Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

IDP Institut für Datenanalyse
und Prozessdesign

Topic-Modeling einer deutschsprachigen Textsammlung

Diplomand
Lukas Baumgartner

Dozent
Oliver Dürr

Die Entwicklung der digitalen Medien nimmt stark zu. Dementsprechend wächst die Menge an Texten ständig. Topic-Modeling dient zur Analyse dieses grossen Kontingentes. Diese Bachelorarbeit führt ein Topic-Modeling von deutschen Texten aus und visualisiert den Zusammenhang der Texte. Dazu verwendet sie die Methode latent Dirichlet allocation (LDA). Die LDA-methode sucht in einem Corpus (Textsammlung) nach P übergeordneten Themen (Topics). Ihr Output ist eine Wahrscheinlichkeitsmatrix (Posterior), die jeden Text durch einen P-dimensionalen Vektor den Topics zuordnet. Abschliessend kann der mehrdimensionale Posterior grafisch dargestellt werden. Ziel dieser Darstellung (Visualisierung) ist die Reduktion der Dimensionen.

Die 911 verwendeten Texte sind deutschsprachige Stiftungszwecke, die von der Firma NonproCons zur Verfügung gestellt sind. Das Ziel der Arbeit liegt in der zweidimensionalen Darstellung des mehrdimensionalen Posteriors der Textsammlung. Texte mit ähnlichem Inhalt sollen räumlich nahe angeordnet werden. Die Methode Principal Component Analysis (PCA) und der Layout Algorithmus Fruchterman-Reingold sind für diese Visualisierung vorhanden. Durch das Ausüben des Text-Mining-Verfahrens verringert sich der Corpus schrittweise auf die wichtigsten Wörter. Der Corpus und die Wahl der P Topics sind der Input der LDA-Methode. Der Output ist der Posterior.

Der Layout Algorithmus Fruchterman-Reingold ist als ein netzwerkbasiertes Layout mächtiger als die PCA. Acht Topics reichen zur Darstellung der Stiftungszwecke. Erfasst sind die Themen folgendermassen: umweltbezogene Ausbildung, soziale Ausbildung, soziokulturelle Ausbildung, musische Ausbildung, akademische Ausbildung, obligatorische Schulbildung, Entwicklungshilfe und Erstausbildung. Längere Texte haben einen stärkeren Einfluss auf die Topics als kürzere. Daher stellt sich die Frage, ob der Umfang der Texte beschränkt werden sollte.

Das Netzwerk visualisiert die Zusammenhänge von Stiftungszwecken mittels Fruchterman-Reingold Layout