

Prototypical Implementation and Assessment of Relatedness Search in Laws, Judgments and Commentaries

Master's Thesis – Final Presentation

Philipp Pickel, 19.12.2016; Munich

Chair of Software Engineering for Business Information Systems (sebis)

Faculty of Informatics

Technische Universität München

www.matthes.in.tum.de

1. Motivation

2. Research Objectives

3. Approaches

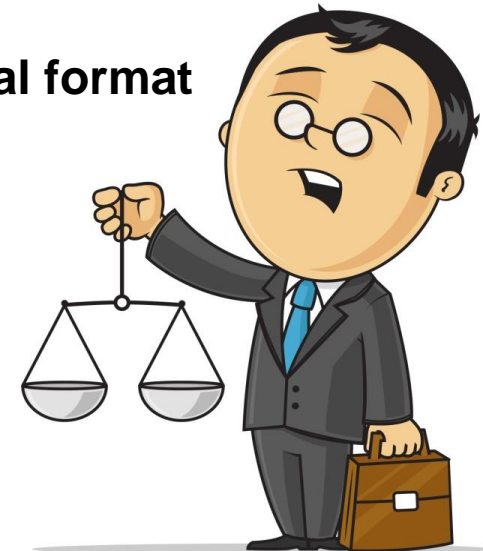
4. Live Demo

5. Assessment

6. Conclusion & Outlook

- **Precedents are important in lawyers' everyday work**
 - Case law in Anglo-American region
 - Continental Europe
 - High Courts give interpretations of laws
 - Past cases can be used as guidance
 - Transformation of precedents to common law
 - All similar cases can't be found without computer aid

- **Huge amounts of legal documents are available in digital format**
 - German federal laws available online
 - BGH publishes about 300 cases each month
 - Database of juris contains > 4 million documents
 - Beck-online provides > 2.5 million documents



1. Motivation

2. Research Objectives

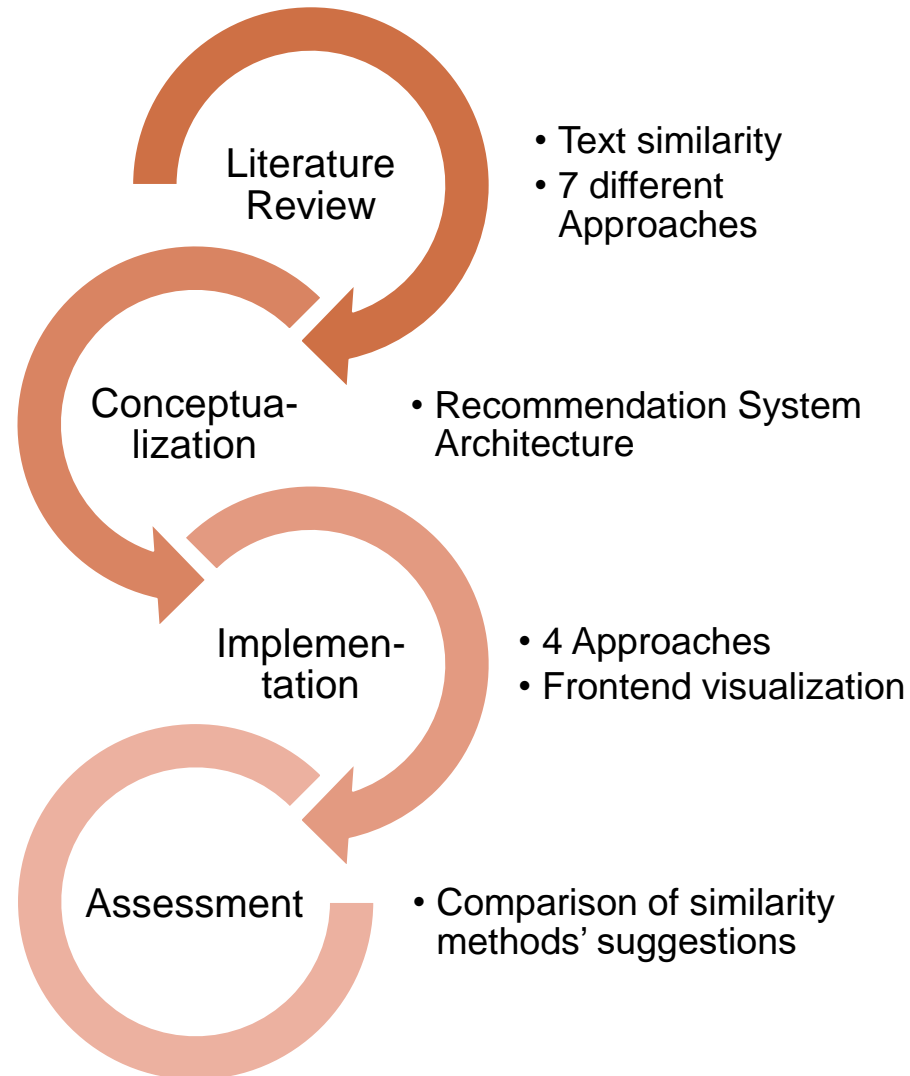
3. Approaches

4. Live Demo

5. Assessment

6. Conclusion & Outlook

- 1** What approaches can be used to measure similarity between two legal documents or parts of it?
- 2** What is a suitable approach to provide such similarity methods?
- 3** How can the results of a similarity search be presented to a user?
- 4** How well do selected similarity methods perform?



1. Motivation

2. Research Objectives

3. Approaches

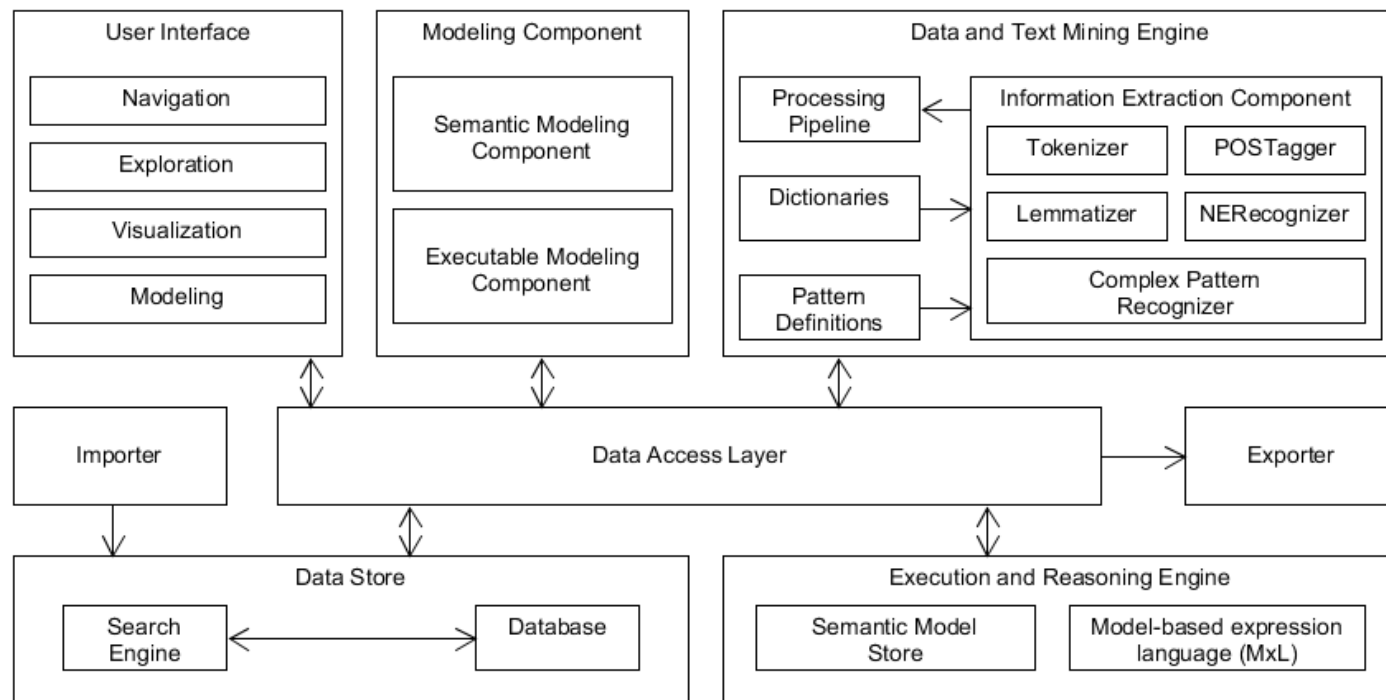
4. Live Demo

5. Assessment

6. Conclusion & Outlook

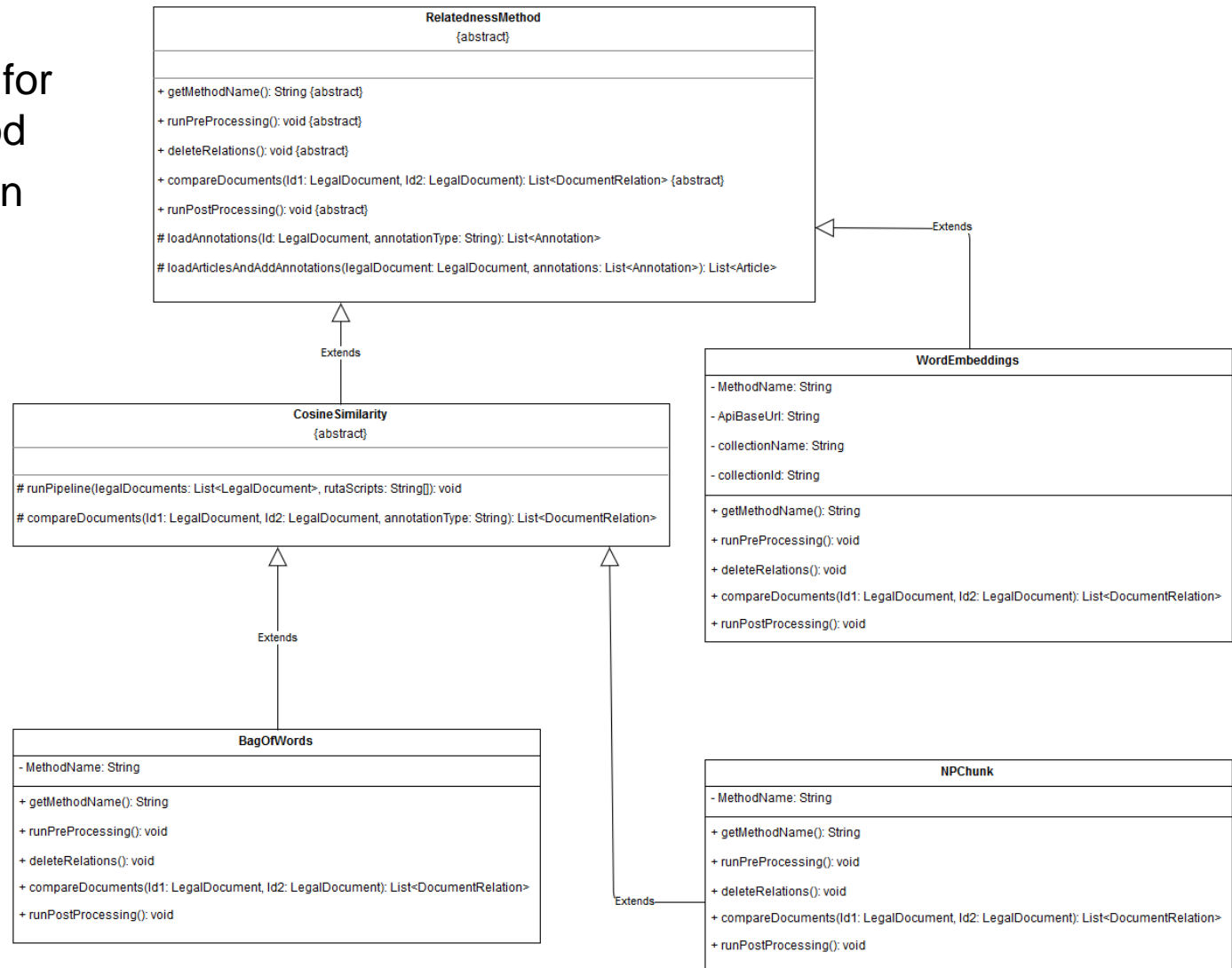
▪ LEXIA

- Part of the interdisciplinary research program Lexalyze
- Data Science Environment for Semantic Analysis of German Legal Texts
- Structured in different components



Architecture

- Same basis for every method
- Easy addition of methods



▪ Bag of Words

- Breaks up texts into set of words
- Represents texts as vectors in Vector Space Model
- Uses cosine similarity to determine similarity of vectors

(1) John likes to play football. Mary likes to watch movies.
(2) John likes to watch football games on TV.

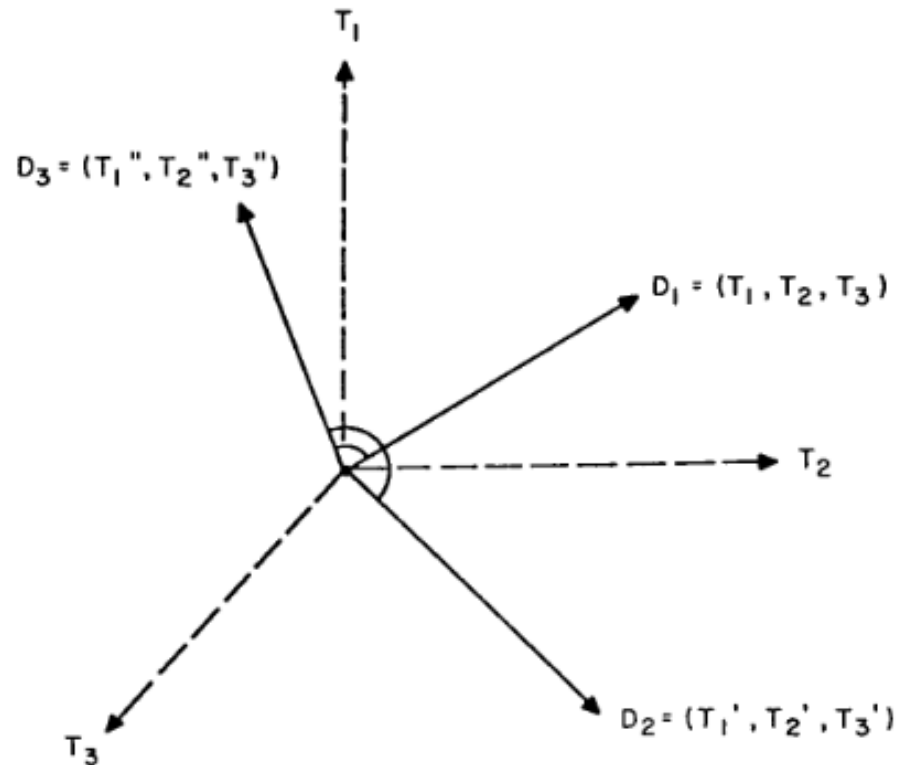


```
[  
  "John",  
  "likes",  
  "to",  
  "play",  
  "football",  
  "Mary",  
  "watch",  
  "movies",  
  "games",  
  "on",  
  "TV"  
]
```

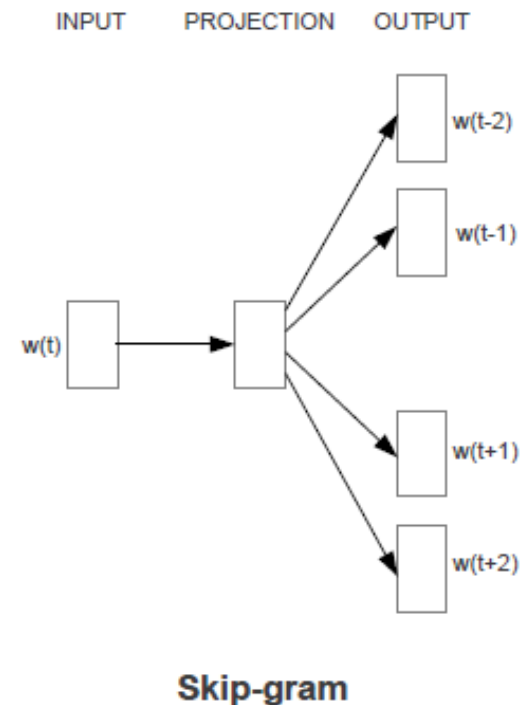
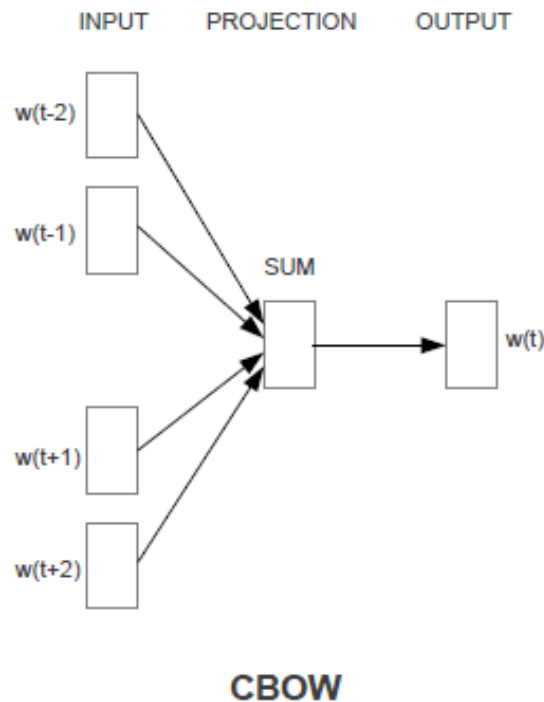


(1) [1, 2, 2, 1, 1, 1, 1, 1, 0, 0, 0]
(2) [1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1]

- **NPCChunk**
 - Uses natural language processing
 - Find Noun Phrases
 - Uses Vector Space Model



- **Word Embeddings**
 - Words are represented as vectors
 - Sum over word vectors results in text vector
 - Vectors are compared via cosine similarity



- **More Like This**
 - Part of elasticsearch Query DSL
 - Search with top n words regarding tf-idf from input text
 - Words are connected with logical OR in query
 - Scoring according Lucene algorithm
 - No a priori computation of relatedness



1. Motivation

2. Research Objectives

3. Approaches

4. Live Demo

5. Assessment

6. Conclusion & Outlook

1. Motivation

2. Research Objectives

3. Approaches

4. Live Demo

5. Assessment

6. Conclusion & Outlook

■ Survey Design

- Relevance rating of 93 norms towards a source norm
 - 4 Source norms (BGB §§ 280, 434, 985 and StGB §242)
 - Top 10 search results from every similarity method for each source norm
- Rating scale from 1 (not relevant) to 4 (very relevant)
- Carried out using google forms

■ Participants

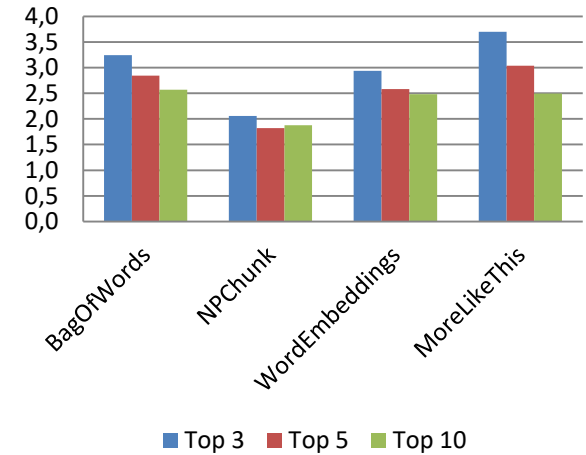
- Experts from legal domain
- Law students
- 11 usable answers

1. BGB §280 Schadensersatz wegen Pflichtverletzung	Bewertung von 1 (nicht relevant) bis 4 (sehr relevant)				
	1	2	3	4	k.A.
§ 282 Schadensersatz statt der Leistung wegen Verletzung einer Pflicht nach § 241 Abs. 2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 281 Schadensersatz statt der Leistung wegen nicht oder nicht wie geschuldet erbrachter Leistung	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 326 Befreiung von der Gegenleistung und Rücktritt beim Ausschluss der Leistungspflicht	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 1243 Rechtswidrige Veräußerung	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 634 Rechte des Bestellers bei Mängeln	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 1165 Freiwerden des Schuldners	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 283 Schadensersatz statt der Leistung bei Ausschluss der Leistungspflicht	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 437 Rechte des Käufers bei Mängeln	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 78 Festsetzung von Zwangsgeld	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 312a Verhältnis zu anderen Vorschriften	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 284 Ersatz vergeblicher Aufwendungen	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 289 Zinseszinsverbot	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 867 Verfolgungsrecht des Besitzers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 825 Bestimmung zu sexuellen Handlungen	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
§ 678 Geschäftsführung gegen den Willen des Geschäftsherrn	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

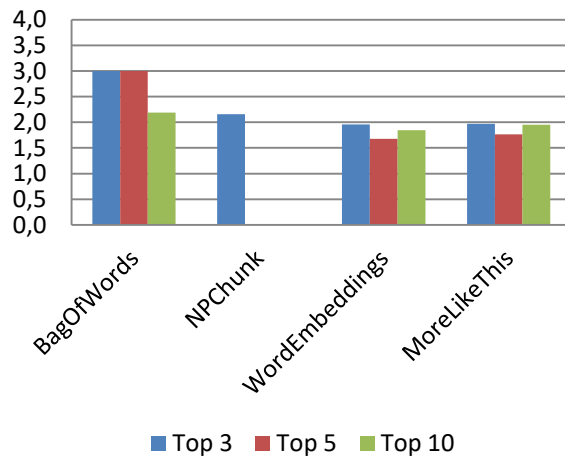
Results

- Ranking within methods works
- No method always superior to others
- Short texts seem to be better input
- NPChunk might have suffered from short texts

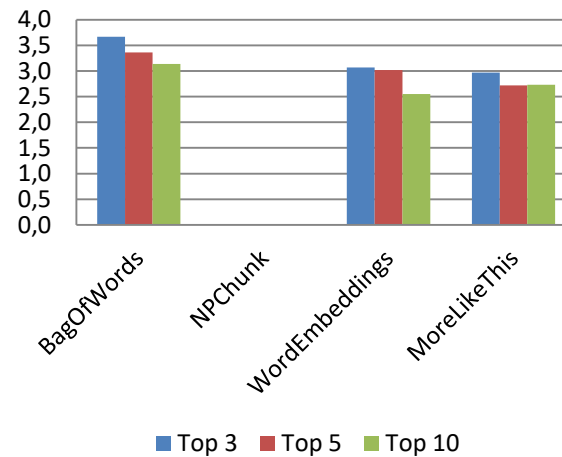
BGB § 280



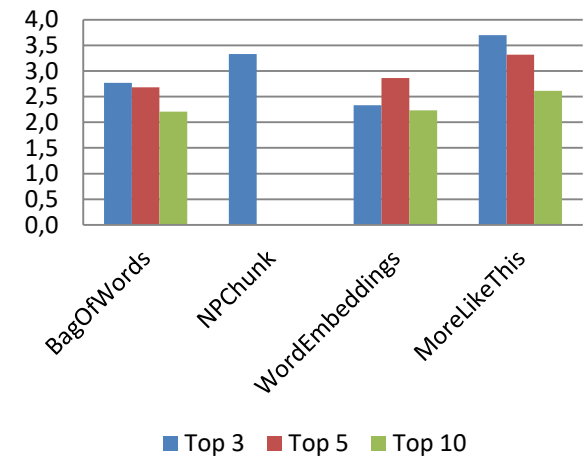
BGB § 434



BGB § 985

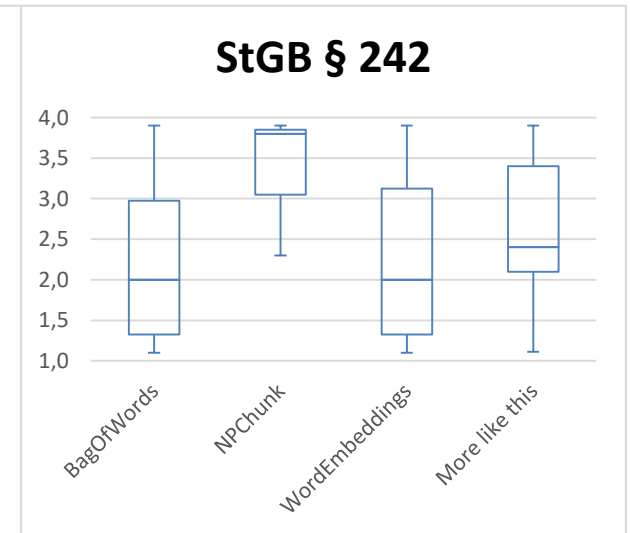
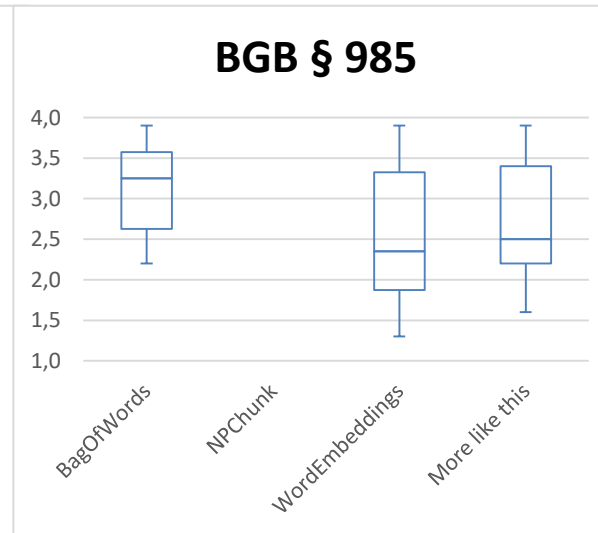
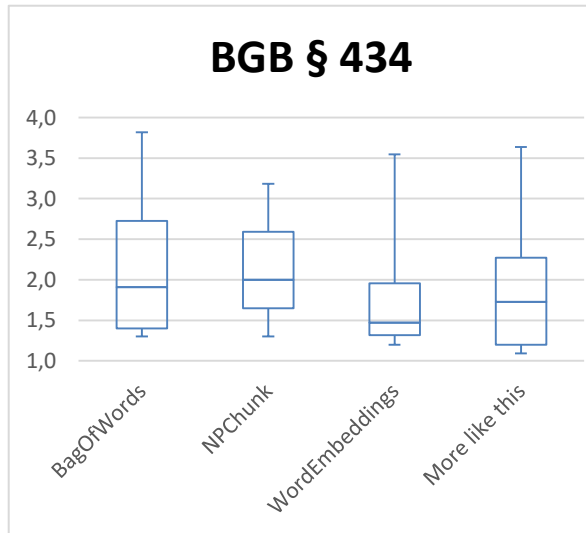
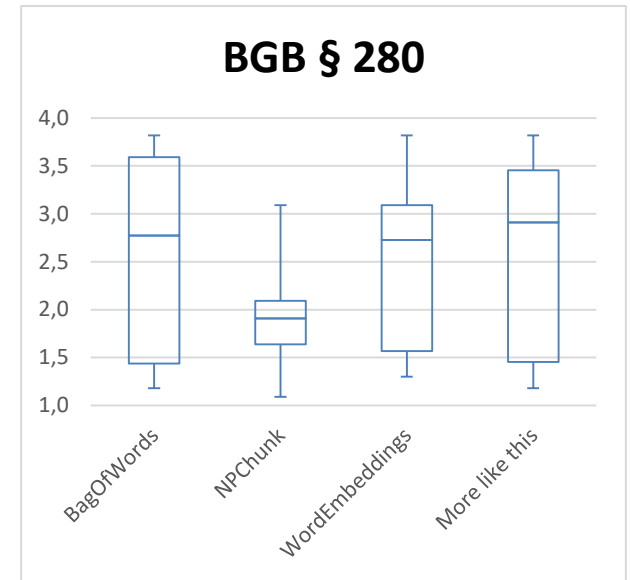


StGB § 242



Results

- Mix of relevant and irrelevant paragraphs
 - High interquartile range
 - Whiskers span almost whole scale
- Highly relevant paragraphs are missing



1. Motivation

2. Research Objectives

3. Approaches

4. Live Demo

5. Assessment

6. Conclusion & Outlook

▪ **Conclusion**

- Different similarity methods were implemented
 - Easy addition of further methods due to architecture
 - Computing similarities is time consuming
- No implemented method could outperform the others

▪ **Outlook**

- Further development of relatedness search in LEXIA
- Several subsequent research possibilities
 - Reduce methods' time consumption, e.g. using parallelization
 - Optimize search results of existing methods
 - Add further methods
 - Use more legal knowledge for similarity methods
- Results of assessment can be used as comparable figures

Thank you for your attention!

Any Questions?

- juris GmbH, "juris.de," [Online]. Available: <http://www.juris.de>. [Accessed 20 06 2016].
- U. Wesel and H. D. Beck, 250 Jahre rechtswissenschaftlicher Verlag C.H.Beck: 1763-2013, C.H.Beck, 2015.
- B. Walzl, F. Matthes, T. Walzl and T. Grass, "LEXIA - A Data Science Environment for Semantic Analysis of German Legal Texts," in *IRIS: Internationales Rechtsinformatik Symposium*, Salzburg, 2016.
- J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, 1977.
- T. Hill, P. Lewicki and P. Lewicki, *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*, StatSoft, Inc., 2006.
- G. Salton and M. McGill, *Introduction to modern information retrieval*, New York: McGraw-Hill, 1983.
- W. Stock, *Information Retrieval*, München: Oldenbourg, 2007.
- M. Levandowsky and D. Winter, "Distance between sets," *Nature*, pp. 34-35, 1971.
- G. Hinton, "Learning distributed representations of concepts," *Proceedings of the eighth annual conference of the cognitive science society*, 1986.
- Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A neural probabilistic language model," *journal of machine learning research*, pp. 1137-1155, Februar 2003.
- T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, pp. 3111-3119, 2013.

- J. S. Greenfield, Distributed programming paradigms with cryptography applications, Springer Science & Business Media, 1994.
- N. Chomsky, Syntactic structures, Walter de Gruyter, 2002.
- Elasticsearch, "More Like This Query | Elasticsearch," [Online]. Available: <https://www.elastic.co/guide/en/elasticsearch/reference/2.0/query-dsl-mlt-query.html>. [Accessed 06 11 2016]
- The Apache Software Foundation, "TFIDF Similarity (Lucene API)," [Online]. Available: https://lucene.apache.org/core/4_9_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html. [Accessed 07 11 2016].
- G. Salton, A. Wong and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, pp. 613-620, 1975.