Pràctica 1: Com podem capturar les dades de la web?

Tipologia i cicle de vida de les dades.

Alumnes: Marc Clupés Peró i Paula Miralles Simó

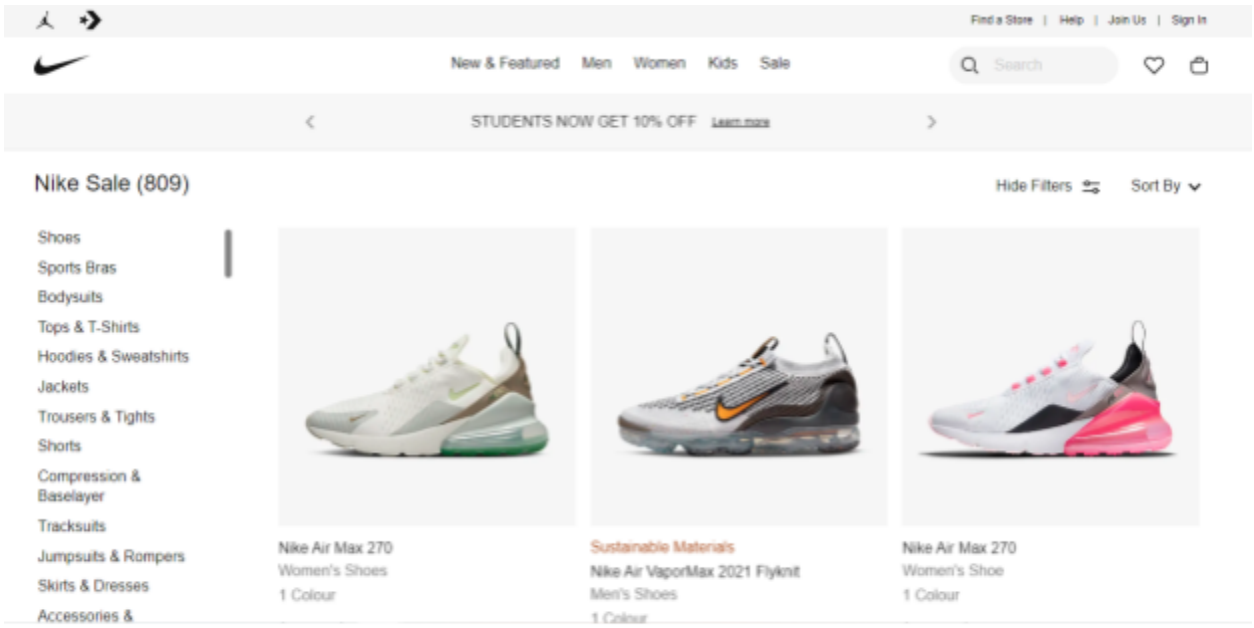
Data: 22/11/2022

Índex.

1. Context

Per tal de portar a terme la pràctica, hem creat un dataset a partir de les dades contingudes en un lloc web. En el nostre lloc, tindrem un dataset de característiques de productes de la marca Nike.

Aquest dataset naix de la recollida de dades de la pàgina web de la marca, en concret, de la secció de productes rebaixats (*https://www.nike.com/ca/w/sale-3yaep*). En el nostre cas, ha sigut l'elecció de la pàgina web la que ens ha portat a triar el tipus de dataset que volíem construir, doncs aquest lloc web ens donava moltes facilitats en quant a extracció de dades i altres facilitats relacionades amb el web scraping.



Imatge 1: Lloc web emprat (https://www.nike.com/ca/w/sale-3yaep)

2. Títol

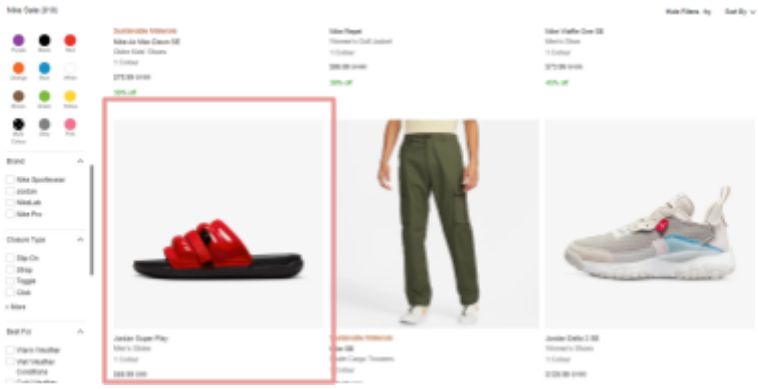
El títol d'aquest dataset es: **Productes rebaixats a Nike.**

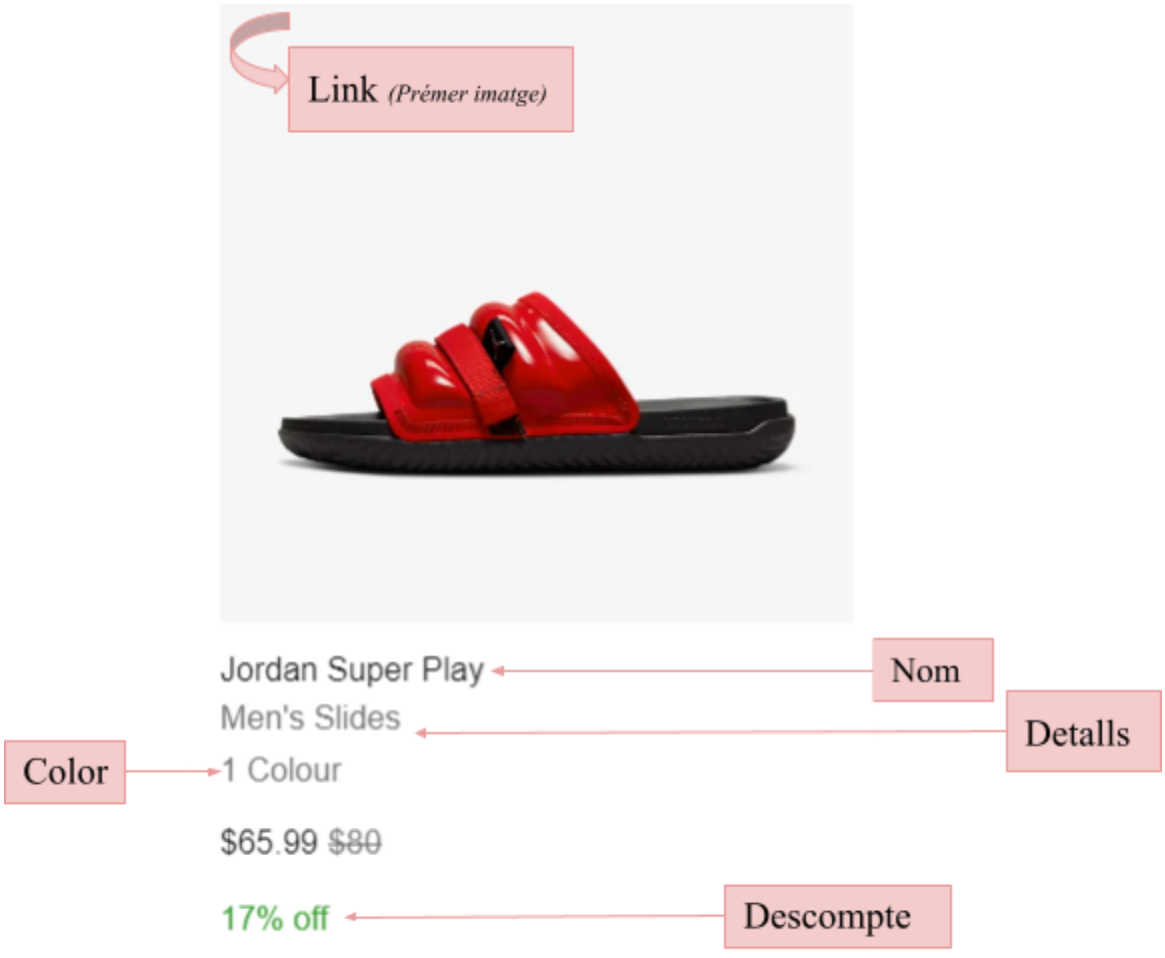
3. Descripció del dataset

El dataset recull especificacions concretes de cada un dels productes en rebaixes que la pàgina de Nike Sale ofereix. Aquestes especificacions es troben relacionades entre si al lloc web i el nostre web scrapping les recull i les posa en comú en un conjunt de dades útils.

4. Representació gràfica

Per a tractar de representar la informació que hem extret, la imatge 2 relaciona cada un dels detalls del producte amb la columna que se li ha assignat.





Imatge 2: Descripció del dataset

5. Contingut

El dataset conté les següents columnes:

Link	Conté el enllaç directe al producte en rebaixes
Nom	Conté el nom que la marca ha donat a este producte
Detalls	Conté especificacions sobre tipus de producte i génere
Colors	Conté el nombre de colors que té el producte
Descompte	Conté el descompte aplicat al producte

Respecte al període de temps de les dades, el dataset té una característica molt especial, i es que és una base de dades molt volàtil en quant al temps, és a dir, quan en un futur es treballi amb el dataset, s'ha de tindre en compte el fet de que els productes en rebaixes de Nike canvien a cada minut, doncs depèn de les compres dels clients i de les decisions de la marca de posar en descompte uns productes o altres.

Per això, este tipus de base de dades s'ha d'utilitzar per a realitzar estudis amb dependència del temps.

6. Propietari

El propietari del lloc web, segons el codi:

```
import whois
print(whois.whois('https://www.nike.com/ca/w/sale-3yaep'))
```

Es:


```
{
  "domain_name": [
    "NIKE.COM",
    "nike.com"
  ],
  "registrar": "MarkMonitor, Inc.",
  "whois_server": "whois.markmonitor.com",
  "referral_url": null,
  "updated_date": "2022-02-01 09:11:14",
  "creation_date": "1995-03-04 05:00:00",
  "expiration_date": [
    "2024-03-05 05:00:00",
    "2024-03-05 00:00:00"
  ],
  "name_servers": [
    "NS-N1.NIKE.COM",
    "NS-N2.NIKE.COM",
```

```
    "NS-N3.NIKE.COM",
    "NS-N4.NIKE.COM",
    "ns-n1.nike.com",
    "ns-n4.nike.com",
    "ns-n3.nike.com",
    "ns-n2.nike.com"
  ],
  "status": [
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited",
    "serverDeleteProhibited https://icann.org/epp#serverDeleteProhibited",
    "serverTransferProhibited https://icann.org/epp#serverTransferProhibited",
    "serverUpdateProhibited https://icann.org/epp#serverUpdateProhibited",
    "clientUpdateProhibited (https://www.icann.org/epp#clientUpdateProhibited)",
    "clientTransferProhibited (https://www.icann.org/epp#clientTransferProhibited)",
    "clientDeleteProhibited (https://www.icann.org/epp#clientDeleteProhibited)",
    "serverUpdateProhibited (https://www.icann.org/epp#serverUpdateProhibited)",
    "serverTransferProhibited (https://www.icann.org/epp#serverTransferProhibited)",
    "serverDeleteProhibited (https://www.icann.org/epp#serverDeleteProhibited)"
  ],
  "emails": [
    "abusecomplaints@markmonitor.com",
    "internet.domain.administrator@nike.com",
    "whoisrequest@markmonitor.com"
  ],
  "dnssec": "unsigned",
  "name": "Internet Domain Administrator",
  "org": "Nike, Inc.",
  "address": "One Bowerman Drive, DF/4",
  "city": "Beaverton",
  "state": "OR",
  "registrant_postal_code": "97005",
  "country": "US"
}
```


Aleshores, considerem que el database té dos propietaris fonamentals: el primer, és el propietari del lloc web que estem analitzant, en este cas, **Nike Inc**.

Per una altra part, els altres propietaris del database seriem aquells que l'hem construït, en este cas, Marc i Paula.

7. Inspiració

Aquesta base de dades respon, principalment, a la pregunta que la defineix i és, quin són els productes que es troben en rebaixes a Nike.

Però com s'explicava en anteriors punts, es tracta de productes en rebaixes en el moment puntual en el que es va construir la base de dades. Per això, és important treballar amb ella sabent esta data i ens inspira pensar que es poden analitzar els productes en rebaixes de Nike segons l'època de l'any en la que ens trobem.

És aquesta part la que ens va inspirar, la possibilitat del poder fer anàlisis del tipus "Quin és el millor mes per a comprar productes de la marca Nike?". Si tenim una base de dades

dels productes en rebaixes de Nike de Març de 2022 i una base de dades dels productes en rebaixes de Nike en Setembre de 2022, podem arribar a algun tipus de conclusió utilitzant les nostres dades.

8. Llicència

El propietaris d'aquesta base de dades elegim seleccionem la llicència **Released Under CC0: Public Domain License**. CC0 ens permet als científics, educadors, artistes i altres creadors i propietaris de bases de dades renunciar a aquests interessos en les nostres obres i, d'aquesta manera, situar-los el més completament possible en el domini públic, de manera que altres puguin construir, millorar i reutilitzar lliurement el nostre treball per a qualsevol propòsit sense restriccions.

9. Codi

El web scraping s'inicia important les llibreries necessàries per a portar a terme la nostra tasca:

```
from bs4 import BeautifulSoup
import pandas as pd
import time
import requests
```

Posteriorment, activarem el driver encarregat de obrir la pàgina principal, i el posarem en funcionament:

```
driver = webdriver.Chrome(
        'C:/Users/marcc/Desktop/UOC/TIPOLOGIA I CICLE DE VIDA/chromedriver_win32')

url = 'https://www.nike.com/ca/w/sale-3yaep'
driver.get(url)
```

Després, farem una petició HTTP per a confirmar que la resposta és del tipus 2XX que, com ens indica a la teoria de l'assignatura, aquesta classe de codi d'estat indica que la petició ha estat rebuda correctament, entesa i acceptada.

```
page = requests.get(url)
print(page)
```

Una vegada a la pàgina, ens trobarem en el primer problema: les cookies. Premem el botó d'acceptar cookies per a poder fer scroll dins de la pàgina:

```
button = driver.find_element(By.XPATH,
'//*[@id="gen-nav-commerce-header-v2"]/div[1]/div/div[2]/div/div[2]/div[2]/button').click()
```

Una vegada les cookies estiguin acceptades i se'ns permeti l'scroll, ho farem fins que arribi al final:

```
last_height = driver.execute_script('return document.body.scrollHeight')
while True:
    driver.execute_script('window.scrollTo(0,document.body.scrollHeight)')
    time.sleep(3)
    new_height = driver.execute_script('return document.body.scrollHeight')
    if new_height == last_height:
        break
    last_height = new_height
```

Importarem el HTML del lloc web al Python i, a més, agafarem també el html de cada producte.

```
soup = BeautifulSoup(driver.page_source, 'lxml')
```

```
product_card = soup.find_all('div', class_ = 'product-card__body')
```

Generarem el dataframe a emplenar definint el nom dels headers.

```
df = pd.DataFrame({'Link':[''], 'Nom':[''], 'Detalls':[''], 'Colors':[''], 'Descompte':['']})
```

I, agafant els detalls de cada producte al lloc web, anem emplenant les diferents files del dataframe.

```
for product in product_card:
    try:
        link = product.find('a', class_ = 'product-card__link-overlay').get('href')
        nom_producte = product.find('div', class_ = 'product-card__title').text
        quantitat_colors = product.find('div', class_ = 'product-card__count-item').text
        detalls = product.find('div', class_ = 'product-card__subtitle').text
        descompte = product.find('div', class_ = 'product-price__perc css-1qwsg2u').text
        df = df.append({'Link':link, 'Nom':nom_producte, 'Detalls':detalls, 'Colors':quantitat_colors,'Descompte':descomp
                    ignore_index = True)
    except:
        pass
```

Finalment, guardarem aquest dataframe com a CSV:

```
df.to_csv('C:/Users/marcc/Desktop/UOC/TIPOLOGIA I CICLE DE VIDA/file_name.csv')
```

10. Dataset

Hem publicat el dataset obtingut en el en format CSV a Zenodo i es pot trobar en el següent enllaç de DOI: *https://doi.org/10.5281/zenodo.7340782*

11. Vídeo

Hem publicat el vídeo de la pràctica a: *https://drive.google.com/file/d/1ijKNhlSN8IfIgrNWOb8ZCg8LVj6lfLGm/view?usp=share_link*

12. Signatures

Investigació Prèvia	Paula Miralles Simó, Marc Clupés Peró
Redacció de les respostes	Paula Miralles Simó, Marc Clupés Peró
Desenvolupament del codi	Paula Miralles Simó, Marc Clupés Peró
Participació al vídeo	Paula Miralles Simó, Marc Clupés Peró