Übung4 - Website scrapen mit BeautifulSoup

1. Wählt eine Website aus, die ihr gerne analysieren wollt. Im eben angeschauten Beispiel, der Watson News-Website, ging es darum, systematisch zu erfassen wieviele Kommentare die Artikel bekommen.

Entrée [25]:
```
import requests
import pandas as pd
from bs4 import BeautifulSoup
```

Entrée [26]:
```
r = requests.get('https://www.24heures.ch/')
content = r.text
soup = BeautifulSoup(content,"html.parser")
```

Entrée [27]:
```
print(soup.prettify())
```

```
<!DOCTYPE html>
<html lang="fr">
 <head>
  <meta content="initial-scale=1.0, width=device-width" name="
viewport"/>
  <link href="https://feed-prod.unitycms.io" rel="preconnect"/
>
  <link href="https://feed-prod.unitycms.io" rel="dns-prefetc
h"/>
  <link href="https://tdn.da-services.ch" rel="preconnect"/>
  <link href="https://tdn.da-services.ch" rel="dns-prefetch"/>
  <link href="https://fonts.gstatic.com/" rel="preconnect"/>
  <link href="https://fonts.gstatic.com/" rel="dns-prefetch"/>
  <link href="https://fonts.googleapis.com/" rel="preconnect"/
>
  <link href="https://fonts.googleapis.com/" rel="dns-prefetc
h"/>
  <link href="https://astronomixus.prod.tda.link/" rel="precon
nect"/>
```

Entrée [28]:
```
titre = soup.find_all('span', {'ArticleTeaser_title__1Xvn1'})
```

Entrée [29]:
```
commentaires1 = soup.find_all('span', {'class':'ArticleTeaserM
```

Entrée [33]:
```
liste1 = []
for comm in commentaires1:
    x = comm.find_all('span')
    liste1.append(x)
```

Entrée [34]:
```
liste1
```

Out[34]:
```
[[<span>1</span>],
 [<span>1</span>],
 [<span>5</span>],
 [<span>17</span>],
 [<span>1</span>],
 [<span>2</span>],
 [<span>2</span>],
 [<span>5</span>],
 [<span>18</span>],
 [<span>4</span>]]
```

Entrée [35]:
```
storybox = soup.find_all('div', {'class':'ArticleTeaser_text__
```

Entrée [36]:
```
storybox[0]
```

Out[36]:
```
<div class="ArticleTeaser_text__lFTLc"><div><h3><span class="A
rticleTeaser_titleheaderbox__2aeik"><span class="ArticleTeaser
_premium__X9wjN">Abo</span><span class="ArticleTeaser_titlehea
der__2-H61">Urbanisme</span></span><span class="ArticleTeaser_
title__1Xvn1">Et voilà la future place du Marché de Vevey</spa
n></h3><p class="ArticleTeaser_lead__2rFcH">Après quatre ans
d'études et de processus participatifs, le projet est à l'enqu
ête publique durant un mois dès ce mardi. La fin de la réalisa
tion est prévue pour 2024.</p></div><div class="ArticleTeaserM
eta_root__3z_C8 ArticleTeaser_article-teaser-meta__Zur0N"><div
class="ArticleTeaserMeta_infowrapper__2pCAc"><time class="Rela
tiveDateTime_root__29Kdb" datetime="2020-09-28T13:23:11.000Z">
il y a 1 heure</time></div><span class="ArticleTeaserMeta_comm
ents__2ILfd ArticleTeaserMeta_whitestroke__TMOy8"><svg height
="17px" viewbox="0 0 17 17" width="17px" xmlns="http://www.w3.
org/2000/svg"><g fill="none" fill-rule="evenodd" stroke="none"
stroke-width="1"><path d="M4.64897677,15.0981244 L8.00357734,1
3.0339463 L8.14518969,13.034 L10.232604,13.034792 C13.1417065,
13.034792 15.5,10.6764985 15.5,7.76739599 C15.5,4.85829351 13.
1417065,2.5 10.232604,2.5 L6.76739599,2.5 C3.85829351,2.5 1.5,
4.85829351 1.5,7.76739599 C1.5,9.77836557 2.63669012,11.585891
3 4.40103623,12.4747297 L4.67993567,12.615233 L4.64897677,15.0
981244 Z" stroke="#808080"></path></g></svg><span>1</span></sp
an></div></div>
```

Entrée [38]:
```
liste2 = []
for elem in storybox:
    try:
        t = elem.find('span', {'ArticleTeaser_title__1Xvn1'}).
    except:
        t = "N/A"

    try:

        k = elem.find('span', {'class':'ArticleTeaserMeta_comm
    except:
        k = "N/A"
```

```
    mini_dict = {'Titre': t,
                 'Commentaires': k}

    liste2.append(mini_dict)
```

Entrée [40]:
```
df = pd.DataFrame(liste2)
```

Entrée [42]:
```
df[df['Commentaires']!="N/A"]
```

Out[42]:

	Titre	Commentaires
0	Et voilà la future place du Marché de Vevey	1
11	Lausanne, capitale des orgues	1
29	L'actualité croquée par nos dessinateurs	5
43	Covid-19: «Aucune raison de se ruer sur l'échi...	17
51	Bastian Baker s'est métamorphosé en gladiateur...	1
61	Le premier repas de Franck Pelux au Lausanne P...	2
62	Ce Vaudois passionné de vélos vintage	2
63	«Je suis un peu stressé, c'est une première re...	5
64	Des cours de «toutou paddle» pour pagayer avec...	18
65	«À un moment donné, on a peut-être trop traité»	4