

SEASR Analytics

Loretta Auvil, University of Illinois at Urbana-Champaign
Boris Capitanu, University of Illinois at Urbana-Champaign
Matthew Jockers, Stanford University
Ted Underwood, University of Illinois at Urbana-Champaign
Ryan Heuser, Stanford University

Abstract

Over the last decade, millions of dollars have been invested in creating large digital library collections of full-text humanities resources. These collections are large enough and rich enough to provide an excellent opportunity to go beyond traditional reading and search to explore tools that enable new ways of learning and teaching via digital scholarship. The SEASR (Software Environment for the Advancement of Scholarly Research) project provides the data synthesis, analytics tools, and supporting infrastructure required to analyze and study such digital collections. This paper describes how the SEASR environment, and in particular the Meandre data analysis framework, facilitates two different analysis techniques -- topic modeling and concept mapping, and a web application using the Google Ngrams data.

Introduction

Over the last decade, millions of dollars have been invested in creating digital library collections. At this point terabytes of full-text humanities resources are publicly available in data repositories, maintained not only in libraries, but also by publishers. These collections, dispersed across many different institutions, are large enough and rich enough to provide an excellent opportunity to go beyond traditional reading and search to explore tools that can enable new ways of learning and teaching via digital scholarship. The SEASR¹ project has developed, tested and applied the data synthesis, analytics tools, and supporting infrastructure required to analyze and study such digital collections.

Background

The SEASR project enables digital humanities developers to rapidly design, build, and share software applications that support research and collaboration. The SEASR project developed an infrastructure (called Meandre²) to support management and integration of basic resources and deployment of analysis in support of digital scholarship. SEASR is the project; Meandre is the data analysis framework. Meandre is a mashup environment for integrating many tools; it is a semantic-web-driven data-intensive environment providing the machinery for assembling and executing data flows which are software applications consisting of software components that process data (such as by accessing a data store, transforming the data from that store, and analyzing or visualizing the transformed results). It provides tools for creating components and

¹ SEASR website, <http://seasr.org>

² X Llorà, Ács B, Auvil L.S., Capitanu B., Welge M.E., Goldberg D.E. (2008). Meandre: Semantic-Driven Data-Intensive Flows in the Clouds, in Proceedings of 4th IEEE International Conference on eScience, 238-245 , IEEE Press.

flows, a high-level language to describe flows, and local and distributed execution environment based on a service-oriented paradigm. Meandre also provides publishing capabilities for flows and components, enabling users to assemble a repository of components for reuse and sharing. This allows users to leverage other research and development efforts by querying and integrating components that have been published previously at other shareable repository locations.

Meandre provides a set of reusable and expandable software components that were developed for digital humanities projects to provide access to data assets, transformation of data, mining of data and text, wrapping of several different natural language processing tools, and mashups with several existing visualizations. In this software demonstration, we describe two different analysis techniques that are provided within Meandre -- topic modeling and concept mapping, and how Meandre can be used as a service for a web application.

Analysis Techniques

Topic Modeling

Topic modeling is an exploratory approach to find patterns in a large text collection by identifying words that frequently occur together. Our collaboration with Matthew Jockers from Stanford led us to integrate the Mallet³ tool for topic modeling within the Meandre environment. The Mallet topic modeling tool contains “an extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyperparameter optimization, and tools for inferring topics in new documents given trained models.”⁴ This integration was implemented by creating several components in Meandre to call the Mallet API to transform the data into the required data structures needed by Mallet and another component to actually perform topic modeling. These Mallet components leverage existing Meandre components that perform natural language processing (NLP), such as part of speech (POS) tagging and filtering. Figure 1 shows the Meandre Workbench with the Mallet components listed in the left pane, an example flow using the Mallet components in the workspace, and properties for the modeling component displayed in the right pane. For more information on this process and analysis, look for Matthew Jockers book forthcoming from UIUC press under the title *Macroanalysis: Methods for Digital Literary History*.

³ McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." <http://mallet.cs.umass.edu>. 2002.

⁴ Topic Modeling, <http://mallet.cs.umass.edu/topics.php>

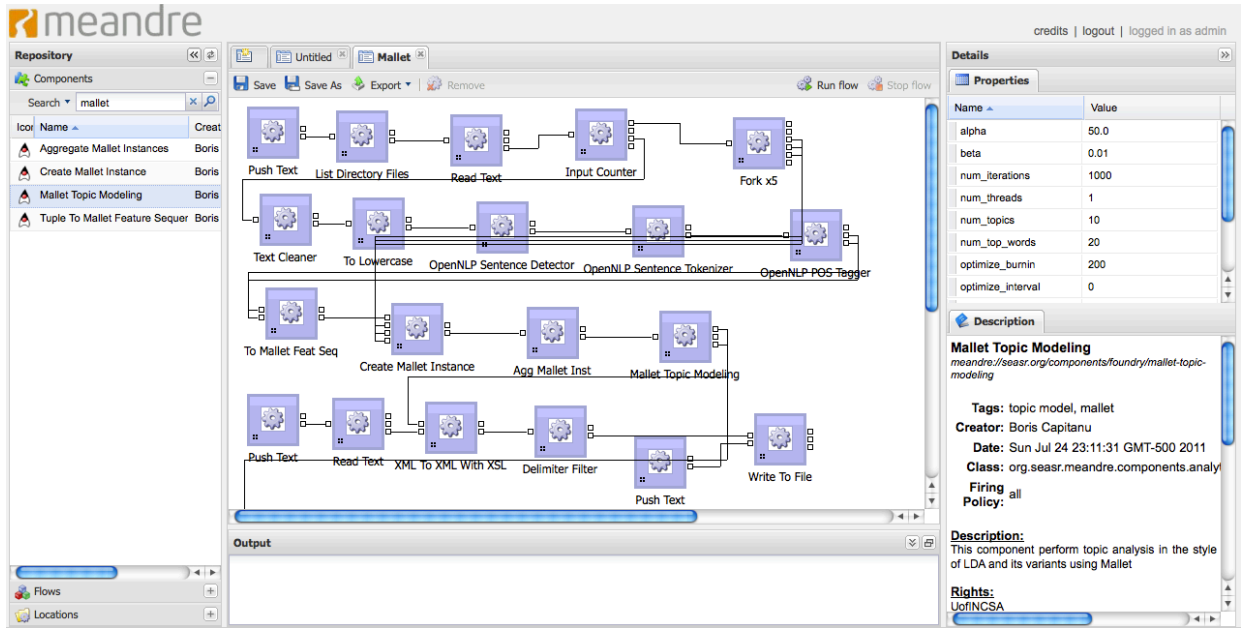


Figure 1. The Meandre flow shows an example of the Mallet components and how they are used with other NLP components.

Topic modeling analysis of a single book can reveal a set of topics that were covered throughout the book. A flow was constructed in Meandre that loads and segments the documents, performs part of speech tagging, constructs the data structure for Mallet, performs Mallet topic modeling, saves the data, and creates a visualization for each topic. Figure 2 shows some of the results of the topic modeling by using a tag cloud to show the relevant words of each topic for The Bible, Old and New Testaments, King James Version⁵. In this case, 8 topics were selected for display with at most 200 words to represent the topic.



Figure 2. Topic modeling results showing 8 topics with each topic shown as a tag cloud of the important words for this topic.

⁵ Project Gutenberg, <http://www.gutenberg.org/ebooks/10>

Concept Mapping

Another exploratory analysis using Meandre is the mapping of concepts across documents or collections, which can be used for sentiment tracking. In this demonstration sentiment is captured by using six emotions (Love, Joy, Surprise, Anger, Sadness, and Fear), as suggested by Parrot⁶, and tracked throughout a book. Documents are segmented by sentences which are subsequently processed through other components, such as part of speech tagging and selection of adjectives. Meandre then attempts to find a path from each adjective to one of the six emotions where each emotion is defined as a concept with a set of seed words. For instance, for the concept “surprise”, seed words include “surprising”, “amazing”, “wonderful”, and “unbelievable”. Meandre attempts to map all adjectives in the document to one of the defined concepts. Each adjective is given a label of the emotion that best fits the metric. In this case the metric used is shortest path, which assigns better ratings for symmetrical paths, ignoring all path lengths greater than 5.

Sentences are aggregated together to form a smoothing in order to highlight trends in a visualization. Figure 3 shows the visualization of the concepts over the course of two different books. One can see that the first book has much higher percentage of fear than the second, while the second has much higher percentage of love.

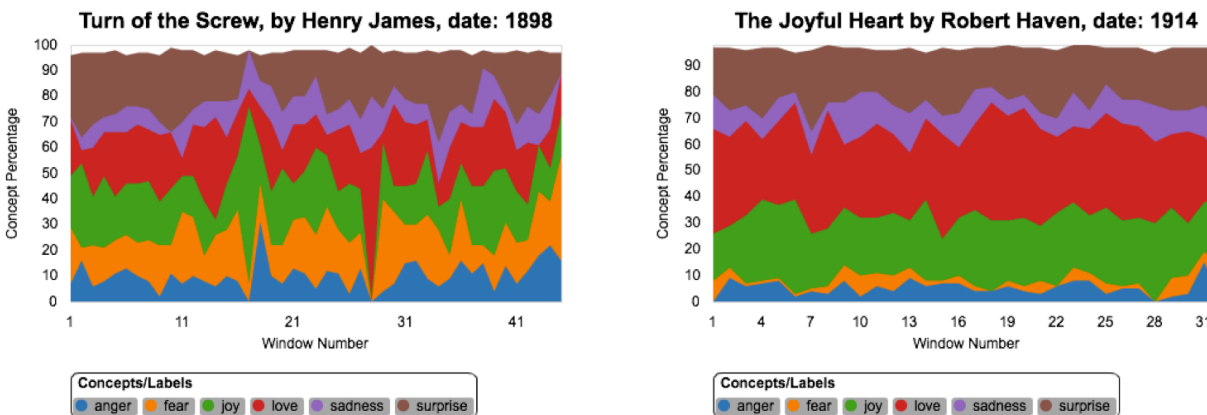


Figure 3. Results of concept mapping on six emotions for two different books.

Web Application Services

Analysis and Visualization of Google Ngrams Data

A simple web application was created using five different Meandre services to provide an exploration tool that combines graphing with correlation findings. Ted Underwood from UIUC collaborated on this project to improve the Google Ngrams⁷ data for using a Pearson⁸ correlation

⁶ Parrot 2001, <http://changingminds.org/explanations/emotions/basic%20emotions.htm>

⁷ Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)

⁸ Pearson Analysis, lib_mysqludf_stat package from <http://www.mysqludf.org>

coefficient analysis to find linear relationships between pairs of ngrams. This new version of Google Ngrams data (for 1-grams) addresses case-sensitivity, period spellings, past-tense syncope ('d), and f/s substitution, as well as many other OCR issues. The list shows how the variations of the original ngrams are mapped to a single ngram. For instance, Figure 4 shows a long list of ngrams that are mapped to “science”. The 1-gram data from Google Ngrams were loaded and then several additional tables were created to clean and normalize the data for further analysis. All the ngrams were run through a Meandre flow that performs spellchecking with some additional transformation rules to identify OCR errors. Meandre was also used to manage all the database queries that were run to compute Pearson coefficients using a database function.

The application has a section for querying or calculating Pearson results, an Ngram Viewer and a list of ngrams that are included in the lines drawn in the graph. Each of these analyses execute a Meandre flow running as a service and then passes the search query, and in some cases other additional parameters, such as minimum and maximum years. For instance, the list of ngrams and their counts displayed in the graph for “she” are returned with a service call in the following format: `http://[machine]/[port]/service/post?query="she"&min_year=1700&max_year=1899`.

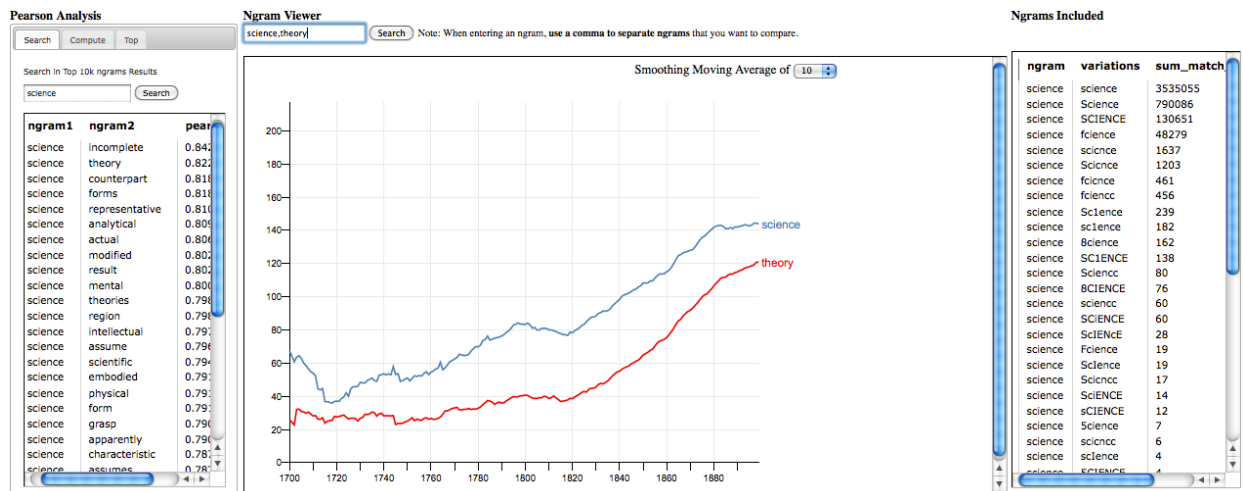


Figure 4. Web application with Meandre executing the services for Pearson Analysis and Ngram viewing.

Conclusion

Through the SEASR project, the Meandre environment and core set of text analysis tools provide a foundation for a variety of analyses and visualizations critical to the meaningful study of digital humanities data. As demonstrated here, not only are a variety of data synthesis and analytical techniques supported (including topic modeling and concept mapping) but the flexibility of the environment enables Meandre to act as a web service that can be leveraged in multiple ways. The power of Meandre comes from its flexibility to easily integrate tools from a variety of sources which can dramatically expand the capabilities researchers can leverage to conduct their research.

Acknowledgements

We thank The Andrew W. Mellon Foundation for funding the research and development of SEASR/Meandre.

Bibliography

- Llorà X, Ács B, Auvil L.S., Capitanu B., Welge M.E., Goldberg D.E. (2008). Meandre: Semantic-Driven Data-Intensive Flows in the Clouds, in Proceedings of 4th IEEE International Conference on eScience, 238-245 , IEEE Press.Parrot 2001, <http://changingminds.org/explanations/emotions/basic%20emotions.htm>
- McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." <http://mallet.cs.umass.edu>. 2002.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books. Science (Published online ahead of print: 12/16/2010)
- Parrot 2001, <http://changingminds.org/explanations/emotions/basic%20emotions.htm>
- Pearson Analysis, <http://www.mysqludf.org>, lib_mysqludf_stat package
- Project Gutenberg, <http://www.gutenberg.org/ebooks/10>
- SEASR, <http://www.seasr.org>
- Topic Modeling, <http://mallet.cs.umass.edu/topics.php>