Erstellung und Visualisierung von Topic-Modellen in WebLicht

,
marie.hinrichs@uni-tuebingen.de
Universität Tübingen, Deutschland

,
cagri.coeltekin@uni-tuebingen.de
Universität Tübingen, Deutschland

Die neueste Erweiterung von besteht aus einer Funktion zur Generierung von Topic-Modellen auf Basis von Nutzerinput. Auch die Visualisierung und Analyse des generierten Modells sind mit WebLicht möglich.

WebLicht (Web-Based Linguistic Chaining Tool) stellt eine virtuelle Forschungsumgebung zur Verfügung, in der Nutzer_innen Verarbeitungsketten zur linguistischen Annotation erstellen und ausführen können und die generierten Annotationen in der Folge visualisieren können. Die WebLicht Webapplikation ist das Frontend eines breiten Frameworks, das darauf abzielt, verbreitete state-of-the-art Annotationswerkzeuge den Forschenden gut zugänglich zur Verfügung zu stellen, ohne dass einzelne Anwendungen heruntergeladen oder installiert werden müssen. Die in WebLicht verfügbaren Werkzeuge werden von den CLARIN Zentren entwickelt und als Webservices zur Verfügung gestellt. Neue Werkzeuge können jederzeit integriert werden, sofern einige Richtlinien befolgt werden. Im Kern muss ein neues Werkzeug hierzu in der CLARIN Center Registry beschrieben werden, als Webservice zur Verfügung gestellt werden und gut definierte Input- und Outputformate benutzen.

Topic-Modelle sind statistische Modelle, die ein Inputdokument in ein Set abstrakter Themen kategorisieren und mit verschiedenen Gewichten oder Prioritäten versehen. Topic-Modelle werden typischerweise automatisch aus einem Set von Dokumenten abgeleitet, ohne dass eine manuelle Annotation notwendig ist. Die resultierenden Modelle können in verschiedensten Aufgaben im Bereich des Natural Language Processing genutzt werden, z. B. zum automatischen Klassifizieren von Dokumenten oder zur Bestimmung verschiedener Wortbedeutungen. Alternativ können sie auch als “Data Mining” Tool genutzt werden, vor allem in Kombination mit passenden Visualisierungen. Auch in den digitalen Geisteswissenschaften ist Topic-Modellierung eine gängige Methode. Insbesondere kann Topic-Modellierung dabei helfen, Muster in großen Textkollektionen zu erkennen. Die Nutzung von Topic-Modellierung in den digitalen Geisteswissenschaften wird unter anderem beschrieben in Jockers (2010) Arbeiten zum Klassifizieren von Blogs beim ‘Day of DH 2010’, in Drouin (2011) Arbeiten zu Proust, in Griffiths und Steyvers' (2004) Arbeiten zu Themen in der Wissenschaft im Verlauf der Zeit und in einer weiteren diachronen Studie von Riddell (2012) zum Thema Topics in der Germanistik.

Da es eines der vorrangigen Ziele von WebLicht ist, Natural Language Processing Tools Geisteswissenschaftlern gut zugänglich zur Verfügung zu stellen, haben wir einen Webservice zur Erstellung und Visualisierung von Topic-Modellen in die WebLicht-Umgebung eingefügt. Das aktuelle Modell nutzt ein Topic-Modell, das im KobRA Projekt entwickelt wurde und auf der weithin bekannte Latent Dirichlet Allocation (LDA) Technik wie von Blei et al. (2003) beschrieben, basiert ist. Das resultierende Topic-Modell wird mit der weit verbreiteten Visualisierungssoftware DFR-browser (cf. Goldstone 2013-2015) visualisiert. DFR-browser bietet vielfältige Visualisierungen, unter anderem von Listen der “Topwörter” zu jedem Topic und von Topic-übergreifenden Worträngen.

Bei der Topic-Modellierung sieht ein üblicher Ablauf wie folgt aus: Der Nutzer lädt eine Textsammlung zu WebLicht hinauf; WebLicht berechnet mit dem oben genannten Webservice ein Topic-Modell mit einer vorgegebenen Anzahl an Topics; die resultierenden Topic-Wort- und Topic-Dokument-Verteilungen werden in das vom DFR-browser benötigte Format konvertiert und im Webbrowser des Nutzers visualisiert. Abbildung 1 enthält eine der Visualisierungsansichten und zeigt die am höchsten eingestuften Wörter für sechs abstrakte Topics in einem Topic-Modell berechnet auf Basis eines großen Zeitungskorpus. Die identifizierten Topics korrespondieren grob mit den Themenfeldern Kultur, Finanzen, Reisen, Politik und Familie. Solch ein Modell kann beispielsweise verwendet werden um Artikel aus einem bestimmten Themenfeld auszuwählen bevor weitere automatische oder manuelle Analysen erfolgen.

Der vorgestellte Webservice arbeitet zur Zeit mit mehreren Dokumenten, die als einzelne Textdatei ohne Metadaten formatiert sein müssen. Der Webservice erlaubt es zur Zeit die gewünschte Anzahl von Themen einzustellen, sowie die häufigsten (zum Beispiel Funktionswörter) oder seltensten (zur Vermeidung von statistischen Störeinflüssen) Wörter herauszufiltern. Die Ergebnisse werden als statische HTML Seite, welche für begrenzte Zeit auf dem Server gespeichert wird, dargestellt. Es ist geplant, zukünftig auch die Metadaten der Dokumente (Autor, Veröffentlichungsdatum usw.) zu verwenden, sowie die Visualisierungen weiter zu verbessern. Abschließend möchten wir bemerken, dass die Integration von Topic-Modellierung in die WebLicht-Umgebung eine Anzahl von Möglichkeiten für die Verarbeitung schafft, wie z.B. das Erstellen von Modellen, die Annotationen (z. B. Part-of-Speech Tags oder syntaktische Relationen) der WebLicht Umgebung nutzen.

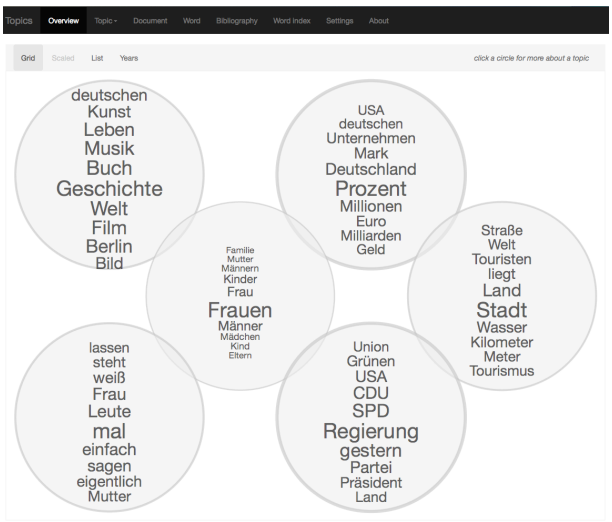

***Abb. 1**: Beispiel Visualisierung mit DFR-browser.*

Bibliographie

Blei, David M. / Ng, Andrew Y. / Jordan, Michael I. (2003): "Latent dirichlet allocation", in: *The Journal of machine Learning research* 3: 993-1022.

CLARIN-D / SfS-Uni. Tübingen (2012): *WebLicht*. Web-Based Linguistic Chaining Tool https://weblicht.sfs.uni-tuebingen.de [letzter Zugriff 11. Februar 2016].

Drouin, Jeff (2011): "Foray Into Topic Modeling", in: *Ecclesiastical Proust Archive* http://www.proustarchive.org/?q=node/35 [letzter Zugriff 11. Februar 2016].

Goldstone, Andrew (2013-2015): *DFR-browser* http://agoldst.github.io/dfr-browser [letzter Zugriff 11. Februar 2016].

Grif#ths, Thomas / Steyvers, Mark (2004): "Finding scientific topics", in: *Proceedings of the National Academy of Sciences* 101 (Suppl 1): 5228–5235 [letzter Zugriff 11. Februar 2016].

Jockers, Matthew (2010): "Who's your DH Blog Mate: Match-Making the Day of DH Bloggers with Topic Modeling", in: Matthew Jockers: *Homepage* http://www.matthewjockers.net/2010/03/19/whos-your-dh-blog-mate-match-making-the-day-of-dh-bloggers-with-topic-modeling [letzter Zugriff 11. Februar 2016].

Storrer, Angelika et al. (2012-2015): *KobRA*. Korpus-basierte Recherche und Analyse mit Hilfe von Data-Mining http://kobra.tu-dortmund.de/mediawiki/ [letzter Zugriff 11. Februar 2016].