Spiral me to the core

Getting a visual grasp on text corpora through clusters and keywords

Maren Scheffel, Katja Niemann, Sarah Leon Rojas,
Hendrik Drachsler, Marcus Specht

Welten Institute
Research Centre for Learning, Teaching and Technology

Open Universiteit
welten-institute.org

Analysis

- Keywords play an important role to get a grasp on text collections
- Clustering of documents also helps getting an overview
- We employ both methods and combine their results

Keyword Extraction

- Keywords given by the papers' authors are often not broad enough, i.e. too narrow a word range

- We used the *AlchemyAPI* to extract keywords from the papers' abstracts and text bodies

- *AlchemyAPI* uses statistical algorithms and natural language processing techniques and ranks keywords accordig to their relevance

- Additionally to *AlchemyAPI*'s stopwords we created our own list

Clustering

- For the clustering we made use of the *carrot2* Java API
- Two of them employed soft clustering techniques with too many overlaps
- The third one is a bisecting k-means algorithm
- Every cluster gets two labels
- Calculations are again based on the papers' abstracts and text bodies

Combination of Results

- Results of the keyword extraction and the clustering are combined into a JSON file

- For every cluster the keywords of its papers are combined and sorted according to their rank, ten highest ones are kept

- Every cluster source file thus contains
  - Two labels
  - Ten keywords
  - List of the respective papers

Visualization

- D3.js framework was used to visualize the results
- Several publication-year combinations are available

Discussion

- Many paper lists contain way more papers from EDM than from LAK

- Only two clusters are dominated by LAK papers (*Social/Network* and *Analytics/Institutions*), five by EDM, rest is split

- LAK papers share their topic range with JETS and EDM

- Some EDM topics seem to be more exclusive and specific (e.g. *Skill/Parameters*, *Detector/Game*)

Discussion

- Many clusters have keywords covering aspects of a domain, an approach, a goal, the data used and stakeholders involved

- The *Analytics/Institutions* cluster:
  - Domain: *higher education*
  - Approach: *social network analysis*, *machine learning*
  - Goal: *learning process*, *student success*
  - Data: *student data*, *online learning environments*, *LMS*

Questions?

Suggestions?

The visualization is available at

http://mitarbeiter.fit.fraunhofer.de/~niemann/LAKchallenge2014/