Text and Data Mining Acquisition and Support

Prepared by: Chris Palazzolo, Erica Bruchko, Chella Vaidyanthan, Sara Palmer
Date: 9 July 2018
Reviewed by: Collection Management

Purpose of this Document

To outline policies for the general acquisition, use, and management of text and data mining corpora.

Parameters & Considerations for Acquiring Text Mining Materials

The following research dataset considerations apply to text-mining as well:

- If the dataset can support multiple researchers across the university, then the Library will be more likely to purchase):
  - Encourage the use of research and grant funds to share costs of datasets of potential use to the entire Emory community, as well as allocated subject funds
  - The option of campus-wide access for a data resource should always be pursued. When data can only be available for usage by specific individuals or departments, purchases should be funded through the resources of those parties.
  - If retaining of data or destruction of data is a requirement (note that this is unenforceable by Emory, and can not be supported)
- Can data that were used for analysis done by a researcher (faculty and/or student) in the research process be publicly shared to comply with journals that have a data sharing policy? Note that the data availability policies of some journals now require authors of submitted manuscripts to provide not just the dataset that was actually analyzed in the manuscript but also the source data that were cleaned and assembled into that analyzed dataset. Journals will make exceptions for data that cannot be shared for terms-of-use reasons, so it will be helpful to have this matter clarified so that the information can be shared with people who wish to use the data.
- Can the data be shared with folks at another university also working on the project? Is this a need for the primary users of the data set? **NOTE: For the most part, vendors will NOT allow for the sharing of the underlying data in their database**
- Does the vendor/provider require any user of the data to sign a confidentiality agreement or the like to make use of the data? Are there any other sorts of administrative overhead associated with the resource?

Other Important Guidelines Specific to TDM:

- Generally, we also do not provide licensing or funding for individual text-mining projects with needs not covered by university wide licenses. Some vendors may require walled garden approaches to control and manage access to large amounts of their data. Frequently, fees are charged by the project. The library is unable to pay for project by

project fees, but will attempt to negotiate with the vendor for a more institutional solution. Therefore, as noted above, we highly encourage scholars to consider grant funding.
- Some vendors require the use of a registered API which does not require any additional funds, but which may need some intervention on the part of the subject librarian or head of collections.
- Emory prefers that all contracts include explicit clauses for the allowance of text/data mining and as a matter of course, should purchase the (physical) TDM data package along with any primary source for which this data is available. Backups will be made locally. Where required, updated licenses will be requested. In addition, if possible, Emory-owned vendor data/content should be available for ingest into Emory's Preservation Repository. Any additional parameters for text and data mining should be acquired at the time of need. As much advance notice as possible is requested.

Storage of Content

- We can store and maintain physical media as necessary for accessing data. In addition, if possible, Emory-owned vendor data/content should be available for ingest into Emory's Preservation Repository. Some metadata treatment may be necessary prior to ingest in order to align with Digital Collections policies. Requests for TDM corpora into the repository should be proposed and processed through regular digital collections channels.
- Not all data from vendors comes in a readily accessible structured form, and may be reformatted.
- We should work with online TDM tools and interfaces as they develop and explore possibilities for the inclusion of Emory locally digitized content.

Housing and Access to Text Mining Materials

Woodruff Library will work closely with the Emory Center for Digital Scholarship (ECDS) to provide access to text mining corpora. Subject librarians will maintain, along with ECDS, a libguide that outlines existing, free corpora for manipulation, along with information regarding library-purchased options. ECDS staff will provide basic instructions and assistance with text mining applications. If more intense assistance is required (e.g., use of programming such as Python), students or faculty should consider assembling a digital project proposal. In many cases, the vendor will assist the patron directly with technical assistance in utilizing a walled garden or API. ECDS and/or subject librarians may be able to assist.

When hard copies of data (e.g. disk drives) are provided by a vendor, ECDS staff may transfer these to local hard drives, and house the original drives in a locked cabinet. Some clean-up (such as file unzipping) may be completed to facilitate access to the content. The local hard drives will be available for check out for a period of two weeks. Users will be required to complete a brief MOU which outlines the terms of use (e.g., see UC Berkeley: https://guides.lib.berkeley.edu/c.php?g=491766&p=5377163#s-lg-box-wrapper-19973930).