MASTERARBEIT

Kontextbezogenes Entity Linking auf Dokumentenebene mit Deep Learning

Verfasser: Denis Martin
Betreuer: Prof. Dr. habil. Alexander Löser
Gutachter: Prof. Christoph Knabe

Übersicht

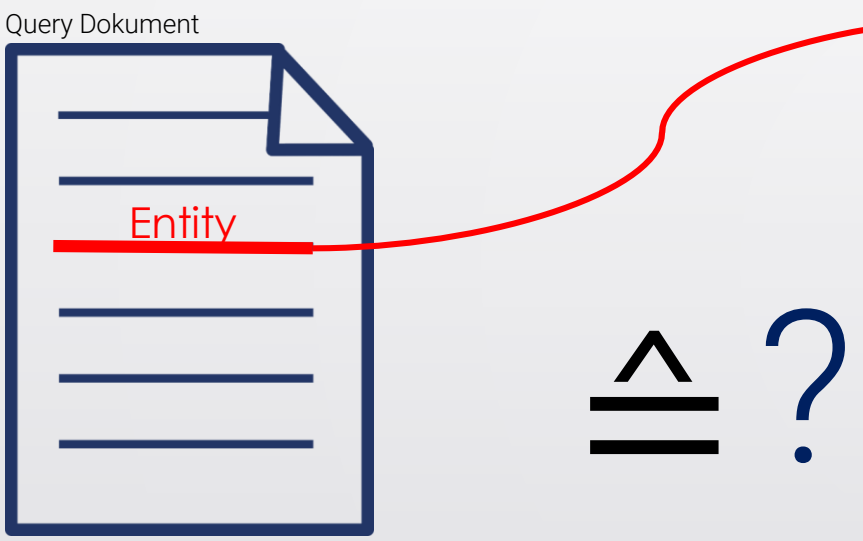
Übersicht

1. Problemstellung
2. Ziel der Arbeit
3. Daten
4. Umsetzung
5. Evaluierung
6. Ergebnisse
7. Ausblick

Problemstellung

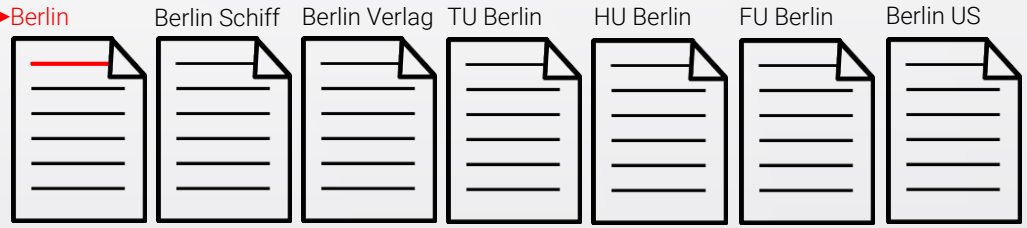
Problemstellung

Entity Linking auf Dokumentenebene



Berlin ist eine wunderschöne Hauptstadt an der Spree.

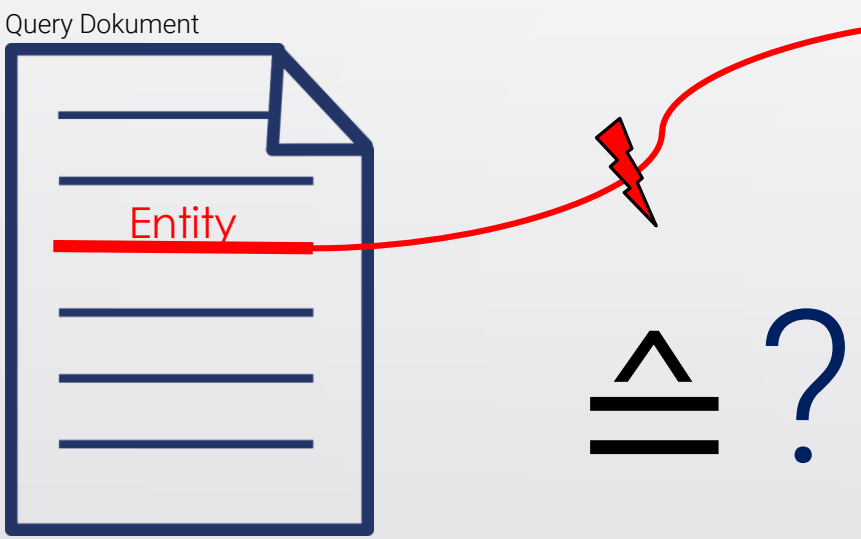
≙ ?



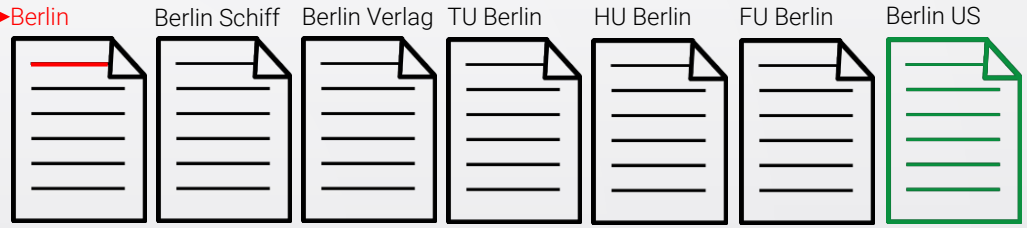
Berlin ist die Hauptstadt der Bundesrepublik Deutschland. Berlin ist geprägt durch viele Fließgewässer und Seen. Im Bezirk Spandau mündet die Spree in die Havel, die den Westen Berlins in Nord-Süd-Richtung durchfließt. Berliner Nebenflüsse der Spree sind die Panke, die Dahme, die Wuhle und die Erpe.

Problemstellung

Entity Linking auf Dokumentenebene



Berlin liegt im Bundesstaat Vermont in der Nähe von Tashmore und dem Tashmore See.



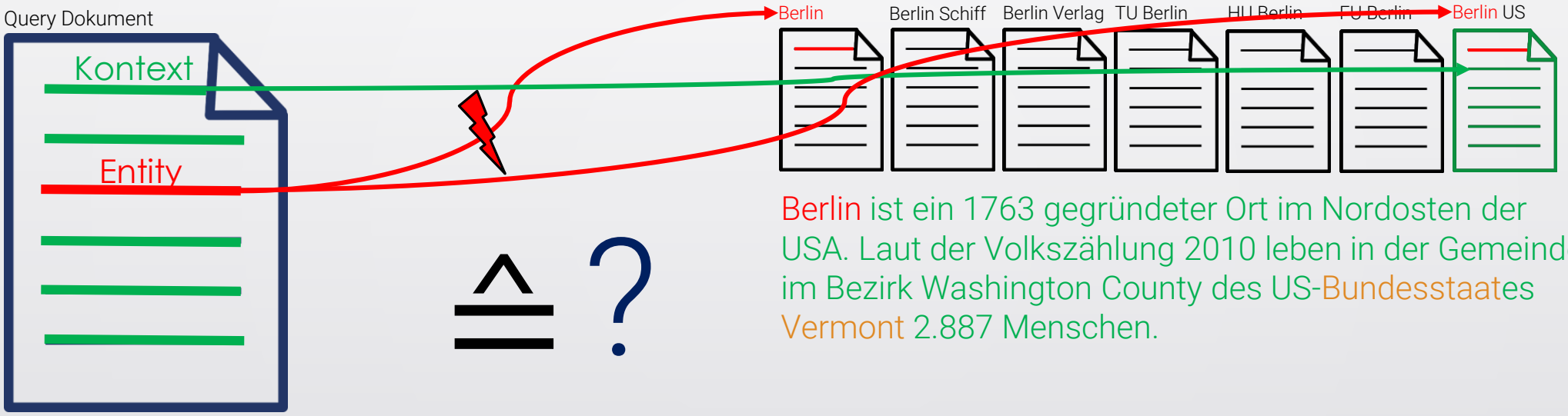
Berlin ist die Hauptstadt der Bundesrepublik Deutschland. Berlin ist geprägt durch viele Fließgewässer und Seen. Im Bezirk Spandau mündet die Spree in die Havel, die den Westen Berlins in Nord-Süd-Richtung durchfließt. Berliner Nebenflüsse der Spree sind die Panke, die Dahme, die Wuhle und die Erpe.

Ein Wort wird durch die Wörter beschrieben, die es begleiten

Zellig Sabbetai Harris (1954 Distributional Structure)

Problemstellung

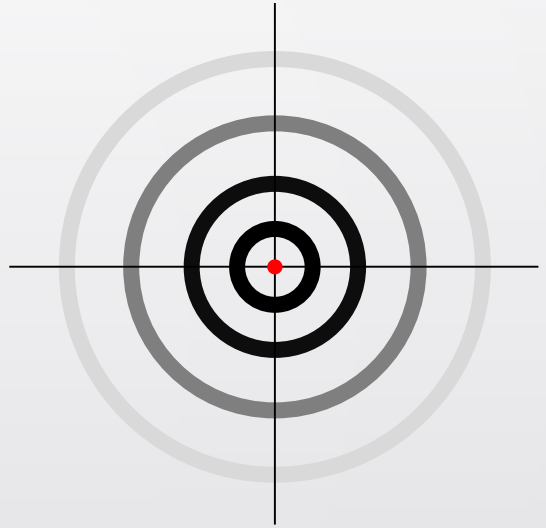
Kontextbezogenes Entity Linking auf Dokumentenebene



Berlin ist ein 1763 gegründeter Ort im Nordosten der USA. Laut der Volkszählung 2010 leben in der Gemeinde im Bezirk Washington County des US-Bundesstaates Vermont 2.887 Menschen.

Berlin liegt im Bundesstaat Vermont in der Nähe von Tashmore und dem Tashmore See.

Ziel der Arbeit



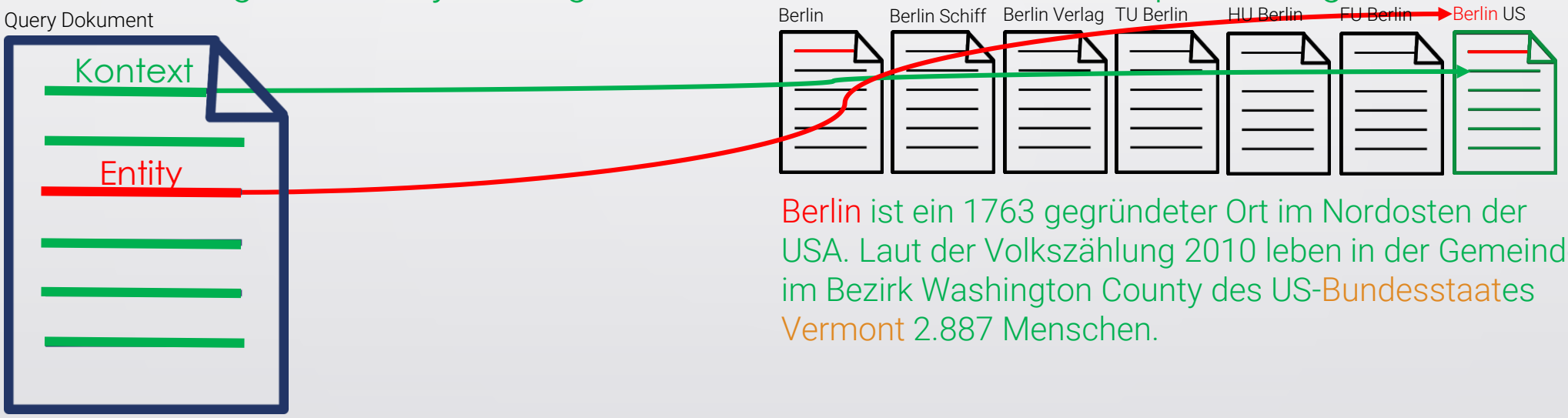
Ziel der Arbeit

Das Ziel dieser Arbeit ist die Zuordnung von einem nicht eindeutigen Begriff zu einem Wikipedia-Artikel, der diesen Begriff beschreibt und ihm dadurch die Eindeutigkeit verleiht, herzustellen.

Die zentrale Frage

Die zentrale Frage, die dabei gestellt wird, kann das neuronale Netz so gut generalisieren, dass es aufgrund von Kontext im Dokument eine korrekte Entscheidung treffen kann.

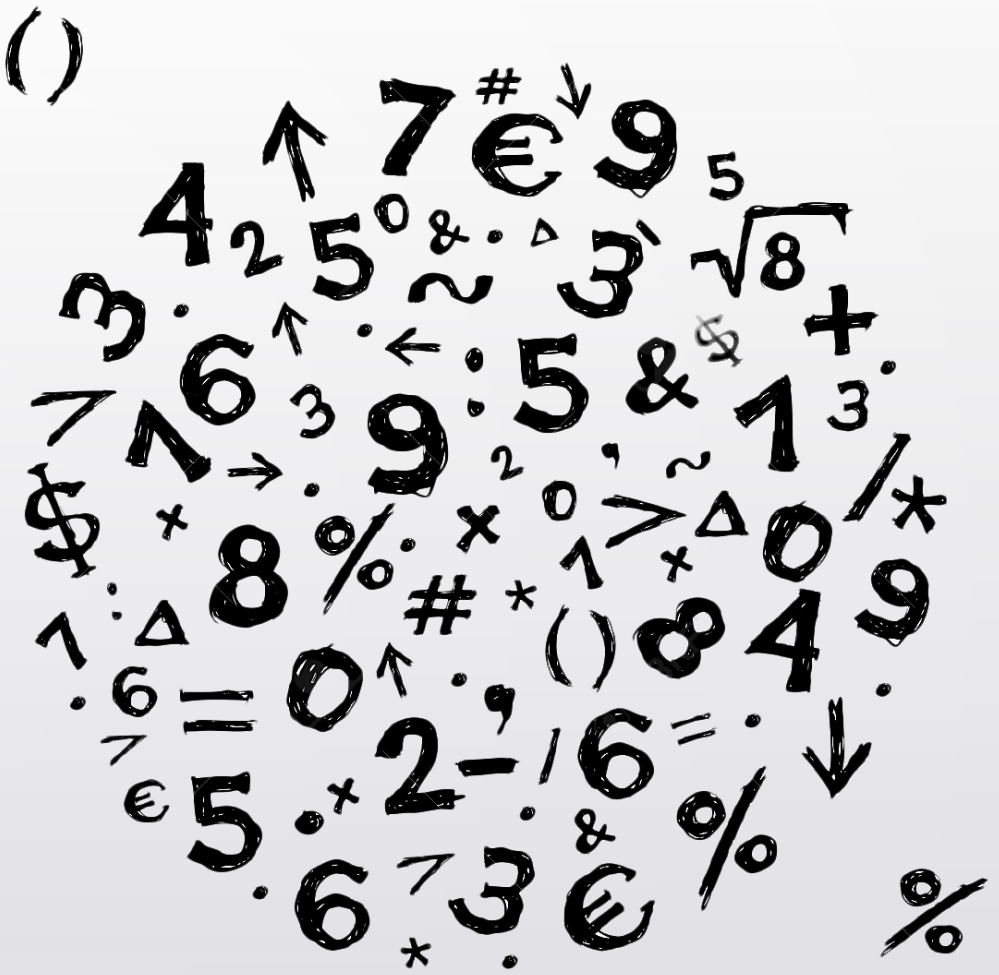
Kontextbezogenes Entity Linking auf Dokumentenebene mit Deep Learning



Berlin ist ein 1763 gegründeter Ort im Nordosten der USA. Laut der Volkszählung 2010 leben in der Gemeinde im Bezirk Washington County des US-Bundesstaates Vermont 2.887 Menschen.

Berlin liegt im Bundesstaat Vermont in der Nähe von Tashmore und dem Tashmore See.

Daten



Deutsche Wikipedia mit 1.556.343 Artikeln

1.556.343 Wikipedia-Artikel mit 15.859.142 Links

Knowledge Base

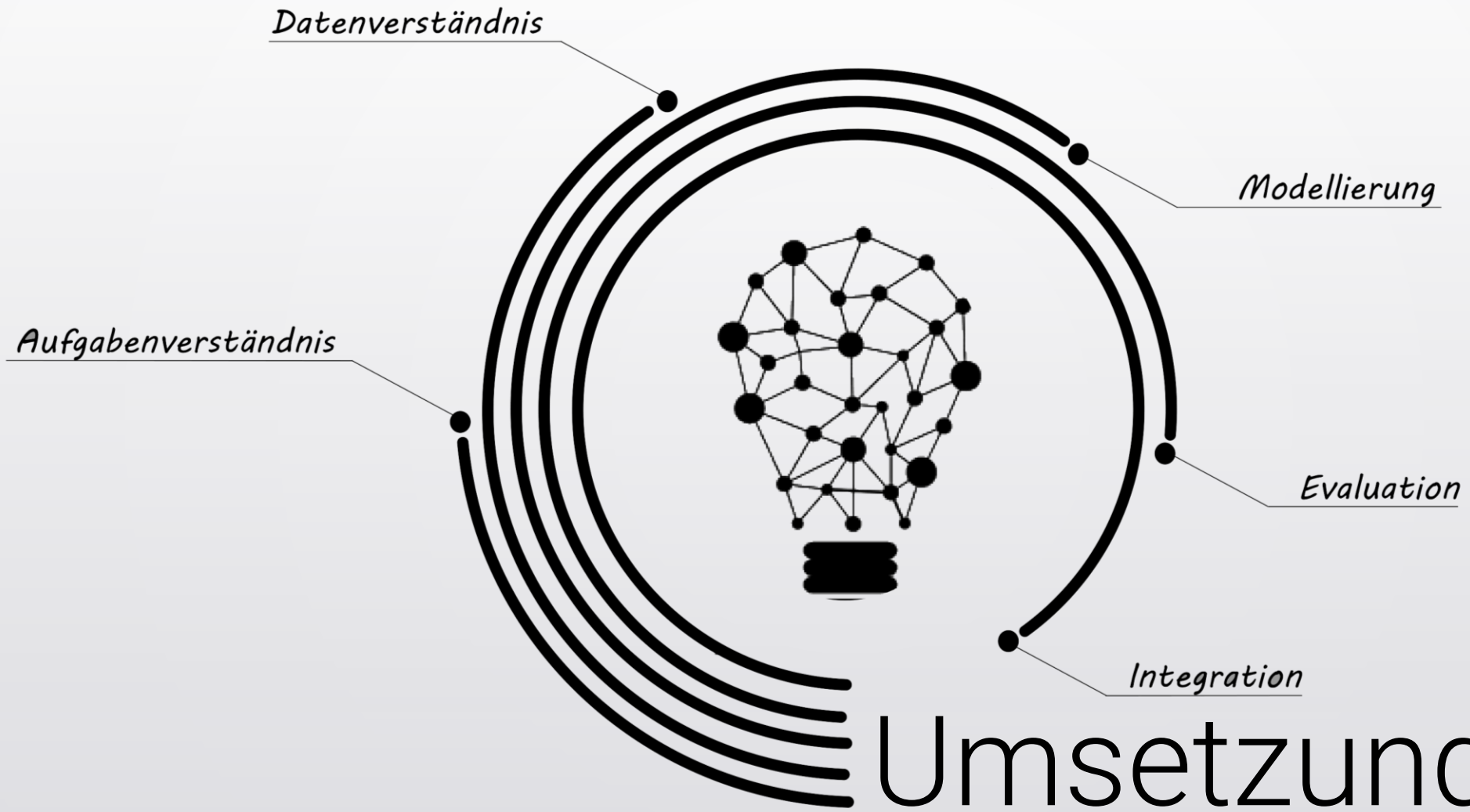
Alle notwendigen Daten werden jeweils in einem Wikipedia-Artikelobjekt zusammengefasst und als JSON-Objekt in einer Knowledge Base Liste gespeichert


```
{
  "class" : "de.datexis.model.article.WikiArticle",
  "id" : "2552494",
  "url" : "http://de.wikipedia.org/wiki/Berlin",
  "title" : "Berlin",
  "text" : "Berlin ( ) ist die Bundeshauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. ...",
  "names" : [ "DE-BE", "Tourismus in Berlin", "Land Berlin", "Berlin"],
  "terms" : [ "BE", "Berlin-Ost", "Berlin-Spandau", "Länder Berlin", "Berlin-Mitte", "Ostberlin", "Hauptstadt" ],
  "dbpediaURL" : "http://de.dbpedia.org/resource/Berlin",
  "popularity" : 31025,
  "score" : 0.0,
  "length" : 1681
}
```


InputData

WikiArticle -> Document

Aus jedem Wikipedia-Artikelobjekt wird jeweils ein Dokument erstellt. Dabei wird der Text auf Sätze und einzelne Wörter zerteilt.


Datenverständnis
Modellierung
Aufgabenverständnis
Evaluation
Integration

Umsetzung

Word2Vec

Berlin (word2vec 150 Dimension, Rundung auf zwei Nachkommastellen)

[-0.10, -0.47, -0.17, -0.03, -0.01, 0.19, -0.16, 0.08, -0.09, -0.35, -0.10, -0.31, -0.34, 0.04, -0.38, 0.34, 0.28, -0.16, 0.42, -0.00, -0.04, 0.42, -0.11, 0.17, 0.27, -0.08, 0.19, 0.12, -0.02, 0.04, 0.01, -0.03, 0.08, -0.05, 0.02, -0.24, 0.05, 0.18, -0.08, 0.06, -0.13, 0.20, -0.32, 0.12, -0.10, -0.33, 0.11, -0.09, 0.03, -0.45, -0.12, -0.38, 0.07, -0.11, 0.37, 0.07, -0.08, 0.56, 0.06, -0.01, 0.25, -0.09, -0.04, 0.21, -0.00, 0.12, -0.11, -0.09, 0.16, 0.12, 0.07, -0.28, 0.10, -0.01, 0.01, 0.12, 0.09, -0.02, 0.53, 0.01, -0.18, 0.20, -0.09, -0.10, 0.56, 0.04, -0.29, 0.23, 0.20, 0.05, 0.01, -0.12, 0.11, -0.52, 0.17, -0.25, 0.39, -0.23, 0.21, -0.04, 0.25, 0.10, -0.02, -0.14, -0.10, -0.10, 0.19, -0.31, 0.01, 0.07, 0.07, -0.16, -0.02, 0.07, -0.02, -0.05, -0.00, 0.42, -0.32, 0.27, 0.27, -0.01, 0.13, -0.16, 0.41, -0.06, 0.16, 0.39, -0.15, 0.04, 0.17, 0.29, 0.19, -0.27, 0.29, 0.02, -0.15, 0.03, 0.10, 0.10, -0.07, 0.10, -0.28, -0.18, 0.10, -0.27, -0.03, 0.05, 0.22, -0.20]

Berlin (word2vec 150 Dimension, ohne Rundung)
[-0.1001083, -0.4657691, -0.17055179, -0.028980851, -0.014070801, 0.19091238, …]

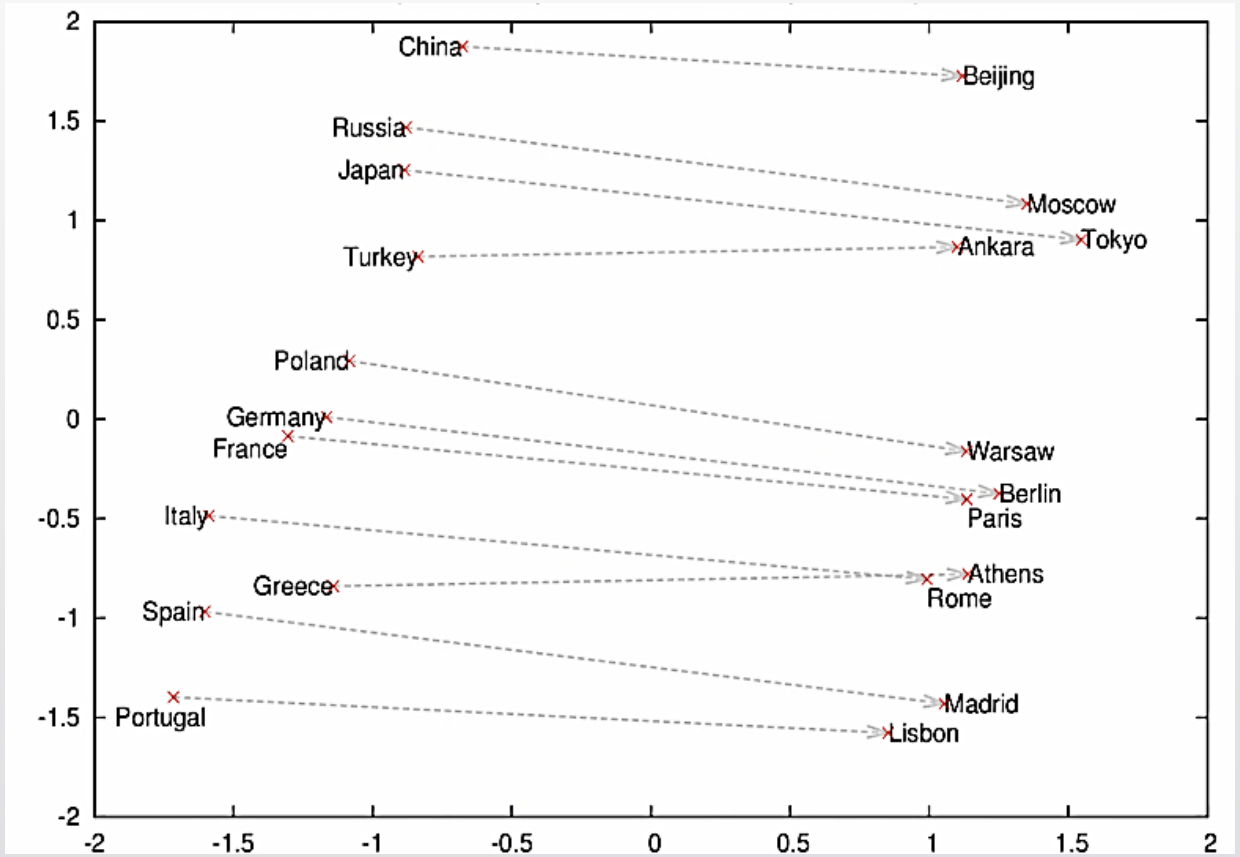
Word2Vec

Berlin = [-0.10, -0.47, -0.17, -0.03, -0.01, 0.19, -0.16, 0.08, -0.09, -0.35, -0.10, -0.31, -0.34, 0.04, -0.38, 0.34, 0.28, -0.16, 0.42, -0.00, -0.04, 0.42, -0.11, 0.17, 0.27, -0.08, 0.19, …]

Wörter wie Rom, Paris, Berlin und Beijing alle diese Hauptstädte befinden sich in der Nähe zueinander, aber sie haben auch jeweils ähnliche Abstände im Vektorraum zu den Ländern, deren Hauptstädte sie sind.

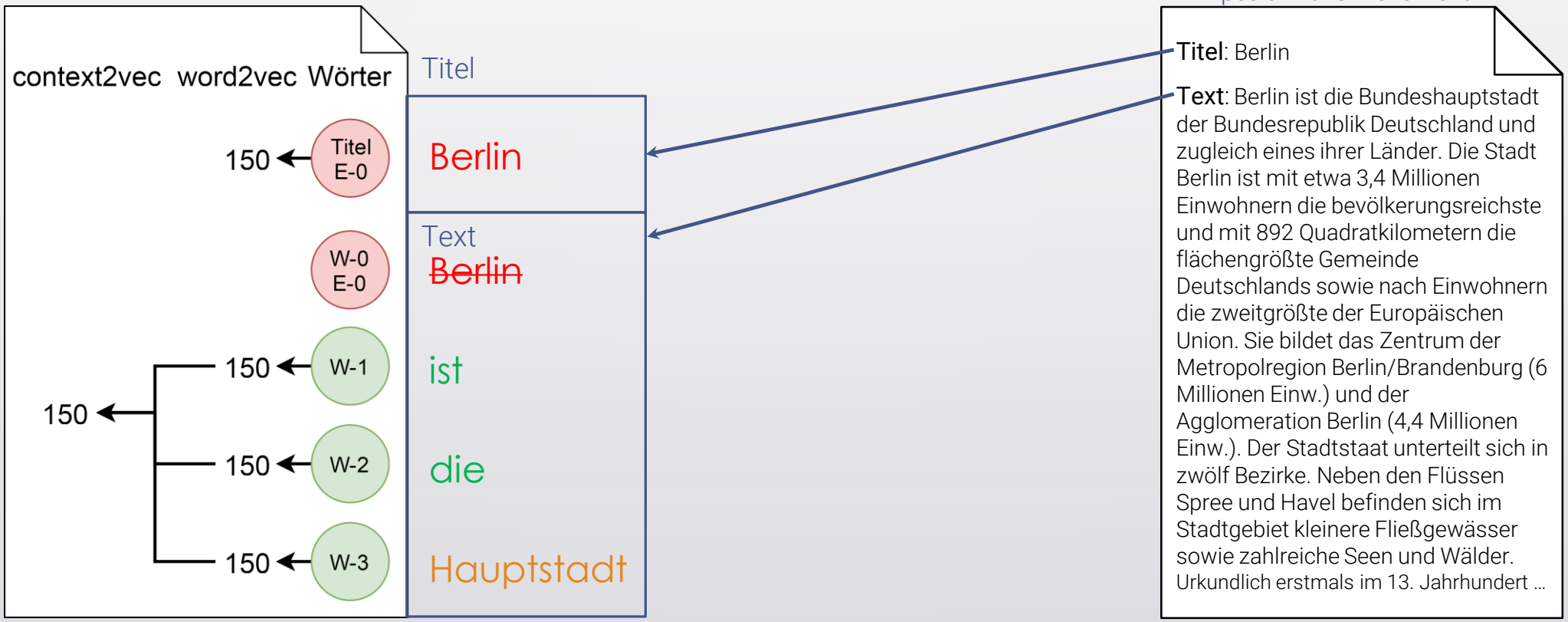
Rom - Italien = Berlin - Deutschland

Rom - Italien + Berlin = Deutschland



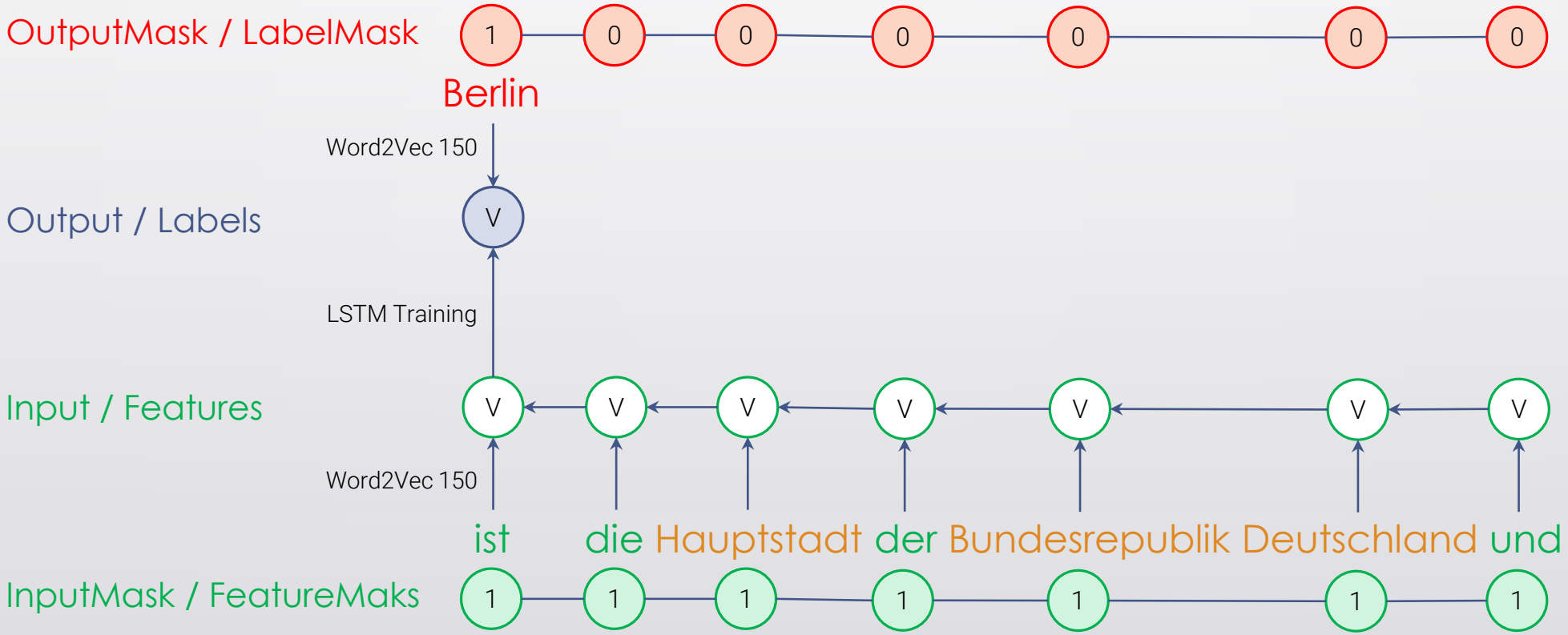
Context2Vec Training mit Wikipedia- Artikeln

Berlin ist die Hauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. …



Training mit Long short-term memory (LSTM)

Berlin ist die Hauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. …



Context2Vec Model

Trainingssettings

gesamte Trainingszeit: 5.408 Minuten (90,13 Stunden oder 3,756 Tage)

trainingSetSize: 200000

learningRate: 0.002

lstmLayerSize: 500

numEpochs: 3

iterations: 1

batchSize: 4

Context2Vec

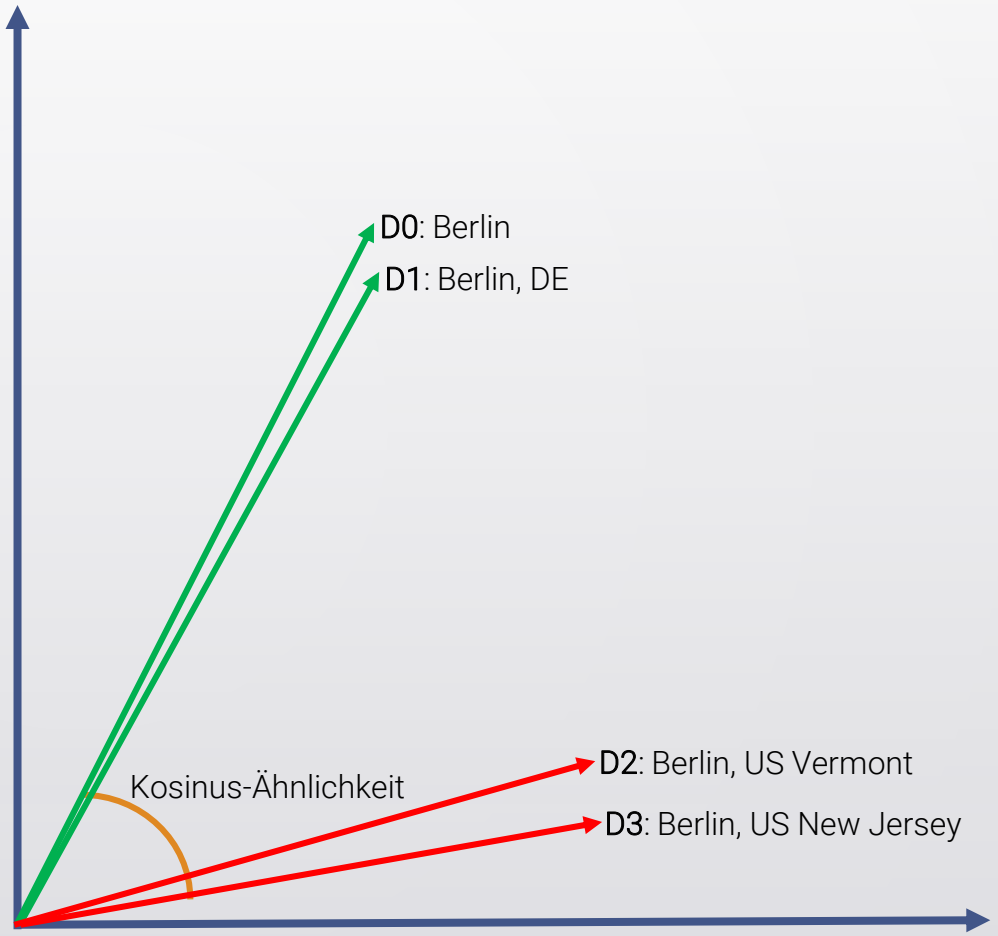
D0: Berlin ist eine wunderschöne Hauptstadt an der Spree.

D1: Berlin ist die Hauptstadt der Bundesrepublik Deutschland.

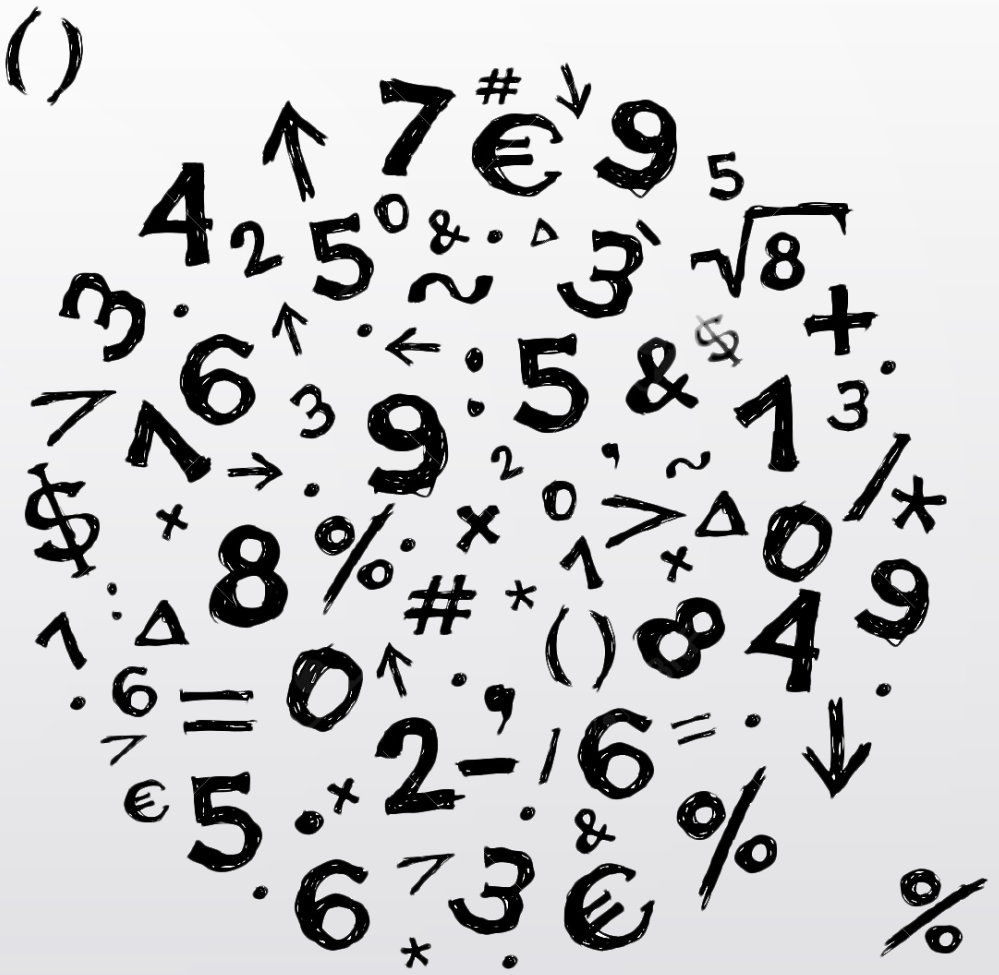
D2: Berlin ist ein 1763 gegründeter Ort im Nordosten der USA.

D3: Berlin ist ein Bezirk in Camden County, New Jersey, USA.

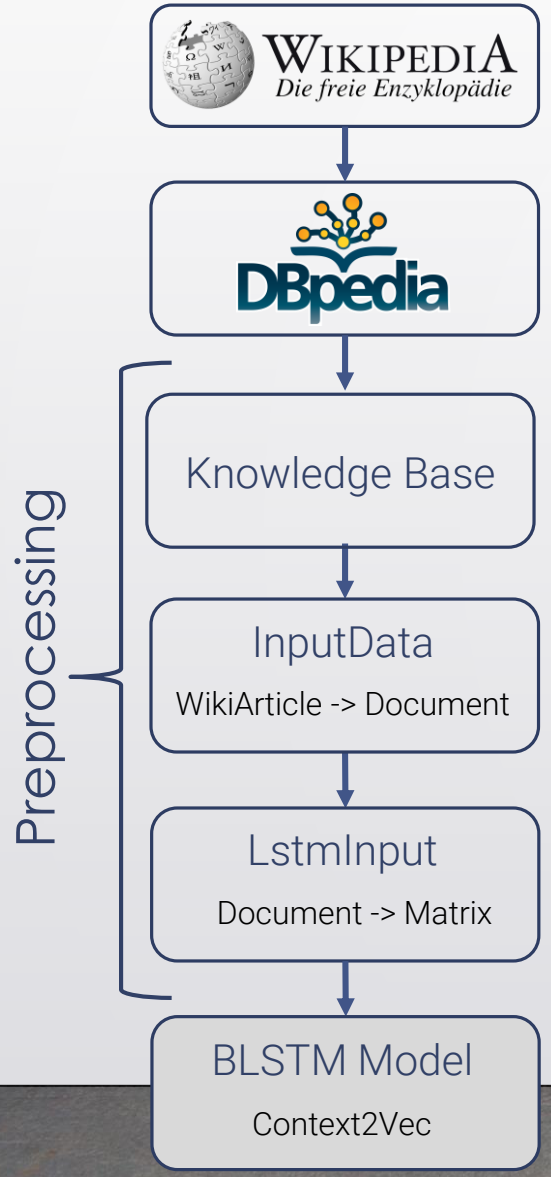
$$\cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$



Evaluierung



Datenfluss



1.556.343 Wikipedia-Artikel mit 15.859.142 Links und 153.626.686 Triples

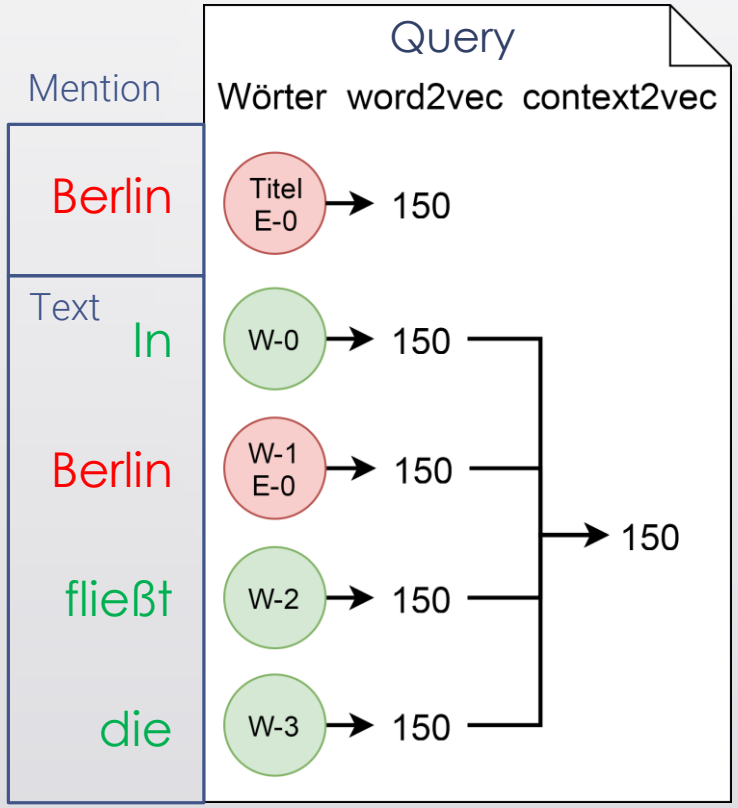
Alle notwendigen Daten werden jeweils in einem Wikipedia-Artikelobjekt zusammengefasst und als JSON-Objekt in einer Knowledge Base Liste gespeichert

Aus jedem Wikipedia-Artikelobjekt wird jeweils ein Dokument erstellt. Dabei wird der Text auf Sätze und einzelne Wörter zerteilt.

Jedes Dokument wird Token für Token mit Word2Vec gewandelt und in eine Features-Matrix angeordnet. Diese Matrix plus drei weitere (LabelMask, Labes, FeatureMask) bekommt das LSTM als Input zum Training.

Context2Vec

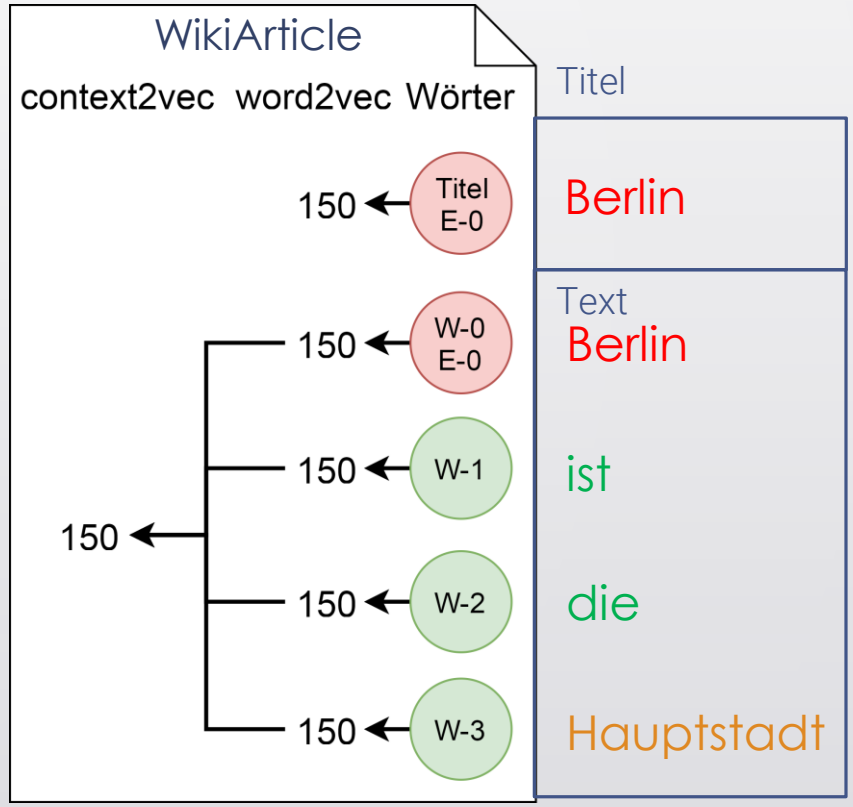
In Berlin fließt die Spree.



Kosinus-Ähnlichkeit

$$\cos(\theta) = \frac{a \cdot b}{\|a\|\|b\|} = \frac{\sum_{i=1}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

Berlin ist die Hauptstadt …



Beschreibung des Testdatensatzes

17 Dokumente

Bei der Festlegung der Artikel wurde darauf geachtet, dass diese eine größere Mehrdeutigkeit aufweisen (z.B. „Berlin“ mit 229 Wikipedia-Artikel aus den 200 Tausend, die „Berlin“ in dem Titel enthalten).

Einschränkungen:

1. der Text darf nicht leer sein
2. der Text darf nicht nur aus dem Titel bestehen
3. der Text muss mindestens aus einem Satz bestehen

Test-Text Berlin

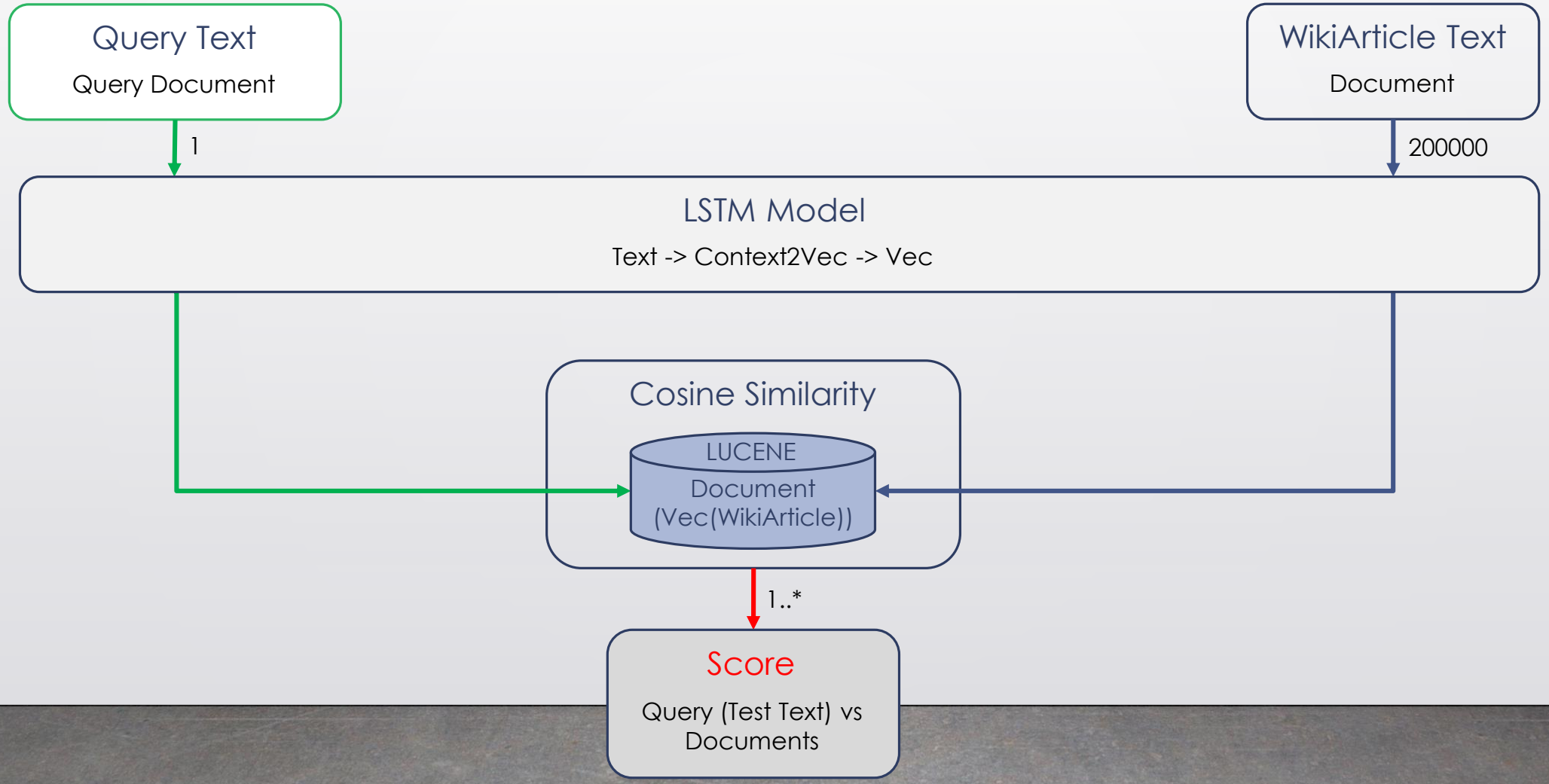
WikiArticle Text - Berlin

Berlin ( ) ist die Bundeshauptstadt der Bundesrepublik Deutschland und zugleich eines ihrer Länder. Die Stadt Berlin ist mit etwa 3,4 Millionen Einwohnern die bevölkerungsreichste und mit 892 Quadratkilometern die flächengrößte Gemeinde Deutschlands sowie nach Einwohnern die zweitgrößte der Europäischen Union. Sie bildet das Zentrum der Metropolregion Berlin/Brandenburg (6 Millionen Einw.) und der Agglomeration Berlin (4,4 Millionen Einw.). Der Stadtstaat unterteilt sich in zwölf Bezirke. Neben den Flüssen Spree und Havel befinden sich im Stadtgebiet kleinere Fließgewässer sowie zahlreiche Seen und Wälder. Urkundlich erstmals im 13. Jahrhundert erwähnt, war Berlin im Verlauf der Geschichte und in verschiedenen Staatsformen Residenz- und Hauptstadt Brandenburgs, Preußens und des Deutschen Reichs. Ab 1949 war der Ostteil der Stadt faktisch Hauptstadt der Deutschen Demokratischen Republik. Mit der deutschen Wiedervereinigung im Jahr 1990 war Berlin wieder gesamtdeutsche Hauptstadt und wurde in der Folge Sitz der Bundesregierung, des Bundespräsidenten, des Deutschen Bundestags, des Bundesrats sowie zahlreicher Bundesministerien und Botschaften. Berlin gilt als Weltstadt der Kultur, Politik, Medien und Wissenschaften. Die Metropole ist ein europäischer Verkehrsknotenpunkt und eines der meistbesuchten Zentren des Kontinents. Die Sportereignisse, Universitäten, Forschungseinrichtungen und Museen Berlins genießen internationalen Ruf. Seit der Jahrhundertwende hat sich die Stadt zu einem Anziehungspunkt für Unternehmensgründer, Kreative und Einwanderer entwickelt. Berlins Architektur, Festivals, Nachtleben und vielfältige Lebensbedingungen sind weltweit bekannt.

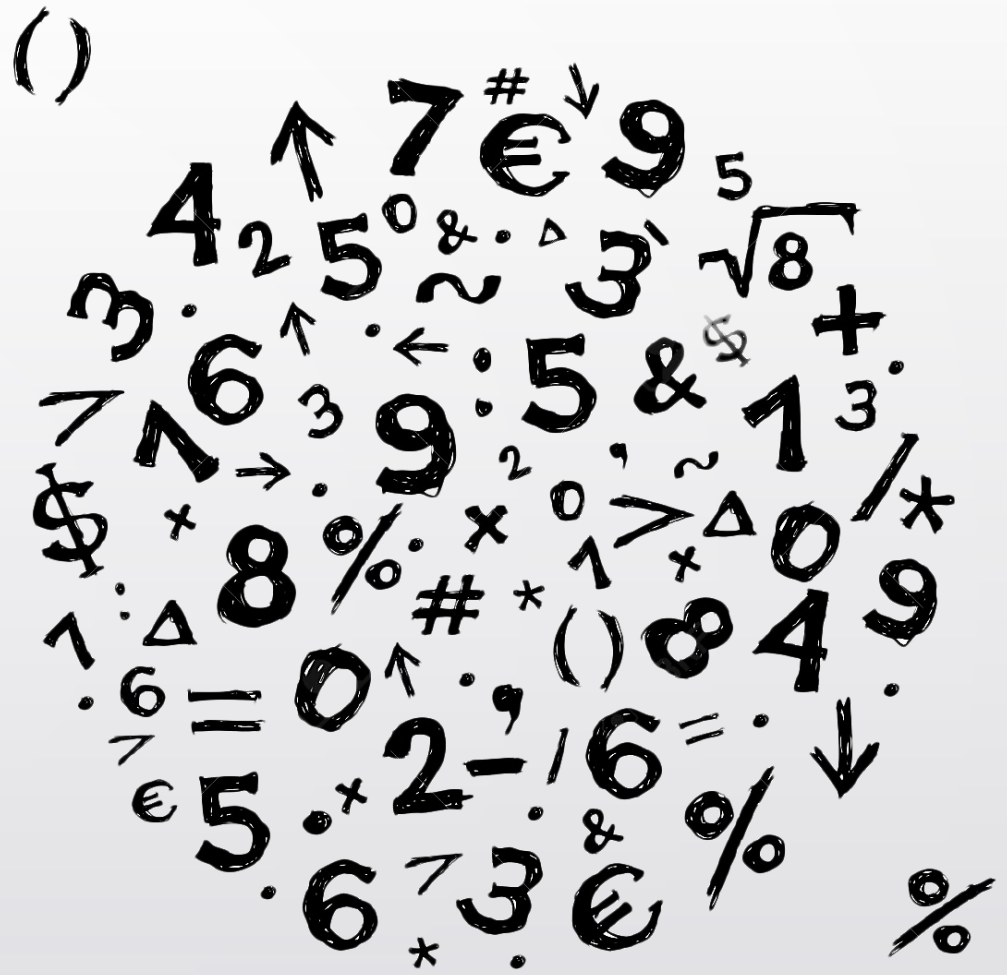
Test Text - Berlin

Eine Bundeshauptstadt an der Spree. Berlin ist eine große Stadt an der Spree und gilt als Weltstadt der Kultur, Politik, Medien und Wissenschaften.

Evaluierung



Ergebnisse



Ergebnisse

Testdokument Nummer	Test-Dokument Label Name	Test-Dokument Entität	Platzierung in der Kandidatenliste	Score	Kosinus-Ähnlichkeit
1	Berlin	Berlin	2	0,990385949611663	0,944547414779663
2	Berlin	Berlin	1	1,0	0,995530128479003
3	Berlin (Vermont)	Berlin	1	1,0	0,912318706512451
4	Berlin (New Hampshire)	Berlin	3	0,978339076042175	0,963497936725616
5	Berlin (Maryland)	Berlin	2	0,999016284942626	0,975255250930786
6	Berlin (Massachusetts)	Berlin	2	0,994660258293151	0,982862353324890
7	Berlin (Kolumbien)	Berlin	2	0,924501001834869	0,772243380546569
8	Berlin (Schiff, 1985)	Berlin	1	1,0	0,883222460746765
9	Berlin (Schiff, 1909)	Berlin	1	1,0	0,943961322307586
10	Berlin Sluggers	Berlin	1	1,0	0,986393392086029
11	Berliner Fernsehturm	Fernsehturm	1	1,0	0,958419859409332
12	Fernsehturm Ostankino	Fernsehturm	1	1,0	0,958419859409332
13	Fernsehturm Dresden	Fernsehturm	1	1,0	0,972606837749481
14	Paris	Paris	1	1,0	0,953020513057708
15	Paris (New York)	Paris	1	1,0	0,944958746433258
16	Paris (Mythologie)	Paris	1	1,0	0,953054547309875
17	Paris (Fernsehserie)	Paris	1	1,0	0,989795446395874
durchschnittliches Ergebnis			1,35	0,993347210042617	0,946476950364954

Kandidatenliste Berlin (Maryland)

Liste mit den 10 Kandidaten mit der größten Kosinus-Ähnlichkeit zu dem Testdokument

Testdokument Titel: "Berlin (Maryland)"

1. 0,9762389659881592 : "Berlin (Wisconsin)"
2. 0,9752552509307861 : "Berlin (Maryland)"
3. 0,9705436825752258 : "Berlin (Massachusetts)"
4. 0,9684509634971619 : "Berlin (New Hampshire)"
5. 0,9654411673545837 : "Berlin (West Virginia)"
6. 0,9495586156845093 : "Berlin (Kentucky)"
7. 0,9428457021713257 : "Berlin Historic District (Nevada)"
8. 0,8983588814735413 : "Berlin Mills Railway"
9. 0,7974255084991455 : "Berlin (Schiff, 1924)"
10. 0,7919023633003235 : "Berlin Observer„

37. 0, 6938755512237549 : "**Berlin**"

Kandidatenliste Berlin (Schiff, 1985)

Liste mit den 10 Kandidaten mit der größten Kosinus-Ähnlichkeit zu dem Testdokument

Testdokument Titel: "Berlin (Schiff, 1985)"

1. 0,8832224607467651 : "Berlin (Schiff, 1985)"
2. 0,8421953916549683 : "Berlin (Schiff, 2012)"
3. 0,8269113302230835 : "Berlin (Schiff, 1889)"
4. 0,7671408653259277 : "Berlin (Schiff, 1906)"
5. 0,7588434815406799 : "Berlin (Schiff, 1924)"
6. 0,7578482031822205 : "Berlin Brigade"
7. 0,7467612624168396 : "Berlin (Radar)"
8. 0,7300619482994080 : "Berlin (Schiff, 1909)"
9. 0,7295939922332764 : "Kommandant des sowjetischen Sektors von Berlin"
10. 0,6869557499885559 : "Landsmannschaft Berlin-Mark Brandenburg„

159. 0,49101918935775757 : "**Berlin**"

Kandidatenliste Paris (Mythologie)

Dabei enthielt der Text von den Testdokumenten keine Entität oder Label. In dem Text von dem Testdokument von Paris als Bruder von Hektor ist das Wort „Paris“ durch „Er“ ersetzt.

Titel: „Paris (Mythologie)“
Nennung: „“
Text: „Er ist der Bruder von Hektor und der Sohn des trojanischen Königs Priamos. Er ist für die Auslösung von dem Trojanischen Krieg verantwortlich.“

Wilipedia-Text: „Paris [ˈpaːrɪs] (griech. Πάρις) ist in der griechischen Mythologie der Sohn des trojanischen Königs Priamos und der Hekabe. Er ist damit Bruder des Hektor und der Kassandra. Insgesamt hat er mehr als 50 Geschwister und Halbgeschwister. Indem er Helena entführt, löst er den Trojanischen Krieg aus. Auffallenderweise, und bisher unerklärt, trägt er in Homers Ilias noch einen zweiten Namen, und diesen sogar häufiger: Alexandros. Überwiegend, wohl wegen des Einflusses Homers, heißt er so auch in den Beischriften der Vasenmalerei. Möglicherweise hängt der Name mit dem König Alaksandu von Wilusa zusammen, der auch in hethitischen Texten vorkommt.“

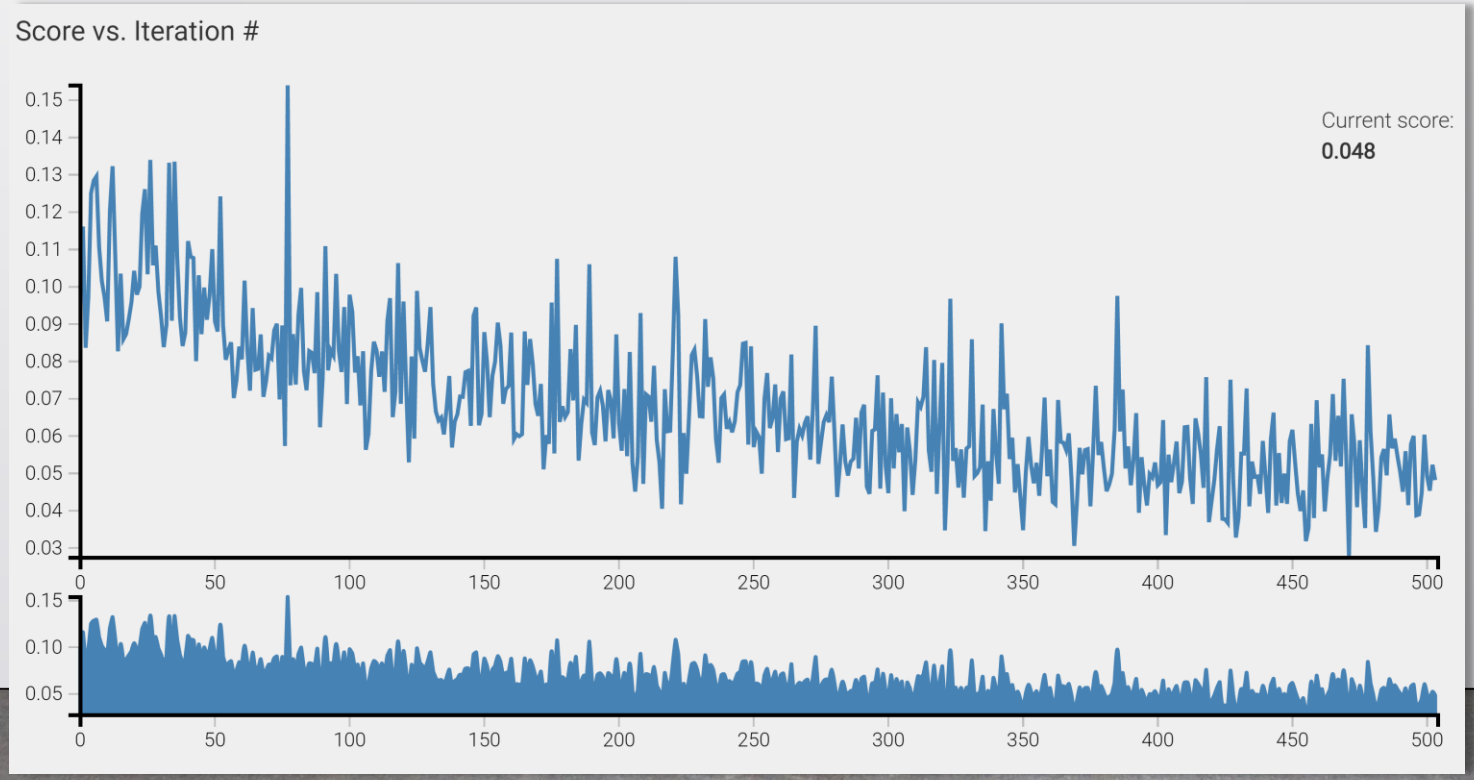
Testdokument Titel: „Paris (Mythologie)"

1. 0,9746598601341248 : "Paris (Mythologie)"
2. 0,919340193271637 : "Urteil des Paris«

Ausblick

Ausblick

- Trainingsdaten optimieren (bei zu kleinen Abstract-Texten sollte der Text durch den restlichen Text ergänzt werden)
- Training optimieren (3 -> 25 Epochen) + paralleles Training



BLST Training Score vs. Iteration (Training mit 1000 Artikeln). Diese Abbildung zeigt, wie der Score sich mit jeder weiteren Iteration verändert. Der Score-Wert ist nicht mit dem Score aus der Evaluierung vergleichbar. Ab der Iteration 450 (was in dem Fall der Epoche 25 entspricht) findet keine signifikante Verbesserung von Score mehr statt (Epoche 3 entspricht 54 Iterationen).

Ohne ein paralleles Training würde die Trainingsdauer 250 Tage betragen, um ein Modell mit 1,6 Millionen Wikipedia-Artikel mit 25 Epochen zu trainieren.

Ausblick

- Erweiterung des Tasks um Entity-Mengen (allen erkannten Entitäten mit dazu errechneten Score für die gesamte Dokumentenbewertung zusammenfassen)
- Wie gut ist Context2Vec auf einem "klassischen" Entity Linking Task?
- Es ergab sich ein neues Grundproblem.

  Eine effiziente (unter 30 ms) Suche mit Cosinus-Similarity auf über 3M Vektoren? Mit INDArrays dauert die Cosinus-Similarity mit 3500K Vektoren über 4000 ms.
- Wo könnte man das Context2Vec Embedding sonst noch einsetzen?

Grundlagen und Literatur

Grundlagen und Literatur

Paper

Entity Linking

- Ben Hachey - Evaluating Entity Linking with Wikipedia (Artificial Intelligence 2013)
- Xiao Ling - Design Challenges for Entity Linking (2015)

Kontextbasiert

- Magnus Sahlgren - The distributional hypothesis (2006)
- Tomas Mikolov - Efficient Estimation of Word Representations in Vector Space (2013)
- Oren Melamud - Learning Generic Context Embedding with Bidirectional LSTM (Cntext2Vec) (CoNLL 2016)

Dokumentenebene

- Tom Kenter - Short Text Similarity with Word Embeddings (CIKM 2015)
- Sourav Dutta undGerhard Weikum - A Joint Model for Cross-Document Co-Reference Resolution and Entity Linking (C3EL 2015)
- Sameer Singh - A Large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia (2012)

Implementierung / Deep Learning

- Sebastian Arnold - Robust Named Entity Recognition in Idiosyncratic Domains (2016)
- Sebastian Arnold - TASTY: Interactive Entity Linking As-You-Type (2016)

Quellen

Denis Martin - Kontextbezogenes Entity Linking auf Dokumentenebene mit Deep Learning (Beuth Hochschule für Technik Berlin 2017)

Bilder

- Word2Vec F16 (https://deeplearning4j.org/img/countries_capitals.png)
- Daten F12, F21, F27 (http://previews.123rf.com/images/kudryashka/kudryashka1111/kudryashka111100073/11264058-Sketch-frame-with-math-symbols-for-your-design-Stock-Vector-math-mathematics-mathematical.jpg)
- Umsetzung F14 (http://www.max-con.de/uploads/assets/methode_big.gif)
- Grundlagen und Literatur F35 (https://weltbild.scene7.com/asset/vgwvpg/vgw/styx-standard-xl/kidoh-schmuckbild-buecher.jpg)