Primera entrega proyecto de curso

Angie Dayana Rincón Mandón

Fonsy Johan Mercado Agudelo

Orlando José Salazar Polo

Introducción a la Inteligencia Artificial para las Ciencias e Ingenierías

Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2023

1. Descripción del problema

La cantidad de información digital disponible se encuentra constantemente en crecimiento lo que ha derivado en una sobrecarga de información que hace que el acceso a elementos de interés en internet para cada individuo sea cada vez más difícil. Lo que hace que exista un aumento en la demanda de sistemas que se encarguen de recomendación.

Los sistemas de recomendación cuentan con elementos de filtrado de información que se encargan de este problema de sobrecarga al limitar los elementos de información más importante de una cantidad enorme de información que se genera dinámicamente de acuerdo con el interés que se observa en el usuario.

El sistema de recomendación tiene la capacidad de predecir si el usuario en cuestión tiene prefiere algún elemento o no según su perfil. Estos son beneficiosos tanto para los proveedores de servicios como para los usuarios, ya que se presenta una reducción en costos de encontrar y seleccionar los elementos que sean de interés para el usuario.

2. Dataset

Para el desarrollo de este proyecto, se seleccionó el *dataset* "*The Movies Dataset*" proporcionado en *Kaggle*. Estos archivos contienen metadatos de 45000 películas enumeradas en el conjunto de datos completo de *MovieLens*. Además, el conjunto de datos también incluye archivos que contienen 26 millones de calificaciones de 270000 usuarios para estas 45000 películas.

Las calificaciones están en una escala de 1 a 5 y se obtuvieron del sitio web oficial de *GroupLens*. El dataset tiene un tamaño total de 900 MB y contiene los siguientes archivos en formato *.csv*

- ***movies_metadata.csv:*** Este es el archivo principal de metadatos de películas, contiene información de 45000 películas, incluyendo carteles, fondos, presupuesto, ingresos, fechas de lanzamiento, idiomas, países de producción y empresas.
- ***keywords.csv:*** Este archivo contiene las palabras clave de la trama de la película en forma de un objeto *JSON* en cadena.
- ***credits.csv:*** Contiene información sobre el reparto y equipo técnico de las películas en forma de objeto JSON en cadena.
- ***links.csv:*** Este archivo contiene los *ID* de *TMDB* e *IMDB* de las películas que aparecen en el conjunto de datos de *MovieLens*.
- ***links_small.csv:*** Este archivo contiene los *ID* de *TMDB* e *IMDB* de un subconjunto de 9000 películas del conjunto de datos completo.
- ***ratings_small.csv:*** Este archivo es un subconjunto de 100000 calificaciones de 700 usuarios en 9000 películas.

3. Métricas

Existen diversas formas de evaluar la calidad de un algoritmo de recomendación, entre las que destacan las medidas de *accuracy* y *coverage*. El *accuracy* mide la proporción de recomendaciones correctas en comparación con el total de recomendaciones posibles, mientras que la *coverage* se refiere a la cantidad de elementos que el sistema es capaz de recomendar en el espacio de búsqueda. Para medir la precisión de los sistemas de filtrado de recomendaciones, se utilizan tanto métricas de precisión estadísticas como de soporte de decisiones.

Métricas de precisión estadística: Evaluar la precisión de una técnica de filtrado comparando las calificaciones pronosticadas directamente con la calificación real del usuario.

El error absoluto medio (MAE), el error cuadrático medio (RMSE) y la correlación se utilizan normalmente como métricas de precisión estadística.

MAE es el más popular y comúnmente utilizado, es una medida de la desviación de la recomendación del valor específico del usuario. Se calcula de la siguiente manera como se observa en la ecuación (1):

$$MAE = \frac{1}{N}\sum_{u,i} |p_{u,i} - r_{u,i}| \quad (1)$$

Donde $p_{u,i}$ es la calificación prevista para el usuario u en el elemento i, $r_{u,i}$ es la calificación real y N es el numero total de calificaciones en el conjunto de elementos. Cuanto menor sea el MAE, con mayor precisión el motor de recomendaciones predice las calificaciones de los usuarios. Además, el error cuadrático medio (RMSE) se observa en la ecuación (2) a continuación:

$$RMSE = \sqrt{\frac{1}{n}\sum_{u,i} (p_{u,i} - r_{u,i})^2} \quad (2)$$

El error cuadrático medio (RMSE) pone más énfasis en un error absoluto más grande y cuanto más bajo es el RMSE, mejor es la precisión de la recomendación.

4. Desempeño

Con el modelo se pretende proporcionar a los usuarios sugerencias de películas que puedan ser de su interés, basándose en su historial de visualización, búsquedas previas o en otros factores relevantes. El objetivo final del modelo de recomendación de películas es mejorar la experiencia del usuario al encontrar películas que le gusten, aumentando así la satisfacción del usuario y su fidelización al servicio de *streaming* o plataforma en línea que ofrezca el servicio. Además, un modelo de recomendación bien desarrollado también puede ayudar a las empresas a aumentar su base de usuarios y a incrementar sus ingresos por publicidad y por el consumo de contenidos.