Web Scrapping

January 31, 2018

0.1 Importar pandas

```
In [2]: import pandas as pd
```

1 Buscar:

2 List of countries by population density

3 URL: https://simple.wikipedia.org/wiki/List_of_countries_by_population_densi

```
In [3]: url = 'https://simple.wikipedia.org/wiki/List_of_countries_by_population_density'
        densidad_df = pd.read_html(url, header=0)
```

```
In [12]: densidad_df = pd.DataFrame(densidad_df[0])
```

```
In [13]: densidad_df.head()
```

```
Out[13]:    Rank Country / dependent territory Population  Date last updated Area (km2)  \
         0     1                  Macau (China)     541200  February 28, 2017       29.2
         1     2                        Monaco      33000               2013       1.95
         2     3                     Singapore    5076700               2000      710.2
         3     4              Hong Kong (China)    7264100               1998       1104
         4     5                 Gibraltar (UK)      31000               2018        6.8

            Area (mi2) Density (km2) Density (mi2)   Notes
         0        11.3         18534         48003     [1]
         1        0.75         16923         43830  [2][3]
         2       274.2          7148         18513     [4]
         3         426          6349         16444     [5]
         4         2.6          4559         11808     [2]
```

```
In [14]: type(densidad_df)
```

```
Out[14]: pandas.core.frame.DataFrame
```

```
In [15]: densidad_df_limpio = densidad_df.copy()
```

```
In [16]: densidad_df_limpio.keys()
```

```
Out[16]: Index(['Rank', 'Country / dependent territory', 'Population',
                'Date last updated', 'Area (km2)', 'Area (mi2)', 'Density (km2)',
                'Density (mi2)', 'Notes'],
               dtype='object')

In [19]: densidad_df_limpio.pop('Area (mi2)')

Out[19]: 0              11.3
         1              0.75
         2             274.2
         3               426
         4               2.6
         5              0.17
         6               290
         7               122
         8                20
         9                13
         10              115
         11            56980
         12               45
         13               30
         14             20.5
         15            13890
         16              790
         17              170
         18               75
         19             2320
         20               24
         21              144
         22            38432
         23              8.1
         24             3427
         25              171
         26             4036
         27              8.1
         28               10
         29            16033
                       ...
         212          148709
         213          239285
         214          489000
         215          830000
         216          478841
         217          132000
         218          188500
         219          119500
         220          424164
         221          496000
```

```
222        6601668
223             100
224          240535
225         1052100
226          103347
227          679360
228           83000
229         3855100
230          224610
231          395960
232           63250
233           40000
234         2966200
235          318261
236           35000
237          103000
238          603909
239            4700
240          840000
241    Area (mi2)
Name: Area (mi2), Length: 242, dtype: object
```

In [18]: densidad_df_limpio.pop('Density (mi2)')

Out[18]: 0   48003
         1   43830
         2   18513
         3   16444
         4   11808
         5   4861
         6   4263
         7   3424
         8   3175
         9   2852
         10   2686
         11   2497
         12   2186
         13   2183
         14   1717
         15   1655
         16   1634
         17   1541
         18   1435
         19   1764
         20   1308
         21   1292
         22   1261
         23   1150

```
24              1163
25              1155
26              1046
27              1041
28              1041
29              1072
            ...
212              36
213              34
214              31
215              31
216              31
217              31
218              26
219              24
220              23
221              23
222              21
223              20
224              18
225              16
226              14
227             9.3
228             9.1
229             8.8
230             8.8
231             8.3
232             8.3
233             8.0
234             8.5
235             6.7
236             5.4
237             4.9
238             4.4
239            0.67
240           0.067
241   Density (/mi2)
Name: Density (mi2), Length: 242, dtype: object
```

In [21]: densidad_df_limpio.pop('Notes')

Out[21]: 0   [1]
         1   [2][3]
         2   [4]
         3   [5]
         4   [2]
         5   [2][6]
         6   [10]
```

```
7      Eurostat estimate
8                    [2]
9                    [2]
10                   [2]
11               [2][7]
12              [9][11]
13                  [10]
14                  [11]
15                  [12]
16                   [2]
17                   [2]
18                   [2]
19                  [13]
20                  [14]
21             [11][15]
22              [2][16]
23                   [2]
24                   [2]
25                   [2]
26                   [2]
27                  [11]
28                   [2]
29                  [17]
              ...
212                 [72]
213                 [67]
214                  [2]
215                  [2]
216                 [73]
217                  [2]
218                  [2]
219                  [2]
220                  [2]
221                 [74]
222                 [75]
223                  [2]
224                  [2]
225                 [76]
226                  [2]
227                  [2]
228                  [2]
229                 [77]
230                  [2]
231                  [2]
232                  [2]
233   Statistics Iceland
234                 [78]
235                  [2]
```

```
236                        NaN
237                        [2]
238                        [2]
239                        [2]
240                    [2][79]
241                      Notes
Name: Notes, Length: 242, dtype: object
```

In [28]: densidad_df_limpio.keys()

Out[28]: Index(['Rank', 'Country / dependent territory', 'Pop', 'Date last updated',
            'Area (km2)', 'Density (km2)'],
          dtype='object')

In [27]: densidad_df_limpio.rename(columns={'Population': 'Pop'},
                                   inplace=True)

In [30]: densidad_df_limpio.rename(columns={'Area (km2)': 'Area'},
                                   inplace=True)

         densidad_df_limpio.rename(columns={'Density (km2)': 'Density'},
                                   inplace=True)

In [31]: densidad_df_limpio.head()

Out[31]:   Rank Country / dependent territory   Pop Date last updated   Area \
         0   1   Macau (China)   541200 February 28, 2017   29.2
         1   2   Monaco   33000   2013   1.95
         2   3   Singapore   5076700   2000   710.2
         3   4   Hong Kong (China)   7264100   1998   1104
         4   5   Gibraltar (UK)   31000   2018   6.8

           Density
         0   18534
         1   16923
         2   7148
         3   6349
         4   4559

In [34]: densidad_df_limpio[['Area', 'Density']].describe()

Out[34]:   Area Density
         count   242   242
         unique 241   170
         top   21   46
         freq   2   5
```

4 Web Scrapping con BS4

```
In [36]: ! pip install bs4
```

```
Collecting bs4
  Downloading bs4-0.0.1.tar.gz
Requirement already satisfied: beautifulsoup4 in /Users/israel/anaconda3/envs/ensitech/lib/pyt
Building wheels for collected packages: bs4
  Running setup.py bdist_wheel for bs4 ... done
  Stored in directory: /Users/israel/Library/Caches/pip/wheels/84/67/d4/9e09d9d5adede2ee1c7b7e8
Successfully built bs4
Installing collected packages: bs4
Successfully installed bs4-0.0.1
```

5 10.14.242.142:8888/?token=b93378280ce2880ccd04c2b6a9f7e916926e18049e17ccfe

```
In [37]: import requests
         from bs4 import BeautifulSoup
```

5.1 DOM (Document Object Mode)

6 Buscar: No Starch Press

7 URL: https://www.nostarch.com/

```
In [41]: # Guardar la URL como String
         url = 'https://www.nostarch.com/'


         # Transformamos la URL como objeto de BS4

         url = BeautifulSoup(requests.get(url).text, 'lxml')
```

```
In [42]: type(url)
```

```
Out[42]: bs4.BeautifulSoup
```

```
In [51]: url.title
```

```
Out[51]: <title>No Starch Press | No Starch Press</title>
```

```
In [55]: url.title.contents
```

```
Out[55]: ['No Starch Press | No Starch Press']
```

Calling a tag like a function is the same as calling its find_all() method. Eg. tag('a') returns a list of all the A tags found within this tag.

```
In [60]: url.find_all('p')

Out[60]: [<p class="cart-block-items collapsed uc-cart-empty">There are no products in your sho
          <p><a href="/sparkfunarduino">The Arduino Inventor's Guide</a> is a hands-on introduc
          <p><a href="https://www.nostarch.com/GrayHatCSharp">Gray Hat C#</a> shows you how to
          <p><a href="/iphoneappsforkids">Coding iPhone Apps for Kids</a> teaches you how to p:
          <p><a href="/packetanalysis3">Practical Packet Analysis, 3rd Ed.</a> teaches you how
          <p><b><a href="/arduinoplayground">Arduino Playground</a></b> is a collection of fun
          <p><b><a href="/hardwarehacker">The Hardware Hacker</a></b> is an illuminating caree!
          <p><a href="/facebookforseniors">Facebook for Seniors</a> offers 12 step-by-step less
          <p><a href="/inventwithpython">Invent Your Own Computer Games with Python</a> teache:
          <p><a href="/scratchcards">Scratch Coding Cards</a> is a deck of 75 activity cards w:
          <p><a href="/beautifulminecraft">Beautiful Minecraft</a> is a stunning collection of
          <p><a href="/wcss2">Wicked Cool Shell Scripts, 2nd Edition</a> is a collection of fu!
          <p><a href="technicbuilder2">The Unofficial LEGO Technic Builder's Guide, 2nd Ed.</a>
          <p><a href="https://www.nostarch.com/gtfo">PoC||GTFO</a> is a compilation of the wild
          <p><a href="/microprocessors">The Manga Guide to Microprocessors</a> is a comic guide
          <p><a href="/arduinohandbook2">Arduino Project Handbook, Vol. 2</a> is a full-color g
          <p>The Python edition of <a href="https://www.nostarch.com/thinkpython">Think Like a
          <p><a href="https://www.nostarch.com/computersforseniors">Computers for Seniors</a> :
          <p><a href="/SeriousCrypto">Serious Cryptography</a> is a practical guide to the past
          <p><a href="https://www.nostarch.com/legoarchitecture">The LEGO Architecture Idea Boo
          <p><a href="/Rust">The Rust Programming Language</a> is the official hands-on, no-nor
          <p>The <a href="/legotrains">LEGO Trains Book</a> is an illustrated guide to the worl
          <p><a href="/networkprotocols">Attacking Network Protocols</a> is a deep-dive into ne
          <p><a href="/azure">Attacking Microsoft Azure</a> is a comprehensive guide to penetra
          <p><a href="/missionpython">Mission Python</a> will teach you how to program an adven
          <p><b><a href="/gameconsole">The Game Console</a></b> is a visual history of video ga
          <p><a href="/mailchimp/subscribe">Want sweet deals? <br/>Sign up for our newsletter.<
          <p>Now available in <a href="/aboutearlyaccess">Early Access</a>!</p>,
          <p> <a href="/arduinohandbook2">Arduino Project Handbook Vol. 2</a><br/>
           <a href="/networkprotocols">Attacking Network Protocols</a><br/>
           <a href="/crackingcodes">Cracking Codes with Python</a><br/>
           <a href="/learnjava">Learn Java the Easy Way</a><br/>
           <a href="/microprocessors">The Manga Guide to Microprocessors</a><br/>
           <a href="/practicalsql">Practical SQL</a><br/>
           <a href="/rootkits">Rootkits and Bootkits</a><br/>
           <a href="/Rust">The Rust Programming Language</a><br/>
           <a href="/seriouscrypto">Serious Cryptography</a><br/>
           <a href="/thinkpython">Think Like a Programmer, Python Edition</a></p>,
          <p><b><a href="/nsm" target="new">The Practice of Network Security Monitoring</a></b>
          <p>Read Andrew bunnie Huang's <b><a href="https://www.reddit.com/r/netsec/comments/5:
          <p><b><a href="http://www.vox.com/videos/2017/2/6/14445516/minecraft-art-business-map
          <p>This is a look inside a mind without peer. <b>Edward Snowden</b> on <b><a href="/!
          <p><b><a href="/facebookforseniors" target="new">Facebook for Seniors</a></b> is one
          <p>Natalie Rusk, creator of the <a href="/scratchcards" target="new"><b>Scratch Codi!
          <p>Merry Brickmas! <a href="/legochristmas" target="new"><b>The LEGO Christmas Orname
          <p>Dave Taylor, author of <b><a href="/wcss2" target="new">Wicked Cool Shell Scripts<
```

```
    <p>Bring your LEGO collection to life: <a href="/legoanimation" target="new"><b>The I
    <p>Need a few geeky summer reads? <b><a href="http://www.wsj.com/articles/three-books
    <p>The author of <b><a href="/tinylegowonders" target="new">Tiny LEGO Wonders</a></b>
    <p><b><a href="http://www.slj.com/2016/05/collection-development/read-watch-alikes/cd
    <p><b><a href="/carhacking" target="new">The Car Hacker's Handbook</a></b> is the <b>
    <p>No Starch Press founder, Bill Pollock, did a Reddit AMA with authors <a href="/me1
    <p>Why publish books about Scratch? The Scratch Foundation and Bill Pollock answer ir
    <p>Be the Corbusier of LEGO: <a href="/legoarchitect" target="new"><b>The LEGO Archit
    <p><a href="/smartgirlsguide" target="new"><b>The Smart Girl's Guide to Privacy</b></
    <p>No Starch Press fans found <b><a href="https://twitter.com/billpollock/status/616]
    <p><a href="feeds/newbooks.xml"><img align="left" alt="RSS Feed" border="0" src="/mis
    <p><a href="feeds/comingsoon.xml"><img align="left" alt="RSS Feed" border="0" src="/r
    <p><a href="/mailchimp/subscribe"><img align="left" alt="Newsletter" border="0" src="
    <p><a href="https://www.twitter.com/nostarch"><img align="left" alt="Twitter" border=
    <p><a href="https://www.facebook.com/home.php?#/pages/No-Starch-Press/57041709833"><i
```

In [61]: url.find_all('a')

Out[61]: [Skip to main cont

    ,
    Catalog,
    Media,
    Write for Us,
    About Us,
    Art & Design,
    General Computing,
    Hacking & Computer Security,
    Hardware / DIY,
    Kids,
    LEGO,
    LEGO^ŏ MINDSTORMS
    Linux & BSD,
    Manga,
    Programming,
    Python,
    Science & Math,
    System Administration,
    EARLY ACCESS,
    <a href="/about_ebooks.htm" style="color: red1;">Free ebook edition with every print
    <a href="/cart"><span class="cart-block-icon-empty" title="View your shopping cart.">
    My account,
    Log in,
    Create account,
    <a href="/scratch"><img alt="Super Scratch Programming Adventure! " class="img-respor
    <a href="/mg_physics.htm"><img alt="The Manga Guide to Physics" class="img-responsive
    <a href="/learnscratch"><img alt="Learn to Program with Scratch" class="img-responsiv
```

```
<a href="/legoarchitect"><img alt="The LEGO Architect" class="img-responsive" height=
<a href="/tlcl"><img alt="The Linux Command Line" class="img-responsive" height="119"
<a id="main-content"></a>,
<a href="/sparkfunarduino">The Arduino Inventor's Guide</a>,
<a href="/sparkfunarduino"><img alt="The Arduino Inventor's Guide" class="img-respons
<a href="/sparkfunarduino">The Arduino Inventor's Guide</a>,
<a href="/grayhatcsharp">Gray Hat C#</a>,
<a href="/grayhatcsharp"><img alt="Gray Hat C#" class="img-responsive" height="225" s
<a href="https://www.nostarch.com/GrayHatCSharp">Gray Hat C#</a>,
<a href="/iphoneappsforkids">Coding iPhone Apps for Kids</a>,
<a href="/iphoneappsforkids"><img alt="Coding iPhone Apps for Kids" class="img-respo
<a href="/iphoneappsforkids">Coding iPhone Apps for Kids</a>,
<a href="/packetanalysis3">Practical Packet Analysis, 3rd Edition</a>,
<a href="/packetanalysis3"><img alt="Practical Packet Analysis, 3rd Edition" class=":
<a href="/packetanalysis3">Practical Packet Analysis, 3rd Ed.</a>,
<a href="/arduinoplayground">Arduino Playground</a>,
<a href="/arduinoplayground"><img alt="Arduino Playground" class="img-responsive" he:
<a href="/arduinoplayground">Arduino Playground</a>,
<a href="/hardwarehacker">Hardware Hacker</a>,
<a href="/hardwarehacker"><img alt="The Hardware Hacker " class="img-responsive" heig
<a href="/hardwarehacker">The Hardware Hacker</a>,
<a href="/facebookforseniors">Facebook for Seniors</a>,
<a href="/facebookforseniors"><img alt="Facebook for Seniors" class="img-responsive"
<a href="/facebookforseniors">Facebook for Seniors</a>,
<a href="/inventwithpython">Invent Your Own Computer Games with Python, 4th Edition</
<a href="/inventwithpython"><img alt="Invent Your Own Computer Games with Python" cla
<a href="/inventwithpython">Invent Your Own Computer Games with Python</a>,
<a href="/scratchcards">Scratch Coding Cards</a>,
<a href="/scratchcards"><img alt="Scratch Coding Cards" class="img-responsive" heigh
<a href="/scratchcards">Scratch Coding Cards</a>,
<a href="/beautifulminecraft">Beautiful Minecraft </a>,
<a href="/beautifulminecraft"><img alt="Beautiful Minecraft" class="img-responsive" l
<a href="/beautifulminecraft">Beautiful Minecraft</a>,
<a href="/wcss2">Wicked Cool Shell Scripts, 2nd Edition</a>,
<a href="/wcss2"><img alt="Wicked Cool Shell Scripts, 2nd Edition" class="img-respons
<a href="/wcss2">Wicked Cool Shell Scripts, 2nd Edition</a>,
<a href="/technicbuilder2">Unofficial LEGO Technic Builder's Guide, 2nd Edition</a>,
<a href="/technicbuilder2"><img alt="Unofficial LEGO Technic Builder's Guide, 2E" cla
<a href="technicbuilder2">The Unofficial LEGO Technic Builder's Guide, 2nd Ed.</a>,
<a class="feed-icon" href="https://www.nostarch.com/feeds/newbooks.xml" title="Subsc:
<a href="/upcoming.htm">see all</a>,
<a href="/gtfo">PoC||GTFO</a>,
<a href="/gtfo"><img alt="" class="img-responsive" height="225" src="https://www.nost
<a href="https://www.nostarch.com/gtfo">PoC||GTFO</a>,
<a href="/microprocessors">The Manga Guide to Microprocessors</a>,
<a href="/microprocessors"><img alt="The Manga Guide to Microprocessors" class="img-
<a href="/microprocessors">The Manga Guide to Microprocessors</a>,
<a href="/arduinohandbook2">Arduino Project Handbook, Vol. 2</a>,
```

```html
<a href="/arduinohandbook2"><img alt="Arduino Project Handbook Vol. 2" class="img-res
<a href="/arduinohandbook2">Arduino Project Handbook, Vol. 2</a>,
<a href="/thinkpython">Think Like a Programmer, Python Edition</a>,
<a href="/thinkpython"><img alt="Think Like a Programmer, Python Edition" class="img-
<a href="https://www.nostarch.com/thinkpython">Think Like a Programmer</a>,
<a href="/computersforseniors">Computers for Seniors</a>,
<a href="/computersforseniors"><img alt="Computers for Seniors" class="img-responsive
<a href="https://www.nostarch.com/computersforseniors">Computers for Seniors</a>,
<a href="/seriouscrypto">Serious Cryptography</a>,
<a href="/seriouscrypto"><img alt="Serious Cryptography" class="img-responsive" heigh
<a href="/SeriousCrypto">Serious Cryptography</a>,
<a href="/legoarchitecture">The LEGO Architecture Idea Book</a>,
<a href="/legoarchitecture"><img alt="The LEGO Architecture Idea Book" class="img-res
<a href="https://www.nostarch.com/legoarchitecture">The LEGO Architecture Idea Book</
<a href="/Rust">The Rust Programming Language</a>,
<a href="/Rust"><img alt="The Rust Programming Language" class="img-responsive" heigh
<a href="/Rust">The Rust Programming Language</a>,
<a href="/legotrains">LEGO Trains Book</a>,
<a href="/legotrains"><img alt="The LEGO Trains Book" class="img-responsive" height="
<a href="/legotrains">LEGO Trains Book</a>,
<a href="/networkprotocols">Attacking Network Protocols</a>,
<a href="/networkprotocols"><img alt="Attacking Network Protocols" class="img-respons
<a href="/networkprotocols">Attacking Network Protocols</a>,
<a href="/azure">Attacking Microsoft Azure</a>,
<a href="/azure"><img alt="Attacking Microsoft Azure" class="img-responsive" height="
<a href="/azure">Attacking Microsoft Azure</a>,
<a href="/missionpython">Mission Python</a>,
<a href="/missionpython"><img alt="Mission Python" class="img-responsive" height="225
<a href="/missionpython">Mission Python</a>,
<a href="/gameconsole">Game Console</a>,
<a href="/gameconsole"><img alt="The Game Console" class="img-responsive" height="225
<a href="/gameconsole">The Game Console</a>,
<a class="feed-icon" href="https://www.nostarch.com/feeds/comingsoon.xml" title="Subs
<a href="/mailchimp/subscribe">Want sweet deals? <br/>Sign up for our newsletter.</a
<a href="/aboutearlyaccess">Early Access</a>,
<a href="/arduinohandbook2">Arduino Project Handbook Vol. 2</a>,
<a href="/networkprotocols">Attacking Network Protocols</a>,
<a href="/crackingcodes">Cracking Codes with Python</a>,
<a href="/learnjava">Learn Java the Easy Way</a>,
<a href="/microprocessors">The Manga Guide to Microprocessors</a>,
<a href="/practicalsql">Practical SQL</a>,
<a href="/rootkits">Rootkits and Bootkits</a>,
<a href="/Rust">The Rust Programming Language</a>,
<a href="/seriouscrypto">Serious Cryptography</a>,
<a href="/thinkpython">Think Like a Programmer, Python Edition</a>,
<a href="/nsm" target="new">The Practice of Network Security Monitoring</a>,
<a href="https://cybercanon.paloaltonetworks.com/" target="new">Cybersecurity Canon<,
<a href="https://www.reddit.com/r/netsec/comments/5xli3d/hi_im_bunnie_author_of_the_
```

```html
<a href="http://www.vox.com/videos/2017/2/6/14445516/minecraft-art-business-maps" tar
<a href="/beautifulminecraft" target="new">Beautiful Minecraft</a>,
<a href="/hardwarehacker" target="new">The Hardware Hacker</a>,
<a href="/facebookforseniors" target="new">Facebook for Seniors</a>,
<a href="http://www.wsj.com/articles/the-best-geeky-books-to-kick-off-the-new-year-14
<a href="/scratchcards" target="new"><b>Scratch Coding Cards</b></a>,
<a href="https://www.iste.org/explore/articleDetail?articleid=864&amp;category=In-the
<a href="/legochristmas" target="new"><b>The LEGO Christmas Ornaments Book</b></a>,
<a href="/download/LEGOChristmasOrnaments_FamilyCircle.pdf" target="new">Family Circl
<a href="http://www.chicagotribune.com/lifestyles/holiday/giftguide/sc-gift-guide-chi
<a href="/wcss2" target="new">Wicked Cool Shell Scripts</a>,
<a href="https://opensource.com/article/16/12/calcshell-interactive-linux-command-lin
<a href="/legoanimation" target="new"><b>The LEGO Animation Book</b></a>,
<a href="http://www.theverge.com/a/holiday-gift-ideas-2016#lego-animation-book" targe
<a href="http://www.wsj.com/articles/three-books-every-geek-should-read-this-summer-1
<a href="/electronicsforkids" target="new">Electronics for Kids</a>,
<a href="/searchtale" target="new">The CS Detective</a>,
<a href="/tinylegowonders" target="new">Tiny LEGO Wonders</a>,
<a href="https://thecreatorsproject.vice.com/blog/diy-lego-engineering-book" target="
<a href="http://www.slj.com/2016/05/collection-development/read-watch-alikes/coding-l
<a href="/scratchjr" target="new">The Official ScratchJr Book</a>,
<a href="/teachkids" target="new">Teach Your Kids to Code</a>,
<a href="/doingmathwithpython" target="new">Doing Math with Python</a>,
<a href="/carhacking" target="new">The Car Hacker's Handbook</a>,
<a href="http://www.wsj.com/articles/3-must-read-books-for-geeks-1458057671?tesla=y"
<a href="/metasploit" target="new">David Kennedy</a>,
<a href="/hacking2.htm target=new">Jon Erickson</a>,
<a href="/malware" target="new">Michael Sikorski</a>,
<a href="/idapro2.htm" target="new">Chris Eagle</a>,
<a href="https://www.reddit.com/r/netsec/comments/3waa8x/im_bill_pollock_no_starch_pr
<a href="https://medium.com/scratchfoundation-blog/why-publish-books-about-scratch-7e
<a href="/legoarchitect" target="new"><b>The LEGO Architect</b></a>,
<a href="http://graphics.wsj.com/off-duty-50/a-guide-to-fall-2015/1278/be-the-corbusi
<a href="/smartgirlsguide" target="new"><b>The Smart Girl's Guide to Privacy</b></a>
<a href="https://twitter.com/violetblue/status/652989002677534720/photo/1" target="ne
<a href="https://twitter.com/billpollock/status/616127920042479616" target="new">our
<a href="https://twitter.com/aCaltum/status/596735294818385920" target="new">Hacking
<a href="feeds/newbooks.xml"><img align="left" alt="RSS Feed" border="0" src="/misc/:
<a href="feeds/comingsoon.xml"><img align="left" alt="RSS Feed" border="0" src="/misc
<a href="/mailchimp/subscribe"><img align="left" alt="Newsletter" border="0" src="/s:
<a href="https://www.twitter.com/nostarch"><img align="left" alt="Twitter" border="0"
<a href="https://www.facebook.com/home.php?#/pages/No-Starch-Press/57041709833"><img
<a class="footer-links" href="/about.htm">About Us</a>,
<a class="footer-links" href="/Jobs.htm">Jobs</a>,
<a class="footer-links" href="/distribution.htm">Sales and Distribution</a>,
<a class="footer-links" href="/rights">Rights</a>,
<a class="footer-links" href="/media.htm">Media</a>,
<a class="footer-links" href="/academic.htm">Academic Requests</a>,
```

```
        <a class="footer-links" href="/conferences.htm">Conferences</a>,
        <a class="footer-links" href="/orderfaq.htm">Order FAQ</a>,
        <a class="footer-links" href="/contactus.htm">Contact Us</a>,
        <a class="footer-links" href="/writeforus.htm">Write for Us</a>,
        <a class="footer-links" href="/privacypolicy.htm">Privacy</a>]

In [87]: lista_href = url.find_all('a')

In [90]: # Lista vacia
         resultados = []


         for item_a in lista_href:
             # Asignamos un valor a la lista
             resultados.append(str(item_a.get_text()))


In [91]: resultados

Out[91]: ['Skip to main content',
          '\n\n',
          'Catalog',
          'Media',
          'Write for Us',
          'About Us',
          'Art & Design',
          'General Computing',
          'Hacking & Computer Security',
          'Hardware / DIY',
          'Kids',
          'LEGO',
          'LEGOŏ MINDSTORMS',
          'Linux & BSD',
          'Manga',
          'Programming',
          'Python',
          'Science & Math',
          'System Administration',
          'EARLY ACCESS',
          'Free ebook edition with every print book purchased from nostarch.com!',
          '',
          'My account',
          'Log in',
          'Create account',
          '',
          '',
          '',
          '',
          '',
```

```
'',
"The Arduino Inventor's Guide",
'',
"The Arduino Inventor's Guide",
'Gray Hat C#',
'',
'Gray Hat C#',
'Coding iPhone Apps for Kids',
'',
'Coding iPhone Apps for Kids',
'Practical Packet Analysis, 3rd Edition',
'',
'Practical Packet Analysis, 3rd Ed.',
'Arduino Playground',
'',
'Arduino Playground',
'Hardware Hacker',
'',
'The Hardware Hacker',
'Facebook for Seniors',
'',
'Facebook for Seniors',
'Invent Your Own Computer Games with Python, 4th Edition',
'',
'Invent Your Own Computer Games with Python',
'Scratch Coding Cards',
'',
'Scratch Coding Cards',
'Beautiful Minecraft ',
'',
'Beautiful Minecraft',
'Wicked Cool Shell Scripts, 2nd Edition',
'',
'Wicked Cool Shell Scripts, 2nd Edition',
"Unofficial LEGO Technic Builder's Guide, 2nd Edition",
'',
"The Unofficial LEGO Technic Builder's Guide, 2nd Ed.",
'',
'see all',
'PoC||GTFO',
'',
'PoC||GTFO',
'The Manga Guide to Microprocessors',
'',
'The Manga Guide to Microprocessors',
'Arduino Project Handbook, Vol. 2',
'',
'Arduino Project Handbook, Vol. 2',
```

```
'Think Like a Programmer, Python Edition',
'',
'Think Like a Programmer',
'Computers for Seniors',
'',
'Computers for Seniors',
'Serious Cryptography',
'',
'Serious Cryptography',
'The LEGO Architecture Idea Book',
'',
'The LEGO Architecture Idea Book',
'The Rust Programming Language',
'',
'The Rust Programming Language',
'LEGO Trains Book',
'',
'LEGO Trains Book',
'Attacking Network Protocols',
'',
'Attacking Network Protocols',
'Attacking Microsoft Azure',
'',
'Attacking Microsoft Azure',
'Mission Python',
'',
'Mission Python',
'Game Console',
'',
'The Game Console',
'',
'Want sweet deals? Sign up for our newsletter.',
'Early Access',
'Arduino Project Handbook Vol. 2',
'Attacking Network Protocols',
'Cracking Codes with Python',
'Learn Java the Easy Way',
'The Manga Guide to Microprocessors',
'Practical SQL',
'Rootkits and Bootkits',
'The Rust Programming Language',
'Serious Cryptography',
'Think Like a Programmer, Python Edition',
'The Practice of Network Security Monitoring',
'Cybersecurity Canon',
'Reddit AMA',
'Vox',
'Beautiful Minecraft',
```

'The Hardware Hacker',
'Facebook for Seniors',
"Wall Street Journal's",
'Scratch Coding Cards',
'ISTE blog',
'The LEGO Christmas Ornaments Book',
'Family Circle Magazine',
'The Chicago Tribune',
'Wicked Cool Shell Scripts',
'Opensource.com',
'The LEGO Animation Book',
"The Verge's Holiday Gift Guide",
'The Wall Street Journal',
'Electronics for Kids',
'The CS Detective',
'Tiny LEGO Wonders',
'The Creators Project',
'School Library Journal',
'The Official ScratchJr Book',
'Teach Your Kids to Code',
'Doing Math with Python',
"The Car Hacker's Handbook",
"Wall Street Journal's",
'David Kennedy',
'Jon Erickson',
'Michael Sikorski',
'Chris Eagle',
'here',
'Medium post',
'The LEGO Architect',
"The Wall Street Journal's guide to Fall's Best",
"The Smart Girl's Guide to Privacy",
'BUST Magazine',
'our sticker',
'Hacking: The Art of Exploitation',
'New Releases',
'Coming Soon',
'Subscribe to Newsletter',
'Twitter',
'Facebook',
'About Us',
'Jobs',
'Sales and Distribution',
'Rights',
'Media',
'Academic Requests',
'Conferences',
'Order FAQ',

```
            'Contact Us',
            'Write for Us',
            'Privacy']

In [94]: serie_pandas = pd.Series(resultados)

In [98]: df_resultados = pd.DataFrame(serie_pandas)

In [102]: df_resultados = df_resultados[df_resultados[0] != 'Skip to main content']

In [106]: df_resultados = df_resultados[df_resultados[0] != '\n\n']

In [111]: df_resultados =  df_resultados[(df_resultados[0] != 'Catalog')
                                       & (df_resultados[0] != 'Media')
                                       & (df_resultados[0] != 'Write for Us')
                                       & (df_resultados[0] != '')
                                       & (df_resultados[0] != 'About Us')
                                       & (df_resultados[0] != 'My account')
                                       & (df_resultados[0] != 'Facebook')]
```

8   Selección por herencia DOM

8.0.1   Herencia DOM: tr > td > a

```
In [114]: hacker_news_url = 'https://news.ycombinator.com/news'

In [116]: hacker_news_url = BeautifulSoup(requests.get(hacker_news_url).text, "lxml")

In [127]: # Seleccionar todos los enlaces (a href='')

          # REGEX
          lista_href = hacker_news_url.select('tr > td > a[href*="."]')

In [128]: resultados = []

          for item in lista_href:
              # Asignamos un valor a la lista
              resultados.append(str(item.get_text()))

In [141]: df_hackernews = pd.DataFrame(pd.Series(resultados))

          df_hackernews = df_hackernews[df_hackernews[0] != '']

In [143]: df_hackernews

Out[143]:                                                  0
          1    SEC Files Fraud Charges in Bitcoin and Office ...
          2                                    Saved by Alice
          3      Building Pixels  A Daily Source of Inspiration
          4    What is skip locked for in PostgreSQL 9.5? (...
```

```
5              Why does Heap's algorithm work? (2016)
6    React, Relay and GraphQL: Under the Hood of Th...
7    WiFi232  An Internet Hayes Modem for Your Ret...
8    Teaching and Learning What Is Mathematics [pdf]
9    Deako (YC W16) Is Hiring a Back End and Front ...
10                  Cicada  Unix shell written in Rust
11   Striking Gemini Images Point Juno Spacecraft T...
12                     Dissecting objc_msgSend on ARM64
13   Myth-Maker of the Brothel: Utamaro's woodblock...
14                              What is a file? (2011)
15             NYPD is canceling its Palantir contract
16   Nevada becomes fifth US state to allow cannabi...
17      What Chinese bosses think of American workers
18                              Dolphin Progress Report
19   Solving the Mystery of Whose Laughter Is On th...
20                 The case of the 500-mile email (2002)
21                              The Accumulibrary (2014)
22   Making a good impression: the British Library'...
23                Microsofts Embedded Learning Library
24   What to Blame for Your Stomach Bug? Not Always...
25   Silicon Valley Women, in Cultural Shift, Frank...
26   I ask 100 information questions to four digita...
27   The Life, Death, and Legacy of iPhone Jailbrea...
28   In the summer of 1962, the U.S. blew up a hydr...
29                              Why We Chose Typescript
```