Entrega 1

Visualización de Información – IIC2026

27 de octubre 2020

Dan Ustilovsky Stifel – 17642396

1. Abstracción de tareas

Luego de analizar el contexto entregado en el enunciado, pude distinguir con claridad dos problemas:

A. Discusión basada en la falta de conocimiento preciso sobre que sexo biológico predomina en los protagonistas de las películas biográficas

Esto se obtiene de la frase:

"… algunos integrantes del grupo creen que la cantidad de películas con protagonistas de sexo masculino es mucho mayor a aquellas con protagonistas de sexo femenino de forma consistente, …"

De esta frase se extra la siguiente tarea: Descubrir (valor individual de) cantidad de personajes de sexo biológico masculino y femenino en las películas biográficas.

La elección de la tarea se basa en que los amigos tienen una creencia sobre cómo sería la comparación entre la cantidad de personajes de cada sexo en la película, por lo que existe una teoría o hipotesis previa, pero no existe conocimiento sustentado de aquello y por ende la acción seria descubrir (Aprender información que antes no era conocida). Por otra parte, dado que la cantidad de personajes de cada sexo es un atributo dado que se puede medir (contar), podemos plantear como objetivo encontrar el valor individual de cada una de esas sumas, una para cada sexo, para así compararlas y poder obtener información a partir de ellas.

En el caso de este par, el único dato del dataset que resulta necesario es la columna "subject_sex".

B. Discusión sobre la evolución de la desigualdad entre la cantidad de películas biográficas con personajes de sexo masculino y femenino.

Lo anterior se desprende de la siguiente frase:

"… otros creen que a medida que pasan los años, esa desigualdad ha ido disminuyendo de forma gradual."

De la frase se decidió seleccionar la siguiente tarea: Comparar la tendencia de la cantidad de películas biográficas con personajes protagonistas a través de los años.

La justificación de la tarea anterior corresponde al hecho que lo que se desea realizar no solo corresponde a descubrir información, si no que se deben analizar la evolución de la proporción entre personajes protagónicos de cada sexo a través del tiempo, lo que supone mas que un simple descubrimiento si no que se trata de agrupar varios objetivos (la cantidad de personajes por sexo y la tendencia). En conjunto con lo anterior, se plantean dos objetivos dado que se deben obtener las cantidades individuales de personajes de cada sexo en cada año y en conjunto analizar la tendencia de las diferentes cantidades a medida que se avanzan los años.

Para realizar lo anterior se necesitan las columnas “year_released” y “subject_sex” del dataset.

2. Abstracción de datos

- Title
    - Categórico, dado que los títulos son frases y por ende no tienen un orden intrínseco.
    - Ordinal, dado que el titulo al ser una frase no tiene comparación aritmética exacta.
    - Valor, dado que dos películas pueden tener el mismo titulo por lo que no permite identificar una película.
- Site
    - Categórico, dado que un URL a una pagina no tiene un orden intrínseco.
    - Ordinal, dado que los URL no pueden ser comparados mediante aritmética exacta.
    - Llave, cada película tiene una única pagina en IMDB, por lo que existe un único link a aquella página, lo que permite identificar una película.
- Country
    - Categórico, dado que un país es un nombre y por ende no tiene un orden intrínseco.
    - Ordinal, un nombre al ser una palabra no tienen comparación mediante aritmética exacta.
    - Valor, dado que varias películas pueden haber sido realizadas en un mismo país, por lo que no permite determinar la película.
- Year_released
    - Ordenado, los años al ser números cuentan con orden intrínseco y numérico.
    - Cuantitativo, los años si cuentan con operaciones aritméticas exactas, la diferencia de años permite apreciar cuanto tiempo ha pasado.
    - Secuencial, los años solo pueden ser valores positivos, pese a que existan años antes/después de cristo, solo son convenciones, pero los años siguen siendo solo positivos.
    - No cíclico, su valor puede aumentar indefinidamente.
    - Valor, mas de una película puede haber sido estrenado un mismo año por lo que no permite identificar una película.
- Box_office
    - Ordenado, dado que un numero como la cantidad de dinero ganada si tiene un orden intrínseco y numérico.
    - Cuantitativo, la cantidad de dinero recaudada si cuenta con operaciones matemáticas exacta como la diferencia, que cuando mas dinero recaudo una película que otra.

    - Secuencial, el dinero recaudado solo puede ser positivo y no existe un valor 0 que divida su rango.
    - No cíclico, ya que la cantidad de dinero podría aumentar indefinidamente.
    - Valor, dado que mas de una película pueden tener el mismo Box office y por ende no permite determinar la película.
- Director
    - Categórico, dado que los nombres no poseen un orden intrínseco.
    - Ordinal, dado que los nombres no cuentan con operaciones aritméticas exactas.
    - Valor, dado que un director probablemente haya dirigido más de una película y por ende no se puede definir.
- Subject
    - Categórico, dado que los nombres no poseen un orden intrínseco.
    - Ordinal, dado que los nombres no cuentan con operaciones aritméticas exactas.
    - Valor, dado que podría haber una más de una película sobre la misma persona y por ende este no permite identificar la película.
- Type_of_subject
    - Categórico, dado que es una frase y por ende no tienen un orden intrínseco.
    - Ordinal, dado que una frase no tiene comparación aritmética exacta.
    - Valor, dado que dos películas pueden tratar sobre personas que son conocidas por lo mismo y por ende no se puede asegurar que defina la película.
- Race_known
    - Categórico, dado que la capacidad de clasificar algo no tienen un orden intrínseco.
    - Ordinal, dado que la información acerca de si se puede determinar la etnicidad de un personaje no tienen comparaciones aritméticas exactas.
    - Valor, dado que más de una persona puede tener o no tener una etnicidad determinable y por ende no asegura definir la película.
- Subject_race
    - Categórico, dado que las etnicidades no tienen un orden intrínseco.
    - Ordinal, dado que las etnicidades no tienen comparaciones aritméticas exactas.
    - Valor, dado que más de una persona puede tener la misma etnicidad y por ende no asegura definir la película.
- Person_of_color
    - Categórico, dado que el color de piel de las personas no tiene un orden intrínseco.
    - Ordinal, dado que el color de piel de las personas no tiene no tienen comparaciones aritméticas exactas.
    - Valor, dado que más de un personaje puede tener piel de color y por ende no asegura definir la película.
- Subject_sex
    - Categórico, dado que el sexo biológico no presenta un orden intrínseco.
    - Ordinal, dado que el sexo biológico no tiene comparación aritmética exacta.
    - Valor, dado que mas de un personaje puede tener el mismo sexo biológico.
- Lead_actor_actress
    - Categórico, dado que un nombre no tiene un orden intrínseco.
    - Ordinal, dado que los nombres no tienen una comparación aritmética exacta.

- Valor, dado que un actor/actriz puede interpretar el rol principal en mas de una película, y por lo tanto no define la película.

¿De qué tipo de dataset se trata biopics.csv?

El tipo es tablas o tabulares, dado que esta en formato csv, el cual es un típico formato para archivos representados por tablas, los cuales se componen de filas y columnas

¿Es un dataset estático o dinámico?

Es un dataset dinámico, dado que contamos con el dataset completo antes de partir a trabajar. Sin embargo, las películas salen constantemente y en caso de que fuéramos actualizando el dataset a medida que esto sucede seria dinámico.

¿Qué tipos de datos singulares hay presentes en el dataset? (Atributo, ítem, enlace, posición o grilla)

Los tipos de datos presentes son:

- Atributo: Esto representa a cada fila de la tabla, dado que cada película es una entidad individual y compleja.
- Ítem: cada columna en la tabla representa un ítem, dado que son elementos que se pueden medir o registrar y que por ende permiten tener su valor en una casilla.