Análisis de Grandes Volúmenes de Datos (Gpo 10)

David Nava Jiménez - A01168501

Edwin David Hernández Alejandre - A01794692

Jorge Fernando Bonilla Diaz - A01793935

Nombre del entregable:

Avance de proyecto 1: Sistema de Recomendación

17 de Mayo de 2024

Contenido

1. Objetivo.

Identificar las diferentes técnicas y herramientas para el manejo de tareas de procesamiento de datos a gran escala en el contexto de un sistema de recomendación utilizando un dataset de Netflix

2. Objetivos específicos.

- Generar un plan de proyecto de acuerdo con la industria elegida para la actividad 2.2 y detallar el plan del proyecto con su cronograma.
- Justificar la selección del conjunto de datos utilizado y describer los pasos de preprocesamiento.
- Realizar al menos un ejercicio de exploración inicial y análisis del conjunto de datos de la industria elegida (la evidencia se debe pone en el repositorio GitHub del equipo).
- Programar al menos un 1 algoritmo de recomendación básico con el conjunto de datos elegido (la evidencia se debe pone en el repositorio GitHub del equipo).

3. Desarrollo.

3.1 Plan de Proyecto.

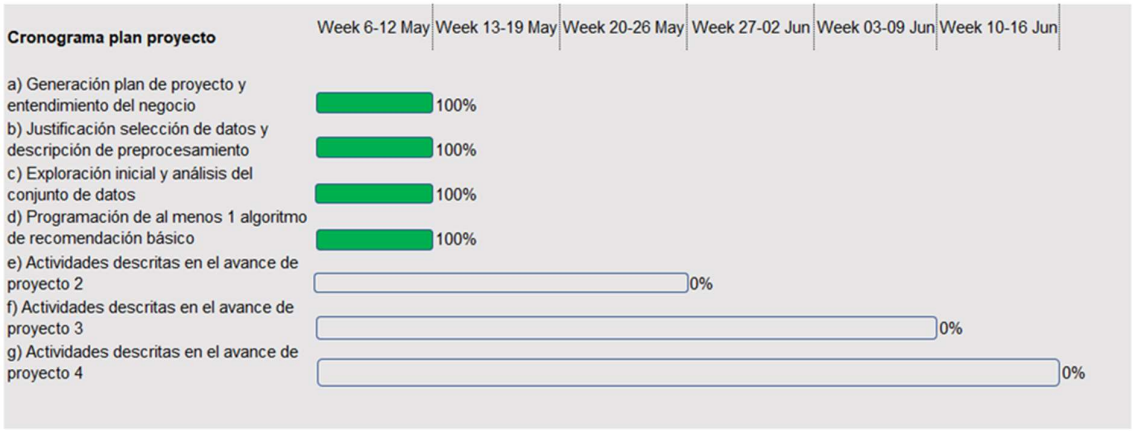
El proyecto que estaremos desarrollando será un sistema de recomendación inteligente, utilizando modelos de aprendizaje automático supervisado, modelos de descomposición de valores singulares (SVD) y usando medidas de similitud como Similitud Euclídea, Coseno o Jaccard. Los cuales nos ayudarán a realizar recomendaciones de películas.

El plan de proyecto se basará en la metodología CRISP-DM (Hotz,2024), la cual consiste en las siguientes fases:

1. Entendimiento del negocio. ¿Qué es lo que necesita el negocio?
2. Entendimiento de los datos. ¿Con qué datos contamos? ¿Los datos están limpios?
3. Preparación de los datos. ¿Cómo organizamos los datos para construir el modelo?
4. ¿Qué técnicas de modelado aplicaremos?
5. Evaluación. ¿Cuál es el mejor modelo que se adapta a los objetivos del negocio?
6. ¿Cómo nuestros stakeholders acceden a los resultados?

3.2 Cronograma de actividades.

El **cronograma** se irá actualizando con las actividades de cada entrega, una vez se tenga visibilidad de estas. Por el momento el cronograma hoy en día es el siguiente:



3.3 Entendimiento del negocio.

Como antecedentes, se realizó la investigación de Netflix y como su modelo de negocio ha revolucionado la industria del entretenimiento en la era digital que vivimos actualmente.

Netflix (de acuerdo con su sitio web) fue fundada en 1997 en California, por Reed Hastings y Marc Randolph, donde generaron la idea para alquilar un DVD por correo, lo que dio pie a la innovación del servicio que hoy conocemos de Netflix, siendo este un sitio para alquilar mediante una suscripción películas y series ilimitadas sin fecha de entrega o penalizaciones por retrasos a cambio de una suscripción mensual.

Posteriormente en el año 2000, realizaron la presentación de un sistema personalizado de recomendación de películas que usa las valoraciones de los suscriptores de títulos anteriores para predecir sus futuras elecciones, en el año 2005 lanzan la función perfiles, que permite a los suscriptores crear listas para distintos usuarios o estados de ánimo

El negocio necesita herramientas de toma de decisiones basadas en datos. Para seguir siendo sostenible, el negocio del streaming de vídeo debe seguir innovando en sistemas de big data para identificar patrones de consumo de películas, lo que conducirá a una mayor satisfacción del cliente, mayores ingresos y una mayor rentabilidad.

3.4 Descripción de datos.

3.4.1 Fuente de datos.

Se va a utilizar una base de datos de Kaggle denominada “The movies dataset”, la cual ocuparemos para el proyecto. Esta Base de Datos, contiene metadatos de las 45.000 películas incluidas en el conjunto de datos completo de MovieLens. El conjunto de datos incluye películas estrenadas en julio de 2017 o antes. Los datos incluyen el reparto, el equipo, las palabras clave del argumento, el presupuesto, los ingresos, los carteles, las fechas de estreno, los idiomas, las productoras, los países, los recuentos de votos de TMDB y los promedios de votos. Este conjunto de datos también contiene archivos con 26 millones de valoraciones de 270.000 usuarios para las 45.000 películas. Las valoraciones están en una escala del 1 al 5 y se han obtenido de la página oficial de GroupLens.

3.4.2 Justificación de selección de dataset.

Para el desarrollo de un sistema de recomendación de películas efectivo y robusto, es fundamental contar con un dataset conformado por un volumen importante de información y representativo de las interacciones reales de los usuarios con los contenidos de la plataforma. En este contexto, hemos seleccionado Them ovies dataset debido a las siguientes razones clave:

Representatividad y Realismo

El Netflix Movies Dataset contiene datos reales de usuarios de Netflix, una de las plataformas de streaming más populares y con una amplia diversidad de usuarios a nivel global. Esto asegura que el sistema de recomendación se entrene con datos que reflejan patrones de comportamiento y preferencias auténticas, lo que aumenta la validez externa del modelo desarrollado.

Complejidad y Riqueza de Datos

Este conjunto de datos contiene varias variables relacionadas con el desarrollo de un sistema de recomendación, incluidas identificaciones de usuarios y películas, calificaciones de usuarios y más. También incluye información sobre géneros cinematográficos, lo que permite implementar y evaluar diferentes enfoques de recomendación basados en contenidos y filtrado colaborativo.

Escalabilidad

El gran volumen de datos del conjunto de datos de películas de Netflix permite entrenar modelos a gran escala, evaluar el rendimiento y la escalabilidad y permitir un análisis detallado para mejorar la personalización de las recomendaciones.

Evaluación Comparativa

El conjunto de datos de películas de Netflix se utiliza ampliamente en investigaciones académicas y competencias de ciencia de datos para comparar resultados y evaluar nuevas metodologías, lo que permite una medición efectiva de nuestro sistema de recomendación en un contexto comparativo.

Disponibilidad y Acceso

El conjunto de datos de películas de Netflix es de fácil acceso y está bien documentado, lo que fomenta la reproducibilidad de los resultados, las colaboraciones y las mejoras continuas.

3.4.3 Estructura del dataset.

El dataset está compuesto por varios archivos que contienen la siguiente información clave:

- **movies_metadata.csv:** El archivo principal de metadatos de películas. Contiene información sobre 45.000 películas incluidas en el conjunto de datos MovieLens. Incluye carteles, fondos, presupuesto, ingresos, fechas de estreno, idiomas, países de producción y empresas.
- **keywords.csv:** Contiene las palabras clave de los argumentos de nuestras películas MovieLens. Disponible en forma de objeto JSON.

- **credits.csv:** Contiene información sobre el reparto y el equipo de todas nuestras películas. Disponible en forma de objeto JSON encadenado.
- **links.csv:** El archivo que contiene los ID de TMDB e IMDB de todas las películas que aparecen en el conjunto de datos de Full MovieLens.
- **Links_small.csv:** Contiene los ID de TMDB e IMDB de un pequeño subconjunto de 9.000 películas del conjunto de datos completo.
- **ratings_small.csv:** El subconjunto de 100.000 valoraciones de 700 usuarios sobre 9.000 películas.

3.4.4 Descripción de los campos.

Para este primer ejercicio, ocuparemos el archivo denominado movies_metadata.csv, el cual contiene las siguientes columnas:

1. Adult (varchar): Es un valor binario que define si la película es +18 o no.
2. belongs_to_collection(varchar): Define si la película pertenece a una serie de películas,
3. budget(big int): Define el presupuesto en USD para la película
4. genres(varchar): Es un valor tipo lista que dicta la(s) categoría(s) que contiene la película
5. homepage(varchar): Pagina web de la película
6. id(int): Valor indice que asigno MovieLens
7. imdb_id(varchar): Valor índice que relaciona esta tabla con links y links_small
8. original_language(varchar): Lenguaje de la película.
9. original_title(varchar): Nombre de la película oficial
10. overview(varchar): Descripcion de la película
11. popularity(float): Indice de popularidad de la película
12. poster_path(varchar): Nombre del archivo que contiene la imagen del poster de la película
13. production_companies(varchar): Valor tipo lista que contiene el nombre de las compañías que produjeron la película.
14. production_countries(varchar): Valor tipo lista que contiene el nombre de los países donde se produjo la película.
15. release_date(date): Fecha en la cual se realizó el estreno de la película
16. revenue(big int): Valor en USD de lo recaudado por la película
17. runtime: N/A
18. spoken_languages(varchar): Valor tipo lista que contiene los idiomas que se hablan en la película:
19. status(varchar): Estatus de la película
20. tagline(varchar): Eslogan de la película
21. title(varchar): Nombre de la película
22. video(varchar): Valor binario que dicta si la película tuvo un trailer.
23. vote_average(float): Promedio de calificacion de la pelicula
24. vote_count(int): Numero de calificaciones que tuvo la pelicula

3.5. Exploración y análisis de los datos.

El EDA se puede encontrar a detalle en la siguiente Notebook:

https://github.com/drekkel/TC4034.10/blob/main/RecommendationSystem/Notebooks/Proyecto_Avance_1.ipynb

3.6 Implementación de un algoritmo básico de recomendación.

La implementación se puede encontrar a detalle en la siguiente Notebook:
https://github.com/drekkel/TC4034.10/blob/main/RecommendationSystem/Notebooks/Proyecto_Avance_1.ipynb

3.7. Siguientes pasos.

Los pasos de preprocesamiento que se seguirán son los siguientes (realizado en Jupyter-Notebook):

1. Identificar si existen valores faltantes en el conjunto de datos.
2. Identificar si existe suficiente cardinalidad de las variables independientes.
3. Para las variables numéricas visualizar la distribución con histogramas y las correlaciones.
4. Se van a escalar las variables numéricas para posteriormente determinar si es viable su transformación.
5. Identificar si es viable la codificación en las variables categóricas.
6. Visualizar las gráficas de variables con las transformaciones antes aplicadas.

Repositorio GitHub privado por equipo con la información de los integrantes del equipo:
https://github.com/drekkel/TC4034.10

4. Conclusiones.

En la primera etapa del desarrollo de nuestro sistema de recomendación de películas, hemos establecido una base sólida para crear un modelo eficaz y sólido. Comenzamos identificando claramente el problema que queremos resolver: mejorar la experiencia del usuario en una plataforma de streaming brindándole recomendaciones personalizadas basadas en sus preferencias y comportamientos.

La selección del conjunto de datos del premio Netflix se justifica en gran medida por su representatividad, riqueza y disponibilidad. Este conjunto de datos no solo ofrece una visión realista de las interacciones entre el usuario y la película, sino que también proporciona una estructura de datos ideal para implementar métodos de filtrado colaborativo y enfoques basados en contenido.

Proporcionamos una descripción detallada de los datos, destacando las características clave de nuestro conjunto de datos, incluidas las variables disponibles y su relevancia para nuestro objetivo. También abordamos las etapas iniciales del preprocesamiento de datos para garantizar que nuestro conjunto de datos esté limpio y preparado para los siguientes pasos en el desarrollo del sistema de recomendación.

En resumen, esta etapa inicial ha sentado las bases necesarias para avanzar hacia la implementación y evaluación de algoritmos de recomendación. Con una comprensión clara del problema, una selección de datos adecuada y una preparación meticulosa de los datos, estamos bien posicionados

para pasar con confianza a las siguientes fases del proyecto, que incluirán el modelado, la validación y la optimización de nuestro sistema de recomendación.

5. Referencias.

- Vergnou, B. (2021). Spotify Recommendation. Kaggle. Retrieved 05 07, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data.
- Hotz, N. (28 de April de 2024). Data Science Process Alliance- What is CRISP DM. Recuperado el 01 de May de 2024, de https://www.datascience-pm.com/crisp-dm-2/
- Sébastien Ronteau, L. M. (2023). Beyond Digital Ubiquity: The Digital Business Model Iron Triangle. In L. M. Sébastien Ronteau, Digital Business Models: The New Value Creation and Capture Mechanisims of the 21st Century (pp. 1-10). Berlin: De Gruyter. Retrieved from https://0-eds-p-ebscohost-com.biblioteca-ils.tec.mx/eds/ebookviewer/ebook/bmxlYmtfXzM0NzI2MzRfX0FO0?sid=43ed574d-f0dc-473c-beab-6fed1e37ae23@redis&vid=1&hid=/&format=EB