

openLegislature

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

Universität Leipzig

Outline

- ① Introduction
 - Korpus
 - Fragestellungen
- ② Methods
- ③ Outlook
- ④ Give us your money!

Introduction

- Korpus
- Fragestellungen

Informationen

Plenarprotokolle des Bundestages

- stenographische Berichte
- ab erste Sitzung des Bundestages September 1949
- vollständig bis zur letzten abgeschlossenen Sitzung
- PDF-Format
- aktuelle Wahlperiode auch als Textdatei

Zugang:

- öffentlich
- frei zugänglich
- <http://suche.bundestag.de/plenarprotokolle/search.form>

Download per Funktionen von:

- *Muster:*
[http://dip21.bundestag.de/dip21/btp/\[Wahlperiode\]/\[Wahlperiode\]\[Sitzung\].pdf](http://dip21.bundestag.de/dip21/btp/[Wahlperiode]/[Wahlperiode][Sitzung].pdf)
- *Bsp.:*
<http://dip21.bundestag.de/dip21/btp/01/01029.pdf>

Größe des Korpus

- Momentan 3895 PDF-Dateien
- Entspricht ca. 10GB

Welche Fragen können die Informationen des Korpus beantwortet werden?

Die ursprüngliche Fragestellung lautete, durch Ähnlichkeitsmessungen von Reden auf die Herkunft (auf dem Autor, z.B. einem Ghostwriter) der Rede schließen zu können. Dieses Ziel kann eventuell durch IR Methoden weiterverfolgt werden.

Statistisch:

- Wie viele Sprecher gab es?
- Wie viele Reden wurden von Mitgliedern einer bestimmten Partei gegeben?

Schlagwortsuche:

- Welcher Redner sprach zu einem bestimmten Sachverhalt?
- Welche Gesetze wurden zu einem Schlagwort verabschiedet?

Warum stellen wir diese Fragen?

Wir brauchen mehr Transparenz! Wir wollen das was schon da ist, einfacher zugänglich machen.

Wer hat schon Zeit sich die Reden des Bundestages anzuhören oder nachzulesen?

- Interesse zur Politik
- Analyse des gesprochenen Wortes
- Geschichte der deutschen Demokratie greifbar zu machen
- Transparenz deutscher Politik zu erhöhen

Die Politik muss sich langsam dem 21. Jahrhundert annähern. 2014, also ungefähr nach 20 Jahren Digitales Zeitalter wie wir es heute kennen, werden Wahlen noch immer mit ZETTEL und STIFT geführt!

Methods



Outlook

- Nächste Schritte
- Ziele Vorlesungszeit
- Ziele Semester
- Erweiterbarkeit
- Einschränkungen

Ausblick I: nächste Schritte

Nächste Schritte:

- erweitern der Metadaten-Datenbank mit Metadaten:
 - aus bestehenden XML Files
 - aus zusätzlichen Quellen (z.B. Sitzverteilungen)
- Überführen der Strukturierten Daten im XML-Format in einen Document Store
- Analyse der Daten:
 - Clustering (Top-Down, Bottom-Up)
 - LDA (Latend Dirichlet Allocation)

Ausblick II: Ziele Vorlesungszeit

Ziele bis Ende Vorlesungszeit:

- XML-Daten in Document-Store ablegen
- Metadaten-Datenbank mit weiteren Metadaten erweitern
- analysieren der Daten mittels mind. zwei Clustering-Verfahren
- Cluster mit wahrscheinlich gleichen Schreiber (aber nicht Redner) finden und darstellen

Ausblick III: Ziele Semester

Ziele bis Ende Semester:

- Analyse mittels LDA
- Visualisierung der Ergebnisse der LDA-Analysen
- weitere Cluster-Verfahren nutzen
- alle (sinnvollen) Ergebnisse vereinen und darstellen
- Untersuchung warum manche Analysen fehlerhafte/schlechte Ergebnisse lieferten

Ausblick IV: Erweiterbarkeit

Erweiterbarkeit wenn uns Zeit bleibt:

- zusätzlichen Metadaten-Quellen auffinden und in die bestehende Metadaten-Datenbank überführen
- weitere Metadaten erzeugen (Bsp. POS-Tagging, N-Gramme und Kookkurrenzen)
- Analyse-Verfahren erweitern
 - andere Cluster-Algorithmen
 - LDA mit anderen Parametern
 - LDA mit anderen Features
- Visualisieren der Ergebnisse

Ausblick V: Einschränkungen I

Einschränkungen wenn wir nicht alle Ziele schaffen:

- weniger Analyseverfahren nutzen (Bsp. nur ein Clusteringverfahren)
- Datenbereinigung verkürzen
- weniger Metadaten als Feature nutzen

Ausblick VI: Einschränkungen II

Wichtigste Ziele:

- Daten in Datenbank strukturiert ablegen
- Clustering-Verfahren auf unsere Daten anwenden
- Interpretation der Ergebnisse

Neue Ziele:

- vermutlich nicht

Give us your money!

We try to achieve reproducible and professional results. Our project could be really interesting in the following sence:

- History / Political Science
- Educational Purposes
- Parties

We could also make this dataset that we just created more human accessible by developping an easy user interface (something like google). Our work would contribute to more transparent german politics, in which every citizen has the power to validate and measure politicians by there speeches.