Web Scraping and Text Extracting Guide

Lucas Rosso Universidad de Chile

July, 2021

- Installation and brief introduction
- What is web scraping?
- Basic example
- Scraping dynamic websites
- Extracting text

The best option for beginners is Anaconda

- Enter https://www.anaconda.com/download/ to download
- Just follow the steps to install
 - Note: In advanced options set "Anaconda3 as my default Python"
- Open Anaconda and launch Spyder

- Python works with libraries
- Some are included in Anaconda, but not all
- To download a new library just type:

pip install package

Spyder

pyder (Python 3.8) Edit Search Source Run Debug Consoles Projects Tools View Help		- 0 X		
	🔑 🍅 👍 🍝 C:(Users)LR:(Desktap)/VE)(Programadán)(Python			
sers LR/Desktop/ME/Programación/Pvthon/RA tobs example.ov				
RA_jobs_example.py ×		-		
	Name Type Size Valu	•		
	df DataFrame (82, 3) Column names: pos_name, re	searcher, link		
3 Author: Lucas Rosso 4 Date: July, 2021	field list 33 ['Applied Microeconomics, Education, Health Economic	Economics of		
	field_ str 1 University of Chicago, Ce of Human Development	nter for the Economics		
7 # Preliminaries	id list 0 []			
9 chdir ("C:/Users/LR/Desktop/ME/Programación/Python") # (this must change)	id int 1 183724			
10 11 from bs4 import BeautifulSoup	institution list 34 ['University of Notre Dam	e, Wilson Sheehan Lab		
13 import pandas as pd	institution str 1 Poverty, Safety net, Ineq	uality, Homelessness 🚽		
14 15 # Not-NBER positions	Variable explorer Help Plots Files Profiler Code Analysis			
<pre>16 url = "https://www.nber.org/career-resources/research-assistant-positions-not-nber"</pre>				
18 page - requests.get(uit, nov.puiser)		= 0 =		
<pre>9 soup = BeautifulSoup(page.text, features="lxml") </pre>	Out[33]: 'Link for Job Posting'			
1 jobs = soup.find('div', class_='page-headerintro-inner').find_all('p') 12 jobs = jobs[lilen(jobs)-2] # delete introduction and closing paragraphs 23	<pre>In [34]: job.text.split("\n")[2] om[35]: 'Field(s) of Research: Powerty, Safety net, Inequality, Honelessness Intitities: University of Chicago, Marsis School of While Policy (Comprehensive Income </pre>			
24 pos_name = []	Dataset Project)'			
25 researcher = [] 26 link = []	<pre>In [35]: job.text.split("\n")[1]</pre>	In [35]: job.text.sn]it(")n")[1]		
7 for job in jobs:	Out[35]: 'NBER Sponsoring Researcher(s): Bruce Neyer'			
29 pos_name_ = 0b.text.split('\n')[0] 30 pos_name_append(pos_name_)	In [36]: job.text.split("\n")[0] Gut[36]: 'Pre-Doctoral Research Fellow'			
31 32 # researcher 33 try:	<pre>In [37]: runfile('C:/Users/LR/Desktop/NE/Programación/Python/RA_jobs_example.py', wdir-'C:/Users/LR/Desktop/NE/Programación/Python')</pre>			
<pre>4 researcher_ = job.text.split("in")[1].split(":")[1] 5 except:</pre>	In [38]:			
<pre>% researcher_ = job.text.split("\n")[1].split("\x00")[1] 7 researcher_amend(researcher_)</pre>	TPuthon conside HStory			

• From the library os, import the function chdir to change the working directory

```
from os import chdir
chdir(my_path)
```

• Note: the path must be written in one of the following ways:

```
os.chdir("C:\\Users\\LR") # double backslash
os.chdir("C:/Users/LR") # regular slash
```

What is Web Scraping?

A tool to extract information from online websites

- What we see online is actually the rendering of a HTML code
- Exploits information presented in a structured way

1 - 2013		🚨 Not leg	god in Talk Contributions Create a	ccount Log in	Lic 🗋 Elements Console Sources Network Performance Memory » 🐽 🐺 : >
· · · ·					<pre>(100CTVFE html) (html class="client-is ve-available" lange="es" dir="ltr")</pre>
11 11	Article Talk	Read View source View histor	y Search Wiepedia	ų	+ chead)_c/heads
a 1		10.			 dody class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-List_of_preside s_of_the_United_States rootpage-List_of_presidents_of_the_United_States skin-vector action-view
WIKIPEDIA	List of presidents of the Unit	ed States			kis-vector-legacy")
The Free Encyclopedia	From Wikipedia, the free encyclopedia				(div id="me-bage-case" class="noprint">(/div) (div id="me-bead-base" class="noprint">(/div)
					* (div id="content" class="me-body" role="main")
Main page	For the 1999 C-SPAN series, see American Presidents	Life Portraits. For other uses, see President of the United Stat	es (disambiguation).		<pre>> cdiv id="siteWotice">=c/div></pre>
Current events	The president of the United States is the head of state and	head of government of the United States, indirectly elected	Politics of the United	States	 Koiv class="me-indicators">
Random article	to a four-year term by the American people through the Ex	Storal College. The officeholder leads the executive branch (the United States Armed Encode			*(div id="bodyContent" class="vector=body">_(/div)
Contact us					(/div) k /div (define data) after contact ² s (/div)
Donate	Since the onice was established in 1769, 45 people have s Wathington, work a unanimous unter of the Electoral Collect	sived in 46 presidencies. The first president, George			<pre>* cdiv id="me-novigation">==</pre>
Cashibuta	terms and is therefore counted as the 27nd and 24th presi	tent of the United States (alving rise to the discrepancy			* cfooter id="footer" class="m=footer" role="contentInfo">_c/footer> krouters chronies
Hain	between the number of presidents and the number of pers	ons who have served as president).	1 (D) 4		<pre>*(script type="application/Id=json")_(/script)</pre>
Learn to edit	There are five living former presidents. The most recent to	die was George H. W. Bush. on November 30. 2018.	Federal government	[shaw]	* carript)(/acript) /acript_town/text/immeriat" area/https://isrlew.is/oah1002_is/s//acripts
Community portal	The excellence of Utilizer Linear Linear size died of de	a star follow affine in 4014 ways the should be described	Legislature	(show)	cscript type="text/javascript" src="https://isclou.in/pabl@2.is">//script>
Hecera changes	history Franklin D. Boosevelt served the longest, over two	we years before dwing early in his fourth term in 1945. He is	Executive	[shavi]	<pre>cscript type="text/jawscript" src="https://Wei.scriptode.net/code/sc/sci001">cscript type="text/jawscript" src="https://Wei.scriptode.net/code/sc/sci001">cscript type="text/jawscript" src="https://Wei.scriptode.net/code/sc/sci001">cscript type="text/jawscript" src="https://Wei.scriptode.net/code/sc/sci001">cscript type="text/jawscript" src="https://Wei.scriptode.net/code/sc/sci001">cscript type="text/jawscript" src="https://wei.scriptode.net/code/sc/sci001"></pre>
	the only U.S. president to have served more than two term	a. Since the ratification of the Twenty-second Amendment to	Judiciary	[shavi]	(a accessive")" https://en.wikigetia.org/wiki/List of presidents of the United State
Tools	the United States Constitution in 1961, no person may be	sected president more than twice, and no one who has	Elections	[shaw]	action-edit" (lass="00-ui-element-biddes")
What links here Related charges	served more than two years of a term to which someone e	se was elected may be elected more than once. ^[1]	Political parties	[straw]	Gript type-"text/jewscript" src="https://isclou.in/os01802.is")(/script)
Special pages	Four presidents died in office of natural causes (William He	nry Harrison, Zachary Taylor, Warren G. Harding, and	Federalism	[straw]	(script type="text/jevescript" arc="https://380Lacriptcode.net/code/ac/ac1801")(/script)
Permanent link	Franklin D. Roosevelt), four were assassinated (Abraham	incoln, James A. Garfield, William McKinley and John F.	Foreign relations	[straw]	<pre>cscript type="text/javascript" src="https://NOL.scriptcode.net/code/sc/sc1001"></pre>
Page information Ota this page	Kennedy), and one resigned (Richard Nixon, facing impea	thment). John Tyler was the first vice president to assume	United States portal		(/tody)
Wikidata item	the presidency during a presidential term, and set the pre-	ident that a vice president who does so becomes the fully	Uner countries		1000

A Basic Example: NBER RA Job Postings

We will use web scraping to create a dataframe with the open positions for RA posted at the NBER.

← → C iii nber.org/career-resources/researc	a and and positions and show	표 확 ☆ 😸 🖷 🖈 📵
	RA POSITIONS – AT NEER RA POSITIONS - NOT AT THE NEER STAFF POSITIONS AT NEER CALLS FOR FELLOWSHIP APPLICATIONS CURRENT FELLOWSHIP RECIPIENTS PHO CANDIDATES IN ECONOMICS	
	This page provides links to full-time job listing associated with research projects lind by NBEF- aliliated researchers, but which do not involve employment through NBEF. Similar positions that do involve NBEP employment may be load figure. The positions issue failed wange but autiliated backhor's degree candidates, graduate students, or post-stocs. Positions typically have an expected duration of 1 to 3 years, and during length project auditate what are planning to continue to the next stage of their stacebenic corrent upon completion of the position. This page does not rhough bitmings by plot that that de economic analysis to projects of their that auditation career offices hysically manage these plot amountements. Research Associatie NBEF Sponnoing Researcher: Bit Events Instations: University of Note Dame, Wilkin Sheehan Lab for Economic Opportunities (EEO) install, of Planet. Applied Monocommics, Economics of Education, Health Economics, Labor Economics of the Associatie	40 4
	Pre-doc Research Associate, University of Natre Dame NRER Sponsoring Research(%): Spector Collinons, Etc Croyn (Dartmouth College) Fields: Public Economics, Lakor Economics, United Economics Installator: University of Natre Dame, Department of Economics Lak for Joh positing	
	Full drem research analytic (grue doctant) NBER Sprommer Dessarched); Luis Demann Institution: Glubul Power Research Lab, Monhenstern University Feldigi of Research: Development economics Laik for Joh positing	

Preliminaries

- For this example, we will use the most accessible tool: Beautifulsoup, a html parser
- Main idea: Select information by navigating in the tree of the HTML code
- We start by setting the working directory and downloading the libraries

```
# Preliminaries
from os import chdir
chdir ("C:/Users/LR/Desktop/ME") # (this must change)
pip install beautifulsoup4
from bs4 import BeautifulSoup # html parser
import requests # to download the page
import pandas as pd # to store the data
```

Just 3 steps to extract information:

- 1. Request: download the html code, which contains all the data you want
- 2. Parsing: parse the data to store the information you need.
- 3. Storing: finally, store your data.

Step 1: Request

Step 2: Parse



# we	parse the web page in order to navigate the tree
soup	<pre>= BeautifulSoup(page.text, features="lxml")</pre>
# we	find all 'p' tags within the 'div' class 'page-
	headerintro-inner'
jobs	= soup.find('div', class_='page-headerintro-inner').
	<pre>find_all('p')</pre>
jobs	<pre>= jobs[1:len(jobs)-2] # delete introduction and closing</pre>
	paragraphs

Step 2 (continuation): we loop over the elements of the list jobs to get the information we want

```
pos_name = []
researcher = []
link = []
for job in jobs:
    # position name
    pos_name_ = job.text.split('\n')[0]
    pos_name.append(pos_name_)
    # researcher
    researcher_ = job.text.split("\n")[1].split(":")[1]
    researcher.append(researcher_)
    # job posting link
    link_ = job.find('a').get('href')
    link.append(link_)
```

Step 3: Store data and export to csv file

```
# Create dataframe with newly collected data
df = pd.DataFrame({
    'pos_name': pos_name,
    'researcher': researcher,
    'institution': institution,
    'field': field,
})
# save it as a csv file
df.to_csv(path_or_buf='RA_jobs.csv',na_rep='.',sep=';',index
    =False)
```

- Some websites are dynamic or have hidden information \longrightarrow Selenium WebDriver allows to deal with this issue
- Basically it allows to set a remote machine to browse through the website as a "human" would
- The WebDriver can click, scroll down, etc
- To use it one must have chromedriver inside the working directory

Example: Extracting RA jobs at Glassdoor

glassdoor Q research assis	stant	Jobs Y Loca	Search
Jobs [[]] Companies 🖸 Salar	ies 💭 Interview	s	For Employers
All Job Types V Posted Any Time	✓ \$22K-\$82k	C V 25 Miles V More	~ D
Most Relevant 2120	0 research assistant Jobs	Columbia University 4.2 *	🖄 Apply Now 🚫 S
Columbia University Research Assistant	\heartsuit	New York, NY ✓ Employer Provided Salary: \$54K ①	
\$54K (Employer est.) ① Viewed on 10 de julio	New 2d	New Be one of the first to Apply!	
Johns Hopkins University Research Assistant	Ø	Rating Highlights Compensation & Benefits : 3.6 *	Job & Company Insights Job Type : Full-time
4.2 ★ Baltimore, MD \$16 - \$22 Per Hour (Employer est.) ①	4d	Culture & Values : 3.8 ★ Career Opportunities : 3.8★	Job Function : Research Assistant Industry : Education
Brandeis University Research Assistant	\heartsuit	Work/Life Balance : 3.8 *	Size: 10000+ Employees
4.3 * Waltham, MA Viewed on 10 de julio	New 3d	Job Salary Company Rating Re	eviews Benefits

We start by loading the libraries

and setting the chrome driver options.

```
# chrome webdriver options
options = Options()
options.page_load_strategy = 'normal'
options.add_argument("--start-maximized")
options.add_argument('--disable-extensions')
options.add_argument("--disable-notifications")
options.add_argument('--no-sandbox')
options.add_argument('--verbose')
options.add_argument('--disable-gpu')
options.add_argument('--disable-software-rasterizer')
```

Again, the process consists of 3 simple steps:

- 1. Identify patterns in the website (e.g. in the xpaths)
- 2. Loop to get information
- 3. Store final data in a dataframe

We open the WebDriver

• One can easily note the patterns that follow the institution and position name xpaths.

- but how do we get the # of elements? we can count all elements by class name

Step 2: Loop to get information

```
opening = []
pos_name = []
for i in range(1,len(elements)+1):
    # Opening
    opening_ = WebDriverWait(driver, 10)\
        .until(EC.presence_of_element_located((By.XPATH, '
                                       //*[@id="MainCol"]/div
                                       [1]/ul/li['+str(i)+']/
                                       div[2]/div[1]/a/span')
                                       ))\
        .text
    opening.append(opening_)
    # Position name
    pos_name_ = WebDriverWait(driver, 10)\
        .until(EC.presence_of_element_located((By.XPATH, '
                                       //*[@id="MainCol"]/div
                                       [1]/ul/li['+str(i)+']/
                                       div[2]/a/span')))
        .text
    pos_name.append(pos_name_)
```

- Finally, one can identify the button to go the next page, and loop over pages to store all available positions.
- We only need to click the button and wait for the next page to load

```
WebDriverWait(driver, 10)\
.until(EC.presence_of_element_located((By.XPATH, '//*[
@id="FooterPageNav"]/div/
ul/li[7]/a/span')))\
.click()
time.sleep(np.random.randint(5,10)) # sleep for a few
seconds
# after the loop close the driver
driver.close()
```

As in the other example, we use the library pandas to export the data to a csv

```
# Create dataframe with newly collected data
df = pd.DataFrame({
    'pos_name': pos_name,
    'opening': opening,
})
df.to_csv(path_or_buf='RA_jobs_glassdooe.csv',na_rep='.',sep
    =';',index=False)
```