Using Digitized Newspapers For Text Analysis In Notebooks

Art Rhyno, Systems Librarian @ University of Windsor, Board Member @ OurDigitalWorld

Introduction

New directions for digital services emerge from positioning large text collections to work with technologies fostered by advances in artificial intelligence and machine learning. This combination is made more accessible by a current trend towards using Jupyter notebooks to access research computing. The Mellon-funded Collections as Data initiative has fostered data literacy and computational analysis using digitized text with Jupyter notebooks, and the COVID-19 Open Research Dataset (CORD-19) has been used with natural language processing to generate new insights in understanding COVID-19. For digitized newspapers, efforts such as the University of Arizona's Newspapers as Data project have demonstrated that Jupyter can enable text data mining and other types of text processing with minimal technical requirements.

Why Jupyter Notebooks?

"Avoid installation woes (and dependency hell) for students (this is hard enough in person)..."

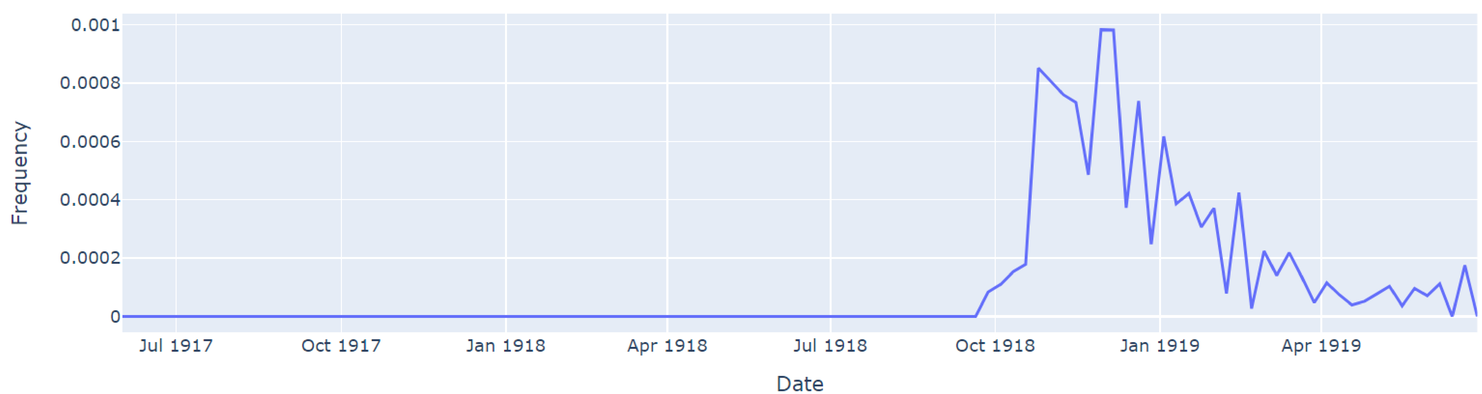
as quoted in Using Newspapers as Data for Collaborative Pedagogy: A Multidisciplinary Interrogation of Borderland. CNI Fall 2020 Virtual Membership Meeting.

Jupyter Notebooks are web-based and run in the browser. They do not require special software to be downloaded or installed. Notebooks also encourage the use of a narrative, rather than presenting just code (and comments).

Jupyter is becoming an accessible entry point for research computing in Canada. The Digital Research Alliance of Canada (formerly Compute Canada), Compute Ontario, and the 2i2c JupyterHub project at the University of Toronto, and other initiatives are adopting a Jupyter approach to optimize shared research computing resources. At the same time, Jupyter has the potential to enhance Research Data Management (RDM) best practices, such as helping to ensure the reproducibility of data and code.

There are also freely available and powerful Jupyter implementations via Binder (which can be integrated into GitHub) and Google Colab (which also works with GitHub in addition to Google Drive).

Example – Spanish Flu in small-town Ontario



Occurrence of the term "influenza" in text of the *Amherstburg Echo*, mid-1917 to mid-1919. Notebook (and data) available on OurDigitalWorld GitHub repository.

This is a very simple example of using digitized text from a newspaper corpus. By creating a Jupyter notebook that can be run interactively, the results can be recreated, and the parameters can be altered (perhaps narrowing in on a smaller data range, for example). Newspaper often represent the most extensive textual record of a community's past, textual analysis and data mining techniques can be used to investigate social trends and other areas of interest in new ways.

Newspaper Digitization

Almost every community in Ontario has had a newspaper covering the events in its past. This unique record of community history is often locked away in microfilm or other dated and difficult formats.

OurDigitalWorld, and its predecessor, OurOntario, has worked with the Leddy Library at the University of Windsor for over a decade to surface newspapers from across the province (and beyond). We have researched Optical Character Recognition (OCR) options for maximizing the text in newspapers resources, built scanner prototypes for providing low-cost alternatives for scanning microform sources (especially microfilm and microfiche), and used Hadoop and other types of distributed processing to tackle the tremendous volume of material that newspapers can represent.

Next Steps

Join us for the upcoming Compute Ontario hybrid workshop series on newspaper text analysis later this month (February 2023).

Talk to us at the OurDigitalWorld booth on the conference floor (or reach out to us virtually).

Acknowledgements: Thanks to Compute Ontario for a Research Data Management grant to help support Notebook research in this area, and to Jess Posgate, Project Manager at OurDigitalWorld, for keeping everything on track.