Einführung in die Computerlinguistik

12: Informationsmanagement

WS 2008/2009

Manfred Pinkal

Informationsmanagement

Große Datenmengen
zugänglich und nutzbar machen

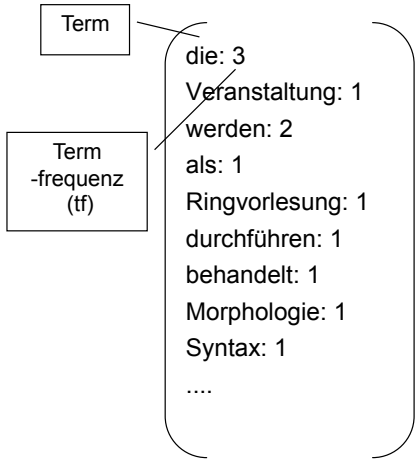
- Konkrete Aufgaben
  - Dokumente klassifizieren: Document Classification
  - Dokumente zusammenfassen: Summarisation
  - Dokumente mit relevanter Information finden: Information Retrieval
  - Relevante Information in Dokumenten finden: Question Answering
  - Relevante Information in Dokumenten finden: Information Extraction

Information Retrieval

- Gegeben: Anfrage (Query)
  - Im einfachsten Fall eine Menge von Suchbegriffen
- Gesucht: Relevante Dokumente
  - Dokumente, deren Informationsgehalt der Anfrage am nächsten kommt.
  - Im IR wird Relevanz modelliert durch die Nähe im semantischen Raum

Beispiel: Vorlesungsankündigung 1

Die Veranstaltung wird als Ringvorlesung durchgeführt. Die jeweilige Lehrkraft für verschiedene Bereiche der Sprachwissenschaft führt in die Ziele und Begriffe des Bereiches ein. Behandelt werden Phonetik und Phonologie, Morphologie und Syntax, Semantik, Pragmatik und Psycholinguistik .



Beispiel: Vorlesungsankündigung 2

Ziel der Veranstaltung ist es, die Teilnehmer mit Grundbegriffen und Grundproblemen der deskriptiven wie theoretischen Syntax und Morphologie vertraut zu machen. Im Vordergrund steht dabei die Syntax des Deutschen, aber auch Phänomene im Englischen oder anderen Sprachen werden diskutiert.

Ziel: 1
die: 4
Veranstaltung: 1
sein: 1
es: 1
Teilnehmer: 1
...
Syntax: 2
Morphologie: 1
...

Beispiel: FAZ-Politik-Artikel

Gegen den Widerstand von Arbeitsminister Clement haben sich Bundeskanzler Schröder und die SPD- Spitze für eine Ausbildungsabgabe ausgesprochen. Eine entsprechende Vorlage wird Montag in der Bundestagsfraktion behandelt.

gegen: 1
der: 1
Widerstand: 1
von: 1
Arbeitsminister: 1
Clement: 1
behandelt: 1
Ausbildungsabgabe: 1
die: 2
...

Query

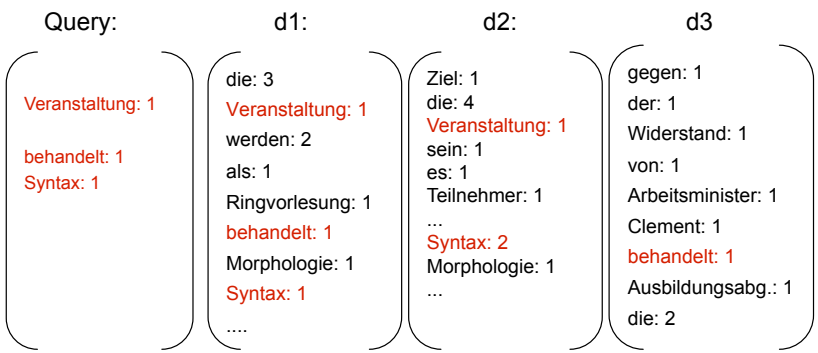
„Welche Veranstaltung behandelt Syntax?"

Veranstaltung: 1
behandelt: 1
Syntax: 1

Termfrequenz in Anfrage und Dokumenten

Query:	d1:	d2:	d3
Veranstaltung: 1 behandelt: 1 Syntax: 1	die: 3 Veranstaltung: 1 werden: 2 als: 1 Ringvorlesung: 1 behandelt: 1 Morphologie: 1 Syntax: 1	Ziel: 1 die: 4 Veranstaltung: 1 sein: 1 es: 1 Teilnehmer: 1 ... Syntax: 2 Morphologie: 1 ...	gegen: 1 der: 1 Widerstand: 1 von: 1 Arbeitsminister: 1 Clement: 1 behandelt: 1 Ausbildungsabg.: 1 die: 2 .....

Termfrequenz in Anfrage und Dokumenten



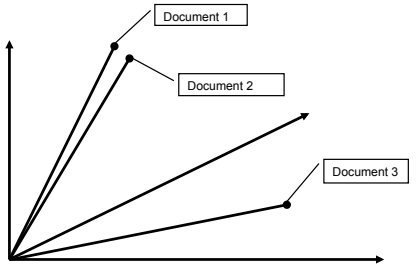
Die Wort-Dokument-Matrix

	d1	d2	d3	...
Veranstaltung	1	1	0	...
Teilnehmer	0	1	0	...
behandelt	1	0	1	...
Widerstand	0	0	1	...
Arbeitsminister	0	0	1	
Clement	0	0	1	...
Syntax	1	2	0	...
Morphologie	1	1	0	...

Vektorraum-Modelle

- Dokumentinformation wird als Muster von Termfrequenzen dargestellt: als Vektor, dessen Dimensionen Wörter sind. Werte sind die jeweiligen Frequenzen.
- Ein Dokument wird repräsentiert als Vektor im vieldimensionalen semantischen Raum, dessen Dimensionen Wörtern entsprechen ("Wortraum" / "word space")
- Informationelle/ semantische Ähnlichkeit von Dokumenten wird modelliert als Nähe ihrer Vektoren.

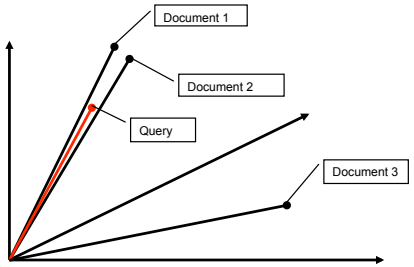
Semantischer Raum



Die Wort-Dokument-Matrix

	d1	d2	d3	q
Veranstaltung	1	1	0	1
Teilnehmer	0	1	0	0
behandelt	1	0	1	1
Widerstand	0	0	1	0
Arbeitsminister	0	0	1	0
Clement	0	0	1	0
Syntax	1	2	0	1
Morphologie	1	1	0	0

Semantischer Raum



Distanz zwischen Vektoren als (inverses) Ähnlichkeitsmaß

- Euklidische Distanz:

$$dist(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- Unser Beispiel:

$$dist(\vec{d}_1,\vec{d}_2) = 1{,}73$$
$$dist(\vec{d}_1,\vec{d}_3) = 2{,}45$$
$$dist(\vec{d}_2,\vec{d}_3) = 3{,}00$$

- Problem: abhängig von der absoluten Häufigkeit der Terme, und damit von der Größe der Dokumente.

Cosinus als Ähnlichkeitsmaß

- Standardmaß für die Ähnlichkeit ist der Cosinus

$$sim_{cosine}(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Wenn Vektoren identische Richtung haben, ist Cosinus 1($cos(0^0)$=1); wenn Vektoren rechtwinklig aufeinander stehen, ist der Cosinus 0 ($cos(90^0)$=0).
- Unser Beispiel:

$$\cos(\vec{d}_1,\vec{d}_2) = 0.76$$
$$\cos(\vec{d}_1,\vec{d}_3) = 0.25$$
$$\cos(\vec{d}_2,\vec{d}_3) = 0.00$$

Vektorraum-Modelle

- Dokumentinformation wird als Muster von Termfrequenzen dargestellt: als Vektor, dessen Dimensionen Wörter sind. Werte sind die jeweiligen Frequenzen.
- Ein Dokument wird repräsentiert als Vektor im vieldimensionalen "Wortraum".
- Informationelle/ semantische Ähnlichkeit von Dokumenten wird definiert als Nähe ihrer Vektoren.
- Die Suchanfrage wird ebenfalls als Vektor dargestellt. Das Dokument, das der Suchanfrage am nächsten ist, ist (entsprechend der Modellierung) das Dokument mit der höchsten Relevanz für die Suchanfrage.

Vorteile von semantischen Räumen

- Nutzung von Frequenzinformation
  - Konzeptuell einfach, effizient
  - Dokumente sind ähnlich, wenn Begriffe gleich häufig vorkommen
- Formalisierung
  - Mathematische Standardverfahren zur Berechnung von Ähnlichkeit / Relevanz (z.B.: euklidische Distanz, Cosinus)
- Erweiterungsmöglichkeiten
  - Mathematische / statistische Methoden (z.B. Googles PageRanking)
  - Linguistische Verfahren: genauere Modellierung der Terme

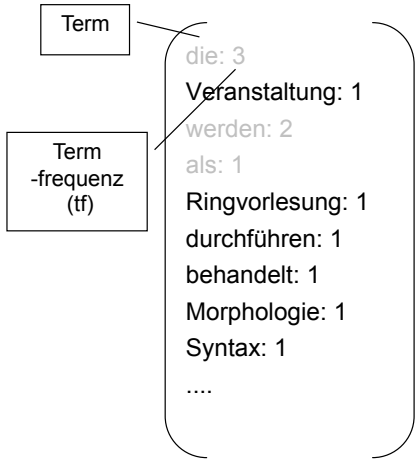
Genauere Modellierung von Termen (1)

Nicht alle Worte sind gleich

- „Stoppworte" komplett entfernen
  - Artikel, Hilfsverben, Präpositionen:
  - Sind semantisch nicht ergiebig, kommen in ähnlicher Verteilung überall vor

Beispiel: Vorlesungsankündigung 1

Die Veranstaltung wird als Ringvorlesung durchgeführt. Die jeweilige Lehrkraft für verschiedene Bereiche der Sprachwissenschaft führt in die Ziele und Begriffe des Bereiches ein. Behandelt werden Phonetik und Phonologie, Morphologie und Syntax, Semantik, Pragmatik und Psycholinguistik .



Genauere Modellierung von Termen (2)

- Informative Worte stärker werten:
  - Dokumentfrequenz n_i von Wort i: Anzahl der Dokumente im Korpus, die das Wort i (mindestens einmal) enthalten
  - Gesamtzahl der Dokumente: *N*
  - idf (inverse Dokumentfrequenz) $idf_i = \log\left(\frac{N}{n_i}\right)$
  - idf ist ein plausibles Maß für die Informativität von Suchtermen
  - **tf * idf**: Termfrequenz * Inverse Dokumentfrequenz

Die Wort-Dokument-Matrix

	d1	d2	d3	q
Veranstaltung	1	1	0	1
Teilnehmer	0	1	0	0
behandelt	1	0	1	1
Widerstand	0	0	1	0
Arbeitsminister	0	0	1	0
Clement	0	0	1	0
Syntax	1	2	0	1
Morphologie	1	1	0	0

$\cos(\vec{q},\vec{d}_1) = 0.77$

$\cos(\vec{q},\vec{d}_2) = 0.65$

$\cos(\vec{q},\vec{d}_3) = 0.29$

Die Wort-Dokument-Matrix

	d1	d2	d3	q
Veranstaltung	5	5	0	5
Teilnehmer	0	5	0	0
behandelt	3	0	3	3
Widerstand	0	0	6	0
Arbeitsminister	0	0	10	0
Clement	0	0	14	0
Syntax	10	20	0	10
Morphologie	11	11	0	0

$\cos(\vec{q},\vec{d}_1) = 0.77$ $\quad$ $\cos(\vec{q},\vec{d}_1) = 0.73$

$\cos(\vec{q},\vec{d}_2) = 0.65$ $\quad$ $\cos(\vec{q},\vec{d}_2) = 0.81$

$\cos(\vec{q},\vec{d}_3) = 0.29$ $\quad$ $\cos(\vec{q},\vec{d}_3) = 0.04$

Query Expansion

- Dokumente können auch dann hoch relevant sein, wenn sie nicht den Suchterm selbst, sondern ein Synonym oder Hyponym enthalten.
- Anfrage-Erweiterung („Query expansion") mit semantisch ähnlichen Wörtern, z.B. WordNet -Information oder Information aus domänenspezifischen Ontologien

- *Syntax > Grammatik, Dependenzgrammatik*, ...

Vektorraum-Modelle der Wortbedeutung (1)

- Die Bedeutung sprachlicher Ausdrücke lässt sich an ihrem Gebrauch ablesen.
- Zum Gebrauch von sprachlichen Ausdrücken gehört ihr Vorkommen in Sätzen und Textdokumenten.
- Bedeutungsähnlichkeit von Wörtern korrespondiert (in gewissem Umfang) der Ähnlichkeit der Verteilungsmuster dieser Wörter über Textdokumente in einem großen Korpus.
- Wenn man die Wort-Dokument-Matrix aus einer anderen Sicht liest, bildet sie eine Grundlage für die Modellierung von Ähnlichkeit der Wortbedeutung:
- Wortbedeutung wird durch Vektoren im „Dokumenten-Raum“ repräsentiert. Semantische Ähnlichkeit z.B. durch den Cosinus gemessen.

Die Wort-Wort-Matrix

	Veranstaltung	Teilnehmer	behandelt	Widerstand	Syntax	Morphologie
Veranstaltung	2	1	1	0	2	2
Teilnehmer	1	1	0	0	1	1
Behandelt	1	0	1	1	1	1
Widerstand	0	0	1	1	0	0
Syntax	2	1	1	0	3	2
Morphologie	2	1	1	0	3	2

Vektorraum-Modelle der Wortbedeutung (2)

- Der Kontext eines Wortes lässt sich statt als Dokument auch als die Menge von Wörtern definieren, die mit dem Wort in einem Kontext kookkurieren.
- Wortbedeutung kann als Vektor im Wortraum repräsentiert, Bedeutungsähnlichkeit als z.B. als Cosinus der Vektoren modelliert werden.

Beurteilung von Information Retrieval

- Gut zur Suche von Dokumenten aus großen Datenmengen
  - einfach zu realisieren
  - schnell
- Problem: Niedrige Präzision
  - Falsche Treffer
  - Ergebnis nur Liste von Dokumenten
- Rolle von sprachlichem Wissen
  - Wenig Wissen nötig
  - Kann zur Optimierung des semantischen Raumes dienen
    - Stoppwörter, Kombination verwandter Wörter