I: WEBSCRAPING MODS203

January 31, 2022

[17]:
```
#Useful imports
import time
time_duration = 0.5
import requests
from bs4 import BeautifulSoup

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt
```

0.1 I/ COLLECTING DATA

[18]:
```
#We import selenium driver
from selenium import webdriver
from selenium.webdriver.chrome.options import Options

options = Options()
options.add_argument("--window-size=1920x1080")
options.add_argument("--verbose")
options.add_argument('-no-sandbox')
options.add_argument('-headless')



from selenium.webdriver.common.by import By

driver = webdriver.Chrome(options=options,)
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
```

0.1.1 I.1/We list all the categories of collects

[19]:
```
#To achieve this, we go on the homepage of gofundme and find the proper soup

r = requests.get('https://www.gofundme.com/fr-fr')
```

```
html = r.text
soup = BeautifulSoup(html, 'html.parser')
a = soup.find_all('a',class_='section-categories-icon cell text-center')
Categories = []
for link in a:
    carac = link.get('href')
    if carac[7]=="s":
        Categories.append(carac[13:])
    elif carac[7]== 'd':
        Categories.append(carac[16:])
    #Page_Links.append(link.get('href'))
    #print(link.get('href'))
    #print(a.get('href'))
for i in range(len(Categories)):
    Categories[i] = Categories[i].replace('fundraising','fundraiser')
Categories

#print(len(Page_Links))
```

[19]:
```
['medical-fundraiser',
 'memorial-fundraiser',
 'emergency-fundraiser',
 'charity-fundraiser',
 'education-fundraiser',
 'animal-fundraiser',
 'environment-fundraiser',
 'business-fundraiser',
 'community-fundraiser',
 'competition-fundraiser',
 'creative-fundraiser',
 'event-fundraiser',
 'faith-fundraiser',
 'family-fundraiser',
 'sports-fundraiser',
 'travel-fundraiser',
 'volunteer-fundraiser',
 'wishes-fundraiser']
```

0.1.2 I.2.1/ Collecting name, amount collected, url and town using Selenium on each category's homepage

We use Selenium and BeautifulSoup in this section because we have to click on a button 'More projects'

[20]:
```
data_name = []
data_collect = []
data_links = []
data_town = []
```

```
### FOR EACH CATEGORY, WE GO ON ITS HOMEPAGE
for i in range (len(Categories)):

    l = "https" + ":"+"//www.gofundme.com/discover/" + Categories[i]
    driver=webdriver.Chrome('chromedriver',options=options)
    driver.get(l)

    #Each click on the button "more" add 12 projects,
    #we click 8 times per categories to have 1728 projects in total
    ButtonMore=driver.find_element(By.CSS_SELECTOR,"div.text-center")
    for i in range(8):
        time.sleep(1.1)
        ButtonMore.click()



    soup = BeautifulSoup(driver.page_source, 'html.parser')


    #AmountCollected and Name
    a = soup.findAll('div',class_="campaign-tile-img--contain js-lazy")
    collect = soup.findAll('div',class_="show-for-medium")
    for text in a:
        data_name.append(text.get('aria-labelledby'))
    for m in range (1,len(collect)+1,2):
        money = collect[m].get_text()
        data_collect.append(money[1:])

    #Links
    Links = soup.findAll('a',class_='fund_tile_card_link')
    for link in Links:
        data_links.append(link.get('href'))

    #Town
    Town = soup.findAll('div',class_="fund-item fund-location␣
 ↪truncate-single-line")
    for town in Town:
        ville = town.find('span',class_="")
        data_town.append(ville.get_text())



###Cleaning text
for i in range (len(data_collect)):
    data_collect[i] = data_collect[i].replace('\xa0','')
```

```
for i in range(len(data_name)):
    data_name[i] = data_name[i].replace('-',' ')
```

[21]:
```
print(len(data_collect))
print(len(data_links))
print(len(data_name))
print(len(data_town))
```

```
1716
1716
1716
1716
```

0.1.3 I.2.2/Collecting Pourcentage of target collected, date creation, category

Now that we have the link of each project, we go on each's specific webpage to collect more features with BeautifulSoup

[22]:
```
Cat = []
Pourcentage_Raised = []
Creation_date = []
Description = []
NumbersOfDonors=[]


### FOR EACH PROJECT, WE GO ON ITS OWN PAGE
for e in data_links :
    r = requests.get(e)
    html = r.text
    soup = BeautifulSoup(html, 'html.parser')

    #Category
    a = soup.find('a',class_='m-campaign-byline-type divider-prefix␣
↪meta-divider flex-container align-center color-dark-gray hrt-tertiary-button␣
↪hrt-base-button hrt-link hrt-link--gray-dark hrt-link--unstyled')
    if not a:
        Cat.append('None')
    else:
        cat = a.get_text()
        Cat.append(cat)

    #Progress
    progress = soup.find('progress',class_='a-progress-bar␣
↪a-progress-bar--green')
    if not progress:
        Pourcentage_Raised.append("None")
    else:
        pourcentage_funding = progress.get('value')
        pourcentage_funding_float = round(float(pourcentage_funding),1)
```

```
        if (pourcentage_funding_float > 100) :
          pourcentage_funding_float = 100
        Pourcentage_Raised.append(pourcentage_funding_float)

    #Date of Creation
    orga = soup.find('span',class_='m-campaign-byline-created a-created-date')
    if not orga:
        Creation_date.append('None')
    else:
        Creation_date.append(orga.get_text())

    #Descritpion
    a = soup.find('div',class_="o-campaign-description")
    description=''
    if not a:
        Description.append('None')
    else:
        for x in a:
            text= x.get_text()
            description += text
        description  = description.replace('\n','')
        description  = description.replace('\xa0','')
        Description.append(description)

    #Number of donors
    a = soup.findAll('script')
    b = str(a)
    indice = b.index('donation_count')
    i=0
    stringNumber = ''
    while b[indice + 16+i] != ",":
        localNumber = b[indice+16+i]
        stringNumber += localNumber
        i = i+1
    NumbersOfDonors.append(int(stringNumber))
```

0.1.4 I.3.1/Gathering in one DF

[23]:
```
data_final = []
for i in range(len(data_links)):
    data_final.append({
        'title' : data_name[i],
        'Categorie' : Cat[i],
        'collect': data_collect[i],
        'town': data_town[i],
        'Pourcentage Raised' : str(Pourcentage_Raised[i]) + "%",
        'Creation Date' : Creation_date[i],
```

```
        'Description' : Description[i],
        'NumberDonors' : NumbersOfDonors[i]


    })
```

[24]:
```
df= pd.DataFrame(data_final)
df
```

[24]:

	title	Categorie	collect	town
0	mammectomie hugo	Medical, Illness & Healing	€2,225 raised of €5,000	Vincennes
1	i miss my life	Medical, Illness & Healing	€506 raised of €10,000	Paris
2	help a mom in coma wake up again	Medical, Illness & Healing	€6,345 raised of €1	Metz
3	aidez gladys pour un ultime traitement	Medical, Illness & Healing	€5,667 raised of €100,000	Héricy
4	aide pour mon papa frais mdicaux opration	Medical, Illness & Healing	€4,889 raised of €10,000	Évreux
…	…	…	…	…
1711	jaimerais raliser le rve de ma cousine	Dreams, Hopes & Wishes	€0 raised of €3,000	Gaillac-d'Aveyron
1712	i need 1 euro from you to be richer	Dreams, Hopes & Wishes	€0 raised of €1.0B	Noisy-le-Sec
1713	ma fille veut sacheter le tlphone de ses rves	Dreams, Hopes & Wishes	€0 raised of €500	Le Bignon
1714	iphone pal ema	Dreams, Hopes & Wishes	€0 raised of €990	Nantes
1715	financement pour acheter haaland les frres	Dreams, Hopes & Wishes	€0 raised of €1.0B	Mulhouse

```
     Pourcentage Raised             Creation Date  \
0                 44.5%        Created 6 days ago
1                  5.1%         Created 1 day ago
2                  100%  Created January 11, 2022
3                  5.7%  Created January 14, 2022
4                 48.9%  Created December 6, 2021
…                 …                         …
1711               1.0%  Created January 21, 2022
1712               1.0%  Created January 21, 2022
1713               1.0%  Created January 20, 2022
1714               1.0%  Created January 20, 2022
1715               1.0%  Created January 20, 2022

                                            Description  NumberDonors
0     Vous me connaissez pour la plus part d'Instagr…           100
1     Hellohope you all doing greati;m Hossam from m…            29
2     Your support is needed !A Mum of six, 58 years…           139
3     «Tout le monde est faible devant la souffrance…            63
4     Chers tous,Mon papa est actuellement en visite…           126
…                                                  …           …
1711  Bonjour, je m'appelle trottein marine je souha…             0
1712  Hello mofos here is the thing . I'm sick of wo…             0
1713  Bonjour, je m'appelle Margot et j'ai 14 ans. J…             0
1714  necesita un iPhone jelpHHHHHHHHHHHHHHHHHHHHHHR…            0
1715  Bonjour je suis Rayane Badaoui si je vous cont…             0

[1716 rows x 8 columns]
```

0.1.5 I.3.2/ Export in csv to clean (II) and analyse (III)

[25]:
```
#Export
df.to_csv("DataFinalfromWebScraping2.csv")
```