NASIG Newsletter
Vol. 31, no. 4 December 2016 eISSN 1542-3417

# NASIG Conference Report: Text Mining 101: What You Should Know

*Ethan Pullman, Carnegie Mellon University; Denise Novak, Carnegie Mellon University; Kristen Garlock, ITHAKA.org; Patricia Cleary, Springer Nature*

Reported by: Marcella Lesher

As promised by the title of their presentation, the speakers provided a comprehensive overview of text mining and how it impacts and provides opportunities to libraries, library service providers, publishers, and, most importantly, researchers. Novak of Carnegie Mellon started the program off by defining text mining as "the automated processing of large amounts of structured digital texts" which enables researchers to analyze and interpret massive amounts of textual data, an impossibility using traditional retrieval methods. Pullman, also of Carnegie Mellon, highlighted examples of text mining projects that use word clouds built from mining large texts, including a class project looking at case documents in the *Authors Guild v. Google* copyright infringement case, and a Carnegie Mellon and Georgetown University joint project called the Six Degrees of Francis Bacon (http://www.sixdegreesoffrancisbacon.com/). This project has allowed researchers to trace the "social connections" between individuals during the time period of Bacon's life.

Pullman described how text mining challenges the traditional roles of library liaisons by going beyond the task of acquiring texts and providing access to them. He noted that "Librarians need to understand how texts are used in the digital age, what tools are available, and what issues impact their acquisition and access." Pullman posed the question of how a librarian can stay informed in order to bring these new tools and methods to faculty and student patrons. He remains informed by reviewing faculty curriculum vitae, publications, syllabi, and research showcases. In general, participation in the research and scholarly communication life of faculty and students is critical.



Novak discussed the acquisition factors associated with text mining. Acquiring text mining services requires knowledge of who will allow text mining, cost information, and licensing that will permit text mining

to take place. At Carnegie Mellon users are presented with library guidance that describes text and data opportunities as well as links to free sources that allow text and data mining (http://www.library.cmu.edu/research/tdm/overview).

Support of text mining of the JSTOR digital library was discussed by Kristen Garlock. She presented information on JSTOR's free Data for Research service (http://about.jstor.org/service/data-for-research) which is "a self-service website for generating datasets from the content on JSTOR." This type of service provides both opportunities and challenges for the organization. Opportunities include development and promotion of new types of scholarship, new partnerships, increased use of publications as scholarly tools, and increased recognition of influential articles. Challenges include staffing and support, keeping up with research trends, and the increasing number of requests for larger and more complex data sets.

Cleary, from Springer Nature, described the publishing side of text mining. She noted that Springer Nature will very shortly be updating their text and data mining (TDM) policy. As noted on her slide presentation, "Springer grants text and data mining rights to subscribed content, providing the purpose is non-commercial research." Individual researchers can download content directly from the SpringerLink platform without going through a registration process. Future SpringerLink subscription agreements and renewals will include a TDM clause; those holding current agreements may also add the TDM clause to take advantage of TDM now. Cleary provided some technical guidance to downloading content, indicating that the CrossRef TDM initiative may be useful. Springer Nature also provides a free metadata API that allows for searching Springer content.