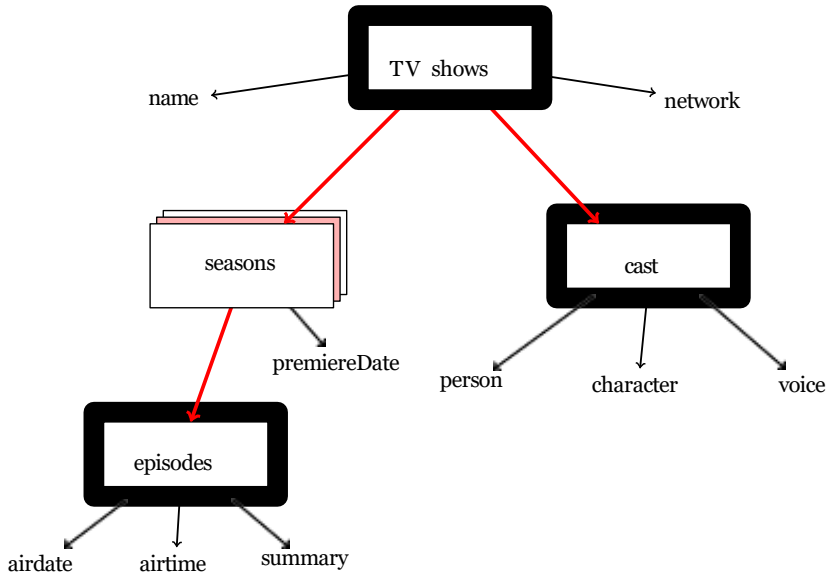
Lecture 18
Web Scraping

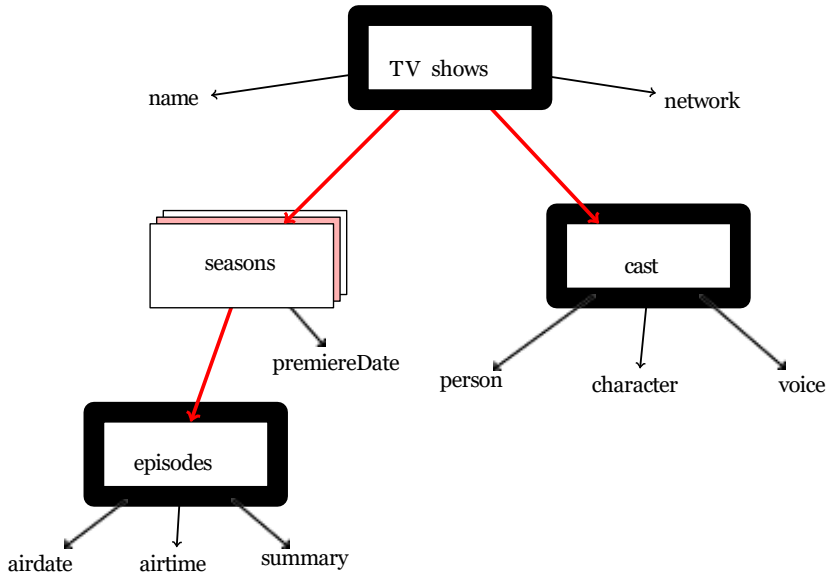
1. Recap
2. HTML Crash Course
3. Web Scraping
4. Ethics of Web Scraping

Hierarchical Data



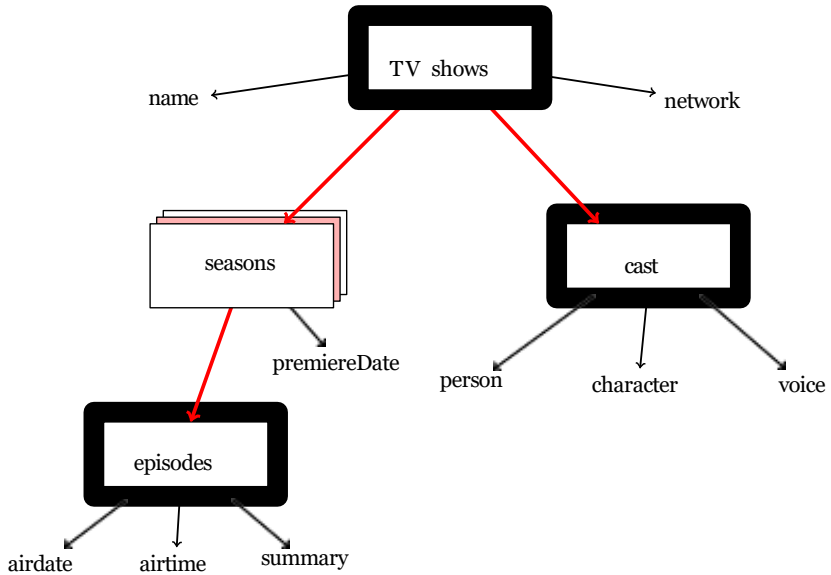
- Hierarchical data can be represented using JSON or XML.

Hierarchical Data



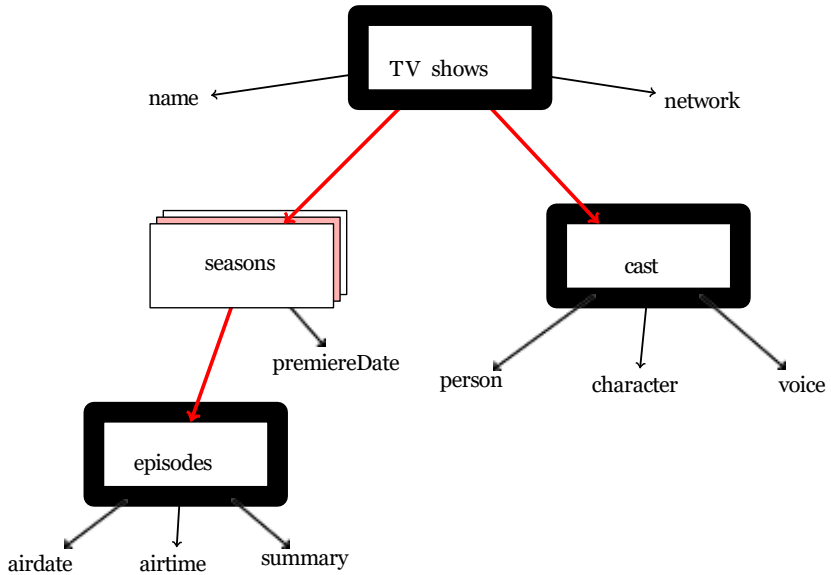
- Hierarchical data can be represented using JSON or XML.
- JSON is just like a Python dictionary.

Hierarchical Data



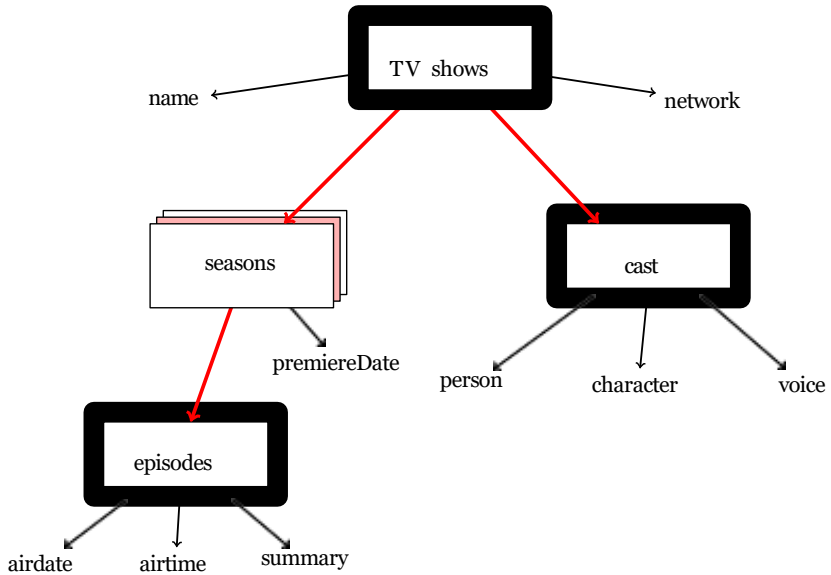
- Hierarchical data can be represented using JSON or XML.
- JSON is just like a Python dictionary.
  - You can use basic Python to extract the information you want.

Hierarchical Data



- Hierarchical data can be represented using JSON or XML.
- JSON is just like a Python dictionary.
  - You can use basic Python to extract the information you want.
  - There are built-in functions like `pd.json_normalize` to "flatten" JSON to tabular data.

Hierarchical Data



- Hierarchical data can be represented using JSON or XML.
- JSON is just like a Python dictionary.
  - You can use basic Python to extract the information you want.
  - There are built-in functions like `pd.json_normalize` to "flatten" JSON to tabular data.
- XML is a different beast.

XML

- Fields are represented by named *tags*.
- Each tag has an open <tag> and a close </tag>.
- Children are represented by nested tags.
- Repeated fields are represented by repeated tags.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <show>
    <name>Girls</name>
    <network>
      <name>NBC</name>
      . . .
    </network>
    <cast>
      <person>...</person>
      <character>....</character>
    </cast>
    <cast>
      . . .
    </cast>
    <season>
      <episode>...</episode>
      <episode>...</episode>
      . . .
    </season>
    <season>
      . . .
    </season>
  </show>
</root>
```

1. Recap
2. HTML Crash Course
3. Web Scraping
4. Ethics of Web Scraping

HyperText Markup Language (HTML)

- HTML is the standard language for describing the layout of webpages.

HyperText Markup Language (HTML)

- HTML is the standard language for describing the layout of webpages.
- It is like XML, with special tags for hyperlinks, tables, images, etc.

HyperText Markup Language (HTML)

- HTML is the standard language for describing the layout of webpages.
- It is like XML, with special tags for hyperlinks, tables, images, etc.
- You don’t need to be an HTML expert to scrape webpages, but you do need to know a few basics.

Hyperlinks

The <a> tag indicates a (hyper)link.

Hyperlinks

The <a> tag indicates a (hyper)link.

- The href= attribute contains the URL.

Hyperlinks

The <a> tag indicates a (hyper)link.

- The href= attribute contains the URL.
- The displayed text is within the <a> tag.

Hyperlinks

The <a> tag indicates a (hyper)link.

- The href= attribute contains the URL.
- The displayed text is within the <a> tag.

Example:

```
Web Scraping<br/>   <br/>: line break
  <a href="lectures/lecture18.pdf">
    slides
  </a> |
  <a href="https://colab.research.google.com/drive/1neQvH5uqoX1j74rgCbp
    colab
  </a>
```

Hyperlinks

The <a> tag indicates a (hyper)link.

- The href= attribute contains the URL.
- The displayed text is within the <a> tag.

Example:

```
Web Scraping<br/>
  <a href="lectures/lecture18.pdf">
    slides
  </a> |
  <a href="https://colab.research.google.com/drive/1neQvH5uqoX1j74rgCbp
    colab
  </a>
```

⇓

Web Scraping
slides | colab

Tables

The <table> tag indicates a table.

Tables

The <table> tag indicates a table.

- The <tr> tag indicates a row.

Tables

The <table> tag indicates a table.

- The <tr> tag indicates a row.
- The <th> and <td> tags indicate a cell within a row.

table header, row, data

Tables

The <table> tag indicates a table.

- The <tr> tag indicates a row.
- The <th> and <td> tags indicate a cell within a row.

table header, data, row

```
<table>
  <tr>
    <th>Rank</th>
    <th>Player</th>
    <th>Saves</th>
  </tr>
  <tr>
    <td>1</td>
    <td>Mariano  Rivera</td>
    <td>652</td>
  </tr>
  <tr>
    <td>2</td>
    <td>Trevor  Hoffman</td>
    <td>601</td>
   </tr>
</table>
```

Tables

The <table> tag indicates a table.

- The <tr> tag indicates a row.
- The <th> and <td> tags indicate a cell within a row.

```
<table>
  <tr>
    <th>Rank</th>
    <th>Player</th>
    <th>Saves</th>
  </tr>
  <tr>
    <td>1</td>
    <td>Mariano  Rivera</td>
    <td>652</td>
  </tr>
  <tr>
    <td>2</td>
    <td>Trevor  Hoffman</td>
    <td>601</td>
   </tr>
</table>
```

$\Rightarrow$

Rank	Player	Saves
1	Mariano Rivera	652
2	Trevor Hoffman	601

1. Recap
2. HTML Crash Course
3. Web Scraping
4. Ethics of Web Scraping

Web Scraping

Let’s use what we’ve just learned to scrape some data!

name → Takım adı
year → Sezon yılı
wins → Galibiyet sayısı
losses → Mağlubiyet sayısı
ot-losses → Uzatma mağlubiyetleri
pct → Kazanma yüzdesi (Win %)
gf → Atılan gol (Goals For)
ga → Yenilen gol (Goals Against)
diff → Gol farkı (+ / -)

```
field = cell.attrs["class"][0]
```

1. Recap
2. HTML Crash Course
3. Web Scraping
4. Ethics of Web Scraping

Ethical Considerations

- Website owners have to pay a small amount each time you visit a webpage.

Ethical Considerations

- Website owners have to pay a small amount each time you visit a webpage.
- This is usually offset by advertising.

Ethical Considerations

- Website owners have to pay a small amount each time you visit a webpage.
- This is usually offset by advertising.
- But when you do web scraping:

Ethical Considerations

- Website owners have to pay a small amount each time you visit a webpage.
- This is usually offset by advertising.
- But when you do web scraping:
  - it is easy to rack up many webpage visits,

Ethical Considerations

- Website owners have to pay a small amount each time you visit a webpage.
- This is usually offset by advertising.
- But when you do web scraping:
  - it is easy to rack up many webpage visits,
  - and you don’t see any ads to offset this cost.

robots.txt

- Most websites have a `robots.txt` file in the home directory that indicate which bots are allowed to scrape and which pages they can scrape.

robots.txt

- Most websites have a `robots.txt` file in the home directory that indicate which bots are allowed to scrape and which
- pages they can scrape.
- Here are a few examples:

robots.txt

- Most websites have a `robots.txt` file in the home directory that indicate which bots are allowed to scrape and which
- pages they can scrape.
- Here are a few examples:
  - http://www.espn.com/robots.txt

robots.txt

- Most websites have a `robots.txt` file in the home directory that indicate which bots are allowed to scrape and which
- pages they can scrape.
- Here are a few examples:
  - http://www.espn.com/robots.txt
  - http://www.nytimes.com/robots.txt

robots.txt

- Most websites have a `robots.txt` file in the home directory that indicate which bots are allowed to scrape and which
- pages they can scrape.
- Here are a few examples:
  - http://www.espn.com/robots.txt
  - http://www.nytimes.com/robots.txt
- However, `robots.txt` is informational only. It doesn't
- *prevent* bots from scraping a webpage.

Preventing Web Scraping

Some websites take more drastic measures to prevent web scraping...

```
[1]  import requests
     from bs4 import BeautifulSoup

     response = requests.get("https://explorecourses.stanford.edu/search?view=catalog&academicYear=&page=0&q=STATS&fil
     soup = BeautifulSoup(response.text)
     soup
```

```
        at org.eclipse.jetty.server.handler.ScopedHandler.handle(ScopedHandler.java:143)
        at org.eclipse.jetty.security.SecurityHandler.handle(SecurityHandler.java:578)
        at org.eclipse.jetty.server.session.SessionHandler.doHandle(SessionHandler.java:221)
        at org.eclipse.jetty.server.handler.ContextHandler.doHandle(ContextHandler.java:1111)
        at org.eclipse.jetty.servlet.ServletHandler.doScope(ServletHandler.java:498)
        at org.eclipse.jetty.server.session.SessionHandler.doScope(SessionHandler.java:183)
        at org.eclipse.jetty.server.handler.ContextHandler.doScope(ContextHandler.java:1045)
        at org.eclipse.jetty.server.handler.ScopedHandler.handle(ScopedHandler.java:141)
        at org.eclipse.jetty.server.handler.HandlerWrapper.handle(HandlerWrapper.java:98)
        at org.eclipse.jetty.server.Server.handle(Server.java:461)
        at org.eclipse.jetty.server.HttpChannel.handle(HttpChannel.java:284)
        at org.eclipse.jetty.server.HttpConnection.onFillable(HttpConnection.java:244)
        at org.eclipse.jetty.io.AbstractConnection$2.run(AbstractConnection.java:534)
        at org.eclipse.jetty.util.thread.QueuedThreadPool.runJob(QueuedThreadPool.java:607)
        at org.eclipse.jetty.util.thread.QueuedThreadPool$3.run(QueuedThreadPool.java:536)
        at java.lang.Thread.run(Thread.java:750)
Caused by: java.lang.RuntimeException: Stop that evilness!  If you want data, all you need to do is ask.  Pegging
our servers isn't very friendly.
        at crsearch.action.SearchAction.execute(SearchAction.java:113)
        at crsearch.frontend.ActionServlet.doGet(ActionServlet.java:58)
        ... 28 more
</pre>
```