Tesseract Steps

1. Clean the image
Refer: https://www.howtoforge.com/ocr_with_tesseract_on_ubuntu704

2. Check if tesseract recognizes the text correctly
$ convert 1440772689.5733.jpg tesseract eng.arial.exp0.tif
$ tesseract tesseract eng.arial.exp0.tif output
Note: It is advisable to have filenames in this format [lang].[fontname].exp[num]
e.g., if font_properties has aerial, then the filename of your image should be eng.aerial.exp0.tif

3. If it didn't, we need to train it

4. create a box file
$ tesseract tesseract eng.arial.exp0.tif tesseract eng.arial.exp0 batch.nochop makebox
Edit the box and correct the characters using jTessBoxEditor
$ java -Xms128m -Xmx1024m -jar jTessBoxEditor.jar
Box Editor --> Open (tif file) --> Edit boxes --> Save
Note: You need a .tif file with a .box file in the same folder to load the boxes
Tip: If some boxes are not getting recognized, you will need to resize and try.
http://stackoverflow.com/questions/9480013/image-processing-to-improve-tesseract-ocr-accuracy

5. Add some standard font properties
<fontname> <italic> <bold> <fixed> <serif> <fraktur>
$ vim font_properties
arial.exp 0 0 0 0 0 0

6. Feed the box back to tesseract
$ tesseract eng.arial.exp0.tif eng.arial.exp0.box nobatch box.train

7. Uncharset extraction
$ unicharset_extractor eng.arial.exp0.box

8. Create clustering data
$ shapeclustering -F font_properties -U unicharset eng.arial.exp0.box.tr

9. mftraining
$ mftraining -F font_properties -U unicharset -O eng.unicharset eng.arial.exp0.box.tr

10. cntraining
cntraining eng.arial.exp0.box.tr

11. combine tessdata
$ mv normproto eng.normproto; mv inttemp eng.inttemp; mv pffmtable eng.pffmtable; mv shapetable eng.shapetable
$ combine_tessdata eng.

12. put the tessdata at a right place
$ sudo cp eng.traineddata /usr/share/tesseract-ocr/tessdata
$ sudo cp eng.pffmtable /usr/share/tesseract-ocr/tessdata
$ sudo cp eng.inttemp /usr/share/tesseract-ocr/tessdata
$ sudo cp eng.normproto /usr/share/tesseract-ocr/tessdata
$ sudo cp eng.unicharset /usr/share/tesseract-ocr/tessdata