OPTICAL CHARACTER RECOGNITION



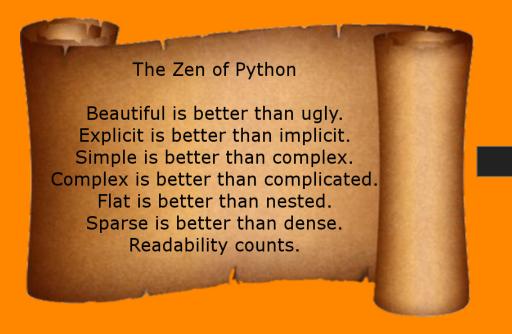
HELLO!

I am Toni Simu.

I am here because I want to discuss about OCR.
(The guitar is for show-off)

What is OCR?





The Zen of Python

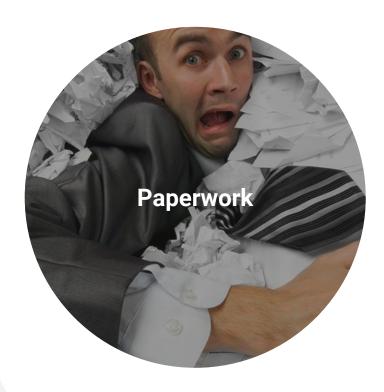
Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.

Extracts text from the images

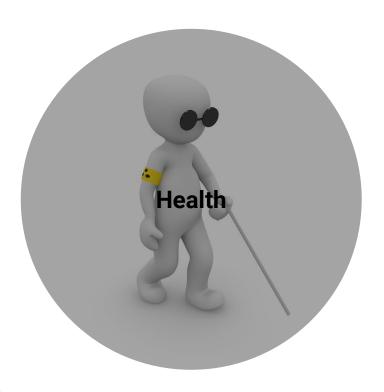


OCR Applications

OCR Applications - Paperwork



OCR Applications - Health

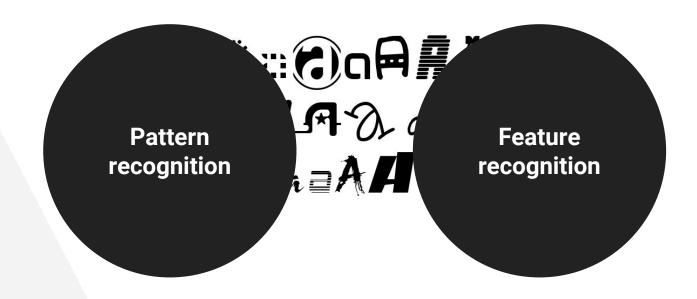


OCR Applications - Translation



How does OCR work?

How does OCR work?



How does OCR work? - Pattern recognition

ABCDEFG HIJKLMNO PQRSTUV WXYZ

1234567890



How does OCR work? - Feature recognition

$$A + 1 + - = A$$

About tesseract

Beautiful is better than ugly.

Explicit is better than implicit.

Simple is better than complex.



Beautiful is better than ugly.



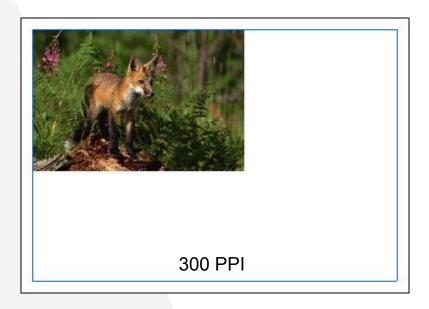
Blela ultli lf lull

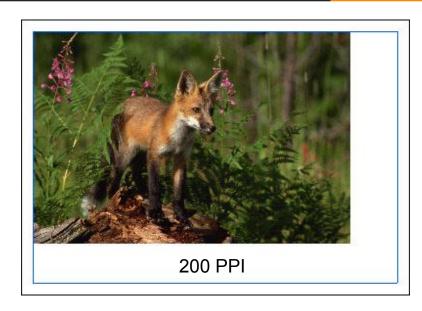
Preparing images for OCR

Image processing



Image processing - Rescaling





The image should be rescaled to have at least 300 ppi for a better accuracy.

Image processing - Binarization





- Turn the image to grayscale
 - Apply Adaptive or Global Thresholding

Image processing - Binarization - Grayscale

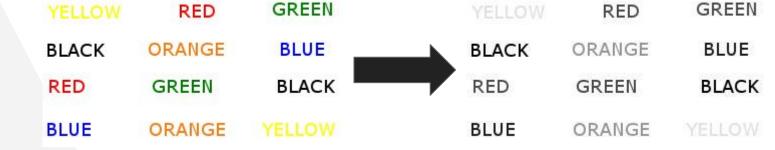


Image processing - Binarization - Thresholding

Less problematic

YELLOW	RED	GREEN
BLACK	ORANGE	BLUE
RED	GREEN	BLACK
BLUE	ORANGE	YELLOW

BLACK

RED

BLUE



RED GREEN
ORANGE BLUE
GREEN BLACK
ORANGE

More problematic





t=180

RED GREEN
ORANGE BLUE

GREEN BLACK

ORANGE

Image processing - Binarization - Adaptive Thresholding

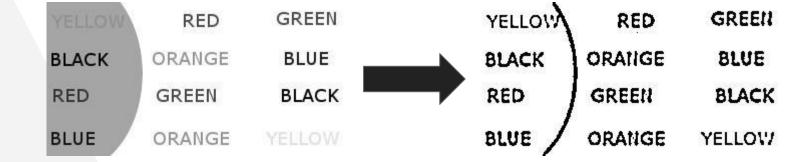


Image processing - Noise Removal

Clarification

In the last edition of the "Catholic New World," an article on the annual conference hosted by the Respect Life Office did not fully reflect the context of Cardinal George's remarks.

During an informal question-and-answer session with the archdiocese's parish Respect Life Coordinators, the cardinal emphasized that the participation by any person in the promotion of abortion, including through his or her political position, is a grave matter. While the issue of withholding Communion from some of these individuals can be complex, Cardinal George said that when any person presents him or herself to receive the Eucharist, they "take their salvation into their own hands." For a more complete explanation of this matter, reference the cardinal's column "Catholic participation in political life, revisited" (CNW, Oct. 10, 2004) online at www.catholicnewworld.com/cnw/issue/2004/cardinal_101004.html.

Clarification

In the last edition of the "Catholic New World," an article on the annual conference hosted by the Respect Life Office did not fully reflect the context of Cardinal George's remarks.

During an informal question-and-answer session with the archdiocese's parish Respect Life Coordinators, the cardinal emphasized that the participation by any person in the promotion of abortion, including through his or her political position, is a grave matter. While the issue of withholding Communion from some of these individuals can be complex, Cardinal George said that when any person presents him or herself to receive the Eucharist, they "take their salvation into their own hands." For a more complete explanation of this matter, reference the cardinal's column "Catholic participation in political life, revisited" (CNW, Oct. 10, 2004) online at www.catholicnewworld.com/cnw/issue/2004/cardinal_101004.html.

Image processing - Rotation / Deskewing

VITA RESIDDI

Ασκρη μέν πατρίς πολυληίος, αλλά θανόντος δοτέα πληξίππων τη Μινων κατέχει Ησιόδου, τού πλείστον έν Έλλδε κόδος Ορείται ἀνδρών κρινομένων έν βασάνωι σορίας.

PLVTARCHI de vita Hesiodi testimonia

Sept. Sapientum Conviv. 162 b. Solo verba facit. » Μιλησίου γάρ, ως ἔσιχεν, ἀνδρός, ωι ξενίας ὁ Ἡσίοδος έχοινώνει καὶ διαίτης ἐν Λοκροίς, τηι τοῦ ξένου θυγατρί κρόφα ουγγενομένου καί φωραθέντος, ὁποφίαν ἔσχεν, ὡς γνοὸς ἀπ' αρχής και συνεπικρόφας το άδικημα, μηδενός ων αίτιος, όργης δε καιρώι και διαβολήι περιπεσών άδίκως. ἀπέκτειναν γάρ ες αυτόν οί της παιδίσχης άδελφοί περί το Λοχρικόν Νέμειον ένεδρεβοαντες καί μετ' αύτου τον ακόλουθον, ώι Τρωίλος ήν δνομα, των δε σωμάτων είς την θάλασσαν ώσθέντων το μέν τού Τρωίλου είς τον Δάφνον ποταμόν έξω φερόμενον έπεσχέθη περικλόστωι χοιράδι μικρόν υπέρ την θάλασσαν άνεχούσης. » καὶ μέχρι νου Τροκλος ή χοιράς καλείται · τοδ δ' Ησιόδου τον νεχρόν εδθός από της οπολαβούσα δελφίνων αγέλη πρός τό 'Ρίον έχόμιζε και την Μολυχρίαν, έτθηχανε δέ Λοχροίς η των Ρίων καθεστώσα θυσία και πανήγορις, ην άγουσεν έτε νον περιφανώς περί τον τόπον έπείνου, ώς δ' ώφθη προσες φερόμενον το σώμα, θαυμάσαντες, ώς είχος, έπι την άκτην κατέδραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἄπαντα δεύτερα του ζητείν τον φόνον έποιούντο διά την δόξαν του Ήσιόδου, καὶ τοῦτο μὲν ταχέως ἔπραξαν εῦρόντες τοὺς φονείς αύτούς τε γαρ κατεπόντισαν ζώντας καί την οίκίαν κατέσκαψαν.

Paus. 18, 10 allatis e Callippo Ceristhio Chesiae potese vesiles especiales de Aspoios val fetipaqua ol "Oppoiose el feti nii "Holleto Reinke (del) especiales or o in a "Ho, trai, Will 23 con "Perios Reinke (del) especiales (del perios vesiles (d

VITA RESIDDE

Ασχρη μέν πατρίς πολυλήτος, άλλά θανόντος δοτέα πληξίππων γή Μινοών κατέχει Ησιόδου, τοδ πλείστον ἐν Ἑλλάδι κόδος όρείται ἀνδρών κρινομένων ἐν βασάνωι σοφίας:

PLVTARCHI de vita Hesiodi testimonia

Sept. Sapientum Conviv. 162 b. Solo verba facit. » Μιλησίου γάρ, ώς ἔσιχεν, ἀνδρός, ώι ξενίας ὁ Ἡσίοδος ἐχοινώνει και διαίτης εν Λοκροίς, τηι του έένου θυγατρί κρύφα συγγενομένου καὶ φωραθέντος, ὁποψίαν ἔσγεν, ὡς γνοὸς ἀπ' άργης καὶ συνεπικρόφας τὸ ἀδίκημα, μηδενὸς ὧν αἴτιος, ὀργης δέ καιρώι και διαβολήι περιπεσών άδίκως, απέκτειναν γάρ ει αύτον οί της παιδίσκης άδελφοί περί το Λοκρικόν Νέμειον ένεδρεύσαντες καί μετ' αύτου τον ακόλουθον, ωι Τρωίλος ήν δνομα, των δὲ σωμάτων εἰς τὴν θάλασσαν ώσθέντων τὸ μὲν τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσγέθη περικλόστωι γοιράδι μικρόν όπερ την θάλασσαν άνεγούσης, » καὶ μέγρι νῶν Τρωίλος ἡ γοιρὰς καλείται· τοῦ δ' Ἡσιόδου τὸν νεχρὸν εὐθὸς ἀπὸ τῆς ὑπολαβούσα δελφίνων ἀγέλη πρὸς το 'Ρίον εχόμιζε και την Μολυκρίαν, επύγγανε δε Λοκροίς ή των Ρίων καθεστώσα θυσία και πανήγυρις, ην άγουσιν έτι νύν περιφανώς περί τὸν τόπον ἐκείνον, ὡς δ' ώφθη προσες φερόμενον το σώμα, θαυμάσαντες, ώς είκος, ἐπὶ τὴν ἀκτὴν κατέδραμον καὶ γνωρίσαντες έτι πρόσφατον τὸν νεκρόν, απαντα δεύτερα τοῦ ζητείν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ Ήσιόδου, καὶ τοῦτο μὲν ταγέως ἔπραξαν εὐρόντες τοὺς φονείς αὐτούς τε γάρ κατεπόντισαν ζώντας καὶ τὴν οἰκίαν κατέσκαψαν.

³ Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τού Χεραίσου καὶ ἐπίτραμμα οἱ 'Όργομένοι τὸ ἐπὶ τὰι 'Hotoδου του μνημονισφούν 10 ἐκ. ὁ 'Ho. trai. Wil 23 τὰν 'Puslaw' Reiske (ἐπὶ) τῶν βίων Ψία.

Image processing - Border removal

@EOA9.PHT

άλλὰ καὶ τῆς ἄκρας εὐκληρίας ἀπολαύειν ὑπολαμβάνουσιν. ήν δέ τις τών ταύτα θεραπεύειν έπισταμένων άλεξέκακον φάρμακον τῷ πάθει προσενεγκεῖν ἐθελήση, ἀποπηδῶσιν αὐτίκα, καθέπερ οί φρενίτιδι κατεγόμενοι νόσω καὶ τὴν σφίσι προσφερομένην αποσείονται θεραπείαν και την Ιατρείαν ώς 5 5 άρρωστίαν αποδιδράσκουσιν. γρή μέντοι τούς τούτην μετιόντας την έπιστήμην και γαλεπαίνοντας φέρειν και λοιδορουμένων ανέχεσθαι, καν πύξ παίωσι, καν λακτίζωσι τοιαθτα γάρ δή ἄττα πλημμελούσιν οί παραπαίοντες καὶ οὐ δυσχεραίνουσε τούτων γενομένων οί λατροί, άλλά καὶ δεσμά 10 προσφέρουσι και καταιονώσι βία τὰς κεφαλάς και πάσαν μηγανήν έπινοούσιν, ώστε το πάθος έξελάσαι και την προ-6 τέραν τῶν μορίων ἀρμονίαν ἀποδοῦναι τῷ όλω, τοῦτο δὰ καὶ ήμεν ποιητέον, καὶ τῶν οθτω διακειμένων ἐπιμελητέον είς δύναμιν, εί γὰο καὶ όλίγοι λίαν είσιν οί τῷ πάθει 15 δεδουλωμένοι καὶ ἐοίκασιν ὑποστάθμη τινὶ παγεία τῶν τοῦ διύλιστήρος οὺ διϊκνουμένη πόρων διὰ παγύτητα, άλλ' οὖν ούκ αμελητέον αυτών ουδέ παροπτέον φθειρομένους ύπο τοῦ πάθους, άλλὰ πάντα πόρον έξευρητέον, ώστε την έπικειμένην αὐτοῖς ὁμίχλην ἀποσκεδάσαι καὶ τοῦ νοεροῦ φωτός 20 p. 5 7 ἐπιδείξαι τὴν αίγλην. | οὐδεὶς γὰρ φιλόπονος γεωργός τὰς μὸν πολλὰς ἀπάνθας ἐπτέμνει, τὰς δὲ ὀλίγας ἐᾶ, ἀλλὰ καν δύο εύρη, καν μίαν, πρόρριζον ανασκά και καθαρόν άποφαίνει το λήτον. πολλώ δή ουν μάλλον τουτο ποιητέον ήμεν ου γάρ επτέμνειν, άλλα μεταβάλλειν τας απάνθας 6 25 8 τῆς ἡμετέρας γεωργίας παραπελεύεται νόμος, φέρε τοίνυν καί ως ακάνθαις την γεωργικήν προσενέγκωμεν δίκελλαν

15 th

βας 25 του

COLL

τῶν litt.

1 etidopoles devalentes: devalentes opsieg C | etidopoles: controlos 8, corr. Spr. èpidies CV: interpleta (Mp. 8 v de B | 3 devançãos 8, corr. Spr. èpidies CV: interpletiones 8°; devançãos sous 6°; corr. Spr. 4 velos ou. CV | 9 de) ou. L | 10 etimes principal correspondentes 8°; corr. Spr. 4 velos ou. CV | 9 de) ou. L | 10 etimes principal correspondentes CV out. Cpc | 4 km LMCV | 13 etimes devancal devance CV, corr. Cpc | 2 km LMCV | 14 termissisches S | 16 etimes the C | 17 delartiges K: devalopripos S: devaluptigos CV | independente CV | inde

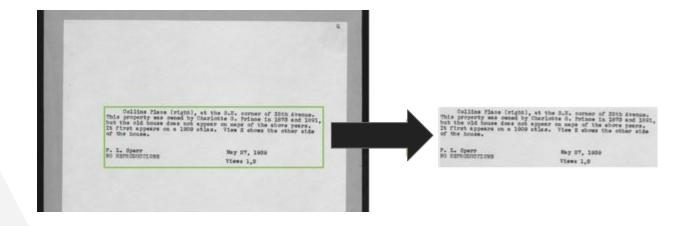
ΘΕΟΔΩΡΗΤΟΥ

άλλὰ καὶ τῆς ἄκρας εὐκληρίας ἀπολαύειν ὑπολαμβάνουσιν* ην δέ τις των ταύτα θεραπεύειν ἐπισταμένων ἀλεξίκακον φάρμακον τῷ πάθει προσενεγκεῖν ἐθελήση, ἀποπηδῶσιν αὐτίκα, καθάπες οί φρενίτιδι κατεγόμενοι νόσφ καὶ τὴν σφίσι προσφερομένην ἀποσείονται θεραπείαν και την Ιατρείαν ώς άρρωστίαν αποδιδράσκουσεν. γρή μέντοι τούς ταύτην μετιόντας την έπιστήμην και χαλεπαίνοντας φέρειν και λοιδοοουμένων ανέγεσθαι, καν πύξ παίωσι, καν λακτίζωσι τοιαθτα γάρ δή άττα πλημμελούσιν οι παραπαίοντες και οὐ δυσχεραίνουσε τούτων γινομένων οι λατροί, άλλά και δεσμά προσφέρουσι και καταιονώσι βία τὰς κεφαλάς και πάσαν μηγανήν έπινοούσιν, ώστε το πάθος έξελάσαι και την προτέραν τῶν μορίων ἀρμονίαν ἀποδοῦναι τῶ ὅλω. τοῦτο δὴ και ήμεν ποιητέον, και των οθτω διακειμένων επιμελητέον είς δύναμεν. εί γὰρ καὶ όλίγοι λίαν εἰσὶν οί τῷ πάθει δεδουλωμένοι καὶ ἐοίκασιν ὑποστάθμη τινὶ παγεία τῶν τοῦ διύλιστήρος οθ διϊκνουμένη πόρων διά παχύτητα, άλλ' οθν ούκ αμελητέου αύτων ούδε παροπτέου αθειρομένους ύπο τοῦ πάθους, άλλα πάντα πόρον έξευρητέον, ώστε την έπικειμένην αθτοίς δμίγλην άποσκεδάσαι και τοῦ νοεροῦ φωτός: 7 ἐπιδείξαι τὴν αϊγλην. | οὐδείς γὰρ φιλόπονος γεωργός τὰς μὸν πολλάς ἀπάνθας ἐπτέμνει, τὰς δὲ ὁλίγας ἔᾳ, ἀλλὰ καν δύο εθρη, καν μίαν, πρόρριζον ανασκά και καθαρόν άποφαίνει το λήϊον. πολλώ δή οθν μάλλον τοθτο ποιητέον ήμεν οθ γάρ έπτέμνειν, άλλά μεταβάλλειν τὰς ἀπάνθας ὁ: της ήμετέρας γεωργίας παραπελεύεται νόμος, φέρε τοίνυν καὶ ώς ἀκάνθαις την γεωργικήν προσενέγκωμεν δίκελλαν

1 τολιχοβικς ἀπαλιαθικεν: ἀπαλιαθικεν ψηθικες Ο | τολιχοβικες αυτορίας Β. ουτο. Τορ.: ψτοξεο Ο V: εθνευχέας Μχε. 9 2 το Β | 3 ἀπαστηθολευ: ἀπαγορεθοκεν Β'; ἀπαγορεθοκεν Β'; ἀπαγορεθοκεν Β'; διαγορεθοκεν Β'; ουτο. Σχρ. | 4 σθερ ουτ. Ο V | 9 δ) ουτ. | 10 ανόστα γενοσένου ουν. V | 11 κατευνδεκ Κι καταιονόδοι ΜV | 13 τάν — Ελές : ψεδεν ἀνακαλθοκούθαι CV, ουτ. Ο γρ. | 6 ΚΕΜΑΟΥ | 14 ἐπαμελεθος Ο | 1 εποθείνειο CV | 1 σόρο C | 2 είναι ΚΕΒΡ, ε | γεκορός ου ουνουργός Β | 25 ἐπερεξεί Β | τὸς ἀκαθτῶς ουν. Μη αδ. Μφ. | 27 παρεστέγρασμε Κ' |

Tips and tricks

Tips and tricks - Crop the text



Try to crop the text part from the image

Tips and tricks - Tesseract configuration

You have a parameter -psm(page segmentation method) that can take the next values:

- O Orientation and script detection (OSD) only.
- 1 Automatic page segmentation with OSD.
- 2 Automatic page segmentation, but no OSD, or OCR.
- 3 Fully automatic page segmentation, but no OSD. (Default)
- 4 Assume a single column of text of variable sizes.
- 5 Assume a single uniform block of vertically aligned text.
- 6 Assume a single uniform block of text.
- 7 Treat the image as a single text line.
- 8 Treat the image as a single word.
- 9 Treat the image as a single word in a circle.
- 10 Treat the image as a single character.

Demo Time



Conclusions / what are you going to do with it?

Resources

- https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality
- https://github.com/danvk/oldnyc/blob/master/ocr/tess/crop_morp hology.py
- http://static.googleusercontent.com/media/research.google.com/r o//pubs/archive/33418.pdf

You can find the code from demos on:

https://github.com/simutoni/ocr_examples

THANKS!

Any questions?

You can find me at simutoni@gmail.com