Text and Data Mining Acquisition and Generative AI Collections Support

Prepared by: Chris Palazzolo, Sara Palmer
Feedback and Input from: Humanities, Social Sciences, and Area and International Studies teams and Collections Steering Committee
Original Date: 9 July 2018
Reviewed by: Collection Management
Update (for AI): Summer 2024

Purpose of this Document

To outline policies for the general acquisition, use, and management of text and data mining and/or AI learning corpora and content at Emory for research and teaching. Text and data mining, or TDM is here defined as means to download, extract, analyze, classify, model, or index information, using computational tools, algorithms, machine learning, artificial intelligence, or automated techniques. TDM may include the use of AI tools, but not always.

The Emory Libraries over policy for licensing TDM and AI training rights in licenses is outlined in **a separate document**.

Parameters & Considerations for Acquiring Materials

The following research dataset considerations apply to requests for corpora for for text and data mining and AI models (e.g. large language models).

- **Availability of data:** The Libraries will be more likely to purchase if available to more than one researcher. When data can only be available for usage by specific individuals or departments, purchases should be funded through the resources of those parties.
- **Funding:** The Libraries 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 and central library funds
- R**etention of data:** If retaining of data or destruction of data is a requirement (note that this is unenforceable by Emory). We can only make reasonable efforts to inform users of this stipulation.
- **Data sharing:** Ideally, data should be made available to co-researchers at other institutions.
- **Confidentiality:** 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?

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.

Other Guidelines Specific to Acquiring, Licensing and Using TDM and AI training data:

- Emory prefers that all contracts include explicit clauses for the allowance of text/data mining and AI training that should not be more restrictive than fair use. Language should be included as to how underlying data should be accessed. See the **AI licensing policy** for more guidance (*under development*).
- Where required, updated licenses will be requested. Any additional parameters for text and data mining should be acquired at the time of need. As much advance notice as possible is requested.
- Most, if not all, licenses restrict automated scraping methods for pulling data. Vendors are concerned about the security of copyrighted data.
    - Some vendors may require walled garden approaches or other secure ,methods to control and manage access to large amounts of their data for TDM or large language model training. The Libraries will inform the researcher and ask that they provide assurance that the data is managed in a secure location.
    - 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.
- Subject librarians will work with Electronic and Continuing Resources (ECR), Scholarly Communication and/or Collections to determine feasibility of using content from our licensed resources for such purposes.
- Generally, we also do not provide licensing or funding for individual text-mining projects or AI training data with needs not covered by university wide licenses. However, the Libraries will make efforts to incorporate text and data mining into existing contracts and/or liaise with the vendor to find an amenable option for the researcher.
- It is not recommended to use **third party** AI tools as (1) their use are restricted in our licenses, as they are not secure and (2) many commercial LLM providers may repurpose user data for training. It is advisable to opt out of any data-sharing.
- When sharing data from TDM or AI Learning models, it should suffice to simply describe your data sources and methodology rather than the underlying data or text, as this would be a license violation. Most publishers make exceptions for terms-of-use restrictions related to data sharing.

Services Provided by Emory Libraries

- Woodruff Library works closely with the Emory Center for Digital Scholarship (ECDS) to provide access to text mining corpora.
- Subject librarians maintains, along with ECDS, a libguide that outlines existing, free corpora for manipulation, along with information regarding library-purchased/licensed options and/or restrictions. Please contact your subject librarian if you have any need for AI training or TDM data.

- ❖ In many cases, text and data mining corpora will be retrievable through APIs provided by the vendor. These may have stipulations. As these corpora can be relatively large, it is up to the researcher to securely store the content per license requirements. Not all data from vendors comes in a readily accessible structured form and may need to be reformatted.
- ❖ ECDS staff may 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.
- Woodruff Library will work closely with the Emory Center for Digital Scholarship (ECDS) to provide access to text mining corpora.
- We will work with online TDM tools and interfaces as they develop and explore possibilities for the inclusion of Emory locally digitized content.
- In some cases, we can store and maintain physical media as necessary for accessing data. The Libraries may also ask for stand-alone purchased corpora in digital format to be locally preserved if feasible and space allow.
- The Libraires do not guarantee storage space for these materials or high-speed computing resources for TDM or AI activities.
- The Libraries may point users to technical contacts at vendors to assist with downloading and/or delivery of data.