Facultad de Ciencias Exactas, Ingeniería y Agrimensura

Tecnicatura en Inteligencia Artificial

Procesamiento del Lenguaje Natural

Trabajo Práctico: Clasificador de Recomendaciones Recreativas.

Alumnos: Kidonakis, Sol K-0624/6
Menescaldi, Brisa M-7184/6

Docentes: Geary, Alan
Manson, Juan Pablo

Fecha de entrega: Miércoles, 06 de Noviembre del 2024.

Índice

Resumen

Este proyecto consiste en desarrollar un sistema de recomendación de ocio para días de mal tiempo en la playa que utiliza procesamiento del lenguaje natural (PLN) para proporcionar actividades basadas en el estado de ánimo del usuario. El objetivo es que el sistema haga sugerencias para ver películas, jugar juegos de mesa o leer libros en función de las emociones y preferencias del usuario.

Para lograr esto, el código implementa un modelo de clasificación que evalúa el estado emocional del usuario en función de señales y lo categoriza como "feliz", "melancólico" o "neutral". Luego se le pide al usuario una oración que describa su interés (por ejemplo, "Historia de amor en la jungla"). Esta entrada se compara semánticamente con tres bases de datos: juegos de mesa (bgg_database.csv), películas (IMDB-Movie-Data.csv) y Proyecto Gutenberg obtenidas mediante web scraping.

Para procesar las preferencias,el sistema utiliza representaciones vectoriales de palabras y cálculos de similitud semántica respaldados por un modelo de reconocimiento de entidades nombradas (NER). El resultado de este programa es una lista personalizada de recomendaciones que se adapta tanto al estado de ánimo del usuario como a sus intereses temáticos. La aplicación fue desarrollada con Google Colab y está disponible en español, pero los datos subyacentes están en inglés.

En resumen, se proporciona un enfoque innovador a las recomendaciones de ocio y demuestra la capacidad de la PNL para personalizar recomendaciones y mejorar la experiencia del usuario durante las vacaciones.

Introducción

El trabajo presentado está dedicado al uso del procesamiento del lenguaje natural (PLN) para desarrollar un sistema de recomendación de entretenimiento diseñado para usuarios que buscan diferentes actividades en función de su estado de ánimo y sus necesidades en el momento del mal tiempo durante las vacaciones. El fin de este proyecto es la necesidad de ofrecer entretenimiento adecuado a situaciones donde el clima limita las actividades al aire libre y donde una experiencia personalizadas pueden aumentar el disfrute del tiempo libre.

Este sistema de recomendación tiene como objetivo principal identificar el estado de ánimo del usuario para proponer actividades recreativas adecuadas, tales como ver películas, jugar juegos de mesa o leer libros. Dónde se procesarán bases de datos en inglés y se adaptarán a una interfaz en español, lo que permite una interacción amigable para el usuario.

Objetivos específicos:

1. Desarrollar un clasificador que identifique el estado de ánimo del usuario.
2. Cree un sistema de búsqueda que, basado en una frase de preferencia, compare intereses con bases de datos de juegos, películas y libros.
3. Aplicar técnicas de embeddings y reconocimiento de entidades para mejorar la relevancia de las recomendaciones.
4. Documentar y estructurar el sistema en un entorno de Google Colab, ofreciendo una interfaz accesible para el usuario.

Metodología

Podemos observar que en este trabajo hemos usado tres conjuntos de datos para realizar recomendaciones de entretenimiento en función del estado de ánimo del usuario y de sus preferencias siendo estos;
- bgg_database.csv : que posee información sobre juegos de mesa, que incluye descripciones y categorías.
- IMDB-Movie-Data.csv : que cuenta con películas, que incluye título, género, actores y descripciones.
- Libros del Proyecto Gutenberg : se hizo una recopilación de los 1000 libros más populares de Proyecto Gutenberg a través de web scraping para obtener información, incluyendo título, autor, descripción, categorías y URL de cada libro.

Para lograr el objetivo principal de la consigna, se compone de un clasificador de estado de ánimo y un sistema de recomendación de entretenimiento.
Clasificación del Estado de Ánimo:
Se clasifican las frases en tres categorías: "Alegre", "Melancólico" y "Ni fu ni fa”.
Se buscó realizar distintos clasificadores para evaluar sus métricas y definir cuál usar. Primero, se probó con un modelo de regresión logística, TF-IDF Vectorizer con un conjunto de datos de 40 frases para cada categoría en dónde se obtuvo como resultado una precisión del modelo del 56% y a su vez identificando que la categoría “Ni fu ni fa” es la que más le cuesta predecir de forma correcta (0.36). Se hizo una prueba del modelo con nuevas frases.
Como segundo clasificador, se creó usando regresión logística con SentenceTransformer donde se utilizan embeddings generados por un modelo de multilenguaje para un conjunto de datos. Dónde se probó con los datos de prueba, obteniendo un 88% de precisión y así obteniendo un mejor valor de métrica al evaluar el modelo; destacando que la categoría “Ni fu ni fa” es la que más le cuesta acertar (0.67) al darle datos nuevos. También se hizo una prueba del modelo con nuevas frases.

Como punto a destacar, siempre se utilizaron el mismo conjunto de datos para los dos modelos de clasificación para así luego hacer una comparación de estos y evaluar mediante la métrica de precisión cual era mejor. En un principio la cantidad de datos no era abundante y los valores eran muy bajos/malos por lo que se optó en ir agrandando dicho tamaño, a su vez se hacía una división 80/20, usando el 80% de los datos para entrenar y el 20% para probar con datos no conocidos y de ahí calcular la evaluación del modelo. Al agregarle valores, llegamos a un total de 40

frases por categoría como tope logrando un buen valor de precisión ya que al querer agregar 50 frases para cada categoría esta métrica empeoraba.
También nos topamos que para el clasificador con SentenceTransformer primero quisimos usar un modelo que estaba entrenado en inglés y por el cual, nos daba un malísimo desempeño. Por eso se terminó optando por uno que tenga variedad de lenguajes.

Debido a las buenas métricas obtenidas, usamos como clasificador del estado de ánimo el modelo que utiliza SentenceTransformer. Ya que los resultados confirmaron que el modelo entrenado con SentenceTransformer ofrecía una mayor precisión en la detección de estados de ánimo que el modelo basado en TF-IDF.

Preferencias:
Se utilizó spaCy con el modelo es_core_news_md para así poder identificar entidades relevantes en las preferencias del usuario, lo que mejora la precisión de las recomendaciones al capturar temas específicos, lugares o personajes históricos y así poder lograr una respuesta personalizada.

Búsqueda de opciones:
Mediante el uso de SentenceTransformer, generamos embeddings para los textos de los diferentes datasets (juegos, películas y libros). Con el objetivo de calcular similitudes entre la preferencia del usuario y las descripciones de los elementos en cada conjunto de datos.

En nuestro caso, usamos Google Colab para lo que fue el desarrollo y la ejecución de código. Usando las siguientes librerías de Python:
• sentence-transformers, para generar embeddings.
• spaCy, para extracción de entidades nombradas (NER).
• nltk, para el manejo de palabras clave.
• ipywidgets para poder crear una interfaz interactiva con el usuario.
• scikit-learn para la clasificación y el cálculo de similitud.

Podemos ver que las recomendaciones de juegos, películas y libros son generadas con base en el estado de ánimo y preferencias específicas brindadas por el usuario, demostrando una alta precisión en la selección de opciones adecuadas. La detección de entidades mejora la precisión al recomendar que coincida con temas específicos.

Desarrollo/Implementación

Así como se implementó el sistema de clasificación de estados de ánimo y el sistema de recomendación basado en procesamiento de lenguaje natural (PNL). Procedemos a dar explicaciones sobre los algoritmos, modelos de IA y técnicas de PNL empleadas, así como una descripción del flujo de trabajo general.

Preparación de los datos
Se comienza con la recopilación y el procesamiento de datos. Utilizamos tres conjuntos de datos diferentes para juegos de mesa, películas y libros:
Libros, se encontraban en la página Project Gutenberg donde se extrajeron los 1000 libros más populares realizando web scraping con BeautifulSoup para extraer los títulos, autores, descripciones y categorías. Implementando la función get_soup que facilita la conexión con la URL y extraer_info que extrae toda la información que quiero de forma estructurada en un diccionario y así estos datos se almacenan en un archivo CSV al que llamamos gutenberg_books.csv.
Pero a este archivo creado, debemos realizarle una limpieza ya que nos encontramos con datos duplicados y filas con valores nulos en todos los conjuntos de datos.
Juegos de mesa, también realizamos modificaciones en donde se reemplazaron caracteres especiales y frases que se encontraban en las descripciones para mejorar su legibilidad.
Películas.

Clasificación del Estado de Ánimo:
Mediante la función predict_mood_st() la cual recibe una frase de entrada y mediante esta, predice el estado de ánimo usando el modelo de regresión logística entrenado utilizando SentenceTransformer con el modelo paraphrase-multilingual-MiniLM-L12-v2p ,el cual capta de una mejor manera las relaciones semánticas entre palabras.
Y así podemos clasificar estados de ánimo con mayor precisión, siendo que se obtuvo un 88% de precisión.

Extracción de Entidades Nombradas
Para lograr que las recomendaciones sean más precisas, utilizamos spaCy para realizar reconocimiento de importancias en la frase de preferencia del usuario. Así poder permite detectar nombres de lugares, personas y conceptos específicos.

Cuando el usuario ingresa sus preferencias mediante la función ingresar_preferencias, la función extraer_entidades identifica términos específicos, ya sea de personas o lugares, para mejorar el contexto de la recomendación.

Generación de Recomendaciones:
Tras limpiar los datos, generamos embeddings para cada entrada (juegos, películas y libros) utilizando SentenceTransformer. Estas permiten calcular similitudes semánticas entre las preferencias del usuario y los elementos disponibles en cada categoría.

Luego se define la función recomendar que toma en cuenta tanto el estado de ánimo como las preferencias del usuario, y busca elementos en los datos que tengan la mayor similitud con la entrada del usuario.
Se comparará la entrada del usuario con las de juegos, películas y libros, utilizando la similitud coseno. Y luego, se devuelve las tres recomendaciones más relevantes en cada categoría.
Para evitar recomendaciones duplicadas, específicamente en la categoría de libros, se agregó un verificador de unicidad donde nos aseguramos que los títulos seleccionados sean únicos. Siempre con el objetivo de lograr una recomendación personalizada y relevante para cada usuario.

Interfaz de usuario
Finalmente, se crea una interfaz interactiva en donde ipywidgets permite al usuario ingresar su estado de ánimo y preferencia. Y la función mostrar_interfaz()gestiona las entradas y presenta las recomendaciones en un formato claro y accesible.

Resultados

Para evaluar el rendimiento del clasificador de estado de ánimo, probamos dos enfoques: TF-IDF con regresión logística y embeddings con SentenceTransformer.

Modelo	Precisión
TF-IDF + Regresión Logística	56%
SentenceTransformer + Regresión Logística	88%

Al decir precisión, nos estamos refiriendo al porcentaje que el modelo predice de forma correcta para datos nuevos.
También podemos ver cual fue la categoría que mejores métricas obtuvo.

TD-IDF

```
Reporte de clasificación Regresión Logística:
              precision    recall  f1-score   support

           0       0.71      0.62      0.67         8
           1       0.71      0.45      0.56        11
           2       0.36      0.67      0.47         6
```

SentenceTransformer

```
Reporte de clasificación Regresión Logística:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         8
           1       1.00      0.73      0.84        11
           2       0.67      1.00      0.80         6
```

Siendo la categoría 0 “Alegre”, 1 “Melancolico” y 2 “Ni fu ni fa”; podemos ver notablemente que a los modelos les cuesta predecir correctamente cuando es “Ni fu ni fa”.

Conclusiones

El sistema de recomendaciones recreativas basado en procesamiento de lenguaje natural (NLP) ha logrado clasificar efectivamente el estado de ánimo del usuario y generar recomendaciones de juegos, películas y libros relevantes a las preferencias ingresadas. Los principales resultados incluyen una precisión del 88% en la clasificación de estados de ánimo utilizando incrustaciones con SentenceTransformer que detecta con gran precisión tres categorías emocionales (Alegre, Melancólico, Ni fu ni fa), y una precisión significativa en las recomendaciones de calidad gracias a la extracción de entidades y el cálculo de similitud de coseno. Estas técnicas han permitido que el sistema identifique adecuadamente contenidos que correspondan tanto al estado emocional como a la temática de interés del usuario.
Por lo que podemos decir que se cumple con los objetivos planteados.