Mixing computers and cultural heritage challenges and opportunities in cross-disciplinary collaborations

Marieke.van.Erp@dh.huc.knaw.nl

Merpeltje
Cross-disciplinary challenges

- Data challenges
- Technology challenges
- Cultural/User challenges
dinsdag
6 ossestaartsoep
HUt *orstjes
L0( * bonen met ananas

t te bonen met ananas
Va0,1 2 blikken witte bonen In 1 uitje
1 eetlepel ?!Wd- 2 eetlepels keuken-
12 knakwostjes, 1 klein
„ ftaLananasDlokJes- SoJrJi het uitJe
en meng dit ,Qoe h bonen met
tomatensaus. Nir,;e groente in een
ingevette ?fd h ste schaal. Roer de
mos- Je hni?or de stroop en giet hier
2* van over de bonen. Dek af met
Algorithmic transparency

KB newspaper archive (100M articles, 6 months server logs) study:

- Reran 1M real user queries through the (black box) search engine
- Counted how often each document appears in top 10 (100, 1000)
- Analysed correlation technical document features with retrieval counts

Findings:

- 96% of the articles never make it into the top 10 (76% never in top 100)
- Engine discriminates against very short and very long documents
- Best scoring articles contain long list of names (local elections, swimming diplomas, ...)

Querylog-based Assessment of Retrievability Bias in a Large Newspaper Corpus

ABSTRACT
For many digital libraries and archives, users are limited to the retrieval system offered by the data creators. It is important for users that all relevant documents are easily able to be retrieved, i.e., that retrieval results are not biased by hidden technological artifacts. However, the bias in the search technique towards the omission of research topics that require extensive searching is not well-understood. In this work, we propose a framework for assessing the retreivability bias of a search engine. We analyze a large newspaper collection and find that the bias is introduced by the search engine itself. To achieve this, we use a combination of data analysis and user studies. The results show that the bias is not introduced by the search engine alone, but rather by the interaction of the search engine and the user. The bias is caused by the search engine's inability to retrieve documents that are not well-indexed, and by the user's inability to find documents that are not well-indexed. The results show that the bias can be reduced by using a combination of data analysis and user studies. The results also show that the bias is not introduced by the search engine alone, but rather by the interaction of the search engine and the user.
Constructing a Recipe Web from Historical Newspapers

Marieke van Erp
Melvin Wevers
Hugo Huurdeman

Image source: https://static.ah.nl/static/recepten/img_006188_890x594_JPG.jpg

@merpeltje
@melvinwevers
@timelessfuture
Butter, salt & pepper

- Analysis of food customs:
  - historians
  - dieticians
  - ethnologists
- 1945 - 1995 Parool, Volkskrant, NRC & Trouw
- Dataset and code available through: https://github.com/DHLab-nl/historical-recipe-web
- Winner National Library - Rijksmuseum - Network Digital Heritage Hacka LOD Hackathon
- You & other researchers are invited to work with us on case studies around food culture
Newspapers as a source for recipes

- perception of a Dutch food culture formed in the 1950s
- newspapers are producer and messengers of public discourse
- newspapers contain views on daily life and customs
- But:
  - keyword search for ‘recepten’ imprecise
  - noise from digitisation process
Newspaper dataset

- Dutch National Library has digitised 90+ million book, newspaper and magazine pages.

- Newspapers published between 1618 - 1995 from the Netherlands, the Dutch Indies (present day Indonesia), the Antilles, the US and Surinam (15% of all newspapers published in the Netherlands).

- Available via website, data dump (until 1876) and API (with agreement).

<table>
<thead>
<tr>
<th>Newspaper</th>
<th>Pages</th>
<th>Articles</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parool</td>
<td>14.194</td>
<td>2,380,697</td>
<td>612,036,106</td>
</tr>
<tr>
<td>Volkskrant</td>
<td>13.628</td>
<td>2,248,652</td>
<td>744,275,792</td>
</tr>
<tr>
<td>NRC</td>
<td>7.199</td>
<td>947.198</td>
<td>489,397,816</td>
</tr>
<tr>
<td>Trouw</td>
<td>13.891</td>
<td>2,578,731</td>
<td>656,941,631</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td>48.912</td>
<td>8,155,278</td>
<td>2,502,651,345</td>
</tr>
</tbody>
</table>

Dinsdag
6 osseestaartsoep
HUt *orstjes
I 0( * bonen met ananas

\t t e bonen met ananas
Va0,1 2 blikken witte bonen ln 1 uitje
1 eetlepel ?!Wd- 2 eetlepels keuken-
12 knakwostjes, 1 klein
,, ftaLananasDlokJes- SoJrJi het uitJe
en meng dit ,Qoe h bonen met
tomatensaus. Nir,;e groente in een
ingevette ?fd h ste schaal. Roer de
mos- Je hni?or de stroop en giet hier
2* van over de bonen. Dek af met
From newspapers to a recipe web

- Allerhande
- Delpher
- Recipe text detection
  - Seed list
  - Text classification
- Structured newspaper recipes
- Recipe articles
- Information Extraction and Multilabel Classification
  - Ingredient and quantity extraction
  - Recipe tags
- Enrichment
  - DBpedia link
  - Scientific name
  - Origin
- Structured and enriched newspaper recipes
What & how much?

- articles cannot automatically be segmented
- OCR errors and non-grammatical sentences are a hurdle for standard NLP pipelines
- lexicon-based extraction of ingredients and quantities
Evaluation

- 100 articles were manually annotated using Recogito

- OCR errors in ingredients or quantities marked separately

- IAA .85 but OCR boundaries difficult: j°ar,anen’ vs o°ar,anen’

- Most precise lexicon: f1 = .67

- More research is needed for out-of-lexicon ingredients

---

Knackwurst White beans with pineapple. Belgian endive salad. Fruit pineapple. 1 or 2 cans (jars) white beans in to-. mato sauce – 1 onion
2 tablespoons syrup – 8 to 12 knackwursts – 1 small can of pineapple mix with the beans in tomato sauce. Put the beans in a with butter mustard through the syrup. This will be easier if the syrup is heated mixture over the beans. Cover the dish with a lid or aluminium foil mins until it’s all warmed up. In the meantime, heat the knackwurst don’t burst. Put them in the dish together with the pineapple chunk are hot and pour on the rest of the syrup. Put the dish back in the c...
## Results ingredients extraction

<table>
<thead>
<tr>
<th></th>
<th>Clean Ingredients</th>
<th></th>
<th>With OCR errors</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision</td>
<td>recall</td>
<td>f₁</td>
<td>correct</td>
</tr>
<tr>
<td>Allerhande</td>
<td>0.70</td>
<td>0.65</td>
<td>0.67</td>
<td>998</td>
</tr>
<tr>
<td>DBpedia</td>
<td>0.60</td>
<td>0.33</td>
<td>0.47</td>
<td>513</td>
</tr>
<tr>
<td>WordNet</td>
<td>0.62</td>
<td>0.50</td>
<td>0.56</td>
<td>764</td>
</tr>
<tr>
<td>AH-DBP-WN</td>
<td>0.56</td>
<td>0.75</td>
<td>0.66</td>
<td>1,154</td>
</tr>
</tbody>
</table>
27,411 new (old) recipes

- 34,479 Tags
- 365,133 ingredients
- >17,000 Links to external sources
- Data and software available at: https://github.com/DHLab-nl/historical-recipe-web
Take home message

• OCR errors can impact information extraction
• OCR post-correction is an active research field, but errors will remain
• Focus on most important elements to extract
Cultural AI
a lab for culturally valued AI
“Cultural AI is as much about using AI for understanding human culture as it is about using knowledge and expertise from the humanities to analyze and improve AI technology.”
“Cultural AI is the study, design and development of AI systems that are implicitly or explicitly aware of the subtle and subjective complexity of human culture.”
Lab organisation and mission

- PhD students and researchers spend time at academic AND heritage partners
- Close collaboration through “data sprints”, monthly reading club, joint conference/workshop visits
- Core partners, Associate partners, Affiliate partners.
Layers of bias

- Bias in the data
- Bias introduced by tool creators
- Technological bias

We do not want to erase bias, we want to make it visible!
The National Innovation Center for Artificial Intelligence (ICAI) has the mission to keep the Netherlands at the forefront of knowledge and talent development in AI.

Creating and nurturing a national AI knowledge and talent ecosystem.

More info: icai.ai
Current projects in the Cultural AI Lab

- BETTER-Mods (funded by NWO) 2 PhD students
- Culturally Aware AI (funded by NWO) 2 PhD students
- SABIO (funded by NDE) 1 researcher
- RE-FRAME (funded by Sound and Vision) 1 PhD student

Upcoming:
- Library AI Principles (funded by National Library) 1 PhD student
- Researcher in Residence (funded by National Library) 1 postdoc
- Transparent pipelines (funded by NWO/NLeSc) 1 postdoc
- Responsible AI in public media (funded by NWO) 1 PhD student
Synaesthesia

And

Cross-modality
Sample Kit by

Mediamatic

For the Odeuropa Workshop
Working with Scent in GLAMs

Odeuropa

Do not open the Sample Kit until the workshop starts on the 20th of May and have a pair of your scissors ready.

In this package you will find a zip-lock bag in which you can dispose of the blotters, to contain their smell.
Conclusions

• Custom domains require custom solutions
• Domain expertise is key
• Take time to get to know each other’s domain & (research) culture
• Where possible, bring physical elements into online meetings