IFLA WLIC 2013
SINGAPORE

Submitted on: July 24, 2013

Text mining newspapers and news content: new trends and research methodologies

Debora Cheney
Library Services to the World Campus/Penn State Online
The Pennsylvania State University Libraries
University Park, PA 16801
USA
dcheney@psu.edu

Abstract:

A growing body of research in the humanities and social sciences seeks to "mine" the text of newspapers and news content. The trend crosses historical newspapers and current newspapers. It also seeks to understand how news travels through social media (Facebook, Twitter, news blogs, etc.). The research method does not rely on the original form of the newspaper or news source, but rather on statistical and word patterns present within the mined text; typically these results are also presented visually in a variety of graphs, word maps, etc. that allow users to visual news text in new ways. This exciting and challenging research methodology presents several challenges for libraries seeking to provide access to news content to support this research method, including access and licensing, copyright, technology and software support, storage and access issues, and reference, instruction, and training needs. This paper will present examples of research conducted using text mined from news sources and present an overview of the challenges libraries and archives will face as they seek to support this new research methodology.

Keywords: newspapers; digital archives; text mining; digital humanities;

Background

A growing body of research in the humanities and social sciences seeks to "mine" the text of government documents, novels/literature, magazines, newspapers and social media content. This new area of research is part of the larger trending interest in "big data"--that is, to use the wealth of data and information generated on a daily basis by and about people, their behavior and interests to "help bring the next thing to market, predict the outcome of elections, and much more." (Rosenbush & Totty, 2013) Companies such as *Ancestry.com* indicate they will crawl digital newspapers in order to make their records more robust and more relevant to their researchers. (Harris, 2012) As interest in text mining newspapers and social media content grows, libraries will be presented with new challenges to provide access to news content to support this research method, including access and licensing, copyright, technology and software support, storage and access issues, and reference, instruction, and training needs. This paper will provide examples of research using text mined from

news sources and discuss the challenges libraries and librarians may face as they support research based using text mining methodologies.

Newspapers have always recorded the first draft of history and many researchers have long sought to understand and gain insight into people, places, and times from their content. Yet, as many researchers know, the research methodologies were cumbersome and slow. Researchers are creating new research methodologies to mine the text of growing amounts of digitized historical newspapers, digital news content (which may appear in a newspaper or only on a website), and social media content. *Google's Ngram* viewer provided a quick way to visualize digitized text resulting from the Google Books project and quickly allowed researchers to visualize words in a new way. (Zimmer, 2012) This research methodology does not rely on the original form of the newspaper, news source, blog posting (Berendt, 2010) or tweet, but rather on statistical and word patterns present within the mined text. Typically these results are then presented visually in a variety of graphs, word maps, and other new visualization methods that allow users to visualize the news in new ways.

Opportunities/Challenges

Because the need is so vast, text mining news content presents new opportunities and challenges for libraries that should be considered as the interest in big data grows and evolves.

A: Access, Licensing and Copyright Issues

Text mining news requires a large corpus of news content which is freely accessible to researchers. As on researcher argues, it is not just about big data, but long-data—that is, "data that has massive historical sweep—taking you from the dawn of civilization to the present day." (Arbesman, 2013) Newspapers certainly don't take us back to the dawn of civilization, but have existed in various forms for many centuries—but are there sufficient amounts and variety freely accessible in digital formats to support text-mining research methods. Text mining also results in a new dataset as a by-product which must be accessed/retrieved and stored for use by the researcher and possibly made accessible future researchers who may wish to duplicate or the research results.

Libraries have primarily focused on digitizing historical news content and they have, for the most part, made this content freely accessible to researchers. However, most present day news and social media content is primarily accessible through social media sites; news database vendors; or from the newspaper corporations themselves. Yet, librarians are beginning to understand the demand and need for news and social media content is broad and wide-ranging and the nature of requests being made of database vendors is becoming increasingly complex. For example, while some research may be focused on a single time period (19th century newspapers), a geographic area (British newspapers), or a specific news source (Twitter), other research may require specific content based on a specific subject (Tweets related to a specific present day election). In this context, the research need/question is as creative as the researchers who use this new research methodology. Tamm-Daniels notes: "understanding how much data needs storing and what kind of access to that data is required also shapes Big Data technology decisions." (Tamm-Daniels, 2013)

Many database vendors monitor the amount of data users are downloading from a newspaper database. Dave Magier (Magier, 2012) has described and argued the concept of "non-consumptive use" of news content should/may influence how we think about news content as it is used for projects using text mining. Although text mining projects result in very large datasets, Magier suggests text mined from news sources is not being used as it was originally intended or in its original form and format (to read a newspaper article, for example), but rather to develop more granular discovery, meaningful correlations, and trends which, with appropriate statistical and visualization tools/methods, may be revealed within the text. The suggestion is that text mined from say a newspaper, would not be restricted by copyright or licensing since it is no longer the original form, but is now data. Meanwhile, researchers want assurances they are able to "mine" content under library license agreements with database vendors.

Key questions:

- Who will provide the large amounts of newspaper content--will newspaper database vendors, newspaper websites, and social media sites open their archives researchers seeking to text mine their news content?
- If large amounts of text is mined from licensed or subscription sources, will Libraries (or researchers) be additional costs to the Library for this content?
- Will copyright law or licensing restrictions require the database vendors to go back to original information providers and determine if their licenses allow for this new form of use? Can such approvals be obtained in a timely manner? (Jockers, Sag, & Schulz, 2012)
- Will researchers be required to obtain a separate license/contract with the database vendor? Researcher? What would the Libraries role be in such negotiations/situations?

B. Storage and Software support

Presently most text mining is achieved through the use of API scripts, web crawlers, and similar approaches. Evidence suggests that authors would like more standardized way to access news content, especially content accessible through licensed databases. They also would like libraries and database vendors to develop focusing on improving and innovations related to the OCR technology. Dirty or incomplete OCR and data born-digital also create challenges for text mining researchers.

Researchers will need software or applications (apps) (for example, the *Google nGram* viewer) that would allow researches access, visualize or import the text data (similar to numeric data which can often be imported into common statistical software packages such as Microsoft Excel or SPSSx. Or, these tools should be built into database interfaces to quickly visualize the data. Will libraries be able to support the development of such tools or provide them, just as other companies are seeing potential opportunities? (Demos, 2013)

In addition researchers would like more support for the searching, retrieving, storing, and re-using these large text datasets mined from the database. Researchers require stable and reliable systems for retrieving and storing content and want some assurances that they will be allowed to retain access to content for extended periods, once it has been mined from the database. Such functionality/features would ensure that datasets can be updated and duplicated as needed, to support future research, or to allow other researchers to duplicate research findings.

Key questions:

- Who will create standardized tools to use across databases and vendors?
- How will digital text be enhanced and sanitized so that it can be searched effectively and with some confidence of consistent outcomes?
- What role can institutional repositories serve being created by many academic libraries serve in providing storage and access to large text mined datasets?

C. Reference, Instruction, and Training Needs

Research which relies on text mined from a wide variety of news sources presents significant challenges for researchers, but it also allows libraries the opportunity to think about ways they can and should support this this exciting and challenging research methodology. Libraries have already supported such research by providing a wealth of digital news content, freely available, and accessible. They continue to develop standards and oversee projects, many of which are funded by national libraries, national foundations; and state libraries. Academic libraries also provide licensed access to a wealth of digital archives and databases which are licensed for use by their students and faculty. What roles remain for libraries in providing reference and collection services?

In “Culturomics 2.0,” Kalev Leetaru (Leetaru, 2011) identifies several factors which influenced which sources were used to decide which databases would be mined. Libraries should expand their traditional research support roles helping researchers locate and identify what news content is needed; strengths and weakness of specific news sources; and whether the source can be used for such projects without additional permissions or payments. Librarians who have subject expertise can help researchers by helping them identify which databases will meet the needs of their research project or where they can obtain open access to key resources; help researchers anticipate challenges they will encounter with the content or database; advise on copyright/licensing process; and advise researchers about the underlying metadata and data structure.

Libraries are also partners in curating data for long-term retention, providing software programing and/or support for developing (or purchasing) specialized tools to use with the text data and provide robust institutional repositories which will ensure long term access to the data. Librarians may be considered partners in text-mining projects and may also provide analytics, data management or project management expertise. (Adams & Gunn, 2013)

Key Questions:

- Will libraries and librarians develop and expand existing services to support text-mining research?
- Will libraries be able to provide access to collections of news content and any appropriate curation, repository, or long-term access needs?
- Will librarians become partners in text-mining projects by developing research tools and other resources?

Conclusion

Text mining news content and sources presents new opportunities and challenges for libraries and librarians. On a larger scale, big data also introduces new challenges and provocations to how researchers use and apply these new research methods. (Boyd & Crawford, 2012) At the most basic level, Libraries will need to begin to think differently about licensing agreements; how to develop collections and provide access to a large corpus of news content; what their role will be in teaching and training researchers about the strengths and weaknesses of news databases and other resources; and how they can develop traditional roles of supporting location and finding information. How will libraries support the researchers, their students, and their long-term scholarly needs as more text-mining research methodologies are integrated into the scholarly community. The challenges are great; the opportunities are significant and libraries will be important partners in many big-data and text-mining research.

WORKS CITED

Adams, J. L., & Gunn, K. B. (2013, April). Keeping up with...digital humanities [internet publication]. *Keeping Up With...*, p. http://www.ala.org/acrl/publications/keeping_up_with/digital_humanities.

Arbesman, S. (2013, January 29). Stop hyping big data and start pyaing attention to 'long data'. *Wired [Internet]*, pp. http://www.wired.com/opinion/2013/01/forget-big-data-think-long-data/.

Berendt, B. (2010). Text mining for news and blogs analysis. In C. Sammut, & G. I. Webb, *Encyclopedia of Machine learning* (pp. 968-972). London: Springer.

Boyd, D., & Crawford, K. (2012, June). Critical questions for big data: provoctions for a cultural, techological, and scholarly phenomenon. *Information, Communication, & Society*, pp. 662-679.

Demos, T. (2013, April 4). Startup Tableau Software plots latest "big data' IPO. *Wall Stree Journal* , p. B4.

Harris, D. (2012, June 12). How Big data helps Ancestry.com map people, places and time. *Gigaom.com [website]*, pp. http://gigaom.com/2012/06/12/how-ancestry-com-is-using-big-data-to-map-time-place-and-people/.

Jockers, M., Sag, J., & Schulz, J. (2012, October 4). Don't let copyright block data mining. *Nature*, pp. 29-30.

Leetaru, K. H. (2011, September 5). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday [Internet Journal]*, p. http://firstmonday.org/ojs/index.php/fm/article/view/3663/3040.

Magier, D. (2012, November 13). Recent trends in text-mining and library services: research library perspective. *Text Mining: Opportunities and Challenges [Webinar]*. Chicago, IL: Center for Research Libraries.

Rosenbush, S., & Totty, M. (2013, March 11). How big data is changing the whole equation for business. *Wall Street Journal*, pp. R1-R2.

Tamm-Daniels, R. (2013, April). The Key to smart big data: know thy technology. *Information Today*, p. 19.

Zimmer, B. (2012, October 18). Bigger, better Ngrams: Brace yourself for the power of grammar. *The Atlantic [website]*, pp. http://www.theatlantic.com/technology/archive/2012/10/bigger-better-google-ngrams-brace-yourself-for-the-power-of-grammar/263487/.