Web Scraping of Citations

Goal: collect citation data of entries on Philpapers.org

Data: the data will be network edges between the citing and cited articles, containing Author, Title, Date of Publication, and Publisher (Journal) data for both the citing and cited articles. The cited article data should be collected on PhilPapers, on the pages for each individual article. The citing data will be collected from Google Scholar, on the Cited By pages for each of the articles from PhilPapers.

Web Pages: To show what I’m seeking, here is a step-by-step guide to the pages you’ll need to access to collect the data.

I’m looking for all and only the citations of articles archived by PhilPapers

<https://philpapers.org/>

On that page, you’ll see Browse by Topic in the middle of the page. Under that, you can see several topics. To access individual articles, you need to go to more specific topics than those listed on the main page. To get started, let’s work with an example, Metaphysics (found under Metaphysics and Epistemology). Click on that link. That will take you to this page:

<https://philpapers.org/browse/metaphysics>

This is still too general. You’ll see subcategories. Click on Theories of Causation, under Causation.

That takes you here:

<https://philpapers.org/browse/theories-of-causation>

Categories at this level of specificity give a list of all the articles categorized under that label. Sorting by impact, the top item I see is Causality: Models, Reasoning, and Inference, at this address:

<https://philpapers.org/rec/PEACMR>

I sorted by impact just to ensure that the example had citations. Not all entries will have any citations.

On this page, you’ll see a Find it on Scholar link in the Options, midway down the page. That will take you to a Google Scholar page:

<https://scholar.google.com/scholar?hl=en&lr=&q=Causality+author%3APearl>

The first entry is the correct one. This is only the case around 97% of the time, but I’m ok with some false positives in the data, so it’s fine to assume that the first entry is the correct one every time. On this link, you’ll see a Cited by 12194 link (the exact number may have changed). Clickong on that link, you get a list of all the articles that have cited the original one.

<https://scholar.google.com/scholar?cites=10996260119229499611&as_sdt=400005&sciodt=0,14&hl=en>

Ok, so those are all of the relevant web pages. On the last link, the list of articles citing the original one, you’ll see author, title, date of publication, and publisher info for each of those 12,194 items, though only a dozen or so per page. I need all that data for each of the entries. Most entries have 0-10 citations, this one that I’ve picked is on the extreme high end. What I’m seeking is a complete list of articles cited and which articles cited them, along with the author, title, date of publication, and publisher info for both the cited and citing articles.