Library Lab: Making Sense of Thousands of Email Messages

Mercè Crosas, Wendy Gogel, Ellen Kraffmiller, Andrea Goethels, Robert Treacy, Brandon Stewart

Project summary

The objective of this Library Lab was to evaluate whether a text analysis tool developed at the Institute for Quantitative Social Science (IQSS) could be used for library projects that involve large amounts of unstructured text. The tool, named **Consilience**, provides an interactive user interface to help you discover different ways of grouping sets of documents, zoom into each cluster within a selected clustering, and zoom in further into each document. In particular, we have used a set of email messages from the Harvard Library (HL) email archiving project to test how Consilience can provide useful clustering and categories to curate and catalog a set of text documents.

Accomplishments

Software Development Accomplishments

During this year, the development team worked on improving the methods to process the text files and pre-calculate a large number of potential clustering solutions. This has helped to improve significantly the usability of the tool – a user can select a point in a space of possible clusterings, and instantaneously is given the clusters generated by that point. The Consilience team has also improved the user interface to allow more easily to zoom into each cluster and each document within a cluster, and compare a clustering point with another. The document viewer also highlights the terms that are common across the documents that form a cluster.

Evaluation Accomplishments

A group of archivists representing Schlesinger Library, Harvard Art Museums, Countway Library and Houghton Library are testing Consilience to provide feedback to the developers on whether it would be useful for processing electronic text-based collections by aiding in the analysis, organization and description of content. The Consilience team also requested that the archivists send suggestions for enhancements and future functionality. The Consilience team loaded several hundred test emails provided by the HL email archiving team as an example of electronic text-based collections, then met with the archivists in September to demonstrate the tool and discuss potential ways to test. The archivists have begun to test and provide feedback and will continue to do so. The archivists from

Countway Library have requested the ability to test on text of some of their larger collections. So far, the archivists are enthusiastic about the potential of the tool to assist them in processing large electronic text collections. More input and refinement is needed for the application to reach v1.0 release, including documentation, searches, and additional usability improvements.

Challenges

We would like to scale the application to process 100,000s text documents. For now, we have proved that the application works well for 10,000 text documents, but we need to continue optimizing our clustering calculations and further parallelize the text processing jobs. It is also a challenge to represent large groups of texts and categorizations in a quick, intuitive, visual way. We would like to refine the existing visualizations and provide additional ones.

Next steps

We plan to expand Consilience to facilitate ingesting document sets. Currently this process is semi-manual and time consuming. We plan to automate it and connect and optimize all the pre-processing steps, which pre-calculate a large number of clustering solutions.
We are also planning to scale the application to take document sets with 100,000 text documents. Currently, it works well with sets of 10,000 documents.
We plan to continue testing with others and collecting feedback during the alpha and beta development phase.

Confirmation that you deposited code, if applicable

The entire source code for Consilience is in GitHub. We have not yet decided on the license we are going to use to distribute the code, so for now the repository in GitHUb is private (https://github.com/IQSS).

Budget spent

Spend full amount: $15,000.

Publicity and Presentations

Since Consilience is still in alpha, we haven't done publicity yet. The Consilience team has met with the Library archivists for internal demos, and Wendy Gogel and Merce Crosas have done a short video for the Library Lab on this project.