/namsel-ocr/myfolder
* out /home/namsel-ocr/out
* ocr_output.txt /home/namsel-ocr/ocr_output.txt
* ocr_results /home/namsel-ocr/ocr_results

4. python namsel.py recognise-page —format-text ./ myfolder/out/“name_of_your_processed_file”

NOTE:

1. Scanned Image should be in Black & White

2. Scanned Image should be in resolution of higher than 300dpi

3.The scanned Image should be in TIFF format

PROBLEMS/BUGs:

* Results not saving in output file
* Only 60% accuracy (spelling)
* encoding and decoding problem

SOLUTION:

Edit the namsel.py

comment out form line no 326 - 331

edit the line 322 as below

with codecs.open(outfilename,’w’) as outfile:

Report by: Thupten Choephel (LTWA)

thupten104@gmail.com