Tecnológico de Monterrey

Avance de proyecto 1: Sistema de Recomendación

Maestría en inteligencia artificial aplicada
Asignatura : Análisis de grandes volúmenes de datos

Equipo:
Abraham Cabanzo Jimenez - A01794355
Ignacio Antonio Ruiz Guerra - A00889972
Moisés Díaz Malagón - A01208580

17 de mayo del 2024

Introducción

Los sistemas de recomendaciones son herramientas computacionales diseñadas para sugerir elementos de interés a los usuarios. Estos elementos pueden ser productos, servicios, contenido multimedia, o cualquier otro tipo de información relevante. Utilizan algoritmos y técnicas de análisis de datos para predecir y recomendar elementos que probablemente sean del agrado o interés del usuario, basándose en su historial de interacciones, preferencias pasadas y similitudes con otros usuarios.

Estos sistemas se utilizan ampliamente en plataformas de comercio electrónico, servicios de streaming de contenido, redes sociales y una variedad de otros contextos digitales. Su objetivo principal es mejorar la experiencia del usuario al proporcionar recomendaciones personalizadas que ayuden a descubrir nuevos elementos relevantes y a optimizar la búsqueda de información.

En este proyecto proponemos mejorar la experiencia de usuarios de servicios de streaming mediante la sugerencia de contenido acorde con sus gustos y preferencias, en específico en este caso y para delimitar el alcance, se hará un sistema de recomendación de películas.

1. Descripción general del plan del proyecto

El proyecto consta de realizar un sistema de recomendación de películas con base en las preferencias del usuario. Este sistema tiene como objetivo predecir la preferencia que un usuario tendrá de una película, y facilitará la toma de decisión del cliente al usar una plataforma de contenido de películas.

Para realizar el sistema de recomendación de películas para una plataforma de streaming, a la que llamaremos “Génesis”, se requerirán los siguientes componentes:

- Una base de datos de películas con sus características: en el punto 2 se detalla la base de datos a utilizar así como el preprocesamiento a aplicar.
- Un algoritmo de recomendación: en esta entrega se plantea un mecanismo de recomendación inicial y relativamente sencillo basado en contenido de las películas, sin embargo es muy probable que se prueben mejoras y algoritmos diversos en posteriores entregas.
- Evaluación del sistema utilizando alguna métrica específica: esto se definirá en una posterior entrega.
- Sistema de despliegue para su utilización: será importante que el sistema desarrollado pueda ser utilizado por otras personas, por medio de un mecanismo como una API web por ejemplo. Por ello en posteriores entregas se trabajará en el despliegue de la solución.

1.1 Cronograma de actividades

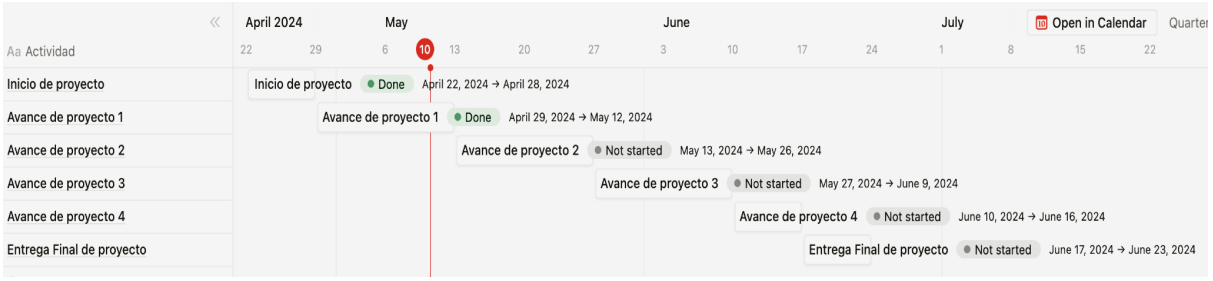


Fig 1. Cronograma de actividades

Como puede observarse en el cronograma de actividades tenemos el proyecto dividido en 4 etapas, un inicio y una entrega final.

Inicio de proyecto: fue entregado el día 28 de abril, se presentó el contexto de los sistemas de recomendación, se explicaron los diferentes tipos de sistemas de

recomendación y se plantearon dos escenarios de uso, el comercio electrónico y los sistemas de streaming de contenido.

Avance de proyecto 1: representa este documento en el que establecemos la base de datos con la que vamos a trabajar, e implementamos una primera aproximación del algoritmo de recomendación. Haremos un entendimiento de los datos.

Avance de proyecto 2 (26 de mayo): no tenemos acceso a la descripción del avance sin embargo, el proyecto requerirá seguir trabajando con la definición y pre-procesamiento de datos

Avance de proyecto 3 (9 de junio): no tenemos acceso a la descripción del avance, planeamos que en este proyecto se trabaje más en el modelado o sistema de recomendación.

Avance de proyecto 4 (16 de junio): no tenemos acceso a la descripción del avance, sin embargo muy probablemente se trabajará en la evaluación de la solución y métricas de desempeño.

Entrega final de proyecto (23 de junio): no tenemos acceso a la descripción del avance, sin embargo se planea que en esta fase se trabaje en el despliegue de la solución a un sistema de nube.

2. Justificación del conjunto de datos a utilizar

Para el desarrollo de este proyecto se utilizará el dataset de películas IMDB que contiene 1000 registros de las mejores películas y shows de televisión (Kaggle, 2021). Se ha seleccionado este conjunto de datos debido a que es de acceso público y tiene licencia CC0 (Creative Commons, s.f.) que permite copiar, modificar, distribuir o realizar trabajo incluso para fines comerciales sin necesidad de solicitar permiso.

Adicionalmente a los beneficios legales y de disponibilidad del dataset, se ha seleccionado este conjunto de datos pues contiene metadatos que pueden ser de utilidad en el desarrollo del algoritmo:

- Poster del link: pudiera ser utilizado en algún algoritmo de clusterización por medio de reconocimiento de imágen.
- Título de la película.
- Año en que se estrenó.
- Certificado ganado por la película.
- Duración.
- Género.
- Calificación de la película en el sitio IMDB.

- Resúmen corto de la película: puede ser utilizado en algoritmos de recomendación utilizando técnicas de procesamiento de lenguaje natural como TF-IDF.
- Calificación ganada de la película.
- Nombre del director: puede ser utilizado como característica de asociación.
- Nombres de los protagonistas.
- Número de votos
- Ganancia generada por la película.

La variedad de tipos de datos disponibles nos da la posibilidad de trabajar con algoritmos diversos de recomendación, desde sistemas de reconocimiento de imágen hasta procesamiento de lenguaje natural o sistemas tradicionales de agrupación para características binarizadas.

En conclusión se selecciona este conjunto de datos por su disponibilidad pública, permisos amplios para trabajar con los datos y contenido rico y variado para desarrollar soluciones creativas de recomendación.

2.1 Pasos de preprocesamiento

Despues de realizar la exploración de los datos se realizarán los siguientes pre-procesamientos:

- La columna Gross, cantidad de moneda inicialmente codificada como cadena de caracteres separados por coma. Será convertida a cantidad numérica, removiendo las comas y convirtiéndo se a tipo flotante.
- La columna Runtime tenía los datos en formato “XXX min”” para indicar la cantidad de minutos de duración de la película. Se utilizó una expresión regular para extraer la cantidad de minutos como número entero.

Después de la exploración de los datos se encontraron valores faltantes en las variables Meta_score, Gross y Certificate.

- Se realizó imputación de valores faltantes en Meta_score utilizando la media de calificación.
- Se realizó la imputación de valores faltantes para Gross (ganancias de la película) reemplazando con la mediana. Se determinó utilizar la mediana y no la media aritmética debido a la distribución fuertemente sesgada a la derecha encontrada en el análisis exploratorio. Aunque se puede argumentar que la proporción de valores faltantes (16%) es relativamente elevada, se considera una variable importante para el sistema de recomendación y por ello se ha conservado.

- Certificate, debido a que no se piensa utilizar en este momento, no se aplicó una imputación específica, sin embargo si fuese necesario se hará más adelante.

Finalmente, debemos convertir las variables categóricas en representaciones numéricas para su utilización en algoritmos de aprendizaje automático.

- Para el género de las películas, debido a su alta cardinalidad encontrada en el EDA (mayor a 200), se ha determinado utilizar codificación binaria.
- Lo mismo sucede con la codificación de director, estrellas y año de publicación, se utilizará codificación binaria debido a la alta cardinalidad. Año de publicación se considera categórica no ordinal pues un año más actual no implica una relación de mejoría o mayor importancia que un menor número.
- En el caso de certificado, se utilizará codificación OneHot, pues su cardinalidad no es elevada (menor de 100), y por lo tanto no implica agregar tantas columnas adicionales.

Al final del documento se propone un ejemplo de algoritmo de recomendación basado en contenido. Para este caso en específico se aplican técnicas del área de NLP. En específico el preprocesamiento aplicado fue:

- Se concatenaron textos contenidos en el dataset, tal como título de la película, género, director y actores principales.
- Se removió todo carácter no alfabético, como signos de puntuación, caracteres especiales y números.
- Se quitaron todas las palabras formadas por una sola letra o caracter.
- Todas las palabras se convirtieron en minúsculas.
- Se eliminaron del texto todas las palabras llamadas stopwords, que representan palabras comunes de un idioma específico, en este caso inglés, o palabras que no aportan demasiada información. Por ejemplo: “the”, “my”, “our”, “it”, etc. En este caso se utilizó el diccionario de stopwords de inglés de la librería NLTK.

3. Exploración inicial y análisis del conjunto de datos:

La exploración de los datos se realizó en el siguiente Jupyter Notebook disponible en github del equipo:

https://github.com/moisesdiazm/sistema-recomendacion-bigdata-mna/blob/main/Proyecto_Avance_1_37.ipynb

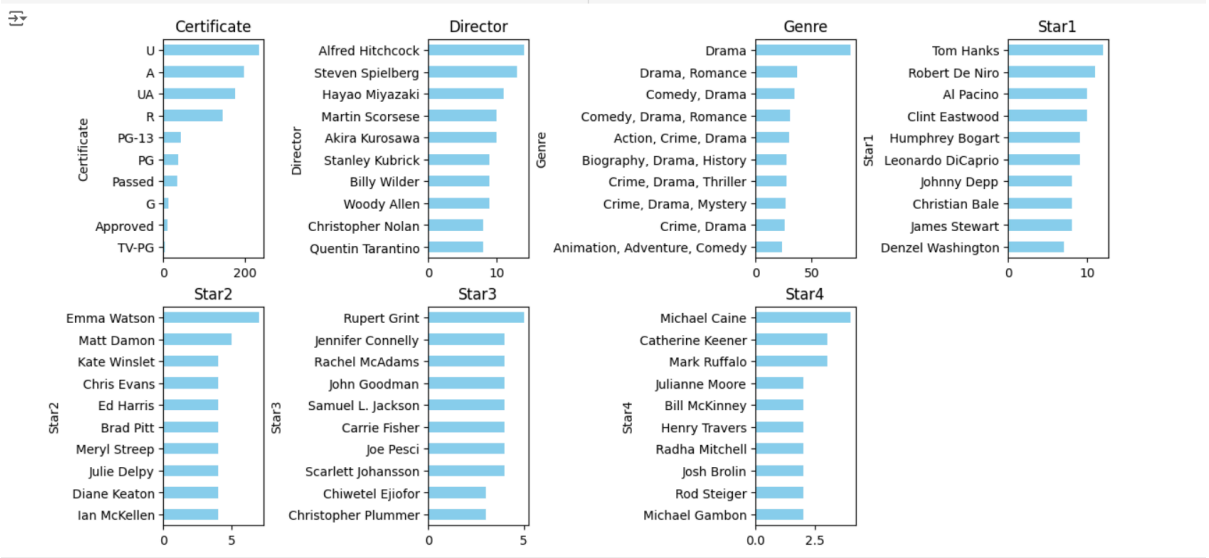


Fig 2. Frecuencia de variables categóricas

En este documento se incluye la exploración completa del conjunto de datos, para variables categóricas y numéricas. Así como el preprocesamiento antes indicado. Las observaciones extensivas de los datos se escribieron en el Jupyter Notebook. Se adjunta la Figura 2 como un ejemplo de gráfica obtenida a partir del análisis extensivo realizado, en este caso muestra las frecuencias de variables categóricas. Más gráficas así como su interpretación se encuentran en el jupyter notebook correspondiente.

4. Experimentación con al menos 1 algoritmo de recomendación básico.

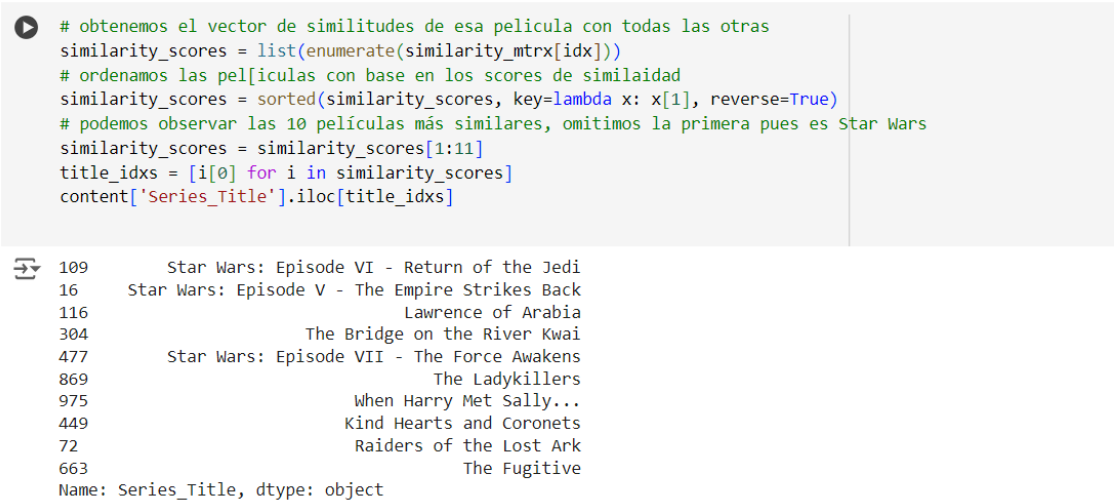
A modo de aproximación inicial se desarrolló un ejemplo básico de sistema de recomendación basado en contenido. Recordemos que este tipo de sistema analiza las características de los productos. De acuerdo con Ruy y Dutta (2022) estos sistemas requieren que los productos estén asociados en grupos de acuerdo con sus características, en este caso en el ejemplo que escribimos están asociados por género, director, actores principales y título. Las agrupaciones en este caso suceden en un espacio vectorial de alta dimensionalidad. Para ello las características de las películas antes mencionadas fueron concatenadas en texto plano y posteriormente convertidas en vectores mediante la técnica TF-IDF (ver la explicación a detalle en el Jupyter Notebook). Cuando una persona califica como positivo una película, las películas similares, calculadas mediante la similitud de coseno entre vectores de todas las películas, son recomendadas. Es necesario para un sistema como este que los artículos tengan suficiente información que los describa, en este caso teníamos riqueza de información en texto plano pero también se podrían hacer técnicas híbridas que consideren los factores numéricos y categóricos mencionados

anteriormente. Las recomendaciones en este sistema tienen la ventaja de ser independientes entre usuarios, otorgando privacidad y seguridad. También puede recomendar productos nuevos, es decir es robusto ante un cold-start.

La evidencia y explicaciones detalladas de la implementación se encuentran en el jupyter notebook en el github del equipo:

https://github.com/moisesdiazm/sistema-recomendacion-bigdata-mna/blob/main/Proyecto_Avance_1_37.ipynb

En la Figura 3 extraída del notebook se puede observar como el sistema le recomienda al usuario películas parecidas en contenido (de acuerdo con la descripción, título, actores, director, etc.) con base en una que película que el usuario ha calificado de forma positiva, en este caso “Star Wars”.



```
# obtenemos el vector de similitudes de esa pelicula con todas las otras
similarity_scores = list(enumerate(similarity_mtrx[idx]))
# ordenamos las pel[iculas con base en los scores de similaidad
similarity_scores = sorted(similarity_scores, key=lambda x: x[1], reverse=True)
# podemos observar las 10 películas más similares, omitimos la primera pues es Star Wars
similarity_scores = similarity_scores[1:11]
title_idxs = [i[0] for i in similarity_scores]
content['Series_Title'].iloc[title_idxs]
```

```
109       Star Wars: Episode VI - Return of the Jedi
16     Star Wars: Episode V - The Empire Strikes Back
116                              Lawrence of Arabia
304                    The Bridge on the River Kwai
477       Star Wars: Episode VII - The Force Awakens
869                                 The Ladykillers
975                          When Harry Met Sally...
449                         Kind Hearts and Coronets
72                           Raiders of the Lost Ark
663                                  The Fugitive
Name: Series_Title, dtype: object
```

Fig 3. Sistema de recomendación en el Jupyter Notebook

Conclusiones:

El proyecto tiene como objetivo desarrollar un sistema de recomendación de películas para la plataforma "Genesis", basado en las preferencias del usuario. Para facilitar las decisiones de los usuarios, se utilizará un dataset de IMDB con 1000 registros de películas y shows de televisión, que es de acceso público y tiene licencia CC0 lo que lo vuelve utilizable para desarrollar nuevos algoritmos incluso para sistemas comerciales.

Inicialmente, el sistema de recomendación se basará en el contenido de las películas, con planes para mejoras futuras. Se definirá una métrica específica para evaluar el sistema en etapas posteriores, y será accesible a través de una API web.

El proyecto se divide en varias etapas, abarcando la definición de la base de datos, preprocesamiento, desarrollo del modelo, evaluación y despliegue del sistema.

El dataset de IMDB fue seleccionado por su disponibilidad pública y riqueza en metadatos, que incluyen títulos, géneros, directores, actores y calificaciones. El preprocesamiento de datos incluye la conversión de valores numéricos, manejo de valores faltantes y codificación de variables categóricas.

La implementación y los resultados se detallan en un Jupyter Notebook que se encuentra alojado en GitHub, donde se describen la exploración de datos, preprocesamiento y desarrollo del algoritmo. Este sistema mejora la experiencia del usuario, al sugerirle películas similares a las que ha calificado positivamente o visto.

Referencias:

Creative Commons. (s.f.). *CC0 1.0 DEED, CC0 1.0 Universal.* Recuperado de: https://creativecommons.org/publicdomain/zero/1.0/

Kaggle. (2021). *IMDB Movies Dataset*. Recuperado de: https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

Roy, D., y Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9*(1), 59. Recuperado de: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00592-5