Seminar

Textdatenanalyse

Carsten Jentsch & Jonas Rieger

Sommersemester 2019

technische universität
dortmund

fakultät
statistik

Was versteht man unter Textdatenanalyse?

Text Mining (Wikipedia): „Text Mining, seltener auch Textmining, Text Data Mining oder Textual Data Mining, ist ein Bündel von Algorithmus-basierten Analyseverfahren zur Entdeckung von Bedeutungsstrukturen aus un- oder schwachstrukturierten Textdaten.“

Besonderheiten:

- praktisch unstrukturierte Daten (im statistischen Sinn),
- zudem (meist) unsaubere Daten,
- Interesse der Beibehaltung bzw. Abstraktion von (Satz-)Strukturen,
- hochdimensionale Daten (Dokumente x Vokabeln).

Anwendungsgebiete:

- Vergleich der Berichterstattung verschiedener Medien,
- Analyse journalistischer Kanäle (twitter, facebook, WhatsApp, ...),
- Spam-Filter, Suchmaschinen, Nutzer-spezifische Werbung, ...

Mögliche Themengebiete

Vorverarbeitung:

- Case Sensitivity, Tokenization, Stopwords, Stemming, Lemmatization,
- Bag-of-Words, N-Gram, tf-idf, ...

Klassifikation & Clustering:

- (un)überwachte Kategorisierung, Recommendation Tools, ...

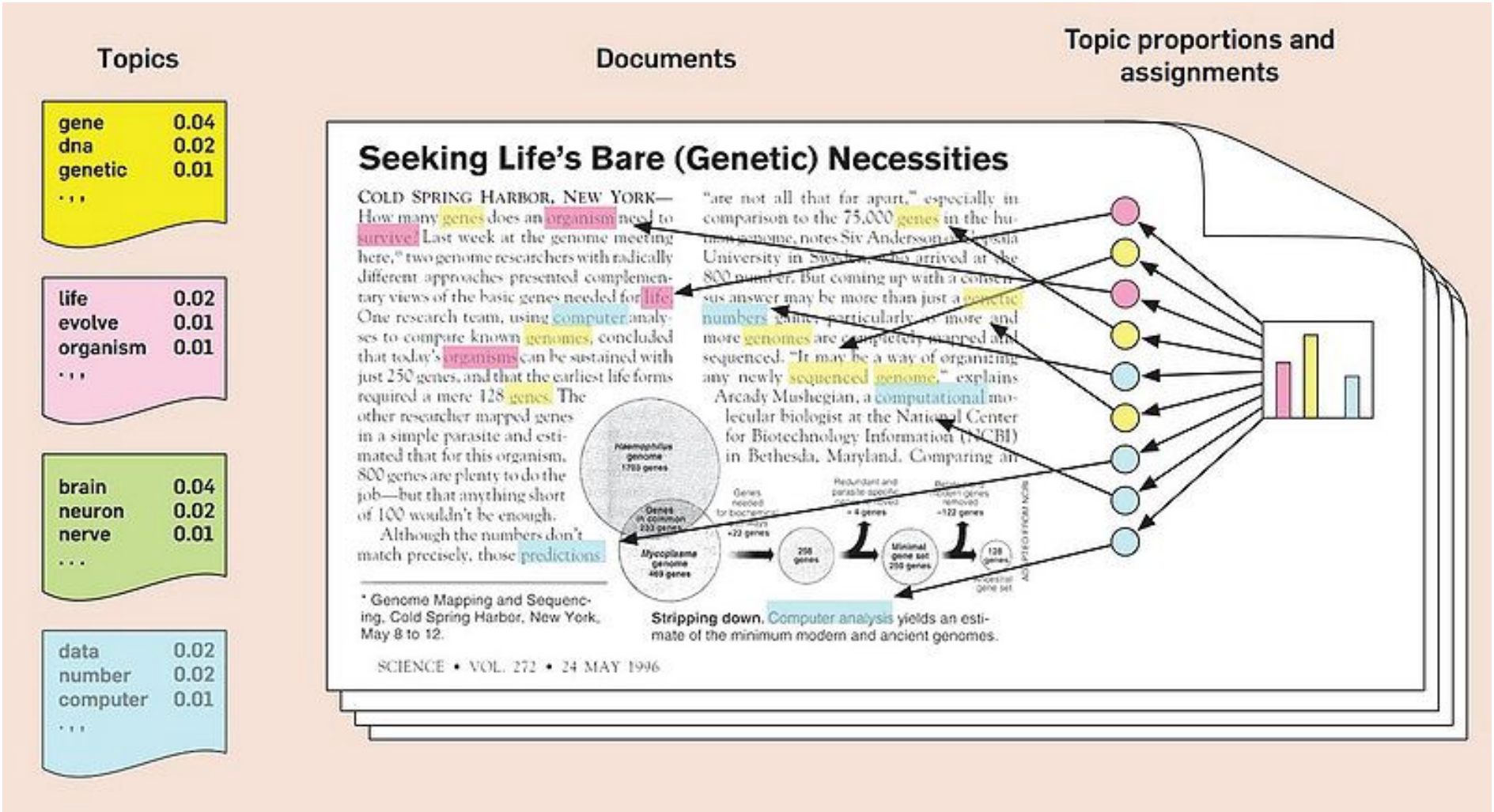
Topic-Modelle:

- Latent Semantic Analysis/Indexing (LSA/LSI),
- Probabilistic Latent Semantic Analysis/Indexing (pLSA/pLSI),
- Latent Dirichlet Allocation (LDA),
- Correlated Topics Model (CTM), ...

Weitere Themen:

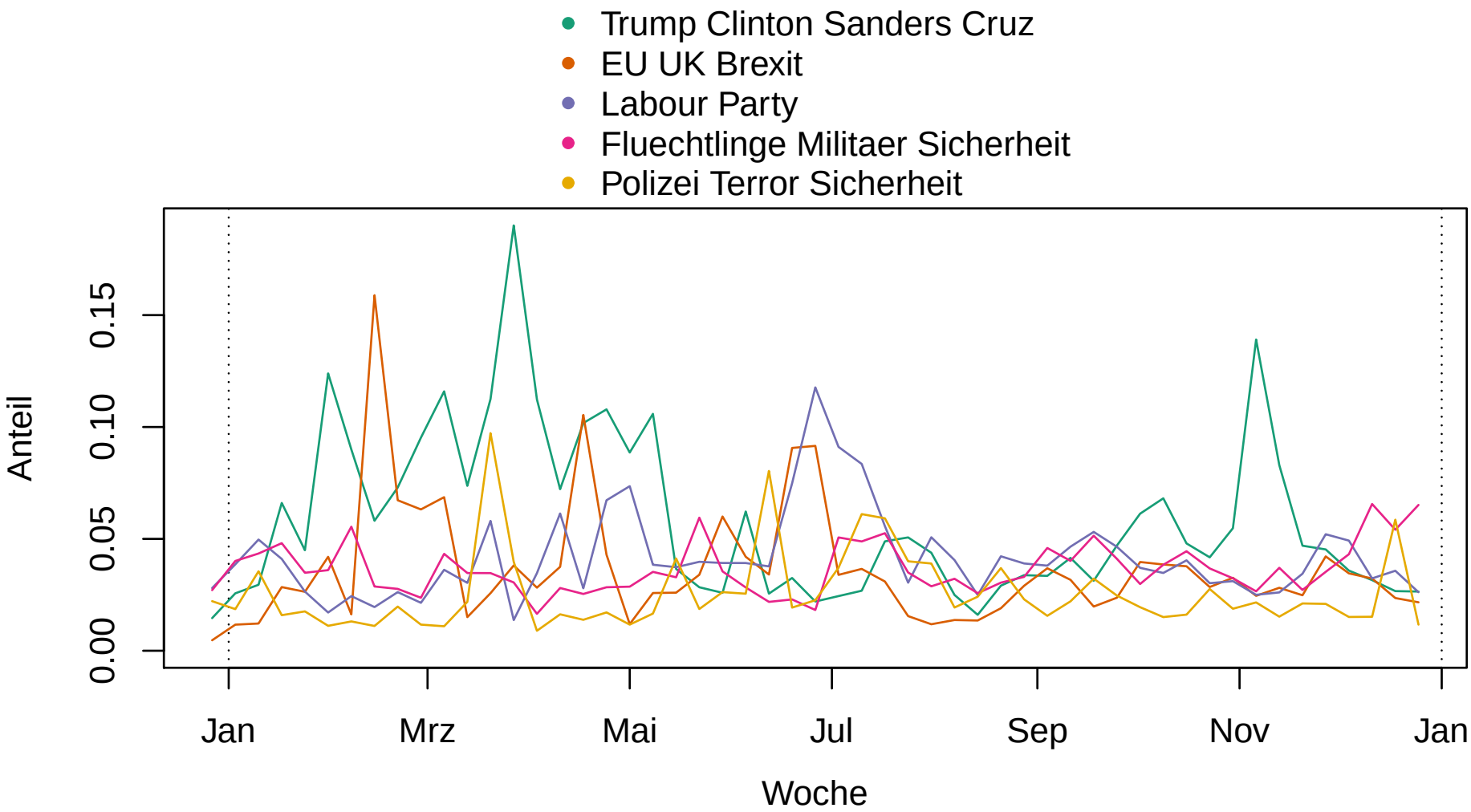
- word2vec, Neural Topic Model (NTM), Topic Intrusion, LDAvis, t-SNE, Wordscore, Wordfish, ...

Beispiel: LDA



Quelle: https://medium.com/@connectwithghosh/topic-modelling-with-latent-dirichlet-allocation-lda-in-pyspark-2cb3ebd5678e

Beispiel: LDA „The Guardian“ 2016



Literatur

- Aggarwal & Zhai (2012) Mining Text Data
- Allahyari et al. (2017) A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques
- Berry & Kogan (2010) Text Mining: Applications and Theory
- Miner et al. (2012) Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications
- Silge & Robinson (2017) Text Mining with R

Mitarbeit im Seminar

Bachelor:

- 30-minütiger Vortrag,
- Handout/Seminarpapier,
- Aktive Teilnahme an Diskussion, Feedback.

Master:

- 45-minütiger Vortrag,
- Handout/Seminarpapier,
- Aktive Teilnahme an Diskussion, Feedback.

Desweiteren:

- Individueller Termin, um über das Handout und die Folien zu sprechen: spätestens eine Woche vor dem Vortragstag.
- Handout und Folien spätestens zwei Tage vor unserem Treffen.

To-Do-Liste

- Einen Überblick über das Themenfeld „Text Mining“ anhand der aufgeführten Literatur verschaffen.
- Favorisierte Themen bestimmen
  (jeder erstellt seine persönliche Rangliste mit seiner Top 5).
- Sollte jemand an einem bestimmten, dem Text Mining zugehörigen Thema, interessiert sein, das nicht aufgeführt ist, kontaktiert mich.

Zeitplan

Termin zur Vorbesprechung:

- **Erste** Semesterwoche (Zeit und Ort wird noch bekannt gegeben, ggf. in Absprache mit Teilnehmern).

Termin zur Themenvergabe:

- **Zweite** Semesterwoche (Zeit und Ort wird noch bekannt gegeben, ggf. in Absprache mit Teilnehmern).

Seminarvorträge:

- Blockseminar **am Ende** des Semesters (Terminfindung in Absprache mit Teilnehmern).

Anmeldung

Unverbindliche Anmeldung für das Seminar per E-Mail an

rieger@statistik.tu-dortmund.de

bis

25.03.2019.