Mirko Wegemann (er/ihm)
Vergleichende Politikwissenschaft
Scharnhorststraße 100
48151 Münster
m.wegemann@uni-muenster.de
Sprechstunde: Mittwoch, 14:00-15:00 (nach Voranmeldung)

Universität Münster

Wintersemester 2024/2025
Quantitative Textanalyse

Kursdaten

Das Seminar findet wöchentlich am Mittwoch von 16 bis 18 Uhr (c.t.) statt. Die erste Sitzung ist am 9. Oktober 2024, die letzte Einheit findet am 22. Januar 2025 statt. Es richtet sich an Bachelor-Studierende und findet in deutscher Sprache statt.

Kursbeschreibung

Dieses Methodenseminar gibt einen Einblick in die quantitative Textanalyse, eine Art der Inhaltsanalyse, welche Texte anhand von numerischen Gemeinsamkeiten untersucht. Dabei werden die Student*innen im Laufe des Seminars lernen, (1) Textdaten von öffentlich zugänglichen Webseiten zu sammeln, (2) das Rohmaterial für verschiedene Analysen vorzubereiten sowie (3) verschiedene Techniken der quantitativen Textanalyse anzuwenden. Zudem werden die Studierenden ein Basisverständnis von neueren Entwicklungen der Textanalyse (wie von Transformer oder Large Language Modellen) entwickeln. Die einzelnen Sitzungen werden dabei sehr praxisorientiert sein und Studierenden die Möglichkeit geben, ein eigenes Projekt im Rahmen des Seminars zu realisieren. Hierbei werden sie eine eigene Forschungsfrage entwickeln, theoretische Erwartungen formulieren, Forschungsdaten erschließen und eine passende Methode der quantitativen Textanalyse anwenden.
Voraussetzung für die Teilnahme sind erste Erfahrungen mit der Statistiksoftware R. Student*innen müssen in allen Sitzungen einen Laptop mitbringen.

Lernziele

Am Ende des Seminars werden Studierende verschiedene Möglichkeiten erlernt haben, um Webinhalte automatisiert herunterzuladen. Darüber hinaus können Student*innen eine Vielzahl von Verfahren anwenden, um den Inhalt von Textdokumenten quantitativ zu analysieren.

In diesem Zusammenhang werden sie den Unterschied sowie die jeweiligen Vor- und Nachteile von Methoden kennenlernen, die mit (*supervised*) und ohne (*unsupervised*) den Input von Forscher*innen funktionieren. Sie lernen, wie Rohdaten in ein nutzbares Format transformiert werden und können verschiedene Themen eines Textdokuments identifizieren. Des weiteren lernen sie, diese Methodenkenntnisse zur Beantwortung einer substantieller politikwissenschaftlicher Frage anzuwenden. Am Ende des Seminars haben die Studierenden ein Grundverständnis von der Funktionsweise neuronaler Netzwerke und Large Language Modellen. Durch stetige Praxisanwendung werden Studierende ihre Kenntnisse in der Statistiksoftware `R` (und R-Studio) vertiefen. Obwohl das Seminar einen klaren Fokus auf das Erlernen einer Methode setzt, werden Student*innen auch einen ersten Einblick in empirische Anwendungsliteratur erhalten.

Voraussetzungen

Die Anzahl der vergebenen Credits hängt von dem Studiengang ab, in dem die Studierenden eingeschrieben sind. Pro ECTS sind etwa 30 Arbeitsstunden angedacht. Die finale Benotung setzt sich aus den folgenden Komponenten zusammen (Stundenaufwand gemessen an 7 ECTS).

- regelmäßige Anwesenheit und aktive Diskussionsbeteiligung — 21 Stunden
- **Seminarleistung**: Anwendung und anschließende Präsentation einer zuvor erlernten Methode (Präsentation in der darauffolgenden Woche)
- **Prüfungsleistung**: Seminararbeit

} 190 Stunden

Anwendung einer Methode

Als verpflichtende Seminarleistung müssen Studierende mindestens eine der erlernten Methoden aus Sitzung 4-6 bzw. 8-12 in der Folgewoche an eigenen/anderen Daten anwenden. Dabei müssen sie eine Arbeitsfrage entwickeln, diese aber nicht weiter erörtern (Literaturüberblicke und Theorieentwicklung ist nicht nötig).
Student*innen müssen ein R-Markdown Skript anwenden, wobei kritische Schritte in der Analyse kommentiert werden sollten. In der Folgesitzung (in der darauffolgenden Woche) soll die Anwendung anhand eines kurzen Impulsvortrags vorgestellt werden. In diesem Vortrag sollten auch kurz auf Herausforderungen und Schwierigkeiten bei der Analyse eingegangen werden.

Das Seminar kann nur bestanden werden, wenn sowohl die Anfertigung eines Skripts sowie die Vorstellung im Plenum in der folgenden Sitzung erfolgt.

Seminararbeit

Diejenigen, welche eine Note für ihr Modul benötigen, müssen am Ende des Semester (spätestens zum 31.03.2025) eine Forschungsarbeit einreichen. Die Arbeit sollte aus circa 6000 Wörter ($\pm$ 10 Prozent) bestehen. Das Literaturverzeichnis sowie ein eventueller Anhang zählen nicht in die Wörtergrenze hinein.
Die Seminararbeit sollte sich an den im Seminar erlernten Methoden orientieren. Es muss **mindestens eine** Methode der quantitativen Textanalyse angewandt werden. Es handelt sich allerdings nicht um ein reines Methodenessay. Stattdessen sollen die Methode(n) auf eine selbst entwickelte wissenschaftliche Fragestellung angewandt werden. Die Struktur der Abschlussarbeit sollte also der eines empirischen Artikels in einer wissenschaftlichen Fachzeitschrift folgen (s. Beispielliteratur).
Dementsprechend sollte die Seminararbeit mit einem Forschungspuzzle beginnen (einem empirischen Phänomen, das durch bestehende Arbeiten nicht erklärt werden kann), die spezifische Forschungsfrage einführen und darlegen, warum es aus wissenschaftlicher und gesellschaftlicher Sicht wichtig ist, diese Frage zu untersuchen. Anschließend sollte eine zielgerichteter Literaturüberblick gegeben werden, auf deren Basis ein eigenes theoretisches Rahmenwerk entwickelt und Forschungshypothesen formuliert werden sollen. Daraufhin sollten die verwendeten Daten und die wichtigsten methodischen Schritte der Analyse erklärt werden. Schlussendlich werden die Ergebnisse präsentiert und kurz in einem Schlussteil eingebettet.
Wir werden in dem Seminar, insbesondere in der letzten Sitzung, die Erwartungen besprechen. Zudem sollten alle Studierenden zur letzten Sitzung eine grobe Idee für ihr Forschungsprojekt haben. Es empfiehlt sich, diese Idee mit mir in einer Sprechstunde abzusprechen.

Plagiate

Plagiarismus und sämtliche Formen des Ghostwritings sind verboten. Schriftliche Arbeiten werden mit Turnitin auf Plagiate überprüft. KI-Tools werden immer elementarer, insbesondere für die Textanalyse (s. unsere vorletzte Sitzung). Dementsprechend ist eine Verwendung von diesen Tools grundsätzlich möglich. Jedoch sollte transparent gemacht werden, in welchem Umfang welche Tools genutzt worden sind (Welche Prompts wurden bspw. genutzt?). Die Abgaben im Seminar müssen auf den Ideen der Student*innen basieren und eigenständig

umgesetzt werden. Es sollte also keine Arbeit eingereicht werden, die von einer KI geschrieben worden ist. Im Verdachtsfall muss die Prüfungsleistung bzw. Abschlussarbeit mündlich verteidigt werden.
Es wird erwartet, dass Student*innen wissenschaftliche Quellen nutzen und diese korrekt zitieren. Um den Workflow zu erleichtern und sich auf künftige Arbeiten (wie bspw. die Bachelorarbeit) vorzubereiten, wird den Studierenden empfohlen, Zitationssoftware zu nutzen. Die Universität Münster bietet über AcadCloud Zugang zu Citavi zu einem reduzierten Preis an. Alternativ kann die Open-Source-Software Zotero verwendet werden. Sollten Zweifel zur korrekten Zitierung bestehen, wenden Sie sich bitte an die Leitlinien zum wissenschaftlichen Arbeiten, die vom Institut für Politikwissenschaft bereitgestellt werden.

Inklusion

Dieses Seminar findet in einem inklusiven Raum statt. Dabei wird der Dozent geschlechtersensible Sprache verwenden. Die Teilnehmenden sind eingeladen, ihre Pronomen mit der Klasse zu teilen.
Des Weiteren wird versucht, eine möglichst dynamische Feedback-Kultur zu etablieren, in der Student*innen bereits während des Semesters Feedback zum Seminarverlauf geben können. Daher werden die Studierenden ermutigt, regelmäßig (anonyme) Rückmeldungen über Google Forms oder per E-Mail an mich zu senden.

Literatur und Ablauf

Lehrbücher

- **Über den generellen Workflow**
  Stoltz, D. S., & Taylor, M. A. (2024, März). *Mapping Texts: Computational Text Analysis for the Social Sciences* (1. Aufl.). Oxford University PressNew York. https://doi.org/10.1093/oso/9780197756874.001.0001

- Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach* (First edition). O'Reilly

- Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press. Verfügbar 27. September 2024 unter https://smltar.com/

- **Für Literaturwissenschaftler*innen, aber mit gutem Einblick in die Datenaufbereitung und einige Analyseformate:**
  Jockers, M. L., & Thalken, R. (2020). *Text Analysis with R: For Students of Literature* (2nd edition). Springer

Daten

Für eure eigenen Projekte (sowohl die Seminar- als auch die Prüfunsleistung) könnt ihr verschiedene Korpora verwenden. Einige Beispielkorpora sind hier gelistet. Im Laufe des Seminars (vor allem in den Scraping-Sitzungen) werden wir aber noch andere Daten erschließen.

- Baumann, M., & Gross, M. (2016). Where Is My Party? Introducing New Data Sets on Ideological Cohesion and Ambiguity of Party Positions in Media Coverage. Verfügbar 30. September 2024 unter http://www.mzes.uni-mannheim.de/d7/en/publications/report/where-is-my-party-introducing-new-data-sets-on-ideological-cohesion-and-ambiguity-of-party-positions-in-media-coverage

- Jankin, S., Baturo, A., & Dasandi, N. (2024, August). United Nations General Debate Corpus 1946-2023. https://doi.org/10.7910/DVN/0TJX8Y

- Lehmann, P., Franzmann, S., Al-Gaddooa, D., Burst, T., Ivanusch, C., Regel, S., Riethmüller, F., Volkens, A., Weßels, B., Zehnter, L., Wissenschaftszentrum Berlin Für Sozialforschung (WZB) & Institut Für Demokratieforschung Göttingen (IfDem). (2024). Manifesto Project Dataset. https://doi.org/10.25522/MANIFESTO.MPDS.2024A

- Rauh, C., & Schwalbach, J. (2020, März). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. https://doi.org/10.7910/DVN/L4OAKN
- Open Knowledge Foundation Deutschland e.V. (2024). kleineAnfragen. Verfügbar 4. Oktober 2024 unter https://kleineanfragen.de/
- Diverse Zeitungskorpora, zugänglich über CLARIN
- Diverse Rechtskorpora

1. Woche: Einführung [9.10.2024]

Es muss nichts vorbereitet werden, Teilnehmende sollten aber ihre Laptops mitbringen.

- Atwell, E. (1999). Computers break the language barrier. *The guardian*. Verfügbar 4. Oktober 2024 unter https://www.theguardian.com/education/0099/oct/17/tefl.news
- Tapper, J. (2023). Authors shocked to find AI ripoffs of their books being sold on Amazon. *The guardian*. Verfügbar 4. Oktober 2024 unter https://www.theguardian.com/technology/2023/sep/30/authors-shocked-to-find-ai-ripoffs-of-their-books-being-sold-on-amazon

2. Woche: R-Crashkurs [16.10.2024]

In dieser Sitzung widmen wir uns der Statistiksoftware `R`. Nach einem kurzen Überblick werden wir die grundlegende Syntax der Softwaresprache gemeinsam erschließen. Sollten keinerlei `R`-Vorkenntnisse bestehen, muss zumindest eine der folgenden Einführungen als Vorbereitung für das Seminar durchgearbeitet werden.

- **Ausführliche, englischsprachige Einführung in `R`:**
  Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023, Juni). *R for Data Science*. O'Reilly Media, Inc. https://r4ds.hadley.nz/
- **Kürzere Einführung in deutscher Sprache:**
  Ellis, A., & Mayer, B. (2024). Einführung in R. https://methodenlehre.github.io/einfuehrung-in-R/Einf%5C%C3%5C%BChrung-in-R.pdf
- Für eine **Kurzanleitung wesentlicher *dplyr*-Funktionen** sind Tutorials auf YouTube hilfreich, z.B. die folgende Playlist.

3. Woche: Was ist quantitative Textanalyse? [23.10.2024]

Wir beschäftigen uns in dieser Sitzung mit der quantitativen Textanalyse als Spezialform der Inhaltsanalyse. Dabei lernen wir u.a. kennen, in welchem Kontext sich quantitative Analysen von Text anbieten.

- **Kapitel 2 zur konzeptionellen Klärung von Inhaltsanalyse und potentiellen Schwierigkeiten der quantitativen Textanalyse:**
  Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology* (Fourth Edition). SAGE

- 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

Optional: Benoit, K. (2020). Text as Data: An Overview. In L. Curini & R. Franzese (Hrsg.), *The SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE Publications Ltd. https://doi.org/10.4135/9781526486387

4. Woche: Automatisierte Datensammlung 1 – Web-Scraping statischer Seiten [30.10.2024]

Diese Woche startet mit dem ersten von insgesamt drei Blöcken, welche sich mit der automatisierten Sammlung von Web-Daten beschäftigt. Wir werden uns in dieser Sitzung viel mit statistischen Webseiten, welche auf HTML basieren, auseinandersetzen.

- **Kapitel 1 und 2**:
  Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley

- Um auf bestimmte Objekte von Webseiten zuzugreifen, benötigen wir ein **Grundverständnis von HTML-Strukturen**. Dementsprechend sollten Student*innen im Vorfeld das folgende Tutorial durchgehen

5. Woche: Automatisierte Datensammlung 2 – Web-Scraping dynamischer Seiten [6.11.2024]

Der zweite Teil des Blocks zur automatisierten Datenbeschaffung widmet sich komplexeren Webseiten, welche auf JavaScript basieren und nicht (allein) über HTML angesteuert werden können.

- **Sektion 9.1.9:**
  Munzert, S. (2015). *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (1st ed.). Wiley
- **Als Hintergrund zu dem Package, welches wir nutzen:**
  Harrison, J., & Ju Yeong, K. (2022, September). RSelenium: R Bindings for 'Selenium WebDriver'. Verfügbar 6. September 2024 unter https://cran.r-project.org/web/packages/RSelenium/index.html

6. Woche: Automatisierte Datensammlung 3 – APIs und XML [13.11.2024]

In diesem dritten Teil der automatisierten Datensammlung kriegen wir einen Einblick in die Funktionsweise von APIs, das sind offizielle Schnittstellen, welche einen direkten Download von Daten ermöglichen. Wir werden dabei verschiedene APIs kennenlernen, u.a. die API des Manifesto Project und der NYT.
Des Weiteren werden wir uns mit der Aufbereitung von Daten im Dateiformat XML (**Ex**tensible **M**arkup **L**anguage) beschäftigen.

7. Woche: Datenaufbereitung und Vorbereitung für Analysen [20.11.2024]

Bevor wir Textdaten für Analysen verwenden können, müssen wir sie meist vorbereiten. Dabei konvertieren wir u.a. Dateien in unterschiedliche Formate oder entfernen bestimmte Features (z.B. störende Symbole, seltene Wörter, usw.). In dieser Sitzung lernen wir den gängigen Workflow für einfache Textanalysen kennen.

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

- Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press. Verfügbar 27. September 2024 unter https://smltar.com/ – insbesondere Kapitel 2-4
- Übersicht zu Regular Expressions in R
- Playground für Regular Expressions in R

8. Woche: Unsupervised Topic Models [27.11.2024]

In dieser Sitzung wenden wir zum ersten Mal eine quantitative Textanalyse an. Anhand von *Topic Models* versuchen wir Texte in verschiedene Kategorien zu klassifizieren.

- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R Package for Structural Topic Models. *Journal of statistical software*, *91*, 1–40. https://doi.org/10.18637/jss.v091.i02
- **Empirischer Anwendungsfall:**
  Bauer, P. C., Barberá, P., Ackermann, K., & Venetz, A. (2017). Is the Left-Right Scale a Valid Measure of Ideology? *Political behavior*, *39*(3), 553–583. https://doi.org/10.1007/s11109-016-9368-2

Optional: Mathematischer Hintergrund zu Latent Dirichlet Allocation:
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *J. mach. learn. res.*, *3*(null), 993–1022

9. Woche: Scaling [4.12.2024]

Während wir in der vorigen Woche einem *unsupervised* Ansatz (d.h. einer Analyse ohne unseren Input als Forscher*innen) gefolgt sind, werden wir in dieser Woche eine semi-supervised Methode anwenden, d.h. wir werden einen gewissen Input in das Modell einspeisen, um Texte anhand einer latenten Dimension zu skalieren.

- Watanabe, K. (2021). Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication methods and measures*, *15*(2), 81–102. https://doi.org/10.1080/19312458.2020.1832976

- **Empirischer Anwendungsfall:**
  Zollinger, D. (2024). Cleavage Identities in Voters' Own Words: Harnessing Open-Ended Survey Responses. *American journal of political science*, *68*(1), 139–159. https://doi.org/10.1111/ajps.12743

10. Woche: Einführung in Supervised Text Analysis und Basics von Machine Learning [11.12.2024]

In dieser Woche lernen wir einige Grundlagen für die *supervised* Analyse von Text. In diesem Zuge beschäftigen wir uns u.a. mit den Basics von maschinellem Lernen.

- **Einführung in Machine Learning, Kapitel 2:**
  James, G., Witten, D., Hastie, T., & Tibshirani. (2021). *An Introduction to Statistical Learning*. Verfügbar 19. September 2024 unter https://www.statlearning.com

- **Empirischer Anwendungsfall**:
  Gessler, T., & Hunger, S. (2022). How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration. *Political science research and methods*, *10*, 524–544. https://doi.org/10.1017/psrm.2021.64

11. Woche: Klassifikation anhand von Supervised Text Analysis [18.12.2024]

Nun nutzen wir eigene Daten, um diese in einen Trainins- und Testdatensatz zu unterteilen und einen Klassifikationsalgorithmus zu trainieren und die Performance dessen im Anschluss zu evaluieren.

- **Kapitel 7:**
  Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press. Verfügbar 27. September 2024 unter https://smltar.com/

- **Empirischer Anwendungsfall:**
  Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting Bots on Russian Political Twitter. *Big data*, *5*(4), 310–324. https://doi.org/10.1089/big.2017.0038

12. Woche: Word Embeddings und neuronale Netzwerke [8.1.2025]

Bisher haben wir uns mit *bags-of-words*-Ansätzen beschäftigt, welche den Kontext eines Wortes im Satz ausblenden. Word Embeddings beziehen diesen (teils) ein. Wir lernen sie hier kennen und wenden sie selbst an.

- **Einführung in Word Embeddings:**
  Rodriguez, P., & Spirling, A. (2021). Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research. *The journal of politics*. https://doi.org/10.1086/715162
- **Empirischer Anwendungsfall:**
  Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding Regression: Models for Context-Specific Description and Inference. *American political science review*, *117*(4), 1255–1274. https://doi.org/10.1017/S0003055422001228

13. Woche: Über R hinaus – Large Language Models [15.1.2025]

In den letzten Jahren gab es dramatische Fortschritte in der computergestützten Textanalyse. Moderne Transformer und Large Language Modelle können nicht nur Text in Kategorien einsortieren, sondern ihn übersetzen oder sogar neuen Text generieren. Wir kriegen hier einen Einblick. Die Sitzung wird dabei in Python umgesetzt.

- **Die Grundidee hinter Transformer-Modellen:**
  Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ukasz Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, *30*. Verfügbar 1. Juni 2024 unter https://proceedings.neurips.cc/paper%5C_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

- Wankmüller, S. (2022). Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis. *Sociological methods & research*, 1–77. https://doi.org/10.1177/00491241221134527

- **Podcast zu der Grundidee von KI:**
  Klein, E. (n. d.). A.I. Could Solve Some of Humanity's Hardest Problems. It Already Has. https://www.youtube.com/watch?v=U8CmAu68z5c

14. Woche: Wrap-Up Session [22.1.2025]

In dieser Sitzung werden wir rekapitulieren, welche Methoden wir im vergangenen Semester gelernt haben. Die Einheit bietet den Studierenden gleichzeitig die Möglichkeit, offene Fragen zu klären und mit der Ideenfindung für die Hausarbeit zu beginnen. Dabei sollen Student*innen einen kurzen 2-minütigen Elevator Pitch des geplanten Forschungsvorhabens vorbereiten.