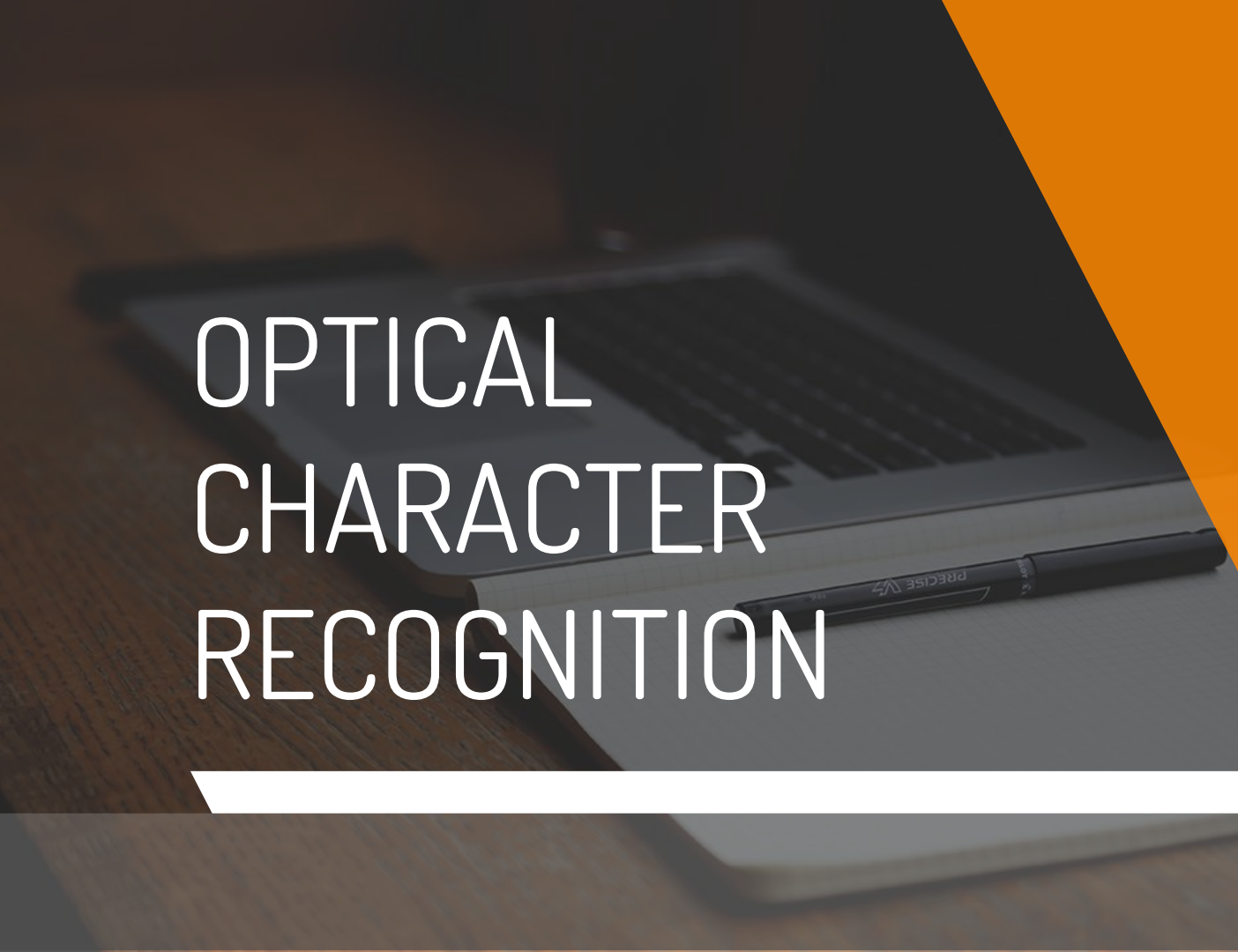
OPTICAL CHARACTER RECOGNITION



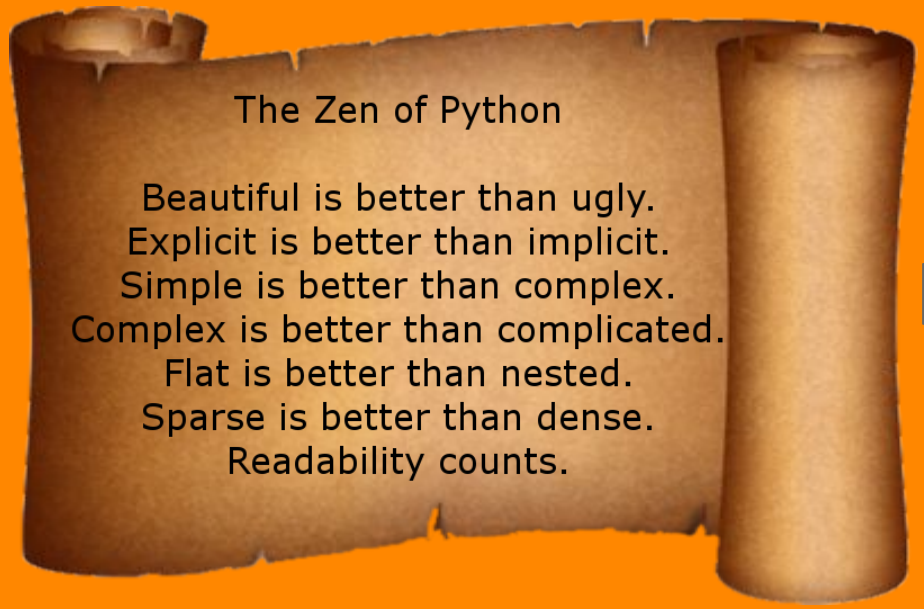
HELLO!

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

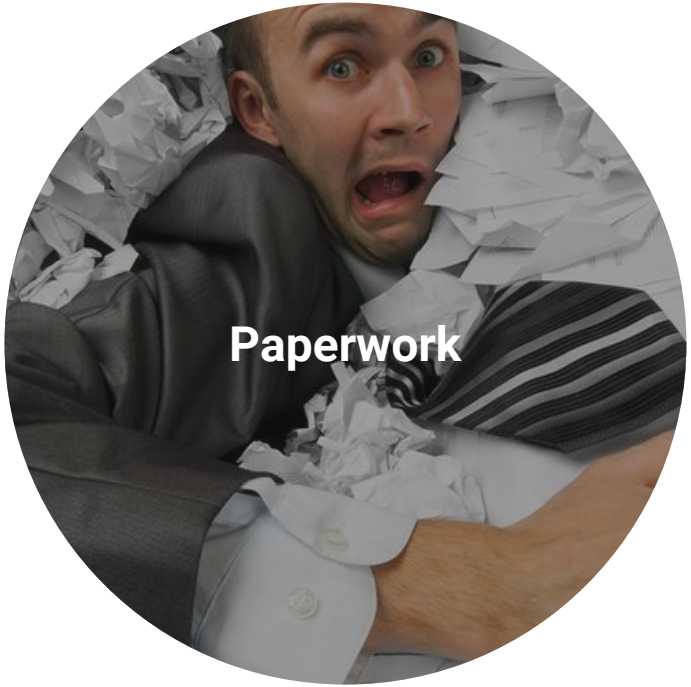
Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.

Extracts text from the images

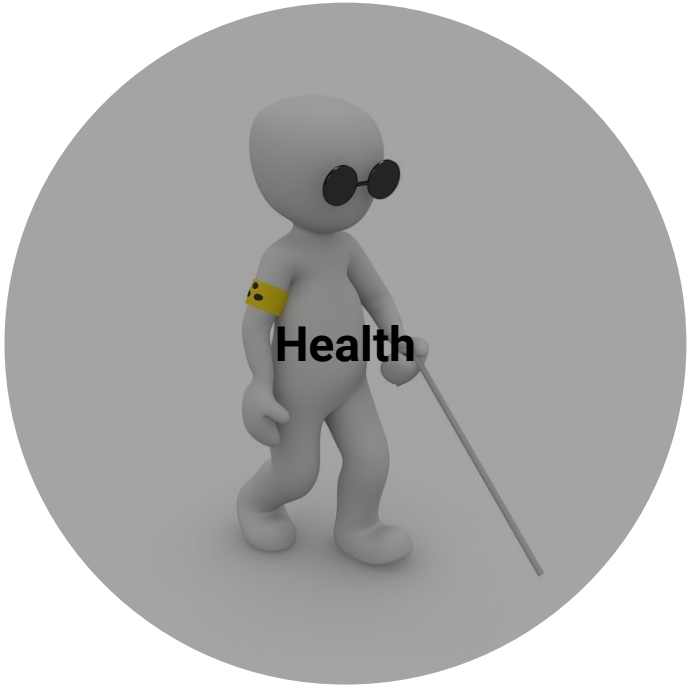
OCR

OCR Applications

OCR Applications – Paperwork



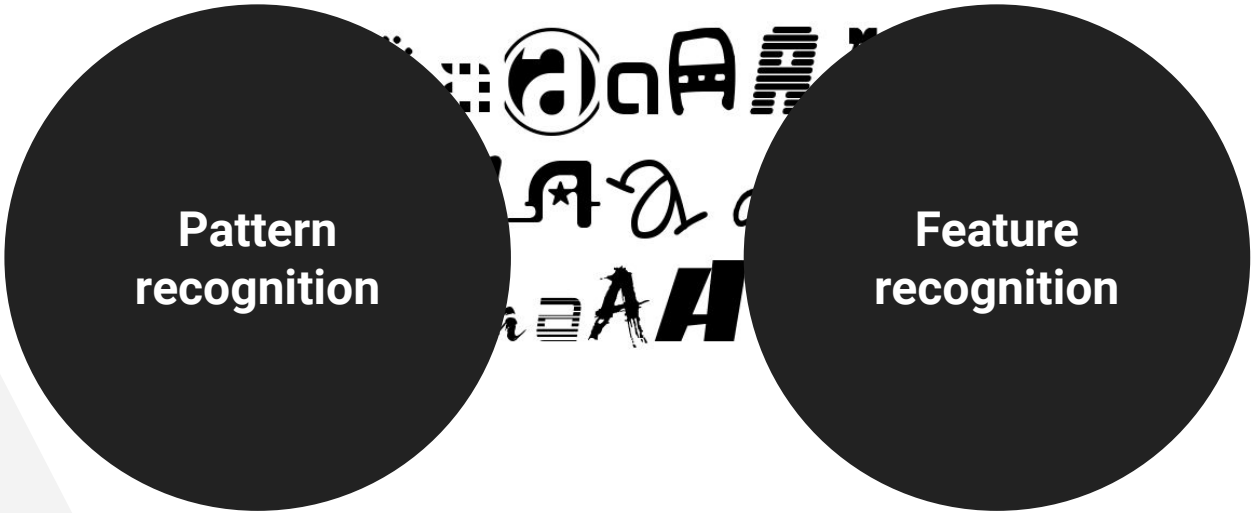
OCR Applications – Health



OCR Applications – Translation

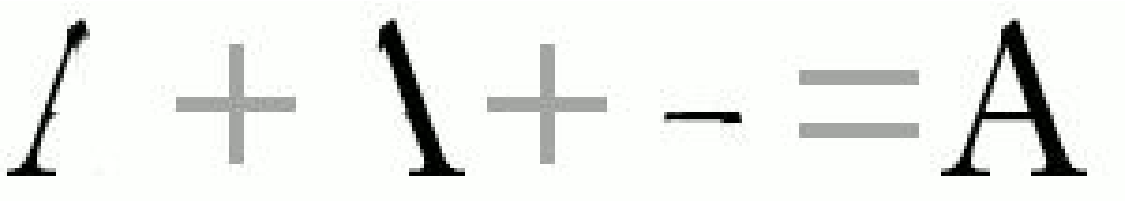
How does OCR work?

How does OCR work?



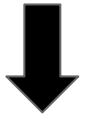
How does OCR work? – Pattern recognition

How does OCR work? - Feature recognition

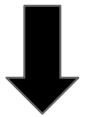


About tesseract

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.



Beautiful|is|better|than|ugly.



B|e|a|u|t|i|f|u|l

Preparing images for OCR

Image processing

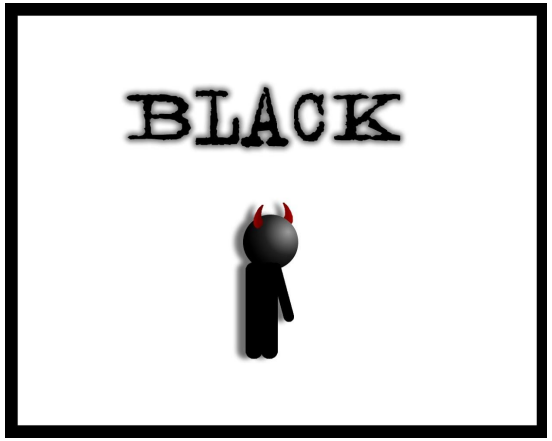
Image processing – Rescaling

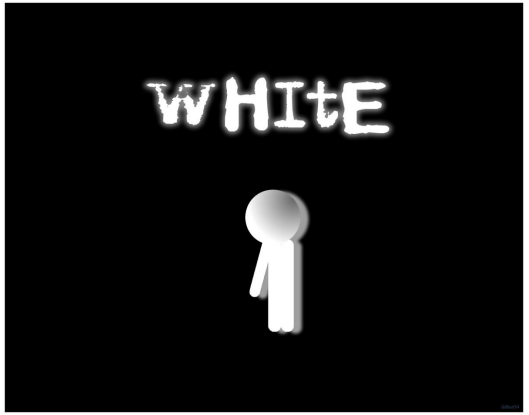
300 PPI

200 PPI

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

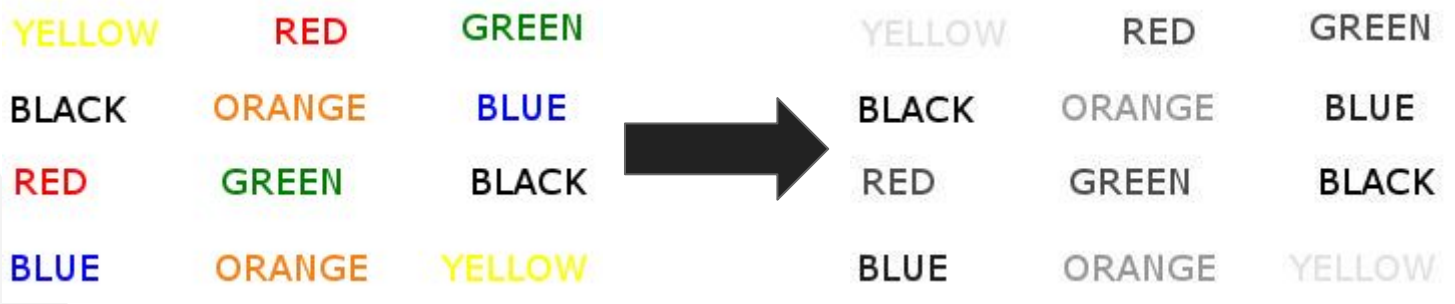
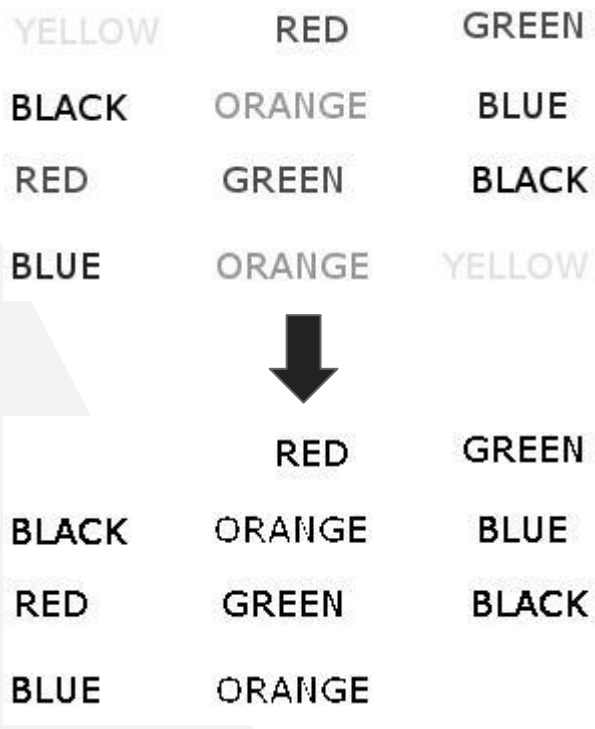


Image processing – Binarization – Thresholding

Less problematic



t=180

More problematic

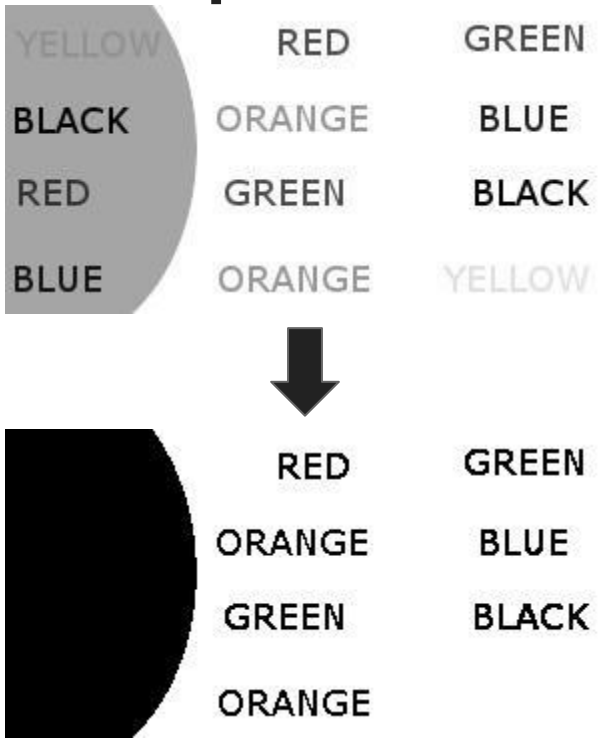
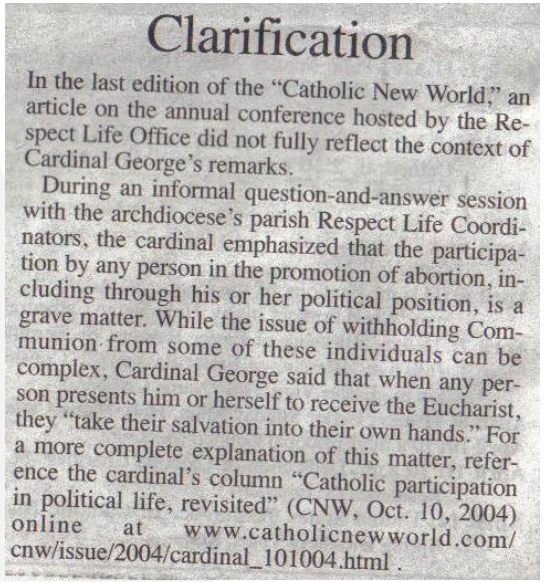


Image processing – Binarization – Adaptive Thresholding

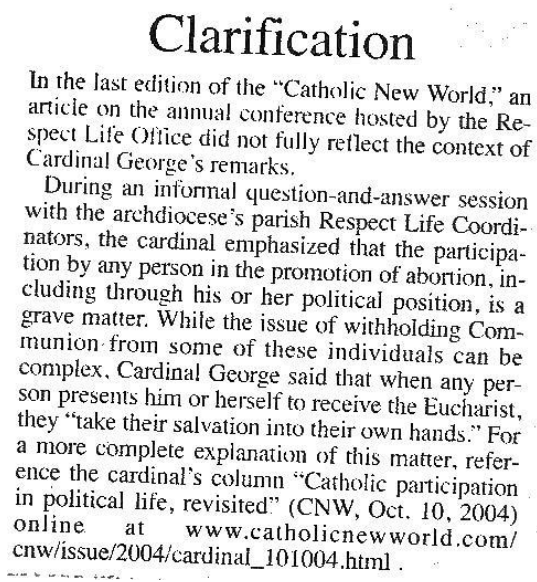
Image processing – Noise Removal



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 .

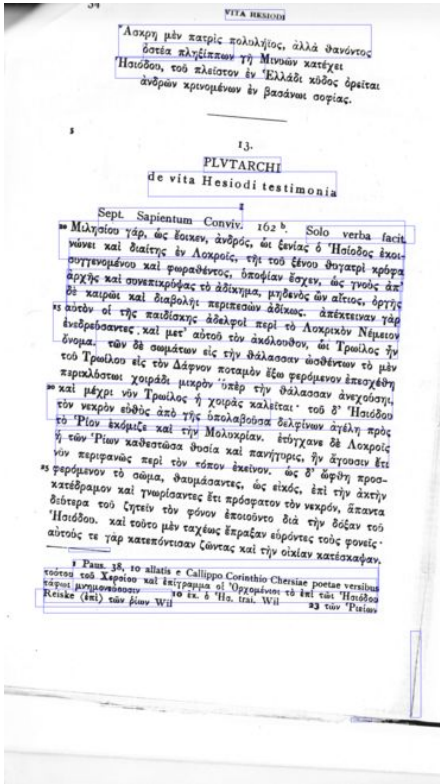


Clarification

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 HESIODI

Ἄσκρη μὲν πατρὶς πολυλήϊος, ἀλλὰ θανόντος
ὀστέα πληξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλείστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνῳ σοφίας.

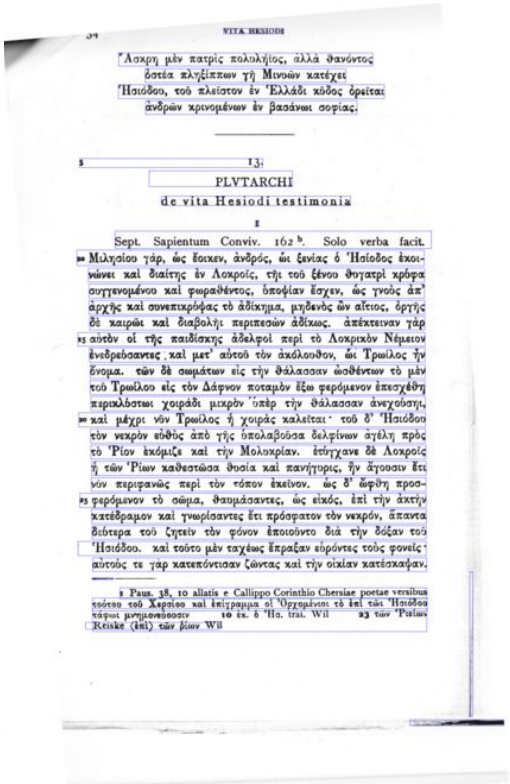
13.

PLVTARCHI
de vita Hesiodi testimonia

Sept. Sapientum Conviv. 162 b. Solo verba facit.

Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὠὶ ξενίας ὁ Ἡσίοδος ἐκοι-
νώνει καὶ διαίτης ἐν Λοκροῖς, τῇ τοῦ ξένου θυγατρὶ κρόφα
συγγενομένου καὶ φωραθέντος, ὑποφίαν ἔσχεν, ὡς γνοὺς ἀπ'
ἀρχῆς καὶ συνεπικρύφας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῇς
δὲ καιρῷ καὶ διαβολῇ περιπεσὼν ἀδίκως. ἀπέκτειναν γάρ
αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον
ἐνεδρεύσαντες καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὠὶ Τρωίλος ἦν
ὄνομα. τῶν δὲ σωμάτων εἰς τὴν θάλασσαν ὠσθέντων τὸ μὲν
τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσχέθη
περικλύστοις χοιράδι μικρὸν ὑπὲρ τὴν θάλασσαν ἀνεχούσῃ,
καὶ μέχρι νῦν Τρωίλος ἡ χοιρὰς καλεῖται· τοῦ δ' Ἡσιόδου
τὸν νεκρὸν εὐθὺς ἀπὸ τῆς ὑπολαβούσης δελφίνων ἀγέλη πρὸς
τὸ Ῥίον ἐκόμιζε καὶ τὴν Μολυκρίαν. ἐτύγχανε δὲ Λοκροῖς
ἡ τῶν Ῥίων καθεστῶσα θυσία καὶ πανήγυρις, ἣν ἄγουσιν ἔτι
νῦν περιφανῶς περὶ τὸν τόπον ἐκεῖνον. ὡς δ' ὤφθη προσ-
φερόμενον τὸ σῶμα, θαυμάσαντες, ὡς εἰκός, ἐπὶ τὴν ἀκτὴν
κατέθραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἅπαντα
δεύτερα τοῦ ζητεῖν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ
Ἡσιόδου. καὶ τοῦτο μὲν ταχέως ἔπραξαν εὑρόντες τοὺς φονεῖς·
αὐτούς τε γὰρ κατεπόντισαν ζῶντας καὶ τὴν οἰκίαν κατέσκαψαν.

1 Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τοῦ Χαρσίου καὶ ἐπίγραμμα οἱ Ὀρχομένιοι τὸ ἐπὶ τῶι Ἡσιόδου τάφωι μνημονεύουσιν 10 ex. ὁ Hs. frat. Wil 23 τῶν Ῥιείων Reiske (ἐπὶ) τῶν ῥίων Wil



VITA HESIODI

Ἄσκρη μὲν πατρὶς πολυλήϊος, ἀλλὰ θανόντος
ὀστέα πληξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλείστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνῳ σοφίας.

13.

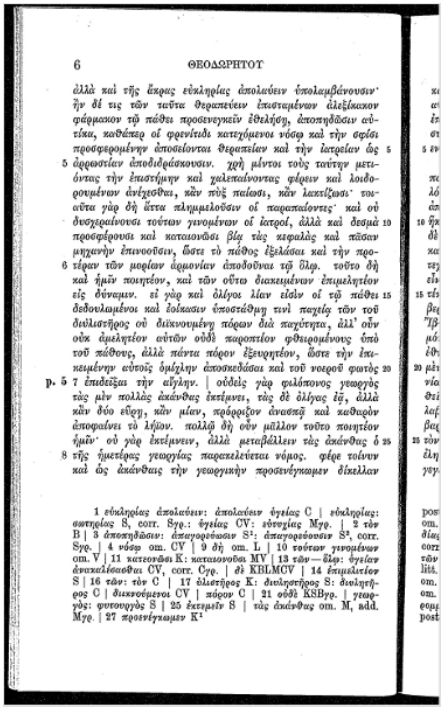
PLVTARCHI
de vita Hesiodi testimonia

Sept. Sapientum Conviv. 162 b. Solo verba facit.

Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὠὶ ξενίας ὁ Ἡσίοδος ἐκοι-
νώνει καὶ διαίτης ἐν Λοκροῖς, τῇ τοῦ ξένου θυγατρὶ κρόφα
συγγενομένου καὶ φωραθέντος, ὑποφίαν ἔσχεν, ὡς γνοὺς ἀπ'
ἀρχῆς καὶ συνεπικρύφας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῇς
δὲ καιρῷ καὶ διαβολῇ περιπεσὼν ἀδίκως. ἀπέκτειναν γάρ
αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον
ἐνεδρεύσαντες καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὠὶ Τρωίλος ἦν
ὄνομα. τῶν δὲ σωμάτων εἰς τὴν θάλασσαν ὠσθέντων τὸ μὲν
τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσχέθη
περικλύστοις χοιράδι μικρὸν ὑπὲρ τὴν θάλασσαν ἀνεχούσῃ,
καὶ μέχρι νῦν Τρωίλος ἡ χοιρὰς καλεῖται· τοῦ δ' Ἡσιόδου
τὸν νεκρὸν εὐθὺς ἀπὸ τῆς ὑπολαβούσης δελφίνων ἀγέλη πρὸς
τὸ Ῥίον ἐκόμιζε καὶ τὴν Μολυκρίαν. ἐτύγχανε δὲ Λοκροῖς
ἡ τῶν Ῥίων καθεστῶσα θυσία καὶ πανήγυρις, ἣν ἄγουσιν ἔτι
νῦν περιφανῶς περὶ τὸν τόπον ἐκεῖνον. ὡς δ' ὤφθη προσ-
φερόμενον τὸ σῶμα, θαυμάσαντες, ὡς εἰκός, ἐπὶ τὴν ἀκτὴν
κατέθραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἅπαντα
δεύτερα τοῦ ζητεῖν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ
Ἡσιόδου. καὶ τοῦτο μὲν ταχέως ἔπραξαν εὑρόντες τοὺς φονεῖς·
αὐτούς τε γὰρ κατεπόντισαν ζῶντας καὶ τὴν οἰκίαν κατέσκαψαν.

1 Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τοῦ Χαρσίου καὶ ἐπίγραμμα οἱ Ὀρχομένιοι τὸ ἐπὶ τῶι Ἡσιόδου τάφωι μνημονεύουσιν 10 ex. ὁ Hs. frat. Wil 23 τῶν Ῥιείων Reiske (ἐπὶ) τῶν ῥίων Wil

Image processing – Border removal



6 ΘΕΟΔΩΡΗΤΟΥ

ἀλλὰ καὶ τῆς ἄκρας εὐκληρίας ἀπολαύειν ὑπολαμβάνουσιν· ἢν δέ τις τῶν ταῦτα θεραπεύειν ἐπισταμένων ἀλεξίκακον φάρμακον τῷ πάθει προσενεγκεῖν ἐθελήσῃ, ἀποπηδῶσιν αὐτίκα, καθάπερ οἱ φρενίτιδι κατεχόμενοι νόσῳ καὶ τὴν σφῶν προσφερομένην ἀποσείονται θεραπείαν καὶ τὴν ἰατρείαν ὡς ἀρρωστίαν ἀποδιδράσκουσιν. χρὴ μέντοι τοὺς τοιούτων μεμισηκότας τὴν ἐπιστήμην καὶ χαλεπαίνοντας φέρειν καὶ λοιδορουμένων ἀνέχεσθαι, κἂν πυῤῥῇ παίσωσι, κἂν λακτίζωσι· τοιαῦτα γὰρ δὴ ἅττα πλημμελοῦσιν οἱ παραπαίοντες· καὶ οὐ δυσχεραίνουσι τούτων γινομένων οἱ ἰατροί, ἀλλὰ καὶ δεσμὰ προσφέρουσι καὶ καταιονῶσι βίᾳ τὰς κεφαλὰς καὶ πᾶσαν μηχανὴν ἐπινοοῦσιν, ὥστε τὸ πάθος ἐξελάσαι καὶ τὴν προτέραν τῶν μορίων ἁρμονίαν ἀποδοῦναι τῷ ὅλῳ. τοῦτο δὴ καὶ ἡμῖν ποιητέον, καὶ τῶν οὕτω διακειμένων ἐπιμελητέον εἰς δύναμιν. εἰ γὰρ καὶ ὀλίγοι λίαν εἰσὶν οἱ τῷ πάθει δεδουλωμένοι καὶ ἐοίκασιν ὑποστάθμῃ τινὶ πηχείᾳ τῶν τοῦ διϋλιστῆρος οὐ διϊκνουμένῃ πόρων διὰ παχύτητα, ἀλλ' οὖν οὐκ ἀμελητέον αὐτῶν οὐδὲ παροπτέον φθειρομένους ὑπὸ τοῦ πάθους, ἀλλὰ πάντα πόρον ἐξευρητέον, ὥστε τὴν ἐπικειμένην αὐτοῖς ὁμίχλην ἀποσκεδάσαι καὶ τοῦ νοεροῦ φωτὸς
p. 5 7 ἐπιδεῖξαι τὴν αἴγλην. | οὐδεὶς γὰρ φιλόπονος γεωργὸς τὰς μὲν πολλὰς ἀκάνθας ἐκτέμνει, τὰς δὲ ὀλίγας ἐᾷ, ἀλλὰ κἂν δύο εὕρῃ, κἂν μίαν, πρόῤῥιζον ἀνασπᾷ καὶ καθαρὸν ἀποφαίνει τὸ λήϊον. πολλῷ δὴ οὖν μᾶλλον τοῦτο ποιητέον ἡμῖν· οὐ γὰρ ἐκτέμνειν, ἀλλὰ μεταβάλλειν τὰς ἀκάνθας ὁ
8 τῆς ἡμετέρας γεωργίας παρακελεύεται νόμος. φέρε τοίνυν καὶ ὡς ἀκάνθας τὴν γεωργικὴν προσενέγκωμεν δίκελλαν

1 εὐκληρίας ἀπολαύειν: ἀπολαύειν ὑγείας C | εὐκληρίας: σωτηρίας S, corr. Sγρ.: ὑγείας CV: εὐσχείας Μγρ. | 2 τὸν B | 3 ἀποσκηδῶσιν: ἀπαγορεύουσιν S²: ἀπαγορεύουσιν S³, corr. Sγρ. | 4 νόσῳ om. CV | 9 δὴ om. L | 10 τούτων γινομένων om. V | 11 καταιονῶσι K: καταιονοῦσι MV | 13 τῶν—ὅλῳ: ὑγείαν ἀναλαβόσθαι CV, corr. Cγρ. | δὴ KBLMCV | 14 ἐπιμελητέον S | 16 τῶν: τὸν C | 17 διϋλιστῆρος K: διυλιστῆρος S: διυλιστῆρος C | διϊκνούμενοι CV | πόρον C | 21 οὐδὲ KSBγρ. | γεωργὸς: φυτουργὸς S | 25 ἐκτεμεῖν S | τὰς ἀκάνθας om. M, add. Μγρ. | 27 προσενέγκωμεν K¹

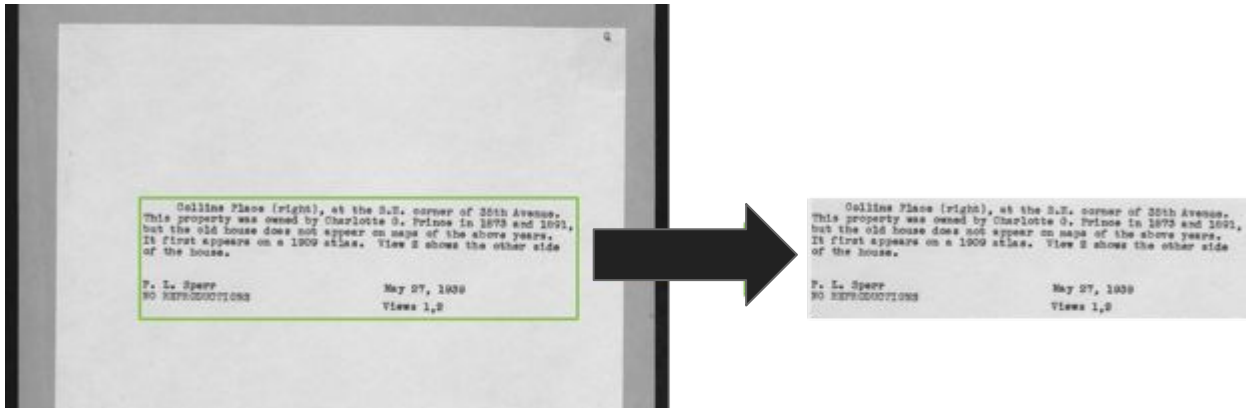
6 ΘΕΟΔΩΡΗΤΟΥ

ἀλλὰ καὶ τῆς ἄκρας εὐκληρίας ἀπολαύειν ὑπολαμβάνουσιν· ἢν δέ τις τῶν ταῦτα θεραπεύειν ἐπισταμένων ἀλεξίκακον φάρμακον τῷ πάθει προσενεγκεῖν ἐθελήσῃ, ἀποπηδῶσιν αὐτίκα, καθάπερ οἱ φρενίτιδι κατεχόμενοι νόσῳ καὶ τὴν σφῶν προσφερομένην ἀποσείονται θεραπείαν καὶ τὴν ἰατρείαν ὡς ἀρρωστίαν ἀποδιδράσκουσιν. χρὴ μέντοι τοὺς τοιούτων μεμισηκότας τὴν ἐπιστήμην καὶ χαλεπαίνοντας φέρειν καὶ λοιδορουμένων ἀνέχεσθαι, κἂν πυῤῥῇ παίσωσι, κἂν λακτίζωσι· τοιαῦτα γὰρ δὴ ἅττα πλημμελοῦσιν οἱ παραπαίοντες· καὶ οὐ δυσχεραίνουσι τούτων γινομένων οἱ ἰατροί, ἀλλὰ καὶ δεσμὰ προσφέρουσι καὶ καταιονῶσι βίᾳ τὰς κεφαλὰς καὶ πᾶσαν μηχανὴν ἐπινοοῦσιν, ὥστε τὸ πάθος ἐξελάσαι καὶ τὴν προτέραν τῶν μορίων ἁρμονίαν ἀποδοῦναι τῷ ὅλῳ. τοῦτο δὴ καὶ ἡμῖν ποιητέον, καὶ τῶν οὕτω διακειμένων ἐπιμελητέον εἰς δύναμιν. εἰ γὰρ καὶ ὀλίγοι λίαν εἰσὶν οἱ τῷ πάθει δεδουλωμένοι καὶ ἐοίκασιν ὑποστάθμῃ τινὶ πηχείᾳ τῶν τοῦ διϋλιστῆρος οὐ διϊκνουμένῃ πόρων διὰ παχύτητα, ἀλλ' οὖν οὐκ ἀμελητέον αὐτῶν οὐδὲ παροπτέον φθειρομένους ὑπὸ τοῦ πάθους, ἀλλὰ πάντα πόρον ἐξευρητέον, ὥστε τὴν ἐπικειμένην αὐτοῖς ὁμίχλην ἀποσκεδάσαι καὶ τοῦ νοεροῦ φωτὸς 7 ἐπιδεῖξαι τὴν αἴγλην. | οὐδεὶς γὰρ φιλόπονος γεωργὸς τὰς μὲν πολλὰς ἀκάνθας ἐκτέμνει, τὰς δὲ ὀλίγας ἐᾷ, ἀλλὰ κἂν δύο εὕρῃ, κἂν μίαν, πρόῤῥιζον ἀνασπᾷ καὶ καθαρὸν ἀποφαίνει τὸ λήϊον. πολλῷ δὴ οὖν μᾶλλον τοῦτο ποιητέον ἡμῖν· οὐ γὰρ ἐκτέμνειν, ἀλλὰ μεταβάλλειν τὰς ἀκάνθας ὁ τῆς ἡμετέρας γεωργίας παρακελεύεται νόμος. φέρε τοίνυν καὶ ὡς ἀκάνθας τὴν γεωργικὴν προσενέγκωμεν δίκελλαν

1 εὐκληρίας ἀπολαύειν: ἀπολαύειν ὑγείας C | εὐκληρίας: σωτηρίας S, corr. Sγρ.: ὑγείας CV: εὐσχείας Μγρ. | 2 τὸν B | 3 ἀποσκηδῶσιν: ἀπαγορεύουσιν S²: ἀπαγορεύουσιν S³, corr. Sγρ. | 4 νόσῳ om. CV | 9 δὴ om. L | 10 τούτων γινομένων om. V | 11 καταιονῶσι K: καταιονοῦσι MV | 13 τῶν—ὅλῳ: ὑγείαν ἀναλαβόσθαι CV, corr. Cγρ. | δὴ KBLMCV | 14 ἐπιμελητέον S | 16 τῶν: τὸν C | 17 διϋλιστῆρος K: διυλιστῆρος S: διυλιστῆρος C | διϊκνούμενοι CV | πόρον C | 21 οὐδὲ KSBγρ. | γεωργὸς: φυτουργὸς S | 25 ἐκτεμεῖν S | τὰς ἀκάνθας om. M, add. Μγρ. | 27 προσενέγκωμεν K¹

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:

0 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_morphology.py
- http://static.googleusercontent.com/media/research.google.com/ro//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