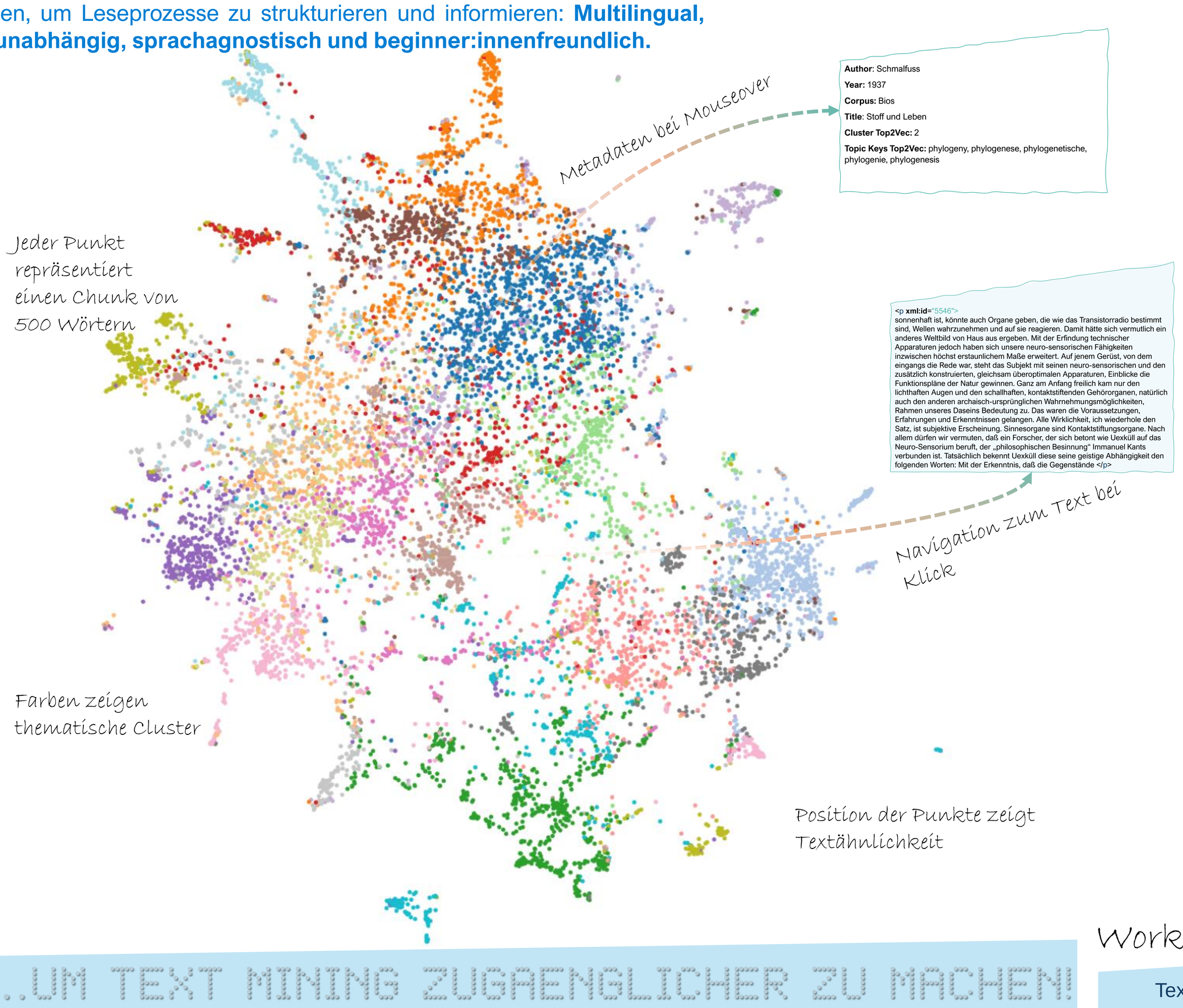
HermeneuTopic

Ein Workflow zur Exploration mehrsprachiger Textkorpora

Stefan Reiners-Selbach
Philosophische Fakultät, HHU Düsseldorf

Zurück zu den Texten...

Text Mining-Techniken wie Topic Modeling und Document Embedding nehmen immer mehr Einfluss auf die Geisteswissenschaften, da sie uns erlauben, große Textmengen zu verarbeiten und so neue Forschungsfragen zu stellen. Aber insbesondere in der Philosophie bleibt die Notwendigkeit, zum Text zurückzukehren: Menschliche Leser:innen bleiben entscheidend, hermeneutische Zugänge wichtig. **HermeneuTopic** zielt darauf ab, die Ergebnisse digitaler Analysen zu nutzen, um Leseprozesse zu strukturieren und informieren: **Multilingual, textlängenunabhängig, sprachagnostisch und beginner:innenfreundlich.**

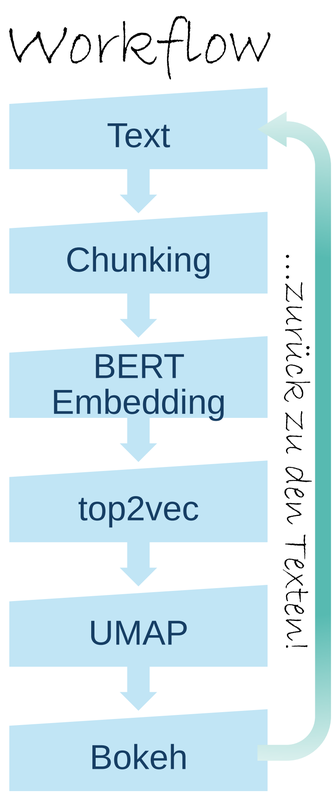


...um Text mithilfe eigener Modelle zu finden!

Die Philosophiegeschichte ist divers – insbesondere, wenn wir an Texten abseits vom Kanon interessiert sind! Text Mining-Techniken ermöglichen uns beispielsweise in der Wissenschaftstheorie und -geschichte Einblicke in die wissenschaftliche Praxis wie niemals zuvor [1]. Doch die für eine Fragestellung relevanten Texte können sich hinsichtlich Sprache, Textlänge und Format stark unterscheiden. Daher braucht es multilinguale Text Mining-Techniken die hinsichtlich dieser Punkte flexibel sind [2]. Solche Techniken können in das herkömmliche philosophische Arbeiten eingebunden werden und auch Beginner:innen ermöglichen, mit deren Ergebnissen umzugehen: Visualisierungen können als intuitive Interfaces zu Texten genutzt werden, um die zugrundeliegenden Textsammlungen thematisch und strukturell zu navigieren und so ohne Blick auf den kanonischen Status eines Textes eine Leseauswahl zu treffen [3].

Um dies zu erreichen, zerteilen wir Texte zunächst in Chunks von 500 Wörtern, auf welchen wir top2vec als embedding-basiertes Topic Modeling [4] anwenden, ohne weitere anspruchsvolle Vorverarbeitung, unter Benutzung eines multilingualen BERT-Models [5]. UMAP wird darauf zur Dimensionsreduktion benutzt [6]. Das resultierende zweidimensionale Streudiagramm wird dann mit Bokeh interaktiv visualisiert [7]: Jeder Punkt wird mit dem durch ihn dargestellten Text-Chunk im Kontext des Ursprungstextes verlinkt und entsprechend seines top2vec-Topics eingefärbt. Diese Topics können als thematische Gruppen gelesen werden. Die resultierende Visualisierung dient so als interaktive Karte des Korpus, welche die zugrundeliegende thematische Struktur aufzeigt. Bei Mouseover können Nutzer:innen Metadaten und die Topic Keys einsehen; bei Klick zum zugehörigen Text-Chunk navigieren.

Workflow



Literatur
[1] See e.g.: O. M. Lean, L. Rivelli, and C. H. Pence, “Digital Literature Analysis for Empirical Philosophy of Science’,” *British Journal for the Philosophy of Science*, vol. 74, 2023, doi: https://doi.org/10.1086/715049.
[2] M. Noichl, “PhilroBERTa: Ein multilinguales Sprachmodell zur Beantwortung philosophiehistorischer Fragestellungen,” in *DHd2023: Open Humanities, Open Culture.*; Q. Dombrowski, “What’s a ‘word’: Multilingual DH and the English Default.” Accessed: Jul. 19, 2023. [Online]. Available: https://www.quinndombrowski.com/blog/2020/10/15/whats-word-multilingual-dh-and-english-default/
[3] S. Reiners-Selbach, J. Baedke, A. Böhm, A. Fábregas Tejeda, and V. Straetmanns, “HermeneuTopic. Ein Workflow zur interaktiven mixed-methods Exploration (philosophie-)historischer Textkorpora,” in *Book of Abstracts DHd 2024*, 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.10698453
[4] D. Angelov, “Top2Vec: Distributed Representations of Topics.” arXiv, Aug. 19, 2020. doi: 10.48550/arXiv.2008.09470.
[5] N. Reimers and I. Gurevych, “Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation.” arXiv, Oct. 05, 2020. doi: 10.48550/arXiv.2004.09813.
[6] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.” arXiv, Sep. 17, 2020. doi: 10.48550/arXiv.1802.03426.
[7] Bokeh Development Team, *Bokeh: Python library for interactive visualization*. 2018. [Online]. Available: https://bokeh.pydata.org/en/latest/

hhu.de