Library Lab - Making Sense of Thousands of Email Messages

Mercè Crosas (Institute for Quantitative Social Science (IQSS)), Andrea Goethals and Wendy Gogel (Library Digital Preservation, Office for Information Services (OIS))

Changes:
In “Introduction”, we’ve added a more detailed description on how the Text Clustering tool works.
In the “Design and Implementation” and “Budget” sections, we’ve changed phase I to be a proof of concept and evaluation to establish if the clustering tool is useful to archivists, before developing the ingest tool.

Summary:

IQSS has developed a research tool that facilitates grouping and organizing large sets of digital documents, helping the user make sense of them through an interactive process. We propose to build an ingest front-end to the tool that will allow the Harvard Library to use this tool to understand, organize and label email archives and similar digital content worth preserving.

Introduction

Libraries and researchers welcome the vast amounts of information that can be easily gathered through the internet, modern media and technologies. Libraries’ holdings and offerings can grow daily by collecting emails, tweets, web sites and archiving them properly. Libraries and researchers can use that data to advance our knowledge of human societies and in general of our world. The flip side is that now we have endless amounts of information to process and try to make sense of.
A first step towards understanding that content is to categorize it and label it properly. But this can take infinite time if each piece of content needs to be read and categorized manually, even if it may seem otherwise (to understand the scale of the problem, a set of only 10 documents can already be categorized into 115975 possible partitions! This is known as the Bell number which increases very rapidly to unmanageable sizes). Decades of research have been devoted to find the right algorithm to calculate fully automatically a desired partition. However, there are many clustering algorithms that can be used and each one could give an interesting grouping and labeling of the documents. At the end, the person interested in categorizing the documents needs to decide what grouping makes more sense in a given context. To achieve this, we need an interactive tool that assists the user to find the right categorization, taking advantage of the cumulative knowledge of existing clustering algorithms. At the Institute for Quantitative Social Science (IQSS), we have implemented such tool. This new clustering tool combines statistics and clustering algorithms to facilitate finding the useful partition of a large set of documents containing unstructured text into subsets that share the same words and concepts. It helps organize and read documents (papers, articles, emails, tweets, etc) in a whole new way that does not only save large amounts of time, but also reveals unthought categorizations and can bring new discoveries.

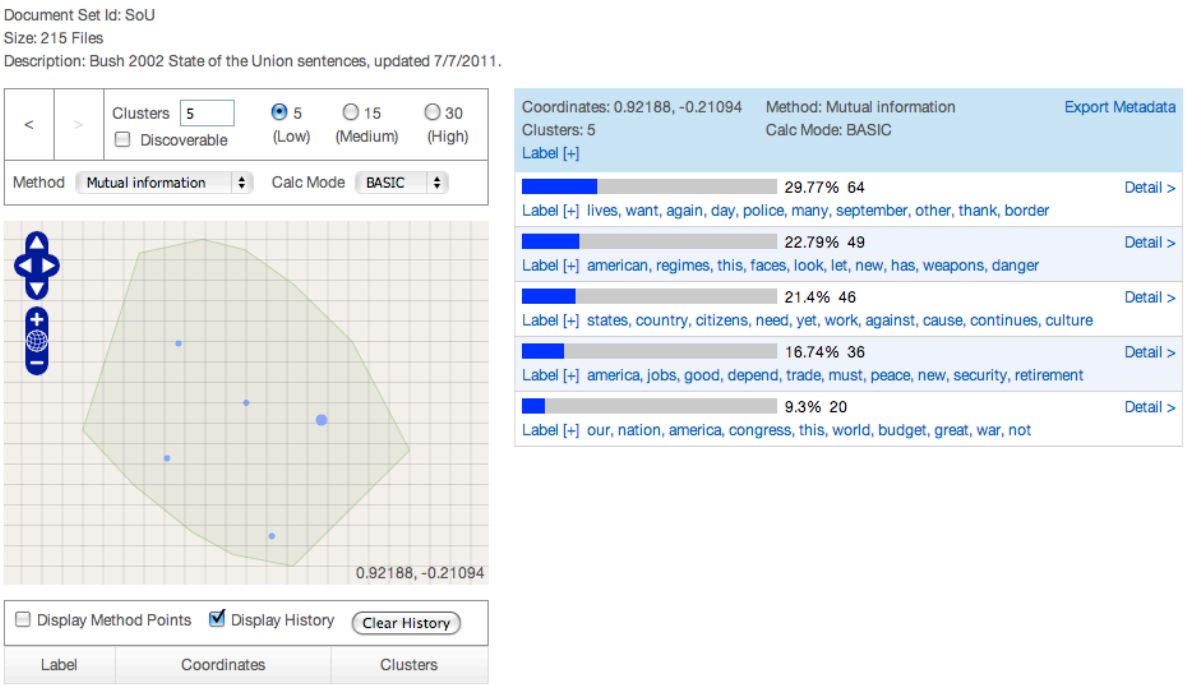

Figure 1. Text Clustering Tool User Interface

How does it work? First, it processes a given document set and calculates initial clusterings using all clustering methods defined in literature (~150 methods). Then it projects these clusterings in a "clustering space" where the distance from one clustering to another is defined by the variation information distance (http://en.wikipedia.org/wiki/Variation_of_information). The tool provides a visualization interface that allows users to choose any point in the clustering space. When the user clicks a point, the application calculates the new clustering for that point, using the distance from that point to all other pre-calculated clustering. Through this interactive process, the user can explore all clusterings near a clustering method of interest (or in a completely region) and choose a preferred categorization. For each clustering, the interface shows all clusters, each one automatically labeled based on the more common words across the documents contained in each cluster. The user can further zoom into a cluster and view each document and useful statistics.

The clustering tool, and the statistics behind it, has already been used to make genuine discoveries by categorizing and help further understanding the nature of press releases by politicians (see http://gking.harvard.edu/files/abs/discov-abs.shtml).

However, currently the tool can only be used by researchers who manually prepare the data into documents, and run multiple preliminary processes in order to "ingest" the documents into the tool. This process is manual, complex and requires strong technical expertise and understanding of the underlying statistics. We propose to develop an ingest tool that will automatically import a set of email messages and similar digital documents into the clustering tool. This work will be done with input from the Library Digital Preservation group, and as a test case of its usefulness within Library services, it will be usable to the on-going Electronic Archiving Services (EAS) application that is being built as part of the Library email archiving

project. EAS supports the archiving process of email messages and their attachments, but eventually will support other types of contents and formats. Adding the clustering tool to EAS will help Library archivists better understand the email content as they process it, by organizing it, grouping it and labeling it with useful metadata.

Design and Implementation

This project will be conducted in two phases. The first phase includes the creation of the tool and an evaluation by archivists of its ability to help understand the information contained within email message collections. Should the first phase result in a positive evaluation, the second phase will integrate the clustering tool into the archivists' workflow for processing email within EAS.

Phase 1 (what is being requested for funding in this proposal)

As an initial proof of concept, a set of email messages will be exported from EAS and uploaded semi-manuallyinto the clustering tool. Archivists will interact with the tool to organize, group and label the content. They will evaluate the tool to see if its use will add benefits to their email archive processing workflows.
This step will only require data parsing and re-formatting, and programming scripts to process and analyze the email messages set so it is usable by the Text Clustering application. It will not require to develop the ingest tool proposed for Phase 2.

Phase 2 (will ask for funding in a later proposal if the first phase is successful)
As a test case of the clustering tool's ability to be used within library services, the tool will be integrated into EAS. Archivists will be able to select sets of content within EAS to be imported into the clustering tool. The organization and labels applied interactively in the clustering tool by the archivists will persist in EAS to be used in the future by researchers exploring the email content.
The ingest tool will be developed in a flexible and scalable way to allow additional formats in the future. It will be implemented in Java and the code will be made open source so others can contribute to it and expand on it. It will accept the current email format used by EAS, EML conforming to RFC 2822.

In future phases the clustering tool could be integrated into additional Library services, for example it could be set up for researchers to explore the web archive content that has been collected in the Library's Web Archiving Collection Service (WAX).

Budget and Timing

Programming parsing, re-formatting and initial statistical calculations - 1.5 months FTE
Test and feedback/bug fixing iterations - 0.5 month FTE
Evaluation - 1 month (archivists with assistance from OIS staff)

Funding Requested for Phase 1:
Total: 2 month FTE (developer) + 1 month (existing Harvard archivists and OIS staff)