

Índice:

Índice …………………………………………………………………………………………………………………………. 1/5

Introducción ………………………………………………………………………………………………………………. 2/5

Metodología ………………………………………………………………………………………………………………. 2/5

Desarrollo/Implementación ………………………………………………………………………………………. 3/5

Resultados …………………………………………………………………………………………………………………. 3/5

Conclusiones ………………………………………………………………………………………………………………. 4/5

# Introducción:

Llevaremos a cabo la resolución de Trabajo Práctico N° 1 de la materia Procesamiento del Lenguaje Natural (NLP).

Implementaremos un Clasificador de Recomendaciones Recreativas utilizando Procesamiento del Lenguaje Natural.

Este trabajo tiene como objetivo desarrollar un programa de Procesamiento de Lenguaje Natural en Python que, según el estado de ánimo del usuario, recomiende ver una película, jugar un juego de mesa o leer un libro. Construiremos un clasificador que categorice el estado de ánimo del usuario, para luego sugerirle el conjunto de recomendaciones basada en una frase de preferencia ingresada por el mismo.

En el contexto de unas vacaciones de playa con días lluviosos, se plantea una solución para recomendar actividades recreativas según el estado de ánimo del usuario en esos días de mal clima.

Este informe detalla las metodologías implementadas, el desarrollo del código, los resultados obtenidos y las conclusiones acerca del trabajo.

# Metodología:

Durante la construcción del proyecto, aplicamos las siguientes metodologías:

## Text Clasification (Clasificación de texto):

Para la clasificación de estados de ánimo, utilizamos métodos de clasificación supervisada de texto, entrenando clasificadores como Regresión Logística, Random Forest y Multinomial Naive Bayes. Estos clasificadores nos permiten categorizar las entradas de texto en estados de ánimo específicos.

## Web Scraping:

Extrajimos información sobre los 1000 libros más populares del proyecto Gutenberg, mediante un enlace proporcionado por la cátedra.

# Desarrollo/Implementación:

## Importación de librerías

## Datasets para entrenamiento del modelo

Para entrenar los modelos predictores del estado de ánimo, le pedimos a la aplicación ChatGPT de OpenAI que nos genere aproximadamente 100 frases acordes a estados de ánimo proporcionados. Los estados de ánimo elegidos fueron: Triste, Melancólico, Neutral, Alegre y Eufórico. Además, solicitamos un dataset extra para testear con frases diferentes a las de entrenamiento, y así poder detectar posible overfitting.

## Entrenamiento de modelos

Fueron realizadas varias pruebas hasta encontrar un modelo con métricas adecuadas. Optamos por el modelo Multinomial Naive Bayes, el cual ofrece un parámetro “alpha” que controla el grado de suavizado. Cuando dicho parámetro está seteado en 0, hace que las palabras nuevas (no observadas en el conjunto de entrenamiento) tengan una probabilidad en cero. Por el contrario, cuando el parámetro es mayor a cero, hace que la probabilidad de que una palabra sea cero se reduzca, por lo cual el modelo así manejaría mejor las palabras nuevas o raras. Concluimos que un buen valor sería 0.1 ya que el modelo se ajusta lo suficiente a los datos de entrenamiento.

## Extracción de información necesaria desde csv y mediante web scraping

Para esta instancia, decidimos extraer de los archivos csv solo información que podría ser relevante a la hora de decidir qué recomendaciones ofrecer.

Para realizar el web scraping, utilizamos la biblioteca BeautifulSoup. Luego de extraer la información y recolectarla en un diccionario, lo convertimos en un DataFrame y posteriormente lo descargamos.

## Calcular recomendaciones

Para convertir nuestro texto a vectores numéricos (embeddings), utilizamos el modelo de Sentence Transformers all-mpnet-base-v2: mapea oraciones en múltiples idiomas a vectores de 768 dimensiones. Esto, para poder medir la similitud semántica entre frases, para encontrar las coincidencias relevantes basadas en el estado de ánimo y la temática elegida por el usuario.

Luego, definimos una función que calcula las recomendaciones de películas, juegos de mesa y libros basadas en los criterios mencionados anteriormente.

Utilizamos el modelo GliNER para identificar entidades relevantes en las descripciones de las posibles recomendaciones. El modelo captura elementos clave como nombres de personas, lugares, eventos, entre otros para clasificar la temática ingresada. Luego convertimos esa descripción en vectores numéricos mediante Sentence Transformers.

El proceso de recomendación mide la similitud mediante coseno entre los embeddings.

# Resultados:

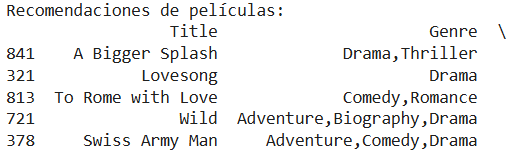
Resultados obtenidos en accuracy y F-1 score (general) entrenando los modelos:

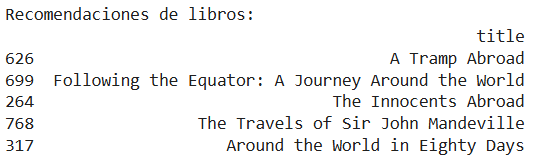
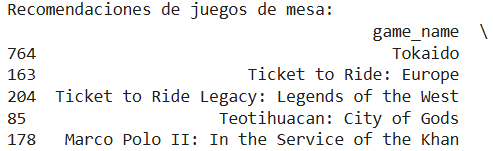
* Regresión Logística: 0.72
* Regresión Logística con modelo pre-entrenado: 0.72
* Random Forest: 0.54
* Multinomial Naive Bayes: 0.85

Ejemplo de entrada:

* Estado de ánimo: Alegre
* Temática**:** “historias de aventuras, con viajes y descubrimientos emocionantes”.

El sistema identificó las entidades relacionadas con la temática y el estado de ánimo y calculó las similitudes de los embeddings entre las descripciones de las películas, juegos de mesa y libros. Posteriormente, las recomendaciones más relevantes fueron:





Este proceso de recomendaciones se realizó utilizando el modelo de similitud coseno entre los embeddings generados por Sentence Transformers. Los resultados mostraron una alta relevancia para el usuario que buscaba actividades recreativas según su estado de ánimo y preferencia temática.

# Conclusiones:

Este trabajo demostró la viabilidad de usar procesamiento de lenguaje natural (PLN) para realizar recomendaciones personalizadas basadas en el estado de ánimo y la temática preferida por el usuario. La combinación de clasificación de texto y similitudes semánticas mediante embeddings resultó en un sistema robusto y eficiente para generar sugerencias en contextos recreativos. Algunas de las conclusiones importantes son:

* Precisión en la clasificación: Los modelos utilizados, especialmente el Multinomial Naive Bayes, lograron una alta precisión en la clasificación de los estados de ánimo, lo que permitió una categorización eficaz de las frases proporcionadas por los usuarios.
* Recomendaciones relevantes: El sistema fue capaz de generar recomendaciones personalizadas que tuvieron en cuenta tanto el estado de ánimo como la temática de las actividades, lo cual mejoró la relevancia de las sugerencias.
* Escalabilidad y mejora: Este modelo podría ampliarse para incorporar más categorías de actividades recreativas y otros estados emocionales, mejorando aún más la precisión y las recomendaciones para un público más diverso.