PRA 1 - Web Scraping

Jordi Puig Ovejero - Master de Ciència de Dades

01-11-2020

1 Context

La web ClinicalTrials.gov és una base de dades d'estudis clínics finançats amb fons privats i públics realitzats a tot el món.

L'objectiu del projecte és extreure un conjunt de proves clíniques realitzades per la COVID-19 arreu del món amb la finalitat de realitzar estudis estadístics i de Data Mining.

A l'hora d'obtenir les dades realitzem un rastreig d'aquesta web. Anem recorrent els diferents elements del resultat de cerca i es descarrega aquesta informació en un document CSV.

2 Títol del dataset

El títol escollit per el dataset és **Covid-19 - Clinical Trials.**

3 Descripció del dataset

El Dataset descarregat conté informació de cada una de les proves clíniques realitzades per la COVID-19 arreu del món, tant d'entitats públiques com privades. Aquestes es troben emmagatzemades a la web ClinicalTrials.gov.

El contingut inclou *dades de l'estudi* (nom, fase, malaltia, tractament), *dades dels participants en l'estudi* (edats, sexe, nombre...) i responsables / patrocinadors de l'estudi.

4 Representació gràfica

5 Contingut i procés

Les dades s'han recollit mitjançant un programa realitzat en Python on es va recorrent les diferents pàgines de la web que conté les proves clíniques i les va descarregant.

Implementació

Selenium: per a realitzar l'scraping s'ha fet servir la llibreria Selenium. Aquesta llibreria l'havia fet servir amb anterioritat per a realitzar Testing en altres plataformes.

Passes a realitzar: les passes realitzades són les següents.

Accedim a la pàgina inicial amb la URL. Aquesta URL filtra les cerques de proves clíniques amb la condició = COVID

Els valors que volem descarregar es troben en una taula, i cada un d'ells, es guarden en la forma:

```
<tr><th>title</th><td>content</td></tr>
```

Per tant, per a cada camp de la pàgina que volem descarregar fem el següent:

i. Primer fem una cerca per th + contains:'title'. És a dir, busquem un th que a més tingui una part del text de 'title'.

```
th = self.browser.find_element_by_xpath("//th[contains(text(), '" +
title + "')]")
```

ii. Després pugem un nivell per anar al tr que conté tant el title com el contingut.

```
tr = th.find_element_by_xpath("./..")
```

iii. Finalment, agafem el contingut del td.

```
td = tr.find_element_by_xpath(".//td[1]")
```

Un cop hem guardat els valors de la pàgina en un objecte de classe Study i l'hem emmagatzemat a una llista anem a la següent página. Les pàgines tenen una paginació amb enllaços de la forma típica (Anterior - Següent). Cerquem la pàgina següent i naveguem a la pàgina:

```
class_name = 'tr-next-link'
try:
  next_link = self.browser.find_element_by_class_name(class_name)
  next_link.click()
except:
  raise Exception('not exists the next link')
```

Anem recorrent totes les pàgines fins que el valor del next_link no existeix. En aquest cas capturarem una excepció ja que es produirà un error i la llençarem per a un tractament en una classe superior. És una excepció controlada.

Logs

Hem posat alguns logs per veure que és el que s'està executant. Per exemple el número de la pàgina.

Altre cosa que m'ha semblat interessant loggar son els camps que no existeixen per a algunes pàgines. No totes les pàgines tenen tots els camps.

```
processing page: 205  ...
not exists element:  Study Population
not exists element:  Study Groups/Cohorts
page: NCT04386668 processed
processing page: 206  ...
not exists element:  Study Population
not exists element:  Study Groups/Cohorts
page: NCT04386668 processed
processing page: 207  ...
```

User-Agent

Per a simular la navegació d'un web browser he fet servir la llibreria fake-useragent. Aquesta llibreria simula un User Agent i ho fa de forma aleatòria per a cada una de les connexions que es realitzen.

Extensió de la pràctica

La práctica está orientada a obtenir les dades dels estudis científics per COVID però he fet una ampliació per a recuperar qualsevol estudi científic amb un criteri de cerca.

Per a obtenir aquests registres realitzem l'execució que hem comentat en el punt previ però de la forma següent:

Per executar fem el següent:

```
python main.py cancer
```

I ens descarregarà els estudis que tenen com a keyword la paraula cancer.

Per defecte la keyword és COVID.

Dades del dataset

Cada un dels registres del dataset està compost per els següents atributs:

1. Dades de l'estudi
    - Id: Identificador del registre (string)
    - Brief Title: Títol (string)
    - Official Title: Títol oficial (string)
    - Brief Summary: Descripció curta (string)
    - Detailed Description: Descripció detallada (string)
    - Study Type: Tipus d'estudi a realitzar (p.ex. COVID) (string)
    - Study Phase: Fase en la qual es troba l'estudi (string)
    - Study Design: Disseny de l'estudi (string)
    - Condition/Disease: Malaltia o condició a tractar (string)
    - Intervention/Treatment: Intervencions realitzades o tractaments (string)
    - Study Arms: Grups d'estudi (string)
    - Start Date: Data d'inici (date)
    - Completion Date: Data de finalització (date)

2. Dades dels participants en l'estudi
    - Estimated Enrollment: Nombre de participants (enter)
    - Eligibility Criteria: Criteri per incloure els participants a l'estudi (string)
    - Sex/Gender: Sexe dels participants (string)
    - Ages: Edats del participants (string)
    - Study Population: Població a estudiar (string)
    - Study Groups/Cohorts: Grups d'estudi (string)
    - Listed Location Countries: Països dels participants en l'estudi (string)

3. Responsables / Sponsors
    - Responsible Party: Responsables del projecte (string)
    - Study Sponsor: Sponsors que participen (string)
    - Collaborators: Col·laboradors de l'estudi (string)
    - Investigators: Grup d'investigadors (string)

6 Agraïments

La pràctica per als estudis universitaris de la UOC la he pogut realitzar gràcies a la base de dades Clinicaltrials.gov, proporcionada per la Biblioteca Nacional de Medicina dels EE. UU.

7 Inspiració

El projecte surt de la necessitat de tenir un dataset dels estudis clínics que s'han realitzat fins ara de la COVID-19.

Amb aquest dataset podem fer estudis estadístics o de data mining amb:

- Quins tipus de intervencions o procediments s'han realitzat o quins medicament s'han emprat.
- Edats i sexe de les persones testades.
- Volum dels individus testats.
- Països participants.
- Fases en les quals es troben els estudis.
- Dates dels estudis.
- …

Aquestes dades són una recopilació sense cap tipus de processament, per tant serà necessari, de cara a realitzar estudis posteriors, fer tractaments i neteja de les dades.

8 Llicència

- Compartir (Share): llibertat per a copiar i redistribuïr el material a qualsevol mitjà o format.
- Atribución (Attribution): s'ha de fer referència a la llicència, donar crèdit i indicar canvis.
- Adaptar (Adapt): barrejar, transformar i construir sobre el material.
- NoComercial (NonCommercial): no es pot fer servir al material amb fins comercials.
- CompartirIgual (ShareAlike): si es barreja, transforma o construeix sobre el material, s'ha de fer servir sota la mateixa llicència.

9 Codi

El codi font està format pels següents fitxers cada un amb una responsabilitat específica, SRP [1] i amb programació orientació a objectes, OOP[2]:

- main.py: és la classe principal que rep la petició i arrenca tot el procés
- scraper.py: és la classe que s'encarrega de realitzar el procés d'scraping, guardar les pàgines en una llista i posteriorment emmagatzemar el fitxer csv.
- browser.py: fa la navegació. Embolcalla un objecte de Selenium i realitza la navegació per les pàgines.
- navigator.py: a partir de la pàgina del navegador permet anar a la següent pàgina o l'anterior.
- study_scraper.py: fa scraping d'una pàgina d'estudi.
- study.py: és l'entitat study.
- data2csv.py: guarda els registes generats de l'scraping realitzat a un fitxer csv.

10 Dataset a Zenodo

El dataset ha estat publicat a Zenodo i el podem trobar en aquesta URL.

[1] "Single-responsibility principle - Wikipedia." https://en.wikipedia.org/wiki/Single-responsibility_principle. Se consultó el 5 nov.. 2020.
[2] "Object-oriented programming - Wikipedia." https://en.wikipedia.org/wiki/Object-oriented_programming. Se consultó el 5 nov.. 2020.

11 Contribucions

Contribucions	Signa
Recerca prèvia	JPO
Redacció de les respostes	JPO
Desenvolupament codi	JPO