Historical legislation

CASE STUDY
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
Content

- Hypothesis
- Sources in Numbers
- From scans to computer-readable text
- Categorisation
- Training Annif
From volumes to individual laws
## Sources in numbers

<table>
<thead>
<tr>
<th>Source</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sources #</td>
<td>108</td>
</tr>
<tr>
<td>Font</td>
<td>88 roman, 20 gothic</td>
</tr>
<tr>
<td>Language</td>
<td>67 Dutch, 26 FR, 1 Lat, 14 mix</td>
</tr>
<tr>
<td>Pages #</td>
<td>75,000</td>
</tr>
<tr>
<td>Characters#</td>
<td>550 mln</td>
</tr>
<tr>
<td>Publication date</td>
<td>17th and 18th century</td>
</tr>
<tr>
<td>Region</td>
<td>Nearly all provinces</td>
</tr>
</tbody>
</table>
Teaching the computer to read and to segment texts
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
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

Precision @1

Category level

- Majority baseline
- Train precision @ 1
- Test precision @ 1
Training Annif

RESULTS

Micro averaged recall & precision

Limit=4, Threshold=0.4

Category level

Test precision micro average
Train precision micro average
Test recall micro average
Train recall micro average
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

• Other work with Annif at KB:
  https://zenodo.org/record/3899723#.X2B2N1bRZJd
Further Reading

FUTURE WORK

• Annemiek 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