**Agrupamiento:**

**Definición:**

El agrupamiento de datos es uno de los métodos de análisis de datos. Su objetivo es dividir un conjunto de datos en diferentes "paquetes" homogéneos, en el sentido de que los datos de cada subconjunto comparten características comunes, que a menudo corresponden a criterios de proximidad (similitud informática) que se definen mediante la introducción de medidas de distancia y clases entre objetos.

**Problema:**

Vamos a agrupar los términos presentes en el conjunto de datos de las reseñas de películas y para ello utilizaremos el algoritmo K-means para el agrupamiento.

**Pasos:**

Extraer palabras clave: la extracción de palabras clave se realiza con los siguientes parámetros:

**-Términos de cada documento:** 3

Este valor fue elegido por dos razones: primero, porque los textos de los comentarios de las películas son bastante cortos y contienen muchas palabras que son comprensibles en lenguaje natural, y segundo, para minimizar la complejidad en tiempo de ejecución.

**-La frecuencia de presencia del término:** 4

El término debe estar presente más de cuatro veces y por lo tanto en cuatro documentos para ser evaluado como un término clave.

**-Cálculo de la matriz de distancias:**

Se calculan los valores de distancias utilizando la distancia euclídea para todas las combinaciones de pares de filas de la tabla de entrada, que contiene los documentos y sus términos. El resultado se añade a la tabla de entrada como una sola columna que contiene los tamaños de los vectores de distancia.

**-Aplicación de K-Means:**

Los parámetros son:

Número de iteraciones: 99

Número de clusters: El valor fue variado de 2 a 3 clusters

**1/ Cuatro clusters:**

La distribución de los clusters se puede ver en el gráfico -1-.

Podemos ver la clasificación por clusters (0, 1, 2 o 3), donde los colores corresponden a los clusters asignados, así como la distancia euclídea y la puntuación de cada término.

**Traducción realizada con el traductor www.DeepL.com/Translator**

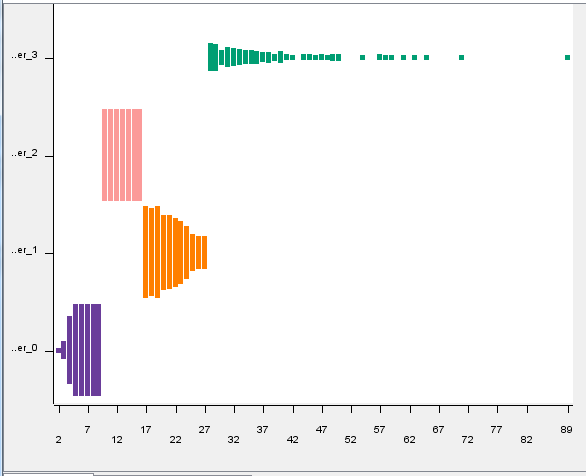


Gráfico -1- Distribución de los clusters (4 clusters con K-media)

**1/Tres clusters:**

La distribución de los clusters se puede ver en el gráfico -2-.

Podemos ver la clasificación por clusters (0, 1 o 2), donde los colores corresponden a los clusters asignados, así como la distancia euclídea y la puntuación de cada término.

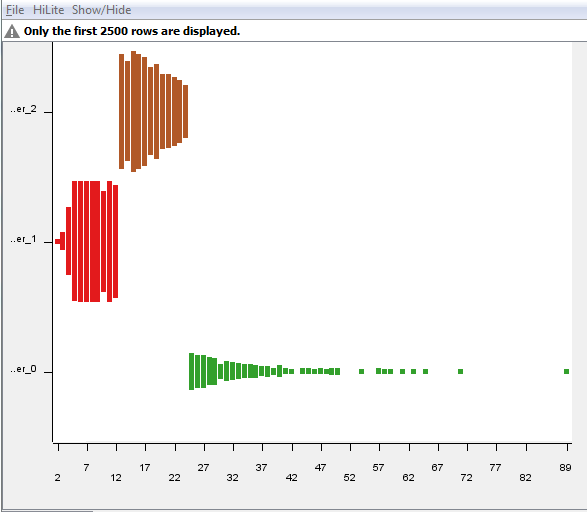


Gráfico -2- Distribución de los clusters (3 clusters con K-media)

**2/ Dos clusters:**

La distribución de los clusters se puede ver en el gráfico -3-.

Podemos ver la clasificación por clusters (0 o 1), donde los colores corresponden a los clusters asignados, así como la distancia euclídea y la puntuación de cada término.



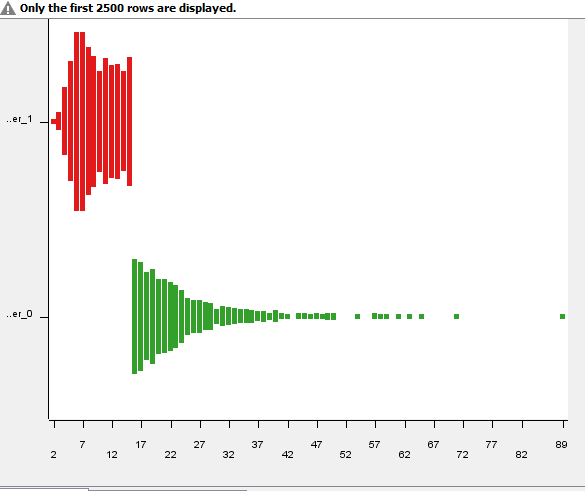


Gráfico -3- Distribución de los clusters (2 clusters con K-media)

**Conclusiones:**

Ya no es posible crear clusters a partir de 6 clusters, debido a los parámetros elegidos o a ya no ser posible dividir más los clusters.

Los términos que no tienen significado generalmente se agrupan en el mismo grupo, esto podría ser una idea para crear una lista de palabras reservadas que sería específica para el corpus estudiado.