Integrant: Joan Codinach Ortiz

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

   L'adreça del lloc web és https://quelibroleo.com/ . He triat aquesta pàgina web perquè estan molt de moda els llocs de vendre contingut digital i fer pagar una subscripció mensual per tenir accés aquest contingut. Hi ha moltes pàgines amb aquesta idea i de diferents tipus de contingut: llibres, pel·lícules, entrevistes, pòdcasts, cursos... . He trobat interessant comprovar que si veig un catàleg en una pàgina web, ja sigui de pel·lícules, llibres..., es pot crear un dataset a partir de totes d'aquestes utilitzant Web Scrapping.

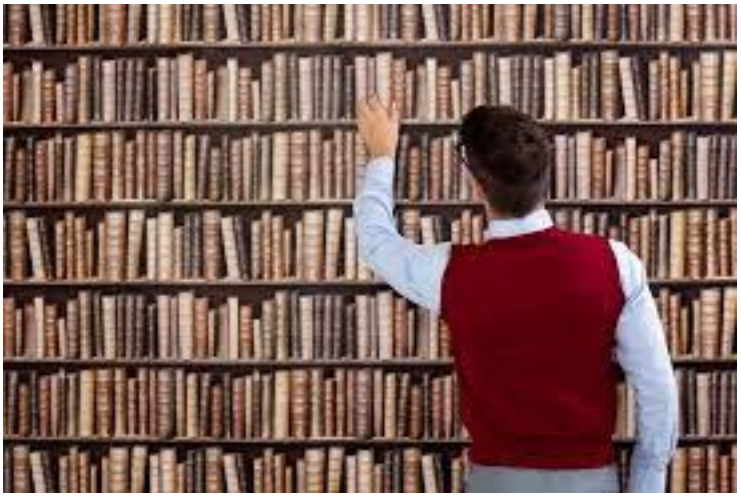
2. **Títol.** Definir un títol que sigui descriptiu pel dataset.

   Valoració dels millors llibres de cada gènere.

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 dataset conté informació dels millors llibres de cada gènere i les seves puntuacions. De cada gènere, hi ha un rànquing de 50 llibres com a màxim dels llibres més ben valorats. Les puntuacions són fetes pels usuaris de la pàgina, la qual poden votar i fer una crítica de cada llibre.

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



   Aquesta imatge, per una banda, representa un dels motius pel qual es crea aquest dataset, que és el de saber quin llibre ens pot agradar més (recomanador) i per altra fa èmfasis que si tenim la informació dels llibres ordenada, serà més fàcil trobar i escollir el que volem llegir.

5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de les dades.

El dataset conté la següent informació:

- Títol: Conté el títol del llibre
- Autor: Conté la persona que ha escrit el llibre
- Gènere: Categoria que pertany el llibre. A continuació mostrem tots els gèneres que s'utilitzen en el lloc web:
    - Actores
    - Arte
    - Autoayuda
    - Autoayuda Y Espiritualidad
    - Biografías, Memorias
    - Ciencias
    - Ciencias Humanas
    - Ciencias Políticas Y Sociales
    - Clásicos De La Literatura
    - Cocina
    - Cómics, Novela Gráfica
    - Deportes Y Juegos
    - Derecho
    - Dietética Y Nutrición
    - Economía
    - Empresa
    - Ensayo
    - Estudios Y Ensayos
    - Fantástica, Ciencia Ficción
    - Ficción Literaria
    - Filología
    - Fotografía
    - Guías De Viaje
    - Historia Del Cine
    - Histórica Y Aventuras
    - Humor
    - Infantil Y Juvenil
    - Informática
    - Juvenil
    - Lecturas Complementarias
    - Literatura Contemporánea
    - Medicina
    - Música
    - Narrativa
    - Narrativa Histórica
    - No Ficción
    - Novela Negra, Intriga, Terror
    - Poesía
    - Poesía, Teatro
    - Psicología Y Pedagogía
    - Romántica, Erótica
    - Varios

- Posició : Conté la posició del rànquing la qual ocupa el llibre dins de la categoria
- Nota mitjana: La mitjana de tots els vots que ha fet els usuaris de la pàgina
- Vots: Nombre de cops que s'ha votat el llibre
- Nota: Valoració del llibre. Aquests són els possibles valors:
    - Pèsimo
    - Malo
    - Regular
    - Bueno
    - Muy bueno
    - Excelente

- Critiques: Nombre de critiques que te el llibre
- Resum: Resum del llibre

El període de temps de les dades pot començar des de que es va escriure els primers llibres. Les valoracions de músiques, llibres i pel·lícules les que estan en el rànquing són les noves. Per tant, tots els llibres que hi ha en el dataset són actuals.

6. **Propietari.** 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.

Mirant la pàgina trobem un apartat anomenat "qui som?", el qual semblava que podria donar informació sobre els creadors de la pàgina web, però no ha sigut el cas. Hem utilitzat la llibreria whois per veure el propietari del lloc web. A continuació hi ha la informació més rellevant de la consulta la qual podem veure el propietari del lloc web:

{
"registrar": "Dinahosting s.l.",
"whois_server": "whois.dinahosting.com",
"emails": "abuse-domains@dinahosting.com",
"dnssec": "unsigned",
"name": "Redacted by Privacy",
"org": "GRUPO RAMIREZ COGOLLOR S.L",
"address": "Redacted by Privacy",
"city": "Redacted by Privacy",
"state": "Madrid",
"registrant_postal_code": "Redacted by Privacy",
"country": "ES"
}

La resta de la informació es troba al executar el codi.

Pel que fa a projectes similars al nostre trobem les dades bastant semblants. En alguns datasets que hem vist a kaggle sobre llibres veiem que hi ha algun atribut més com pot ser el preu o el país de l'autor. També hem observat datasets que fan referència a un catàleg de pel·lícules d'una plataforma de contingut streaming. En aquest cas, hem vist atributs molt semblants els nostres i atributs que al parlar de llibres no hi són, com són el cas de la duració i els actors que participen. Per tant, veiem que és un dataset molt semblant a anàlisis anteriors. A continuació podem veure un exemple dels quals hem mirat per fer aquest anàlisis.

https://www.kaggle.com/datasets/sootersaalu/amazon-top-50-bestselling-books-2009-2019

Pel que fa als aspectes legals sabem que es troba en plena evolució i les lleis presenten certa complexitat. Nosaltres per actuar de forma correcta hem mirat les principals disposicions en les quals es basen els casos judicials i hem intentat no fer els mateixos passos. A continuació es mostren algunes de les pautes que hem anat seguint.

- No hem incomplert cap norma que ha imposat la pàgina web.
- No utilitzarem aquesta informació per finalitats comercials.
- No hem interferit en la propietat personal d'un individu i aquesta causa una pèrdua de valor.
- Hem mirat el fitxer robots.txt per veure les restriccions que cal tenir en compte quan es pretén rastrejar-les. Sabem que només és un suggeriment, però hem de tenir en compte que el principal objectiu és reduir les possibilitats de ser bloquejats.
- Hem vist que la millor pràctica és obtenir un permís per escrit però en aquest cas pràctic no era necesari.
- En les condicions d'ús no deien res sobre l'extracció automàtica de dades.

A més hem seguit les pràctiques per fer web scrapping fetes per en Vanaden Broucke I Baeysens:

- Hem verificat que existeixi una API (en aquesta pràctica tampoc la podíem utilitzar).
- Hem utilitzat llibreries com **BeautifulSoup** que facilita la tasca.
- En agafar les dades No hem causat d'anys, ja que hem fet un nombre petit de peticions i no hem sobrecarregat el servidor. Hem fet ús de la llibreria time.

- Hem modificat els **headers** quan fèiem una petició i sobretot hem modificat el **user agent**, que és la capçalera per prevenir web scrapping.
- Hem mirat si les **cookies** del navegador es trobaven buides.

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.

Aquest dataset podem extreure informació rellevant. Amb aquesta informació la gent pot escollir un llibre i hi ha molta més probabilitat que li agradi que no per atzar o una portada atractiva. Podrà llegir les crítiques de cada llibre i un resum per saber si és el que està buscant.

A la mateixa pàgina web segur que podríem ajudar a partir de les dades amb la finalitat de millorar el funcionament de la pàgina i augmentes el nombre de visites. Veuríem si una categoria hi ha poca interacció i es pot esborrar, si n'hi ha massa i potser es pot dividir en dues subcategories...

Si anem a un enfocament més empresarial, el dataset pot convertir-se en coneixement important per les llibreries, ja que sabran els llibres que agraden més i els podran tenir un stock molt més precís que a nivell econòmic significa un augment de les vendes. Pot ser interessant pels autors dels llibres, poden veure els llibres que han fet que tenen més èxit i poden extreure conclusions a partir de les crítiques. D'aquesta manera poden evitar coses que el públic lector no els acaba de convèncer.

També és interessant per les plataformes de contingut digital. Estem d'acord que quan més bo sigui el contingut de la plataforma, hi haurà més subscriptors. Per tant, les plataformes volen saber el que agrada els clients per incorporar-ho al seu catàleg.

En definitiva, veiem que hi ha diversos casos els quals els hi pot interessa molt el valor que nosaltres podem treure d'aquest dataset.

A part de poder respondre totes aquestes preguntes que hem anat esmentant per les empreses, el dataset també pot respondre preguntes més bàsiques com per exemple: Quins són els millors llibres?, Els més ben valorats?, Els més criticats?, Els millors autors?, Quines categories són les més aclamades?...

Pel que fa als anàlisis anteriors, encara que tinguem les mateixes dades sempre podem enfocar de diferent manera. En l'exemple del enllaç d'abans el qual el dataset s'anomena "Amazon Top 50 Bestselling Books 2009 – 2019" serà un dataset que l'estudi anirà relacionat en trobar els motius els quals són els més venuts, fer èmfasis en el preu per veure si és un atribut important, o realment és més determinant el contingut del llibre... En canvi, en el nostre dataset tot i sé bastant semblant podem enfocar altres tipus d'estudis com els que ja hem comentat.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció. Exemples de llicències que poden considerar-se:
    - 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.
    - Altres (especificar quina).

La llicència seleccionada és Released Under CC BY-NC-SA 4.0 License. Els motius els quals s'ha triat aquesta llicència són els següents:

• S'ha de proveir el nom del creador del dataset, indicant els canvis que s'han realitzat. D'aquesta manera queda estructurat la feina del creador i la feina del que vol utilitzar les dades. Així també es veurà l'aportació nova en relació amb el treball original i es podrà opinar si l'ha millorat o no.
• No es permet l'ús comercial. Volem que aquest dataset no es faci servir per fins comercials. Tot i que això reportaria cert reconeixement l'autor original.

• Els canvis duts a terme després del treball publicat estan sota aquesta. D'aquesta manera se segueix els mateixos passos que l'autor original.

9. **Codi.** Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
   - El codi haurà de situar-se a la carpeta **/source** del repositori.
   - S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python poden obtenir-se mitjançant la comanda
     `pip3 freeze > requirements.txt`
   - Al document PDF s'han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recol·lecció de dades, quines dificultats presenta el lloc web triat, i com les heu resolt.

A continuació es mostren les llibreries i versions utilitzades:

beautifulsoup4==4.11.1
certifi==2022.9.24
charset-normalizer==2.1.1
future==0.18.2
idna==3.4
numpy==1.23.4
pandas==1.5.1
python-dateutil==2.8.2
python-whois==0.8.0
pytz==2022.6
requests==2.28.1
six==1.16.0
soupsieve==2.3.2.post1
urllib3==1.26.12

Com el codi realitza la recol·lecció de dades és la següent. Un cop es troba a la pàgina fa un filtratge per els enllaços que acaben per mejores-generos, d'aquesta manera només navega per la secció que ens interessa. A partir d'aquí recorrem els enllaços de cada categoria que hi ha el lloc web i anem agafant la informació de cada llibre. Com que hi ha paginació, quan arribem al final d'una pàgina mirem si hi ha més llibres d'aquella categoria, si n'hi ha, anem a la següent pàgina, si no canviem de categoria.

La pàgina hi ha un sistema d'inici de sessió, però no ha sigut necessari utilitzar el request.Session per accedir a les dades, ja que teníem accés sense haver d'estar registrats.

La paginació ha sigut complicada, però ho hem resolt fent un find d'un tag el qual contingués l'atribut rel="next", el qual suposava que hi havia un botó per poder anar a la següent pàgina i que encara hi havia llibres.

El fet que no hi hagi poques classes el lloc web triat fa que dificulti utilitzar web scrapping. Ho hem solucionat fent servir més el find com per exemple en la següent línia.

book.find('div', class_="col-lg-8").find('small').find('a').text

En el lloc web hi havien molts enllaços I això ha complicat la navegació, ho hem resolt filtrant I agafant els enllaços que necessitavem per crear el dataset.

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/...). El dataset també haurà d'incloure's a la carpeta **/dataset** del repositori.

https://doi.org/10.5281/zenodo.7338488

11. Vídeo. Realitzar un breu vídeo explicatiu de la pràctica **(màxim 10 minuts)**, que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo s'haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure les dades. Indicar l'enllaç del vídeo (https://drive.google.com/...), que haurà d'estar al Google Drive de la UOC.

https://drive.google.com/file/d/1F_uHBIEMD0mgrd6f54H5we3vu79hRoGV/view?usp=sharing