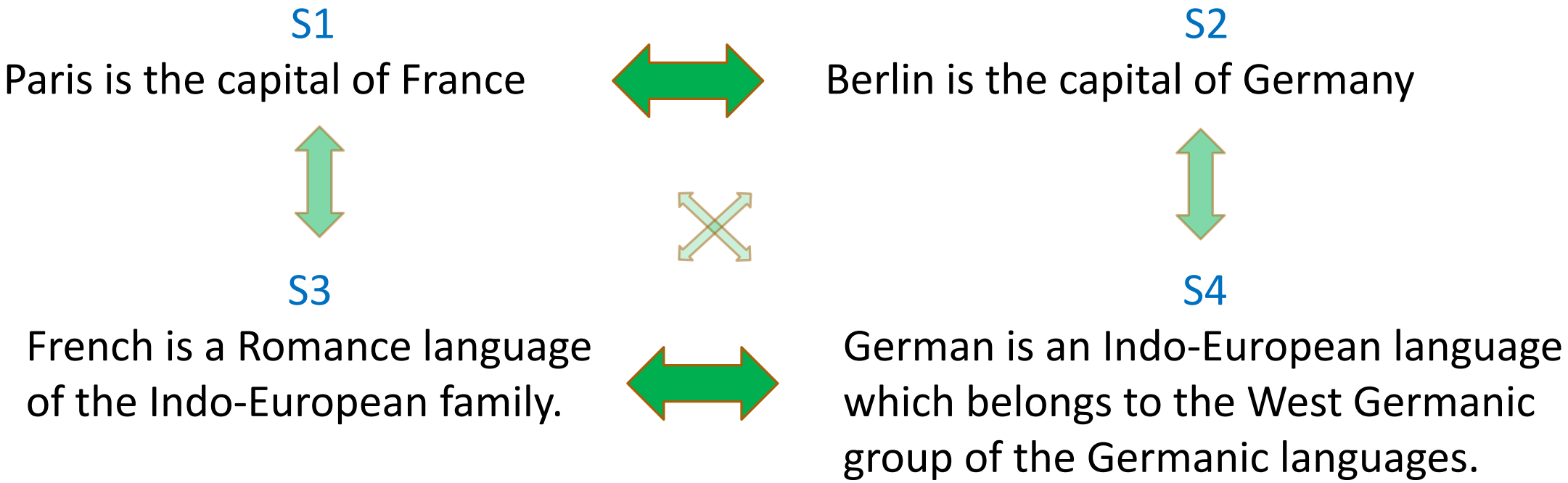
Semantische Ähnlichkeit

Nehmen wir an, wir haben einen Dataset mit Texten über Frankreich und Deutschland zu verschiedenen Themen (z. B. Hauptstadt, Wirtschaft, Küche, Demografie usw.). Wir möchten Informationen zu bestimmten Themen zusammenführen. Wie können wir dies erreichen und für viele weitere Länderkombinationen wiederholen?

Semantische Ähnlichkeit

Einfache Beispiele



Wir verwenden englische Sätze, weil wir die Embeddings für englische Wörter schon früher im Kurs gesehen haben

Semantische Ähnlichkeit

Von Wörtern zu Bedeutung und Kontext

Semantische Ähnlichkeit

Semantische Repräsentation eines Satzes

Wenn wir eine semantische Repräsentation von Sätzen finden könnten, könnten wir sie vergleichen und die Ähnlichkeit einstufen

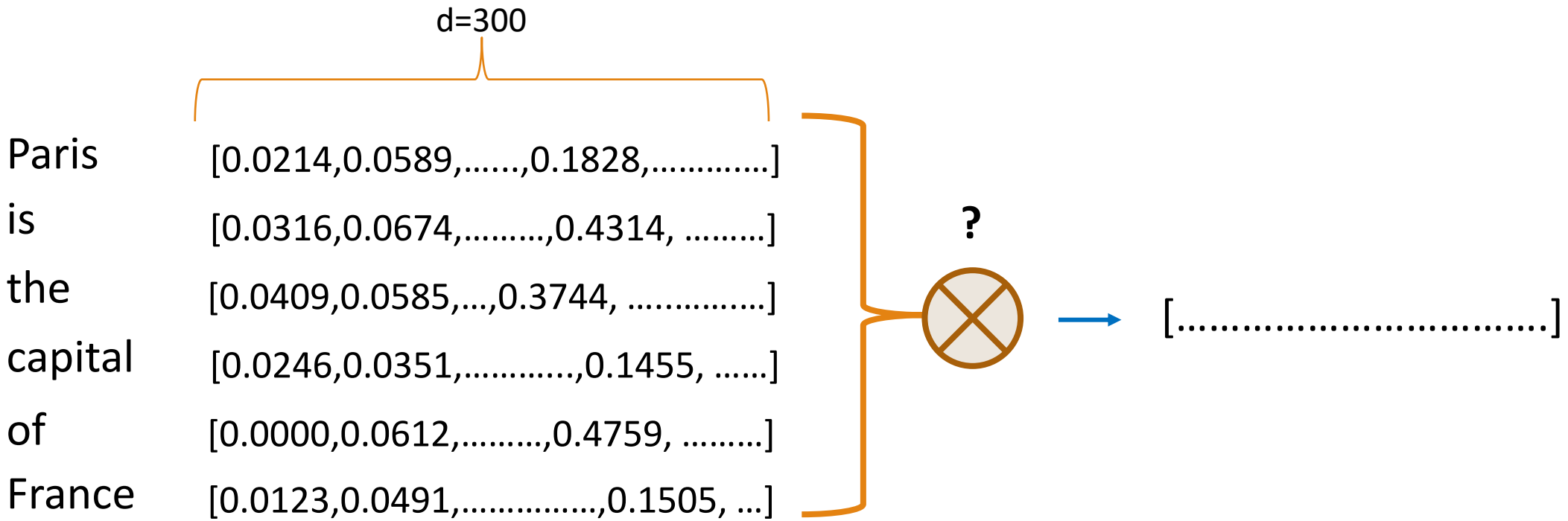
Können wir etwas verwenden, das wir bereits im Kurs gelernt haben?

Semantische Ähnlichkeit

Word-Embeddings -> Sentence-Embeddings

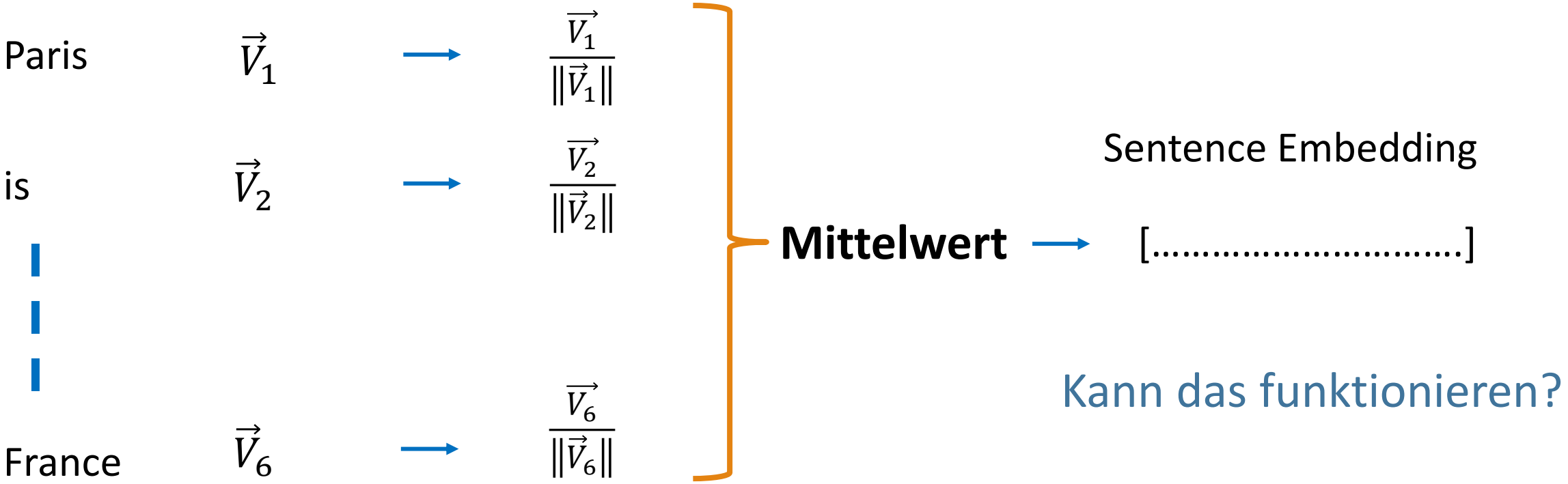
fastText
Library for efficient text classification and representation learning

https://fasttext.cc/docs/en/english-vectors.html



Semantische Ähnlichkeit

Word-Embeddings -> Sentence-Embeddings

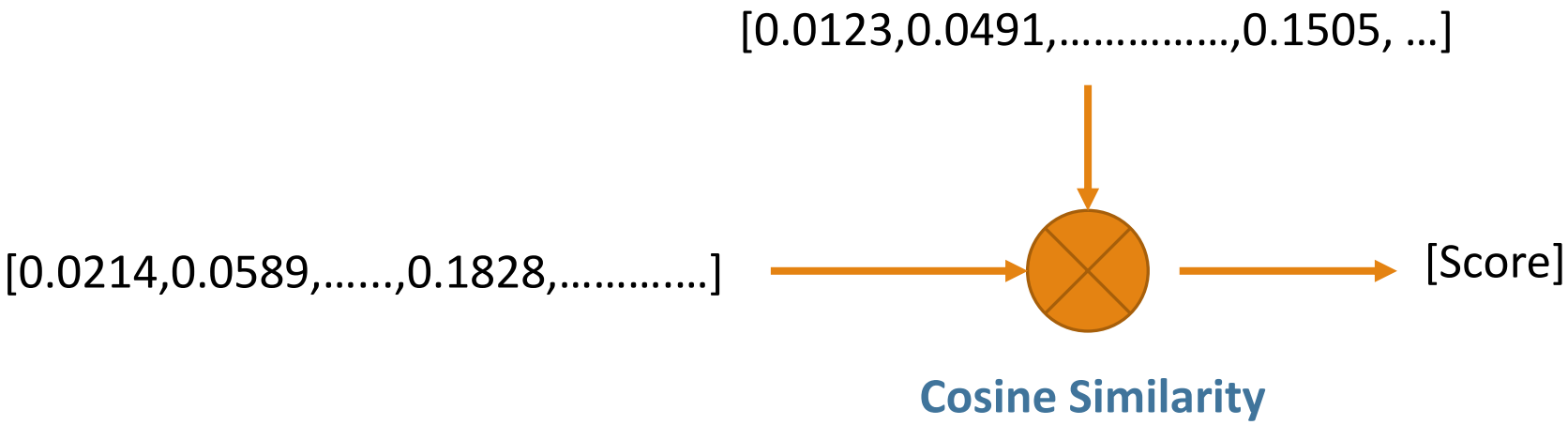


Kann das funktionieren?

Semantische Ähnlichkeit

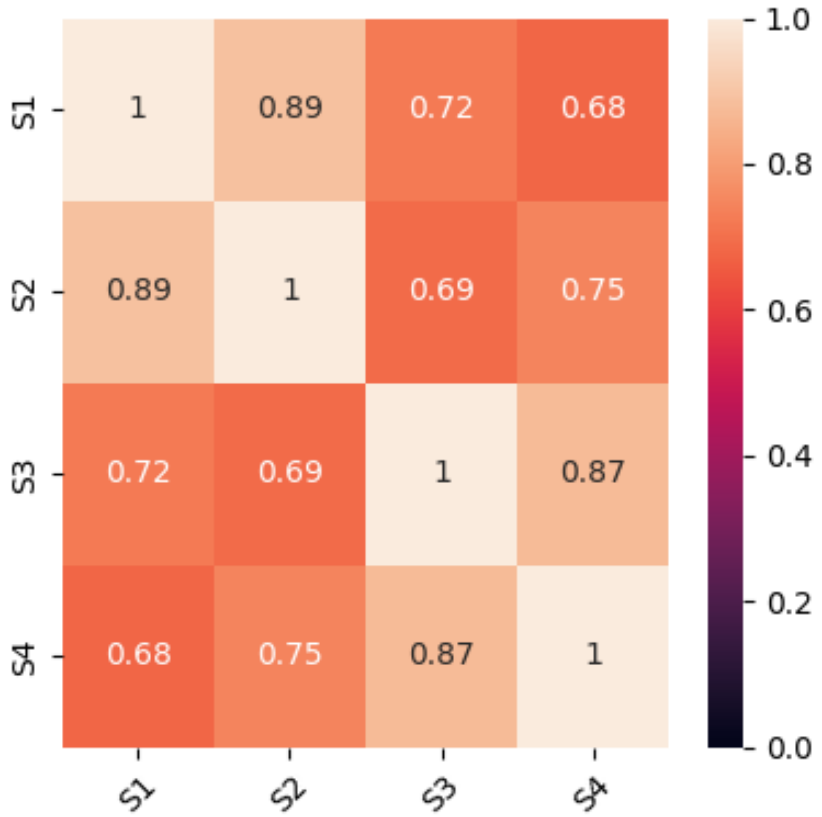
Ähnlichkeit

Wiederholung



Semantische Ähnlichkeit

Ist das gut genug?



Jupyter Notebook:
https://github.com/saurabhkumar/lecture1_semantic_similarity/blob/main/embeddings_and_semantics.ipynb

Semantische Ähnlichkeit

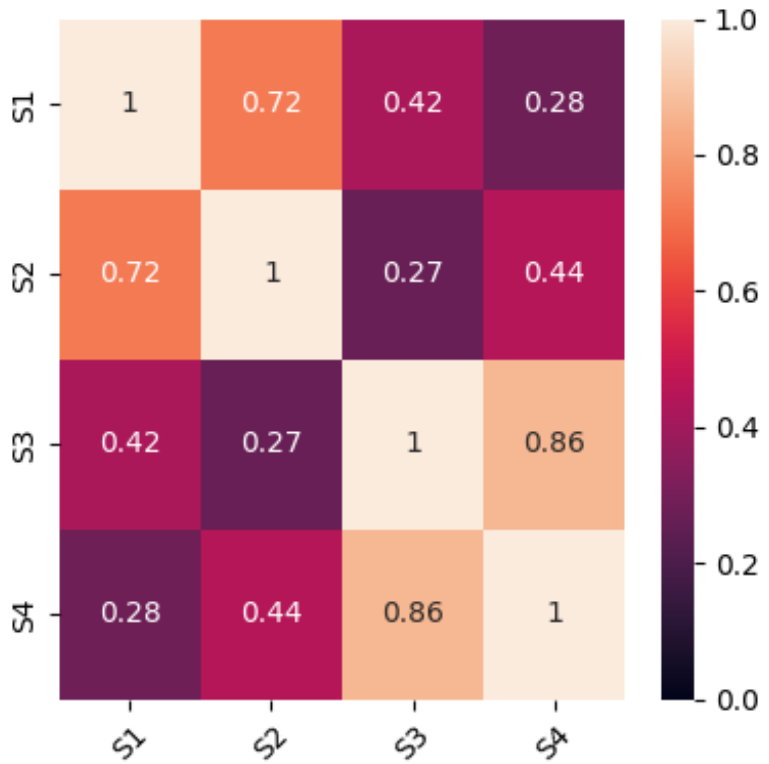
Was wäre, wenn wir nur die wichtigen Wörter verwenden könnten?

S1 Paris capital France

S2 Berlin capital Germany

S3 French language

S4 German language



Dieser Artikel zeigt, warum diese Methode sehr wirksam sein kann: Arora, S, Liang, Y & Ma, T 2019, 'A simple but tough-to-beat baseline for sentence embeddings', Paper presented at 5th International Conference on Learning Representations, ICLR 2017

Semantische Ähnlichkeit

Aber ….

Methoden zur Auswahl nur "wichtiger" Wörter oder zur "unterschiedlichen Gewichtung" von Wörtern für ungesehene bereichsspezifische Texte sind oft schwierig.

Semantische Ähnlichkeit

Die Hoffnung ….

Später in diesem Kurs werden wir Methoden wie „Attention", "Transformers" und "BERT" begegnen.

Wir werden versuchen, einen ‚*Sneak-Peek*‘ darauf zu bekommen, was möglich ist, wenn eine Methode diese als Grundlage für die Lösung der semantischen Ähnlichkeitsaufgabe verwendet

Semantische Ähnlichkeit

Zwei Methoden

Universal Sentence Encoder

Daniel Cer[a], Yinfei Yang[a], Sheng-yi Kong[a], Nan Hua[a], Nicole Limtiaco[b], Rhomni St. John[a], Noah Constant[a], Mario Guajardo-Céspedes[a], Steve Yuan[c], Chris Tar[a], Yun-Hsuan Sung[a], Brian Strope[a], Ray Kurzweil[a]

[a]Google Research
Mountain View, CA

[b]Google Research
New York, NY

[c]Google
Cambridge, MA

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

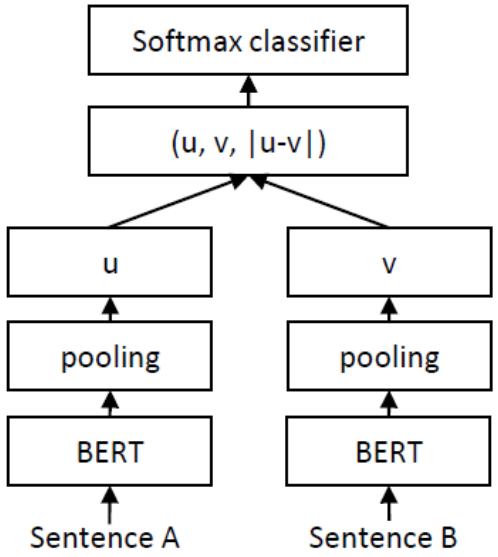
Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Semantische Ähnlichkeit

Sentence-BERT

[Sentence-BERT paper]

[Website mit Erklärungen und Link zu Modellen]



Semantische Ähnlichkeit

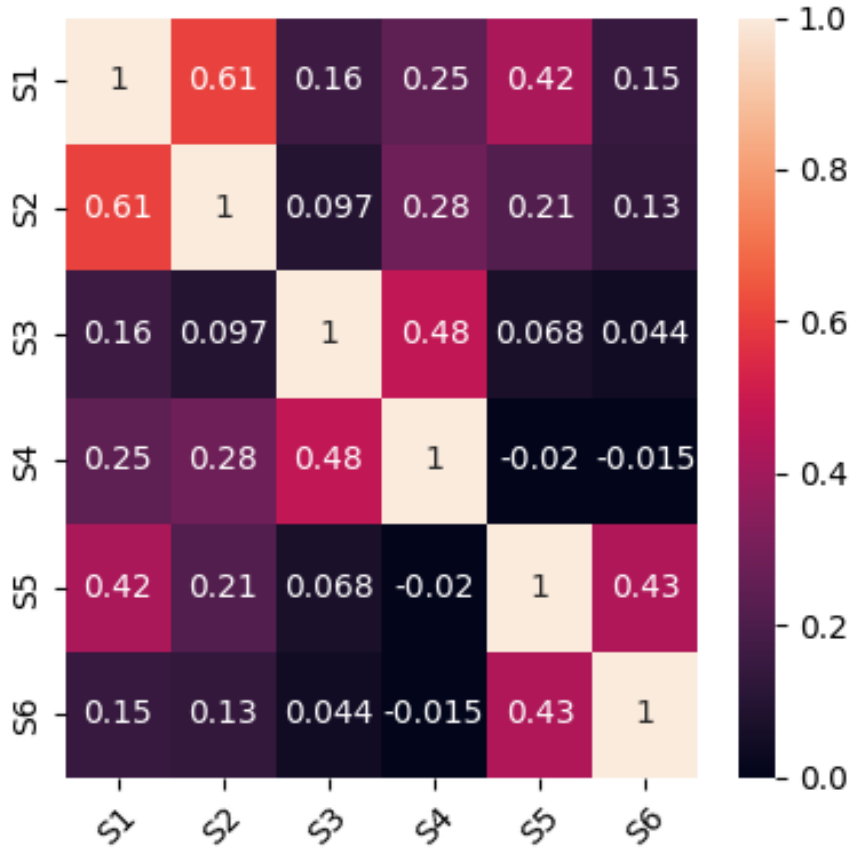
Wie die realen Daten aussehen könnten

Wichtiger Hinweis: Diese Texte wurden aus Wikipedia entnommen

S1	France has a developed high-income mixed economy characterised by sizeable government involvement, economic diversity, a skilled labour force, and high innovation. For roughly two centuries the French economy has consistently ranked among the ten largest globally
S2	French economy is the world's seventh-largest economy by nominal GDP
S3	Germany is a federal, parliamentary, representative democratic republic. Federal legislative power is vested in the parliament consisting of the Bundestag (Federal Diet) and Bundesrat (Federal Council), which together form the legislative body
S4	With a population of 80.2 million according to the 2011 German Census, rising to 83.7 million as of 2022, Germany is the most populous country in the European Union, the second-most populous country in Europe after Russia, and the nineteenth-most populous country in the world
S5	Each region of France has iconic traditional specialties: cassoulet in the Southwest, choucroute in Alsace, quiche in the Lorraine region, beef bourguignon in Burgundy, provençal tapenade, etc
S6	A typical French Christmas dish is turkey with chestnuts

Semantische Ähnlichkeit

Ergebnisse



Semantische Ähnlichkeit

Können wir es verbessern?

Wenn wir mehr Anpassung an unseren Kontext wünschen, können wir ***Finetuning*** verwenden

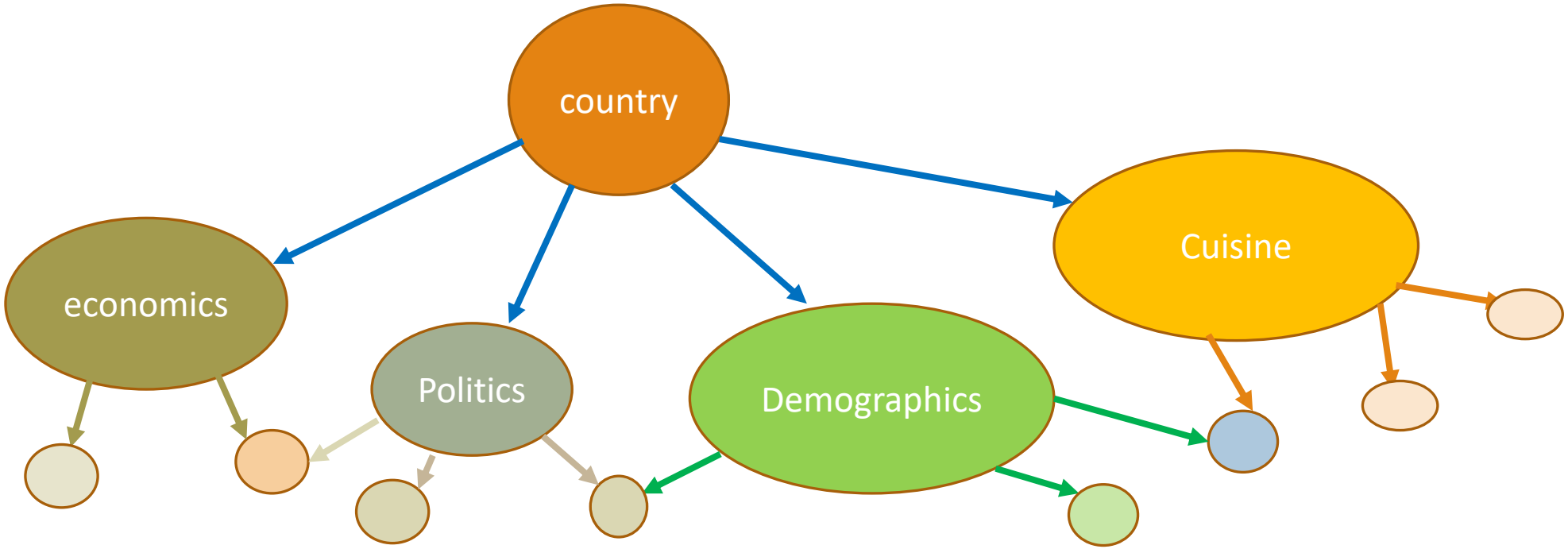
Wir werden hier nicht auf die Details eingehen, aber ein Beispiel ist im Jupyter Notebook für diese Präsentation verfügbar.

So pay ***ATTENTION*** to the **next lectures** from Prof. Zarcone ☺

Semantische Ähnlichkeit

Was sonst noch möglich ist

Wenn wir einem solchen Graph automatisch Elemente hinzufügen wollten



Semantische Ähnlichkeit

Was sonst noch möglich ist

Wenn wir Sätze in verschiedenen Sprachen haben

Multilingual Universal Sentence Encoder for Semantic Retrieval

Yinfei Yang[a†], Daniel Cer[a†], Amin Ahmad[a], Mandy Guo[a], Jax Law[a], Noah Constant[a], Gustavo Hernandez Abrego[a], Steve Yuan[b], Chris Tar[a], Yun-Hsuan Sung[a], Brian Strope[a], Ray Kurzweil[a]

[a]Google AI
Mountain View, CA

[c]Google
Cambridge, MA

Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de