Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik
Abteilung Automatische Sprachverarbeitung

Wortschatz Zeitgeist

Seminararbeit

Autoren: TBA
Kießling, Max
Otto, Wolfgang (2885214)
Modul: Anwendungen Linguistische Informatik (10-202-2307)
Abgabe: 19. Mai 2015. (Sommersemester 2015)
Betreuer: Maciej Janicki
Seminarleiter: Prof. Dr. Uwe Quasthoff

Inhaltsverzeichnis

1 Einleitung

1.1 Aufgabenstellung

Es soll untersucht werden, welche Maße sinnvoll für die Generierung der Wörter des Tages sind. Eine Implementierung soll in SQL und für Zeitreihenuntersuchungen in R erfolgen.
Desweiteren soll ein regelbasiertes Verfahren implementiert werden um Datumsangaben und andere strukturelle Worte zu filtern.

1.2 Status quo

1.3 Vergleichbare Ansätze

Tagesaktuelle Wikiartikel
google trends?

2 Das finden von Tagesaktuellen Wörter

2.1 Maße zur Trend-Detection

2.1.1 Relatives Vorkommen (Referenz)

Das Maß wird durch die Differenz des relativen Anteils eines Tokens an allen Tokens eines Tages und dem relativen Anteil eines Tokens an allen Tokens eines Vergleichskorpus bestimmt.

$$sig_{referenz}(w) = p_{tag} - p_{jahr} \tag{2.1}$$

2.1.2 Poisson-Maß

Die Formel leitet sich aus der Poissonverteilung ab und beschreibt wie Wahrscheinlich es ist, dass die gemessene Tagesfrequenz beobachtet werden kann.

$$sig_{poisson}(w) = \frac{k(\log(k) - \log(n \cdot p) - 1))}{\log(n)} \tag{2.2}$$

k:= Anzahl der Token von w in Tagesbericht
n := Anzahl der Tokens in Tagesbericht
p := relativer Anteil eines Tokens am Jahreskorpus
Es ist das gleiche Maß wie in [8, S. 338-340] beschrieben und hergeleitet. Hier aber nicht zum auffinden von signifikanten Kookurenzen, sondern zum auffinden von signifikanten Nennungen im Tageskorpus gegenüber einem Vergleichskorpus.

2.1.3 Termfrequenz inverse Dokumentenfrequenz (tf-idf)

$$sig_{tfidf}(w) = \frac{k}{\max(K)} \cdot \log(\frac{365}{|documentdays(w)|}) \tag{2.3}$$

2.1.4 Z-Score

Benattar et al. beschreiben in [3] einen Ansatz zur Trend-Erkennung basierend auf dem Z-Score. Dabei beziehen Sie neben der relativen Worthäufigkeit noch die Anzahl der Tage ein an denen ein Wort mindestens einmal auftritt, um so das 0-Frequenz-Problem zu umgehen.

Berechnung

- Wortfrequenz
  $f(w)_d :=$ Anzahl der Vorkommen von Wort w an Datum d
- relative Worthäufigkeit
  Die relative Worthäufigkeit $p(w)_d$ berechnet sich durch:
  $t_d :=$ Anzahl verschiedener Worte an Datum d

$$p(w)_d = \frac{f(w)_d}{t_d}$$

- Erwartungswert
  Der Erwartungswert $\bar{w}$ berechnet sich durch:
  $N :=$ Anzahl der Tage in der betrachteten Zeitspanne

$$\bar{w} = \frac{1}{N} \sum p(w)_d$$

- Standartabweichung
  Die Standartabweichung σ_w berechnet sich durch:

$$\sigma_w = \sqrt{\frac{1}{N} \sum (p(w)_d - \bar{w}^2}$$

- ZScore
  Der Zscore $Z(w)_d$ misst die Abweichung der relativen Worthäufigkeit vom Erwartungswert in Vielfachen der Standartabweichung.

$$Z(w)_d = \frac{p(w)_d - \bar{w}}{\sigma_w}$$

- Auftrittshäufigkeit
  Die Auftrittshäufigkeit $Po(w)$ gibt an wie vielen Tagen innerhalb des betrachteten Zeitraums das Wort mindestens einmal auftritt:
  $nbD(w) :=$ Anzahl der Tage an denen w vorkommt// $c_d :=$ Anzahl der Tage innerhalb des betrachteten Zeitraums

$$Po(w) = \frac{nbD(w}{c_d}$$

- Schwellwerte Zur besseren Unterscheidung echter Trends von statistischen Anomalien schlagen Benattar et. al. vor die Worte anhand ihrer Auftrittshäufigkeit zu clustern. Den Clustern werden dabei Z-Score-Schwellwerte zugeordnet. Überschreitet der Z-Score eines Wortes den Schwellwert seines Clusters wird dieses Wort als signifikant und somit als Trend eingestuft. Cluster mit niedriger Auftrittshäufigkeit erhalten dabei höhere Schwellwerte. Je häufiger ein Wort auftritt desto niedriger wird der Schwellwert.

Vorgehen

1. Für jedes Wort in w in d berechne $Z(w)$
2.

2.1.5 Weitere Maße

Einbeziehung der Anzahl von Quellen.

2.1.6 Bewertung der Maße

Kriterien: Anteil niederfrequenter Wörter in der Top-Liste

2.2 Zeitreihenanalysen

2.3 Cleaning

Es sollen Datumsangaben und evtl. neu auftauchende strukturelle Angaben ausgefiltert werden.
Ansatz: Regelbasiert.
Gibt es Maße, die solche Angaben strukturell ausschließen?

3 Implementierungen in SQL und R

4 Ein empirischer Vergleich

5 Bewertung und Zusammenfassung

Literaturverzeichnis

[1] Aggarwal, Charu C.: Mining text streams. In: *Mining Text Data*. Springer, 2012, S. 297–321

[2] Aggarwal, Charu C.: Mining text and social streams: a review. In: *ACM SIGKDD Explorations Newsletter* 15 (2014), Nr. 2, S. 9–19

[3] Benattar, Gary ; Trébuchet, Philippe u. a.: Trend Analysis in Polls, Topics, Opinions and Answers. (2011)

[4] Benhardus, James ; Kalita, Jugal: Streaming trend detection in twitter. In: *International Journal of Web Based Communities* 9 (2013), Nr. 1, S. 122–139

[5] 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

[6] 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

[7] 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

[8] Heyer, Gerhard ; Quasthoff, Uwe ; Wittig, Thomas: *Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse*. W3L, 2006

[9] 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

[10] 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

[11] 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