



# Tecnicatura Universitaria en Inteligencia Artificial

## Procesamiento del Lenguaje Natural

---

# Trabajo Práctico 1

### Integrantes:

- Herrera Morena - (H-1187/8)
- Zorzolo Rubio Juana - (Z-1217/3)

### Profesores:

- Juan Pablo Manson.
- Alan Geary.

Fecha de entrega: 06/11/24

## ÍNDICE

Resumen .....	3
Introducción.....	3
Metodología.....	4
Desarrollo/Implementación.....	5
Resultados.....	8
Pruebas y Ejemplos de Ejecución.....	9
Conclusiones .....	15
Referencias.....	15
Anexos.....	16

## Resumen

Este trabajo práctico presenta el desarrollo de un clasificador de recomendaciones recreativas basado en el procesamiento de lenguaje natural (NLP). El objetivo es ofrecer opciones de entretenimiento (como películas, juegos de mesa y libros) para días de lluvia, según una frase de preferencia ingresada por el usuario.

El sistema clasifica el estado de ánimo en categorías predefinidas, utilizando un modelo de clasificación supervisado para comprender el contexto emocional del usuario y métodos de embeddings y reconocimiento de entidades nombradas (NER) para calcular la similitud semántica entre la frase ingresada y una base de datos de opciones recreativas.

Los datos utilizados provienen de bases de datos de películas y juegos y del Proyecto Gutenberg con técnicas de web scraping para extraer la información de libros. Los resultados muestran recomendaciones personalizadas y coherentes, logrando el objetivo de crear un sistema útil y adaptado al idioma español.

En el futuro, se propone optimizar el sistema para mejores resultados y menor uso de recursos. Además, explorar nuevas fuentes de datos para ampliar el catálogo de recomendaciones.

## Introducción

### Contexto y Justificación

El entretenimiento ganó relevancia en situaciones donde las actividades al aire libre son limitadas por condiciones climáticas adversas. En este trabajo, se desarrolla una solución que el NLP, analiza el estado de ánimo del usuario y proporciona recomendaciones recreativas personalizadas basadas en las preferencias ingresadas. Esta herramienta puede ser de utilidad durante días de lluvia en vacaciones, donde las personas buscan alternativas para pasar el tiempo dentro de sus alojamientos. El uso de NLP permite una interacción natural en español, facilitando una experiencia más accesible para usuarios hispanohablantes.

### Propósito del Proyecto

El propósito de este proyecto es construir un sistema de recomendación que pueda comprender las preferencias del usuario y proponer opciones de entretenimiento (películas, juegos de mesa y libros). Para lograrlo, el sistema utiliza un clasificador supervisado que identifica el estado de ánimo, combinado con técnicas de embeddings y reconocimiento de entidades nombradas para optimizar la búsqueda y relevancia de las recomendaciones.

### Objetivos Específicos

1. **Desarrollar un clasificador de estados de ánimo** que categorice el estado emocional del usuario a partir de una frase ingresada en español, con categorías como "Alegre", "Melancólico" o "Ni fu ni fa".
2. **Implementar una técnica de procesamiento semántico** mediante embeddings y reconocimiento de entidades nombradas para calcular la similitud entre la preferencia ingresada por el usuario y las descripciones de las opciones de entretenimiento.
3. **Construir un sistema de scraping** para extraer información sobre libros desde el Proyecto Gutenberg y estructurarla en un dataset de fácil acceso.

4. **Integrar y estructurar los datos de las distintas fuentes** (juegos de mesa, películas y libros) para permitir recomendaciones adaptadas a los gustos del usuario.
5. **Proporcionar un sistema de recomendaciones en español** que logre una interacción natural con el usuario hispanohablante, facilitando el entendimiento y la usabilidad de la herramienta.

## Metodología

### Fuentes de Datos

- **Juegos de Mesa:** La base de datos de juegos de mesa utilizada es “bgg\_database.csv,” que contiene información detallada sobre diferentes tipos de juegos, proporcionando una amplia gama de opciones.
- **Películas:** Los datos sobre películas se obtuvieron de “IMDB-Movie-Data.csv,” que contiene datos sobre películas de diversas temáticas y géneros.
- **Libros:** Para los libros, se realizó web scraping en el sitio Proyecto Gutenberg (<https://www.gutenberg.org>), extrayendo datos de los 1000 libros más populares. El proceso se realizó con BeautifulSoup para analizar el HTML, y las listas de libros fueron filtradas y estructuradas en un dataframe para su posterior análisis.

### Técnicas de Procesamiento de Lenguaje Natural

- **Clasificación Supervisada del Estado de Ánimo:** Se utilizó un modelo de clasificación supervisada, implementado con el paquete Hugging Face Transformers y un modelo pre entrenado de BERT, específicamente [nlptown/bert-base-multilingual-uncased-sentiment](#). Este clasificador fue diseñado para identificar el estado de ánimo en frases en español, categorizando el estado emocional en etiquetas del 1 al 5. Luego en base a la etiqueta, se clasifica el sentimiento del usuario como "Alegre", "Melancólico", o "Ni fu ni fa". Además, se utilizó un pipeline de clasificación para simplificar el proceso de inferencia. El resultado es el sentimiento clasificado.
- **Traducción:** Debido a que la frase ingresada está en español, pero los datos en inglés, se utilizó el modelo de traducción automática NMT con su tokenizador. En específico el modelo es [Helsinki-NLP/opus-mt-es-en](#), el cual está diseñado para traducir textos de español a inglés.
- **Embeddings y Similitud Semántica:** Se implementó un modelo de embeddings usando SentenceTransformer, específicamente el modelo [paraphrase-multilingual-MiniLM-L12-v2](#). Este convierte tanto las descripciones como la preferencia ingresada en vectores que capturan el significado semántico. Luego, se calculó la similitud semántica utilizando la similitud del coseno entre la frase de preferencia y cada descripción. Este enfoque permite una visión basada en el contexto.
- **Reconocimiento de Entidades Nombradas (NER):** Para identificar entidades específicas (personas, años, locaciones, etc.) dentro de la frase de preferencia y las descripciones de las opciones recreativas, se utilizó un modelo NER, [gliner\\_multi-v2.1](#), cargado a través de la biblioteca GLINER. Este proceso asegura que las recomendaciones sean coherentes con los intereses específicos del usuario.

## Herramientas y Tecnologías Empleadas

- **Software:** Python, junto con bibliotecas como pandas, numpy, BeautifulSoup, requests, transformers de Hugging Face y Gliner.
- **Hardware:** El desarrollo se llevó a cabo en Google Colab, aprovechando la infraestructura en la nube para la carga y el procesamiento de modelos pesados de NLP.

## Pasos del Proceso de Implementación y Cumplimiento de Objetivos

1. **Carga y Exploración de Datos:** Se comenzó con la carga de los datasets, inspeccionando su estructura y realizando ajustes necesarios en el pre procesamiento, como cambio en el nombre de las variables y eliminación de columnas innecesarias para el sistema.
2. **Web Scraping y Preparación del Dataset de Libros:** Se extrajo y procesó la lista de libros populares desde el Proyecto Gutenberg. Esto incluyó la limpieza y formateo de los datos para su inclusión en el sistema de recomendaciones.
3. **Construcción del Clasificador de Estado de Ánimo:** Se implementó un clasificador utilizando el modelo BERT, evaluando su desempeño en frases de prueba en español para asegurar una clasificación de alta precisión. En base a estas pruebas, se eligió el rango numérico para categorizar el estado de ánimo (alegre, melancólico o ni fu ni fa).
4. **Implementación de la Similitud Semántica y NER:** Con el estado de ánimo clasificado y la frase de preferencia ingresada, se calcularon las similitudes entre las frases del usuario y las descripciones de las opciones de entretenimiento, complementado con NER para capturar entidades clave.
5. **Validación y Pruebas:** El sistema fue sometido a pruebas con diferentes frases de entrada para evaluar la efectividad y coherencia de las recomendaciones.

## Desarrollo/Implementación

Nota: para mejor visualización y comprensión del código ejecutado, abrir el mismo con la opción *Open in colab* ya que desde GitHub no se muestra bien el archivo .ipynb



## Preparación y Pre procesamiento de Datos

- **Exploración Inicial de los Datasets:** En primer lugar, cada dataset fue cargado y explorado utilizando pandas, lo cual permitió identificar estructuras y tipos de datos. En particular, el dataset de libros, se generó a partir de un script de web scraping en el Proyecto Gutenberg. Utilizando BeautifulSoup, el proceso comenzó con la obtención de la lista de los 1000 libros más populares. A través de la búsqueda de enlaces específicos, se identificaron aquellos que contenían "/ebooks/" seguido de un número, garantizando así la validez de los libros seleccionados. De cada libro se extrajo información clave y se realizó una limpieza básica para asegurar un mejor procesamiento de los datos.

Para evitar sobrecargar el servidor, se incluyó un mecanismo de control de flujo de una pausa de 0.01 segundos entre cada solicitud. Una vez completada la recolección de los 1000 libros, los datos obtenidos se almacenaron en un dataframe.

- **Pre Procesamiento de los Datasets y del texto:** Para mayor facilidad, se renombraron columnas importantes dentro de los tres datasets y eliminaron las variables que no se precisaban para la recomendación. Además, para asegurar consistencia, se realizaron transformaciones en el texto, como por ejemplo convertirlo a minúsculas, eliminar acentos y caracteres no deseados y reemplazar cualquier signo que no sea alfanumérico por un espacio. Esto mejora la calidad de la entrada y la efectividad de los modelos en las tareas de NER y embeddings.

	title	author	link	genre	description
0	Frankenstein; Or, The Modern Prometheus	Mary Wollstonecraft Shelley	<a href="https://www.gutenberg.org/ebooks/84">https://www.gutenberg.org/ebooks/84</a>	InScience fictionIn	"Frankenstein; Or, The Modern Prometheus" by M...

	name	director/author	link	description
0	Frankenstein; Or, The Modern Prometheus	Mary Wollstonecraft Shelley	<a href="https://www.gutenberg.org/ebooks/84">https://www.gutenberg.org/ebooks/84</a>	"Frankenstein; Or, The Modern Prometheus" by M...

### Clasificador del Estado de Ánimo

- **Modelo de Clasificación Supervisada:** Para la clasificación del estado de ánimo en categorías se utilizó un modelo BERT. La entrada es una frase breve del usuario que expresa su estado emocional actual, la misma es procesada y clasificada por el modelo para obtener una etiqueta correspondiente entre 1 y 5. De esta manera se la clasifica y define al usuario como "Alegre", "Melancólico", y "Ni fu ni fa".

```
Frase del usuario: 'cansado'
Sentimiento: 1 star
Estado de ánimo del usuario: Melancólico
```

### Traducción y procesamiento de preferencia

- **Procesamiento de traducción:** A la preferencia ingresada por el usuario se la traduce con NMT. Dentro del código, se elige ir mostrando la frase ingresada, la frase limpia y su traducción. Posteriormente, se obtienen las entidades de la traducción y se calculan los embeddings, utilizando los modelos y funciones previamente cargados y definidos. Finalmente, se retornan los embeddings de la preferencia junto con las entidades encontradas para la futura búsqueda de similitudes con las descripciones de los dataframe.

```
Preferencia ingresada: una historia futurista donde la inteligencia artificial invade al mundo
Frase parseada: una historia futurista donde la inteligencia artificial invade al mundo
Traducción: a futuristic story where artificial intelligence invades the world
Entidades encontradas: [('artificial intelligence', 'organization'), ('world', 'location')]
```

### Embeddings y Comparación Semántica

- **Generación de Embeddings:** Una vez determinado el estado de ánimo, el usuario ingresa una frase de preferencia. A esta frase se la traduce y se aplica un modelo de embeddings, obteniendo representaciones vectoriales para la frase del usuario.
- **Cálculo de Similitud:** Con estos embeddings, se calcula la distancia entre la preferencia del usuario y los elementos en las bases de datos de películas, juegos de mesa y libros. La similitud se mide con la distancia coseno, de modo que las recomendaciones con mayor valor son las que el sistema prioriza.

## Reconocimiento de Entidades Nombradas (NER)

- **Identificación de Entidades Clave:** Utilizando un modelo NER, el sistema detecta entidades en la frase de preferencia (por ejemplo, “estados unidos” “locación” o “19/2/1983” “date”). Estas se consideran al momento de realizar la búsqueda de recomendaciones, filtrando y priorizando elementos que contengan entidades matcheadas.

## Generación de Recomendaciones

- **Calculo de coincidencias y distancias:** Iterando sobre las filas del dataset que se recibe como parámetro, se extrae el embedding de cada fila y se calcula la distancia que tiene con el embedding de la preferencia, utilizando la similitud del coseno. A continuación, se comparan las entidades que se encuentran en la preferencia del usuario con las que están presentes en la descripción de la fila del dataset. Este paso se realiza mediante la búsqueda de coincidencias, contando cuántas entidades de la preferencia aparecen en la descripción del elemento.
- **Integración de los Resultados:** Para cada fila, se crea un diccionario que contiene el nombre de la película, juego o libro, la cantidad de coincidencias de entidades y la similitud entre los embeddings. Si existen columnas adicionales en el dataset, como el nombre del autor/director y el link, también se agregan al diccionario para enriquecer la recomendación.
- **Filtrado de recomendaciones:** Finalmente, el sistema genera un conjunto de recomendaciones en función a la similitud semántica y la coincidencia de entidades de la frase de preferencia con la descripción. De esta manera, se proporciona una lista personalizada de 3 películas, juegos de mesa y libros.

Top 3 juegos sugeridos:

name	link	matches	similarity
Sentinels of the Multiverse	<a href="https://boardgamegeek.com/boardgame/102652/sentinels-of-the-multiverse">https://boardgamegeek.com/boardgame/102652/sentinels-of-the-multiverse</a>	1	0.5007
Euphoria: Build a Better Dystopia	<a href="https://boardgamegeek.com/boardgame/133848/euphoria-build-a-better-dystopia">https://boardgamegeek.com/boardgame/133848/euphoria-build-a-better-dystopia</a>	1	0.4525
Smartphone Inc.	<a href="https://boardgamegeek.com/boardgame/246684/smartphone-inc">https://boardgamegeek.com/boardgame/246684/smartphone-inc</a>	1	0.4419

Top 3 películas sugeridos:

name	director/author	matches	similarity
Spy	Paul Feig	1	0.535
Suicide Squad	David Ayer	1	0.4134
World War Z	Marc Forster	1	0.402

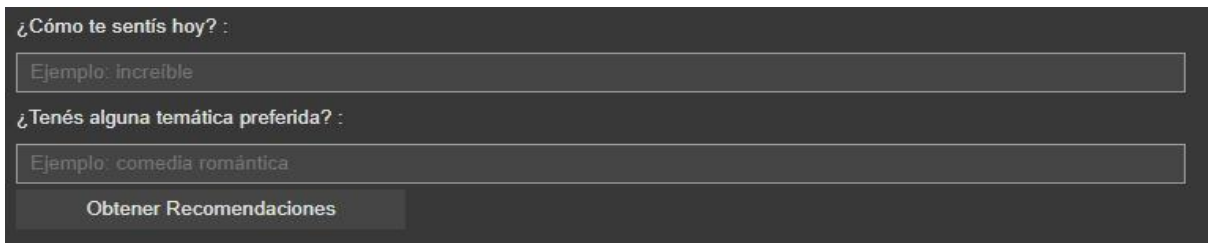
Top 3 libros sugeridos:

name	director/author	link	matches	similarity
Anthem	Ayn Rand	<a href="https://www.gutenberg.org/ebooks/1250">https://www.gutenberg.org/ebooks/1250</a>	1	0.4653
A Child's History of the World	V. M. Hillyer	<a href="https://www.gutenberg.org/ebooks/67149">https://www.gutenberg.org/ebooks/67149</a>	1	0.3618
The Protocols and world revolution :	Sergie Nilus	<a href="https://www.gutenberg.org/ebooks/64977">https://www.gutenberg.org/ebooks/64977</a>	1	0.3308

## Programa Principal

Se solicita al usuario que ingrese como se siente y la temática de preferencia que tiene. En base a eso se llaman a las funciones que evalúan el estado de ánimo y procesan la preferencia, la cual posteriormente se compara con las descripciones de los datasets.

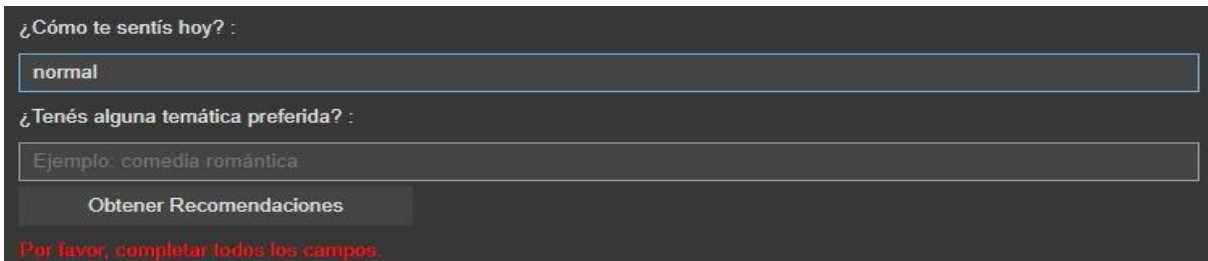
Se agregó una validación de entrada, la cual verifica que ambos campos estén completos por el usuario, en caso contrario, muestra un mensaje de error. Además, se hizo una pequeña interfaz interactiva, con etiquetas, campos de texto y un botón, el cual se acciona con la tecla 'Enter'.



¿Cómo te sentís hoy? :

¿Tenés alguna temática preferida? :

Obtener Recomendaciones



¿Cómo te sentís hoy? :

¿Tenés alguna temática preferida? :

Obtener Recomendaciones

Por favor, completar todos los campos.

En base a las comparaciones, se genera un top 3 de recomendaciones para los tres tipos de contenido, mostrando la información más relevante. El usuario recibe contenido personalizado.

## Resultados

### Desempeño del Clasificador de Estado de Ánimo

- El modelo de clasificación de estado de ánimo mostró un desempeño satisfactorio en la identificación de emociones, con un alto nivel de precisión en la categorización entre "Alegre," "Melancólico," y "Ni fu ni fa".
- Se llevaron a cabo pruebas adicionales utilizando un conjunto de frases en español para asegurar que el modelo BERT pudiera interpretar adecuadamente las emociones en el idioma, lo que reforzó su precisión en el contexto de uso.

### Evaluación de Recomendaciones Basadas en Similitud Semántica

- Para evaluar la efectividad del sistema de recomendación, se realizaron pruebas con frases de preferencia relacionadas a temas específicos. Por ejemplo, una frase como "una historia de amor en la selva" resultó en recomendaciones adecuadas de películas románticas y libros de aventura que contienen temas similares.
- Los resultados fueron presentados al usuario en orden de relevancia, calculada según la similitud semántica. En la mayoría de los casos, las recomendaciones más relevantes lograron captar el contexto de la preferencia ingresada, presentando opciones alineadas con las expectativas del usuario.

### Análisis de Resultados con NER (Reconocimiento de Entidades Nombradas)

- Las pruebas que incluyeron términos específicos como locaciones o fechas en las frases de preferencia mostraron que el modelo de NER fue efectivo en identificar palabras clave, mejorando la calidad de las recomendaciones.
- En los casos donde el modelo NER no detectó coincidencias exactas, las recomendaciones aún se mantuvieron coherentes gracias al cálculo de similitud semántica, mostrando resultados más precisos en general.



## Interacción con el Usuario

- El sistema permite al usuario ingresar su estado de ánimo y una preferencia temática para generar recomendaciones personalizadas. Después de verificar que ambos campos estén completos, se realiza el análisis de sentimientos y el reconocimiento de entidades nombradas para identificar emociones y temas clave.
- En caso de campos incompletos, el sistema muestra un mensaje de error claro, ayudando al usuario a ajustar su entrada. Finalmente, se presentan las tres recomendaciones principales en cada categoría (juegos, películas, libros), ordenadas por relevancia según la preferencia indicada, logrando una interacción intuitiva y efectiva.

## Pruebas y Ejemplos de Ejecución

En los siguientes ejemplos se observa, por un lado, la emoción y la preferencia que ingresa el usuario. Además, se muestra la frase parseada, su traducción y las entidades encontradas en la misma.

Por otro lado, se ven las 3 recomendaciones de juegos, películas y libros que más se adaptan a la preferencia del usuario.

- ✚ Sentimiento: muy bien.
- ✚ Preferencia: una mujer en un musical de estados unidos.

```
Frase del usuario: 'muy bien'
Sentimiento: 5 stars
Estado de ánimo del usuario: Alegre
-----
Preferencia ingresada: una mujer joven en un musical de estados unidos

Frase parseada: una mujer joven en un musical de estados unidos
Traducción: a young woman in a musical from the United States
Entidades encontradas: [('young woman', 'person'), ('united states', 'location')]
```

Top 3 juegos sugeridos:

name	link	matches	similarity
Freedom: The Underground Railroad	<a href="https://boardgamegeek.com/boardgame/119506/freedom-the-underground-railroad">https://boardgamegeek.com/boardgame/119506/freedom-the-underground-railroad</a>	1	0.1518
Twilight Struggle	<a href="https://boardgamegeek.com/boardgame/12333/twilight-struggle">https://boardgamegeek.com/boardgame/12333/twilight-struggle</a>	1	0.1472
Carnegie	<a href="https://boardgamegeek.com/boardgame/310873/carnegie">https://boardgamegeek.com/boardgame/310873/carnegie</a>	1	0.0906

Top 3 películas sugeridos:

name	director/author	matches	similarity
It Follows	David Robert Mitchell	1	0.346
Safe Haven	Lasse Hallström	1	0.3021
Jupiter Ascending	Lana Wachowski	1	0.2827

Top 3 libros sugeridos:

name	matches	similarity	director/author	link
Slave Narratives: A Folk History of Slavery in the United States from Interviews with Former Slaves, (56)	1	0.2179	Desconocido	<a href="https://www.gutenberg.org/ebooks/36020">https://www.gutenberg.org/ebooks/36020</a>
The Souls of Black Folk	1	0.2157	W. E. B. Du Bois	<a href="https://www.gutenberg.org/ebooks/408">https://www.gutenberg.org/ebooks/408</a>
History of Woman Suffrage, Volume I (50)	1	0.2134	Desconocido	<a href="https://www.gutenberg.org/ebooks/28020">https://www.gutenberg.org/ebooks/28020</a>

- ✚ Sentimiento: triste.
- ✚ Preferencia: una historia de amor en la selva en 1980.

```

Frase del usuario: 'triste'
Sentimiento: 1 star
Estado de ánimo del usuario: Melancólico
-----
Preferencia ingresada: una historia de amor en la selva en 1980

Frase parseada: una historia de amor en la selva en 1980

Traducción: a love story in the jungle in 1980

Entidades encontradas: [('1980', 'time')]

```

Top 3 juegos sugeridos:

name	link	matches	similarity
The Lost Expedition	<a href="https://boardgamegeek.com/boardgame/216459/the-lost-expedition">https://boardgamegeek.com/boardgame/216459/the-lost-expedition</a>	0	0.4855
Endless Winter: Paleoamericans	<a href="https://boardgamegeek.com/boardgame/305096/endless-winter-paleoamericans">https://boardgamegeek.com/boardgame/305096/endless-winter-paleoamericans</a>	0	0.4766
Everdell	<a href="https://boardgamegeek.com/boardgame/199792/everdell">https://boardgamegeek.com/boardgame/199792/everdell</a>	0	0.4159

Top 3 películas sugeridos:

name	director/author	matches	similarity
Argo	Ben Affleck	1	0.1461
The Bad Batch	Ana Lily Amirpour	0	0.6009
Vincent N Roxxy	Gary Michael Schultz	0	0.5201

Top 3 libros sugeridos:

name	director/author	link	matches	similarity
The Jungle	Upton Sinclair	<a href="https://www.gutenberg.org/ebooks/140">https://www.gutenberg.org/ebooks/140</a>	0	0.6622
The Beast in the Jungle	Henry James	<a href="https://www.gutenberg.org/ebooks/1093">https://www.gutenberg.org/ebooks/1093</a>	0	0.619
The Jungle Book	Rudyard Kipling	<a href="https://www.gutenberg.org/ebooks/236">https://www.gutenberg.org/ebooks/236</a>	0	0.5997

- 🌈 Sentimiento: exhausto pero feliz
- 🌈 Preferencia: un verano en nueva york.

```

Frase del usuario: 'exhausta pero feliz'
Sentimiento: 4 stars
Estado de ánimo del usuario: Ni fu ni fa
-----
Preferencia ingresada: un verano en nueva york

Frase parseada: un verano en nueva york

Traducción: a summer in new york

Entidades encontradas: [('summer', 'event'), ('new york', 'location')]

```

Top 3 juegos sugeridos:

name	link	matches	similarity
A Fake Artist Goes to New York	<a href="https://boardgamegeek.com/boardgame/135779/a-fake-artist-goes-to-new-york">https://boardgamegeek.com/boardgame/135779/a-fake-artist-goes-to-new-york</a>	1	0.2583
Tammany Hall	<a href="https://boardgamegeek.com/boardgame/30645/tammany-hall">https://boardgamegeek.com/boardgame/30645/tammany-hall</a>	1	0.2364
Railways of the World	<a href="https://boardgamegeek.com/boardgame/17133/railways-of-the-world">https://boardgamegeek.com/boardgame/17133/railways-of-the-world</a>	1	0.1704

Top 3 películas sugeridos:

name	director/author	matches	similarity
Cloverfield	Matt Reeves	1	0.4189
Carol	Todd Haynes	1	0.3596
Blue Jasmine	Woody Allen	1	0.3333

Top 3 libros sugeridos:

name	director/author	link
matches	similarity	
How the Other Half Lives: Studies Among the Tenements of New York	Jacob A. Riis	<a href="https://www.gutenberg.org/ebooks/45502">https://www.gutenberg.org/ebooks/45502</a>
1	0.3785	
The Age of Innocence	Edith Wharton	<a href="https://www.gutenberg.org/ebooks/541">https://www.gutenberg.org/ebooks/541</a>
1	0.2782	
The Federalist Papers	Alexander Hamilton and John Jay and James Madison	<a href="https://www.gutenberg.org/ebooks/1404">https://www.gutenberg.org/ebooks/1404</a>
1	0.2464	

- 🚩 Sentimiento: tranquilo.
- 🚩 Preferencia: una historia del siglo 19.

```

Frase del usuario: 'tranquilo'
Sentimiento: 5 stars
Estado de ánimo del usuario: Alegre
-----
Preferencia ingresada: una historia del siglo 19

Frase parseada: una historia del siglo 19

Traducción: a history of the 19th century

Entidades encontradas: [('19th century', 'time')]

```

Top 3 juegos sugeridos:

name	link	matches	similarity
Trains	<a href="https://boardgamegeek.com/boardgame/121488/trains">https://boardgamegeek.com/boardgame/121488/trains</a>	1	0.4319
Great Western Trail	<a href="https://boardgamegeek.com/boardgame/193738/great-western-trail">https://boardgamegeek.com/boardgame/193738/great-western-trail</a>	1	0.3547
Great Western Trail: Second Edition	<a href="https://boardgamegeek.com/boardgame/341169/great-western-trail-second-edition">https://boardgamegeek.com/boardgame/341169/great-western-trail-second-edition</a>	1	0.3547

Top 3 películas sugeridos:

name	director/author	matches	similarity
Lady Macbeth	William Oldroyd	1	0.363
Les Misérables	Tom Hooper	1	0.2381
Australia	Baz Luhrmann	0	0.4847

Top 3 libros sugeridos:

name	matches	similarity	director/author	link
The Best of the World's Classics, Restricted to Prose, Vol. X (of X) - America - II, Index (161)	1	0.5106	Desconocido	<a href="https://www.gutenberg.org/ebooks/29145">https://www.gutenberg.org/ebooks/29145</a>
Personal Memoirs of U. S. Grant, Complete	1	0.3767	Ulysses S. Grant	<a href="https://www.gutenberg.org/ebooks/4367">https://www.gutenberg.org/ebooks/4367</a>
The Thousand and One Nights, Vol. I. (141)	1	0.3429	Desconocido	<a href="https://www.gutenberg.org/ebooks/34206">https://www.gutenberg.org/ebooks/34206</a>



- 🚩 Sentimiento: contento pero con malestar físico.
- 🚩 Preferencia: algo relacionada a star wars.

```

Frase del usuario: 'contento pero con malestar físico'
Sentimiento: 3 stars
Estado de ánimo del usuario: Ni fu ni fa
-----
Preferencia ingresada: algo relacionado a star wars

Frase parseada: algo relacionado a star wars

Traducción: something related to star wars

Entidades encontradas: []

```

Top 3 juegos sugeridos:

name	link	matches	similarity
Star Realms	<a href="https://boardgamegeek.com/boardgame/147020/star-realms">https://boardgamegeek.com/boardgame/147020/star-realms</a>	0	0.7015
Star Wars: The Clone Wars	<a href="https://boardgamegeek.com/boardgame/370913/star-wars-the-clone-wars">https://boardgamegeek.com/boardgame/370913/star-wars-the-clone-wars</a>	0	0.6638
Star Wars: Rebellion	<a href="https://boardgamegeek.com/boardgame/187645/star-wars-rebellion">https://boardgamegeek.com/boardgame/187645/star-wars-rebellion</a>	0	0.6563

Top 3 películas sugeridos:

name	director/author	matches	similarity
Spectral	Nic Mathieu	0	0.5187
Pacific Rim	Guillermo del Toro	0	0.5165
Guardians of the Galaxy	James Gunn	0	0.5162

Top 3 libros sugeridos:

name	director/author	link	matches	similarity
The War of the Worlds	H. G. Wells	<a href="https://www.gutenberg.org/ebooks/36">https://www.gutenberg.org/ebooks/36</a>	0	0.5503
The History of Herodotus - Volume 1	Herodotus	<a href="https://www.gutenberg.org/ebooks/2707">https://www.gutenberg.org/ebooks/2707</a>	0	0.5441
Ritual of the Order of the Eastern Star (84)	Desconocido	<a href="https://www.gutenberg.org/ebooks/61130">https://www.gutenberg.org/ebooks/61130</a>	0	0.531

- 🚦 Sentimiento: bien.
- 🚦 Preferencia: carreras de alta velocidad.

```

Frase del usuario: 'bien'
Sentimiento: 4 stars
Estado de ánimo del usuario: Ni fu ni fa
-----
Preferencia ingresada: carreras de alta velocidad

Frase parseada: carreras de alta velocidad

Traducción: high-speed racing

Entidades encontradas: []

```

Top 3 juegos sugeridos:

name	link	matches	similarity
Downforce	<a href="https://boardgamegeek.com/boardgame/215311/downforce">https://boardgamegeek.com/boardgame/215311/downforce</a>	0	0.6293
Formula Dé	<a href="https://boardgamegeek.com/boardgame/173/formula-de">https://boardgamegeek.com/boardgame/173/formula-de</a>	0	0.6185
Heat: Pedal to the Metal	<a href="https://boardgamegeek.com/boardgame/366013/heat-pedal-to-the-metal">https://boardgamegeek.com/boardgame/366013/heat-pedal-to-the-metal</a>	0	0.5616

Top 3 películas sugeridos:

name	director/author	matches	similarity
The Fast and the Furious: Tokyo Drift	Justin Lin	0	0.4504
Unstoppable	Tony Scott	0	0.4444
Need for Speed	Scott Waugh	0	0.4412

Top 3 libros sugeridos:

name	director/author	link	matches	similarity
The Time Machine	H. G. Wells	<a href="https://www.gutenberg.org/ebooks/35">https://www.gutenberg.org/ebooks/35</a>	0	0.1776
The Practice and Science of Drawing	Harold Speed	<a href="https://www.gutenberg.org/ebooks/14264">https://www.gutenberg.org/ebooks/14264</a>	0	0.1704
The History of Herodotus – Volume 1	Herodotus	<a href="https://www.gutenberg.org/ebooks/2707">https://www.gutenberg.org/ebooks/2707</a>	0	0.1697

## Conclusiones

### Evaluación del Cumplimiento de los Objetivos

- El sistema logró cumplir con los objetivos iniciales, proporcionando recomendaciones basadas en la preferencia de temática del usuario. El clasificador de emociones y el sistema de similitud semántica demostraron ser herramientas efectivas para capturar la intención del usuario.
- La implementación de NER también contribuyó a mejorar la precisión en la búsqueda de coincidencias, aumentando la relevancia de las recomendaciones en función de palabras clave específicas en las frases del usuario.

### Limitaciones y Desafíos

- **Scraping para el dataset de libros:** Dificultades en cuanto a disponibilidad de datos y estandarización del contenido. Sin embargo, se realizaron ajustes para filtrar y limpiar la información, asegurando que los datos fueran utilizables en el sistema de recomendación.
- **Tiempo del procesamiento de las descripciones:** Debido al gran volumen de datos en los tres datasets (películas, juegos de mesa y libros), el tiempo necesario para parsear, obtener emdeddings y entidades nombradas de todo el contenido fue considerable. El parsing involucró el manejo de textos extensos en descripciones, lo cual incrementó el tiempo de procesamiento y el consumo de recursos.
- **Longitud de descripciones:** Algunos textos requerían ajustes adicionales para asegurar que fueran interpretados correctamente por el sistema de recomendación, sin afectar su rendimiento.

### Recomendaciones para Trabajos Futuros

Para futuras implementaciones, sería beneficioso ampliar las opciones de entretenimientos y realizar un ajuste más específico para reconocer y analizar más datos disponibles de cada actividad, teniendo información adicional de las recomendaciones. Por ejemplo, si permite multijugadores, los idiomas disponibles, características especiales, entre otros detalles que puedan ser relevantes para el usuario.

La implementación de un modelo de NLP ajustado a varios idiomas podría mejorar aún más la utilidad del sistema, especialmente en la clasificación de emociones y el reconocimiento de entidades.

Dado el tiempo considerable que requiere el scraping y el parsing de textos largos en los datasets (películas, juegos de mesa y libros), sería recomendable investigar técnicas de optimización para acelerar este proceso, o modelos más efectivos.

De igual forma, mejorar la velocidad y el rendimiento general del sistema, sin comprometer la calidad de la recomendación. Además, se podría evaluar el impacto de la longitud de los textos en el modelo y ajustar los pre procesamiento para mejorar la eficiencia en futuras implementaciones.

## Referencias

Unidad 1 - Extracción y Procesamiento de Texto: <https://gentle-cress-e61.notion.site/Unidad-1-Extracci-n-y-Procesamiento-de-Texto-25f15e55f7ac4e1096376b060df1837f>

Unidad 2 - Representación Vectorial de Texto: <https://gentle-cress-e61.notion.site/Unidad-2-Representaci-n-Vectorial-de-Texto-6ad0dcf3b18a447e8b2989b325b57d3f>

Unidad 3 - Procesamiento del Lenguaje: <https://gentle-cress-e61.notion.site/Unidad-3-Procesamiento-del-Lenguaje-48b3f630e08a49e59bbcabfa39273e0c>

## Anexos

### Código Fuente

El código Python completo, incluyendo el clasificador de emociones, el procesamiento de embeddings, y el scraping para el dataset de libros, se incluye en el archivo adjunto del informe.

### Datos Brutos

Datasets pre procesados y archivos CSV generados durante el scraping también se adjuntan para su referencia y replicación del proceso.

**Link Google Drive:** <https://drive.google.com/drive/folders/1zcbWzUJZDywuySo-YuFyQbqOrgZaFWDI?usp=sharing>