R CASE STUDY: WEB SCRAPING

John R Little

Duke University

DUKE UNIVERSITY: LAND ACKNOWLEDGEMENT

I want to take a moment to honor the land in Durham, NC. Duke University sits on the ancestral lands of the

DEMONSTRATION GOALS

- Building on earlier Rfun workshops
- Web scraping is fundamentally a deconstruction process
- Introduce just enough HTML/CSS
- Introduce the `library(rvest)` package for harvesting websites/HTML
- Tidyverse iteration with `purrr::map` - Point out useful documentation & resources

This is a demonstration of leveraging the Tidyverse. This is not a research design or HTML design

CAVEATS

- You will be as successful as the web author(s) were consistent
- Read and follow the *Terms of Use* for any target web host
- Read and honor the host's `robots.txt` | `https://www.robotstxt.org`
- Always **pause** to avoid the perception of a *Denial of Service* (DOS) attack

SCRAPING

Step one: **Gather**

ingest web page data for analysis

`rvest::read_html()`

Step two: **Crawling**

systematically (iterating) through a website, gathering data from more than one page (URL)

`purrr::map()`

Step three: **Parsing**

Separating the syntactic elements of a web page into meaningful data

```
rvest::html_nodes()
rvest::html_text()
rvest::html_attr()
```

HTML

Hypter Text Markup Language

```
<html>
  <body>

    <h1>My First Heading</h1>
    <p>My first paragraph contains a
    <a href="https://www.w3schools.com">link</a> to
    W3schools.com
    </p>

  </body>
</html>
```

HTML + CSS

Cascading Style Sheets

```
<html>
<body>

  <div class="abc"> ... </div>

  <div id="xyz">
    <span class="foo"> ... </span>
  </div>

  <span id="bar"> ... </span>

</body>
</html>
```

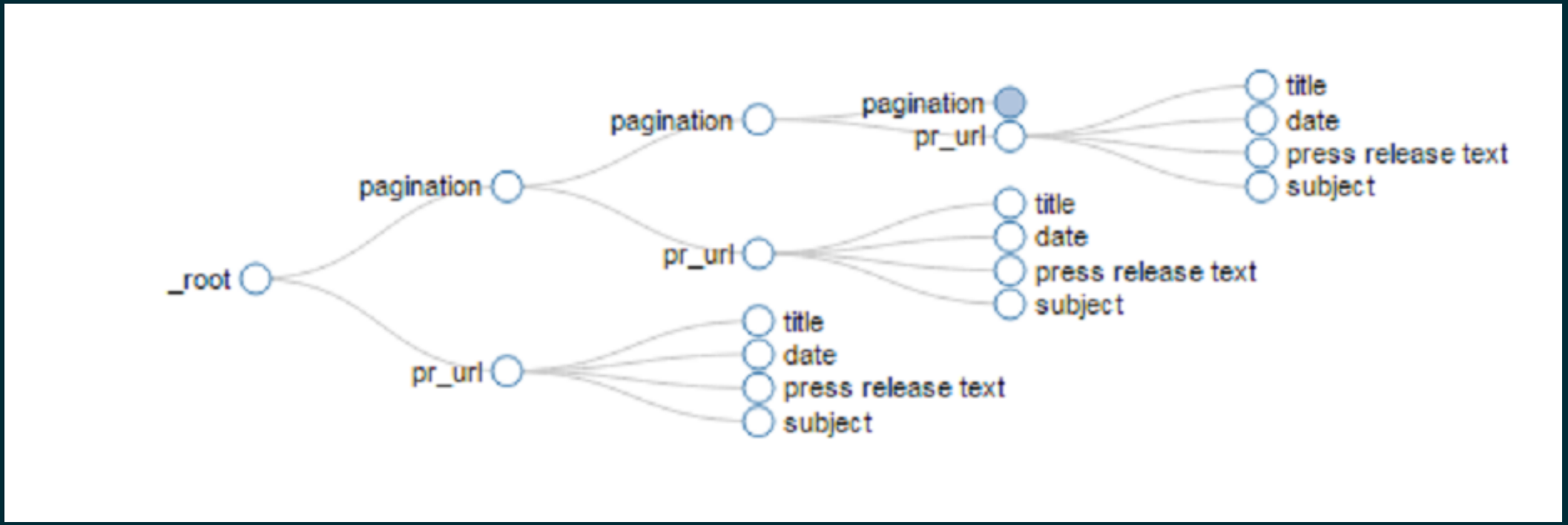
for example: https://www.vondel.humanities.uva.nl/style.css

PROCEDURE

The basic workflow of web scraping is

1. Development
    - Import raw HTML of a single target page (page detail: a lea
    - Parse the HTML of the test page and gather specific data
    - Check *robots.txt* and *Terms Of Use* (TOU)
    - In a web browser, manually browse and understand the tar
    - *Parse* the site navigation and develop an *iteration* plan
    - *Iterate*: orchestrate/automate page crawling
    - Perform a dry run with a limited subset of the target web si
    - Construct pauses: avoid the posture of a DNS attack
2. Production
    - Iterate/Crawl the site (navigation: branches)

SITE TREE


_root
pagination
pagination
pagination
pr_url
title
date
press release text
subject
pr_url
title
date
press release text
subject
pr_url
title
date
press release text
subject