



Evaluación del algoritmo K-Means en la asignación de centroides: Desarrollo de un prototipo simulador

Osvaldo Sposito, Julio Bossero, Viviana Ledesma, Lorena Matteo, Sebastián Quevedo

RESUMEN

El presente trabajo de investigación se centra en un estudio comparativo de la calidad del agrupamiento entre el algoritmo K-Means respecto a K-Means++, que es una extensión del primero. La evaluación de la comparativa entre los métodos se realizó empleando la métrica del “coeficiente de silueta” (Silhouette) y los resultados indican que ambos algoritmos tienen coeficientes positivos, lo que evidencia una correcta agrupación de los objetos.

CONTEXTO

Financiado por la UNLaM, el trabajo es la continuación de los proyectos “Estudio del proceso de recuperación de información en corpus voluminosos mediante una modificación del algoritmo K-Means para universos de gran dimensión y múltiples segmentos”, perteneciente al Programa de Incentivos (2023-2024), y “Justicia Tecnológica: Reforzando Experticia con Redes Neuronales para una toma de decisiones más eficiente” (2024-2025)

LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

- Estudiar diferentes librerías para C#, para graficar las salidas de K-Means .
- Analizar distintas métricas para evaluar la calidad de agrupamiento.
- Estudiar otras modificaciones de K-Means como: K-Medoides, Fuzzy C-Means, Mini Batch K-means, etc

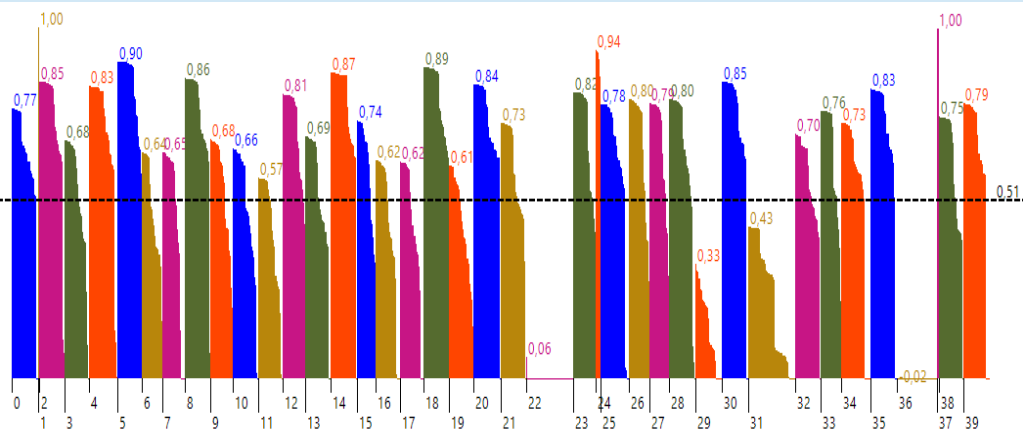
Modelo Vectorial

matriz_resultante.txt: Bloc de notas
Archivo Edición Formato Ver Ayuda
0.000000 0.000000 0.000000 0.000000 0.000000 0.049421
0.000000 0.000000 0.552939 0.125486 0.000000 0.000000
0.469207 0.000000 0.000000 0.000000 0.089008 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.377602
0.000000 0.389278 0.000000 0.000000 0.000000 0.165931
0.000000 0.000000 0.000000 0.295097 0.000000 0.000000
0.000000 0.000000 0.000000 0.000000 0.000000 0.000000

Coeficiente de silueta

$$\text{Silueta} = \frac{1}{N} \sum_{i=1}^N \left(\frac{b_i - a_i}{\max\{b_i, a_i\}} \right)$$

Resultado del Coeficiente de silueta K-Means++ de los 40 grupos



Método	Silhouette
k-Means	0,035278
K-Means++	0,505137

El “Coeficiente de Silueta” (Silhouette) es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering . La métrica puede ser calculada con cualquier fórmula de distancia, como la distancia euclidiana o la distancia Manhattan.

Su coeficiente va de -1 a +1, donde: un valor positivo indica que el objeto está bien emparejado con su propio grupo o clúster, si es 0 la observación i está entre dos grupos y si el valor es negativo significa que está mal asignada a su grupo o clúster.

RESULTADOS OBTENIDOS/ ESPERADOS

- Se empleó el coeficiente de Silueta (Silhouette) como métrica para esta comparación. El resultado obtenido del promedio de los 40 grupos empleados en la prueba muestra que el valor promedio del K-Means++ es casi de 0.51, mejorando al promedio que tiene de K-Means que es un poco superior a 0.03.
- Durante ambos resultados indicaron que las agrupaciones obtenidas mediante los algoritmos son adecuadas, porque en el resultado de los valores del coeficiente no hay cifras cercanas a cero ni negativas.

FORMACIÓN DE RECURSOS HUMANOS

- Un miembro dedicado a la elaboración de su tesis doctoral, aportando al equipo tanto sus habilidades investigativas como su compromiso con el avance del conocimiento en el campo.
- Dos integrantes que recientemente han obtenido sus Maestrías en Informática en la UNLaM, consolidando así su formación académica y su capacidad para abordar desafíos investigativos de alto nivel.
- Cuatro integrantes han sido incorporados en la "Carrera de Docente Investigador UNLaM" en la categoría Investigador en Formación. Esta designación reconoce su potencial como investigadores emergentes y les brinda la oportunidad de desarrollarse tanto en el ámbito académico como en el de la investigación.