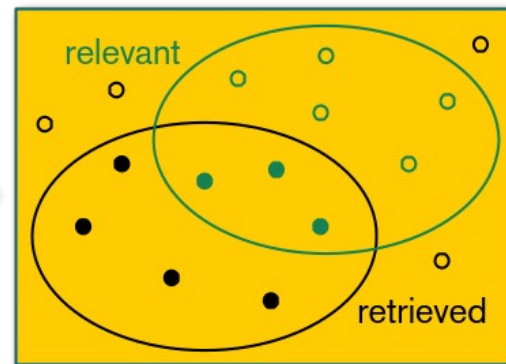


Vektorbasierte Erschließung von Bundesgerichtsurteilen

13.11.2024, Thomas Murphy

- 1) Einführung: Information Retrieval Prozess
- 2) Bundesgerichtsurteile
- 3) Versuchsaufbau: Kombination von Volltextsuche mit vektorbasierter Suche
- 4) Vektoren, Sprachmodelle, Vektordatenbanken
- 5) Evaluation
- 6) Schlussfolgerungen



1960er
Cyril Cleverdon



Cranfield experiments

1992
Donna K. Harman



TREC

2000

Text REtrieval Conference (TREC)

...to encourage research in information retrieval
from large text collections.



NTCIR

etc.



- **Verbalisierung des Informationsbedürfnisses:**
 - Voraussetzung: Verstehen des Informationsbedürfnisses: richtige Begriffe, Vollständigkeit
 - Paradox: muss Informationsbedürfnis verstanden haben, um es richtig zu verbalisieren
- **Codierung des verbalisierten Informationsbedürfnisses:**
 - Voraussetzung: Verstehen des Systems: Facetten, richtige Operatoren, etc.
 - Paradox: muss die Resultate kennen, um Anfrage richtig zu codieren
- Das System kann a priori nur die **explizite** Information, welche in der „codierten“ Anfrage enthalten ist, auswerten.
- Das Resultat bevorzugt also Entscheide, welche bestmöglich zu dieser „codierten Anfrage“ passen. Nicht explizit formulierte Präferenzen oder Hintergrundwissen des Benutzers können nicht in die Resultatfindung einfließen.

Darstellungsform	Vergleichbarkeit
Aufstellungssystematik	durch räumliche Nähe
Katalogisat mit Erschliessungsinformationen	mit exakten Datenbankabfragen
Volltext	mit Kosinus- und TF-IDF-basierten Abfragen
Einbettungsvektoren	mit Vektor-Datenbanken (Ähnlichkeit)
...	...



- Bundesgericht: Oberstes Gericht der Schweiz mit Sitz in Lausanne
- Sprachen: Deutsch, Französisch oder Italienisch
- 2 Kollektionen:
 - Urteile ab 2000
 - Leitentscheide ab 1954:
 - Inhaltserschliessung mit Normen und Deskriptoren (TDS)
 - Entscheide, die bedeutsam für die Weiterentwicklung der Rechtsprechung sind
- Struktur der Urteile (vereinfacht)
 - Urteilkopf
 - Sachverhalt
 - Erwägungen
 - Entscheid (Dispositiv)
- Sachverhalt, Erwägungen und Entscheid sind in Paragraphen / Abschnitte unterteilt

[Zurück zur Einstiegsseite](#)

BGE ab 1954 (publizierte Leitentscheide) und Urteile EGMR ab 1983

[Löschen](#) [Hilfe](#)

Suchen in:

- ☒ Bundesgericht
☐ Richterliche Behörden des Europarats
☐ allen Urteilen
[mehr...](#)

[Suchen](#)

Zwischen: 1954 und 2024

2 exakte Treffer gefunden für **abgebrochener Zahn**

1. **127 IV 236** ○○○○○
39. Auszug aus dem Urteil des Kassationshofes vom 24. August 2001 i.S. A. gegen B. (Nichtigkeitsbeschwerde)
Regeste [D, E, I] Art. 270 lit. g BSIP; Legitimation des Privattraftklägers zur eidgenössischen Nichtigkeitsbeschwerde. Nur der prinzipale Privattraftkläger ist zur Nichtigkeitsbeschwerde legitimiert (E. 2b/aa). Art. 270 lit. e Ziff. 1 BSIP; Legitimation des Opfers zur e...
2. **103 V 177** ○○○○○
40. Urteil vom 27. Dezember 1977 i.S. Michel gegen Eidgenössische Militärversicherung und Versicherungsgericht des Kantons Basel-Landschaft
Regeste [D, E, I] Art. 4 und 5 MVG. Zur Haftung der Militärversicherung für Zahnschäden (Zusammenfassung und Präzisierung der Rechtsprechung).

1 sinngemässer Treffer gefunden für **abgebrochener Zahn**

3. **114 V 169** ○○○○○
35. Urteil vom 15. August 1988 i.S. H. gegen Schweizerische Krankenkasse Helvetia und Versicherungsgericht des Kantons Zürich
Regeste [D, E, I] Art. 9 Abs. 1 UVV, Art. 14 Abs. 2 Vo III: Kausalzusammenhang beim unfallbedingten Zahnschaden. Die Adäquanz des Kausalzusammenhangs zwischen dem als Unfall zu qualifizierenden Abbrechen eines Zahnes beim Beissen auf eine Nusschale im Nussbrot und dem e...

11453 partielle Treffer gefunden für **abgebrochener Zahn**

4. **93 I 455** ○○○○○
57. Urteil vom 29. September 1967 i.S. Hoesen gegen eidg. Steuerverwaltung.
Regeste [D, E, I] Warenumsatzsteuer: Das Bestimmen von Brillengläsern und das Anpassen von Kontaktlinsen durch den Optiker ist verschieden von der Tätigkeit, die der Zahnarzt ausübt, wenn er die für den Patienten bestimmten Zahnprothesen bearbeitet. Es verstösst deshalb ...
5. **112 V 201** ○○○○○
36. Urteil vom 19. Juni 1986 i.S. Zürich Versicherungsgesellschaft gegen Koller und ...

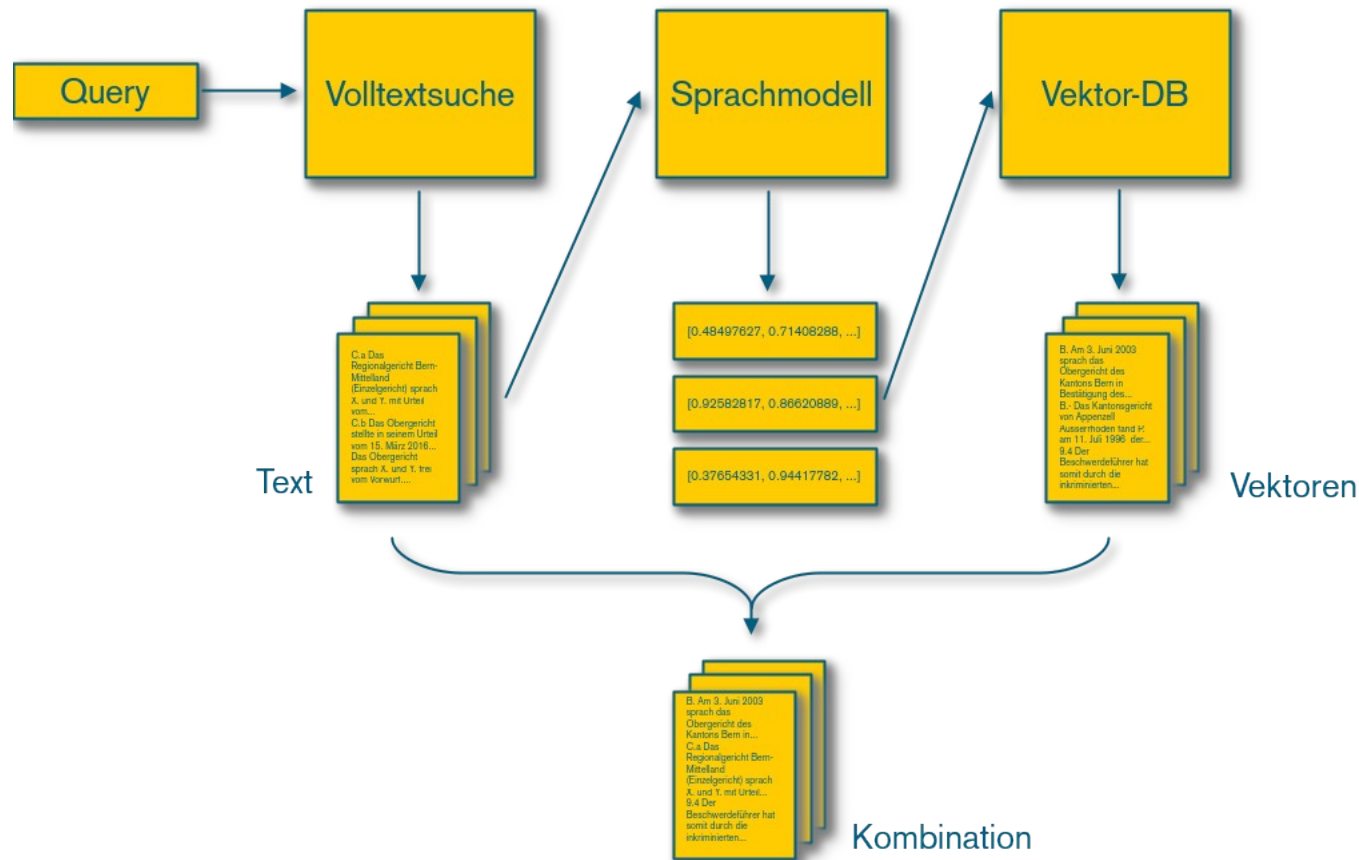


Suchtipp

Sie können direkt nach BGE Referenzen suchen, indem Sie die Referenz mit einem führenden 'BGE' eingeben. Beispiel: **BGE 129 III 31**.

© 2008 Eurospider Information Technology AG

- Basis: Datendump vom Dezember 2023
- Kollektion "Leitentscheide"
 - 13'810 Urteile (deutsch)
 - 202'712 Erwägungsabschnitte
- Kollektion "Weitere Urteile ab 2000"
 - 96'591 Urteile (deutsch)
 - 979'544 Erwägungsabschnitte





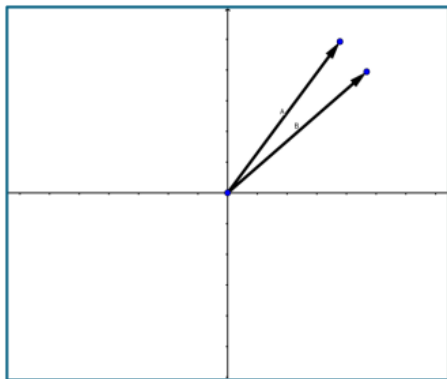
- Ein Textembedding ist ein Vektor, welcher eine bestimmte sprachliche Sequenz (Wort, Satz, Absatz, etc.) numerisch abbildet
- Vektoren sind eine Datenrepräsentation, mit denen ein Computer arbeiten kann
- Idee: Vektor so erstellen, dass er den Kontext und die Bedeutung der zugrunde liegenden Sequenz repräsentiert
- Generierung eines Vektors durch ein Sprachmodell
- Ein Sprachmodell modelliert die Abfolge von sprachlichen Sequenzen
- Ein Modell wird anhand Sprachdaten trainiert, heutzutage riesige Datenmengen (LLM, "large language model")

```
[0.9227, 0.8595, 0.6257, 0.6412, 0.1145, 0.6246, 0.1327, 0.4044, 0.6417, 0.5415, 0.7726, 0.1039,  
0.8558, 0.3519, 0.2290, 0.6679, 0.6720, 0.8095, 0.1226, 0.4887, 0.9439, 0.0308, 0.4969, 0.4534,  
0.4330, 0.2569, 0.6473, 0.0653, 0.4108, 0.2042, 0.6788, 0.7211, 0.3764, 0.5494, 0.8018, 0.5228,  
0.3182, 0.0227, 0.7964, 0.5031, 0.5176, 0.9413, 0.0388, 0.2114, 0.8818, 0.6722, 0.4239, 0.9606,  
0.1280, 0.4163, 0.6434, 0.4811, 0.1984, 0.7387, 0.7294, 0.0230, 0.1691, 0.3699, 0.7916, 0.6904,  
0.1290, 0.7514, 0.8234, 0.9500, 0.6390, 0.0438, 0.2377, 0.4646, 0.4923, 0.0460, 0.5760, 0.8383,  
0.3015, 0.9673, 0.9107, 0.4176, 0.3336, 0.6802, 0.6472, 0.6358, 0.2314, 0.6713, 0.3085, 0.4895,  
0.8968, 0.3050, 0.7804, 0.2378, 0.4344, 0.1610, 0.6881, 0.6189, 0.9249, 0.1128, 0.4408, 0.7194,  
0.8081, 0.4321, 0.3448, 0.0360, 0.9064, 0.7872, 0.8238, 0.4606, 0.8784, 0.1115, 0.3650, 0.4879,  
0.4603, 0.4811, 0.1417, 0.7602, 0.0620, 0.1139, 0.1961, 0.4663, 0.4988, 0.2565, 0.6976, 0.5650,  
0.6245, 0.7742, 0.3307, 0.9049, 0.0882, 0.6438, 0.2553, 0.3545]
```

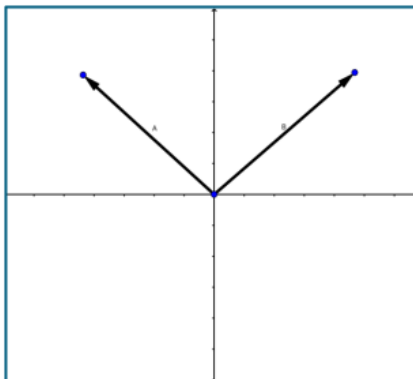


- Vektoren sind vergleichbar
- Verschiedene Ähnlichkeitsmasse: Kosinus-Ähnlichkeit, Skalarprodukt, Euklidischer Abstand

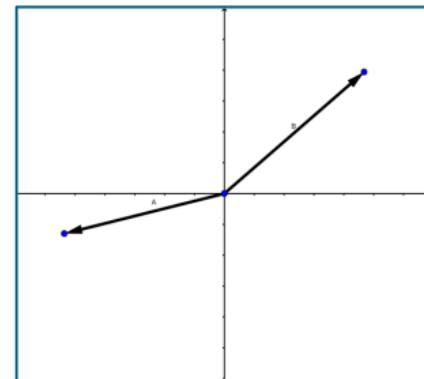
ähnlich ($\cos \approx 1$)



nicht verwandt ($\cos \approx 0$)



entgegengesetzt ($\cos < 0$)



- Vortrainiertes BERT-Modell + zusätzliche Pooling-Operation, um ein Embedding von fixer Länge zu erhalten
- Training u.a. auf klassifizierten Satzpaaren: contradiction, neutral, entailment
- 1 Mio. Satzpaare

A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

(Bowman et al. 2015)



- Lernen von qualitativ hochwertigen Embeddings anhand kontrastiven Satzpaaren
- Semi-strukturierte Daten als Trainingsmaterial
- Beispiele: Beitrag-Kommentar-Paare, Frage-Antwort-Paare, Name + Artikeltitel - Passage-Paare aus Wikipedia
- Anschliessend Fine-Tuning auf geringer Menge strukturierter Daten
- 270 Mio. Satzpaare

data source	type of text pairs	random example	# of pairs
Wikipedia	(entity+section title, passage)	q: Lexden History p: The site on which Lexden now stands was crossed by the fortifications of iron age Colchester. ...	24M
Reddit	(post, upvoted comment)	q: What makes a client good quality to you? p: I'm putting together my ideal client ... p: Respectful of schedules. And pays on time...	60M
Common Crawl	(title, passage)	q: Central Intake Unit Broome County p: Caseworkers from Central Intake assess the household and risk of placement. If eligible...	69M
Stackexchange	(title, answer) (title+description, answer)	q: Will killing Python made problems for Apache p: Python and Apache aren't related, unless your app is making use of Python. ...	19M
S2ORC	(title, abstract) (title, citation title) (abstract, citation abstract)	q: Constructive Dual DP for Reservoir Optimization p: Dynamic programming (DP) is a well established technique for optimization of reservoir manage...	90M
News	(title, passage) (highlight, passage)	q: LG Display reports Q1 operating loss as ... p: April 25 (Reuters) - South Korea's LG Display Co Ltd reported its first quarterly operating loss...	3M
Others	misc.	misc.	6M
All above	-	-	~ 270M

(Wang et al. 2022)



- Spezialisiertes System, um effizient mit hoch-dimensionalen Vektoren umzugehen
- Indexieren und Abfragen von Vektoren
- Operationen, welche herkömmliche Datenbanken nicht einfach auf Vektoren durchführen können
- Vektoren bilden eigenen Suchraum im Gegensatz / Ergänzung zu Volltext und Metadaten
- Herkömmliche Datenbanken nicht darauf ausgelegt, Vektoren effizient zu speichern und abzufragen.
- Ähnliche Vektoren suchen anhand Ähnlichkeitsmassen: Kosinus-Ähnlichkeit, Skalarprodukt, Euklidischer Abstand



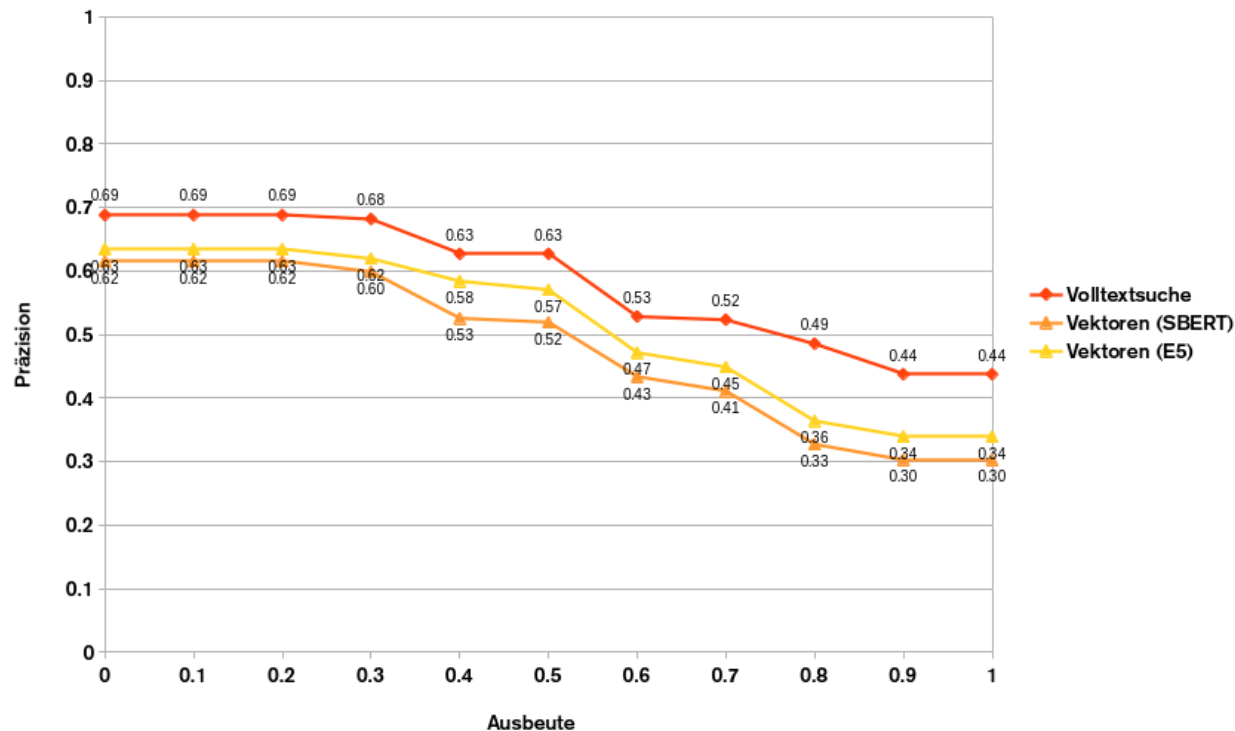


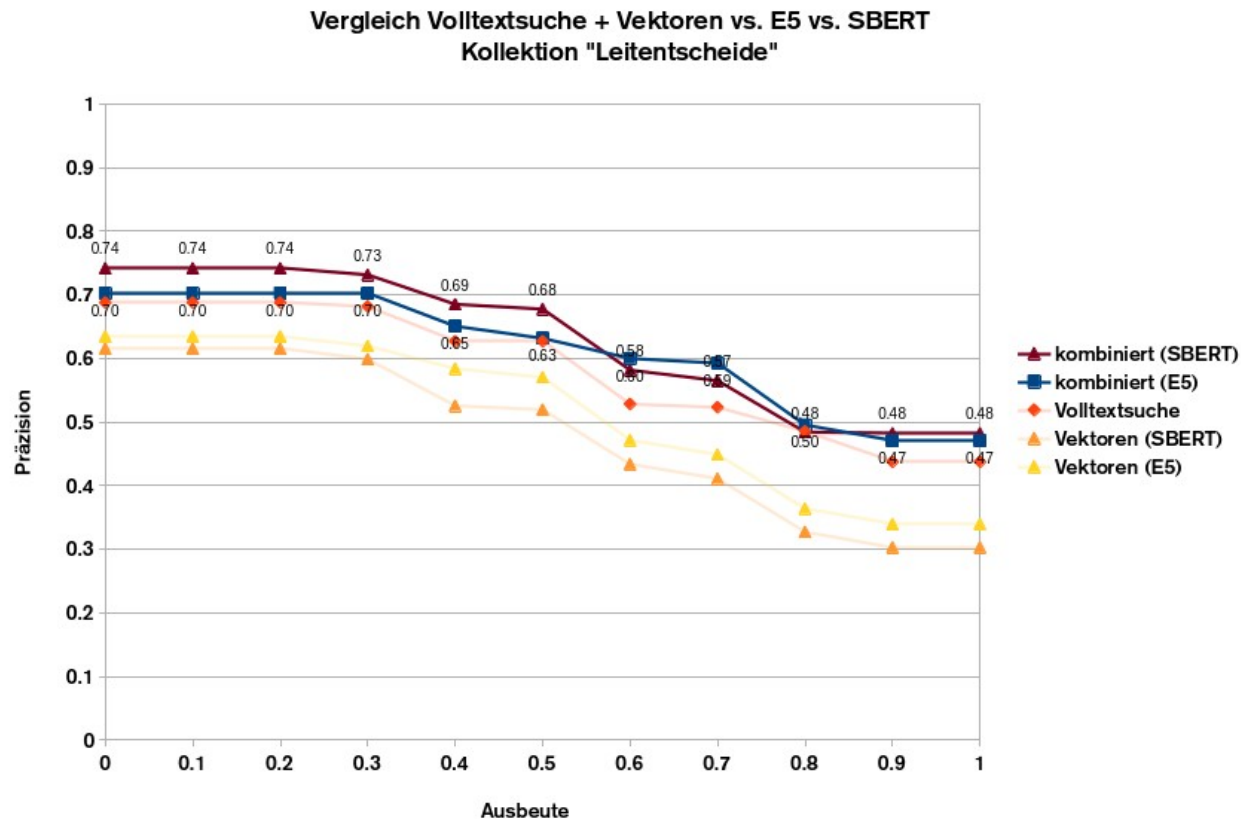
- Total 123 Queries
 - 38 Testqueries "Leitentscheide"
 - 85 Testqueries "Weitere Urteile ab 2000"
- Jede Testquery verfügt über:
 - Queryformulierung
 - Hintergrund: Informationsbedürfnis, Absicht der suchenden Person
 - relevante Urteile (1-n)
 - irrelevante Urteile (0-n)
 - Begründung der Relevanz / Irrelevanz

Suchanfrage	BGE mit Blutprobe Zwangsmassnahme
Typ	Leitentscheid
Hintergrund	Ist eine Blutprobe für einen Alkoholttest eine Zwangsmassnahme nach StPO?
Relevante Urteile	BGE 146 IV 88, BGE 143 IV 313
Begründung Relevanz	Im Bundesgerichtsentscheid wird die Blutprobe als Zwangsmassnahme angesehen.
Irrelevante Urteile	BGE 95 IV 144, BGE 121 II 134, BGE 94 IV 11, BGE 95 IV 121, BGE 103 IV 49
Begründung Irrelevanz	Im Bundesgerichtsentscheid ging es nur um Blutproben, nicht aber um Zwangsmassnahmen.



Vergleich Volltextsuche vs. E5 vs. SBERT
Kollektion "Leitentscheide"







- Text-basiertes Information Retrieval kann auf mehr als 30 Jahre Optimierung zurückblicken (TREC-1 1992).
- KI-Methoden mit Aufkommen von Transformers ("Attention is all you need", Vaswani et al. 2017) und LLMs in den letzten Jahren rasant an Bedeutung gewonnen
- KI-Methoden performen im Allgemeinen nicht per se
- Optimierung von KI-Methoden nicht trivial
- Erschliessung von Informationsobjekten mit Vektoren erfordert andere Vorgehensweisen und andere Systeme
- Rechtsinformationen wie beispielsweise Bundesgerichtsurteile sind keine typischen Texte (Arni 2010)



Arni, Thomas and Samir Abdou. "Cross-Language Information Retrieval in the Legal Domain." SIGIR 2010 Industry Track, 2010.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. "A large annotated corpus for learning natural language inference." In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics, 2015.

Peters, Carol, Martin Braschler, and Paul Clough. "Multilingual information retrieval: From research to practice." Heidelberg: Springer, 2012.

Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. "Attention is all you need." In Advances in Neural Information Processing Systems, 2017.

Wang, Liang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, Furu Wei. "Text embeddings by weakly-supervised contrastive pre-training." arXiv preprint arXiv:2212.03533, 2022 [12.11.2024].