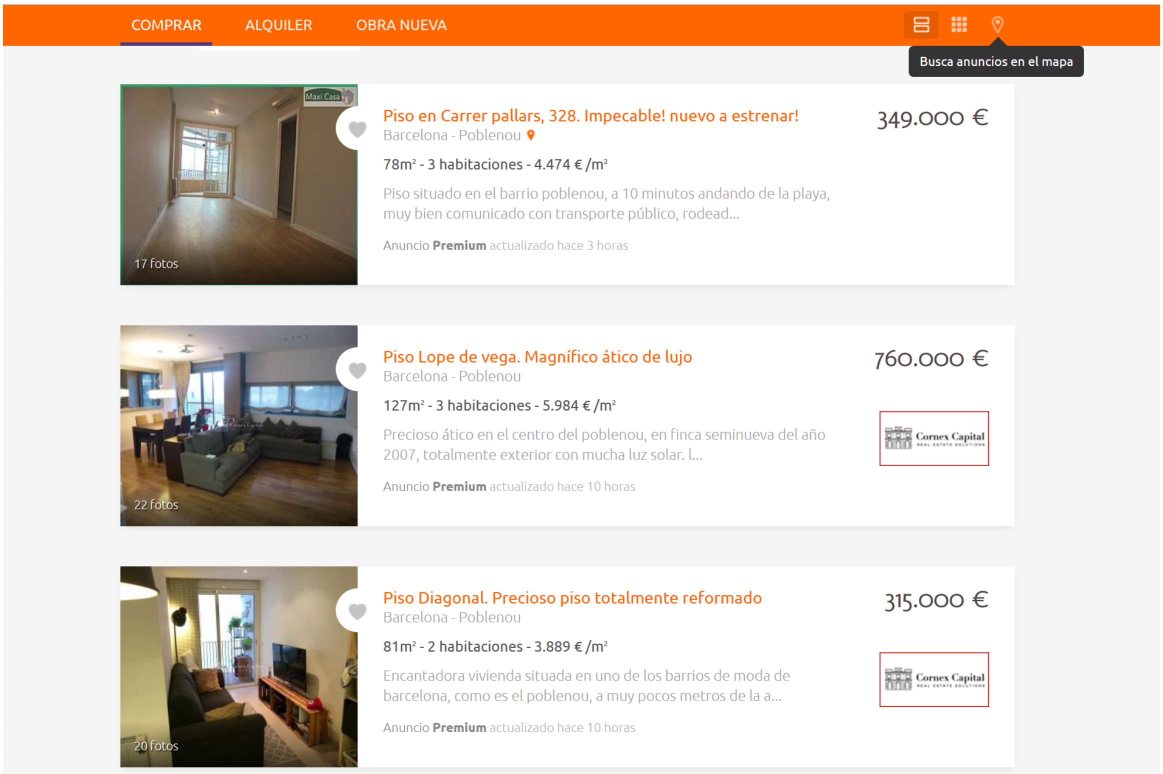
Tips for Web Scraping in Python with `bs4`

José María Lago Alonso

I've seen in the seminars that the less intuitive part of web-scraping via `bs4` appears when we try to scrap in the very deep of a nested structure, like in the following example:



17 fotos

Piso en Carrer pallars, 328. Impecable! nuevo a estrenar!
Barcelona - Poblenou
78m² - 3 habitaciones - 4.474 € /m²
Piso situado en el barrio poblenou, a 10 minutos andando de la playa, muy bien comunicado con transporte público, rodead...
Anuncio **Premium** actualizado hace 3 horas
349.000 €

22 fotos

Piso Lope de vega. Magnífico ático de lujo
Barcelona - Poblenou
127m² - 3 habitaciones - 5.984 € /m²
Precioso ático en el centro del poblenou, en finca seminueva del año 2007, totalmente exterior con mucha luz solar. l...
Anuncio **Premium** actualizado hace 10 horas
760.000 €

20 fotos

Piso Diagonal. Precioso piso totalmente reformado
Barcelona - Poblenou
81m² - 2 habitaciones - 3.889 € /m²
Encantadora vivienda situada en uno de los barrios de moda de barcelona, como es el poblenou, a muy pocos metros de la a...
Anuncio **Premium** actualizado hace 10 horas
315.000 €

Then the extra tools to scrap this data are:

```python
from urllib.request import urlopen
from bs4 import BeautifulSoup, NavigableString, Comment

url = "http://www.habitaclia.com/comprar-vivienda-en-barcelona-barrio_poblenou/provincia_barcelona-barcelones-area_6-sant_marti/listainmuebles.htm"

##Make the soup
page = urlopen(url) ## is up to your connection the URLError: <urlopen error [WinError 10060]
soup = BeautifulSoup(page,"lxml")

casts = soup.find_all('ul', attrs={'class': 'enlista'})

cast = casts[0]
lista = []
for a in cast.find_all('a'):
    for child in a.children:
        if isinstance(child, NavigableString) and not isinstance(child, Comment) and str(child).strip() != "":
            lista.append('{}'.format(str(child).strip()))
```

```python

newlist = []
for i in range(len(lista)):
    if "si baja" in lista[i] or "a favoritos" in lista[i] or "Ver mapa" in lista[i]:
        ""
    else:
        newlist.append(lista[i])
```

With this we capture the names of the adds in habitaclia.
The underlying idea of this example is to see the new objects `NavigableString` and also the `Comment`, so finally we have attributes like `children` that helps us to scrap in a nested way.