Text Mining für News-Sites

Nina Hälker

Department Informatik, HAW Hamburg
Sommersemester 2014

Ablauf

A Motivation

- Aufbauprojekt „Was sagt das Ausland?“
- Fokus der Masterarbeit: Text Mining für News-Sites

B Drei Papers: Fokus, Ergebnisse, eigenes Fazit

- Differentiating Data- and Text-Mining Terminology
- Blogosphere: Research Issues, Tools, and Applications
- Open Domain Extraction from Twitter

C Wie weiter?

Text Mining für News-Sites

Nina Hälker

Literatur / Papers

Nitin Agarwal and Huan Liu. 2008. Blogosphere: research issues, tools, and applications. SIGKDD Explor. Newsl. 10, 1 (May 2008), 18-31

Jan H. Kroeze, Machdel C. Matthee, and Theo J. D. Bothma. 2003. Differentiating data- and text-mining terminology. In Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology (SAICSIT '03), Jarr Eloff, Andries Engelbrecht, Paula Kotzé, and Mariki Eloff (Eds.). South African Institute for Computer Scientists and Information Technologists, Republic of South Africa, 93-101

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 1104-1112

Text Mining für News-Sites

Nina Hälker

Motivation

Aufbauprojekt:

„Was sagt das Ausland?“ – Extrahieren und Verknüpfen von Artikeln aus News-Sites weltweit

Planung Masterarbeit:

Text Mining bei unstrukturierten Daten – Auf der Grundlage eines englischsprachigen Artikelarchivs

Text Mining für News-Sites

Nina Hälker

Aufbauprojekt: „Was sagt das Ausland?“

Ausgangspunkt: News-Sites werden international unverbunden präsentiert.

Interesse:

Extrahieren und Verknüpfen thematisch ähnlicher Artikel durch

→ Text-Mining Verfahren

?

Text Mining für News-Sites

Nina Hälker

Masterarbeit: Text Mining für News-Sites

Ausgangspunkt: Textkorpus mit ca. 3.500 englischsprachigen Texten

Interesse:

Aufbereitung des Archivs durch

→ New Storytelling

→ Text-Mining Verfahren

?

Text Mining für News-Sites

Nina Hälker

Paper 1: Differentiating Data- and Text-Mining Terminology (Kroeze, Matthee, Bothma. 2003)

Definition: „The discovery of knowledge from databases sources containing free text is called text mining.“

Ausgangspunkt: Durch die rasche Expansion des Webs werden Methoden des Text Mining wichtiger und schwieriger.

Aufgabe: „Tell me something I didn't know but would like to know.“

Text Mining für News-Sites

Nina Hälker

Paper 1: Differentiating Data- and Text-Mining Terminology (Kroeze, Matthee, Bothma. 2003)

„If **non-novel** text-mining investigation ist **information retrieval**,

and if **semi-novel** text-mining investigation is **knowledge discovery**,

then **novel** text-mining investigation should be **knowledge creation**.“*

Creating new knowledge** → **Intelligent Text Mining**

* vgl. Kroeze, Matthee, Bothma 2003

** durch Text-Mining-Verfahren

Text Mining für News-Sites

Nina Hälker

Paper 1: Differentiating Data- and Text-Mining Terminology (Kroeze, Matthee, Bothma. 2003)

Funktion des N**atural Language Processing** (NLP) im Intelligent Text Mining:

- **inhärente Struktur von Texten** aufdecken und darunter liegende **linguistische Strukturen** zu erforschen, um daraus syntaktische und semantische Darstellungen des Textes zu ermöglichen

Fazit:

Natural Language Processing im Text Mining nutzen, um

- Diskurse bzw. Diskussionsverläufe zu entdecken
- Struktur der Texte jenseits ihres offensichtlichen Gehalts zu untersuchen (Popkultur versus Wissenschaft, Osteuropa- versus Westeuropa-“Sprech“, ...)

Text Mining für News-Sites

Nina Hälker

Paper 2: Blogosphere: Research Issues, Tools, and Applications (Argawal, Liu. 2008)

Definition: Blogosphere – Das Blog-Universum (bestehend aus single-author- und multi-authored Blogs)

Ausgangspunkt: Mit dem Web 2.0 wurden aus Content-Konsumenten Content-Produzenten

→ Open source intelligence

→ Zunahme kollektiver Wissensproduktion

Frage: Wie kann dieses kollektive Wissen geborgen werden?

Text Mining für News-Sites

Nina Hälker

Paper 2: Blogosphere: Research Issues, Tools, and Applications (Argawal, Liu. 2008)

Forschungsfacetten (u.a.):

- Modeling the Blogosphere
- Blog Clustering
- Community discovery
- Influence in Blogs
- Trust and Reputation

Text Mining für News-Sites

Nina Hälker

Paper 2: Blogosphere: Research Issues, Tools, and Applications (Argawal, Liu. 2008)

Fazit:

- tf-idf (*top three most famous words* eines jeden Artikels extrahieren) für Aufbereitung von Zeitungs-Archive nutzbar?

  Frage: wonach bemisst sich *top three most famous*?

- Influencer-Untersuchungen für die Aufbereitung von Archiven auch hinsichtlich New Storytelling von Interesse? (Wer sind die network-mover?, Was sind die Themen – und wie haben sie sich entwickelt?)

Text Mining für News-Sites

Nina Hälker

Paper 3: Open Domain Event Extraction from Twitter (Ritter, Mausam, Etzioni, Clark. 2012)

Definition: Open Domain = Daten sind auslesbar

Ausgangspunkt: Tweets liefern aktuellste Informationen und Kommentare über/zu gerade stattfindenden Events.

Herausforderungen:

- Informationen sind oft zerstückelt und ungenau („noisy“)
- Unwichtige Informationen und verknappte Sprache müssen erkannt und herausgefiltert werden
- Es gibt viele unerwartbare und unplanbare Themen, denn: Twitterer schreiben über alles

Text Mining für News-Sites

Nina Hälker

Paper 3: Open Domain Event Extraction from Twitter (Ritter, Mausam, Etzioni, Clark. 2012)

Möglichkeiten:

- Tweets sind kurz, einfach geschrieben, pragmatisch strukturiert
- Volumen der anfallenden Tweets viel höher als bei z.B. Newssites

→ Struktur von Tweets und inhaltliche Redundanz in Form von Dopplungen unterstützen die Auswertungsmöglichkeiten

Frage: Was muss ein Werkzeug leisten, das Events extrahieren, verbinden, kategorisieren (und im besten Fall auch noch die Qualität der Events bemessen) kann?

Text Mining für News-Sites

Nina Hälker

Paper 3: Open Domain Event Extraction from Twitter (Ritter, Mausam, Etzioni, Clark. 2012)

Studie über **TwiCal:** dem ersten o*pen domain event-extraction and categorization system* für Twitter

- 4-Tupel representation of events: entity, event phrase, date, type, weil Tweets oft genau diese Aspekte enthalten:

Entity	Event Phrase	Date	Type
Steve Jobs	died	10/06/11	Death
iPhone	announcement	10/04/11	Product launch
Amanda Knox	verdict	10/03/11	Trial

Vgl. Ritter, Mausam, Etzioni, Clark 2012

Text Mining für News-Sites

Nina Hälker

Paper 3: Open Domain Event Extraction from Twitter (Ritter, Mausam, Etzioni, Clark. 2012)

Verfahren (u.a.):

- Tweets werden POS-getaggt („part of speech“-Taggen als ein Abstraktionsschritt)
- Begriffsentitäten und Event-Phrases werden extrahiert, zeitliche Ausdrücke beseitigt, Events in Typen kategorisiert

→ Aussagekraft eines Tweets berechnet sich durch die (berechnete) Stärke der Verbindung zwischen jeder Begriffsentität, dem Datum und der Anzahl der gesendeten Tweets

Text Mining für News-Sites

Nina Hälker

Paper 3: Open Domain Event Extraction from Twitter (Ritter, Mausam, Etzioni, Clark. 2012)

Fazit:

- Entsprechend der Spezifika von Twitter Spezifika von für den Artikelkorpus herausarbeiten (Herausforderungen, Möglichkeiten)
- POS-Tagger für Auslesen von News-Sites nutzen
- Algorithmus erarbeiten für die Aussagekraft von Artikeln, z.B.

$$\frac{\text{Keywords Titel/Text}}{\text{Anzahl Beiträge/Autor}} \times \frac{\text{Sichtbarkeit des Themas in News}}{\text{Veröffentlichungsdatum}}$$

Text Mining für News-Sites

Nina Hälker

Wie weiter?

Text Mining für News-Sites

Nina Hälker

Soweit für heute.

Danke.

Gibt es Fragen?