Pràctica 1 – Web scraping

1. Títol del dataset. Cal que poseu un títol que sigui descriptiu.

Compra per al·lèrgics.

2. Subtítol del dataset. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.

Llista de productes indicant els al·lergènics o traces que conté en els seus ingredients, per tal de facilitar la compra a al·lèrgics.

3. Imatge. Agregueu una imatge que identifiqui el vostre dataset visualment.

http://www.ticbeat.com/salud/alergicos-tardan-87-minutos-en-hacer-la-compra/

4. Context. Quina és la matèria del conjunt de dades?

Per una persona al·lèrgica, anar a comprar no és senzill. D'al·lèrgies n'hi ha a diversitat de productes, així que la única opció és comprovar producte a producte la llista d'ingredients. Això fa que sigui molt farregós i un acaba comprant sempre els mateixos productes, per no haver d'anar llegint tots els ingredients de tot el supermercat. El que es pretén amb aquest dataset és tenir els al·lèrgens de cada producte per facilitar el filtratge d'aliments apte per a cada consumidor.

5. Contingut. Quins camps inclou? Quin és el període de temps de les dades i com s'ha recollit?

Obtenim el següent fitxer:

alergenos.csv: llistat de productes de la secció de congelats

- Categoria: classificació dels productes
- Tipu: classificació dels productes
- Producte: nom descriptiu del producte, inclou marca i pes.
- Imatge: nom del fitxer de la imatge del producte.
- Alergens: llista d'al·lergènics que conté.

NOTA: he restringit el llistat de productes només a la secció de congelats com a mostra del procés per restringir el volum de productes exportats. Només he exportat aquells productes que en el web tenen informat el llistat d'ingredients.

Imatges:

- Una carpeta amb totes les imatges dels productes del fitxer anterior. El nom de cada fitxer correspon en l'indicat en el CSV per a cada producte.

NOTA: el nombre d'imatges exportades no coincideix exactament amb el nombre de registres del CSV (hi ha 2 imatges menys). Això és perquè un mateix producte es troba al web classificat en Tipus diferents. Són:

"*pizza de chocolate con topping de chocolates variados*" es troba dins el tipus "*Helados, postres y nata"* i també dins del tipus "*Pizza, Lasaña y Canelones".*

"Guisantes finos al vapor listos en 3 minutos" es troba dins el tipus *"Platos preparados congelados*" i també dins del tipus *"Verduras, hortalizas y frutas".*

Les dades s'han obtingut el dia 03/11/2018 mitjançant webscrapping de la pàgina de venda online: https://www.elcorteingles.es/supermercado. Són els productes disponibles aquest dia en la compra online. El procés simula la navegació manual de

l'usuari per arribar a obtenir el detall dels ingredients de cada producte i la seva imatge.

6. Agraïments. Qui és propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.

Les dades pertanyen al supermercat online de El Corte Inglés. Vull agrair el detall dels ingredients publicats en la web.

És complicat trobar supermercats amb els ingredients publicats dels productes. Hi ha pàgines que publiquen fitxes tècniques on apareix el llistat de productes indicant els al·lergènics que conté. Però el format amb el que estan publicats (llistat en pdf en forma de taula amb creus per cada al·lergen que conté), dificulta la tasca de recerca, havent de consultar marca per marca, producte per producte, el seu contingut. Per exemple, la marca Nestlé utilitza aquest format comentat:

https://www.helados.nestle.es/cliente/alergenos.aspx
https://www.helados.nestle.es/cliente/Content/pdfs/alergenos/listado-alergenos-2018-la-lechera.pdf

O altres com la cadena BonPreu que ofereix un llistat de productes que no contenen gluten, o no contenen llet o no contenen ou. Però no apareix cap més guia pels mutlial·lèrgics:

https://www.bonpreuesclat.cat/es/inici?p_p_id=101&p_p_lifecycle=0&p_p_state=maximized&p_p_mode=view&_101_struts_action=%2Fasset_publisher%2Fview_content&_101_returnToFullPageURL=%2F%3Fp_p_id%3D3&_101_assetEntryId=36529&_101_type=content&_101_urlTitle=fleca_qualitat_que_oferim&inheritRedirect=true

Hi ha botigues especialitzades per al·lèrgics, però només es troba una petita varietat de productes, no és per fer compres setmanals. Tot i així la pàgina no facilita la recerca de productes filtrats per al·lèrgens segons les necessitats de l'usuari:

https://jotambesocallergic.cat/wordpress_4/web-ca/botiga/xocolata/piruleta-xocolata-cor/

També hi ha APPs on es pot configurar les al·lèrgies i al escanejar el codi de barres t'indica si és apte o no pel consum configurat. No hi són tots els productes, però és una bona iniciativa. Però fa que s'hagi d'anar escanejant producte a producte.

https://www.allergeneat.com/

És per això que m'he decantat per la pàgina de El Corte Inglés pel detall de la informació dels ingredients de cada producte que ofereix. Tot i que en aquest web hi ha una opció de filtrat, només permet filtrar o per sense gluten o sense ou o sense llet. No hi ha opció de multial·lèrgies ni cap altre ingredient més a filtrar.

https://www.elcorteingles.es/supermercado

7. Inspiració. Per què és interessant aquest conjunt de dades? Quines preguntes li agradaria respondre la comunitat?

La meva inspiració és per la meva experiència personal en aquest tema. La meva filla, ja de ben petita se li ha detectat al·lèrgia a l'ou, llegums, cefal·lòpodes i marisc. Durant tots aquests anys he après molt de les composicions dels productes, com tenen ingredients o traces inimaginables a priori. Molts anys he hagut primer de trucar directament a les empreses per saber exactament els continguts de cada producte consumit. A partir del 2014 la tasca ha estat més senzilla, al aplicar-se una norma europea d'obligar a les empreses informar de tots els ingredients incloses les traces i a més remarcar una llista d'al·lèrgens de forma destacada (negreta o majúscules per exemple). Però tot i així no és fàcil trobar productes nous. I a més hi ha ingredients que trobo que encara no estan ben indicats, que no estan a la llista de mínims marcats per la Comunitat Europea (i que la meva filla no pot consumir).

La pregunta que voldria respondre és la que es fa qualsevol al·lèrgic: quins productes del supermercat puc optar a comprar? La idea és poder donar una resposta ràpida sense haver de llegir ingredients producte a producte, o haver de pensar primer el producte i després comprovar si és apte o no és apte. Fustra molt veure una llista llarga de gelats, llegir ingredients un a un fins a descobrir que no hi ha cap que pugui consumir. És per això que una aplicació on s'indiquin els ingredients a filtrar, facilitaria enormement la tasca a aquest col·lectiu.

8. Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir perquè l'heu seleccionada:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**

○ **Unknown License**

Escullo la llicència **CC BY-SA 4.0 License**, ja que crec que és la que més s'ajusta a la pràctica. Compleix els termes que es demana. Resum de la llicència, com apunta creativecommons.org:

Sou lliure de:

- **Compartir**: copiar i redistribuir el material en qualsevol mitjà i format
- **Adaptar:** remesclar, transformar i crear a partir del material per a qualsevol finalitat, fins i tot comercial.

Aquesta llicència és acceptable per a obres culturals lliures.

El llicenciador no pot revocar aquestes llibertats, sempre que seguiu els termes de la llicència.

Termes:

- **Reconeixement**: Heu de reconèixer l'autoria de manera apropiada, proporcionar un enllaç a la llicència i indicar si heu fet algun canvi. Podeu fer-ho de qualsevol manera raonable, però no d'una manera que suggereixi que el llicenciador us dóna suport o patrocina l'ús que en feu.
- **CompartirIgual**: Si remescleu, transformeu o creeu a partir del material, heu de difondre les vostres creacions amb la mateixa llicència que l'obra original.
- **No hi ha cap restricció addicional**: No podeu aplicar termes legals ni mesures tecnològiques que restringeixin legalment a altres de fer qualsevol cosa que la llicència permet.

Avisos:

- No heu de complir amb la llicència per als elements del material en el domini públic o quan el seu ús està permès per l'aplicació d'una excepció o limitació dels drets d'autor.
- No es donen garanties. La llicència pot no ser suficient per autoritzar la utilització que en voleu fer. Per exemple, altres aspectes com la publicitat, la privacitat, o els drets morals poden limitar la forma d'utilitzar el material.

El text complet el podem trobar a:

https://creativecommons.org/licenses/by-sa/4.0/legalcode

9. Codi: Cal adjuntar el codi amb el que heu generat el dataset, preferiblement amb R o Python, que us ha ajudat a generar el data set

S'adjunta el codi en el notebook **dcampilloca_Practica1_webscraping.ipynb**

10. Dataset: Dataset en format

S'adjunta el resultat en **alergenos.csv** i carpeta **imagenes**

Bibliografia:

Subirats, Laia; Calvo, Mireia; (2018); *Web Scraping.* Editorial UOC.

Richard, Lawson; (2018-10-02); *Web Scraping with Python.* PACKT Publishing open source

Creativecommons.org: https://creativecommons.org/licenses/by-sa/4.0/deed.ca

https://creativecommons.org/licenses/by-sa/4.0/legalcode

ticbeat.com: http://www.ticbeat.com/salud/alergicos-tardan-87-minutos-en-hacer-la-compra/

El Corte Inglés: https://www.elcorteingles.es/supermercado

Nestlé: https://www.helados.nestle.es/cliente/alergenos.aspx

https://www.helados.nestle.es/cliente/Content/pdfs/alergenos/listado-alergenos-2018-la-lechera.pdf

Bonpreu Esclat: https://www.bonpreuesclat.cat/es/inici?p_p_id=101&p_p_lifecycle=0&p_p_state=maximized&p_p_mode=view&_101_struts_action=%2Fasset_publisher%2Fview_content&_101_returnToFullPageURL=%2F%3Fp_p_id%3D3&_101_assetEntryId=36529&_101_type=content&_101_urlTitle=fleca_qualitat_que_oferim&inheritRedirect=true

Jo també sóc al·lèrgic: https://jotambesocallergic.cat/wordpress_4/web-ca/botiga/xocolata/piruleta-xocolata-cor/

Allergeneat: https://www.allergeneat.com/

Stackoverflow: https://stackoverflow.com/questions/24226781/changing-user-agent-in-python-3-for-urrlib-request-urlopen/24226797

https://stackoverflow.com/questions/48906246/extract-json-from-html-in-python-beautifulsoup?rq=1

Python.org: https://docs.python.org/3/howto/regex.html

Dataquest.io: https://www.dataquest.io/blog/web-scraping-tutorial-python/

https://www.dataquest.io/blog/web-scraping-beautifulsoup/

Toddhayton.com: http://toddhayton.com/2015/01/16/scraping-by-example-json-data/