Text Mining

Joachim Schole

Fakultät Technik und Informatik
Hochschule für angewandte Wissenschaften Hamburg

Grundseminar, WS 2014

Agenda

1. Motivation und Anwendungsfälle
   - Begriffserklärung
   - Anwendungsfälle
2. Schnittstellen/Relevante Arbeiten
3. Text-Mining-Prozess
4. Algorithmen und Anwendungsbeispiele
   - Naive Bayes
   - Strukturerkennung
5. Mögliche Projektaufgaben

Agenda

1. Motivation und Anwendungsfälle
   - Begriffserklärung
   - Anwendungsfälle
2. Schnittstellen/Relevante Arbeiten
3. Text-Mining-Prozess
4. Algorithmen und Anwendungsbeispiele
   - Naive Bayes
   - Strukturerkennung
5. Mögliche Projektaufgaben

Agenda

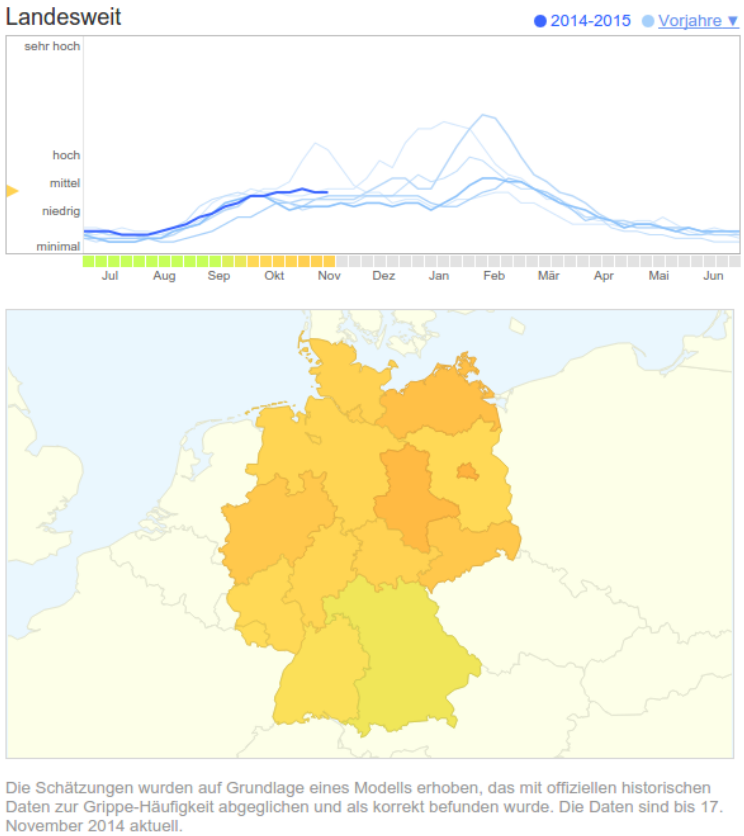
Begriffserklärung

- Automatische Extraktion von Informationen aus Texten
- Interdisziplinär - Techniken aus
  - Natural Language Processing
  - Künstliche Intelligenz
  - Information Retrieval/Extraction
- 80% der Daten liegen in Form von Text vor [HR06]
  - Dadurch kommerziell höheres Potential als Data Mining
- Persönliche Motivation durch Bachelorarbeit

Agenda

Anwendungsfälle

- Trenderkennung: Google Flutrends
- Unterstützung von Berufsgruppen
  - Robot Journalism [Roo14]
  - Juristen [MP14]
- Social Media Monitoring
- Ad-Placement/Targeting
- Sicherheit/Terror



Google Flutrends[1]

[1] www.google.org/flutrends

Agenda

1. Motivation und Anwendungsfälle
   - Begriffserklärung
   - Anwendungsfälle
2. Schnittstellen/Relevante Arbeiten
3. Text-Mining-Prozess
4. Algorithmen und Anwendungsbeispiele
   - Naive Bayes
   - Strukturerkennung
5. Mögliche Projektaufgaben

Schnittstellen/Relevante Arbeiten

- Second Screen - Ivan Demin
  - Generierung von Anfragen aus Dokumentationsbeschreibungen an Informationsquellen [Dem14]
- Trenderkennung - Marcel Schöneberg
  - Früherkennung von Twitter-Trends Anhand von Weak Signals [Sch14]
- Rezeptempfehlungen - Sigurd Sippel
  - Rezeptempfehlungen für Cocktails [Sip14]

Agenda

Text-Mining-Prozess

- Festlegung eines Ziels des Prozesses
- Was soll erreicht werden?

Text-Mining-Prozess

- Festlegung einer oder mehrerer Datenquelle(n)
- Sicherstellung der Anbindung

Text-Mining-Prozess

- Dokumente in weiterverwendbare Form bringen
- Stemming/Lemmatizing
- Stopwords entfernen
- Tagging von Satzteilen für syntaktische Analyse
- Auflösung von Doppeldeutigkeiten

Text-Mining-Prozess

- Anwendung von Analysemethoden auf die aufbereiteten Daten
- Stimmungsanalyse
- Dokumentvergleiche
- Automatische Textzusammenfassung

Text-Mining-Prozess

- Abschließende Interpretation der Analyseergebnisse
- Neu gewonnene Informationen

Text-Mining-Prozess

Text-Mining-Prozess, in Anlehnung an [HR06]

- Verwendung der Ergebnisse
- Unternehmensentscheidungen
- Ad-Placement/Targeting

Agenda

1. Motivation und Anwendungsfälle
   - Begriffserklärung
   - Anwendungsfälle
2. Schnittstellen/Relevante Arbeiten
3. Text-Mining-Prozess
4. Algorithmen und Anwendungsbeispiele
   - Naive Bayes
   - Strukturerkennung
5. Mögliche Projektaufgaben

Agenda

- Ordnet den einzelnen Wörtern eines Satzes die Wahrscheinlichkeiten zu, in einem positiven oder negativen Satz zu stehen
  - bildet Summen der Wahrscheinlichkeiten
  - höchster Wert bestimmt, ob der Satz positiv oder negativ ist
- Beispiel Erkennung von Spam-Accounts anhand des Namens [Fre13]
  - Statt ganzer Namen Buchstabenfolgen benutzt
  - Kann genau so auf Email-Adressen übertragen werden

Agenda

Strukturerkennung

Algorithmen und Anwendungsbeispiele

- Syntaktische Analyse / Part-of-Speech (PoS) Tagging
  - Versehen von Wörtern mit Tags, welche den Satzteil angeben
  - Ermöglicht tiefergehende Analysen
- Ermittlung von Stopwords (the, of, and, ...)
  - Normalerweise entfernt, da nicht relevant
  - Können verwendet werden, um Plagiate zu erkennen [Sta11]
    - Stopwords können schwer ersetzt werden, andere Begriffe dagegen schon
    - Gleiche Stopword-Strukturen können Hinweis auf Plagiat sein

Agenda

1. Motivation und Anwendungsfälle
   - Begriffserklärung
   - Anwendungsfälle
2. Schnittstellen/Relevante Arbeiten
3. Text-Mining-Prozess
4. Algorithmen und Anwendungsbeispiele
   - Naive Bayes
   - Strukturerkennung
5. Mögliche Projektaufgaben

Mögliche Projektaufgaben

- Wahrscheinlich im Bereich Digitaljournalismus
  - Automatisierte Dossier-Erstellung
  - Weiterführen des Second-Screen-Projekts
- Alternativ Robot Journalism möglich

Quellen I


Demin, Ivan
Text Mining for Second Screen. HAW Hamburg, 2014


Freeman, David Mandell
Using Naive Bayes to Detect Spammy Names in Social Networks.
http://doi.acm.org/10.1145/2517312.2517314
Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security, ACM, 2013


Hippner, Hajo; Rentzmann, René
Text Mining.
http://dx.doi.org/10.1007/s00287-006-0091-y
Informatik-Spektrum 29, 287-290, Springer-Verlag, 2006

Quellen II

McGinnis, John O.; Pearce, Russell G.
The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436937
82 Fordham Law Review 3041, 2014.

Roose, Kevin
Robots Are Invading the News Business, and It's Great for Journalists.
http://nymag.com/daily/intelligencer/2014/07/why-robot-journalism-is-great-for-journalists.html
New York Magazine, 11.07.2014, letzter Abruf am 20.11.2014.

Schöneberg, Marcel
Ansätze zur Trenderkennung in Texten . HAW Hamburg, 2014

Sippel, Sigurd
Recommendations for cocktail recipes. HAW Hamburg, 2014

Quellen III

Stamatatos, Efstathios

Plagiarism Detection Based on Structural Information.

http://doi.acm.org/10.1145/2063576.2063754

Proceedings of the 20th ACM International Conference on Information and Knowledge Management, ACM, 2011