

Quantitative Wissenschaftsforschung und Computational Social Science

Saïd Unger^{1*}, Lukas Erhard¹, Christian Koß¹, Oliver Wieczorek², Jan Riebling³, Mark Lutter³, Richard Münch², and Raphael H. Heiberger¹

¹ *Universität Stuttgart*

² *Universität Bamberg*

³ *Universität Wuppertal*

* *Kontakt: [Saïd Unger](#)*

16 & 17. September

1 Inhalt

Die quantitative Erforschung der Wissenschaft nutzt immer stärker die Potentiale umfangreicher Datenquellen und computerwissenschaftlicher Methoden. Eine daraus entstehende *Science of Science* greift dabei zur Analyse der sozialen Prozesse des wissenschaftlichen Systems auf Millionen an Datenpunkten zurück. Die große Zahl verfügbarer Daten macht deshalb gerade die (quantitative) Wissenschaftsforschung zu einem idealen Gegenstandsbereich interdisziplinärer Methoden einer entstehenden *Computational Social Science (CSS)*.

Während dieser Prozess international bereits fortgeschritten ist, gibt es in Deutschland bislang nur vereinzelte Standorte und Projekte, die die Möglichkeiten von CSS nutzen. Der Workshop des Projektverbunds "[Aufbruch, Abbruch oder Durchbruch](#)" (ABD) möchte zu diesem Zweck eine *praktische Einführung* zu dem Thema geben, mit Schwerpunkt auf relevante Daten Techniken und Fragestellungen aus der Wissenschaftsforschung. Zu diesem Zweck stellen wir die Möglichkeiten (und Probleme) der Akquise und Analyse großer Datenmengen vor, insbesondere aus dem *Web of Science*. Dabei wird der gesamte Prozess, vom Datenherunterladen, Extrahieren bis zu fortgeschrittenen Analysemethoden dargestellt. Unter Bezugnahme auf das Projektziel - wissenschaftliche Karriereverläufe - liefert der Workshop konkret eine Einführung in die folgenden Teilbereiche:

1. Abruf großer Mengen Publikationsdaten aus dem Web of Science mit *Python*

2. Anwendung von Techniken des *Natural Language Processing* auf wissenschaftliche Abstracts
3. Analyse von Koautoren-Netzwerken

2 Programm

Part I: WoS-Handling mit Python.

16.09.2021: *Lukas Erhard (Universität Stuttgart)*

1. *Begrüßung und kurze Programmvorstellung, 14.00 - 14.15*
2. *Datenerhebung, 14.15 - 15.45*
Einführung in die Arbeit mit verschiedenen Datenbanken in Python. Ziel ist die prototypische Erstellung eines Datensatzes und ein Einblick in *SQLAlchemy*. Die erhobenen Daten werden in den folgenden Workshops verwendet.
3. *Herausforderungen der Datendisambiguation, 16.15 - 17.45*
Nach einer kurzen Vorstellung der erhobenen Daten des Projektes ABD wird auf KB/WoS spezifische Herausforderungen bei der Disambiguierung von Autoren und Institutionen eingegangen.

Part II: Natural Language Processing mit WoS-Daten in Python.

17.09.2021: *Christian Koß (Universität Stuttgart)*

1. *Named Entity Recognition und Entity Matching, 9.00 - 10.30*
Es wird gezeigt wie Texte speziell für die Erkennung von Entitäten aufbereitet werden:
 - Muster-Erkennung von spezifischen Eigenheiten in Datensätzen/Dokumenten
 - Text-Cleaning-Techniken wie Stemming, PoS-Tagging und Noise removal
 - Matching von Entitäten über deren Features mit der Cosinus-similarity
 - Test und Validierung von Matchingverfahren
2. *Text Preprocessing for Topic Models + LDA Topic Modeling, 11.00 - 12.30*
Es wird gezeigt wie Texte speziell für die Erkennung von Topics aufbereitet werden:
 - Techniken zum Aufbereiten von größeren Texten
 - Verbesserung von Topic Models durch TF-IDF und Word2Vec
 - Anwendung des LDA Topic Models auf einen WoS-Korpus
 - Validierung und Parameter-Boosting von LDA

Part III: Soziale Netzwerkanalyse von Koautorschaft.
17.09.2021: Saïd Unger (Universität Stuttgart)

1. *Theorie, 14.00 - 15.30*

- Relevanz von Kollaborationen: Warum lohnt es sich Kollaborationen zu betrachten (Matthew Effect, steigende Relevanz von Teams etc.)
- Relevanz von Netzwerkforschung: Wie funktioniert Netzwerkforschung (Nodes und Edges, Zentralität, Dichte, Communities)

2. *Hands on Networks, 16.00 - 17.30*

- Netzwerke einladen, Rahmenbedingungen prüfen
- Metriken berechnen
- Visualisierung
- Ausblick (Modellierung)

3 Teilnahme und Voraussetzungen

Der Workshop wird vollständig online stattfinden. Ein entsprechender Link wird den Teilnehmenden spätestens eine Woche vorher per E-Mail zugesendet. Der Workshop ist für alle Interessierte zugänglich, wobei jedoch Mitglieder des BMBF-Förderschwerpunkts *Quantitative Wissenschaftsforschung* bei großer Nachfrage Teilnahmepriorität besitzen werden.

Zur Anmeldung bitte einfach nur eine Email unter Angabe des Namens, Institution und (falls zutreffend) Projekttitel im Förderschwerpunkt an [Saïd Unger \(said.unger@sowi.uni-stuttgart.de\)](mailto:said.unger@sowi.uni-stuttgart.de) senden.

Kenntnisse in der Programmiersprache *Python* werden vorausgesetzt, da eine Einführung darin den Rahmen des Kurses übersteigen würde. Dafür existieren zahlreiche Online-Ressourcen, von denen einige [hier](#) aufgeführt sind.

Zudem empfehlen wir, sich bereits vor dem Workshop einen Überblick über CSS und Wissenschaftsforschung zu verschaffen. Geeignete Literatur hierzu wäre [Evans and Aceves \(2016\)](#); [Fortunato et al. \(2018\)](#); [Heiberger and Riebling \(2016\)](#).

Evans, J. A. and Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, 42(1):21–50.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L. (2018). Science of science. *Science*, 359(6379):eaao0185.

Heiberger, R. H. and Riebling, J. R. (2016). Installing computational social science: Facing the challenges of new information and communication technologies in social science. *Methodological Innovations*, 9:1–11.