laSalle ENG
Universitat Ramon Llull

Máster en Big Data

Asignatura: Casos de analítica
Fecha: 30/03/2023

Sebastian Cuevas, Pol Gràcia Login _____ Curso: **2022/23**

Caso 1: Book-Crossing analysis

Este segundo caso gira entorno al mundo de la lectura. Específicamente, se quiere hacer un análisis de un dataset de libros para, en última instancia, crear un modelo que recomiende nuevas lecturas. Para llevarlo a cabo, partiremos del conjunto de datos *Book-Crossing dataset*, que puede descargarse del siguiente enlace: http://www2.informatik.uni-freiburg.de/~cziegler/BX/.

Con estos datos, se os propone que apliquéis técnicas de estadística, analítica, minería de datos y visualización para responder a las siguientes preguntas. No hay restricciones acerca de las técnicas ni tecnologías a utilizar siempre y cuando los resultados sean reproducibles y estén debidamente justificados. Explicitad y detallad todos los pasos hechos para responder a cada pregunta y las conclusiones que podáis derivar de ellas.

Para responder todas las preguntas se creo un repositorio que contendrá todo el codigo en el siguiente enlace: https://github.com/sebasfire3/Book-Crossing-

PRIMERA PARTE: ANÁLISIS CUANTITATIVO.

1.1 Primer examen preliminar del *dataset*. ¿En qué formato está el dataset? ¿Cómo podemos leerlo correctamente? ¿Qué campos hay en cada fichero del *dataset*? ¿Cuál es su significado? ¿Existen valores aparentemente incorrectos?

El dataset BX-CSV-Dump.zip contiene tres archivos CSV:

- BX-Books.csv - Este archivo contiene información sobre los libros, como su ISBN, título, autor, fecha de publicación, editor y número de páginas.
- BX-Users.csv - Este archivo contiene información sobre los usuarios del sistema, como su ID de usuario, nombre, edad, género y ubicación.
- BX-Book-Ratings.csv - Este archivo contiene información sobre las calificaciones de los libros dadas por los usuarios, como el ISBN del libro, el ID de usuario que lo calificó y la puntuación que le dio al libro.

Para leer estos archivos CSV, se puede utilizar cualquier lenguaje de programación que soporte la lectura de archivos CSV, como Python, R, Java, etc. En Python, por ejemplo, se puede utilizar la biblioteca Pandas para leer estos archivos CSV.

Los campos en cada archivo tienen los siguientes significados:

- En BX-Books.csv:
  - ISBN: el número de identificación único del libro.
  - Title: el título del libro.
  - Author: el autor del libro.
  - Year of Publication: el año de publicación del libro.
  - Publisher: el editor del libro.
  - Image-URL-S: la URL de la imagen pequeña del libro.
  - Image-URL-M: la URL de la imagen mediana del libro.

    - Image-URL-L: la URL de la imagen grande del libro.
- En BX-Users.csv:
    - User-ID: el número de identificación único del usuario.
    - Location: la ubicación del usuario.
    - Age: la edad del usuario. Puede estar en blanco.
- En BX-Book-Ratings.csv:
    - User-ID: el número de identificación único del usuario que calificó el libro.
    - ISBN: el número de identificación único del libro que fue calificado.
    - Book-Rating: la calificación del libro dada por el usuario. El rango de calificación es de 1 a 10. Un valor de 0 significa que el usuario no calificó el libro.

En cuanto a los valores aparentemente incorrectos, es posible que haya algunos valores nulos o en blanco en los campos de edad de los usuarios o de año de publicación de los libros. Además, algunos ISBN o ID de usuario pueden estar duplicados o no ser válidos. Por lo tanto, es importante limpiar y validar los datos antes de utilizarlos para cualquier análisis o modelado.

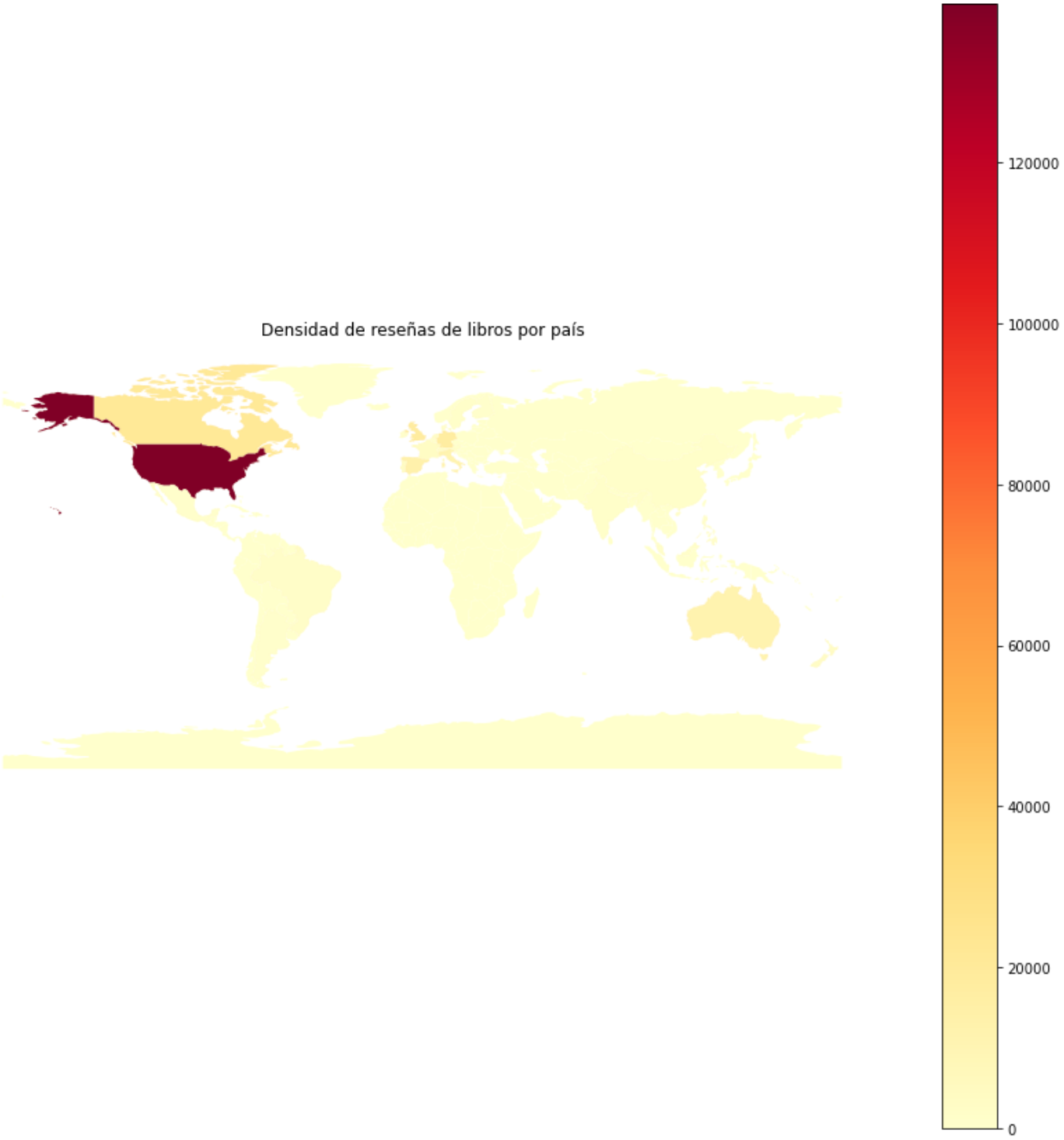
Sí, existen algunos valores aparentemente incorrectos o incompletos en el dataset. Por ejemplo:

- En el archivo BX-Books.csv, algunos libros tienen el valor "NULL" o "0" en el campo "Year of Publication", lo cual no es un valor válido para este campo. Además, algunos libros tienen información incompleta o incorrecta en los campos "Author" o "Publisher".
- En el archivo BX-Users.csv, algunos usuarios tienen el valor "NULL" o en blanco en el campo "Age", lo cual puede ser un indicador de que la información de edad de estos usuarios no se proporcionó o no se registró correctamente.
- En el archivo BX-Book-Ratings.csv, algunos registros tienen el valor "0" en el campo "Book-Rating", lo cual puede significar que el usuario no calificó el libro o que la calificación se perdió o no se registró correctamente.

Es importante tener en cuenta estos valores aparentemente incorrectos o incompletos y tomar medidas para limpiar y validar los datos antes de utilizarlos para cualquier análisis o modelado.

1.2 Empezamos por visualizar el origen geográfico desde donde se han hecho las contribuciones que componen *dataset*. Haced un *plot* geográfico que muestre desde dónde se han hecho las reseñas de los libros y, de alguna manera, habilite ver la densidad de reseñas por país.

Para visualizar el origen geográfico desde donde se han hecho las contribuciones que componen el dataset, podemos utilizar el archivo "BX-Users.csv" que contiene información sobre los usuarios, incluyendo su ubicación. Para mostrar la densidad de reseñas por país, podemos utilizar un mapa de calor (heatmap) que muestre el número de usuarios y las reseñas por país.



Podemos apreciar que la mayoría de los libros se encuentran en Europa y Norte America, con algunas menciones en Ocenaia. Asia, Africa y Sudamérica casi no tienen.

SEGUNDA PARTE: ANÁLISIS CUALITATIVO.

2.1 ¿Cuál fue el año en el que se publicaron más libros? Muéstralo en un gráfico ¿Y el autor más plorífico? ¿Cuántos libros suyos hay en el dataset?

```
Año con más libros publicados: 2002
Autor más prolífico: Agatha Christie
Cantidad de libros del autor más prolífico: 632
```

En resumen del código cargaría los datos del archivo "BX-Books-clean.csv" en un DataFrame de pandas, y luego utilizaría algunas de las funciones de esta librería para responder a las preguntas planteadas. El primer gráfico mostraría la cantidad de libros publicados por año, mientras que las líneas finales del código imprimirían el año con más libros publicados, el autor más prolífico y la cantidad de libros escritos por dicho autor en el dataset.

2.2 ¿Cuáles son los orígenes geográficos y la edad de los reseñadores más jovenes?

```
#Unimos usuarios que hacen rates

raters = pd.merge(ratings, users, on = 'User-ID', how = 'inner')

# Encontrar la edad del reseñador más joven
raters = raters[(~raters['Age'].isna()) & (raters['Age']>0)]
youngest_age = raters['Age'].min()

# Encontrar los orígenes geográficos de los reseñadores más jóvenes
youngest_reviewers = raters[raters['Age'] == youngest_age]['country'].unique()

print(f"Los orígenes geográficos de los reseñadores más jóvenes son: {',
'.join(youngest_reviewers)}")
print(f"La edad del reseñador más joven es {youngest_age}")
```

Se deben unir los datasets de users y users ratings para obtener un dataset combinado y poder responder la pregunta.

Los orígenes geográficos de los reseñadores más jóvenes son: usa, greece, india, canada, spain, , italy, germany, united kingdom, japan, new zealand, australia, costa rica, portugal, switzerland

La edad del reseñador más joven es 1.0

2.3 Busca los mejores libros del año 2000 según Goodreads (https://www.goodreads.com/) utilizando técnicas de *web scrapping* ¿Cuáles de los autores que aparecen en la lista están también en el dataset? ¿Cuál fue el género más popular?

En el codigo se puede apreciar como se realizo la creación de libros.csv con la información de la pagina Goodreads de todos los libros del 2000, en resumen primeramente se creo un dataset con el autor y el libro por ejemplo:

	Título	Autor
0	Harry Potter and the Goblet of Fire (Harry Pot...	J.K. Rowling
1	Angels & Demons (Robert Langdon, #1)	Dan Brown
2	A Storm of Swords (A Song of Ice and Fire, #3)	George R.R. Martin
3	The Amber Spyglass (His Dark Materials, #3)	Philip Pullman
4	The Amazing Adventures of Kavalier & Clay	Michael Chabon

Despues se agrego la columna genero y con la librería selenium el RPA se ingresa a cada libro obtiene todos los géneros respectivos y lo ingresa en la columna del dataset para obtener como resultado lo siguiente:

0	Harry Potter and the Goblet of Fire (Harry Pot...	J.K. Rowling	['Fantasy', 'Young Adult', 'Fiction', 'Magic',...
1	Angels & Demons (Robert Langdon, #1)	Dan Brown	['Fiction', 'Mystery', 'Thriller', 'Mystery Th...
2	A Storm of Swords (A Song of Ice and Fire, #3)	George R.R. Martin	['Fantasy', 'Fiction', 'Epic Fantasy', 'Scienc...
3	The Amber Spyglass (His Dark Materials, #3)	Philip Pullman	['Fantasy', 'Young Adult', 'Fiction', 'Childre...
4	The Amazing Adventures of Kavalier & Clay	Michael Chabon	['Fiction', 'Historical Fiction', 'Novels', 'L...
5	Me Talk Pretty One Day	David Sedaris	['Nonfiction', 'Humor', 'Memoir', 'Essays', 'S...
6	On Writing: A Memoir of the Craft	Stephen King	['Nonfiction', 'Writing', 'Memoir', 'Biography...
7	Persepolis: The Story of a Childhood (Persepol...	Marjane Satrapi	['Graphic Novels', 'Nonfiction', 'Memoir', 'Co...
8	Kitchen Confidential: Adventures in the Culina...	Anthony Bourdain	['Nonfiction', 'Food', 'Memoir', 'Biography', ...

Con este nuevo dataset podremos responder las siguientes preguntas.

- ¿Cuáles de los autores que aparecen en la lista están también en el dataset?

La lista de autores que aparecen en ambos dataset:

Se encuentra en el codigo debido a su gran numero de autores.

```
['tanya huff', 'susan mallery', 'marc levy', 'diana wynne jones', 'tim powers', 'david liss', 'julie garwood', 'philip roth', 'todd strasser', 'nina bouraoui',
```

- ¿Cuál fue el género más popular?

```
El género más popular es: Fiction con 233 libros.
```

3.1 Elige tres autores del dataset y calcula la probabilidad de que una nueva obra suya guste a los lectores.

Para calcular la probabilidad de que una obra nueva de un autor guste a los lectores, debemos basarnos en sus libros anteriores.

En este caso, consideraremos que un libro gusta a los lectores si la puntuación es igual o superior a 7.

Seleccionamos 3 autores: Vonda n. Mcintyre, Simon Mawer y Stuart Cohen.

La fórmula consiste en dividir el total puntuaciones con 7 o más y dividirlo entre el total de puntuaciones, así sabríamos cuál es la probabilidad de que los lectores puntuen con un 7 o más a una obra nueva de cada autor respectivamente.

```
Probabilidad de que una obra de vonda n. mcintyre guste a los lectores: 0.6571428571428571
Probabilidad de que una obra de simon mawer guste a los lectores: 0.6363636363636364
Probabilidad de que una obra de stuart cohen guste a los lectores: 0.6785185185185185
```

3.2 Diseña un modelo que, a partir de un libro de entrada, te recomiende una nueva lectura. Puedes utilizar o bien el dataset proporcionado o bien un dataset creado por ti mismo (por ejemplo, utilizando técnicas de *web scrapping* 😉) y con más características (o una combinación de ambos). Respecto a este sistema, a modo de ejemplo, explica las recomendaciones que proporcionaría el modelo si entráramos los siguientes libros:

- *A Court of Thornes and Roses de Sarah J. Maas*
- *Hamlet de William Shakespeare*
- *Don Quijote de la Mancha de Miguel de Cervantes*

Para diseñar un modelo de recomendación de libros se puede utilizar un algoritmo de filtrado colaborativo basado en vecinos más cercanos (KNN). Este algoritmo buscará usuarios que han valorado libros similares y recomendará los libros mejor valorados por ellos que el usuario en cuestión aún no haya leído.

Para aplicar este algoritmo, se utilizarían las valoraciones que los usuarios han dado a los diferentes libros del dataset.

Es importante tener en cuenta que este modelo de recomendación tendría limitaciones y no sería perfecto. Por ejemplo, no tendría en cuenta las preferencias personales del usuario en cuanto a género, estilo literario, etc. Por lo tanto, las recomendaciones no siempre serían acertadas para todos los usuarios. Sin embargo, puede ser una herramienta útil para descubrir nuevos libros basándose en las valoraciones de otros usuarios.

Más concretamente, se va a usar un algoritmo de filtrado colaborativo.

El filtrado colaborativo se basa en la idea de que a las personas similares (según los datos) suelen gustarles cosas similares. Predice qué artículo le gustará a un usuario basándose en las preferencias de artículos de otros usuarios similares.

El filtrado colaborativo utiliza una matriz usuario-artículo para generar recomendaciones. Esta matriz contiene los valores que indican la preferencia de un usuario por un artículo determinado. Estos valores pueden representar opiniones explícitas (valoraciones directas de los usuarios) o implícitas (comportamientos indirectos de los usuarios, como escuchar, comprar, ver...).

Para hacer esta matriz, se hace una matriz sparse que relaciona usuarios, los libros y ratings con la librería de scipy. De esta manera, el modelo de KNN puede entender y relacionar el id del usuario con los ratings y ver los libros similares al de entrada.

El libro de *'A Court of Thornes and Roses de Sarah J. Maas'* no se encuentra en el dataset, lo substituimos por harry potter.

```
book id: 0174434693
Los 1 libros que te recomendamos después de leer Hamlet son: ['William Shakespeare: King Lear']
book id: 9504604730
Los 1 libros que te recomendamos después de leer Don Quijote de La Mancha son: ['Palestina, Ocupacio I Resistencia: Manual
Practic Sobre La Questio Palestina I El Conflicte Arabo-Israelia']
book id: 1594130000
Los 1 libros que te recomendamos después de leer Harry Potter and the Sorcerer's Stone son: ['Not of War Only']
```

Viendo las recomendaciones:

- Para Hamlet, la recomendación tiene sentido, ya que se trata de otra obra épica de William Shakespeare.
- Para Don Quijote, obtenemos un libro sobre Palestina. Realmente podemos ver cómo no hay mucha relación entre ambos libros. Quizá tal y cómo se ha indicado al principio, la falta de contecxto, clústering por género hace que el modelo pueda ser impreciso.
- Para Harry Potter, la recomendación es Not of War only, lo que parece tener sentido ya que se trata de otro libro de aventuras y fantástico. Aunque quizá esperaríamos alguna recomendación cómo otro de los libros de HP.