Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

A High performance OCR System For Medieval Documents

(e.g. "Narrenschiffe" Novels of 15th Century)

Dr. Saqib Bukhari & Prof. Dr. Andreas Dengel

What we were cable of before the start of this project?

What we were cable of before the start of this project?

We have been developing 

for a decade

What we were cable of before the start of this project?

We have been developing OCROPUS for a decade

Google Sponsored Project to DFK

"A project to develop advanced OCR technologies in DFK"I, Google Code Blog, 2007

Complete OCR Work-Flow for Document Analysis

Breuel: *The OCRopus Open Source OCR System*. Proceedings SPIE 20th Annual Symposium 2008

Preprocessing, Layout Analysis, OCR, ...

- Shafait, Keysers, Breuel: Performance Evaluation and Benchmarking of Six-page Segmentation Algorithms. IEEE TPAMI, 2008.
- Bukhari, Shafait, Breuel: High Performance Layout Analysis of Arabic and Urdu Document Images. IEEE ICPR, 2011.

What we were cable of before the start of this project?

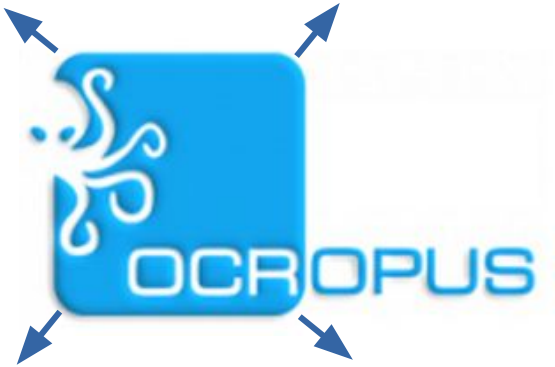
We have been developing OCROPUS for a decade

Google Sponsored Project to DFK

"A project to develop advanced OCR technologies in DFK"I, Google Code Blog, 2007

Complete OCR Work-Flow for Document Analysis

Breuel: *The OCRopus Open Source OCR System*. Proceedings SPIE 20th Annual Symposium 2008



Preprocessing, Layout Analysis, OCR, ...

- Shafait, Keysers, Breuel: Performance Evaluation and Benchmarking of Six-page Segmentation Algorithms. IEEE TPAMI, 2008.
- Bukhari, Shafait, Breuel: High Performance Layout Analysis of Arabic and Urdu Document Images. IEEE ICPR, 2011.

OCR Processing for Historical Documents

Breuel, Ul-Hasan, Azawi, Shafait: High Performance OCR for Printed English and Fraktur using LSTM Networks. IEEE ICDAR, 2013.

What we were cable of before the start of this project?

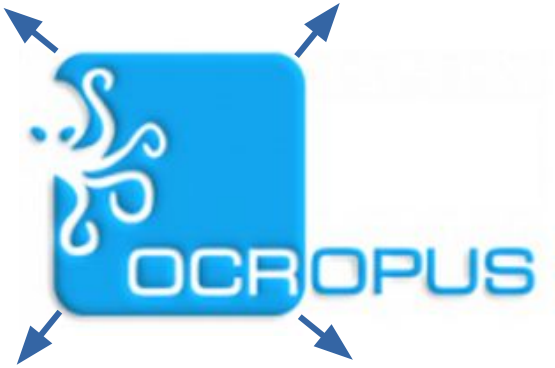
We have been developing OCROPUS for a decade

Google Sponsored Project to DFK

"A project to develop advanced OCR technologies in DFK"I, Google Code Blog, 2007

Complete OCR Work-Flow for Document Analysis

Breuel: *The OCRopus Open Source OCR System*. Proceedings SPIE 20th Annual Symposium 2008



Preprocessing, Layout Analysis, OCR, ...

- Shafait, Keysers, Breuel: Performance Evaluation and Benchmarking of Six-page Segmentation Algorithms. IEEE TPAMI, 2008.
- Bukhari, Shafait, Breuel: High Performance Layout Analysis of Arabic and Urdu Document Images. IEEE ICPR, 2011.

OCR Processing for Historical Documents

Breuel, Ul-Hasan, Azawi, Shafait: High Performance OCR for Printed English and Fraktur using LSTM Networks. IEEE ICDAR, 2013.

So far six PhD theses have been completed on OCRopus, and three are on the way!

What we were cable of before the start of this project?

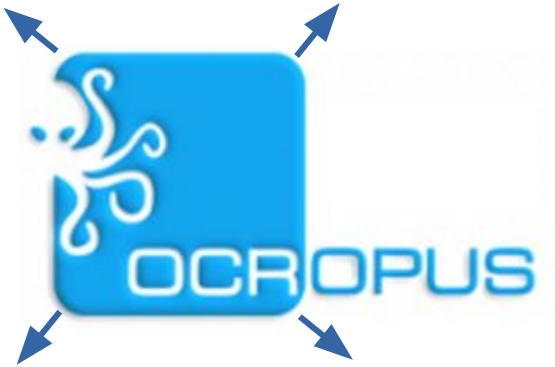
We have been developing OCROPUS for a decade

Google Sponsored Project to DFK

"A project to develop advanced OCR technologies in DFK"I, Google Code Blog, 2007

Complete OCR Work-Flow for Document Analysis

Breuel: *The OCRopus Open Source OCR System*. Proceedings SPIE 20th Annual Symposium 2008



Ocrosic

Springmann: Ocrocis, a project manager for Ocropus, Ludwig-Maximilians-University, Munich, 2015.

Preprocessing, Layout Analysis, OCR, ...

- Shafait, Keysers, Breuel: Performance Evaluation and Benchmarking of Six-page Segmentation Algorithms. IEEE TPAMI, 2008.
- Bukhari, Shafait, Breuel: High Performance Layout Analysis of Arabic and Urdu Document Images. IEEE ICPR, 2011.

OCR Processing for Historical Documents

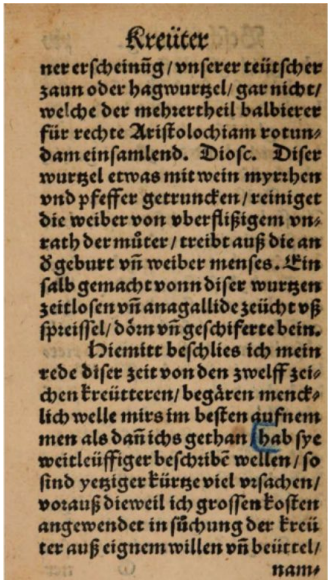
Breuel, Ul-Hasan, Azawi, Shafait: High Performance OCR for Printed English and Fraktur using LSTM Networks. IEEE ICDAR, 2013.

So far six PhD theses have been completed on OCRopus, and three are on the way!

What we were cable of before the start of this project?

An Example: The OCROPUS Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]



Kreüter

ner erſcheinūg / vnſerer teütſcher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Ariſtolochiam rotundam einſamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberflüßigem vnrath der mūter / treibt auß die andgeburt vñ weiber menſes. Ein ſalb gemacht vonn diſer wurtzen zeitloſen vñ anagallide zeücht vß ſpreiſſel / dörn vñ geſchifferte bein.

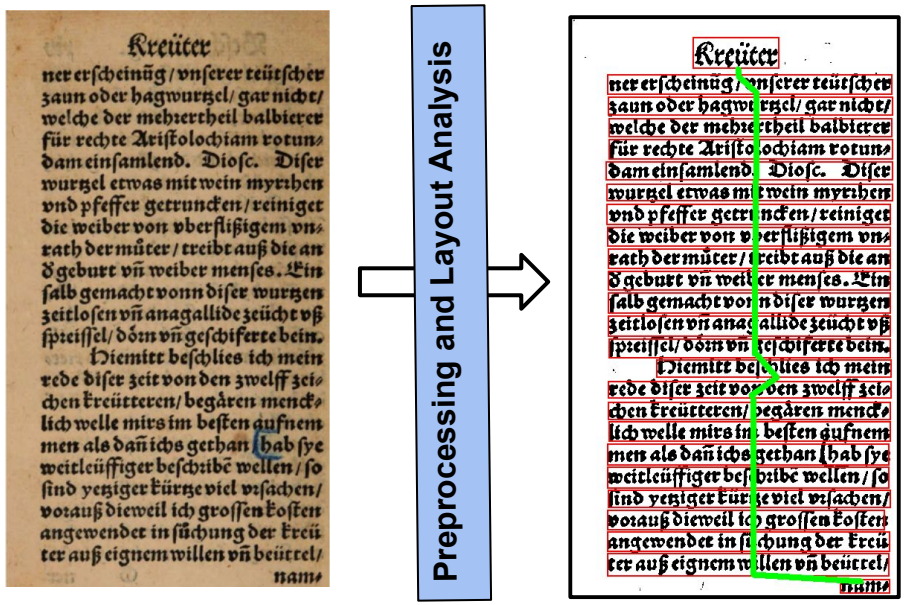
Hiemitt beſchlies ich mein rede diſer zeit von den zwelff zeichen kreütteren / begären menckclich welle mirs im beſten aufnemmen als dañ ichs gethan hab ſye weitleüffiger beſchribē wellen / ſo ſind yetziger kürtze viel vrſachen / vorauß dieweil ich groſſen koſten angewender in ſüchung der kreüter auß eignem willen vñ beüttel / nam-

[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

What we were cable of before the start of this project?

An Example: The OCROPUS Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]

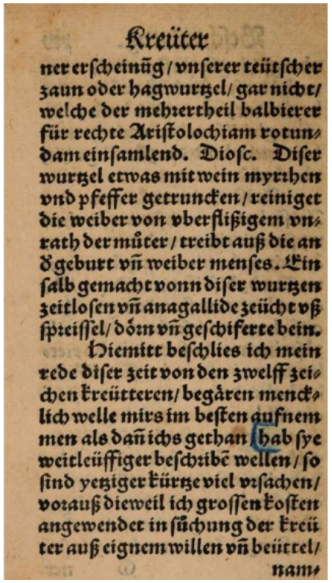


[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

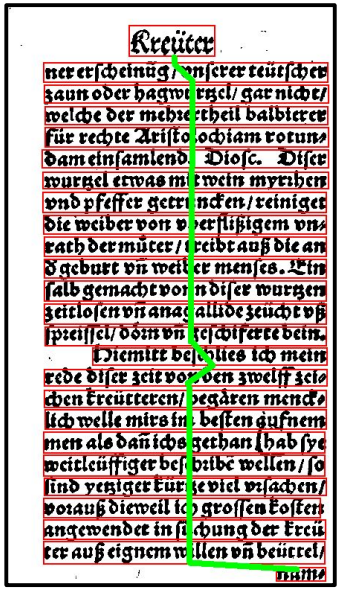
What we were cable of before the start of this project?

An Example: The OCROPUS Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]



Preprocessing and Layout Analysis



Trained OCR Line Recognizer

OCRopus Results [1]

Kreüter
ner erfcheinüg/ vnferer teütfcher
zaun oder hagwurtzel/ gar micht/
welche der mehrertheil balbierer
für rechte Ariftolochiam rotun-
dam einfamlend. Dioſc. Difer
wurtzel etwas mit wein myrrhen
vnd pfeffer getruncken/ reiniget
die weiber von vberflifzigem vn-
rath der můter/ treibt aufz die an
d geburt vñ weiber menſes. Ein
falb gemacht vonn difer wurtzen
zeitlofen vñ anagallide zeücht vfz
fpreiffel/ dörnvñ gefchiferte bein.
Hiemitt befchlies ich mein
rede difer zeit von den zwelff zei-
chen kreütteren/ begären menck-
lich welle mirs im beften aufnem
men als dañ ichs gethan/ hab fye
weitleüffiger befchribē wellen/ fo
find yetziger kürtze viel vrfach_en/
voraufz dieweil ich groffen koften
angewendet in fůchung der kreü
ter aufz eignem willen vñ beüt_tel/
ncn-

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

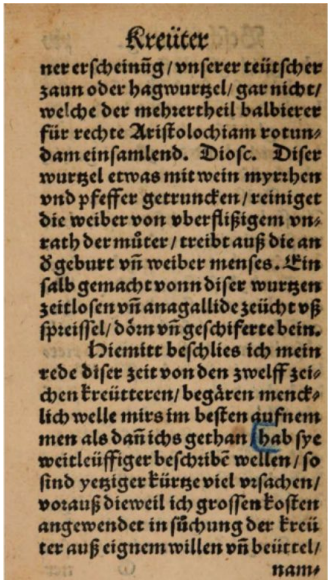
[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

What we were cable of before the start of this project?

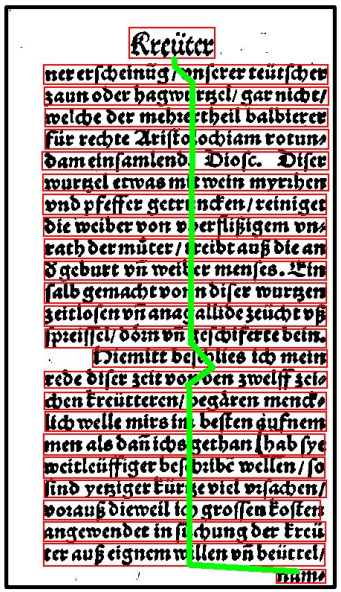
An Example: The OCROPUS Work-Flow For 15th Century Historical Documents

On Test images from Adam von Bodenstein (1557), the performance:
Ocropus 99%, **ABBYY 85%** and and **Tesseract 78%** [1].

Adam von Bodenstein (1557) [1]



Preprocessing and Layout Analysis



Trained OCR Line Recognizer

OCRopus Results [1]

Kreüter
ner erfcheinüg/ vnferer teütfcher
zaun oder hagwurtzel/ gar micht/
welche der mehrertheil balbierer
für rechte Ariftolochiam rotun-
dam einfamlend. Diofc. Difer
wurtzel etwas mit wein myrrhen
vnd pfeffer getruncken/ reiniget
die weiber von vberflüfzigem vn-
rath der mûter/ treibt aufz die an
d geburt vñ weiber menfes. Ein
falb gemacht vonn difer wurtzen
zeitlofen vñ anagallide zeücht vfz
fpreiffel/ dörnvñ gefchiferte bein.
Hiemitt befchlies ich mein
rede difer zeit von den zwelff zei-
chen kreütteren/ begären menck-
lich welle mirs im beften aufnem
men als dañ ichs gethan/ hab fye
weitleüffiger befchribē wellen/ fo
find yetziger kürtze viel vrfach_en/
voraufz dieweil ich groffen koften
angewendet in fûchung der kreü
ter aufz eignem willen vñ beüt_tel/
ncn-

ABBYY Results [1]

Kreücev
ner erscheinng / vnserer teutscher
zäun oderhagwuryel/ garnicht/
welche der mehrertheil baldierer
fiir rechte Aristolochiam rotun-
dameinsamlend. Diosc. Diser
würget etwas mit wein Myrrhen
vnd Pfeffer getrvncken/ reiniget
die weider von vderssissigem vm
rath dermäter/ treibt auß die an
Lgedurt vn weider menses. Ein
faldgemachtvonndiser würgen
zeitlosenvnanagallidezeüchtvß
spreissel/ dörnvngeschiferte dein.
iZiemitt deschlies ich mein
rede diser zeitvonden zwelffzei-
chenkreütteren/degären menck-
ltch welle mirs im dessen aufnem
men als dan ichs gethan / had fye
weitleüfftgerdeschridc wellen / f»
sind yegigerkürgeviel vrsachen/
voraußdieweil ichgrosicnkosten
angewendet insüchungder kreü
ter auß eignem willen vndeücrcl/

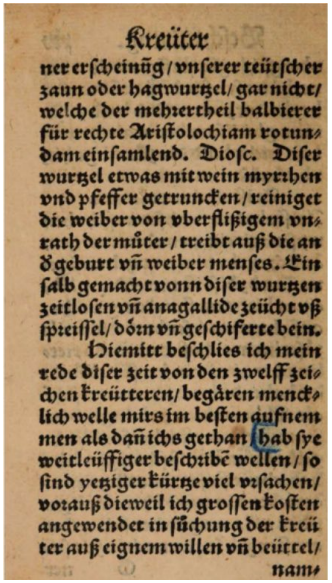
[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

What we were cable of before the start of this project?

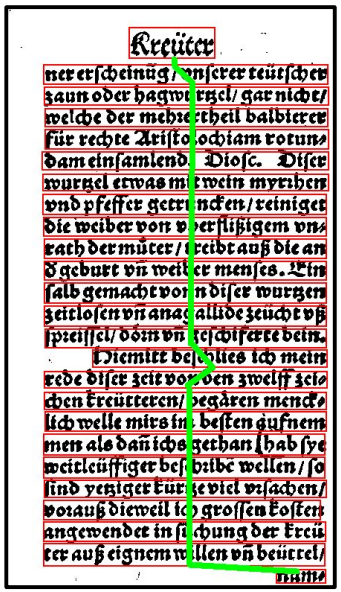
An Example: The OCROPUS Work-Flow For 15th Century Historical Documents

On Test images from Adam von Bodenstein (1557), the performance:
Ocropus 99%, **ABBYY 85%** and and **Tesseract 78%** [1].

Adam von Bodenstein (1557) [1]



Preprocessing and Layout Analysis



Trained OCR Line Recognizer

OCRopus Results [1]

Kreüter
ner erſcheinüg/ vnſerer teütſcher
zaun oder hagwurtzel/ gar micht/
welche der mehrertheil balbierer
für rechte Ariſtolochiam rotun-
dam einſamlend. Dioſc. Difer
wurtzel etwas mit wein myrrhen
vnd pfeffer getruncken/ reiniget
die weiber von vberflifzigem vn-
rath der mûter/ treibt aufz die an
d geburt vñ weiber menſes. Ein
falb gemacht vonn difer wurtzen
zeitlofen vñ anagallide zeücht vfz
ſpreiffel/ dörnvñ gefchiferte bein.
Hiemitt befchlies ich mein
rede difer zeit von den zwelff zei-
chen kreütteren/ begären menck-
lich welle mirs im beften aufnem
men als dañ ichs gethan/ hab fye
weitleüffiger befchribē wellen/ fo
find yetziger kürtze viel vrfach_en/
voraufz dieweil ich groffen koften
angewendet in fûchung der kreü
ter aufz eignem willen vñ beüt_tel/
ncn-

ABBYY Results [1]

Kreücev
ner erscheinng / vnserer teutscher
zäun oderhagwuryel/ garnicht/
welche der mehrertheil baldierer
fiir rechte Aristolochiam rotun-
dameinsamlend. Diosc. Diser
würget etwas mit wein Myrrhen
vnd Pfeffer getrvncken/ reiniget
die weider von vderssissigem vm
rath dermäter/ treibt auß die an
Lgedurt vn weider menses. Ein
faldgemachtvonndiser würgen
zeitlosenvnanagallidezeüchtvß
spreissel/ dörnvngeschiferte dein.
iZiemitt deschlies ich mein
rede diser zeitvonden zwelffzei-
chenkreütteren/degären menck-
ltch welle mirs im dessen aufnem
men als dan ichs gethan / had fye
weitleüfftgerdeschridc wellen / f»
sind yegigerkürgeviel vrsachen/
voraußdieweil ichgrosicnkosten
angewendet insüchungder kreü
ter auß eignem willen vndeücrcl/

so what else is required for the 15th Century “Narrenschiffe” novels of this project?

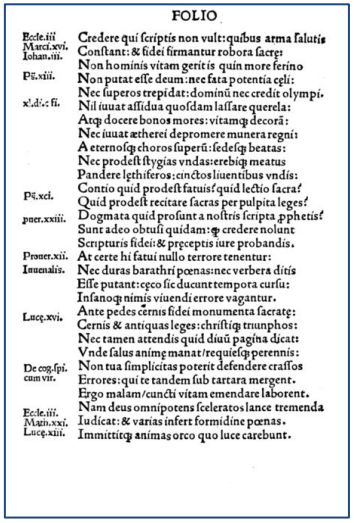
What are the limitations of the OCRopus?

What are the limitations of the OCRopus?

Layout Analysis

What are the limitations of the OCRopus?

Layout Analysis



FOLIO

Ecclesiastici
Marci.xvi.
Iohan.iii.
Credere qui scriptis non vult: quibus arma salutis
Constant: & fidei firmantur robora sacre:
Non hominis vitam gerit is quin more ferino
Ps.xiii.
Non putat esse deum: nec fata potentia celi:
Nec superos trepidat: dominū nec credit olympi.
xl.dic fi.
Nil iuuat assidua quosdam lassare querela:
Atq; docere bonos mores: vitamq; decorā:
Nec iuuat ætherei depromere munera regni:
A eternosq; choros superū: sedesq; beatas:
Nec prodest stygias vndas: erebiq; meatus
Pandere lgthiferos: cinctos liuentibus vndis:
Ps.xci.
Contio quid prodest fatuis? quid lectio sacra?
Quid prodest recitare sacras per pulpita leges?
puer.xxiii.
Dogmata quid prosunt a nostris scripta pphetis?
Sunt adeo obtusi quidam: & credere nolunt
Scripturis fidei: & preceptis iure probandis.
Prouer.xii.
At certe hi fatui nullo terrore tenentur:
Iuuenalis.
Nec duras barathri pœnas: nec verbera ditis
Esse putant: ceco sic ducunt tempora cursu:
Insanoq; nimis viuendi errore vagantur.
Lucę.xvi.
Ante pedes cernis fidei monumenta sacratę:
Cernis & antiquas leges: christiq; triunphos:
Nec tamen attendis quid diuū pagina dicat:
Vnde salus animę manat/requiesq; perennis:
De cog.spi. cum vir.
Non tua simplicitas poterit defendere crassos
Errores: qui te tandem sub tartara mergent.
Ergo malam/cuncti vitam emendare laborent.
Nam deus omnipotens sceleratos lance tremenda
Ecclesiastici.
Matth.xxi.
Lucę.xiii.
Iudicat: & varias infert formidine pœnas.
Immittitq; animas orco quo luce carebunt.

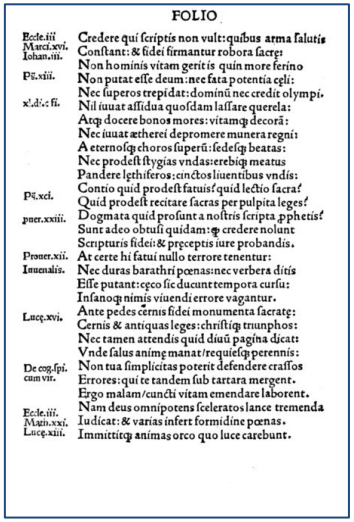
FOLIO

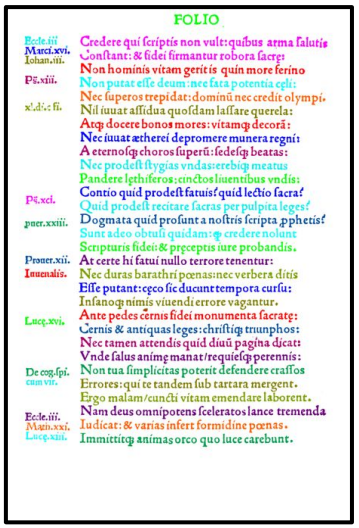
Ecclesiastici
Marci.xvi.
Iohan.iii.
Credere qui scriptis non vult: quibus arma salutis
Constant: & fidei firmantur robora sacre:
Non hominis vitam gerit is quin more ferino
Ps.xiii.
Non putat esse deum: nec fata potentia celi:
Nec superos trepidat: dominū nec credit olympi.
xl.dic fi.
Nil iuuat assidua quosdam lassare querela:
Atq; docere bonos mores: vitamq; decorā:
Nec iuuat ætherei depromere munera regni:
A eternosq; choros superū: sedesq; beatas:
Nec prodest stygias vndas: erebiq; meatus
Pandere lgthiferos: cinctos liuentibus vndis:
Ps.xci.
Contio quid prodest fatuis? quid lectio sacra?
Quid prodest recitare sacras per pulpita leges?
puer.xxiii.
Dogmata quid prosunt a nostris scripta pphetis?
Sunt adeo obtusi quidam: & credere nolunt
Scripturis fidei: & preceptis iure probandis.
Prouer.xii.
At certe hi fatui nullo terrore tenentur:
Iuuenalis.
Nec duras barathri pœnas: nec verbera ditis
Esse putant: ceco sic ducunt tempora cursu:
Insanoq; nimis viuendi errore vagantur.
Lucę.xvi.
Ante pedes cernis fidei monumenta sacratę:
Cernis & antiquas leges: christiq; triunphos:
Nec tamen attendis quid diuū pagina dicat:
Vnde salus animę manat/requiesq; perennis:
De cog.spi. cum vir.
Non tua simplicitas poterit defendere crassos
Errores: qui te tandem sub tartara mergent.
Ergo malam/cuncti vitam emendare laborent.
Nam deus omnipotens sceleratos lance tremenda
Ecclesiastici.
Matth.xxi.
Lucę.xiii.
Iudicat: & varias infert formidine pœnas.
Immittitq; animas orco quo luce carebunt.

OCRopus 100%

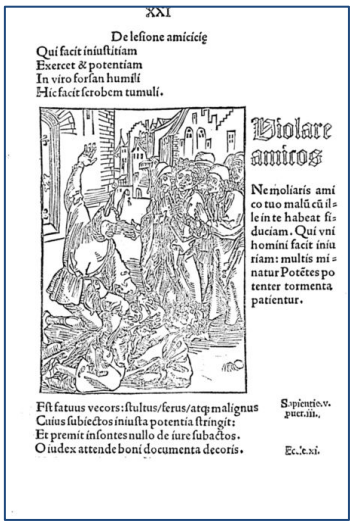
What are the limitations of the OCRopus?

Layout Analysis





OCRopus 100%

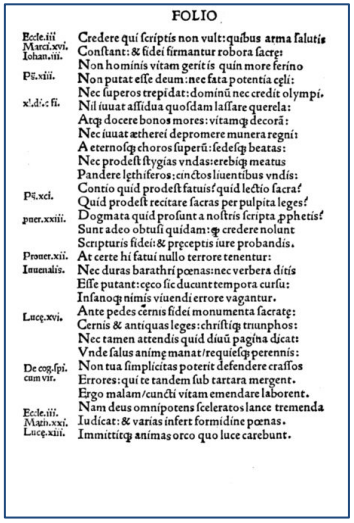


OCRopus 75%

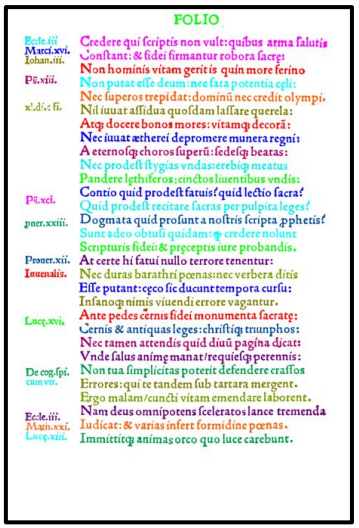
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

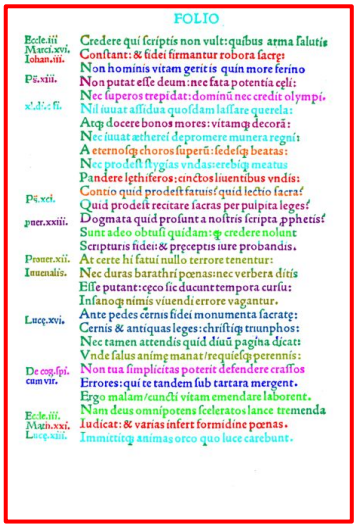
What are the limitations of the OCRopus?

Layout Analysis

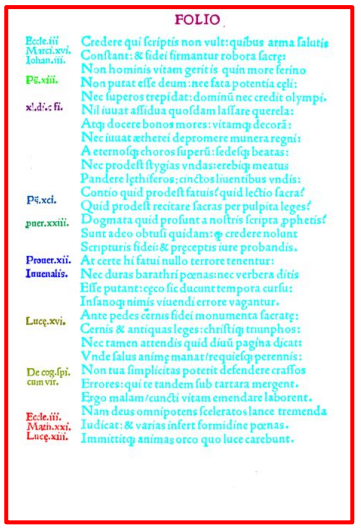


OCRopus 100%

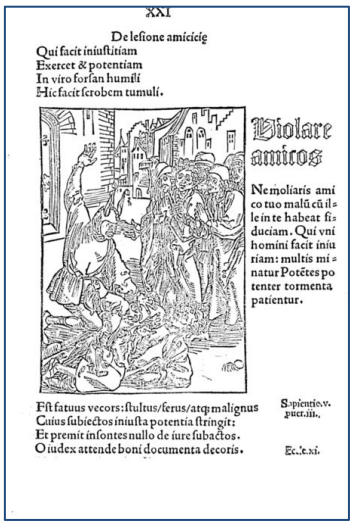




XY Cut < 5%

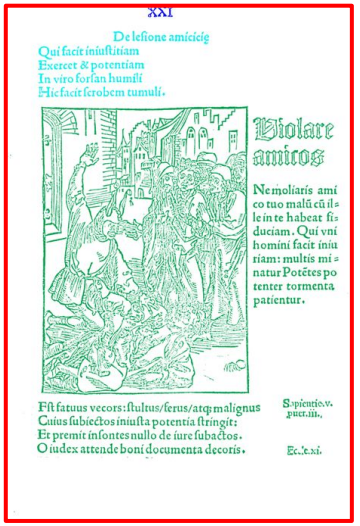


Voronoi < 5%



OCRopus 75%

XY Cut < 5%



Voronoi < 5%

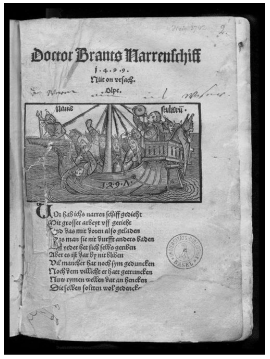
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

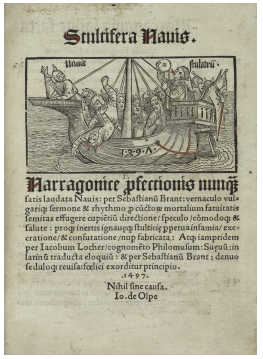
Dr. Syed Saqib Bukhari

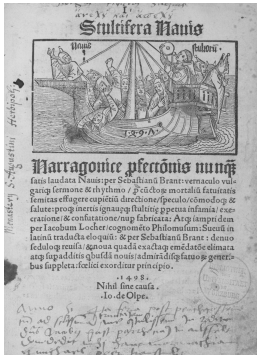
What are the limitations of the OCRopus?

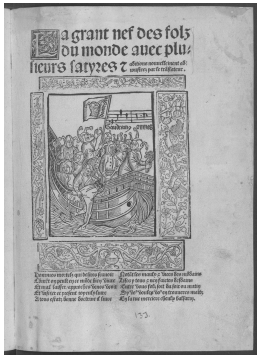
Layout Analysis

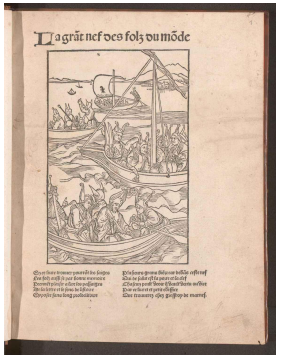
What are we dealing in this project?

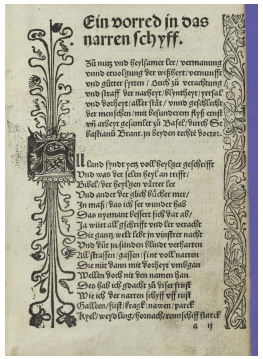


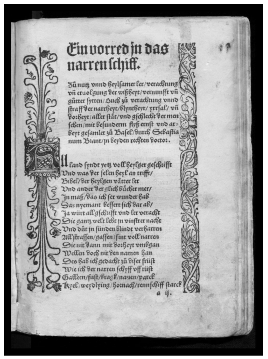


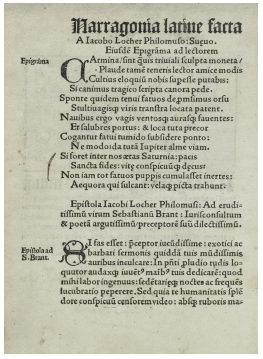


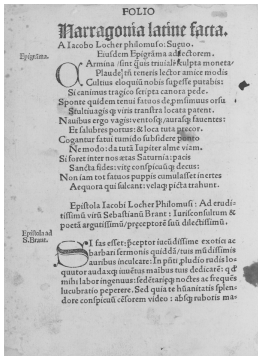


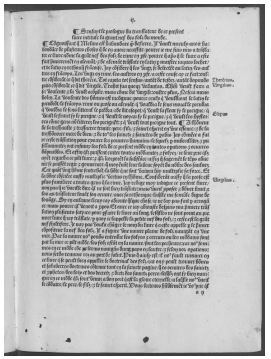


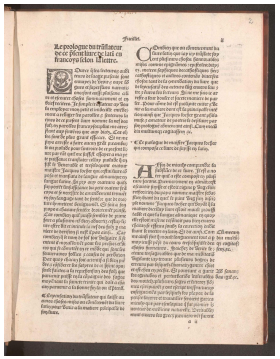












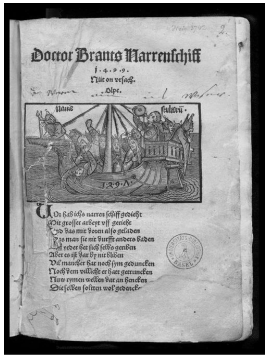
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

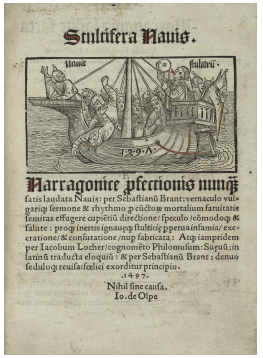
Dr. Syed Saqib Bukhari

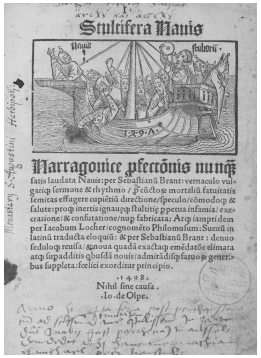
What are the limitations of the OCRopus?

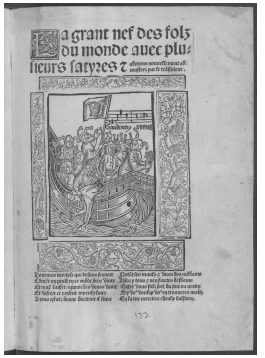
Layout Analysis

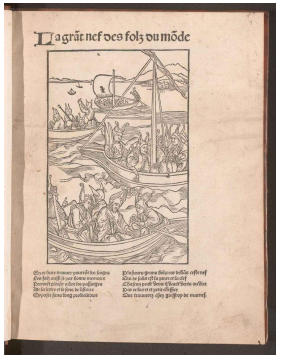
What are we dealing in this project?

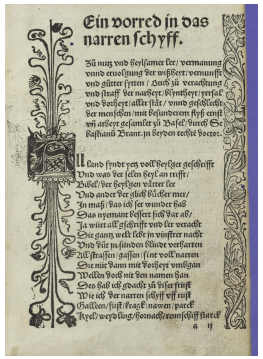


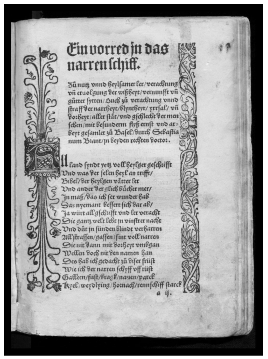


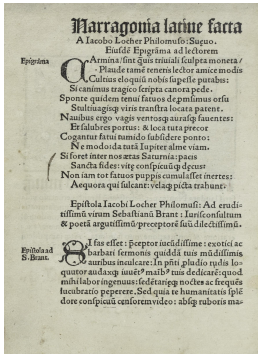


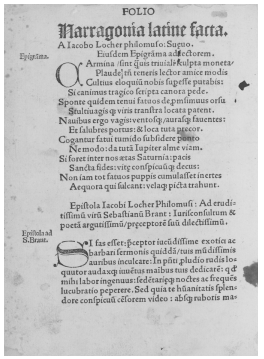


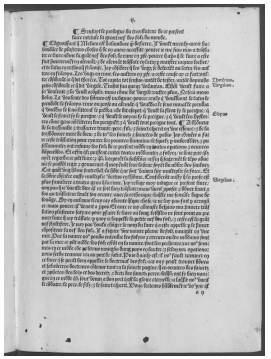
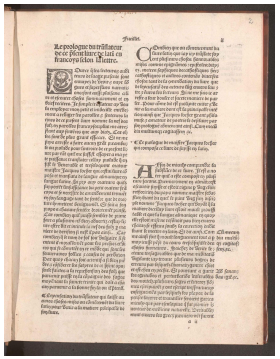










so what else is required for the 15th Century “Narrenschiffe” novels of this project?

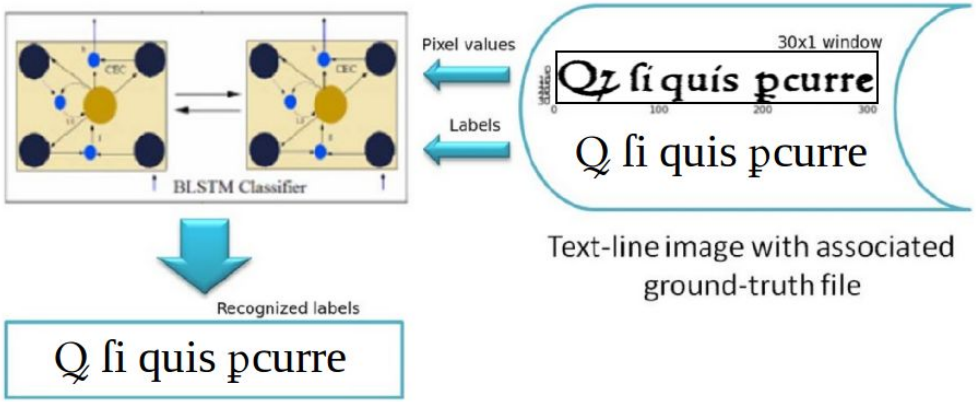
(i) Advanced Layout Analysis

What are the limitations of the OCRopus?

OCR Model

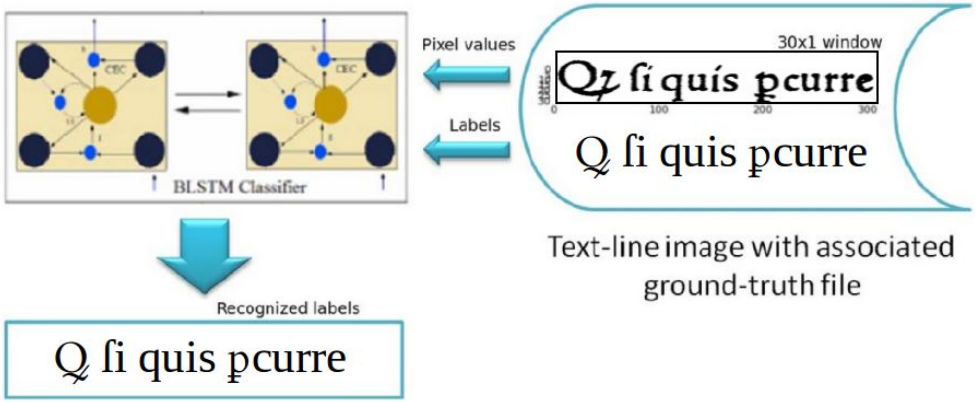
What are the limitations of the OCRopus?

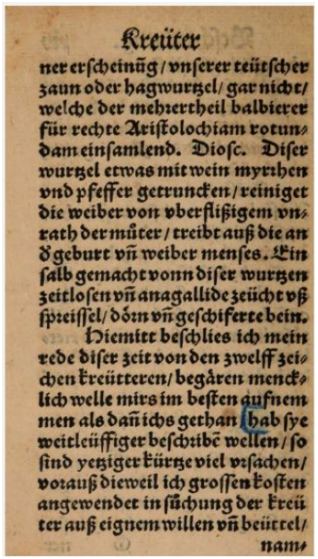
OCR Model



What are the limitations of the OCRopus?

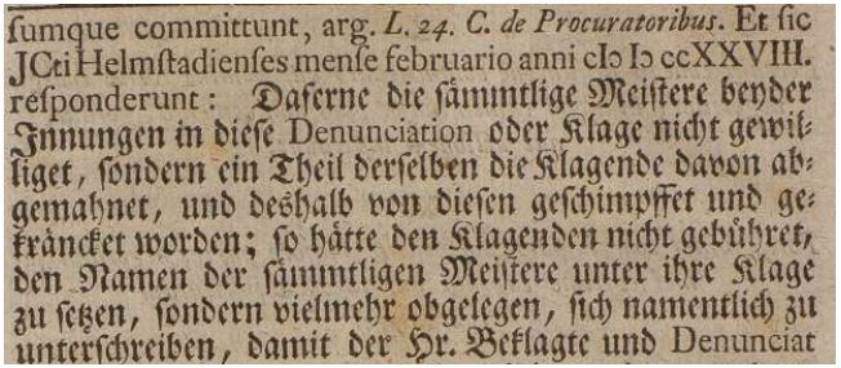
OCR Model



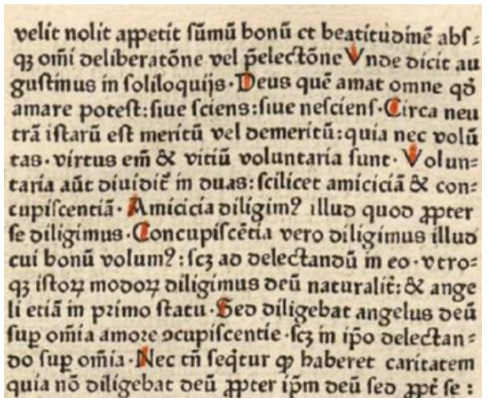

Kreüter
ner erſcheinūg / vnſerer teütſcher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Ariſtolochiam rotundam einſamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberflißigem vnrath der mūter / treibt auß die andgeburt vñ weiber menſes. Ein ſalb gemacht vonn diſer wurtzen zeitloſen vñ anagallide zeücht vß ſpreiſſel / dörn vñ geſchiferte bein.

Diemitt beſchlies ich mein rede diſer zeit von den zwelff zeichen kreütteren / begären menckclich welle mirs im beſten aufnemmen als dañ ichs gethan hab ſye weitleüffiger beſchribē wellen / ſo ſind yetziger kürtze viel vrſachen / vorauß dieweil ich groſſen koſten angewendet in ſüchung der kreüter auß eignem willen vñ beüttel / nam-

Adam von Bodenstein 1557 [1]
(**OCRopus: 99%**, ABBYY: 85%. Tesseract: 78%)


ſumque committunt, arg. *L. 24. C. de Procuratoribus.* Et ſic JCti Helmſtadienſes menſe februario anni cIↃ IↃ ccXXVIII. reſponderunt: Daferne die ſämmtlige Meiſtere beyder Innungen in dieſe Denunciation oder Klage nicht gewilliget, ſondern ein Theil derſelben die Klagende davon abgemahnet, und deshalb von dieſen geſchimpffet und gekräncket worden; ſo hätte den Klagenden nicht gebühret, den Namen der ſämmtligen Meiſtere unter ihre Klage zu ſetzen, ſondern vielmehr obgelegen, ſich namentlich zu unterſchreiben, damit der Hr. Beklagte und Denunciat

Augustinus Leyer 1735 [1]
(**OCRopus: 97%**, ABBYY: 77%. Tesseract: 82%)

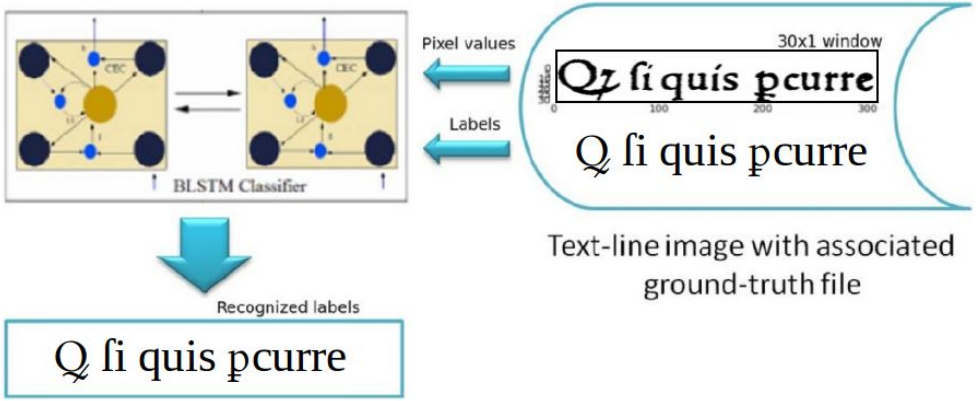

velit nolit appetit ſūmū bonū et beatitudinē abſq; oñi deliberatōne vel ꝑelectōne Vnde dicit augustinus in ſoliloquijs · Deus quē amat omne qō amare poteſt: ſiue ſciens: ſiue neſcienſ · Circa neutrā iſtarū eſt meritū vel demeritū: quia nec volūtas · virtus eñ & vitiū voluntaria ſunt · Voluntaria aūt diuidit in duas: ſcilicet amiciciā & concupiſcentiā · Amicicia diligim? illud quod ꝓpter ſe diligimus · Concupiſcētia vero diligimus illud cui bonū volum? : ſcz ad delectandū in eo · vtroq; iſtoꝝ modoꝝ diligimus deū naturaliē: & angeli etiā in primo ſtatu · Sed diligebat angelus deū ſup oñia amore ocupiſcentie · ſcz in ipo delectando ſup oñia · Nec cñ ſequtur q̄ haberet caritatem quia nō diligebat deū ꝓpter ipm deū ſed ꝓpē ſe :

Augsburg before 1476 [1]
(**OCRopus: 98%**)

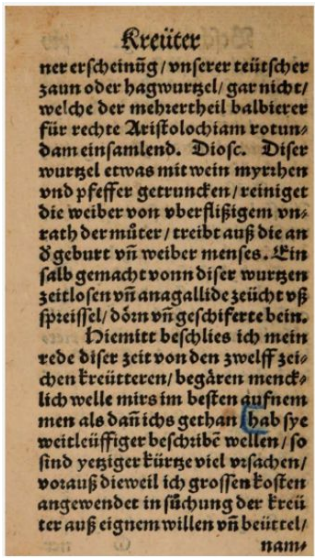
[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

What are the limitations of the OCRopus?

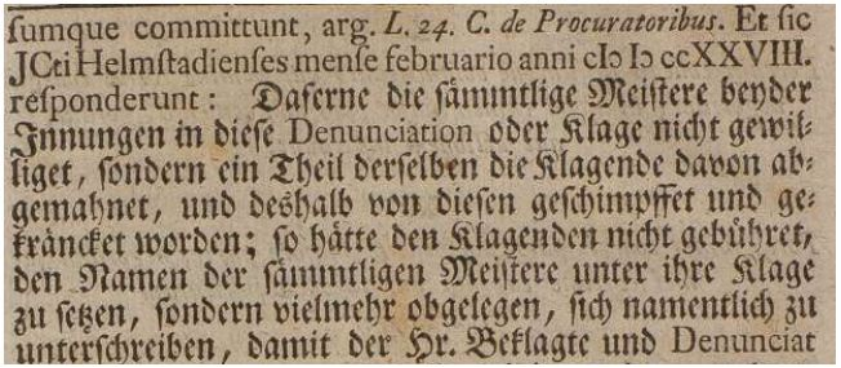
OCR Model



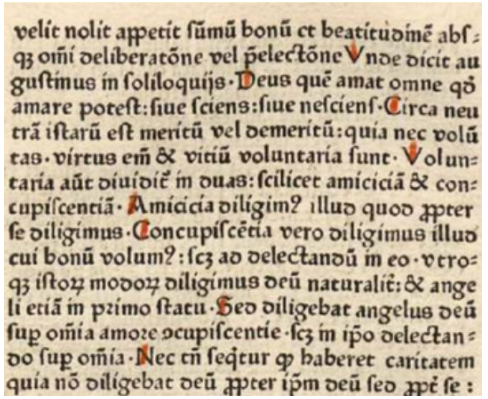
Training Data 50K Lines



Adam von Bodenstein 1557 [1]
(**OCRopus: 99%**, ABBYY: 85%. Tesseract: 78%)



Augustinus Leyer 1735 [1]
(**OCRopus: 97%**, ABBYY: 77%. Tesseract: 82%)



Augsburg before 1476 [1]
(**OCRopus: 98%**)

[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

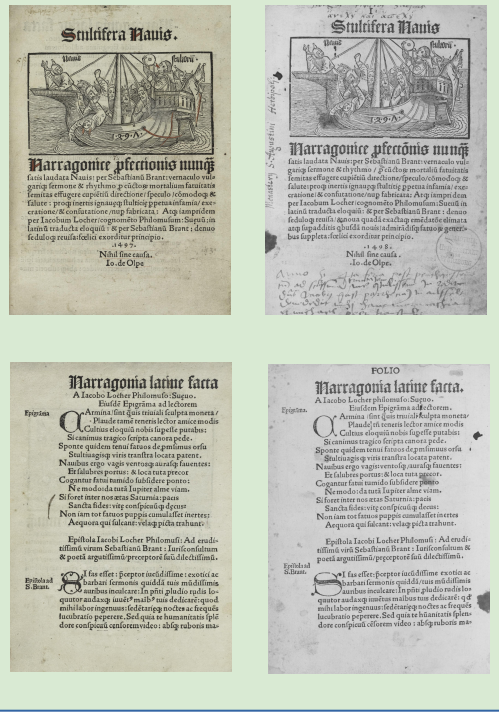
What are the limitations of the OCRopus?

OCR Model

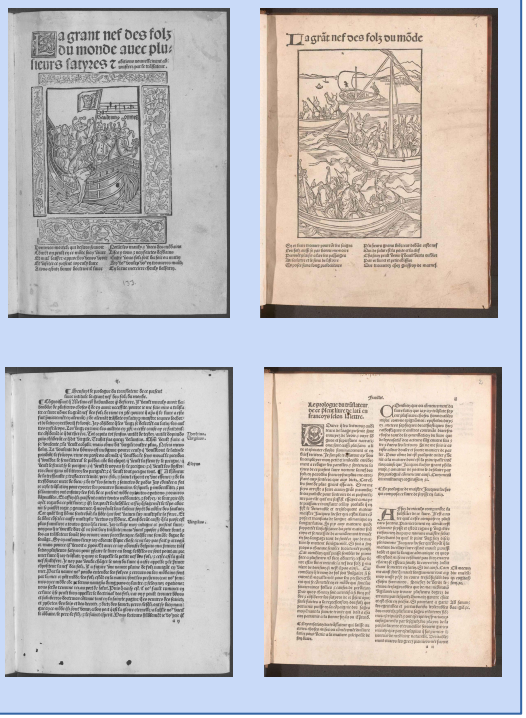
What are we dealing in this project?

German

Latin



French

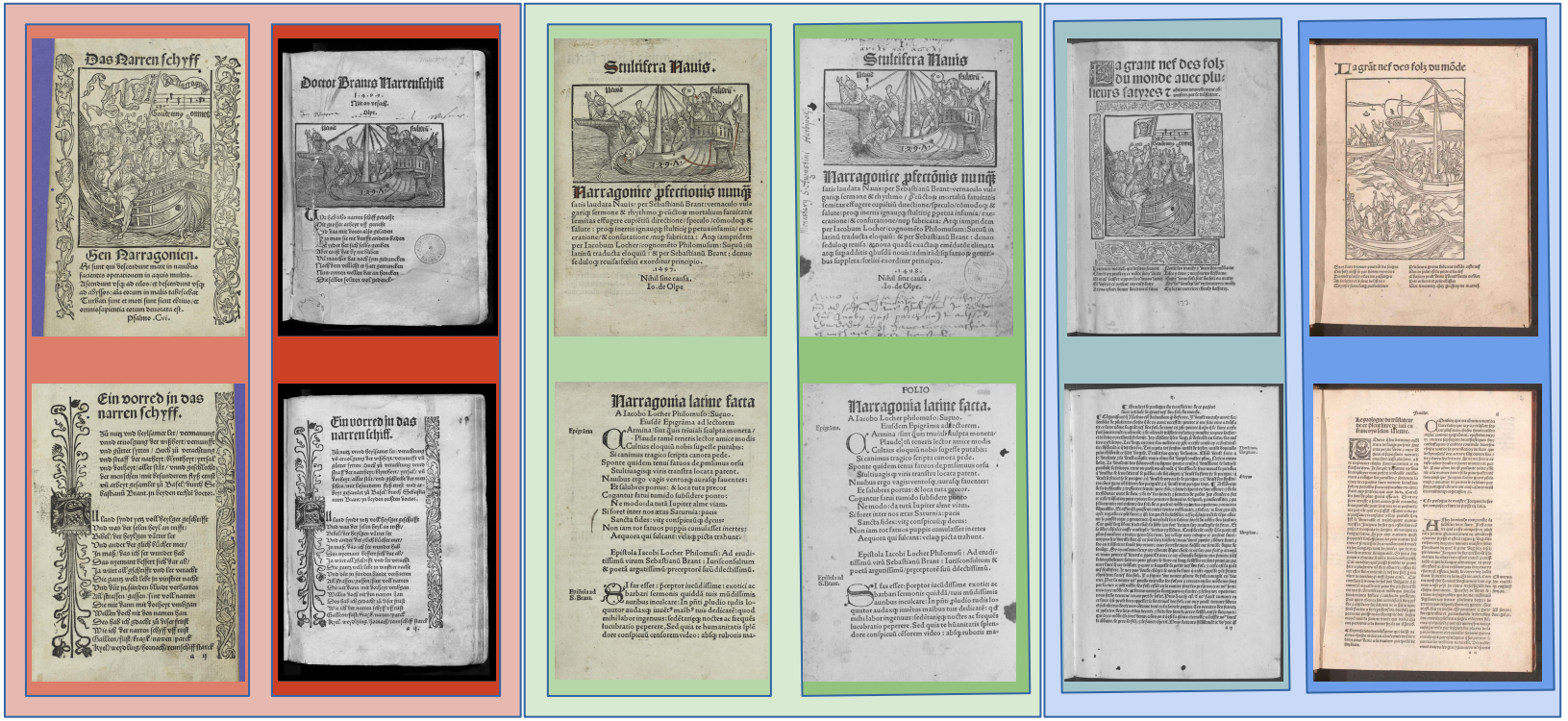


What are the limitations of the OCRopus?

OCR Model

What are we dealing in this project?

German | Latin | French



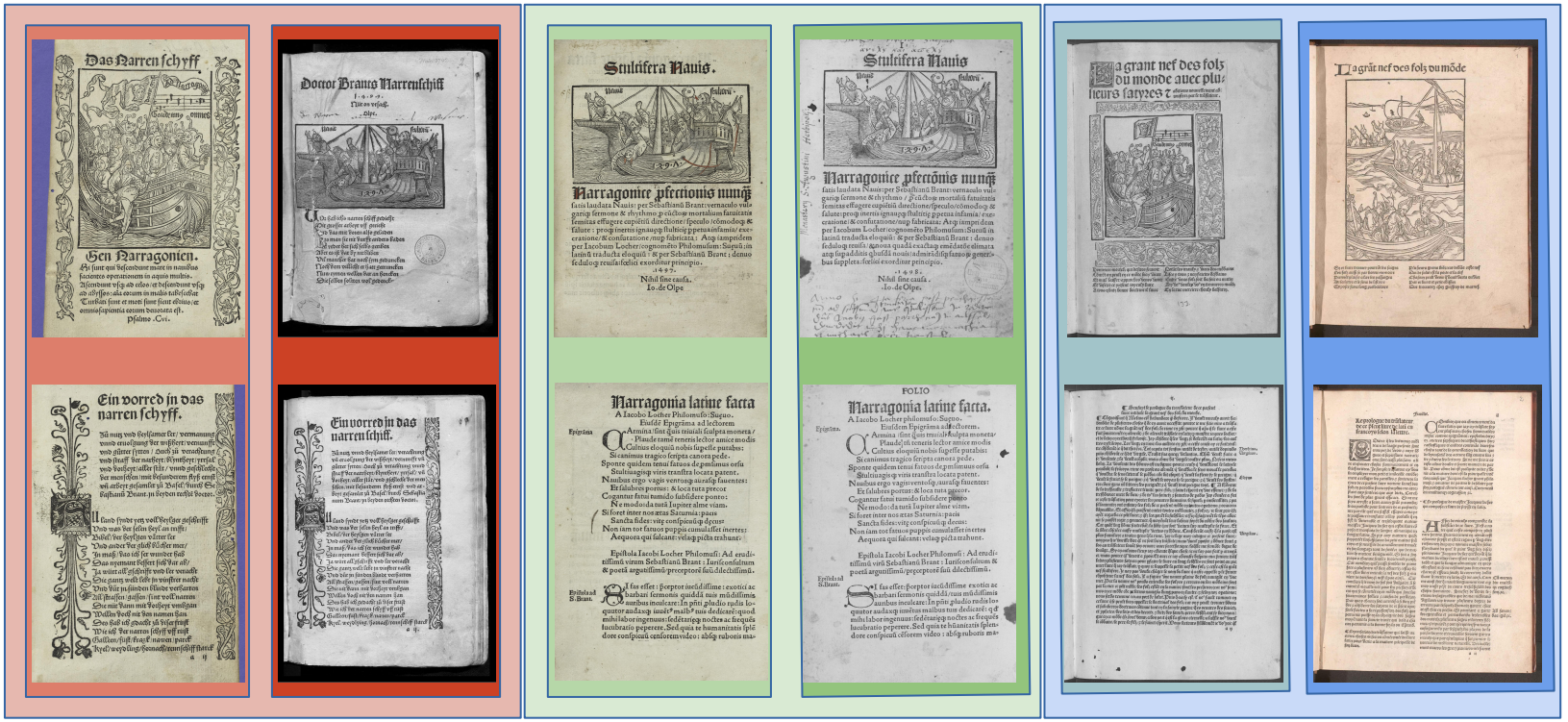
Basel 1495 (5046) | Basel 1499 (5047) | Basel 1497 (5054) | Basel 1498 (5062) | Lyon 1499 (5060) | Paris 1499 (5065)

What are the limitations of the OCRopus?

OCR Model

What are we dealing in this project?

German | Latin | French



Basel 1495 (5046) | Basel 1499 (5047) | Basel 1497 (5054) | Basel 1498 (5062) | Lyon 1499 (5060) | Paris 1499 (5065)

so what else is required for the 15th Century “Narrenschiffe” novels of this project?

(ii) Automatic OCR Model (anyOCR!)

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

Our Main Goals in This Project!

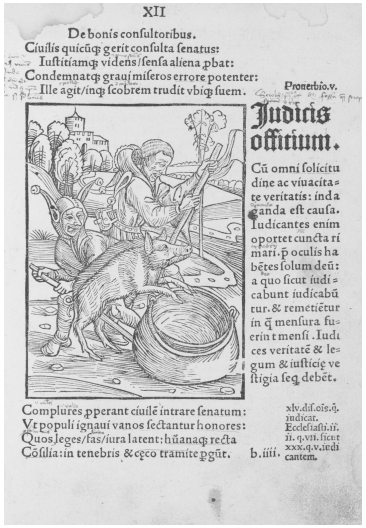
Our Main Goals in This Project!

Advanced Layout Analysis

Automatic OCR Model (anyOCR)

Our Main Goals in This Project!

Advanced Layout Analysis


XII

De bonis conſultoribus.

Ciuilis quicūq; gerit conſulta ſenatus:
Iuſtitiamq; videns/ſenſa aliena ꝓbat:
Condemnatq; graui miſeros errore potenter:
Ille agit/inq; ſcobrem trudit vbiq; ſuem. Prouerbio.v.

Judicis officium.

Cū omni ſolicitudine ac viuacitate veritatis: inda ganda eſt cauſa. Iudicantes enim oportet cuncta rimari.p̄ oculis habētes ſolum deū: a quo ſicut iudicabūtur.& remetiētur in q̄ menſura fuerint menſi .Iudices veritatē & legum & iuſticię vestigia ſeq debēt.

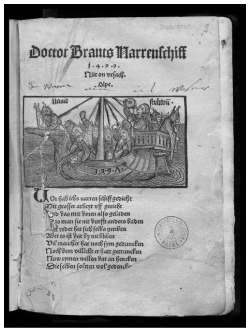
Complures ꝓperant ciuilē intrare ſenatum: xlv.dif.oīs.q̄. iudicat.
Vt populi ignaui vanos ſectantur honores: Eccleſiaſti.ii. n.q.vii.ſicut
Quos leges/fas/iura latent: hūanaq; recta xxx.q.v.iudi cantem.
Cōſilia:in tenebris & cęco tramite pgūt. b.iiii.

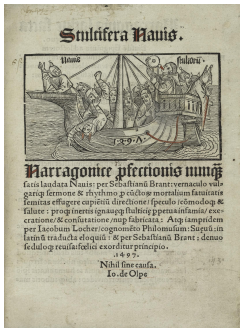
XII
De bonis conſultoribus.
Ciuilis quicūq; gerit conſulta ſenatus:
Iu itiamq; videns/ſenſa aliena ꝓbat:
Condemnatq; graui miſeros errore potenter:
Prouerbio.v.
Ille agit/inq; ſcrobem trudit vbiq; ſuem.
Iudicis
offitium.
Cū omni ſolicitu
dine ac viuacita=
te veritatis: inda
ganda e cauſa.
Iudicantes enim
oportet cun a ri
mari.p̄ oculis ha
bētes ſolum deū:
a quo ſicut iudi=
cabunt iudicabū
tur.& remetiētur
in q̄ menſura fu=
erint menſi .Iudi
ces veritatē & le=
gum & iu icię ve
igia ſeq debēt.
Complures ꝓperant ciuilē intrare ſenatum:
xlv.dif.oīs.q̄.
iudicat.
Eccleſia i.ii.
ii.q.vii.ſicut
xxx.q.v.iudi
cantem.
Vt populi ignaui vanos ſe antur honores:
Quos leges/fas/iura latent: hūanaq; re a
Cōſilia:in tenebris & cęco tramite pgūt. b.iiii.

Automatic OCR Model (anyOCR)

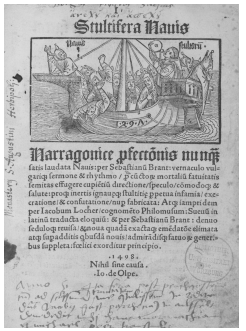

Das Narren ſchyff

Gen Narragonien.

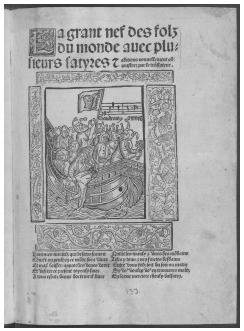

Doctor Brants Narrenſchiff

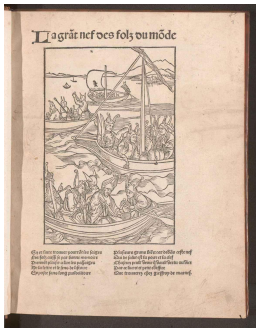

Stultifera Nauis.

Narragonice ꝓfectionis nunq̄


Stultifera Nauis

Narragonice ꝓfectionis nunq̄

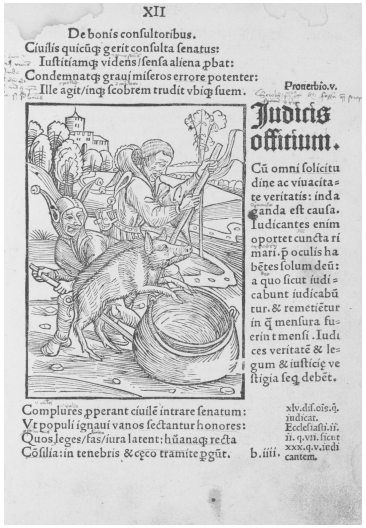

a grant nef des folz
du monde avec plu=
sieurs sayzes

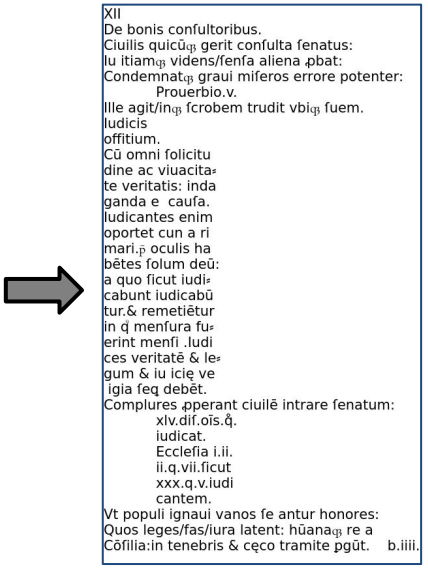

La grāt nef des folz du mōde

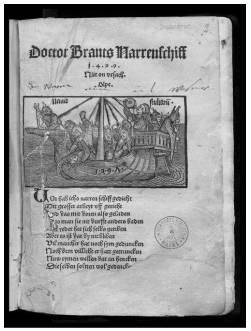
Our Main Goals in This Project!

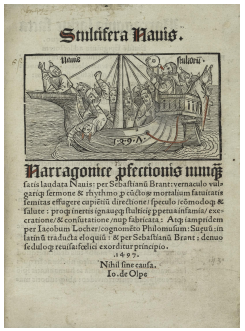
Advanced Layout Analysis

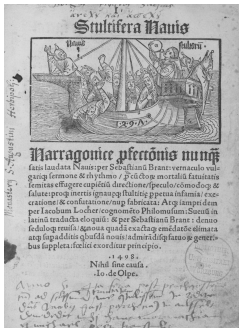
Automatic OCR Model (anyOCR)

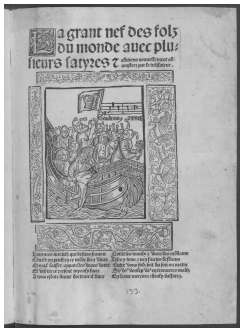


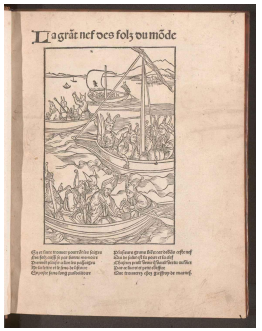








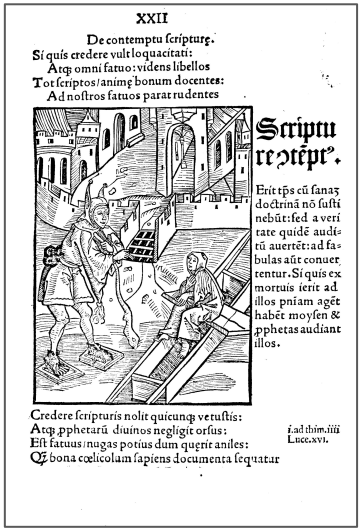




Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Advanced Layout Analysis

Text and Non-Text Segmentation Method


XXII

De contemptu ſcripturę.

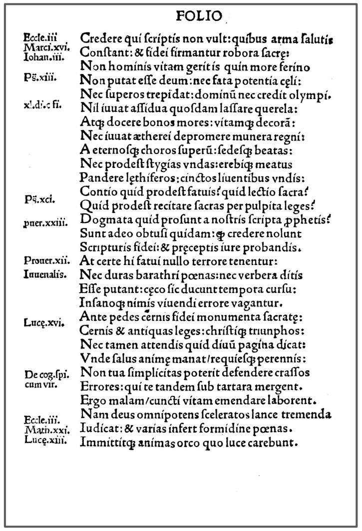
Si quis credere vult loquacitati:
Atq; omni fatuo: videns libellos
Tot ſcriptos/animę bonum docentes:
Ad noſtros fatuos parat rudentes

Scriptu re ↄtẽpt⁹.

Erit tp̃s cũ ſanã doctrinã nõ ſuſtinebũt: ſed a veritate quidẽ auditũ auertẽt: ad fabulas aũt conuertentur. Si quis ex mortuis ierit ad illos pniam agẽt habẽt moyſen & pphetas audiant illos.

Credere ſcripturis nolit quicunq; vetuſtis:
Atq; pphetarũ diuinos negligit orſus:
Eſt fatuus/nugas potius dum quęrit aniles:
Q̃ȝ bona cœlicolum ſapiens documenta ſequatur

i.ad thim.iiii
Luce.xvi.


FOLIO

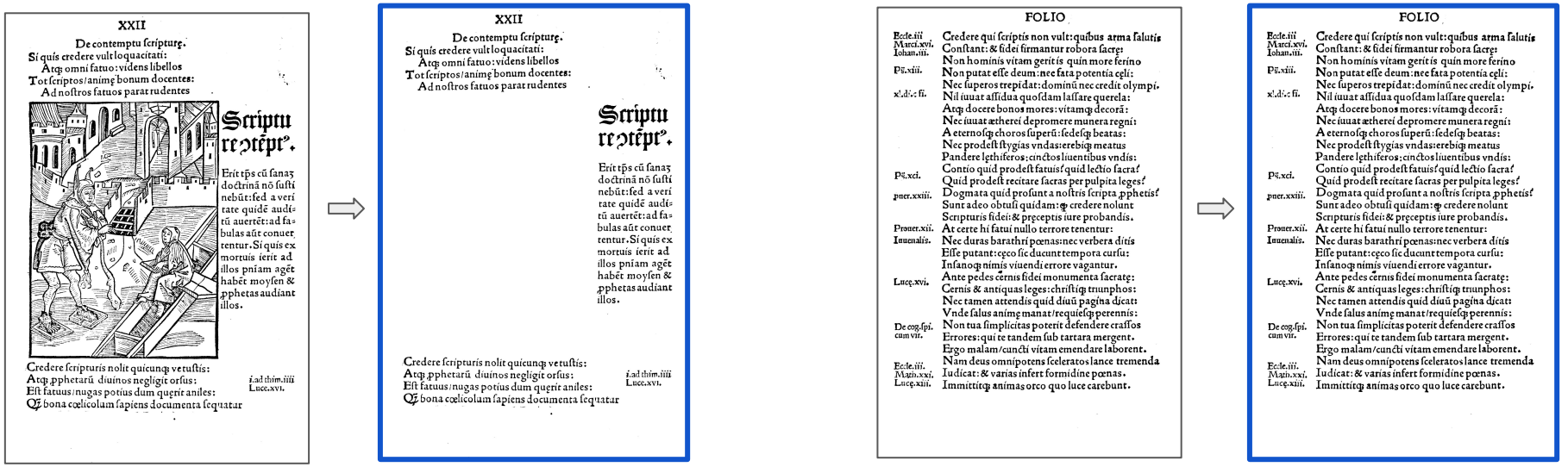
Eccle.iii / Marci.xvi. / Iohan.iii. — Credere qui ſcriptis non vult: quibus arma ſalutis
Conſtant: & fidei firmantur robora ſacrę:
Non hominis vitam geritis quin more ferino
Ps.xiii. — Non putat eſſe deum: nec fata potentia cęli:
Nec ſuperos trepidat: dominũ nec credit olympi.
xl.di.c fi. — Nil iuuat aſſidua quoſdam laſſare querela:
Atq; docere bonos mores: vitamq; decorã:
Nec iuuat ætherei depromere munera regni:
A eternoſq; choros ſuperũ: ſedeſq; beatas:
Nec prodeſt ſtygias vndas: erebiq; meatus
Pandere lęthiferos: cinctos liuentibus vndis:
Contio quid prodeſt fatuis? quid lectio ſacra?
Ps.xci. — Quid prodeſt recitare ſacras per pulpita leges?
pruer.xxiii. — Dogmata quid proſunt a noſtris ſcripta pphetis?
Sunt adeo obtuſi quidam: q̃ credere nolunt
Scripturis fidei: & pręceptis iure probandis.
Prouer.xii. — At certe hi fatui nullo terrore tenentur:
Iuuenalis. — Nec duras barathri pœnas: nec verbera ditis
Eſſe putant: cęco ſic ducunt tempora curſu:
Inſanoq; nimis viuendi errore vagantur.
Lucę.xvi. — Ante pedes cernis fidei monumenta ſacratę:
Cernis & antiquas leges: chriſtiq; triunphos:
Nec tamen attendis quid diuũ pagina dicat:
Vnde ſalus animę manat/requieſq; perennis:
De cog.ſpi. cum vir. — Non tua ſimplicitas poterit defendere craſſos
Errores: qui te tandem ſub tartara mergent.
Ergo malam/cuncti vitam emendare laborent.
Nam deus omnipotens ſceleratos lance tremenda
Eccle.iii. / Math.xxi. — Iudicat: & varias infert formidine pœnas.
Lucę.xiii. — Immittitq; animas orco quo luce carebunt.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

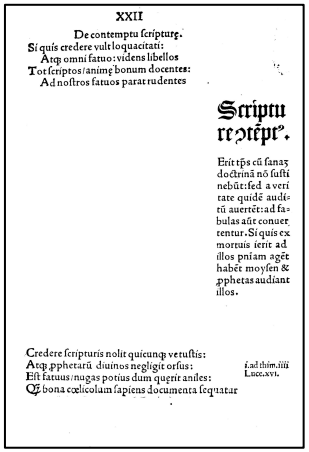
Dr. Syed Saqib Bukhari

Advanced Layout Analysis

Text and Non-Text Segmentation Method



Text Line Segmentation



XXII

De contemptu scripturę.

Si quis credere vult loquacitati:
Atq; omni fatuo: videns libellos
Tot scriptos/animę bonum docentes:
Ad nostros fatuos parat rudentes

Scriptu
re ꝓtēpr̄.

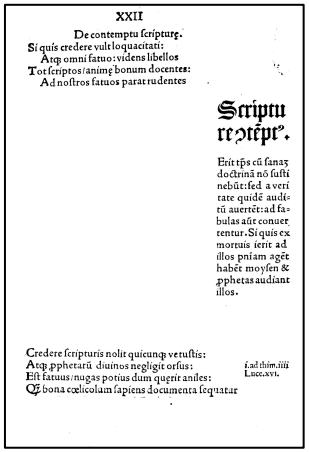
Erit tp̄s cū sanaꝫ
doctrinā nō sufti
nebūt: sed a veri
tate quidē audi-
tū auertēt: ad fa-
bulas aūt conuer-
tentur. Si quis ex
mortuis ierit ad
illos pniam agēt
habēt moysen &
ꝓphetas audiant
illos.

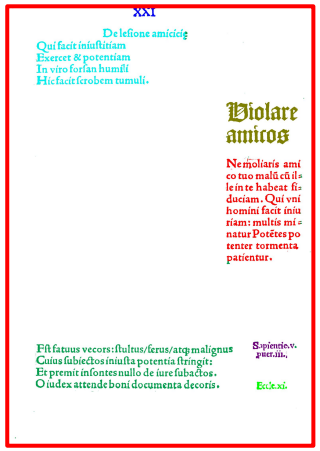
Credere scripturis nolit quicunq; vetustis:
Atq; ꝓphetarū diuinos negligit orsus:
Elt fatuus/nugas potius dum quęrit aniles:
Qꝫ bona cœlicolum sapiens documenta sequatur

i.ad thim.iiii
Luce.xvi.

Advanced Layout Analysis

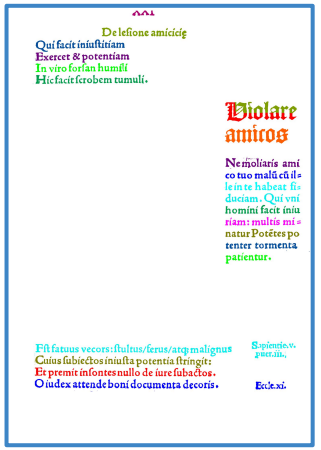
Text Line Segmentation





XY Cut

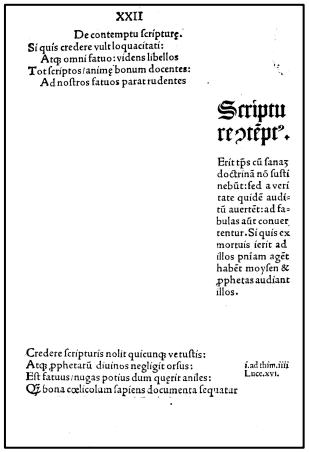
Voronoi



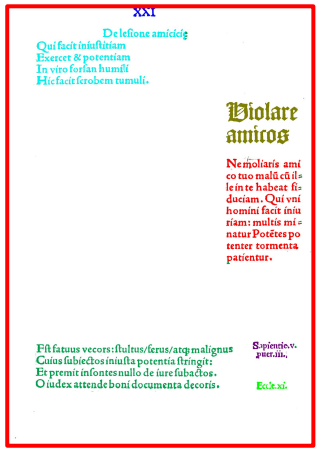
OCRopus

Advanced Layout Analysis

Text Line Segmentation

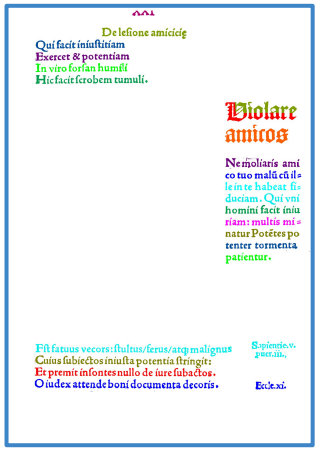


XY Cut



Voronoi

OCRopus



OCRopus++

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

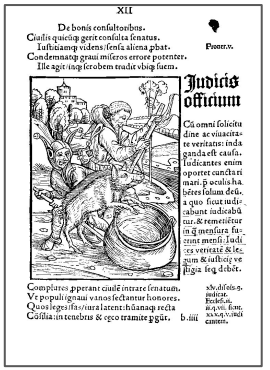
Dr. Syed Saqib Bukhari

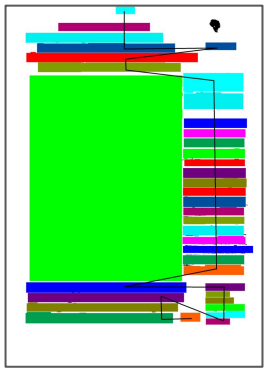
Advanced Layout Analysis

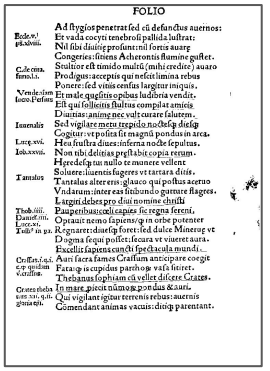
Performance Evaluation of **OCRopus++**:

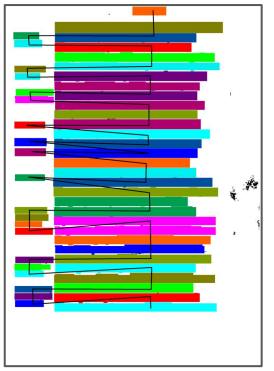
Advanced Layout Analysis

Performance Evaluation of **OCRopus++**:







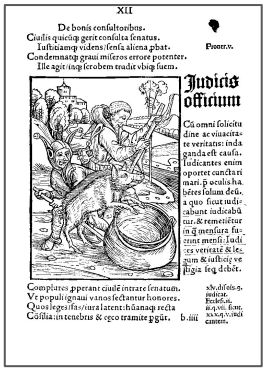
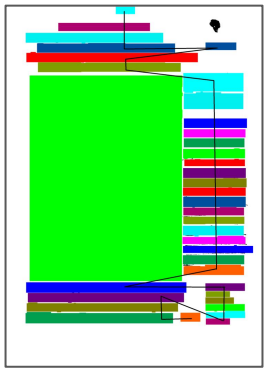
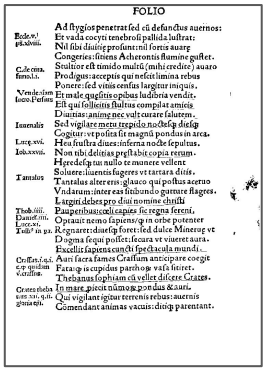
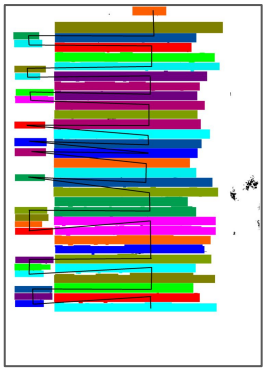


Advanced Layout Analysis

Performance Evaluation of OCRopus++:

- Text and Non-Text Segmentation Accuracy: **99.34%**
- Text Line Segmentation Accuracy: **87%***

*as compared to state-of-the-art methods ~10% and **OCRopus 80%**.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

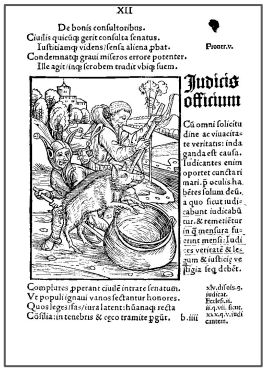
Dr. Syed Saqib Bukhari

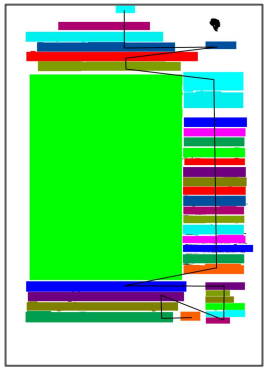
Advanced Layout Analysis

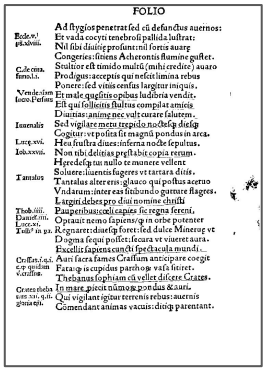
Performance Evaluation of OCRopus++:

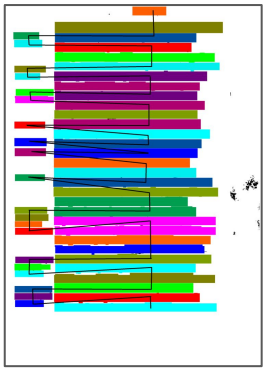
- Text and Non-Text Segmentation Accuracy: **99.34%**
- Text Line Segmentation Accuracy: **87%***

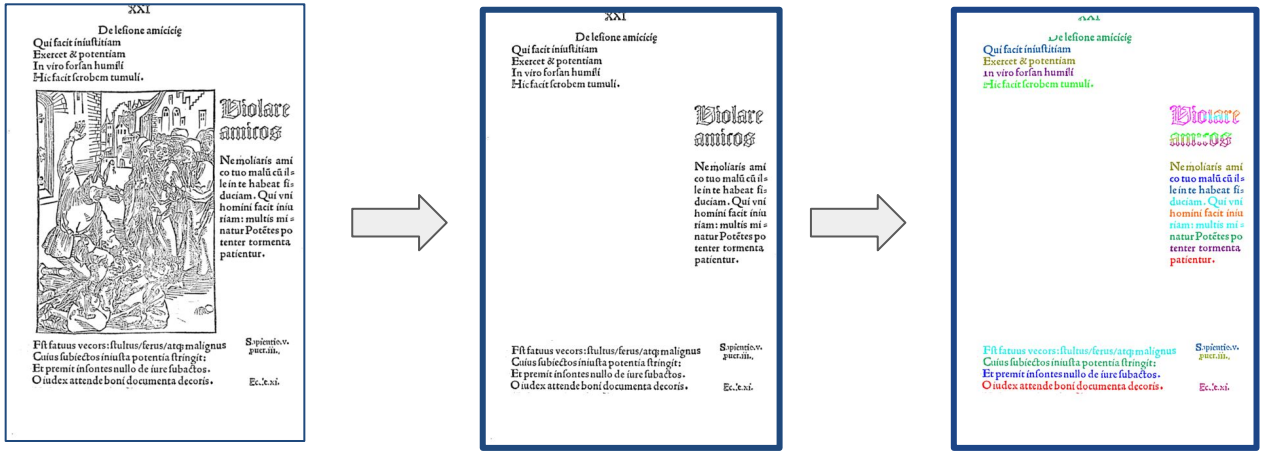
*as compared to state-of-the-art methods ~10% and **OCRopus 80%**.











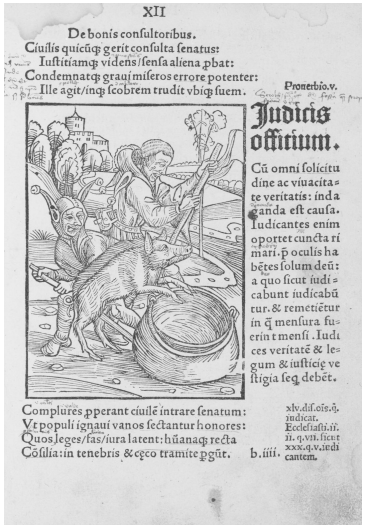
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

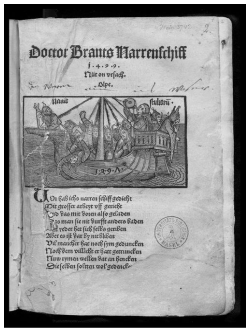
Our Main Goals in This Project!

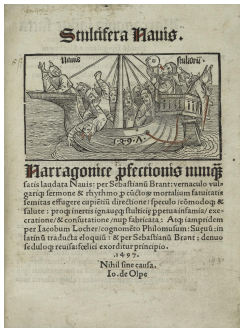
Advanced Layout Analysis

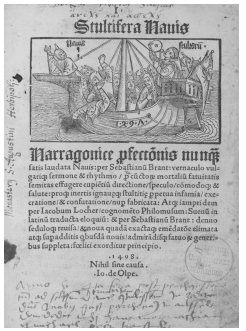
Automatic OCR Model (anyOCR)

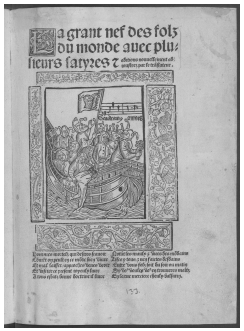


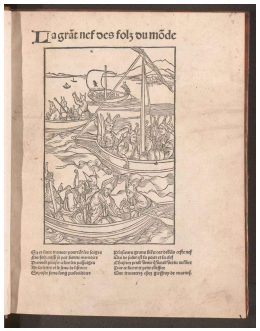
XII
De bonis conſultoribus.
Ciuilis quicūq; gerit conſulta ſenatus:
Iu itiamq; videns/ſenſa aliena ꝓbat:
Condemnatq; graui miſeros errore potenter:
Prouerbio.v.
Ille agit/inq; ſcrobem trudit vbiq; ſuem.
Iudicis
offitium.
Cū omni ſolicitu
dine ac viuacita
te veritatis: inda
ganda e cauſa.
Iudicantes enim
oportet cun a ri
mari.p̄ oculis ha
bētes ſolum deū:
a quo ſicut iudi
cabunt iudicabū
tur.& remetiētur
in q̄ menſura fu
erint menſi .Iudi
ces veritatē & le
gum & iu icię ve
igia ſeq debēt.
Complures ꝓperant ciuilē intrare ſenatum:
xlv.dif.oīs.q̄.
iudicat.
Eccleſia i.ii.
ii.q.vii.ſicut
xxx.q.v.iudi
cantem.
Vt populi ignaui vanos ſe antur honores:
Quos leges/fas/iura latent: hūanaq; re a
Cōſilia:in tenebris & cęco tramite pgūt. b.iiii.











Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

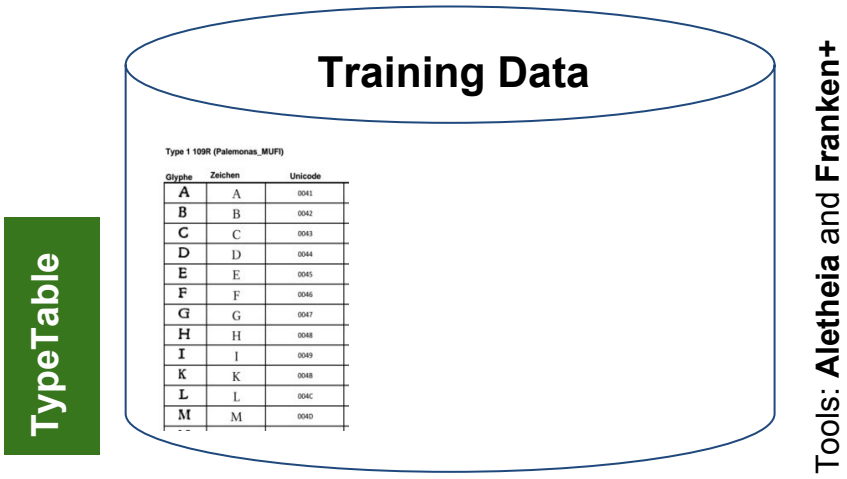
Background: OCR Training Models can broadly be classified as:

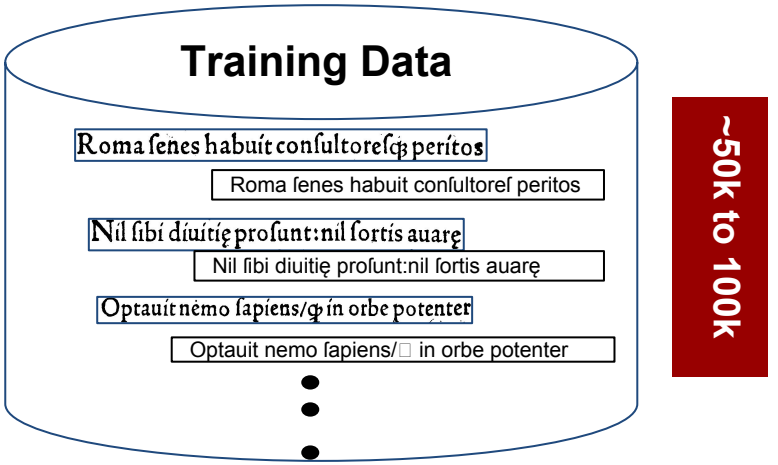
- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
- **Segmentation-Free OCR (OCRopus)**
  - line recognizer

anyOCR - Automatic OCR Model

Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
- **Segmentation-Free OCR (OCRopus)**
  - line recognizer





Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

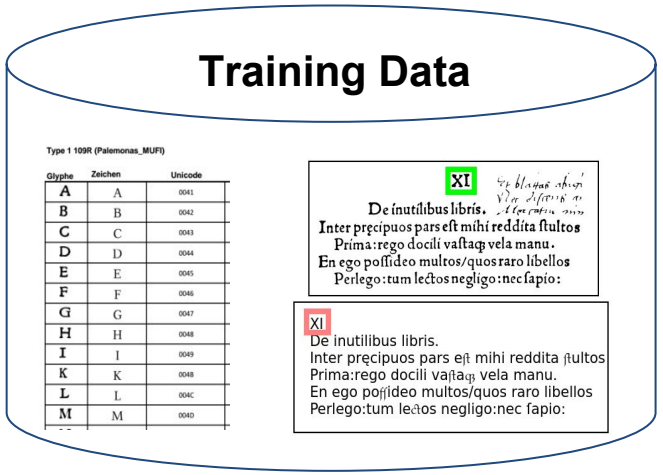
Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

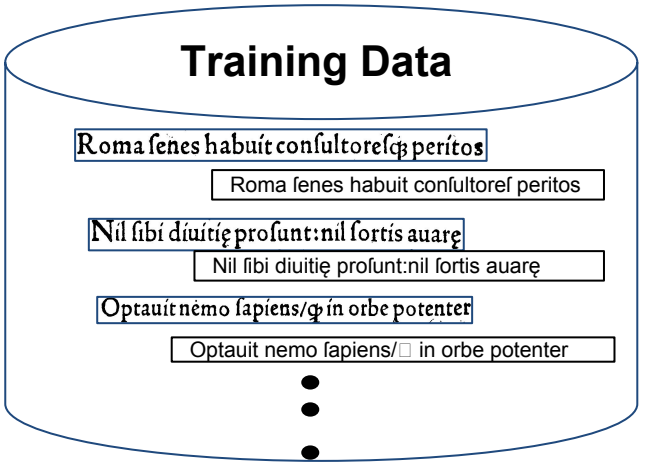
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
- **Segmentation-Free OCR (OCRopus)**
  - line recognizer

TypeTable / Few Pages



Tools: **Aletheia** and **Franken+**



~50k to 100k

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

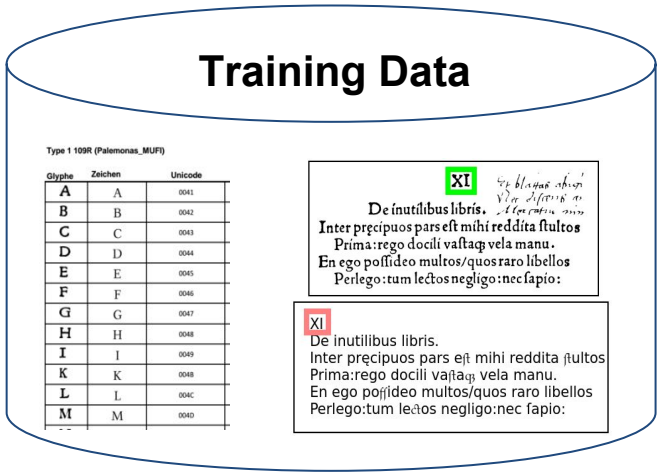
Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

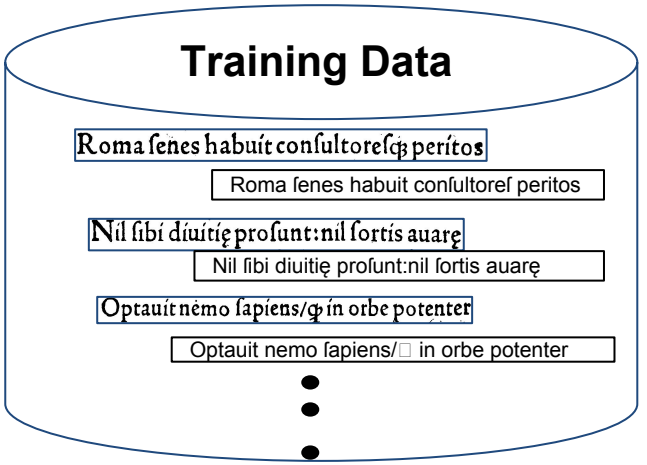
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
  - Lower Performance
    - Sensitive to Noise
    - Character Segmentation Errors
    - No Language Model
- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
  - Higher Performance
    - Insensitive to Noise
    - Independent to Noise & Character Seg.
    - Implicit Language Model

TypeTable / Few Pages



Tools: **Aletheia** and **Franken+**



~50k to 100k

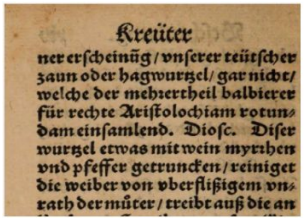
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

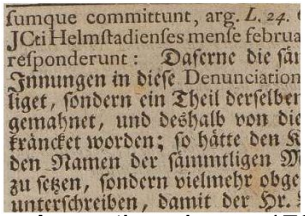
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
  - Lower Performance
    - Sensitive to Noise
    - Character Segmentation Errors
    - No Language Model
- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
  - Higher Performance
    - Insensitive to Noise
    - Independent to Noise & Character Seg.
    - Implicit Language Model



Adam von Bodenstein 1557

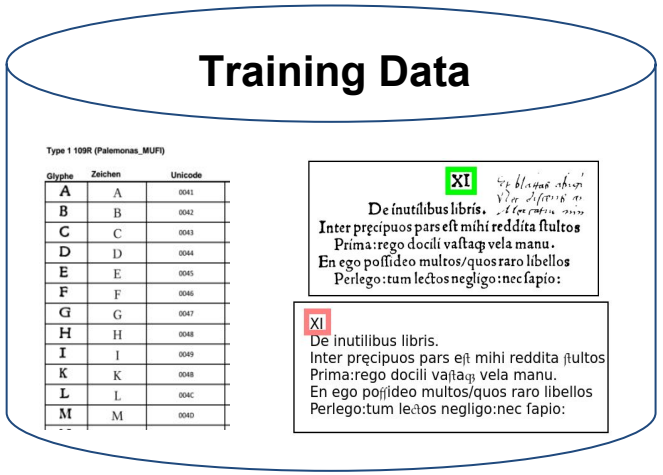
(Ocropus: 99%, Tesseract: 78%)



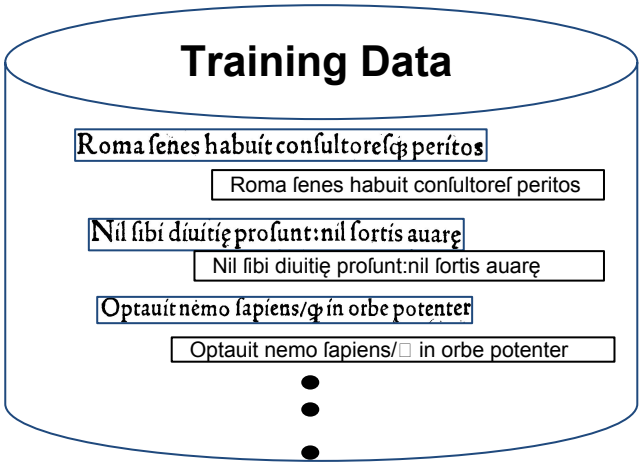
Augustinus Leyer 1735

(OCRopus: 97%, Tesseract: 82%)

TypeTable / Few Pages



Tools: **Aletheia** and **Franken+**



~50k to 100k

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

Background: OCR Training Models can broadly be classified as:

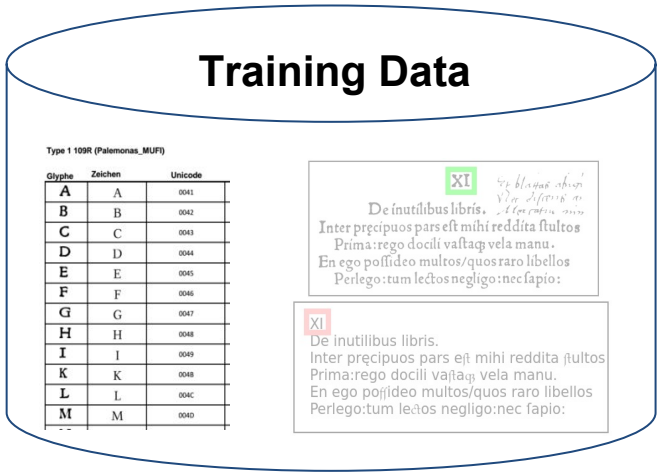
- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
  - Lower Performance
    - Sensitive to Noise
    - Character Segmentation Errors
    - No Language Model

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
  - Higher Performance
    - Insensitive to Noise
    - Independent to Noise & Character Seg.
    - Implicit Language Model

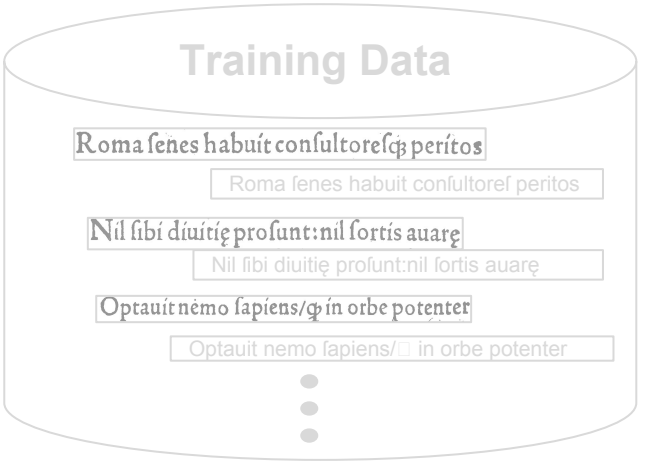
Adam von Bodenstein 1557
(Ocropus: 99%, Tesseract: 78%)

Augustinus Leyer 1735
(OCRopus: 97%, Tesseract: 82%)

TypeTable / Few Pages



Tools: Aletheia and Franken+



~50k to 100k

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable

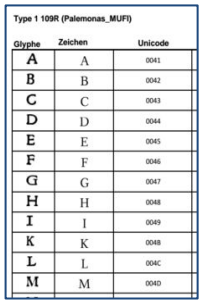
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



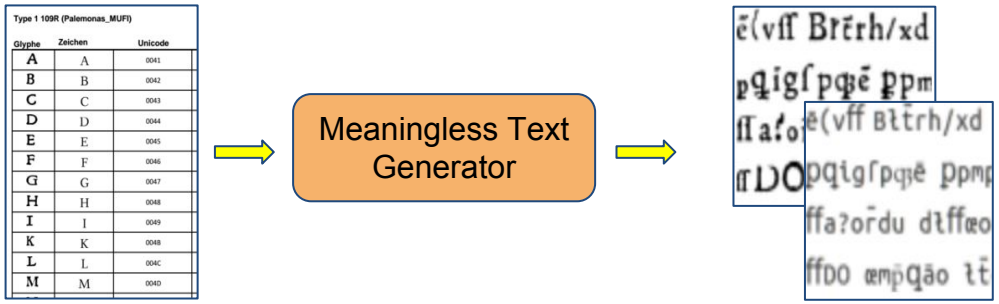
Type 1 109R (Palemonas_MUFI)

Glyphe	Zeichen	Unicode
A	A	0041
B	B	0042
C	C	0043
D	D	0044
E	E	0045
F	F	0046
G	G	0047
H	H	0048
I	I	0049
K	K	0048
L	L	004C
M	M	004D

anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

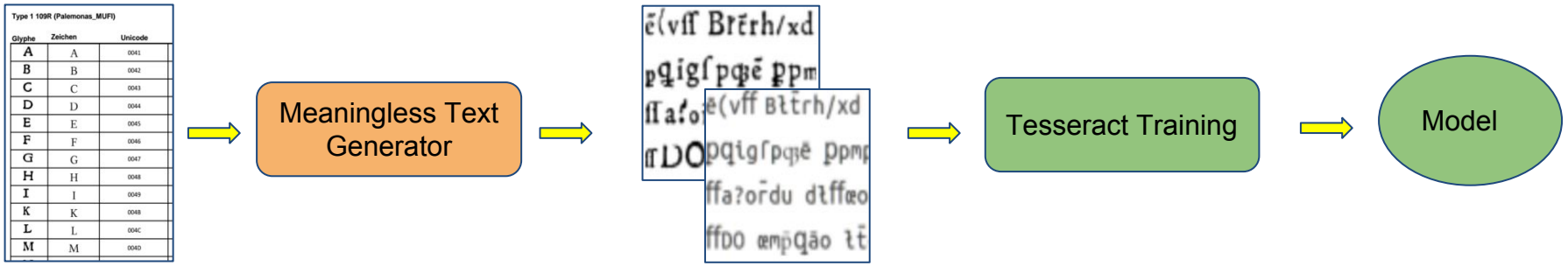
- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

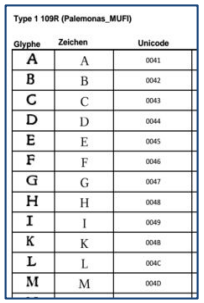
- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



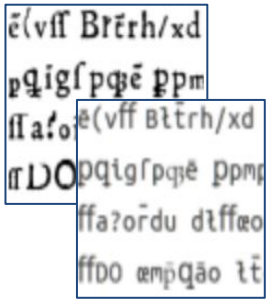
anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



Meaningless Text Generator



Tesseract Training

Layout Analysis

Line Images

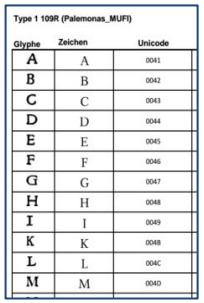
Model

Erroneous Text Lines

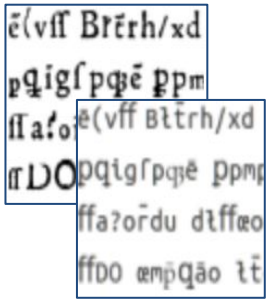
anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



Meaningless Text Generator



Tesseract Training

Model

Layout Analysis

Line Images

Erroneous Text Lines

Ocropus Training

OCRoRECT Model

Less Erroneous Text Lines

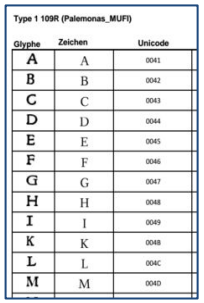
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

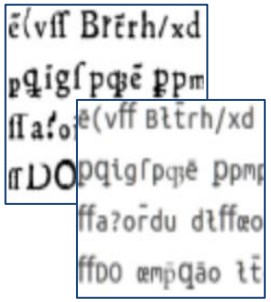
anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



Meaningless Text Generator



Tesseract Training

Model

Layout Analysis

Line Images

Erroneous Text Lines

Ocropus Training

OCRoRECT Model

Less Erroneous Text Lines

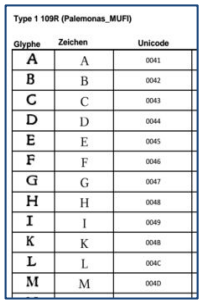
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

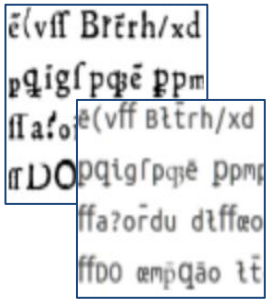
anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



Meaningless Text Generator



Tesseract Training

Model

Layout Analysis

Line Images

Erroneous Text Lines

Ocropus Training

OCRoRECT Model

Correct Text Lines

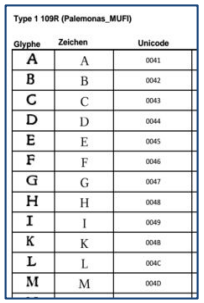
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

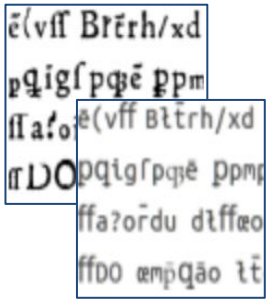
anyOCR - Automatic OCR Model

anyOCR: *A Combination of* ***OCRo****pus and tesse****RECT*** *(OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model



Meaningless Text Generator



Tesseract Training

Model

Layout Analysis

Line Images

Erroneous Text Lines

Ocropus Training

OCRoRECT Model

Correct Text Lines

[Publication] Bukhari, Ul-Hasan, Dengel: OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters, DAS 2016 [accepted for Oral presentation]

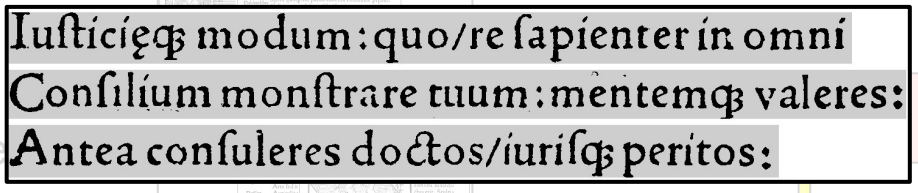
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

*anyOCR: A Combination of **OCR**opus and tesse**RECT** (OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Se
  - Implicit Language Model



Meaningless Text Generator

Tesseract Training

Model

Line Images

Erroneous Text Lines

Ocropus Training

OCRoRECT Model

Correct Text Lines

[Publication] Bukhari, Ul-Hasan, Dengel: OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters, DAS 2016 [accepted for Oral presentation]

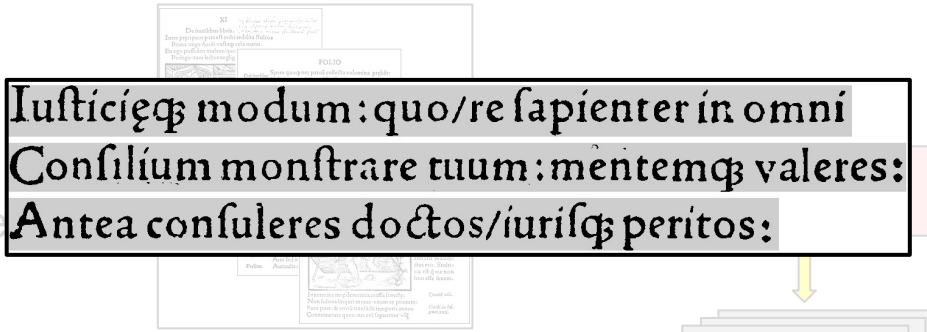
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

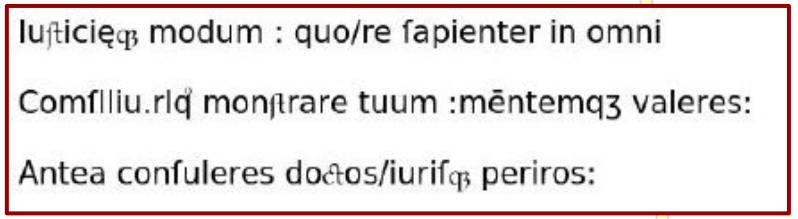
*anyOCR: A Combination of **OCR**opus and tesse**RECT** (OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Se
  - Implicit Language Model



Line Images

Meaningless Text Generator



Erroneous Text Lines

OCRoRECT Model

Ocropus Training

Correct Text Lines

[Publication] Bukhari, Ul-Hasan, Dengel: OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters, DAS 2016 [accepted for Oral presentation]

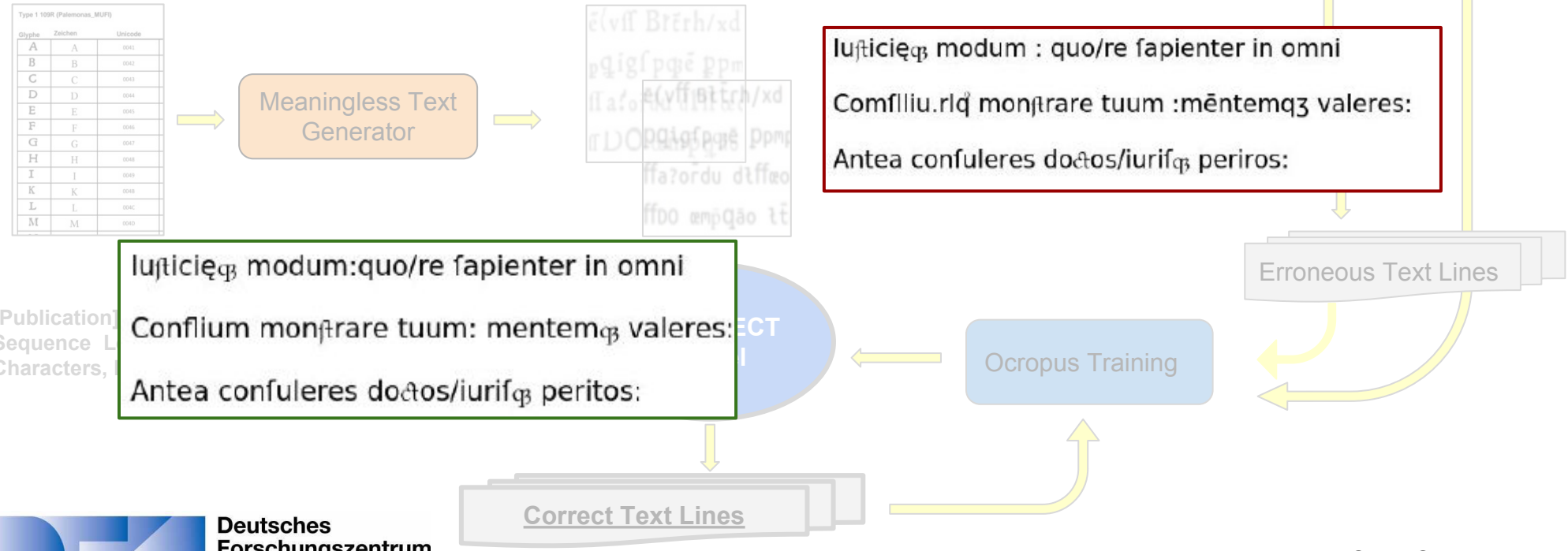
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

***anyOCR:** A Combination of **OCRo**pus and tesse**RECT** (OCRoRECT)*

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Se
  - Implicit Language Model

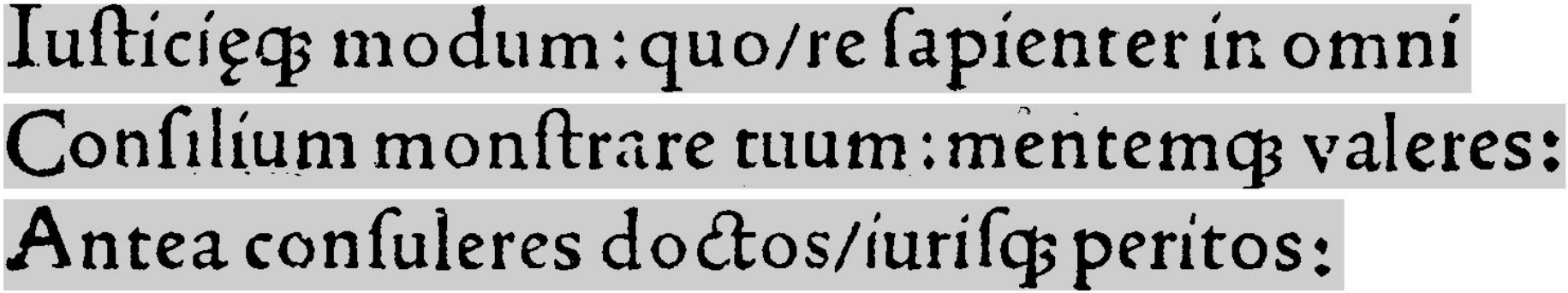


Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

Qualitative Performance Evaluation of OCRopus++



Tesseract

```
luſticięqꝫ modum : quo/re ſapienter in omni
Comflliu.rlq̊ monſtrare tuum :mēntemqʒ valeres:
Antea conſuleres doctos/iuriſqꝫ periros:
```

anyOCR: OCRoRECT

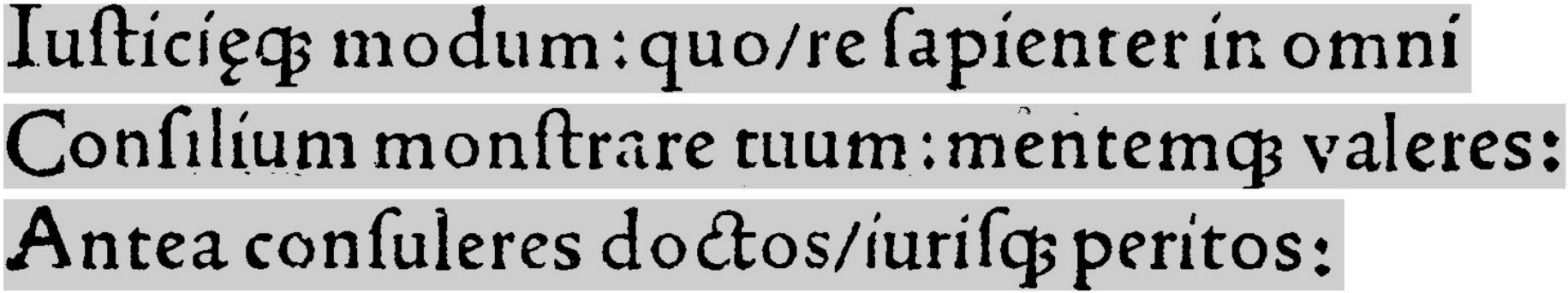
```
luſticięqꝫ modum:quo/re ſapienter in omni
Conflium monſtrare tuum: mentemqꝫ valeres:
Antea conſuleres doctos/iuriſqꝫ peritos:
```

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Dr. Syed Saqib Bukhari

anyOCR - Automatic OCR Model

Qualitative Performance Evaluation of OCRopus++



Tesseract	OCRopus	anyOCR: OCRoRECT
luſticięꝫ modum : quo/re ſapienter in omni	luſticięꝫ modum:quo/re ſapienter in omni	luſticięꝫ modum:quo/re ſapienter in omni
Comflliu.rlꝙ monſtrare tuum :mēntemqʒ valeres:	Conſilium monſtrare tuum:mentemꝫ valeres:	Conflium monſtrare tuum: mentemꝫ valeres:
Antea conſuleres doctos/iuriſꝫ periros:	Antea conſuleres doctos/iuriſꝫ peritos:	Antea conſuleres doctos/iuriſꝫ peritos:

anyOCR - Automatic OCR Model

Performance Evaluation of OCRopus++:

Dataset \ OCR Model	Tesseract	OCRopus	anyOCR - OCRoRECT
Adam von Bodenstein 1557	78% [1]	**99%** [1]	-
Augustinus Leyer 1735	82% [1]	**97%** [1]	-

[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf

anyOCR - Automatic OCR Model

Performance Evaluation of OCRopus++:

Dataset \ OCR Model	Tesseract	OCRopus	anyOCR - OCRoRECT
Adam von Bodenstein 1557	78% [1]	**99%** [1]	-
Augustinus Leyer 1735	82% [1]	**97%** [1]	-
Basel 1497 (Narrenschif, Latin)	*77% [2]*	***98%** [2]*	

[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf
[2] Bukhari, Ul-Hasan, Dengel: OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters, DAS 2016 [accepted for Oral presentation]

anyOCR - Automatic OCR Model

Performance Evaluation of OCRopus++:

Dataset \ OCR Model	Tesseract	OCRopus	anyOCR - OCRoRECT
Adam von Bodenstein 1557	78% [1]	**99%** [1]	-
Augustinus Leyer 1735	82% [1]	**97%** [1]	-
Basel 1497 (Narrenschif, Latin)	*77% [2]*	***98%** [2]*	***95% [2]***

[1] Springmann, Ocrocis Tutorial: cistern.cis.lmu.de/ocrocis/tutorial.pdf
[2] Bukhari, Ul-Hasan, Dengel: OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters, DAS 2016 [accepted for Oral presentation]

Conclusion

Conclusion

- Layout Analysis is an open-challenging problem for complex documents
- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)

Conclusion

- Layout Analysis is an open-challenging problem for complex documents
- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)
- Presented **OCRopus++** OCR System
  - advanced layout analysis
  - anyOCR - OCRoRECT recognizer

Conclusion

- Layout Analysis is an open-challenging problem for complex documents
- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)
- Presented **OCRopus++** OCR System
  - advanced layout analysis
  - anyOCR - OCRoRECT recognizer
- In future, add more features in **OCRopus++**

Conclusion

- Layout Analysis is an open-challenging problem for complex documents
- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)
- Presented **OCRopus++** OCR System
  - advanced layout analysis
  - anyOCR - OCRoRECT recognizer
- In future, add more features in **OCRopus++**
- Publications:
  - Bukhari, Ul-Hasan, Dengel, “OCRoRACT: A Sequence Learning OCR System Trained on Isolated Characters”, DAS 2016
  - Bukhari, Ul-Hasan, Dengel, "Meaningless Text OCR Model for Medieval Scripts", 2nd International Conference on Natural Sciences and Technology in Manuscript Analysis 2016, Germany.
  - Bukhari, Jenckel, Dengel, "Clustering Benchmark for Characters in Historical Documents", DAS 2016.
  - Bukhari, Nunamaker, Borth, Dengel, "A Tesseract Based OCR Framework For Historical Document Lacking Ground-Truth", ICIP 2016 [Under Review]

Thank you, ... Questions?

Address:
Dr.-Ing. Syed Saqib Bukhari
DFKI GmbH
P.O. Box 2080
D-67608 Kaiserslautern
email: saqib.bukhari@dfki.de
http://www.dfki.de/~bukhari