Pràctica 1 – Tipologia i cicle de vida de les dades

COMPONENTS DEL GRUP: **Arnau Santos Ribelles** i **Ferran Pintó Haro**.

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

El conjunt de dades correspon a una extracció amb Web Scraping de productes a la pàgina web de Back Market. Back Market és una plataforma online que es dedica a vendre productes reacondicionats. S'ha recollit informació sobre els Iphones que apareixen a la seva pàgina en un moment donat, amb la opció de poder substituir la cerca dels "Iphone" per qualsevol altre producte, ja que el canvi implicaria el canvi d'una variable en el codi i alguns retocs més.

El lloc web proporciona aquesta informació perquè es dedica a vendre productes reacondicionats, entre els quals Iphones; mostrant el model, preu, puntuació, etcètera. Així doncs, s'han aprofitat aquestes dades dels productes a la venta (concretament d'Iphones) a la pàgina web de Back Market.

2. **Títol**. Definir un títol que sigui descriptiu pel dataset.

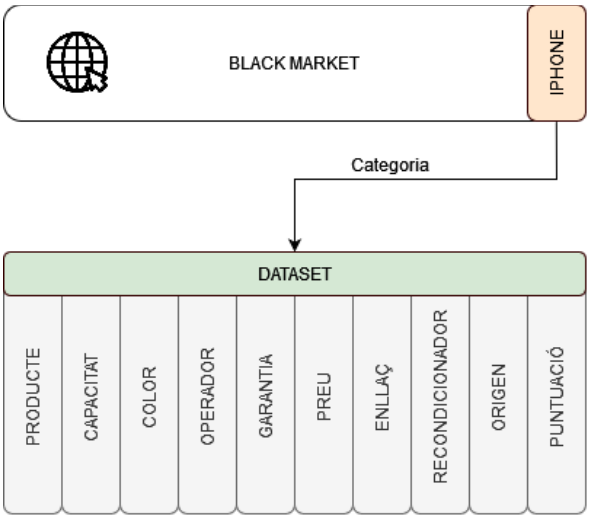
Iphones publicats a la pàgina web de Back Market en un moment donat, amb les seves característiques.

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 escollit

El conjunt de dades extret recull tots els productes que hi ha a la venda en la tenda online "Black market" en funció d'una categoria específica. En aquest cas, la categoria cercada és "iphone". Per tant, el dataset inclou tots els telèfons mòbils "iphone" reacondicionats disponibles en la tenda online en un moment concret. Per cada producte extret s'obtenen les seves característiques principals *(veure punt 5)*, entre d'altres l'enllaç per accedir a la pàgina de venda del producte concret.

Les dades han passat un cert procés de neteja al extreure-les, filtrant els productes que realment són Iphone, separant les característiques en capacitat, color i operador i transformant el preu i la puntuació a variables decimals. Tot i així, encara poden existir inconsistències i el format no és necessàriament el més adequat per una anàlisi directa. Per exemple, els "strings" estan emmagatzemats com a "object" i la variable capacitat no està emmagatzemada com a numèric ja que conté la unitat de mesura (per exemple, "128 GB"), igual que la variable garantia (per exemple, "Garantía: 24 meses"). El dataset un cop extret no s'ha netejat/preprocessat.

4. **Representació gràfica**. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.



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

El dataset inclou els camps següents:

- El producte que és (***Producte***) [tipus *string*. Exemple: "iPhone 11"]
- La seva capacitat en Gigues (***Capacitat***) [*string*. Exemple: "128 GB"]
- El seu color (***Color***) [*string*. Exemple: "Negro"]
- Si és lliure d'operador (***Operador***) [*string*. Exemple: "Libre"]
- El preu del producte (***Preu***) [*float*. Exemple: "258.58"]
- La puntuació del producte (***Puntuacio***) [*float*. Exemple: "4.1"]
- L'empresa que ha reacondicionat el producte (***Reacondicionador***) [*string*. Exemple: "NomoPhone"]
- Des d'on s'envia el producte (***Origen_enviament***) [*string*. Exemple: "Francia"]
- Els mesos de garantia (***Garantia***) [*string*. Exemple: "Garantía: 24 meses"]
- La pàgina web de cada producte (***Url***) [*string*. Exemple: "https://www.backmarket.es/iphone-xr-128-gb-negro-libre-segunda-mano/199289.html#l=12"] .

Els productes són variants a la pàgina web en funció del moment i, per tant, les dades obtingudes a l'executar el script corresponen als Iphone en venda en aquell moment a la pàgina. Les dades del dataset són corresponents al dia 25/02/2022 a les 19:56; moment en que es va executar el Script per tal de "scrapejar" la pàgina web i obtenir les dades.

Les dades s'han recollit mitjançant un script de Python (i amb l'ajuda de la llibreria BeautifulSoup) que realitza un procés de web scraping, extraient els productes de la pàgina. Es troba primer, el número de pàgines que contenen Iphones al buscar-ho i cerca diferents valors a totes les pàgines (els productes, la seva garantia, preu i url de cada producte). Posteriorment busca a la url de cada producte, altres variables que no estan a la pàgina general on es mostren tots els productes: les característiques (Giges, color i lliberat d'operador), el país des d'on s'envia el producte, l'empresa que va reacondicionar el producte i la puntuació. Finalment es crea un dataframe que s'exporta en un CSV. Abans d'exportar-lo hi ha una petita neteja de dades en la que, entre d'altres, es filtren només els

productes que siguin Iphones, ja que al buscar “iphone” a la pàgina es mostren com a resultat fundes i altres telèfons mòbils, i que per tant també s’han registrat.

6. **Agraïments**. Presentar el propietari del conjunt de dades. És necessari incloure cites d’anàlisis anteriors o, en cas de no haver-n’hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s’han seguit per actuar d’acord amb els principis ètics i legals en el context del projecte.

Agraïments al propietari del conjunt de dades: Back Market, malgrat exclou a tots els robots d'alguns índexs de la seva pàgina. L'ús de les dades serà exclusivament acadèmic.

No s’han trobat anàlisis anteriors en aquesta pàgina web, només serveis que ofereixen el web scraping en aquesta pàgina *1. Del recurs 2 s’han agafat algunes idees ja que presenta un anàlisi similar: la construcció d’un programa de Python que mitjançant web scraping entra a Ebay, recorre totes les publicacions d’un telèfon mòbil concret (Iphone X) i les extreu en un Excel. Una altre pàgina visitada per observar anàlisis similars i agafar idees ha estat el recurs 3, que fa Web Scraping d’Amazon per emmagatzemar en un csv els monitors de pantalla que hi apareixen en realitzar aquesta cerca en les diferents pàgines. D’aquest recurs s’ha obtingut la idea general de com fer web scraping d’una pàgina de venta de productes i de com passar pàgina quan al realitzar una cerca apareixen diverses pàgines de resultats. I per últim el recurs 4 estudiat fa un Web Scraping a Amazon d’un sol producte per guardar el preu en un txt i rebre un correu electrònic quan baixi més de 10€.

Tant Ebay com Amazon també són plataformes especialitzades en venta online, on s’especifiquen les característiques de cada producte que tenen publicat per a la venta.

Pel que fa als principis ètics i legals en el context del projecte, el primer pas realitzat per actuar d’acord amb els principis ètics i legals ha estat verificar l’arxiu *robots.txt* de la pàgina. Aquests indiquen que exclouen tots els robots de l’accés al directori "*/search/*" i, per tant, a la cerca de productes dins la pàgina. Les restriccions són un suggeriment i malgrat la voluntat del propietari de no ser rastrejat a continuació s’indiquen els motius del rastreig.

Cal apuntar que el rastreig que es fa es tracta d’informació pública, és un lloc web que exposa informació públicament, sense acceptar termes i condicions; i no és necessari iniciar sessió en el lloc web. No s’accedeix a l’ordinador d’una altra persona ni s’interfereix en la propietat personal d’un individu o empresa. D’altra banda, no es tracta de dades delicades ni sensibles. També cal matisar que la informació extreta no s’utilitza ni utilitzarà amb finalitats comercials, sinó merament acadèmiques. Així, es compleix un ús limitat del material protegit per drets d’autors sense el permís explícit del titular dels drets: la recerca és considerat un ús legítim.

Cal remarcar la modificació del user agent per prevenir la revisió d’aquesta capçalera i el bloqueig de la cerca. S’ha intentat no causar dany i no saturar de peticions el servidor web, malgrat es fan força peticions (entre 400 i 500 aproximadament) per obtenir de manera acadèmica la informació més útil possible.

*1 https://webautomation.io/pde/backmarketes-extractor/326/

*2 https://www.youtube.com/watch?v=mbjZNTjHh0Y

*3 https://www.youtube.com/watch?v=_AeudsbKYG8

*4 https://www.youtube.com/watch?v=TcEiCMpEynU

7. **Inspiració**. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

El conjunt de dades permet tenir tots els Iphones en venta publicats a la pàgina de Back Market en un dataset, per poder visualitzar fàcil i ràpidament les diferents característiques d'aquests. Veient la funcionalitat, la utilitat que més ràpid emergeix com a idea és que podria ser rellevant per persones que cerquin comprar-se un telèfon i vulguin comparar preus i característiques de manera senzilla i manejable.

Malgrat el seu ús no sigui aquest, podria ser útil per analitzar la competència (si ho utilitzés una empresa que es dediqués a vendre iphones, per comparar productes i preus de Back Market), per resumir la informació de la pàgina web que està distribuïda en diferents pàgines, i, com s'ha comentat, una persona individual que en la situació de comprar-se un Iphone ho podria fer servir per analitzar els diferents iphones a la venta en un moment concret a la pàgina, comparant les seves característiques i preus.

Par altra banda, aquest conjunt de dades podria ser d'especial interès per realitzar models d'estimació de preus de diferents productes reacondicionats. Així es podria emprar per construir tot un model de preus que optimitzi tant en la venda com en la compra d'aquest tipus de productes. Per tant, amb aquesta finalitat podria ser emprat pel venedor com pel particular.

La única diferència entre l'anàlisi presentat i les cites presentades a l'apartat 6 utilitzades com a "inspiració" és la plataforma en la qual s'aplica el Web Scraping. Les tècniques d'anàlisi utilitzades són semblants, però s'han adaptat a les particularitats del nostre cas d'estudi.

La cita que inclou el Web Scraping de Ebay (cita 2) fa la cerca amb la finalitat principal de trobar les millors ofertes de manera senzilla. La cita 3 és acadèmica però no busca respondre més preguntes que la de simplement extreure informació de productes i preus d'Amazon. D'altra banda la cita 4, referent al Web Scraping d'un sol producte a Amazon fa la cerca amb la finalitat de saber quan el producte assenyalat baixa de preu (una càmera de fotos). Guarda cada hora el preu i fa que si en el moment que fa la cerca a la pàgina, el preu és inferior (10€ menys en aquest cas), li enviï un correu avisant-lo.

Per tant els interessos dels anàlisis presentats a l'apartat 6 són a ús individual, per trobar la millor oferta. En el primer cas i el segon tenint totes les dades dels productes en un Excel/csv per poder-les comparar i en l'altre scrapejant un producte concret i rebent un correu quan baixa de preu. L'anàlisi presentat en el present document pretén apuntar més utilitats, com ja ha estat especificat, tant a nivell individual com empresarial (tant pel venedor com empreses competidores).

8. **Llicència**. Selecciona una d'aquestes llicències pel dataset resultant i justifica el motiu de la seva selecció.

La llicència escollida per la publicació d'aquest conjunt de dades és **CC BY-SA 4.0 License**. Aquesta ha estat la llicència escollida pels següents motius:

- Atribució: Es proporciona el nom/s del/s creador/s del conjunt de dades i s'indiquen els canvis realitzats envers les dades originals. D'aquesta manera es pot observar fàcilment en quina mesura s'han realitzat aportacions en relació al treball original.

- Ús no comercial: no es pot utilitzar el dataset amb fins comercials.

- Si es transforma o construeix a partir del material, s'han de distribuir les contribucions mitjançant la mateixa llicència. Això fa que el dataset corresponent al present estudi s'hagi de distribuir, en cas de modificació, sota la mateixa llicència, ja que la voluntat de la pàgina on s'han obtingut les dades és l'ús no comercial d'aquest material.

En resum, el dataset es podria compartir (copiar i redistribuir) i adaptar (transformar i crear) sota les condicions de no fer-ne us comercial, indicant els canvis si se'n fan i distribuint-ho sota la mateixa llicència.

9. **Codi**. Adjunta al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

El codi Python amb el que s'ha generat el dataset està adjuntat al repositori Git, en la carpeta **`codi`** i el nom "**`Practica1-WebScraping.ipynb`**".

10. **Dataset**. Publicar el dataset obtingut en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

El dataset obtingut ha estat publicat en format CSV a Zenodo amb el nom de "PRAC1_Iphone-Backmarket.csv" i una breu descripció. Els camps es separen amb ";". S'adjunta, a part de a l'arxiu README, el DOI: **10.5281/zenodo.6385536**

L'enllaç DOI és **https://doi.org/10.5281/zenodo.6385536**.

TAULA DE CONTRIBUCIONS

Contribucions	Signatura
Investigació prèvia	FPH, ASR
Redacció de les respostes	FPH, ASR
Desenvolupament del codi	FPH, ASR

Recursos

1. Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC

2. Masip, D. (2010). El lenguaje Python. Editorial UOC.