Historical legislation

CASE STUDY

Annemieke Romein
Sara Veldhoen

huygens ing

KB nationale bibliotheek

About us

Annemieke Romein

- Researcher
- KNAW Huygens ING
- Researcher-in-Residence at KB in 2019

Sara Veldhoen

- Research Software Engineer
- KB, national library of the Netherlands

huygens ing

KB nationale bibliotheek

Content

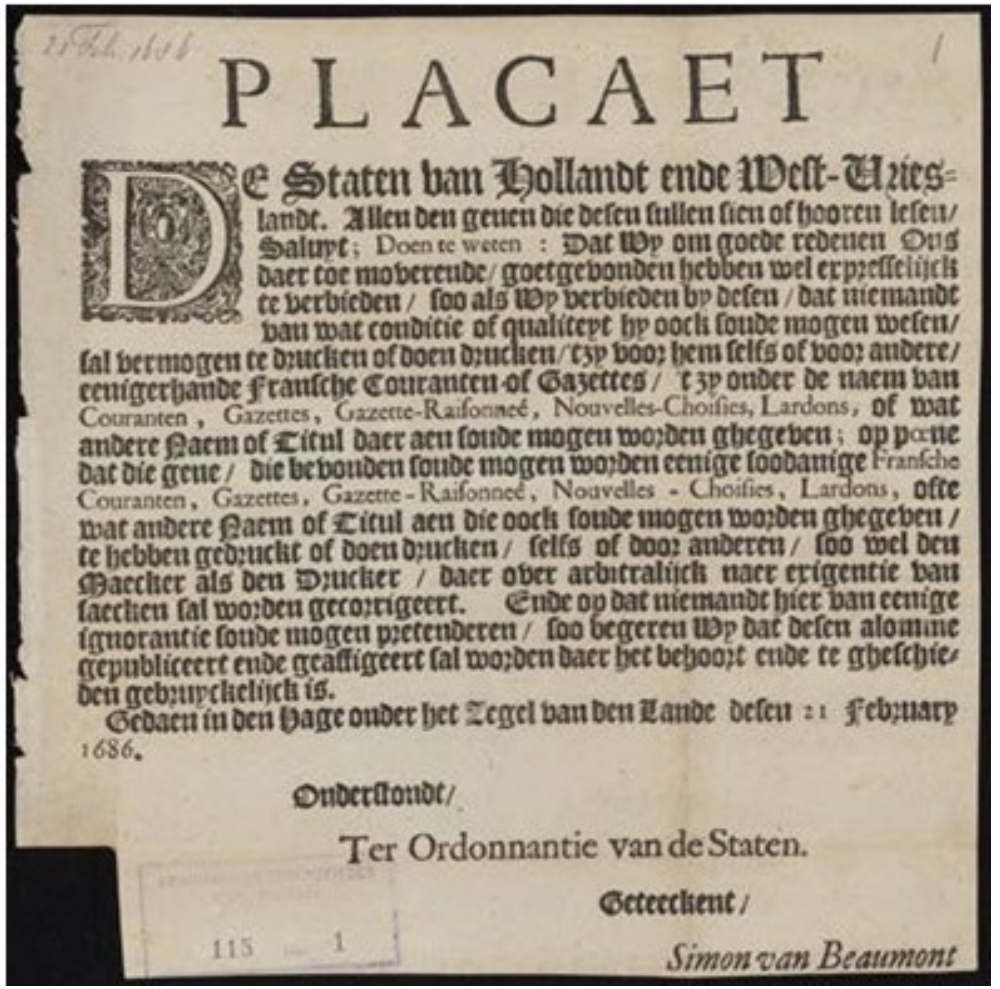
- Hypothesis
- Sources in Numbers
- From scans to computer-readable text
- Categorisation
- Training Annif

Paalhuis,
Amsterdam
(ca. 1660)

huygens ing

KB nationale bibliotheek

From volumes to individual laws



PLACAET

De Staten van Hollandt ende West-Vries-
landt. Allen den geenen die desen sullen sien of hooren lesen/
Salupt; Doen te weten : Dat Wy om goede redenen Ons
daer toe moverende/ goetgevonden hebben wel expresselijck
te verbieden / soo als Wy verbieden by desen / dat niemandt
van wat conditie of qualitept hy oock sonde mogen wesen/
sal vermogen te drucken of doen drucken/ tzy voor hem selfs of voor andere/
eenigerhande Fransche Couranten of Gazettes / tzy onder de naem van
Couranten, Gazettes, Gazette-Raisonneé, Nouvelles-Choisies, Lardons, of wat
andere Naem of Titul daer aen sonde mogen worden gegeven; op pœne
dat die gene / die bevonden sonde mogen worden eenige soodanige Fransche
Couranten, Gazettes, Gazette-Raisonneé, Nouvelles-Choisies, Lardons, ofte
wat andere Naem of Titul aen die oock sonde mogen worden gegeven /
te hebben gedruckt of doen drucken / selfs of door anderen / soo wel den
Maecker als den Drucker / daer over arbitralijck naer exigentie van
saecken sal worden gecorrigeert. Ende op dat niemandt hier van eenige
ignorantie sonde mogen pretenderen / soo begeren Wy dat desen alomme
gepubliceert ende geaffigeert sal worden daer het behoort ende te ghelschie-
den gebruyckelijck is.

Gedaen in den Hage onder het Segel van den Lande desen 21 February
1686.

Onderstondt/

Ter Ordonnantie van de Staten.

Geteeckent/

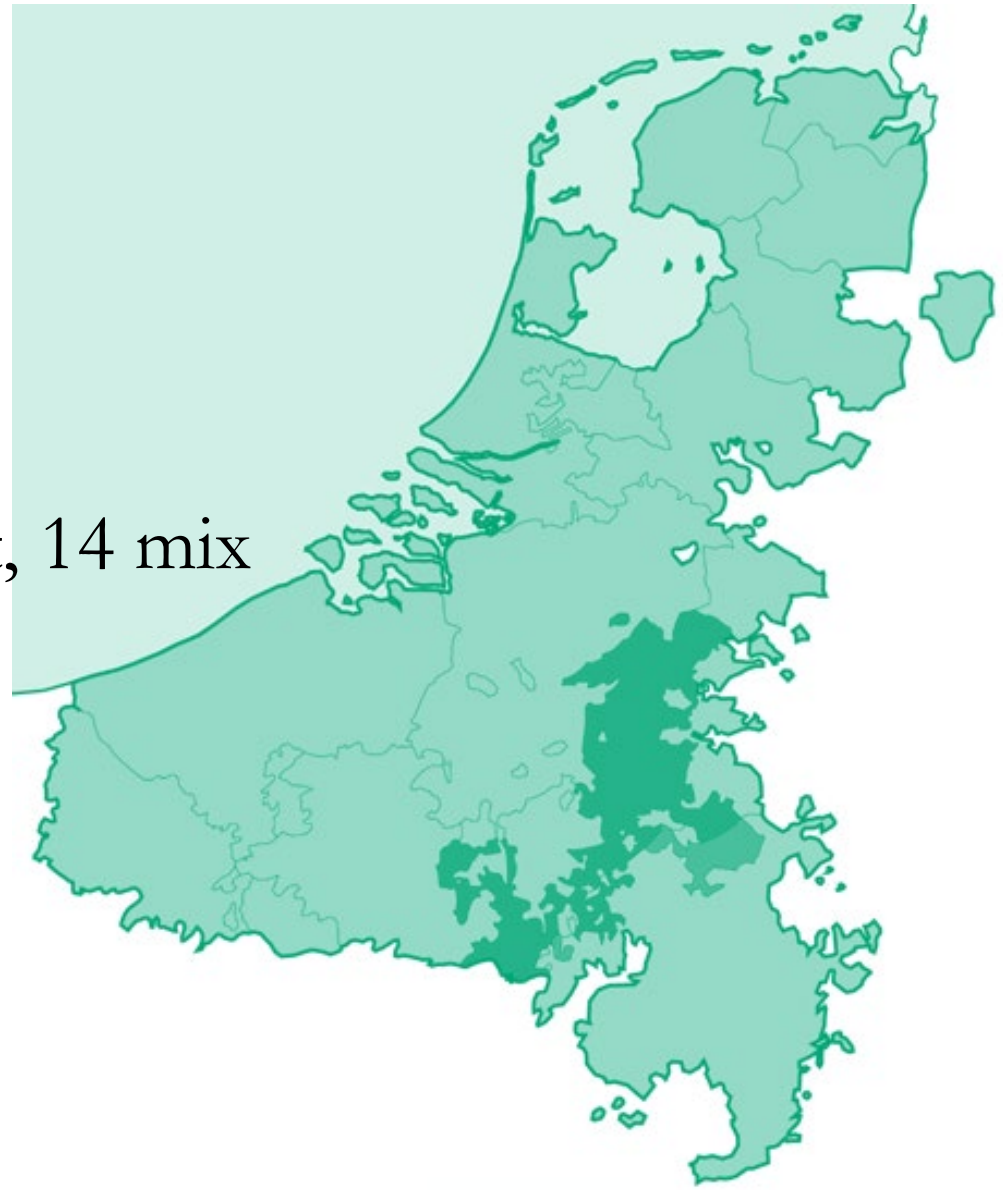
Simon van Beaumont

huygens ing

KB nationale bibliotheek

Sources in numbers

Sources #	108
Font	88 roman, 20 gothic
Language	67 Dutch, 26 FR, 1 Lat, 14 mix
Pages #	75.000
Characters#	550 mln
Publication date	17th and 18th century
Region	Nearly all provinces



huygens ing

KB nationale bibliotheek

Teaching the computer to read and to segment texts

Transkribus


DE Staten van Stadt Groningen ende Ommelanden: DOEN TE WETEN. Alhoewel wy verhoopt hadden dat door het redres eñ correctie van de Lijſte van verpachtinge by ons in het laeſt verlopen Jaer 1622. gedaen ende alomme gepubliceert / eenighſins gheremedieert ende belet ſulden zijn geweeſt / de veelvuldige frauden ende dieverpen / Item liſtige ende ſchadelijcke practiquen / compoſitien eñ andere abuyſen / ſtreckende tot groten aff-breuck van de Generale Middelen / eñ merckelijcke præjuditie van alle goede getrouwe Ingeſetenen ende oprechte Liefhebbers van 't Vaderlant / dewelcke deur alſulcke publicque Dieven ende Pranden van 't Gemeene beſte / niet alleene in haere neeringe becortet worden / maer oock de ſchade ende 't achterhepdt van 't Gemeene by den ſelven verooſaeckt

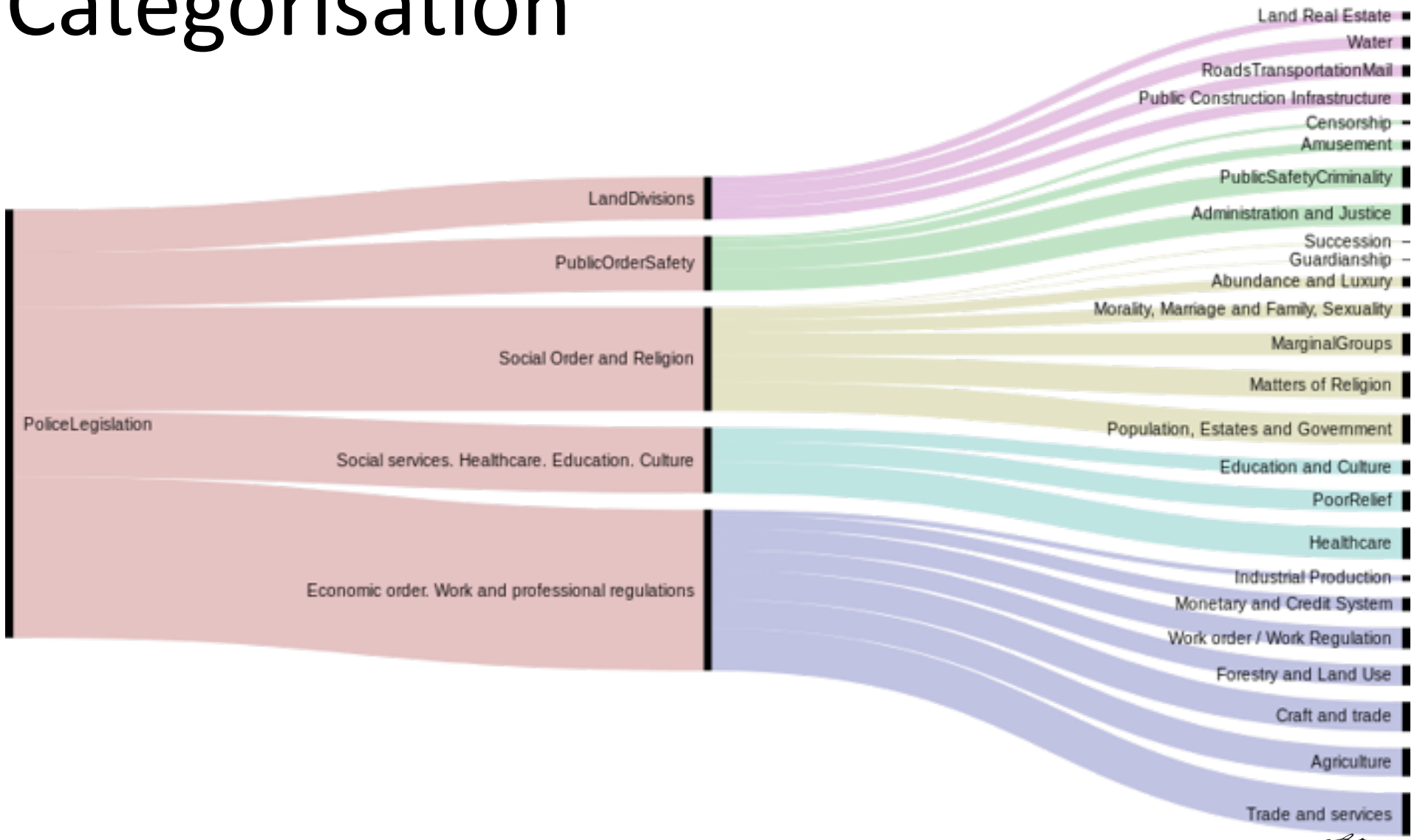
huygens ing

KB nationale bibliotheek

Categorisation

- Manually labelled ±3000 laws.
  - Used in the casestudy: 470.
- Hierarchical categorisation – 4(+1) layers (MPIeR).
  - Converted to SKOS format (Short Knowledge Organisation System)
  - https://doi.org/10.5281/zenodo.3564586

Categorisation



See http://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html

huygens ing

KB nationale bibliotheek

Training Annif

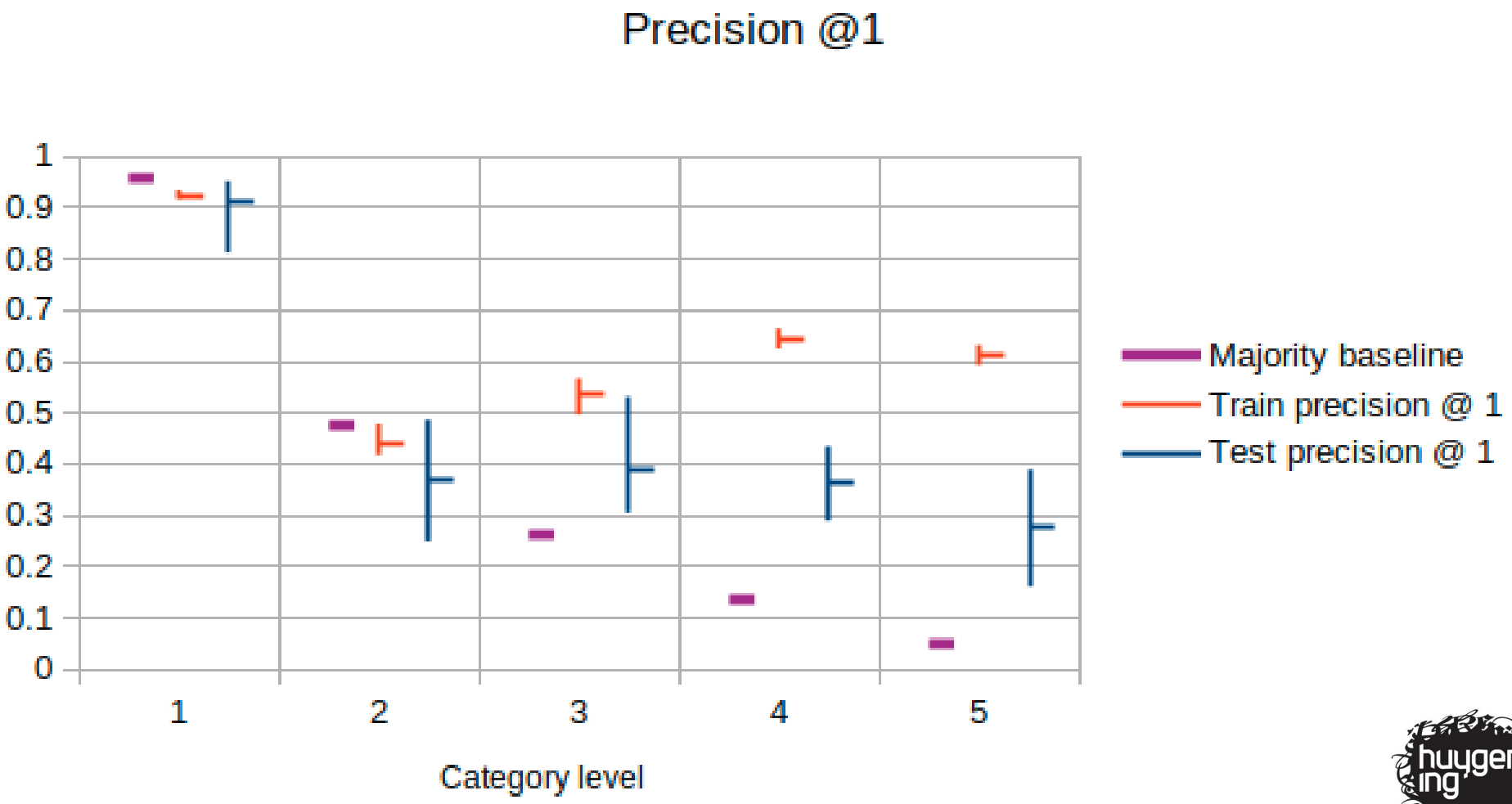
SET-UP

- Proof of concept: 470 labeled 'documents' (laws)
- <10 subjects each, 3.3 on average
- 69% and 28% of annotations at level 5 (deepest) and 4

- Training per category level
- 10-fold Cross Validation
- Backend: TF-IDF
- Limit = 4, threshold = 0.4

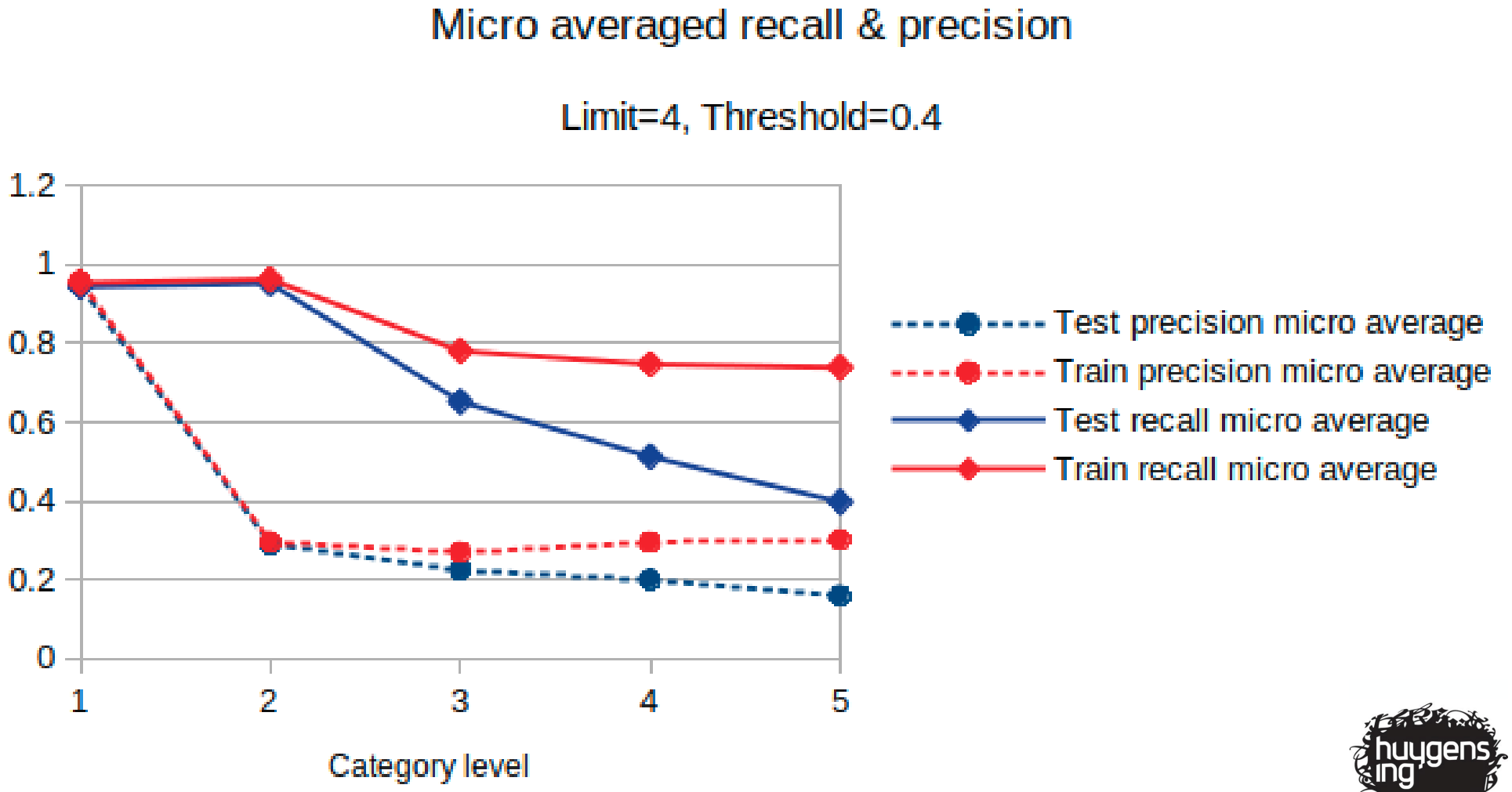
Training Annif

RESULTS



Training Annif

RESULTS



Training Annif

DISCUSSION

- Limited amount of data
- Only 'dumb' TF-IDF – curious about other backends
- Impact of OCR quality?
- Stemmer: Snowball Dutch
  Impact of historical language & spelling variation?
- Making use of the hierarchy in categories

Further Reading

POINTERS

- DH Benelux journal publication:
  http://journal.dhbenelux.org/journal/issues/002/article-23-romein/article-23-romein.html
- Info on KB lab:

  Dataset: https://lab.kb.nl/dataset/entangled-histories-ordinances-low-countries

  Blogs: https://lab.kb.nl/about-us/blog/entangled-histories-bumps-road-and-bursts-success
- Other work with Annif at KB:
  https://zenodo.org/record/3899723#.X2B2NIbRZJd

huygens ing

KB nationale bibliotheek

Further Reading

FUTURE WORK

- Annemieke continues to work on early-modern legislation (https://en.huygens.knaw.nl/projecten/game-of-thrones/)

- At the KB, we're experimenting with Annif in a tool to aid the cataloguing department: 'Demosaurus'.
  See https://zenodo.org/record/3899723#.X2B6Q4bRZJc

KB nationale bibliotheek