Exploring Data Curation: Tools, Techniques, and AI in Summer Internship at NCDS

During the summer of 2024, I had the opportunity to intern with the NNLM National Center for Data Services (NCDS) internship in partnership with the Data Curation Network (DCN). Working closely with Mikala and Shawna, along with two other interns, along with two other interns, we focused on cleaning, analyzing, and visualizing data related to data curator job postings and trends over time. Our project utilized the data pulled from the International Association for Social Science Information Service and Technology (IASSIST) job repository, containing job posts from 2005 through April 2024.

The goal of the internship was to explore data curation related job trends over this time, while gaining hands-on experience with various tools used in data analysis and curation. Over the course of the summer, my fellow interns and I learned new programs and tools, such as Tableau, Akkio, and Voyant, while enhancing my skills in familiar tools such as Excel and OpenRefine.

OpenRefine

OpenRefine is a program for cleaning and transforming data. One of the most useful functions I discovered in OpenRefine is its clustering function, which helps normalize data by grouping together similar terms and fixing inconsistencies such as misspellings or case sensitivity. The facet function is also immensely helpful since, once the data is imported, faceting can help focus on viewing and editing separate columns at the same time.

Tableau

One of the reasons I was eager to start this internship was to learn how to visualize data. I took a data organization course in my master's program and one of the assignments was to visualize data of our choice (based on some guidelines). I used Microsoft Excel to do this, but no matter how much I edited visualizations made in Excel, I did not think they turned out as well as it could have. This drove me to figure out best practices or tips and tricks for visualization so I could improve.

Although Tableau looked like an interesting alternative to Excel to me, it seemed daunting to use during the NCDS demonstration. One of the best ways to learn is to experience it yourself, so after the demonstration I spent time testing how to use the program. I thought I had the hang of it at one point, but when I wanted to visualize different data points, I realized I needed to have my dataset edited differently which, to my knowledge, you cannot do in Tableau itself. Instead, you must edit your spreadsheet outside of the program and import this new file back in. I sometimes found this a bit inconvenient, but the resulting visualizations were more polished than anything I had previously created in Excel.

Akkio

Before this internship, I hadn't given much thought to the role of AI in data curation, aside from its rising prominence in discussions around art, writing, and information retrieval. I was curious to see how AI could be applied to the more technical aspects of our project. My fellow intern Kimberly covered this in a previous blog post, so I will focus on AI with regards to my project.

We utilized an AI tool, Akkio, as an experiment, just to test how it would manipulate and possibly curate the dataset, which I found interesting and had not considered as something an AI tool could do. I enjoyed trying to determine how to phrase my prompts to ensure the resulting transformation was what I had in mind. I liked comparing what was found when using OpenRefine to what I found through Akkio transformations. I appreciated using Akkio, as I did not know the extent of what AI programs could do; of course, there are so many debates over the use of AI, so I am curious to learn more about the two sides and the possible solutions.

Voyant

Voyant, a text mining program, was the last tool we learned during the internship. After uploading or copying text into the program, different windows appear showing various analyses of the data. With Voyant, I was able to use the Contexts feature to see what words preceded and followed a specific word or phrase. For example, I focused on seeing what surrounded words related to *preferred* and *required qualifications*. Though I didn't have much time to explore all of Voyant's functionalities, the experience showed me how text mining can be a powerful tool for analyzing patterns in large bodies of text, which is essential when working with datasets that include qualitative information.

Conclusion

Reflecting on my summer internship, I found it interesting to see what skills, qualifications, and experience employers preferred or required of potential applicants in research data or data curation roles. Prior to this experience, I assumed that any jobs involving working with data meant needing to be an expert in coding or statistics, but learning these tools and other data curator duties over the summer made data curation and data librarianship seem like viable career paths.