**Sobre el set de datos**

**Introducción**

El continuo crecimiento de la industria del comercio electrónico ha hecho que el problema de la recuperación de productos sea particularmente importante. A medida que más empresas trasladan sus actividades a la Web, el volumen y la diversidad de la información relacionada con los productos aumenta rápidamente. Estos factores dificultan que los usuarios identifiquen y comparen las características de sus productos deseados. El agrupamiento, la clasificación y la coincidencia de productos son algoritmos útiles que pueden contribuir a la organización de la información relacionada con el producto y, en consecuencia, mejorar la eficacia de la recuperación.

Este repositorio está diseñado para proporcionar múltiples conjuntos de datos que son adecuados para dichos algoritmos. Cada conjunto de datos va acompañado de su correspondiente archivo de verdad del terreno que se puede utilizar con fines de evaluación.

**Contenido**

Este repositorio incluye 18 conjuntos de datos del mundo real de diferentes categorías de productos, adquiridos de dos plataformas de comparación de productos en línea: PriceRunner y Skroutz. En particular, rastreamos parcialmente estas dos plataformas y construimos 8 conjuntos de datos de cada una. Cada uno de estos 16 conjuntos de datos representa una categoría de producto específica. Las categorías fueron seleccionadas con dos criterios, con el fin de: i) estudiar la diferencia de rendimiento de los mismos métodos en productos similares proporcionados por diferentes proveedores, y ii) examinar la efectividad de los algoritmos en productos de diversas categorías. Por esta razón, incluimos productos de categorías idénticas y diferentes. Además, creamos un conjunto de datos agregado por plataforma que contiene todos los productos de las 8 categorías combinadas. Estos conjuntos de datos permiten el examen del rendimiento en conjuntos de datos heterogéneos.

Los conjuntos de datos se proporcionan en formatos CSV y XML estándar. Cada entrada CSV/XML incluye la siguiente información:

* identificación de producto
* Título del producto tal como aparece en la respectiva plataforma de comparación de productos (pero en minúsculas y sin puntuación)
* Identificación del proveedor: esta es la identificación de la tienda electrónica que proporciona (vende) el producto. La identificación del proveedor se puede usar con fines de refinamiento, como el algoritmo de verificación que desarrollamos en [1].
* ID de clúster: este es el ID del clúster al que pertenece el producto. Útil para tareas de agrupación y coincidencia de entidades.
* Etiqueta de clúster: El título del clúster antes mencionado.
* ID de categoría: es el ID de la categoría a la que pertenece el producto. Es significativo principalmente en los dos conjuntos de datos agregados que contienen productos de múltiples categorías. Útil para tareas de clasificación y categorización.
* Etiqueta de categoría: El título de la categoría antes mencionada.