Versteckte Variablen – Modelle für spezielle Retrieval – Aufgaben

Christof Bräutigam

22.5.2008

Gliederung

1. Motivation
2. Termbasiertes Modell
3. Modelle mit versteckten Variablen
4. Evaluation

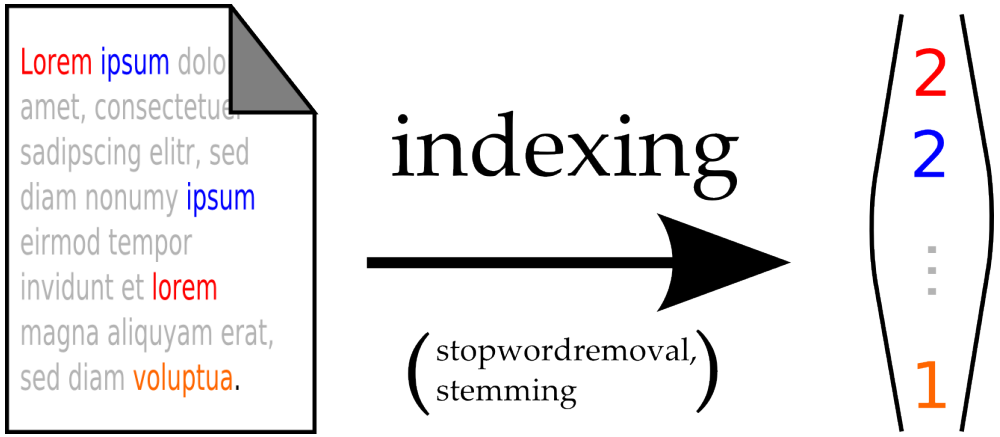
Motivation

- Problem: umfangreiche, unstrukturierte, unkategorisierte Datensammlungen
- Ziel: Informations(rück)gewinnung (Retrieval)
- speziell: Clustering (automatische Kategorisierung)

Termbasiertes Modell – Vektorraummodell

Vektorraummodell (Salton et.al. 1975)

- Dokumente (und Anfragen) werden als Vektoren von Indextermen repräsentiert



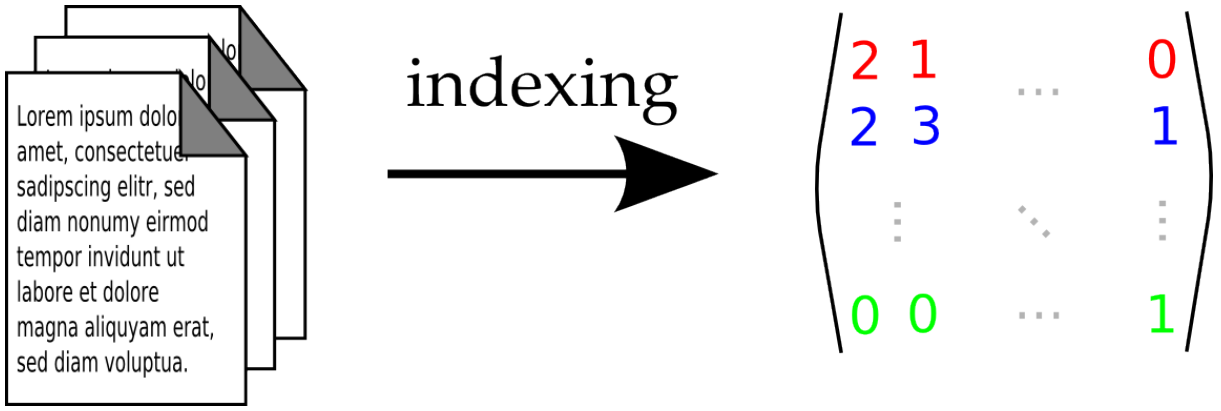
Reales Dokument *d*

Repräsentation **d**
(Dokumentvektor)

Termbasiertes Modell – Vektorraummodell

Vektorraummodell (Salton et.al. 1975)

- Dokumente (und Anfragen) werden als Vektoren von Indextermen repräsentiert



Kollektion Term-Dokument-Matrix (TDM)

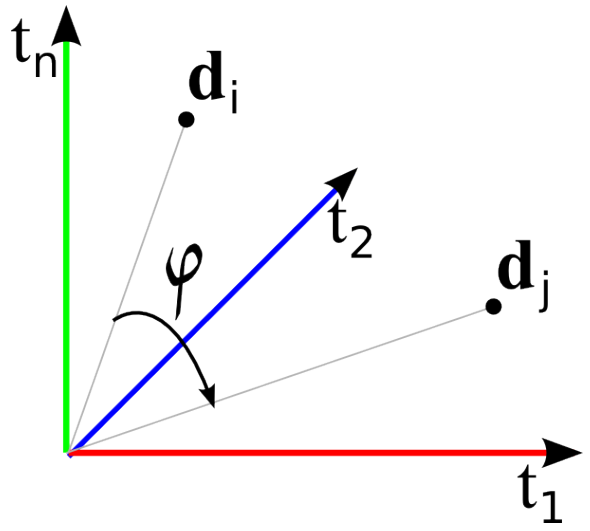
Termbasiertes Modell – Vektorraummodell

Vektorraummodell (Salton et.al. 1975)

- Dokumente (und Anfragen) werden als Vektoren von Indextermen repräsentiert
- Retrievalfunktion: der Cosinus des Winkels zwischen Vektoren wird als Relevanzmaß verwendet

$$\cos(\varphi) = \frac{d_i^T d_j}{\|d_i\| \cdot \|d_j\|}$$

$$0 \leq \cos(\varphi) \leq 1$$



Termbasiertes Modell – Vektorraummodell

Kritik

+ Indexterme direkt automatisch extrahierbar
+ einfach, effektiv, gute Informationswiedergabe
+ flexibel (viele Optimierungsmöglichkeiten, z.B. verschiedene Termgewichtungsverfahren wie *tf*, *tfidf*)

- hohe Dimension des Vektorraumes
- dünn besetzte TDM
- begrenzt auf Terme, kein semantischer Vergleich
- Performanz verringert durch *Synonyme* und *Homonyme*

Modelle mit versteckten Variablen

Ziel

- Ausnutzen der semantischen Information im Retrieval

Idee

- Semantik eines Textes äußert sich in den Termen
- Semantik ist verknüpft mit den *Konzepten* „hinter“ dem Text
- Analyse der TDM
- transformation des (hochdimensionalen) Termindex in einen (niedrigdimensionalen) Konzeptindex
- dabei Erhaltung der vorhandenen Information des Vektorraummodells

Modelle mit versteckten Variablen - LSI

Latent Semantic Indexing (Deerwester et. al. 1990)

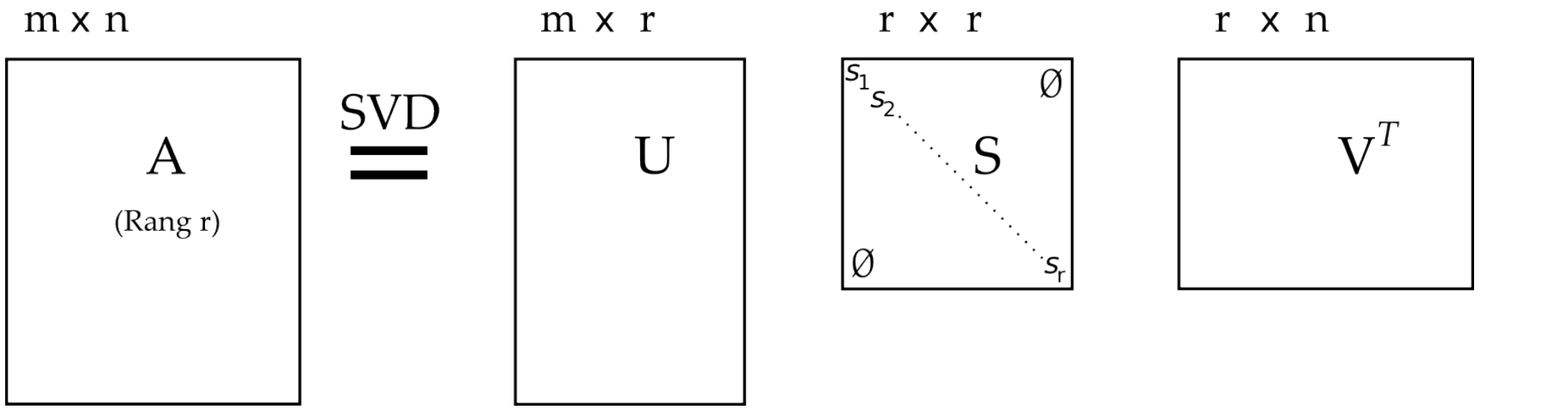
- basiert auf einer Faktoranalyse der TDM
- Verfahren: Singulärwertzerlegung (SVD)

$$\mathbf{A} \overset{\text{SVD}}{=} \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{T}$$

Modelle mit versteckten Variablen - LSI

Latent Semantic Indexing (Deerwester et. al. 1990)

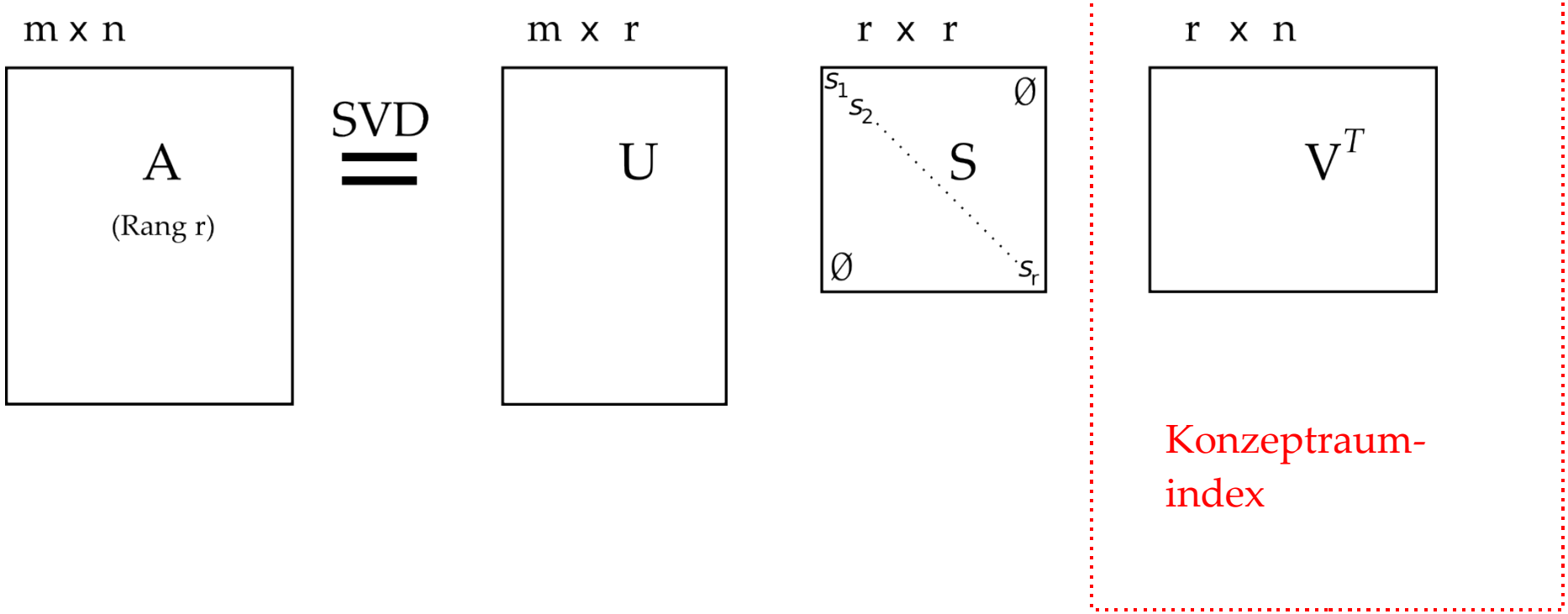
- basiert auf einer Faktoranalyse der TDM
- Verfahren: Singulärwertzerlegung (SVD)



Modelle mit versteckten Variablen - LSI

Latent Semantic Indexing (Deerwester et. al. 1990)

- basiert auf einer Faktoranalyse der TDM
- Verfahren: Singulärwertzerlegung (SVD)

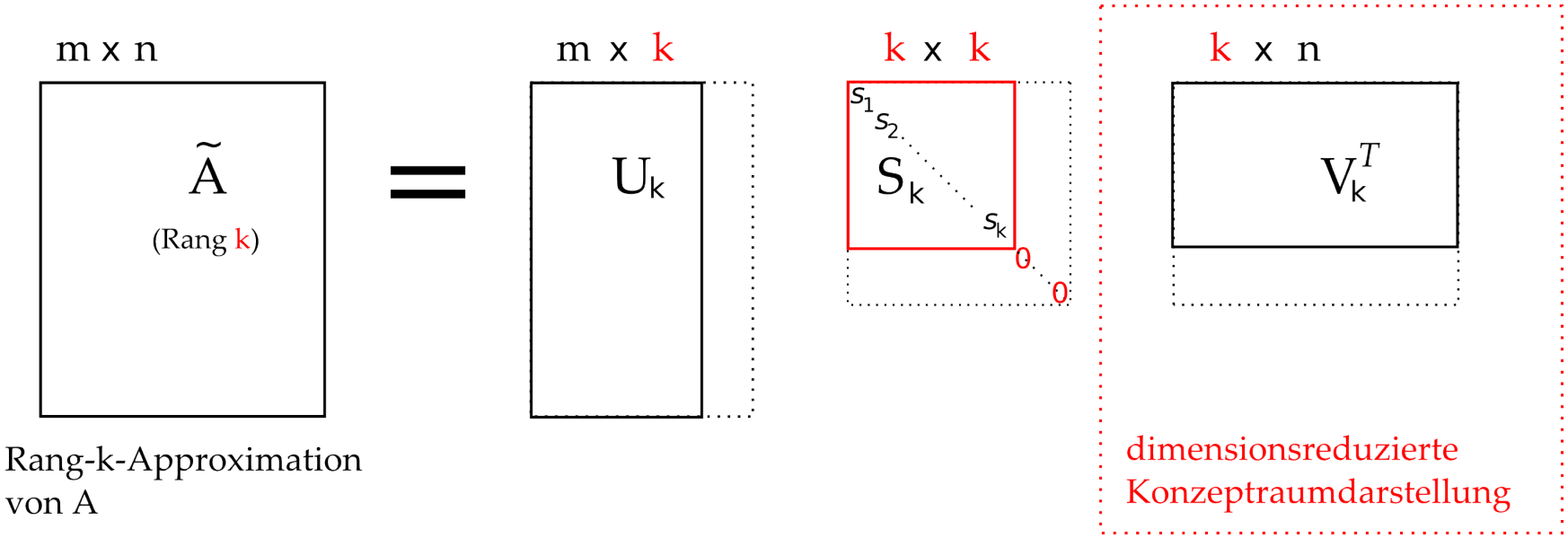


Modelle mit versteckten Variablen - LSI

- SVD ermöglicht Approximation der TDM
- Approximation wird zur Dimensionsreduktion des Konzeptraumes eingesetzt

Modelle mit versteckten Variablen - LSI

- SVD ermöglicht Approximation der TDM
- Approximation wird zur Dimensionsreduktion des Konzeptraumes eingesetzt



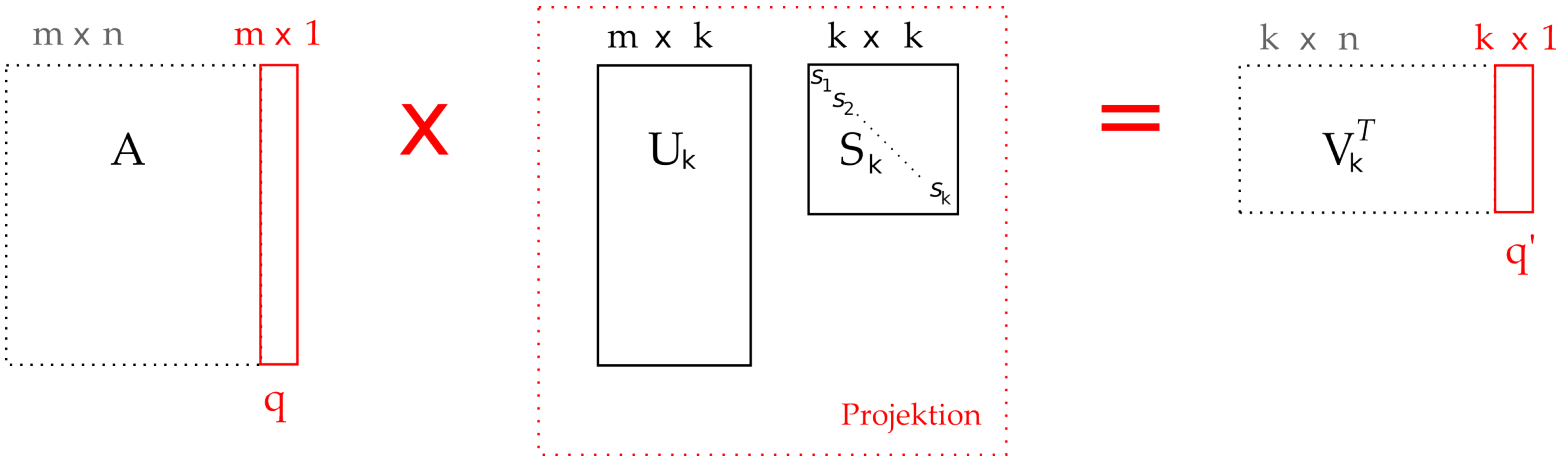
Rang-k-Approximation von A

- Approximation ist optimal hinsichtlich der Frobenius-Norm

Modelle mit versteckten Variablen - LSI

- Projektion neuer Dokumente in den Konzeptraum

$$q' = q^T U_k S_k^{-1}$$



- analog Projektion neuer Terme in den Konzeptraum
- ermöglicht Erweiterung des Konzeptraumes ohne komplette Neuberechnung (aber: Reskalierung, Updateformeln nötig)

Modelle mit versteckten Variablen - LSI

- Retrieval im dimensionsreduzierten Konzeptraum analog zum Vektorraummodell (Konzeptvektoren + Cosinusähnlichkeit)

- Wahl der Dimension beeinflusst das Retrievalergebnis
- es sind keine Regeln für die Wahl einer Dimension basierend auf einer Analyse der Kollektion bekannt

Modelle mit versteckten Variablen - LSI

Kritik

+ semantische Information wird nutzbar
+ gute Ergebnisse schon in sehr geringen Dimensionen
+ Synonyme werden erkannt
+ impliziter Thesaurus

- hoher Aufwand beim Modelltraining (SVD: ~$O(N^3)$)
- Funktionsweise nur in Ansätzen wissenschaftlich erschlossen
- keine gute Erkennung von Homonymen
- ungeeignet für heterogene, dynamische Kollektionen (WWW)

Modelle mit versteckten Variablen - PLSI

Probabilistic Latent Semantic Indexing (Hofmann 1999)

- Grundlage:
  generatives statistisches Sprachmodell – Aspektmodell

Idee

- Dokumente generieren Worte mit best. Wahrscheinlichkeit
- direkter Zusammenhang von Worten und Dokumenten (beobachtbar und quantifiziert in TDM) wird entkoppelt über eine unbeobachtete Variable – Konzepte

Modelle mit versteckten Variablen - PLSI

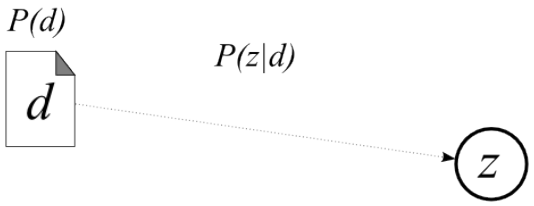
Aspektmodell

- wähle ein Dokument *d* mit der A-priori-Wkt. *P(d)*

Modelle mit versteckten Variablen - PLSI

Aspektmodell

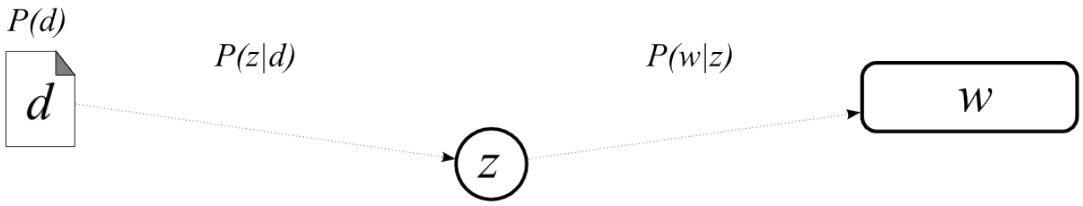
- wähle ein Dokument *d* mit der A-priori-Wkt. *P(d)*
- generiere ein Konzept *z* mit der bedingten Wkt. *P(z|d)*



Modelle mit versteckten Variablen - PLSI

Aspektmodell

- wähle ein Dokument *d* mit der A-priori-Wkt. *P(d)*
- generiere ein Konzept *z* mit der bedingten Wkt. *P(z|d)*
- generiere ein Wort *w* mit der bedingten Wkt. *P(w|z)*

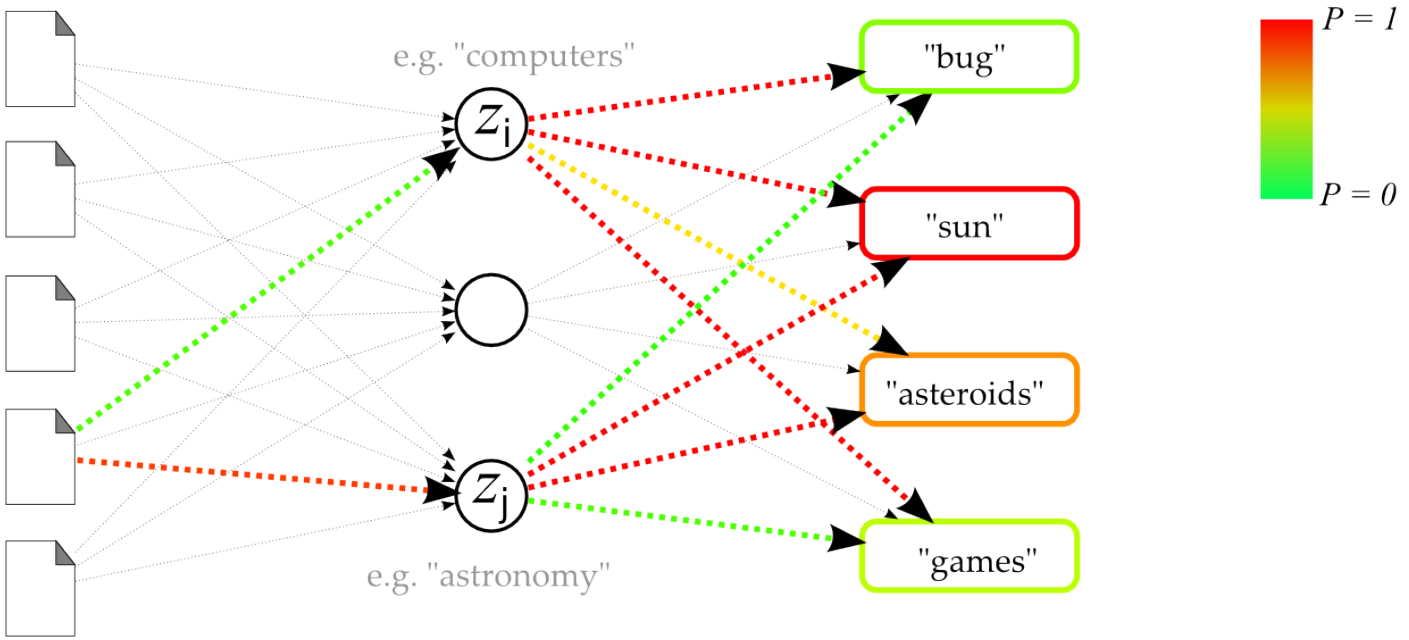


Modelle mit versteckten Variablen - PLSI

Aspektmodell

- formal:

$$P(d,w) = P(d) \sum_{z \in Z} P(z|d) P(w|z)$$



Quelle: G. Weikum, „Advanced IR Models“, Lectures IRDM WS 2005

Modelle mit versteckten Variablen - PLSI

Aspektmodell

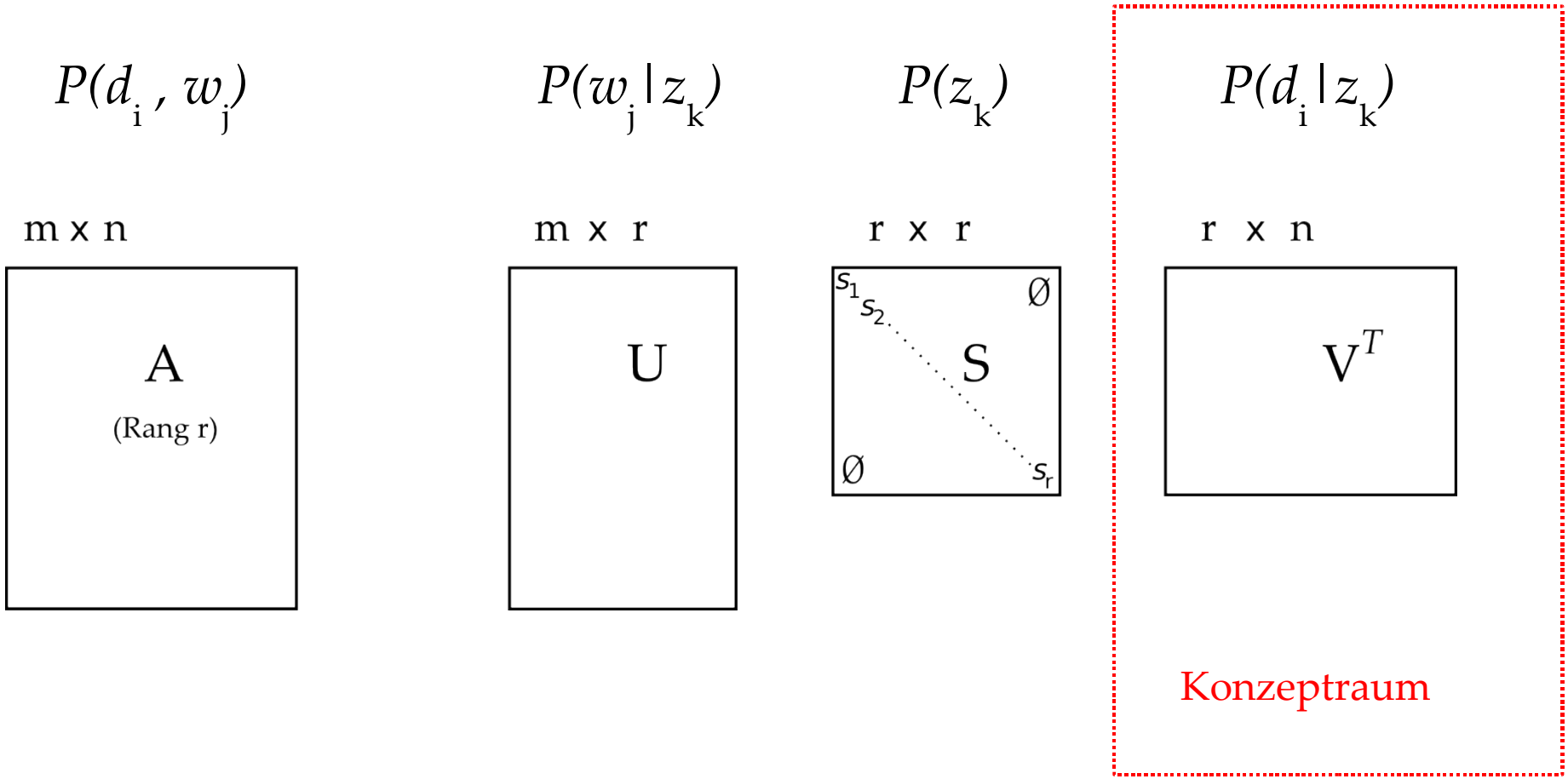
- formal:

$$P(d,w) = P(d) \sum_{z \in Z} P(z|d) P(w|z) = \sum_{z \in Z} P(z) P(w|z) P(d|z)$$

- $P(d, w)$ ist ein beobachtbarer Zusammenhang (Daten)
- $P(z)$, $P(w|z)$, $P(d|z)$ sind Parameter eines Modells, welches die beobachteten Daten generiert
- Anpassung der Modellparameter mit ***Maximum Likelihood Estimation*** (MLE)
- unbeobachtete Parameter erfordern Optimierverfahren zur Schätzung des ML: ***Expectation Maximization*** (EM)

Modelle mit versteckten Variablen - PLSI

- Vergleich PLSI – Parameter mit LSI – Matritzen



Modelle mit versteckten Variablen - PLSI

- Konzeptraumdarstellung neuer Dokumente durch Verwendung der Parameter $P(z)$ und $P(w|z)$ eines optimierten Modells und berechnen der $P(d|z)$ mittels EM
- analog lassen sich neue Terme einfügen

- Retrieval im Konzeptraum analog zu LSI und Vektorraummodell

- Dimension (= Anzahl der Konzepte) wird vor dem Training festgelegt (vgl. LSI – Reduktion nach SVD)
- Wahl der Dimension beeinflusst das Ergebnis (aber weniger stark als bei LSI)

Modelle mit versteckten Variablen - PLSI

Kritik

+ solides statistisches Sprachmodell
+ Synonyme und Homonyme werden erkannt

- (sehr) hoher Trainingsaufwand (EM-Algorithmus: ~$O(N^3)$)
- ungeeignet für heterogene, dynamische Kollektionen (WWW)
- generatives Modell beschränkt auf die Trainingsmenge (neuer Ansatz vorgestellt mit LDA (Blei 2003))

Evaluation

- Clusteringexperimente
- Vergleich Vektorraummodell – LSI – PLSI
- Performanzmaß: F-Measure

Evaluation

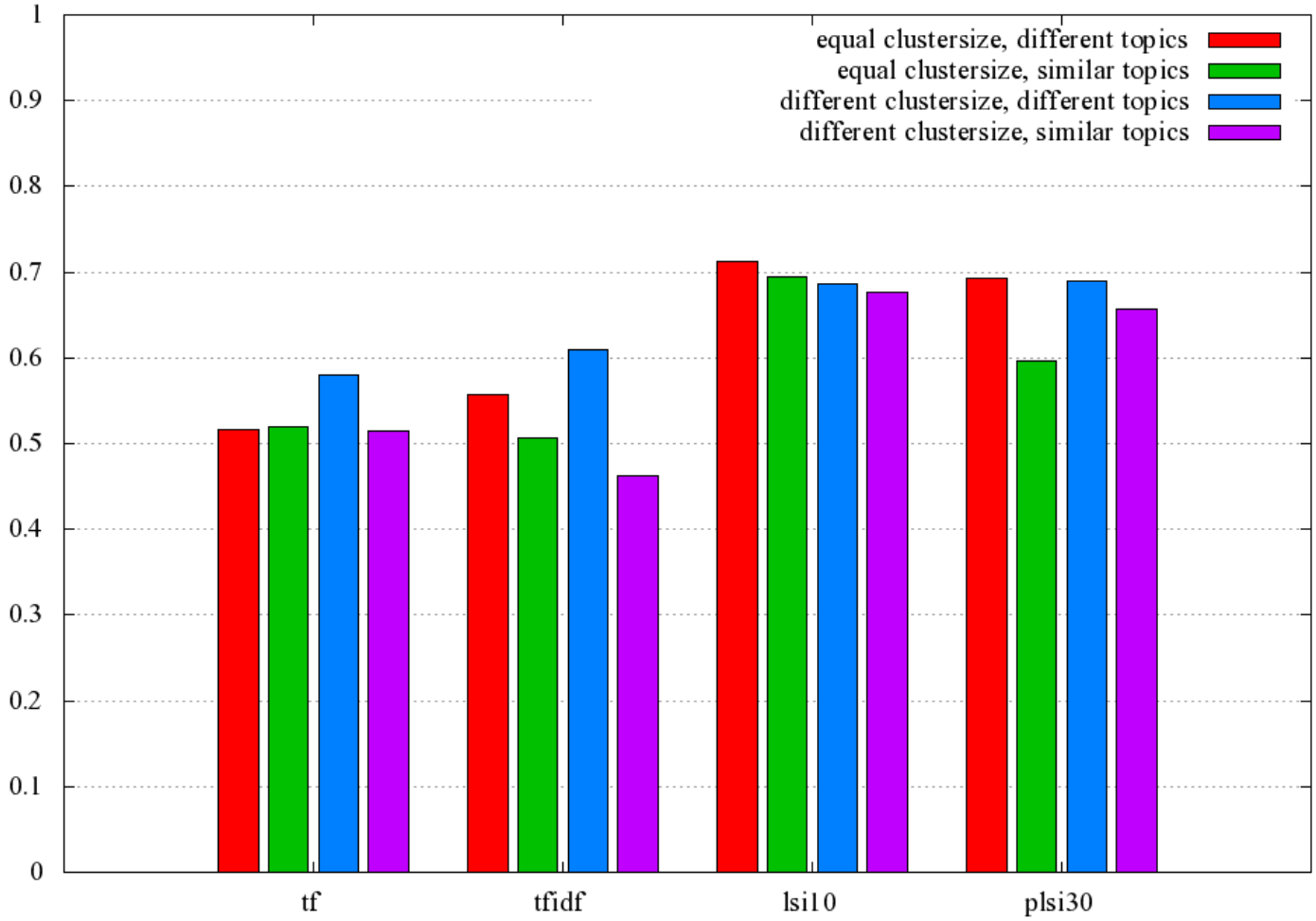
Experimentkollektionen

- Reuters
  - geschlossen, statisch – gute Ergebnisse sind zu erwarten
  - thematisch kategorisiert
  - untersucht wurden verschiedene Charakteristika bzgl. Ähnlichkeit der Themen und Clustergröße (in den Abb. farbig codiert)

- Spock-Trainingskorpus
  - heterogen (WWW) – schlechte Ergebnisse sind zu erwarten
  - kategorisiert nach Personen

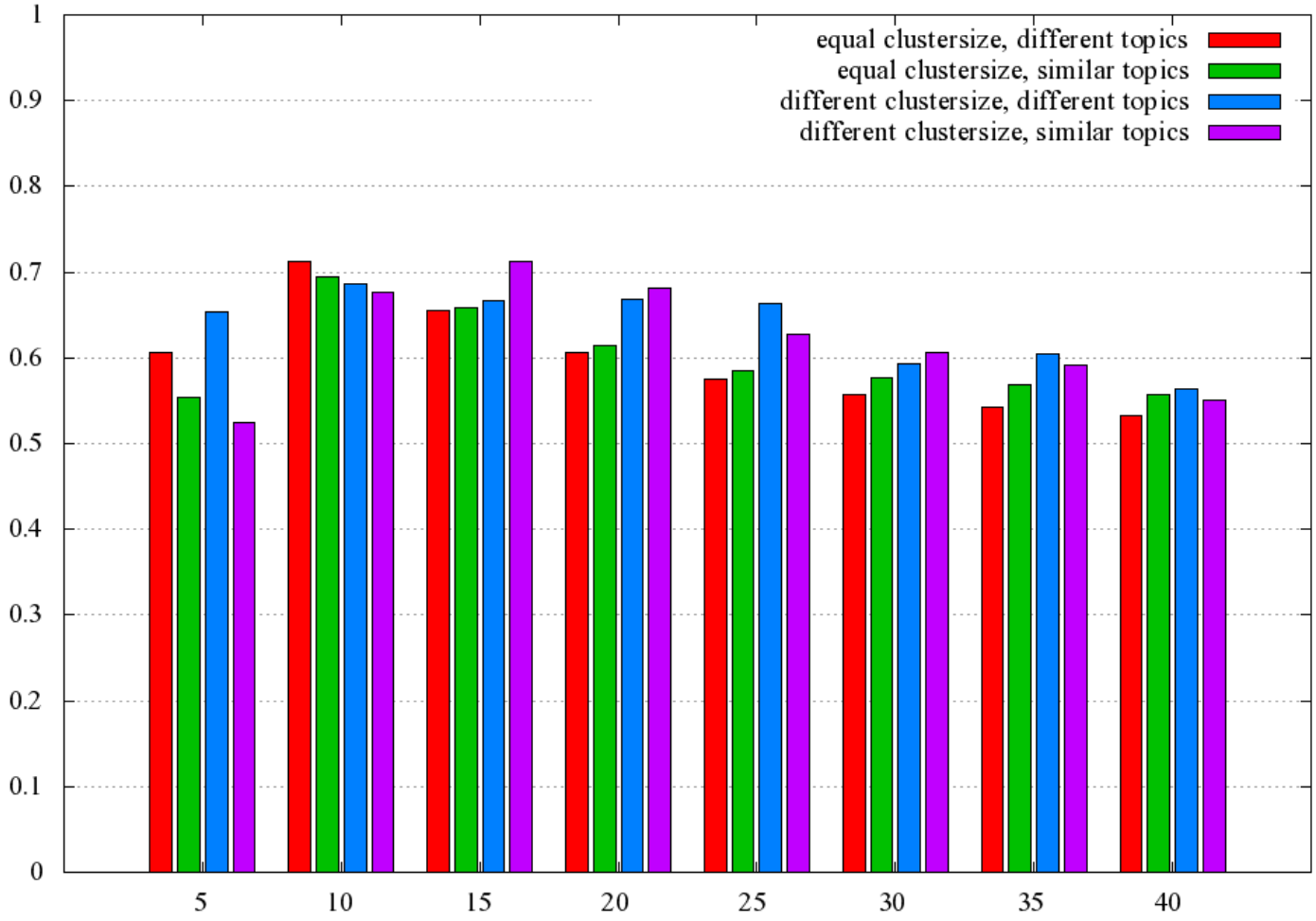
Evaluation

- F-Measures Reuters (Vergleich)



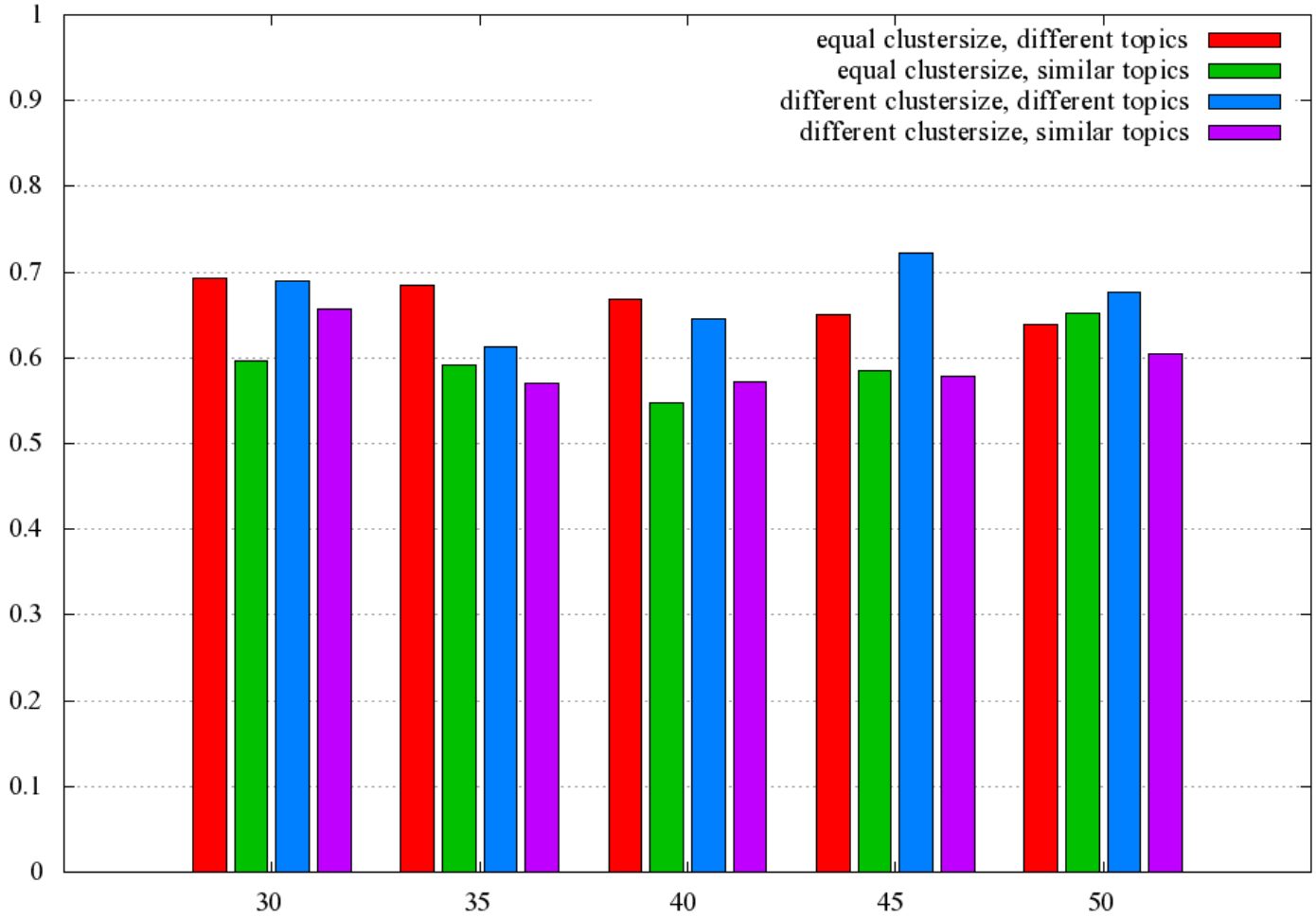
Evaluation

- F-Measures Reuters (LSI, versch. Dimensionen)



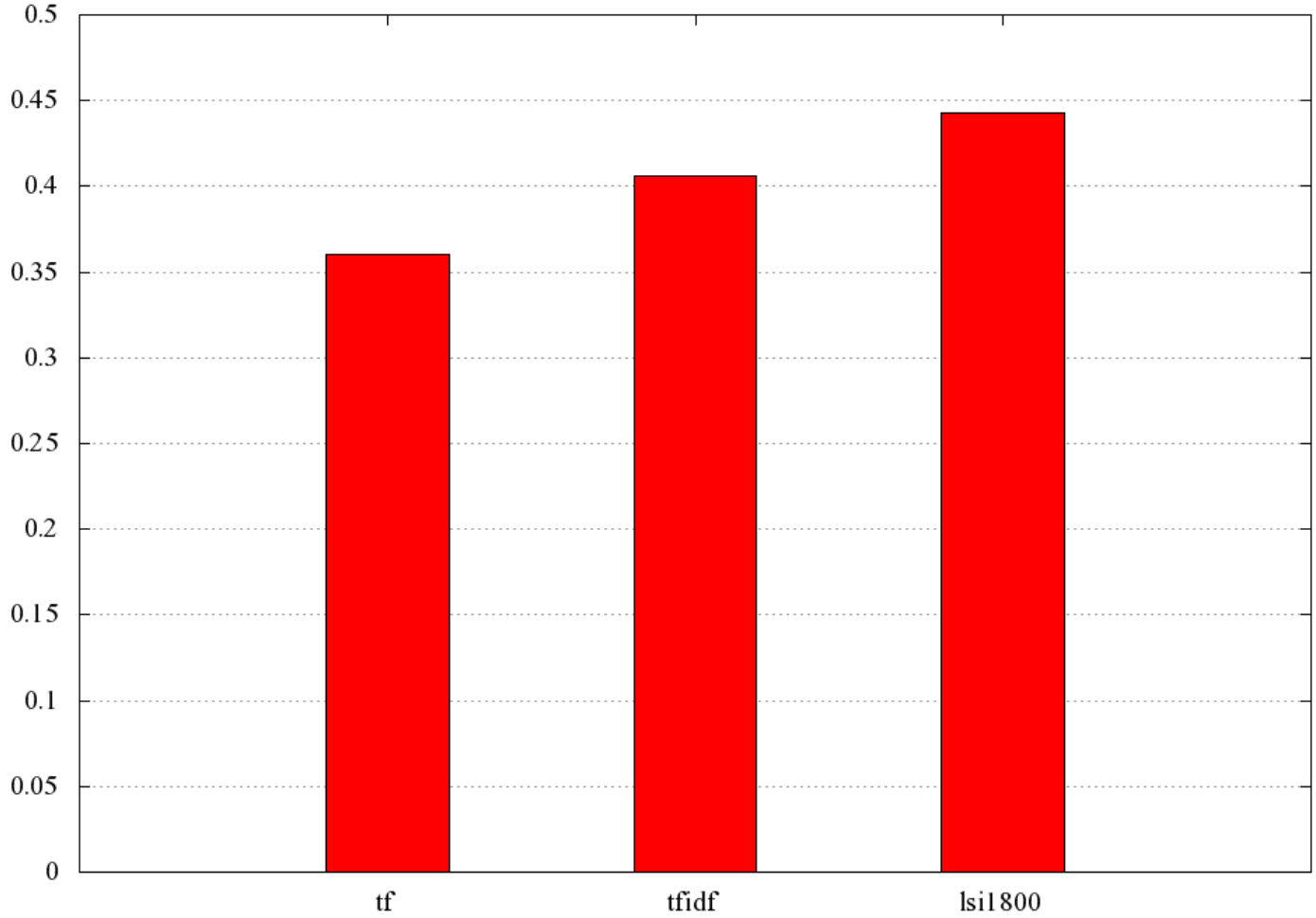
Evaluation

- F-Measures Reuters (PLSI, versch. Dimensionen)



Evaluation

- F-Measures Spock (Vergleich)



Schluss

Danke für Ihre Aufmerksamkeit!