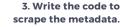
SEARCHING FOR DATA

Searching for data can often be cumbersome. By searching with more meaningful variables, you can often get better datasets for teaching concepts, training machine learning models, and more.

1. Brainstorm fields to include.



This can include author, author affiliation, repository size, etc.

Make sure variables are meaningful and are descriptive about the datasets

I wrote this code with functions in R that accept the webpage URL as a parameter and returns the scraped metadata attributes for a dataset. After successfully obtaining metadata for a single dataset, write the code to iterate over all datasets in a repository. This can be completed with a nested for-loop that scrapes the links for the dataset pages before the nested loop.

5. Clean the metadata

As explained before, we decided to write a function that "merges and drops" a certain pair of columns.

This is often two columns that contain similar contents. If two columns are virtually identical, cells of those two columns cannot both be occupied. Thus, one should shift the contents of one column over to the other column, if space is available.

Lastly, drop the original column.

Identify repositories you can find your targeted fields. Organize it into a data frame to aid in the iteration process.

Examples include Kaggle, Zenodo, and Harvard Dataverse.

2. Find repositories to obtain the metadata

It is now time to merge your work into one dataset. I wrote a function that would contain every variable from every repository and bind them into one dataset. This can be a temporary solution but the data will need to be cleaned moving forward since it will contain redundant columns.

4. Combine metadata from every repository..

Once you are done cleaning and curating your metadata, you are ready to share it. This can be uploading to an existing repository or creating your own. Be sure to incorporate the user experience well when designing a repository that searches through

6. Share your metadatayour data.

