UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Exploring Significant Interactions in Live News

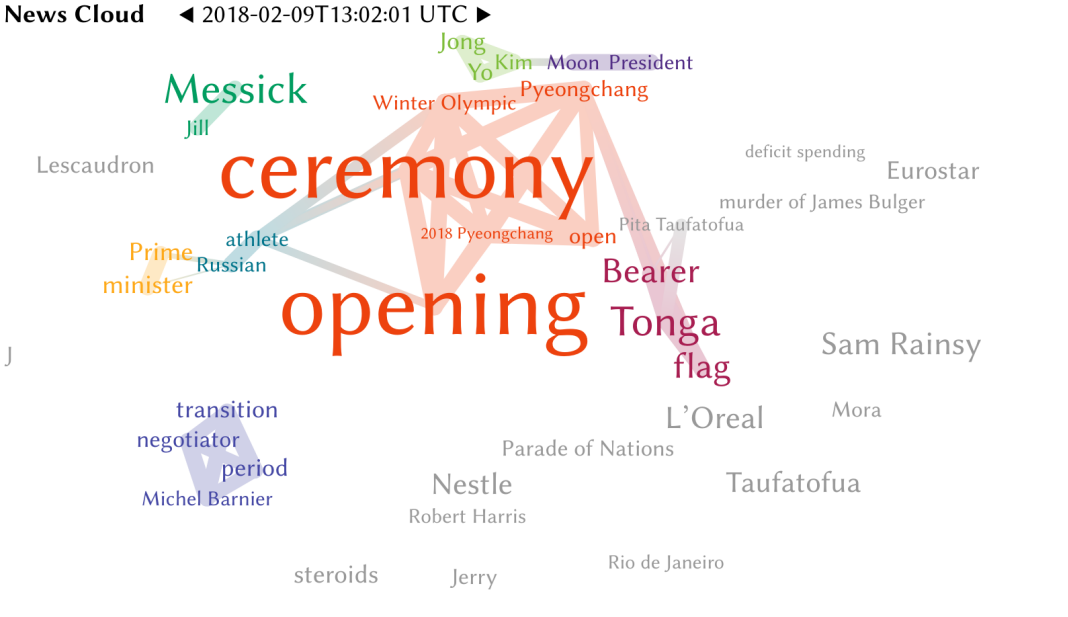
Erich Schubert, **Andreas Spitz**, Michael Gertz

March 26, 2018 — NewsIR'18 Workshop at ECIR 2018

Heidelberg University, Germany
Database Systems Research Group

What is in the news right now?

Example: Olympic Games Opening Ceremony



Core Idea: Cooccurrences

Focussing on the participating entities

- politicians, countries, companies, and celebrities are always in the news
- what *changes* is how they *interact*

See also: A. Spitz and M. Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *ACM SIGIR*. 2016

Core Idea: Cooccurrences

Focussing on the participating entities

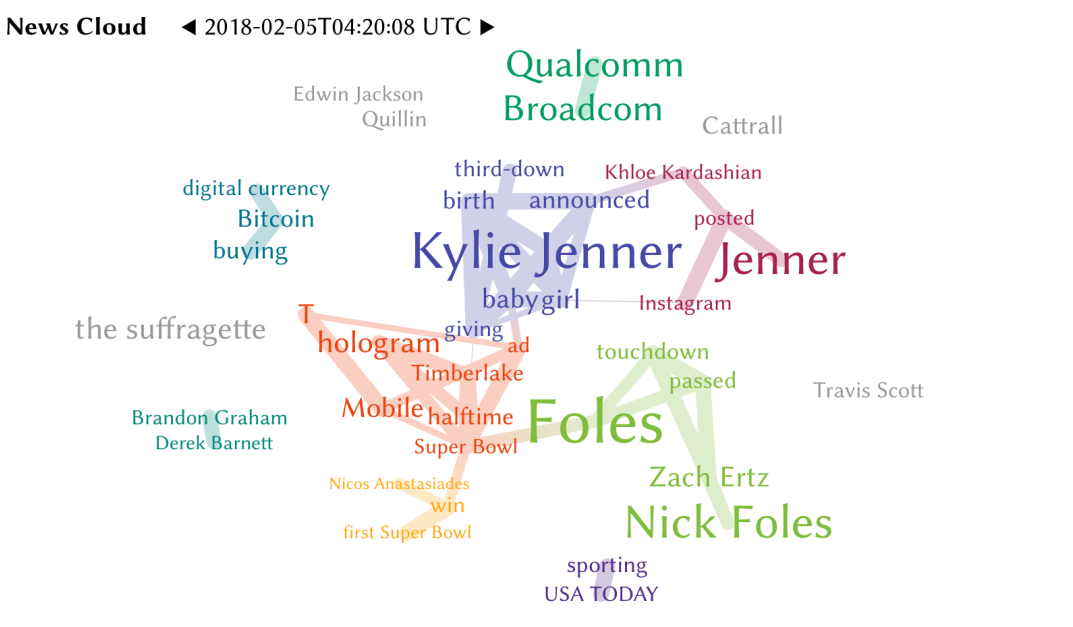
- politicians, countries, companies, and celebrities are always in the news
- what *changes* is how they *interact*

Capturing interactions

- it is not sufficient to look at one thing at a time
- instead, look at the **cooccurrences** of terms and entities

See also: A. Spitz and M. Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *ACM SIGIR*. 2016

Example: Superbowl



Core Ideas: Significance

Counting is not enough:

- many methods use word counts
- certain words are always frequent, others always rare
- it is interesting if a *rare* term or entity suddenly becomes *frequent*

Significance: compare frequency to *expected* frequency!

Details on our significance measure are in the arXiv predecessor:
E. Schubert, A. Spitz, M. Weiler, J. Geiß, and M. Gertz. "Semantic Word Clouds with Background Corpus Normalization and t-distributed Stochastic Neighbor Embedding". In: *CoRR* abs/1708.03569 (2017). URL: http://arxiv.org/abs/1708.03569

Prototype Overview: Data Preparation

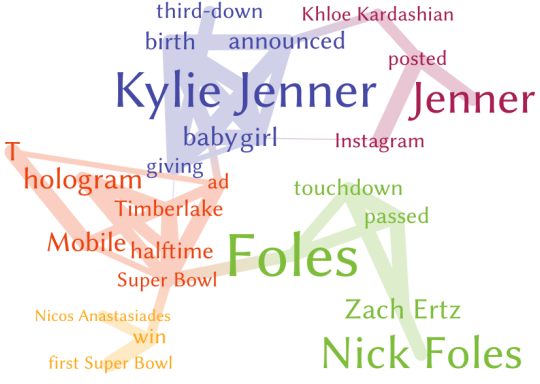
1. monitor live news (push notifications & RSS)
2. group articles in microbatches (25 articles)
3. crawl and extract text
4. tokenize text, detect and link entities
5. aggregate weighted cooccurrences
6. score significance based on estimated frequencies
7. update estimates for next micro-batch

Use count-min style sketches for estimation:

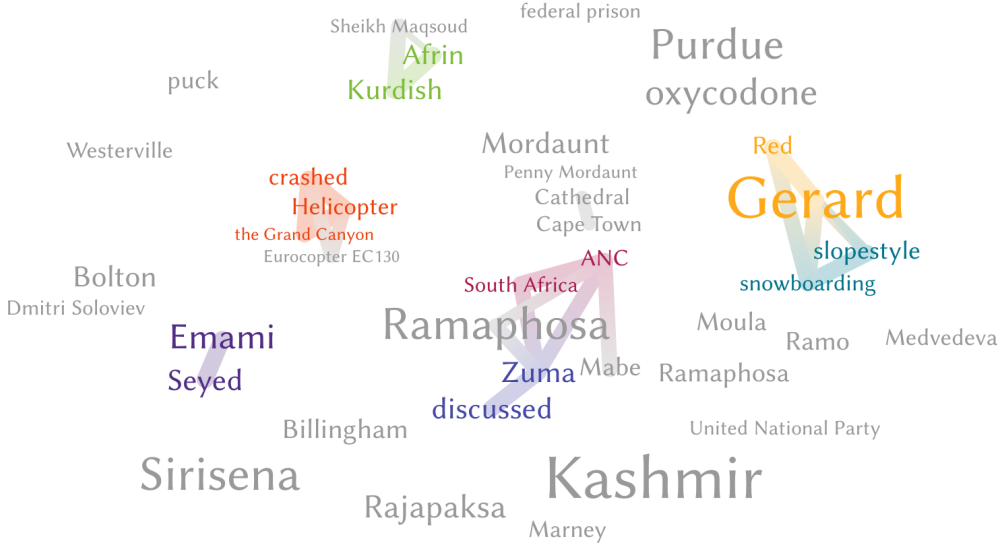
E. Schubert, M. Weiler, and H.-P. Kriegel. "SigniTrend: Scalable Detection of Emerging Topics in Textual Streams by Hashed Significance Thresholds". In: *ACM KDD*. 2014

Prototype Overview: Visual Layout

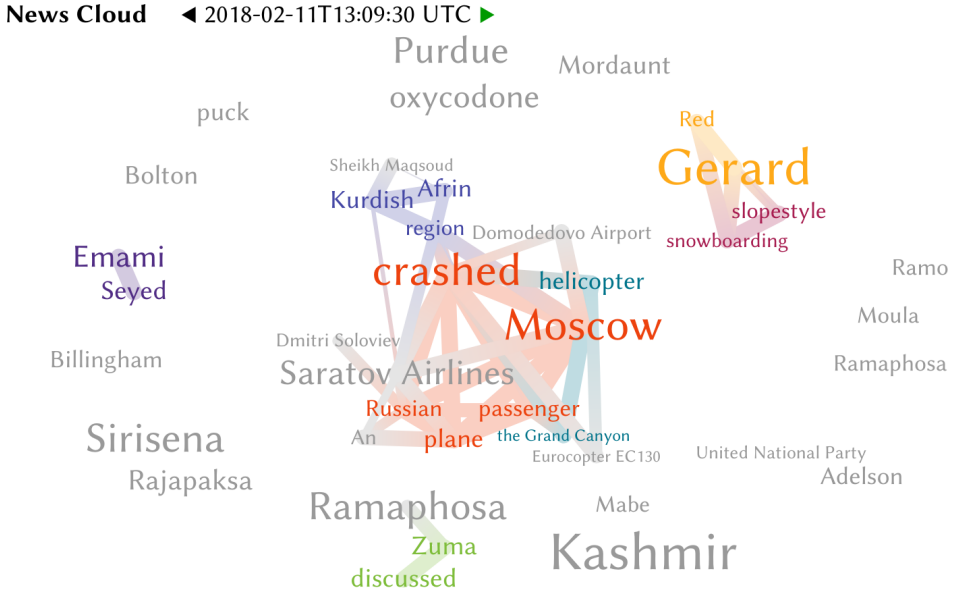
1. select and cluster top (co-) occurrences based on significance
2. visualize as word-cloud in the browser with significance-based SNE
3. edges visualize significant cooccurrences
4. colors denote clusters
5. currently supported languages:
   English and German



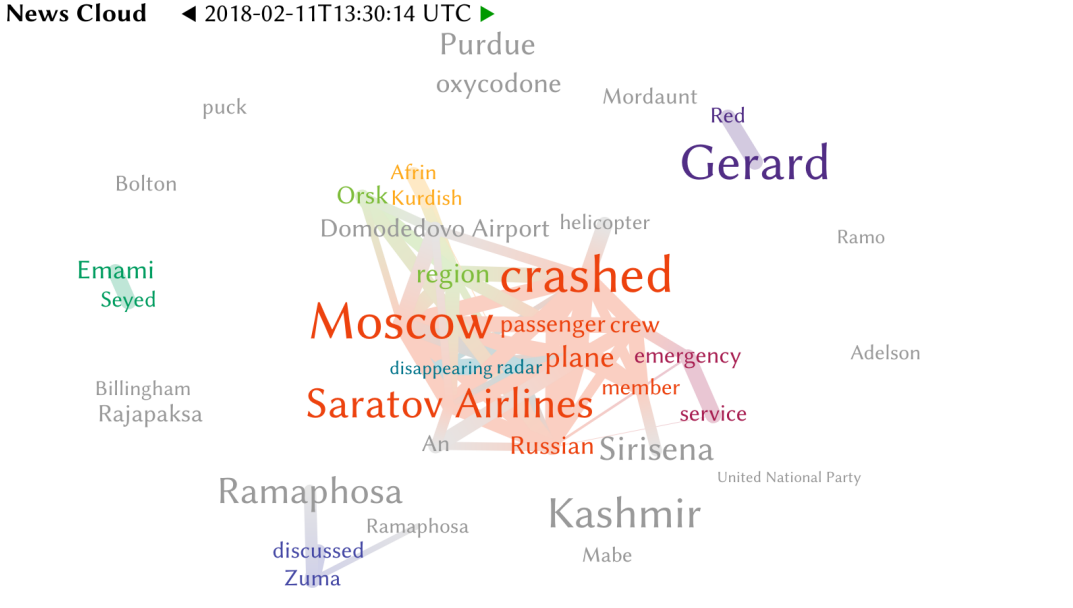
Topic Example: Moscow Plane Crash (prior)



Topic Example: Moscow Plane Crash (emerging)



Topic Example: Moscow Plane Crash (dominant)



Try the live demo:

newsir-demo.ifi.uni-heidelberg.de

Try the live demo:

newsir-demo.ifi.uni-heidelberg.de

Thank you!

Questions & Discussion