Quantitative Textanalyse

Sitzung 14: Abschlussitzung

Mirko Wegemann

22. Januar 2025

Ablaufplan der heutigen Sitzung

- Vorstellung euer Projektideen
- Bewertungskriterien
- Review der gelernten Inhalte
- Feedback

Eure Projektideen

Findet euch in Paaren zusammen, wobei jeweils eine Person eine Abschlussarbeit schreibt. Diese Person stellt in maximal zwei Minuten die Idee der Arbeit vor. Im nächsten Schritt mischen sich die Paare und die Person, welche zuvor zugehört hat, stellt den Pitch einer weiteren Person vor. Habt ihr potentielle Fragen zu der Projektidee? Was findet ihr besonders gut? Was könnte herausfordernd sein?

Die Bewertung der Projektarbeit I

Formale Kriterien

Zur Erinnerung:

- Abgabe der Hausarbeit bis zum **31.03.2025**
- Abgabe einer .pdf-Datei sowie einem R-Skript. Die Analyse muss replizierbar sein. Falls hierfür ein Datensatz benötigt wird, reicht diesen mit ein.
- Länge der Arbeit: 5000-6000 Wörter exkl. Abstract und Literaturverzeichnis)
- Korrekte Zitation, Transparenz bei KI-Nutzung
- Wurden alle Tabellen und Abbildungen beschriftet und referenziert?
- Alle weiteren formalen Kriterien, insbesondere das zu nutzende Word-Template findet ihr hier

Inhaltliche Kriterien I

- entspricht die Gliederung der einer wissenschaftlichen Arbeit (s. Dokument)
- Fokus insbesondere auf diesen Kriterien:
  - Verständlichkeit der Forschungsfrage
  - Herleitung der Relevanz der Forschungsfrage
  - Systematischer Literaturüberblick
  - Nutzung passender Datenquelle und gute Beschreibung dieser
  - Methodenauswahl und Erklärung der Umsetzung sowie Abwägung der Vor- und Nachteile einer Methode
  - Validierung der Methode (Klassifikationsmetriken, Möglichkeiten nach Quinn et al. (2010), Beispiele)
  - Korrekte Beschreibung der Ergebnisse
  - Rückkopplung der Ergebnisse zur Forschungsfrage
  - Eingehen auf Stärken und Limitierungen der vorliegenden Arbeit

Strukturelle Kriterien

- Leiten die Einzelteile der Arbeit gut ineinander über?
- Wurden Wiederholungen vermieden?

Literatur

Powner, L. C. (2015). *A political science student's practical guide*. Sage/CQ Pres

- Einstieg in akademisches Schreiben in der Politikwissenschaft
- Überblick über alle Unterschritte einer empirischen Arbeit
- besonders hilfreich sind die ersten 3 Kapitel (Forschungsfrage, Theorie und Literaturüberblick)

Was habt ihr gelernt?

Nehmt euch ein paar Minuten Zeit und denkt darüber nach, was ihr aus dem Seminar mitnehmt.

R-Crashkurs

Wir haben mit einem zweiwöchigen Crash-Kurs zu `R` begonnen. Unser Fokus lag dabei vor allem auf...

- Objekten
- Import von Datensätzen
- Datenaufbereitung über das `tidyverse`
- im Zuge des Seminars haben wir auch gelernt, Regressionsanalysen zu interpretieren (vgl. Sitzung zu Topic Models)

Konzeptklärung I

Wir haben quantitative Textanalyse als eine Form von Inhaltsanalyse kennengelernt, bei der wir versuchen, Beziehungen und Regularitäten zwischen verschiedenen Textfragmenten entdecken.

"Computational text analysis (also called Quantitative Text Analysis, Automated Content Analysis, Text Mining, Text as Data etc.) draws on techniques developed in natural language processing and machine learning to analyse textual documents." (Chun Ting-Ho 2021)

Datenbeschaffung

Ihr habt gelernt, Daten zu erschließen.

- Zugriff auf existierende Datenquellen und Import in `R`
- Scraping statischer Webseiten über `rvest`
- Scraping dynamischer Webseiten mithilfe von `RSelenium`
- API-Zugriff durch API-Wrapper oder `jsonlite`

Datenvorbereitung I

Wir sind durch die konventionelle Pipeline der Datenvorbereitung von *bags-of-words*-Ansätzen durchgegangen:

- Import des Datensatzes
- ggf. Filtern der Daten
- Umwandlung in ein *corpus*-Objekt
- ...weitere Umwandlung in ein *tokens*-Objekt
- ...bis zur Transformation in eine **d**ocument-**f**requency-**m**atrix

Datenvorbereitung II

Dabei haben wir oftmals einige Veränderungen an den Daten vorgenommen:

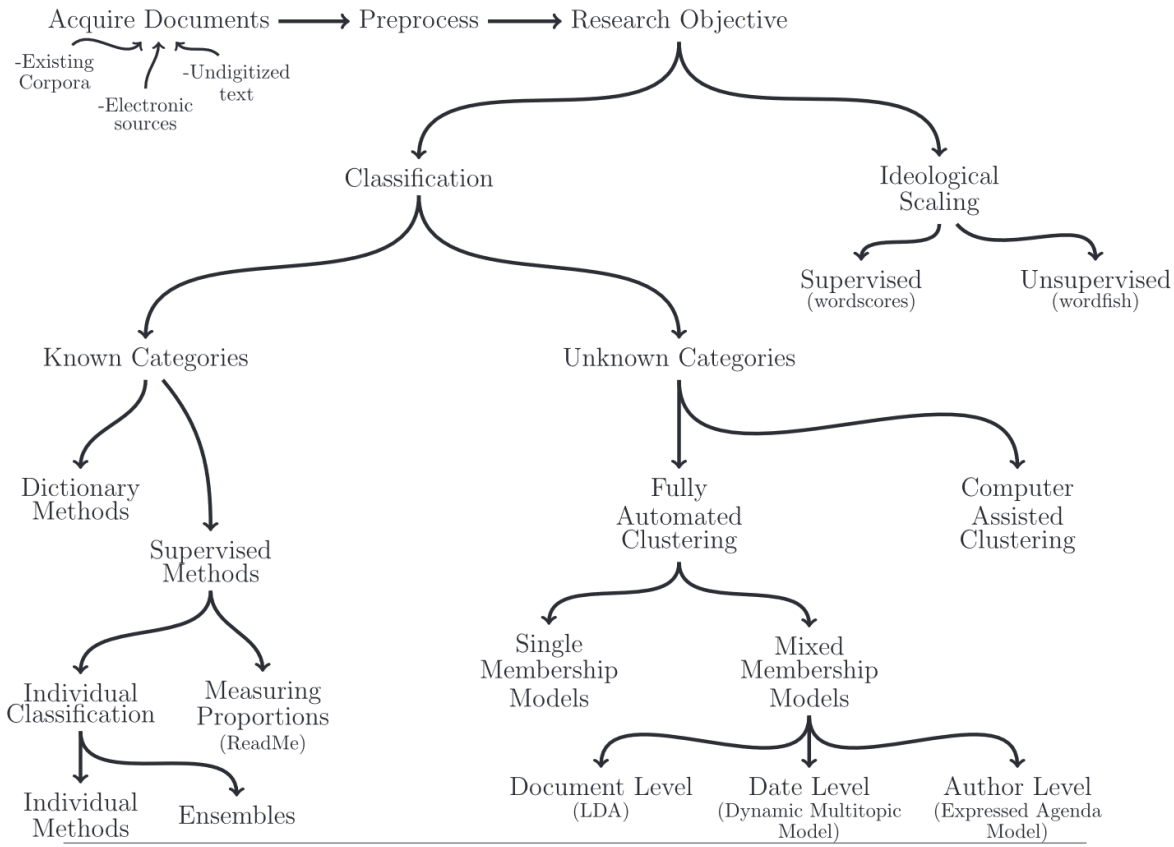
- Lower-Casing
- Entfernen von Features ohne semantischen Wert (meist Satzzeichen, Zahlen, Stopwörter, Symbole)
- Entfernen von besonders häufigen/besonders seltenen Features
- n-grams
- Stemming oder Lemmatisierung

Datenvorbereitung III

Wir haben aber auch über potentielle Kritik an diesen pre-processing Steps gesprochen (Denny and Spirling 2017)

- Modelle können sehr sensitiv sein
- Das Paket `preText` hilft uns, den Effekt von pre-processing auf unseren Korpus besser abzuschätzen (wenn es denn korrekt installiert ist)
- Pre-Processing muss immer theoretisch begründet werden

Datenanalyse I



Methoden der Textanalyse

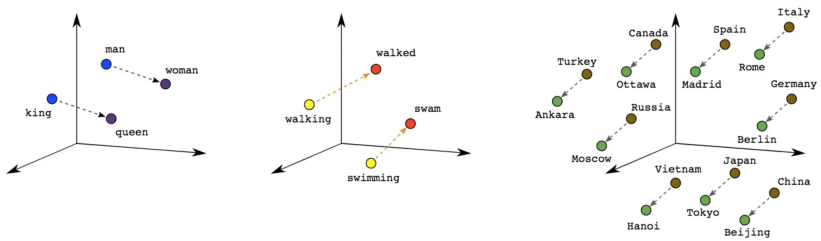
Datenanalyse II

Wir haben uns mit folgenden Analysemethoden beschäftigt:

- Wörterbuch-Analyse
- Unsupervised topic models
- Semi-supervised scaling (Latent Semantic Scaling)
- Supervised classification

Datenanalyse III

Zudem haben wir Word Embeddings kennengelernt und gelernt, wie wir diese *intrinsisch* (*nearest neighbors*, *Embedding Regression*) sowie instrumentell (als Input von *neural networks*) nutzen können.



Male-Female Verb Tense Country-Capital

Noch etwas zu ergänzen?

- Dinge, die ihr gelernt habt?
- Fragen, die noch offen bleiben?
- Offenes Feedback?

Feedback

Biite geht nun ins Learnweb und füllt die Abschlussevaluation aus.

Ausblick

- Diejenigen, die eine Abschlussarbeit schreiben: Sprechstunde vereinbaren
- Für alle anderen, danke für Eure Mitarbeit während des Seminars!
- Besucht gerne die Seminare unseres Lehrbereichs im nächsten Semester:
  - 'Einführung in die Vergleichende Politikwissenschaft' von Prof. Daniel Bischof
  - 'Was wir von wissenschaftlichen Studien lernen können (Kausale Inferenz I)' von Prof. Daniel Bischof
  - 'Applied Introduction to R for Political Scientists' von Elena Leuschner
  - 'Gender and Political Representation' von mir

References I

Denny, M., & Spirling, A. (2017). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. https://doi.org/10.2139/ssrn.2849145

Powner, L. C. (2015). *A political science student's practical guide*. Sage/CQ Pres.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science*, *54*(1), 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x