Pràctica 1: Web scraping

1. Context:

El joc de dades que presentem s'ha recol·lectat en el marc de la Pràctica 1 de l'assignatuara **Tipologia i Cicle de Vida de les Dades** del **Màster Universitari de Ciència de Dades** de la **Universitat Oberta de Catalunya (UOC)** durant el primer semestre del Curs 20/21.

L'objectiu de la pràctica és la construcció d'un joc de dades a partir de la informació present en una pàgina web a través de *web scraping*.

Per a desenvolupar el projecte s'ha escollit la pàgina web *goodreads* com a pàgina objectiu. *Goodreads* és una web social de catalogació, referència i recomanació de llibres mantinguda per la comunitat d'usuaris i amb més de 10.000.000 de llibres [1]. La quantitat i diversitat d'informació present a la seva extensa base de dades fa que sigui un bon candidat per al desenvolupament de la pràctica.

L'objectiu del projecte és, doncs, la creació d'una base de dades de llibres amb informació sobre els llibres (títol, autor, pàgines, editoral, isbn, etc.), sobre el seu contingut (gènere, protagonistes, localització, etc.), i informació provinent de la interacció social dels usuaris (puntuació, número de vots, percentatge de *likes*, etc.). A més, s'aprofitarà per recollir informació de preus dels llibres des del portal IberLibro, un mercat online de llibres propietat de AbeBooks el qual, al mateix temps, és una filial d'Amazon i pioner en el mercat online de venta de llibres [2]. S'utilitzarà la informació del ISBN extreta anteriorment de la pàgina de goodreads per buscar el llibre en el portal IberLibro i recuperar el preu d'aquest.

Per tal de concretar el desenvolupament del projecte s'ha decidit fer *web scraping* XX.XXX primers llibres de la llista Best Books Ever que contè, segons els indicadors de *goodreads* els millors llibres de la història.

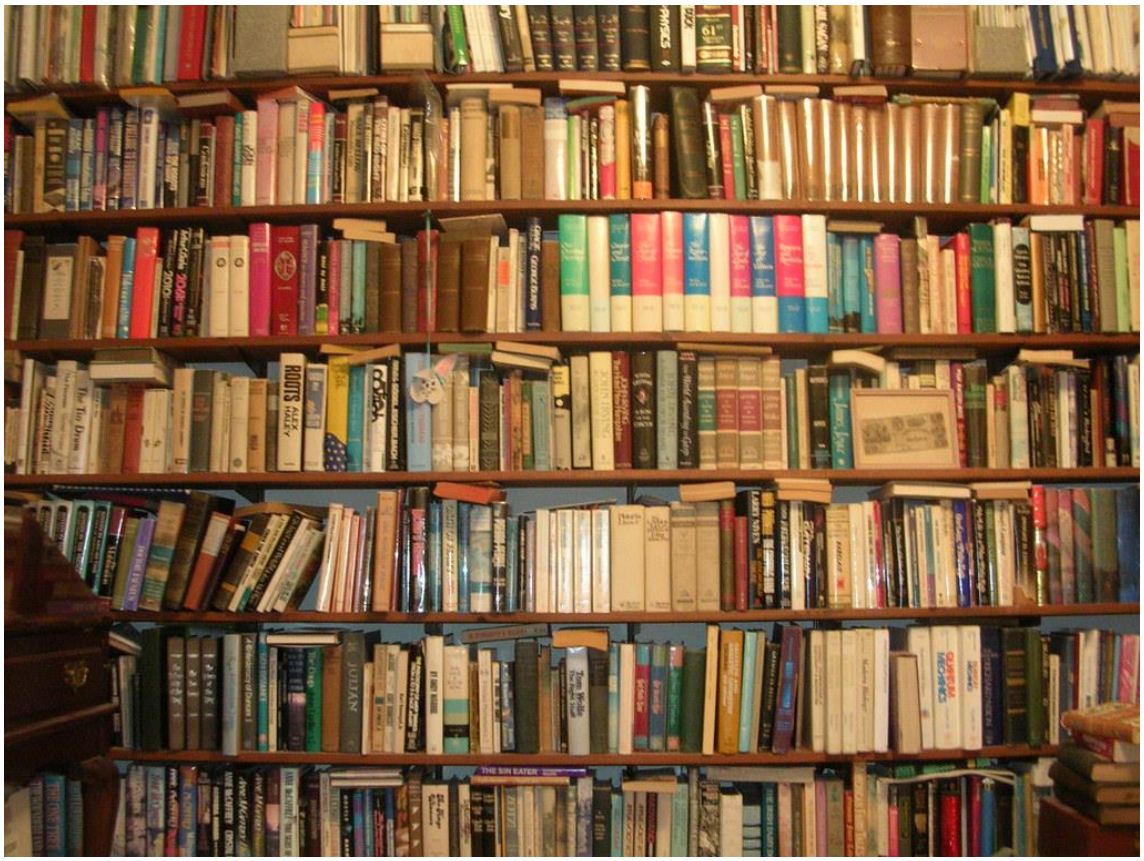
2. Definir un títol pel joc de dades

PROPOSTA: Best Books Ever dataset

3. Descripció del joc de dades

El joc de dades conté informació sobre els XX.XXX primers llibres de la llista *Best Books Ever*.

4. Representació gràfica



[3] sota llicència CC BY-SA 2.0

5. Contingut

El joc de dades conté 22 atributs sobre xx.xxx instàncies de llibres descarregades el XX de novembre de 2020 mitjançant un *web scraper* basar en *Selenium* i desenvolupat en *python*.

Els camps contingut al dataset són els següents:

- **bookId:** Identificador del llibre a goodreads.com
- **title:** Títol del llibre
- **series:** Nom de la saga a la quan pertany el llibre
- **author:** Autor/s del llibre
- **description:** Descripció del llibre (contraportada)
- **edition:** Tipus d'edició (p.ex: *Anniversary Edition*)
- **pages:** Número de pàgines
- **publisher:** Editorial

- **publishDate:** Data de publicació
- **firstPublishDate:** Data de primera publicació
- **language:** Idioma
- **bookFormat:** Tipus d'encuadernació
- **ISBN:** Referència ISBN de 13 xifres
- **awards:** Llistat de premis
- **genres:** Gèneres als quals pertany el llibre, per ordre de votació
- **characters:** Nom dels protagonistes principals
- **setting:** Localització de la història
- **rating:** Puntuació global del llibre a goodreads.com
- **numRatings:** Nombre total de valoracions a goodreads.com
- **ratingsByStars:** Nombre de valoracions desglossat per puntuació (5-1 estrelles).
- **likedPercent:** Percentatge de lectors als quals ha agradat el llibre.
- **bbeScore:** Puntuació del llibre a la llista *Best Books Ever*
- **bbeVotes:** Nombre de vots a la llista *Best Books Ever*
- **coverImg:** Imatge de portada en format png.
- **price:** Preu del llibre (WIP)

6. Agraïments:

Les dades han estat recollides de la pàgina web https://www.goodreads.com/

7. Inspiració:

El principal interès del joc de dades creat en aquesta pràctica és el de disposar de la informació sobre llibres present a la base de dades de goodreads.com en un format coherent i estructurat que permeti l'anàlisi d'aquesta i la extracció efectiva de coneixement.

Les varietat d'atributs recollits permeten de fer multitud d'anàlisi, entre les quals podríem destacar:

- Identificació dels atributs que tenen més influència en la valoració dels llibres, en el nombre de vots, l'score o el preu.

- Anàlisi d'agrupament i clústering: Analitzar si els atributs recollits permeten d'agrupar els llibres en diferents grups de característiques similars i identificar i definir les característiques de cada grup.
- Anàlisis segregats per anys de publicació, per localització de la història, etc.
- Autors més prolífics o amb llibres millor valorats.
- Regressions sobre variables com *score* o valoració dels llibres.
- En última instància el joc de dades també permetria de desenvolupar un sistema de recomanació de llibres a partir d'una entrada de llibres llegits.

8. Llicència:

TODO

9. Codi

GoodReads_scraper.ipynb : Notebook amb el codi necessari per extraure la informació de cada llibre i crear gran part del dataset.

get_princeFromISBN.ipynb : Notebook amb el codi de la funció per extraure el preu de cada llibre a partir de l'ISBN al web IberLibro.

GoodReads_linkScraper.ipynb : Notebook amb el codi necessari per extraure les URLs de cada llibre present a la llista objectiu *Best Books Ever*

download_images.ipynb : Notebook amb el codi necessari per descarregar les imatges de portada de cada llibre a partir de l'adreça emmagatzemana en dels camps del dataset.

TODO Organitzar codi, homogeneitzar estils, etc.

[4]–[9]

10. Dataset

list_data_all_BBE.txt : Llistat d'enllaços, scores i votacions de la llista *Best Books Ever* -> output de ***GoodReads_linkScraper.ipynb***

book_db_10000.txt : Database preliminar amb 10000 llibres -> output de ***GoodReads_scraper.ipynb***

11. Bibliografia i Webgrafia

[1] "Goodreads - Viquipèdia, l'enciclopèdia lliure." [Online]. Available: https://ca.wikipedia.org/wiki/Goodreads. [Accessed: 30-Oct-2020].

[2] "IberLibro.com: Información de la Empresa." [Online]. Available: https://www.iberlibro.com/docs/CompanyInformation/. [Accessed: 30-Oct-2020].

[3] “Wall of Books | wall of books in a friend’s house in Michiga… | benuski | Flickr.” [Online]. Available: https://www.flickr.com/photos/36986477@N05/3502143020. [Accessed: 30-Oct-2020].

[4] H. Brody, “The Ultimate Guide To Web Scraping,” 2013.

[5] B. S. Mózo, *Python Automation Cookbook*, vol. 53, no. 9. 2017.

[6] A. . Fallis, *Data Visualization with Python and Javascript*, vol. 53, no. 9. 2013.

[7] M. Heydt, *Python Web Scrapping Cookbook: Over 90 Proven Recipes to Get Your Scraping with Python, Microservices, Docker and AWS.* 2018.

[8] S. vanden Broucke and B. Baesens, *Practical Web Scraping for Data Science: Best Practices and Examples with Python*. 2018.

[9] V. Nair, *Getting Started with Beautiful Soup*. 2014.