

METHODEN UND FORSCHUNGSLOGIK
DER KOMMUNIKATIONSWISSENSCHAFT

Andreas Niekler

Automatisierte Verfahren
für die Themenanalyse
nachrichtenorientierter Textquellen

HERBERT VON HALEM VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Andreas Niekler

*Automatisierte Verfahren für die Themenanalyse
nachrichtenorientierter Textquellen*

Methoden und Forschungslogik der Kommunikationswissenschaft, 13
Köln: Halem, 2018

Andreas Niekler, geb. 21.7.1979, Dr. Ing., ist wissenschaftlicher Mitarbeiter an der Universität Leipzig. Derzeit arbeitet er an der Dynamik semantischer Kontexte in diachronen Korpora und der Extraktion von Aktivitäts- und Prozessinformationen aus großen Dokumentenquellen. Andreas Niekler studierte Medientechnik an der HTWK Leipzig und der University of West Scotland. Nach zwei Jahren als freier Programmierer und Dozent wechselte er an die Universität Leipzig. In seinem Dissertationsprojekt arbeitete er mit automatisierten Methoden zur Inhalts- und Themenanalyse.

Die Reihe *Methoden und Forschungslogik der Kommunikationswissenschaft* wird herausgegeben von Prof. Dr. Werner Wirth.

ISSN 1863-4966

Alle Rechte, insbesondere das Recht der Vervielfältigung und Verbreitung sowie der Übersetzung, vorbehalten. Kein Teil des Werkes darf in irgendeiner Form (durch Fotokopie, Mikrofilm oder ein anderes Verfahren) ohne schriftliche Genehmigung des Verlages reproduziert oder unter Verwendung elektronischer Systeme (inkl. Online-Netzwerken) gespeichert, verarbeitet, vervielfältigt oder verbreitet werden.

© 2018 by Herbert von Halem Verlag, Köln

ISBN (Print): 978-3-86962-261-3

ISBN (PDF): 978-3-86962-262-0

Den Herbert von Halem Verlag erreichen Sie auch im Internet unter <http://www.halem-verlag.de>
E-Mail: info@halem-verlag.de

SATZ: Herbert von Halem Verlag

LEKTORAT: Imke Hirschmann, Rabea Wolf

DRUCK: docupoint GmbH, Magdeburg

GESTALTUNG: Claudia Ott Grafischer Entwurf, Düsseldorf

Copyright Lexicon ©1992 by The Enschedé Font Foundry.

Lexicon® is a Registered Trademark of The Enschedé Font Foundry.

Inhalt

1.	EINLEITUNG	15
1.1	Ausgangslage	19
1.2	Problemstellung und Ziele	21
1.3	Aufbau der Arbeit	22
2.	TECHNISCHE UND THEORETISCHE GRUNDLAGEN FÜR DIE AUTOMATISCHE INHALTSANALYSE VON THEMENSTRUKTUREN	24
2.1	Inhaltsanalyse	25
2.1.1	Methodik und Eigenschaften	26
2.1.1.1	<i>Qualitative und quantitative Inhaltsanalysen</i>	27
2.1.1.2	<i>Deskription und Inferenz</i>	28
2.1.1.3	<i>Deduktiv und Induktiv</i>	29
2.1.2	Planung, Struktur und Ablauf	30
2.1.2.1	<i>Wichtige Begriffe der Inhaltsanalyse</i>	30
2.1.2.2	<i>Methodik der Kategorienbildung bei Inhaltsanalysen</i>	33
2.1.3	Themenanalysen	36
2.1.3.1	<i>Synthese linguistischer Themenanalysen</i>	42
2.1.3.2	<i>Das Thema im zeitlichen Verlauf</i>	45
2.1.3.3	<i>Nachrichtenfaktoren</i>	46
2.1.3.4	<i>Ökonomische Probleme der Inhaltsanalyse</i>	50

2.2	Computergestützte Analyse digitaler Textquellen	51
2.2.1	Verarbeitung und Repräsentation	52
2.2.1.1	<i>Quellen, Zeichensätze und Sprachen</i>	52
2.2.1.2	<i>Vorbereitung der Texte</i>	54
2.2.1.3	<i>Speicherung verarbeiteter Texte und Metadaten</i>	57
2.2.2	Maschinelles Lernen (Machine-Learning) und Text-Mining	60
2.2.2.1	<i>Statistik und maschinelles Lernen mit Text</i>	61
2.2.2.2	<i>Überwachtes und unüberwachtes Lernen</i>	62
2.2.2.3	<i>Information Retrieval und explorative Suche</i>	63
2.3	Zusammenfassung	66
2.4	Konkretisierung der Forschungsfragen	69
3.	ALGORITHMEN UND METHODEN FÜR DIE AUTOMATISCHE THEMENANALYSE	73
3.1	Topic Detection and Tracking	74
3.1.1	Clustermethode	77
3.1.2	Anwendung	78
3.2	Topic-Modelle	87
3.2.1	Latent Dirchlet Allocation	88
3.2.2	Erweiterungen und alternative Modelle	93
3.2.3	Berechnung und Inferenz	98
3.2.3.1	<i>Wie viele Themen hat ein Korpus? – Exkurs zu Dirichlet-Verteilung und -Sampling und deren Bedeutung für die latenten Variablen im LDA-Modell</i>	102
3.2.4	Anwendung	107
3.3	Signifikante Kookkurrenzen	117
3.4	Häufigkeiten, Messgrößen und Zeitreihen in Themen	125
3.4.1	Themenhäufigkeit	125
3.4.2	Worthäufigkeit	128
3.5	Zusammenfassung	131

4.	EXEMPLARISCHE ANALYSE	132
4.1	Vorbereitung und Verarbeitung	134
4.2	Bestimmung relevanter Themen	136
4.2.1	Explorative Analyse mit Textdateien	139
4.2.2	Explorative Analyse mit grafischen Oberflächen	151
4.2.3	Evaluation der explorativen Themenselektion	156
	4.2.3.1 <i>Validität der Themenverkettung</i>	156
	4.2.3.2 <i>Reliabilität in unterschiedlichen Korpora</i>	157
4.3	Themenhäufigkeiten	160
4.3.1	Häufigkeiten ohne Beachtung der Zeitstempel	161
4.3.2	Häufigkeiten mit Beachtung der Zeitstempel und Evaluation	162
	4.3.2.1 <i>Reliabilität</i>	164
	4.3.2.2 <i>Validität</i>	167
4.3.3	Zwischenfazit	180
4.4	Wort- und Akteurshäufigkeiten in Themen	181
4.4.1	Themenabhängige Häufigkeiten von Wörtern	183
4.4.2	Themenabhängige Häufigkeiten von Eigennamen	188
4.4.3	Abgrenzung zu Worthäufigkeitsanalysen	191
4.4.4	Zwischenfazit	193
4.5	Analyse des Aussagegehalts in Themen durch Kookkurrenzanalysen	194
4.5.1	Analyse von Schlüsselbegriffen	194
4.5.2	Analyse der Auswirkungen von Schlüsselereignissen	201
4.5.3	Zwischenfazit	204
4.6	Zusammenfassung und weitere Analysemöglichkeiten	206
5.	DISKUSSION DER FORSCHUNGSFRAGEN ZU AUTOMATISIERTEN THEMENANALYSEN	211
5.1	Grundsätzliche Fragen	211
5.1.1	Anschlussfähigkeit an die Methodik der Inhaltsanalyse	212
5.1.2	Automatisierung der Inhalts- bzw. Themenanalyse	215

5.2	Erweiterte Fragen	217
5.2.1	Qualitative und quantitative Aspekte	217
5.2.2	Deduktive und induktive Charakteristiken	218
5.2.3	Validität und Reliabilität	219
5.2.4	Weiterverarbeitung, Analyse und Anwendung von Ergebnissen	221
5.2.4.1	<i>Diachrone Themenanalyse</i>	221
5.2.4.2	<i>Häufigkeitsverläufe und Zyklen von Themen</i>	222
5.2.4.3	<i>Nachrichtenfaktoren</i>	223
5.2.4.4	<i>Vergleichbarkeit unterschiedlicher Quellen</i>	224
5.2.5	Datenhaltung und Datenverarbeitung	224
5.3	Fazit und Ausblick	227
6.	Anhang	230
7.	Literaturverzeichnis	247