UNIVERSIDAD NACIONAL DE COLOMBIA - SEDE MEDELLÍN

FACULTAD DE MINAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN Y DE LA DECISIÓN

Eq. 20

- **Inteligencia Artificial,** Código: **3007855**
- **Introducción a la Inteligencia Artificial,** Código: **3010476**

Semestre**: 02/2024**

Prof. Demetrio Arturo Ovalle Carranza, Ph.D. (e-mail: dovalle@unal.edu.co)

Monitores:
Daniel Metaute Medina (**e-mail: dmetaute@unal.edu.co**)
Felipe Muñoz Echeverri (e-mail: fmunoze@ unal.edu.co)

ENUNCIADO DEL MICRO-PROYECTO2 (12%) - CLUSTERING

Fecha: Martes 17 de diciembre
Fecha de entrega: Lunes 27 de enero, entrega digital; sustentación martes 28 y jueves 30 de enero de 2024

Estudiantes:

Leonard David Vivas Dallos
Tomás Escobar Rivera
Nicolás Orozco Medina
Equipo 20

OBJETIVO – Analizar las películas clasificadas en el top 1000 de IMDB.

DATASET: 2 archivos, IMDB_movie_reviews_details.csv & IMDB_movie_reviews_details_summary.txt

Usando un algoritmo similar al presentado en clase y en la práctica, realizar un programa en Python que determine los clústeres en los datos del archivo .csv a partir de las tres variables que según su análisis mejor segmentan los datos.

Se sugiere mezclar los datos de forma aleatoria antes de realizar cualquier otro tipo de preprocesamiento.

Indicaciones de entrega: sólo debe entregarse el archivo descargado de Google Colaboratory con el siguiente formato de nombre: “Equipo#_Microproyecto2” *(ejemplo: Equipo20_Microproyecto2*). Los trabajos serán probados con el conjunto de datos original, por lo tanto, todas las modificaciones hechas al dataset deben realizarse desde el código.

Los integrantes deben aparecer en la libreta (.ipynb) para que se les pueda asignar la calificación obtenida.

1) Explique si tuvo que realizar algún tipo de pre-procesamiento en el archivo .csv.

Para este ejercicio se usará el dataset IMDB_movie_reviews_details.csv y su descripción en IMDB_movie_reviews_details_summary.txt.

2) Explique las técnicas utilizadas para establecer el número de clústeres del conjunto de datos. (Método 1. Curva de codo, Método 2. Estadístico de gap, Método 3. Análisis de la Silueta, etc.). Analice los resultados obtenidos y justifique su respuesta.

3) Realizar un agrupamiento jerárquico de los datos a través de un dendograma, según se explicó en el taller y en clase, donde se visualicen los clústeres del dataset entregado. Este código debe hacer parte de la libreta (.ipynb) y ser debidamente documentado.

4) Basado en los clústeres establecidos, describa qué características son más representativas para cada grupo o clúster, realizando la mayor caracterización de los datos posible.

NOTA: Recuerde explicar qué está haciendo en cada parte del código. Para esto debe entregar la libreta (.ipynb) correspondiente, debidamente documentada. La falta de esto afectará la calificación final.

5) A partir del análisis de resultados describa 5 tendencias encontradas en los datos, justifique su respuesta.