Text Segmentation

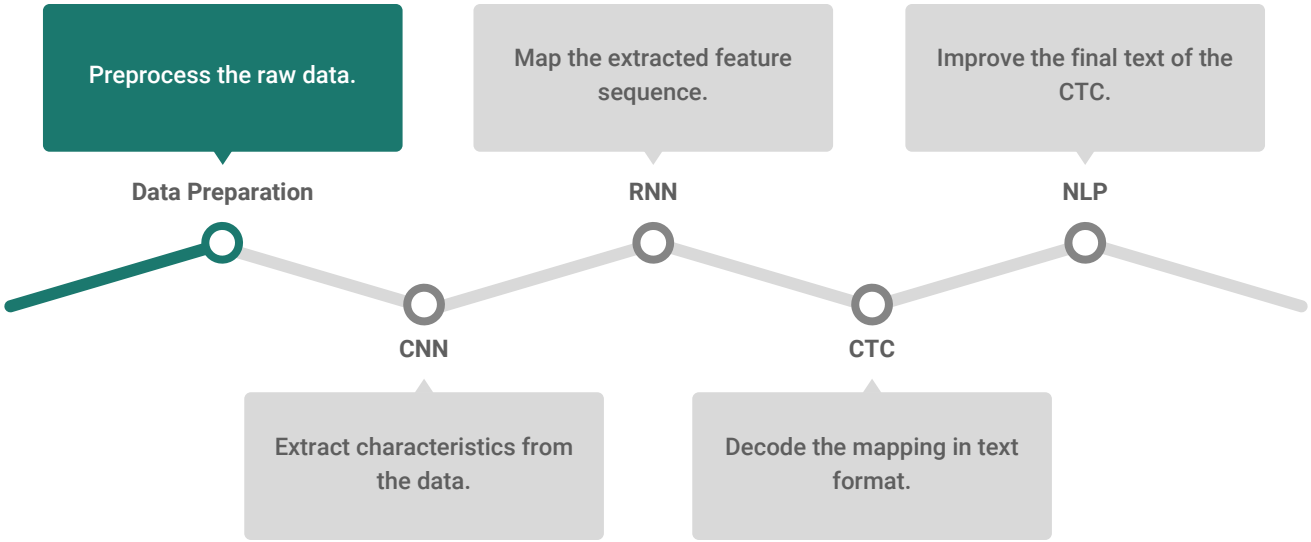
Arthur Flor

Content

1. Introduction
2. Studied Content
3. Applied Techniques
4. Conclusions
5. Upcoming Activities
6. References

Introduction

Study related to the Handwritten Text Recognition project using Natural Language Processing (HTR + NLP).



Studied Content

- ➔ Document Scanner;
- ➔ Binarization with illumination compensation;
- ➔ Line Segmentation with deslanting;
- ➔ Word Segmentation.

Document Scanner

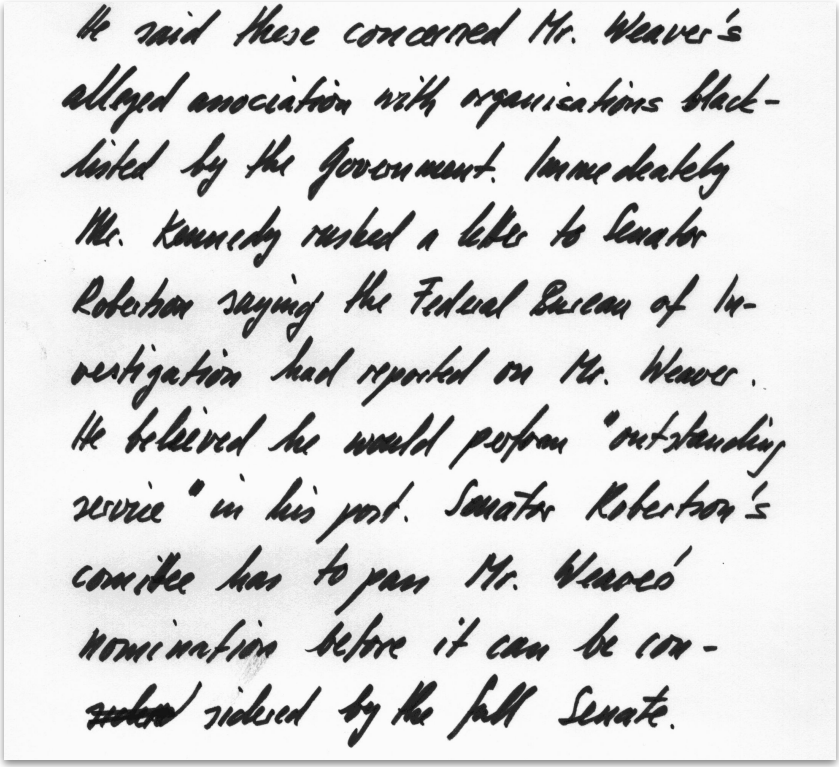
Process of detecting the predominant contour in the image and segment using a four-point perspective transformation.

Image test with the document scanner process

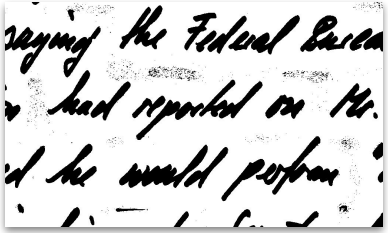
Binarization

- ➔ 1979: A Threshold Selection Method from Gray-Level Histograms (**Otsu**);
- ➔ 1986: An Introduction to Digital Image Processing (**Niblack**);
- ➔ 1997: Adaptive Document Binarization (**Sauvola**);
- ➔ 2002: Text Localization, Enhancement and Binarization in Multimedia Documents (**Wolf**);
- ➔ 2010: Binarization of Historical Document Images Using the Local Maximum and Minimum (**Su**).

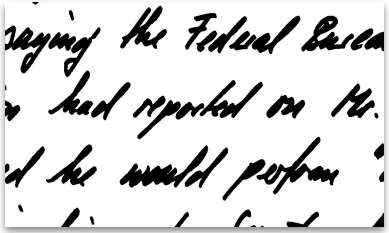
Binarization



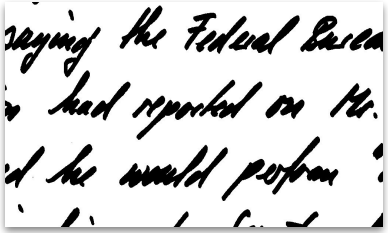
Original image (cropped)



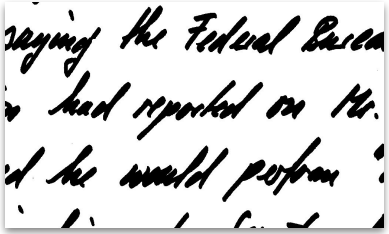
Niblack



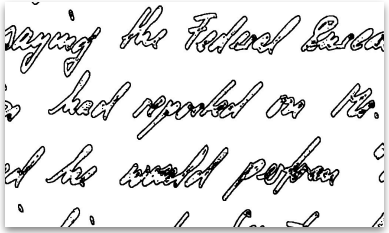
Otsu



Sauvola

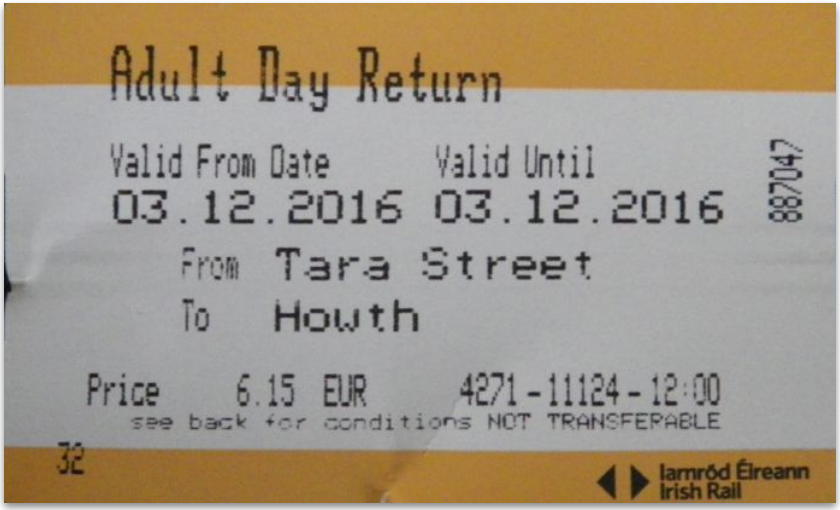


Wolf

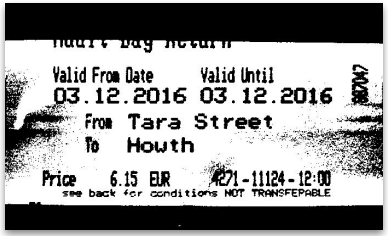


Su

Binarization

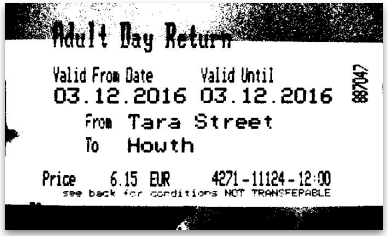


Original image (cropped)

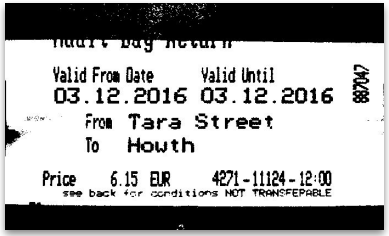


Niblack

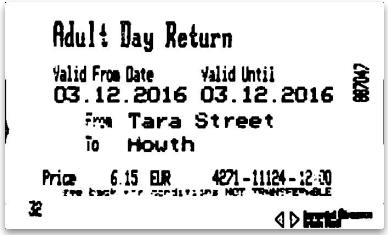
Otsu



Sauvola

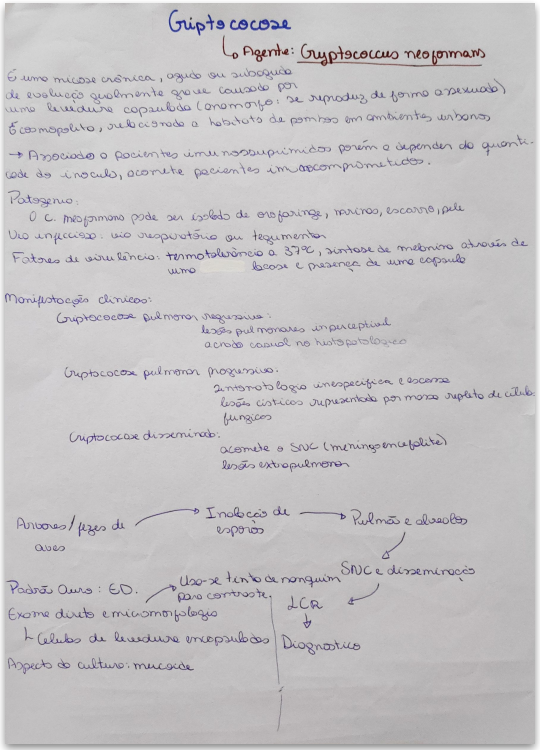


Wolf

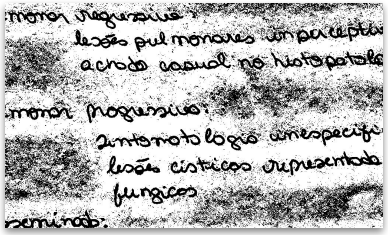


Su

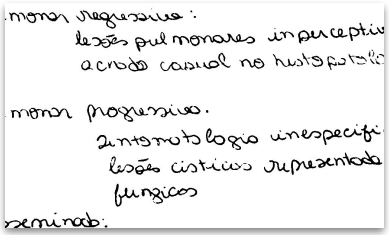
Binarization



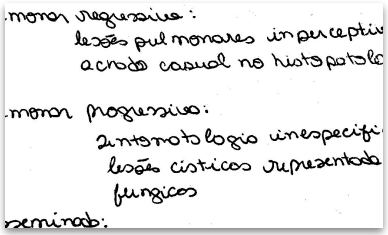
Original image (cropped)



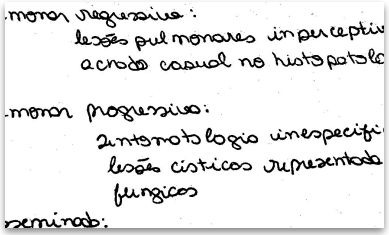
Niblack



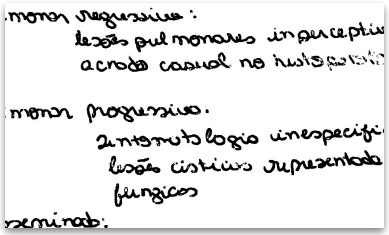
Otsu



Sauvola

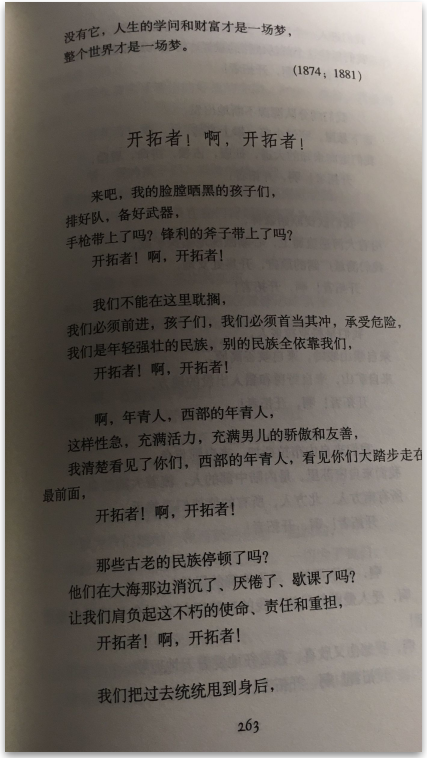


Wolf



Su

Binarization


没有它，人生的学问和财富才是一场梦，
整个世界才是一场梦。

(1874，1881)

开拓者！啊，开拓者！

来吧，我的脸膛晒黑的孩子们，
排好队，备好武器，
手枪带上了吗？锋利的斧子带上了吗？
开拓者！啊，开拓者！

我们不能在这里耽搁，
我们必须前进，孩子们，我们必须首当其冲，承受危险，
我们是年轻强壮的民族，别的民族全依靠我们，
开拓者！啊，开拓者！

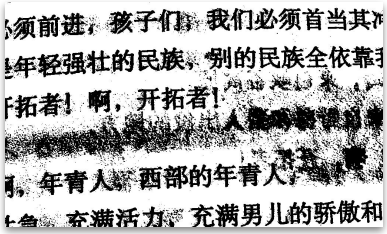
啊，年青人，西部的年青人，
这样性急，充满活力，充满男儿的骄傲和友善，
我清楚看见了你们，西部的年青人，看见你们大踏步走在
最前面，
开拓者！啊，开拓者！

那些古老的民族停顿了吗？
他们在大海那边消沉了、厌倦了、歇课了吗？
让我们肩负起这不朽的使命、责任和重担，
开拓者！啊，开拓者！

我们把过去统统甩到身后，

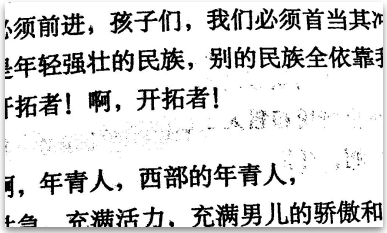
263

Original image (cropped)

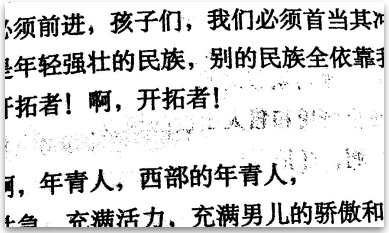


Niblack

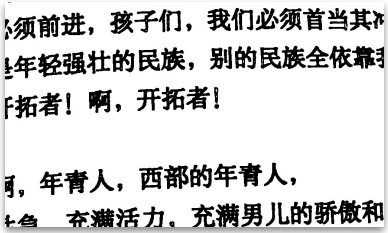
Otsu



Sauvola



Wolf

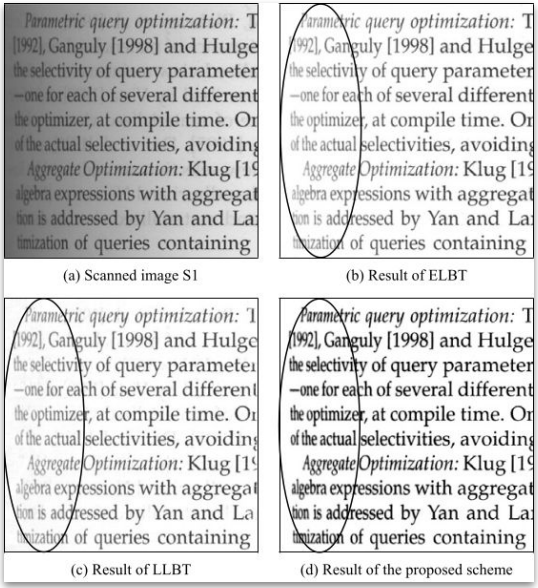


Su

Illumination Compensation

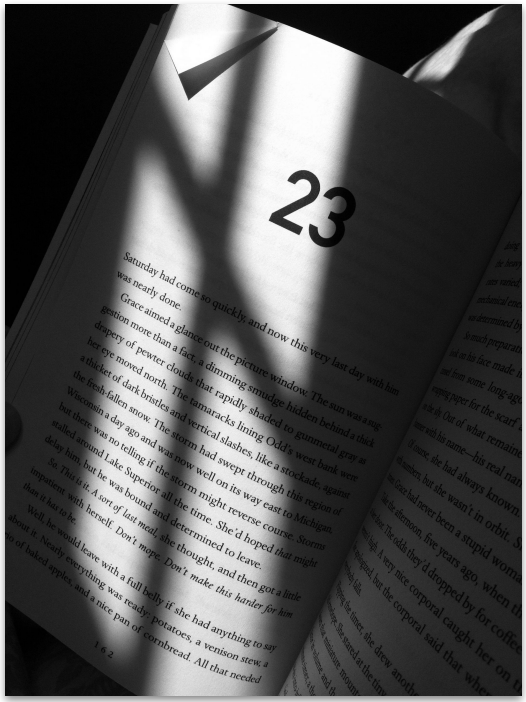
Image illumination leveling process to aid in the process of adaptive binarization.

Paper: Efficient illumination compensation techniques for text images (2012).



Result with physical scanned image

Illumination Compensation



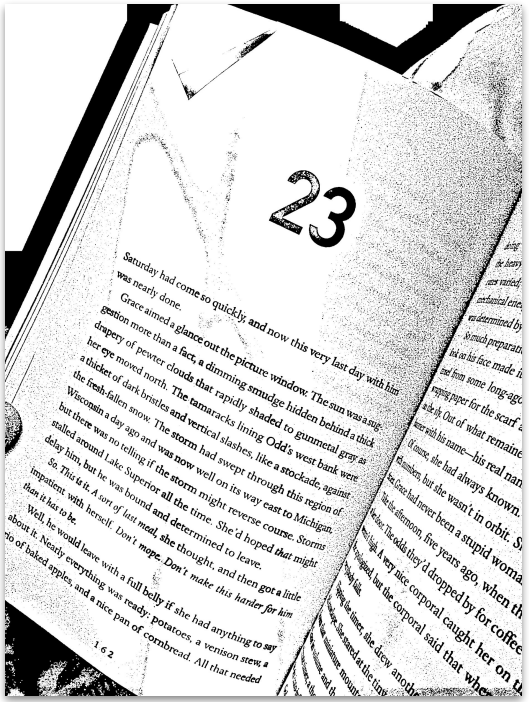
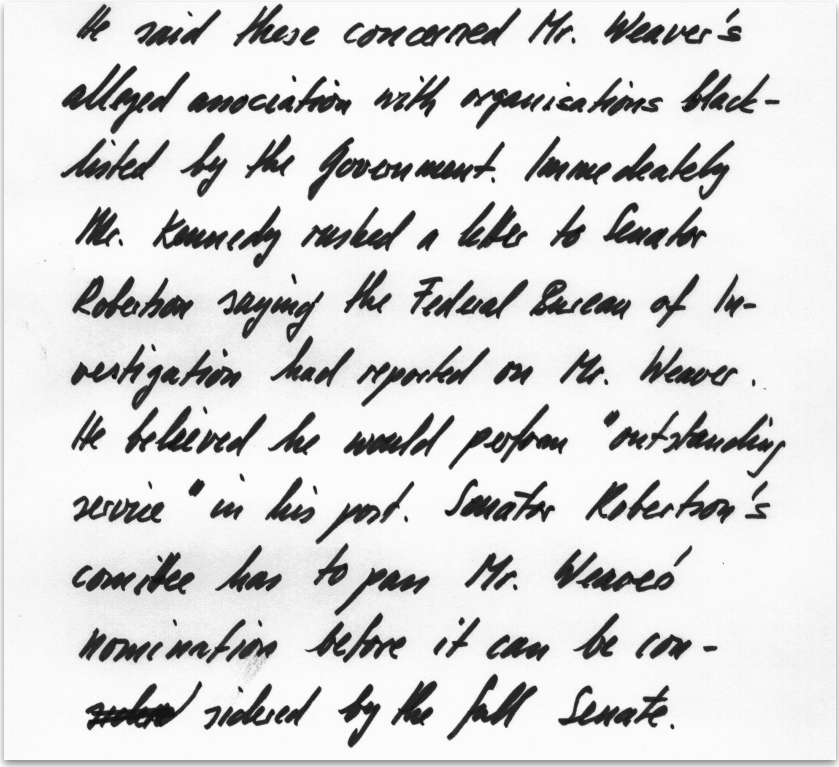
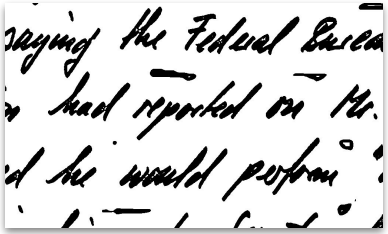


Image test with lots of shadow

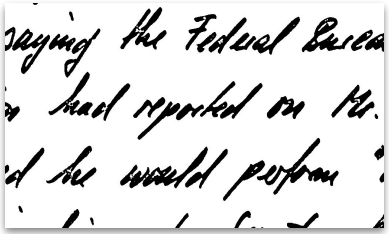
Binarization + illu. comp.



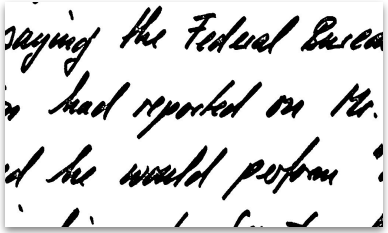
Original image (cropped)



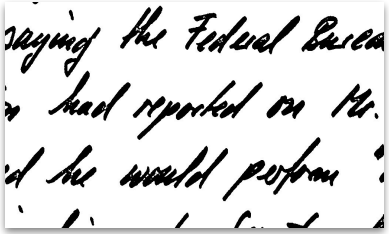
Niblack



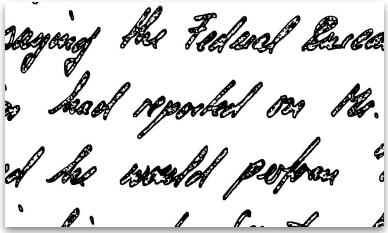
Otsu



Sauvola

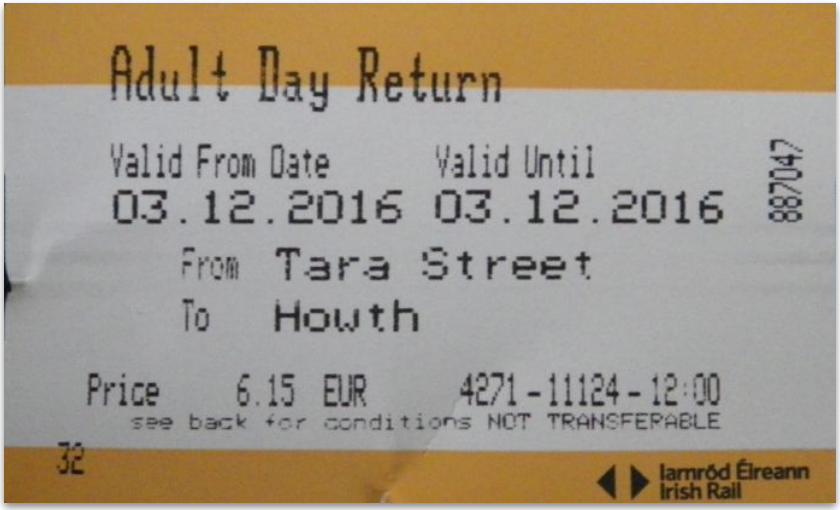


Wolf

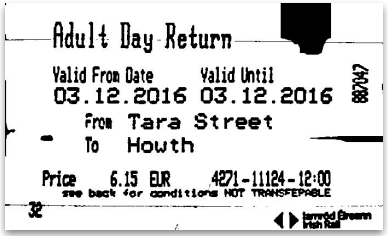


Su

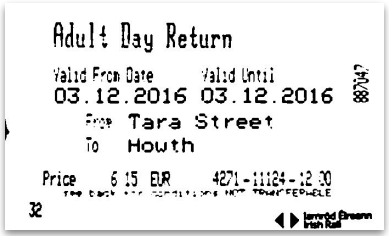
Binarization + illu. comp.



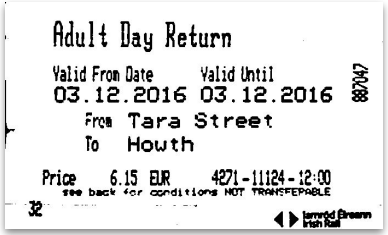
Original image (cropped)



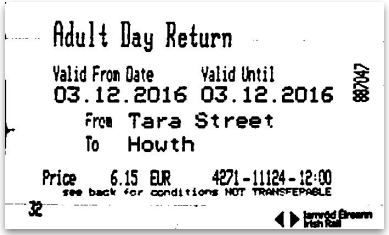
Niblack



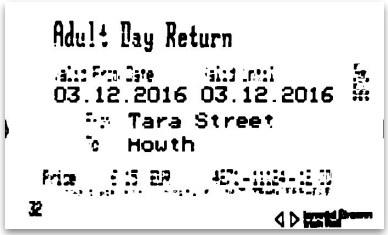
Otsu



Sauvola

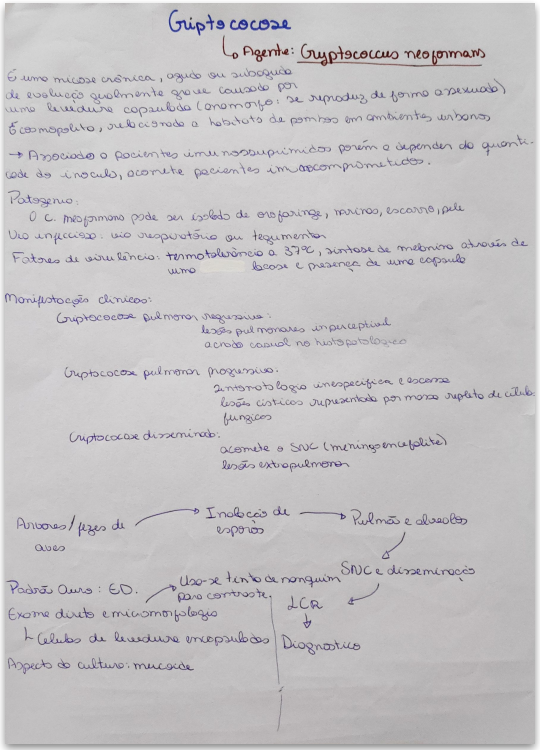


Wolf

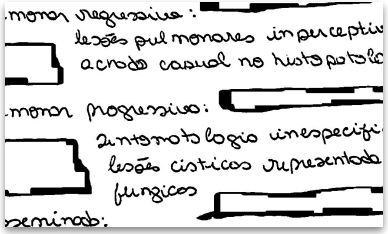


Su

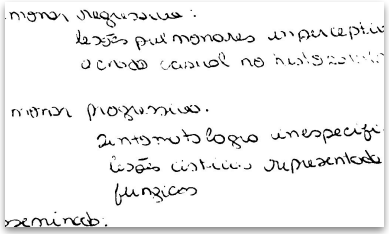
Binarization + illu. comp.



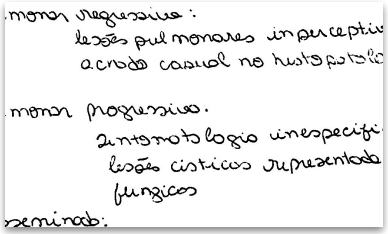
Original image (cropped)



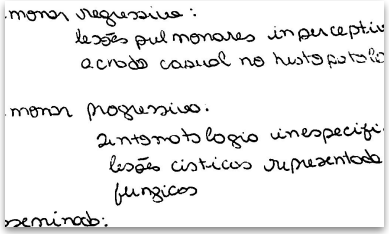
Niblack



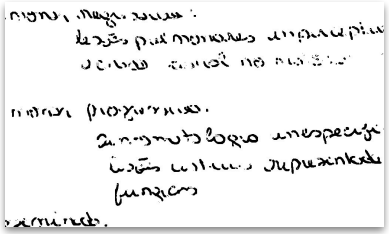
Otsu



Sauvola

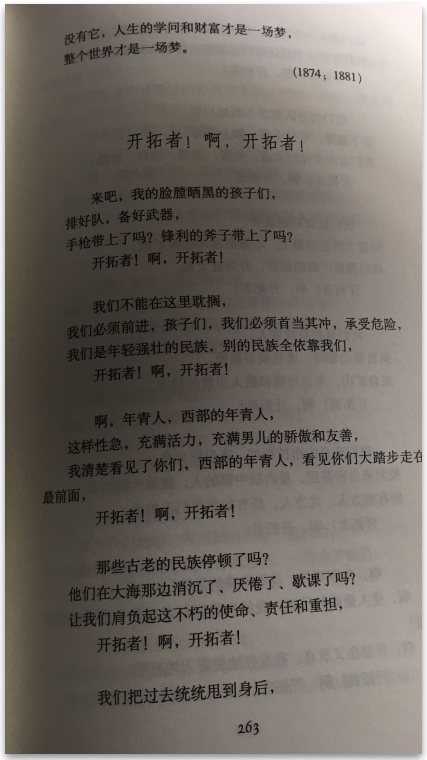


Wolf



Su

Binarization + illu. comp.



没有它，人生的学问和财富才是一场梦，
整个世界才是一场梦。

（1874，1881）

开拓者！啊，开拓者！

来吧，我的脸膛晒黑的孩子们，
排好队，备好武器，
手枪带上了吗？锋利的斧子带上了吗？
开拓者！啊，开拓者！

我们不能在这里耽搁，
我们必须前进，孩子们，我们必须首当其冲，承受危险，
我们是年轻强壮的民族，别的民族全依靠我们，
开拓者！啊，开拓者！

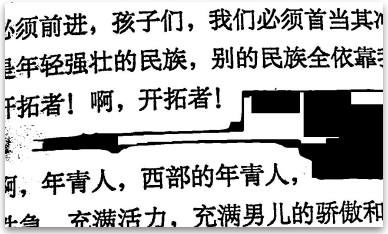
啊，年青人，西部的年青人，
这样性急，充满活力，充满男儿的骄傲和友善，
我清楚看见了你们，西部的年青人，看见你们大踏步走在
最前面，
开拓者！啊，开拓者！

那些古老的民族停顿了吗？
他们在大海那边消沉了、厌倦了、歇课了吗？
让我们肩负起这不朽的使命、责任和重担，
开拓者！啊，开拓者！

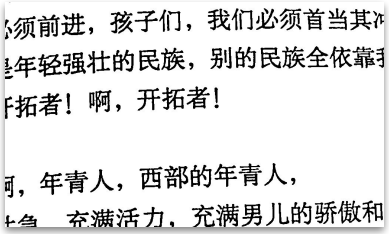
我们把过去统统甩到身后，

263

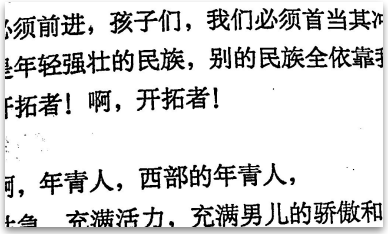
Original image (cropped)



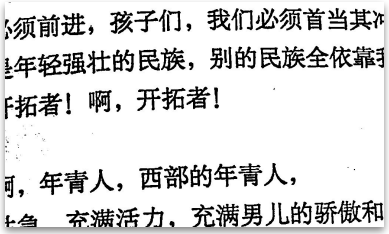
Niblack



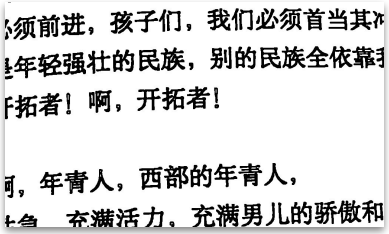
Otsu



Sauvola



Wolf



Su

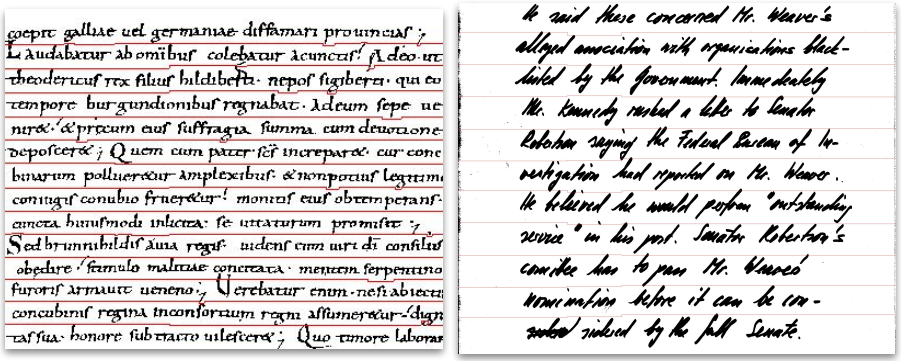
Segmentation

Paper: Text-Based Image Segmentation Methodology, 2014.

- ➔ Levels of text segmentation (line, word, character);
- ➔ Segmentation Methodologies (pixel counting approach, histogram approach, Y histogram projection, text line separation, false line exclusion, line region recovery, smearing approach, stochastic approach, water flow approach);

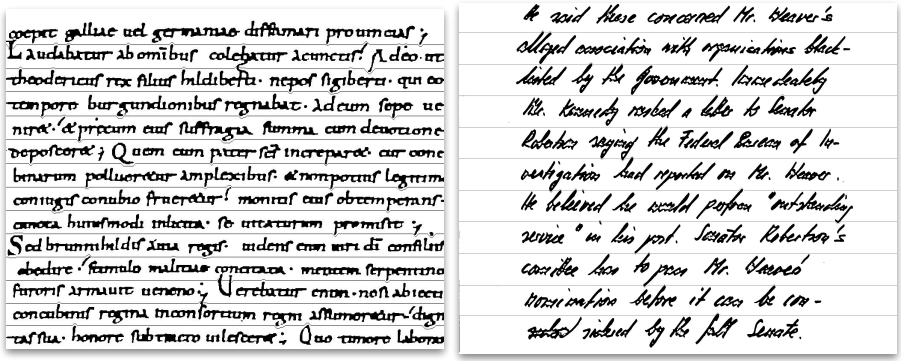
Line Segmentation

A Statistical approach to line segmentation in handwritten documents, 2007



Left image (400x310), binarization and line segmentation;
Right image (2479x1985), binarization and line segmentation;
Total time: 3s.

A* Path Planning for Line Segmentation of Handwritten Documents, 2014



Left image (400x310), binarization and line segmentation;
Right image (2479x1985), binarization and line segmentation;
Total time: 50s.

Deslanting Image

Removes the cursive writing style. It is used as a preprocessing step for handwritten text recognition.
Paper: A New Normalization Technique For Cursive Handwritten Words, 2001.

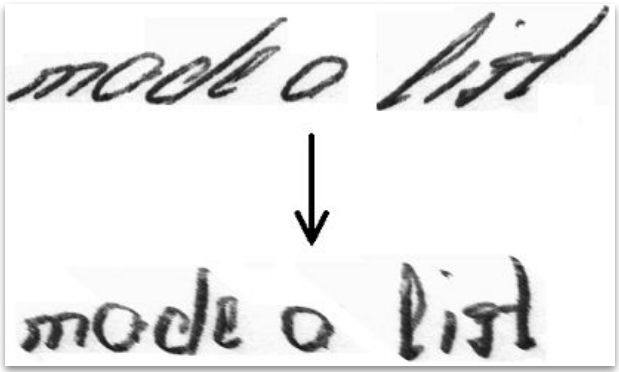


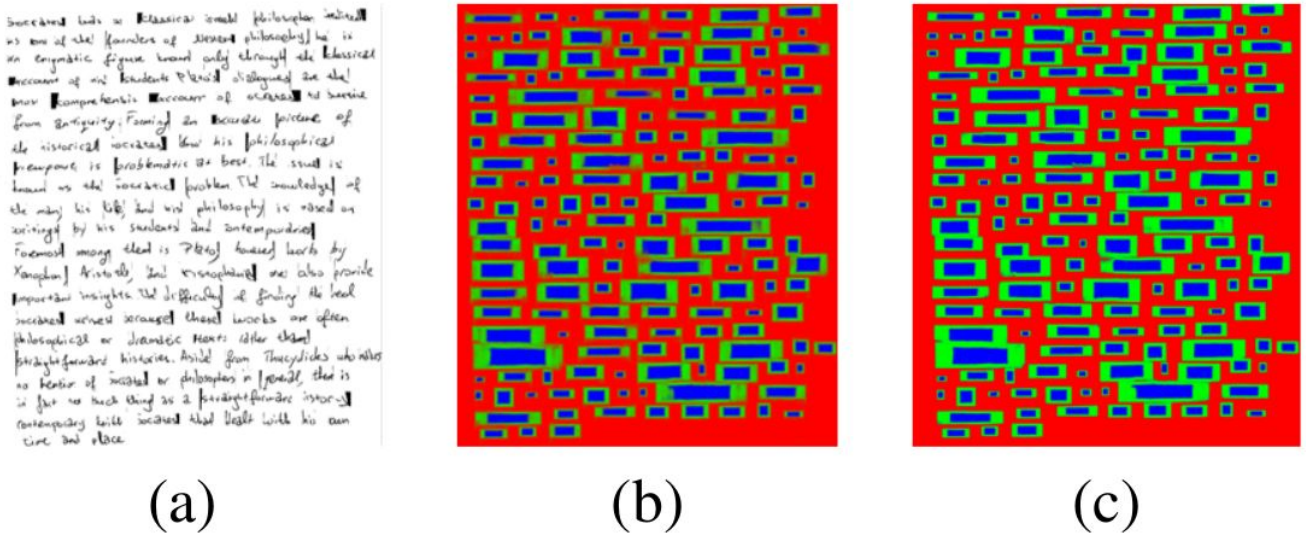
Image test, input (top) and output (bottom)

Word Segmentation

- ➔ 1999: Scale Space Technique for Word Segmentation in Handwritten Documents;
- ➔ 2015: Word Segmentation Method for Handwritten Documents based on Structured Learning;
- ➔ 2018: Keyword spotting in historical handwritten documents based on graph matching;
- ➔ 2018: Toward a Dataset Agnostic Word Segmentation Method.

Word Segmentation

Toward a Dataset Agnostic Word Segmentation Method (2018)



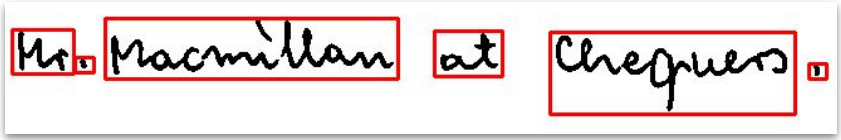
(a) (b) (c)

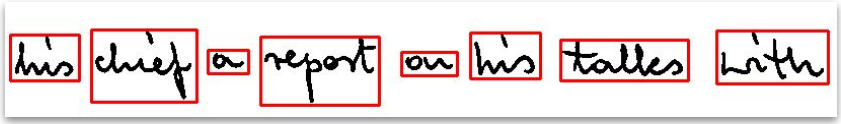
A sample document from the ICDAR dataset (a), heatmap generated by the heatmap network (b), and a smooth heatmap generated by the smoother network (c).

Word Segmentation

Implementation of scale space technique for word segmentation as proposed by R. Manmatha and N. Srimal. Even though the paper is from 1999, the method still achieves good results, is fast, and is easy to implement.

Harald's implementation approach
(https://github.com/githubharald/WordSegmentation)





Harald's implementation approach with adjusts

Segmentation

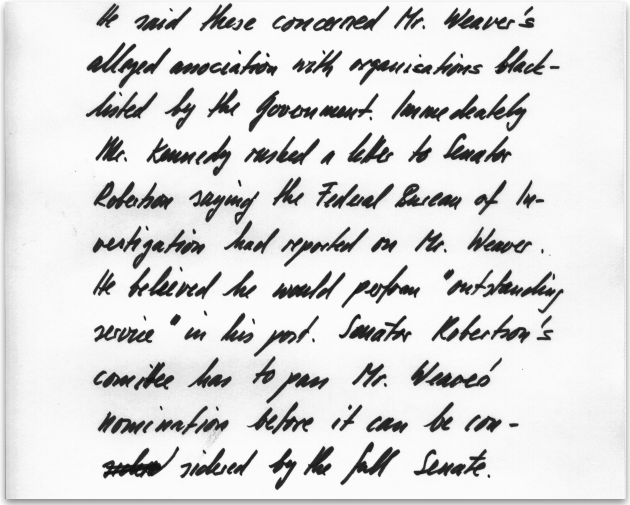
Compressed image segmentation approach:

- ➔ 2013: Extraction of Line-Word-Character Segments Directly from Run-Length Compressed Printed Text-Documents;
- ➔ 2018: Text line Segmentation in Compressed Representation of Handwritten Document using Tunneling Algorithm;
- ➔ 2018: Word Segmentation Directly in Run-Length Compressed Handwritten Document Images using Connected Component Analysis.

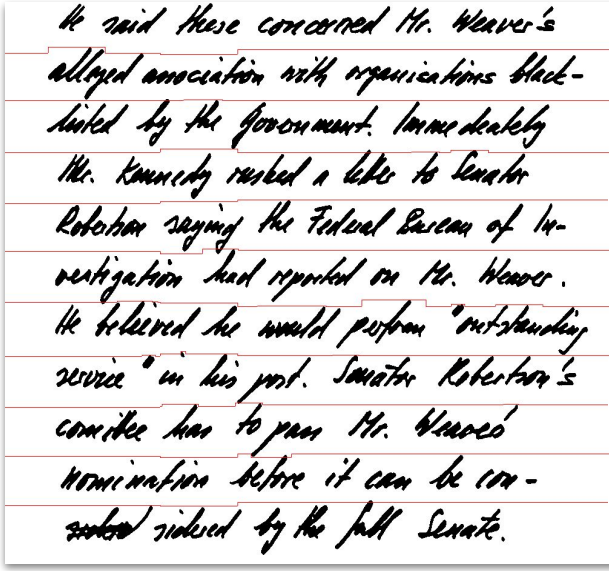
Results

- ➔ 10 images have been tested;
- ➔ Each of the images had a different challenge to deal with;
- ➔ The techniques were implemented to obtain good results in all.

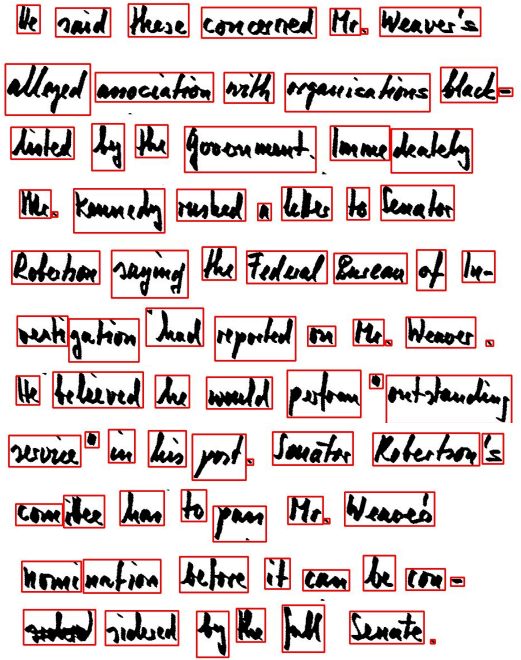
Results



Original image

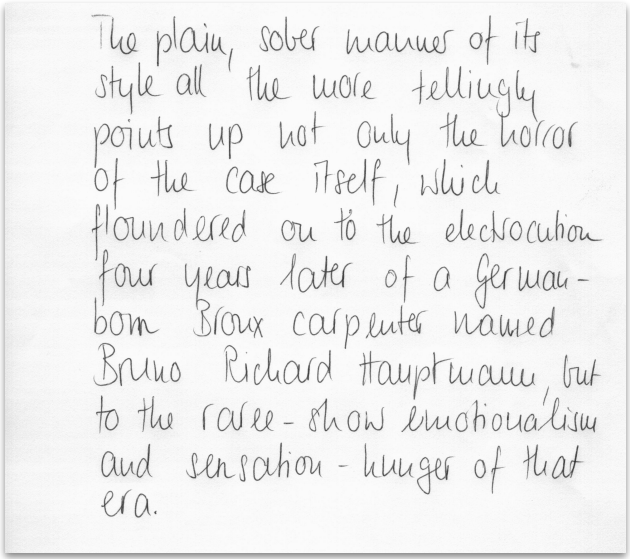


Line Segmentation

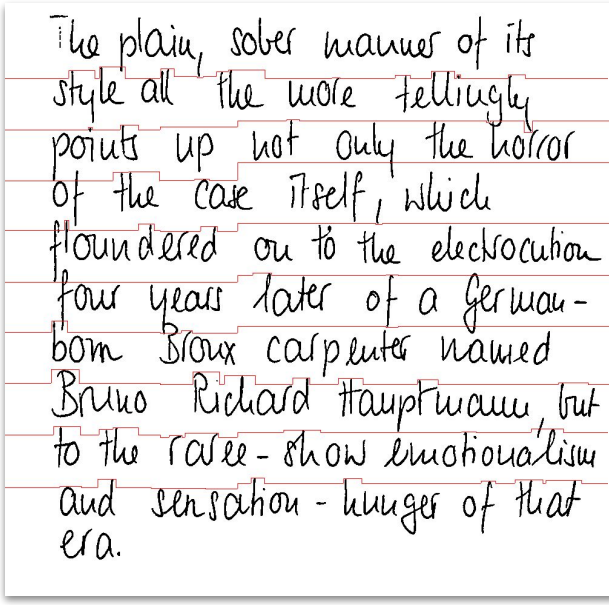


Word Segmentation

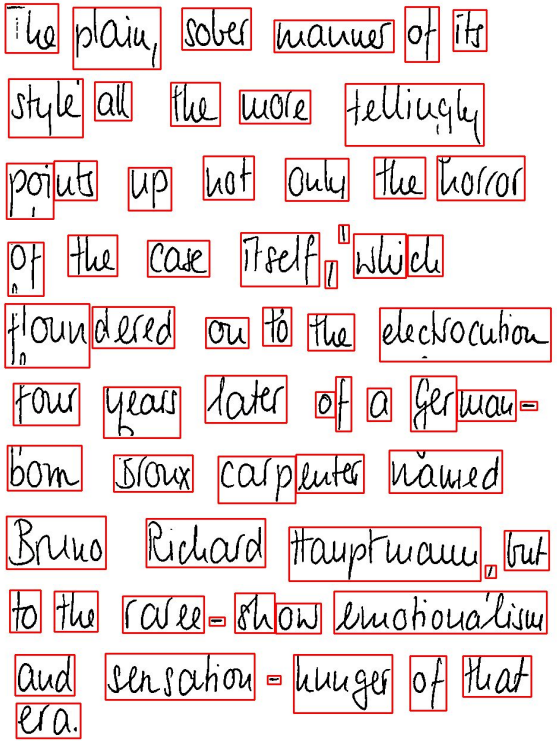
Results



Original image

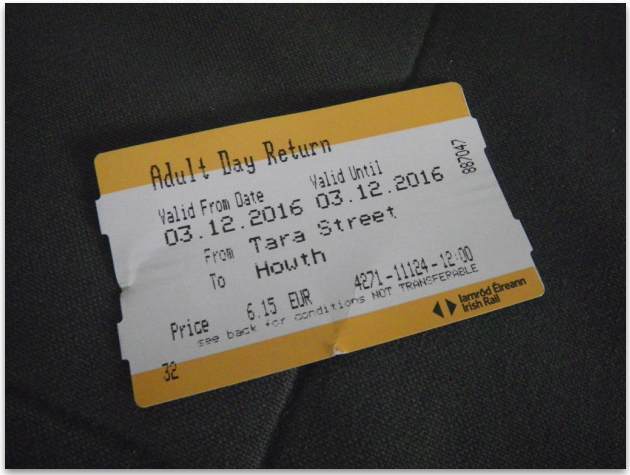


Line Segmentation

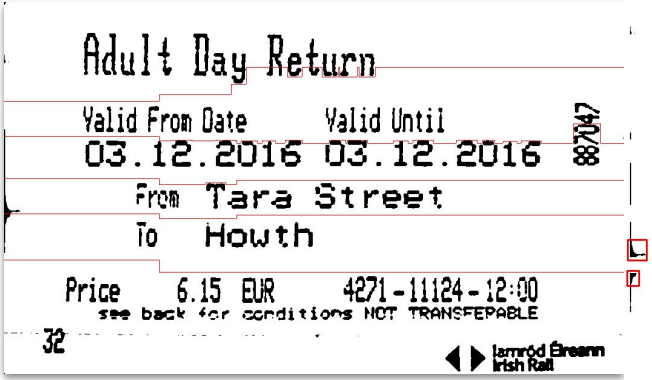


Word Segmentation

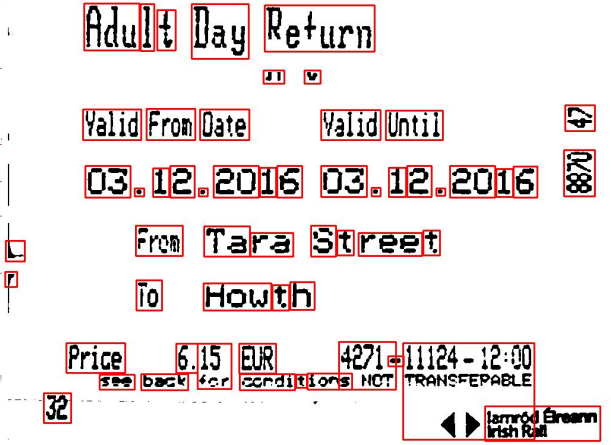
Results



Original image

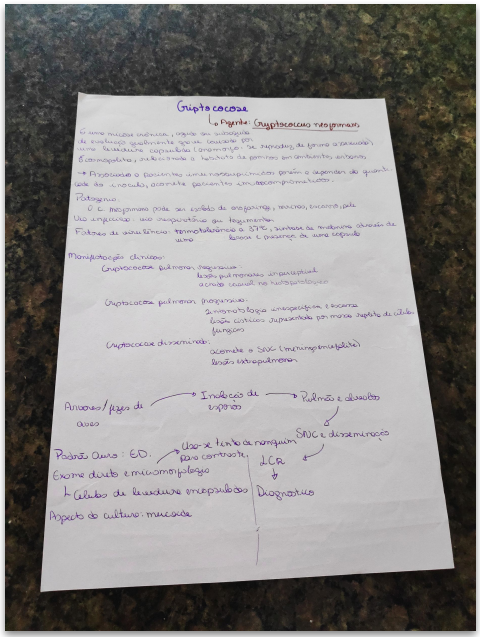


Line Segmentation

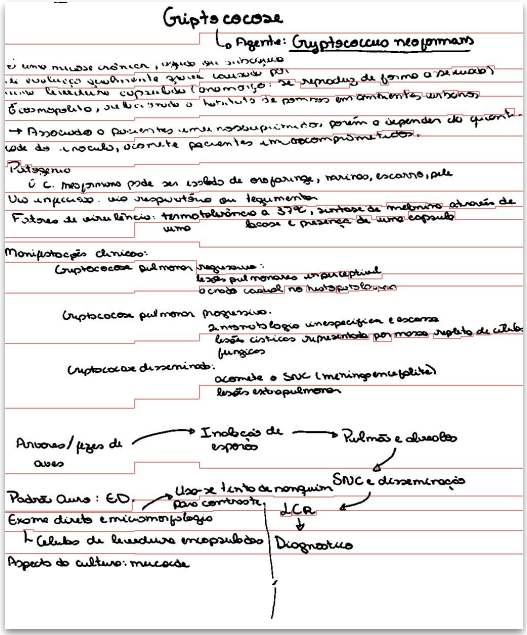


Word Segmentation

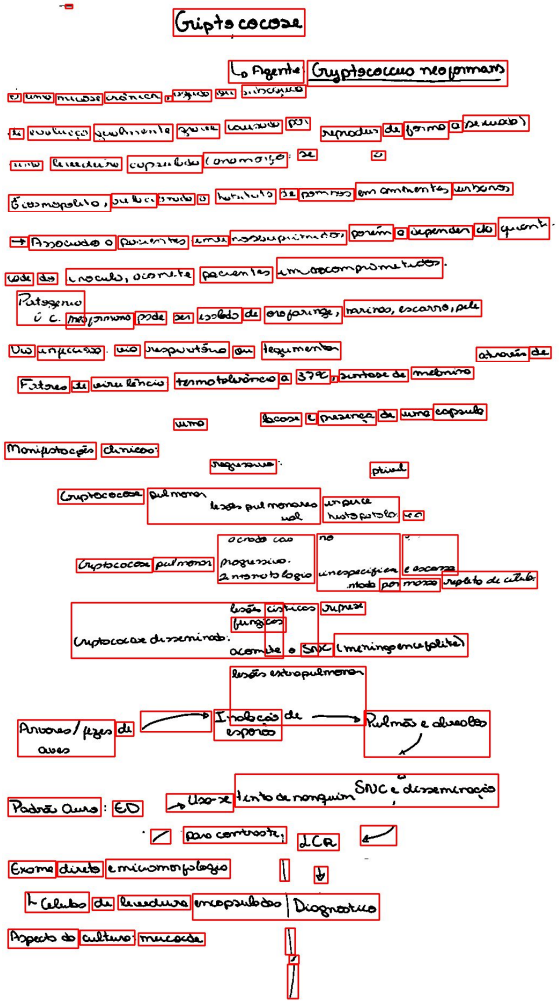
Results



Original image



Line Segmentation



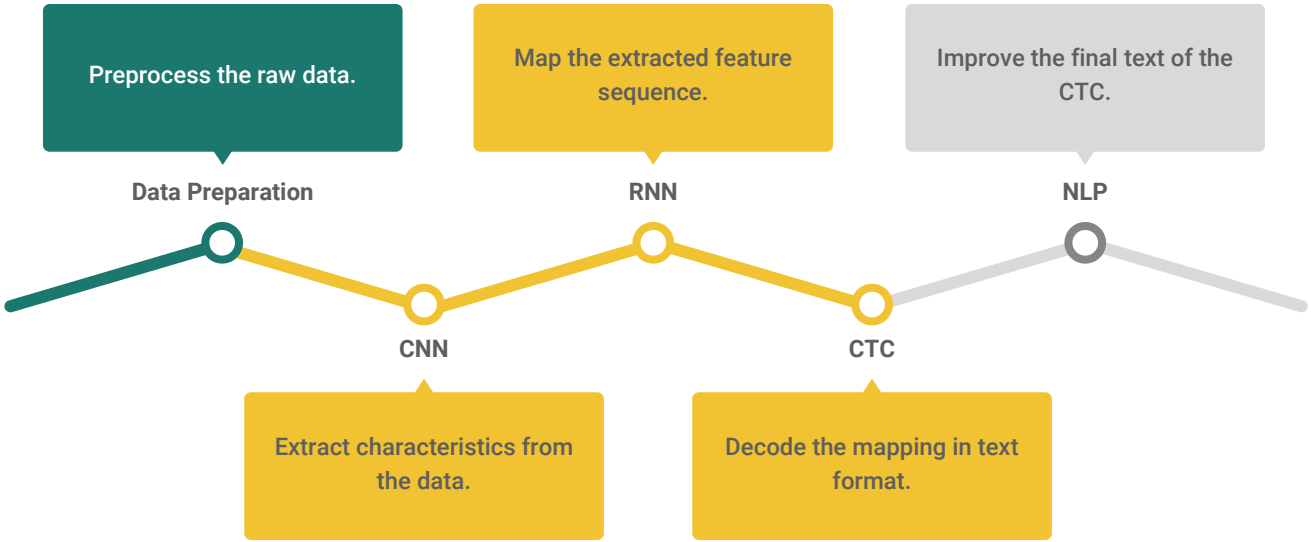
Word Segmentation

Conclusions

- ➔ Good results for simple manuscripts (like IAM dataset), with static code to line and word segmentation;
- ➔ Ambient noise problems (i.e. book border);
- ➔ Bad segmentations (line or/and word) compromises the dissertation project focus (NLP).

Upcoming Activities

Study and implement with more detail the steps of CNN, RNN and CTC.



References

- ARIVAZHAGAN, Manivannan; SRINIVASAN, Harish; SRIHARI, Sargur. **A Statistical approach to line segmentation in handwritten documents**. The International Society for Optical Engineering, 2007;
- AXLER, Gregory; WOLF, Lior. **Toward a Dataset Agnostic Word Segmentation Method**. 2018 25th IEEE International Conference on Image Processing, 2018;
- CHEN, Kuo-Nan; CHEN, Chin-Hao; CHANG, Chin-Chen. **Efficient illumination compensation techniques for text images**. Digital Signal Processing, v. 22, p. 726-733, 2012;
- JAVED, Mohammed; NAGABHUSHAN, P.; CHAUDHURI, B. **Extraction of Line-Word-Character Segments Directly from Run-Length Compressed Printed Text-Documents**. 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, 2013;
- LAZZARA, Guillaume; GÉRAUD, Thierry. **Efficient Multiscale Sauvola's Binarization**. 2014 International Journal on Document Analysis and Recognition manuscript, 2014;

References

- MANMATHA, R.; SRIMAL, Nitin. **Scale Space Technique for Word Segmentation in Handwritten Documents**. SCALE-SPACE '99 Proceedings of the Second International Conference on Scale-Space Theories in Computer Vision, p. 22-33, 1999;
- MEHUL, Gupta; ANKITA, Patel; NAMRATA, Dave; RAHUL, Goradia; SHETH, Saurin. **Text-Based Image Segmentation Methodology**. 2nd International Conference on Innovations in Automation and Mechatronics Engineering, 2014;
- NIBLACK, Wayne. **An Introduction to Digital Image Processing**. Strandberg Publishing Company Birkeroed, Denmark, 1986;
- NTIROGIANNIS, K.; GATOS, B.; PRATIKAKIS, I. **A Combined Approach for the Binarization of Handwritten Document Images**. Pattern Recognition Letters 35, 2014;
- OTSU, Nobuyuki. **A Threshold Selection Method from Gray-Level Histograms**. IEEE Transactions on Systems, Man and Cybernetics 9, 62-66, 1979;

References

- ROSEBROCK, Adrian. **Build a Kick-Ass Mobile Document Scanner**, 2014. Disponível em <https://www.pyimagesearch.com/2014/09/01/build-kick-ass-mobile-document-scanner>. Acessado em 27 de fevereiro de 2019;
- RYU, Jewoong; KOO, Hyung Il; CHO, Nam Ik. **Word Segmentation Method for Handwritten Documents based on Structured Learning**. IEEE Signal Processing Letters, v. 22, 2015;
- SAUVOLA, Jaakko; SEPPANEN, Tapio; HAAPAKOSKI, Sami; PIETIKAINEN, Matti. **Adaptive Document Binarization**. IEEE Computer Society Washington, 1997;
- STAUFFER, Michael; FISCHER, Andreas; RIESEN, Kaspar. **Keyword Spotting in Historical Handwritten Documents Based on Graph Matching**. Pattern Recognition 81, p. 240-253, 2018;
- SU, Bolan; LU, Shijian; TAN, Chew Lim. **Binarization of Historical Document Images Using the Local Maximum and Minimum**. Document Analysis Systems, 2010;

References

- SURINTA, Olarik; HOLTKAMP, Michiel; KARABAA, Faik; OOSTEN, Jean-Paul van; SCHOMAKER, Lambert; WIERING, Marco. **A* Path Planning for Line Segmentation of Handwritten Documents**. 14th International Conference on Frontiers in Handwriting Recognition, 2014;
- VEDERE, Amarnath R.; NAGABHUSHAN, P. **Text line Segmentation in Compressed Representation of Handwritten Document using Tunneling Algorithm**. International Journal of Intelligent Systems and Applications in Engineering, p. 251-261, 2018;
- VEDERE, Amarnath R.; NAGABHUSHAN, P.; JAVED, Mohammed. **Word Segmentation Directly in Run-Length Compressed Handwritten Document Images using Connected Component Analysis**. IET Image Processing, 2018;
- VINCIARELLI, Alessandro; LUETTIN, Juergen. **A New Normalization Technique for Cursive Handwritten Words**. Pattern Recognition Letters 22, 2001;

References

➔ WOLF, Cristian; JOLION, Jean-Michel; CHASSAING, Françoise. **Text Localization, Enhancement and Binarization in Multimedia Documents**. International Conference on Pattern Recognition (ICPR), v. 4, p. 1037-1040, 2002.