


Annotator instructions for paper-level data science extraction

The goal of the annotation is to extract from the paper the **instructions** that are implicitly given for recreating the dataset that the paper uses for its analysis.

What to annotate:

- Paper Title
- Paper Abstract
- Links to any code or data that is associated with the paper
- Any datasets used in the paper
 - Links to the datasets
 - Include any rationale for why these datasets were chosen, if they are given
 - Include links to the datasets if publicly available.
 - References for any private datasets
 - Variables that are used in each of the datasets. Be thorough and descriptive
 - For example, if data includes all tweets that include a set of keywords between two dates, explain this with enough detail that the dataset could be reproduced if we had access to the raw tweets.
 - Include rationale for why particular variables were chosen, if they are given.
 - If the dataset is doing geospatial inference, specify the spatial granularity of the data:
 - For example: County vs State vs Country vs Census Block.
 - Make sure to specify all locations with enough detail that the dataset can be exactly reproduced. So if a dataset includes data from 196 counties in the US,, please give explicitly the names of all of the counties.
 - If the dataset does *temporal analysis*, ie uses datasets that vary in time, specify the time range and the temporal granularity of the data
 - For example "weeks between January 1 2020 and January 1, 2022"

- If the dataset carries out *transformations* of datasets then please list them explicitly with enough granularity that the work could be reproduced. For example *the most recent values from Google Searches are available up to 36 hours before the current date...thus, for Google searches, data reported at time t were shifted to time $t + 2$ to address the 36-hour delay. ..or epidemiological reports suffer from backfilling and reporting delays due to postprocessing so epidemiological data reported at time t were shifted to time $t + 7$.*
- Summary of any data cleaning procedures or any ways of eliminating bias in the data.
 - Copy directly the text of the paper, and note the section of the paper where this discussion occurs
 - The goal is to have enough detail to reproduce the paper. Thus it is better to note leave out details.
- Open source code repositories used to do the analysis of the paper

 [10-5-23]Annotation_Data_Science_Paper_Example