Marc González
Maria Sunyer

Web Scraping: Ulabox

Marc González
Maria Sunyer

Context

Hem decidit realitzar *scraping* a la pàgina web del supermercat “Ulabox”. Ens hem decantat per fer scraping d’un supermercat perquè considerem que els últims mesos s’estan disparant alarmes socials en torn al preu de la cistella de compra[1], per tant, el fet de poder obtenir les dades d’un supermercat ens pot permetre aplicacions que cerquin ofertes, tendències de preus, etc.

A més a més, l'experiència obtinguda durant la realització de la pràctica és reutilitzable al fer scraping d’altres supermercats o pàgines basades amb productes.

En quant a supermercat, hem decidit fer servir Ulabox com a exemple degut a que la pàgina[2] robots.txt, permeteix fer scraping de qualsevol element de la pàgina https://www.ulabox.com/robots.txt

L’adreça del supermercat és al següent:

- https://www.ulabox.com/

Títol

El dataset hem decidit anomenar-lo “**data_productes_alimentacio_ulabox**” el fet d’incloure el tag “data” és degut a que l’objectiu seria poder extreure dades de forma periodica i no tan sols extreure l’informació d’un sol día.

Per la realització de la pràctica, el dataset resultat s’ha anomenat:

- Nov-19-2022_productes_alimentacio_ulabox.csv

Descripció del dataset

Aquest dataset agrupa informació referent a tots els productes d’alimentació disponibles al supermercat Ulabox. Per cada producte, obtenim l’informació mostrada en la página web: el seu nom, imatge, preu, informació nutricional i fabricant. A més a més, guardem l’enllaç a on es troba el producte i la categoria i subcategoria dins del supermercat.

Hem separat el dataset en dues parts, la primera conté les dades dels productes (descripció, preu…) i la segona, conté les imatges dels productes. Per tant el dataset conté de dues parts separades pero relacionades per el camp “ID”.

Les parts es troben al projecte:

- /csv/ : Conté el dataset normal.
- /img/ : Conté les imatges dels productes evaluats.

---

[1] Noticia sobre el preu dels aliments, web: https://elpais.com/economia/2022-10-13/el-subidon-de-la-cesta-de-la-compra-en-los-ultimos-siete-meses-resumido-en-un-hilo-de-twitter.htm

[2] Ulabox robots, web: https://www.ulabox.com/robots.txt

Marc González
Maria Sunyer

Representació gràfica

El dataset resultant es pot representar en la següent taula:

Atributs	Tipus de Dades
Id	Numèric
Categoría	Text
Subcategoria	Text
Enllaç	Text
Nom Producte	Text
Preu	Text
PreuBase	Text
Ingredients	Text
Valor Energètic Kj	Text
Valor Energetic KC	Text
Grases	Text
Hidrats	Text
Sucre	Text
Proteines	Text
Sal	Text
Fabricant	Text
Valor Energètic KC	Text

Marc González
Maria Sunyer

Contingut

El període de temps de les dades per tots els camps que inclou el dataset és d'el día en que s'executa el programa, en el nostre cas, divendres divuit de novembre.

Els camps que inclou el dataset són els següent:

- Hi ha quatre camps inicials que és generen al detectar els productes a analitzar.
    - Id: Atribut númeric que ens serveix per controlar el nombre de productes del dataset.
    - Categoría: Atribut en el que guardem la categoria principal d'un producte (per exemple, Olis i condiments).
    - Subcategoría: Atribut en el que guardem la subcategoria d'un producte (per exemple, Olis d'oliva).
    - Enllaç: Enllaç al producte.
- Hi ha quatre camps secundaris que descriuen els productes i són camps que gairebé sempre trobem disponibles:
    - Nom Producte: Atribut en el que guardem el nom "comercial" del producte.
    - Preu: Atribut amb el preu del producte.
    - Preu Base: Atribut amb el preu base del producte (per exemple, el preu per kilogram)
    - Ingredients: Atribut que no sempre és disponible, guardem els ingredients que contenen els productes.
- Finalment, com a informació complementària guardem la taula nutricional i el fabricant del producte. Aquesta informació no sempre está disponible.
    - Taula nutricional: Atributs que guarden la informació nutricional del producte (Valor Energètic KJ, Valor Energetic KC, Greixos, Hidrats, Sucre, Proteines, Sal)
    - Fabricant: Empresa responsable de la manufactura del producte.

Propietari

Les dades provenen del supermercat “Ulabox” que és el responsable de la pàgina web. Al tractar-se d’una empresa privada, no proporciona datasets públics i per tant, no hem dispossat d’anàlisis anteriors.

Per aquest motiu hem cercat datasets similars de supermercats, decidint utilitzar categories similars als trobats amb l’empresa “Tesco”, que és una cadena gran d’Estats Units.

El dataset de refèrencia pot trobar-se:

- https://data.world/crawlfeeds/tesco-groceries-dataset

Per seguir amb els principis legals i ètics durant el projecte, ens hem preocupat per:

- Principis legals:Hem seguit les següents consideracions.
    - Hem cercat als termes i condicions per seccions o paraules clau referents a la prohibició de l’extracció de les dades, sense trobar cap prohibició[3].
    - A nivells de LGDPDD, considerem que el dataset i les dades utilitzades entren en la categoría: “Son fuentes de acceso público o los datos se recaban por un fin de interés público general.” al tractar-se de dades d’accès públic.
    - No hem comes frau al accedir a les dades, totes són públiques.
    - Hem llegit el fitxer robots.txt per asegurar que la pàgina permetia l’ús d’agents.
    - No hem fet ús indegut d’ordinadors, al realitzar la pràctica amb els nostres recursos.
    - No hem utilitzat dades protegides per drets de propietat intel·lectual.
- Principis ètics: Els nostres principis ètics s’han centrat en dos camps.
    - El primer ha estat seguir bones pràctiques en el transcurs de la pràctica, hem treballat sobre html’s estàtics i hem intentat no accedir a la pàgina si no era estrictament necessari.
    - En segon punt, hem realentitzat l’execució del spider, per tal de no colapsar de peticions el servidor. Triant el ralentir l’obtenció del dataset a paralelitzar l’obtenció de dades i arriscar-nos a generar problemes a l’organització.

      També ens hem limitat a rastrejar dades públiques i que eren permeses tant a nivell legal com a nivell de termes i condicions (descartant altres supermercats com El Corte Inglés, per exemple).

      Finalment, la creació del dataset s’utilitza amb finalitats educatives i no comercials.

[3] Termes i condicions: https://www.ulabox.com/terminos-y-condiciones?ula_src=front_index&ula_mdm=footer_block

Marc González
Maria Sunyer

Inspiració

Hem considerat interessant aquest conjunt de dades perque en la situació actual d'inflació economica, és important poder analitzar l'evolució del preu dels productes. Si obtenim aquestes dades durant un període de temps podem respondre les següents preguntes:

- Ha augmentat el preu dels productes de forma generalitzada?
- Quins productes estan més afectats per la inflació?
- Hi ha productes que han reduit la quantitat per evitar haver d'augmentar el preu?
- Hi ha productes que han variat la seva composició per reduir costos?

A més a més, dades com les contingudes a les taules nutricionals, ens permetrien alimentar models que generen "dietes" equilibrades i basades en un pressupost baix, o reduint un tipus de macronutrient (o ampliant-lo). També podriem utilitzar les dades com a base d'un programa per crear receptes i generar una llista de compra automática, etc.

Llicència

La llicència que hem considerat és la següent:

- Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

Considerem que el nostre dataset pot ser interessant per els projectes que hem citat, per tant, volem facilitar que es distribueixi, modifiqui i adapti el dataset. Tot i així, per prevenir, preferim limitar l'ús a estrictement aplicacions no comercials.

Aquesta limitació la hem decidit per reduir el risc de que s'utilitzi el projecte per analitzar els preus de la botiga i és faci una competencia "deslleial". Sabem que totes les empreses utilitzen eines similars per coneixer el mercat, però com l'esperit del nostre projecte és educatiu i neix de la curiositat de tenir unes dades base desde les que seguir aprenent i extreure'n conclusions, preferim que segueixi així.

Marc González
Maria Sunyer

Codi

Per la realització del codi hem utilitzat Python i es pot trobar en el següent repositori de GitHub:
Enllaç al repositori: PRAC1_Tipologia
Enllaç al dataset: https://zenodo.org/record/7343170

En el nostre repositori es pot trobar la següent estructura:

- img/: Aquí es poden trobar les imatges dels productes, el nom de la imatge fa referència a l'ID del producte.
- pdf/: Document de la pràctica
- dataset/: dataset resultant de l'execució del codi.
- src/: codi Python.

Per la realització de la pràctica hem utilitzat les següents llibreries:

- Beautiful Soup: hem utilitzat aquesta llibreria per elaborar l'anàlisi de les pàgines, obtenir l'estructura d'Ulabox i per poder obtenir la informació dels productes.
- Requests: hem usat aquesta llibreria per descarregar-nos les pàgines web dels productes, ja que no era necessari l'execució prèvia del codi per l'obtenció de la informació.
- Selenium: hem fet servir aquesta llibreria per descarregar-nos les pàgines web necessàries per a l'obtenció de l'estructura d'Ulabox, i per l'obtenció dels enllaços dels productes. Inicialment, vam intentar dur a terme aquesta part fent servir Requests, però vam trobar que els fitxers aconseguits mancaven la informació que necessitàvem i vam arribar la conclusió que era necessari una execució inicial del JavaScript de la pàgina web perquè completes aquesta la informació que necessitàvem.
- Pandas: hem fet servir aquesta llibreria per la construcció del dataset.
- Shuttil: hem fet ús aquesta llibreria per guardar les imatges.
- Re: hem utilitzat aquesta llibreria per cercar informació en el document mitjançant expressions regulars.
- Time: hem utilitzat aquesta llibreria per espaiar les peticions HTTP i així evitar saturar el servidor.
- Json: hem fet servir aquesta llibreria per llegir fitxers json. En el nostre cas, com que considerem que l'obtenció de l'estructura de la pàgina i l'obtenció dels enllaços dels productes és una informació que no cal actualitzar cada vegada que es vulgui obtenir un nou dataset, de manera que donem l'opció de carregar un JSON amb aquesta informació i així reduir el temps necessari per obtenir el dataset.

El nostre codi es pot dividir en 3 parts. Inicialment, hi ha la part d'obtenció dels enllaços dels productes conjuntament amb l'estructura de la pàgina. En aquesta part

Marc González
Maria Sunyer

és a on fem servir Selenium per l'obtenció de la informació. En la segona part hi hauria l'obtenció de la informació del producte i per últim hi hauria la creació del dataset i l'obtenció i emmagatzematge de les imatges dels productes.

Vídeo

Enllaç al video:
https://drive.google.com/file/d/1pwPw2BZQqYoS_qOangnqGEfAZDlL1VIg/view?usp=sharing

Enllaç al dataset: https://zenodo.org/record/7343170