HIGH PERFORMANCE DATA ANALYTICS

Pacific Northwest
NATIONAL LABORATORY

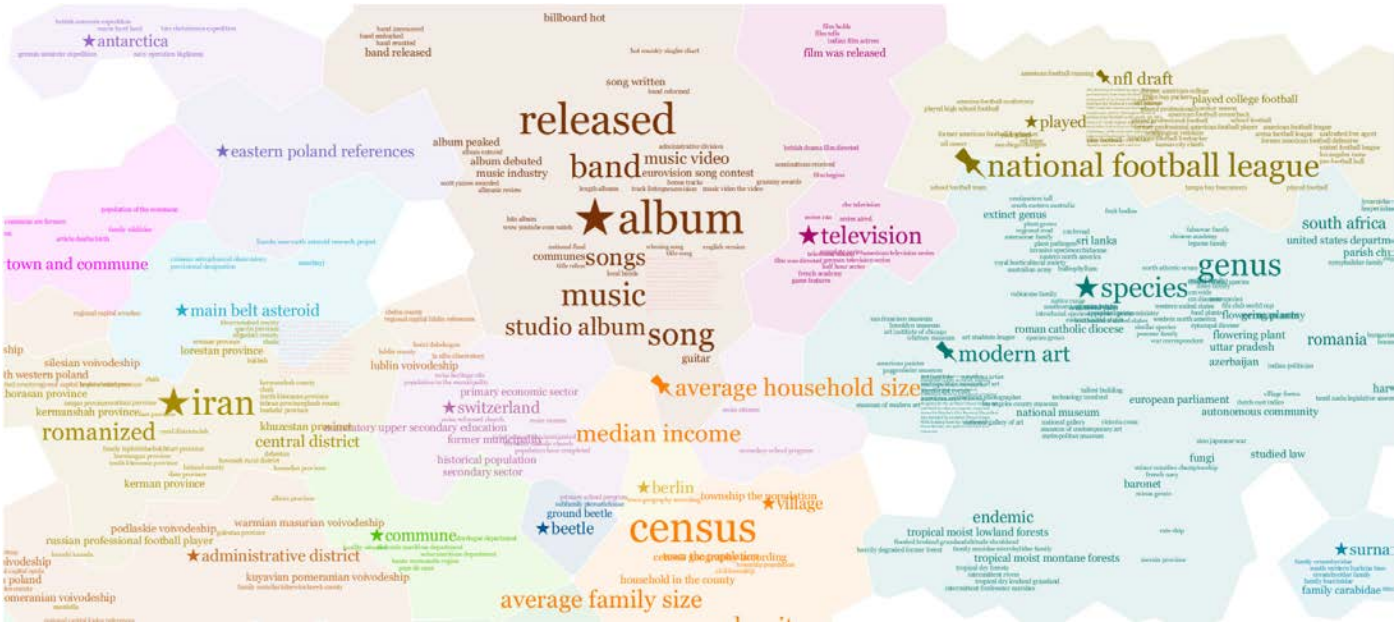
Proudly Operated by **Battelle** *Since 1965*

Typograph for Exploring Text Algorithm Outputs

CHALLENGE

The field of data mining has produced a number of modeling techniques, such as latent Semantic indexing (LSI) and latent Dirichlet allocation (LDA), to derive higher-level concepts, or topics, from a corpus of documents. These models ingest a corpus and create an index of frequently occurring terms in those documents. The terms and their co-occurrences are used to assemble terms into topics. While generating topics from large corpora is computationally challenging, user analysis of these topics can be equally daunting. This user analysis challenge is the focus of our research effort.

Developing an efficient mechanism to visualize the output of topic models, providing new insights into the content and semantic attributes of very large data sets.



The Typograph visualization of Rapid Automatic Keyword Extraction (RAKE) keywords from the English Wikipedia corpus.

U.S. DEPARTMENT OF ENERGY

CURRENT PRACTICE

When generating LSI or LDA models, researchers are interested in understanding the generated topic models to better comprehend the topics created by their model. Current practice involves manually inspecting spreadsheets and other data structures. Using spreadsheets highly constrains the exploratory analysis capability and provides researchers with a sub-optimal representation of their corpus data model. This approach also makes it difficult to see a more global perspective of topics and relations to other topics.

Typograph currently exists as a prototype system running at Pacific Northwest National Laboratory. Typograph was initially designed to use PNNL's Rapid Automatic Keyword Extraction (RAKE) algorithm[1], which extracts single and multi-term keywords from each document. Research into the design of the Typograph user interface and Typograph processing pipeline was performed at PNNL using a corpus of 4.3 million English articles from Wikipedia. This prototype system employs the 10,000 most frequent RAKE keywords to find their associations based on co-occurrence within documents. Given these associations, the Typograph system clusters these keywords, placing them into a hierarchy structure that can be displayed visually and explored interactively in the Typograph user interface.

Previous attempts to visualize LSI and LDA topic models successfully demonstrated what a gensim topic model might look like within the Typograph user interface, but weaknesses were exposed in the pipeline and Typograph data structures, making it more difficult to visualize these models. Based on in-depth knowledge of the Typograph engine's inner workings, PNNL was able to incorporate new algorithms. A limited number of integration points are currently available, and those are restricted in the types of model output that can be provided to Typograph.

TECHNICAL APPROACH

Using lessons learned in the current research, the pipeline will be redesigned and engineered to permit the results of other algorithms, such as LSI and LDA, to be used to provide necessary Typograph data artifacts. PNNL recognizes that not all alternative algorithms can produce every required artifact. Where possible, the software will be designed so that some data artifacts can be considered optional, and functionality requiring those artifacts will be disabled when they are not available.

Specifically, this processing pipeline will be split from large monolithic components into smaller, interchangeable components, exposing well-defined interfaces and data structures that connect the pieces together. This new pipeline will provide the flexibility to incorporate artifacts from LSI and LDA topic models, as well as future analytics, for visualization in the Typograph user interface.

IMPACT

Upon successful completion of this work, an efficient mechanism for visualizing the output of topic models will be available, providing new insights into the content and Semantic attributes of very large data sets.

[1] See Rose et al. "Automatic keyword extraction from individual documents." Accessed October 11, 2016 at http://media.wiley.com/product_data/excerpt/22/04707498/0470749822.pdf.

Contacts

David Gillen
Principal Investigator
(509) 375-5935
david.gillen@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965