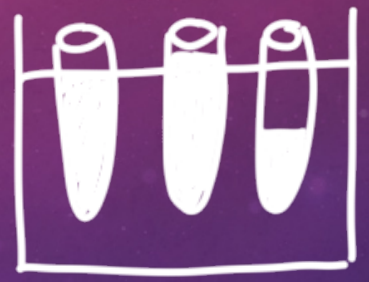
GENDERGERECHTE SPRACHE UND MACHINE LEARNING

ERFAHRUNGSAUSTAUSCH

Larissa Haas

Motivation

Hintergrund:
Suchmaschinen und
Machine Learning



Experiment 1:
Preprocessing

Experiment 2:
Word Embeddings

Elea Brandt, queerfeministische Shitstormtrooperin
@EleaBrandt

Das mit der Aussprache ist echt hart, ich weiß.

Bei der Gesellschaft deutscher Sprache werden Leute auch nie beurlaubt, kochen nie Spiegelei und kreieren keine schönen Dinge.

Weil das kann einfach NIEMAND aussprechen!

> **MDR AKTUELL** @MDRAktuell · 22 Std.
> Die Gesellschaft für deutsche Sprache rät von der Nutzung des Gendersternchens ab. Der Verein teilte mit, das Sternchen sei mit Grammatik und Rechtschreibung nicht vereinbar. Zudem sei die Aussprache unklar.

2:59 nachm. · 13. Aug. 2020 · Twitter Web App

192 Retweets **685** „Gefällt mir"-Angaben **11** Zitate

James A. Sullivan
@fantasyautor

Die Gesellschaft für deutsche Sprache hat also was gegen das Gendersternchen.🙄

Die Positionen, die da vertreten werden, zeigen nicht gerade ihre Aufgeschlossenheit. Dort scheint kaum ein Interesse an einer lebendigen Sprache vorhanden zu sein.

gfds.de/gendersternche...

9:57 nachm. · 13. Aug. 2020 · Twitter for iPhone

45 Retweets **177** „Gefällt mir"-Angabe

Roxane Bicker | Autorin ✏️
@roxane_bicker

Zwecks "kann man ja nicht aussprechen" vom Genderstern, ne?
Vielleicht sollten sich alle mal etwas mehr mit sprachlichen Hintergründen beschäftigen, Sprache ist nämlich sowas von inkonsequent 😂
1/?

> **Elea Brandt, queerfeministische Shitstormtrooperin** @EleaBr... · 20 Std.
> Das mit der Aussprache ist echt hart, ich weiß.
>
> Bei der Gesellschaft deutscher Sprache werden Leute auch nie beurlaubt, kochen nie Spiegelei und kreieren keine schönen Dinge.
>
> Weil das kann einfach NIEMAND aussprechen! twitter.com/MDRAktuell/sta...
> Diesen Thread anzeigen

7:39 vorm. · 14. Aug. 2020 · Twitter for Android

15 Retweets **37** „Gefällt mir"-Angaben **1** Zitat

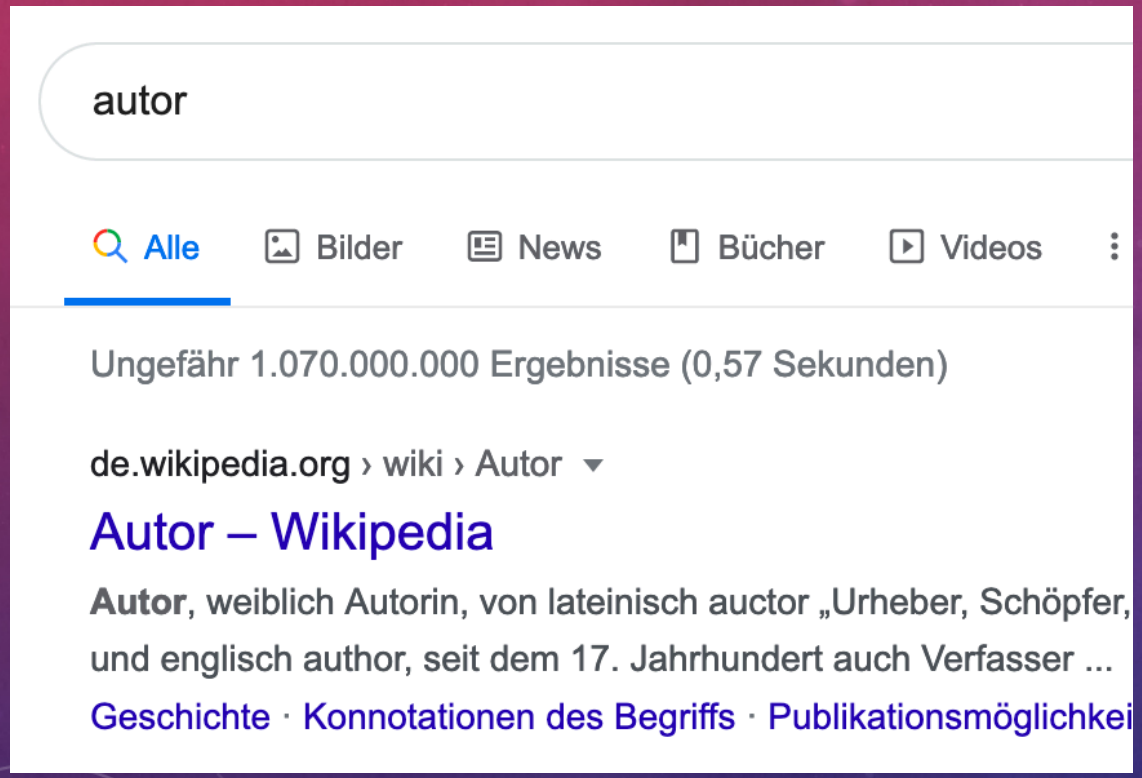

Google

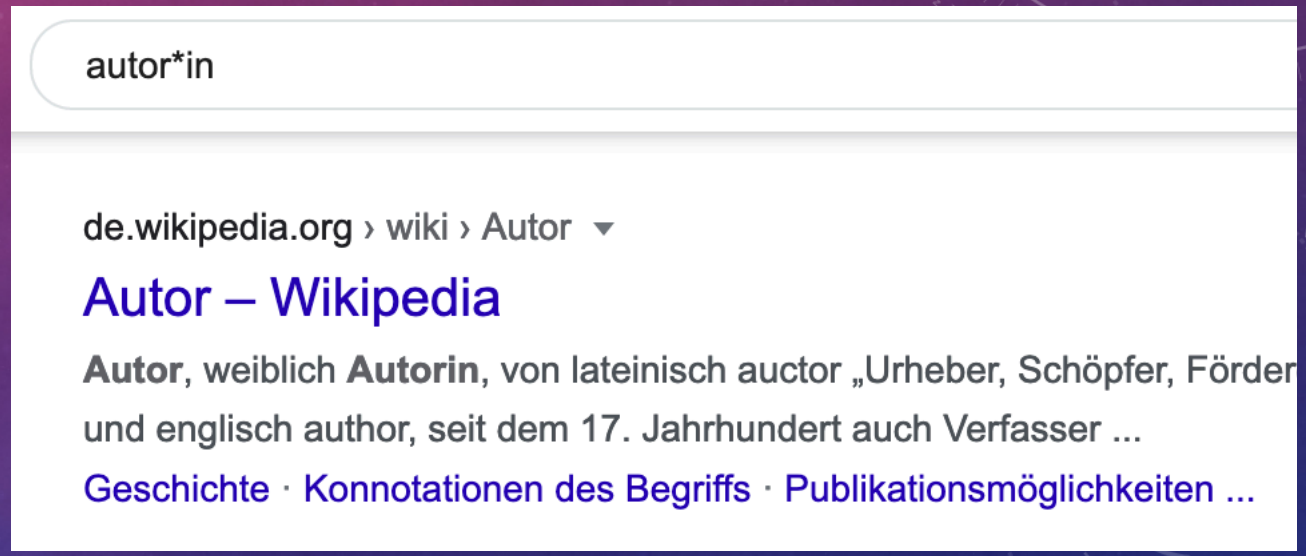
HINTERGRUND

- Artikel: Bezeichne ich mich als Autorin/Lektorin/… auf meiner Homepage, werde ich nicht so gut gefunden, als Autor/Lektor/… dagegen besser
- Zwei Gründe
  - Potentiell immer noch die häufigere Schreibweise/Nutzung
  - Grundlagen, wie Suchmaschinen funktionieren

HINTERGRUND

- Suchmaschinen gehören inzwischen zum Bereich Data Science und Machine Learning
- Früher mit einfachen Keyword Übereinstimmungen
  - Je seltener ein Wort, desto wichtiger ist es für die Relevanz einer Webseite
  - Dieses Wort wird aber auch nicht häufig gesucht, deshalb sind “Buzzwords” und “Keywords”, die oft gesucht werden, wichtig für eine Webseite
- Außerdem: PageRank
  - Wie sind Webseiten untereinander verlinkt (bezeichnet Relevanz)
  - Und wie oft werden sie bei bestimmten Suchen gefunden und angeklickt?

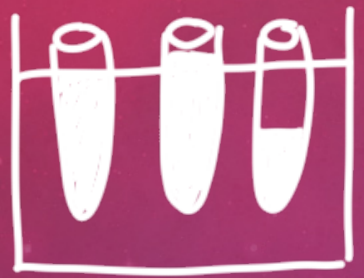

autor
Alle
Bilder
News
Bücher
Videos
Ungefähr 1.070.000.000 Ergebnisse (0,57 Sekunden)
de.wikipedia.org › wiki › Autor
Autor – Wikipedia
Autor, weiblich Autorin, von lateinisch auctor „Urheber, Schöpfer,
und englisch author, seit dem 17. Jahrhundert auch Verfasser ...
Geschichte · Konnotationen des Begriffs · Publikationsmöglichkei


autor*in
de.wikipedia.org › wiki › Autor
Autor – Wikipedia
Autor, weiblich Autorin, von lateinisch auctor „Urheber, Schöpfer, Förder
und englisch author, seit dem 17. Jahrhundert auch Verfasser ...
Geschichte · Konnotationen des Begriffs · Publikationsmöglichkeiten ...

HINTERGRUND

- Nun: auch klassisches Machine Learning (Natural Language Processing) im Einsatz, um über direkte Wort-Entsprechungen hinaus zu gehen
- Sowohl Suchtext als auch Webseitentexte werden “preprocessed”, also vorbereitet, damit die Maschine besser damit umgehen kann
- Das passiert nicht nur bei Suchmaschinen, sondern bei so ziemlich allen Arten von Machine Learning, die mit Text zu tun haben
  - Beispiele: QA-Systeme

EXPERIMENT 1



- Preprocessing mit Gendergerechter Sprache
- Weil: Hier wird auf Zeichen zurückgegriffen, die zwar in der Schriftsprache gebräuchlich sind, also erstmal kein Problem für die Maschine darstellen, die aber inhaltlich anders interpretiert werden.
- Symbole und Zeichen bedeuten meist eine inhaltliche Trennung, ein Sinnabschnitt und warden entweder auch so interpretiert (also immer wenn ein Punkt kommt, trenne ich den Text davor von dem Text danach), oder einfach ignoriert und entfernt, damit der Text von ungewollten Zeichen ”gesäubert” ist

EXPERIMENT 1

- Standard Preprocessing kommt aus dem Englischen Sprachgebiet, wo Probleme mit generischem Maskulinum gar nicht erst auftauchen bzw. die Notwendigkeit für Gendergerechte Sprache eher gering ist

EXPERIMENT 1

- Wie verhalten sich nun Standard-Vorgehensweisen von Machine Learning zu Wörtern wie Autor*in, Lektor:in, Journalist_in?

M↓

```
text = "Autorinnen und Autoren, Autor:innen Autor_innen, AutorInnen Autor/innen, Autor*innen, Autor(innen)"
pronomen = "jede*r, jede:r, jede_r, jede/r, eine*r, eine:r, eine_r, eine/r"
```

EXPERIMENT 1

- Doppelpunkt fast immer als Trennungszeichen interpretiert, Stern und Unterstrich wurden beibelassen

```
lemmas = []
for token in doc:
  if token.is_punct is False and token.is_stop is False:
    print(token.lemma_)
    lemmas.append(token.lemma_)
```

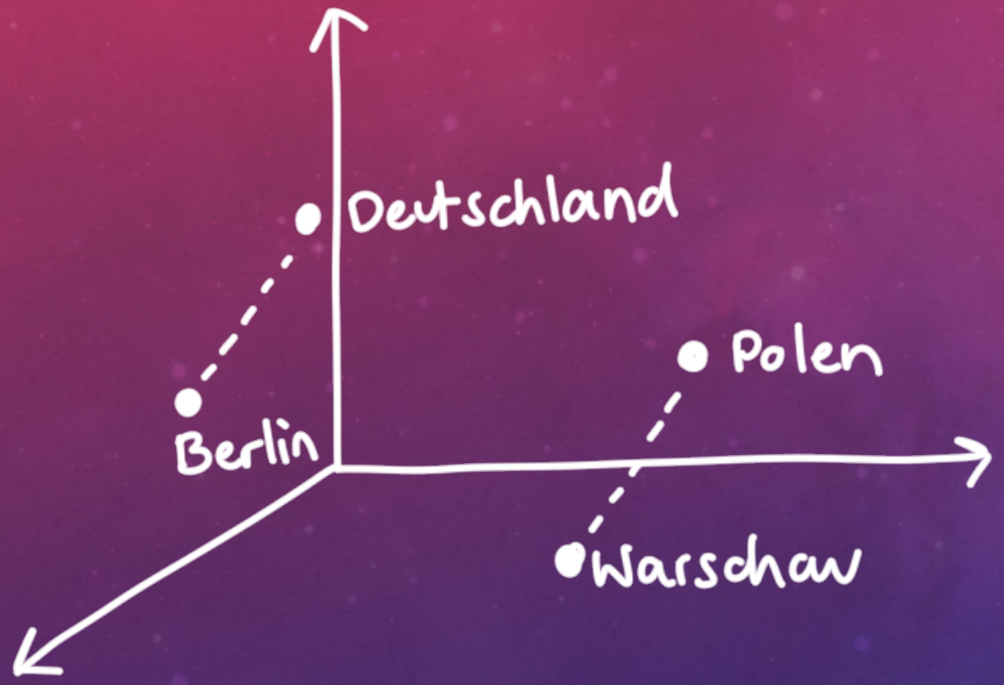
```
Autorinnen
Autoren
Autor
innen
Autor_innen
AutorInnen
Autor/innen
Autor*innen
Autor
innen
```

EXPERIMENT 2

- In vielen Anwendungen werden sogenannte Word Embeddings benutzt

- Texte werden in ein Neuronales Netz gespeist, je nach Netz wird dabei nur das Wort oder auch der Kontext, in dem das Wort vorkommt, in Gewichtungen umgewandelt.
- Jedes Wort wird dann schließlich durch einen Vektor repräsentiert, also eine Reihe von Zahlen. Man kann sich das so vorstellen:
  - …

EXKURS WORD EMBEDDINGS



[45]

```
berlin = model.get_vector('Berlin')
deutschland = model.get_vector('Deutschland')
polen = model.get_vector('Polen')
result = berlin - deutschland + polen
```

[47]

```
model.similar_by_vector(result)
```

```
[('Polen', 0.8874868154525757),
 ('Warschau', 0.7945132851600647),
 ('Berlin', 0.7723734974861145),
 ('Breslau', 0.7452571988105774),
 ('Krakau', 0.7378164529800415),
 ('Stettin', 0.708252429962158 2),
```

EXPERIMENT 2

- Was machen die State-of-the-Art Word Embeddings aus Wörtern mit Gendergerechter Sprache?

```
[61]  M↓
    for w in words:
        print(w, [w in model.words])

Autorinnen [True]
Autoren [True]
Autorin [True]
Autor [True]
Autor:innen [False]
Autor_innen [False]
AutorInnen [True]
Autor/innen [False]
Autor*innen [False]
Autor(innen) [False]
```

FAZIT

- Umfgang / Einheitliches Vorgehen schwierig
- Ihr als Textschaffende – was würdet ihr euch wünschen?
- Wo seht ihr noch Probleme, die ich vielleicht gar nicht auf dem Schirm habe?
- Habt ihr noch Fragen?