Bachelorarbeit

Norman Zänker/Christian Zietzsch

Text Mining
und dessen Implementierung

Diplom.de

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.

ISBN: 9783842806283

Norman Zänker, Christian Zietzsch

Text Mining und dessen Implementierung

Diplom.de

Bachelorarbeit

Norman Zänker/Christian Zietzsch

Text Mining
und dessen Implementierung

Diplom.de

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

Inhaltsverzeichnis

1 Einleitung **10**
1.1 Zielsetzung . . . 11
1.2 Aufbau der Arbeit . . . 11

2 Grundlagen **12**
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

3 Text Mining-Prozess **27**
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

4 Text Mining-Prozess anhand des Zalazar Text Miner **64**
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

5 Schlusswort **87**

Literatur **89**

Abbildungsverzeichnis

1 Überblick: Aufbau eines Textes (vgl. *G. Heyer* [5]) 15
2 Prozess des Text Mining 27
3 Vergleich Text Mining- und Data Mining Prozess 28
4 Aufbau eines IR-Systems (vgl. *T. Gottron* [2]) 30
5 Aufbau eines Hidden-Markov-Modells 42
6 Komplettes Gitter für „The design of the car is great.“ (in Anlehnung an *K. Haenelt* [4]). 44
7 Reduziertes Gitter für „The design of the car is great.“ mithilfe des Viterbi-Algorithmus (in Anlehnung an *K. Haenelt* [4]). 46
8 Wortnetz der Wortform „technology“ (vgl. *Uni-Leipzig, Projekt Wortschatz* [12]) 50
9 Clustering mit k-means (vgl. *G. Heyer* [5]) 52
10 Hierarchische Verfahren: agglomerativ (Links), divisiv (Rechts) 53
11 Schmetterlingsproblem (vgl. *S. Grossmann* [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. *G. Heyer* [5]) 19
2 Häufigkeitssortierte Liste des Romans „Tom Sawyer“ (vgl. *B. Hoffmann* [6]) 37
3 Hidden-Markov-Matrix für den Satz „The design of the car is great“ (Matrizenanordnung: A-Π-B) 42
4 Beispiele für Signifikanzwerte der Kookkurrenz zweier Wortformen (vgl. *G. Heyer* [5]) 50
5 Term-Dokument-Matrix (vgl. *G. Heyer* [5]) 56
6 Dokument-Dokument-Matrix (vgl. *G. Heyer* [5]) 57
7 Dokument-Dokument-Matrix (Beispiel) 59
8 Auszug aus dem Penn Treebank Tagset 76
9 Softwarelösungen zur Informationsextraktion 87