

Text-Mining – Eine Einführung

Projekt FIDIUS

Fynn Feldpausch Dominik Menke

Universität Bremen, Fachbereich 3

31. Oktober 2009

Überblick

- 1 Grundlagen
- 2 Repräsentation und Modellierung
- 3 Klassifikation und Clusterbildung
- 4 Zusammenfassung

Grundlagen

Begriffsdefinition

»[Jede] Operation, die sich mit dem Sammeln und der Analyse von Text auseinandersetzt.«

— Carsten Siegmund, nach Richard Hackathorn, 11/1998

Begriffsdefinition

»[Jede] Operation, die sich mit dem Sammeln und der Analyse von Text auseinandersetzt.«

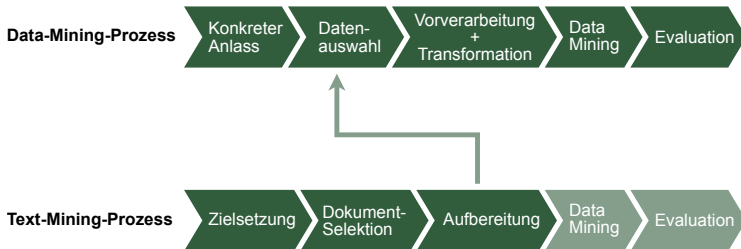
— Carsten Siegmund, nach Richard Hackathorn, 11/1998

»Text Mining is the discovery [. . .] of new, previously unknown information, by automatically extracting information from different written resources.«

— Marti Hearst, 10/2003

Abgrenzung zu Data-Mining

- Datenbasis für Data-Mining (atomar) strukturierte Datenbanken
- Datenbasis für Text-Mining natürlichsprachliche Texte
- Data-Mining kann Teil des Text-Minings sein:



Methoden

■ Dokumentaufbereitung

- morphologische Analyse (z.B. Stammformreduktion)
- syntaktische Analyse (z.B. *Part of Speech Tagging*)
- semantische Analyse (z.B. *Word Sense Disambiguation*)

Methoden

■ Dokumentaufbereitung

- morphologische Analyse (z.B. Stammformreduktion)
- syntaktische Analyse (z.B. *Part of Speech Tagging*)
- semantische Analyse (z.B. *Word Sense Disambiguation*)

■ Text-Mining-Methoden

- Informationsextraktion
- Verfolgung von Themen (*Topic Tracking*)
- Zusammenfassen (*Summarization*)
- Kategorisieren (*Categorization*)
- Clusterbildung (*Clustering*)
- Informationsvisualisierung (*Information Visualizing*)
- Frage-Antwort-Systeme (*Question Answering*)

Anwendungsszenarien

■ Wissenschaft

- Herstellen von (bisher unbekanntem) Zusammenhängen
- Analyse gesellschaftlicher Problemfelder
- ...

Anwendungsszenarien

■ Wissenschaft

- Herstellen von (bisher unbekanntem) Zusammenhängen
- Analyse gesellschaftlicher Problemfelder
- ...

■ Wirtschaft

- Analyse von Käufer- und Wählerverhalten
- Auswertung von Newsgroups
- Spam-Filtering
- Logfile-Analyse
- Kommunikationsüberwachung
- Marktanalyse
- ...

Repräsentation und Modellierung

Bag-of-Words

- Wortreihenfolge für *echtes* Verständnis von Bedeutung
- Reihenfolge kann für grobe Aufgaben vernachlässigt werden

Bag-of-Words

- Wortreihenfolge für *echtes* Verständnis von Bedeutung
- Reihenfolge kann für grobe Aufgaben vernachlässigt werden

- Dokumente werden als *Bag-of-Words* interpretiert
 - vorhandene Wortmenge identifizieren
 - Vokabular V bilden
 - ggf. Größe $m = |V|$ reduzieren
 - beliebige Ordnung auf Worten festlegen

Bag-of-Words

- Wortreihenfolge für *echtes* Verständnis von Bedeutung
- Reihenfolge kann für grobe Aufgaben vernachlässigt werden

- Dokumente werden als *Bag-of-Words* interpretiert
 - vorhandene Wortmenge identifizieren
 - Vokabular V bilden
 - ggf. Größe $m = |V|$ reduzieren
 - beliebige Ordnung auf Worten festlegen

- oft werden Stopp-Worte eliminiert
 - Pronomen
 - Prepositionen
 - ...

Vektorrepräsentation

- Dokument wird als Vektor x repräsentiert
- x besteht aus $m = |V|$ ganzzahligen Einträgen
- j -te Komponente x_j von x gibt Auftrittshäufigkeit des Wortes j an

Vektorrepräsentation

- Dokument wird als Vektor x repräsentiert
- x besteht aus $m = |V|$ ganzzahligen Einträgen
- j -te Komponente x_j von x gibt Auftrittshäufigkeit des Wortes j an
- Länge des Dokumentes ist $n = \sum_{j=1}^m x_j$
- typischerweise gilt:
 - n ist viel kleiner als m
 - $x_j = 0$ für die meisten j

Vektorrepräsentation

- Dokument wird als Vektor x repräsentiert
- x besteht aus $m = |V|$ ganzzahligen Einträgen
- j -te Komponente x_j von x gibt Auftrittshäufigkeit des Wortes j an
- Länge des Dokumentes ist $n = \sum_{j=1}^m x_j$
- typischerweise gilt:
 - n ist viel kleiner als m
 - $x_j = 0$ für die meisten j
- Dokumentkollektionen lassen sich als zweidimensionale Matrix repräsentieren

Multinomialverteilung

- Modellierung der Dokumentmengen als Wahrscheinlichkeitsverteilung:

$$p(x; \theta) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j} \quad \text{mit} \quad \sum_{j=1}^m \theta_j = 1$$

Multinomialverteilung

- Modellierung der Dokumentmengen als Wahrscheinlichkeitsverteilung:

$$p(x; \theta) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j} \quad \text{mit} \quad \sum_{j=1}^m \theta_j = 1$$

- Beide Vektoren x und θ besitzen die Länge m
 - θ_j ist die Wahrscheinlichkeit des Wortes j
 - x_j ist die Auftrittshäufigkeit des Wortes j

Multinomialverteilung

- Modellierung der Dokumentmengen als Wahrscheinlichkeitsverteilung:

$$p(x; \theta) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j} \quad \text{mit} \quad \sum_{j=1}^m \theta_j = 1$$

- Beide Vektoren x und θ besitzen die Länge m
 - θ_j ist die Wahrscheinlichkeit des Wortes j
 - x_j ist die Auftrittshäufigkeit des Wortes j
- Parameter θ so wählen, dass Dokumente aus Trainingsmenge hohe Wahrscheinlichkeiten erhalten
- weitere Dokumente gegen dieses Modell testen

Multinomialverteilung

- Summe der Verteilung über alle Dokumente muss 1 ergeben
- Anzahl der Dokumente wächst exponentiell zu ihrer Länge: m^n
⇒ Einzelwahrscheinlichkeiten sehr klein

Multinomialverteilung

- Summe der Verteilung über alle Dokumente muss 1 ergeben
- Anzahl der Dokumente wächst exponentiell zu ihrer Länge: m^n
⇒ Einzelwahrscheinlichkeiten sehr klein
- Berechnung mit logarithmischen Wahrscheinlichkeiten:

$$p(x; \theta) = \frac{n!}{\prod_{j=1}^m x_j!} \prod_{j=1}^m \theta_j^{x_j}$$

⇓

$$\log p(x; \theta) = \log n! - \left[\sum_{j=1}^m \log x_j! \right] + \left[\sum_{j=1}^m x_j \cdot \log \theta_j \right]$$

Multinomialverteilung

- Parameter ergeben sich durch $\theta_j = \frac{1}{T} \sum_x x_j$
(Summe der Trainingsdokumente)
- Normalisierungskonstante $T = \sum_x \sum_j x_j$
(Summe der Längen aller Trainingsdokumente)

Multinomialverteilung

- Parameter ergeben sich durch $\theta_j = \frac{1}{T} \sum_x x_j$
(Summe der Trainingsdokumente)
- Normalisierungskonstante $T = \sum_x \sum_j x_j$
(Summe der Längen aller Trainingsdokumente)
- Wahrscheinlichkeiten gleich 0 sind unerwünscht
- $\theta_j > 0$ für alle j erwünscht
- Parameter ergeben sich durch:

$$\theta_j = \frac{1}{T'} \left(c + \sum_x x_j \right)$$

Klassifikation und Clusterbildung

Klassifikation vs. Clusterbildung

■ Klassifikation

- Zuordnung eines Dokuments zu einer Kategorie durch inhaltliche Analyse
- Beispieldokumente als Trainingsbasis

Klassifikation vs. Clusterbildung

■ Klassifikation

- Zuordnung eines Dokuments zu einer Kategorie durch inhaltliche Analyse
- Beispieldokumente als Trainingsbasis

■ Clusterbildung

- Einteilung von Dokumenten in unterschiedsarme »Cluster«
- vollautomatischer Prozess
- keine vordefinierten Kategorien

Bayes-Klassifikatoren

- Zuordnung eines Objektes zu einer Klasse mit der höchsten Wahrscheinlichkeit bzw. den geringsten Kosten
- Basiert auf Satz von Bayes:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

Bayes-Klassifikatoren

- Zuordnung eines Objektes zu einer Klasse mit der höchsten Wahrscheinlichkeit bzw. den geringsten Kosten
- Basiert auf Satz von Bayes:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

- x Beispieldokument
- y Klasse des Beispieldokuments
(Klassen numeriert von 1 bis K)

Bayes-Klassifikatoren

- Berechnung von $p(y = k|x)$:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

Bayes-Klassifikatoren

- Berechnung von $p(y = k|x)$:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

$p(x|y = k)$ beliebige Verteilung mit trainierten Parametern

Bayes-Klassifikatoren

- Berechnung von $p(y = k|x)$:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

$p(x|y = k)$ beliebige Verteilung mit trainierten Parametern

$p(y = k)$ Berechnung durch $n_k / \sum_{k=1}^K n_k$,

wobei n_k Anzahl der Trainingsdokumente der Klasse k

Bayes-Klassifikatoren

- Berechnung von $p(y = k|x)$:

$$p(y = k|x) = \frac{p(x|y = k)p(y = k)}{p(x)}$$

$p(x|y = k)$ beliebige Verteilung mit trainierten Parametern

$p(y = k)$ Berechnung durch $n_k / \sum_{k=1}^K n_k$,

wobei n_k Anzahl der Trainingsdokumente der Klasse k

$p(x)$ Berechnung durch $\sum_{k=1}^K p(x|y = k)p(y = k)$

Häufungen

- Multinomialverteilung nimmt für ein Wort j stets Wahrscheinlichkeit θ_j an

Häufungen

- Multinomialverteilung nimmt für ein Wort j stets Wahrscheinlichkeit θ_j an
- Worte treten allerdings oft in Häufungen auf:

Deutsche Autobauer Vorreiter bei flexibler Produktion

*Lange war **Toyota** das Vorbild für effiziente Autoproduktion. Vorreiter dieser Entwicklung ist mittlerweile allerdings nicht mehr **Toyota**, sondern Porsche. Anfang der Neunzigerjahre, als Porsche kurz vor der Pleite stand, holte man sich dort **Toyota**-Manager als externe Berater zur Rettung ins Boot. Porsche kopierte das **Toyota**-Konzept nicht einfach, sondern verfolgte eine eigene Strategie. Im Gegensatz zu **Toyota** machen die Zuffenhausener nicht alles im eigenen Haus, sondern sie lassen extern produzieren und setzen auf Entwicklungskooperationen.*

<http://www.heise.de/newsticker/meldung/>

Deutsche-Autobauer-Vorreiter-bei-flexibler-Produktion-218705.html

DCM – *Dirichlet compound multinomial*

- Urne mit Kugeln in $|V|$ verschiedenen Farben
- Zufallsziehung mit Zurücklegen
- Hinzufügen einer Kugel gleicher Farbe

DCM – *Dirichlet compound multinomial*

- Urne mit Kugeln in $|V|$ verschiedenen Farben
- Zufallsziehung mit Zurücklegen
- Hinzufügen einer Kugel gleicher Farbe

- initiale Menge an Kugeln mit Farbe j ist β_j
- Parametervektor β der Länge $|V|$ keiner Beschränkung unterworfen
- zusätzlicher Freiheitsgrad modelliert Häufungen
- je kleiner die Parameter β_j , desto mehr Worte neigen zu Häufungen

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

K Anzahl der Komponenten der Mischverteilung

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

K Anzahl der Komponenten der Mischverteilung

$p(x; \theta_k)$ Verteilung der Komponente k

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

- K Anzahl der Komponenten der Mischverteilung
- $p(x; \theta_k)$ Verteilung der Komponente k
- α_k Anteil der Komponente k

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

- K Anzahl der Komponenten der Mischverteilung
- $p(x; \theta_k)$ Verteilung der Komponente k
- α_k Anteil der Komponente k

Clusterbildung

- Anpassung einer Mischverteilung an gegebene Menge von Dokumenten:

$$p(x) = \sum_{k=1}^K \alpha_k p(x; \theta_k)$$

- K Anzahl der Komponenten der Mischverteilung
- $p(x; \theta_k)$ Verteilung der Komponente k
- α_k Anteil der Komponente k

- jede Komponente ist ein Cluster
- ein Cluster z.B. ein Themengebiet

Topic Models

- Clustering erlaubt lediglich Zuordnung zu einem Themengebiet
- Zuordnung zu mehreren Themengebieten meist plausibler
- *Topic Models* machen diese Annahme

Topic Models

- Clustering erlaubt lediglich Zuordnung zu einem Themengebiet
- Zuordnung zu mehreren Themengebieten meist plausibler
- *Topic Models* machen diese Annahme

- LDA (*Latent Dirichlet allocation*) weit verbreitetes Topic Model
- jedes Wort kann mehreren Themen zugeordnet werden
- Themen sind dokumentübergreifend identisch
- Anteil der Themen in Dokumenten verschieden

Zusammenfassung

Zusammenfassung

- Methoden des Text-Mining
Analyse, Kategorisieren, Clusterbildung, ...

Zusammenfassung

- Methoden des Text-Mining
Analyse, Kategorisieren, Clusterbildung, ...
- Anwendung in Wissenschaft und Wirtschaft
Zusammenhänge erkennen, Spam-Filterung, Marktanalyse, ...

Zusammenfassung

- **Methoden des Text-Mining**
Analyse, Kategorisieren, Clusterbildung, ...
- **Anwendung in Wissenschaft und Wirtschaft**
Zusammenhänge erkennen, Spam-Filterung, Marktanalyse, ...
- **Repräsentation und Modellierung von Dokumenten**
Bag-of-Words, Vektorrepräsentation, Multinomialverteilung, ...

Zusammenfassung

- **Methoden des Text-Mining**
Analyse, Kategorisieren, Clusterbildung, ...
- **Anwendung in Wissenschaft und Wirtschaft**
Zusammenhänge erkennen, Spam-Filterung, Marktanalyse, ...
- **Repräsentation und Modellierung von Dokumenten**
Bag-of-Words, Vektorrepräsentation, Multinomialverteilung, ...
- **Zuordnung von Dokumenten zu Kategorien und Clusterbildung**
Bayes-Klassifikatoren, Mischverteilung, Topic Models ...

Zusammenfassung

- **Methoden des Text-Mining**
Analyse, Kategorisieren, Clusterbildung, ...
- **Anwendung in Wissenschaft und Wirtschaft**
Zusammenhänge erkennen, Spam-Filterung, Marktanalyse, ...
- **Repräsentation und Modellierung von Dokumenten**
Bag-of-Words, Vektorrepräsentation, Multinomialverteilung, ...
- **Zuordnung von Dokumenten zu Kategorien und Clusterbildung**
Bayes-Klassifikatoren, Mischverteilung, Topic Models ...
- **Umgang mit Worthäufungen**
Häufungswahrscheinlichkeiten, DCM, ...

Zusammenfassung

- Methoden des Text-Mining
Analyse, Kategorisieren, Clusterbildung, ...
- Anwendung in Wissenschaft und Wirtschaft
Zusammenhänge erkennen, Spam-Filterung, Marktanalyse, ...
- Repräsentation und Modellierung von Dokumenten
Bag-of-Words, Vektorrepräsentation, Multinomialverteilung, ...
- Zuordnung von Dokumenten zu Kategorien und Clusterbildung
Bayes-Klassifikatoren, Mischverteilung, Topic Models ...
- Umgang mit Worthäufungen
Häufungswahrscheinlichkeiten, DCM, ...

Gibt es Fragen?

Quellen

- C. Siegmund: Einführung in Text Mining. In: *Text Mining – Wissensgewinnung aus natürlichsprachlichen Dokumenten*, Universität Karlsruhe, R. Witte und J. Mülle [Hrsg.], 2006.
- M.A. Hearst. What is Text Mining?
<http://www.sims.berkeley.edu/~hearst/text-mining.html>, 2003.
- H. Hippner und R. Rentzmann: Text Mining.
http://www.gi-ev.de/no_cache/service/informatiklexikon/informatiklexikon-detailansicht/meldung/text-mining-137.html, GI e.V. [Hrsg.], 2006.
- L. Gotter. Text Mining – Wissensgewinnung aus Texten.
<http://www.wissensexploration.de/textmining.php>, 2008.