Tobias Gumpp, London und Marc Schneider, München*

Online-Appendix zu „Informatische Methoden in der Rechtswissenschaft – ein interdisziplinärer Beitrag zur Theorie der Gesetzgebung“

I. Zur Benutzung der frei zugänglichen Daten und des Quellcodes

Der Code wurde auf Python version 3.7 geschrieben und die notwendigen Pakete können mittels pip installiert werden. Der Datensatz wurde hier heruntergeladen: https://zenodo.org/record/3831112.

II. Zum Datensatz und zur Datensatzbereinigung

Der für alle Analysen verwendete Datensatz[1] stammt aus einer kompilierten Version des (nach dortigen Auskünften „nahezu“) gesamten deutschen Bundesrechts von der Website *gesetze-im-internet* des BMJV.[2] Da für die hiesige Analyse Verordnungen keine Relevanz aufweisen, da sie für eine Kodifikation zu speziell und kleinteilig sind, haben wir den Datensatz auf Gesetze reduziert, indem wir nur Dateien mit der Endung G, O, oder B als Daten verwendet haben (Gesetz, Ordnung oder Buch). Durch diese und weitere[3] Einschränkungen wurde der Datensatz von 6487 Dateien auf 1663 Gesetzestexte reduziert.

Ferner wurden die Daten durch in der *data science* übliche Schritte weiter aufbereitet.[4] Insbesondere wurden (für das *Topic Model*) sogenannte *Stopwörter*[5] wie Artikel, Konjunktionen und Präpositionen entfernt, da diese nicht zur Identifikation des Themas beitragen und aufgrund ihrer Vielzahl die Komplexität unnötig erhöhen.[6] Für die Verweiserkennung und die Erstellung des Netzwerks wurde automatisiert aus dem Gesetzestitel ein Lang-, ein Kurztitel und eine Abkürzung ausgelesen (wie etwa: „Gesetz betreffend die Erwerbs- und Wirtschaftsgenossenschaften“ als Lang- und „Genossenschaftsgesetz“ als Kurztitel,

---

* Jurist (Univ.) *Tobias Gumpp* ist M.Sc.-Student an der LSE, London; *Marc Schneider*, B.Sc. (TopMath), B.Ed. ist M.Sc.- und M.Ed.-Student an der TU München.

[1] Vgl. allgemein zur juristischen Datensammlung *Coupette/Fleckner*, JZ 2018, 379 (380–383).

[2] Abrufbar unter https://zenodo.org/record/3934926. Der Stand der benutzten Daten ist 18. Mai 2020. Zum Datensatz von *gesetze-im-internet* siehe die Erklärung unter https://www.gesetze-im-internet.de.

[3] Es wurden außerdem Verordnungen aus dem Datensatz entfernt, deren Kürzel auf „VO“, „VollzO“, „BO“ oder „AnO“ enden, die mit „VO“ beginnen oder die Zeichenfolgen „_DV-„ enthalten. Desweiteren wurden die Namen untersucht und Texte mit „Verordnung“ innerhalb des Namens ignoriert, falls der Name nicht die Zeichenfolge „Gesetz“ enthält.

[4] Einführung mit Herkunftserläuterung (wenn auch bezogen auf überwachtes maschinelles Lernen, was jedoch i.E. keinen Unterschied macht) bei *Hvitfeldt/Silge*, Supervised Machine Learning for Text Analysis in R, sub 3 (bzgl. Stopwörtern).

[5] Hierzu wurden die deutschen Stopwörter der Bibliothek des Python Natural Language Toolkits (https://www.nltk.org/) verwendet.

[6] Für LDA ebenso *Blei*, Commun. ACM 2012, 77 (78).

„GenG“ als Abkürzung). Da vornehmlich explizite externe Verweise[7] für die vorliegende Forschungsfrage maßgeblich sind, somit also nur Verweise auf andere Gesetze unter Gesetzesnennung erfasst werden sollen, konnte so nicht nur das Bestehen eines Verweises, sondern auch dessen Verweisziel identifiziert werden. Da indes Gesetze typischerweise im Genitiv zitiert werden, war darüberhinaus das sog. *stemming*[8] des Textes notwenidg, d.h. die Wortstammanalyse, sodass nicht nur das „Bürgerliche Gesetzbuch“ sondern auch „des Bürgerlichen Gesetzbuchs“ erkannt werden kann. Dies geschah mittels der von *Weißweiler/Fraser* bereitgestellten Bibliothek *cistem* zur Bildung deutscher Wortstämme.[9]

III. Auswahlmethodik der Textsammlung Beck-Texte im dtv zum Gesellschaftsrecht für Abbildung 1

Es wurde das Inhaltsverzeichnis (S. V-VII) der Textsammlung „Gesellschaftsrecht: GesR“ der Beck-Texte im dtv, 17. Aufl. 2020, genutzt. Die dortigen aufgeführten Texte wurden dann dadurch reduziert, dass alle Nichtgesetze, alle nichtbundesdeutschen Texte wie etwa europäische Verordnungen, sowie alle Gesetze, die bloß im Auszug abgedruckt sind, aussortiert wurden. Dies ist darin begründet, dass sich unser Datensatz auf bundesdeutsche Gesetze beschränkt; Auszugsgesetze sollten ausgeschieden werden, um nur die Kerntexte zu beinhalten (was jedoch auch das BGB und HGB ausscheiden ließ).

[7] Für eine Taxonomie der Verweisarten siehe aus informatischer Sicht *Waltl u. a.*, Jusletter IT Flash 2017, Rz. 5 ff. Nach dieser Einordnung beziehen wir uns auf *fully-explicit references* (jedoch weiter verfeinert: nur auf vom Verweisgesetz verschiedene Gesetze und keine Einzelnormen).

[8] Gute Einleitung bei *Hvitfeldt/Silge*, Supervised Machine Learning for Text Analysis in R, sub 4.

[9] Abrufbar unter https://github.com/LeonieWeissweiler/CISTEM, Erläuterung hierzu bei *Weißweiler/Fraser*, Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers.