

MAGÍSTER EN DATA SCIENCE

Scraper de Biobio Chile

Proyecto 22 por cechiang, Carloslugook e iggypoprl

Pasos a seguir

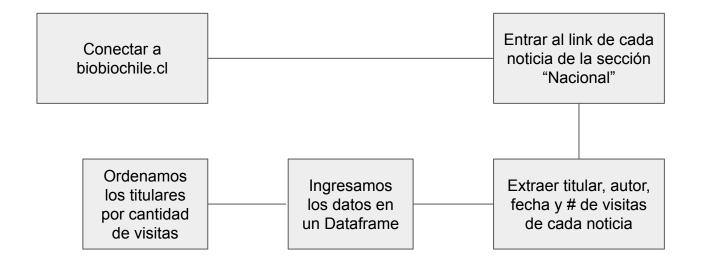
En el presente trabajo realizaremos web scraping de la página web https://www.biobiochile.cl/ con Beautiful Soup y Selenium.

Se programará un scraper con python para descargar información de las noticias nacionales con más vistas cada día, identificando el autor, la fecha de la noticia, las visitas y el titular con su link.





Pasos a seguir



Todo el proceso de web scraping y manipulación de datos lo vamos a realizar con Python con la librería de Beautiful Soup y Selenium.

El scraper podrá servir de base para usuarios que quieran realizar análisis de las noticias más visitadas de biobiochile.cl, así como también identificar quienes son los autores más populares y los días con más visitas: qué días generan más vistas, qué autores son los más visitados, a qué hora se generan las visitas, etc.

Pseudocódigo del scraper de BioBioChile

request biobiochile' url

click headline

extract headline link

for articulo in len(articulos):

extract these variables: amount of visits, date, headline, author

save the values in a list

arrange variables and values as pandas data frame

sort rows by amount of visits

download as csv

Parte principal del código

```
# Creamos un ciclo for para extraer la variables que queremos. Utilizamos Beautiful Soup y Selenium
  d = []
 for articulo in range(len(Articulos)):
      try:
          headline = Articulos[articulo].find('h2',{'class':"article-title"}).get_text().replace('"',"")
          Autor = Articulos[articulo].find('a',{'class':'article-author'}).get_text()
          Fecha = Articulos[articulo].find('div',{'class':'article-date-hour'}).get text().strip()
          Link = Articulos[articulo].find('a', href = re.compile(r'[/]([a-z]|[A-Z])\w+'))['href']
          driver = webdriver.Chrome(service=s)
          driver.get(Link)
          visitas = driver.find element(By.XPATH,"//span[@class='post-visits']").get attribute('innerHTML')
          driver.close()
          d.append((headline,Autor,visitas,Fecha,Link))
          #print(headline,"//",Autor,'//',Fecha,"//",Link,"//",visitas)
      except:
         pass
```

Primeras siete filas del data frame sin ordenar

	Headline	Autor	visitas	Link	Hora	Año	Dia_semana	Dia_numerico	Mes
0	Trabajadores de Inacap de Valdivia se manifest	Tamara Rojas	121	https://www.biobiochile.cl/noticias/nacional/r	21:23	2022	Lunes	12	Septiembre
1	Detienen a hombre acusado de asaltar a menor d	Tamara Rojas	142	https://www.biobiochile.cl/noticias/nacional/r	21:15	2022	Lunes	12	Septiembre
2	Chile Vamos desmiente acuerdo para proceso con	Diego Vera	1,272	https://www.biobiochile.cl/noticias/nacional/c	21:09	2022	Lunes	12	Septiembre
3	PDI confirma vínculo entre dos homicidios en C	Tamara Rojas	274	https://www.biobiochile.cl/noticias/nacional/r	21:02	2022	Lunes	12	Septiembre
4	Cámara aprueba prórroga del Estado de Excepció	Florencia Ortiz	1,369	https://www.biobiochile.cl/noticias/nacional/c	19:54	2022	Lunes	12	Septiembre
5	¿Se venció tu cédula de identidad? Habilitan p	Florencia Ortiz	2,315	https://www.biobiochile.cl/noticias/nacional/c	19:43	2022	Lunes	12	Septiembre
6	Reportan nuevo homicidio al interior de cárcel	Paola Valenzuela	3,270	https://www.biobiochile.cl/noticias/nacional/r	19:31	2022	Lunes	12	Septiembre
7	Roban \$30 millones en equipos de computación a	Paola Valenzuela	395	https://www.biobiochile.cl/noticias/nacional/r	19:16	2022	Lunes	12	Septiembre

Noticias ordenadas por cantidad vistas

Headline	Autor	visitas	Link	Hora	Año	Dia_semana	Dia_numerico	Mes
Roban \$30 millones en equipos de computación a	Paola Valenzuela	395.000	https://www.biobiochile.cl/noticias/nacional/r	19:16	2022	Lunes	12.0	Septiembre
PDI confirma vínculo entre dos homicidios en C	Tamara Rojas	274.000	https://www.biobiochile.cl/noticias/nacional/r	21:02	2022	Lunes	12.0	Septiembre
Detienen a hombre acusado de asaltar a menor d	Tamara Rojas	142.000	https://www.biobiochile.cl/noticias/nacional/r	21:15	2022	Lunes	12.0	Septiembre
Trabajadores de Inacap de Valdivia se manifest	Tamara Rojas	121.000	https://www.biobiochile.cl/noticias/nacional/r	21:23	2022	Lunes	12.0	Septiembre
Tohá anuncia ajuste a Estado de Excepción para	Felipe Reyes	6.022	https://www.biobiochile.cl/noticias/nacional/c	19:10	2022	Lunes	12.0	Septiembre
Baleado en El Bosque tiene 12 años: terminó he	Florencia Ortiz	3.935	https://www.biobiochile.cl/noticias/nacional/r	18:51	2022	Lunes	12.0	Septiembre
Reportan nuevo homicidio al interior de cárcel	Paola Valenzuela	3.270	https://www.biobiochile.cl/noticias/nacional/r	19:31	2022	Lunes	12.0	Septiembre
¿Se venció tu cédula de identidad? Habilitan p	Florencia Ortiz	2.315	https://www.biobiochile.cl/noticias/nacional/c	19:43	2022	Lunes	12.0	Septiembre
Cámara aprueba prórroga del Estado de Excepció	Florencia Ortiz	1.369	https://www.biobiochile.cl/noticias/nacional/c	19:54	2022	Lunes	12.0	Septiembre
	Roban \$30 millones en equipos de computación a PDI confirma vínculo entre dos homicidios en C Detienen a hombre acusado de asaltar a menor d Trabajadores de Inacap de Valdivia se manifest Tohá anuncia ajuste a Estado de Excepción para Baleado en El Bosque tiene 12 años: terminó he Reportan nuevo homicidio al interior de cárcel ¿Se venció tu cédula de identidad? Habilitan p Cámara aprueba prórroga del Estado de	Roban \$30 millones en equipos de computación a PDI confirma vínculo entre dos homicidios en C Detienen a hombre acusado de asaltar a menor d Trabajadores de Inacap de Valdivia se manifest Tohá anuncia ajuste a Estado de Excepción para Baleado en El Bosque tiene 12 años: terminó he Reportan nuevo homicidio al interior de cárcel ¿Se venció tu cédula de identidad? Habilitan p Cámara aprueba prórroga del Estado de Elorencia Ortiz	Roban \$30 millones en equipos de computación a PDI confirma vínculo entre dos homicidios en C Detienen a hombre acusado de asaltar a menor d Trabajadores de Inacap de Valdivia se manifest Tohá anuncia ajuste a Estado de Excepción para Baleado en El Bosque tiene 12 años: terminó he Reportan nuevo homicidio al interior de cárcel ¿Se venció tu cédula de identidad? Habilitan p Cámara aprueba prórroga del Estado de Elorencia Ortiz Elorencia Ortiz 1 395.000 Tamara Rojas 274.000 Tamara Rojas 121.000 Felipe Reyes 6.022 Florencia Ortiz 3.935 Florencia Ortiz 2.315	Roban \$30 millones en equipos de computación a PDI confirma vínculo entre dos homicidios en C Tamara Rojas 274.000 https://www.biobiochile.cl/noticias/nacional/r Detienen a hombre acusado de asaltar a menor d Trabajadores de Inacap de Valdivía se manifest Tohá anuncia ajuste a Estado de Excepción para Baleado en El Bosque tiene 12 años: terminó he Reportan nuevo homicidio al interior de cárcel ¿Se venció tu cédula de identidad? Habilitan p Cámara aprueba prórroga del Estado de El Paola Valenzuela Cámara aprueba prórroga del Estado de El Paola Critiz Paola Valenzuela 395.000 https://www.biobiochile.cl/noticias/nacional/r https://www.biobiochile.cl/noticias/nacional/r Paola Valenzuela 3.270 https://www.biobiochile.cl/noticias/nacional/r Eliprencia Ortiz 2.315 https://www.biobiochile.cl/noticias/nacional/c	Roban \$30 millones en equipos de computación a Paola Valenzuela 395.000 https://www.biobiochile.cl/noticias/nacional/r 19:16 PDI confirma vínculo entre dos homicidios en C Tamara Rojas 274.000 https://www.biobiochile.cl/noticias/nacional/r 21:02 Detienen a hombre acusado de asaltar a menor d Tamara Rojas 142.000 https://www.biobiochile.cl/noticias/nacional/r 21:15 Trabajadores de Inacap de Valdivia se manifest Tamara Rojas 121.000 https://www.biobiochile.cl/noticias/nacional/r 21:23 Tohá anuncia ajuste a Estado de Excepción para Felipe Reyes 6.022 https://www.biobiochile.cl/noticias/nacional/c 19:10 Baleado en El Bosque tiene 12 años: terminó he Florencia Ortiz 3.935 https://www.biobiochile.cl/noticias/nacional/r 18:51 Reportan nuevo homicidio al interior de cárcel Valenzuela Valenzuela 3.270 https://www.biobiochile.cl/noticias/nacional/r 19:31 Cámara aprueba prórroga del Estado de Elorencia Ortiz 1.369 https://www.biobiochile.cl/noticias/nacional/c 19:54	Roban \$30 millones en equipos de valenzuela a 395.000 https://www.biobiochile.cl/noticias/nacional/r 19:16 2022 PDI confirma vínculo entre dos homicidios en C Tamara Rojas 274.000 https://www.biobiochile.cl/noticias/nacional/r 21:02 2022 Detienen a hombre acusado de asaltar a menor d Tamara Rojas 142.000 https://www.biobiochile.cl/noticias/nacional/r 21:15 2022 Trabajadores de Inacap de Valdivia se manifest Tamara Rojas 121.000 https://www.biobiochile.cl/noticias/nacional/r 21:23 2022 Tohá anuncia ajuste a Estado de Excepción para Felipe Reyes 6.022 https://www.biobiochile.cl/noticias/nacional/r 19:10 2022 Baleado en El Bosque tiene 12 años: terminó he Florencia Ortiz terminó he Valenzuela Valenzuela Valenzuela Valenzuela 2.315 https://www.biobiochile.cl/noticias/nacional/r 19:31 2022 Cámara aprueba prórroga del Estado de Florencia Ortiz 1.369 https://www.biobiochile.cl/noticias/nacional/c 19:54 2022	Roban \$30 millones en equipos de computación a Paola Valenzuela 395.000 https://www.biobiochile.cl/noticias/nacional/r 19:16 2022 Lunes PDI confirma vínculo entre dos homicidios en C Tamara Rojas 274.000 https://www.biobiochile.cl/noticias/nacional/r 21:02 2022 Lunes Detienen a hombre acusado de asaltar a menor d Tamara Rojas 142.000 https://www.biobiochile.cl/noticias/nacional/r 21:15 2022 Lunes Trabajadores de Inacap de Valdivía se manifest Tamara Rojas 121.000 https://www.biobiochile.cl/noticias/nacional/r 21:23 2022 Lunes Tohá anuncia ajuste a Estado de Excepción para Felipe Reyes 6.022 https://www.biobiochile.cl/noticias/nacional/r 19:10 2022 Lunes Baleado en El Bosque tiene 12 años: terminó he Florencia Ortiz 3.935 https://www.biobiochile.cl/noticias/nacional/r 18:51 2022 Lunes & Valenzuela 2.327 https://www.biobiochile.cl/noticias/nacional/r 19:31 2022 Lunes Câmara aprueba prórroga del Estado de Florencia Ortiz 1.369 https://www.biobiochile.cl/noticias/nacional/c 19:43 2022 Lunes	Roban \$30 millones en equipos de computación a PDI confirma vínculo entre dos homicidios en C Tamara Rojas 274.000 https://www.biobiochile.cl/noticias/nacional/r 21:02 2022 Lunes 12.0 12.0 12.0 12.0 12.0 12.0 12.0 12.0

csv final



Cantidad de filas: 19 por día

Cantidad de columnas: 9

El paso siguiente es scrapear diariamente e ir añadiendo estas observaciones al DataFrame inicial