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 data in the

observations.

3. Write the code to scrape the metadata.

We 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.

After you are done iterating over many repositories of metadata,, it is now time to merge your work into one dataset. We 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.

5. Clean the metadata

Since the original metadata contains redundant columns, we decided to write a function that "merges and drops" a certain pair of columns. This is oftentimes 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 contents of one column over to the other column, if a space is available. Lastly, drop the original column. This task can be accomplished by writing a function.

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

your data.

6. Share your metadata.

METADATA

After you have identified repositories that contain your desired fields, write code to scrape the data from the webpage. Organize this data into a data frame to aid in the iteration process.

This can often be a simple Google search if you are not very familiar with various metadata repositories. Some examples include Kaggle, Zenodo, and Harvard

Dataverse.

2. Find repositories to obtain the metadata

4. Combine metadata from every repository..