Osman Kpaka, David Xiong, Daniel Zhang
INLS 490 - North Carolina Gazetteer
Spring 2022

Final Report

Documentation of CSV file

We have added the following columns: URL_US_News, inception, longitude, and latitude. URL_US_News contains the urls linking with U.S. News website for each school. Inception contains the year the program or the school was launched. Longitude and latitude are the geographic coordinates of the school.

Potential data values for accreditation status are SACS, SACS and State, and State. SACS accreditation is approval of a higher education institution by the Southern Association of Colleges and Schools (SACS). State approval to operate signifies that institutions have satisfied certain minimum requirements established by a state.

Locale type desc indicates the type of geographic area where a school is located. Potential values for it are City, Large; City, Midsize; City, Small; Rural, Distantt; Rural, Fringe; Rural, Remote; Suburb, Large; Suburb, Midsize; Suburb, Small; Town, Distant; Town, Fringe; and Town, Remote.

Comma

Stores tabular data and best viewed through a spreadsheet program, we can have multiple data and nodes and helps update data. We can view a CSV File as a plain text in the editor.Each line of the file is a data record and stored in one line of the text file. T

Documentation of the resulting RDF

Resource description Framework, is the standard for modeling data it is flexible and simple yet structured. It uses a subject predicate and object called triples. Iri's eliminate ambiguity when data comes from different sources. We used the rdf to to filter down data into sizeable amount of information. We then compared this information with other information so that users will have access to the data with the best schools fit for their child.

Documentation of research process

To build up our dataset, we first found a huge database of all schools in North Carolina from the North Carolina Department of Public Instruction. The database consists of tons of properties of every single school in North Carolina. Luckily, the database has its own filter function, so we only keep schools whose current grade level contains high school level, which is 09-10-11-12. After searching other two or three databases of NC high schools, we found this one is actually the most detailed and satisfactory database we could find, so we decided to use it as the basic dataset. As there are so many irrelevant and confusing properties that make the dataset too redundant, we use both Excel and Openrefine to get rid of columns that seem unnecessary and combine some columns together.

The next step was adding more properties to make our dataset ample. As learnt from class, we used Openrefine to link the dataset with Wikidata and get each school's inception. Then we used an add-on tool on Google Sheet called Geocode by Awesome Table to create each school's coordinate according to the documented address. Besides these additional columns, we shifted our focus to the main objective, which is helping middle school, high school students and their parents to choose school. Usually do people choose a school based on the school's average test score, student:teacher ratio, admission rate, etc. While these data change every year, and properties in a gazetteer are supposed to be stable, at least not changing every year. Therefore, we found US News, a really good resource for people to choose a school; it contains the school's state ranking, total employment, test proficiency, student diversity and so on. More importantly, these data are kept updated at least annually. To ensure our gazetteer users can easily access US News website, we made a python crawler app, its basic running logic is to search the school's name on US News search engine one by one to get to the page of each school on US News website and copy and paste the url to the dataset. Another useful resource we found is the North Carolina School Report Cards, it also has detailed information about each school. However, since it is a new database, there are a limited number of schools documented, so we decided not to refer to it for now, instead we can wait until there is enough number of schools included in the database.

Future works

We can link our dataset with north carolina school report cards from North Carolina Department of Public Instruction. For example, the link of the report card for Apex High is https://ncreports.ondemand.sas.com/src/school?school=920316&year=2021&lng=en. The only changing part is school=920316. This number is the school id we have in our dataset. It should be easy to incorporate school report cards into our dataset. The problem is that there are more than 400 schools available with report cards while we have more than 800 schools in our dataset. We still need to figure out which schools in our dataset contain report cards.

Also, since US News adds and updates schools into their database very frequently, it would be good for us to run the python crawler once a month to make sure all the urls are new.

Credit and acknowledgements

Daniel (Hao) Zhang:

I filtered data from North Carolina Department of Public Instruction and used OpenRefine to further clean the data, combining several columns together, deleting unnecessary columns, and using reconciliation to get the inception year and others from wikidata. I also found this dataset called North Carolina school report cards, ready to incorporate into our dataset.

David (Haolin) Xiong:

I processed the data after downloading it from the database, including rearranging the columns to make the dataset look more logical, deleting redundant columns, and adding new columns like full address. I also wrote the python crawler program to help us get school urls on the US News website, and used Geocode to get each school's coordinate.

Osman Kpaka:

I refined the data from our database and updated the columns of information. Additionally, I kept track of redundant information to not display on our database and make our interface more user-friendly. This allows our user-face to be accessible to a multiude of individuals.