
# Práctica 1 - Tipología y Ciclo de Vida de los Datos

| | |
|---|---|
| Autors | Pol López Vidaller<br>Gemma Vendrell Tarrés |
| Web triada | https://www.abacus.coop/es/libros/libros-recomendados/libros-mas-leidos |
| Repositori | https://github.com/PLVgit/abacus-python-web-scraper |
| Enllaç DOI Zenodo | https://zenodo.org/records/10981789 |
| Vídeo de la presentació | https://drive.google.com/file/d/19wYHsOK5VwVx9JprbhGOmC06jnho2DK7/view?usp=drive_link |

## 1. Context

Per la realització d'aquest treball hem utilitzat la pàgina web d'Abacus, empresa catalana especialitzada en la distribució de material educatiu, editorial i d'oficina.Només entrar a la seva pàgina web trobem un apartat on ens indica un seguit de llibres recomanats.

En els últims anys l'habit de comprar per internet a augmentat considerablement i evidentment també ha passat amb els llibres, cada vegada més persones cerquen per internet recomanacions i opinions sobre els llibres que volen llegir. Actualment gràcies a internet és força senzill escollir un llibre que compleixi les expectatives ja que amb les opinions de molts dels usuaris i la seva puntuació es pot trobar el llibre adient a cada ocasió.

Per aquest motiu i aprofitant que arriba Sant Jordi ens va sembla interessant recollir informació sobre els títols, autors, preu, any d'aquests llibres. Al ser la web oficial d'Abacus podem estar segurs que la informació és fiable.

Pàgina web utilizada:
https://www.abacus.coop/es/libros/libros-recomendados/libros-mas-leidos

## 2. Títol

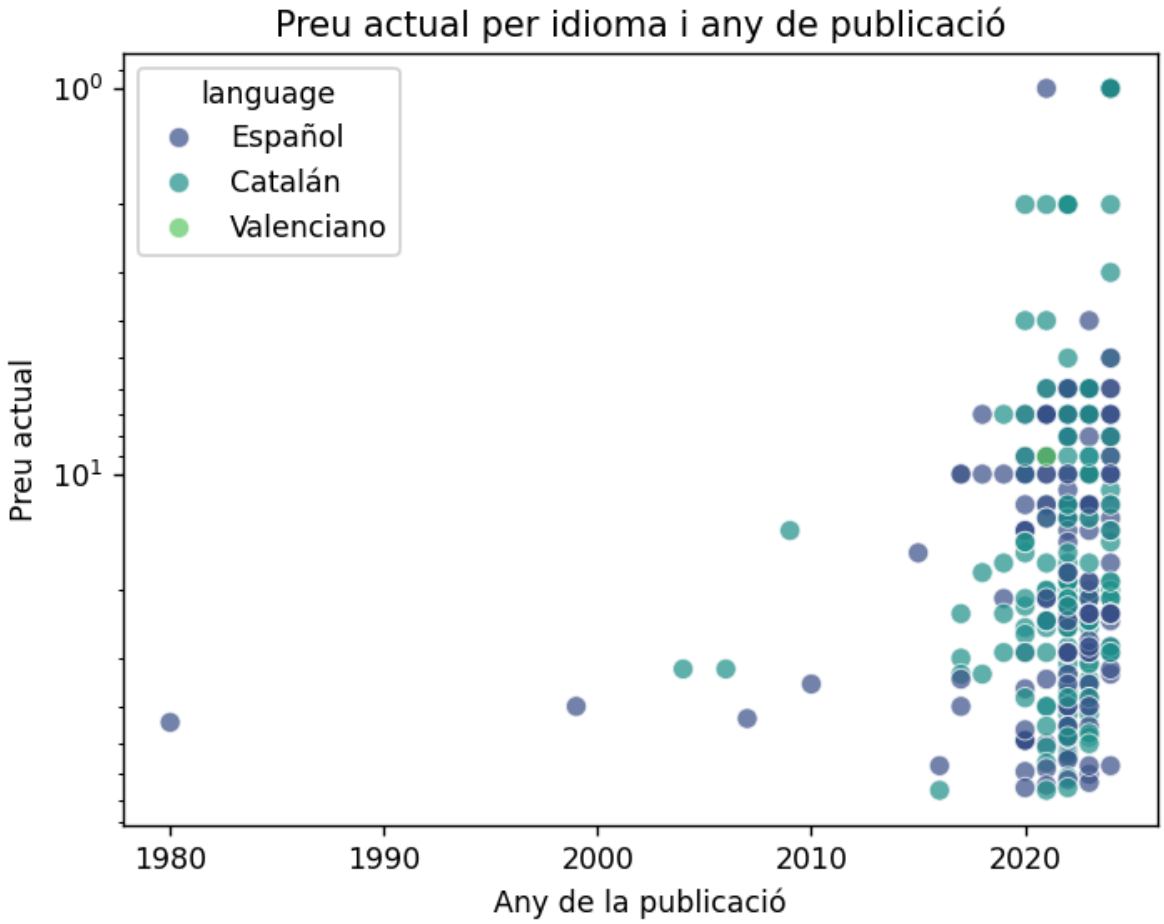
El títol triat per al dataset ha sigut *"abacus_llibres_recomanats.csv"*, degut a que l'objectiu principal ha sigut el scraping de la web de llibres recomanats, tot i que es poguès fer servir per qualsevol apartat de la web d'abacus.

## 3. Descripció

El conjunt de dades recopilades per aquest treball conté informació rellevant sobre una selecció de llibres recomanats a la botiga Abacus, que representen els més llegits i populars entre la seva clientela.A part del títol i l'autor podem trobar l'any de publicació de cada llibre i el idioma en que està escrit. Per acabar-ho de complementar, el dataset també inclou el link que ens proporcioa lla imatge de la portada de cada llibre, d'aquesta manera es pot reconèixer el llibre més fàcilment.

És important destacar que, encara que el codi s'hagi desenvolupat per extreure les dades dels llibres també el podríem utilitzar en altres seccions del lloc web, com per exemple que ens generhi un dataset amb els jocs infantils que hi ha disponibles de manera online.

## 4. Representació gràfica




El gràfic proporciona una representació visual de la relació entre la popularitat dels llibres i el seu any de publicació, així com el preu associat. Observem una tendència que indica que la majoria dels llibres més populars són més recents, amb una concentració significativa de llibres publicats a partir del 2015 i fins al 2024. Aquesta concentració s'intensifica a mesura que ens acostem al 2024, suggerint una tendència de preferència per les obres més noves.

Un aspecte interessant a destacar és que el preu no sembla ser un indicador directe de la popularitat dels llibres. Tot i que alguns dels llibres més antics tendeixen a tenir un preu més baix, això no és necessàriament correlacionat amb la seva popularitat. Podria ser intuïtiu inferir que els llibres més antics són menys populars, però cal considerar altres factors. Per exemple, és possible que molts d'aquests llibres antics amb preus més baixos siguin considerats clàssics o tinguin una demanda persistent al llarg del temps, la qual cosa els manté rellevants i populars malgrat el seu any de publicació.

En resum, el gràfic ens ofereix una perspectiva sobre la dinàmica temporal de la popularitat dels llibres, indicant una preferència pels llibres més nous. Tanmateix, ens recorda la importància de considerar diversos factors, com ara la naturalesa del contingut i la demanda històrica, abans de fer conclusions sobre la popularitat dels llibres basant-nos únicament en el seu preu o l'any de publicació.

## 5. Contingut

El dataset consta de deu variables que fan referència als diferents llibres.
Les variables que tenim són les següents:

- Title: Títol del llibre
- Author: Autor del llibre
- Description: Descripció
- Actual_price: Preu actual
- Prev_price: Preu abans del descompte
- Pages: Quantitat de pàgines
- Year: Any de publicació
- Language: Idioma
- Publisher: Editorial
- Image: Link de la imatge de la portada

Tenim tres variables quantitatives, que corresponden al preu actual i antic del llibre i el número de pàgines, la resta de variables son tipus string.

## 6. Propietari

En el panorama actual, la compra de llibres en línia ha experimentat un creixement significatiu, com ho demostren diversos estudis de recerca (Curcic, 2023; Grand View Research, 2020). Aquesta tendència es veu accentuada per dates assenyalades com Sant Jordi, on la demanda de llibres s'intensifica considerablement.

En aquest context, el projecte de web scraping de la secció de llibres d'Abacus es presenta com una iniciativa oportuna i rellevant per diversos motius.

El projecte de web scraping se centra en la secció de llibres de la pàgina web d'Abacus, amb l'objectiu d'obtenir dades sobre els llibres disponibles, com les mencionades anteriorment. Aquestes dades s'extrauen de manera automatitzada mitjançant tècniques de web scraping, respectant les directrius establertes al fitxer robots.txt d'Abacus. El fitxer robots.txt indica que no hi ha restriccions per a l'accés a les pàgines públiques de la secció de llibres, excepte les àrees relacionades amb dades personals dels usuaris.

Per tal de realitzar el nostre projecte de web scraping de manera responsable i ètica, hem seguit 4 principis fonamentals:

- **Respecte a la privacitat:** S'ha evitat l'extracció de dades personals dels usuaris o informació confidencial com senyalava el robots.txt de la web.

- **Transparència i traçabilitat:** La metodologia de web scraping, l'anàlisi de dades i els resultats obtinguts es documenten de manera clara i detallada, permetent la revisió i el seguiment del procés per part d'altres investigadors o persones interessades.

- **Ús responsable de les dades:** Les dades obtingudes s'utilitzen exclusivament per a finalitats de recerca i anàlisi, sense cap objectiu comercial.

- **Respecte a la propietat intel·lectual:** S'ha verificat que el contingut de la pàgina web d'Abacus no està subjecte a drets d'autor exclusius o restriccions d'ús que impedeixin el web scraping. En aquest sentit, s’han scrapejat només variables que són públiques.

## 7. Inspiració

Primerament, amb aquest projecte intentem omplir un buit a la recerca. Si bé existeixen projectes de recerca sobre la compra online de llibre, no hi ha estudis equivalents centrats en llibreries cooperatives online com Abacus. Aquest projecte aporta una perspectiva única i complementària a la comprensió d'aquest sector.

L'objectiu principal es aportar iinformació rellevant pels usuaris locals. Els resultats del web scraping  poden proporcionar a lectors catalans dades rellevants sobre la seva oferta de llibres, els gustos dels lectors locals i les tendències del mercat. Aquesta informació pot ser útil o interessant tant per clients com per llibreries per a la presa de decisions estratègiques i la millora dels seus serveis.

Una manera d'aconseguir-ho és millorant la transparència i l'accés a la informació. El projecte es basa en l'ús de dades públiques disponibles a la web d'Abacus, respectant els principis d'ètica i legalitat del web scraping. Això garanteix la transparència del procés i facilita la verificació de la informació per part de tots els actors interessats.

Els resultats del projecte de web scraping poden tenir diverses aplicacions beneficioses.

La informació obtinguda pot ser utilitzada per a analitzar la varietat de llibres disponibles, les categories més populars, els autors més destacats i les tendències de preus. El projecte pot aportar informació sobre els llibres més comprats, les ressenyes dels clients i els comportaments de compra online.

L'automatització de l'accès a aquesta informació pot servir de benchmarking amb altres llibreries. Les dades obtingudes poden ser comparades amb les d'altres llibreries cooperatives o online per a identificar diferències i similituds en l'oferta i la demanda.

Tot això comporta la millora d'experiència dels usuaris. Abacus pot utilitzar els resultats del projecte per a personalitzar les recomanacions de llibres, oferir promocions més ajustades als gustos dels clients i optimitzar la seva pàgina web per a facilitar la navegació i la compra.

## 8. Llicència

Per seleccionar un llicència adecuada pel dataset s'han de tenir diferents aspectes en compte, com per exemple la naturalesa de les dades, les restriccions que s'han d'aplicar i el proposit per el que s'ha fet el dataset.

La Released Under CC0: Public Domain License s'utilitzaria quan el titular renuncia a tots els drets de l'obra, es posa el dataset en domini públic i per tant qualsevol persona el pot utilitzar, modificar i utilitzar fins i tot per mitjans comercials. Un altre exemple de llicència és Released Under CC BY-SA 4.0 License que permet adaptar i construir part de l'obra fins i tot per usos comercials sempre que es doni crèdit a l'autor original i apliqui la mateixa llicència a les noves creacions. Seria una bona llicència per el nostre dataset, tot i així nosaltres aplicariem la llicència Released Under CC BY-NC-SA 4.0 ja que permet als usuaris copiar i redistribuir el material en qualsevol format sempre i quan es doni crèdit a l'autor original, no es pot utilitzar el material per a finalitats comercials i sempre que es modificquin o es crei qualsevol obra a partir d'aquest material s'ha de difondre amb la mateixa llicència, de la mateixa manera, no es poden aplicar termes legals que restringeixi legalment a altres fer el que la llicència permet.

Aquesta llicència és una bona opció perquè així es té un control de l'ús comercial i garanteix que la resta d'usuaris comparteixin la informació amb les mateixes restriccions.

## 9. Codi

Aquest script Python implementa un scraper web orientat a objectes per extreure informació sobre llibres de la secció "Llibres més llegits" del lloc web d'Abacus (https://www.abacus.coop/es/libros/libros-recomendados/libros-mas-leidos).

**Classe AbacusScraper:**

La classe AbacusScraper encapsula la lògica del scraping web. Les seves propietats i mètodes permeten gestionar el procés de manera modular i reutilitzable.

**9.1. def __init__ inicialització (constructor):**

Aquesta funció constructora inicialitza les variables d'instància necessàries per al scraping:

- **base_url:** URL base de la pàgina de llibres més llegits.
- **page_urls:** Llista per emmagatzemar les URLs de totes les pàgines paginades.
- **book_urls:** Llista per emmagatzemar les URLs individuals de cada llibre.
- **book_data:** Llista per guardar els diccionaris amb la informació extreta de cada llibre.
- **df:** Variable per guardar el DataFrame de Pandas que contindrà les dades dels llibres.

**9.2. get_next_page_urls():**

Aquesta funció recupera les URLs de totes les pàgines paginades de la secció de llibres més llegits. Realitza una petició GET a la URL base i, a continuació, utilitza BeautifulSoup per analitzar el contingut HTML de la resposta. Busca la secció del paginador i extreu les URLs de les pàgines restants mitjançant els atributs data-page dels enllaços de paginació.

Per assegurar el funcionament d'aquesta funció, s'ha extret la lògica darrere dels diferents enllaços de les pàgines i s'ha construït l'enllaç complert utilitzant una concatenació.

**9.3. get_book_urls():**

Aquesta funció recorre cada pàgina obtinguda prèviament i recupera les URLs individuals de cada llibre. Analitza el contingut HTML de cada pàgina i identifica els enllaços que condueixen

a la pàgina de detall del llibre. Extrau l'adreça relativa de l'enllaç i construeix la URL completa afegint la base de la URL.

**9.4. scrape_book_details():**

Aquesta funció és la responsable d'extreure la informació específica de cada llibre. Per a cada URL de llibre a la llista book_urls, realitza una petició GET i analitza el contingut HTML resultant. A continuació, busca elements específics dins de l'estructura HTML de la pàgina de detall per extreure informació de les variables explicades anteriorment.

L'ús de dues línies per extreure el text d'un element HTML (element.text.strip() if element else None) evita errors de tipus ValueError en cas que l'element buscat no existeixi a la pàgina. La primera línia comprova si l'element està present i, en cas afirmatiu, n'extreu el text i l'elimina dels espais inicials i finals. La segona línia assigna None si l'element no es troba, evitant així l'error.

**9.5. create_dataframe():**

Aquesta funció converteix la llista de diccionaris book_data en un DataFrame de Pandas. El DataFrame permet una estructuració i manipulació eficients de les dades extretes. L'objectiu d'aquesta funció és tan la manipulació de les dades amb Python en cas de que sigui l'objectiu de l'usuari, però també la extracció en forma de excel.

**9.6. export_to_csv():**

Aquesta funció permet a l'usuari exportar en forma de .csv la informació scrapejada. Aquest format és molt utilitzat per repositoris de dades i pot ser útil depenent de l'usuari.

## 10. Dataset

https://zenodo.org/records/10981789

## 11. Vídeo

https://drive.google.com/file/d/19wYHsOK5VwVx9JprbhGOmC06jnho2DK7/view?usp=drive_link

## 12. Referències

*Grand View Research. (2020). Online Book Services Market Size, Share & Trends Analysis Report By Product (Trade, Education), By Region, And Segment Forecasts, 2020 - 2027. Recuperat de* https://www.grandviewresearch.com/industry-analysis/online-book-services-market

*Curcic. (2023). Book Sales Statistics - An Updated Overview of the Market. Recuperat de* https://wordsrated.com/book-sales-statistics/

| Contribucions | Signatura |
|---|---|
| Investigació prèvia | PLV,GVT |
| Redacció de les respostes | PLV,GVT |
| Desenvolupament del codi | PLV,GVT |
| Participació al vídeo | PLV,GVT |