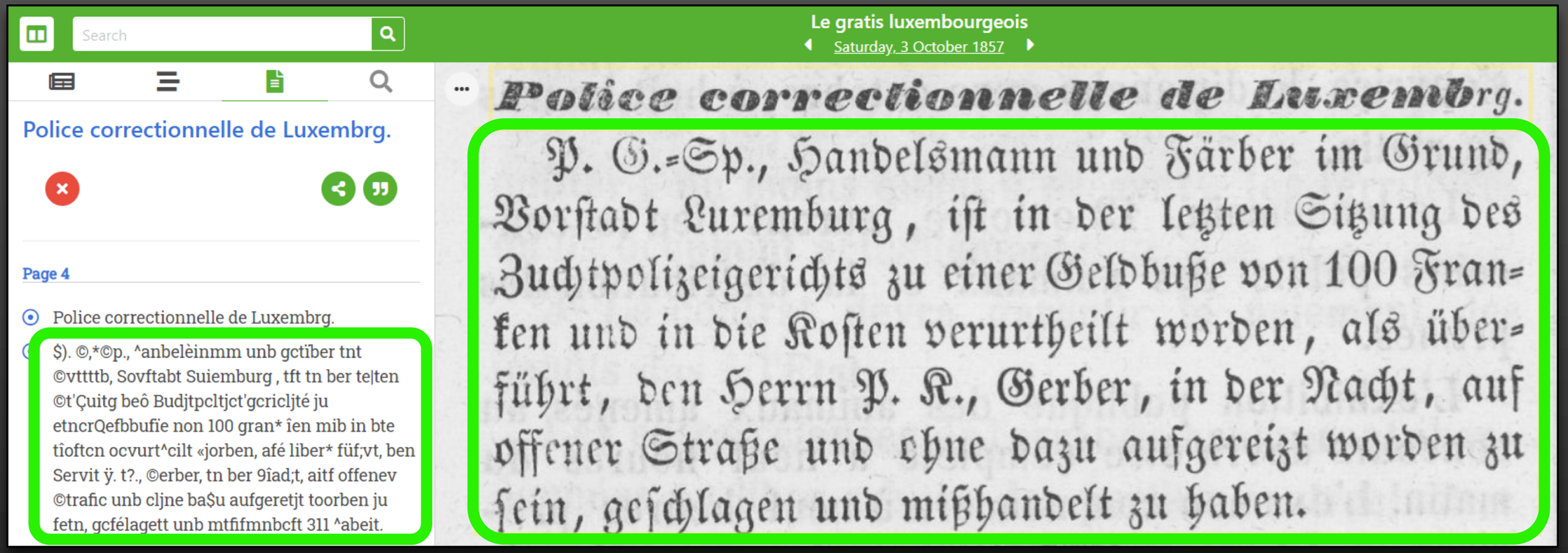
ARTIFICIAL INTELLIGENCE

at BnL

Pit Schneider | 11/12/2020

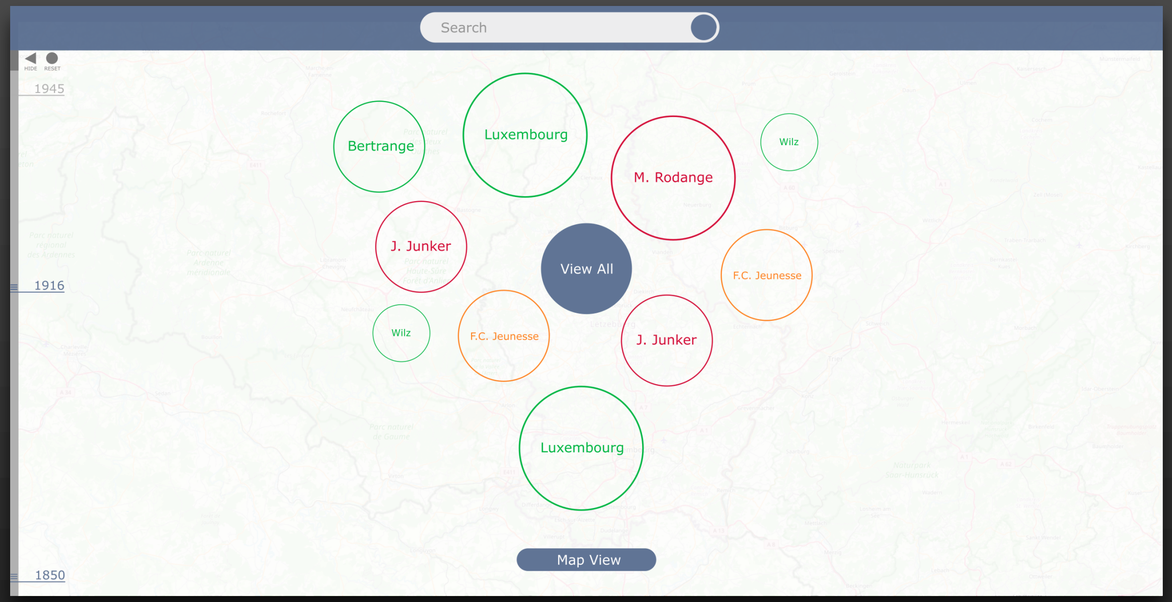
Bibliothèque nationale du Luxembourg
Nationalbibliothéik

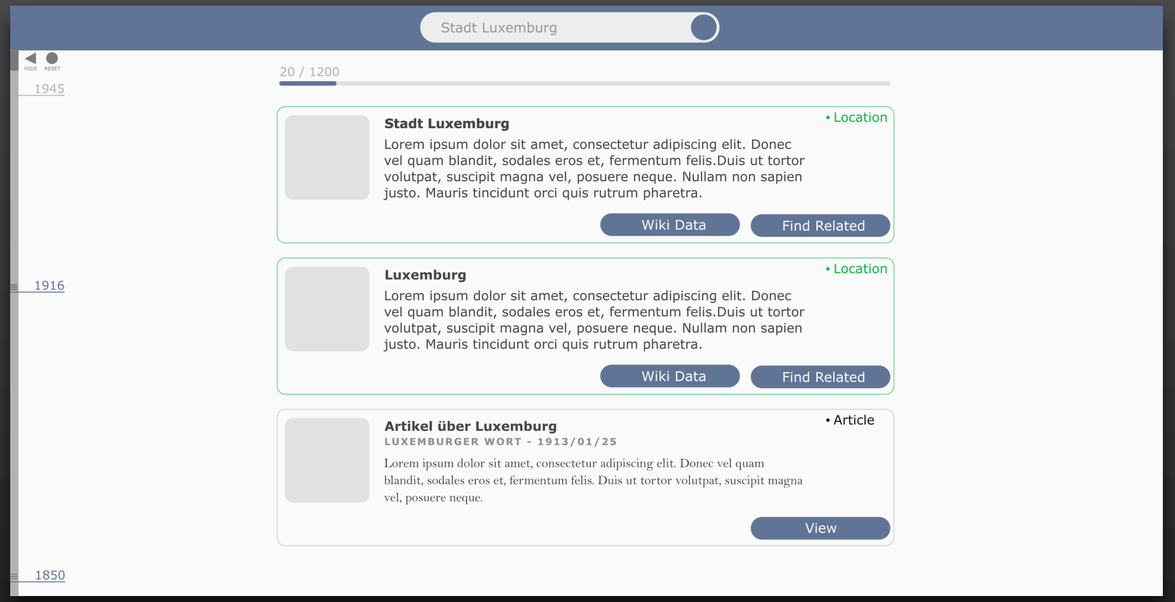


eLuxemburgensia

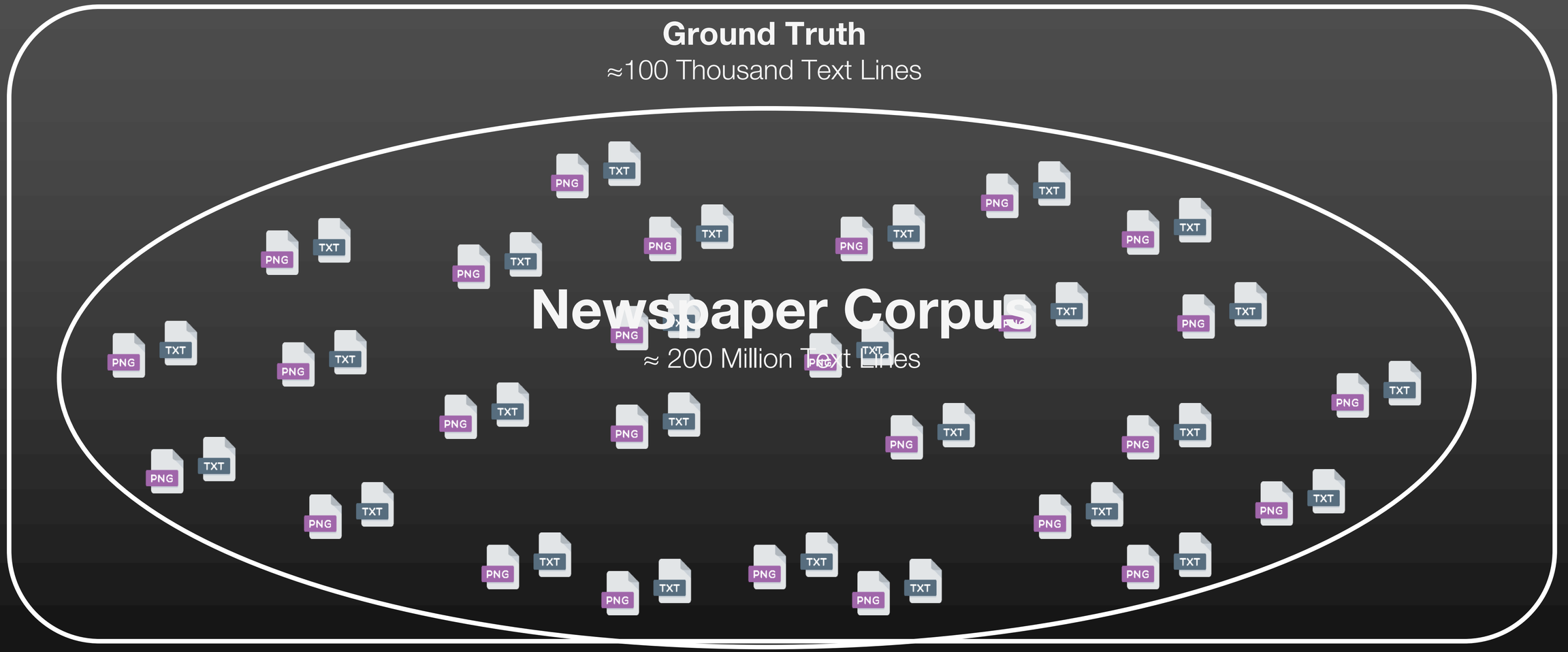
1) Improve Optical Character Recognition (OCR)

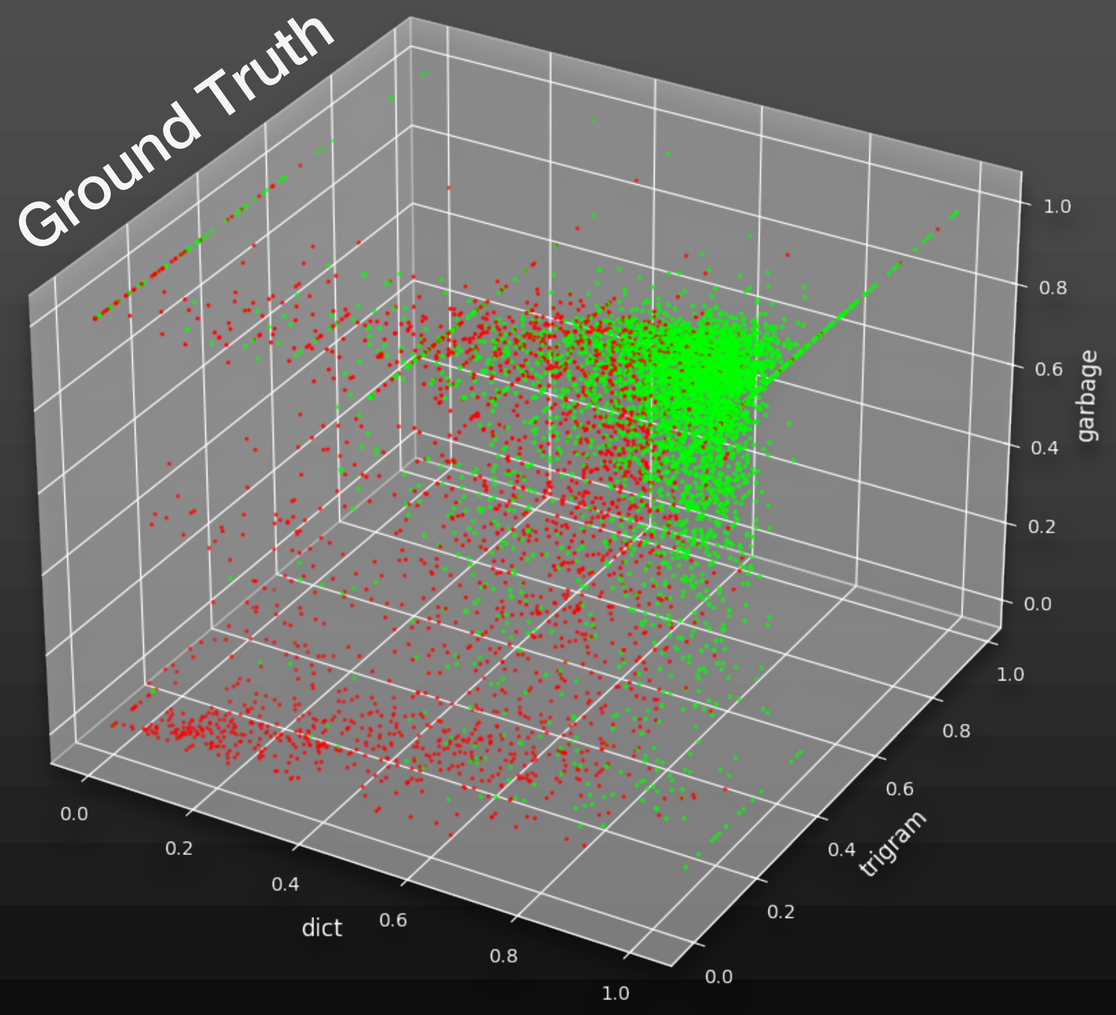
2) Improve Newspaper Exploration





named entities • entity relations • timeline • maps • wikidata


Ground Truth
≈100 Thousand Text Lines
Newspaper Corpus
≈ 200 Million Text Lines



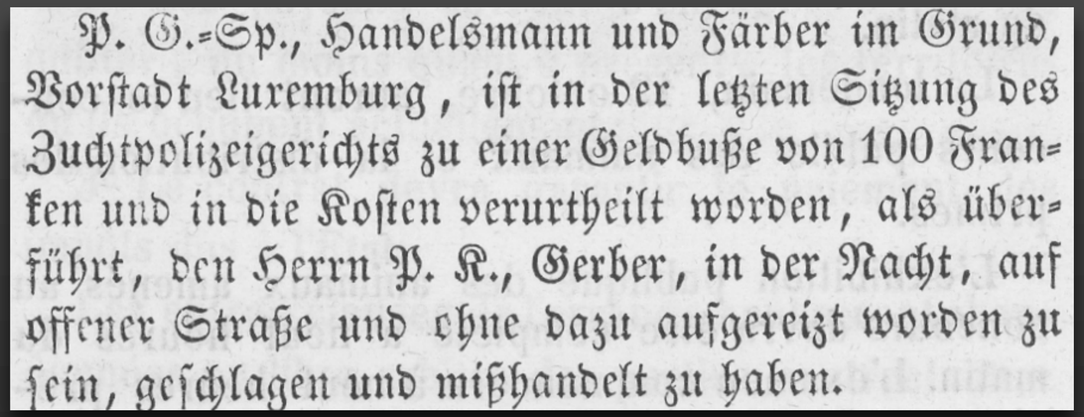
score >= 0.95

else

score = 1 - (editDistance / |chars|)

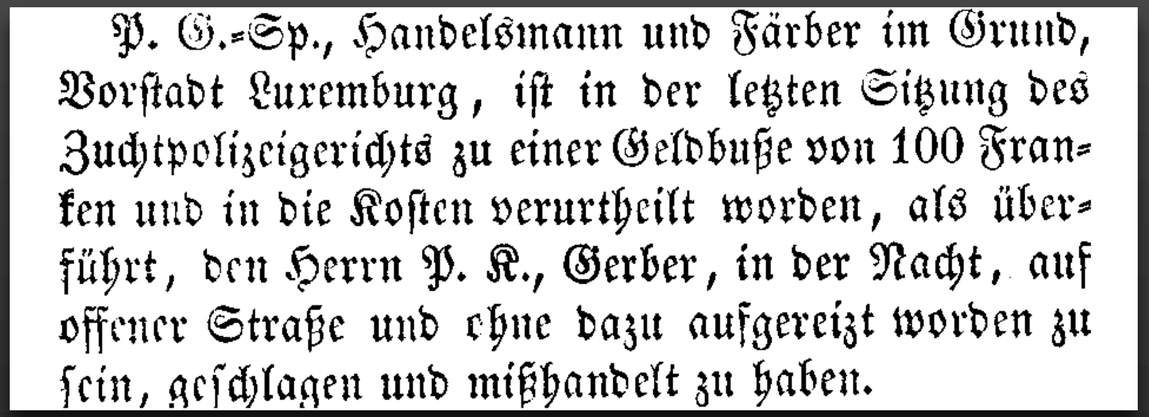
85% accuracy
on test set

unsupervised quality evaluation • kNN algorithm


P. G.-Sp., Handelsmann und Färber im Grund, Vorstadt Luxemburg, ist in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Kosten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu sein, geschlagen und mißhandelt zu haben.

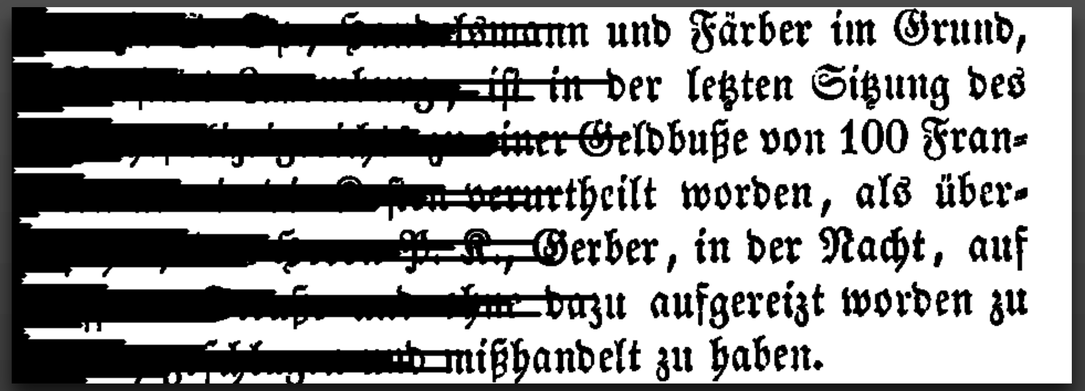
Original Text Block

OpenCV


P. G.-Sp., Handelsmann und Färber im Grund, Vorstadt Luxemburg, ist in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Kosten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu sein, geschlagen und mißhandelt zu haben.

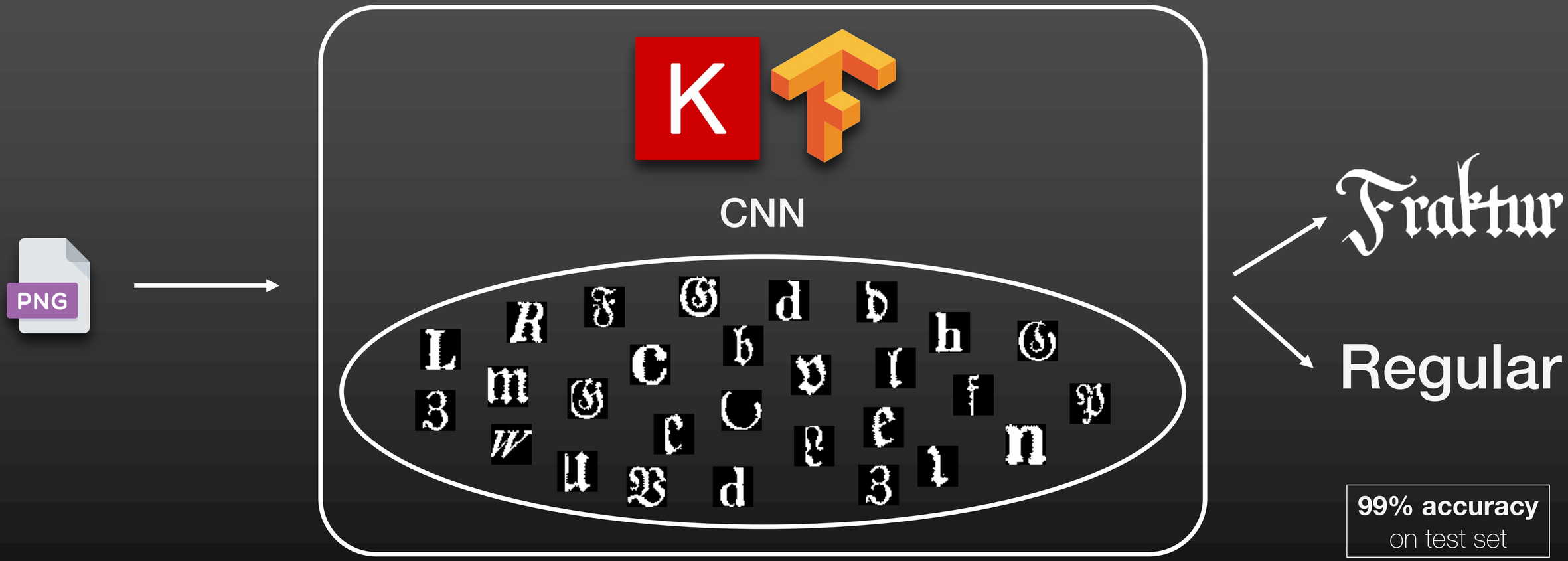
Binarized Text Block

cleaning • dilation • padding • inversion • white on black detection



P. G.-Sp., Handelsmann und Färber im Grund, Vorstadt Luxemburg, ist in der letzten Sitzung des Zuchtpolizeigerichts zu einer Geldbuße von 100 Franken und in die Kosten verurtheilt worden, als überführt, den Herrn P. K., Gerber, in der Nacht, auf offener Straße und ohne dazu aufgereizt worden zu sein, geschlagen und mißhandelt zu haben.

morphology • connected components • horizontal histogram projection



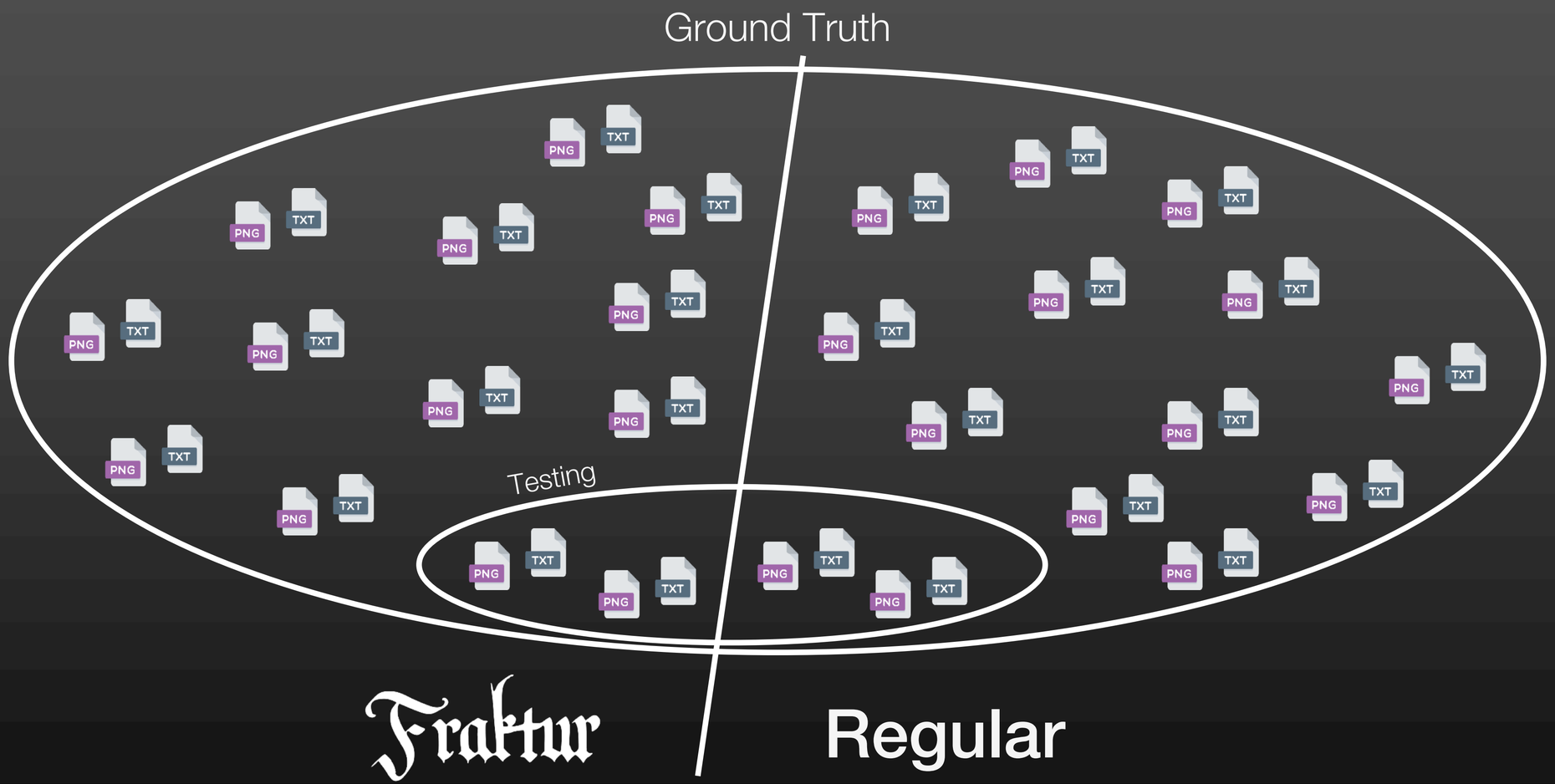
convolutional neural network • font recognition • binary classifier

kraken

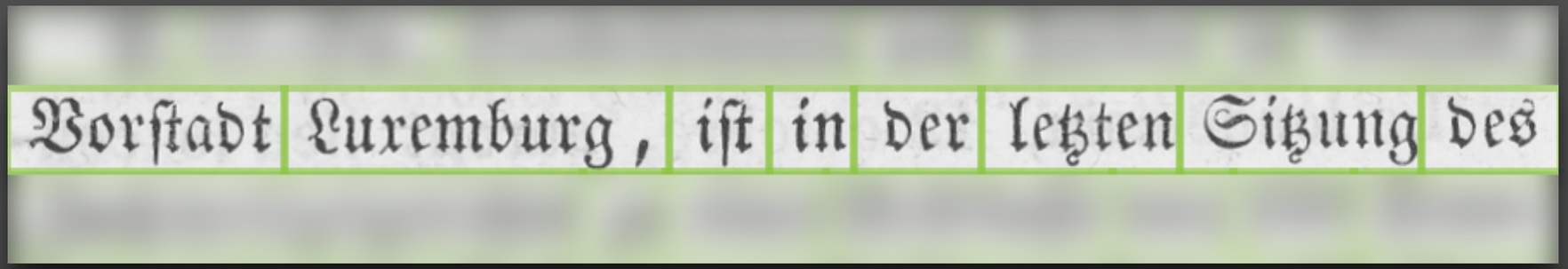
Tesseract

Calamari

open source software • custom model training

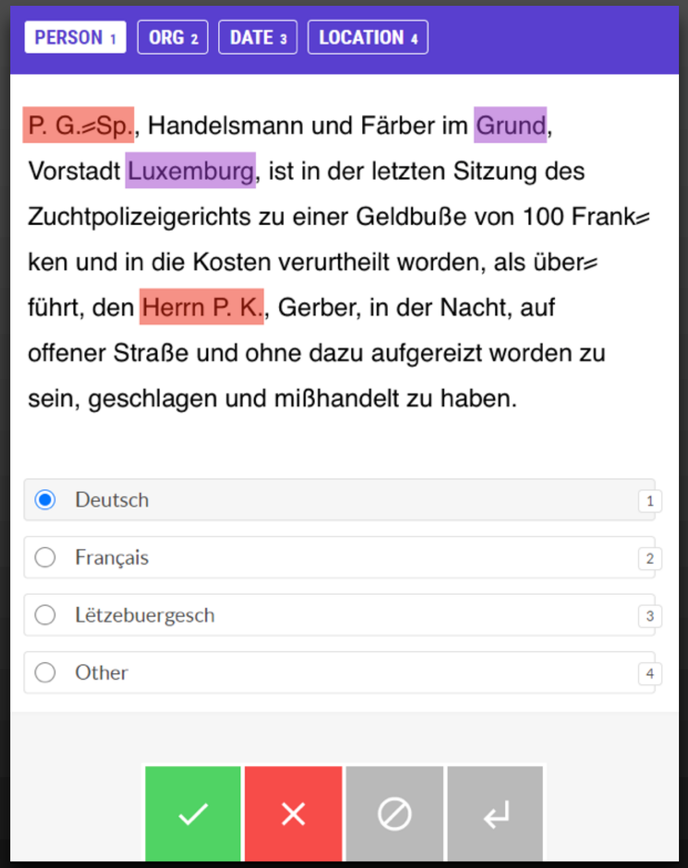


train/test sets • line/text pairs



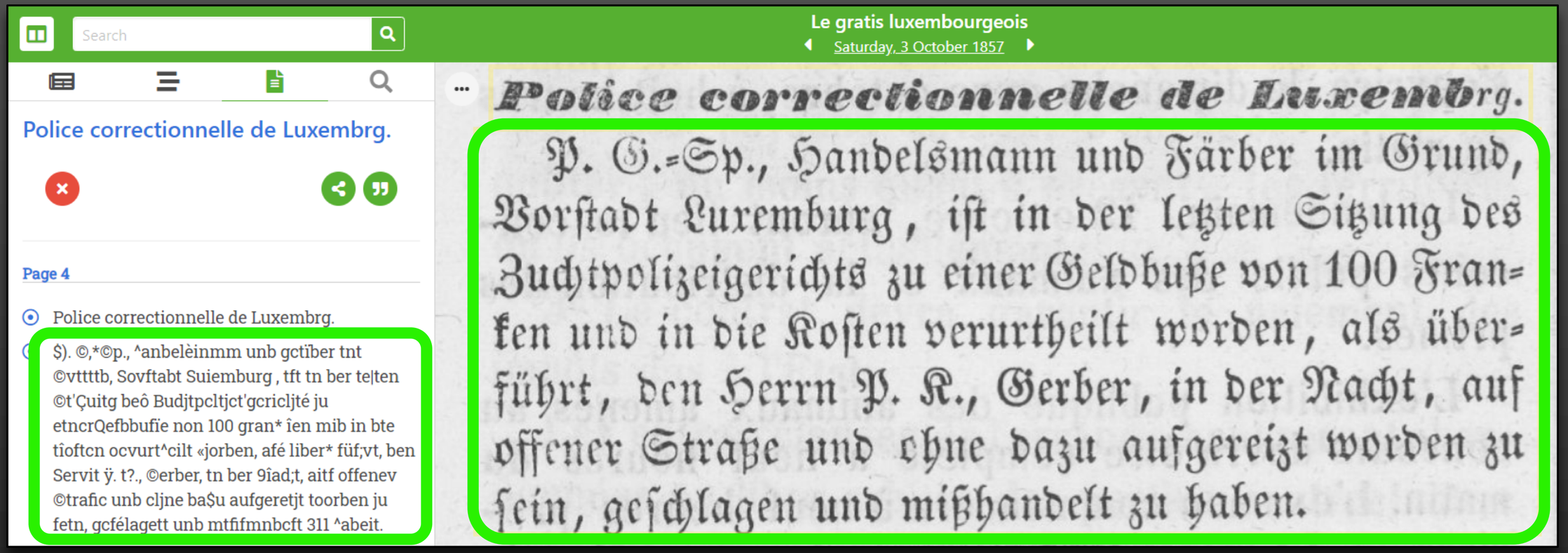

```
<TextLine HPOS="0" VPOS="41" WIDTH="761" HEIGHT="41">
    <String CONTENT="Vorstadt" CC="00000000" HPOS="0" VPOS="41" WIDTH="135" HEIGHT="41"/>
    <SP/>
    <String CONTENT="Luxemburg," CC="0000000000" HPOS="135" VPOS="41" WIDTH="189" HEIGHT="41"/>
    <SP/>
    <String CONTENT="ist" CC="000" HPOS="324" VPOS="41" WIDTH="49" HEIGHT="41"/>
    <SP/>
    <String CONTENT="in" CC="00" HPOS="373" VPOS="41" WIDTH="42" HEIGHT="41"/>
    <SP/>
    <String CONTENT="der" CC="000" HPOS="415" VPOS="41" WIDTH="63" HEIGHT="41"/>
    <SP/>
    <String CONTENT="letzten" CC="0000100" HPOS="478" VPOS="41" WIDTH="97" HEIGHT="41"/>
    <SP/>
    <String CONTENT="Sitzung" CC="1000000" HPOS="575" VPOS="41" WIDTH="118" HEIGHT="41"/>
    <SP/>
    <String CONTENT="des" CC="000" HPOS="693" VPOS="41" WIDTH="68" HEIGHT="41"/>
</TextLine>
```

word bounding boxes • ALTO XML schema



spaCy

ground truth generation • named entity recognition (NER)



eLuxemburgensia

1) Improve Optical Character Recognition (OCR)

2) Improve Newspaper Exploration

1) Improve Optical Character Recognition

Accuracy
Improvement for an estimated 30% of text blocks

Production
OCR application on entire corpus in coming weeks

2) Improve Newspaper Exploration

Named Entity Recognition
Finished ground truth generation - observing first results

New User Interface
Started front end development for new BNL Labs platform

data.public.lu

Bibliothèque nationale du Luxembourg
Open Data

ARTIFICIAL INTELLIGENCE

at BnL

Pit Schneider | 11/12/2020

Bibliothèque nationale du Luxembourg
Nationalbibliothéik