

Wortschatz Zeitgeist

Wolfgang Otto, Thomas Döring, Max Kießling

Seminar Anwendungen der linguistischen Informatik

16. Juni 2015

Wörter des Tages

Wörter des Tages – Portal

wortschatz lexikon

Die »Wörter des Tages« zeigen, welche Begriffe heute besonders aktuell sind. Dazu werden verschiedene Tageszeitungen und Newsdienste täglich ausgewertet. Die »Wörter des Tages« stehen morgens ab etwa 7 Uhr zur Verfügung.

Die Aktualität eines Begriffs ergibt sich aus seiner Häufigkeit heute, verglichen mit seiner durchschnittlichen Häufigkeit über längere Zeit hinweg.

Wortschatz · Wörter des Tages · 25.03.2015	
Sportler, Trainer, Funktionäre	Barrichello · Bastian Schweinsteiger · Baur · Beiersdorfer · Diego Simeone · Dietmar Beiersdorfer · Holger Badstuber · Joachim Löw · Loeb · Lukas Podolski · Marc Marquez · Michel Platini · Petkovic · Platini · Raphael · Riis · Sepp Blatter · Valentino Rossi · Weidenfeller · Wolfgang Niersbach
Sport	DFB-Präsident · EM-Qualifikationsspiel · Feyenoord · UEFA · Uefa · VfR Aalen
Politiker	Barack Obama · Cameron · Erdogan · Frank-Walter Steinmeier · Hermann Gröhe · Hollande · Jazencuk · Kretschmann · Mussolini
Organisation	AKP · Airbus · Amazon · BNP Paribas · Bawag · Charles Vögele · Citigroup · DFL · Dell · EuGH · Euro-Gruppe · Europäischen Fußball-Union · Evonik · Evotec · Exekutivkomitee · Fatah · Foundation · Fresenius · GDL · GNU · Germanwings · GfK · HDI · HGST · Hornbach · IAB · IG BCE · IWC · Ican · Juniper · Kantonalbank · Komax · Le Figaro · Logitech · Lokführergewerkschaft · Monsanto · Morgan Stanley · Paribas · Pixar · Repsol · Roche · SKODA · Santander · Skyguide · Swisscom · Tui · Universal · Universal Music · Valentino
Ereignis	Antrittsbesuch · Bohrungen · Exekution · Rückenwind
Schlagwort	Abschaltung · Absturz · Altersarmut · Android-Geräte · Antibiotika · Batman · Biogas · Consumer · Cyanogenmod · Dividenden · Domain · Domains · EFSF · EM-Qualifikation · Ecopop · Einkaufsmanagerindex · Elektroautos · Flugschreiber · Flugsicherung · Fracking · GT · Giftspritze · Google Maps · Internetnutzer · Kauflaune · Kriminalstatistik · Krisenstab · Lkw-Maut · MMS · MacBook · Mehrwertsteuersatz · Migrationshintergrund · NPD-Verbotsverfahren · Notruf · Pkw-Maut · Pollen · Raspberry Pi · Reach · Reformpaket · Schadcode · Sicherheitslücke · Sinkflug · Stimmrechtsbeschränkung · Swift · Süddeutsche · TecDax · WhatsApp · Widder · Zombies
Ort	Aalen · Alexandria · Bam · Barcelona · Kansas · Kansas City · Landebahn · Liechtenstein · Minnesota · Palo Alto · Sepang · Silicon Valley · Südrfrankreich · Tiflis · Utah
Personen aus Kunst, Kultur und Wissenschaft	Angelina Jolie · Boulez · Euler · Glass · Monroe
sonstige Personen	Allergiker · Angelina · Assange · Beitz · Carsten Spohr · Ex-Präsidenten · Fluglotsen · Gardner · Hannelore Kraft · Jürgen Großmann · Lagerfeld · May · Oleg · Papst Franziskus · Pete · Rahmstorf · Teammanager · Theo Zwanziger · V-Leute · Winterthur

- Sammelt tagesaktuelle Begriffe aus täglich gecrawlten Newsseiten
- Begriffe nach Themen-/Sachgruppen geordnet

Wörter des Tages

Fragestellungen

- Was macht ein Wort des Tages aus?
- Wie erkenne ich ein Wort des Tages
- Wie kann es berechnet werden?

Algorithmen

Relative Häufigkeit

Idee: Tokens, deren relatives Auftreten am gewählten Tag im Verhältnis zum relativen Auftreten im Referenzzeitraum (2014) besonders groß ist, sind interessante Wörter.

Formel:

$$\text{sigfreqratio}(w) = \frac{\frac{k_{\text{day}}}{n_{\text{day}}}}{\frac{k_{2014}}{n_{2014}}} \quad (1)$$

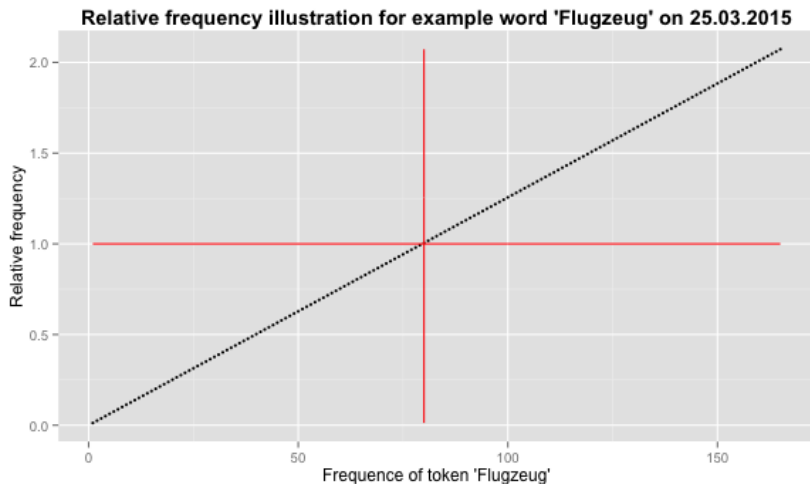
k_{day} : Frequenz des Tokens an einem Tag

n_{day} : Summe der Frequenzen aller Tokens eines Tages

k_{2014} : Frequenz des Tokens im Referenz Zeitrahmen (2014)

n_{day} : Summe der Frequenzen aller Tokens im Referenzzeitrahmen (2014)

Relative Häufigkeit



Relative Häufigkeiten - Bemerkungen

- Erster Ansatz
- Einfache Implementierung
- Selten Auftretende Wörter werden gegenüber anderen interessanten Wörtern bevorteilt
- Positiv: Hochfrequente Worte werden selten hoch gerankt

TF/IDF

Idee: Wir gewichten die Auftretensfrequenz eines Token an einem Tag mit dem Inversen einer Maßzahl, die Angibt an wie vielen Tagen im Referenzjahr das Wort erwähnt wurde.

Modifikationen:

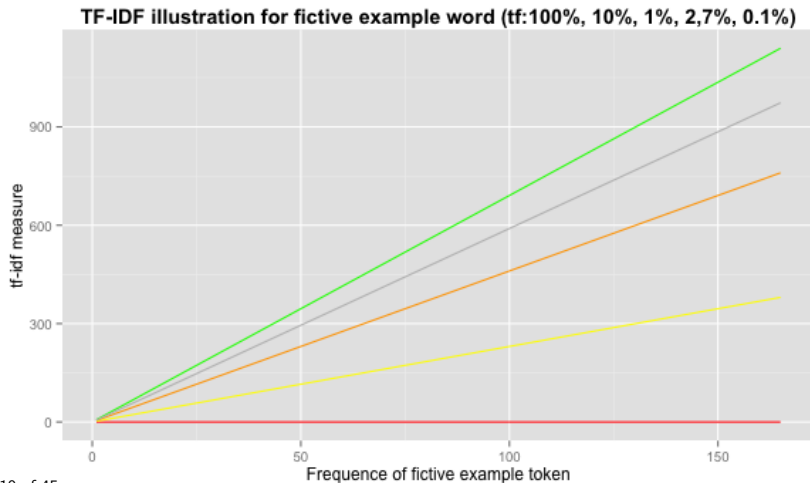
- Relativierung der Frequenz auf Frequenz des häufigsten Tokens am Tag (Vergleichbarkeit)
- Logarithmieren des IDF-Wertes

Formel:

$$sig_{tfidf}(w) = \frac{k}{\max(K)} \cdot \log\left(\frac{365}{documentdays(w)}\right) \quad (2)$$

k : Frequenz eines Tokens an einem Tag K : Alle Frequenzen an einem Tag

TF-IDF Beispiel



Poisson als Maß

Idee: Modellierung der Wahrscheinlichkeit eine bestimmte Frequenz eines Tokens zu sehen. Wenn die Tagesfrequenz eines Tokens sehr unwahrscheinlich ist, ist das Token interessant.

Annahme: Diese Wahrscheinlichkeiten sind Poisson-Verteilt.

Formel der Poisson-Verteilung allgemein:

$$P_{\lambda}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (3)$$

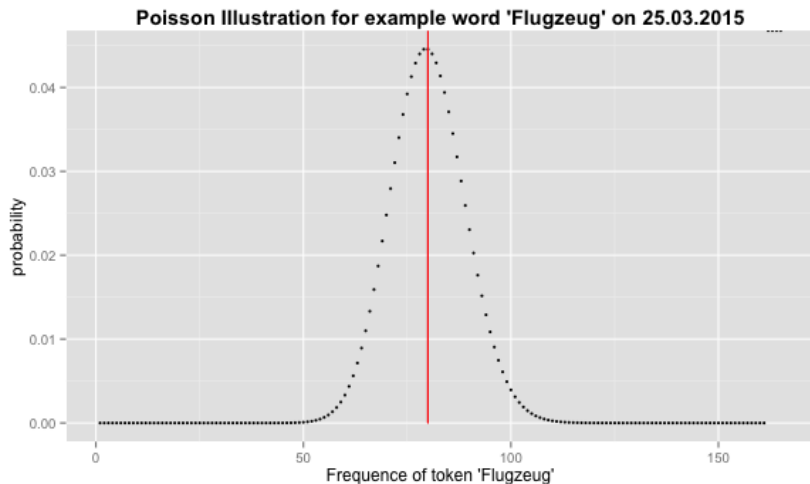
λ : Welche Frequenz wird erwartet

(relativer Anteil im Referenzkorpus · Umfang des Tageskorpus)

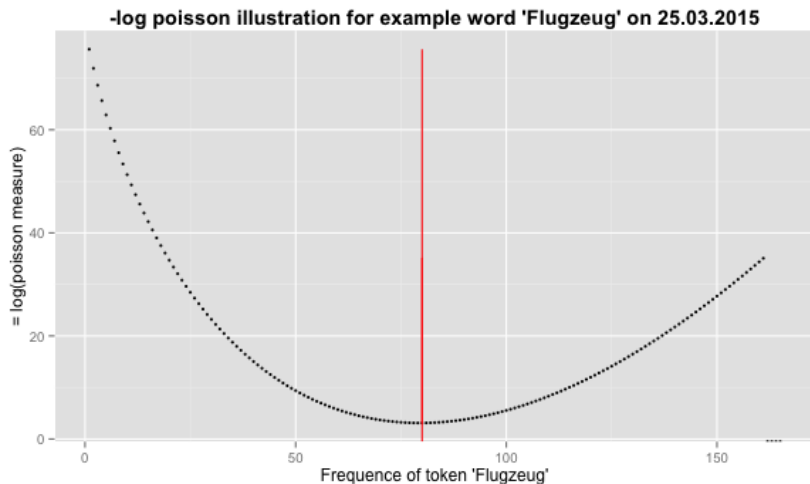
k : tatsächliches Auftreten von einem Wort k

$P_{\lambda}(k)$: Erwartete Wahrscheinlichkeit meine Beobachtung k

Poisson Verteilung



Poisson Verteilung II



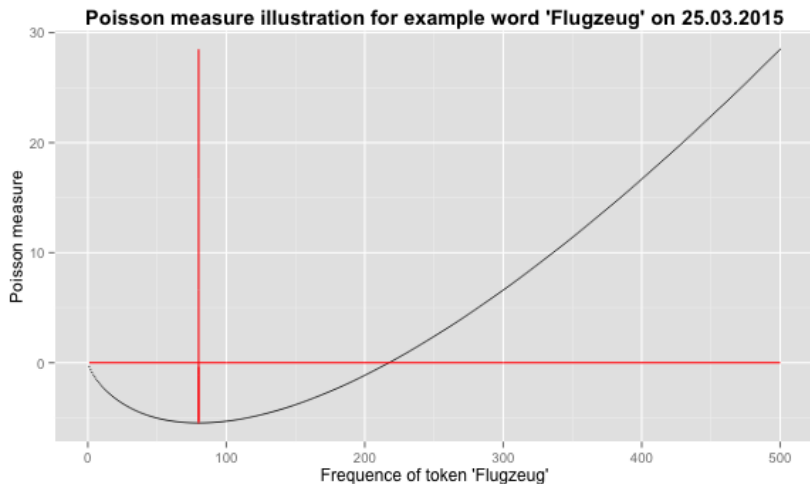
Poisson als Maß: Implementierung

- Problem: Berechnung der Fakultät
- Vergleichbarkeit der Werte einzelner Tage untereinander
- Ziel: Hoher Rang soll einen hohen Wert haben ($-\log$ -Methode)
- Wenn die Frequenz unterdurchschnittlich ist, soll kein hoher Wert erzeugt werden

Formel:

$$\text{sig}_{\text{poisson}}(w) = \frac{k(\log(k) - \log(n \cdot p) - 1))}{\log(n)} \quad (4)$$

Poisson-Verteilung



Einschub – Wortzahl vs. Satzzahl

- Bei der Referenz wird mit Satzzahlen gearbeitet
- Jeder Satz hat im Schnitt gleiche Anzahl von Wörtern (≈ 10)

$$\frac{Satz_{heute}}{Satz_{jahr}} \approx \frac{Token_{heute}}{Token_{jahr}} \quad (5)$$

- Zur Überprüfung später mehr

Z-Score

Grundidee: Messe die Abweichung der Auftretenshäufigkeit vom Erwartungswert in Vielfachen der Standardabweichung.

Definition (ZScore)

$$sig_{zscore}(w) = \frac{f_{rel}(w) - E(w)}{\sigma(w)}$$

$f_{rel}(w)$: relative Häufigkeit des Wortes w an diesem Tag

$E(w)$: Erwartungswert der relativen Häufigkeit von w

$\sigma(w)$: Standardabweichung der relativen Häufigkeit von w

Z-Score – Bracketing

Problem: Verdoppelung der Frequenz für Worte mit durchschnittlich kleiner relativer Frequenz einfacher \Rightarrow Hohe Z-Score Werte können leichter erreicht werden

Lösung:

1. Gruppierung der Worte anhand ihrer relativen Dokumentenhäufigkeit
2. Zuweisen eines individuellen Korrekturwerts für jede Gruppe:

[0-5[[5-10[[10-20[[20-30[[30-40[
-20	-25	-15	-12	-10
[40-50[[50-60[[60-70[[70-80[[80-100[
-10	-9	-8	-6	-5

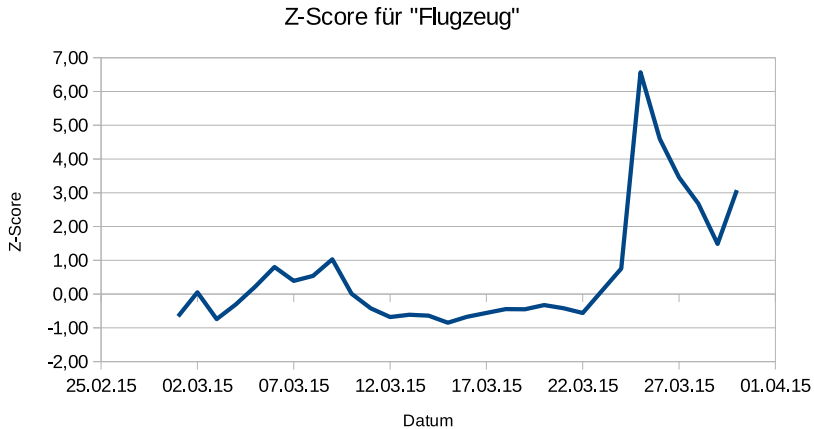
Z-Score – Zero-Frequency-Problem

Problem: Worte im ersten Bracket haben bestenfalls an 5% der Tage Frequenzen $> 0 \Rightarrow$ sehr geringer Erwartungswert und Standardabweichung

Lösungsmöglichkeiten:

- Laplace-Smoothing
- Ignoriere alle Worte mit relativer Frequenz $<$ Schwellwert t
 t : Durchschnittliche relative Frequenz aller Worte mit relativer Dokumentenfrequenz < 0.1

Z-Score Beispiel



Zeitreihenanalyse

Definition (Zeitreihenanalyse)

Unter einer Zeitreihe versteht man die Entwicklung einer bestimmten Größe, deren Werte im Zeitablauf zu bestimmten Zeitpunkten oder für bestimmte Zeitintervalle erfasst und dargestellt werden

Maß: gleitender Mittelwert

- Glättet Zeit oder Datenreihen
- Erfolgt durch glätten hoher Frequenzanteile
- Es gibt ein Raster der Größe n
- Es werden n Tage zusammenaddiert und dann durch n geteilt

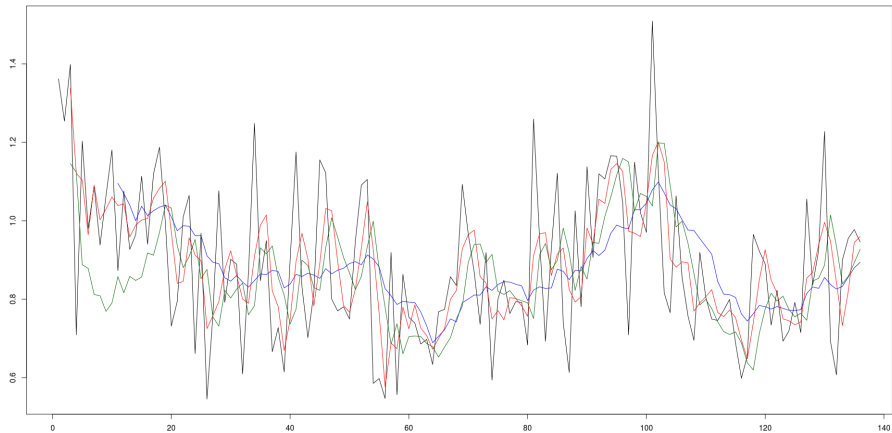
Wie hilft uns das weiter?

- Tritt ein Wort häufiger als sein Durchschnittswert an dem Tag auf kann das interessant sein.

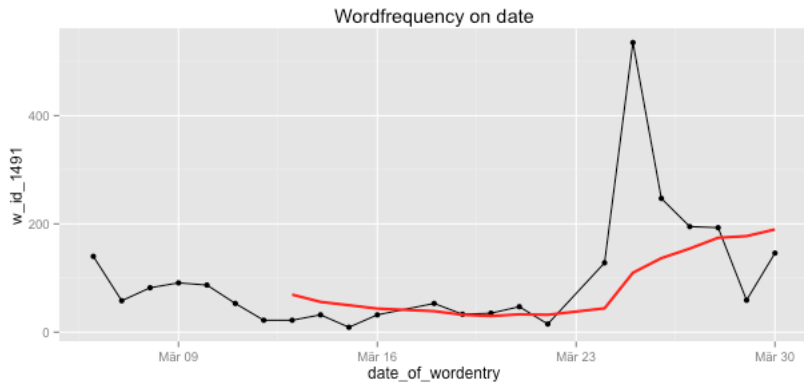
Erster Ansatz: R

- Der erste Ansatz war ein R Programm welches den gleitenden Mittelwert ausrechnen sollte
- Problem: R verarbeitet Wörter einzeln
- 3 Mio. Wörter \rightarrow 3 Mio. Transaktionen = MySQL Overkill
- Ausführungszeit würde mehrere Tage beanspruchen

Beispiel: Haus



Beispiel: Flugzeug



Zweiter Ansatz: MySQL

- Der Zweite Ansatz ist es direkt in MySQL zu berechnen
- Problem: Inner Join auf selbe Tabelle (ca. 20 Mio Zeilen)
- Jeder Eintrag muss geprüft werden ob die Join Tabelle den Eintrag in der Größe des Rasters hat
- Eine Datums Differenz Tabelle kann das ganze jedoch beschleunigen

Finaler Ansatz: R BigTable

- Diesmal reshape der Tabelle
- Spalten = Wörter, Zeilen = Datumfelder, Wert = freq
- Darüber kann man das effizient einzeln berechnen
- Danach überführung in alte Struktur und Speicherung

Vergleich und Auswertung

Qualitative vs. Quantitative Auswertung

- Schwierigkeit einer quantifizierbaren qualitativen Evaluierung
- Quantitative vergleiche möglich, aber keine Aussage über Qualität
- Im Rahmen des Projektes möglich:
 - **“Evaluierung durch draufschaun”**
 - Geeignetes Maß zum quantitativen Vergleich nutzen

Qualitative Auswertung (25.3.2015)

	poisson	tf_idf
1	Germanwings	Germanwings-Maschine
2	Absturz	Milke
3	Germanwings-Maschine	Germanwings-Airbus
4	A320	9525
5	Airbus	Germanwings-Flug
6	25. März	Germanwings
7	Haltern	Germanwings-Chef
8	Tsipras	Tsipras
9	Alpen	Barcelonnette
10	Südfrankreich	A320

Qualitative Auswertung (25.3.2015)

	fregratio	z_score
1	Barcelonnette	Haltern
2	Germanwings-Airbus	Aden
3	Germanwings-Chef	Südfrankreich
4	Rajana	Sinkflug
5	Germanwings-Maschine	Akte X
6	Dalkurd	A320
7	9525	Eierstöcke
8	Fire-TV-Stick	Hadi
9	ArtikelPolitik	10.53
10	18.03.2015	Ja Nein

Qualitative Auswertung (25.3.2015)

	poisson	tf_idf
31	Germanwings-Flug	Eierstöcke entfernen
32	Unglück	Kolomoiski
33	abgestürzt	Flugschreiber
34	Germanwings-Airbus	Bürokratiebremse
35	Jemen	Sollecito
36	Flugschreiber	Dalkurd
37	2015	Haltern am See
38	4U	Akte X
39	KAC	Hadis
40	S6	Bloodborne

Qualitative Auswertung (25.3.2015)

	freqratio	z_score
31	Eierstöcke entfernen	4Players.de
32	Germanwings-Flug	57,5
33	Germanwings-Flugzeug	Bassbariton
34	22.03.15	Alkoholiker
35	Feuerwehr-Leutnant	Debra
36	Gehenna	hervorragendem
37	Grabetz	XF
38	Höchstbefristungsdauer	25. März
39	Luciano Moggi	Angehörigen
40	Schultreppe	Crews

Quantitative Auswertung

Problemstellung: Vergleich von sortierten Listen mit potentiell unterschiedlichem Inhalt.

- Der Vergleich von Wortpaaren nicht sauber möglich.
- Schwierigkeit eines Mengenbasierter Ansatzes:
Reihenfolge wird nicht beachtet

Quantitative Auswertung: Maximum Overlap

Idee: Es wird ein Mengenbasierter Ansatz für Teillisten genutzt und dann gemittelt.

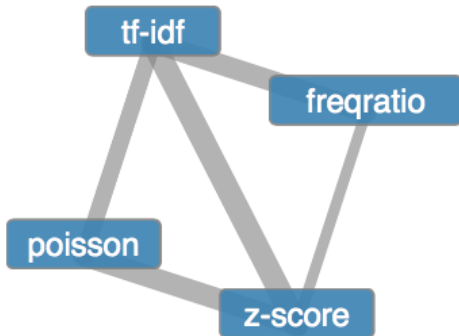
Für jeden Rang der Listen wird eine Teilliste (Rang 1 bis betrachteter Rang) verglichen.

Beispiel: Tafelbild

Quantitative Auswertung: Maximum Overlap Ergebnisse

	List	List_to_compare	average_overlap
1	tf-idf	z-score	0.41
2	tf-idf	poisson	0.21
3	tf-idf	fregratio	0.33
4	z-score	poisson	0.35
5	z-score	fregratio	0.09
6	poisson	fregratio	0.00

Quantitative Auswertung: Maximum Overlap Ergebnisse



Einschub II: Wortzahl vs. Satzzahl

	List	List_to_compare	average_overlap
3	poisson	poisson_old	0.99961
4	fregratio	fregratio_old	1.00000


Zusammenfassung

- Es wurden bestehende Verfahren untersucht
- Es wurden weitere Verfahren ausprobiert
- Es wurden die Ergebnisse quantitativ und qualitativ verglichen
- Es wurden MySQL und R Implementierungen umgesetzt.
- Es werden noch Musterbasierte Verfahren zum Cleaning der Listen implementiert
- Es wird noch ein weiteres Vergleichsmaß mit Berücksichtigung der Anzahl der Quellen in denen ein Token erwähnt wird untersucht.



Quellen (1)

-  AGGARWAL, Charu C.:
Mining text streams.
In: *Mining Text Data*.
Springer, 2012, S. 297–321
-  AGGARWAL, Charu C.:
Mining text and social streams: a review.
In: *ACM SIGKDD Explorations Newsletter* 15 (2014), Nr. 2, S.
9–19
-  BENATTAR, Gary ; TRÉBUCHET, Philippe u. a.:
Trend Analysis in Polls, Topics, Opinions and Answers.
(2011)



Quellen (2)

-  BENHARDUS, James ; KALITA, Jugal:
Streaming trend detection in twitter.
In: International Journal of Web Based Communities 9 (2013),
Nr. 1, S. 122–139
-  CHEN, Chaomei:
CiteSpace II: Detecting and visualizing emerging trends and
transient patterns in scientific literature.
*In: Journal of the American Society for information Science and
Technology* 57 (2006), Nr. 3, S. 359–377

Quellen (3)

-  GAO, Yan ; LIU, Jin ; MA, PeiXun:
The hot keyphrase extraction based on tf* pdf.
In: Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on IEEE, 2011, S. 1524–1528
-  GUPTA, Manish ; GAO, Jing ; AGGARWAL, Charu ; HAN, Jiawei:
Outlier detection for temporal data.
In: Synthesis Lectures on Data Mining and Knowledge Discovery 5 (2014), Nr. 1, S. 1–129

Quellen (4)

-  HEYER, Gerhard ; QUASTHOFF, Uwe ; WITTIG, Thomas:
Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse.
W3L, 2006
-  JAHNAVI, Y ; RADHIKA, Y:
Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents.
In: Advanced Computing Technologies (ICACT), 2013 15th International Conference on IEEE, 2013, S. 1–6

Quellen (5)



KONTOSTATHIS, April ; GALITSKY, Leon M. ; POTTENGER, William M. ; ROY, Soma ; PHELPS, Daniel J.:

A survey of emerging trend detection in textual data mining.

In: *Survey of Text Mining*.

Springer, 2004, S. 185–224




VIERMETZ, Maximilian ; SKUBACZ, Michal ; ZIEGLER, Cai-Nicolas ; SEIPEL, Dietmar:

Tracking topic evolution in news environments.

In: *E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services, 2008 10th IEEE Conference on IEEE*, 2008, S. 215–220

Quellen (6)

-  WEBBER, William ; MOFFAT, Alistair ; ZOBEL, Justin:
A similarity measure for indefinite rankings.
In: *ACM Transactions on Information Systems (TOIS)* 28 (2010),
Nr. 4, S. 20