Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Text Mining

Ausarbeitung im Rahmen des Aufbauseminars im Studiengang Next Media

Xenia Sataev

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Inhaltsverzeichnis

1 Einleitung und Begriffsbestimmung 1

2 Text Mining-Prozess 3

3 Methoden 5

3.1 Information Extraction 5

3.2 Clustern 7

3.3 Klassifikation 7

3.4 Information Retrieval 7

3.5 Identifikation von Trends und Events 8

4 Anbieter von Text Mining Systemen 8

5 Fazit und Ausblick 9

Literaturverzeichnis 11

1 Einleitung und Begriffsbestimmung

„Bei der Berlinale ist der Dokumentationsfilm "Fuocoammare" von Gianfranco Rosi als bester Film ausgezeichnet worden. In dem Werk wird der Alltag auf der italienischen Insel Lampedusa gezeigt. Dort kommen seit Monaten zahlreiche Flüchtlinge an.
Der beste Film auf der diesjährigen Berlinale heißt "Fuocoammare" - übersetzt "Feuer auf See". Die Dokumentation von Gianfranco Rosi wurde mit dem Goldenen Bären ausgezeichnet. Der auf der italienischen Insel Lampedusa gedrehte Film nähert sich der Flüchtlingskatastrophe auf dem Mittelmeer durch stille Beobachtung, abseits traditioneller Berichterstattungsformen.
Rosi kontrastiert idyllische Szenen vom Alltag der Inselbewohner mit Momenten, die das Grauen der Flucht auf kleinen, völlig überfüllten Booten zeigen. Auch vor Bildern des realen Sterbens schreckt er nicht zurück.
In seiner Dankesrede vor rund 1600 Gästen sagte Rosi, er hoffe, dass sein Werk ein Bewusstsein dafür schaffe, dass das Sterben der Menschen auf dem Mittelmeer nicht zu akzeptieren sei. "Ich widme diesen Preis den Menschen von Lampedusa, die ihr Herz den Menschen öffnen, die dort ankommen", sagte Rosi.“ (tagesschau.de, Februar 2016)

Untersucht man diesen Zeitungsartikel ohne technische Hilfsmittel nach den Wörtern, die er enthält, und ordnet diese in Kategorien ein, lässt sich daraus eine thematische Gewichtung ableiten. Die am häufigsten vorkommenden Kategorien sind Medien (lila) und Politik/Gesellschaft (grün), gefolgt von Ortsangaben (blau) sowie positiv aufgeladenen Wörtern (braun). Seltener vertreten sind Wörter aus den Bereichen Natur (orange), Zeit- und Zahlenangaben (rot), Namen (gelb) und negativ aufgeladene Wörter (türkis). Begriffe, die keine besondere Aussagekraft haben und als Füllwörter dienen, wurden ausgegraut. Die Häufigkeit der Wörter spiegelt das Grundthema des Artikels wider. Es geht um ein positives Ereignis, eine Filmpreisverleihung, die mit verschiedenen Orten in Beziehung steht und gleichzeitig auch ein politisches Thema betrifft. Ohne den Text chronologisch zu lesen oder ihn inhaltlich in Gänze zu erfassen, lässt sich das Thema des Artikels dennoch durch die Dominanz der Wörter in diesem erfassen.

Um große Textmengen zu untersuchen, reicht ein manuelles Vorgehen jedoch nicht aus. Dies ist allerdings nötig, um die riesige Anzahl an Textdokumenten, die Unternehmen vorliegen, zu untersuchen, Bibliotheken zu verwalten oder Blog- oder Tweets inhaltlich zu analysieren. Zu diesem Zweck wurde die Methode des Text Mining entwickelt, die als Sonderform des Data Mining verstanden werden kann. Die Besonderheiten des Text Mining sind zum einen die Struktur der Daten, zum anderen die Herkunft dieser. Der zu untersuchende Datentyp ist un- bzw. semistrukturierter Text aus internen oder externen Quellen.

Im Folgenden wird der Begriff des Text Mining geklärt sowie die unterschiedlichen Prozessschritte des behandelt. Anschließend werden im Kapitel 3 die verschiedenen Methoden des Text Mining wie Information Extraction, Clustern, Klassifikation, Information Retrieval sowie Identifikation von Trends und Events beleuchtet. Folglich werden die Anbieter von Text Mining-Programmen kategorisiert und schließlich ein Ausblick zu dem behandelten Thema gegeben.

Seit über dreißig Jahren ist Text Mining ein relevantes Thema in der Forschung, obwohl zu Beginn nur in wenigen Fachrichtungen, wie beispielsweise Biowissenschaften, verbreitet (vgl. Upshall 2014: 91). Den Begriff Text Mining beziehungsweise Knowledge Discovery in Text (KDT) prägten erstmals Feldman und Dagan im Jahr 1995 (vgl. Feldman/Dagan 1995: 112).
„Text mining entails automatically analyzing a corpus of text documents and discovering previously hidden information. The result might be another piece of text or any visual representation“ (Hashimi/ Hafez/ Mathkour 2015: 729). Text Mining dient der Entdeckung neuer Informationen und Muster in Textdokumenten mittels spezifischer Algorithmen. Laut Schätzungen bilden Texte 80% der Informationsbasis eines Unternehmens. Sie besitzen zwar eine Semantik, sind jedoch unstrukturierte Daten, die sich von den strukturierten Daten in Datenbanken unterscheiden. Häufig soll ein Dokument mit Hilfe des Text Mining beispielsweise nach Themengebieten klassifiziert werden. Während beim Data Mining vor allem auf interne Daten zurückgegriffen wird, werden beim Text Mining auch Daten externer Quellen ausgewertet. (Vgl. Gabriel et al. 2009: 142ff.) Neben herkömmlichen Textdokumenten spielen zunehmend internetbasierte eine wichtige Rolle.
Aufgrund der riesigen Textdatenmengen, die aus verschiedenen Quellen zur Verfügung stehen, ist es nicht möglich diese vollständig zu lesen und eine manuelle Analyse durchzuführen. So werden Text Mining-Techniken benötigt, um Text in Daten zu konvertieren, die anschließend mit Hilfe verschiedener Data Mining Analysetechniken untersucht werden können. Ebenso wie mit Hilfe von Data Mining können mittels Text Mining unter anderem neue Muster entdeckt, zukünftige Ereignisse vorausgesagt sowie zunehmende oder rückläufige Entwicklungen überwacht werden. Auch das Beobachten von internetbasierten Textquellen, die in Blogs, Social Media etc. zu finden sind, kann mit Text Mining stattfinden. Im biomedizinischen Bereich hilft Text Mining dabei, die Literatursuche in Datenban-

ken zu verbessern. Auch die Analyse, das Speichern und die Verfügbarkeit von Informationen auf verschiedenen Webseiten und Suchmaschinen werden durch diese Technik effizienter und genauer gemacht. Im linguistischen Feld ist die lexikalische Analyse und die Mustererkennung, mit der man die Verbreitung von Worthäufigkeiten entdecken kann, zu nennen. (Vgl. Hashimi/ Hafez/ Mathkour 2015: 729)

Im Text Mining werden verschiedene Prozessstufen unterschieden, die mit dem KDD-Prozess[1] vergleichbar sind. Ein wichtiger Unterschied besteht jedoch in der Datenaufbereitung. So ist beim Text Mining eine „zusätzliche linguistische Datenaufbereitung erforderlich, um die fehlende Datenstruktur zu rekonstruieren“ (Hippner/ Rentzmann 2006: 287)

2 Text Mining-Prozess

Hippner und Rentzmann unterscheiden sechs Stufen des Text Mining-Prozesses (vgl. Abb. 1):

- *Aufgabendefiniton*
- *Dokumentenselektion*
- *Dokumentenaufbereitung*
- *(Text) Mining Methoden*
- *Interpretation/ Evaluation*
- *Anwendung*

(Abb. 1: Text Mining Prozess nach Hippner/ Rentzmann 2006: 288)

Nachdem im ersten Schritt das Ziel und die Aufgaben der Analyse festgelegt wurden (*Aufgabendefinition*), werden im zweiten Schritt die Dokumente beziehungsweise Texte identifiziert, die für die Fragestellung von Bedeutung sind (*Do-*

[1] Knowledge Discovery in Databases (KDD)-Prozess setzt sich aus folgenden Schritten zusammen: Selektion der Daten, Datenvorverarbeitung, Transformation, Data Mining, Interpretation der Ergebnisse (vgl. Fayyad et al. 1996: 29)

kumentenselektion). In einem speziellen Document Warehouse können verschiedene Dokumenttypen, wie etwa E-Mails oder Berichte, verschiedener Quellen zusammengeführt werden. (Vgl. Hippner/ Rentzmann 2006: 287f.)

Der dritte Schritt Datenbereinigung und Datenvorverarbeitung (*Dokumentaufbereitung*) hat im Text Mining-Verfahren eine besondere Bedeutung. Nachfolgend werden einige der vielen möglichen Techniken thematisiert. In Kapitel 3.1 werden einige dieser Verfahren nochmals aufgegriffen und genauer erläutert.

„Die Herausforderung des Text Mining liegt [.] darin, die in einem Text sprachlich wiedergegebene Information für die maschinelle Analyse zu erschließen" (ebd. 287). Da Texte unstrukturierte Daten sind, müssen vor der eigentlichen Analyse die relevanten Informationen aus dem Text herausgezogen werden. Um die Menge der zu untersuchenden Wörter zu reduzieren, müssen nicht relevante Begriffe mit Hilfe von Methoden wie *Filtering*, *Lemmatization* oder *Stemming* (Hotho 2005: 6f.) entfernt werden. Mit dem Filtering werden Füllwörter, die wenig oder keinen Inhalt tragen, wie Artikel, Konjunktionen oder Präpositionen, entfernt. Das Lemmatization dient der Umwandlung von Verbformen in Infinitive sowie von Substantiven in ihre Singularform. Dafür wird jedes Wort in dem Text mit der zutreffenden Wortart, wie Verb oder Nomen, markiert beziehungsweise getaggt. Mit dem *Stemming* werden Wörter auf ihren Wortstamm gekürzt, um Begriffe mit unterschiedlicher Schreibweise miteinander vergleichen zu können. (Vgl. Hotho 2005: 6f.) Die Groß- und Kleinschreibung wird meist ignoriert. Abkürzungen und Synonyme müssen erkannt werden. Um die relevanten Informationen herauszuziehen, werden Schlüsselwörter, die repräsentativ für die Texte sind, oder die Häufigkeitsverteilung von Begriffen identifiziert.

Das Ergebnis des Schrittes Dokumentaufbereitung ist „eine reduzierte Menge von Wörtern (bag of words), die man zusätzlich noch gewichten kann. Auf der Basis dieser Wortmengen findet dann die eigentliche Datenanalyse statt" (ebd.).

Im vierten Schritt können die aufbereiteten Daten nun mit Hilfe von Verfahren verarbeitet werden, die auch beim klassischen Data Mining zum Einsatz kommen. So können Texte beispielweise in Cluster gruppiert (vgl. Kapitel 3.2) oder verschiedenen Klassen zugeordnet werden (vgl. Kapitel 3.3).

Im fünften Schritt werden die Ergebnisse interpretiert, bewertet und abschließend in der Praxis angewendet. So können beispielsweise mittels einer Text Mining-Analyse von Verwendungszwecken bei Banktransaktionen Kundendaten mit den

neu gewonnenen Informationen angereichert werden. (Vgl. Hippner/ Rentzmann 2006: 289)

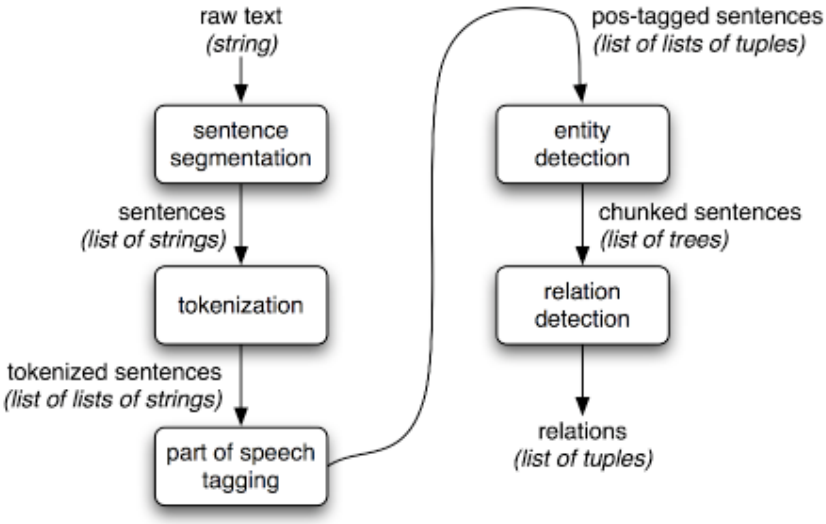
Im nachfolgenden Kapitel werden einige der unterschiedlichen Methoden erläutert, die beim Text Mining zum Einsatz kommen.

3 Methoden

In der Wissenschaft existiert eine Vielzahl verschiedener Aufgabenbereiche beziehungsweise Methoden des Text Mining. Die vorliegende Arbeit beschränkt sich auf die fünf am häufigsten genannten Methoden. Müller und Lenz (2013) unterscheiden zwischen Information Extraction (Kapitel 3.1), Clustern (Kapitel 3.2), Klassifikation (Kapitel 3.3) und Information Retrieval (Kapitel 3.4). Cramer, Engel und Whitney sprechen zudem von der Aufdeckung von Events und Trends in Text Streams mittels Text Mining (Kapitel 3.5). Da Information Extraction ein bedeutsamer Bestandteil des Text Mining ist, wird diese Methode detaillierter behandelt.

3.1 Information Extraction

Bei der Information Extraction werden strukturierte Informationen aus Texten herausgezogen. „Der Vorgang der Informationsextraktion entspricht der Überführung eines unstrukturierten Textes in auswertbare Datenbankfelder“ (Meier/ Beckh 2000: 166). Das Ziel ist die Extraktion von Fakten und Beziehungen aus unstrukturierten und semistrukturierten Texten. Die Hauptaufgabe ist das Herausziehen von Textteilen und das Zuordnen von spezifischen Attributen zu diesen (vgl. Hotho 2005: 22). Dabei kommen üblicherweise Verfahren der Computerlinguistik zum Einsatz.
Bird/ Klein/ Loper (2009: 136ff.) unterscheiden fünf Schritte im Information Extraction Prozess (Abb. 2).



(Abbildung 2: Information Extraction Prozess nach Bird/ Klein/ Loper 2009: 136)

Im ersten Schritt *Sentence Segmentation* wird der unverarbeitete Text, mittels Satzendzeichen wie „.“, „!“, „?“, in Sätze aufgespalten. Mit Hilfe eines sogenannten Tokenizers (lexikalischer Scanner) wird im folgenden Prozessschritt *Tokenization* jeder Satz des Dokuments in Wörter (Token) unterteilt. Dabei werden alle Satzzeichen und andere nicht als Text identifizierten Bestandteile innerhalb eines Satzes durch Leerzeichen ersetzt. Das Ergebnis sind Listen von Wörtern in unveränderter Reihenfolge, die zur weiteren Verarbeitung genutzt werden.
Bei dem *Part-of-speech Tagging* wird jedes Wort im Satz mit einer Wortart (part of speech tag) versehen. Diese können unter anderem Verben, Adjektive, Nomen oder Präpositionen sein. Entscheidend ist dabei der Kontext, denn ein und demselben Wort können in unterschiedlichen Sätzen verschiedene Wortarten zugeordnet werden. Im nächsten Schritt *Entity Detection* wird nach potentiell relevanten Entitäten in einem Satz gesucht. Diese sind zum Beispiel Personennamen, Ortsnamen oder Zeitangaben. Durch Zuweisung von Wortarten und damit der Musterdefinition im vorangegangenen Schritt, können in diesem Prozessschritt Satzteile bestimmt werden, die extrahiert werden sollen. Im letzten Schritt *Relation Detection* werden die Beziehungen zwischen den unterschiedlichen Entitäten in einem Text entdeckt. Die Relationen werden mittels Fragen „Wer?“, „Wen?“, „Was?“, „Wann?“, „Wo?“ und „Warum?“ extrahiert. In dem in der Einleitung dargestellten Zeitungsartikel hat Gianfranco Rosi (Wer?) einen Filmpreis (Was?) im Februar (Wann?) bei der Berlinale (Wo?) für den besten Film (Warum?) bekommen.

In das Programm wird also ein unverarbeiteter Text eingefügt (Input), den das Programm in eine Liste von Tupeln (Entität, Beziehung, Entität) zergliedert (Output). (Vgl. Bird/ Klein/ Loper 2009: 136ff.; Vgl. Müller/ Lenz 2013: 111f.)

3.2 Clustern

Im Clustering geht es darum Strukturen in den Daten zu finden und die Daten aufzuteilen. Mittels Algorithmen werden bei diesem Verfahren Textdokumente mit ähnlichem Inhalt automatisch in Gruppen (Cluster) eingeteilt. Aber auch innerhalb eines Textdokumentes können Cluster gebildet werden, um Wörter zu gruppieren. Dabei werden Objekte mit ähnlichen Eigenschaften in Gruppen zusammengefasst, die Objekte unterschiedlicher Cluster unterscheiden sich hingegen voneinander. (Vgl. Hotho 2005: 15)

3.3 Klassifikation

Ähnlich wie bei dem Clustering zielt die Klassifikation darauf ab, Textdokumente verschiedenen Gruppen (Klassen) zuzuordnen. Der Unterschied zum Clustering ist, dass in der Clusteranalyse die Gruppen automatisch gefunden werden, in der Klassifikation jedoch bereits bekannt sind. Dabei werden in dem Prozess Regeln gesucht, um die Klasse eines Objekts zu bestimmen. (Vgl. Müller/ Lenz 2013: 95f.) Dies geschieht zum einen anhand von Merkmalen des Dokuments zum anderen anhand bereits klassifizierter Texte (vgl. ebd. 112). Auf diese Weise werden beispielsweise Nachrichtentexte in Klassen wie Politik, Kunst oder Sport eingeteilt.

3.4 Information Retrieval

Bei der Methode Information Retrieval werden Texte mithilfe von Stichwortanfragen gesucht. Ein alltägliches Beispiel ist die Stichwortsuche über Suchmaschinen wie Google, Yahoo! oder Bing. So soll mittels Schlüsselwörter die Menge relevanter Dokumente gefunden werden (vgl. ebd. 113).
Dabei geht es weniger um die Inhalte der Dokumente als viel mehr um das Identifizieren der Dokumente selbst: „Information retrieval is the finding of documents

which contain answers to questions and not the finding of answers itself“ (Hearst 1999 nach Parhizkar 2010: 4)

3.5 Identifikation von Trends und Events

Cramer, Engel und Whitney sprechen von der Aufdeckung von Events und Trends in Text Streams mittels Text Mining. Mit Text Streams sind Sammlungen von Dokumenten oder Nachrichten gemeint, die im Zeitverlauf erzeugt und beobachtet werden. In diesen Text Streams werden durch Text Mining Themenveränderungen entdeckt und charakterisiert, um zum einen kurzzeitige atypische Ereignisse, zum anderen langfristige Verlagerungen thematischer Schwerpunkte zu identifizieren. (Vgl. Engel/ Whitney/ Cramer 2010: 167)

Mit der zunehmenden Bedeutung von Text Mining nimmt auch die Zahl der Anbieter verschiedener Text Mining Programme zu. Ein Vergleich der unterschiedlichen Anbieter stellt sich aufgrund der Bandbreite verschiedener Funktionalitäten schwierig dar. Die Unterschiede in der Softwareentwicklung aufgrund fehlender standardisierter Regeln verstärken das Vergleichsproblem.
Dennoch wird im folgenden Kapitel eine mögliche Kategorisierung von Anbietern von Text Mining Programmen vorgestellt.

4 Anbieter von Text Mining Systemen

Softwarepakete reichen von Unternehmensebene – wie etwa *Temis* oder *SmartLogic* – bis hin zu Lösungen für individuelle Nutzung – zum Beispiel *Nvivo* oder *Dedoose*. Andere Anbieter haben sich auf einige Funktionen des Text Mining spezialisiert. So bietet *Zemanta* das Taggen von individuellen Blogposts an, *GATE* oder *RapidMiner* haben Versionen zur Analyse von Tweets herausgebracht. (Vgl. Upshall 2014: 98).
Hippner und Rentzmann unterscheiden zwischen reinen Text-Mining-Anbietern wie beispielsweise *Clearforest, Inxight* oder *Temis*. Indirekte Anbieter, die bereits Data Mining Systeme anbieten und diese um Text Mining-Funktionen erweitert haben, sind unter anderem *IBM, SAS* oder *SPSS*. Neben den oben erwähnten, zäh-

len auch *Fast*, das sich auf Suchtechnologie spezialisiert hat oder *Verity*, das seinen Focus im Bereich Information Retrieval hat, zu den sogenannten Teil-Anbietern. (Vgl Hippner/ Rentzmann 2005: 289)
Neben bezahlten Programmen gibt es auch Open Source-Lösungen, etwa von *R* oder *KNIME*.

5 Fazit und Ausblick

Bücher, Verträge, Blogs, Twitter, Wikipedia, ob analog oder digital, Text ist überall. Angesichts der durch die Digitalisierung immer mehr steigenden Datenmassen sowie der Speicherung dieser, reichen manuelle Auswertungsmethoden nicht aus, um sich den Inhalten von Textdokumenten zu nähern. Das Identifizieren und Extrahieren relevanter Informationen und Daten von unstrukturierten Texten gewinnt gleichzeitig immer mehr an Bedeutung. Doch die computerbasierte Analyse von großen Textmengen steht erst an den Anfängen.
Verglichen mit dem klassischen Data Mining-Prozess ist der Prozess des Text Mining langwieriger und komplexer. Die Datenbereinigung sowie die Datenaufbereitung sind ein großer Bestandteil des gesamten Prozesses. Die Kenntnis über den Umgang mit diesen Techniken und das Bedienen der zahlreichen auf dem Markt angebotenen Tools und Programmen ist demnach unabdingbar und erfordert Erfahrung.
Das Potenzial von Text Mining ist groß. Zum einen liegt der Großteil der Informationen in Textform vor, zum anderen spielen Kenntnisse und Informationen über Kunden, Märkte und die Konkurrenz eine sehr große Rolle für den Erfolg eines Unternehmens (vgl. Hippner/ Rentzmann 2006: 289).
„By sampling the sentiments expressed in the torrent of blog posts, tweets and Facebook updates, you can gain unprecedented insights into the mood of the world and use it to predict what is to come“ (Giles 2011: 34). Forscher haben Wege gefunden um mittels Text Mining die Stufe von Angst einer Nation zu messen, um mit den Ergebnissen Vorhersagen von Aktienmarktentwicklungen zu verbessern. Andere haben die besondere Häufung von Stichworten zum Thema Job in Google-Suchen als Indikator für den Anstieg von Arbeitslosigkeit identifiziert. In vielen Bereichen haben Analysen das Potenzial durch neu generiertes Wissen be-

stimmten Entwicklungen entgegenzuwirken. (Vgl. ebd.) Jedoch gibt es natürlich auch die Schattenseite. „The blog posts and tweets in which we share our thoughts and feelings are all now a target for advertisers. We are all part of a vast market research project, whether we like it or not.“ (ebd.)

Literaturverzeichnis

Bird, Steven/ **Klein**, Ewan/ **Loper**, Edward: Natural language processing with Python (2009). O'Reilly.

Engel, Dave/ **Whitney**, Paul/ **Cramer**, Nick (2010): Events and trends in text streams. In: Berry, Michael/ Kogan, Jacob (2010): Text Mining: Applications and Theory: 167-182.

Fayyad, Usama/ **Piatetsky-Shapiro**, Gregory/ **Smyth**, Padhraic (1996): The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM, Nov. 1996, Vol. 39, Nr. 11: 27-34.

Feldman, Ronen/ **Dagan**, Ido (1995): Knowledge Discovery in Textual Databases (KDT). In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95): 112–117.

Gabriel, Roland/ **Gluchowski**, Peter/ **Pastwa**, Alexander (2009): Data Warehouse & Data Mining. Herdecke/ Witten.

Giles, Jim (2011): Text Mining. In: NewScientist. Vol. 210, Nr. 2812, 2011: 34.

Hashimi, Hussein/ **Hafez**, Alaaeldin/ **Mathkour**, Hassan (2015): Selection criteria for text mining approaches. In: Computers in Human Behavior, 2015, Vol. 51: 729–733.

Hippner, Hajo/ **Rentzmann**, René (2006): Text Mining. In: Informatik Spektrum, Volume 29, Nr. 4, 2006: 287-290.

Hotho, Andreas (2005): A Brief Survey of Text Mining. http://www.kde.cs.uni-kassel.de/hotho/pub/2005/hotho05TextMining.pdf (abgerufen am 21. Februar 2016).

Meier, Marco/ **Beckh**, Michael (2000): Text Mining. In: WIRTSCHAFTSINFORMATIK, Vol. 42, Nr. 2, 2000: 165-167.

Parhizkar, Behrang/ **Kourouma**, Kerfalla/ **Fazilah,** Siti/ **Nian**, Yap Sing/ **Navartnam**, Sujata/ **Ng Giap Wen**, Edmund (2010): Intelligent Online Course evaluation system Using NLP Approach. In: International Journal of Computer and Network Security, Vol. 2, Nr. 10, 2010: 1-17.

Upshall, Michael (2014): Text mining: Using search to provide solutions. In Business Information Review, 2014, Vol. 31, Nr. 2: 91–99.

Internetquellen

tagesschau.de:
http://www.tagesschau.de/kultur/berlinale-goldener-baer-101.html (abgerufen am 21. Februar 2016)

Abbildungsverzeichnis

Abbildung 1:
Text Mining Prozess nach Hippner/ Rentzmann 2006: 288. In:
Hippner, Hajo/ Rentzmann, René (2006): Text Mining. In: Informatik Spektrum, Volume 29, Nr. 4, 2006: 287-290.

Abbildung 2:
Information Extraction Prozess nach Bird/ Klein/ Loper 2009: 136. In:
Bird, Steven/ Klein, Ewan/ Loper, Edward: Natural language processing with Python (2009). O'Reilly.

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

_____

(Datum, Unterschrift)