Machine Learning Aplicado
Taller Práctico: Similitud del Coseno

profesor Aaron Ponce Sandoval
Nombre: Fabian Alexis Vidal Torres

1. Instrucciones: La actividad consiste en la aplicación del cálculo de similitud del coseno a un conjunto de datos de Peliculas.csv. El cálculo debe ser realizado de una pelicula vs todas, por cada comparación crear un gráfico de similitud del coseno.
Librerías: Pandas, Numpy

Objetivo: Aprender a codificar y utilizar la medida de similitud del coseno para encontrar la similitud entre películas en un conjunto de datos csv.

2. Descripción de Datos:

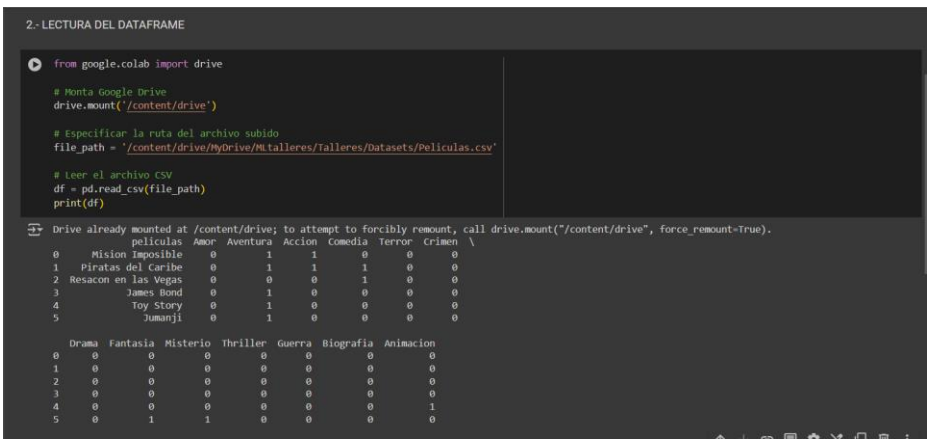
Este conjunto de datos contiene información sobre películas y su clasificación en diferentes géneros. Cada fila representa una película, y cada columna representa un género cinematográfico. Los géneros incluidos en el conjunto de datos son: Amor, Aventura, Accion, Comedia, Terror, Crimen, Drama, Fantasía, Misterio, Thriller, Guerra, Biografia y Animacion. Cada entrada en la tabla representa si una película pertenece o no a un género particular, siendo 1 para verdadero y 0 para falso.

3. Desarrollo de Actividad

Para cada uno de los siguientes ítems, adjuntar imagen del código realizado.

1. Importación de librerías

2. Lectura de dataframe



3. Procesamiento de datos

3.- PROCESAMIENTO DE LOS DATOS

```
genres = df.iloc[:, 1:]
print(genres)
```

```
   Amor  Aventura  Accion  Comedia  Terror  Crimen  Drama  Fantasia  Misterio  \
0     0         1       1        0       0       0      0         0         0
1     0         1       1        1       0       0      0         0         0
2     0         0       0        1       0       0      0         0         0
3     0         1       0        0       0       0      0         0         0
4     0         1       0        0       0       0      0         0         0
5     0         1       0        0       0       0      0         1         1

   Thriller  Guerra  Biografia  Animacion
0         0       0          0          0
1         0       0          0          0
2         0       0          0          0
3         0       0          0          0
4         0       0          0          1
5         0       0          0          0
```

Extreamos todas las columnas las cuales son los generos que compararemos a expecion de la primera

4. Descripcion de función de similitud

4.- DESCRIPCION DE FUNCION DE SIMILITUD

```
cosine_sim = cosine_similarity(genres)

cosine_sim_df = pd.DataFrame(cosine_sim, index=df['peliculas'], columns=df['peliculas'])

# Mostrar el dataframe de similitudes
print("Similitud del Coseno entre Películas:")
print(cosine_sim_df)
```

```
Similitud del Coseno entre Películas:
peliculas            Mision Imposible  Piratas del Caribe  \
peliculas
Mision Imposible             1.000000            0.816497
Piratas del Caribe           0.816497            1.000000
Resacon en las Vegas         0.000000            0.577350
James Bond                   0.707107            0.577350
Toy Story                    0.500000            0.408248
Jumanji                      0.408248            0.333333

peliculas            Resacon en las Vegas  James Bond  Toy Story   Jumanji
peliculas
Mision Imposible                  0.00000    0.707107   0.500000  0.408248
Piratas del Caribe                0.57735    0.577350   0.408248  0.333333
Resacon en las Vegas              1.00000    0.000000   0.000000  0.000000
James Bond                        0.00000    1.000000   0.707107  0.577350
Toy Story                         0.00000    0.707107   1.000000  0.408248
Jumanji                           0.00000    0.577350   0.408248  1.000000
```

La fórmula de similitud del coseno está dada por: sim (A, B) = (A·B) / ‖A‖‖B‖, donde A y B son las películas para comparar. Los resultados describen una matriz que muestra la similitud que presentan las películas donde 1 es el valor máximo y 0 el mínimo. Algunos resultados son sorprendentes, puesto que se puede apreciar que existen películas como Toy Story y James bond donde la similitud es del 0,7. Esto se ve reflejado en la coincidencia de sus géneros, ambas son de aventura y presentan este género en común.

5. Gráfico comparativo de similitud entre las películas.

5.- GRAFICO COMPARATIVO DE SIMILITUD ENTRE PELÍCULAS

```
[14] #Gráficos de similitud para cada película
     for idx, movie in enumerate(df['peliculas']):
         similarities = cosine_sim[idx]
         plt.figure(figsize=(10, 6))
         plt.barh(df['peliculas'], similarities, color='skyblue')
         plt.xlabel('Similitud del Coseno')
         plt.title(f'Similitud del Coseno de "{movie}" con otras películas')
         plt.axvline(x=0.5, color='red', linestyle='--')  # Línea de referencia para similitud del coseno 0.5
         plt.xlim(0, 1)
         plt.tight_layout()
         plt.show()
```

A continuación, se muestran algunos gráficos generados por el Código.

Grafico 1

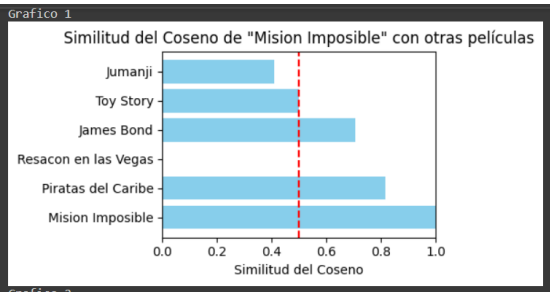
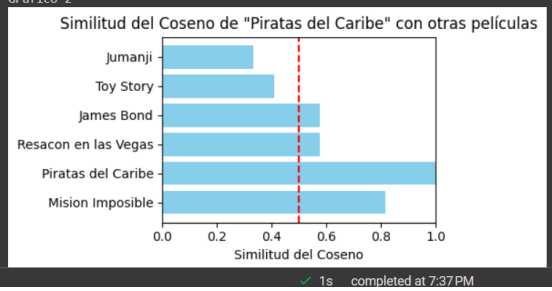


Grafico 2



✓ 1s completed at 7:37PM

Grafico 3

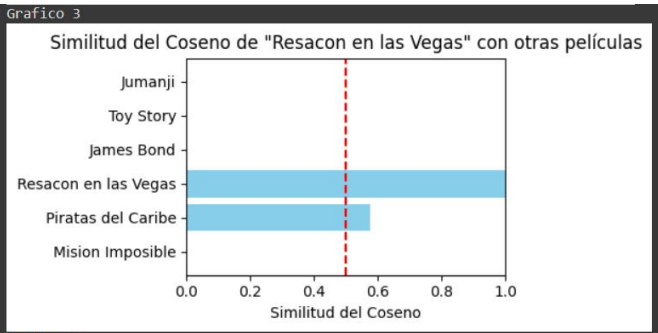


Grafico 4

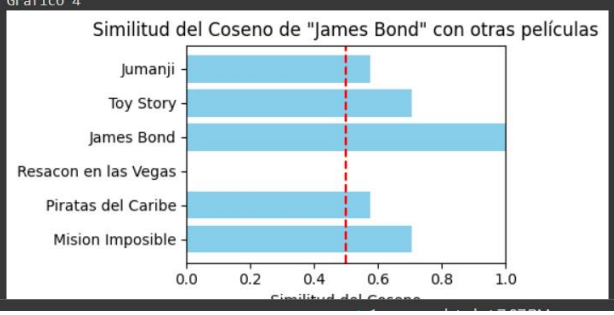


Grafico 5

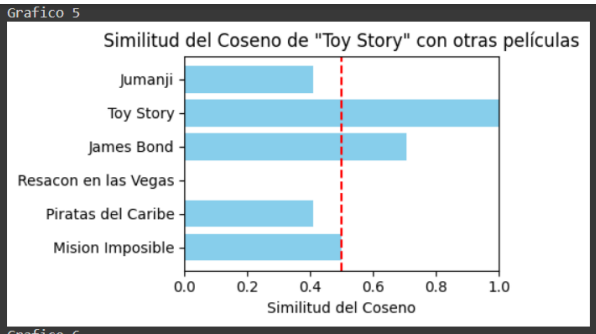


Grafico 6