Erica Zhang

LIS452

5/11/17

Final Project Writeup: Visualization of Library Statistics for The Career Center[[1]](#footnote-1)

For my final project, I chose to write a program that would read data about The Career Center’s circulation transactions, and create some graphs that would help us understand how that activity has fluctuated over time, and perhaps identify any interesting trends or patterns that we could consider as we work to better promote our print collection. Specifically, this program takes in a csv file with the following columns: type, datetime, itemcallnumber, title, and subjects. With this data, particularly the datetime and subjects columns, I can then create tables and graphs to visualize the following:

* number of circulation transactions over this time period by date
* number of circulation transactions over this time period grouped by month
* number of circulation transactions over this time period grouped by year
* volume of transactions of overall for a given month
* top 20 most popular subjects

To write out the code for this project, I decided to borrow from Shub’s lecture before Spring Break, and try out Jupyter Notebook, since I thought that being able to have an inline plotting tool would be very helpful, as well as the easy incorporation of comments amidst code. I also wanted to see if I could learn some basics about the Pandas library, which made our finance assignment a lot more digestible. My general strategy for each graph was creating a dictionary with dates as keys and the count of those dates as values, and then converting the dictionary into a pandas series that would then allow me to create a graph using pyplot.

Though Shub was able to accomplish my project with just a few lines of code in about 10 minutes, I found a number of aspects challenging. Understanding the Pandas documentation was definitely an obstacle. Similar to reading Python documentation, it was hard to understand both the syntax of any particular function or concept, and the way it ran, without any examples. Stackoverflow was fairly helpful in this regard, as were a number of “data science for beginners” tutorials.

Similarly, another challenge in putting together this program was understanding data types, both with respect to Pandas, and just in general. Every time I used something new in Pandas, I had to check the type of the result to understand what I could do with it next, or if I had to then try and convert it again to the data type that I was expecting to work with. Understanding the difference between series and dataframes was really key to helping me understand what parameters or methods I could use. This seemed to get especially complicated with date data types. I tried using date types to utilize the handy conveniences that come with using dates, but I think there were other times where it was just easier to stay with string formats, such as my “transactions over time by year” plot. When I first plotted the graph, the x-axis labels were returned in scientific notation. As hard as I tried to change the strings into dates, in the end, it was a lot easier to just modify the labels to turn off scientific notation.

Finally, for the subjects popularity portion, I found that data cleaning was really essential to getting a good visualization. The first challenge was trying to extract the subjects column and converting it into a usable list, which involved moving it from a list of lists to a master list. Then I realized that there were a lot of empty strings, NaN values, and strings preceded by whitespaces, so I had to figure out how to make all of the terms consistent, so that “job hunting” was counted with “Job hunting” and “ job hunting”.

Overall, I thought this project offered a good mix between using concepts we learned in class and trying to incorporate something new that took awhile to learn. In terms of application, I hope that these graphs can provide some insight into the way students use the Career Center’s library. For instance, there are a number of days that have significantly more activity than the rest, so I may go back and investigate those days. In addition, who knew October was such a popular month for circulation activity?

On the other hand, there are a lot of issues with these graphs that I hope with more practice and experience, I can eventually resolve. The graphs themselves could do with more helpful labels, both on the x and y axis, title, and legends, but also at key data points, such as those sharp spikes in the date graph. I also realize that it may take a more thoughtful approach to the subjects popularity to really gain any insight into popular subjects, since subjects are both very general, like “job hunting”, or incredibly specific, such as “resumes (employment) retired military personnel employment united states”. This graph could thus spark conversation about how we can improve upon our current copy cataloging practices to have better keywords, not just to make students’ searching easier, but also our data analysis. Something to keep me practicing Python in the semesters to come…

1. All files can be found in Moodle as well as my Github repository: <https://github.com/ezhang15/lis452>. Final deliverables include: koha\_initial\_data.csv, Zhang\_Erica\_LIS452\_Final\_Project.py, Zhang\_Erica\_Final\_Writeup.docx, Circulation Statistics for The Career Center Resource Library.pdf. [↑](#footnote-ref-1)