Text Mining

OS Datamining SS 10

Thomas Boy

25. Mai 2010

1 Gliederung
2 Einleitung
- Motivation
- Konkretisierung

3 Allgemeines
- Definiton Text Mining
- Ablaufschema

4 Anwendungen
- funktionale Anwendungen

5 Bedeutungsanalyse
- Verarbeitung des Rohtextes
- Grundlagen
- Kookkurrenten

6 Verfahren
- Differenzanalyse
- Clustering
- Musteranalyse

Abbildung: Quelle: [DatFlut]

"It has been estimated that the amount of Information in the world doubles every 20 months" [ArcKnow]

"Im Internet veranschlagt man einen Zuwachs von ca. 1 Million neuer Dokumente pro Tag !" [WiRo08]

"It has been estimated that the amount of Information in the world doubles every 20 months" [ArcKnow]

"Im Internet veranschlagt man einen Zuwachs von ca. 1 Million neuer Dokumente pro Tag !" [WiRo08]

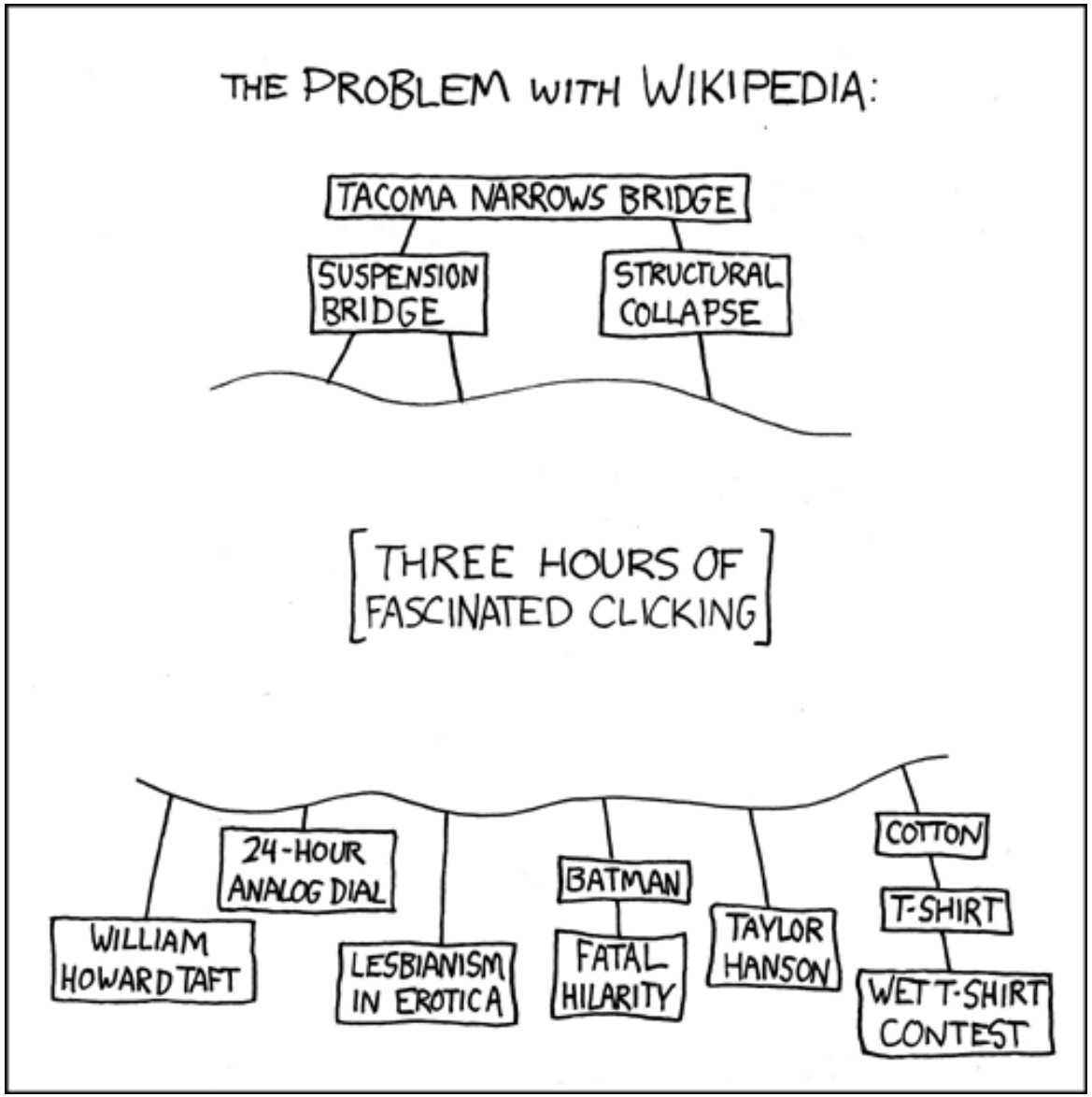


Abbildung: Quelle: [xkcd]

Problem

- große Menge an textuellen Daten
- untschiedliches Format
- geringe bis kein Struktur der Texte

Lösung - Text Mining

- Werkzeuge zur Strukturierung der Daten
- Aufdecken von Zusammenhängen in und zwischen Texten
- ansprechende Darstellung neuer und relevanter Informationen
- Verfahren zur Kategorisierung von Texten

Problem

- große Menge an textuellen Daten
- untschiedliches Format
- geringe bis kein Struktur der Texte

Lösung - Text Mining

- Werkzeuge zur Strukturierung der Daten
- Aufdecken von Zusammenhängen in und zwischen Texten
- ansprechende Darstellung neuer und relevanter Informationen
- Verfahren zur Kategorisierung von Texten

Definition Text Mining nach [TeMiHa08]

„Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous through the identifikation and exploration of interesting patterns.“ [TeMiHa08, S.1]

Definition Text Mining nach [WiRo08]

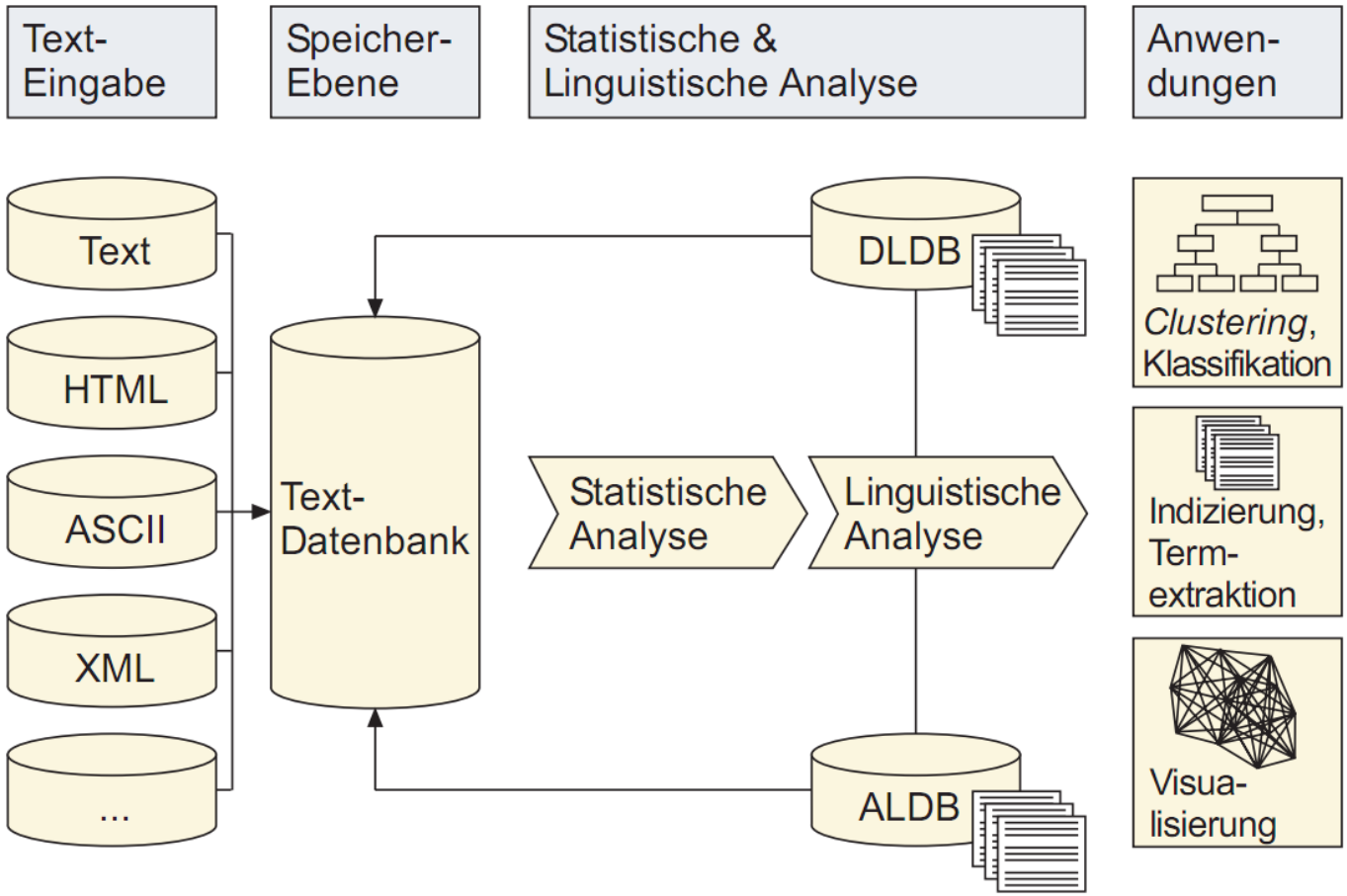
„Mit dem Terminus **Text Mining** werden computergestützte Verfahren für die semantische Analyse von Texten bezeichnet, welche die automatische bzw. semi-automatische **Strukturierung** von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.“ [WiRo08, S. 3]

Definition Text Mining nach [TeMiHa08]

„Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous through the identifikation and exploration of interesting patterns.“ [TeMiHa08, S.1]

Definition Text Mining nach [WiRo08]

„Mit dem Terminus **Text Mining** werden computergestützte Verfahren für die semantische Analyse von Texten bezeichnet, welche die automatische bzw. semi-automatische **Strukturierung** von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.“ [WiRo08, S. 3]



Legende:
DLDB = Domänenspezifische linguistische Datenbank
ALDB = Allgemeine linguistische Datenbank (Referenzwortschatz)

Abbildung: grundlegendes Ablaufschema nach [WiRo08]

funktionale Anwendungen

Text Mining dient dazu :

- Identifikation relevanter fachspezifischer Ausdrücke
- semantische Relationen zwischen einzelnen Ausdrücken berechnen und damit inhaltliche Strukturen in Texten offenzulegen
- Ähnlichkeiten zwischen Begriffen zu finden
- ähnliche Dokumente zu finden
- Definitionen, Erläuterungen und Referenzen in Texten aufzufinden

praktische Anwendung am Beispiel **Unternehmen** :

- effiziente und hochselektive Recherche in Textbeständen
- automatische Filterung von Nachrichten (Bsp. E-Mails gruppieren, Spam identifizieren)

Abbildung: Quelle: [Abendblatt]

- automatischer Aufbau von firmen- oder fachspezifischen Glossaren und Thesauren
- teilautomatische Erstellung von semantischen Netzen für das Wissensmanagment

praktische Anwendung am Beispiel **Unternehmen** :

- effiziente und hochselektive Recherche in Textbeständen
- automatische Filterung von Nachrichten (Bsp. E-Mails gruppieren, Spam identifizieren)

Abbildung: Quelle: [Abendblatt]

- automatischer Aufbau von firmen- oder fachspezifischen Glossaren und Thesauren
- teilautomatische Erstellung von semantischen Netzen für das Wissensmanagment

Arbeitsschritte:

- Konvertierung der Quelldokumente in „reine“ Texte
- Segmentierung des Textes auf verschiedenen linguistischen Ebenen (Sätze, Phrasen und Wörter)
- Herausfiltern von Stoppwörtern (optional)
- Bildung von Wortstämmen (Stemming, Lemmatisierung) Beispiele:
  - lachte ▷ lach (Stemming)
  - lachte ▷ lachen (Lemmatisierung)
  - bekannnter Stemming Algorithmus für englische Sprache: Porter Stemmer
  - deutsche Sprache meist lexikon-basierte Ansätze
- Zuordnung von Wortarten (POS-Tagging) mit Hilfe des Hidden-Markov-Modell
- Einpflegen der Daten in Textdatenbank

Arbeitsschritte:

- Konvertierung der Quelldokumente in „reine“ Texte
- Segmentierung des Textes auf verschiedenen linguistischen Ebenen (Sätze, Phrasen und Wörter)
- Herausfiltern von Stoppwörtern (optional)
- Bildung von Wortstämmen (Stemming, Lemmatisierung) Beispiele:
  - lachte ▷ lach (Stemming)
  - lachte ▷ lachen (Lemmatisierung)
  - bekannnter Stemming Algorithmus für englische Sprache: Porter Stemmer
  - deutsche Sprache meist lexikon-basierte Ansätze
- Zuordnung von Wortarten (POS-Tagging) mit Hilfe des Hidden-Markov-Modell
- Einpflegen der Daten in Textdatenbank

Ziel der Bedeutungsanlayse

- das Wissen, welches im Text enthalten ist, extrahieren
- den Inhalt den der Textes repräsentiert aus Wörtern und Sätzen ableiten

Grundlage bilden **Relationen** :

- **syntagmatische** Relation
  - gemeinsames Auftreten zweier Wortformen in einem *Text*
  - exemplarisches Beispiel: „Das *schöne Wetter* in Leipzig.“
  - Betrachtung von Wörtern in einem **lokalen** Kontext (Satz oder linker, rechter Nachbar)
  - gemeinsames Auftreten mit gewissem Signifikanzmaß führt zum Begriff **signifikante Kookkurrenten**
  - weitere Beispiel: Aufzählungen, feste Wendungen, Mehrfachwortbegriffe

Ziel der Bedeutungsanlayse

- das Wissen, welches im Text enthalten ist, extrahieren
- den Inhalt den der Textes repräsentiert aus Wörtern und Sätzen ableiten

Grundlage bilden **Relationen** :

- **syntagmatische** Relation
  - gemeinsames Auftreten zweier Wortformen in einem *Text*
  - exemplarisches Beispiel: „Das *schöne Wetter* in Leipzig.“
  - Betrachtung von Wörtern in einem **lokalen** Kontext (Satz oder linker, rechter Nachbar)
  - gemeinsames Auftreten mit gewissem Signifikanzmaß führt zum Begriff **signifikante Kookkurrenten**
  - weitere Beispiel: Aufzählungen, feste Wendungen, Mehrfachwortbegriffe

weitere Arten von **Relationen** :

- **paradigmatische** Relation
  - gemeinsames Auftreten von zwei Wortformen in einem ähnlichen *Kontext*

Beispiel

In der Satzform „Die X scheint“ werden nur Belegungen zugelassen, die meist mit dem Verb „scheinen“ (in der Bedeutung „Licht aussenden“) gemeinsam auftreten:
„Sonne“,“Lampe“,“Kerze“,“Laterne“,...

- Betrachtung von Wörtern in einem **globalen** Kontext (Menge aller signifikante Kookkurrenten)

weitere Arten von **Relationen** :

- **paradigmatische** Relation
  - gemeinsames Auftreten von zwei Wortformen in einem ähnlichen *Kontext*

Beispiel

In der Satzform „Die X scheint“ werden nur Belegungen zugelassen, die meist mit dem Verb „scheinen“ (in der Bedeutung „Licht aussenden“) gemeinsam auftreten:
„Sonne“,“Lampe“,“Kerze“,“Laterne“,...

  - Betrachtung von Wörtern in einem **globalen** Kontext (Menge aller signifikante Kookkurrenten)

weitere Arten von **Relationen** :

- **semantische** Relation
  - nur semantische Relation, wenn syntagmatische oder paradigmatische Relation
  - Beispiel für semantische Relationen benachbarter Wortformen: Kategorie bzw. Funktionsangabe, Maßeinheit oder Qualifizierung wie Teil-von-Beziehungen, Instrument-für-Beziehung, Ober- Unterbegriff
  - oftmals Analyse linker, rechter Nachbarn einer Wortform
  - Benutzung von Mustern
  - Beispiel Kategorie- oder Funktionsangabe:

⟨*NOMEN*⟩	⟨*EIGENNAMEN*⟩
Stadt	Leipzig
Stadt	Hamburg
Stadt	München
Bundeskanzler	Schröder
Ministerpräsident	Schröder
Parteivorsitzender	Schröder

weitere Arten von **Relationen** :

- **semantische** Relation
  - nur semantische Relation, wenn syntagmatische oder paradigmatische Relation
  - Beispiel für semantische Relationen benachbarter Wortformen: Kategorie bzw. Funktionsangabe, Maßeinheit oder Qualifizierung wie Teil-von-Beziehungen, Instrument-für-Beziehung, Ober- Unterbegriff
  - oftmals Analyse linker, rechter Nachbarn einer Wortform
  - Benutzung von Mustern
  - Beispiel Kategorie- oder Funktionsangabe:

$\langle NOMEN \rangle$	$\langle EIGENNAMEN \rangle$
Stadt	Leipzig
Stadt	Hamburg
Stadt	München
Bundeskanzler	Schröder
Ministerpräsident	Schröder
Parteivorsitzender	Schröder

signifikante Kookkurrenten

- Idee: häufiges gemeinsames Auftreten von Wörter ▷ semantische Beziehung der Wörter
- Unterscheidung: Nachbarschaftskookkurrenten, Satzkookkurrenten
- das Signifikanzmaß soll dem intuitiven Gefühl von Zusammengehörigkeit von Wortformen entsprechen
- Beispiel: Polizei - verhaftet, berittene - Polizei

Abbildung: Quelle: [Police]

- Berechnung eines Signifikanzwertes

signifikante Kookkurrenten:

Berechnung eines Signifikanzwertes

- a,b: Anzahl der Sätze die A, B enthalten
- k: Anzahl Sätze die A *und* B enthalten
- n: Gesamtanzahl Sätze
- $\lambda = \frac{a \cdot b}{n}$

Signifikanz:

$$sig(A,B) = \frac{-\log\left(1 - e^{-\lambda} \cdot \sum_{i=1}^{k-1} \frac{1}{i!} \cdot \lambda^i\right)}{\log n}$$

signifikante Kookkurrenten:

Berechnung eines Signifikanzwertes

- a,b: Anzahl der Sätze die A, B enthalten
- k: Anzahl Sätze die A *und* B enthalten
- n: Gesamtanzahl Sätze
- $\lambda = \frac{a \cdot b}{n}$

Signifikanz:

$$sig(A,B) = \frac{-\log\left(1 - e^{-\lambda} \cdot \sum_{i=1}^{k-1} \frac{1}{i!} \cdot \lambda^i\right)}{\log n}$$

weiter mit signifikanten Kookkurrenten:

Näherungsformel:

falls, $\frac{k+1}{\lambda} > 2,5$

$$sig(A,B) \approx \frac{\lambda - k \cdot \log \lambda + \log k!}{\log n}$$

falls, $\frac{k+1}{\lambda} > 2,5$ und $k > 10$

$$sig(A,B) \approx \frac{k \cdot (\log k - \log \lambda - 1)}{\log n}$$

weiter mit signifikanten Kookkurrenten:

Näherungsformel:

falls, $\frac{k+1}{\lambda} > 2,5$

$$sig(A,B) \approx \frac{\lambda - k \cdot \log \lambda + \log k!}{\log n}$$

falls, $\frac{k+1}{\lambda} > 2,5$ und $k > 10$

$$sig(A,B) \approx \frac{k \cdot (\log k - \log \lambda - 1)}{\log n}$$

weiter mit signifikanten Kookkurrenten:

A	B	a	b	k	sig(A,B)
Romeo	Julia	343	1080	124	51.85
Stadt	Einwohner	37053	2611	54	30.47
Steuer-gelder	Verschwendung	251	373	54	25.58
Polizei	verhaftet	20550	1928	131	16.06
Unfall	Krankenhaus	1987	2250	11	1.01

Tabelle: [WiRo08, S.140]

weiter mit signifikanten Kookkurrenten:

- Beispielanwendung: Entdeckung von Polysemie
- Visualisierung mit „simulated annealing„

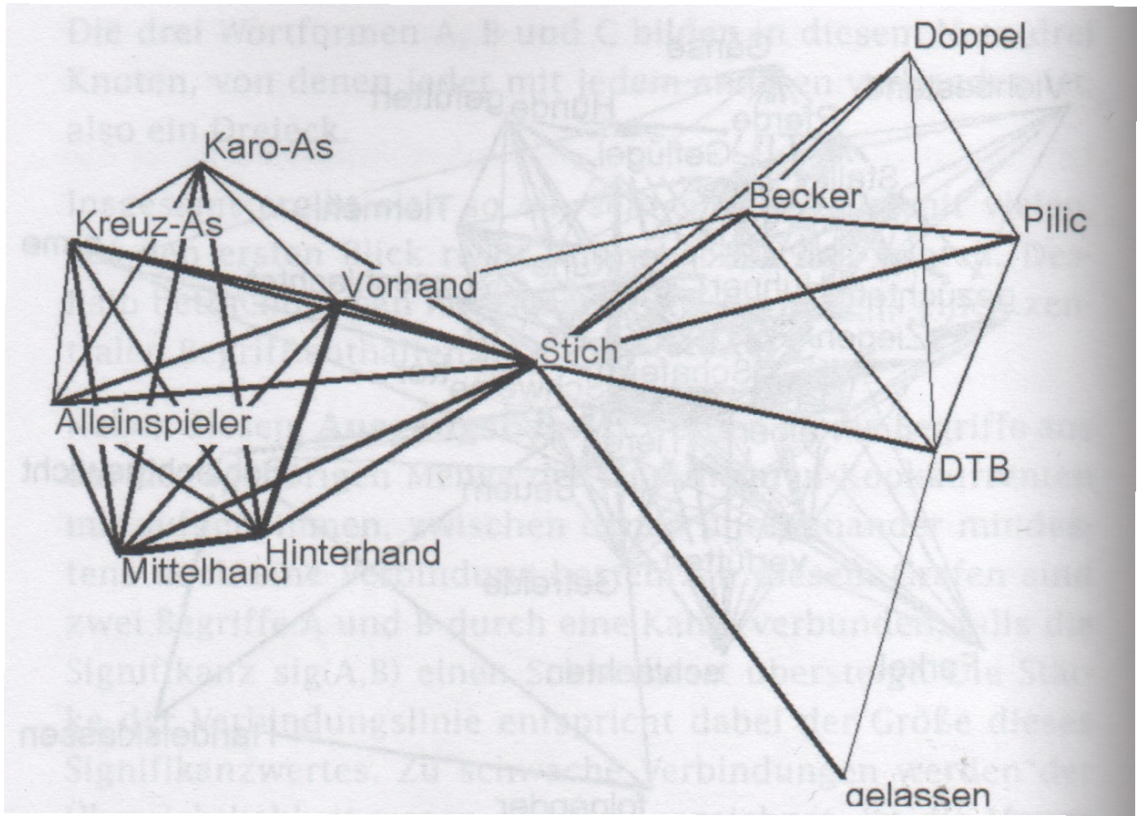
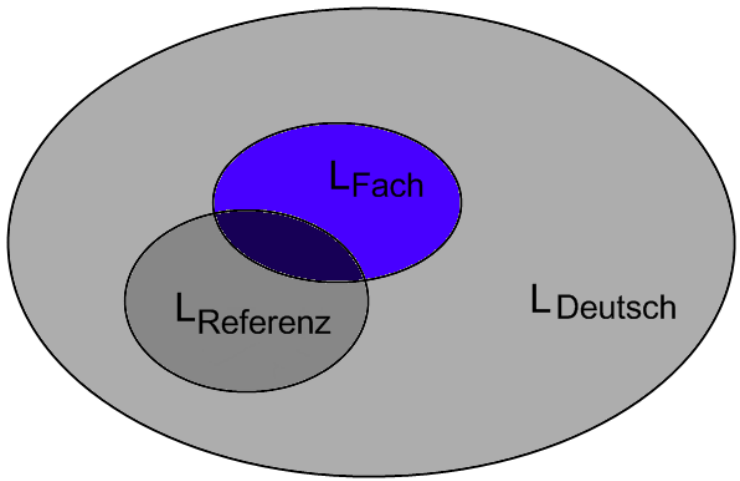


Abbildung: Quelle [WiRo08, S. 152]

Verfahren - Differenzanalyse

- statistisches Verfahren
- Ermittlung von diskimnierenden Termen
- Untersuchung der unterschiedlichen Verteilungen von Wortformen bzw. Wortkombinationen
- Anwendung bei Terminologieextraktion, Beschlagwortung und Sachgebietsklassifikation



- Grundlage bilden zwei Textkorpora
  - Analysekorpus
  - Referenzkorpus

- Ermittlung der Auftrittswahrscheinlichkeiten von Wortformen und deren Kombinationen
- Vergleich der Verteilung und Einordnung in **Klassen**:
  - Klasse 1: Wortformen, die nicht im Referenzkorpus vorkommen
  - Klasse 2: Wortformen, die relativ häufiger im Analsysekorpus vorkommen, als im Referenzkorpus
  - Klasse 3: Wortformen, die mit etwa gleicher relativer Häufigkeit in beiden Textkorpora vorkommen
  - Klasse 4: Wortformen, die seltener im Fachtext auftauchen als im Analysekorpus
- Bsp. Einteilung in Häufigkeitsklassen im Projekt Deutscher Wortschatz:

$$HKL(w) = ganzer\ Anteil\left(\log_2 \frac{|„der“|}{|w|}\right)$$

- Ermittlung der Auftrittswahrscheinlichkeiten von Wortformen und deren Kombinationen
- Vergleich der Verteilung und Einordnung in **Klassen**:
  - Klasse 1: Wortformen, die nicht im Referenzkorpus vorkommen
  - Klasse 2: Wortformen, die relativ häufiger im Analsysekorpus vorkommen, als im Referenzkorpus
  - Klasse 3: Wortformen, die mit etwa gleicher relativer Häufigkeit in beiden Textkorpora vorkommen
  - Klasse 4: Wortformen, die seltener im Fachtext auftauchen als im Analysekorpus
- Bsp. Einteilung in Häufigkeitsklassen im Projekt Deutscher Wortschatz:

$$HKL(w) = ganzer\ Anteil\left(\log_2 \frac{|„der“|}{|w|}\right)$$

Verfahren - Clustering - Dokumentenähnlichkeit

- Ziel: Menge von Dokumenten nach thematischer Ähnlichkeit einordnen
- nötige Arbeitsschritte:
  1. Identifikation der charakteristischen Merkmale bzw. Eigenschaften (Indexterme)
  2. Erzeugen von Dokumentenvektoren
  3. Auswahl eines Ähnlichkeitsmaßes
  4. Erzeugen der Ähnlichkeitsmatrix
  5. Cluster-Analyse

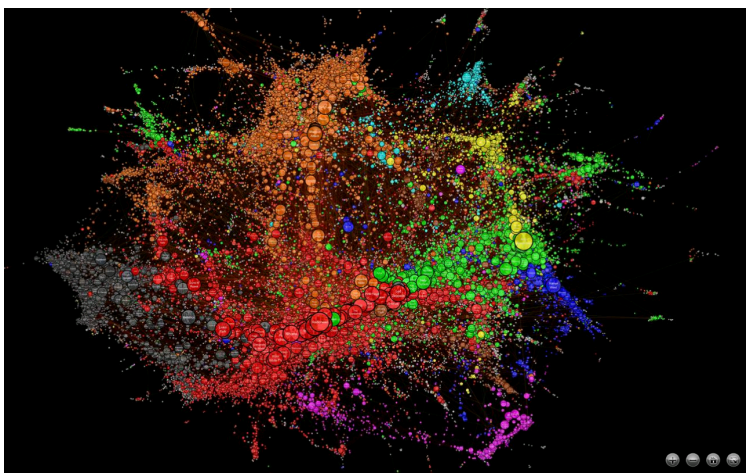


Abbildung: Quelle: [ClusVisu]

Vorgehen am Beispiel

Dokument 1

Ein Vertrag ist ein Vertrag ist ein Vertrag.

Dokument 2

Je riskanter der Weg, desto größer der Profit.

Dokument 3

Es führen viele Wege zum Profit.

Dokument 4

Die Rechtfertigung von Profit ist Profit.

1 Identifikation der charakteristischen Terme

- Segmentierung des Textes in Terme
- Wortbeugungen auf Wortstamm zurückführen

Terme - ohne Stoppwörter

t_1 =Vetrag, t_2 =riskant, t_3 =Weg, t_4 =groß,
t_5 =Profit, t_6 =führen, t_7 =Rechtfertigung

2 Erzeugen des Dokumentenvektoren

- Annahme: häufig vertretene Wortformen repräsentieren Dokument gut
- Berechnung **Termfrequenz** $f_{i,m}$ des Terms t_i im Dokument d_m

Beispiel - Termfrequenzen

$f_{1,1} = 3, f_{1,2} = 0, f_{5,2} = 1, f_{5,4} = 2$

- unterschiedliche Länge von Dokumenten ▷ Normalisierung
- unterschiedliche Vorgehensweise
- Beispiel: relative Häufigkeit $nf_{i,m} = \frac{f_{i,m}}{\sum_{t_j \in d_m} f_{j,m}}$

2 zu Dokumentenvektoren

Beispiel - normalisierte Termfrequenzen

$nf_{1,1} = \frac{3}{3} = 1, nf_{1,2} = \frac{0}{4} = 0, nf_{5,2} = \frac{1}{4} = 0.25, nf_{5,4} = \frac{2}{3} = 0,\overline{6}$

- weitere Forderung: Vorkommen des Terms sollte in wenigen Dokumenten besonders häufig sein
- Aussage liefert **inverse Dokumentfrequenz** $idf_i = \log \frac{|d|}{|d:t_i \in d|}$

Beispiel - inverse Dokumentfrequenz

$idf_1 = \log \frac{4}{1} \approx 0.602, idf_5 = \log \frac{4}{3} \approx 0.125$

- Zusammen mit der normalisierten Termfrequenz lässt sich ein Maß der **Wichtigkeit** $w_{i,m}$ in Abhängigkeit zum Term berechnen

2 zu Dokumentenvektoren

- Wichtigkeit des Terms $w_{i,m} = nf_{i,m} \cdot idf_i$

Beispiel - Wichtigkeit

$$w_{1,1} = nf_{1,1} \cdot idf_1 = 1 \cdot \log \tfrac{4}{1} \approx 0.602,$$
$$w_{1,2} = nf_{1,2} \cdot idf_1 = 0 \cdot \log \tfrac{4}{1} = 0,$$
$$w_{5,2} = nf_{5,2} \cdot idf_5 = 0.25 \cdot \log \tfrac{4}{3} \approx 0.0301,$$
$$w_{5,4} = nf_{5,4} \cdot idf_5 = \tfrac{2}{3} \cdot \log \tfrac{4}{3} \approx 0.083$$

- Bildung Term-Dokument-Matrix

2 zu Dokumentenvektoren

- Term-Dokument-Matrix, beinhaltet Dokumentenvektoren $\vec{d_m}$

Beispiel - Term-Dokument-Matrix

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
d_1	0.602	0	0	0	0	0	0
d_2	0	0.151	0.075	0.151	0.031	0	0
d_3	0	0	0.1	0	0.042	0.201	0
d_4	0	0	0	0	0.083	0	0.201

- Beispiel Dokumentenvektor:
  $\vec{d_2} = (0, 0.151, 0.075, 0.151, 0.031, 0, 0)$

3 Auswahl eines Ähnlichkeitsmaßes

- verschiedene Varianten
- Euklidische Distanz:

$$dist_{Euklid}(\vec{d}_i, \vec{d}_j) = \sqrt{\sum_{k=1}^{n}(w_{k,i} - w_{k,j})^2}$$

- Skalarprodukt
- Cosinus-Maß

$$sim_{Cos}(\vec{d}_i, \vec{d}_j) = \frac{\sum_{k=1}^{n}(w_{k,i} \cdot w_{k,j})}{\sqrt{\sum_{k=1}^{n}(w_{k,i})^2} \cdot \sqrt{\sum_{k=1}^{n}(w_{k,j})^2}}$$

4 Erzeugen der Ähnlichkeitsmatrix

- Berechnung Dokument-Dokument-Matrix

Beispiel - Dokument-Dokument-Matrix mit $sim_{Cos}(\vec{d_i}, \vec{d_j})$

	d_1	d_2	d_3	d_4
d_1	1	0	0	0
d_2	0	1	0.169	0.052
d_3	0	0.169	1	0.07
d_4	0	0.052	0.07	1

- $1 \approx$ stärkste Ähnlichkeit
- $0 \approx$ keine Ähnlichkeit

5 Clusteranalyse

- Verwendung: *bottom up* (agglomerativ) / single-link Clustering
- Vorgehen:
  - einzelne Dokumente bilden seperaten Cluster
  - Ähnlichkeit des Clusters entsteht aus ähnlichsten Elementen des Clusters
    $sim_{slink}(c_1, c_2) = \max_{x \in c_1, y \in c_2} (sim(x, y))$
  - es ergeben sich $n \cdot (n-1)$ Cluster mit n Anzahl der Dokumente

5 zu Clusteranalyse

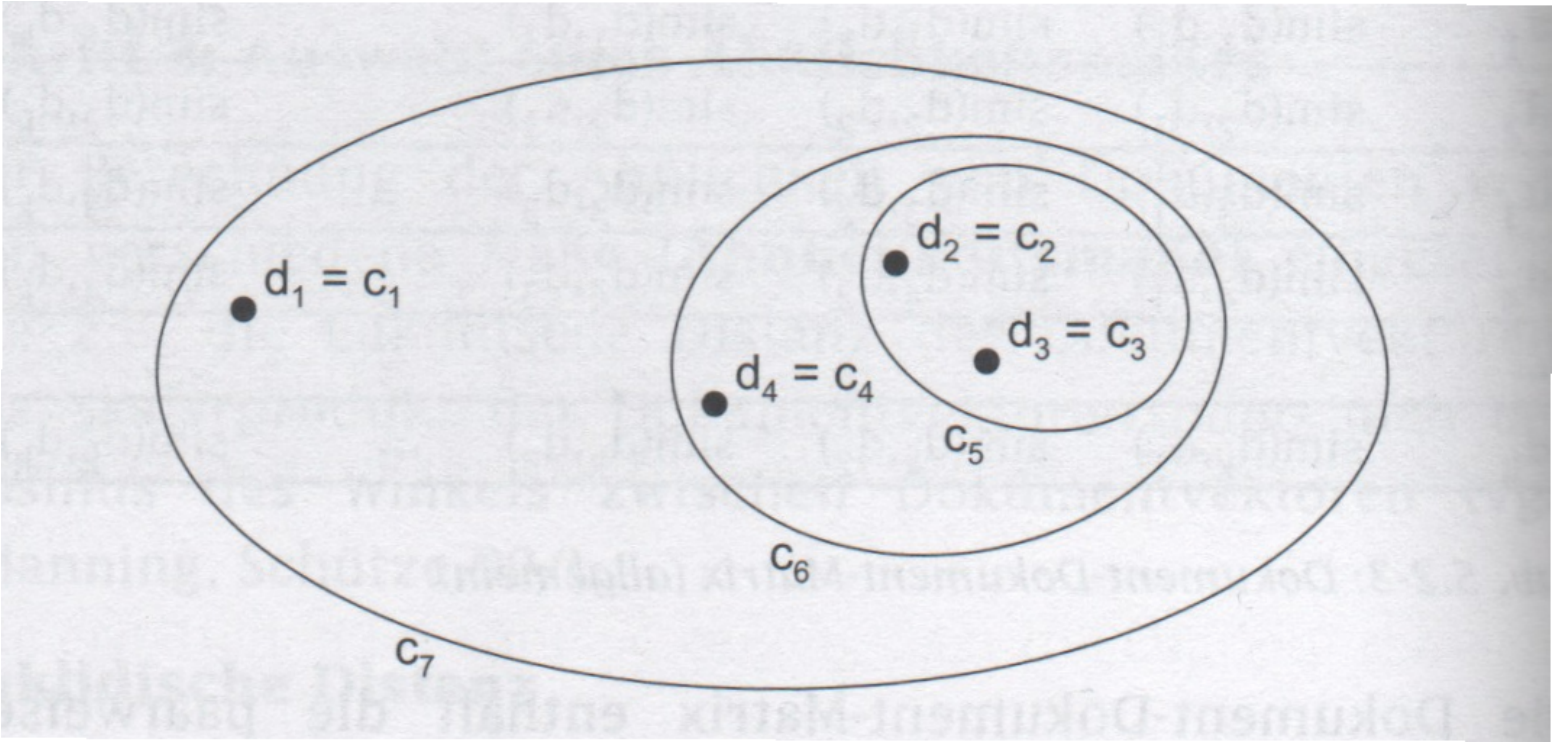


Abbildung: Dendrogramm

5 zu Clusteranalyse

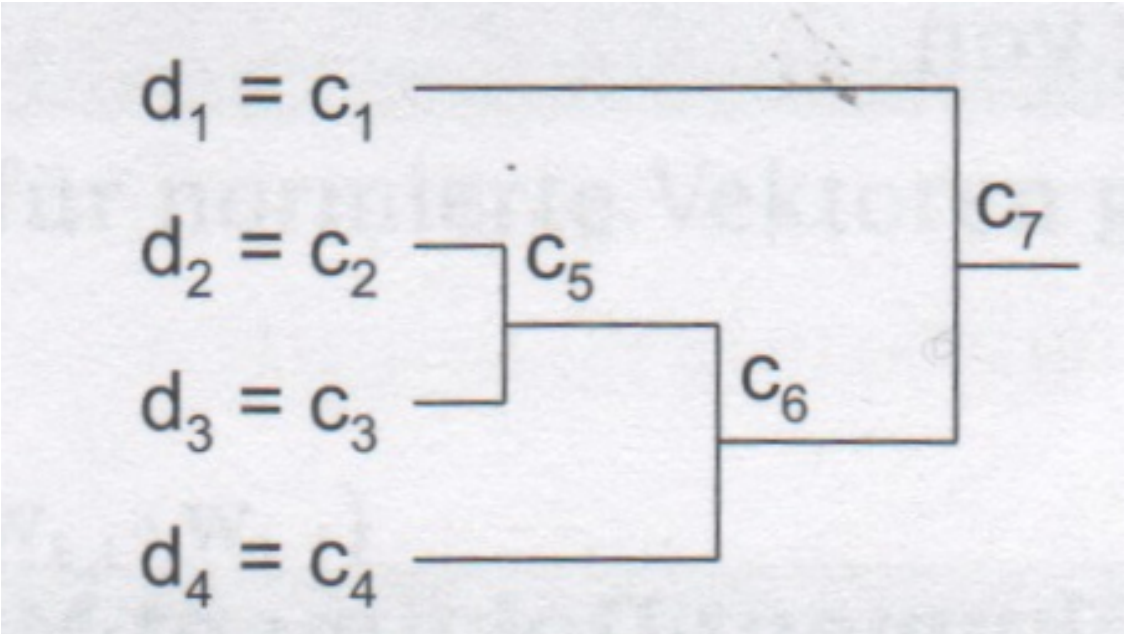


Abbildung: Dendrogramm

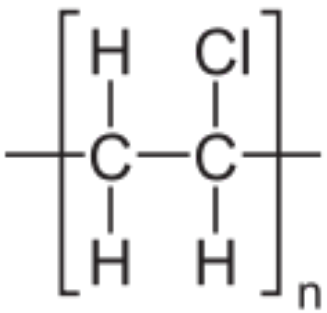
Verfahren - Musteranalyse

- Suchen und Endecken von Mustern im Textkorpus / Textsammlung
- Benutzung von Regulären Ausdrücken für Abfragen in Textdatenbank

Beispiel - Suche nach Wortarten im Textkorpus

Dies[ART] ist[VERB] ein[ART] Beispiel[NOMEN].
Suche: *[ART] *[NOMEN]

- Anwendung:
  - Entdeckung von Morphenmuster in Medezin oder Chemie
  - (Präfix*(Stamm Fugenelemente? Suffix?)+ Suffix*)
  - Suche: Polyvinylchlorid (PVC)



Literatur


Gerhard Heyer, Uwe Quasthoff, Thomas Wittig
Text Mining: Wissensrohstoff Text
1. korrigierte Auflage, W3L-Verlag, 2008.


Michael W. Berry, Malu Castellanos
Survey of Text Mining II: Clustering, Classification, and Retrieval: No. 2.


Ronen Feldman, James Sanger
The Text Mining Handbook
Cambridge University Press 2008


K.-U Carsten, Ch. Ebert, E. Endriss, S. Jekat, R. Klabunde, H. Langer
Computerlinguistik und Sprachtechnologie
Spektrum Akademischer Verlag 2004

Bildquellen


William J. Frawley, Gregory Piatetsky-Shapiro, Christopher J. Matheus
Knowledge Discovery in Databases
http://www.aaai.org/ojs/index.php/aimagazine/article/view/1011
1992


http://www.regioit-aachen.de


http://imgs.xkcd.com/comics


http://www.abendblatt.de/multimedia/


http://ostfussball.com/


http://sixdegrees.hu/last.fm/interactive_map.html

Danke für Ihre Aufmerksamkeit
Fragen ??