Fakultät für
Mathematik und Informatik

Semantic Textual Similarity von textuellen Lernmaterialien

Fakultät für **Mathematik und Informatik**
Lehrgebiet Kooperative Systeme
Moritz C. Rieger & Niels Seidel

FernUniversität in Hagen

Agenda

1. Problemstellung und Zielsetzung
2. Ansatz
3. Evaluation
4. Ausblick

1. Problemstellung und Zielsetzung

Problemstellung

Wie lassen sich für textuelle Lernmaterialien semantisch ähnliche Materialien bestimmen?

Großes Angebot an Kursen

- > 1600 Kurse an der FernUniversität
- 134 Kurse an der Fakultät für Mathematik und Informatik
- bis zu 60 Wahlmöglichkeiten je nach Studiengang

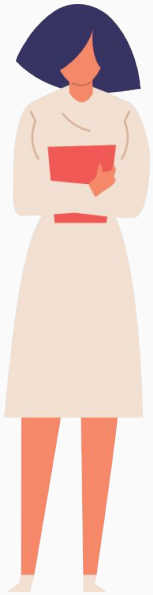
Aktuelle Entscheidungshilfen

- Modulhandbücher
- Veranstaltungs-Webseiten
- Probekapitel
- Skripte mit 300--500 Seiten

Zielgruppe

Studieninteressierte

Studierende



Lehrende

Grafik: pikisuperstar / Freepik

Stand der Forschung

Empfehlungssysteme für die Wahl von Lehrveranstaltungen [Li18, AL13, ZS10, LM14]

Visualisierung von sich ähnelnden Lehrveranstaltungen [Br16, BE15]

Natural Language Processing

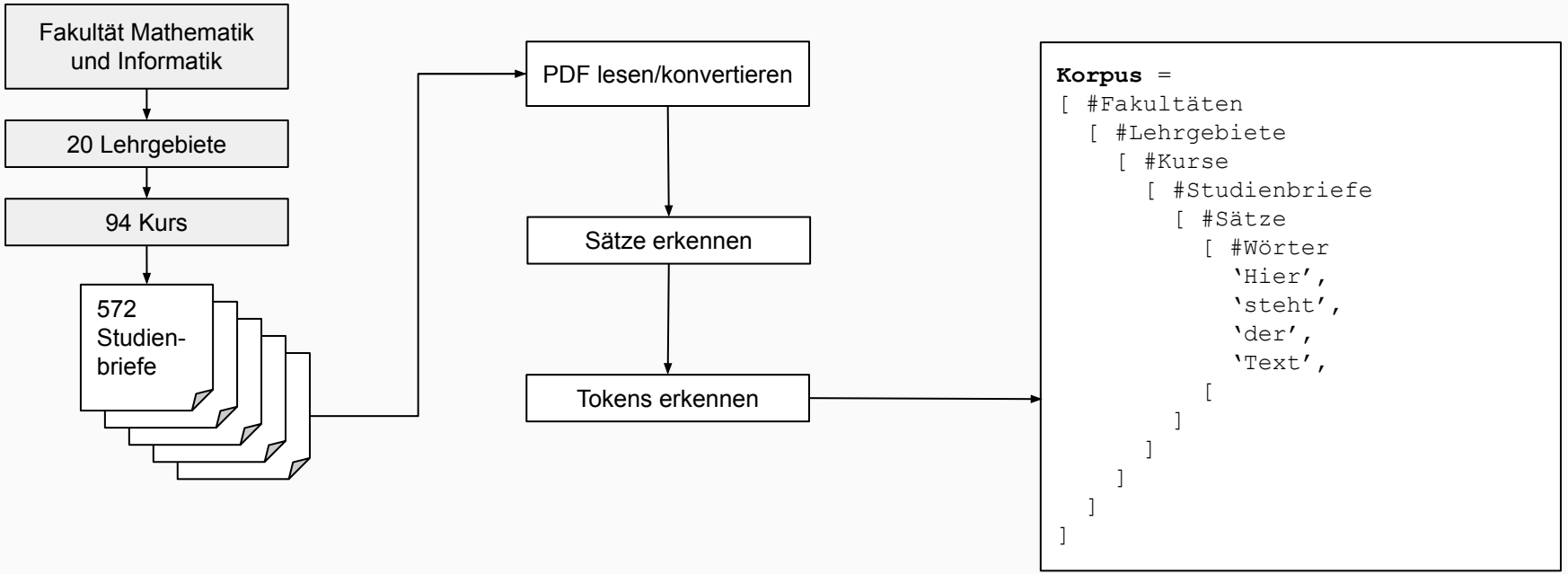
- local representations: N-grams [SS11], Bag-of-words, 1-of-N-coding
- continuous representations: LSA, LDA, Distributed Representations [DOL15]
- **Distributed Representations**
  - erfassen mehrere Grade von Ähnlichkeit [Mi13]
  - auf Dokumente übertragbar [LM14]
- (Semantische) Ähnlichkeit größerer Textmengen
  - encoder-decoder model [Zh15]
  - N-grams/ MinHash [SS11]
  - SemEval2018 Task 7 [Gal17]

Zielsetzung

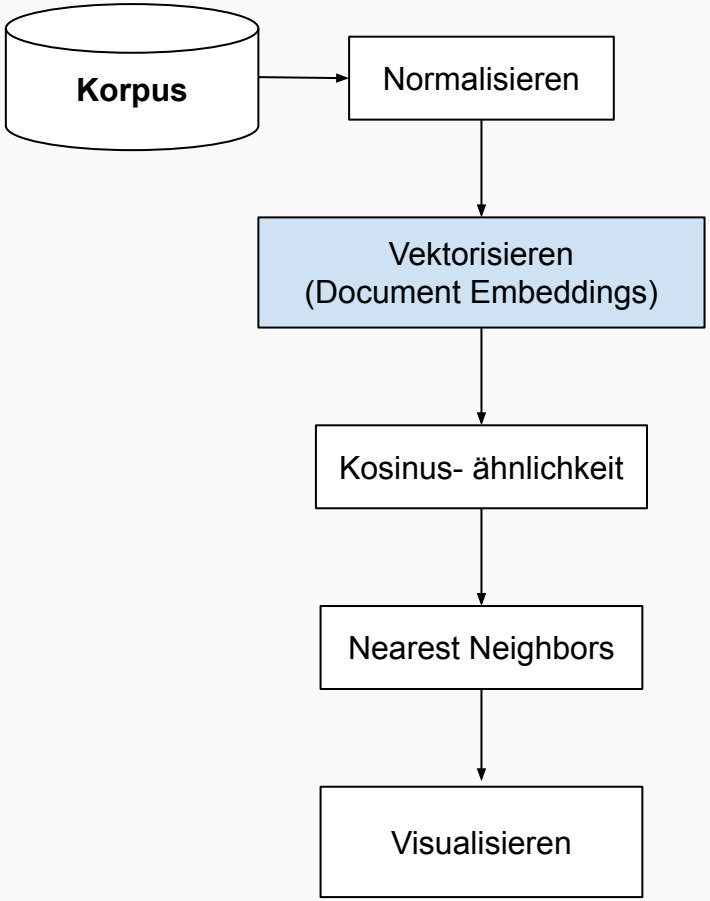
Entwicklung eines automatischen Verfahrens zur Analyse der semantischen Ähnlichkeit von Studientexten der Fakultät Mathematik und Informatik der FernUniversität in Hagen.

1. Korpus aus 94 Kursen mit 572 Studienbriefen erstellen
2. Ermittlung der semantische Ähnlichkeiten der Studienbriefe
3. Evaluation des Verfahrens
4. Nutzung der Ähnlichkeitsrelationen für verschiedene Anwendungen

2. Ansatz



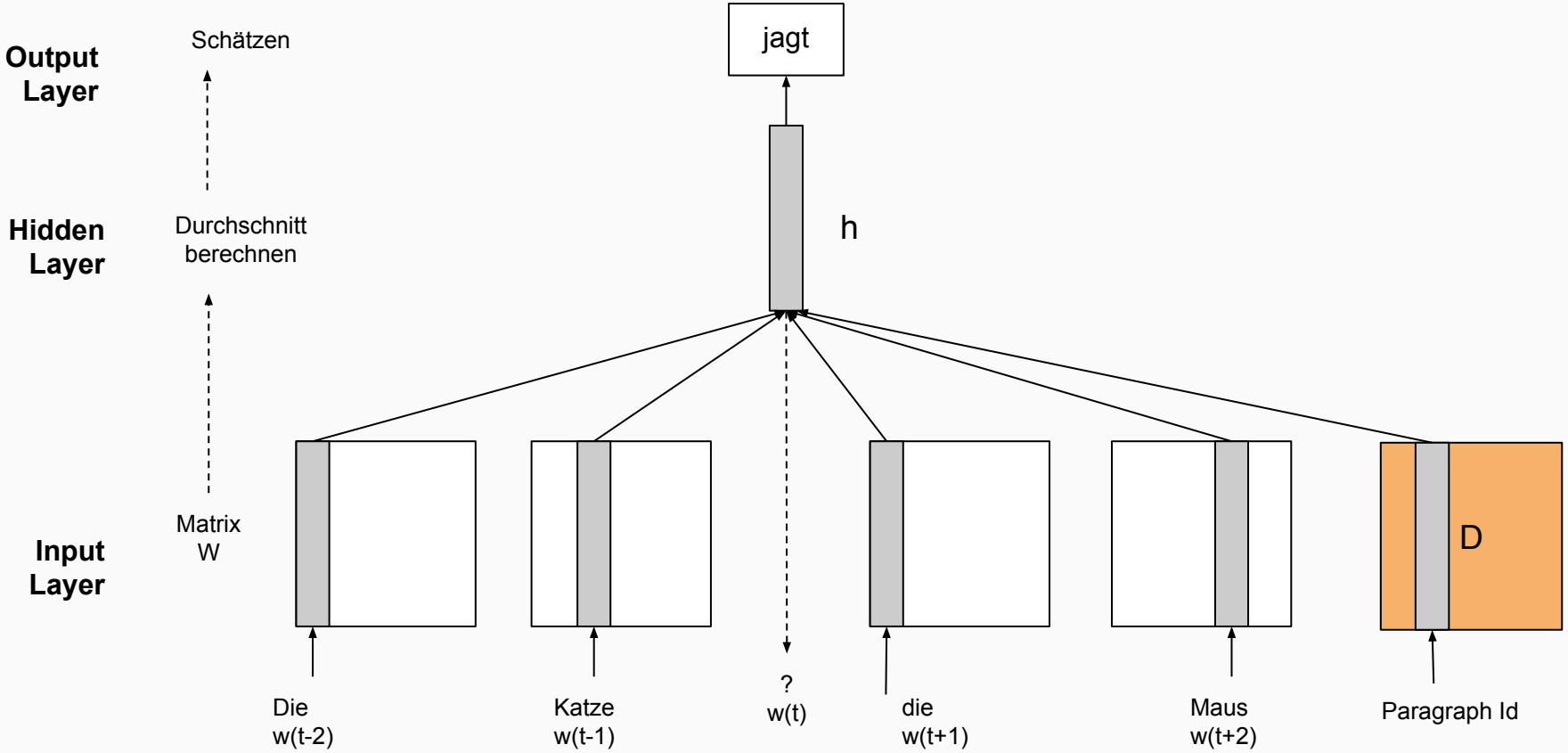
Erstellung des Korpus der Studienkurse



Ablauf der Analyse

"Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache."

– Ludwig Wittgenstein [Wi53]


Output Layer
Schätzen
jagt
Hidden Layer
Durchschnitt berechnen
h
Input Layer
Matrix W
D
Die w(t-2)
Katze w(t-1)
? w(t)
die w(t+1)
Maus w(t+2)
Paragraph Id

Document Embeddings

- enthalten Semantik als indirekte Folge der Schätzaufgabe
- geringe Dimensionalität (50-300)
- kontinuierliche Darstellung ermöglicht Vergleich von allen Wörtern
- Analogien zwischen Vektoren möglich
- Semantische Ähnlichkeit über Kosinusähnlichkeit messbar

- keine Information über den Inhalt der Dokumente
- keine Information weshalb Vektoren ähnlich sind
- Homonyme bleiben unberücksichtigt

3. Evaluation

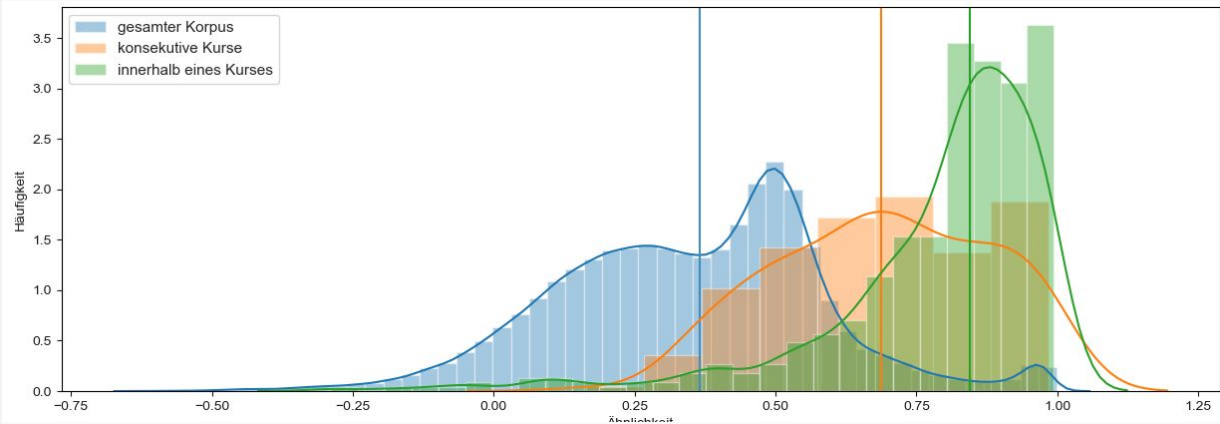
Plausibilitätsprüfung

Hypothese A

Studienbriefe eines Kurses sind sich ähnlicher, als Studienbriefe anderer Kurse.

Hypothese B

Studienbriefe von konsekutiven Kursen sind sich ähnlicher, als Studienbriefe anderer Kurse.

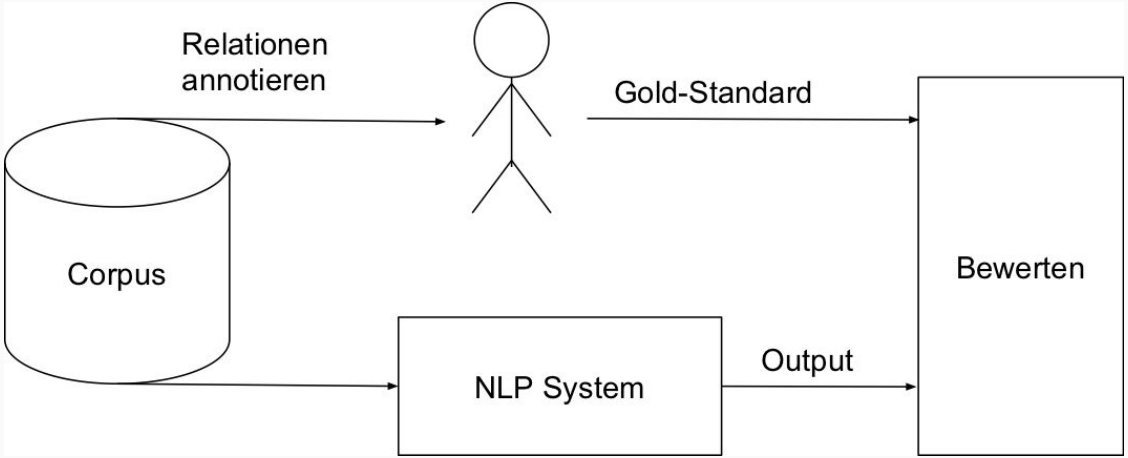


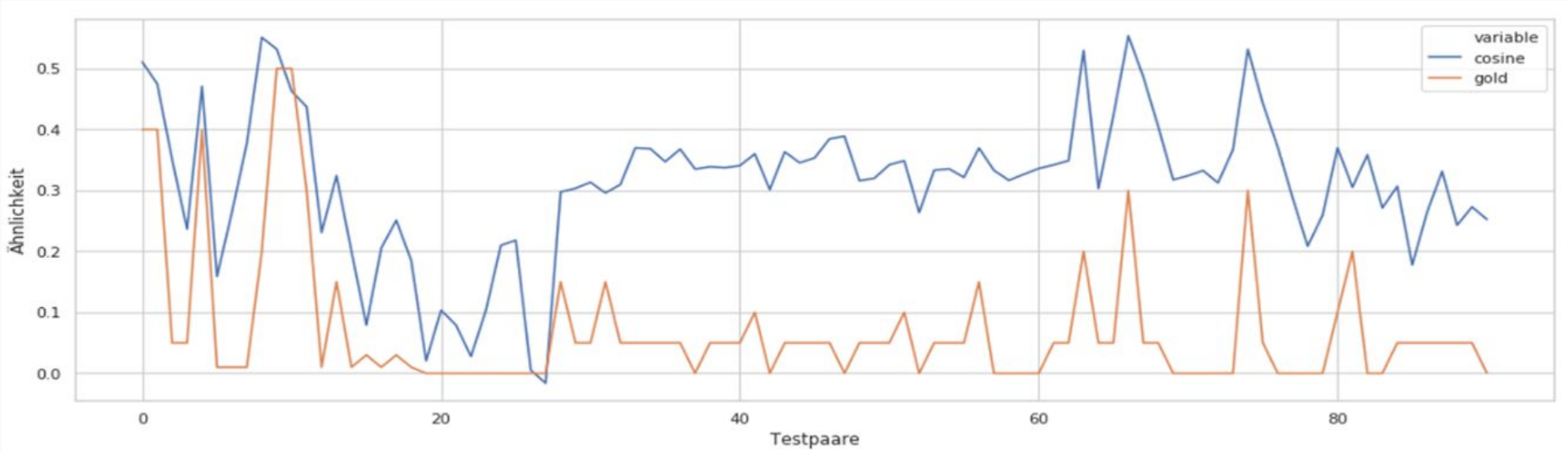
Goldstandard

Drei Autoren verglichen Kurseinheiten ihres Kurses mit Kurseinheiten eines ähnlichen Kurses.

kontinuierliche Bewertung: 0 → 100

3 x 28 Vergleiche





Korrelation mit dem Goldstandard

Pearson's r = 0,598

Kendalls τ = 0,451

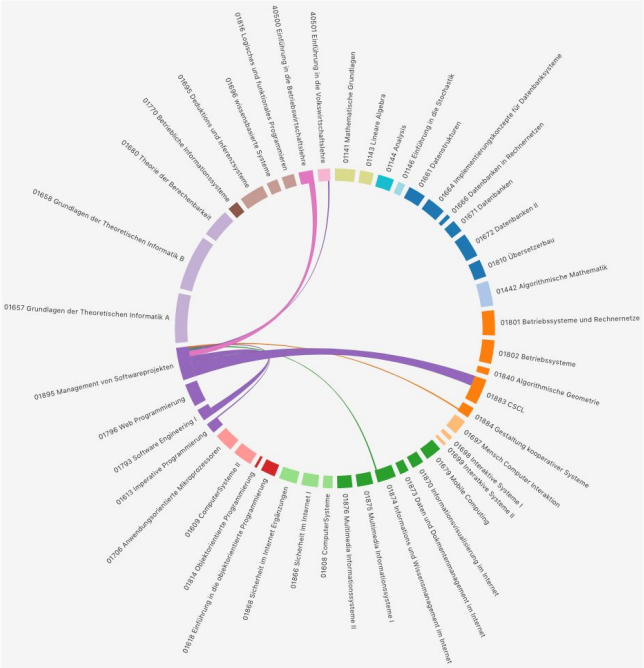
4. Ausblick

Weiterentwicklung des Verfahrens

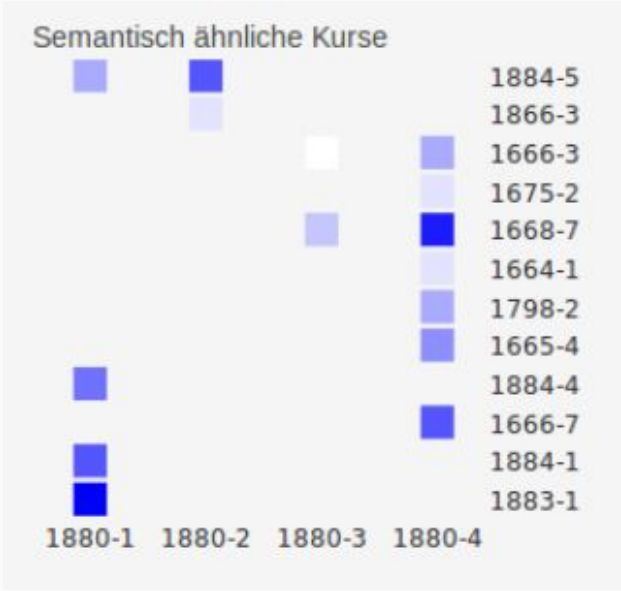
- Erklärbarkeit durch kapitelgenaue und seitenweise Analyse [SS11]
- Evaluation weiterer Embedding Modelle (BERT, ELMo, SpaCy)
- Erweiterung des Goldstandards
- Clustering von Themenschwerpunkten

Erste Anwendung des Verfahrens

Explorative Informationsplattform



Dashboard für Kursbetreuer

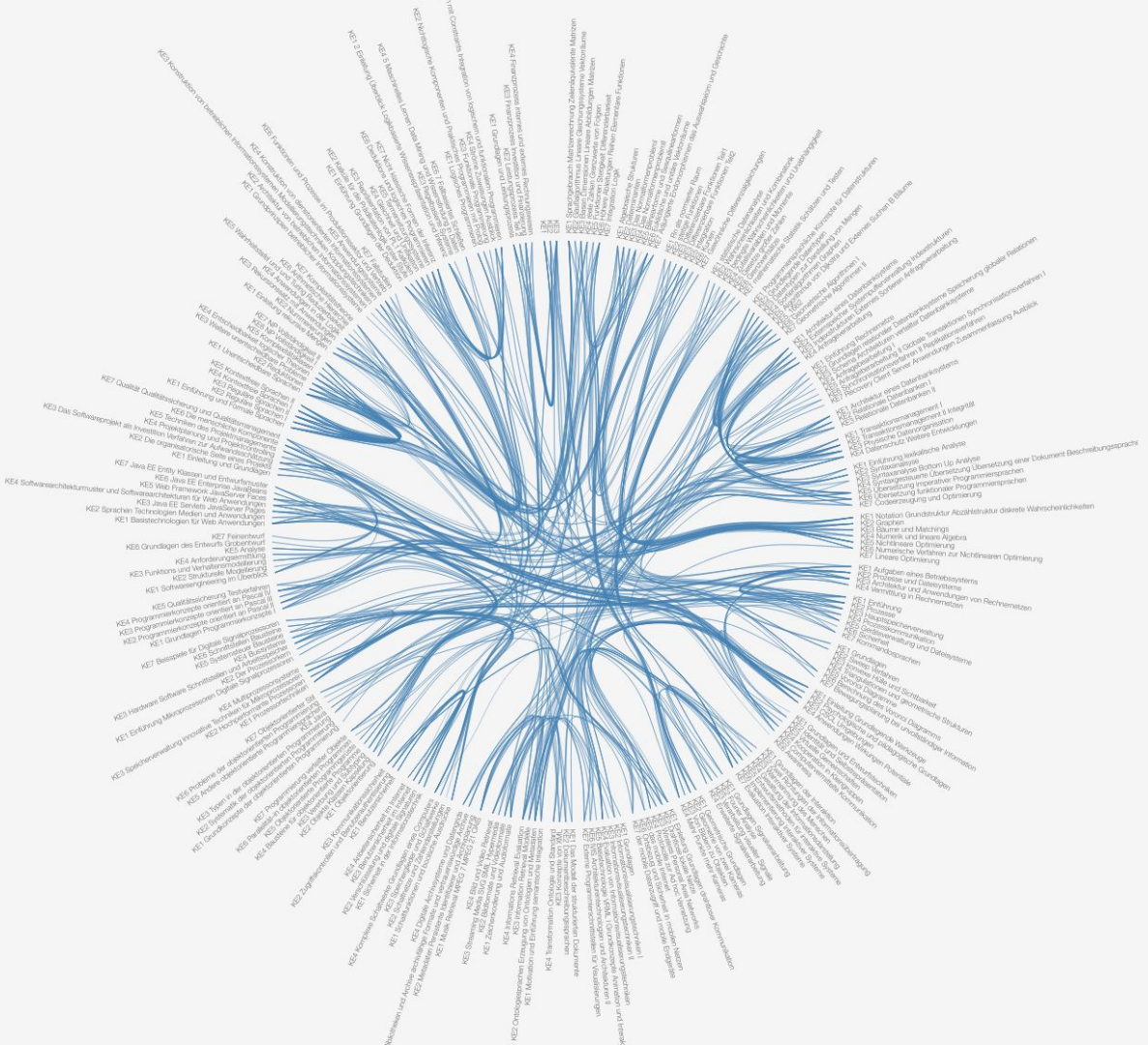


Vielen Dank!

Kontakt:

Moritz C. Rieger
moritz.rieger@posteo.de

Dr. Niels Seidel
niels.seidel@fernuni-hagen.de
Lehrgebiet Kooperative Systeme
Fakultät Mathematik und Informatik
FernUniversität in Hagen



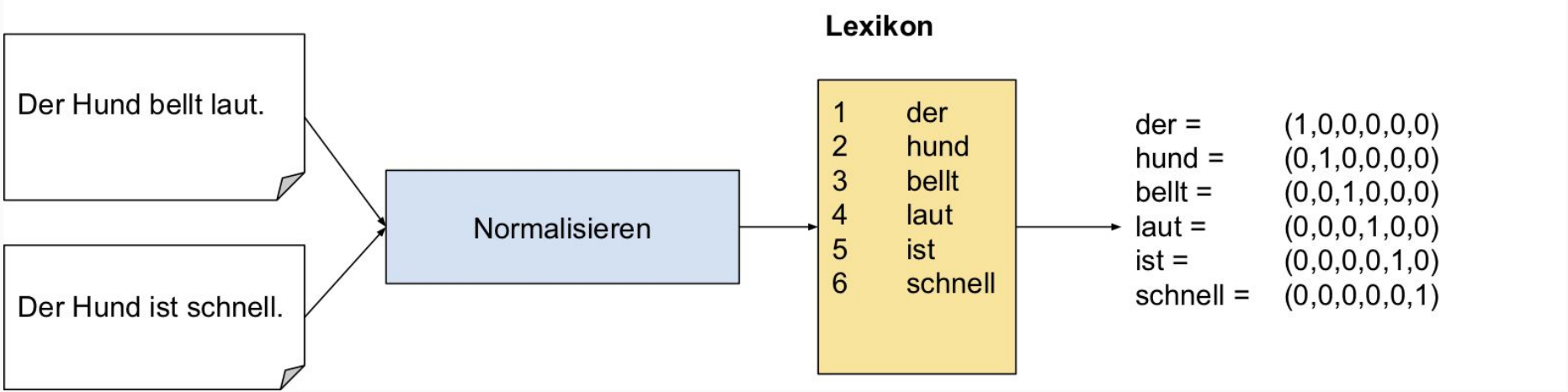
Literaturverzeichnis

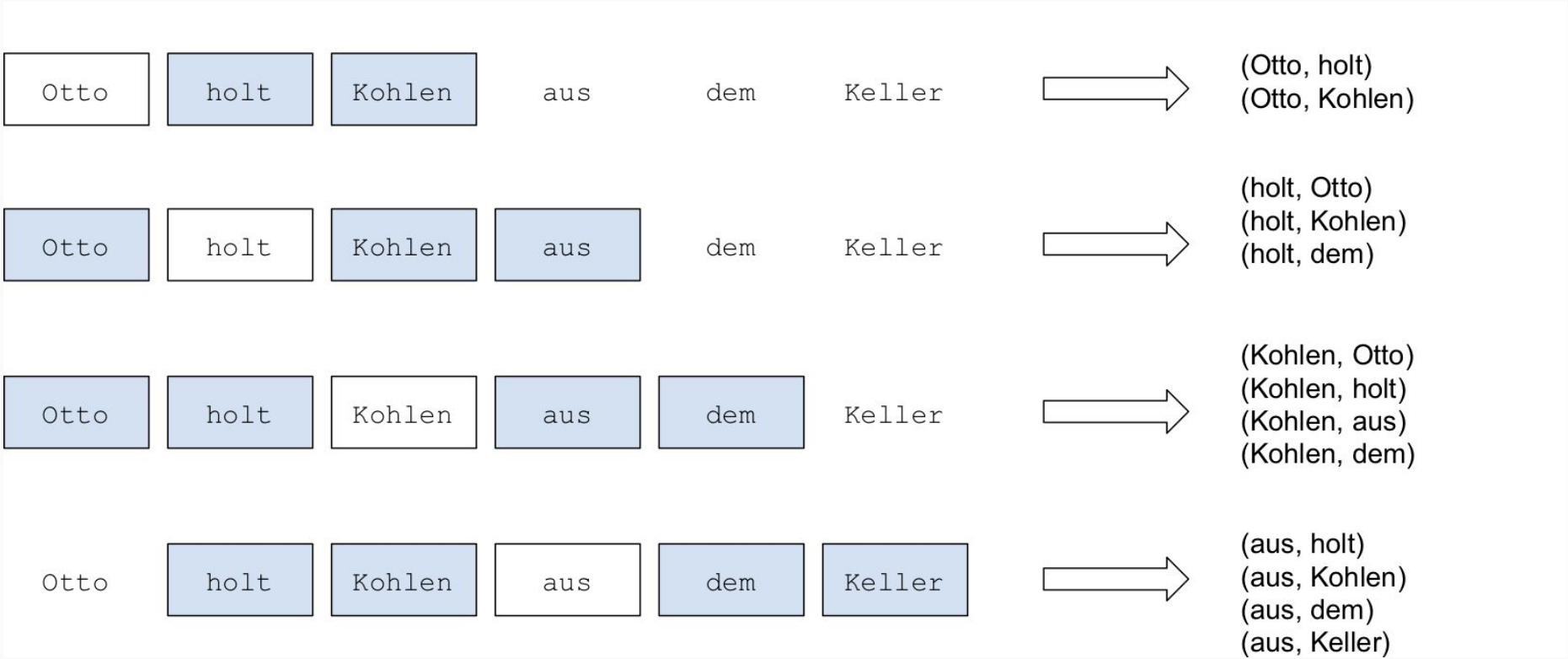
[AC18] Askinadze, A.; Conrad, S.: Development of an Educational Dashboard for the Integration of German State Universities' Data. In: Educational Datamining. S. 527–528, 2018.
[Ag12] Agirre, E. et al.: SemEval-2012 task 6: a pilot on semantic textual similarity. In: Proc. Of the first joint conf. On lexical and computational semantics. S. 385-393, 2012.
[AL13] Aher, S.B.; Lobo, L.M.R.J.: Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. Knowledge-Based Syst. 51, S. 1–14, 2013.
[BE15] Baumann, A; Endraß, M.A.A.: Visual Analytics in der Studienverlaufsplanung. In: Mensch und Computer 2015 Tagungsband. S. 467–469, 2015.
[BLK09] Bird, S.; Loper, E.; Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc., 2009.
[BOH11] Bostock, M.; Ogievetsky, V.; Heer, J.: D3 Data-Driven Documents. {IEEE} Trans. Vis.Comput. Graph. 17, S. 2301–2309, 2011.
[Br07] Brusilovsky, P.: Adaptive Navigation Support. In: The adaptive web. S. 263–290,2007.
[Br16] Brackhage, C. et al.: ModuleBase: Hochschulübergreifende Datenbank von Studiengangsmodulen, https://github.com/nise/moduleBase, 2016.
[Ce17] Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L.: SemEval-2017 Task 1:Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation, 2017.
[DJ13] D'Aquin, M.; Jay, N.: Interpreting Data Mining Results with Linked Data for Learning Analytics: Motivation, Case Study and Directions. In: Learning Analytics and Knowledge (LAK'13). S. S. 155–164, 2013.
[DOL15] Dai, A.M.; Olah, C.; Le, Q. V: Document Embedding with Paragraph Vectors. CoRR. Abs/1507.0, 2015.
[Ga17] Gábor, K. et al.: Exploring Vector Spaces for Semantic Relations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. S. 1814–1823, 2017.
[Gr09] Grefenstette, E.: Analysing Document Similarity Measures, https://www.cs.ox.ac.uk/publications/publication3348-abstract.html, 2009.
[Ka13] Kardan, A.A. et al.: Prediction of student course selection in online higher education institutes using neural network. Comput. Educ. 65, S. 1–11, 2013.
[Ki15] Kiros, R. et al.: Skip-Thought Vectors. CoRR. Abs/1506.0, 2015.
[Li18] Liu, B. et al.: Matching Long Text Documents via Graph Convolutional Networks.CoRR. Abs/1802.0, 2018.
[Li18b] Lin, J. et al.: Intelligent Recommendation System for Course Selection in Smart Education. Procedia Comput. Sci. 129, S. 449–453, 2018.
[LM14] Le, Q. V; Mikolov, T.: Distributed Representations of Sentences and Documents. CoRR. Abs/1405.4053, 2014.

Literaturverzeichnis

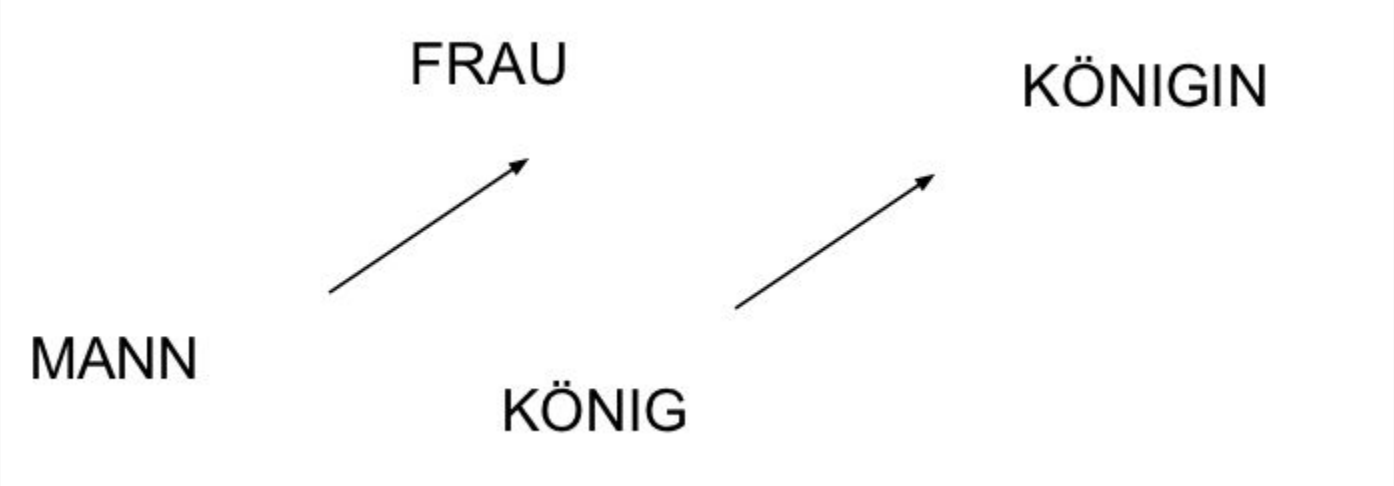
[Mi13] Mikolov, T.; Chen, K.; Corrado, G.; Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: ICLR. , Scottsdalte, Arizona, 2013.
[OGD16] Ognjanovic, I.; Gasevic, D.; Dawson, S.: Using institutional data to predict student course selections in higher education. Internet High. Educ. 29, S. 49–62, 2016.
[PSM14] Pennington, J.; Socher, R.; Manning, C.D.: GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP). S. 1532–1543, 2014.
[Ri18] Rieger, M.C.: Semantische Relationen in Studienbriefen der FernUniversität in Hagen. Universitätsbibliothek Hagen. http://nbn-resolving.de/urn:nbn:de:hbz:708-dh9737,2018.
[Sc16] Schwendimann, B. et al.: Perceiving learning at a glance: A systematic literature review of learning dashboard research. IEEE Trans. Learn. Technol. PP, 2016.
[SS11] Spasojevic, N.; Poncin, G.: Large Scale Page-Based Book Similarity Clustering. In: ICDAR 2011, 2011.
[Wi53] Wittgenstein, L.: Philosophical Investigations. In: New York: The Macmillan Company. S. 272. Blackwell, 1953.
[WIZ15] Weiss, S.M.; Indurkhya, N.; Zhang, T.: Fundamentals of Predictive Text Mining. Springer London, London, 2015.
[Zh15] Zhu, Y. et al.: Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. arXiv e-prints, 2015.
[ZS10] Zare-ee, A.; Shekarey, A.: The effects of social, familial, and personal factors on students' course selection in Iranian technical schools. Procedia - Soc. Behav. Sci. 9, S. 295–298, 2010.

Anhang


Lexikon
Der Hund bellt laut.
Der Hund ist schnell.
Normalisieren
1 der
2 hund
3 bellt
4 laut
5 ist
6 schnell
der = (1,0,0,0,0,0)
hund = (0,1,0,0,0,0)
bellt = (0,0,1,0,0,0)
laut = (0,0,0,1,0,0)
ist = (0,0,0,0,1,0)
schnell = (0,0,0,0,0,1)



Word2Vec: Erstellen eines Trainingssets mit Fenstergröße 2 [Ri18]

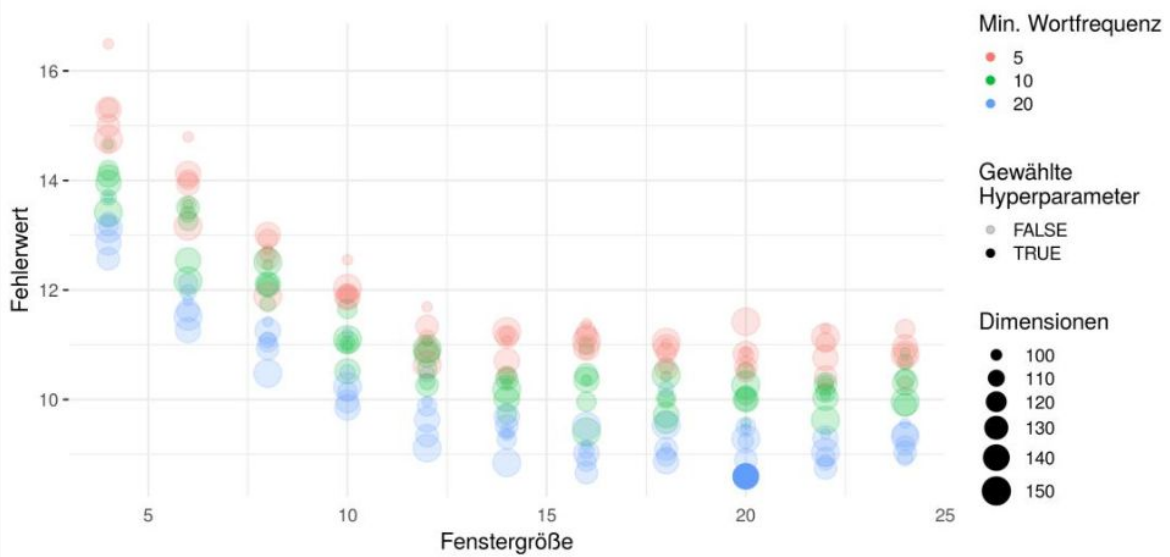


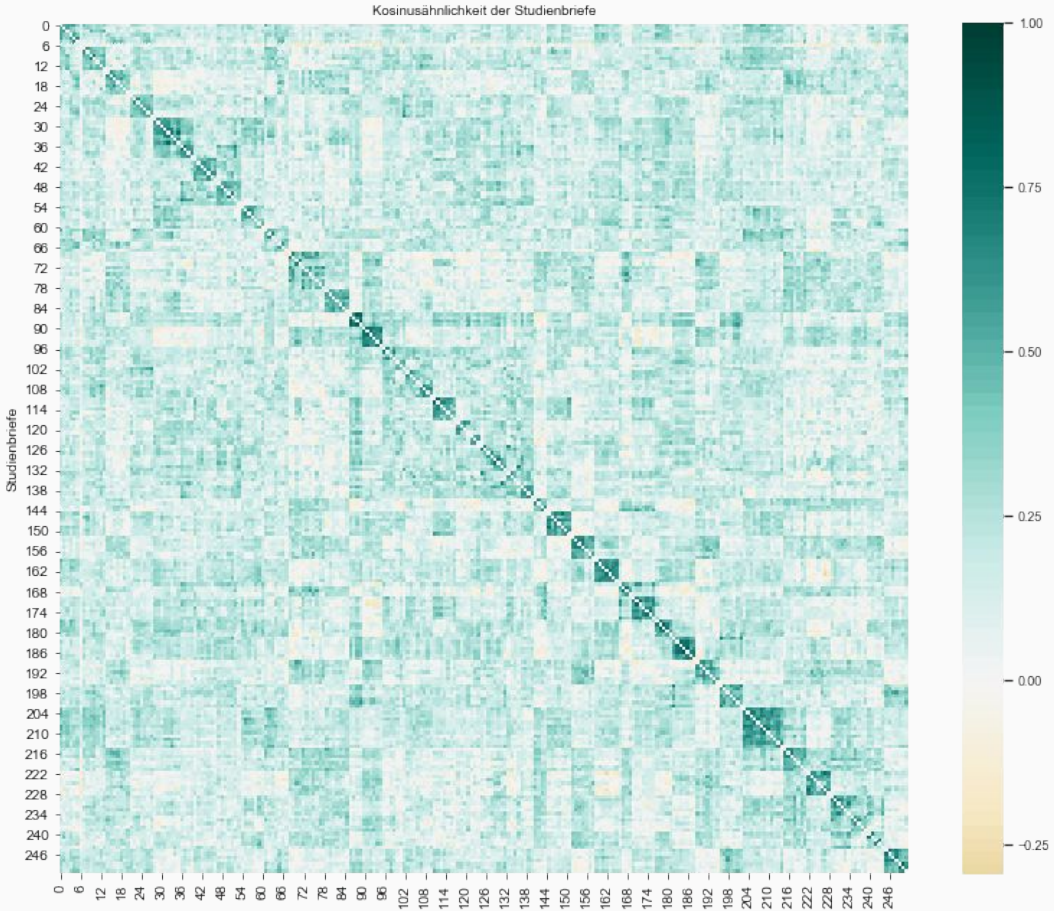
Semantische Relation von Wortvektoren [LM14]

Hyperparametertuning

mittlerer quadratischer Fehler zum Goldstandard bei:

Fenstergröße: 20
Wortfrequenz: 20
Dimensionen: 140





Kosinusähnlichkeit aller Dokumente im Korpus zueinander