

Current Topics in Digital Philology

The openLegislature project

Dan Häberlein, Peggy Lucke, J. Nathanael Philipp, Alexander
Richter

Universität Leipzig

Gliederung

- 1 Introduction
 - Corpus
 - Questions
- 2 Methods
 - Methods
- 3 Results
- 4 Outlook
 - next Steps
- 5 Take a chance on us!

Informations

Plenary Protocols from Bundestag

- stenographic reports in PDF
- open to the public
- siehe [**bundestag**]
- size of corpus circa 10GB → more than 3900 PDF

Questions to the information in the corpus

Statistic:

- How many speakers are in one legislative period/total?
- How many speeches from one party/speaker?

Keyword-search:

- Which speaker spoke to a special topic?

Why this questions?

We want more transparency! The answers are there, but too difficult to reach for all other people. That will be changed!

Gliederung

- 1 Introduction
 - Corpus
 - Questions
- 2 Methods**
 - **Methods**
- 3 Results
- 4 Outlook
 - next Steps
- 5 Take a chance on us!

Methods

Preprocessing:

- stop word filter
- lower case transformation
- word count for SLDA
- tf-idf for log-likelihood
- cooccurrences per speaker

Methods

Algorithms:

- Log-likelihood
- Topic Modell / SLDA

"Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.", siehe [blatent]

In supervised latent Dirichlet allocation (sLDA), we add to LDA a response variable associated with each document,
[BleiSupervisedTopicModels]

SLDA Methods

- ① single step approach
 - create single dataset for an election period
 - calculate top words for each speaker
- ② two step approach
 - create dataset for each protocol
 - calculate top words for each speaker
 - merge results for an election period
 - calculate top words for each speaker

- Usage of Listener Patterns [**javainsel9**]
- Usage of Github-Library Async [**async**] for easy creation of concurrent process chains

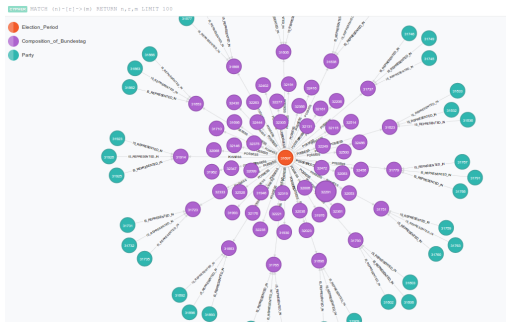


Gliederung

- 1 Introduction
 - Corpus
 - Questions
- 2 Methods
 - Methods
- 3 Results
- 4 Outlook
 - next Steps
- 5 Take a chance on us!

temporary results

- unstructured Textfiles available (PDF / TXT) for all election periods
- semi-structured XML files processed from PDF
- Metadatabase (NEO4J) with data of all election periods
- XPath query's on XML-Files



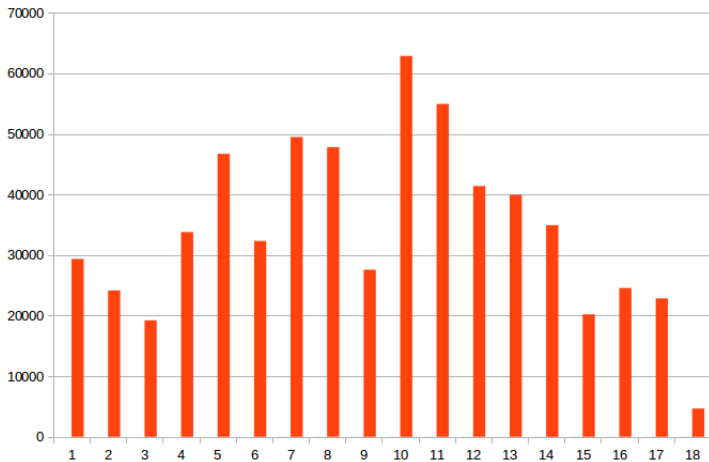
temporary results 2

- NoSQL Database:
 - all speeches
 - Speaker with all Speeches
 - appearance of words by speech
 - Speakerstatistics over all election periods
 - Partystatistics over all election periods
- website for browsing corpusdata/-statistics with visualisations
- input files for slda (arff-files) generated (for 18th election period)
- slda output: significant words for each speech of the 18th election period

Statistics

- 18 election periods
- 7004 speaker
- 679910 speeches
- 39 partys
- But: data not as clean as possible
 - typing errors: e.g. "CSU/CSU", "Angelika Me rtens", ..
 - parsing problems

Statistics 2: Speeches pro election period



significant words for speeches

18th election period, second session: Thomas Oppermann

- 1. staat
- 2. verhandeln
- 3. snowden
- 4. nsa
- 5. praxis
- 6. geheimdienste
- 7. ausspioniert
- 8. hören
- 9. möglichkeit
- 10. schutz

significant words for speeches 2

18th election period, third session: Oskar Lafontaine

- 1. waffenexporte
- 2. währung
- 3. ökonomisch
- 4. zukunftsaufgaben
- 5. währungsspekulation
- 6. übernachtungen
- 7. verteilung
- 8. waggons
- 9. zug
- 10. schneller

Gliederung

- 1 Introduction
 - Corpus
 - Questions
- 2 Methods
 - Methods
- 3 Results
- 4 Outlook**
 - next Steps
- 5 Take a chance on us!

What are our next tasks we need to accomplish?

- finalize our results, make them human accessible
- provide (easy) query interface for everyday users
- extend statistical webpage

Last steps until the end of the semester

- reprocess xml parsing
- finish result visualization
- connect our data with the meta data database (e.g. match every speaker to gouvernement / opposition)
- Log Likelihood on the GPU with our corpus

Gliederung

- 1 Introduction
 - Corpus
 - Questions
- 2 Methods
 - Methods
- 3 Results
- 4 Outlook
 - next Steps
- 5 Take a chance on us!

Take a chance on us!

Our project could be really interessting in the following sence:

- History / Political Science
- Educational Purposes
- Parties

We would try to contact the listed stakeholder to learn more about there needs.

Also, we would like to distribute our results in a way that everybody can consume them. This could involve the enhancement of your current result page (namely by a query interface for keywords) or the administration of a virtual machine with our results and software artifacts to recreate them.

What does it tell about human history and society?

- We could create deeper insights of how humans interact with each other, especially when they argue.
- We can follow and study important milestones of history (from the European Coal and Steel Community to the European Union)
- We can analyse which party and even which politician represents a specific opinion.

We think in the time of NSA and total surveillance a project like ours could contribute as a weapon against irresponsible politicians. The information about what the elected leaders actually do and say could be one step to a more enlightened society.