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 [Ii]mmigran?t.*, group by party and "plot" the development of these words over the last 100 years, andcompare 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.**
  - Period: Dutch: 1814-2014, UK 1935-2014, CA: from 1994
  - three 8 linked datasets: proceedings, politicians, parties
  - Data format: XML (and derived from that, HTML, RDF, and now also JSON)

- **Numbers (summer 2016)**

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

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