Quantitative Textanalyse

Einführung

Mirko Wegemann

Universität Münster
Institut für Politikwissenschaft

09. Oktober 2024

Zu meiner Person

Mirko Wegemann (er/ihm)

- Seit Oktober 2024: Wissenschaftlicher Mitarbeiter am Lehrbereich für Vergleichende Politikwissenschaft bei Prof. Daniel Bischof
- zuvor: Doktorand am Europäischen Hochschulinstitut in Florenz
- Interessensschwerpunkte
  - politische Parteien
  - politische Kommunikation
  - politische Kultur von Geschlechterrollen

Ziele des Seminars

- Crashkurs in der Statistiksoftware R und das grafische Interface RStudio
- Einführung in grundlegende Konzepte und Methoden der quantitativen Textanalyse und ihre Anwendung in R
- Diskussion von empirischen Forschungsartikeln
- Entwicklung eigener Forschungsideen (Fragen, Hypothesen und Designs)

Unser Semesterplan I

1. Woche Einführung
2. Woche R-Crashkurs
3. Woche Quantitative Textanalyse als Form der Inhaltsanalyse
4. Woche Automatisierte Datensammlung I: Web-Scraping
5. Woche Automatisierte Datensammlung II: Dynamische Websites
6. Woche Automatisierte Datensammlung III: APIs
7. Woche Datenaufbereitung

Unser Semesterplan II

8. Woche	Unsupervised Topic Models
9. Woche	Scaling
10. Woche	Grundlagen des Maschinellen Lernens
11. Woche	Klassifikation anhand von Supervised Text Analysis
12. Woche	Word Embeddings und neuronale Netzwerke
13. Woche	Large Language Models
14. Woche	Wrap-Up und Präsentationen eigener Projektideen

Insgesamt: recht wenig Literatur, im späteren Verlauf teils Anwendungstexte

Genereller Ablauf der Sitzungen

1. (Vorstellung der Methodenanwendung durch Studierende)
2. Fragen zur letzten Sitzung
3. Input
4. Anwendung in R
5. (Diskussion von Anwendungstexten)

Organisation des Kurses

- Kommunikation und Literatur über Learnweb (Abruf von Mails wird erwartet)
  - Literaturrecherche über ULB Münster, Google Scholar oder Web of Science
  - Zugriff auf die meisten Artikel über das WLAN der Universität Münster, von Zuhause per VPN
  - für einen Einblick in die Uni-Bibliothek: Audiotour der Bibliothek
- Slides über meine Website

Anforderungen I

Arbeitsaufwand

Ein ECTS-Punkt steht für max. 30 Stunden à 60 Minuten an tatsächlichem Arbeitsaufwand seitens des/der Studierenden.[1]
Beispielrechnung:

$$\rightarrow \ 30 \times 7 = 210 \ \textit{Stunden}$$

Teilnahme an der Lehrveranstaltung:

$$1.5 \times 14 = 21 \ \textit{Stunden}$$

Vor- und Nachbereitung des Kurses sowie Entwicklung der Prüfungsleistungen

$$210 - 21 = 190 \ \textit{Stunden}$$

[1] Recht.NRW

Anforderungen II

- Regelmäßige Anwesenheit und Partizipation
- Vorige Auseinandersetzung mit der Sitzungsliteratur
- Anwendung einer Methode anhand von eigenen Forschungsdaten
- Ausarbeitung eines Forschungsprojekts

Anforderungen III

Leistungen

- Studienleistung: Anwendung einer Methode
- Prüfungsleistung: Abschlussarbeit

Teilleistungen I

Anwendung einer Methode

In den Sitzungen 4-6 und 8-12 lernen wir Methoden der Datensammlung und -analyse kennen. Alle Studierenden sollten zu **einer** dieser Sitzungen die erlernte Methode anwenden und im Rahmen eines kurzen Inputs in der nächsten Sitzung vorstellen.

- Problemstellung definieren
- Methode in R umsetzen und kommentieren
- Kurzvortrag (max. 3 Minuten) über Erfahrungen bei der Umsetzung

Teilleistungen II

Abschlussarbeit

- 6000 Wörter $\pm$ 10 Prozent (exkl. Inhalts- und Literaturverzeichnis und Deckblatt), abzugeben bis spätestens zum 31.03.2025
- selbstgewähltes Thema, welches mit mindestens einer erlernten Methode empirisch analysiert wird
- Standardformat einer empirischen Arbeit: Einleitung (inkl. Forschungsfrage Motivation), Literaturüberblick, Theorieteil (inkl. Hypothesenentwicklung), Fallauswahl und Methode, Resultate, Schlussteil)
- Achten auf die richtige Zitierweise (Leitlinien zum Beispiel hier); Schulung zu Citavi bspw. nächste Woche durch die ULB
- Abgabe von schriftlicher Arbeit (PDF-Format) und Skript (im .R- oder .RMD-Format)

Plagiate I

"Ein Plagiat stellt im prüfungsrechtlichen Sinn einen vorsätzlichen Täuschungsversuch dar.

Der Tatbestand des Plagiats ist erfüllt, wenn in schriftlichen Studien- und Prüfungsleistungen 'Texte Dritter ganz oder teilweise, wörtlich oder nahezu wörtlich übernommen und als eigene wissenschaftliche Leistung' ausgegeben werden. Dieses Vorgehen 'widerspricht nicht nur guter wissenschaftlicher Praxis, es ist auch eine Form geistigen Diebstahls und damit eine Verletzung des Urheberrechts'."

IfPol: Resolution des Deutschen Hochschulverbandes

Plagiate II

Plagiatstypen

- Vollplagiat
- Zitat ohne Beleg
- Übersetzungsplagiat
- Selbstplagiat
- Ghostwriting

Prüfungsleistungen werden über die Software *Turnitin* auf Plagiate untersucht.

Inklusion

Wir pflegen eine offene inklusive Kultur in diesem Seminar.

- In Seminarkommunikation wird eine geschlechtergerechte Sprache verwendet.
  - Studien, warum es sinnvoll ist, das generische Maskulinum aufzubrechen: Stahlberg and Sczesny (2001) and Vervecken et al. (2013)
- Es gibt hier keinen Platz für Rassismus, Sexismus, Homo- oder Transphobie.
- Wir lernen gemeinsam: inhaltliche Fragen sind legitim und erwünscht, Wissenslücken sind kein Grund, sich zu schämen.

Kontakt

- Nach dem Seminar oder in der Sprechstunde (nach voriger Mailanmeldung)
- E-Mail: m.wegemann@uni-muenster.de
- Adresse
  Institut für Politikwissenschaft
  Lehrbereich: Vergleichende Politikwissenschaft
  Raum
  48151 Münster
- Feedback über Google Forms oder per Mail

Fragen?

Und ihr?

Eine kleine Umfrage:
https://pingo.coactum.de/872203/

Was ist Quantitative Textanalyse? I

Habt ihr eine Idee?

Was ist Quantitative Textanalyse? II

"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)

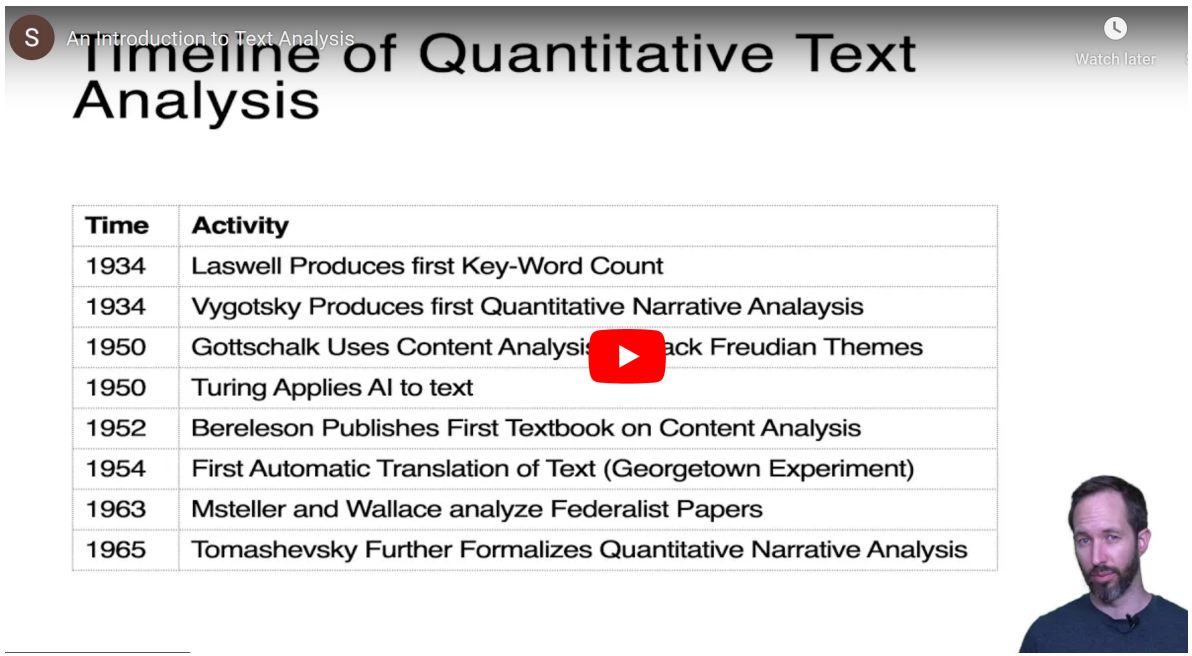
Was ist Quantitative Textanalyse? III

Es handelt sich um...

- eine Form der Inhaltsanalyse, in der wir Text bzw. dessen Bestandteile in numerische Vektoren umwandeln, um Beziehungen und Regularitäten zwischen dem Text zu entdecken (Benoit 2009)
- QTA kann sowohl explorativ als auch erklärend genutzt werden, ist aber nie atheoretisch (Bonikowski and Nelson 2022)

Ist das wirklich neu? I

Nein, quantitative Textanalyse gibt es schon länger.



Time	Activity
1934	Laswell Produces first Key-Word Count
1934	Vygotsky Produces first Quantitative Narrative Analaysis
1950	Gottschalk Uses Content Analysis ack Freudian Themes
1950	Turing Applies AI to text
1952	Bereleson Publishes First Textbook on Content Analysis
1954	First Automatic Translation of Text (Georgetown Experiment)
1963	Msteller and Wallace analyze Federalist Papers
1965	Tomashevsky Further Formalizes Quantitative Narrative Analysis

Figure: Geschichte quantitativer Textanalyse

Ist das wirklich neu? II

Ansätze, die darauf basieren, Wörter in Texten zu zählen (key-word count) gibt es schon lange.

- Durchbruch durch General Inquirer (Stone et al. 1962)
- Idee: Kategorisierung von 164 verschiedenen Kategorien mithilfe eines Wörterbuchs

$\rightarrow$ hierauf liegt der Fokus vieler Methoden, die wir kennenlernen werden

Ist das wirklich neu? III

Können Maschinen denken?

> *'I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about* 10^9*, to make them play the imitation game so well that an average interrogator will not have more than 70 per cent, chance of making the right identification after five minutes of questioning.' (Turing 1950, p. 442)*

Ist das wirklich neu? IV

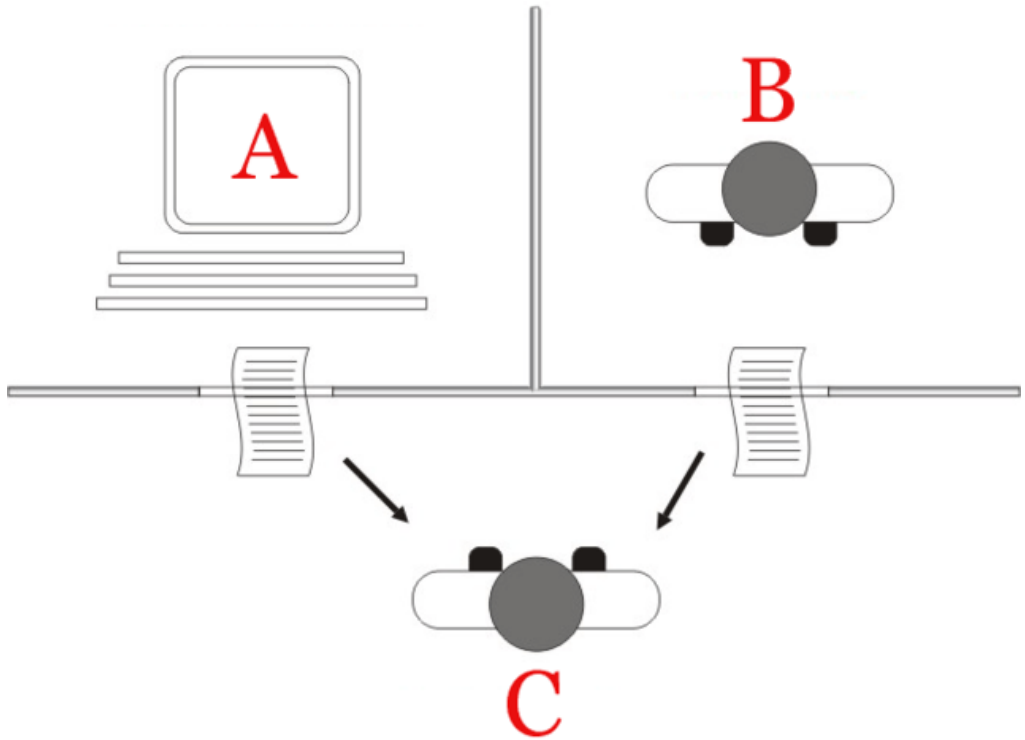
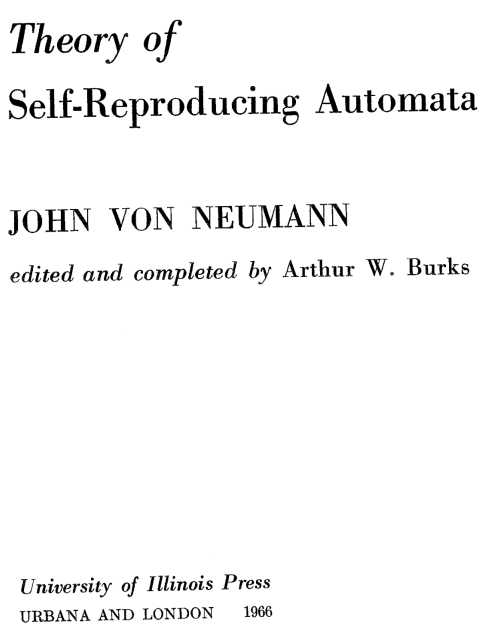


Figure: The Imitation Game

Ist das wirklich neu? V


Theory of
Self-Reproducing Automata

JOHN VON NEUMANN

edited and completed by Arthur W. Burks

University of Illinois Press
URBANA AND LONDON 1966

Figure: Neumann's Idee von selbst-replizierenden Maschinen

Evolution via Mutation (ähnlich DNA), Architektur eines Encoder und Decoder

Ist das wirklich neu? VI

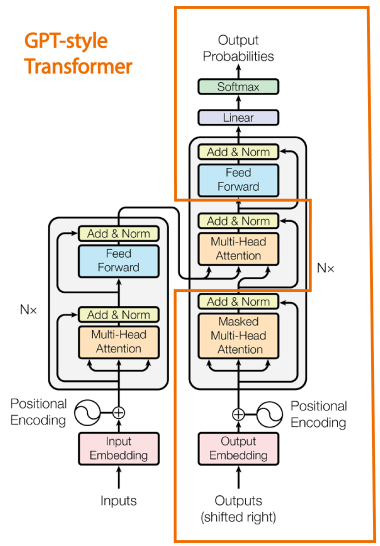


Figure: Die Transformer-Architektur

In den Sitzungen 12-13 werden wir uns mit neuronalen Netzwerken beschäftigen

Warum quantitativ?

Figure: Big Data – World Economic Forum

Warum quantitativ?

Figure: Big Data – World Economic Forum

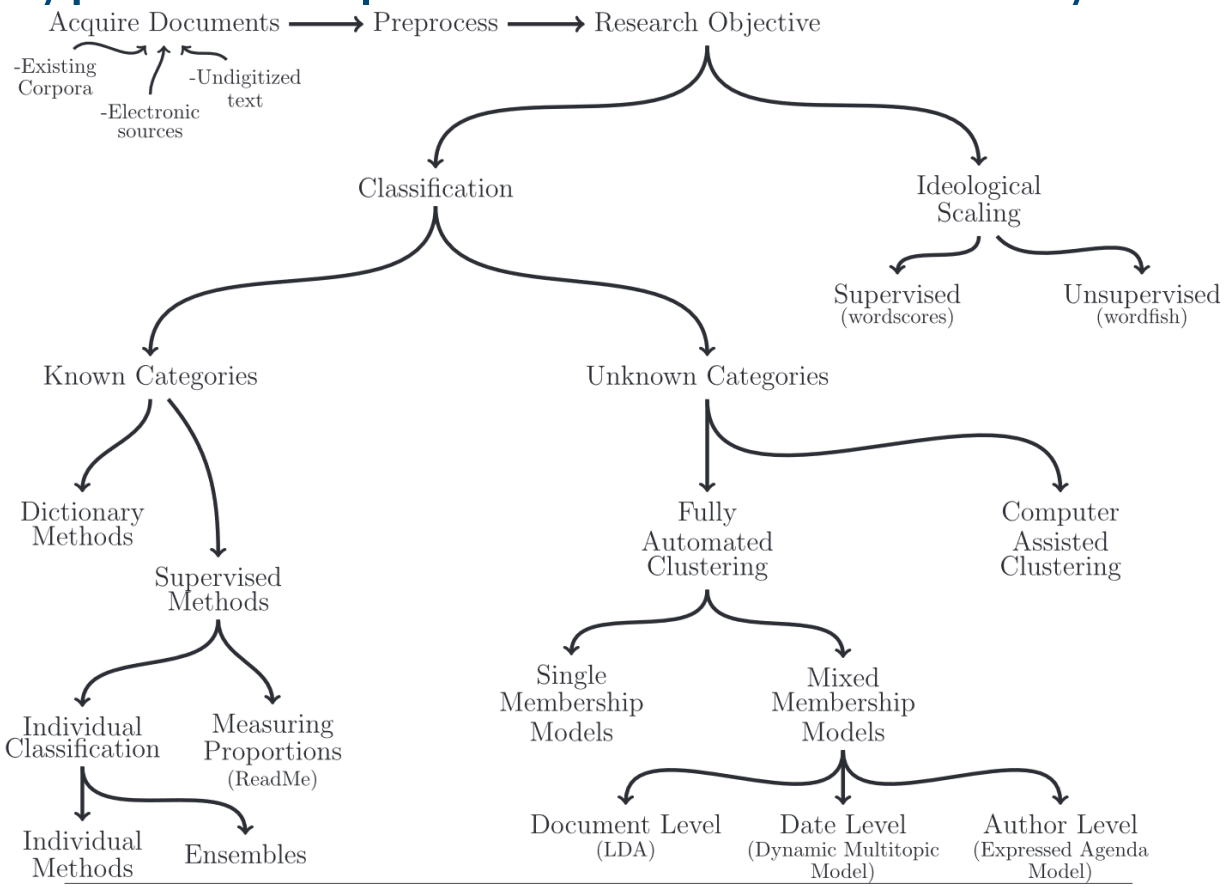
Quantitative Analyse ermöglicht es uns, Muster in großen Datenmengen zu erkennen, welche wir ansonsten nur schwerlich erschließen könnten.

Typen der quantitativen Textanalyse I

Es gibt verschiedene Klassifikationsversuche:

- unsupervised vs. supervised (Grimmer and Stewart 2013)
- regelbasiert, supervised, unsupervised und hybrid (Baden et al. 2020)

Typen der quantitativen Textanalyse II



Typen der quantitativen Textanalyse III

Neben diesen Typen unterscheiden wir im Seminar auch zwischen der Art, wie Textbestandteile analysiert werden...

- Bags-of-Words Ansätze $\rightarrow$ kein Kontext, Wörter werden in einen Topf/eine Tasche geworfen
- Embeddings $\rightarrow$ kontextabhängig

Ab wann quantitative Textanalyse?

Keine Konvention, abhängig von der Methode...:

- bei bags-of-words benötigen wir oft große Datenmengen, um valide und reliable Ergebnisse zu erhalten
- bei Embeddings reichen oftmals schon geringe Mengen an Daten

Bis zur nächsten Woche... I

- Lesen des Syllabus
  - Fehlt euch ein Thema?
  - Habt ihr weitere Literaturvorschläge?
  - Wisst ihr von einer Methode bereits, dass ihr sie nutzen möchtet?
  - Habt ihr noch Fragen zu den Seminarleistungen?

Bis zur nächsten Woche... II

- Vorbereitung des R-Crashkurses
  - Download von R sowie RStudio
  - Durcharbeiten der Tutorials
    - Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, June). *R for Data Science*. O'Reilly Media, Inc. https://r4ds.hadley.nz/
    - Ellis, A., & Mayer, B. (2024). Einführung in R. https://methodenlehre.github.io/einfuehrung-in-R/Einf%5C%C3%5C%BChrung-in-R.pdf

plain

Danke für eure Aufmerksamkeit!
Noch Fragen?

Literatur I

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020).Hybrid Content Analysis: Toward a Strategy for the Theory-driven, Computer-assisted Classification of Large Text Corpora. *Communication Methods and Measures*. Retrieved June 2, 2024, from https://www.tandfonline.com/doi/abs/10.1080/19312458.2020.1803247

Benoit, K. (2009). Introduction to quantitative text analysis.

Bonikowski, B., & Nelson, L. K. (2022).From Ends to Means: The Promise of Computational Text Analysis for Theoretically Driven Sociological Research. *Sociological Methods & Research*, *51*(4), 1469–1483. https://doi.org/10.1177/00491241221123088

Literatur II

Chun Ting-Ho, J. (2021). Introduction to Computational Text Analysis and Social Media Research using R. Retrieved October 5, 2024, from https://www.sciencespo.fr/ecole-recherche/en/news/introduction-computational-text-analysis-and-social-media-research-using-r

Ellis, A., & Mayer, B. (2024). Einführung in R. https://methodenlehre.github.io/einfuehrung-in-R/Einf%5C%C3%5C%BChrung-in-R.pdf

Grimmer, J., & Stewart, B. M. (2013).Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Literatur III

Stahlberg, D., & Sczesny, S. (2001).Effekte des generischen Maskulinums und alternativer Sprachformen auf den gedanklichen Einbezug von Frauen. *Psychologische Rundschau*, *52*(3), 131–140. https://doi.org/10.1026//0033-3042.52.3.131

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962).The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, *7*(4), 484–498. https://doi.org/10.1002/bs.3830070412

Turing, A. M. (1950).Computing Machinery and Intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Literatur IV

Vervecken, D., Hannover, B., & Wolter, I. (2013).Changing (S)expectations: How gender fair job descriptions impact children's perceptions and interest regarding traditionally male occupations. *Journal of Vocational Behavior*, *82*(3), 208–220. https://doi.org/10.1016/j.jvb.2013.01.008

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, June). *R for Data Science*. O'Reilly Media, Inc. https://r4ds.hadley.nz/