PlagScan Results of plagiarism analysis from 2017-09-30 12:15 UTC

Sem7-Project.docx

12.5%

Date: 2017-09-30 12:07 UTC

✱ All sources **40** | Internet sources **9**

- [35] https://misteroleg.wordpress.com/2012/12...pre-processing-task/
  3.9% 12 matches
- [37] https://stackoverflow.com/questions/37378052/how-can-i-ocr-a-file-in-farsi
  3.8% 7 matches
- [39] https://www.quora.com/Where-can-I-get-th...e-with-documentation
  2.8% 7 matches
- [40] https://www.scribd.com/document/258117998/REport
  2.8% 4 matches
- [41] https://medium.com/@hdinhofer/optical-ch...-vision-76887e1d6ab0
  2.4% 3 matches
- [42] https://www.quora.com/How-does-the-Tesseract-API-for-OCR-work
  2.4% 2 matches
- [43] https://www.coursehero.com/file/pgpmjj/N...Tessesract-We-first/
  2.1% 4 matches
- [45] enlighten-ing.com/wp/
  0.4% 1 matches
- [46] ufology.wikia.com/wiki/Charles_Brown
  0.3% 1 matches

17 pages, 1803 words

⚠ A very light text-color was detected that might conceal letters used to merge words.

PlagLevel: selected / overall

133 matches from 47 sources, of which 45 are online sources.

Settings

Data policy: *Compare with web sources, Check against my documents*
Sensitivity: *Medium*
Bibliography: *Consider text*
Citation detection: *Reduce PlagLevel*
Whitelist: --

Introduction

What is a Personal Info Assistant (PIA)?
Following the idea of a "Personal Digital Assistant" (PDA), a PIA can be defined as a hardware which runs applications that provide quick reference to lists and processed data through proper links.

Why PIA?
We live in the age of information.
In the entire length of time between waking up to the sound of an alarm set on our branded smartphones and setting the same alarm before going to bed at night, we encounter a wide variety of tasks every day.
Common among these activities is the fact that each of these activities expects us to be informed. Using a washing machine needs us to know how to operate the buttons. Using an air-conditioner needs us to know what buttons to press on the remote in order to get the right setting.

Well, these are simple, aren't they?
Yes, because manufacturers make their products easy to use by hiding their inner features.

What's your response when your washing machine wouldn't run no matter how many buttons you press or your AC won't cool at the right temperature?
We all get irritated, don't we? The best response we have is to call customer care or a repairman.
Right at that moment something crosses my mind. Given our dependency on these, how little do we know about the things around us!
How much do we know about our 32" LCD Plasma TV; about the 5-speed automatic washing machine; about the 4-star rated AC; about the RO Water Purifier – all of which we operate on a daily basis at our homes?
"I'm not a techie guy", you may say, and you're correct. It's not necessary to knew everything about them, but a working knowledge wouldn't hurt.
This issue is serious for students & professionals.

Not getting the desired output on the CRO no matter how well you have connected the circuits? Does it bother you why it happened?
Been there. Done that.
[45]
The more we go out, the more we learn how little we know. But that is no excuse to have no idea about not having any idea about things we use on a daily basis.

The only problem is – How would it be done?

No one has the time or desire to leaf through the pages of a user guide; or to search for authentic documentation on the device.
Well, what if I told you that you do not need to do any of these; the information will reveal itself to you!
What would your response be?

Well, this project may do just that.[37]

Summary of Tesseract OCR Engine

Tesseract was originally developed at Hewlett-Packard Laboratories Bristol and at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994, with some more changes made in 1996 to port to Windows, and some C++izing in 1998. In 2005 Tesseract was open sourced by HP. Since 2006 it is developed by Google.

- The 'tesseract-ocr'package contains an OCR engine - libtesseract and a command line program - tesseract.
- The lead developer is Ray Smith. The maintainer is Zdenko Podobny.
- Tesseract has unicode (UTF-8) support, and can recognize more than 100 languages "out of the box".
- Tesseract supports various output formats: plain-text, hocr(html), pdf, tsv, invisible-text-only pdf.
- You should note that in many cases, in order to get better OCR results, you'll need to improve the quality of the image you are giving Tesseract.
- This project does not include a GUI application. If you need one, please see the 3rdParty wiki page.
- Tesseract can be trained to recognize other languages. See Tesseract Training for more information.

Improving the quality of the output:

There are a variety of reasons you might not get good quality output from Tesseract. It's important to note that unless you're using a very unusual font or a new language retraining Tesseract is unlikely to help.

Image processing
- Rescaling
- Binarisation
- Noise Removal
- Rotation / Deskewing
- Border Removal
- Tools / Libraries

Page segmentation method
Dictionaries, word lists, and patterns

Image processing

Tesseract does various image processing operations internally (using the Leptonica library) before doing the actual OCR. It generally does a very good job of this, but there will inevitably be cases where it isn't good enough, which can result in a significant reduction in accuracy.

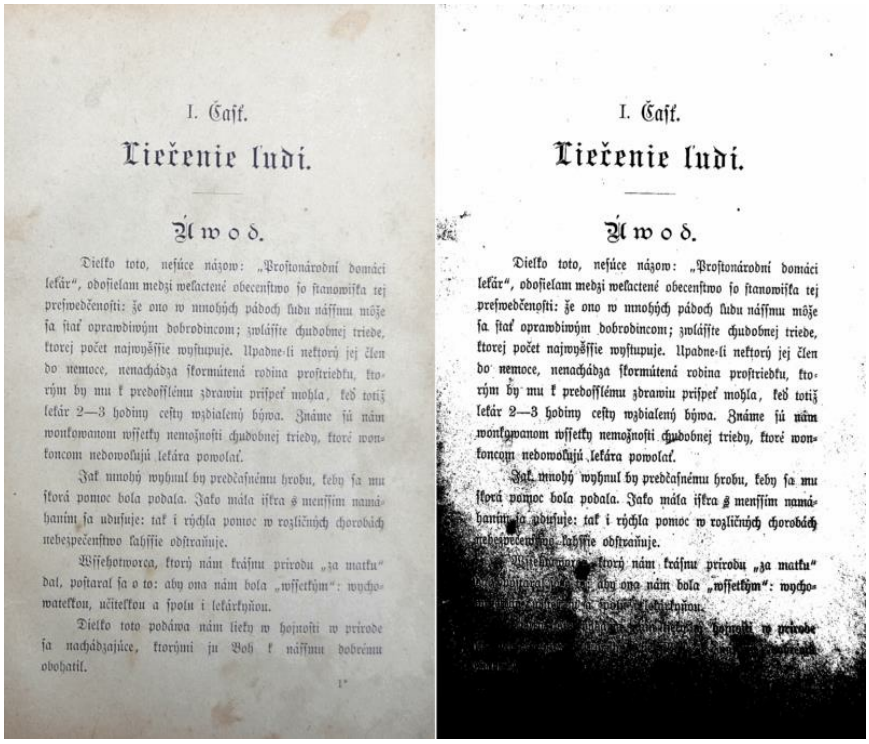
You can see how Tesseract has processed the image by using the configuration variable tessedit_write_images to true when running Tesseract. If the resulting tessinput.tif file looks problematic, try some of these image processing operations before passing the image to Tesseract.

Rescaling

Tesseract works best on images which have a DPI of at least 300 dpi, so it may be beneficial to resize images.

Binarisation

This is converting an image to black and white.[40] Tesseract does this internally, but the result can be suboptimal, particularly if the page background is of uneven darkness.



I. Časť.

Liečenie ľudi.

Úvod.

Dielko toto, nesúce názov: „Prostonárodní domáci lekár", odosielam medzi welactené obecenstwo so stanowiska tej prespwedčenosti: že ono w mnohých pádoch ľudu nášmu může sa stať oprawdiwým dobrodincom; zwlášste chudobnej triede, ktorej počet najwyššie wystupuje. Upadne-li nektorý jej člen do nemoce, nenachádza skormútená rodina prostriedku, ktorým by mu k predošlému zdraviu prispeť mohla, keď totiž lekár 2—3 hodiny cesty wzdialený býwa. Známe sú nám wonkajšnom wšetky nemožnosti chudobnej triedy, ktoré wonkoncom nedowolujú lekára powolať.

Jak mnohý wyhnul by predčasnému hrobu, keby sa mu skorá pomoc bola podala. Jako mála iskra s menšším namáhaním sa udušuje: tak i rýchla pomoc w rozličných chorobách nebezpečenstwo ľahšie odstraňuje.

Wšehotworca, ktorý nám krásnu prirodu „za matku" dal, postaral sa o to: aby ona nám bola „wšetkým": wychowatelkou, učitelkou a spolu i lekárkyňou.

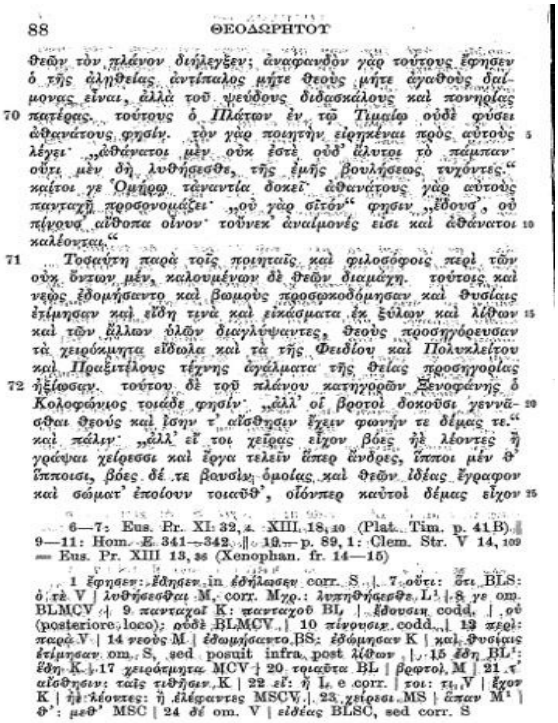
Dielko toto podáwa nám lieky w hojnosti w prirode sa nachádzajúce, ktorými ju Boh k nášmu dobrému obohatil.

1*

[40]

Noise Removal

Noise is random variation of brightness or colour in an image, that can make the text of the image more difficult to read.[40] Certain types of noise cannot be removed by Tesseract in the binarisation step, which can cause accuracy rates to drop.



88 ΘΕΟΔΩΡΗΤΟΥ

θεῶν τὸν πλάνον διήλεγξεν; ἀναφανδὸν γὰρ τούτους ἔφησεν
ὁ τῆς ἀληθείας ἀντίπαλος μήτε θεοὺς μήτε ἀγαθοὺς δαι-
μονας εἶναι, ἀλλὰ τοῦ ψεύδους διδασκάλους καὶ πονηρίας
70 πατέρας. τούτους ὁ Πλάτων ἐν τῷ Τιμαίῳ οὐδὲ φύσει
ἀθανάτους φησίν. τὸν γὰρ ποιητὴν εἰρηκέναι πρὸς αὐτοὺς 5
λέγει· „ἀθάνατοι μὲν οὐκ ἐστὲ οὐδ' ἄλυτοι τὸ πάμπαν·
οὔτι μὲν δὴ λυθήσεσθε, τῆς ἐμῆς βουλήσεως τυγόντες.“
καίτοι γε Ὁμήρῳ τἀναντία δοκεῖ· ἀθανάτους γὰρ αὐτοὺς
πανταχῇ προσονομάζει· „οὐ γὰρ σίτον“ φησιν „ἔδουσ', οὐ
πίνουσ' αἴθοπα οἶνον· τοὔνεκ' ἀναίμονές εἰσι καὶ ἀθάνατοι 10
καλέονται.“

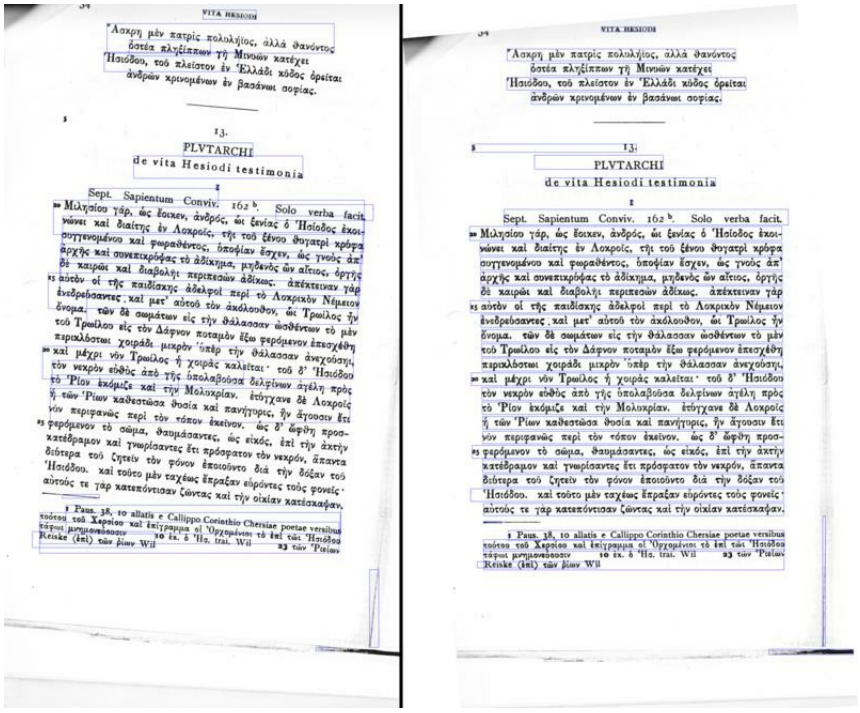
71 Τοσαύτη παρὰ τοῖς ποιηταῖς καὶ φιλοσόφοις περὶ τῶν
οὐκ ὄντων μέν, καλουμένων δὲ θεῶν διαμάχη. τούτοις καὶ
νεὼς ἐδομήσαντο καὶ βωμοὺς προσῳκοδόμησαν καὶ θυσίας
ἐτέλεσαν καὶ εἴδη τινὰ καὶ εἰκάσματα ἐκ ξύλων καὶ λίθων 15
καὶ τῶν ἄλλων ὑλῶν διαγλύψαντες, θεοὺς προσηγόρευσαν
τὰ χειρόκμητα εἴδωλα καὶ τὰ τῆς Φειδίου καὶ Πολυκλείτου
καὶ Πραξιτέλους τέχνης ἀγάλματα τῆς θείας προσηγορίας
72 ἠξίωσαν. τούτου δὲ τοῦ πλάνου κατηγορῶν Ξενοφάνης ὁ
Κολοφώνιος τοιάδε φησίν· „ἀλλ' οἱ βροτοὶ δοκοῦσι γεννᾶ- 20
σθαι θεούς καὶ ἴσην τ' αἴσθησιν ἔχειν φωνήν τε δέμας τε.“
καὶ πάλιν· „ἀλλ' εἴ τοι χεῖρας εἶχον βόες ἠὲ λέοντες ἢ
γράψαι χείρεσσι καὶ ἔργα τελεῖν ἅπερ ἄνδρες, ἵπποι μὲν θ'
ἵπποισι, βόες δέ τε βουσὶν ὁμοίας καὶ θεῶν ἰδέας ἔγραφον
καὶ σώματ' ἐποίουν τοιαῦθ', οἷόνπερ καὐτοὶ δέμας εἶχον 25

6—7: Eus. Pr. XI 32, 4. XIII 18, 10 (Plat. Tim. p. 41 B). 9—11: Hom. E 341—342. || 10. — p. 89, 1: Clem. Str. V 14, 109 = Eus. Pr. XIII 13, 36 (Xenophan. fr. 14—15)

1 ἔφησεν: ἔδηλωσεν in ἐσήλωσεν corr. S | 7 οὔτι: ὅτι BLS: ὁ τὲ V | λυθήσεσθαι M, corr. Mχρ.: λυπηθήσεσθε L¹ | 8 γε om. BLMCV | 9 παντοχοῖ K: παντοχοῦ BL | ἔδουσιν codd. | οὐ (posteriore loco): οὐδὲ BLMCV | 10 πίνουσιν codd. | 12 περὶ: παρὰ V | 14 ποιηταῖς M | ἐδομήσαντο BS: ἐδόμησαν K | καὶ θυσίας ἐτέλεσαν om. S, sed posuit infra post λίθων | 15 ἔδη BL¹: ἔδη K | 17 χειρόκμητα MCV | 20 τοιαῦτα BL | βροτοὶ M | 21 τ' αἴσθησιν: ταῖς τιθήσιν K | 22 εἴ: ἢ L e corr. | τοι: τι V | ἔχον K | ἠὲ λέοντες: ἢ ἐλέφαντες MSCV | 23 χείρεσι MS | ἅπαν M¹ | θ': μεθ' MSC | 24 δέ om. V | εἰδέας BLSC, sed corr. S

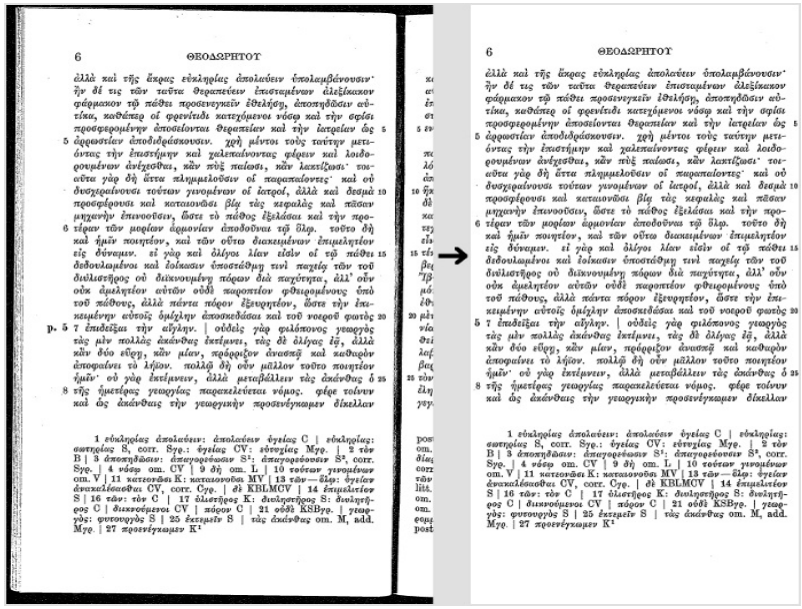
Rotation / Deskewing

A skewed image is when an page has been scanned when not straight.[40] The quality of Tesseract's line segmentation reduces significantly if a page is too skewed, which severely impacts the quality of the OCR. To address this rotating the page image so that the text lines are horizontal.



Border Removal

Scanned pages often have dark borders around them. These can be erroneously picked up as extra characters, especially if they vary in shape and gradation.



Tools / Libraries

- Leptonica
- OpenCV
- Scan Tailor
- ImageMagick
- unpaper
- ImageJ
- Gimp

Page segmentation method

By default Tesseract expects a page of text when it segments an image. If you're just seeking to OCR a small region try a different segmentation mode, using the -psm argument. Note that adding a white border to text which is too tightly cropped may also help, see issue 398.

To see a complete list of supported page segmentation modes, use tesseract -h. Here's the list as of 3.21:

[35]
0 Orientation and script detection (OSD) only.[35]

1 Automatic page segmentation with OSD.[35]

2 Automatic page segmentation, but no OSD, or OCR.[35]

3 Fully automatic page segmentation, but no OSD.[35] (Default)

4 Assume a single column of text of variable sizes.[35]

5 Assume a single uniform block of vertically aligned text.[35]

6 Assume a single uniform block of text.[35]

7 Treat the image as a single text line.[35]

8 Treat the image as a single word.[35]

9 Treat the image as a single word in a circle.[35]

10 Treat the image as a single character.

11 Sparse text. Find as much text as possible in no particular order.

12 Sparse text with OSD.

13 Raw line. Treat the image as a single text line, bypassing hacks that are Tesseract-specific.

Dictionaries, word lists, and patterns

By default Tesseract is optimized to recognize sentences of words. If you're trying to recognize something else, like receipts, price lists, or codes, there are a few things you can do to improve the accuracy of your results, as well as double-checking that the appropriate segmentation method is selected.

Disabling the dictionaries Tesseract uses should increase recognition if most of your text isn't dictionary words. They can be disabled by setting the both of the configuration variables load_system_dawg and load_freq_dawg to false.

It is also possible to add words to the word list Tesseract uses to help recognition, or to add common character patterns, which can further help to improve accuracy if you have a good idea of the sort of input you expect. This is explained in more detail in the Tesseract manual.

If you know you will only encounter a subset of the characters available in the language, such as only digits, you can use the tessedit_char_whitelist configuration variable. See the FAQ for an example.

Implementation

1. Block Diagram
2. Progress Remarks
3. Softwares
4. Source Code Listing
    - pytesseract.py
    - ocr_pia_1.py
    - ocr_pia_2.py
    - thresholding_adaptive_naive.py
    - thresholding_global_naive.py
    - thresholding_global_naive2.py
    - thresholding_global_otsu.py
    - thresholding_global_otsu2.py

Progress Remarks

- OCR Pre-processing
  The first step towards a mobile-platform implementation for our project was to identify the key areas to be resolved to improve text recognition.
  We identified these areas to be:

1. Choice of lossless compression image format (eg. PNG)
2. Adaptive Binarization of image
3. Fragmentation
4. Filter Application

- Source Code
  The source-code will be made available (alongwith necessary datasets) on my Github page:
  http://www.github.com/CRT13/Projects/

Results

Basic Dataset

Our basic dataset consisted of 5 images, captured using a Lenovo K6 Power smartphone. The testing system was a low-end laptop running Windows, with the following specs:

imple Thresholding using OpenCV

OpenCV offers 5 options for "Simple Thresholding". These are listed below:

- **THRESH_BINARY**

$$\texttt{dst}(x,y) = \begin{cases} \texttt{maxval} & \text{if } \texttt{src}(x,y) > \texttt{thresh} \\ 0 & \text{otherwise} \end{cases}$$

- **THRESH_BINARY_INV**

$$\texttt{dst}(x,y) = \begin{cases} 0 & \text{if } \texttt{src}(x,y) > \texttt{thresh} \\ \texttt{maxval} & \text{otherwise} \end{cases}$$

- **THRESH_TRUNC**

$$\texttt{dst}(x,y) = \begin{cases} \texttt{threshold} & \text{if } \texttt{src}(x,y) > \texttt{thresh} \\ \texttt{src}(x,y) & \text{otherwise} \end{cases}$$

- **THRESH_TOZERO**

$$\texttt{dst}(x,y) = \begin{cases} \texttt{src}(x,y) & \text{if } \texttt{src}(x,y) > \texttt{thresh} \\ 0 & \text{otherwise} \end{cases}$$

- **THRESH_TOZERO_INV**

$$\texttt{dst}(x,y) = \begin{cases} 0 & \text{if } \texttt{src}(x,y) > \texttt{thresh} \\ \texttt{src}(x,y) & \text{otherwise} \end{cases}$$

The details can be referred from pg. 294 of "The OpenCV Reference Manual".
So, we generated sample plots for each of these.

Original Image

BINARY

BINARY_INV

TRUNC

TOZERO

TOZERO_INV

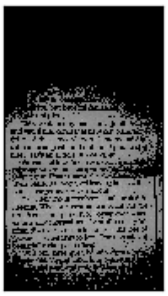
Original Image

BINARY

BINARY_INV

TRUNC

TOZERO



TOZERO_INV

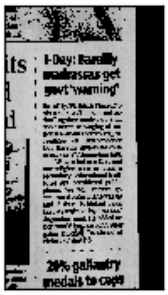
Original Image

BINARY

BINARY_INV

TRUNC

TOZERO



TOZERO_INV

Original Image

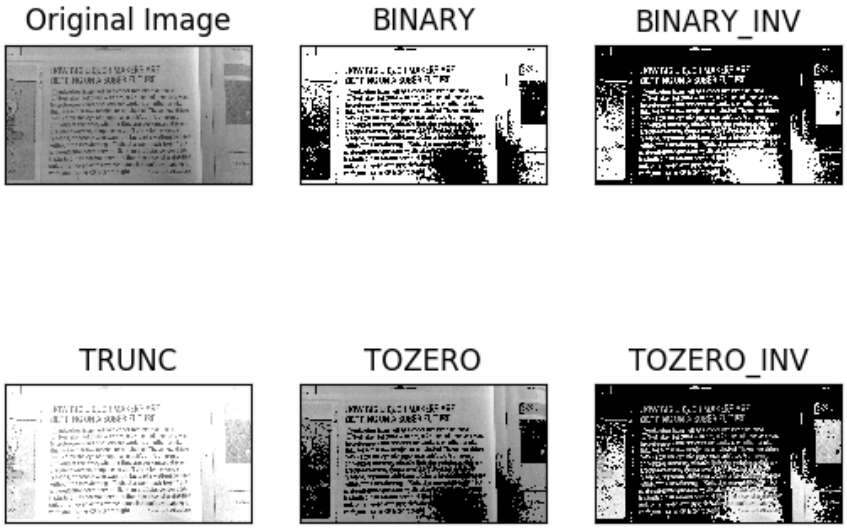
BINARY

BINARY_INV

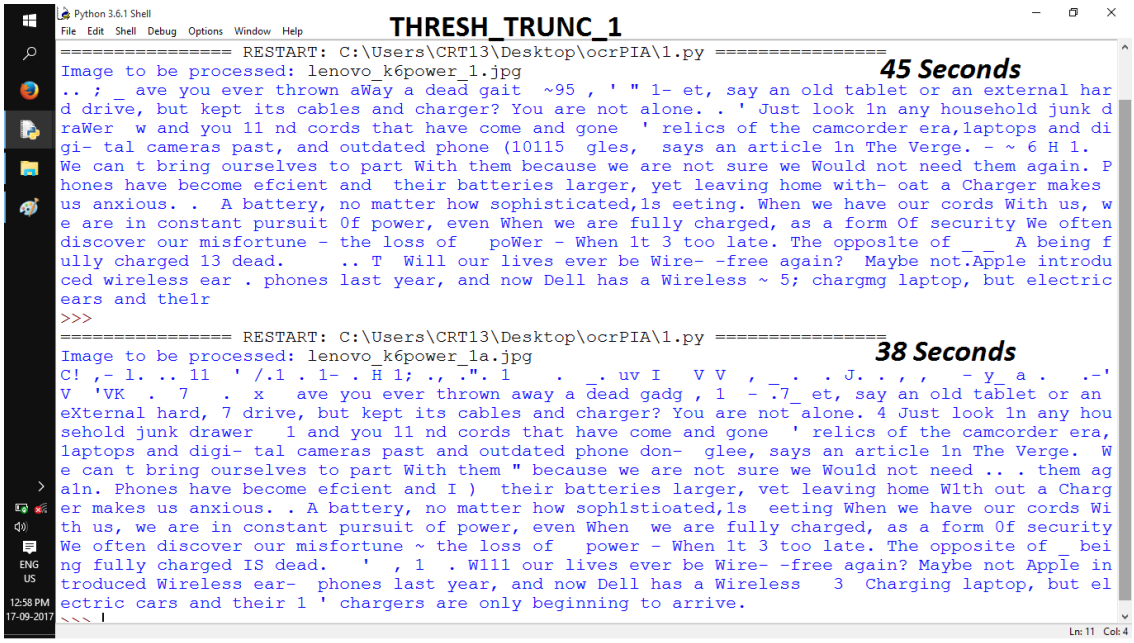
TRUNC

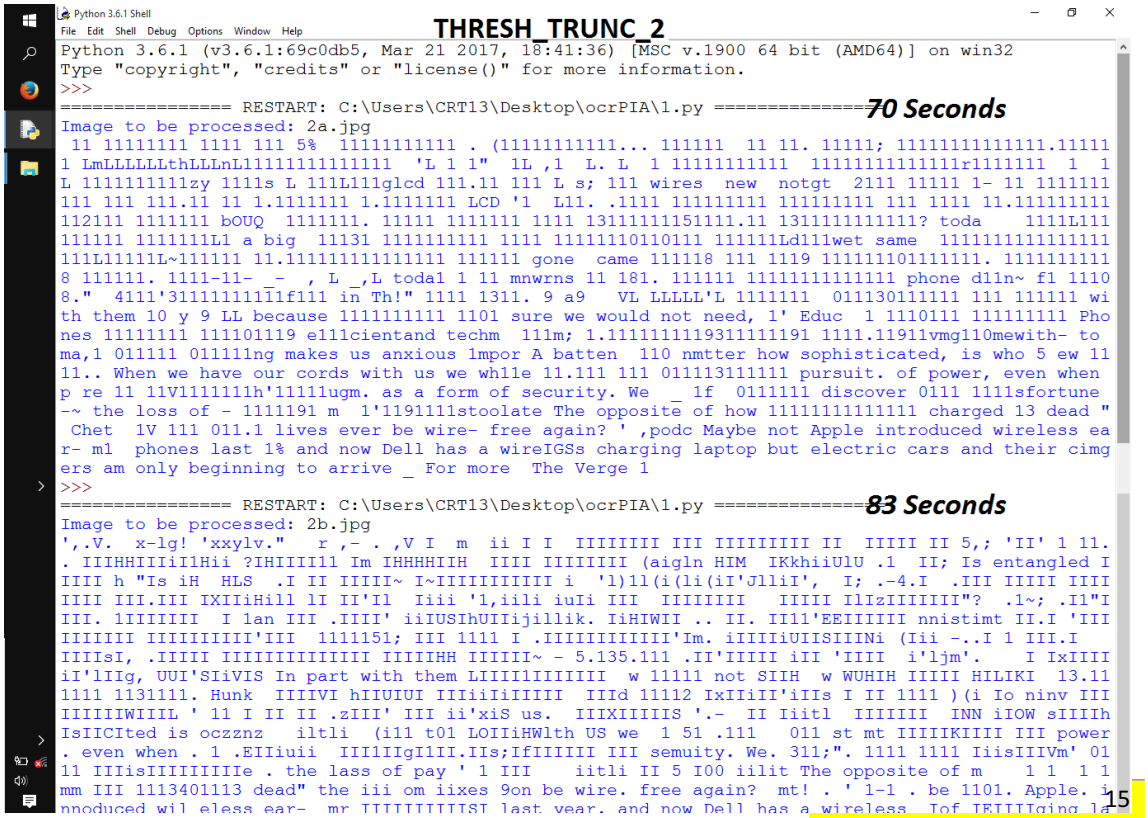
TOZERO

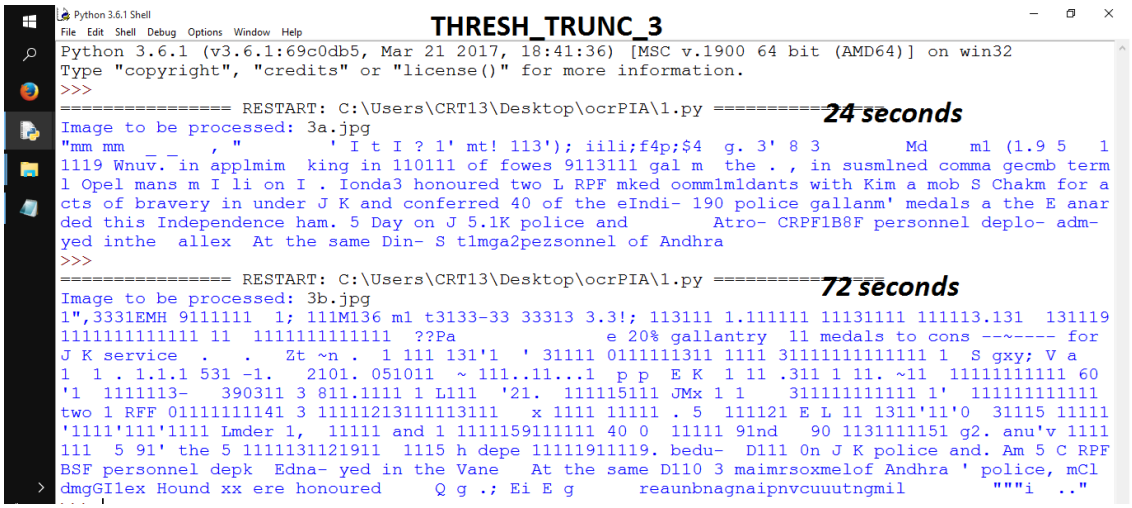
TOZERO_INV

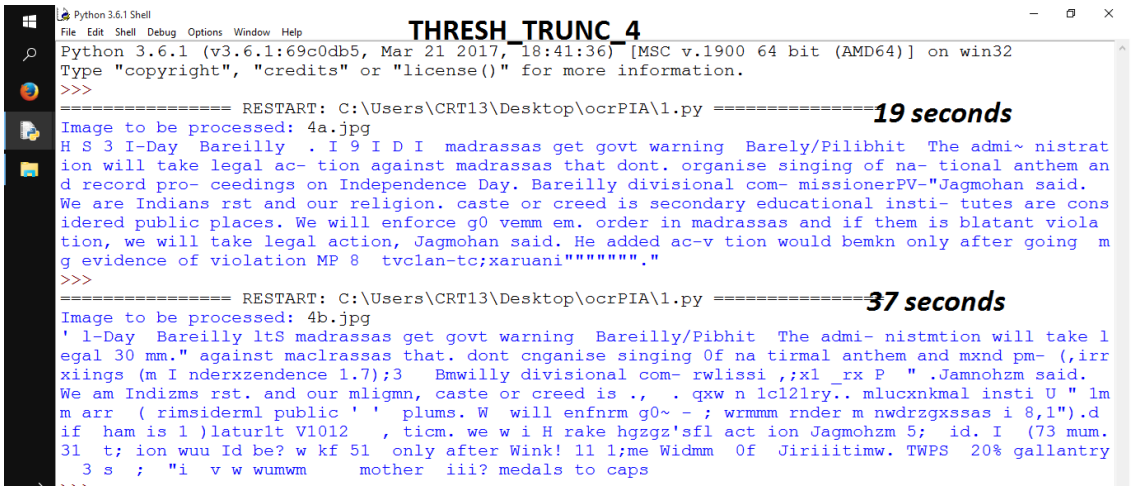


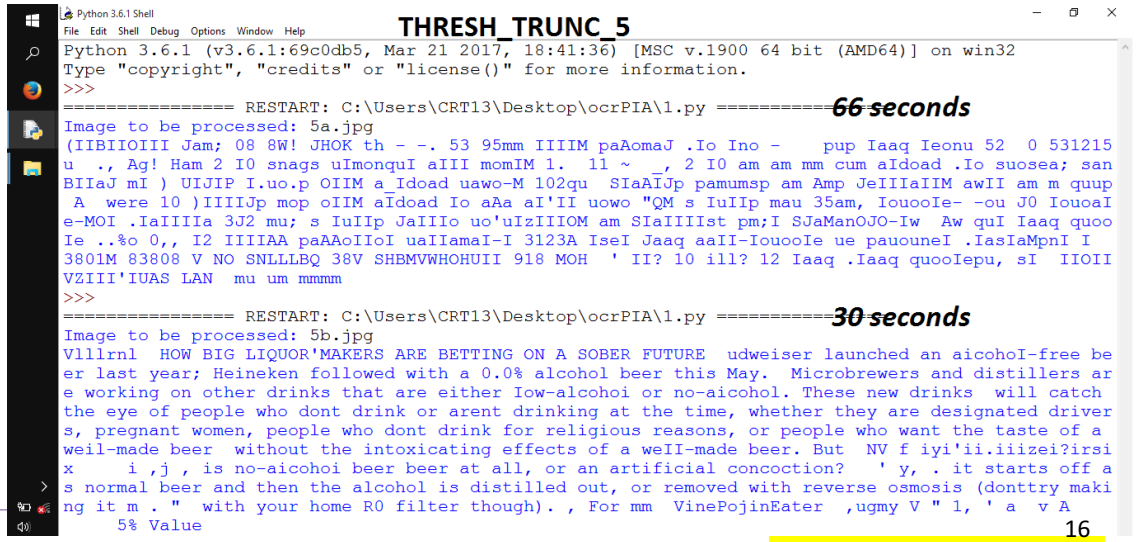
Next, we tried to observe the impact of “Simple Thresholding” on Tesseract's OCR output.
We present the results for THRESH_TRUNC. First things first, there was an obvious reduction in file-size, but the runtimes were arbitrary, possibly due to incosistencies caused by the lossy nature of JPEG images.


Python 3.6.1 Shell
File Edit Shell Debug Options Window Help
THRESH_TRUNC_1
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: lenovo_k6power_1.jpg
45 Seconds
.. ; _ ave you ever thrown aWay a dead gait ~95 , ' " 1- et, say an old tablet or an external har
d drive, but kept its cables and charger? You are not alone. . ' Just look 1n any household junk d
raWer w and you 11 nd cords that have come and gone ' relics of the camcorder era,laptops and di
gi- tal cameras past, and outdated phone (10115 gles, says an article 1n The Verge. - ~ 6 H 1.
We can t bring ourselves to part With them because we are not sure we Would not need them again. P
hones have become efcient and their batteries larger, yet leaving home with- oat a Charger makes
us anxious. . A battery, no matter how sophisticated,1s eeting. When we have our cords With us, w
e are in constant pursuit 0f power, even When we are fully charged, as a form Of security We often
discover our misfortune - the loss of poWer - When 1t 3 too late. The opposite of _ _ A being f
ully charged 13 dead. .. T Will our lives ever be Wire- -free again? Maybe not.Apple introdu
ced wireless ear . phones last year, and now Dell has a Wireless ~ 5; chargmg laptop, but electric
ears and the1r
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
Image to be processed: lenovo_k6power_1a.jpg
38 Seconds
C! ,- 1. .. 11 ' /.1 . 1- . H 1; ., .". 1 . _. uv I V V , _. . J. . , , - y_ a . .-'
V 'VK . 7 . x ave you ever thrown away a dead gadg , 1 - .7_ et, say an old tablet or an
eXternal hard, 7 drive, but kept its cables and charger? You are not alone. 4 Just look 1n any hou
sehold junk drawer 1 and you 11 nd cords that have come and gone ' relics of the camcorder era,
laptops and digi- tal cameras past and outdated phone don- glee, says an article 1n The Verge. W
e can t bring ourselves to part With them " because we are not sure we Would not need .. . them ag
a1n. Phones have become efcient and I) their batteries larger, vet leaving home W1th out a Charg
er makes us anxious. . A battery, no matter how soph1stioated,1s eeting When we have our cords Wi
th us, we are in constant pursuit of power, even When we are fully charged, as a form 0f security
We often discover our misfortune ~ the loss of power - When 1t 3 too late. The opposite of _ bei
ng fully charged IS dead. ' , 1 . W111 our lives ever be Wire- -free again? Maybe not Apple in
troduced Wireless ear- phones last year, and now Dell has a Wireless 3 Charging laptop, but el
ectric cars and their 1 ' chargers are only beginning to arrive.
>>>
ENG
US
12:58 PM
17-09-2017
Ln: 11 Col: 4


Python 3.6.1 Shell
File Edit Shell Debug Options Window Help
THRESH_TRUNC_2
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
70 Seconds
Image to be processed: 2a.jpg
11 11111111 1111 111 5% 11111111111 . (11111111111... 111111 11 11. 11111; 111111111111111.11111
1 LmLLLLLLthLLLnL111111111111111 'L 1 1" 1L ,1 L. L 1 11111111111 1111111111111r1111111 1 1
L 1111111111zy 1111s L 111L111glcd 111.11 111 L s; 111 wires new notgt 2111 11111 1- 11 1111111
111 111 111.11 11 1.1111111 1.1111111 LCD '1 L11. .1111 111111111 111111111 111 1111 11.111111111
112111 1111111 bOUQ 1111111. 11111 1111111 1111 13111111151111.11 1311111111111? toda 1111L111
111111 1111111L1 a big 11131 1111111111 1111 11111110110111 111111Ld111wet same 11111111111111111
111L11111L~111111 11.111111111111111 111111 gone came 111118 111 1119 111111101111111. 1111111111
8 111111. 1111-11- _- , L _,L toda1 1 11 mnwrns 11 181. 111111 11111111111111 phone d11n~ f1 1110
8." 4111'3111111111f111 in Th!" 1111 1311. 9 a9 VL LLLLL'L 1111111 011130111111 111 111111 wi
th them 10 y 9 LL because 1111111111 1101 sure we would not need, 1' Educ 1 1110111 111111111 Pho
nes 11111111 111101119 e111cientand techm 111m; 1.1111111119311111191 1111.11911vmg110mewith- to
ma,1 011111 011111ng makes us anxious 1mpor A batten 110 nmtter how sophisticated, is who 5 ew 11
11.. When we have our cords with us we wh11e 11.111 111 011113111111 pursuit. of power, even when
p re 11 11V1111111h'11111ugm. as a form of security. We _ 1f 0111111 discover 0111 1111sfortune
-~ the loss of - 1111191 m 1'1191111stoolate The opposite of how 11111111111111 charged 13 dead "
Chet 1V 111 011.1 lives ever be wire- free again? ' ,podc Maybe not Apple introduced wireless ea
r- m1 phones last 1% and now Dell has a wireIGSs charging laptop but electric cars and their cimg
ers am only beginning to arrive _ For more The Verge 1
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
83 Seconds
Image to be processed: 2b.jpg
',.V. x-lg! 'xxylv." r ,- . ,V I m ii I I IIIIIIII III IIIIIIIII II IIIII II 5,; 'II' 1 11.
. IIIHHIIIiI1Hii ?IHIIII11 Im IHHHHIIH IIII IIIIIIII (aigln HIM IKkhiiUlU .1 II; Is entangled I
IIII h "Is iH HLS .I II IIIII~ I~IIIIIIIIII i 'l)ll(i(li(iI'JlliI', I; .-4.I .III IIIII IIII
IIII III.III IXIIiHill lI II'Il Iiii '1,iili iuIi III IIIIIIII IIIII IlIzIIIIIII"? .1~; .I1"I
III. 1IIIIIIII I 1an III .IIII' iiIUSIhUIIijillik. IiHIWII .. II. II11'EEIIIIII nnistimt II.I 'III
IIIIIII IIIIIIIII'III 1111151; III 1111 I .IIIIIIIIIII'Im. iIIIIiUIISIIINi (Iii -..I 1 III.I
IIIIsI, .IIIII IIIIIIIIIIIII IIIIHH IIIIII~ - 5.135.111 .II'IIIII iII 'IIII i'ljm'. I IxIIII
iI'lIIg, UUI'SIiVIS In part with them LIIII1IIIIIII w 11111 not SIIH w WUHIH IIIII HILIKI 13.11
1111 1131111. Hunk IIIIVI hIIUIUI IIIiiIiIIIII IIId 11112 IxIIiII'iIIs I II 1111)(i Io ninv III
IIIIIIWIIIL ' 11 I II II .zIII' III ii'xiS us. IIIXIIIIIS '.- II Iiitl IIIIIII INN iIOW sIIIIh
IsIICIted is oczznz iltli (i11 t01 LOIIiHWlth US we 1 51 .111 011 st mt IIIIIKIIII III power
. even when . 1 .EIIiuii IIIIIIgI1II.IIs;IfIIIIII III semuity. We. 311;". 1111 1111 IiisIIIVm' 01
11 IIIisIIIIIIIIIe . the lass of pay ' 1 III iitli II 5 I00 iilit The opposite of m 1 1 1 1
mm III 1113401113 dead" the iii om iixes 9on be wire. free again? mt! . ' 1-1 . be 1101. Apple. i
nnoduced wil eless ear- mr IIIIIIIIIIISI last year, and now Dell has a wireless Iof IEIIIIging la


Python 3.6.1 Shell
File Edit Shell Debug Options Window Help
THRESH_TRUNC_3
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
24 seconds
Image to be processed: 3a.jpg
"mm mm _ _ , " ' I t I ? 1' mt! 113'); iili;f4p;$4 g. 3' 8 3 Md ml (1.9 5 1
1119 Wnuv. in applmim king in 110111 of fowes 9113111 gal m the . , in susmlned comma gecmb term
l Opel mans m I li on I . Ionda3 honoured two L RPF mked oommlm1dants with Kim a mob S Chakm for a
cts of bravery in under J K and conferred 40 of the eIndi- 190 police gallanm' medals a the E anar
ded this Independence ham. 5 Day on J 5.1K police and Atro- CRPF1B8F personnel deplo- adm-
yed inthe allex At the same Din- S tlmga2pezsonnel of Andhra
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
72 seconds
Image to be processed: 3b.jpg
1",3331EMH 9111111 1; 111M136 m1 t3133-33 33313 3.3!; 113111 1.111111 11131111 111113.131 131119
1111111111111 11 1111111111111 ??Pa e 20% gallantry 11 medals to cons --~---- for
J K service . . Zt ~n . 1 111 131'1 ' 31111 0111111311 1111 31111111111111 1 S gxy; V a
1 1 . 1.1.1 531 -1. 2101. 051011 ~ 111..11...1 p p E K 1 11 .311 1 11. ~11 11111111111 60
'1 1111113- 390311 3 811.1111 1 L111 '21. 111115111 JMx 1 1 311111111111 1' 111111111111
two 1 RFF 01111111141 3 11111213111113111 x 1111 11111 . 5 111121 E L 11 1311'11'0 31115 11111
'1111'111'1111 Lmder 1, 11111 and 1 1111159111111 40 0 11111 91nd 90 1131111151 g2. anu'v 1111
111 5 91' the 5 1111131121911 1115 h depe 11111911119. bedu- D111 0n J K police and. Am 5 C RPF
BSF personnel depk Edna- yed in the Vane At the same D110 3 maimrsoxmelof Andhra ' police, mCl
dmgGIlex Hound xx ere honoured Q g .; Ei E g reaunbnagnaipnvcuuutngmil """i .."


Python 3.6.1 Shell
File Edit Shell Debug Options Window Help
THRESH_TRUNC_4
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
19 seconds
Image to be processed: 4a.jpg
H S 3 I-Day Bareilly . I 9 I D I madrassas get govt warning Barely/Pilibhit The admi~ nistrat
ion will take legal ac- tion against madrassas that dont. organise singing of na- tional anthem an
d record pro- ceedings on Independence Day. Bareilly divisional com- missionerPV-"Jagmohan said.
We are Indians rst and our religion. caste or creed is secondary educational insti- tutes are cons
idered public places. We will enforce g0 vemm em. order in madrassas and if them is blatant viola
tion, we will take legal action, Jagmohan said. He added ac-v tion would bemkn only after going m
g evidence of violation MP 8 tvclan-tc;xaruani"""""""."
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
37 seconds
Image to be processed: 4b.jpg
' l-Day Bareilly ltS madrassas get govt warning Bareilly/Pibhit The admi- nistmtion will take l
egal 30 mm." against maclrassas that. dont cnganise singing 0f na tirmal anthem and mxnd pm- (,irr
xiings (m I nderxzendence 1.7);3 Bmwilly divisional com- rwlissi ,;x1 _rx P " .Jamnohzm said.
We am Indizms rst. and our mligmn, caste or creed is ., . qxw n 1c121ry.. mlucxnkmal insti U " 1m
m arr (rimsiderml public ' ' plums. W will enfnrm g0~ - ; wrmmm rnder m nwdrzgxssas i 8,1").d
if ham is 1)latur1t V1012 , ticm. we w i H rake hgzgz'sfl act ion Jagmohzm 5; id. I (73 mum.
31 t; ion wuu Id be? w kf 51 only after Wink! 11 1;me Widmm 0f Jiriiitimw. TWPS 20% gallantry
3 s ; "i v w wumwm mother iii? medals to caps


Python 3.6.1 Shell
File Edit Shell Debug Options Window Help
THRESH_TRUNC_5
Python 3.6.1 (v3.6.1:69c0db5, Mar 21 2017, 18:41:36) [MSC v.1900 64 bit (AMD64)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
66 seconds
Image to be processed: 5a.jpg
(IIBIIOIII Jam; 08 8W! JHOK th - -. 53 95mm IIIIM paAomaJ .Io Ino - pup Iaaq Ieonu 52 0 531215
u ., Ag! Ham 2 I0 snags uImonquI aIII momIM 1. 11 ~ _, 2 I0 am am mm cum aIdoad .Io suosea; san
BIIaJ mI) UIJIP I.uo.p OIIM a_Idoad uawo-M 102qu SIaAIJp pamumsp am Amp JeIIIaIIM awII am m quup
A were 10)IIIIJp mop oIIM aIdoad Io aAa aI'II uowo "QM s IuIIp mau 35am, IouooIe- -ou J0 IouoaI
e-MOI .IaIIIIa 3J2 mu; s IuIIp JaIIIo uo'uIzIIIOM am SIaIIIIst pm;I SJaManOJO-Iw Aw quI Iaaq quoo
Ie ..%o 0,, I2 IIIIAA paAAoIIoI uaIIamaI-I 3123A IseI Jaaq aaII-IouooIe ue pauouneI .IasIaMpnI I
3801M 83808 V NO SNLLLBQ 38V SHBMVWHOHUII 918 MOH ' II? 10 ill? 12 Iaaq .Iaaq quooIepu, sI IIOII
VZIII'IUAS LAN mu um mmmm
>>>
========== RESTART: C:\Users\CRT13\Desktop\ocrPIA\1.py ==========
30 seconds
Image to be processed: 5b.jpg
Vlllrnl HOW BIG LIQUOR'MAKERS ARE BETTING ON A SOBER FUTURE udweiser launched an aicohoI-free be
er last year; Heineken followed with a 0.0% alcohol beer this May. Microbrewers and distillers ar
e working on other drinks that are either Iow-alcohoi or no-aicohol. These new drinks will catch
the eye of people who dont drink or arent drinking at the time, whether they are designated driver
s, pregnant women, people who dont drink for religious reasons, or people who want the taste of a
weil-made beer without the intoxicating effects of a weII-made beer. But NV f iyi'ii.iiizei?irsi
x i ,j , is no-aicohoi beer beer at all, or an artificial concoction? ' y, . it starts off a
s normal beer and then the alcohol is distilled out, or removed with reverse osmosis (donttry maki
ng it m . " with your home R0 filter though). , For mm VinePojinEater ,ugmy V " 1, ' a v A
5% Value

References

Patents:
[1] US6577762B1, Background Surface thresholding
https://patents.google.com/patent/US6577762B1
[2] US7400768B1, Enhanced optical recognition of digitized images through selective bit-insertion.
https://patents.google.com/patent/US7400768B1
[3] US9298980B1, Image preprocessing for character recognition.
https://patents.google.com/patent/US9298980B1
[4] US20120063690A1, Object-Based Optical Character Recognition Pre-Processing Algorithm.
https://patents.google.com/patent/US20120063690A1
[5] US7106905B2, Systems and methods for processing text-based electronic documents.
https://patents.google.com/patent/US7106905B2
[6] US20130329023A1, Text recognition driven functionality.
https://patents.google.com/patent/US20130329023A1

Literature:
[7] Eugene Borovikov, A survey of modern optical character recognition techniques
https://arxiv.org/abs/1412.4183
[8] M Seeger, C Dance, Binarising camera images for OCR (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/953754/
[9] Ranjith Unnikrishnan, Ray Smith, Combined Script and Page Orientation Estimation using the Tesseract OCR engine (ICDAR '07 Proceedings of the Ninth International Conference on Document Analysis and Recognition)
https://dl.acm.org/citation.cfm?id=1304846
[10] Ray Smith, An Overview of the Tesseract OCR Engine (MOCR '09 Proceedings of the International Workshop on Multilingual OCR)
https://dl.acm.org/citation.cfm?id=1577809
[11] Ray Smith, Daria Antoniva, Dar-Shyang Lee, Adapting the Tesseract Open Source OCR Engine for Multilingual OCR (MOCR '09 Proceedings of the International Workshop on Multilingual OCR)
https://dl.acm.org/citation.cfm?id=1577804
[12] Zheng Zhang, CL Tan, Binarizing document image using coplanar prefilter (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/953750/
[13] Zheng Zhang, CL Tan, Correcting document image warping based on regression of curved text lines (ICDAR 2003, Proceedings of the 9th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/1227732/
[14] Zheng Zhang, CL Tan, Recovery of distorted document images from bound volumes (ICDAR 2001, Proceedings of the 6th International Conference on Document Analysis and Recognition)
http://ieeexplore.ieee.org/document/953826/