Text Mining

Jan Bartels

Agenda

- Was ist Text Mining
- Unterschied zu Data Mining
- Herausforderungen
- Ablauf von Text Mining
- Wofür wird es verwendet?
- Text Mining an der HAW

Was ist Text Mining?

Text mining [is a] process in which a user interacts with a document collection over time by using a suite of analysis tools.
[...] text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. (Feldman & Sanger, 2007) [4]

- Sonderform des Data Mining
- Analyse von Textdokumenten

Unterschiede zum Datamining

- Vorbereitung der Daten
  - Erzeugen einer nutzbaren Struktur

- Herausforderungen
  - Wortstamm
  - Synonyme
  - Schreibfehler
  - Satzumformungen

Herausforderungen - Wortstämme

Reduzierung eines Wortes auf den Wortstamm

- Connection, connected, connecting -> connect
- ging, geht, gehst -> gehen

- *Stemming* - Entfernung von Pre / Suffixen

  ➢ **Porter Stemming(Entfernung von Suffixen) [2]**

- Lemmatization - Wörterbuch mit verschiedenen Formen

[2] Martin F. Porter - An algorithm for suffix stripping

Herausforderungen - Synonyme

- Synonyme / Semantische Gleichheit
  - Sonnabend -> Samstag
  - Handlung, Operation, Tat -> Aktion
  - wandeln, mutieren, wechseln -> ändern

Ablauf des Text Mining Processes

Ablauf nach Hajo Hippner & René Rentzmann

Abb. 1 Der Prozess des Text Mining

[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)

Aufbereitung

- Morphologische Analyse:
  - Stammformreduktion
  - Ziel: Reduzierung der Komplexität
- Syntaktischen Analyse
  - Annotation der Wörter
    - Einteilung in Wortarten (Verb, Adjektiv, Substantiv)
    - Basiert auf Lexikon sowie Wort-Kontext-Definition
  - Satzbau analyse
    - Einteilung in Subjekt, Prädikat, Objekt usw.
- Semantische Analyse
  - Verarbeitung von kontextuellen Wissen
    - Ist eine Bank ein Geldinstitut oder eine Sitzgelegenheit

[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)

(Text) Mining

- Klassifikation
  - Einordnung in vorgegebene Kategorien
- Segmentierung
  - Gruppierung mit ähnlichen Texten
- Abhängigkeitsanalyse
  - Analyse von gemeinsam auftretenden Termen

[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)

Anwendung der Ergebnisse

- Competitive Intelligence
  - Frühzeitige Entdeckung von Trends
    - Kunden Entwicklung
    - Marktentwicklung
    - Konkurrenz Entwicklung

[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)

Wofür wird Text Mining verwendet?

- Customer care
  - Automatische Support Antworten, Hilfe der Support Mitarbeiter
- Social Media
  - Erkennen von Stimmungsbildern aus Nutzerkommentaren
- Sicherheit / Anti Terror
  - Analyse von Blogs und Social Media
- Spam Erkennung

Software Lösungen

- Reine Text Mining Software
  - Clearforest
  - Inxight
- Indirekte Anbieter (Data Mining Software mit Text Mining support)
  - IBM SAS
- Software welche Text Mining inter nutzt
  - Fast (Suchtechnologie)
  - Verity (Information retrieval)

[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)

Text Mining - An der HAW

Joachim Schole: Strukturierung des Gegenstandsbereichs Whiskysorten mit Hilfe von Textmining

Marcel Schöneberg: Automatisierte Erstellung von Pressedossiers durch Textmining

Ivan Demin: Entwicklung einer Plattform für Second Screen Experimente

Ausblick: Masterstudium

Grundseminar: Text Mining

Grundprojekt: Anwendung v. Text Mining

Hauptseminar & Hauptprojekt: Machine / Deep - Learning im Ecommerce-Bereich

Masterarbeit: KI im E Commerce

Quellen

[1] Data Mining, Jürgen Cleve & Uwe Lämmel, 2014, De Gruyter Verlag
[2] An algorithm for suffix stripping, M.F.Porter, 1980, MCB UP Ltd
[3] Chung-Chian Hsu and Chien-Hsing Chen. 2010. Mining Synonymous Transliterations from the World Wide Web. 9, 1, Article 1 (March 2010)
[4] The Text Mining Handbook, Ronen Feldman & James Sanger, 2007, Cambridge University Press
[5] Text Mining, Hippner, H. & Rentzmann, R. Informatik Spektrum (2006)