UNIVERSITÄT KOBLENZ · LANDAU

Topic Time

Inhalt

- Einleitung
- Definitionen
- Vorgehensweise
- Konzepte
- Forschungsergebnisse
- Fazit

Einleitung

„Erkläre die Vergangenheit, erkenne die Gegenwart, sage die Zukunft voraus.“

Hippokrates (460-377 v.Chr.)

- Gegenwart erkennen 
- Vergangenheit erklären ?
- Zukunft voraussagen 

Definitionen

- Wort *w*: Basiseinheit
- Dokument *d*: Sequenz aus *N* Wörtern
- Korpus *c*: Sammlung von *M* Dokumenten
- Topic *z*: Themengebiet
  - $w \rightarrow z$
  - $z \rightarrow d$
  - Multinominale Verteilungen

Vorgehensweise

Vorgehensweise
-Datensammlung-

- Große Datenmenge
- Google News Archive
  - 32 Ländernamen
  - 61 unterschiedliche Suchanfragen
    - Landernamen, Firmennamen, Personen, anderes
- Artikel von 1990 – 2010
- Speichern der Webseite / Kurzfassung

Vorgehensweise -Pre-Processing-

- Löschen von:
  - HTML-Tags
  - Java-Script Code
  - Kopf- und Fußzeilen
  - andere nichtinhaltliche Elemente
- Extraktion des größten Textabschnitts
- nur englischsprachige Artikel
  - n-Gramm Matching

Vorgehensweise
-Extraktion temporaler Ausdrücke-

- Zeitpunkte in Vergangenheit oder Zukunft
- GUTime Tagger
  - absolut („09. Mai 2012“, „Oktober 1986“)
  - relativ („vor 10 Jahren“, „in 5 Monaten“)
    - benötigt Anker
- TimeML Markup Language

Vorgehensweise
-Extraktion temporaler Ausdrücke-

- Granularität:
  - Jahr
  - Jahres-, Monats-, Tagesangaben → ein Wert
  - Jahre, Monate, Tage (kleinste Einheit der Referenz)
  - Zeitpunkte und Zeitintervalle
  - „nach“, „zwischen“, „innerhalb“, „ab“, „um“
  - „zu Beginn“, „Mitte“, „am Ende“

Konzepte

Konzept zur Analyse vergangenheitsbezogener Informationen

- Latent Dirichlet Allocation
  - „kompliziertes“ Verfahren
- Verknüpfung mit zeitlicher Verteilung
  - einfaches Zählen

Latent Dirichlet Allocation

- Wähle die Anzahl *N* von Wörtern, die das Dokument *d* enthält
- Wähle eine Menge von Topics aus einer Auswahl von *K* Topics
- Generiere jedes Wort *w* in dem Dokument:
  - Wähle ein Topic *z* gemäß der multinominalen Verteilung
  - Wähle ein spezifisches Wort aus der multinominalen Verteilung

Collapsed Gibbs Sampling

- Ordne jedem Wort *w* in jedem Dokument *d* eines der *K* Topics zu
- Nehme jedes Dokument *d*
  - Nehme jedes Wort *w* aus *d*
    - Berechne $P(z|d)$
    - Berechne $P(w|z)$
    - Berechne $P(z|d) \cdot P(w|z)$
    - Ordne dem Wort *w* ein neues Topic *z* gemäß der berechneten Verteilung zu.
- Wiederhole diesen Schritt „genügend oft“
- Annähernd stabiler Zustand

Konzept zur Analyse vergangenheitsbezogener Informationen

- Topicverteilung eines Jahres

$$P(z|y) = \frac{1}{|D_y|} \sum_{d \in D_y} P(z|d)$$

- Nennung eines Topics

$$P(p|y,z) = \frac{P(p,y,z)}{P(y,z)} = \frac{P(z,|p,y)P(p,y)}{P(z|y)P(y)}$$

Konzept zur Analyse zukunftsbezogener Informationen

- Zu LDA ähnliches Modell
  - Gruppierung über Topics
  - und Zeitverteilungen
- Referenzen nicht exakt
  - Wahrscheinlichkeitsverteilung entlang der Zeit
  - Zeitpunkt → Gaußsche Normalverteilung
  - Enddatum → wachsende Exponentialverteilung
  - Anfangsdatum →abnehmende Exponentialverteilung
  - Periode → Gleichverteilung

Grundmodell

- Dokument *d* wird generiert durch

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

- Ähnlich zu LDA

Einbezug temporaler Ähnlichkeit

- $G_d(t)$ Wahrscheinlichkeitsfunktion eines Dokuments *d*
- $G_z(t)$ Wahrscheinlichkeitsfunktion eines Topics *z*

$$H(d|z) = \frac{h(d|z)}{\sum_z h(d|z)} \qquad h(d|z) = \frac{1}{D_{KL}(G_d||G_z) + 1}$$

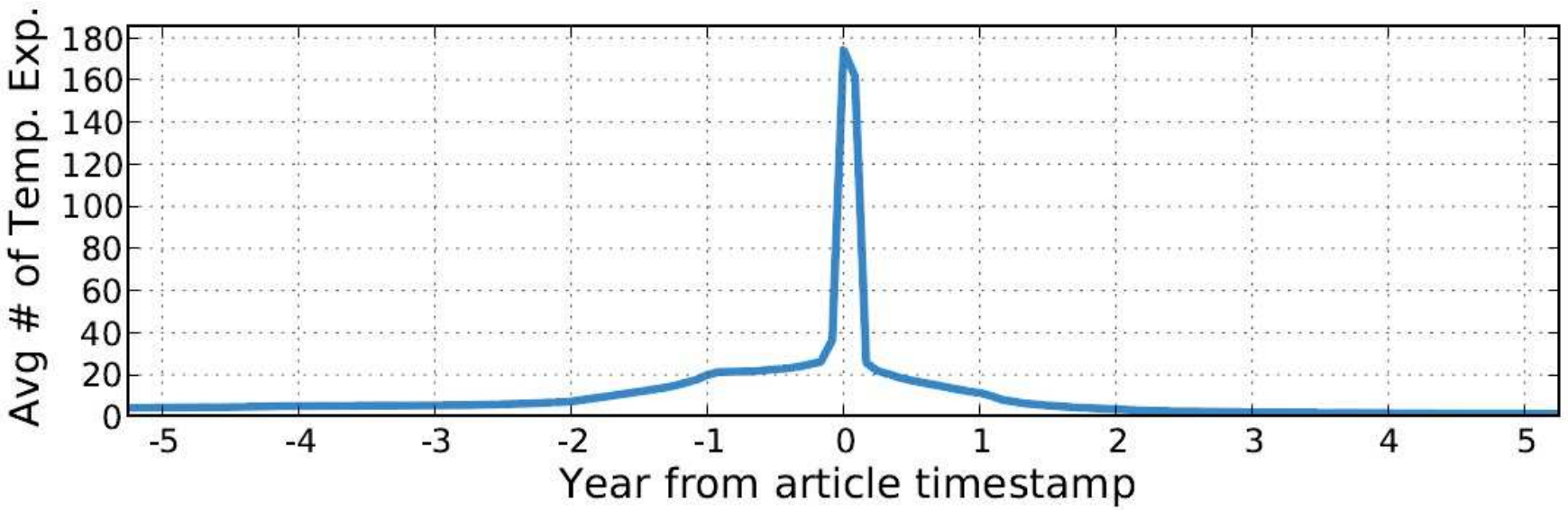
- $D_{KL}(G_d||G_z)$ Kullback-Leibler Divergenz

Gesamtmodell

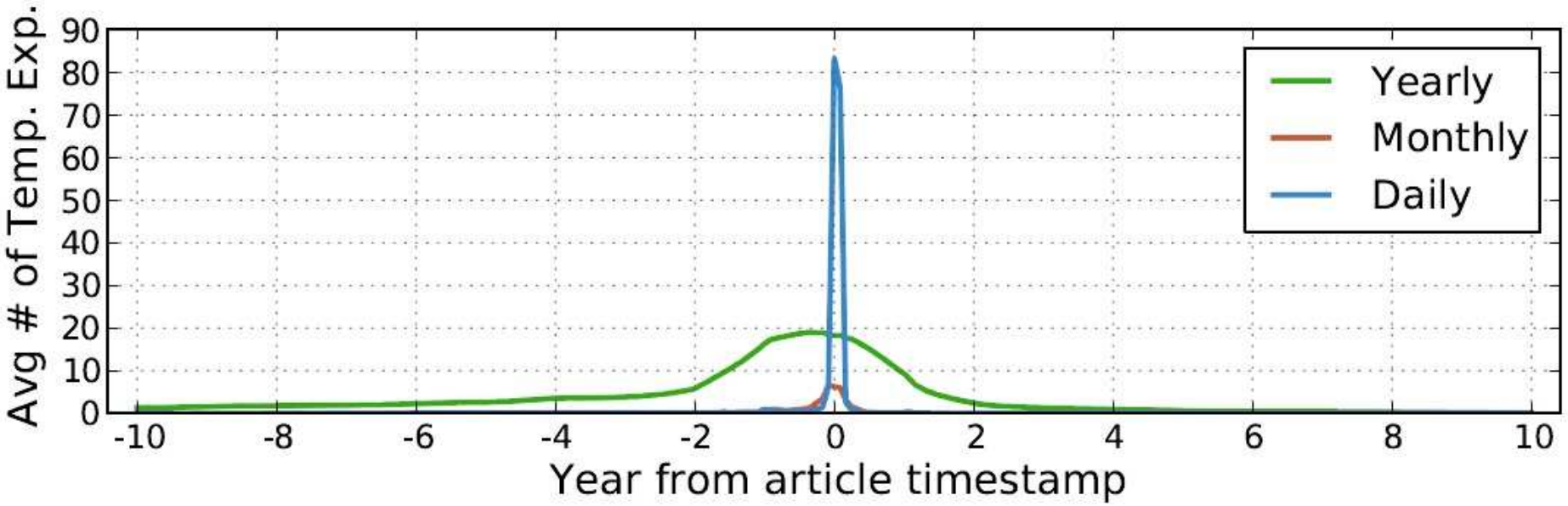
$$P(d) = \sum_{z \in Z} P(z)(\prod_{w \in W_d} P(w|z)^{N_{w,d}} \times H(d|z)^{\alpha})$$

Forschungsergebnisse

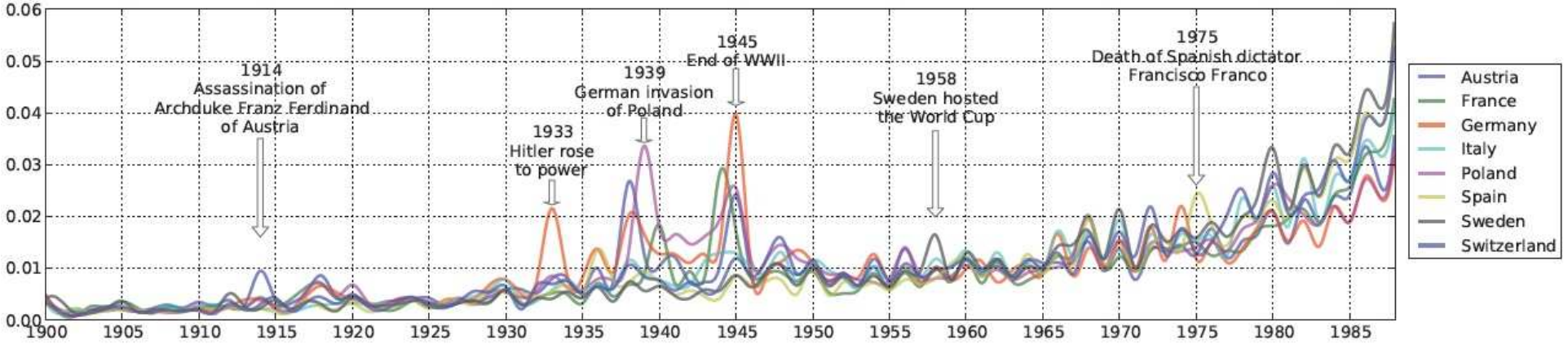
Allgemeine Feststellungen



Allgemeine Feststellungen



Analyse von Verweisen in die Vergangenheit -Verteilung der Referenzen-



Analyse von Verweisen in die Vergangenheit -Signifikante Jahre und Topics-

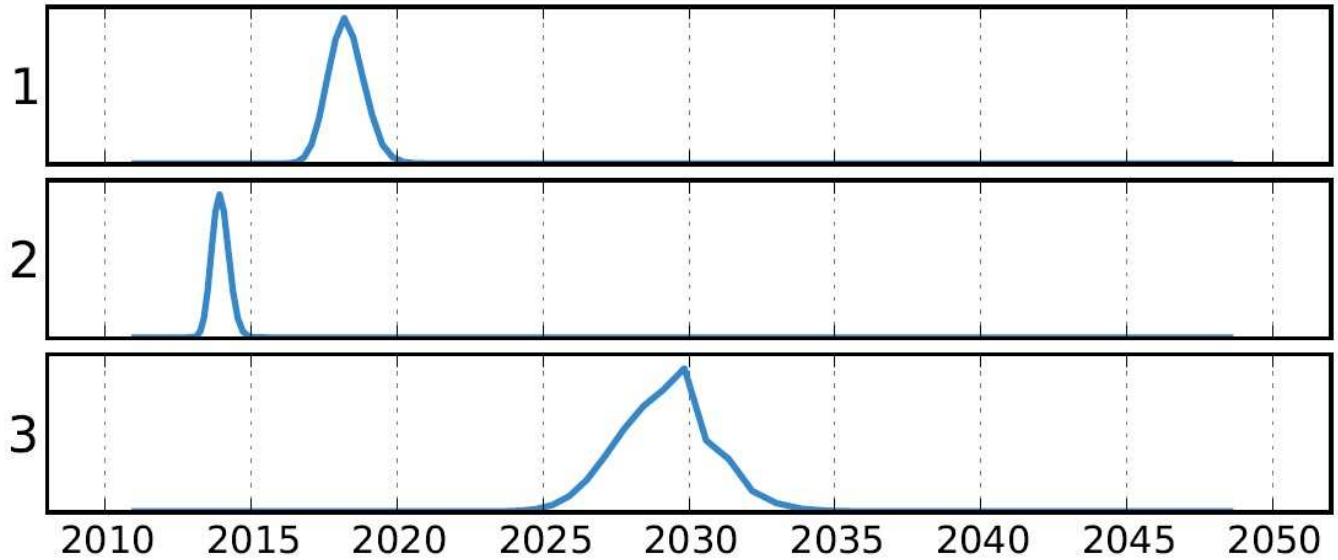
Land	Jahr	Schlagwörter
Deutschland	1945	war, world, end, day, second, declared, allies, europe, first, empire
	1939	soviet, poland, union, europe, war, eastern, western, czechoslovakia, polish, invaded
	1974	world, cup, final, england, team, first, won, second, win, football
Frankreich	1944	war, world, army, american, battle, soldiers, legion, served, veterans, french
	1940	war, french, german, germany, hitler, occupation, world, resistance, nazi, britain
	1968	killed, people, group, spain, attack, eta, basque, region, year, police
Polen	1939	war, hitler, germany, invasion, britain, invaded, france, german, september, world
	1945	camp, concentration, auschwitz, camps, nazi, death, nazis, sent, january, prisoners
	1980	communist, solidarity, walesa, gdansk, workers, movement, union, leader, government, lech

Analyse von Verweisen in die Vergangenheit -Auslöser-

- Deutschland, Polen
  - 1945: Ende des zweiten Weltkriegs
  - 1995, 2005: 50. und 60. Jahrestag
- Japan:
  - 1972: Olympische Winterspiele
  - 1998: Olympische Winterspiele

Analyse von Verweisen in die Zukunft
-Fallstudie: NASA-

ID	Anzahl	Häufige Ausdrücke
1	276	moon, space, astronauts, return, mars, agency, president, lunar, program, new
2	139	space, launch, rst, mission, shuttle, agency, flight, spacecraft, orion
3	82	earth, asteroid, space, apophis, mars, chance, hit, mission, propulsion, scientists



Fazit

- Gute Ansätze
- Abschätzung des Aufbaus des Textkorpus möglich
  - Wahrscheinlichkeiten
- Plausible Forschungsergbnisse
  - Hintergrundwissen notwendig

- Nur Zeitungsartikel
- Nur englischsprachig

  Soll behoben werden

- Nutzen?

Quellen

- Ching-man Au Yeung and Adam Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM ’11, pages 1231–1240, New York, NY, USA, 2011. ACM.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- Edwin Chen. Introduction to latent dirichlet allocation. Webseite, 02. Mai 2012. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/.
- William B. Cavnar and John M. Trenkle. N-gram-basestext categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.
- Adam Jatowt and Ching-man Au Yeung. Extracting collective expectations about the future from large text collections. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM ’11, pages 1259–1264, New York, NY, USA, 2011. ACM.
- Todd K. Moon. The expectation-maximization algorithm. In *IEEE Signal Processing Magazine*, pages 47–60, 1996.
- Inderjeet Mani and George Wilson. Robust temporal processing of news. In *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL’2000)*, pages 69–76, 2000.
- Jonathan Shlens. Notes on kullback-leibler divergence and likelihood theory. Webseite, 20. August 2007. http://www.snl.salk.edu/~shlens/.

Vielen Dank für die Aufmerksamkeit!