KI für Alle 2: Verstehen, Bewerten, Reflektieren

Themenblock Clustering: vom Sortieren bis zum Explorieren:
06_05Evaluation_TopicModeling

Evaluation und Interpretation Topic Models

Erarbeitet von
Dr. Katarina Boland

Lernziele

- Du kannst erklären, wie man die Interpretierbarkeit von Topics bestimmen kann

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

- Du kannst die Model Precision berechnen
- Du kannst das Prinzip des Topic Log Odds Maßes verstehen
- Du kannst verschiedene Aspekte von Topic Models aufzählen, deren Visualisierung für die Evaluation hilfreich ist
- Du kannst verschiedene Visualisierungen von Topic Models verstehen und interpretieren

Inhalt

Einstieg

Ähnlich wie beim Clustering gibt es auch für Topic Models keinen allgemeingültigen Standard bzgl. der Evaluationsmaße, die angewendet werden sollten. Und auch hier ist die Bewertung der Güte subjektiv und stark vom individuellen Anwendungsfall abhängig.

Es finden sich verschiedene Gütekriterien, die unterschiedliche Eigenschaften mehr oder weniger stark gewichten.

Wir beschränken uns hier auf eine Erklärung des Kriteriums der Interpretierbarkeit von erzeugten Topics oder Themenclustern.

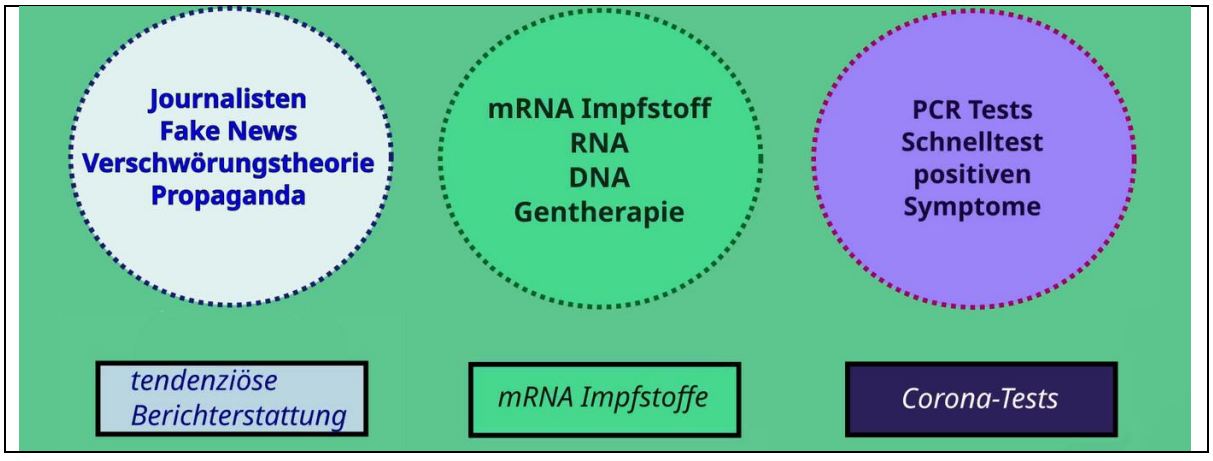
Interpretierbarkeit: Kohärenz und Diversität

Ähnlich wie beim Clustering sollten sich gute Topics oder Themencluster dadurch auszeichnen, dass es eine geringe Varianz innerhalb eines jeden Topics gibt, also die Wörter, Sätze oder Dokumente innerhalb eines Topics einander ähnlich sind.

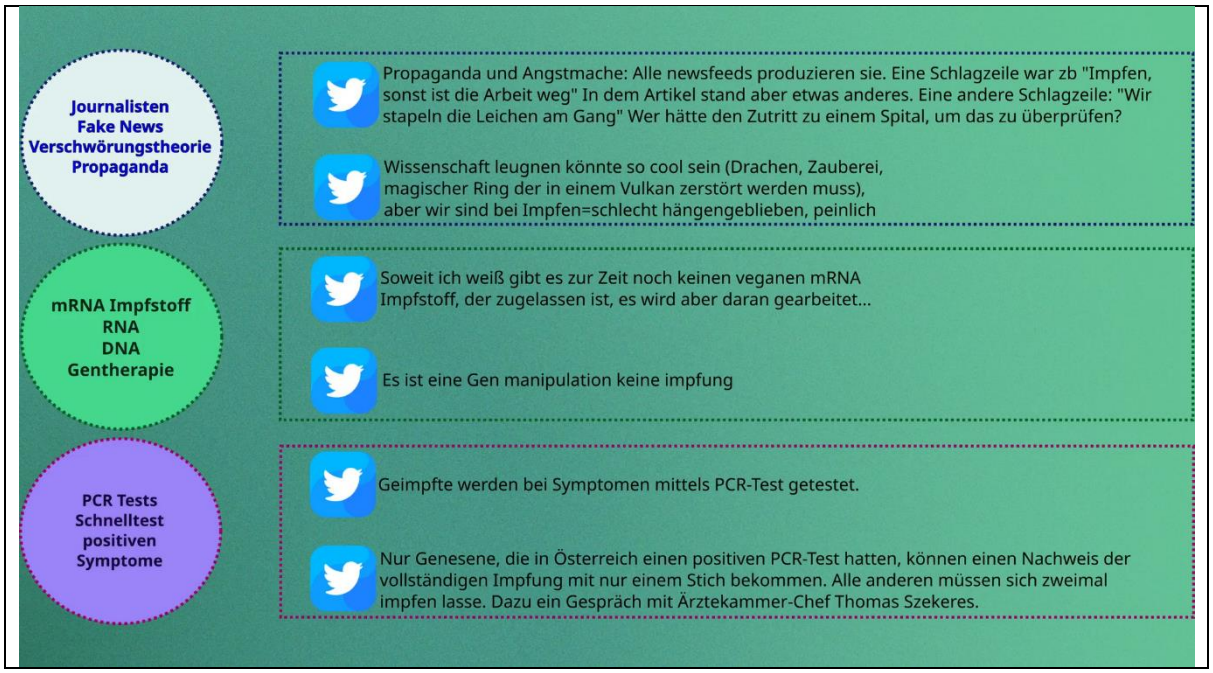
Gleichzeitig sollten die Topics zueinander trennscharf sein, also die Abstände zwischen ihnen groß und die enthaltenen Wörter, Sätze oder Dokumente eines Themenclusters möglichst unähnlich zu denen innerhalb jedes anderen Clusters sein. Wenn beides gegeben ist, sind die Topics interpretierbar, d. h. Menschen können ihre Bedeutung erkennen und voneinander unterscheiden.

Quelle [1]
Quelle [2]

Zur Veranschaulichung hier ein Ausschnitt der generierten Topics auf unserem Datensatz.



In diesem Bild siehst du die repräsentativen Wörter dreier generierter Themencluster. Die Wörter innerhalb eines Clusters sollten semantisch ähnlich genug sein, dass es dir leichtfallen sollte, ein Label dafür zu finden. Gelingt dir das, sind die Themen kohärent. Die Cluster sollten gleichzeitig divers sein. D. h., die Labels, die du findest, sollten sich voneinander unterscheiden. Es sollten nicht mehrere Cluster dasselbe Label bekommen können.



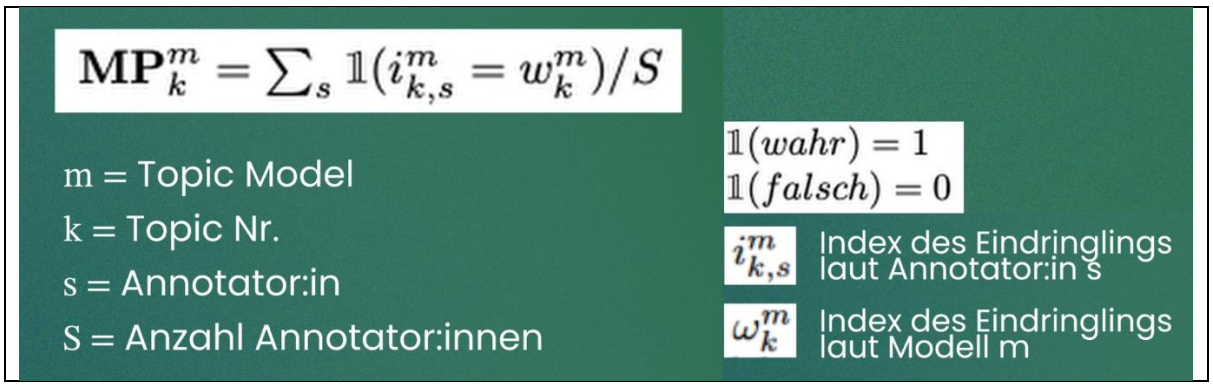
Dieses Bild zeigt die Zuordnung von Dokumenten (Tweets) zu Themenclustern. Hier sollte es leichtfallen, die Tweets den Topics zuzuordnen. Wenn ein Dokument in mehrere oder in keins der Topics passen sollte, spräche dies für eine geringe Diversität.

Wie aber lässt sich das messen? Wir stellen dir im Folgenden zwei Versuchsdesigns vor, um die Interpretierbarkeit von generierten Topics durch menschliche Annotator*innen zu bestimmen.

Word Intrusion

Annotator*innen bekommen eine Menge von Wörtern pro Topic. Dieses sind die repräsentativsten, wahrscheinlichsten Wörter für jedes der Topics, plus ein zusätzliches, das nicht dazu gehört: der „Eindringling” (englisch: intruder).

Die Aufgabe ist es, den Eindringling zu erkennen. Dieser wird zufällig aus einem Pool von Wörtern, die nur mit geringer Wahrscheinlichkeit zum Topic gehören, ausgewählt. Aus den Annotationen kann dann die sogenannte Model Precision errechnet werden:



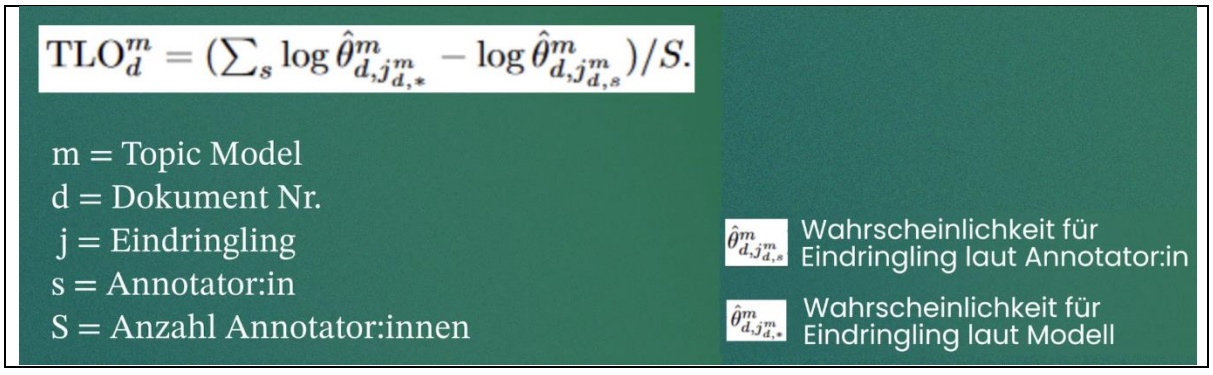
Quelle [1]

Sie ist definiert als der Anteil an Proband*innen, die mit dem Modell übereinstimmen.

Topic Intrusion

Hier bekommen die Annotator*innen ein Dokument, oder einen Ausschnitt daraus mit seinem Titel, wenn die Dokumente sehr lang sind. Zusätzlich bekommen sie eine Liste von Wörtern, die jeweils aus den drei wahrscheinlichsten und einem zufällig ausgewählten wenig wahrscheinlichem Topic stammen.

Das letzte, unwahrscheinliche Topic ist der Eindringling, der erkannt werden muss. Aus der Übereinstimmung der menschlichen Urteile und denen des Algorithmus lässt sich das sogenannte topic log odds Maß berechnen:



Quelle [1]

Je höher der topic log odds Wert, desto näher ist er an den menschlichen Urteilen und desto besser ist somit die erzeugte Zuordnung von Dokumenten zu Topics.

Hierbei ist zu beachten, dass die Anzahl von Topics diese Maße beeinflusst: Eine zu große Anzahl kann dazu führen, dass sich ähnliche Wörter, Sätze oder Dokumente in mehreren Topics wiederfinden. Ein Thema muss sozusagen über mehrere Cluster verteilt werden, um die gewünschte Anzahl an Themencluster zu erreichen.

Eine zu geringe Anzahl an Topics kann dazu führen, dass die Themencluster sehr heterogen sind, weil verschiedene Themen zusammengeworfen werden müssen.

Hieran siehst du, dass es sinnvoll ist, sich nicht nur auf eine Metrik zu verlassen, um die Güte von Topic Models zu bewerten, sondern ihre Güte am besten an der Kombination verschiedener Metriken abzulesen ist.

Visuelle Inspektion und „face validity"

Für einen Überblick über die generierten Topics und ihre Interpretierbarkeit sind Visualisierungen sehr hilfreich.

Quelle [2]

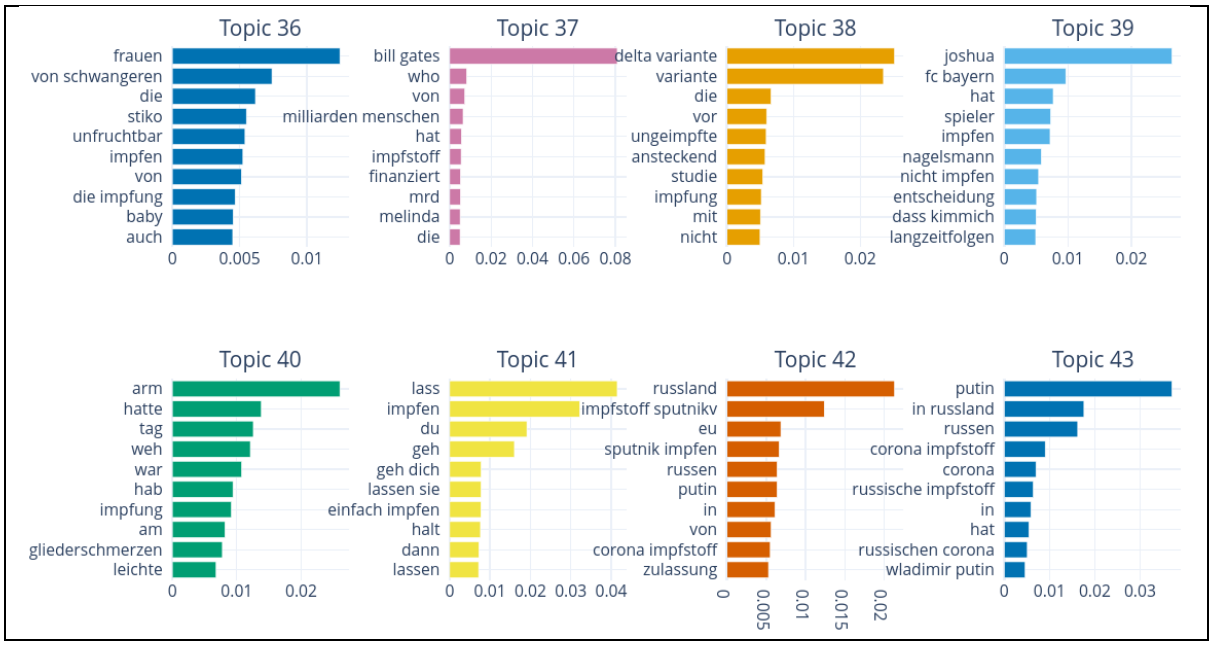
Im BERTopic-Framework sind viele Visualisierungen direkt integriert: beispielsweise „Barchart", ein Balkendiagramm für alle Topics und ihre repräsentativen Terme, eine „Intertopic Distance Map", die die Verteilung von Topics hinsichtlich verschiedener Dimensionen und ihre Abstände zueinander abbildet und ein Dendrogramm.

Quelle [3]

Wir schauen uns diese im Folgenden genauer an, du solltest aber auch mit den übrigen Visualisierungen experimentieren, wenn du mit BERTopic Models arbeitest.

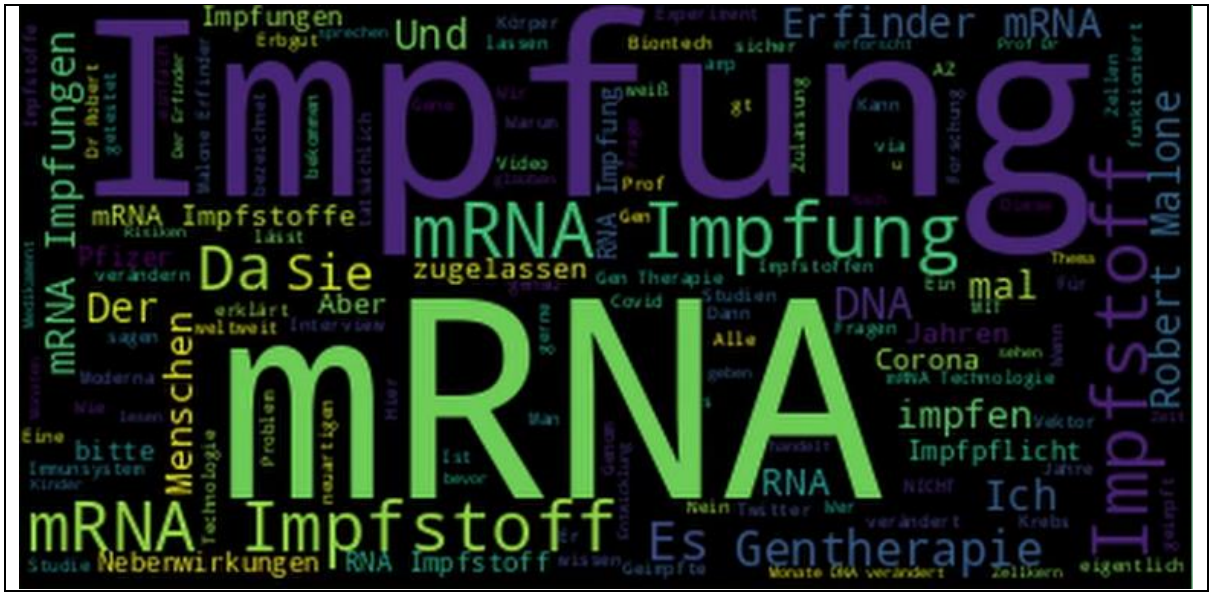
Barchart

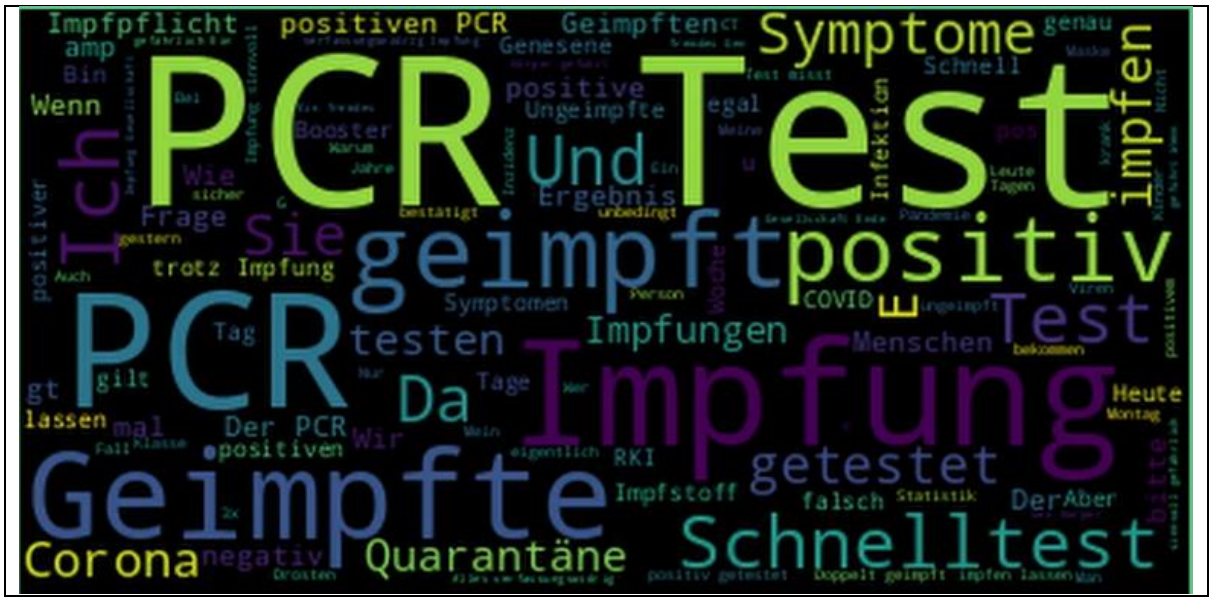
Hier ist ein Ausschnitt der Topics aus unseren Tweet-Daten. Dargestellt ist jeweils die Nummer des Topics, sowie die repräsentativen Begriffe (Unigramme und Bigramme, also einzelne Wörter oder Kombinationen aus zwei Wörtern) mit ihren Relevanzscores.



Aus der Darstellung können wir ableiten, dass Topic 36 von Impfung und Schwangerschaft handelt, 37 von Bill Gates, 38 von Corona Varianten, 39 von Fußballspielern und dem Fall Kimmich, 40 von Impfnebenwirkungen und 41 scheint Aufrufe zum Impfen zu beinhalten. 42 und 43 handeln beide vom russischen Impfstoff.

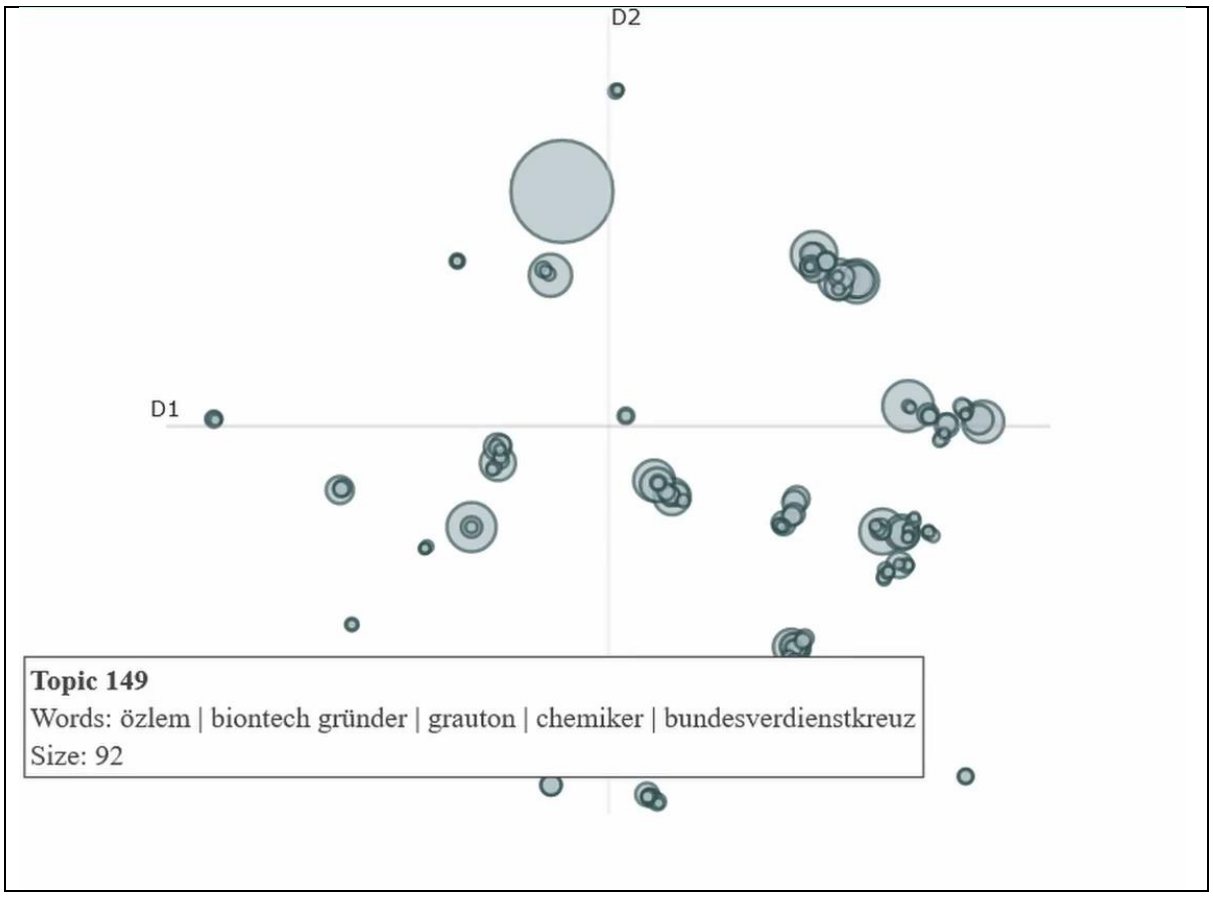
Alternativ oder zusätzlich kannst du auch Wordclouds für die Visualisierung der Themenclusters erstellen, beispielsweise mit der wordcloud Bibliothek von Python.





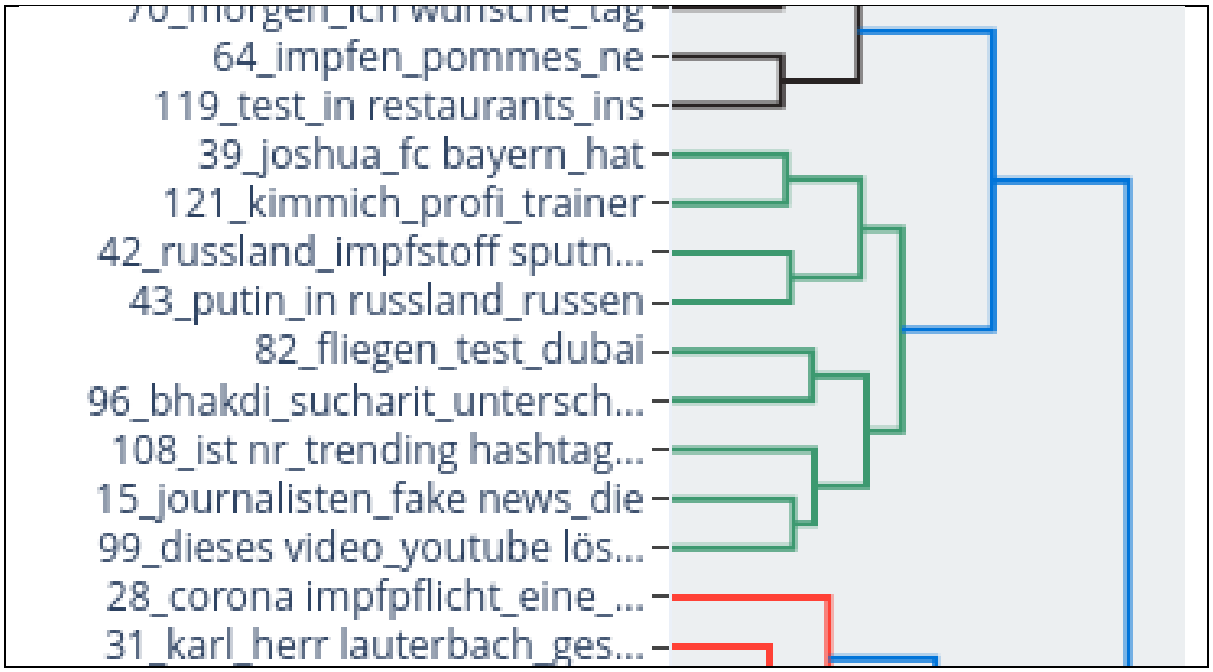
Intertopic Distance Map

In dieser Darstellung finden wir Cluster von ähnlichen Topics, die also nah beieinander sind. Hierfür wurden die Dimensionen auf zwei reduziert. Mit dieser Darstellung können wir uns außerdem die Größe der Topics, also die Anzahlen der enthaltenen Dokumente, anzeigen lassen.



Dendrogramm

Die Beziehung zwischen Topics lassen sich auch in einem Dendrogramm genauer untersuchen. Dieser Ausschnitt aus unseren generierten Topics zeigt, dass in der Tat die Topics 42 und 43 als ähnlich eingestuft werden. Bei einer reduzierten Anzahl an Topics würde der Algorithmus beide in einem Topic zusammenfassen.



Repräsentative Dokumente

In BERTopic sind das Clustering der Dokumente und die Selektion von repräsentativen Begriffen verschiedene Schritte, die auch getrennt evaluiert werden können. Erscheinen Topics nicht trennscharf, so könnte dies auch daran liegen, dass die repräsentativen Begriffe nicht gut gewählt sind. Vielleicht sind die Themencluster an sich aber durchaus sinnvoll.

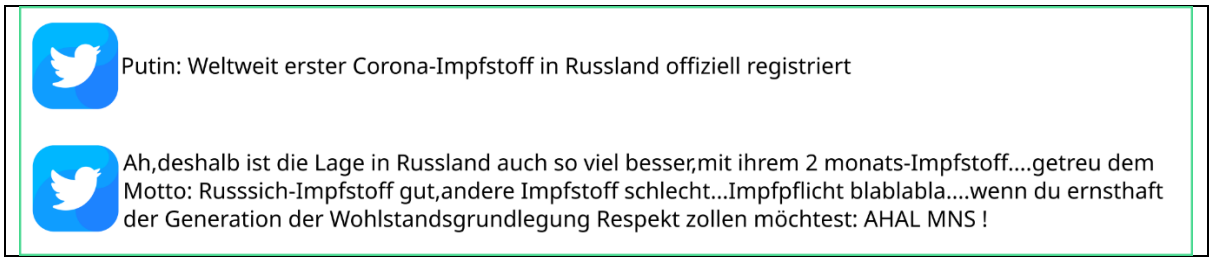
Um dies einschätzen zu können, kannst du dir alle, oder eine Auswahl repräsentativer Dokumente für jedes Themencluster anzeigen lassen.

Für unser Topic 42 finden wir u.a. folgende Dokumente:

Österreich will Corona-Impfstoff Sputnik V einsetzen

Slowakische Arzneimittelkontrolle kritisiert Hersteller von Sputnik V: Bratislava – In der Slowakei hat das staatliche Institut für Arzneimittelkontrolle SUKL einen kritischen Bericht über den russischen Impfstoff... #Slowakei #SputnikV #Kritik

Für Topic 43 diese:



Wie du siehst, sind beide Topics also durchaus unterschiedlich: während das eine primär von der Entwicklung des russischen Impfstoffs und der Situation in Russland spricht, behandelt das andere vor allem die Zulassung des russischen Impfstoffs und seinen Nutzen in der EU. Ob dieser Unterschied relevant ist oder ob lieber beide Topics zu einem zusammengeschmolzen werden sollten, hängt davon ab, was deine Untersuchungsfrage ist.

Abschluss

Die Evaluation von Topic Models ist komplex. Mit Hilfe von Visualisierungen lässt sich ein guter Überblick über generierte Topics bzw. Themencluster erreichen. Zusätzlich können mit Hilfe von quantitative Qualitätsmaße errechnet werden. Wir haben zwei dieser Maße kennengelernt, die sich auf Experimente mit manuell erzeugten Annotationen stützen und die du für die Evaluation verwenden kannst. Wie beim Clustering sind auch hier extrinsische Qualitätsmaße verwendbar, wenn du den Nutzen deiner Topics für eine Anwendung bemessen möchtest.

Quellen

Quelle [1] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. Proceedings of the 22nd International Conference on Neural Information Processing Systems, 288–296.

Quelle [2] Morstatter, F., & Liu, H. (2018). In Search of Coherence and Consensus: Measuring the Interpretability of Statistical Topics. Journal of Machine Learning Research, 18(169), 1–32. http://jmlr.org/papers/v18/17-069.html

Quelle [3] Grootendorst, Maarten. (o. J.). Https://maartengr.github.io/BERTopic/. Abgerufen 27. August 2024, von https://maartengr.github.io/BERTopic/

Weiterführendes Material

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112, 102131. https://doi.org/10.1016/j.is.2022.102131

Boland, K., Starke, C., Bensmann, F., Marcinkowski, F., & Dietze, S. (2024). A computational analysis of German online discourses about COVID-19 vaccinations to inform policy-making in times of crisis (Preprint). https://doi.org/10.2196/preprints.63909

Disclaimer

Transkript zu dem Video „06 Clustering: vom Sortieren bis zum Explorieren: Evaluation und Interpretation Topic Models“, Katarina Boland.