Exposé
Max Prantz 11096202

Th Köln 9.11.2022

Inhaltsverzeichniss

Problemstellung und Motivation

Im Rahmen der Literaturdatenbank der ZB-Med[1] werden Publikationen bereitgestellt um Forschenden einen schnellen und informationsreichen Einblick in verschiedenste Themen bereit zu stellen. Um die gespeicherte Literatur bei einer Suche schnell und nach Relevanz gut einschätzbar darstellen zu können werden die Texte bzw. die Abstracts und Überschriften der Publikationen klassifiziert. So werden die Dokumente momentan mit einer Software eines Drittanbieters in mehrere übergeordneten Themen unterteilt und mit einem Symbol versehen. Das soll Suchenden dabei helfen schnell zu erkennen, ob das gefundene Dokument grob in das gewünschte Themengebiet fällt. Die vier Hauptkategorien sind, Medizin, Ernährung, Umweltwissenschaften und Landwirtschaft.
Da die momentan genutzte Software keine Möglichkeit bietet in den Prozess der Klassifizierung Einblick zu erhalten und somit eine „Blackbox" ist, welche nicht einmal fein eingestellt werden kann, soll Abhilfe mit einer eigenen Pipeline zur Klassifizierung der Dokumente entstehen. Das Ziel ist die neue Software weiter entwickeln und modifizieren zu können, ohne auf dritte Anbieter zurück greifen zu müssen. Auch soll die neue Software mit den alten Systemen genauso zusammenspielen, wie die vorherige, um weitere umfassendere Modifikationen an der bestehenden Suchmaschine zu vermeiden.

Methodik

Die Umsetzung des Projekts beinhaltet mehrere Lösungswege, bei jedem dieser ist unklar wie gut das Gewünschte Ergebnis erzielt wird, auch ist es möglich mehrere dieser zu Kombinieren. Der Endgültige Aufbau der Pipeline ist somit noch nicht ente gültig festgelegt und wird ggfs. während der Umsetzung mehrfach verändert.
Als „Golden Record" bzw. benchmark wird die schon bestehende Klassifizierung genutzt. Da diese Klassifizierung nicht perfekt ist, ist es das Ziel mindestens genauso gut oder besser Dokumente klassifizieren zu können als die vorherige Software.
Um dieses Ziel zu erreichen werden, nach dem Laden und Bauen eines Trainingsdatensets aus der Datenbank der ZB-Med, mehrere Methodiken verwendet.
Aus dem entstandenen Text-Corpus der Datenbank werden Titel und Abstracts in einem „Word to Vector Space" verglichen um einen Trainings-, Test- und Validierungsdatensatz zu erstellen. Je nach Umfang der erstellten Sets werden diese gekürzt werden müssen um mit den Rechenzeiten des Modell-trainings nicht in Zeitliche Schwierigkeiten zu kommen.
Mithilfe der erstellten Sets können dann verschiedene Maschine-learning-Modelle trainiert werden. So können „Decision-Tree-Models[2]", „K-Nearest-Neighbour-Algorythms[3]" oder Transformer-Modelle[4] trainiert werden, diese werden iterativ verbessert bis Entweder das erwünschte Ergebnis erreicht ist, oder aber klar wird, dass die verwendete Methodik zu keinem zufriedenstellenden Ergebnis führt. Validiert werden die Ergebnisse der Modelle mit dem zuvor erhaltenem „Golden Record" der momentanen Klassifizierungsmethode.
Um einen besseren Überblick über die vorhandenen Publikationen und deren Themengebiete erhalten zu können werden diese zusätzlich noch mithilfe von TF-IDF[5] und

---

[1] https://www.zbmed.de letzter zugriff 07.11.2022
[2] https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575 Zuletzt besucht am 07.11.2022
[3] https://www.ibm.com/de-de/topics/knn Zuletzt besucht am 07.11.2022
[4] https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/ Zuletzt besucht am 07.11.2022
[5] http://www.tfidf.com Zuletzt besucht am 07.11.2022

LDA-Methodiken[6] untersucht. Auch diese beiden Ansätze sollen mit dem „Golden Record“ verglichen werden.

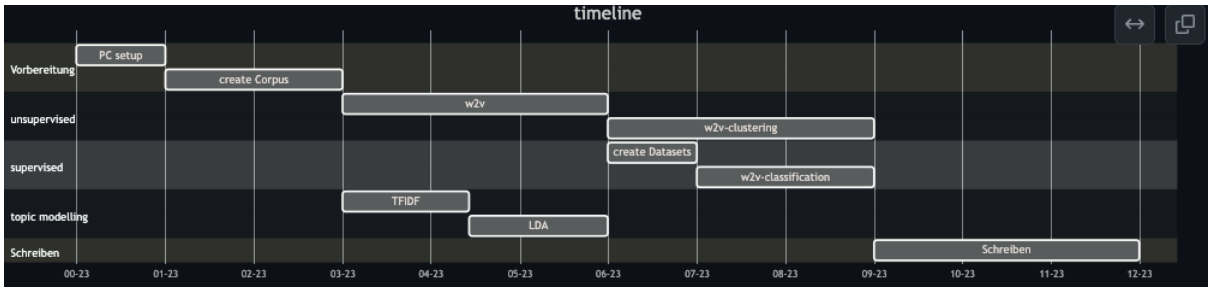
Aufbau und Inhalt der Arbeit

Zu Beginn wird das Thema, Zielsetzung und Aufbau der Bachelorarbeit vorgestellt. Im Anschluss werden die theoretischen Grundlagen und der Momentane „Ist-Zustand“ des Systems diskutiert und veranschaulicht.
Anschließend wird gezeigt nach welchen Kriterien die Klassifizierung der Texte erfolgen soll und wie die dann in der Suchmaschine dargestellt werden.
Aufbauend darauf wird darauf eingegangen, wie der Textkorpus erstellt wurde und mit welchen Methoden des Textmining bzw. der Textklassifizierung die Publikationen den Überthemen zugeordnet werden.
Durch die iterative Entwicklung des Klassifizierungsmodells werden anschließend die Ergebnisse der verschiedenen Entwicklungsstufen diskutiert und schlussendlich das Endergebnis präsentiert. In diesem Schritt wird auch darauf eingegangen in wie weit sich das alte und neue Klassifizierungsmodell in der Genauigkeit ihrer Ergebnisse unterscheiden.

[6] https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd Zuletzt besucht am 07.11.2022

Gliederung

Zeitplan


timeline
Vorbereitung
unsupervised
supervised
topic modelling
Schreiben
PC setup
create Corpus
w2v
w2v-clustering
create Datasets
w2v-classification
TFIDF
LDA
Schreiben
00-23
01-23
02-23
03-23
04-23
05-23
06-23
07-23
08-23
09-23
10-23
11-23
12-23