Der Event Crawl als Ansatz für den Aufbau von Webarchiven

Markus Eckl, Simon Donig
Sebastian Gassner, Florence Reiter
Daniel Göler, Malte Rehbein

Universität Passau

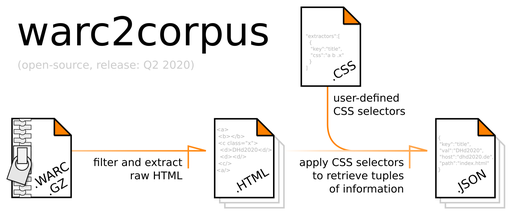
Webarchive als historische Quelle

- Webarchive sind Kopien des online Web, geprägt durch Fragmentierung (räumlich und zeitlich inkonsistent), Multimodalität und Hypermedialität.
- Analyseverfahren müssen diesen Herausforderungen Rechnung tragen.

Der Event Crawl

- Gecrawlt werden Webseiten bezüglich ihrer Relevanz auf ein Ereignis, bspw. Naturkatastrophen, Sportereignisse, politische Wahlen.
- Event Crawls bilden diachrone Dynamiken über das Archiv hinweg ab.

Erschließung eines Webarchives



Entwicklung des Werkzeuges **warc2corpus** um wohlgeformte Informationen aus einem Webarchiv zu extrahieren.

Webarchive sind massive Datenbestände, die granularer Datenextraktions- und Auswertungsverfahren bedürfen.

warc2corpus erzeugt ein Korpus durch Extraktion wohlgeformter Informationen aus Webarchiven.

Multivariate Textminingverfahren ermöglichen die Indizierung und subsequente Erschließung der Derivate für geistes- und sozialwissenschaftliche Fragestellungen.

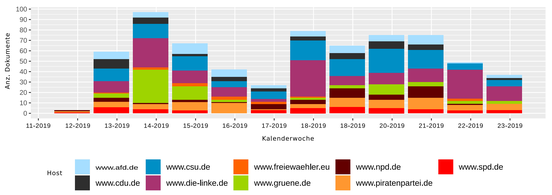
…zur Projektwebseite.

Analyse des Europawahlkampfs 2019

Gecrawlt wurden Webpages von

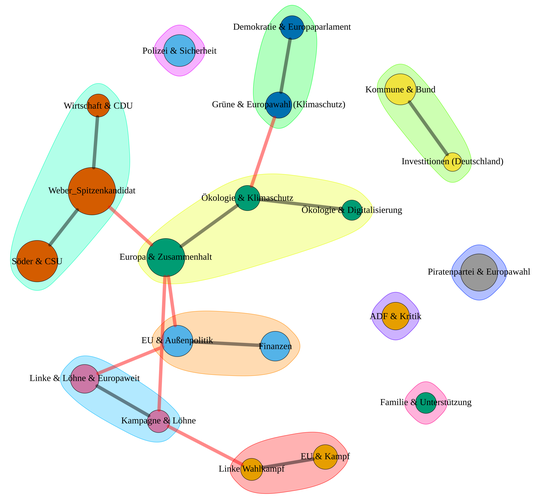
- 14 deutsche Parteien.
- 8 europäische Parteien & Fraktionen.
- 12 Spitzenkandidaten.
- 7 Newsmedien.

Publikationen der Parteien im Wahlkampf



Identifikation von Wahlkampfthemen der Parteien

- **Forschungsfrage:** Inwieweit kann die Europawahl 2019 als eine “second order election” verstanden werden?
- **Methode:** Latent Dirichlet Allocation & Structural Topic Modeling (STM).
- **Erste Ergebnisse:** Korrelationsnetzwerk der Topics & Modularity Clustering:



UNIVERSITÄT PASSAU
BSB Bayerische StaatsBibliothek Information in erster Linie
Gefördert durch DFG Deutsche Forschungsgemeinschaft