Universidad Nacional de Rosario
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Tecnicatura Universitaria en Inteligencia Artificial
Procesamiento de Imágenes I - IA 4.4

TRABAJO PRÁCTICO N°1

Clasificador de Recomendaciones Recreativas utilizando Procesamiento de Lenguaje Natural (NLP)

Integrantes:

Cima, Nancy Lucía
Sumiacher, Julia

Docentes:

Manson, Juan Pablo
Geary, Alan

Fecha de entrega: 06/11/2024

Índice

Resumen

En este trabajo, se desarrolló un sistema de recomendaciones recreativas orientado a sugerir actividades de entretenimiento en días de mal clima, utilizando técnicas de Procesamiento de Lenguaje Natural (NLP). Este sistema clasifica el estado de ánimo del usuario en categorías como "Alegre", "Melancólico" o "Ni fu ni fa" y, basándose en esta clasificación junto con una frase de preferencia proporcionada por el usuario, genera recomendaciones personalizadas entre películas, juegos de mesa y libros.

La implementación utiliza un clasificador supervisado y técnicas de embeddings para evaluar similitudes semánticas, logrando resultados que demuestran la efectividad de NLP en aplicaciones de recomendaciones, ya que el sistema resultante es capaz de realizar recomendaciones acertadas y adaptadas a los intereses y emociones del usuario.

Introducción

Durante el periodo vacacional, es común enfrentarse a días de mal clima que limitan las actividades al aire libre. En este contexto, surge la necesidad de contar con alternativas recreativas en interiores, adaptadas al estado de ánimo del usuario. Este escenario genera la necesidad de un sistema que ofrezca opciones de entretenimiento en interiores, adaptadas a los intereses y estado de ánimo del usuario, para enriquecer la experiencia vacacional incluso en días de mal clima.

El uso de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) ha permitido grandes avances en la personalización de servicios de recomendación. Mediante técnicas de NLP, es posible analizar tanto el estado emocional del usuario como sus preferencias temáticas, logrando que un sistema de recomendación no solo sugiera opciones, sino que también lo haga de forma adaptada a cómo se siente la persona en un momento particular. En este proyecto, se exploran y aplican estas técnicas para desarrollar un sistema capaz de realizar recomendaciones recreativas personalizadas en función de dos factores clave: el estado de ánimo del usuario y una preferencia ingresada por él.

Este sistema de recomendación utiliza tres fuentes de datos: una base de datos de juegos de mesa, una colección de películas y un conjunto de libros del Proyecto Gutenberg. Estas fuentes fueron elegidas para cubrir un amplio espectro de opciones de entretenimiento en interiores, considerando que podrían ser las actividades preferidas durante días de mal clima. Para ello, el proyecto se centra en desarrollar un clasificador de estado de ánimo y en aplicar técnicas de embeddings, que permiten medir la similitud semántica entre los textos de preferencias del usuario y las descripciones de las opciones recreativas en cada dataset.

El objetivo del trabajo es mostrar cómo el uso de técnicas de NLP, como la clasificación de texto y la similitud semántica, puede facilitar la creación de sistemas de recomendación personalizados.

Este informe documenta cada paso del desarrollo del sistema, desde la recolección y preparación de datos, pasando por la construcción y entrenamiento del clasificador de estado de ánimo, hasta la implementación de las técnicas de NLP para realizar recomendaciones personalizadas.

Metodología

El desarrollo del sistema de recomendaciones se estructuró en varias etapas, desde la recopilación de datos hasta la implementación y validación de modelos de clasificación y recomendación. A continuación, se detallan las mismas.

Fuente y Preparación de los Datos

Para construir un sistema de recomendaciones se utilizaron tres fuentes de datos:

- **Juegos de Mesa**: La base de datos 'bgg_database.csv' de BoardGameGeek (BGG) fue empleada como fuente para las recomendaciones de juegos de mesa. BGG es una plataforma ampliamente reconocida que proporciona información detallada sobre juegos de mesa, incluyendo descripciones, géneros y popularidad.
- **Películas**: Se utilizó la base de datos 'IMDB-Movie-Data.csv', que contiene información de películas de IMDb, una de las plataformas más completas y confiables para información sobre cine. Este dataset incluye detalles como el género, el año de estreno, la sinopsis y la calificación de cada película.
- **Libros**: Los libros fueron seleccionados mediante **web scraping** de la página del Proyecto Gutenberg, específicamente de la sección que lista los 1000 libros más populares. Esta fuente es ideal debido a la amplia disponibilidad de textos de dominio público, lo que permite cubrir una gran variedad de géneros y temáticas. El scraping se realizó primero de los libros más populares, obteniendo el título y otros datos de cada libro, entre ellos un link para ver más información relevante. Al obtener este link, se realizó otro proceso de web scraping sobre la página de cada libro y así obtener su resumen. Algunos de los libros no tenían resumen pero sí un título más extenso, así que de no encontrar resumen, se colocó el título como tal. En caso de tampoco haber título en la página del libro, se incorporó el texto "Resumen no disponible".

Se chequeo la presencia de datos faltantes en los datasets, y solo se noto que la columna "Título Secundario" del dataset de libros tiene una cantidad significativa de datos faltantes. Por esto mismo, esta columna es tenida en cuenta para el programa.

Descripción de Métodos y Técnicas Utilizadas

El desarrollo del sistema implicó una serie de técnicas de NLP y aprendizaje automático, que se pueden agrupar en tres grandes áreas.

Primero, tenemos la clasificación del Estado de Ánimo. Se desarrolló un clasificador supervisado para detectar el estado de ánimo del usuario a partir de un texto corto. Se tomó la decisión de utilizar un modelo de clasificación ya que el análisis de sentimientos se puede ver como un problema de clasificación binaria o multinomial, donde cada texto se asigna a una de las posibles categorías de sentimiento.

Para esto, primero se aplicó TfidfVectorizer para convertir el input del usuario en una representación numérica usando la técnica TF-IDF (Term Frequency-Inverse Document Frequency) y así identificar términos relevantes y reducir el peso de palabras comunes que no aportan mucho al análisis, como las stopwords. Seguido de esto, se aplicó el modelo Multinomial Naive Bayes (MultinomialNB) para clasificar el estado de ánimo del usuario en categorías como “Alegre”, “Triste” o “Ni Fu Ni Fa”. Este modelo fue seleccionado por su simplicidad y efectividad en problemas de clasificación de texto.

Luego, la generación de embeddings y el análisis de la similitud semántica entre ellos. Para comparar la frase de preferencia del usuario con las descripciones de cada recomendación, se emplearon técnicas de embeddings, que convierten textos en vectores de alta dimensionalidad que representan el significado semántico. Específicamente, se usó el modelo Universal Sentence Encoder (USE) para generar embeddings y se calculó la similitud coseno entre vectores. Esto permite medir qué tan parecida es la frase ingresada por el usuario con los textos de los juegos, películas o libros, asegurando que las recomendaciones sean coherentes con sus preferencias. Así, se eligieron las opciones con mayor valor de similitud coseno para cada tipo de entretenimiento.

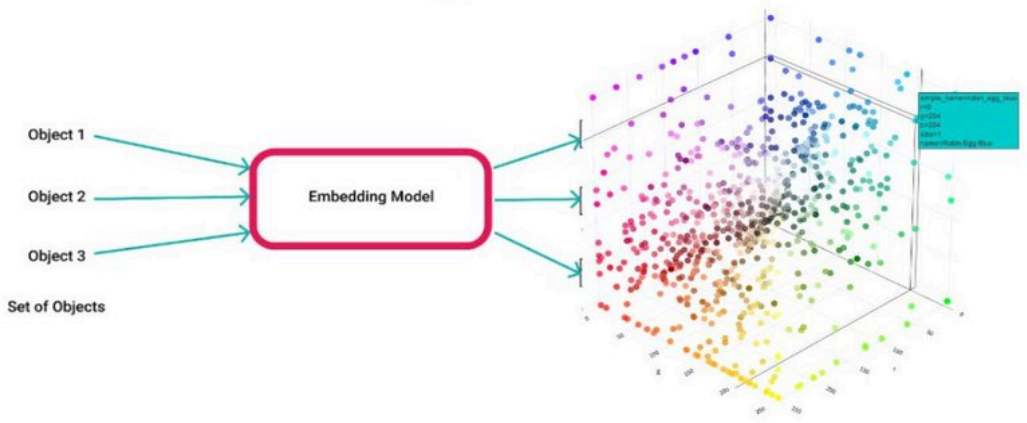

Figura 1: Ilustración del proceso de generación de embeddings.

Cabe destacar que, los datasets se encuentran en inglés por lo que se tomó la decisión de traducir los inputs del usuario a este idioma antes de realizar el embedding. Esta opción nos pareció más óptima dada el volumen de datos de los datasets y a que los inputs suelen ser oraciones cortas y sencillos que pueden traducirse correctamente con mayor precisión. Para dicha traducción, se usó el traductor de Google, que se encuentra disponible desde la librería GoogleTranslator. Otra opción es utilizar la versión multilingüe del Universal Sentence Encoder, pero en este caso se consideró suficiente traducir los inputs del usuario.

Por último, se implementó el reconocimiento de Entidades Nombradas (NER) como técnica complementaria. Se utilizó NER para extraer palabras clave importantes de la frase de preferencia del usuario, lo cual ayudó a ajustar mejor las recomendaciones a las intenciones del usuario. Gracias a esta técnica, se puede identificar si el usuario expresa alguna preferencia, por ejemplo, de actor y sumar a las recomendaciones las que contengan a las entidades recomendadas.

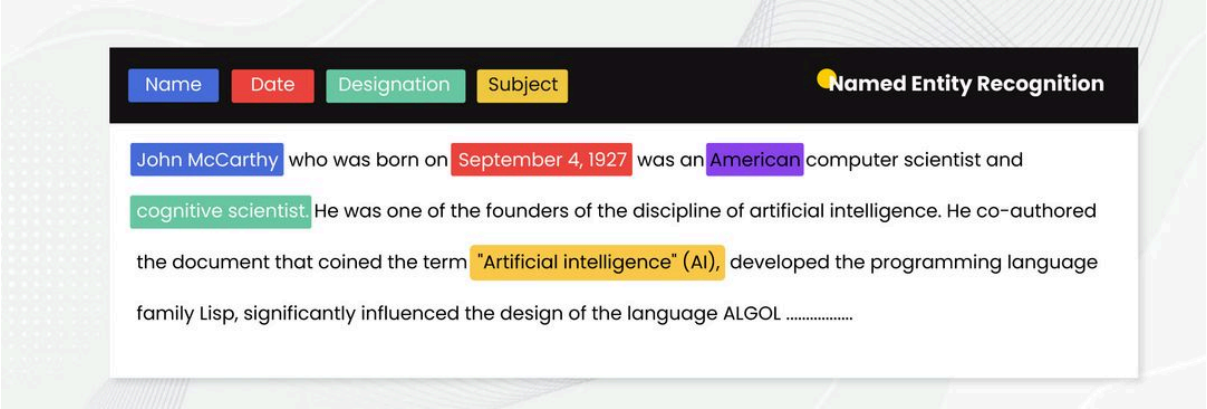

Figura 2: Ilustración del proceso de reconocimiento de entidades nombradas.

Pasos del desarrollo

A continuación, se detallan los pasos seguidos en el desarrollo del sistema de recomendaciones:

1. **Recopilación y Preparación de Datos**:

Se obtuvieron los tres datasets principales. El dataset de libros se obtuvo mediante la técnica de web scraping como se menciona anteriormente y los dataset restantes provienen de archivos en formato csv.

2. **Entrenamiento del Clasificador de Estado de Ánimo**:

Se recopiló un conjunto de datos etiquetados con ejemplos de frases en español que representarán los estados de ánimo requeridos. Este dataset de entrenamiento se utilizó para entrenar el clasificador supervisado, que fue ajustado y evaluado mediante métricas de precisión y recall.

3. **Construcción de Embeddings y Similitud Semántica**:

Se eligió el modelo Universal Sentence Encoder (USE) que tiene como objetivo convertir oraciones completas en representaciones vectoriales de longitud fija. Estos embeddings, pueden capturar en cierto modo, la esencia semántica de la oración.

Este modelo se utilizó para generar embeddings para la combinación de los dos inputs del usuario y para la información de los datasets. Para esto último, se concateno la información de las columnas más relevantes de cada dataset en una nueva columna “Info”. Con esto realizado, se calcularon las similitudes mediante coseno entre embedding correspondiente a los inputs del usuarios y las diferentes opciones de entretenimiento, lo cual permitió obtener recomendaciones acordes a los intereses del usuario.

4. **Implementación del Sistema de Recomendaciones**:

Se integraron las recomendaciones finales basadas en la combinación del estado de ánimo del usuario y los resultados de similitud semántica. De acuerdo con el estado de ánimo detectado, el sistema selecciona entre libros, películas o juegos que tienen una alta similitud con la frase de preferencia. Además, se sumó el reconocimiento de Entidades Nombradas (NER) como segunda capa de recomendación.

5. **Validación del Sistema**:

Se realizaron pruebas con frases de diferentes tipos para verificar la precisión del clasificador y la coherencia de las recomendaciones generadas. Estas pruebas mostraron que el sistema puede capturar el contexto emocional del usuario y proporcionar recomendaciones que se ajustan a sus intereses.

Resultados

El sistema de recomendaciones recreativas desarrollado en este proyecto se probó en múltiples casos de uso para evaluar su efectividad en dos áreas clave: la precisión del clasificador de estado de ánimo y la relevancia de las recomendaciones generadas. A continuación, se detallan los resultados obtenidos en cada fase y su interpretación.

El clasificador de estado de ánimo es el primer componente del sistema. El modelo fue entrenado y evaluado para medir su precisión, exactitud y capacidad de generalización. Los resultados fueron los siguientes:

	precision	recall	f1-score	support
Alegría	1.00	0.69	0.82	13
Ni fu ni fa	0.94	0.94	0.94	17
Tristeza	0.64	0.90	0.75	10
accuracy			0.85	40
macro avg	0.86	0.84	0.84	40
weighted avg	0.89	0.85	0.85	40

Figura 3: Métricas del modelo de análisis de sentimientos.

- **Alegría**:
  - *Precisión*: 100%, lo que indica que todas las predicciones de "Alegría" fueron correctas cuando el modelo identificó esta clase.
  - *Recall*: 69%, reflejando una menor capacidad para identificar correctamente todos los casos de "Alegría" en comparación con las otras clases.
  - *F1-score*: 0.82, evidenciando un rendimiento aceptable, aunque con oportunidades de mejora en el recall.
- **Ni fu ni fa**:
  - *Precisión*: 94%, mostrando un alto nivel de acierto en las predicciones de esta clase.
  - *Recall*: 94%, indicando una excelente capacidad para identificar todos los casos de "Ni fu ni fa" en el conjunto de prueba.
  - *F1-score*: 0.94, con el mejor desempeño general entre las tres clases.
- **Tristeza**:
  - *Precisión*: 64%, lo que sugiere que el modelo tuvo dificultades para identificar adecuadamente todos los casos de "Tristeza".
  - *Recall*: 90%, demostrando una buena capacidad para captar esta clase.
  - *F1-score*: 0.75, destacando un área de mejora en la precisión de la clase.

Métricas Generales

- **Precisión General**: 85%, lo cual indica que el modelo realiza una clasificación acertada en la mayoría de los casos.
- **Macro Promedio (Macro Avg)**: La precisión, recall y F1-score promedio son de 0.86, 0.84 y 0.84 respectivamente, mostrando un balance razonable entre las clases.
- **Weighted Avg**: Estos valores ponderados son similares, con una precisión y F1-score promedio de 0.89 y 0.85, respectivamente.

Aunque el modelo muestra una precisión y recall elevados en general, se observan oportunidades de ajuste en la clase "Tristeza", donde la precisión fue relativamente baja en comparación con las demás clases. Esto sugiere que una mejora futura podría involucrar la incorporación de datos adicionales o técnicas de aprendizaje profundo para comprender mejor las variaciones semánticas en expresiones emocionales más sutiles.

La segunda parte del sistema consiste en generar recomendaciones basadas en la frase de preferencia ingresada por el usuario. El sistema demostró que podía adaptar sus recomendaciones de acuerdo con el estado de ánimo detectado, además de poder detectar bien las entidades presentes en los inputs en el general de los casos.

Estos resultados reflejan que el modelo es efectivo en encontrar similitudes semánticas significativas y puede ofrecer recomendaciones de alta calidad. Sin embargo, algunos aspectos podrían mejorarse, como la mejora del rendimiento del análisis de sentimientos o la incorporación de más capas de sentimientos. Además, se noto que no siempre NER identifica bien las entidades.

También, se puede visualizar el embedding generado para entender cómo las palabras están representadas en un espacio vectorial. Para eso, se utilizó PCA (Principal Component Analysis) como técnica de reducción de dimensionalidad para visualizarlo en un espacio tridimensional.

Aquí se ve que los diferentes tipos de entretenimiento se agrupan de manera muy marcada, con una mayor cercanía entre libros y películas, que con juegos. (Este gráfico se aprecia mejor desde el Colab al poder moverlo y así verlo desde diferentes ángulos).

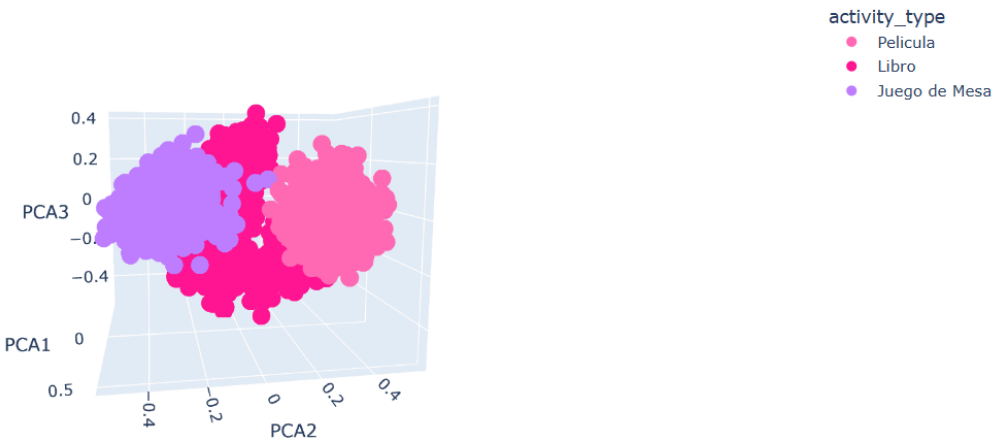

Figura 4: Embeddings 3D coloreados por tipo de actividad.

Conclusiones

En resumen, los resultados obtenidos validan la efectividad del sistema y demuestran que, a través de técnicas de NLP y clasificación, es posible desarrollar un recomendador recreativo que ofrezca sugerencias personalizadas y relevantes. Si bien se identificaron áreas de mejora, los resultados generales muestran que el sistema es capaz de adaptar sus recomendaciones de acuerdo con los estados emocionales y preferencias de los usuarios, enriqueciendo su experiencia de entretenimiento en interiores.

Se proponen como mejorar futuras, ampliar el análisis de sentimientos para una mayor variedad de emociones y también mejorar su precisión.

El sistema desarrollado cumple satisfactoriamente con el objetivo de ofrecer recomendaciones recreativas basadas en el estado de ánimo y las preferencias del usuario. A través del uso de NLP, el sistema puede identificar estados de ánimo y analizar frases para luego generar recomendaciones adaptadas a cada contexto.

Anexos

Como material de apoyo para la realización del trabajo práctico se usaron los apuntes de clase.