Pràctica 1: Web scraping UOC M2.951 Tipologia i cicle de vida de les dades

Adreça del repositori de Github:
https://github.com/jcarles/webcrawler

Descripció
Aquesta pràctica s'ha realitzat sota el context de l'assignatura Tipologia i cicle de vida de les dades, del Màster en Ciència de Dades de la UOC. S'apliquen tècniques de web scraping per extreure dades mitjançant Python de la web nuforc.org i generar un dataset.

Membres de l'equip
Aquesta activitat ha estat realitzada de manera individual per **Joan Carles Badia Purroy**.

Fitxers del codi font
- download.py : Funció per a descarregar el codi html d'una url donada
- crawl_url.py : funció per a extreure de la pàgina principal, els links a les url on hi ha la informació i baixar cadascuna de les pàgines web i cridar al callback que les interpreti .
- ScrapeCallback.py : Conté la classe que es crida quan s'ha extret els anteriors links i que, mitjançant lxml cerca dins de la taula html, les dades. Tot seguit les guarda en arrays i les va volcant en un fitxer .csv . El fitxer també conté el punt d'entrada principal main (per executar tot el procés)

Fitxer csv:

https://github.com/jcarles/webcrawler/blob/master/ufo.csv

Estructura del dataset

Mostra del dataset

El dataset està format per un recull històric d'observacions ufològiques dels Estats Units . Consta dels següents camps:

- Date/ Time: Data i hora en que es va produïr el fenòmen
- City: Ciutat dels EEUU en que es va produïr
- State: Estat dels EEUU en que es circumscriu el fenòmen
- Shape: Categoria de forma que tenia el fenòmen, Pot tenir els següents valors: (Flash/Light/Triangle/Rectangle/Unknown/ Other)
- Duration: Temps que va durar el fenòmen. (en minuts/segons, hores...)(caldria uniformitzar aquesta dada)
- Summary: Text descriptiu amb el testimoni del fenòmen.
- Posted: Data en que es va registrar a la base de dades.
- L'arxiu consta de 122.570 registres. Les dades mes antigues son de l'any 1948, tot i que s'hauria de normalitzar la columna Date / Time per fer un estudi, ja que s'hi barrejen diferents formats de data. Les dades s'han recollit mitjançant alertes de testimonis de fenòmen ufològic.
- El propietari de les dades és NUFORC.ORG. Son les sigles de NATIONAL UFO REPORTING CENTER. És un centre dels EEUU dedicat a recol·lecció i exposició de Dades UFO objectives. És una corporació sense ànim de lucre. Està suportada mitjançant mitjançant subscripcions i venda de vídeos i merxandatge i altres aportacions privades. Té la seva seu social a Seattle, WA i va ser fundat l'any 1974 per l'investigador ufològic Robert J. Gribble. La principal funció al llarg de les passades dues dècades ha estat rebre, registrar, fins al més alt grau de detall possible, corroborar i documentar informes d'individus que han estat testimonis de successos possiblement relacionats amb OVNIS.
- El principal medi utilitzat pel centre per a rebre informes d'avistaments és la línea telefònica, que ha estat operant de forma gairebé contínua des de 1974. Durant aquest període,

la línia ha processat milers de trucades, i el Centre ha distribuït la informació a milers d'individus.
- La línia ha estat disponible les 24 hores del dia . La independència del Centre de qualsevol altra organització ufològica, combinada amb la seva política d'anonimat garantit als informadors, l'ha convertit en la organització més popular i amplament acceptada arreu.
- la línia és ben coneguda per agencies de la llei, aeroports, Serveis meteorològics, instal·lacions militars, NASA, i centres d'emergències 911 arreu dels Estats Units i part del Canada.
- El que distingeix les seves operacions de moltes altres organitzacions ufològiques, és que posa a disposició del públic totes les seves dades, oferint informació detallada als investigadors Ufològics.

Estudis previs

De les dades que ofereix públicament nuforc.org se n'ha fet estudis previs, com per exemple :

- https://greenet09.github.io/datasophy/2018/07/03/UFO.html
- http://metrocosm.com/map-of-ufo-sightings/
- La idea d'un estudi seria intentar trobar una pauta, un patró, que permeti donar una explicació o almenys, establir un patró del fenòmen ufològic, en base al tipus d'avistament, la localització, la data, i, d'alguna manera, establir una correlació entre aquests paràmetres i situacions cojunturals que es puguin desprendre.

Llicència

Les dades es troben sota la llicència **Open Data Commons Open Database License (ODbL)** que permet la reutilització de les dades sempre que es reconegui l'autoria de la informació original; es mantindrà la mateixa llicència en les obres derivades les quals puguin restringir el seu ús si, a més, es distribueix una versió sense aquestes restricció d'ús.

- M'ha semblat la més adequada ja que les dades no les he generat jo si no que provenen d'una altra font a la qual com a mínim crec que s'ha de fer referència.

Contingut audiovisual

- En cas de que hagués hagut de tractar contingut audiovisual a l'hora de tractar una foto es faria de la següent manera:

```
import urllib resource =
urllib.urlopen("nomdelfitxer.jpg")
output = open("file01.jpg","wb")
output.write(resource.read()) output.close()
```

- Es a dir, recollint del html la url de la foto i obrint un fitxer d'escriptura on escriure el contingut de la url.

Selenium

Selenium és una eina que automatitza la interacció amb una web. Combinada amb Python permet la programació de web scrapping de pàgines més complicades, que inclouen links Javascript per exemple, ja que es pot interactuar amb els links (com ara prement un botó emulant la interacció de l'usuari, per accedir a una part diferent de la web)

- Per exemple:

```
driver = webdriver.Firefox() driver.implicitly_wait(30)
driver.get(url)
python_button =
driver.find_element_by_id('MainContent_uxLevel1_Agencies_u
xAgencyBtn_33') #FHSU
python_button.click() #click fhsu link
```

Contribucions	Signa
Recerca prèvia	Joan Carles Badia
Redacció de les respostes	Joan Carles Badia
Desenvolupament codi	Joan Carles Badia