## Descripción: C:\Documents and Settings\admin\Escritorio\WorkSpace\Gestio\Logos\logo laSalle_URL_enginyeria_ positiu_CAT.jpg

#### Máster en Big Data

## Asignatura: Casos de analítica

## Fecha: 07/05/2024

Apellidos: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Nombre: \_\_\_\_\_\_\_\_\_\_ Login \_\_\_\_\_\_\_\_\_\_ Curso: **2023/24**

**Caso 3: Book 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 de dos conjuntos de datos:

1. *Book-Crossing dataset*, que puede descargarse del siguiente enlace: <https://www.kaggle.com/datasets/somnambwl/bookcrossing-dataset>.
2. *LSGoodreads dataset:* que puede descargarse del eStudy.

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.

Primera Parte: Análisis Cuantitativo.

* 1. Primer examen preliminar del de los datos. ¿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?
  2. Utiliza técnicas de webscrapping para enriquecer el dataset de *Goodreads,* ya sea con libros nuevos o añadiendo características. ¿cuántos libros nuevos o características has podido añadir? ¿Qué criterio has seguido para escoger los libros que querías añadir al dataset? Detalla todo el proceso y las elecciones de diseño.

Segunda Parte: Análisis Cualitativo.

* 1. ¿Cuál fue el año en el que se publicaron más libros? Muéstralo en un gráfico ¿Y el autor más prolífico? ¿Cuántos libros suyos hay en el dataset?
  2. Analiza los orígenes geográficos y la edad de los reseñadores.

Tercera Parte: Análisis Predictivo.

* 1. Una editorial nos ha contactado para ver qué parámetros debería tener un libro para que fuera existoso. A partir del dataset y su análisis, orienta a la editorial sobre qué parámetros deben seguir a la hora de publicar un nuevo libro.
  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 enriquecerlo (por ejemplo, utilizando técnicas de *web scrapping* 😉o añadiendo más atributos a los libros actuales). 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*
* *La apología de Sócrates de Platón*

*Consideraciones*

Deben contestarse todas las preguntas para considerar el trabajo como entregado. Podéis realizarlo en grupos de máximo 2 personas.

*Entrega*

Para que el trabajo conste como entregado, debéis subir al eStudy ambos miembros del grupo tanto un informe del trabajo como el código utilizado. Los datasets que se hayan modificado también deberán ser entregados. La fecha de entrega es el 14 de junio de 2024.

**Anexo**

Para hacer este caso, es necesario hacer webscrapping de Goodreads. Para ello, se os ha facilitado un seguido de scripts de Python que os pueden ayudar con la tarea. Se recomienda utilizar el software PyCharm para correr los códigos, aunque luego en análisis del dataset se haga en Google Colab.

Los scripts más interesantes son los siguientes:

* Script **get\_books\_from\_list.py**: en este script, podéis pasarle un enlace que contenga una lista de libros de Goodreads (por ejemplo, esta: <https://www.goodreads.com/list/show/8599.Magical_and_Mystical_Creatures_Titles_ONLY_>) y se descargará el título y el identificador de Goodreads para guardarlo en una lista en formato .txt. Esta lista nos servirá para hacer obtener la información de los libros utilizando el siguiente script.
* Script **get\_books.py**: la función de este script es descargarse la información individual de cada libro gracias a una lista obtenida por el script get\_books\_from\_list.py. De este modo, abrirá la página individual de cada libro (por ejemplo, esta: <https://www.goodreads.com/book/show/5907.The_Hobbit> ) y se guardará información en referencia al título, el autor, las valoraciones, etc. Para cada libro analizado, se guardará un fichero .json con la información encontrada. Además, cuando se hayan terminado de leer todos los libros de la lista, se creará un fichero .CSV con la información de todos los libros.

“Reading brings us unknown friends.” — Honoré de Balzac