Web scraping for research

a brief introduction

Jason T. Kiley, Oklahoma State University
Overview

• What is web scraping?
• When is web scraping most effective?
• What is the process for web scraping data?
• What are the best practices?
Web scraping defined

• Web scraping is simply extracting usable data from web pages.

• Semi-structured data: there is a structure, but the computer needs us to help see it.
When to scrape (or not)

- If a site has an application programming interface (API) that works for you, use it.
- If not, consider the number of sources and the amount of data.
- Web scraping has very high fixed costs (per source) and relatively low variable costs.

<table>
<thead>
<tr>
<th></th>
<th>One (few) sources</th>
<th>Many sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small data</td>
<td>Manual(-ish) gathering</td>
<td>Manual(-ish) gathering</td>
</tr>
<tr>
<td>Big data</td>
<td>Web scraping (if no API)</td>
<td>Look for a database or aggregator</td>
</tr>
</tbody>
</table>
Web scraping
the process

1. **Identify content**
2. **Pilot test**
3. **Retrieve content**
4. **Process content**
5. **Merge w/ archival data**
Best practices
and hard-won lessons

• Be cool: it’s fairly easy to block web scraping, but most sites only block aggressive scraping (e.g., too fast; too many connections, not targeted). I like to request no more than one page every 10-20 seconds.

• Pilot study: do it. This is a heavy lift in terms of project management, and you’ll save time later by proving that your process works now.

• Retrieve, then process: if you download all of the pages, you can simply reprocess them if (i.e. when) you need to change something.

• Filter first: many sites have pages with links that lead to the full pages that you want. Use the link page data to filter down to only what you need.
See you in the Q&A!