Sistema de Recomendación de Libros de Biblioteca

Mamani Ayala, Brandon (2015052715), Quispe Mamani, Angelo (2015052826), Vizcarra Llanque, Jhordy (2015052719), Ordoñez Quilli, Ronald (2015052821), Rodriguez Mamani, Juan (2017057862)

Tacna, Perú

Abstract

The recommendation systems are part of an information filtering system, which present different types of topics or information items (movies, music, books, news, images, web pages, etc.) that are of interest to a user. particular. In this short article we will try to provide a general explanation about the different data models, such as Content Filtering and Collaborative Filtering. Then we will apply these models in an example with the help of Colab, where we will filter data and show the results as well as their different graphs. Then we will give some appreciations and conclusions of what was our Final Work.

1. Resumen

Los sistemas de recomendación forman parte de un sistema de filtrado de información, los cuales presentan distintos tipos de temas o ítems de información (películas, música, libros, noticias, imágenes, páginas web, etc.) que son del interés de un usuario en particular. En este breve artículo intentaremos brindar una explicación general acerca de los diferentes modelos de datos, tales como el Filtrado por Contenido y Filtrado Colaborativo. Luego se aplicarán dichos modelos en un ejemplo con ayuda de Colab, donde filtraremos datos y mostraremos los resultados así como sus diferentes gráficas. Luego daremos unas apreciaciones y conclusiones de lo que fue nuestro Trabajo Final.

2. Introduccion

Todos hemos estado expuestos a los famosos "filtros colaborativos": cuando estabas comprando online y automáticamente el sitio web te recomienda otro artículo similar, o cuando le diste "me gusta" a algo y de repente te empezaron a salir recomendaciones similares en una red social. Los sistemas de recomendación evalúan patrones de tu comportamiento y de miles de usuarios a la vez para emitir nuevas recomendaciones. Ésta es una de las aplicaciones más comunes de "Machine Learning", porque te da la sensación de estar navegando en un sitio web programado únicamente para tí.

La mayoría de sistemas de recomendación de productos usan filtros colaborativos. Estamos en una era donde nosotros como usuarios generamos más información que nunca antes en la historia. Esta información es usada por los servicios que usamos a diario para darnos una experiencia mas personalizada.

Un claro ejemplo de esto es Netflix, ellos usan la información que nosotros generamos para hacernos recomendaciones y mostrarnos categorías y películas según nuestros gustos. Otro ejemplo es Amazon, que hace recomendaciones de productos que nos podrían interesar, basándose en los productos que ya compramos.

El filtro colaborativo es una técnica usada por los sistemas de recomendación, se basa en el hecho de que si dos personas X y Y en un mismo sistema tienen gustos similares, entonces recomendarle a la persona X cosas que le gusten a la persona Y será de gran relevancia, en contraste a simplemente darle una opción cualquiera.

Hay una variación de filtro colaborativo que se basa en comparar los items en vez de los usuarios para arrojar los items relacionados, de modo que si un usuario entra a ver un determinado item, le podemos recomendar otro.

3. Marco Teorico

3.1 Filtrado Contenido

El filtro de contenido web es una solución de software y/o hardware que tiene como finalidad actuar como un intermediario entre los accesos de los colaboradores a internet, posibilitando la aplicación de políticas definidas por la empresa.

Esto significa que el acceso a Internet deja de ser hecho directamente por el ordenador y pasa a ser realizado a través de la solución de filtrado de contenido web. Para el usuario esta transacción puede ser transparente, o mediante el uso de credenciales para el reconocimiento del usuario durante el primer intento de acceso a Internet. Es importante resaltar que la conexión con el sitio remoto es hecha por el filtro de contenido web, que aplica las reglas, y si el intento de acceso está en conformidad el sitio se brinda al navegador.

Este modelo garantiza que las peticiones sean tunelizadas y, conociendo detalles de quien está realizando la solicitud, puedan ser permitidas o no. Esto garantiza también, en muchos casos, la protección contra phishing y otras contaminaciones a través de páginas en Internet.

Existen diversas soluciones de filtrado de contenido web, desde las más simples que permiten crear reglas de acceso con contenidos que pueden o no ser accedidos, hasta soluciones más complejas y completas, que poseen bases de datos voluminosas con clasificación de sitios basados en categorías de interés.

¿Cómo aumentar la productividad utilizando el filtro de contenido web?

- Después del concepto de filtrado de contenido web, el incremento de productividad en ambientes corporativos se da a través de la implementación de esta solución, que traerá gran visibilidad sobre los accesos y consumo de Internet.

- Una buena estrategia en la implantación de estas soluciones es permitir todo el tráfico, si la empresa no posee ninguna regla anterior, de esta forma es posible identificar el perfil de consumo y posibles abusos por parte de los colaboradores.
- El más importante, y factor crítico de éxito, es buscar entender con usuarios, líderes de sector, o personas en cargo de confianza, lo que es primordial en Internet para la realización de su trabajo. Esto debe funcionar siempre, de lo contrario su negocio será afectado.
- Las medidas radicales que restringen totalmente el Internet para determinados usuarios deben ser muy bien analizadas, porque además de no ser positivo en términos motivacionales, puede estimular a que el colaborador busque alternativas para burlar el acceso.
- Cada empresa tiene sus necesidades particulares de acceso a Internet, y respetar esto es fundamental para aumentar la productividad y la seguridad en su entorno. Busque evitar radicalismos, entender lo que es primordial para el trabajo, pero también crear accesos en determinados horarios para que se pueda leer noticias, accederse a las redes sociales, estimulando el ocio creativo.
- Esto es totalmente posible y saludable con tecnologías de filtrado de contenido web, pues además de poder controlar horarios, direcciones, usuarios, sitios y etc. Es también visible en la mayoría de las soluciones el consumo de Internet. Con base en ello, y detectando abusos, se pueden tomar medidas de concientización y bloqueos.

3.2 Filtrado por Colaborativo

El filtrado colaborativo es una técnica utilizada por los sistemas de recomendación para solventar los problemas derivados de la sobreinformación que los consumidores sufren en Internet. Esta tendencia crece cada día más, debido a su enorme funcionalidad son más los usuarios que se valen de esta herramienta en sus búsquedas.

Antes del nacimiento de Internet el consumidor no tenía ninguna fuente de información salvo la propia publicidad del producto. El mercado ha pasado de esta escasez de información a la saturación de los mismos. En

este contexto, surgen los filtrados colaborativos. Las empresas incorporan estas herramientas en su página y los propios usuarios construyen una inteligencia colectiva mediante un sistema de recomendaciones que son luego estudiados y traducidos mediante algoritmos estadísticos.

Una de las empresas pioneras en incorporar esta herramienta dentro de su web fue la famosa tienda online Amazon.com, que informa a sus usuarios de los productos que podían interesarles partiendo de los que ya había clicado.

Existen diferentes tipos de filtrado a la hora de establecer las recomendaciones, se pueden clasificar en cuatro:

- Filtrados basado en contenido: las recomendaciones se hacen según los contenidos que puedan gustar o interesar.
- Filtrados demográficos: se realizan por las características de los usuarios (edad, sexo, estudios...).
- Filtrados colaborativos: las recomendaciones están basadas en las búsquedas con votos positivos de usuarios similares.
- Filtrados híbridos: mezclan los dos o tres de los filtrados anteriores para una mejor experiencia.

Los filtrados colaborativos sirve para hacer predicciones automáticas sobre los intereses de un usuario mediante la recopilación de preferencias o gustos del mismo consumidor u otros consumidores con intereses comunes.

Los sistemas de filtrado poseen muchas variantes con algoritmos que se utilizan para su elaboración:

- Algoritmos de filtrado colaborativo basados en memoria, o algoritmos de vecinos cercanos (Nearest Neighbour): utilizan los datos de recogidos para calcular la similitud entre los usuarios o elementos comunes. Fue de los primeros en usarse y es sencillo y eficaz. Funcionan buscando usuarios con patrones de evaluación similares con

el usuario activo, para el que se está haciendo la selección. También utilizan técnicas estadísticas para encontrar vecinos con un historial de búsqueda parecido al usuario actual. Su principal inconveniente es que necesitan un número mínimo de usuarios para realizar la recomendación.

- Algoritmos de filtrado colaborativo basados en Modelo: se utilizan algoritmos de aprendizaje automático para encontrar patrones. Mejora el rendimiento en cuanto a la predicción porque da un fundamento más intuitivo. Funcionan usando las evaluaciones de los usuarios afines para calcular la elección del usuario activo. Primero elaboran un modelo de las búsquedas del usuario pero este proceso necesita un aprendizaje largo e intensivo.

También existen algoritmos híbridos que combinan ambos modelos pero son complejos y costosos de implementar.
Dificultades que podemos encontrar cuando usamos el filtrado colaborativo:

- Escasez de datos: los sistemas de filtrado colaborativo se basan en conjuntos de datos. Si esta muestra de datos es escasa puede ser muy costosa y poco eficaz. En ocasiones un problema común es empezar de cero, ya que no se pueden recopilar preferencias con precisión y fiabilidad.
- Sinónimos: la diversidad de etiquetas con nombres similares a veces no son reconocidos por los sistema de filtrados cuando en realidad el usuario está buscando el mismo elemento y se pierde información. Por ejemplo: un usuario que busca ordenadores o computadoras, son sinónimos pero el buscador no los relaciona.
- Haters o black sheep: otra dificultad que afecta a los sistemas de filtrados son las opiniones de los usuarios que no están de acuerdo con nada y todas sus recomendaciones son negativas, empeoran la calidad de las filtraciones.
- Shilling attacks: en los sistemas de recomendación cualquiera puede hacer evaluaciones, pudiendo un usuario votar positivamente sólo

a sus productos y servicios y dar negativo a sus competidores, falseando la eficacia de esta herramienta.

- Diversidad: los filtros intentan buscar una diversidad para poder recomendar entre múltiples opciones. En ocasiones este filtro van reduciendo esta variedad dando sólo visibilidad a los productos con mayor popularidad.

El filtrado colaborativo es una gran herramienta para mejorar la visibilidad y la mejor forma de dar a conocer nuevos productos a más clientes ¿Quieres continuar aprendiendo técnicas para tu negocio? No dudes y échale un vistazo a nuestro Postgrado en e-Commerce Omnichannel de IEBS Business School donde aprenderás todo lo que necesitas para dar el mejor impulso a tu empresa.

4. Ejemplos

Pandas es un paquete de Python que proporciona estructuras de datos similares a los dataframes de R. Pandas depende de Numpy, la librería que añade un potente tipo matricial a Python. Los principales tipos de datos que pueden representarse con pandas son:

- Datos tabulares con columnas de tipo heterogéneo con etiquetas en columnas y filas.Series temporales.
- Series temporales.

Pandas proporciona herramientas que permiten:

- Leer y escribir datos en diferentes formatos: CSV, Microsoft Excel, bases SQL y formato HDF5
- Seleccionar y filtrar de manera sencilla tablas de datos en función de posición, valor o etiquetas
- Fusionar y unir datos
- Transformar datos aplicando funciones tanto en global como por ventanas

- Manipulación de series temporales
- Hacer gráficas.

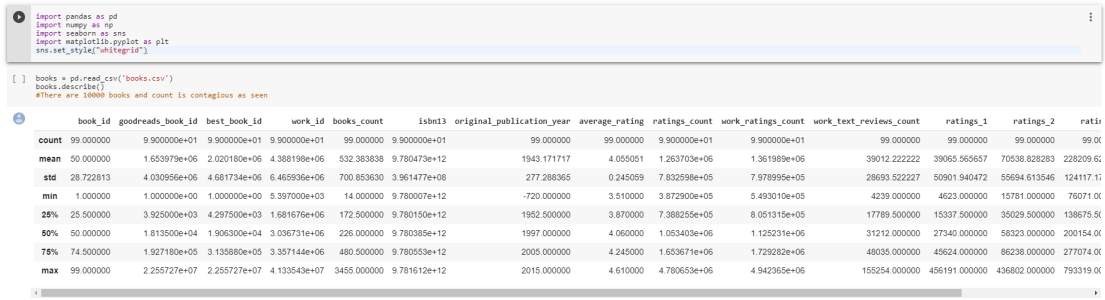
En pandas existen tres tipos básicos de objetos todos ellos basados a su vez en Numpy:

- Series (listas, 1D)
- DataFrame (tablas, 2D)
- Panels (tablas 3D)

En Python la biblioteca más extendida para gráficas 2D y 3D es matplotlib. Permite obtener gráficas de muy buena calidad, con una gran capacidad de control y una curva de aprendizaje moderada. Todos los aspectos de la figura pueden controlarse mediante código.

Vistas:

- Importacion de librerias y datos del archivo book.csv



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_style("whitegrid")
```

```
books = pd.read_csv('books.csv')
books.describe()
#There are 10000 books and count is contagious as seen
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn13	original_publication_year	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratir
count	99.000000	9.900000e+01	9.900000e+01	9.900000e+01	99.000000	9.900000e+01	99.000000	99.000000	9.900000e+01	9.900000e+01	99.000000	99.000000	99.000000	99.0
mean	50.000000	1.653979e+06	2.020180e+06	4.388198e+06	532.383838	9.780473e+12	1943.171717	4.055051	1.263703e+06	1.361989e+06	39012.222222	39065.565657	70538.828283	228209.62
std	28.722813	4.030956e+06	4.681734e+06	6.465936e+06	700.853630	3.961477e+08	277.288365	0.245059	7.832598e+05	7.978995e+05	28693.522227	50901.940472	55694.613546	124117.17
min	1.000000	1.000000e+00	1.000000e+00	5.397000e+03	14.000000	9.780007e+12	-720.000000	3.510000	3.872900e+05	5.493010e+05	4239.000000	4623.000000	15781.000000	76071.00
25%	25.500000	3.925000e+03	4.297500e+03	1.681676e+06	172.500000	9.780150e+12	1952.500000	3.870000	7.388255e+05	8.051315e+05	17789.500000	15337.500000	35029.500000	138675.50
50%	50.000000	1.813500e+04	1.906300e+04	3.036731e+06	226.000000	9.780385e+12	1997.000000	4.060000	1.053403e+06	1.125231e+06	31212.000000	27340.000000	58323.000000	200154.00
75%	74.500000	1.927180e+05	3.135880e+05	3.357144e+06	480.500000	9.780553e+12	2005.000000	4.245000	1.653671e+06	1.729282e+06	48035.000000	45624.000000	86238.000000	277074.00
max	99.000000	2.255727e+07	2.255727e+07	4.133543e+07	3455.000000	9.781612e+12	2015.000000	4.610000	4.780653e+06	4.942365e+06	155254.000000	456191.000000	436802.000000	793319.00

```
books.head()
```

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	title	language_code	average_rating	ratings_count	work_ratings_count	work_text_reviews_count	ratings_1
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	The Hunger Games (The Hunger Games, #1)	eng	4.34	4780653	4942365	155254	66715
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	Harry Potter and the Sorcerer's Stone (Harry P...	eng	4.44	4602479	4800065	75867	75504
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	Twilight (Twilight, #1)	en-US	3.57	3866839	3916824	95009	456191
3	4	2657	2657	3275794	487	61120081	9.780061e+12	Harper Lee	1960.0	To Kill a Mockingbird	To Kill a Mockingbird	eng	4.25	3198671	3340896	72586	60427
4	5	4671	4671	245494	1356	743273567	9.780743e+12	F. Scott Fitzgerald	1925.0	The Great Gatsby	The Great Gatsby	eng	3.89	2683664	2773745	51992	86236

- Nombre de los autores del archivo book.csv

```
[ ]  print(books_filter['authors'])
```

```
0                                      Suzanne Collins
1                            J.K. Rowling, Mary GrandPré
2                                    Stephenie Meyer
3                                         Harper Lee
4                                  F. Scott Fitzgerald
5                                         John Green
6                                       J.R.R. Tolkien
7                                       J.D. Salinger
8                                           Dan Brown
9                                        Jane Austen
10                                     Khaled Hosseini
11                                      Veronica Roth
12              George Orwell, Erich Fromm, Celâl Üster
13                                      George Orwell
14     Anne Frank, Eleanor Roosevelt, B.M. Mooyaart-D...
15                           Stieg Larsson, Reg Keeland
16                                     Suzanne Collins
17              J.K. Rowling, Mary GrandPré, Rufus Beck
18                                      J.R.R. Tolkien
19                                     Suzanne Collins
20                           J.K. Rowling, Mary GrandPré
21                                       Alice Sebold
22                           J.K. Rowling, Mary GrandPré
23                           J.K. Rowling, Mary GrandPré
24                           J.K. Rowling, Mary GrandPré
25                                          Dan Brown
26                           J.K. Rowling, Mary GrandPré
27                                     William Golding
28         William Shakespeare, Robert           Jackson
29                                       Gillian Flynn
                               ...
69                                     Orson Scott Card
70     Mary Wollstonecraft Shelley, Percy Bysshe Shel...
71                                       Stephen King
72                                     Stephenie Meyer
73                                          John Green
74                                       Helen Fielding
75              Jane Austen, Tony Tanner, Ros Ballaster
76                           Louis Sachar, Louis Sachar
77                                    Lauren Weisberger
78     Homer, Robert Fagles, E.V. Rieu, Frédéric Mugl...
79     Antoine de Saint-Exupéry, Richard Howard, Dom ...
80                                     Jeannette Walls
81                                        Jon Krakauer
82     Charles Dickens, Richard Maxwell, Hablot Knigh...
83                                     Michael Crichton
84                                     Shel Silverstein
85                                        John Grisham
86                           Elie Wiesel, Marion Wiesel
87                                          John Green
88                                     William Goldman
89                                         S.E. Hinton
90                                       James Dashner
91                  Steven D. Levitt, Stephen J. Dubner
92                               Frances Hodgson Burnett
93              Gabriel García Márquez, Gregory Rabassa
94                        Oscar Wilde, Jeffrey Eugenides
95                                       E.L. James
96           Bram Stoker, Nina Auerbach, David J. Skal
97                           Stieg Larsson, Reg Keeland
98                                       E.L. James
```

- Titulos de los libros del archivo book.csv

```
[ ]  print(books_filter['title'])
```

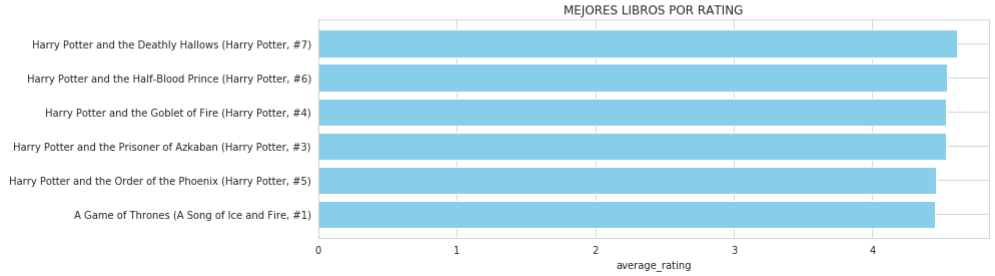
```
0                The Hunger Games (The Hunger Games, #1)
1      Harry Potter and the Sorcerer's Stone (Harry P...
2                               Twilight (Twilight, #1)
3                                 To Kill a Mockingbird
4                                      The Great Gatsby
5                                The Fault in Our Stars
6                                            The Hobbit
7                               The Catcher in the Rye
8                 Angels & Demons  (Robert Langdon, #1)
9                                   Pride and Prejudice
10                                      The Kite Runner
11                           Divergent (Divergent, #1)
12                                                1984
13                                         Animal Farm
14                           The Diary of a Young Girl
15     The Girl with the Dragon Tattoo (Millennium, #1)
16                Catching Fire (The Hunger Games, #2)
17     Harry Potter and the Prisoner of Azkaban (Harr...
18     The Fellowship of the Ring (The Lord of the Ri...
19                    Mockingjay (The Hunger Games, #3)
20     Harry Potter and the Order of the Phoenix (Har...
21                                   The Lovely Bones
22     Harry Potter and the Chamber of Secrets (Harry...
23     Harry Potter and the Goblet of Fire (Harry Pot...
24     Harry Potter and the Deathly Hallows (Harry Po...
25              The Da Vinci Code (Robert Langdon, #2)
26     Harry Potter and the Half-Blood Prince (Harry ...
27                                    Lord of the Flies
28                                     Romeo and Juliet
29                                            Gone Girl
                            ...
69                     Ender's Game (Ender's Saga, #1)
70                                          Frankenstein
71                        The Shining (The Shining #1)
72                             The Host (The Host, #1)
73                                   Looking for Alaska
74           Bridget Jones's Diary (Bridget Jones, #1)
75                                Sense and Sensibility
76                                    Holes (Holes, #1)
77     The Devil Wears Prada (The Devil Wears Prada, #1)
78                                         The Odyssey
79                                    The Little Prince
80                                     The Glass Castle
81                                        Into the Wild
82                                A Tale of Two Cities
83                  Jurassic Park (Jurassic Park, #1)
84                                       The Giving Tree
85                                        A Time to Kill
86                        Night (The Night Trilogy #1)
87                                          Paper Towns
88                                   The Princess Bride
89                                       The Outsiders
90                     The Maze Runner (Maze Runner, #1)
91     Freakonomics: A Rogue Economist Explores the H...
92                                    The Secret Garden
93                         One Hundred Years of Solitude
94                           The Picture of Dorian Gray
95                Fifty Shades Freed (Fifty Shades, #3)
96                                               Dracula
97        The Girl Who Played with Fire (Millennium, #2)
98                Fifty Shades Darker (Fifty Shades, #2)
```

- Consultas de Mejores Libros

```
[ ]  #Graphics of book
     pop = books.sort_values('average_rating', ascending=False)
     import matplotlib.pyplot as plt
     plt.figure(figsize=(12,4))

     plt.barh(pop['title'].head(6),pop['average_rating'].head(6), align='center',color='skyblue')
     plt.gca().invert_yaxis()
     plt.xlabel("average_rating")
     plt.title("MEJORES LIBROS POR RATING")
```

```
Text(0.5, 1.0, 'MEJORES LIBROS POR RATING')
```



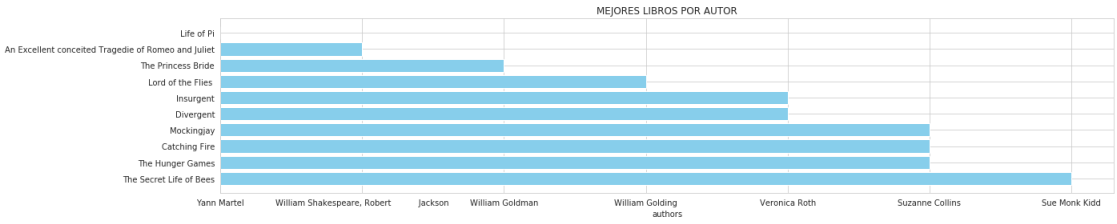
- Consultas de Mejores Libros por Autor

  Este objetivo se hace necesario para poder iniciar una valoracion de los autores.

```
[ ]  #Graphics of book
     pop = books.sort_values('authors', ascending=False)
     import matplotlib.pyplot as plt
     plt.figure(figsize=(20,4))

     plt.barh(pop['original_title'].head(10),pop['authors'].head(10), align='center',color='skyblue')
     plt.gca().invert_yaxis()
     plt.xlabel("authors")
     plt.title("MEJORES LIBROS POR AUTOR")
```

```
Text(0.5, 1.0, 'MEJORES LIBROS POR AUTOR')
```



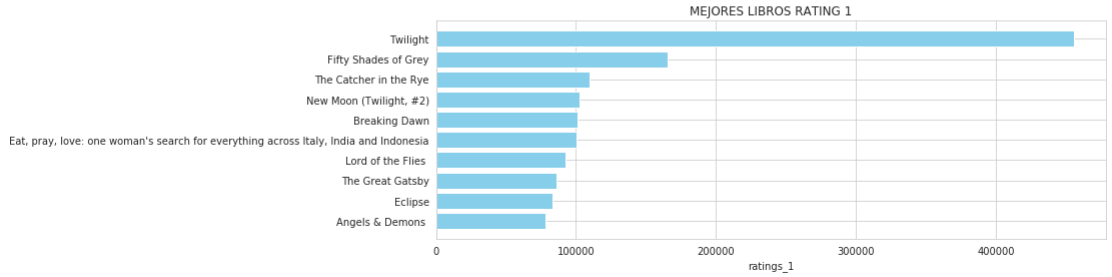
- Consultas de Mejores Libros por Rating 1

  Este objetivo se hace necesario para poder iniciar una valoracion de los libros.

```
[ ]  #Graphics of book
     pop = books.sort_values('ratings_1', ascending=False)
     import matplotlib.pyplot as plt
     plt.figure(figsize=(12,4))

     plt.barh(pop['original_title'].head(10),pop['ratings_1'].head(10), align='center',color='skyblue')
     plt.gca().invert_yaxis()
     plt.xlabel("ratings_1")
     plt.title("MEJORES LIBROS RATING 1")
```

```
Text(0.5, 1.0, 'MEJORES LIBROS RATING 1')
```



5. Conclusion

El método Kimball está orientado a la consulta de la información, por lo que su estructura interna está especialmente diseñada para garantizar una explotación de los datos rápida y sencilla, no requiriendo usuarios especializados para ello. Por el contrario, el método Inmon persigue la integración de todos los datos de la compañía, estando orientado hacia el almacenaje de grandes volúmenes de datos, por lo que su estructura interna normalizada se diseña para evitar la redundancia de datos, simplificar las labores de mantenimiento, etc. cuestiones que complican las consultas de la información, requiriendo que los usuarios finales estén mucho más especializados. Así, podríamos decir que el enfoque de Kimball se ajusta más a proyectos pequeños en los que se persiga un sistema fácilmente explotable y entendible por el usuario y de rápido desarrollo, siendo el modelo de Inmon más apropiado para sistemas complejos de mayor importancia.

Referencias
Referencias

[1]

[2]

[3] http://tdan.com/data-warehouse-design-inmon-versus-kimball/20300

[4] https://blog.bi-geek.com/arquitectura-comparativa-inmon-y-kimball/

[5] https://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/

[6] https://twooctobers.com/blog/8-data-storytelling-concepts-with-examples/

[7] https://www.ucasal.edu.ar/htm/ingenieria/cuadernos/archivos/5-p56-rivadera-formateado.pdf