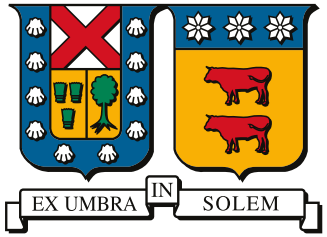


UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

DEPARTAMENTO
DE MATEMÁTICA

# Sistema de recomendación de libros utilizando modelos de ML

## Aplicaciones de la Matemática en la Ingeniería

**J. Martínez – B. Muñoz – J.L Nanjari**
**Y. Parra – M. Solla – G. Vega**

2$^{do}$ semestre 2022

# Contenidos




Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Descripción del Problema y Objetivos

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Motivación

¿Alguna vez has terminado un libro, o una serie, o una película, y has querido aventurarte en una nueva historia de la misma temática anterior, pero no sabías qué hacer?

Los sistemas recomendadores sirven para que estas situaciones se resuelvan inmediatamente, ¡y puedas seguir disfrutando del contenido que más disfrutas!

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Nuestro problema y objetivos

Nuestro problema particular se basa en cómo recomendar libros que se encuentran en el conjunto de datos de Goodreads Books.

**Objetivos:**

1. Crear un sistema basado en modelos de Machine Learning, identificando los parámetros más importantes para este.

2. Que el sistema creado no deje de recomendar libros menos populares por sobre otros.

3. Comparar diversos métodos de comparación para ver cuál es el mejor.

# Análisis Exploratorio de Datos

Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Tipos datos del DataFrame

**bookID:** Número identificador del libro en la data
**title:** Título del libro
**authors:** Autores del libro
**average_rating:** Rating promedio
**isbn:** Número identificador Internacional de 11 dígitos
**isbn13:**Número identificador Internacional de 13 dígitos
**language_code:** Idioma del libro
**num_pages:** Número de páginas
**ratings_count:** Cantidad de ratings
**text_reviews_count:** Cantidad de reviews escritas
**publication_date:** Fecha de publicación
**publisher:** Editorial

# Número de Páginas

| | raitings | num_pages | reviews |
|---|---|---|---|
| mean | 3,93 | 336 | 542 |
| min | 0 | 0 | 0 |
| max | 5 | 6576 | 94265 |

Figura: Descripción del DataFrame

| Nombre | Páginas | Publisher |
|---|---|---|
| Murder by Moonlight & Other Mysteries | 0 | Simon Schuster Audio |

Figura: Ejemplo de Audiolibro

# Idiomas

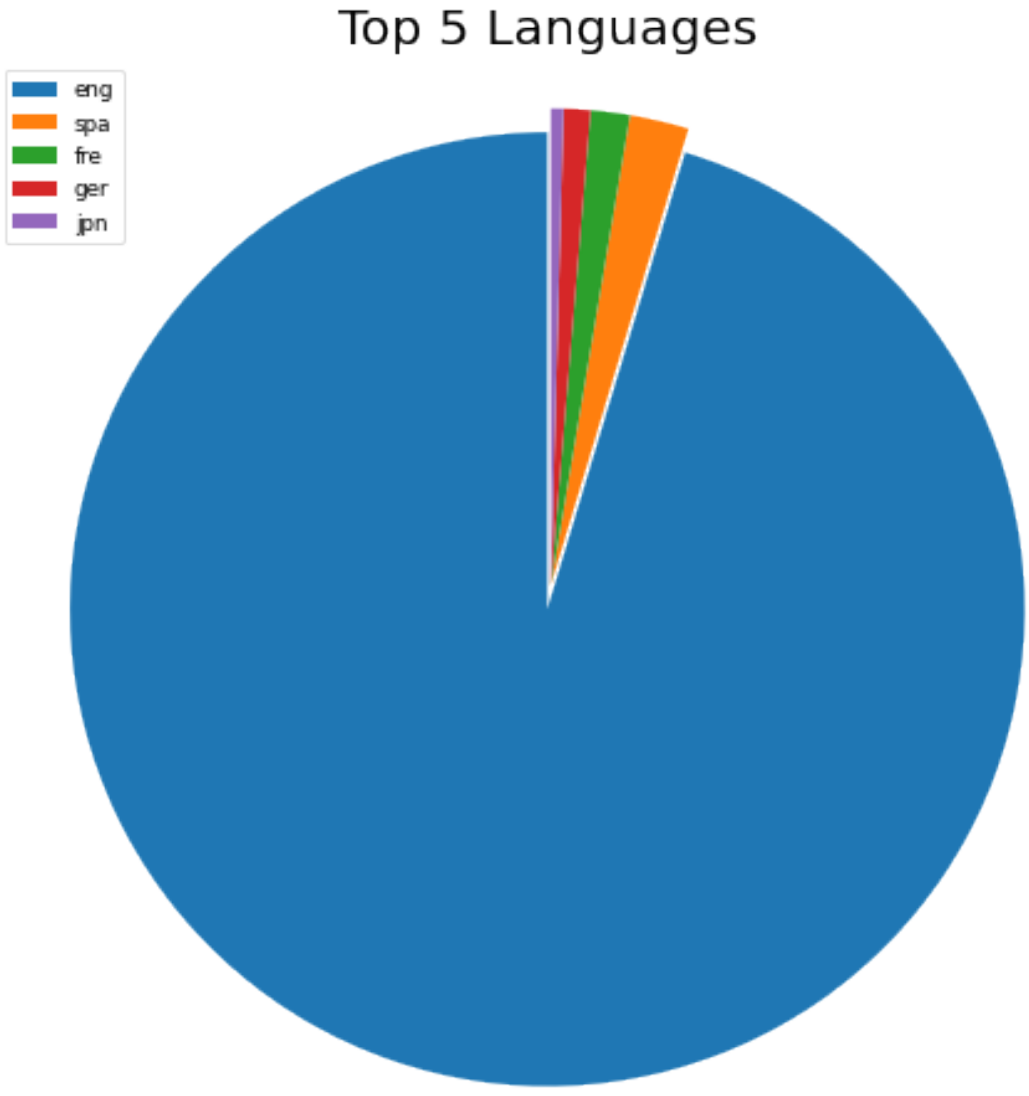


Figura: Distribución libros por idioma

# Correlación

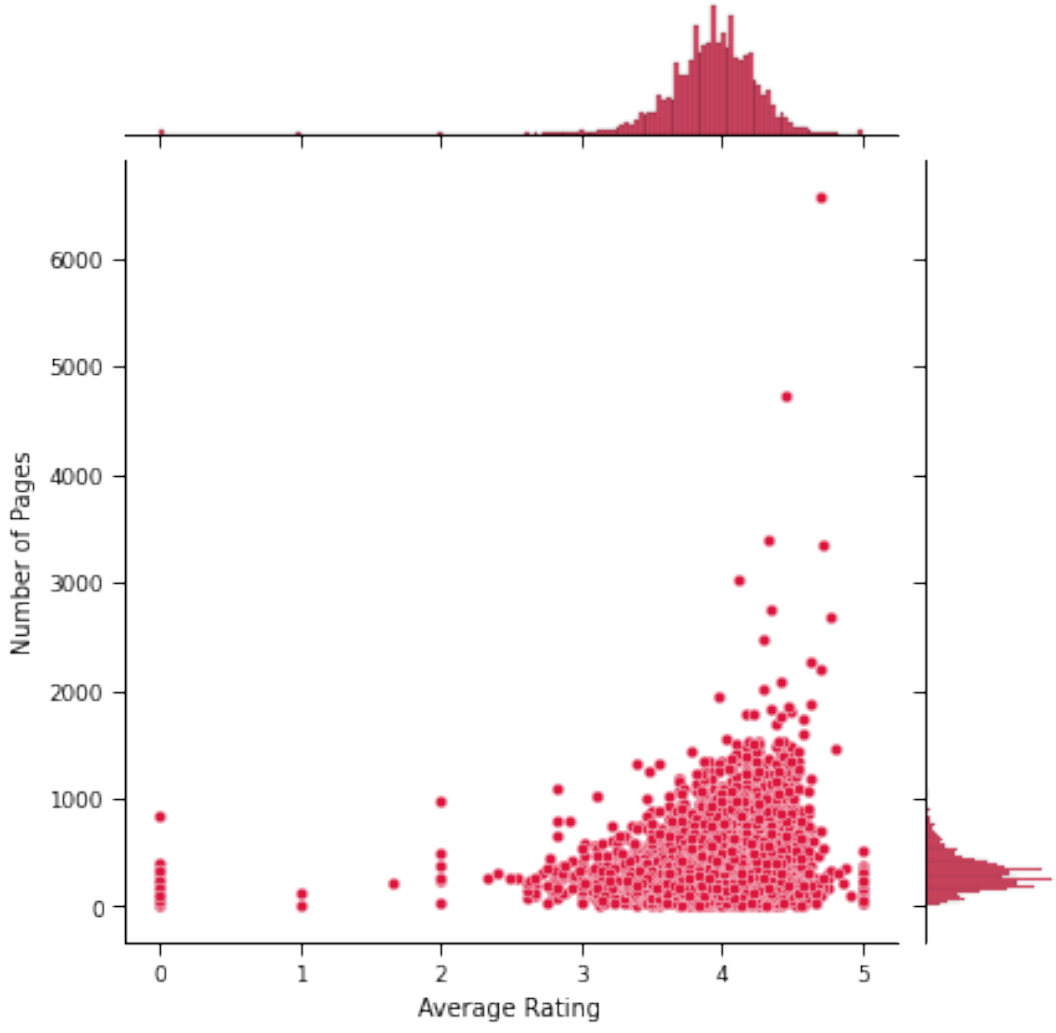


Figura: Distribución rating promedio por n° de páginas

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Correlación

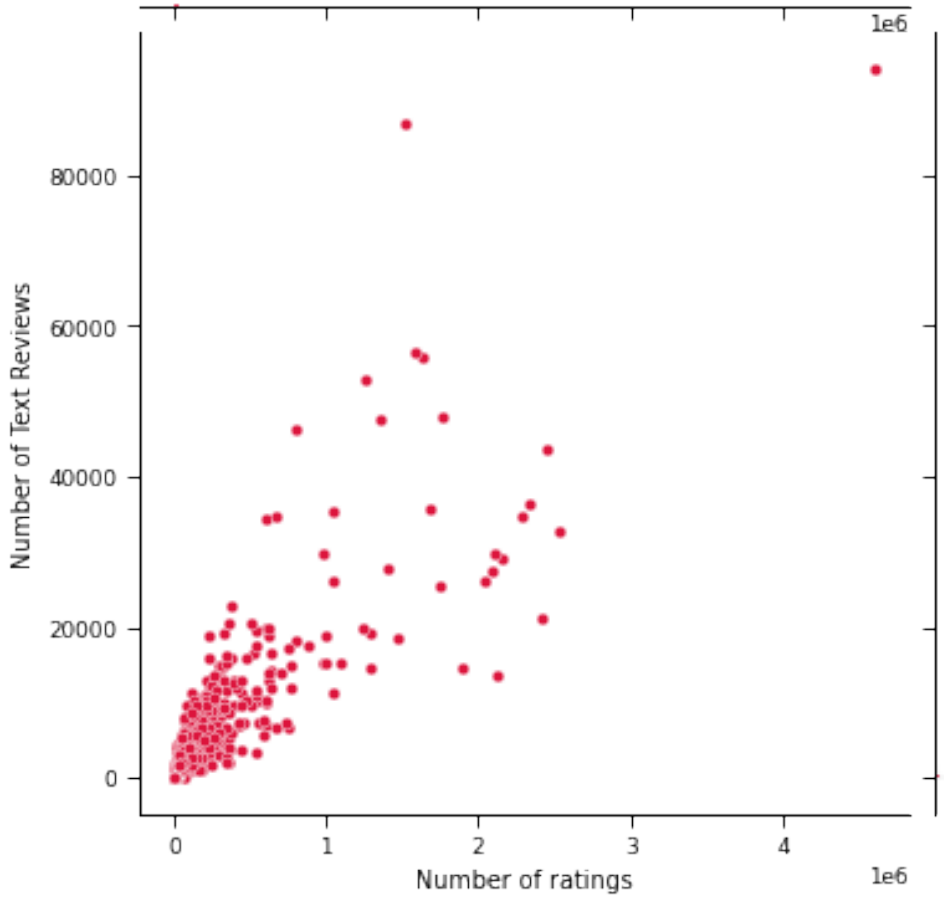


Figura: Distribución n° de ratings por n° de reviews

# Autores y Rating Ajustado

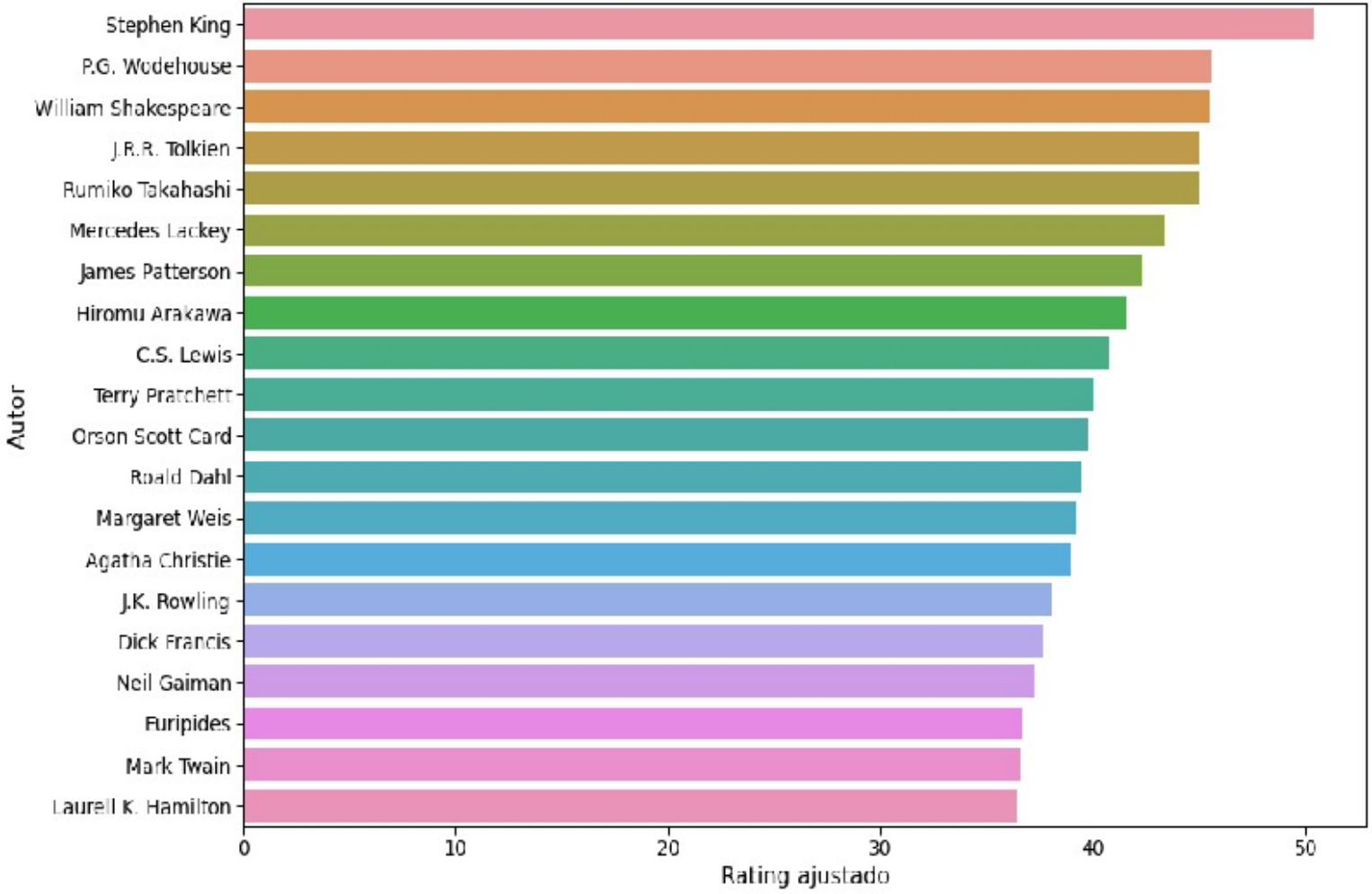


Figura: Distribución rating ajustado por autores

# Años

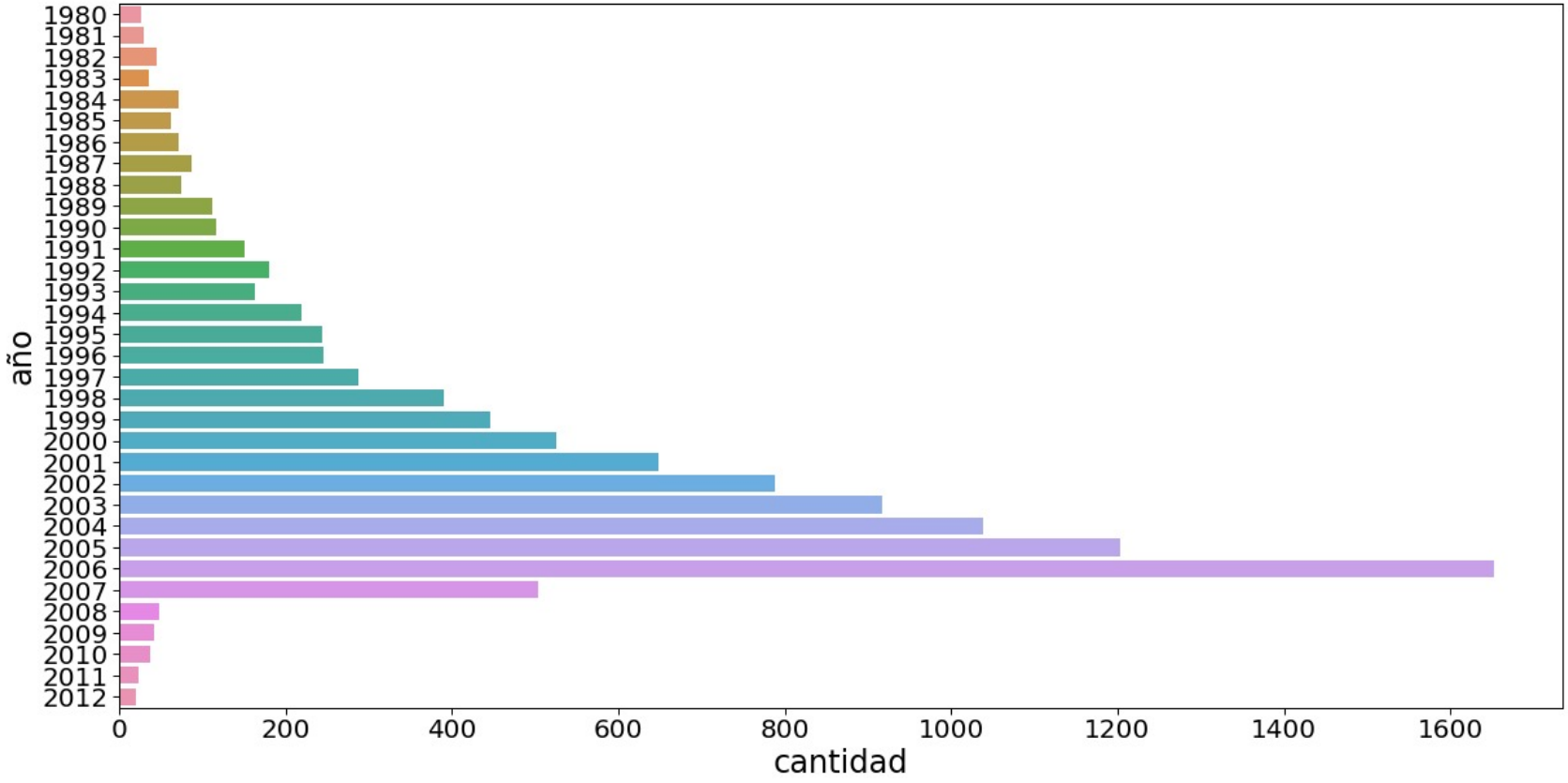


Figura: Distribución libros por año

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Conclusiones del Análisis

Los parámetros más importantes para realizar el modelo:

1. Idioma
2. Cantidad de páginas
3. Rating
4. Autor
5. Año de publicación

# Modelos de Recomendación

Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Vecinos más cercanos

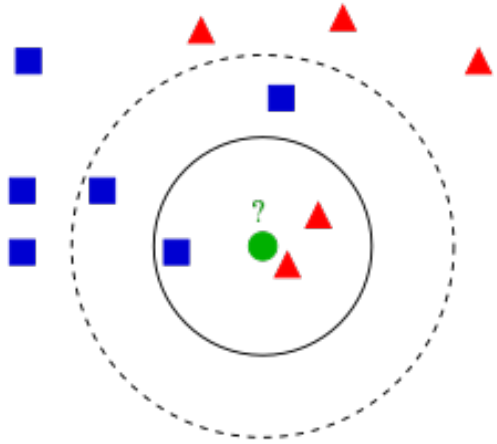


Figura: Esquema de Vecinos Más Cercanos

Sean $x, y$ vectores:

Euclideana: $\sqrt{\langle x-y, x-y \rangle}$

Coseno similitud: $\frac{\langle x,y \rangle}{||x||||y||}$

Kernel sigmoide: $\tanh(\alpha \langle x, y \rangle + c)$

Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Preprocesamiento

1. Variables numericas estandarizadas (StandarScaler)
2. Metodo dummie para los idiomas
3. Codigo autor segun adjusted rating
4. df_procesado

# Resultados del Modelo: Euclideana

| Euclideana | num_pages | codigo_autor | Año | text_reviews | ratings_count |
|---|---|---|---|---|---|
| Harry Potter #6 | 652 | 14 | 2006 | 27k | 2M |
| Harry Potter #5 | 870 | 14 | 2004 | 29k | 2M |
| Animal Farm | 122 | 114 | 2003 | 29k | 2M |
| Lord of the Flies | 182 | 3609 | 1999 | 26k | 2M |
| Harry Potter #2 | 341 | 14 | 1999 | 34k | 2M |
| Harry Potter #3 | 435 | 14 | 2004 | 36k | 2M |

Figura: Resultados del modelo utilizando Distancia Euclideana

# Resultados del Modelo: Coseno Similitud

| Coseno | num_pages | codigo_autor | Año | text_reviews | ratings_count |
|---|---|---|---|---|---|
| Harry Potter #6 | 652 | 14 | 2006 | 27k | 2M |
| Harry Potter #5 | 870 | 14 | 2004 | 29k | 2M |
| Little Women | 449 | 158 | 2004 | 18k | 1,5M |
| The Hobbit  or There and Back Again | 366 | 3 | 2002 | 32k | 2,5M |
| Memoirs of a Geisha | 434 | 891 | 2005 | 19k | 1,3M |
| Harry Potter #3 | 435 | 14 | 2004 | 36k | 2,3M |

Figura: Resultados del modelo utilizando Coseno Similitud

# Resultados del Modelo: Distancia Sigmoidal

| Sigmoide | num_pages | codigo_autor | Año | text_reviews | ratings_count |
|---|---|---|---|---|---|
| Harry Potter #6 | 652 | 14 | 2006 | 27k | 2M |
| The Book Thief | 552 | 1292 | 2006 | 86k | 1,5M |
| Twilight | 501 | 3747 | 2006 | 94k | 1,5M |
| The Hobbit or There and Back Again | 366 | 3 | 2002 | 32k | 2,5M |
| Harry Potter #3 | 435 | 14 | 2004 | 36k | 2,3M |
| Harry Potter #2 | 341 | 14 | 1999 | 34k | 2,3M |

Figura: Resultados del modelo utilizando Distancia Sigmoidal

# Resultados

1. General: falla el genero, acierta el autor
2. Euclideana: Atributos en general similares (excepciones)
3. Coseno: No sobrestima valores grandes
4. Sigmoidal: Libros vecinos de libros distintos
5. Sigmoidal: Recomendaciones distintas a Euclideana y Coseno

# Conclusiones

Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Conclusiones

1. El sistema de recomendación es excelente para agrupar libros por idioma y popularidad, y también incorpora adecuadamente el resto de los datos en su recomendación.
2. Algunas recomendaciones son muy malas respecto al género del libro (pero era algo esperable, dado el dataset utilizado)
3. Al implementar un sistema recomendador, son fundamentales tanto el entendimiento y manejo de los datos como el uso adecuado de modelos de ML.

DMAT
Departamento de Matemática
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

# Trabajos futuros

1. Refinar el modelo, creando un buen sistema de comparación de resultados para identifcar una métrica óptima.
2. Obtener una mejor base de datos, que contenga información sobre el contenido de los libros.
3. Explorar otros paradigmas para sistemas recomendadores (algoritmo colaborativo y sistemas mixtos).