Ciencia de Datos e Inteligencia de Negocios

Tarea 2. Medidas de Similitud

Habiendo hecho un estudio preliminar de la calidad de los datos. Normalmente conviene definir el tipo de dato de cada variable. La clasificación de los datos, normalmente se hace en las siguientes categorías: Variables Binarias o de doble estado, variables cualitativas o multiestado, variables cuantitativas y variables genéticas.

En clase se revisaron con varios ejemplos de los índices de similitud de datos binarios, la obtención de los índices de similitud más comunes y su forma de aplicación a bases de datos.

Del archivo de las películas en donde se lograron hacer comparaciones entre los alumnos del curso en función de que si les gusto la película o no les gusto (0 ó 1 basados en un umbral.)

Los datos reales de las películas están calificadas en una escala de número de estrellas y se estableció un umbral para convertirlos a binarios o doble estado. Esto puede sesgar el resultado real de las recomendaciones, y también puede estar afectado por el índice de similitud que se esté utilizando. Considerando lo anterior, se propone la presente tarea.

La realización de esta tarea considera las siguientes actividades:

1. Del ejercicio de clase donde se hacia el mini recomendador de películas, elegir un índice de similitud y un umbral deseado para generar las recomendaciones de las películas para ustedes. Es decir, ustedes son el usuario a quien le recomendaran películas y se tienen que comparar con todos los demás usuarios. Capturar los resultados de las dos versiones de recomendador.
2. Del archivo de las películas obtener las variables dummy o auxiliares para convertir las calificaciones de las películas en variables dummy. (No se tiene que establecer un umbral).
3. Repetir las dos versiones de los recomendadores pero ahora usando el mismo índice de similitud del punto 1.
4. Capturar y comparar los resultados de las dos versiones de recomendadores (el primero es convirtiendo las variables a valores binarios y el segundo es considerando las variables como multiestado y usando variables dummy).
5. Entregar un pequeño reporte donde se muestren los resultados de la comparación y comentarios sobre si mejoró o no mejoró la recomendación.