2022

PRÀCTICA 1

WEB SCRAPING

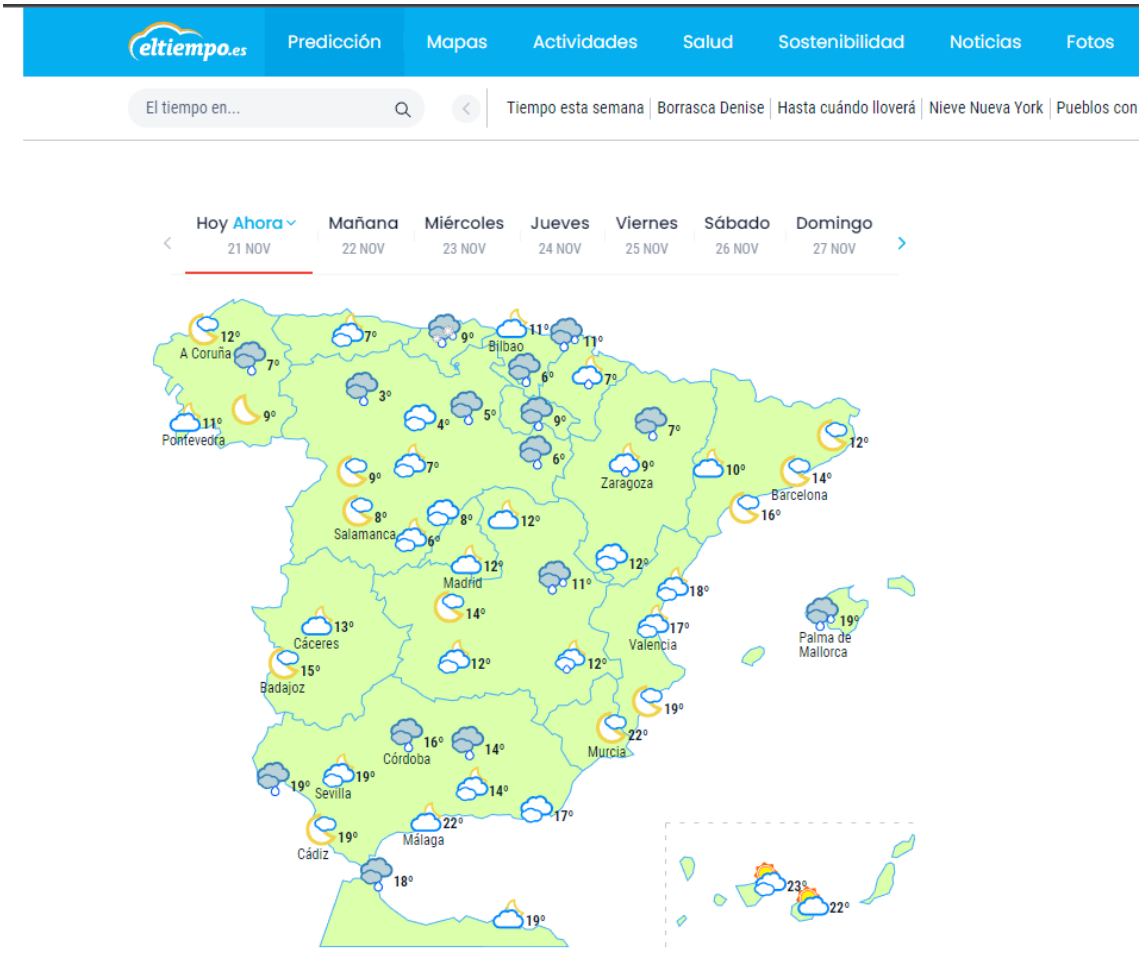
ÀLEX LÓPEZ DIAZ

1. Context

La meva primera idea per aquest projecte era diferent al que s'ha acabat realitzant. La meva primera idea era recopilar informació referent al temps.

Volia crear una eina que utilitzant web-scraping pogués extreure les temperatures i prediccions de temps.

Diverses pàgines semblaven potencialment utilitzables, i em vaig decantar per https://www.eltiempo.es/.



A priori, les dades semblen incrustades a l'html:

```
<text x="476.2905728426534" y="114.75910635764153" class="temp-tooltip temp-tooltip-9" orig-temp="7">
  <a href="/aragon">
    <tspan x="476.2905728426534" y="114.75910635764153" class="et-tooltip" data-temp="7">7º</tspan> == $0
  </a>
</text>
```

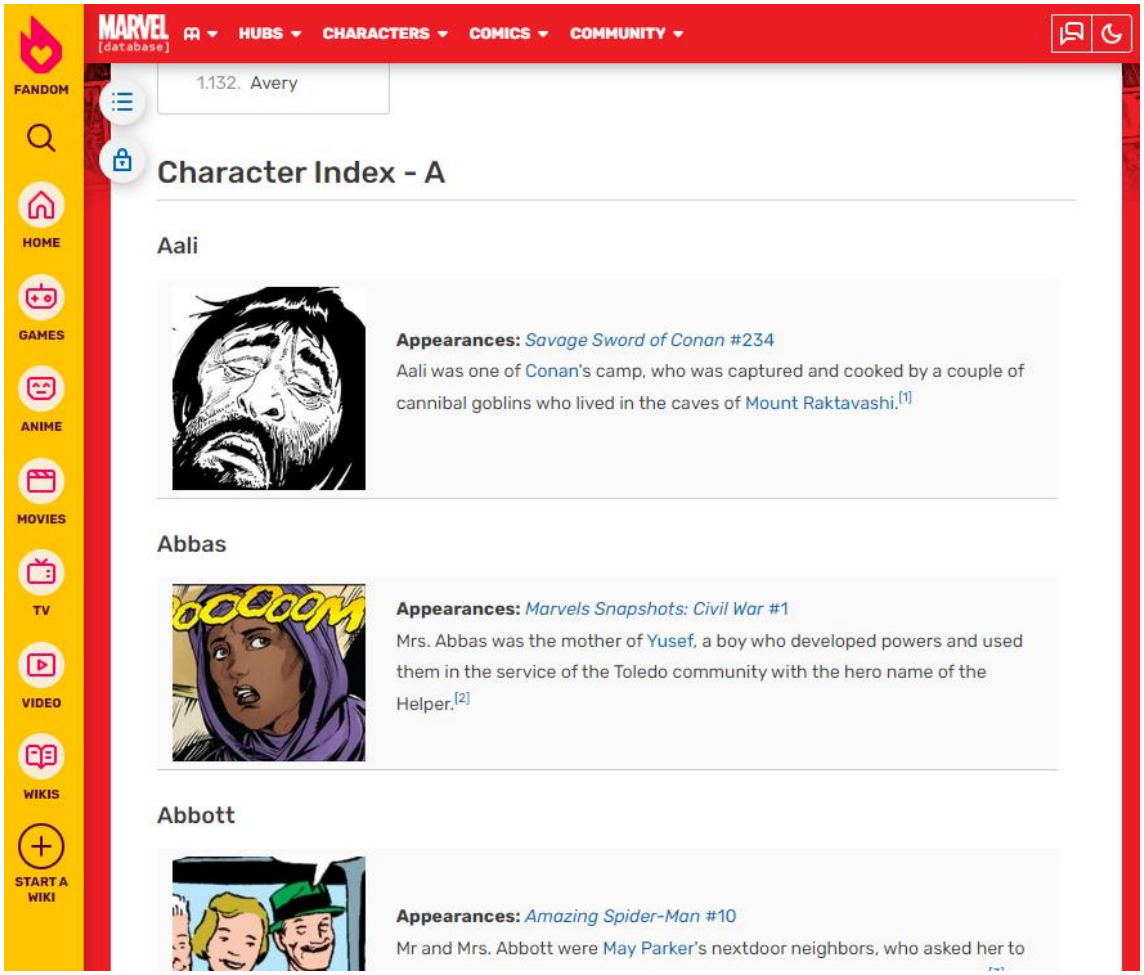
Però un cop l'utilitzem per buscar-ne les dades, ens trobem que no hi ha cap registre amb el nom "temp-tooltip temp-tooltip-9". La pàgina utilitza fitxers JavaScript que realitzen consultes a una API que retorna la informació. Sembla una pràctica estesa a les pàgines que retornen informació sobre el temps.

Al trobar-me aquests impediments, ja que no està permès la utilització de consultes a APIs o webs que tinguin la informació en taules, vaig decidir buscar pàgines que servissin com a

"repositoris" d'informació que la gent pogués consultar.

Traient una mica la meva vena "friki" he topat amb una pàgina web que conté informació sobre personatges de còmics de MARVEL: https://marvel.fandom.com/wiki/Character_Index. Conté les dades de diversos personatges de la franquícia.

Si entrem a qualsevol dels punts de l'índex podem veure la pàgina següent:



Amb aquesta informació podem fer un recull de dades dels personatges de Marvel, extraient-ne informació rellevant com podria ser quantes vegades ha aparegut aquest personatge, quin aspecte té o amb quins altres personatges pot estar relacionat.

Si ens posem en la pell d'una empresa creadora i divulgadora de còmics, podria ser interessant veure quins personatges tenen més rellevància dins el món de Marvel i, amb la seva descripció, intentar replicar-los a la nostra empresa.

Comentar també que dins aquesta pàgina hi ha personatges que conten amb pàgines per ells mateixos, però el nombre de peticions a la pàgina m'ha comportat a bloquejos per part de la pàgina web, per tant, m'he decantat només per utilitzar aquesta pàgina només. És suficient per muntar un sistema de web-scrapping i per familiaritzar-nos amb l'entorn i les eines.

2. Títol

El títol del dataset podria ser “Marvel’s Characters Information”.

3. Descripció del dataset

La pàgina tampoc ens brinda molta informació, però és suficient per treure’n informació rellevant. També trobo interessant extreure’n informació en format d’imatge.

Exportarem informació refent a: nom del personatge, imatges, referencies a altres còmics i una descripció dels personatges. Hem de ser curosos perquè depenent de l’entrega i/o aparicio, un mateix personatge pot haver canviat el seu context, per tant, per cada aparició crearem un registre.

4. Representació gràfica

La millor representació gràfica és un exemple de cada article:

Sergei

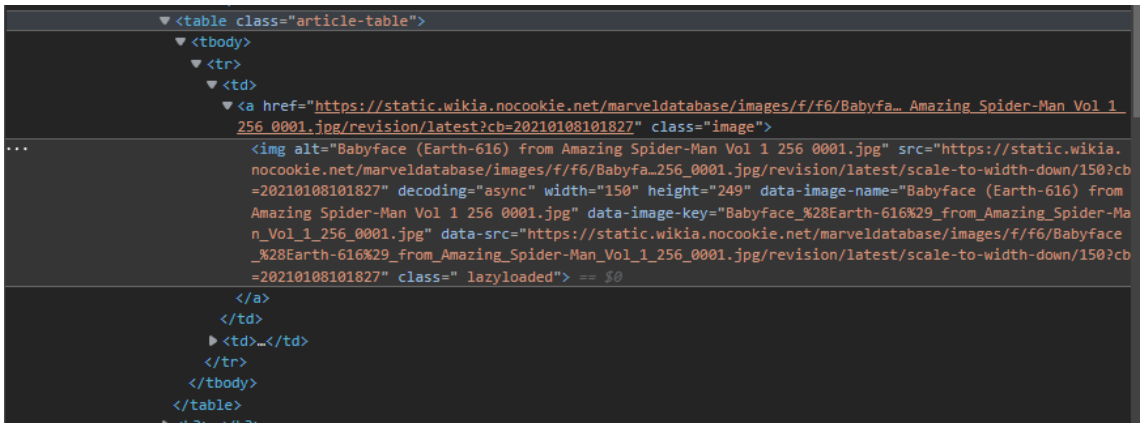
Appearances: *Captain America* #618

“A Russian diplomat who showed Gyrich confidential documents convincing him that Barnes should be extradited to Russia. After making Gyrich confess about Barnes's extradition, Rogers went to see Sergei but by the time he arrived at his home it was too late as he had been murdered.”

5. Contingut

Primerament, anem a analitzar l'estructura de la web a treure la informació.

Cada element que nosaltres extraurem com a registre està dividit dins la web per "articles".



En la figura anterior podem veure un exemple que es correspon a aquest article:

El títol es troba en un grup separat. Ho comento perquè quan s'hagi de fer l'extracció s'hauran de tractar per separat.

Aquestes webs que contenen la informació es troben indexades en una pàgina que serveix com a índex. La web https://marvel.fandom.com/wiki/Character_Index . Veiem una foto:

L'objectiu és muntar una eina que recorri els elements del índex i reculli la informació de cada pàgina automàticament.

Estructura

Veiem com quedaria l'estructura del dataset amb les dades de cada article.

Podríem dir que aquest dataset contindrà informació com si fos un arxiu, no un dataset operacional que pot contenir informació que canvia cada poc temps.

Les columnes del dataset son:

Nom de la columna	**Descripció**	**Tipus**	**Exemple**
id	identificador	int	12
name	Nom del personatge	string	Abrams
image	Url de l'imatge del personatge	string	https://static.wikia.nocookie.net/marveldatabase/images/4/46/...
appearences	Referencies al còmic on ha aparegut	string	Iron Man Vol 1 230
description	Descripció del personatge	string	Dr. Abrams coached Firepower until Edwin Cord convinced Senator Boynton Firepower was capable of concluding his testing without external interference.

El període de temps de les dades és molt llarg en aquest cas. Tots aquests personatges ja han aparegut en algun còmic i aquests no canviaran amb el temps.

6. Propietari

El propietari de les dades és el propietari de la propietat intel·lectual d'aquests personatges. Atesos a que la marca MARVEL és una marca registrada no tenim la potestat d'apropiar-nos d'aquestes dades.

A part, si investiguem a la pàgina d'on anem a extreure les dades ens trobem que indica explícitament el següent:

You agree not to:

- *Except as expressly permitted by the Company (for example with respect to the use of text content that is submitted to particular Fandom communities as permitted as set forth at our licensing page), you may not modify, publish, transmit, reproduce, scrape, create derivative works from, distribute, perform, adapt, aggregate, sell, transfer or in any way exploit any of the content, in whole or in part,*

- *Use any robot, spider, site search and/or retrieval application, or other device to scrape, extract, retrieve or index any portion of the content;*

Extret de https://www.fandom.com/terms-of-use

Exposa que es requereix permís explícit i firmat per la web per extreure'n la informació a través de qualsevol eina d'scrap, del qual no disposem.

En conclusió, deduïm que el propietari del conjunt de dades és la pàgina d'on hem extret les dades.

Anem a extreure el propietari de la web. Python conté una llibreria que ens brida aquesta opció. La llibreria es "whois". Ens retorna la següent informació:


```
{
  "domain_name": "FANDOM.COM",
  "registrar": "TUCOWS, INC.",
  "whois_server": "whois.tucows.com",
  "referral_url": null,
  "updated_date": "2020-10-20 17:10:37",
  "creation_date": "1996-10-11 04:00:00",
  "expiration_date": "2023-10-10 04:00:00",
  "name_servers": [
    "NS1.P11.DYNECT.NET",
    "NS2.P11.DYNECT.NET",
    "NS3.P11.DYNECT.NET",
    "NS4.P11.DYNECT.NET",
    "ns1.p11.dynect.net",
    "ns3.p11.dynect.net",
    "ns2.p11.dynect.net",
    "ns4.p11.dynect.net"
  ],
  "status": [
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited"
  ],
  "emails": [
    "domainabuse@tucows.com",
    "fandom.com@contactprivacy.com",
    "help@hover.com"
  ],
  "dnssec": "unsigned",
  "name": "Contact Privacy Inc. Customer 0141730466",
  "org": "Contact Privacy Inc. Customer 0141730466",
  "address": "96 Mowat Ave",
  "city": "Toronto",
  "state": "ON",
  "registrant_postal_code": "M6K 3M1",
  "country": "CA"
}
```


El propietari és TUCOWS, INC.

Anàlisis similars

No he trobat anàlisis que utilitzessin datasets similars. Al final he escollit un tema molt específic i més com una "excusa" per familiaritzar-me amb les eines i procediments del web-scrapping.

El que si he trobat han sigut datasets similars, amb dades i columnes similars a les meves. Veure l'apartat 7.

User agent

El user agent és una part de la capçalera de les consultes HTTP que conté informació sobre la petició: quin navegador s'està utlitzant, el sistema operatiu...

En aquesta pràctica no hem necessitat modificar el user agent que incorpora la llibreria requests de Python, però per curiositat aquí adjunto el meu.

El meu user agent:

Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36 Edg/107.0.1418.52

I el que hem utilitzat per la pràctica, que extraiem amb la consulta següent:

requests.utils.default_headers()

```
{'User-Agent': 'python-requests/2.28.1', 'Accept-Encoding': 'gzip, deflate, b
r', 'Accept': '*/*', 'Connection': 'keep-alive'}
```

7. Inspiració

La inspiració va venir donada per, principalment, gustos personals. És un món que sempre m'ha agradat. A part, vaig trobar un parell de datasets que contenien informació similar tot i que amb informació més completa.

Els datasets en qüestió son:

- Dataset de cómics

  https://www.kaggle.com/datasets/leonardopena/marvel-vs-dc

- Dataset characters

  data/avengers at 9e9cee37d0695ccc6866c67f38373675231758ab · fivethirtyeight/data · GitHub

Exemples de preguntes que pretén respondre son:

- Qui era x personatge?
- En quants títols va aparèixer?
- Quina relació pot tenir entre personatges?
- Més aparicions equival a més rellevant?
- Quin aspecte tenia x personatge?
- Algun personatge ha canviat de bàndol o ha pogut tenir un paper rellevant?

Son preguntes força específiques.

8. Llicència

Utilitzaria la llicència CC BY-NC-SA 4.0.
Tota la informació recollida no em pertany i tots aquests noms i personatges són propietat de Marvel, per tant, citaria de quina pàgina s'han extret les dades.

9. Codi

El codi que he realitzat està pensat per ser executat seqüencialment. És indispensable executar tots els blocs de codi per ordre.

Faig cinc cèntims de com funciona el codi per poder-lo entendre. La majoria de línies dins el codi estan comentades per entendre'l millor.

El codi consta de dues parts, una que he anomenat "cas base" i l'altra "procediment principal".

Cas base

En aquesta execució mostro el procés que he seguit per extreure la informació d'una pàgina web de l'índex. En aquest cas, la primera pàgina de totes, la dels personatges que comencen amb la lletra A.

Aprofitant que estava investigant la pàgina, he anat creant diverses funcions per cada funcionalitat que ha de presentar aquesta extracció de dades en concret. Les he creat pensant en generalitzar-les per poder-les utilitzar en el procediment principal.

Les funcions son:

a. **extract_characters_names (soup):** serveix per extreure exclusivament els noms de totes els personatges. Retorna una llista amb els noms. Per paràmetre s'ha de passar la pàgina sencera.
b. **extract_character_info (name,tr):** serveix per extreure les dades dels blocs amb el tag "tr". Retorna un registre a afegir al dataset final amb les columnes name, imatge, appearences i description. Per paràmetre s'ha d'indicar el nom del personatge que estem registrant.
   He hagut d'afegir comprovacions perquè alguns registres no tenien foto, appearences o descripció, per tant, quan trobava un cas on no n'hi ha simplement guardo un string buit.
c. **exctract_all_character_info (name, soup):** aquesta funció s'encarrega d'iterar tots els blocs de dades (tr) que pot presentar el personatge i cridem a la funció anterior. Per paràmetre s'ha d'indicar el nom del personatge i el contingut d'una taula que dins del contigut de la pàgina web té el tag de "table".
d. **scrap_web_with_info (soup):** aquesta funció s'encarrega d'extreure tota la informació de la pàgina, cridant al mètode anterior. Per paràmetre es passa el contingut de la pàgina sencera.

e. **export_data (result):** Funció que exporta una llista de dades en un fitxer csv. Per paràmentre indiquem la llista a guardar.

Procediment principal

En aquest procediment ens centrarem en iterar totes les pàgines de l'ìndex per extreure la informació de cada una.

El primer pas és extreure els sufixes que hem d'afegir a la URL de la pàgina. L'url de cada pàgina "cas base" està formada per

https://marvel.fandom.com/wiki/Character_Index + /A

Un cop extrets els sufixes, els iterem i cridem a la funció scrap_web_with_info() que hem creat al cas base.

Comentar que la pàgina web bloqueja la nostra sessió quan hem fet un nombre x de consultes. Per evitar això, hem aplicat un sistema d'espaiat de peticions creant una pausa random entre 1 i 5 segons per evitar ser bloquejats.

Per comprovar que tot funciona correctament he afegit missatges de consola per comprovar el procediment.

Exemple d'execució:

```
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/A
Sleep for: 4 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/B
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/C
Sleep for: 4 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/D
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/E
Sleep for: 2 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/F
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/G
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/H
Sleep for: 2 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/I
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/J
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/K
Sleep for: 4 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/L
Sleep for: 5 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/M
Sleep for: 1 seconds.
```

Continua a la següent pàgina

```
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/Y
Sleep for: 1 seconds.
Scrapping page: https://marvel.fandom.com/wiki/Character_Index/Z
Sleep for: 1 seconds.
TOT OK
```

El codi està penjat al repositori:

https://github.com/Alopezd24/PRAC1_Web_Scrapping

10. Dataset

Com s'ha explicat a l'apartat 6, la pàgina web especifica que no es poden realitzar extraccions de dades utilitzant eines d'scrap explícitament, per tant, per evitar problemes legals, he optat per considerar el dataset de "risc" i seguiré el procediment indicat per l'enunciat de la pràctica.

A la plataforma Zenodo i al repositori Git hi haurà un dataset fictici amb les mateixes columnes però no amb les mateixes dades.

El dataset estarà penjat a la mateixa carpeta del Drive que es penjarà el vídeo.

Link: https://zenodo.org/record/7349044#.Y31HZUmZOUk

11. Video

Link del vídeo:

https://drive.google.com/file/d/1i1V86rqguPLeQspM3Wow31N2-pch6dzH/view?usp=share_link

12. Contribucions

La pràctica ha sigut realitzada per una sola persona.