Cápsula 3: Reducción de dimensionalidad

Hola, bienvenidxs a una cápsula del curso Visualización de Información. En esta hablaré de una última forma de agregación: la reducción de dimensionalidad.

La reducción de dimensionalidad es un caso especial de reducción mediante agregación de atributos. En su caso más general, esto significa combinar múltiples atributos de un *dataset* y reemplazarlos por un atributo representante. Una forma simple de hacerlo es agrupando atributos que sean similares, y calcular un atributo derivado de ambos atributos, como el promedio.

Una forma más compleja de agregación de atributos es la reducción de dimensionalidad. Donde el objetivo es preservar el significado estructural del conjunto de datos utilizando menos atributos para representar los ítems.

El término reducción de dimensionalidad es común en áreas de estadística y minería de datos. Si bien usa la palabra reducción que podría relacionarse con la idea de filtración antes presentada, en realidad por lo general hace referencia a agregación de atributos en otros.

Más específicamente, la reducción de dimensionalidad como acción busca transformar datos multidimensionales a datos de pocas dimensiones. Podemos pensarlo de forma geométrica donde nuestros ítems viven en un espacio multidimensional determinado por los atributos que contamos para describirlos, y que se desean llevar a dos dimensiones o tres dimensiones, pero sin perder tanta información que diferencie y describa los ítems originales.

No es lo mismo que simplemente proyectar ítems, ignorando ciertos atributos. Para efectos de visualización, es común usar un método de reducción de dimensionalidad que permita mostrar espacialmente y en dos dimensiones un *dataset*. De tal forma que ítems próximos realmente sean similares según sus todos sus atributos originales, y qué ítems lejanos sean realmente diferentes.

Hay muchos métodos computacionales-matemáticos para producir reducción de dimensionalidad. No iremos en detalle de esta naturaleza en este curso, pero es bueno saber que existen, y son un campo de estudio importante en *data science* y estadística.

A modo de ejemplo, un caso común de estudio es el análisis de colecciones de grandes documentos de texto. Comparar textos en una colección gigante de documentos no es trivial, ya que no es suficiente concentrarnos en un par de atributos simples para comparar.

Un documento es un ítem complejo. Suelen representarse como bolsas de palabras, que son cuentas de las palabras que presenta el documento. Pensándolo así, contamos con un valor

cuantitativo para cada palabra solo para un texto. ¡Pero la cantidad de palabras es enorme! ¡Entonces son muchos atributos! Por eso aplicar reducción de dimensionalidad en este caso y traer la cantidad de atributos a algo manejable es muy apropiado.

En pantalla se muestra un resultado de un procesamiento así. Es un gráfico de dispersión que muestra miles de documentos desplegados por dos atributos generados por algún método de reducción de dimensionalidad en base a sus bolsas de palabras.

Los colores utilizados codifican los resultados de un proceso de clustering que agrupa documentos por la similaridad de los atributos calculados. Se logra apreciar que las posiciones visuales otorgadas están de acuerdo con este agrupamiento, en la mayor parte.

Cuando el número de dimensiones generadas en una reducción es dos, usualmente los valores se representan mediante un gráfico de dispersión. Cuando más atributos surgen, una matriz de gráficos de dispersión es buena opción. Han habido estudios que indican que estos dos *idioms* son las representaciones más seguras para inspeccionar datos reducidos dimensionalmente.

En este contexto no se usa el gráfico de dispersión con la intención de encontrar correlación, sino que con la intención de hallar cómo distribuyen y qué agrupamientos se generan en esta transformación.

Otras tareas que se pueden vincular son la poder revisar la información de alta dimensionalidad original de cada dato, probablemente mediante interacción, y así apreciar si efectivamente hay una correspondencia en menos dimensiones. Esto mismo realiza la herramienta Moodplay, que distribuye espacialmente artistas a través de reducción de dimensionalidad, y permite ver el detalle de los artistas desplegados mediante interacción.

Con eso termina el contenido de esta cápsula. Recuerda que si tienes preguntas, puedes dejarlas en los comentarios del video para responderlas en la sesión en vivo de esta temática. ¡Chao!