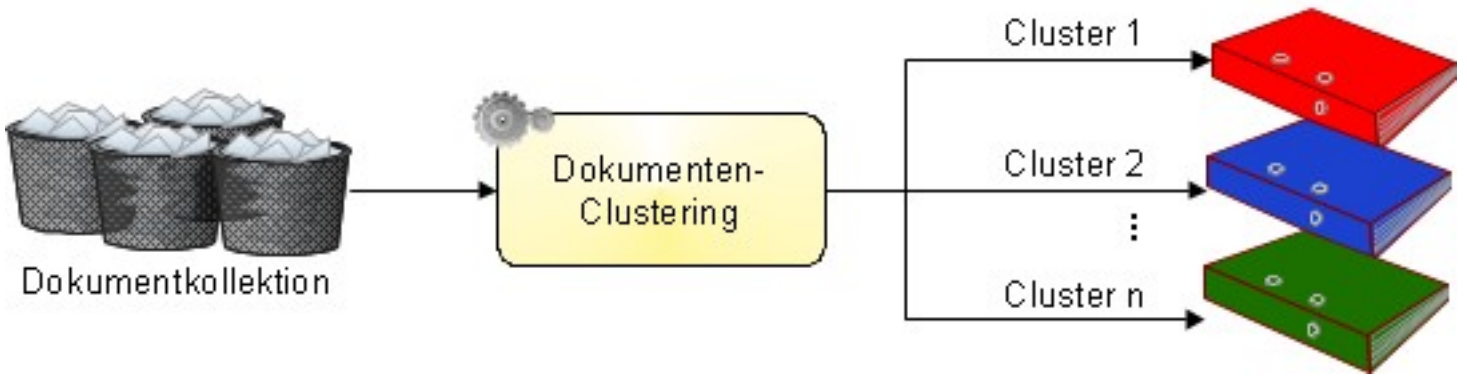
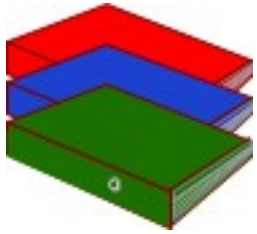
Dokumenten-Clustering





Externe Qualitätsindizes

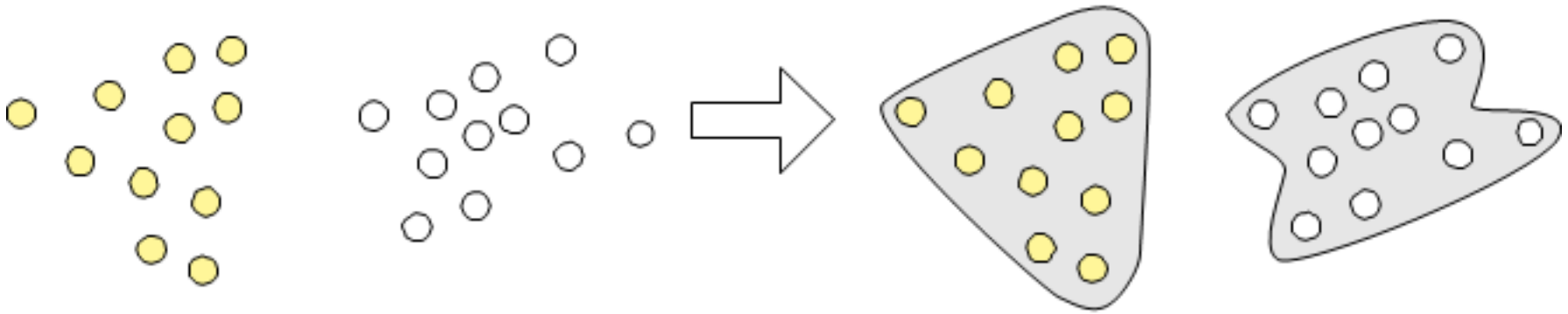
- Vergleich des erzeugten Clusterings mit der Referenzlösung
- F-Measure
- Precision & Recall
- Werte zw. 0 und 1.

Cluster-Analyse

- Ähnlichkeitsmatrix erstellen

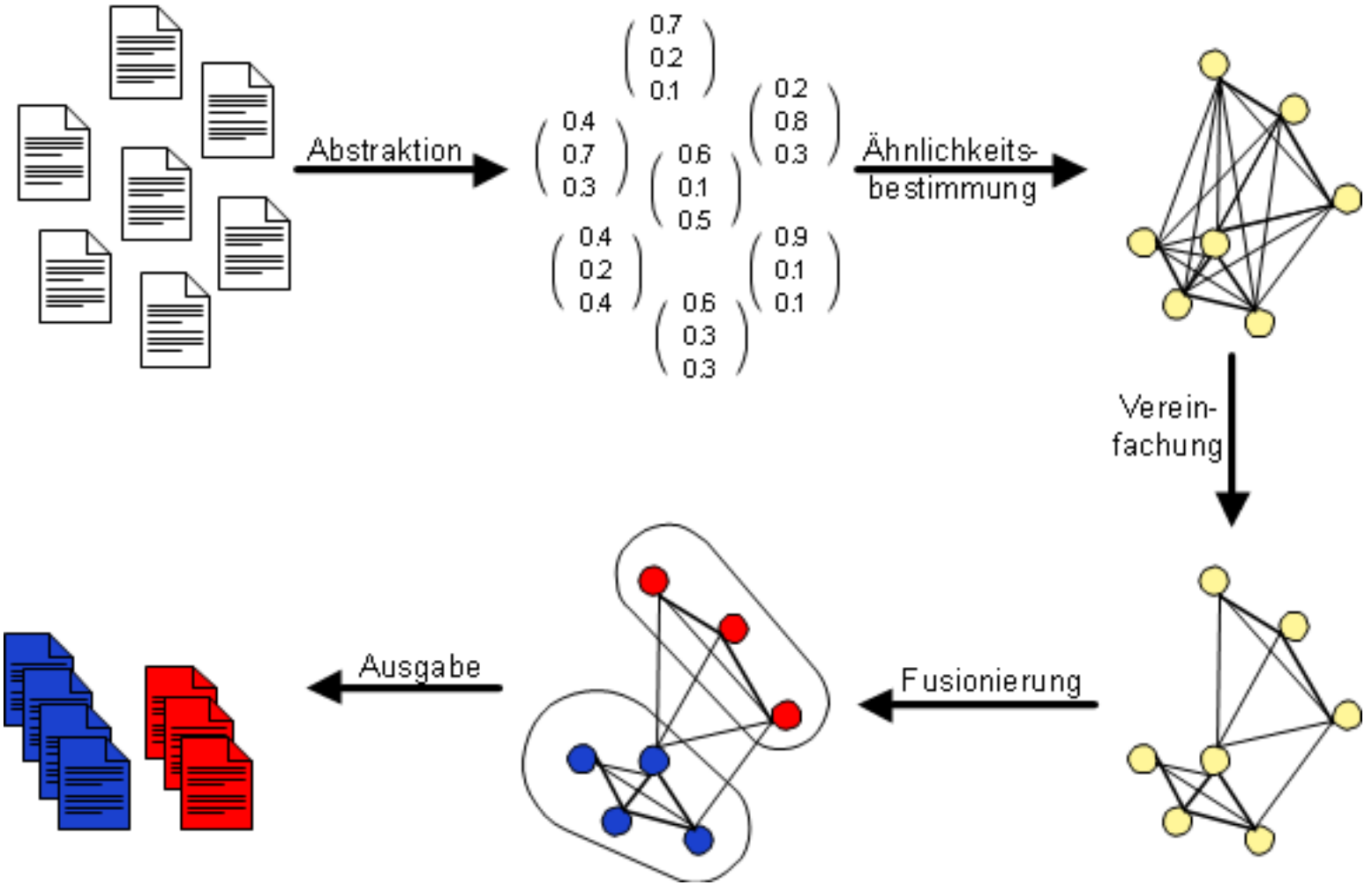
Ähnlichkeitsmatrix	Dokument 1	Dokument 2	$\cdots$	Dokument N
Dokument 1	1	$\varphi(\mathbf{d_1}, \mathbf{d_2})$		$\varphi(\mathbf{d_1}, \mathbf{d_N})$
Dokument 2	$\varphi(\mathbf{d_2}, \mathbf{d_1})$	1		$\varphi(\mathbf{d_2}, \mathbf{d_N})$
$\vdots$			$\ddots$	$\vdots$
Dokument N		$\cdots$	$\cdots$	1

- Fusionierung zu Gruppen



Verfahren zur Modellbildung für das Dokumenten-Clustering

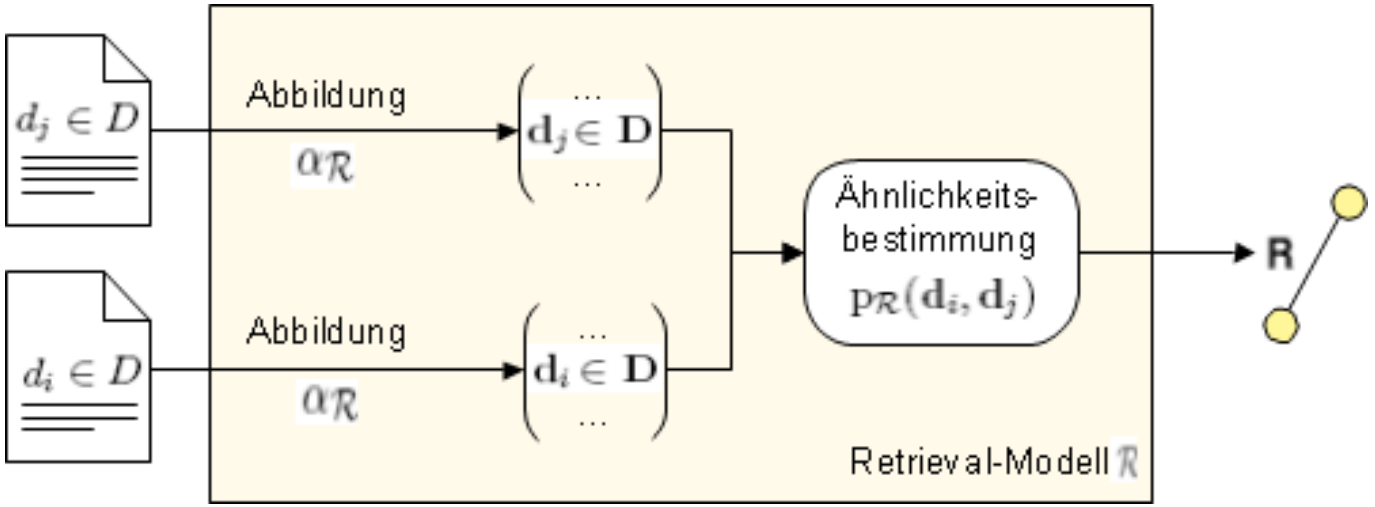
Dokumenten-Clustering



Retrieval-Modell

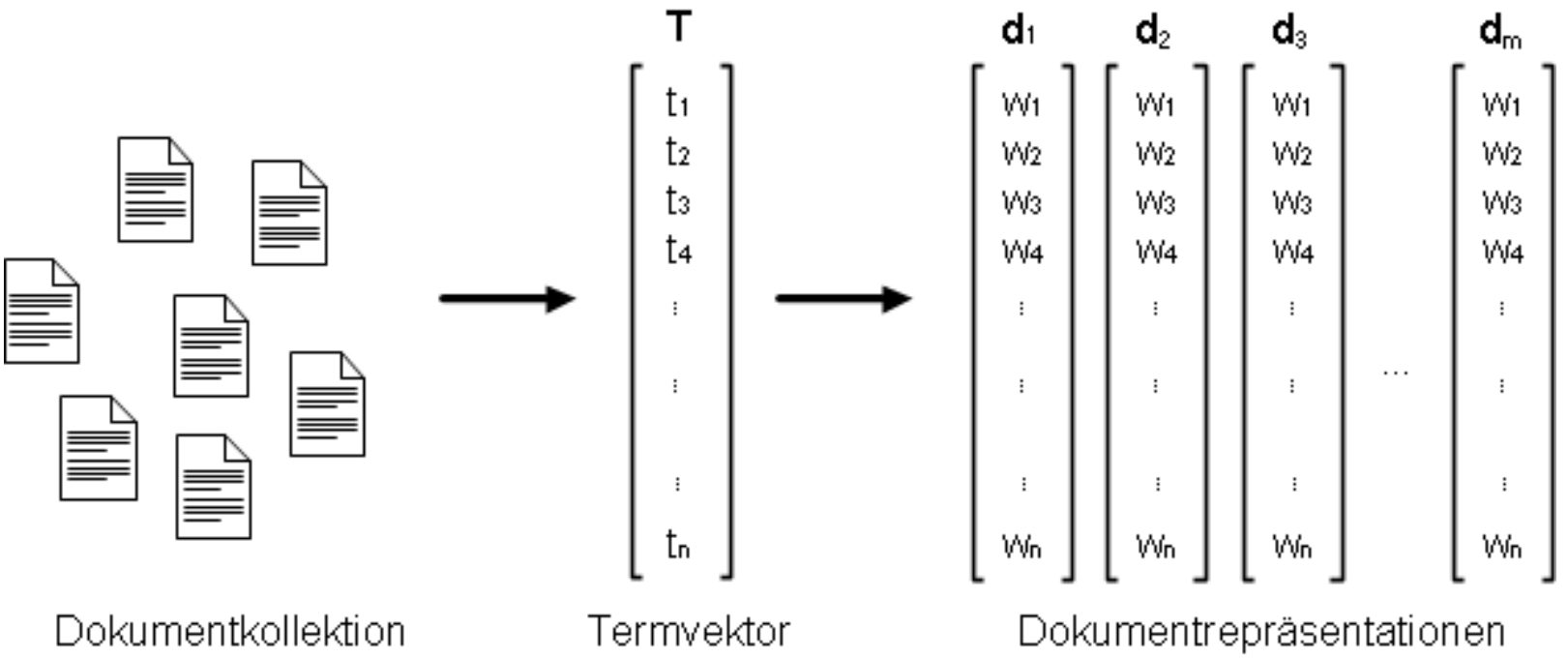
Eine Retrieval-Modell formalisiert eine linguistische Theorie, auf deren Grundlage die thematische Verwandtschaft zwischen zwei Dokumenten bestimmt werden kann.

Konzeptueller Aufbau:



Dokumentrepräsentationen

Indexierung der Dokumente und Gewichtung der Repräsentationen:



Retrieval-Modelle

- Vektorraummodell
- Modell „Divergence From Randomness“
- Best-Match-Modell
- Statistisches Sprachmodell von Ponte & Croft
- Explizit semantisches Modell (ESA)

Vektorraummodell

Termhäufigkeit:

$$w_{tf}(\mathrm{t}, \mathbf{d}) = tf(\mathrm{t}, d)$$

Inverse Dokumenthäufigkeit:

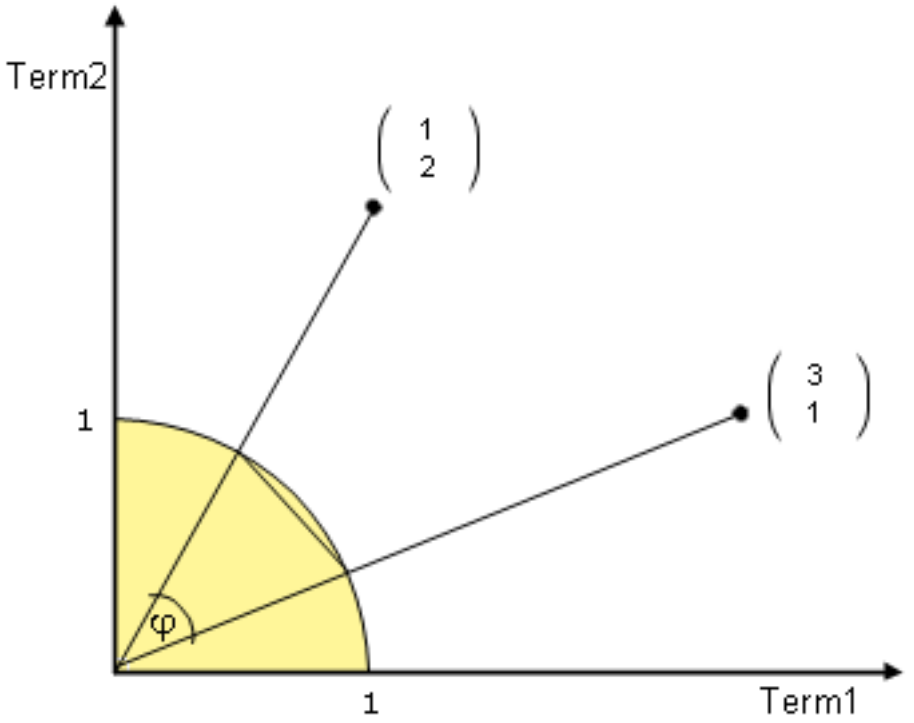
$$w_{Idf}(t) = ld\left(\frac{\mathrm{N}}{\mathrm{n}_t}\right)$$

Tf-Idf:

$$w_{tfIdf}(t, \mathbf{d}) = w_{tf}(t, \mathbf{d}) \cdot w_{IdF}(t)$$

Kosinusähnlichkeit:

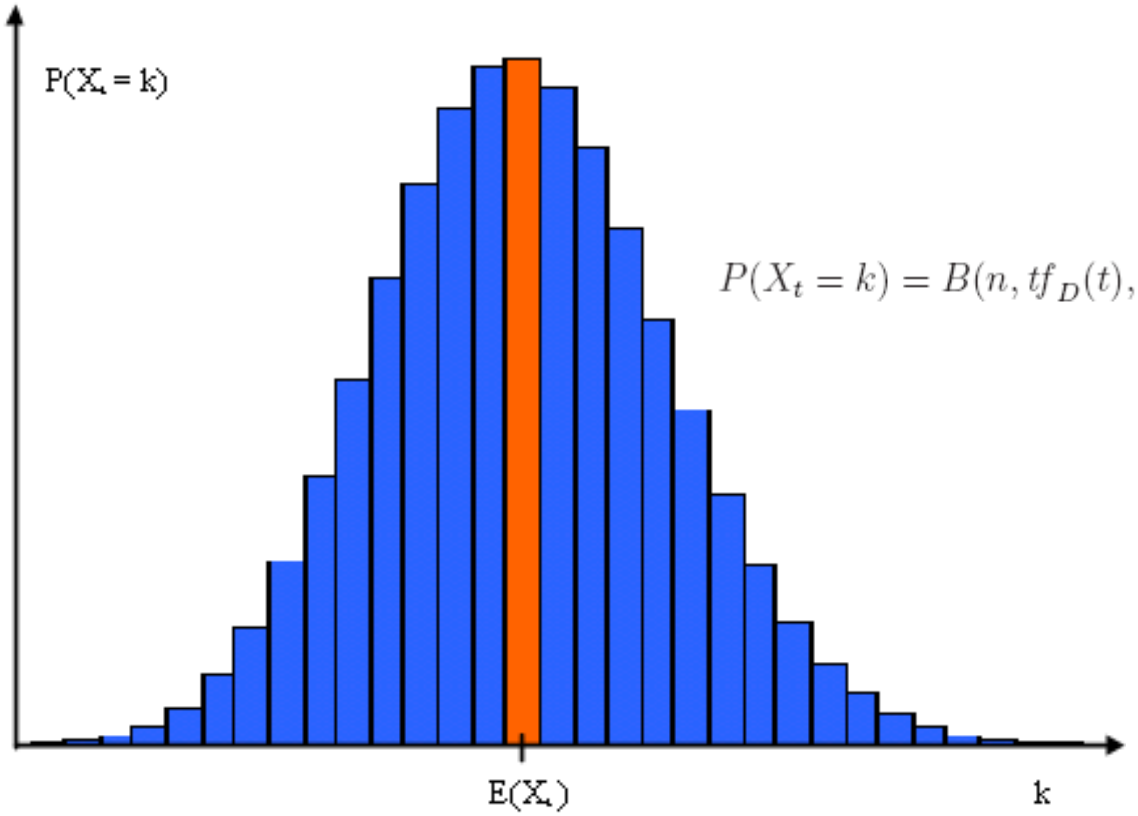
$$\varphi_{cos}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\langle \mathbf{d}_i, \mathbf{d}_j \rangle}{\|\mathbf{d}_i\| \cdot \|\mathbf{d}_j\|}$$



Divergence From Randomness

Annahme: Die Terme wurden zufällig auf die einzelnen Dokumente verteilt.

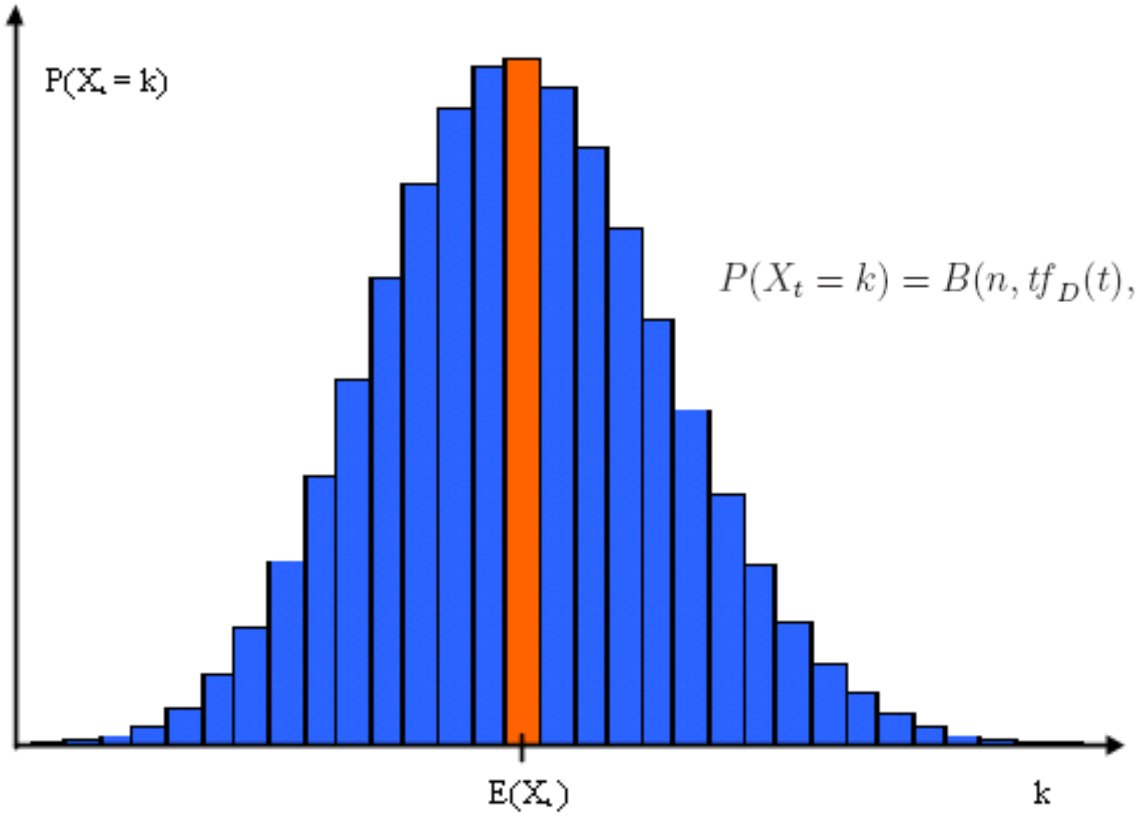
Die Binomialverteilung modelliert die Wahrscheinlichkeit, dass ein Dokument einen Term k Mal zugeordnet bekommen hat:



$$P(X_t = k) = B(n, tf_D(t), k) = \binom{tf_D(t)}{k} p^k q^{tf_D(t)-k}$$

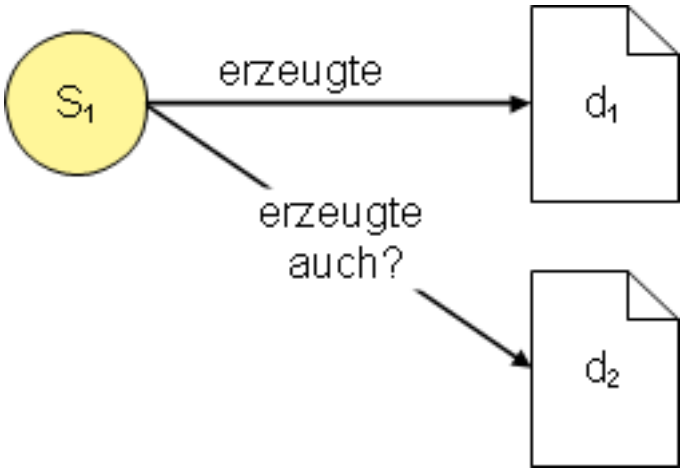
Divergence From Randomness

Betrachtet man die tatsächlich in einem Dokument vorgefundene Termhäufigkeit, gibt die Unwahrscheinlichkeit dieser Beobachtung X einen Hinweis auf die Bedeutsamkeit des Terms im Dokument.



$$P(X_t = k) = B(n, tf_D(t), k) = \binom{tf_D(t)}{k} p^k q^{tf_D(t)-k}$$

Statistisches Sprachmodell



1. Abschätzung:

$$p_{ml}(t|\mathcal{S}_\mathrm{d}) = \frac{tf_{t,d}}{l_d}$$

2. Abschätzung:

$$p_{avg}(t|\mathcal{S}_\mathrm{d}) = \frac{\left(\sum_{d \in D_t} p_{ml}(t|\mathcal{S}_\mathrm{d})\right)}{\mathrm{n}_t}$$

Retrieval-Funktion:

$$\rho_{\mathcal{R}_{\mathrm{Ponte}}}(\mathrm{d}_1, \mathrm{d}_2) = p(\mathrm{d}_2|\mathcal{S}_{\mathrm{d}_1}) =$$

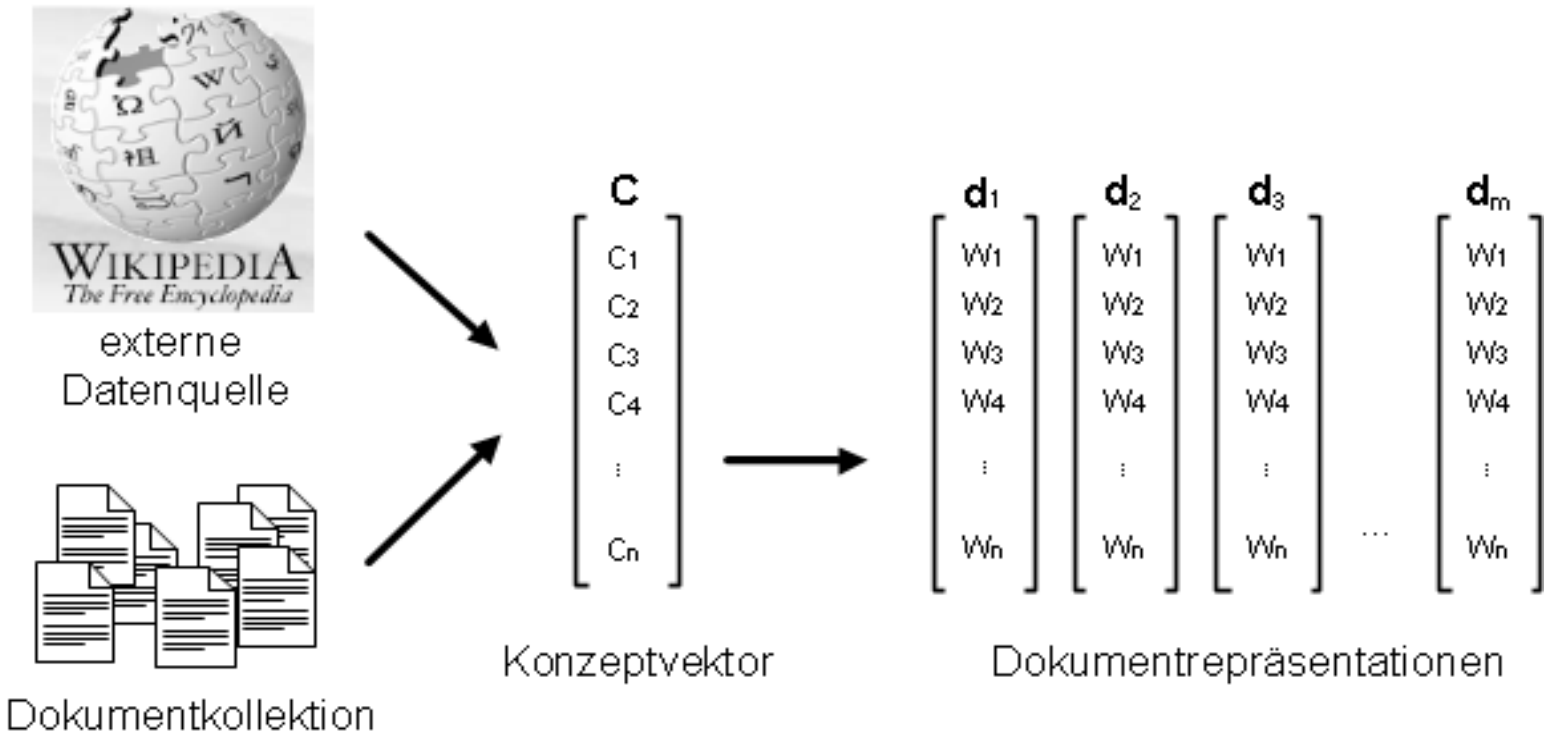
$$\prod_{t \in d_2} p(t|\mathcal{S}_{\mathrm{d}_1}) \cdot \prod_{t \notin d_2} 1 - p(t|\mathcal{S}_{\mathrm{d}_1})$$

Kombination der Schätzer:

$$p(t|\mathcal{S}_\mathrm{d}) = p_{ml}(t|\mathcal{S}_\mathrm{d})^{(1-R_{t,d})} \cdot p_{avg}(t|\mathcal{S}_\mathrm{d})^{R_{t,d}}$$

Retrieval-Modelle

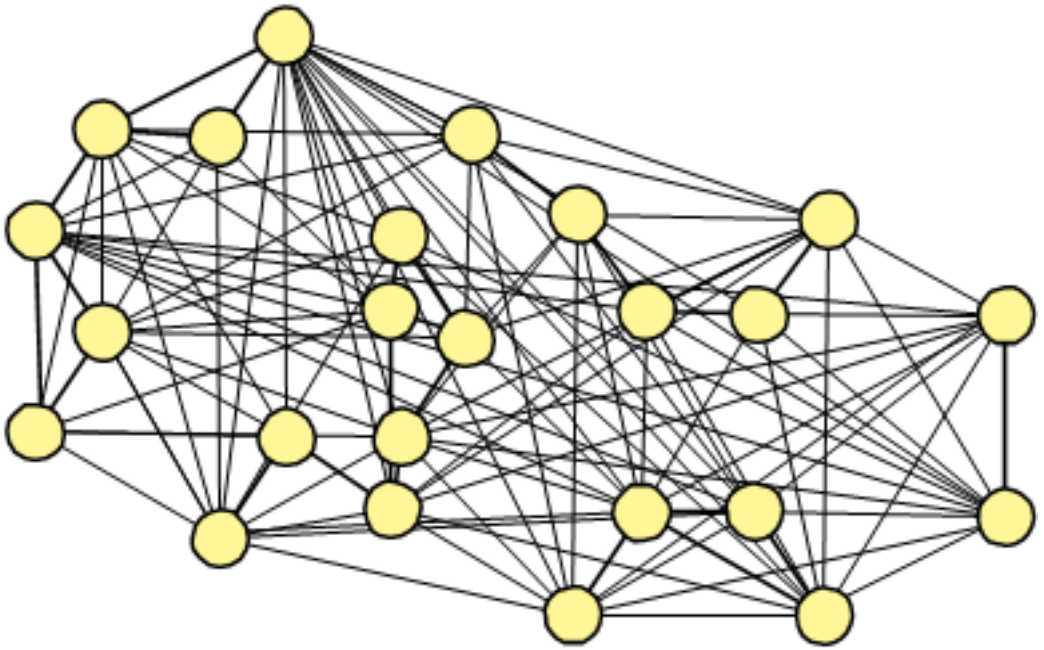
- Explizit semantisches Modell (ESA)



Modellvereinfachung

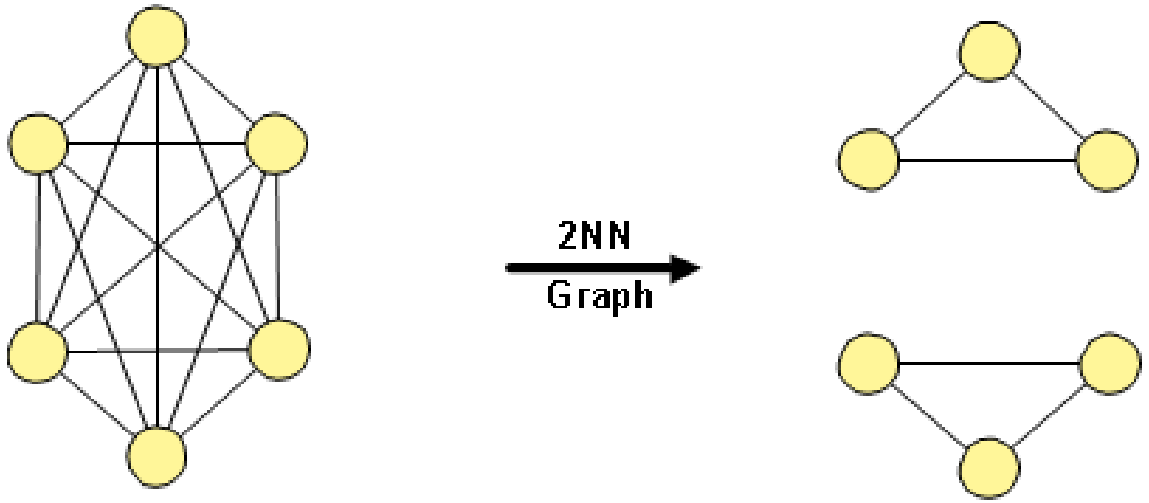
Beobachtung:

Es entstehen viele sehr kleine Ähnlichkeitswerte zwischen Dokumenten ohne thematische Verwandtschaft.



k-Nearest-Neighbor-Graph

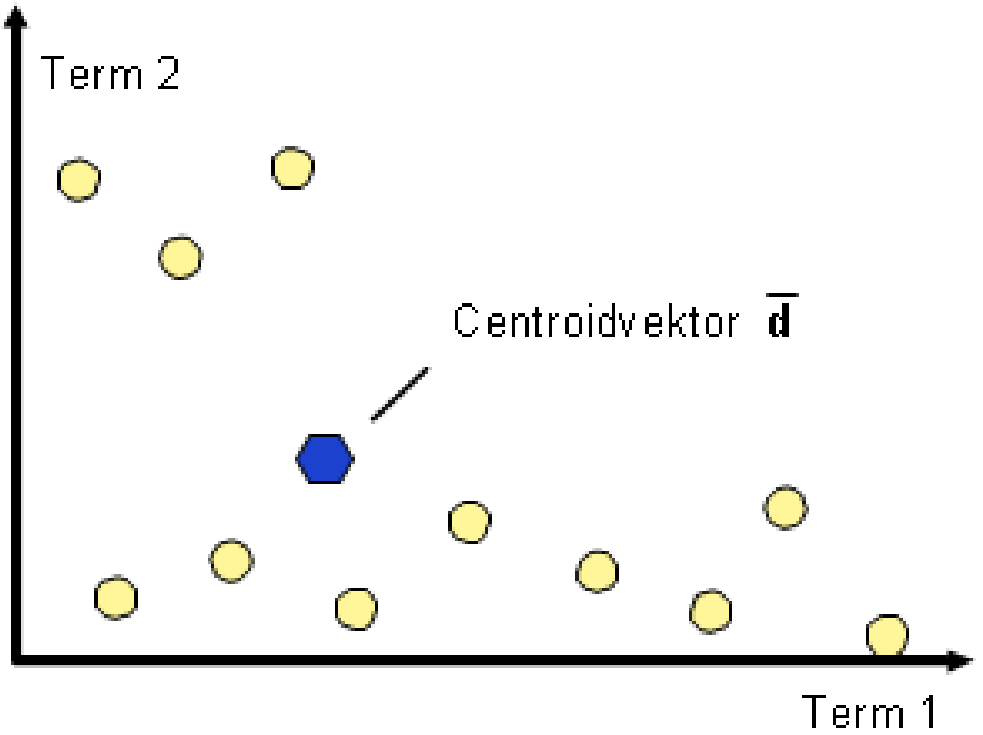
Reduktion der Kanten pro Dokument auf die k schwersten.



Problem: Bestimmung des Parameters k.

Expected Similarity

Nur die Ähnlichkeitswerte, die signifikant größer sind als die erwarteten Ähnlichkeiten, bleiben erhalten.



Centroidbestimmung:

$$[\overline{\mathrm{d}}]_t = \frac{\sum_{i=1}^{\mathrm{N}} w(t, \mathbf{d}_i)}{\mathrm{N}}$$

Erwartete Ähnlichkeit:

$$\overline{\rho_{\mathrm{d}}}(\mathbf{d}_i) = \rho_{\mathcal{R}}(\mathbf{d}_i, \overline{\mathrm{d}})$$

Signifikanzbestimmung:

$$[\mathrm{d}^*]_t = \frac{2 \cdot w(t, \overline{\mathrm{d}}) \cdot w_{max}(t, \mathrm{D})}{w(t, \overline{\mathrm{d}}) + w_{max}(t, \mathrm{D})}$$

Ergebnisse

Mittleres F-Measure der Clusterings, die für die verschiedenen Gewichtungsvorschriften in den Experimenten erzeugt wurden:

Gewichtungsvorschrift	keine Vereinfachung	optimaler kNN-Graph	Expexected Similarity
Sprachmodell	0.22	0.86	0.86
Tf	0.22	0.85	0.83
BestMatch25	0.2	0.85	0.83
Ltu	0.23	0.83	0.81
Lnu	0.23	0.83	0.8
Div. from Rand. 1	0.32	0.82	0.76
TfIdf	0.3	0.80	0.73
ESA	0.32	0.78	0.77
Div. from Rand. 2	0.29	0.67	0.64

Das Verfahren Expected Similarity erreicht parameterlos die Qualität der besten Vereinfachung, die mit einem kNN-Graphen möglich ist.

Zusammenfassung

- Retrieval-Modelle implementiert und experimentell verglichen.
- Interne Qualitätsmaße auf ihre Leistungsfähigkeit hin untersucht.
- Verfahren zur Modellvereinfachung entwickelt und evaluiert.

Vielen Dank für die Aufmerksamkeit!

Testkollektionen

Reuters-Kollektionen	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
# Kategorien	3	3	3	5	5	5	6
# Dokumente	300	1500	1500	500	1500	2500	600
gleichverteilt	ja	ja	nein	ja	ja	nein	ja

LATimes-Kollektionen	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
# Cluster	3	3	3	5	5	5	6
# Dokumente	300	1500	1500	500	1500	2500	600
gleichverteilt	ja	ja	nein	ja	ja	nein	ja

Modellvereinfachung

- Expected Similarity

Exp. Sim.	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.9	0.96	0.97	0.62	0.56	0.68	0.43
LM	0.92	0.97	0.97	0.66	0.64	0.69	0.6
Lnu	0.81	0.92	0.91	0.55	0.56	0.61	0.43
TfIdf	0.88	0.97	0.96	0.51	0.62	0.69	0.42
Ltu	0.79	0.96	0.96	0.6	0.65	0.66	0.55
INBH2	0.75	0.96	0.92	0.63	0.69	0.73	0.59
PLH2	0.7	0.94	0.83	0.27	0.13	0.4	0.13
BM25	0.88	0.97	0.97	0.68	0.55	0.69	0.65
ESA	0.9	0.90	0.9	0.69	0.68	0.75	0.61

Unmittelbare Ergebnisse

Roh	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.17	0.17	0.3	0.07	0.07	0.16	0.05
LM	0.17	0.33	0.3	0.07	0.07	0.16	0.05
Lnu	0.17	0.17	0.38	0.07	0.07	0.16	0.05
TfIdf	0.33	0.49	0.54	0.07	0.07	0.16	0.05
Ltu	0.33	0.17	0.3	0.07	0.07	0.16	0.05
INBH2	0.49	0.33	0.47	0.07	0.07	0.16	0.05
PLH2	0.17	0.17	0.38	0.07	0.07	0.16	0.05
BM25	0.17	0.17	0.3	0.07	0.07	0.16	0.05
ESA	0.44	0.51	0.62	0.07	0.07	0.38	0.05

Ergebnisse

Mittleres F-Measure der Clusterings, die für die verschiedenen Gewichtungsvorschriften in den Experimenten erzeugt wurden:

Gewichtungsvorschrift	keine Vereinfachung	optimaler kNN-Graph
Sprachmodell	0.22	0.86
Tf	0.22	0.85
BestMatch25	0.2	0.85
Ltu	0.23	0.83
Lnu	0.23	0.83
Div. from Rand. 1	0.32	0.82
TfIdf	0.3	0.80
ESA	0.32	0.78
Div. from Rand. 2	0.29	0.67

Mit dem kNN-Graphen ergibt sich eine deutliche Qualitätsverbesserung.

Modellvereinfachung

- Expected Similarity

Exp. Sim.	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.90	0.98	0.99	0.95	0.89	0.89	0.86
LM	0.94	0.98	0.98	0.94	0.9	0.9	0.87
Lnu	0.91	0.98	0.99	0.97	0.8	0.96	0.86
TfIdf	0.76	0.95	0.89	0.89	0.54	0.39	0.78
Ltu	0.87	0.98	0.99	0.93	0.79	0.70	0.88
INBH2	0.68	0.92	0.83	0.83	0.72	0.67	0.7
PLH2	0.75	0.88	0.86	0.84	0.75	0.66	0.75
BM25	0.79	0.91	0.97	0.93	0.84	0.83	0.90
ESA	0.93	0.92	0.89	0.85	0.58	0.38	0.75

Erzeugung der Ähnlichkeitsmatrix

- Indexierung und Gewichtung

  Verwendete Informationen:

  - Termhäufigkeit $tf(t,d)$.
  - Dokumenthäufigkeit $df(t)$.

Modellvereinfachung

- k-Nearest-Neighbor-Graph

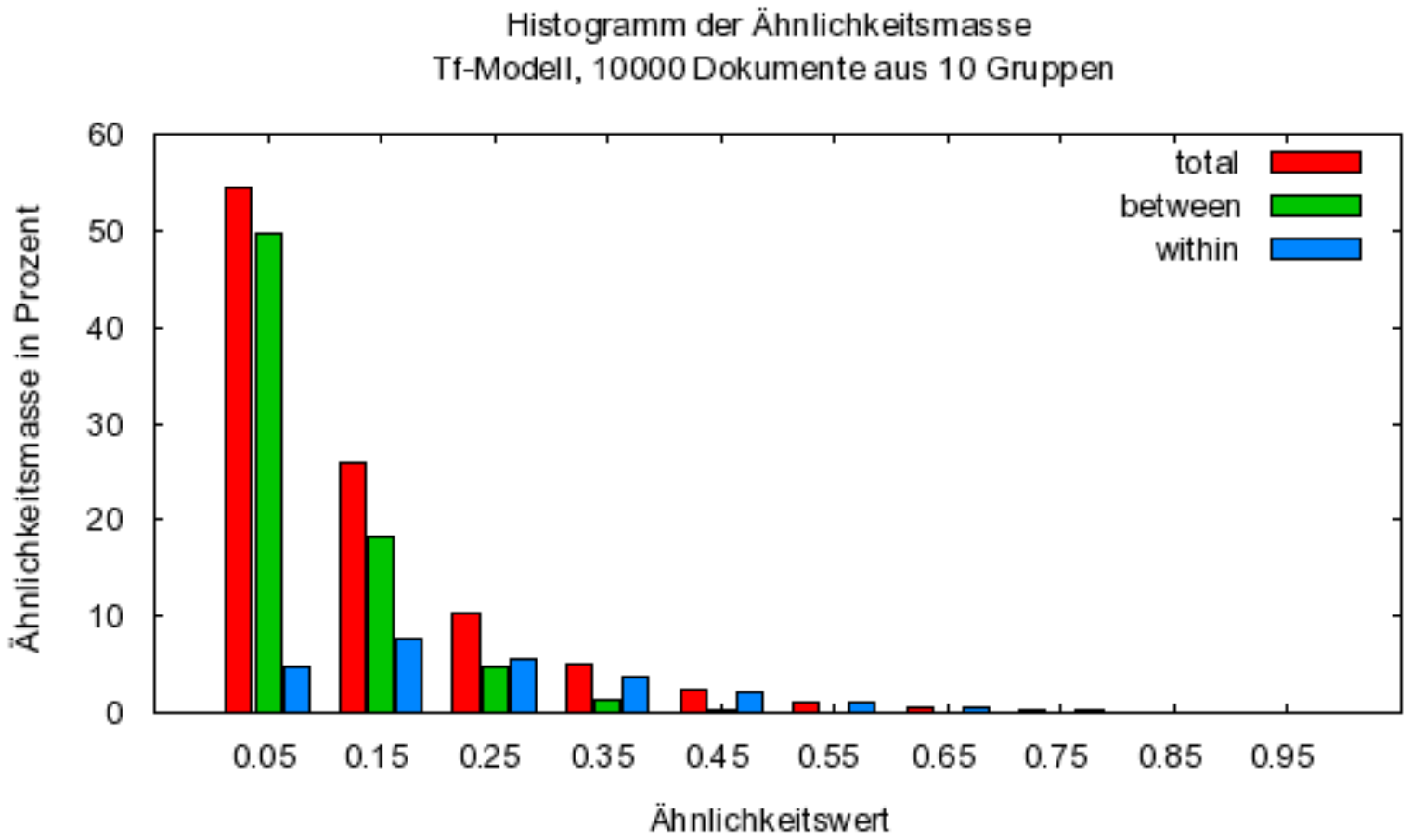
k-Max	LC3E300	LC3E1500	LC3U1500	LC5E500	LC5E1500	LC5U2500	LC6E600
lnc	0.95	0.96	0.97	0.61	0.56	0.69	0.59
LM	0.95	0.97	0.97	0.67	0.66	0.69	0.59
Lnu	0.84	0.8	0.91	0.58	0.6	0.7	0.56
TfIdf	0.97	0.98	0.98	0.72	0.5	0.68	0.48
Ltu	0.87	0.94	0.95	0.58	0.66	0.74	0.59
INBH2	0.94	0.96	0.97	0.57	0.71	0.74	0.61
PLH2	0.93	0.95	0.9	0.2	0.15	0.35	0.25
BM25	0.95	0.97	0.97	0.68	0.66	0.69	0.62
ESA	0.93	0.92	0.93	0.69	0.69	0.59	0.54

Modellvereinfachung

- k-Nearest-Neighbor-Graph

k-Max	RC3E300	RC3E1500	RC3U1500	RC5E500	RC5E1500	RC5U2500	RC6E600
lnc	0.98	0.98	0.99	0.95	0.9	0.8	0.89
LM	0.95	0.98	0.99	0.96	0.92	0.74	0.92
Lnu	0.98	0.99	0.99	0.98	0.85	0.93	0.86
TfIdf	0.97	0.98	0.97	0.93	0.6	0.74	0.75
Ltu	0.98	0.99	0.99	0.95	0.80	0.72	0.85
INBH2	0.95	0.98	0.96	0.95	0.74	0.61	0.78
PLH2	0.95	0.98	0.92	0.88	0.7	0.61	0.74
BM25	0.96	0.96	0.97	0.91	0.77	0.83	0.91
ESA	0.96	0.95	0.94	0.84	0.58	0.58	0.74

Zufällige Termüberschneidungen



Expected Similarity

- Ermittlung von Signifikanz über das harmonische Mittel:

$$[\mathbf{d}_{harm}]_t = \frac{2 \cdot w(t, \overline{\mathbf{d}}) \cdot w_{max}(t, \mathbf{D})}{w(t, \overline{\mathbf{d}}) + w_{max}(t, \mathbf{D})}$$