PoliticalMashup Search Engine

Diachronical comparative analysis on parliamentary proceedings made easy

Maarten Marx, Informatics, University of Amsterdam
Goal

Create a repository of ALL European parliamentary proceedings in one machine readable format
Two main use cases:

**Easy search and exploration**
- Quickly find parts of proceedings, then "deep read" them, and explore further.
- Very easy, exact reference to immutable sources (by means of a URL)
  - a URL exists for every speech, even every paragraph, in the context of a debate

**Data science like applications**
- "Machine shallow reading" of large amounts of data.
- Example:
  Look at neighbouring words of [li]mmigran?t.*, group by party and "plot" the development of these words over the last 100 years, and compare this for Canada, the UK and the Netherlands.
Data: Parliamentary Proceedings

What's in the data?

Actors: Politicians and parties

- Names, abbreviations, different spellings, and identifiers (URI's) used in the proceedings
- Immutable attributes: gender, date of birth, wikipedia pages, links to DBpedia, etc
- Mutable attributes: membership of parties, constituencies, occupation, etc
Data: Parliamentary Proceedings

- **Proceedings**
- Nested data with metadata at each level, and most text in the bottom level
- topic
  - speech
    - paragraph

**Meaning**
- topic: a debate on ......
- speech: a 'non-interrupted' sequence of words spoken by one person
Data: numbers and sizes

- Data: parliamentary proceedings CA, NL, UK. Only debates.
  - three 8 linked datasets: proceedings, politicians, parties
  - Data format: XML (and derived from that, HTML, RDF, and now also JSON)

- Numbers (summer 2016)

<table>
<thead>
<tr>
<th>State</th>
<th>Topics</th>
<th>Scenes</th>
<th>Speeches</th>
<th>GB XML</th>
<th>GB Index</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>332675</td>
<td>573477</td>
<td>6274707</td>
<td>11.0</td>
<td>27.0</td>
<td>1803-2014</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>105734</td>
<td>198909</td>
<td>2616865</td>
<td>6.7</td>
<td>15.2</td>
<td>1814-2004</td>
</tr>
<tr>
<td>Canada</td>
<td>211236</td>
<td>392598</td>
<td>4425249</td>
<td>5.1</td>
<td>14.7</td>
<td>1901-2014</td>
</tr>
<tr>
<td>Sweden</td>
<td>51316</td>
<td>0</td>
<td>310740</td>
<td>1.1</td>
<td>1.9</td>
<td>1990-2014</td>
</tr>
<tr>
<td>Denmark</td>
<td>18946</td>
<td>0</td>
<td>549053</td>
<td>0.8</td>
<td>1.4</td>
<td>1999-2014</td>
</tr>
<tr>
<td>Norway</td>
<td>13262</td>
<td>88238</td>
<td>188825</td>
<td>0.5</td>
<td>1.2</td>
<td>1998-2014</td>
</tr>
<tr>
<td>European Union</td>
<td>20302</td>
<td>0</td>
<td>282831</td>
<td>0.6</td>
<td>1.1</td>
<td>1999-2014</td>
</tr>
<tr>
<td>Belgium</td>
<td>24423</td>
<td>0</td>
<td>139371</td>
<td>0.3</td>
<td>0.5</td>
<td>1995-2014</td>
</tr>
</tbody>
</table>
Information needs we want to support

Exploratory search instead of known-item search
• known-item search is what most existing search engines on this data cater for

Used techniques
• facets
• allow different rankings
• search at multiple granularities: topic, scene, speech
• Aggregations
  • time lines, also grouped by actors
  • word-cloud summaries, also grouped by actors.
• Allow queries which combine content with debate network constraints:
  • return documents about moslims in which Wilders speaks but where he is not interrupted by Pechtold.
Demo

search.politicalmashup.nl

Highlights

• Entry point retrieval
  • user chooses granularity

• Data exploration:
  • facets
  • different sorting options
  • 3D histograms (# hits per actor per year)
  • 2D "summaries" (related terms per actor)

• Network type queries
  • inclusion and exclusion of actors

• Ngram viewer
Conclusions

1. ElasticSearch engines scales well to really large collections
2. Prototype development on a serious (complete) dataset is feasible
   • reindexing the dutch (15Gb) collection is done in 90 minutes

Next steps
1. Stress testing
2. Add more datasets