Emerging Pathways: Supporting Text & Data Mining Business Projects

Joel B. Thornton, Head of Research & Instruction Services
Austin Wilkins, Data Services Specialist
University of Arkansas Libraries

Introduction

For business researchers, the need to obtain access to a large corpus of textual data for data mining and large-scale computational analysis has emerged as a significant need. Recently, the library at the University of Arkansas partnered with researchers to support text and data mining projects. Due to commercial vendor restrictions limiting the amount of content that researchers can download from their platforms, researchers have turned to the library to find solutions to their text and data mining issues. This article will highlight a recent text and data mining project that the University of Arkansas library supported in partnership with a vendor.

The Project

In July 2019, a faculty member from the business college sought our expertise for assistance with a text and data mining project. Driven by an article revision deadline, the project timeline was two months. The faculty member needed full-text articles (2001 to 2018) from select sources, i.e., *The New York Times*, *Financial Times*, and *Fortune*, discussing any of the more than 2,200 large publicly traded U.S. firms.

The researcher was able to find the relevant articles for a given firm using the search functionality of the library databases. However, given the numerous firms, sources, and dates desired, efficiently collecting the articles proved difficult due to daily download restrictions.

Project Support

We contacted our vendors to find a solution to this researcher's data need and were informed of a pilot program that would allow him to analyze a large corpus of documents on the vendor's servers. This solution bypassed the download limits while also allowing for more computationally intensive analyses to be run on cloud servers.

We determined that the researcher had a perfect use case for this pilot program. Thus we contacted him for an onboarding discussion. He possessed advanced Python coding skills and expertise in text and computational analysis. We began working with a vendor to develop a corpus of documents that exceeded 1 million news articles. This study had been conducted previously in the accounting field, but only by hand, and only to approximately 800 documents at maximum. This researcher planned to complete the project as a proof of concept in using programmatic tools in this analysis style while

completing a project many orders of magnitude larger than similar projects in the accounting field.

Our project involvement led to an ongoing relationship with the researcher. The immediate issues were navigating the hosted programming environment, downloading and processing the text data, and handling data tagging issues in the provided corpus of documents. We kept regular communication with the researcher to ensure that issues were discussed and solved quickly, and provided sample scripts, walkthroughs, and advice on processing the large dataset.

We discovered that friendly and informal communications helped increase the connection between the researcher and our support team during this process. While the project was ongoing, we communicated regularly with him, discussing opportunities to showcase the research, supporting issues with the analysis, and providing an interested and helpful ear for any problems that may have developed.

Outcomes

The researcher completed the project with minimal difficulties, analyzing documents on a scale many orders of magnitude larger than previously seen in his field. The final project analyzed 1.2 million documents between 2001 and 2018, using a combination of word search and sentiment analysis. The analysis provided evidence to support the hypothesis while also producing additional insights that may be used in future research. This researcher has since presented his work in several workshops and presentations and is currently in the final publishing phase. Overall, we identified several factors that contributed to the success of this program: (1) the technical expertise of the researcher, (2) a clear and outlined research plan, and (3) the regular communications between the researcher and our library partners.

Conclusion

With the emergence of researcher expectations, tools, and computational analysis capabilities, the demand to support researchers with text and data mining projects will increase. Academic libraries must develop robust, scalable services, which requires building extensive partnerships with content providers to assist researchers with complex projects. This article highlights one approach to supporting text and data mining business projects. Additionally, this project exhibits the value of possessing in-house expertise in modern analytical techniques to support local researchers.