Introduction to Python and Web Scraping

Soubhik Barari

MIT Political Science Methods Workshop
March 2017

Introduction to Python and Web Scraping

Repository for code: https://github.com/soubhikbarari/MITMethods2017_python (https://github.com/soubhikbarari/MITMethods2017_python)

Overview

Goal: Teach you how to use Python for common purpose web scraping in social science.

Skills:

1. Python Fundamentals
2. Web Scraping
    - `HTML` / `XML` / `JSON`
    - Working with Twitter data
    - Working with PDF data

Python Fundamentals : Introduction

Motivations

- How do I perform text analysis on a collection of documents?
- How do I work with network data?
- *How do I build an automated web scraping pipeline?*
- How do I write my own machine learning routines?

... How can I flexibly and efficiently do cutting-edge computational social science?

Python Fundamentals : Introduction

Why use Python?

- Highly general purpose
- Easily customizable
- Functional *and* object-oriented

Python vs. R

- Building tools vs. doing analysis
- Flexibility vs. convenience
- Speed vs. parallelization
- Computational vs. statistical

Python Fundamentals : Best Practices

Installation recommendations

- For the best community support and third-party accessibility... install **Version 2.7**.
- For an off-the-shelf, all-inclusive data science environment... install **Anaconda**.

Python Fundamentals : Best Practices

Usage recommendations

- For quick and dirty Python commands ... use a **Python interpreter**.
- For an interactive Python document ... use a **Jupyter notebook**.
- For development ... use an **IDE**
  - "*build your own IDE*" ------ **Sublime Text**
  - "*most like RStudio*" ------ **Spyder**

Python Fundamentals : Best Practices

Package recommendations

- **`pandas`** : basic R-style data structures
- **`numpy, statsmodels`** : numerical / statistical computing
- **`mechanize, beautifulsoup4, pdfminer`** : web scraping
- **`scikit-learn`** : machine learning
- **`networkx, graphx`** : network analysis
- **`bokeh, seaborn`** : data visualization

Python Fundamentals : Best Practices

The **Zen of Python**:

> **Beautiful** is better than ugly
>
> **Explicit** is better than implicit
>
> **Simple** is better than complex
>
> **Complex** is better than complicated
>
> **Readability** counts

Python Fundamentals : Building Blocks

(Demo)

Python Fundamentals : Python vs. R

Python Fundamentals : Python vs. R

Importing a CSV

From https://www.dataquest.io/blog/python-vs-r/

R

```r
nba <- read.csv("nba_2013.csv")
```

Python

```python
import pandas
nba = pandas.read_csv("nba_2013.csv")
```

Python Fundamentals : Python vs. R

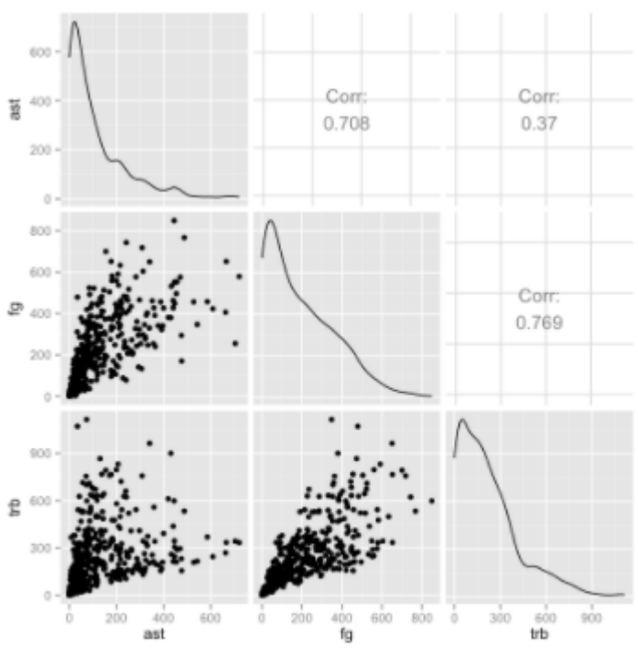
Pairwise plots

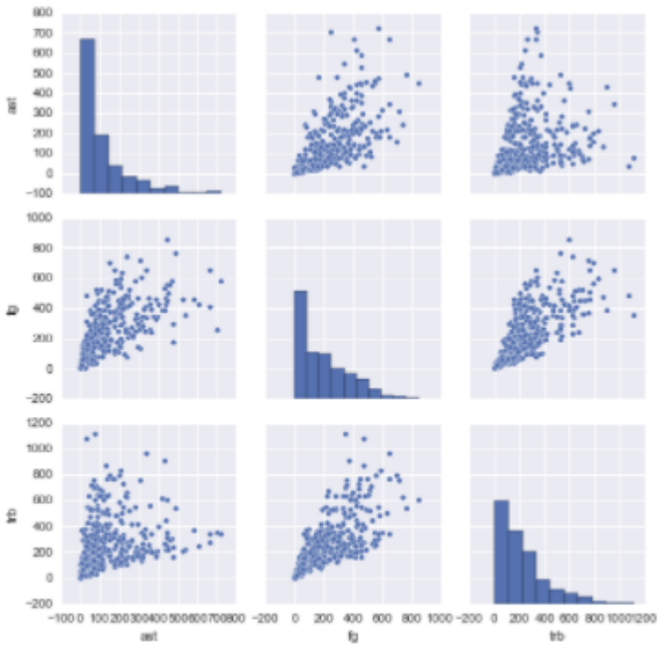
R

```r
library(GGally)
ggpairs(nba[,c("ast", "fg", "trb")])
```

Python

```python
import seaborn as sns
import matplotlib.pyplot as plt
sns.pairplot(nba[["ast", "fg", "trb"]])
plt.show()
```





Python Fundamentals : Python vs. R

Summary statistics

R

```
summary(fit)
```

```
Call:
lm(formula = ast ~ fg, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-228.26  -35.38  -11.45   11.99  559.61
[output truncated]
```

Python

```
import statsmodels.formula.api as sm
model = sm.ols(formula='ast ~ fga', data=train)
fitted = model.fit()
fitted.summary()
```

```
OLS Regression Results
==============================
Dep. Variable:                    ast
R-squared:                      0.568
Model:                            OLS
Adj. R-squared:                 0.567
[output truncated]
```

Python Fundamentals : Python vs. R

Univariate regression

R

```
fit <- lm(ast ~ fg, data=train)
predictions <- predict(fit, test)
```

Python

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(train[["fg"]], train["ast"])
predictions = lr.predict(test[["fg"]])
```

Python Fundamentals : Python vs. R

Web scraping

R

```r
library(rvest)
page <- read_html(url)
table <- html_nodes(page, ".stats_table")[3]
rows <- html_nodes(table, "tr")
cells <- html_nodes(rows, "td a")
teams <- html_text(cells)

extractRow <- function(rows, i){
    if(i == 1){
        return
    }
    row <- rows[i]
    tag <- "td"
    if(i == 2){
        tag <- "th"
    }
    items <- html_nodes(row, tag)
    html_text(items)
}

scrapeData <- function(team){
    teamData <- html_nodes(page, paste("#",team,"_ba
    rows <- html_nodes(teamData, "tr")
    lapply(seq_along(rows), extractRow, rows=rows)
}

data <- lapply(teams, scrapeData)
```

Python

```python
from bs4 import BeautifulSoup
import re
soup = BeautifulSoup(data, 'html.parser')
box_scores = []
for tag in soup.find_all(id=re.compile("[A-Z]{3,}_ba
    rows = []
    for i, row in enumerate(tag.find_all("tr")):
        if i == 0:
            continue
        elif i == 1:
            tag = "th"
        else:
            tag = "td"
        row_data = [item.get_text() for item in row.
        rows.append(row_data)
    box_scores.append(rows)
```

Python Fundamentals : Python vs. R

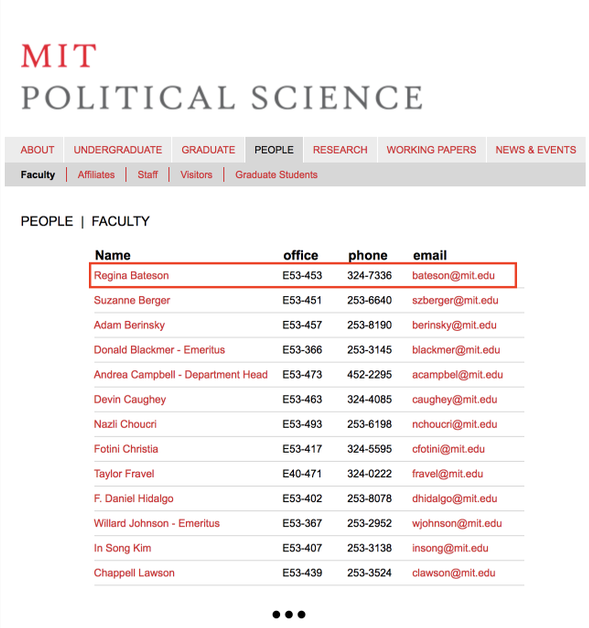
> "Colors of the Python's skin are usually similar to the colors of its habitat. Snakes blend easily with their environment."

Web Scraping

Web Scraping : Site structures

Web Scraping : Site structures

HTML - "a block-based display language."

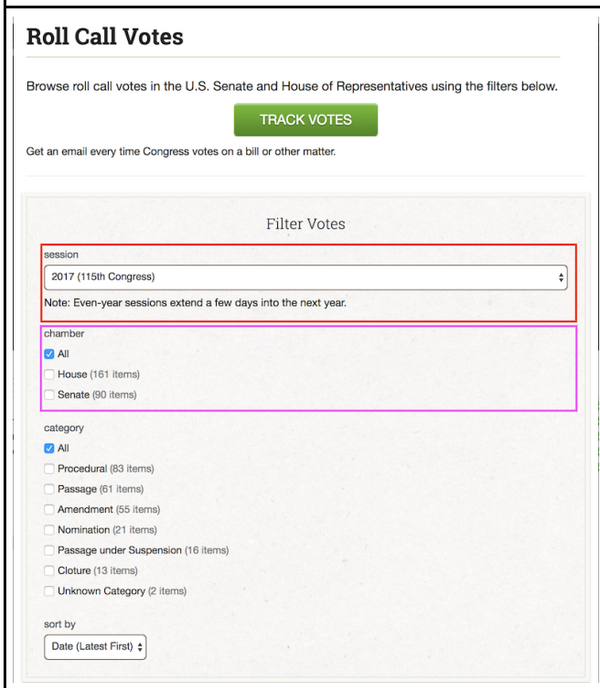
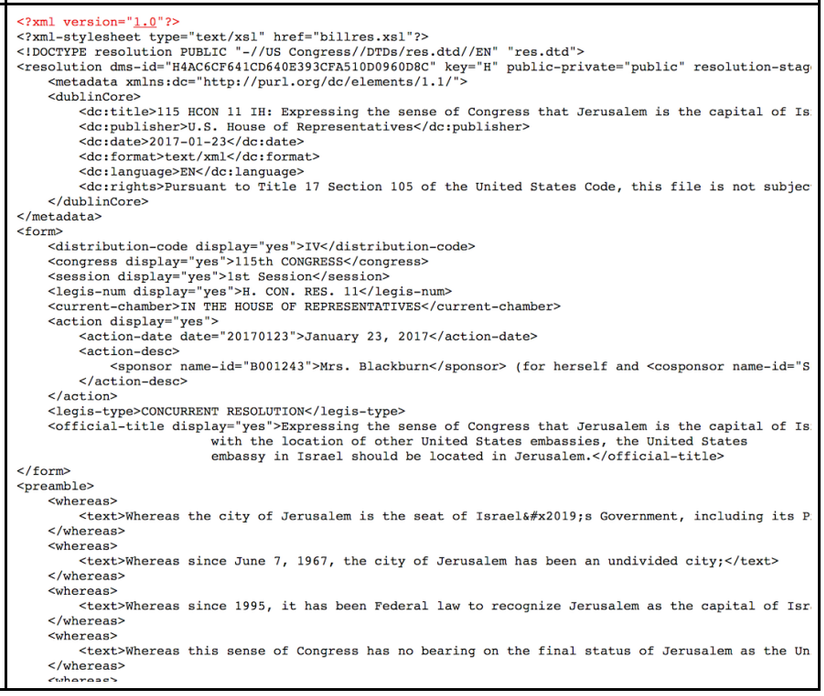
Display	Source
	


```
  <div id="content" class="clearself">

    <div id="aboutMain">
      <h2><a href="/polisci/people/faculty/index.html">PEOPLE</a>  | 
Faculty</h2>
      <table id="listings">
        <thead>
          <tr>
            <td class="name">Name</td>
            <td class="office">office</td>
            <td class="phone">phone</td>
            <td class="email">email</td>
          </tr>
        </thead>
        <tbody>
          <tr>
            <td class="name"><a href="/polisci/people/faculty/gina-bateson.html"><span
class="name">Regina Bateson</span></a></td>
            <td class="office"><span class="office">E53-453</span></td>
            <td class="phone">324-7336</td>
            <td class="email"><a href="mailto:bateson@mit.edu"><span
class="email">bateson@mit.edu</span></a></td>
          </tr>
          <tr>
            <td class="name"><a href="/polisci/people/faculty/suzanne-berger.html"><span
class="name">Suzanne Berger</span></a></td>
            <td class="office"><span class="office">E53-451</span></td>
            <td class="phone"><span class="phone">253-6640</span></td>
            <td class="email"><a href="mailto:szberger@mit.edu"><span
class="email">szberger@mit.edu</span></a></td>
          </tr>
```

Web Scraping : Site structures

Form - "a user interface for modifying web displays."

Display	Source
	


```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="billres.xsl"?>
<!DOCTYPE resolution PUBLIC "-//US Congress//DTDs/res.dtd//EN" "res.dtd">
<resolution dms-id="H4AC6CF641CD640E393CFA510D0960D8C" key="H" public-private="public" resolution-stag
    <metadata xmlns:dc="http://purl.org/dc/elements/1.1/">
    <dublinCore>
        <dc:title>115 HCON 11 IH: Expressing the sense of Congress that Jerusalem is the capital of Is
        <dc:publisher>U.S. House of Representatives</dc:publisher>
        <dc:date>2017-01-23</dc:date>
        <dc:format>text/xml</dc:format>
        <dc:language>EN</dc:language>
        <dc:rights>Pursuant to Title 17 Section 105 of the United States Code, this file is not subjec
    </dublinCore>
</metadata>
<form>
    <distribution-code display="yes">IV</distribution-code>
    <congress display="yes">115th CONGRESS</congress>
    <session display="yes">1st Session</session>
    <legis-num display="yes">H. CON. RES. 11</legis-num>
    <current-chamber>IN THE HOUSE OF REPRESENTATIVES</current-chamber>
    <action display="yes">
        <action-date date="20170123">January 23, 2017</action-date>
        <action-desc>
            <sponsor name-id="B001243">Mrs. Blackburn</sponsor> (for herself and <cosponsor name-id="S
        </action-desc>
    </action>
    <legis-type>CONCURRENT RESOLUTION</legis-type>
    <official-title display="yes">Expressing the sense of Congress that Jerusalem is the capital of Is
                        with the location of other United States embassies, the United States
                        embassy in Israel should be located in Jerusalem.</official-title>
</form>
<preamble>
    <whereas>
        <text>Whereas the city of Jerusalem is the seat of Israel&#x2019;s Government, including its P
    </whereas>
    <whereas>
        <text>Whereas since June 7, 1967, the city of Jerusalem has been an undivided city;</text>
    </whereas>
    <whereas>
        <text>Whereas since 1995, it has been Federal law to recognize Jerusalem as the capital of Isr
    </whereas>
    <whereas>
        <text>Whereas this sense of Congress has no bearing on the final status of Jerusalem as the Un
    </whereas>
```

Web Scraping Techniques : Site structures

XML - "a hierarchical file type to store raw data" (usually displayed in source format)

Display	Source
115TH CONGRESS 1ST SESSION **H. CON. RES. 21** Reaffirming a strong commitment to the United States-Australia alliance relationship. IN THE HOUSE OF REPRESENTATIVES FEBRUARY 6, 2017 Mr. ENGEL (for himself, Mr. ISSA, Mr. SHERMAN, Mr. KEATING, Mrs. NAPOLITANO, Ms. GABBARD, Mr. DEUTCH, Mr. BERA, Mr. TED LIEU of California, Mr. CASTRO of Texas, Ms. KELLY of Illinois, Mr. SUOZZI, Mr. MEEKS, Mrs. TORRES, Mr. SIRES, Mr. BRENDAN F. BOYLE of Pennsylvania, Ms. SPEIER, Mr. CONNOLLY, Ms. HANABUSA, Ms. BORDALLO, Mr. HASTINGS, Mr. EVANS, Mr. SMITH of Washington, Mr. ESPAILLAT, Mr. COURTNEY, Mr. CROWLEY, Mr. HIMES, Mr. SCHNEIDER, Ms. TITUS, Mr. COHEN, and Mr. MCGOVERN) submitted the following concurrent resolution; which was referred to the Committee on Foreign Affairs **CONCURRENT RESOLUTION** Reaffirming a strong commitment to the United States-Australia alliance relationship. Whereas Australia is a vital partner of the United States; Whereas the United States and Australia share core values as well as deep cultural, security, and people-to-people ties; Whereas Australia has been a treaty ally of the United States since the signing of the Australia-New Zealand-United States (ANZUS) Treaty in 1951; Whereas an alliance bond is a sacred vow of friendship and trust, and Australia has always been a faithful and reliable partner to the United States; Whereas United States-Australia defense and intelligence ties and cooperation are exceptionally close, and Australian forces have fought together with the United States military in every significant conflict since World War I and over 100,000 Australian servicemembers have paid the highest price in the course of their service alongside American allies; • • •	<?xml version="1.0"?> <?xml-stylesheet type="text/xsl" href="billres.xsl"?> <!DOCTYPE resolution PUBLIC "-//US Congress//DTDs/res.dtd//EN" "res.dtd"> <resolution dms-id="H5BF9F0D117914E848F426BCDCB70163D" key="H" public-private="public" resolution-stage="Introduced-i <metadata xmlns:dc="http://purl.org/dc/elements/1.1/"> <dublinCore> <dc:title>115 HCON 21 IH: Reaffirming a strong commitment to the United States-Australia alliance relationshi <dc:publisher>U.S. House of Representatives</dc:publisher> <dc:date>2017-02-06</dc:date> <dc:format>text/xml</dc:format> <dc:language>EN</dc:language> <dc:rights>Pursuant to Title 17 Section 105 of the United States Code, this file is not subject to copyright </dublinCore> </metadata> <form> <distribution-code display="yes">IV</distribution-code> <congress display="yes">115th CONGRESS</congress> <session display="yes">1st Session</session> <legis-num display="yes">H. CON. RES. 21</legis-num> <current-chamber>IN THE HOUSE OF REPRESENTATIVES</current-chamber> <action display="yes"> <action-date date="20170206">February 6, 2017</action-date> <action-desc> <sponsor name-id="E000179">Mr. Engel</sponsor> (for himself, <cosponsor name-id="I000056">Mr. Issa</cospo </action-desc> </action> <legis-type>CONCURRENT RESOLUTION</legis-type> <official-title display="yes">Reaffirming a strong commitment to the United States-Australia alliance relationshi </form> <preamble> <whereas> <text>Whereas Australia is a vital partner of the United States;</text> </whereas> <whereas> <text>Whereas the United States and Australia share core values as well as deep cultural, security, and peopl </whereas> <whereas> <text>Whereas Australia has been a treaty ally of the United States since the signing of the Australia-New Ze </whereas> <whereas> <text>Whereas an alliance bond is a sacred vow of friendship and trust, and Australia has always been a faith </whereas> <whereas> <text>Whereas United States-Australia defense and intelligence ties and cooperation are exceptionally close, </whereas> <whereas> <text>Whereas Australia was one of the first countries to commit troops to United States military operations </whereas> <whereas> <text>Whereas Australia is a close partner of the United States, sharing information essential to the defense </whereas> <whereas> <text>Whereas the United States-Australia alliance is an anchor for peace and stability in the Indo-Asia Paci </whereas> <whereas> <text>Whereas Australia has welcomed proposals to reposition United States Marines to maintain Marine forces </whereas> • • •

Web Scraping Techniques : Site structures

JSON - "an unstructured file type to store raw data" (the most common web API type)

Display


```
{
 "meta": {
  "limit": 441,
  "offset": 0,
  "total_count": 432
 },
 "objects": [
  {
   "created": "2011-03-16T17:49:00",
   "id": 28927519,
   "option": {
    "id": 426366,
    "key": "+",
    "value": "Aye",
    "vote": 1
   },
   "person": {
    "bioguideid": "A000022",
    "birthday": "1942-11-19",
    "cspanid": 1002061,
    "firstname": "Gary",
    "gender": "male",
    "gender_label": "Male",
    "id": 400003,
    "lastname": "Ackerman",
    "link": "https://www.govtrack.us/congress/members/gary_ackerman/400003",
    "middlename": "L.",
    "name": "Rep. Gary Ackerman [D-NY5, 1993-2012]",
    "namemod": "",
```


• • •

Web Scraping Techniques : Site structures

PDF - "a document presentation file type"

Display	Source
	

Web Scraping : A Case Study in Congressional Data

I want to study Congress using "big data":

- I want a list of all current Congress members.
- I want to match them to their Wikipedia pages.
- I want a database of Congressional Twitter rhetoric
- I want to parse Congressional bills.

... how can I do this using standard Python tools?

Web Scraping : A Case Study in Congressional Data

First, we need some set up --
dev.twitter.com/resources/signup

(break)

Web Scraping : A Case Study in Congressional Data

(demo)

Web Scraping : Intermediate to Advanced Topics

- Stateful browsing using `mechanize`
- Text parsing using Regular expressions (`ReGeX`)
- Natural language processing / text mining
- Streaming / listening applications