

ARTIFICIAL INTELLIGENCE

at BnL

Pit Schneider | 11/12/2020



Le gratis luxembourgeois
Saturday, 3 October 1857

Police correctionnelle de Luxemburg.

Page 4

Police correctionnelle de Luxemburg.

§). ©,*©p., ^anbelèinmm unb gctüber tnt
©vttttb, Sovftabt Suiemburg, tft tn ber tejten
©t'Çuitg beô Budjtpcltjct'gcricljté ju
etncrQefbbuffie non 100 gran* îen mib in bte
tïoftcn ocvurt^cilt «jorben, afé liber* fûf;vt, ben
Servit y. t?, ©erber, tn ber 9iad;t, aitif offenev
©trafic unb cljne ba\$u aufgeretjt toorben ju
fetn, gcfélagett unb mtfifmnbctt 311 ^abeit.

Police correctionnelle de Luxemburg.

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

eLuxemburgensia

1) Improve Optical Character Recognition (OCR)

2) Improve Newspaper Exploration

Search

1945

1916

1850

Map View

Stadt Luxemburg

20 / 1200

1945

1916

1850

Stadt Luxemburg • Location
 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec vel quam blandit, sodales eros et, fermentum felis. Duis ut tortor volutpat, suscipit magna vel, posuere neque. Nullam non sapien justo. Mauris tincidunt orci quis rutrum pharetra.
 Wiki Data Find Related

Luxemburg • Location
 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec vel quam blandit, sodales eros et, fermentum felis. Duis ut tortor volutpat, suscipit magna vel, posuere neque. Nullam non sapien justo. Mauris tincidunt orci quis rutrum pharetra.
 Wiki Data Find Related

Artikel über Luxemburg • Article
 LUXEMBURGER WORT - 1913/01/25
 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec vel quam blandit, sodales eros et, fermentum felis. Duis ut tortor volutpat, suscipit magna vel, posuere neque.
 View

named entities • entity relations • timeline • maps • wikidata

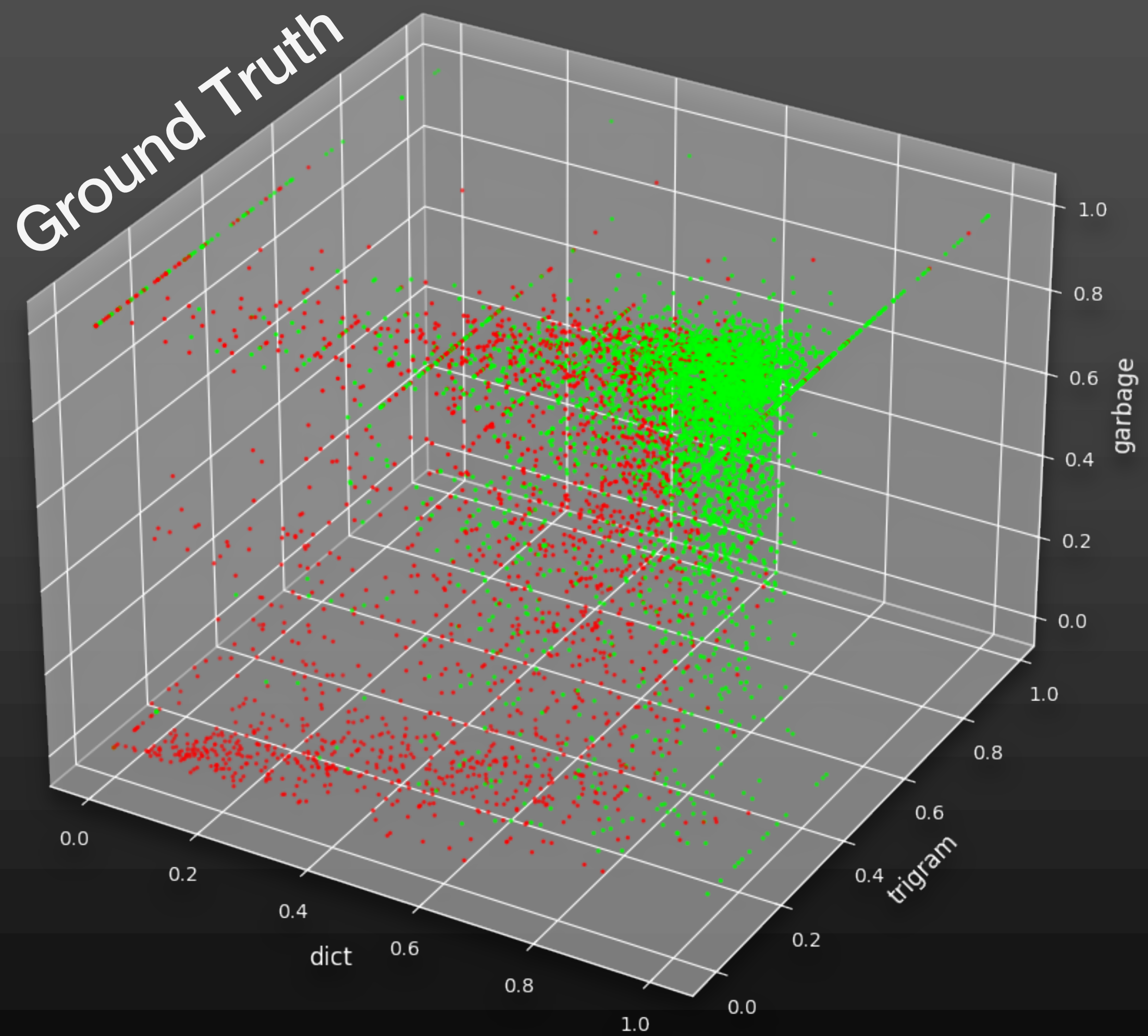
Ground Truth

≈100 Thousand Text Lines

Newspaper Corpus

≈ 200 Million Text Lines





● score ≥ 0.95
 ● else

score = $1 - (\text{editDistance} / |\text{chars}|)$

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.

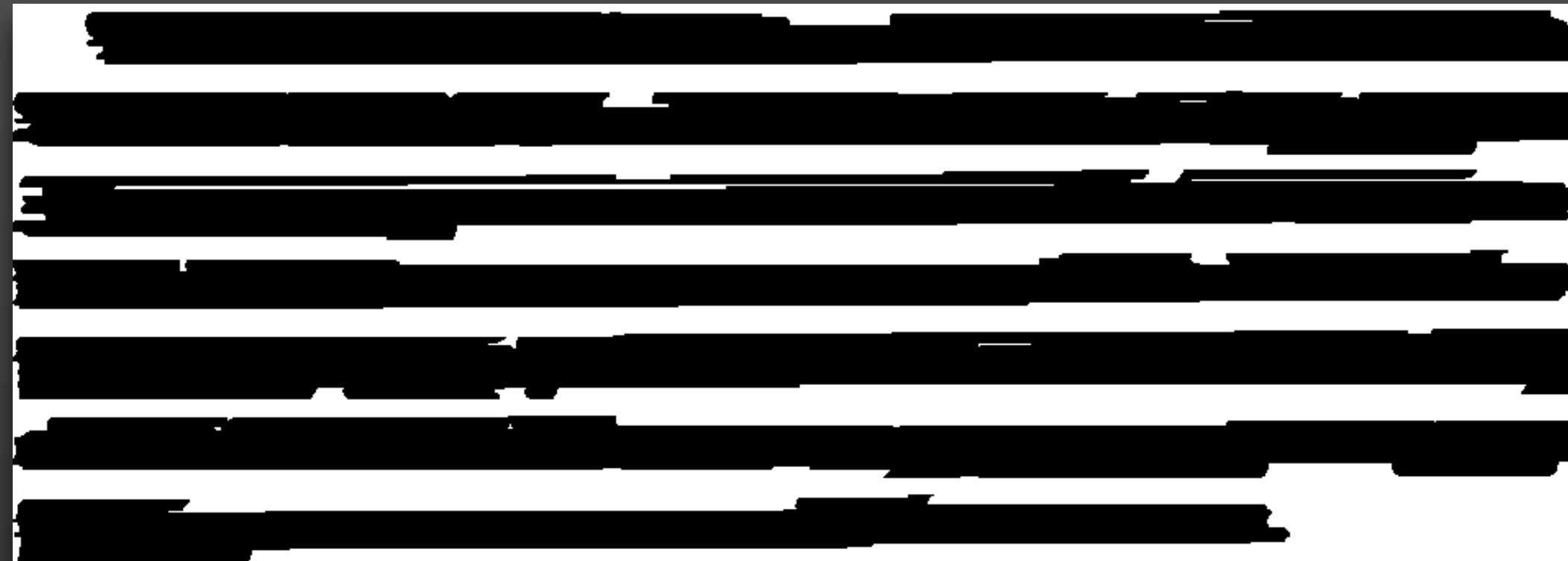
Original Text Block

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



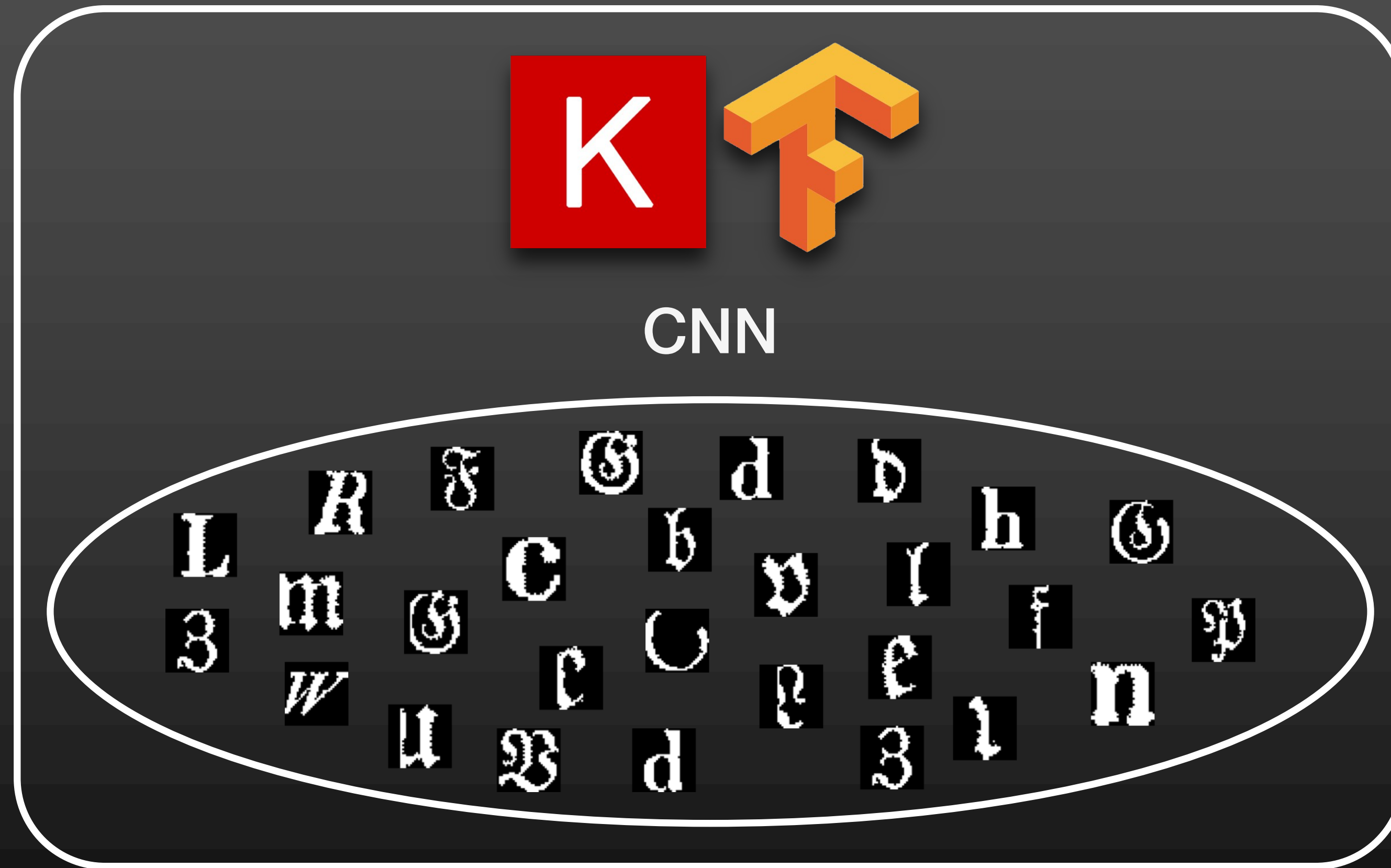
cleaning • dilation • padding • inversion • white on black detection



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

P. G.=Sp., Handelsmann und Färber im Grund,
Vorstadt Luxemburg, ist in der letzten Sitzung des
Zuchtpolizeigerichts zu einer Geldbuße von 100 Fran-
ken und in die Kosten verurtheilt worden, als über-
führt, den Herrn P. R., 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



Fraktur

Regular

99% accuracy
on test set

convolutional neural network • font recognition • binary classifier

Quality

Binarization

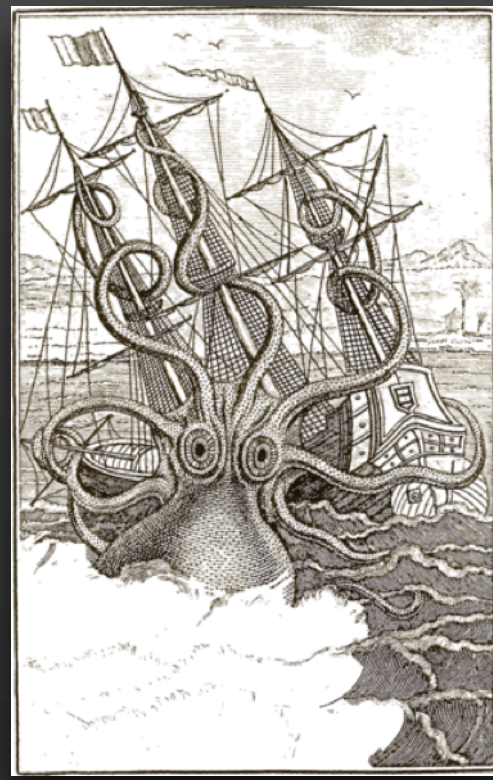
Segmentation

Font

OCR

ALTO

Entities



kraken



Tesseract



Calamari

open source software • custom model training

Quality

Binarization

Segmentation

Font

OCR

ALTO

Entities

Ground Truth



Fraktur

Regular

train/test sets • line/text pairs

Vorstadt Luxemburg, ist in der letzten Sitzung des

```
18 <TextLine HPOS="0" VPOS="41" WIDTH="761" HEIGHT="41">
19   <String CONTENT="Vorstadt" CC="00000000" HPOS="0" VPOS="41" WIDTH="135" HEIGHT="41"/>
20   <SP/>
21   <String CONTENT="Luxemburg," CC="0000000000" HPOS="135" VPOS="41" WIDTH="189" HEIGHT="41"/>
22   <SP/>
23   <String CONTENT="ist" CC="000" HPOS="324" VPOS="41" WIDTH="49" HEIGHT="41"/>
24   <SP/>
25   <String CONTENT="in" CC="00" HPOS="373" VPOS="41" WIDTH="42" HEIGHT="41"/>
26   <SP/>
27   <String CONTENT="der" CC="000" HPOS="415" VPOS="41" WIDTH="63" HEIGHT="41"/>
28   <SP/>
29   <String CONTENT="letzten" CC="0000100" HPOS="478" VPOS="41" WIDTH="97" HEIGHT="41"/>
30   <SP/>
31   <String CONTENT="Sitzung" CC="1000000" HPOS="575" VPOS="41" WIDTH="118" HEIGHT="41"/>
32   <SP/>
33   <String CONTENT="des" CC="000" HPOS="693" VPOS="41" WIDTH="68" HEIGHT="41"/>
34 </TextLine>
```

word bounding boxes • ALTO XML schema

PERSON 1 ORG 2 DATE 3 LOCATION 4

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.

Deutsch 1

Français 2

Lëtzebuergesch 3

Other 4

✓ ✗ ⌕ ↩

spaCy

ground truth generation • named entity recognition (NER)

Le gratis luxembourgeois
Saturday, 3 October 1857

Police correctionnelle de Luxemburg.

Page 4

Police correctionnelle de Luxemburg.

§). ©,*©p., ^anbelèinmm unb gctüber tnt
©vttttb, Sovftabt Suiemburg, tft tn ber tejtē
©t'Çuitg beô Budjtpcltjct'gcricljté ju
etncrQefbbuffie non 100 gran* îen mib in bte
tïoftcn ocvurt^cilt «jorben, afé liber* fûf;vt, ben
Servit y. t?, ©erber, tn ber 9iad;t, aitif offenev
©trafic unb cljne ba\$u aufgeretjt toorben ju
fetn, gcfélagett unb mtfifmnbctt 311 ^abeit.

Police correctionnelle de Luxemburg.

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

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

ARTIFICIAL INTELLIGENCE

at BnL

Pit Schneider | 11/12/2020

