PRÀCTICA 1 Tipologia i cicle de vida de les dades

Laura Guzman

1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Per a la realització de la pràctica s'ha triat la pàgina web d'Ikea. Concretament ens centrarem en el producte "cadira" (https://www.ikea.com/es/es/cat/sillas-fu002/)
Aquest dataset és interessant per als venedors de cadires (competidors del sector de mobles), ja que es poden comparar els preus segons el tipus de producte i accedir a una petita descripció del producte.

S'ha triat aquesta pàgina web ja que és estructurada, de manera que ens ha semblat adequada per iniciar-se en el web scraping, ja que compleix els objectius:
- Aplicar coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous
- Identificar dades rellevants que el seu tractament aporten valor a una empresa i la identificació de nous projectes analítics
- Capturar dades
- Actuar amb principis ètics i legals
- Desenvolupar capacitat de cerca, gestió i ús de la informació i recursos.

A part, s'ha verificat que a l'arxiu robots.txt (https://www.ikea.com/robots.txt) no aparagués la prohibició per a robots per a entrar al link seleccionat.

2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

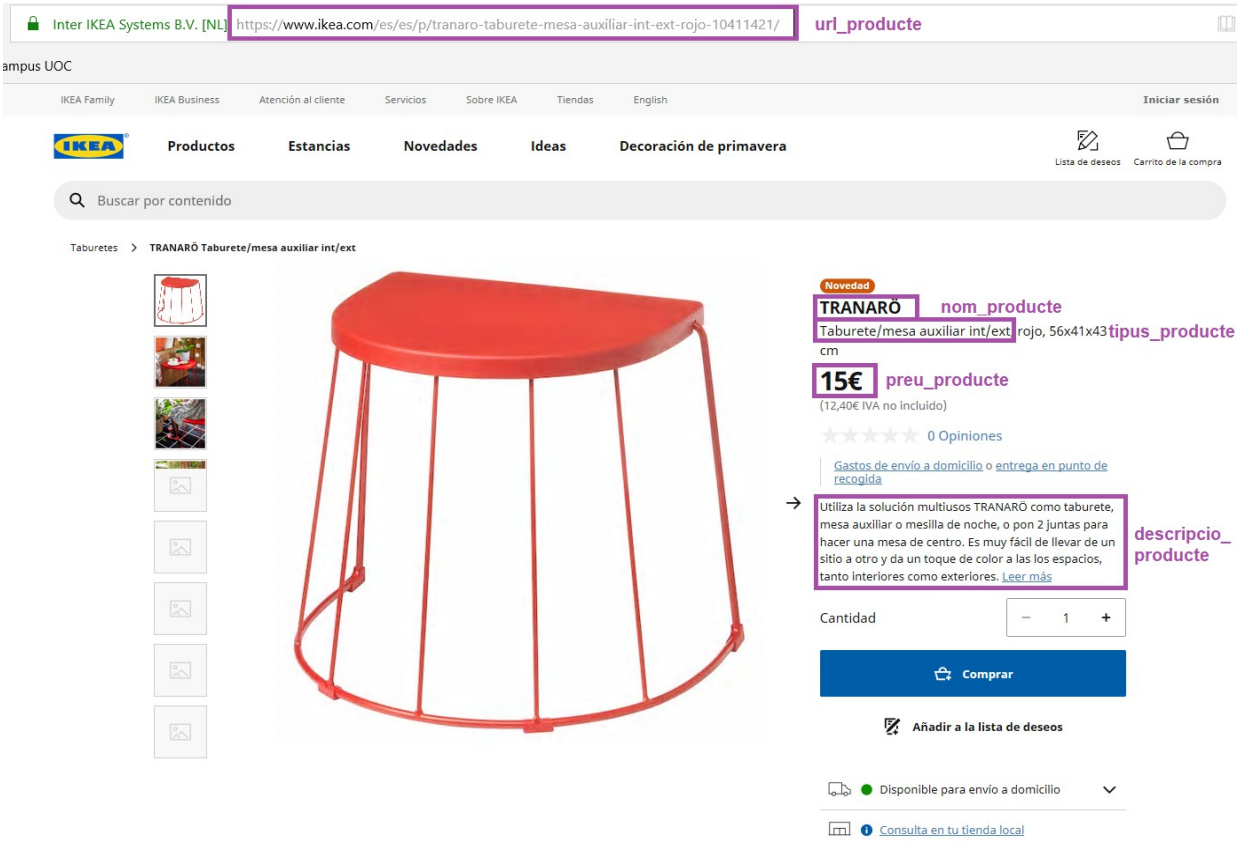
El títol per les dades extretes (dataset) és : cadires_ikea , ja que aquest nom conté el producte seleccionat i el nom de l'empresa d'on s'ha agafat.

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

En el dataset s'hi troben les dades extretes de la pàgina web d'Ikea en relació als productes cadires.
Al dataset hi els següents atributs del producte:

nom_producte: nom comercial del producte
tipus_subproducte: tipus de subproducte: cadira, tamburet, silló, banc...
preu_producte: preu actual del producte.
url_producte: Url que permet accedir a la pàgina del producte
descripcio_producte: Descripció del producte, si n'hi ha

4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment
El dataset és integrament extret de les pàgines web de producte silla de ikea i dels links de cada producte. Per resumir-ho, en una imatge de la pàgina web d'ikea, s'han marcat els atributs del dataset perquè així es veu de manera global la informació extreta per cada producte.

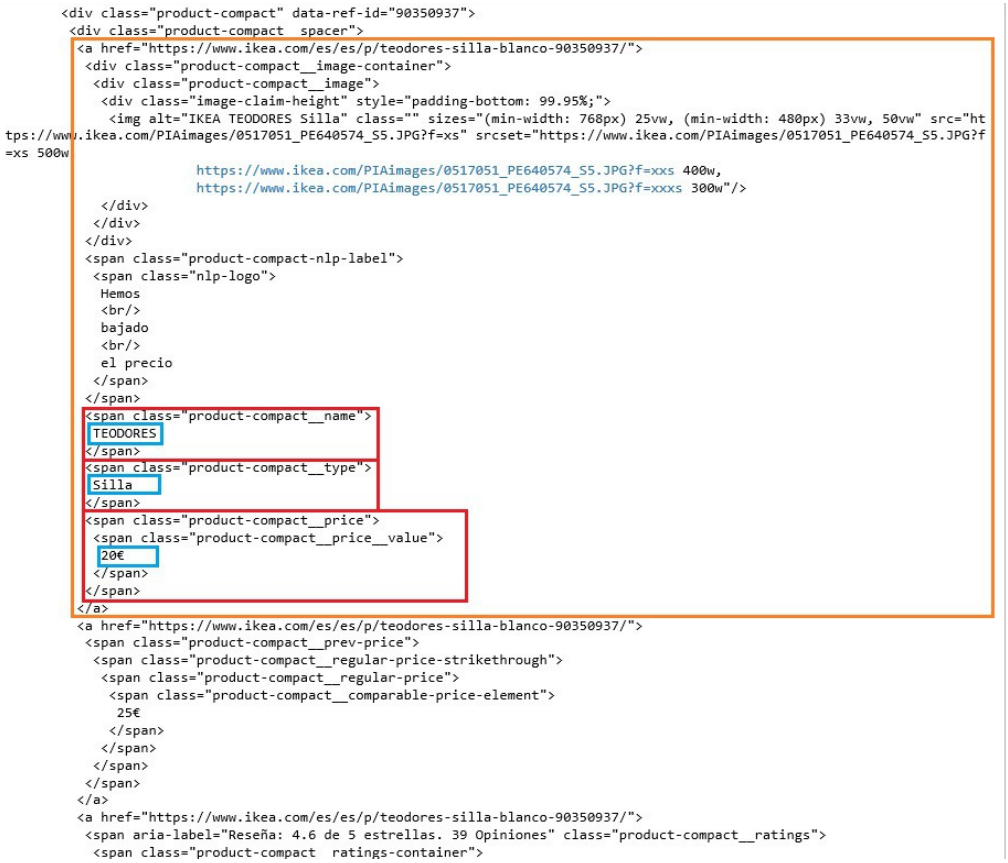


5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

Per a extreure el dataset, s'han seguit varis passos:
Primer de tot, s'ha buscat l'enllaç que ens porta a la secció de cadires d'ikea. La secció de cadires consta de 23 pàgines d'articles, de manera que per agafar tots els articles, hem de recòrrer aquestes 23 pàgines. Per a fer-ho, hem buscat al codi de la web d'ikea el link per anar a la següent pàgina.
Per tant, la url d'inici és https://www.ikea.com/es/es/cat/sillas-fu002/ i llavors dins d'un for es carrega la url, s'extreuen els articles d'aquesta pàgina i llavors busca la següent url per tornar a carregar. Simplificant seria:

for i in range(n_pag_inici,n_pag_final):
carregar url a beautiful soup
extreure atributs productes
agafar la següent url (que es troba a <link href="https://www.ikea.com/es/es/cat/sillas-fu002/page-#/" rel="next"/>)

Els atributs que s'extreuen en aquest bucle són :
nom_producte
tipus_subproducte
preu_producte
url_producte
on la url es troba dins de <a href="url"> i el nom, tipus i preu, es troben dins de en les classes product-compact__name, product-compact__type i product-compact__price__value, com es pot veure en la següent imatge:



Cada atribut és una llista, i llavors quan tenim tots els atributs, els ajuntem en un dataframe.

A part d'aquests atributs mencionats, també s'ha trobat important agafar la descripció de cada producte, que es troba accedint al link propi de cada article (llista que hem recollit url_producte). Per agafar la descripció de cada producte, hem fet un bucle que reorre tots els links i de cadascun en carrega la pàgina i n'extreu la descripció, que es troba a < div class=product-pip__benefit-summary>. Alguns productes no tenen descripció, de manera que llavors introduim "" a la llista.

Ara ja tenim tots els atributs que ens interessen pel dataset i amb pandas, creem un dataframe que conté totes les llistes creades (nom, tipus, preu, url i descripció) i mostrem la taula per veure que tota la informació agafada és correcte.

Finalment, carreguem el dataframe creat en un csv .

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Els propietaris de la web on s'extreu informació, www.ikea.es, són Ikea Ibérica,sa i Ikea Norte,sl. El grup IKEA va ser fundat al 1943 per Ingvar Kamprad, començant per un petit negoci de venta per correspondència i actualment és una marca mundial .

Els propietaris de la web on es troba el conjunt de dades, no han aprovat ni autoritzat aquest treball.

En les condicions d'ús de la web, ens indica que "La visita o utilización del Portal por tu parte deberá hacerse en todo momento de forma responsable y ajustada a la legalidad vigente, las reglas de la buena fe, y respetando en todo caso los derechos de propiedad intelectual e industrial titularidad de las Sociedades, cualquier otra sociedad mercantil del grupo IKEA, u cualesquiera otras terceras personas, físicas o jurídicas."

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Ikea ha revolucionat el món dels negocis, i pot ser un referent per molts venedors de mobles. Gràcies a aquest conjunt de dades es poden saber els productes que té ikea i els seus preus actuals, d'aquesta manera, gràcies a aquest dataset, es pot analitzar la competència, perquè es poden comparar productes i preus del principal competidor.

1 8. Llicència. **Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:**
○ 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

En aquest cas, al haver utilitzat les dades d'una pàgina web que no és pública, triaria una llicència released under CC BY-NC-ND, que és una llicència restrictiva que tant sols permet la descàrrega i que es pugui compartir sempre que se'n reconegui l'autoria però no poden ser modificades ni utilitzades per a finalitat comercial.

9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement enPython o, alternativament, en R.

Arxiu "codi_ikea_5.ipynb"

10. Dataset. Presentar el dataset en format CSV

Arxiu "cadires_ikea.csv"