TEXT AND DATA MINING – SEEKING TRACTION

Thomas Padilla

Visiting Digital Research Services Librarian

University of Nevada Las Vegas

thomaspadilla.org

Working in libraries affords many opportunities to engage challenges associated with text and data mining (TDM) limited access datasets. Challenges afford themselves across interactions with diverse disciplinary communities at institutions with varying resources. Increasingly, data requested by research communities fall outside the traditional purview of library collection development and research support. Solutions to TDM challenges are few given gaps in understanding and misaligned values. Continued activity in light of these factors fosters an environment where weaknesses and threats are many and seemingly tractable opportunities are few. In the space of this brief statement I will introduce four challenges: **(1)** underdeveloped and inconsistent content provider effort to meet text and data mining need, **(2)** misaligned values reinforced by ambiguous and/or overly restrictive content provider terms, **(3)** debt incurred by technical abstraction, and **(4)** purposeful technical opacity.

To a large extent content provider efforts to meet TDM needs are underdeveloped. Wide variation in mechanisms for provisioning limited access datasets evidence that underdevelopment. From one provider to the next there is little indication of concerted development activity which leads to a wild and woolly data access environment. This situation places undue burden on users and library service providers. Variation in development of application programming interfaces, where they are available at all, creates user learning curves and library service debts that are difficult to anticipate let alone bear. Variable positions on data documentation inclusion – that is to say which accompanying information a content provider deems in scope for data acquisition can be difficult to anticipate. For example, there have been instances of content providers refusing to share document type definition files along with a bulk data purchase given a claim of “intellectual property”. In a similar vein, content providers have been hesitant to provide documentation that details the component parts of a dataset where they are constituted from multiple publications. Where that data is provided upon request, post time delay, it can appear in forms that needlessly obfuscate data evaluation – e.g. a PDF of tabular data describing the dataset. Further, data documentation arising from a user paradigm wherein bibliographic description is fit for purpose is less well matched to a data driven paradigm.

Underdeveloped and/or misaligned content provider values relative to TDM are readily apparent at the level of terms associated with limited access datasets. It is not uncommon to be referred by a content provider to a provision for sharing ‘snippets of text’ when inquiring on the ability to share data utilized and/or generated in the course TDM research. The ability to share data has a direct correspondence to the ability to conduct transparent research oriented toward reproducibility or replicability. Reference to a “snippets of text” provision serves to inhibit library guidance in the use of resources for TDM given the ambiguous nature of

the provision. Some institutions will move forward with that ambiguity and others will not. Guidance of this kind further compounds the long-term ability of university, domain, and scholarly repositories to acquire and preserve TDM research. In a slightly different vein, engagements with non-traditional content providers that vend social media data entail onerous and cost prohibitive terms that impact the viability of library support and broader research potential. For example, a National Science Foundation proposal focused on studying linguistic change in the Midwest was pursued by 3 institutions. In order to support this research, the principal investigator sought to acquire data through a social media data vendor. The data vendor required each participating institution to acquire the data and when asked about sharing data for the purposes of peer review suggested that an additional acquisition would need to occur. According to these terms the possibility of moving forward was essentially dead on arrival.

Finally, efforts to enable TDM with limited access datasets have been challenged by degrees of technical abstraction or intentional technical opacity that can be difficult to weather. For example, HTRC's data capsule project is a wonderful sign of progress but abstraction via the non-consumptive paradigm holds the potential to create a technical and methodological debt that compounds researcher effort in the course of peer review and library effort to promote the viability of the service. On the content provider side, a related tension with very different motivations, can be observed in proprietary machine learning approaches to generating open data and scholarly metrics. Without better understanding of the method generating the data and the metrics serious questions arise as to the viability of that data. Furthermore, this data constitutes a potential threat insofar as administrative pressure is set to push building upon those data in the pursuit of demonstrating scholarly impact. This essentially creates a lock in with seemingly open data vended on an opaque, proprietary foundation. It is akin to a "DPLA as platform" approach without the corresponding commitment to openness.

These are just a few, high level glosses of issues I have encountered in the library community. I look forward very much to learning from my peers at the IMLS supported, Data Mining Research Using In-copyright and Limited-access Text Datasets national forum.

Acknowledgement

My thanks to Devin Higgins and Alix Keener for reviewing and commenting on this statement.