

Norman Zänker/Christian Zietzsch

## **Text Mining und dessen Implementierung**

## **Bibliografische Information der Deutschen Nationalbibliothek:**

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de/> abrufbar.

Dieses Werk sowie alle darin enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsschutz zugelassen ist, bedarf der vorherigen Zustimmung des Verlanges. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen, Auswertungen durch Datenbanken und für die Einspeicherung und Verarbeitung in elektronische Systeme. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe (einschließlich Mikrokopie) sowie der Auswertung durch Datenbanken oder ähnliche Einrichtungen, vorbehalten.

Copyright © 2010 Diplomica Verlag GmbH  
ISBN: 9783842806283

**Norman Zänker, Christian Zietzsch**

## **Text Mining und dessen Implementierung**



Norman Zänker/Christian Zietzsch

## **Text Mining und dessen Implementierung**

Norman Zänker/Christian Zietzsch  
**Text Mining und dessen Implementierung**

ISBN: 978-3-8428-0628-3

Herstellung: Diplomica® Verlag GmbH, Hamburg, 2011

Zugl. Technische Universität Bergakademie Freiberg, Freiberg, Deutschland,  
Bachelorarbeit, 2010

---

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden und der Verlag, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

© Diplomica Verlag GmbH  
<http://www.diplomica.de>, Hamburg 2011

## Inhaltsverzeichnis

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Einleitung</b>  | <b>10</b> |
| 1.1      | Zielsetzung . . . . .  | 11        |
| 1.2      | Aufbau der Arbeit . . . . .  | 11        |
| <b>2</b> | <b>Grundlagen</b>  | <b>12</b> |
| 2.1      | Was ist Text Mining? . . . . .   | 12        |
| 2.2      | Aufbau und Struktur von Text . . . . .                                       | 13        |
| 2.3      | Linguistischer Strukturalismus als Grundlage zur Bedeutungsanalyse . . . . . | 15        |
| 2.3.1    | Die Linguistik und ihre Ebenen . . . . .                                     | 15        |
| 2.3.2    | Syntagmatische und Paradigmatische Relationen . . . . .                      | 20        |
| 2.3.3    | Semantische Relationen . . . . .   | 24        |
| <b>3</b> | <b>Text Mining-Prozess</b>   | <b>27</b> |
| 3.1      | Unterschied Text Mining und Data Mining . . . . .                            | 28        |
| 3.2      | Dokumentsuche . . . . .  | 29        |
| 3.2.1    | Information Retrieval . . . . .  | 29        |
| 3.2.2    | Aufbau und Funktion eines Information Retrieval Systems . . . . .            | 30        |
| 3.3      | Dokumentaufbereitung . . . . .   | 32        |
| 3.3.1    | Textressourcen . . . . .   | 32        |
| 3.3.2    | Aufbau eines Analysekorpus . . . . .   | 32        |
| 3.3.2.1  | Satzsegmentierung . . . . .  | 34        |
| 3.3.2.2  | Wortsegmentierung . . . . .  | 35        |
| 3.4      | Text Mining - Statistische Analysemethoden . . . . .                         | 37        |
| 3.4.1    | Zipfsches Gesetz . . . . .   | 37        |
| 3.4.2    | Differenzanalyse . . . . .   | 37        |
| 3.4.3    | Part-of-Speech Tagging . . . . .   | 39        |
| 3.4.3.1  | Regelbasierte Tagger . . . . .   | 39        |
| 3.4.3.2  | Stochastische Tagger . . . . .   | 41        |
| 3.4.3.3  | Regelbasierte Tagger vs. Stochastische Tagger . . . . .                      | 47        |
| 3.4.4    | Kookkurrenzanalyse . . . . .   | 47        |
| 3.5      | Text Mining - Clusteranalyse . . . . .                                       | 51        |
| 3.5.1    | Nicht-hierarchische Verfahren . . . . .                                      | 51        |
| 3.5.2    | Hierarchische Verfahren . . . . .  | 53        |
| 3.5.3    | Fuzzy-Clusteranalyse . . . . .   | 54        |
| 3.5.4    | Dokumentähnlichkeit . . . . .  | 55        |
| 3.5.5    | Anwendungsbeispiel . . . . .   | 58        |
| 3.6      | Text Mining - Musteranalyse . . . . .  | 60        |
| 3.6.1    | Reguläre Ausdrücke . . . . .   | 60        |
| 3.6.2    | Syntaktische Muster . . . . .  | 62        |
| <b>4</b> | <b>Text Mining-Prozess anhand des Zalazar Text Miner</b>                     | <b>64</b> |
| 4.1      | Programmaufbau . . . . .   | 65        |

---

|          |   |           |
|----------|---|-----------|
| 4.2      | Dokumentaufbereitung . . . . .                                  | 67        |
| 4.2.1    | Formatierung des zu analysierenden Textdokuments . . . . .      | 67        |
| 4.2.2    | Aufbau des Analysekorpus anhand der Satz- und Wortsegmentierung | 68        |
| 4.3      | Dokumentanalyse . . . . .                                       | 75        |
| 4.3.1    | Part-of-Speech Tagging . . . . .                                | 75        |
| 4.3.2    | Differenzanalyse . . . . .                                      | 78        |
| 4.3.3    | Musteranalyse . . . . .   | 79        |
| 4.4      | Ergebnisevaluation . . . . .                                    | 81        |
| 4.5      | Handhabung des Zalazar Text Miner . . . . .                     | 82        |
| 4.5.1    | Öffnen einer neuen Mail . . . . .                               | 83        |
| 4.5.2    | Durchführen der Textanalyse . . . . .                           | 84        |
| 4.5.3    | Speichern der Ergebnisse . . . . .                              | 85        |
| 4.5.4    | Laden der Ergebnisse einer bereits analysierten Mail . . . . .  | 86        |
| <b>5</b> | <b>Schlusswort</b>  | <b>87</b> |
|          | <b>Literatur</b>  | <b>89</b> |



## Abbildungsverzeichnis

|    |  |    |
|----|--|----|
| 1  | Überblick: Aufbau eines Textes (vgl. <i>G. Heyer</i> [5]) . . . . .  | 15 |
| 2  | Prozess des Text Mining . . . . .  | 27 |
| 3  | Vergleich Text Mining- und Data Mining Prozess . . . . .   | 28 |
| 4  | Aufbau eines IR-Systems (vgl. <i>T. Gottron</i> [2]) . . . . .   | 30 |
| 5  | Aufbau eines Hidden-Markov-Modells . . . . .   | 42 |
| 6  | Komplettes Gitter für „The design of the car is great.“ (in Anlehnung an <i>K. Haenelt</i> [4]). . . . .                                   | 44 |
| 7  | Reduziertes Gitter für „The design of the car is great.“ mithilfe des Viterbi-Algorithmus (in Anlehnung an <i>K. Haenelt</i> [4]). . . . . | 46 |
| 8  | Wortnetz der Wortform „technology“ (vgl. <i>Uni-Leipzig, Projekt Wortschatz</i> [12]) . . . . .  | 50 |
| 9  | Clustering mit k-means (vgl. <i>G. Heyer</i> [5]) . . . . .  | 52 |
| 10 | Hierarchische Verfahren: agglomerativ (Links), divisiv (Rechts) . . . . .  | 53 |
| 11 | Schmetterlingsproblem (vgl. <i>S. Grossmann</i> [3]) . . . . .   | 54 |
| 12 | Verfahren zur Ähnlichkeitsbestimmung zweier Cluster: single-link (Links), complete-link (Mitte), group-average (Rechts) . . . . .          | 58 |
| 13 | Beispiel: Clustering thematisch ähnlicher Dokumente . . . . .  | 59 |
| 14 | Schematischer Aufbau des Zalazar Text Miner . . . . .  | 65 |
| 15 | Klassendiagramm des Zalazar Text Miner . . . . .   | 66 |
| 16 | Analysekorpus des Zalazar Text Miner . . . . .   | 69 |
| 17 | Hauptfenster des Zalazar Text Miner . . . . .  | 82 |
| 18 | FileDialog zum Öffnen einer Mail . . . . .   | 83 |
| 19 | Analysefenster des Zalazar Text Miner . . . . .  | 84 |
| 20 | FileDialog zum Speichern der gewonnenen Ergebnisse . . . . .   | 85 |
| 21 | FileDialog zum Laden bereits gespeicherter Analyseergebnisse . . . . .   | 86 |

## Tabellenverzeichnis

|   |  |    |
|---|--|----|
| 1 | Linguistische Ebenen und ihre Teildisziplinen (vgl. <i>G. Heyer</i> [5]) . . . . .                       | 19 |
| 2 | Häufigkeitssortierte Liste des Romans „Tom Sawyer“ (vgl. <i>B. Hoffmann</i> [6]) . . . . .               | 37 |
| 3 | Hidden-Markov-Matrix für den Satz „The design of the car is great“ (Matrizenanordnung: A-II-B) . . . . . | 42 |
| 4 | Beispiele für Signifikanzwerte der Kookkurrenz zweier Wortformen (vgl. <i>G. Heyer</i> [5]) . . . . .    | 50 |
| 5 | Term-Dokument-Matrix (vgl. <i>G. Heyer</i> [5]) . . . . .  | 56 |
| 6 | Dokument-Dokument-Matrix (vgl. <i>G. Heyer</i> [5]) . . . . .  | 57 |
| 7 | Dokument-Dokument-Matrix (Beispiel) . . . . .  | 59 |
| 8 | Auszug aus dem Penn Treebank Tagset . . . . .  | 76 |
| 9 | Softwarelösungen zur Informationsextraktion . . . . .  | 87 |