Web Scraping

INFO 370

Learning Objectives

Practice developing data science questions

Perform a web-scraping task in Python

- Request and parse content from the web
- Write functions and process data in Python

Developing Data Science Questions

Data science is a process.

Data Science Processes

1. Identify a domain/field of interest
2. Perform background research on the topic of your choice
3. Pose a question for your field
4. Identify/collect data (*class today)
5. Process/organize/clean data
6. Explore your data
7. Answer question of interest using the appropriate technique
8. Share results

What makes a good question?

This is never linear or straightforward.

In small groups, write ~3 **measurable**, **testable** data science questions related to gender equity in the entertainment industry (broadly defined).

- While a count can be powerful, answering the question needs to be more than summation

- It may be helpful to frame your question in terms of "trends" and/or "relationships"

I'm the only woman to get the best director award [at the Golden Globes]. You know that was 1984 — that was 34 years ago

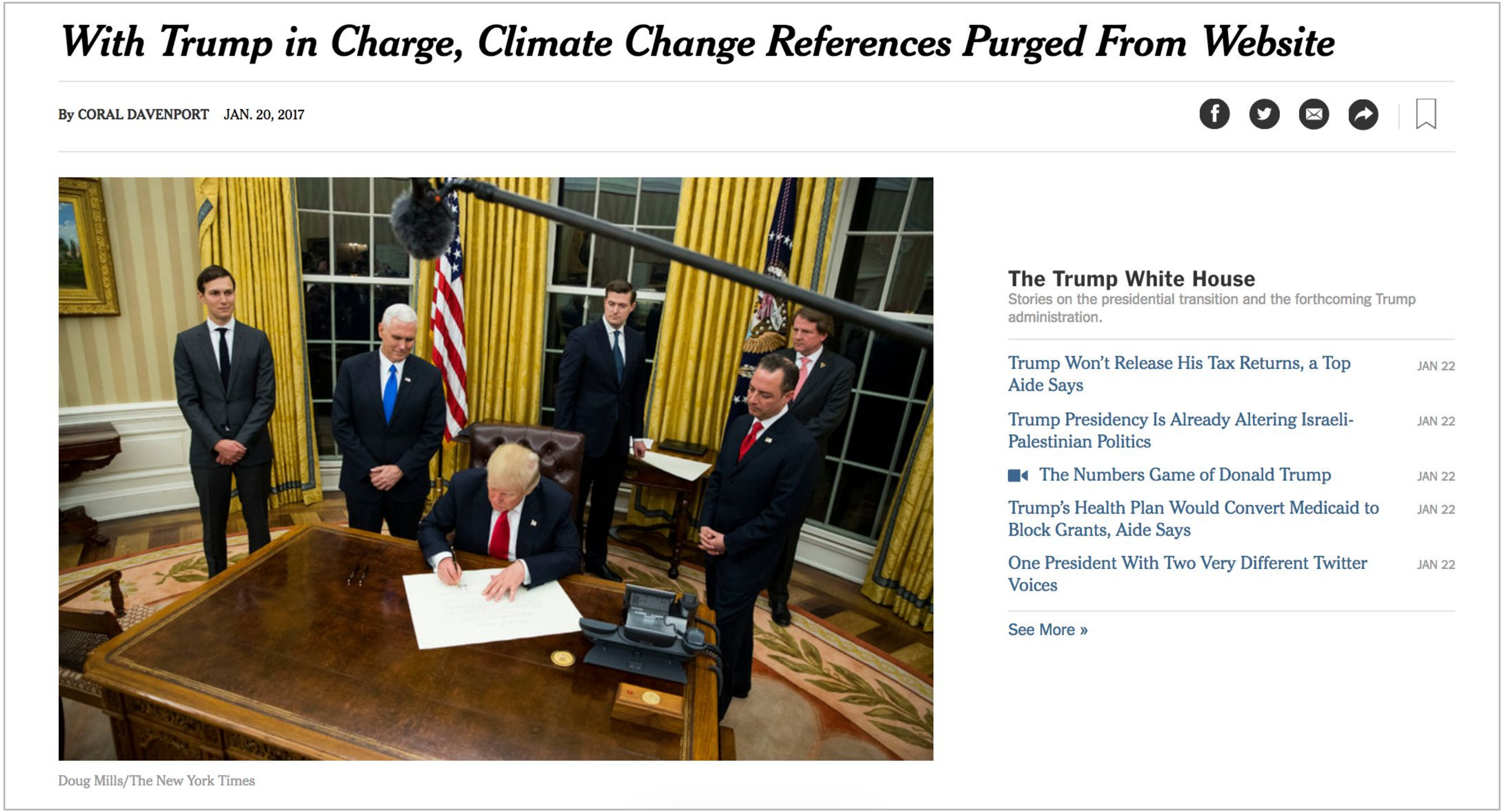
Gender and the Entertainment Industry

Web scraping

SCRAPE IT ALL

Hackers downloaded US government climate data and stored it on European servers as Trump was being inaugurated

Web-scraping in action (link)



With Trump in Charge, Climate Change References Purged From Website

By CORAL DAVENPORT JAN. 20, 2017

Doug Mills/The New York Times

The Trump White House

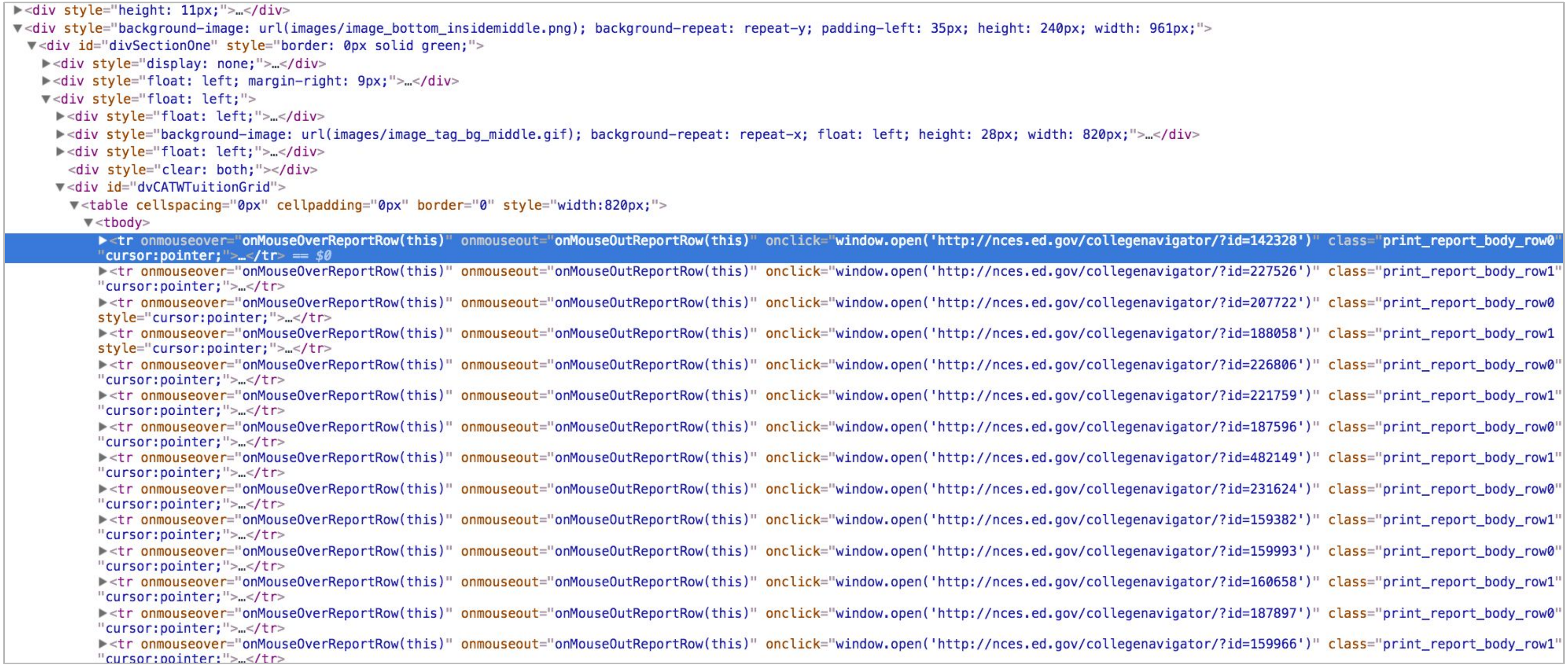
Stories on the presidential transition and the forthcoming Trump administration.

Trump Won't Release His Tax Returns, a Top Aide Says	JAN 22
Trump Presidency Is Already Altering Israeli-Palestinian Politics	JAN 22
The Numbers Game of Donald Trump	JAN 22
Trump's Health Plan Would Convert Medicaid to Block Grants, Aide Says	JAN 22
One President With Two Very Different Twitter Voices	JAN 22

See More »

We needed it (link)

How is content organized on a webpage?



```
▶<div style="height: 11px;">…</div>
▼<div style="background-image: url(images/image_bottom_insidemiddle.png); background-repeat: repeat-y; padding-left: 35px; height: 240px; width: 961px;">
  ▼<div id="divSectionOne" style="border: 0px solid green;">
    ▶<div style="display: none;">…</div>
    ▶<div style="float: left; margin-right: 9px;">…</div>
    ▼<div style="float: left;">
      ▶<div style="float: left;">…</div>
      ▶<div style="background-image: url(images/image_tag_bg_middle.gif); background-repeat: repeat-x; float: left; height: 28px; width: 820px;">…</div>
      ▶<div style="float: left;">…</div>
       <div style="clear: both;"></div>
      ▼<div id="dvCATWTuitionGrid">
        ▼<table cellspacing="0px" cellpadding="0px" border="0" style="width:820px;">
          ▼<tbody>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=142328')" class="print_report_body_row0"
            "cursor:pointer;">…</tr> == $0
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=227526')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=207722')" class="print_report_body_row0
            style="cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=188058')" class="print_report_body_row1
            style="cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=226806')" class="print_report_body_row0"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=221759')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=187596')" class="print_report_body_row0"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=482149')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=231624')" class="print_report_body_row0"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=159382')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=159993')" class="print_report_body_row0"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=160658')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=187897')" class="print_report_body_row0"
            "cursor:pointer;">…</tr>
            ▶<tr onmouseover="onMouseOverReportRow(this)" onmouseout="onMouseOutReportRow(this)" onclick="window.open('http://nces.ed.gov/collegenavigator/?id=159966')" class="print_report_body_row1"
            "cursor:pointer;">…</tr>
```

We need mechanisms for **downloading**, **parsing,** and **searching** the DOM

Requesting webpage information

Import the **requests** package

Request information from a URL

```
import requests as r
page = r.get(url)
print(page.content) # where the HTML tree structure is stored
```

Parsing webpage information

Most popular HTML parsing package is **beautifulsoup**

```
soup = bs(page.content, 'html.parser')
container = soup.findAll('div', { "class" : "container"}) # find divs with class container

# Find elements inside of container
headers = container.findAll('h1')
```

May want to use other packages such as *regular expressions* (**re**) notebook-set-1

Upcoming...

Notebook set 1 **due Friday** night

Assignment 1 **due Tuesday** before class

I will be offline throughout the holiday weekend

Next week

- Exploratory Data Analysis
- Metric Development
- Notebook set 2