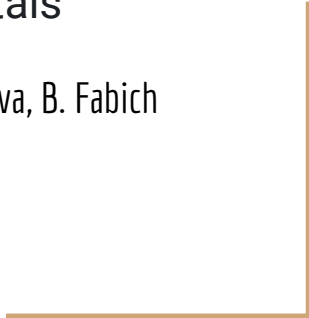




Latent Dirichlet Allocation

Machine Learning Fundamentals

N. Elsässer, J. Schneeberg, A. Stöhrer da Silva, B. Fabich

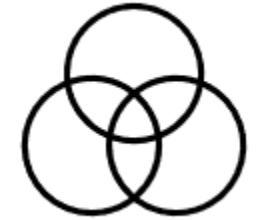


Ursprung



- Durch die Digitalisierung stieg/steigt die Anzahl von digitalen Texten
 - Deutsches Textarchiv umfasst um die 145 Mio Texte
- Diese Masse muss **verwaltet** werden, um **gefunden** und **abgerufen** werden zu können

Topic Modelling



- große Textmengen können automatisch in Themenbereiche unterteilt werden
 - Annahme: Jede **Wortform** ist zu einem **Themenbereich** zugehörig
- Durch die **Verteilung** der Wortformen sollen die Themenbereiche des Textes abgeleitet werden können

Stärken, Schwächen



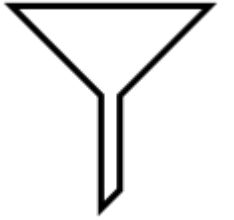
- + große Texte können zusammengefasst und klassifiziert werden
 - + Der Vorgang geschieht automatisch
 - Informationen wie die Struktur der Syntax oder Regelmäßigkeiten von Aussagen resultieren nicht
- Kann als Vorarbeit für weitere Analysen genutzt werden

Grundlagen



- Das Latent Dirichlet Allocation (LDA) Verfahren ist ein 3 Stufiges Bayesian Modell um Topic Modelling durchzuführen
- **Latent:** Die Themen sind unbekannt, müssen aus den Dokumenten und dessen **Wortverteilung** abgeleitet werden
- **Dirichlet:** Verteilung stellt die Prioriverteilung (Anfangswahrscheinlichkeit) für die Themenverteilung dar. Parameter bestimmen die Anhäufung der Themen
- **Allocation:** Beschreibt, wie Wörter aufgrund von deren Wahrscheinlichkeit den Themen zugeordnet sind

Datenquelle



- Art: 210.000 Schlagzeilen der HuffPost
- Zeitraum: 2012 bis 2022
- Verfahren: Webscraping
- Quelle: [kaggle.com](https://www.kaggle.com)

	link	headline	category	short_description	authors	date
0	https://www.huffpost.com/entry/covid-boosters-...	Over 4 Million Americans Roll Up Sleeves For O...	U.S. NEWS	Health experts said it is too early to predict...	Carla K. Johnson, AP	2022-09-23
1	https://www.huffpost.com/entry/american-airlin...	American Airlines Flyer Charged, Banned For Li...	U.S. NEWS	He was subdued by passengers and crew when he ...	Mary Papenfuss	2022-09-23
2	https://www.huffpost.com/entry/funniest-tweets...	23 Of The Funniest Tweets About Cats And Dogs ...	COMEDY	"Until you have a dog you don't understand wha...	Elyse Wanshel	2022-09-23
3	https://www.huffpost.com/entry/funniest-parent...	The Funniest Tweets From Parents This Week (Se...	PARENTING	"Accidentally put grown-up toothpaste on my to...	Caroline Bologna	2022-09-23
4	https://www.huffpost.com/entry/amy-cooper-lose...	Woman Who Called Cops On Black Bird-Watcher Lo...	U.S. NEWS	Amy Cooper accused investment firm Franklin Te...	Nina Golgowski	2022-09-22

Datenbereinigung

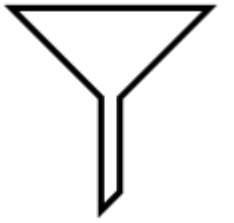


- Maßnahmen: stopwords, tokenize, lemmatizer, Satzzeichen, Kleinbuchstaben
- Neue Spalte für Weiterverarbeitung

```
print(preprocess_text("This is a very important text mentioning the President, the Health Department and Germany"), sep=" ")  
['important', 'text', 'mentioning', 'president', 'health', 'department', 'germany']
```

short_description	processed_text
I had never heard of "Law & Order: Special Vic...	[olivia, benson, belief, never, heard, law, or...
Thiel, a tech billionaire, is an outspoken sup...	[new, lgbtq, club, tie, peter, thiel, people, ...
According to a newly released report, the Unit...	[country, win, olympic, medal, 247, wall, st, ...
Libby was convicted of lying about how he lear...	[trump, pardon, scooter, libby, dick, cheneys,...
A man was stabbed to death and his wife was se...	[deadly, stabbing, attack, maryland, prayer, c...

Datenbereinigung - Besonderheit



- Zu breite Kategorien

headline	category	short_description
93-Year-Old Woman Goes Viral When She Tells In...	WEIRD NEWS	On Monday, Coors Light dropped off 150 cans of...
Shooting At Brooklyn Community Event Leaves 1 ...	U.S. NEWS	New York City Mayor Bill de Blasio vowed to "d...
Woman Accused Of Commandeering Bus After Drive...	WEIRD NEWS	She didn't get very far.
Now I Know What Fear Is	WORLD NEWS	Everywhere I looked on my way home, someone wa...
Reinventing Europe Along These 7 Points	WORLDPOST	In its present form, the EU is weak, particula...

Gensim-Dictionary und Bag of Words



Beispiel mit Tieren:

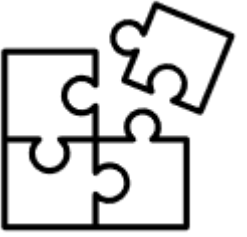
Daten = [['cat', 'meow'], ['dog', 'bark'], ['cat', 'dog']]

Gensim-Dictionary = {0: 'cat', 1: 'meow', 2: 'dog', 3: 'bark'}

Korpus =

```
[  
    [(0, 2), (2, 1)], # "I have a cat, another cat and a dog"  
    [(2, 1), (3, 1), (1, 1)] # "My dog barks and does not meow"  
]
```

Data Splitting



nicht implementiert, weil...

...keine Vorhersagen

...keine unbekannten
Daten

...mehr Daten = bessere
Qualität

Simple LDA



Wir brauchen...



Ablauf

1. Daten laden
2. Daten bereinigen
3. Daten aufbereiten
4. Dictionary und Corpus erstellen
5. Modell trainieren
6. 50 Minuten warten... (unlucky)
7. Freudensprung machen wenn das Modell fertig ist
8. Visualisierung

```
import time
start_time = time.time()
lda_model = LdaModel(corpus=corpus, id2word=id2word,
                     num_topics=unique_categories,
                     random_state=42,
                     passes=10,
                     alpha="auto",
                     per_word_topics=True)

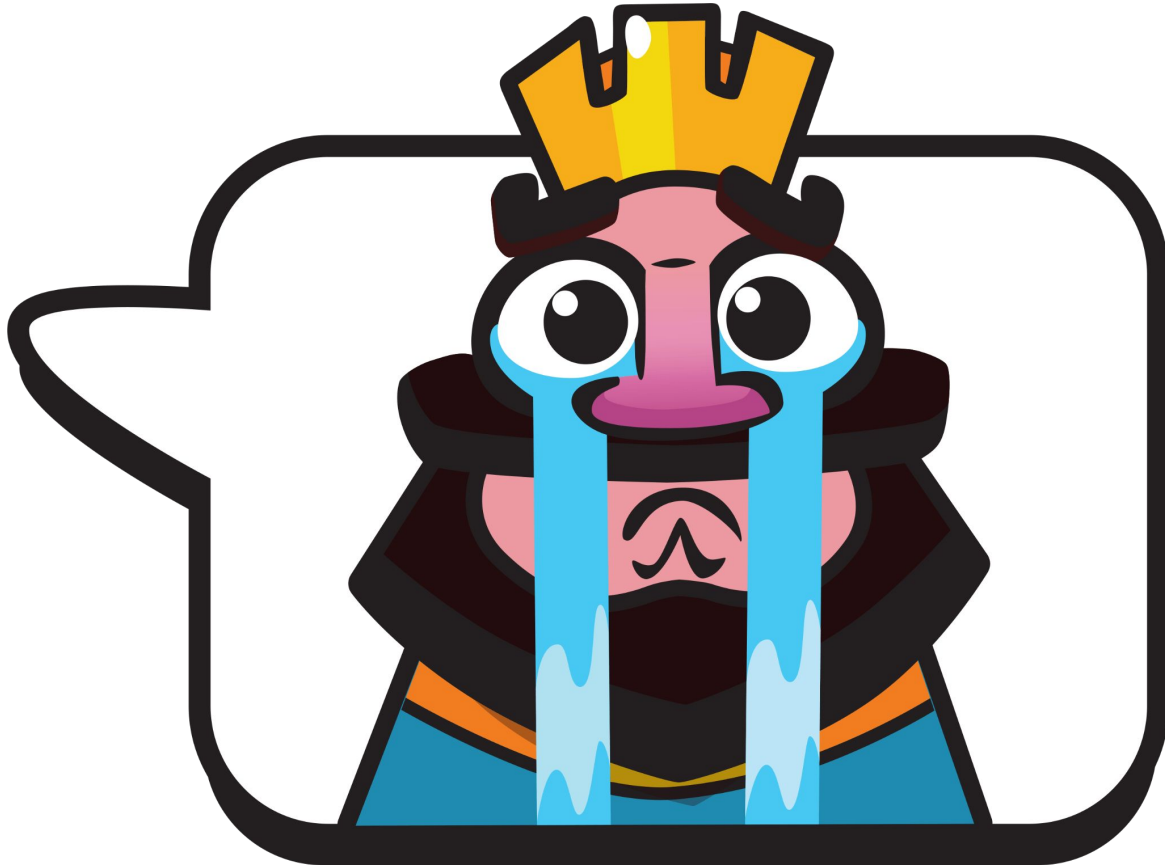
stop_time = time.time()
run_time = stop_time - start_time
print("Wall time:", run_time/60, "m")
```

Wall time: 50.23810725212097 m

Die Evaluierung

1. Dominantes Topic für jedes Dokument finden
2. Nach Kategorien filtern, häufigstes Topic mappen

Die Evaluierung - 2a.



```
pprint(category_topic_mapping)
```

```
{'ARTS': 11,  
 'ARTS & CULTURE': 11,  
 'BLACK VOICES': 11,  
 'BUSINESS': 11,  
 'COLLEGE': 11,  
 'COMEDY': 11,  
 'CRIME': 11,  
 'CULTURE & ARTS': 11,  
 'DIVORCE': 11,  
 'EDUCATION': 11,  
 'ENTERTAINMENT': 11,  
 'ENVIRONMENT': 11,  
 'FIFTY': 11,  
 'FOOD & DRINK': 11,  
 'GOOD NEWS': 11,  
 'GREEN': 11,  
 'HEALTHY LIVING': 11,  
 'HOME & LIVING': 11,  
 'IMPACT': 11,  
 'LATINO VOICES': 11,  
 'MEDIA': 11,  
 'MONEY': 11,  
 'PARENTING': 11,  
 'PARENTS': 11,  
 'POLITICS': 11,  
 'QUEER VOICES': 11,  
 'RELIGION': 11,  
 'SCIENCE': 11,  
 'SPORTS': 11,  
 'STYLE': 11,  
 'STYLE & BEAUTY': 11,  
 'TASTE': 11,  
 'TECH': 11,  
 'TRAVEL': 11,  
 'WEDDINGS': 11}
```

Die Evaluierung

1. Dominantes Topics für jedes Dokument finden
- ~~2. Nach Kategorien filtern, häufigstes Topic mappen~~
 - a. Schlechte Idee
3. Nach Topics filtern, häufigste Kategorie mappen
 - a. Besser

Die Evaluierung - 3a.



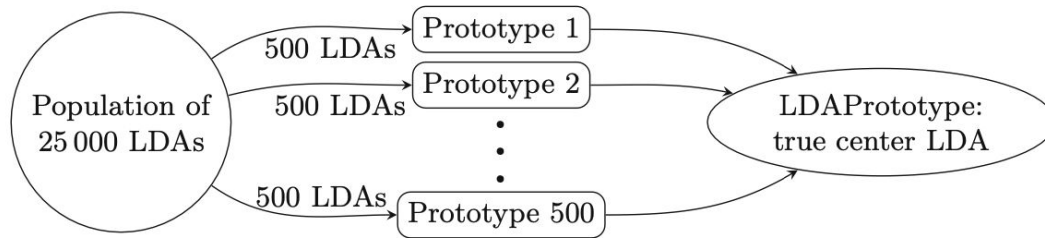
```
pprint(topic_category_mapping)
```

```
{1: 'BUSINESS',  
2: 'IMPACT',  
3: 'POLITICS',  
6: 'COMEDY',  
7: 'ENTERTAINMENT',  
9: 'MONEY',  
10: 'DIVORCE',  
11: 'SPORTS',  
12: 'ENVIRONMENT',  
13: 'BLACK VOICES',  
14: 'FOOD & DRINK',  
16: 'MEDIA',  
18: 'STYLE & BEAUTY',  
19: 'TRAVEL',  
20: 'WELLNESS',  
21: 'PARENTING',  
22: 'PARENTS',  
26: 'CRIME',  
27: 'GREEN',  
28: 'HOME & LIVING',  
29: 'WEDDINGS',  
30: 'RELIGION',  
36: 'QUEER VOICES'}
```

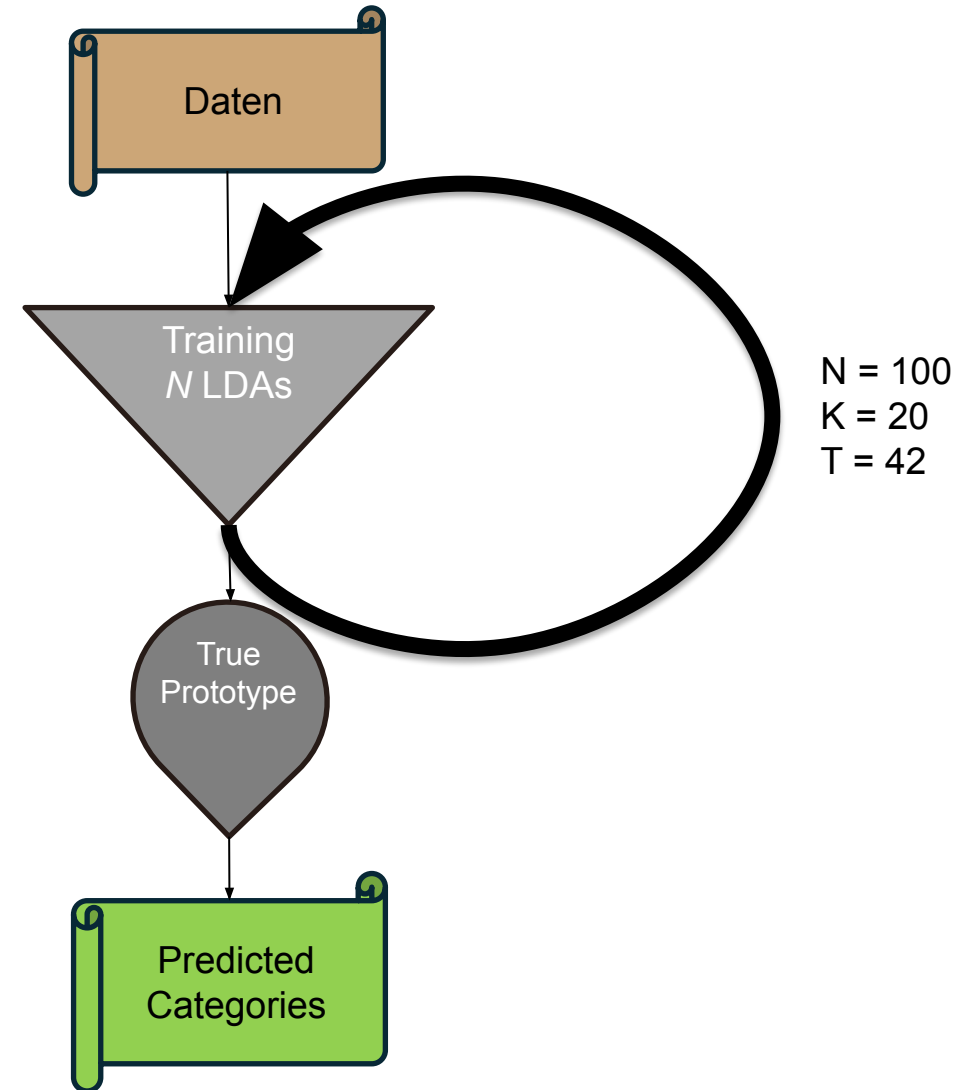
Die Evaluierung

1. Dominantes Topics für jedes Dokument finden
2. Nach Kategorien filtern, häufigstes Topic mappen
 - a. Schlechte Idee
3. Nach Topics filtern, häufigste Kategorie mappen
 - a. Besser
4. Daten matchen und exportieren zum Vergleich

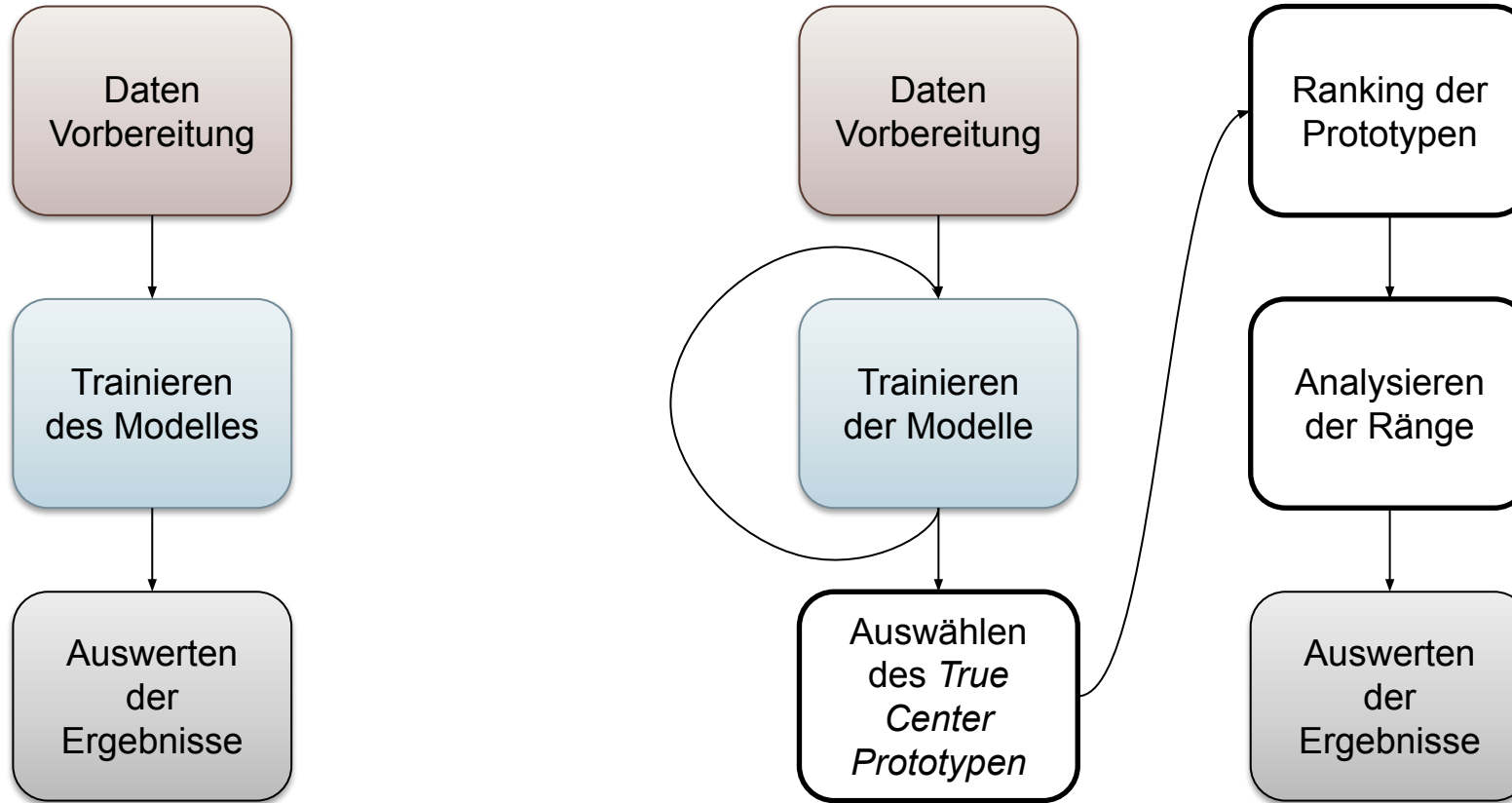
Improved LDA



(Rieger et al. (2020))



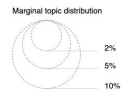
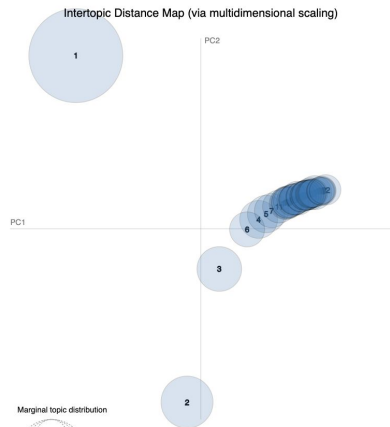
Ablauf Unterschiede



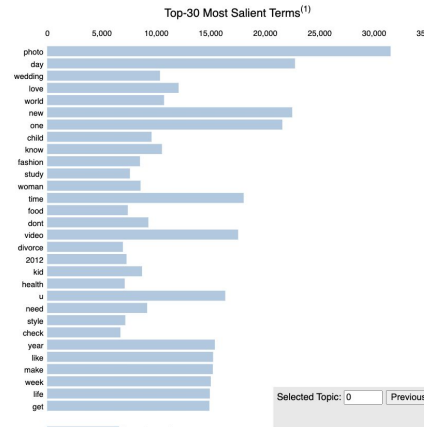
Visualisierung



Selected Topic: 0 Previous Topic Next Topic Clear Topic

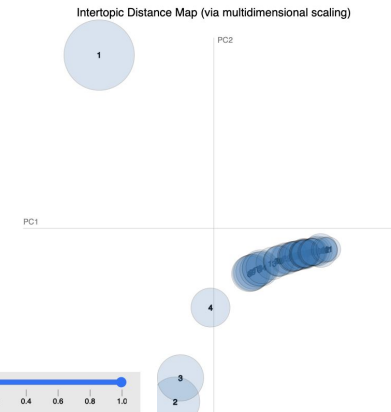


Slide to adjust relevance metric:⁽²⁾
 $\lambda = 1$ 0.0 0.2 0.4 0.6 0.8 1.0

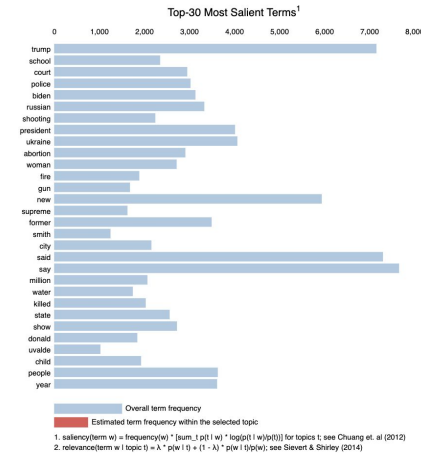


e?

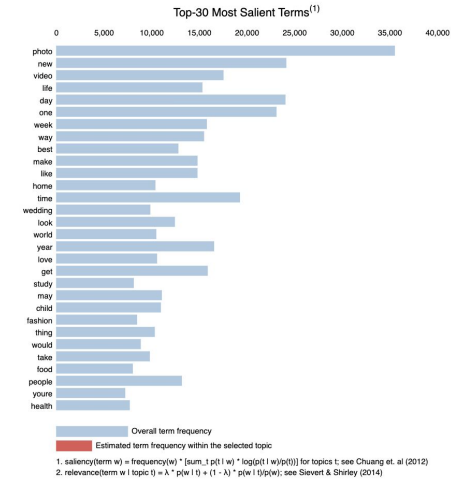
Selected Topic: 0 Previous Topic Next Topic Clear Topic



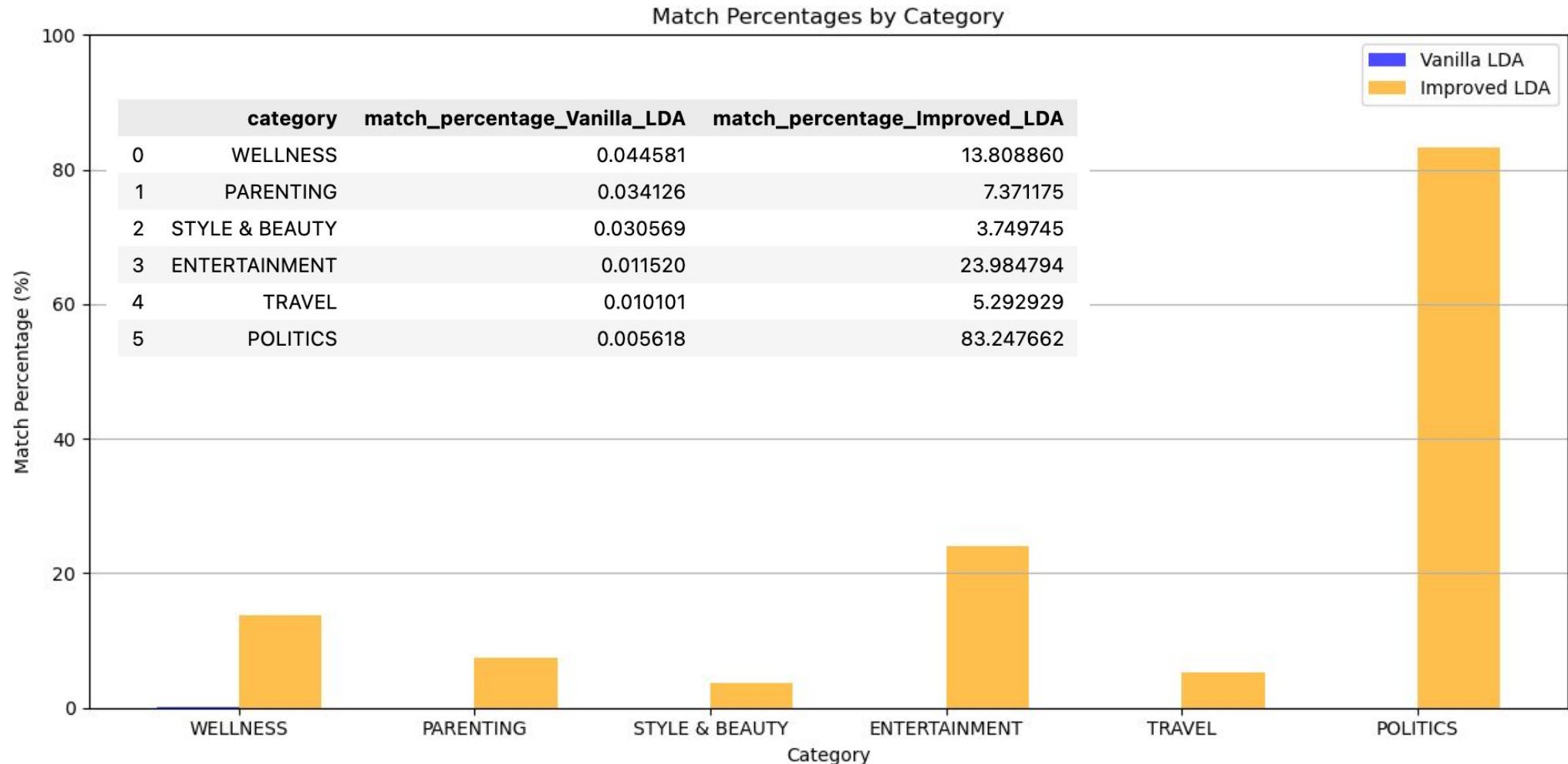
Slide to adjust relevance metric:⁽²⁾
 $\lambda = 1$ 0.0 0.2 0.4 0.6 0.8 1.0



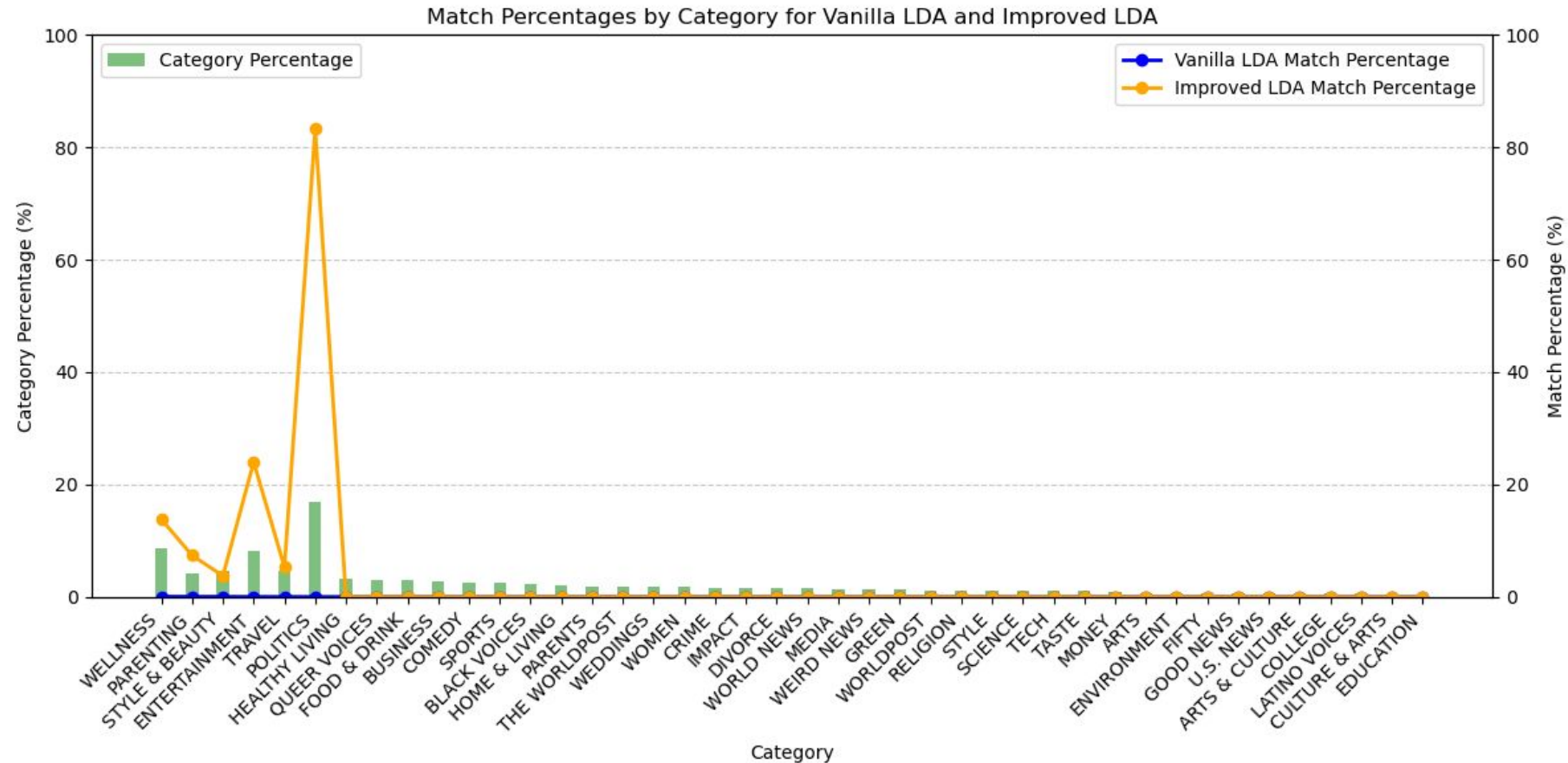
Slide to adjust relevance metric:⁽²⁾
 $\lambda = 1$ 0.0 0.2 0.4 0.6 0.8 1.0



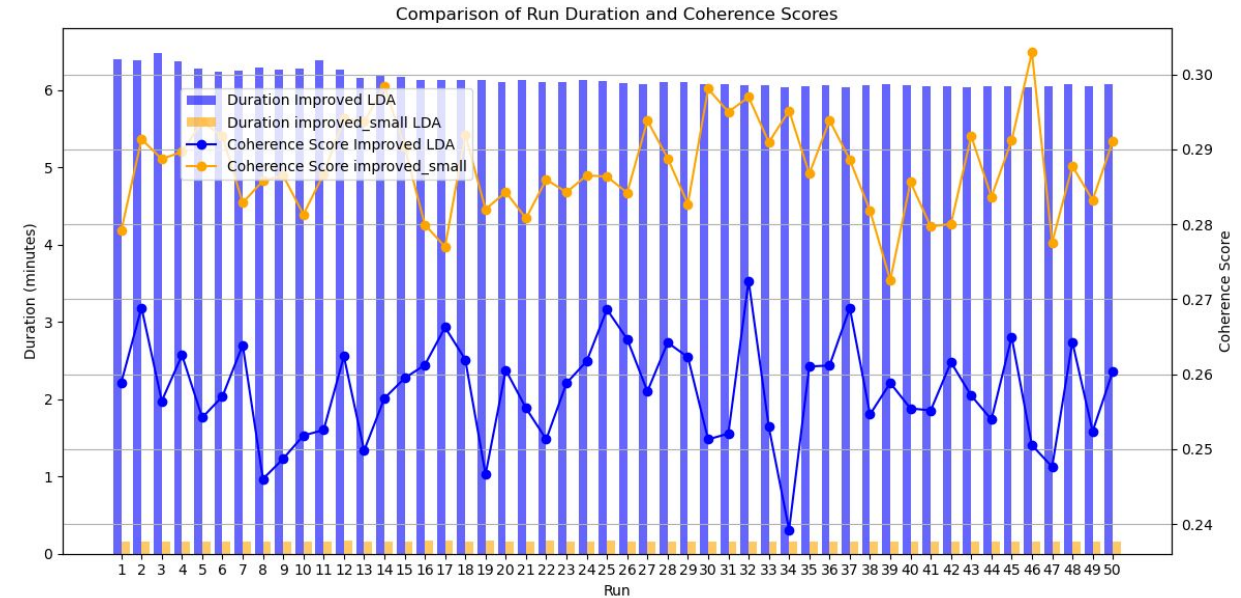
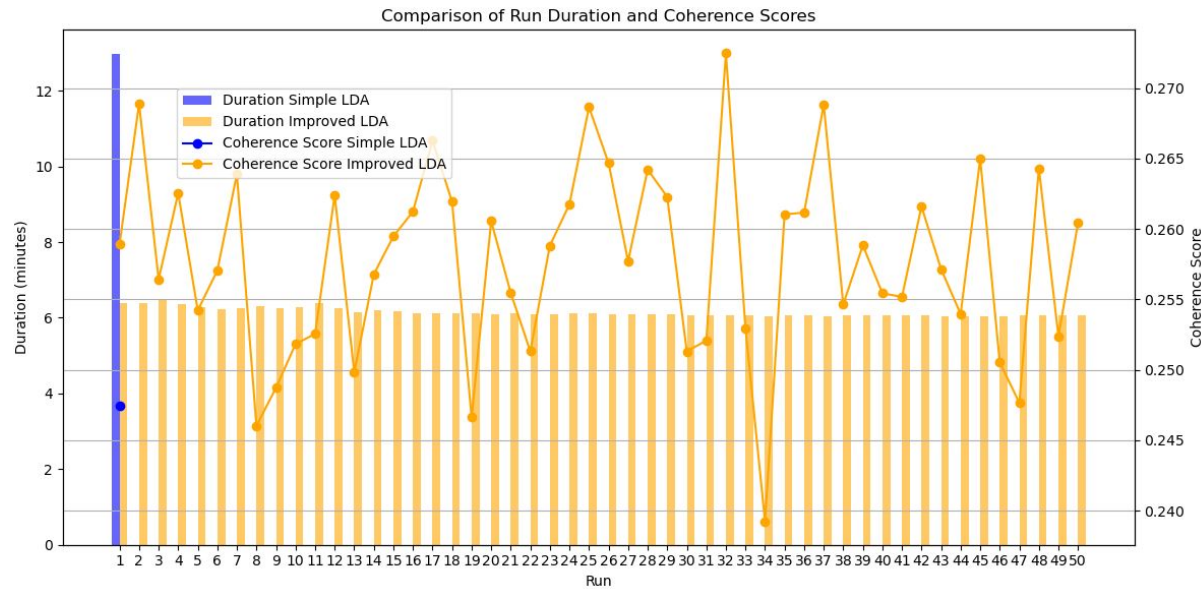
Konfidenz in den Kategorien



Vergleich der Modellperformance



Laufzeit und Kohärenz



Vergleich der Modelle



Herkömmliches LDA

- + einfache Implementierung
- + Kurzes Training mit einem kleinen Datensatz
- kann schnell falsch genutzt werden
- underfitting ist ein Problem

Verbessertes LDA

- + verbesserte Form
- + bessere Einsicht in die Daten
- + Mehr Implementierungsschritte
- Komplexer in der Implementierung
- lange Laufzeit