text2ddc meets Literature - Ein Verfahren für die Analyse und Visualisierung thematischer Makrostrukturen

Preprint

Alexander Mehler
Goethe-Universität Frankfurt
mehler@em.uni-frankfurt.de

Tolga Uslu
Goethe-Universität Frankfurt
uslu@em.uni-frankfurt.de

Rüdiger Gleim
Goethe-Universität Frankfurt
gleim@em.uni-frankfurt.de

Daniel Baumartz
Goethe-Universität Frankfurt
baumartz@stud.uni-frankfurt.de

April 3, 2019

***K*eywords** Text Klassifikation · Dewey Decimal Classification · DDC · Wikipedia

Abstract

In diesem Poster geht es um die thematische Analyse und Visualisierung literarischer Werke mithilfe automatisierter Klassifikationsalgorithmen. Hierfür wird ein bereits entwickelter Algorithmus namens text2ddc [3, 1] verwendet, um die Themenverteilungen literarischer Werke zu identifizieren. Darüber hinaus thematisiert der Beitrag, wie diese Verteilungen von Themen und deren Abhängigkeiten untereinander visualisiert werden können.

Bei *text2ddc* handelt es sich um einen Klassifikator auf Basis *neuronaler Netze*, der Texte einer bestimmten Anzahl von Sprachen nach der *Dewey-Dezimalklassifikation* (DDC) kategorisiert. Die DDC ist ein internationaler Standard für die Themenklassifikation im Bereich von (digitalen) Bibliotheken. Um text2ddc zu trainieren, wurde die *Wikipedia* verwendet. Da viele Artikel der Wikipedia mit der *Gemeinsamen Normdatei* (GND) verlinkt sind und die GND Informationen zu den entsprechenden DDC-Kategorien hinterlegt, war es möglich, ein vergleichsweise großes und zugleich breites DDC-orientiertes Trainingskorpora für das Deutsche aufzubauen. Am Beispiel dieses Korpus erreicht unser Algorithmus einen F-Score von 87,4%. Da die Artikel der Wikipedia auch über Sprachgrenzen hinweg untereinander verlinkt sind, war es zudem möglich, text2ddc für über 40 Sprachen zu trainieren.

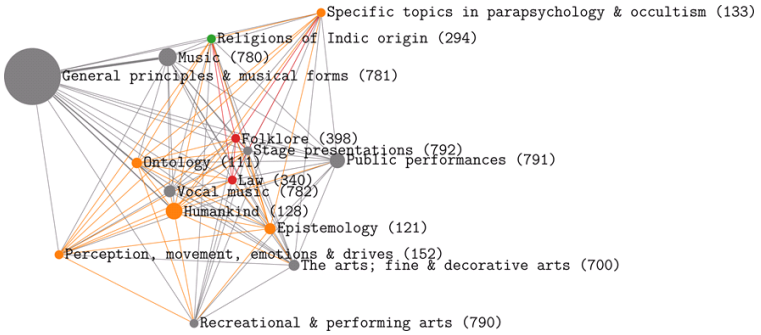

Figure 1: Friedrich Nietzsche (Die Geburt der Tragödie aus dem Geiste der Musik).

TEXT2DDC wurde auf Korpora verschiedener Genres angewandt, um deren Themenverteilungen zu analysieren. Zum einen betrifft dies die Wikipedia selbst, aber auch Korpora basierend auf StadtWikis, anhand derer bestimmt wurde, welche Themen dominant sind und wie diese zusammenhängen. Ein drittes Beispiel betrifft literarische Texte bzw. historische Texte der Wissenschaft. Abbildung 1 zeigt etwa die Themenverteilung von Die Geburt der Tragödie aus dem Geiste der Musik von Friedrich Nietzsche. In dieser Abbildung repräsentieren die Knoten die DDC-Kategorien, wobei die Knotenfarbe dazu dient, die jeweilige DDC-Hauptkategorie zu identifizieren. Kanten zwischen den Knoten repräsentieren den Zusammenhang der jeweiligen Themen. Hierfür wurde die semantische Ähnlichkeit von Sektionen, Paragraphen und Sätzen ausgewertet. Ein alternatives Beispiel bildet Massenpsychologie und Ich-Analyse von Sigmund Freud (siehe Abbildung 2). Die Beispiele verdeutlichen nicht nur die erwarteten Unterschiede beider Werke, sondern zeigen zugleich makrostrukturelle thematische Zusammenhänge auf.

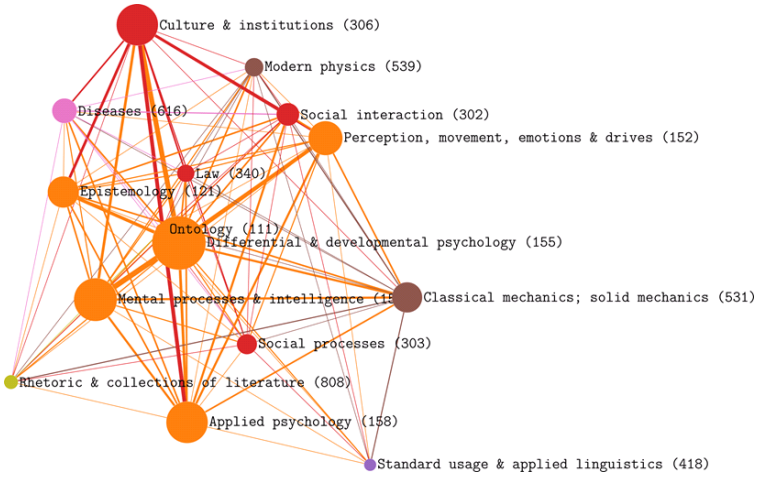

Figure 2: Sigmund Freud (Massenpsychologie und Ich-Analyse).

Der zugrundeliegende Algorithmus basiert auf folgendem Prozedere: Zunächst wird der Inputtext in Sektionen untergliedert, wofür die jeweilige logische Dokumentstruktur ausgewertet wird. Anschließend werden verschiedene NLP-Methoden angewendet, um Informationen über Lemmata und Named Entities zu gewinnen, was wiederum auf einer automatischen Disambiguierung basiert. Mittels dieser Methoden erreichen wir eine höhere Genauigkeit bei der Klassifikation mit TEXT2DDC. Im nächsten Schritt werden die Sektionen unter Verwendung der DDC als Zielklassifikation kategorisiert. Je mehr Sektionen auf dasselbe DDC-Thema abgebildet werden, desto höher ist das Gewicht des entsprechenden Zielknotens und desto größer kann dieser dargestellt werden. Da bei dieser Vorgehensweise nicht auf Linkstrukturen zurückgegriffen werden kann, erfolgt die Induktion von Themenkanten nach der inhaltlichen Ähnlichkeit der den Themen zugeordneten Sektionen. Hierfür werden Texteinbettungsalgorithmen aus dem Bereich neuronaler Netze angewandt.

Das Poster untersucht anhand der Werke einer Reihe von deutschsprachigen Autoren (u.a. Karl Marx, Sigmund Freud, Franz Kafka, Friedrich Nietzsche, Thomas Mann und Martin Heidegger) die Möglichkeiten und Grenzen von Themenkarten zur Erfassung makrostruktureller Themenzusammenhänge von Texten, wie sie unser Algorithmus erfasst. Auf diese Weise soll eine Alternative zu den in den DH omnipräsenten *topic models* aufgezeigt werden. Zu diesem Zweck experimentiert der Beitrag mit alternativen Visualisierungstechniken basierend auf interaktiven konzentrischen Netzwerken (POLYVIZ [2]) und alternativ basierend auf klassischen Netzwerkdarstellungen.

References

[1] Daniel Baumartz, Tolga Uslu, and Alexander Mehler. LTV: Labeled topic vector. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics: System Demonstrations, August 20-26*, Santa Fe, New Mexico, USA, 2018. The COLING 2018 Organizing Committee.

[2] Tolga Uslu and Alexander Mehler. PolyViz: a visualization system for a special kind of multipartite graphs. In *Proceedings of the IEEE VIS 2018*, IEEE VIS 2018, 2018.

[3] Tolga Uslu, Alexander Mehler, Andreas Niekler, and Daniel Baumartz. Towards a DDC-based topic network model of wikipedia. In *Proceedings of 2nd International Workshop on Modeling, Analysis, and Management of Social Networks and their Applications (SOCNET 2018), February 28, 2018*, 2018.