Digitizing Rubenstein Library's Subject Card Catalog

Lade Laniyan and Margaret Wolfe
Duke University

Abstract

In 2012, the Rubenstein Library at Duke University destroyed and digitized their card catalog collection, which had been hand-written and typed between the late 1940s and the early 2000s. As of May 2023, the subject file cards were available only by request and in the format of over 100,000 individual JPEG files. The goal of this Data+ team was to create an easily searchable and sortable dataset from these files in order to make them more readily available to researchers as well as more easily accessible for library staff to perform analysis. The team's work is in collaboration with the Rubenstein Library and their broader initiative to find and describe historically marginalized voices within their collection.

Objectives

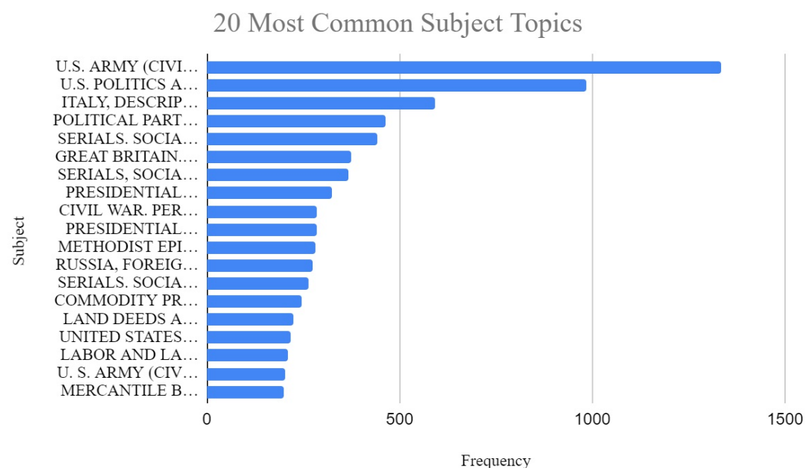
- Make the digitized cards more readily available to researchers by uploading them to the internet archive
- Create a centralized dataset containing important data from the cards, including date, location, and subject
- Clean this dataset and use it to perform analysis
- If possible, cross reference the subject cards with their main entry counterparts analyzed in a 2021 Data+ team

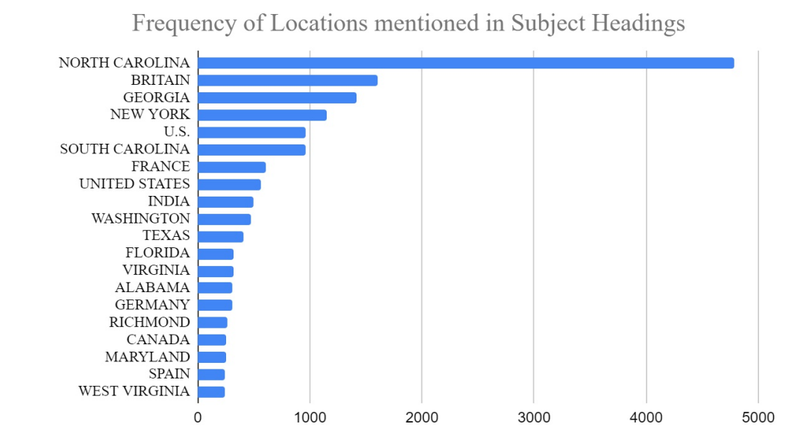
Methods

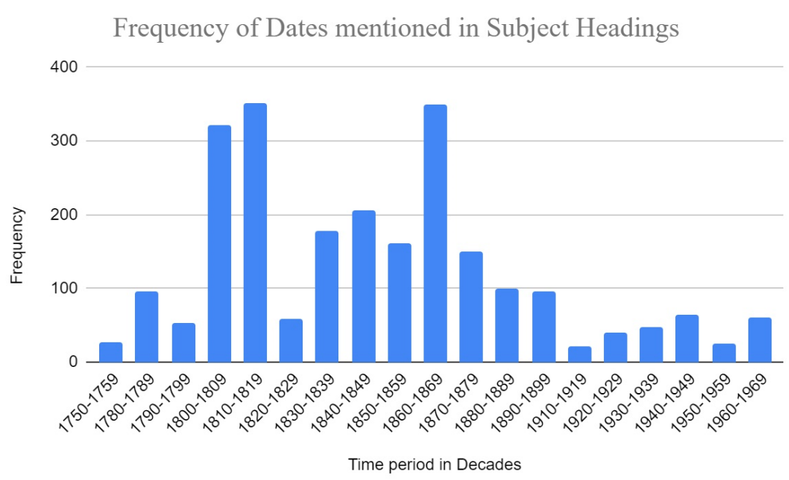
We primarily used the Python coding language throughout the project. Packages used included tesseract in order to perform an OCR conversion of digitized subject file JPEGs into TXT files; spaCy in order to convert this 5GB txt file into a comprehensive dataset; Pandas in order to clean this dataset and easily perform analysis; the internetarchive package and API in order to upload to archive.org; and finally csv, re, and nltk to perform final analysis of our dataset. We used a github repository to store our code, which is publicly available both for future projects and for later review.

Results

In addition to making the digitized cards more easily accessible using the Internet Archive, where we created a collection for the cards to reside within the Rubenstein Library's account, we created a comprehensive dataset of over 159,000 subject cards. This dataset created a space which is easily searchable by subject, location, and date, in addition to being compiled using a common format and location. We performed an analysis of these searchable topics in order to visualize the most common topics covered in the Duke archival collections. In order to do this, we had to condense and additionally clean the dataset from its original format, since the cards were handwritten and typed and did not contain consistent wording or format. We confirmed a lack of BIPOC and female subjects, and found extensive collections on both the U.S. and the C.S.A. militaries and the Civil War.







Conclusion

We encountered many challenges throughout our project, specifically because we were working in a format with which we were unfamiliar. The cards' lack of a consistent typography or format posed challenges from the beginning, as we went through the process of trial and error in order to OCR the JPEG files into TXT files. The data we obtained from this process had to be exceptionally clean in order to create an easily searchable database on both Internet Archive and within our own dataset. We did succeed in creating a clean dataset of over 159,000 rows, but this process took extensive time and effort. We performed cursory analysis using this dataset, mostly involving frequency of subjects, dates, and locations, the three most commonly mentioned points of data in the cards. However, in the event of future project teams or work on the subject file cards, a more extensive analysis could be performed, specifically within the most common subject types and locations. Duke University has an extensive archival collection of Civil War and Confederate documents, and an analysis of this subset of cards and their origins would be valuable. In addition, an analysis of the language of the cards using machine learning and natural language processing would be beneficial in order to visualize the use of arcane language and phrasing based upon the dates the cards were created and the subjects they involve. Finally, if given additional time we would liked to have performed a cross-analysis with the data from the 2021 Data+ project team.

References

- https://archive.org/
- https://library.duke.edu/rubenstein
- https://digitalcommons.unf.edu/cgi/viewcontent.cgi?article=1018&context=bliss
- https://realpython.com/python-data-cleaning-numpy-pandas/