Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Universidad Nacional de Rosario

Procesamiento del Lenguaje Natural
Trabajo Práctico No. 1

05/11/2024

Docentes: Juan Pablo Manson, Alan Geary
Julieta Lopez Ceratto (L-3311/1)
Giuliano Crenna (C-7438/1)

Contenidos

Resumen

Este informe presenta el desarrollo de un sistema de recomendación recreativa, diseñado para sugerir opciones de entretenimiento en función del estado de ánimo del usuario. Utilizando técnicas de **Procesamiento de Lenguaje Natural** (NLP), el sistema clasifica el estado de ánimo del usuario y, a partir de una frase de preferencia, sugiere películas, juegos de mesa o libros relevantes. Se aplicaron modelos de clasificación y métodos de embeddings para detectar similitudes semánticas en bases de datos de juegos, películas y libros. Los resultados demuestran que el sistema es capaz de realizar recomendaciones acertadas y adaptadas a los intereses y emociones del usuario.

1 Introducción

Durante las vacaciones, el mal tiempo puede limitar las opciones de entretenimiento al aire libre, generando la necesidad de alternativas recreativas. Este proyecto propone el desarrollo de un sistema de recomendación basado en **Procesamiento de Lenguaje Natural** (NLP), que utiliza el estado de ánimo del usuario y una frase de preferencia para sugerir actividades recreativas. El objetivo es diseñar un modelo que clasifique emociones y recomiende películas, juegos o libros. Este informe documenta el proceso de construcción del sistema, las metodologías empleadas y los resultados obtenidos.

2 Metodología

2.1 Datos Utilizados

Para el desarrollo del sistema se utilizaron las siguientes fuentes de datos:

- **bgg_database.csv**: Base de datos de juegos de mesa.
- **IMDB-Movie-Data.csv**: Base de datos de películas.
- **Proyecto Gutenberg**: Dataset generado mediante web scraping, obteniendo información de los 1000 libros más populares.

2.2 Clasificación del Estado de Ánimo

Se aplicó el modelo *Multinomial Naive Bayes* (MultinomialNB) para clasificar el estado de ánimo del usuario en categorías como “Alegre”, “Melancólico” o “Ni Fu Ni Fa”. Este modelo fue seleccionado por su simplicidad y efectividad en problemas de clasificación de texto.

2.3 Embeddings y Similitud Semántica

Para la búsqueda de coincidencias, se utilizaron embeddings generados con *Universal Sentence Encoder Multilingual*, que permite comparar descripciones en inglés con frases en español. Esto facilitó la recomendación semántica sin necesidad de traducir datos o consultas.

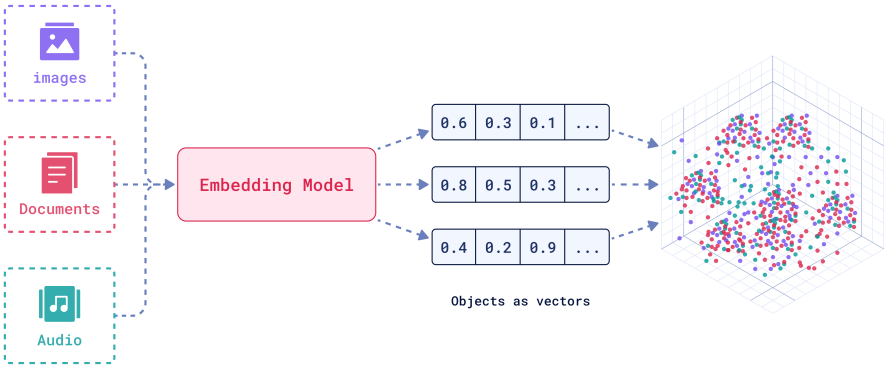

Figure 1: Representación gráfica de embeddings.

2.4 Modelo de Recomendación

Se seleccionó el modelo *k-Nearest Neighbors* (KNN) para realizar recomendaciones personalizadas. Este modelo basa sus recomendaciones en la similitud entre puntos de datos, ideal para sugerir elementos similares a las preferencias expresadas por el usuario.

3 Justificacion de Decisiones Tomadas

3.1 Utilizacion de MultinomialNB

Este modelo fue elegido para la clasificación del estado de ánimo del usuario ya que se trata de un modelo con dos características muy importantes: simpleza y efectividad.

Se vio que los resultados en cuanto a la clasificación eran más que aceptables ya que el análisis de sentimientos se ve como un problema de clasificación binaria o multinomial, donde cada texto se asigna a una de las posibles categorías de sentimiento.

Además, tenía bajos niveles de requerimientos en lo que respecta a tiempo y cómputo (dos factores esenciales para nuestro equipo y en general para cualquier persona que quiera implementar un modelo).

Como otro punto importante, es un modelo que funciona bien en altas dimensiones, cualidad inherente al procesamiento del lenguaje natural.

"El modelo Naive Bayes Multinomial fue seleccionado debido a su rapidez, simplicidad y eficacia demostrada en tareas de clasificación de texto, como el análisis de sentimientos." Navarro Bellido, J. (2023). Modelos de aprendizaje automático en análisis de sentimiento: Comparativa de rendimiento (Trabajo de fin de grado, p. 37). Universitat Politècnica de València, Escuela Técnica Superior de Ingeniería Informática.

3.2 Modificación del Dataset de Libros

Se decidió modificar el dataset de libros obtenido mediante web scraping debido a varias razones importantes:

- **Desbalance de Clases**: La cantidad de datos en el dataset de libros era considerablemente mayor en comparación con los datasets de películas y juegos, lo cual creaba un desbalance de clases. Aproximadamente, el dataset de libros contaba con 4000 filas, lo cual podía sesgar las recomendaciones hacia libros y limitar la diversidad y precisión de las recomendaciones.

- **Limitaciones de Cómputo**: Dado que el proyecto requiere la generación de embeddings de las descripciones textuales para medir la similitud semántica entre la frase ingresada por el usuario y los elementos de los datasets, el tamaño del dataset de libros resultaba inadecuado considerando las limitaciones de cómputo y tiempo del equipo. Reducir la cantidad de datos facilitó una implementación más rápida y eficiente.

- **Descripciones Nulas**: Muchos de los libros carecían de descripciones completas o eran nulas. La recomendación depende de la similitud semántica entre el texto de entrada del usuario y las descripciones de los elementos; por lo tanto, la falta de información descriptiva hacía que ciertos libros fueran menos relevantes y menos útiles en el sistema de recomendaciones.

Estas modificaciones resultaron en un dataset más balanceado y manejable, optimizando el rendimiento del modelo y la relevancia de las recomendaciones generadas.

3.3 Encoder elegido

Se eligió universal-sentence-encoder-multilingual. Esto se debe a que la información de los dataset se encontraba en inglés y el usuario se comunica en idioma español. De esta forma se evitaba una elección de vecino por asociación del lenguaje y no de la semántica.

Además simplificaba el proceso. No se eligió traducir en ninguno de los dos sentidos (de ingles a español el dataset o de español a inglés la consulta) ya que es de amplio conocimiento cómo se pierde parte de la intención del usuario y del significado del texto mediante las traducciones, problema inherente a la ambigüedad del lenguaje.

3.4 ¿Por qué KNN?

Se eligió este modelo ya que:

1. KNN, al basarse en la distancia entre puntos, es ideal para recomendaciones personalizadas. Si la frase del usuario describe una temática o emoción específica, KNN puede encontrar elementos similares en la base de datos que respondan a esos mismos temas o estilos, proporcionando recomendaciones que encajan mejor con los gustos del usuario.

2. Capacidad para Manejar Multicategoría (Películas, Libros, Juegos): , puede realizar recomendaciones cruzadas, sugiriendo tanto películas como libros o juegos que coincidan con la frase de entrada del usuario, lo que permite un enfoque multimodal.

3. Capacidad para Recomendar Nuevos Elementos sin Reentrenamiento Completo: Una ventaja de KNN es que, al no requerir un proceso de entrenamiento extenso, permite añadir nuevos elementos al conjunto de recomendaciones sin necesidad de reconstruir o reentrenar un modelo completo. Al agregar nuevas películas, libros o juegos con sus embeddings, puede incluirlos de inmediato en sus recomendaciones.

4. No necesita suposiciones de normalidad o linealidad en los datos, lo cual es una ventaja en un problema de recomendación donde la relación entre los gustos del usuario y el contenido es altamente no lineal y depende de la similitud semántica, no de funciones específicas entre los datos.

3.5 Elecciones de columnas para embeddings

No se eligió la columna relacionada al título o nombre ya que nuestra visión es que el usuario busque por aquello que tiene ganas de ver y no por un título (no habría una recomendación realmente si busca por título).

Se eligieron columnas de "descripción" o "resúmen" es decir que describen el elemento, y otras propias de cada categoría: para libros autor, películas género y para juegos cantidad de jugadores. Además se agrega el tipo de elemento ya sea película, libro o juego dado que es importante si el usuario expresa que quiere alguno en especifico o relacionado (si menciona leer, jugar, película, etc.)

4 Desarrollo o Implementación

La implementación del sistema se realizó en Python, empleando bibliotecas de NLP como `scikit-learn` y `TensorFlow`. A continuación, se describen los pasos principales:

1. **Preprocesamiento de Datos**: Se limpiaron y organizaron los datasets, eliminando datos nulos y descripciones repetidas.

2. **Clasificación de Emociones**: Se entrenó el modelo MultinomialNB utilizando un conjunto de datos etiquetado para determinar la categoría de emoción.

3. **Generación de Embeddings**: Se aplicó el modelo Universal Sentence Encoder Multilingual para convertir descripciones y frases de usuario en vectores numéricos.t

4. **Recomendación de Elementos**: Se utilizó KNN para buscar elementos con embeddings similares a los de la frase ingresada por el usuario.

5 Resultados

Los resultados indican que el sistema es capaz de identificar el estado de ánimo del usuario con un nivel aceptable de precisión. Las recomendaciones generadas a partir de las frases de preferencia reflejan una alta coherencia con el sentimiento y temática deseados. A continuación, se presentan algunos ejemplos de recomendaciones obtenidas:

- Para el estado de ánimo "Alegre" y la frase "aventura en el mar", se recomendaron películas y libros de temática similar.

- Para el estado de ánimo “Melancólico” y la frase “historia de amor”, el sistema sugirió opciones alineadas con la emoción y género.

Podemos asimismo observar cómo se agrupan de manera marcada cada clase mediante PCA.

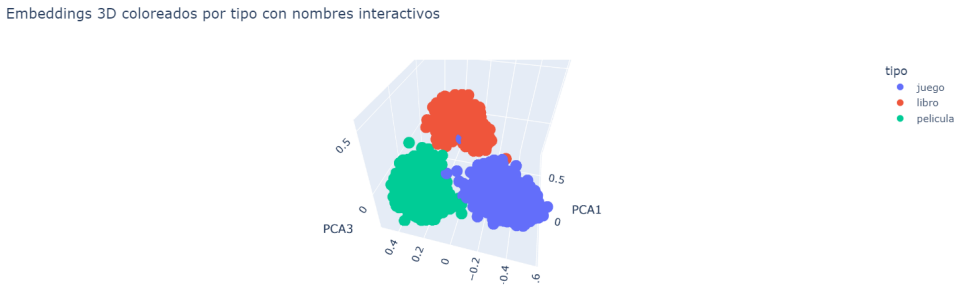

Figure 2: Visualización mediante PCA de las clases

6 Conclusiones

El sistema de recomendación recreativa desarrollado demuestra la efectividad del uso de NLP en la personalización de recomendaciones en función del estado de ánimo. La combinación de clasificación de emociones, embeddings multilingües y KNN permite generar opciones relevantes y diversas. Sin embargo, se podrían implementar mejoras en la clasificación de emociones mediante el uso de modelos más complejos. Además, futuras investigaciones podrían explorar la adaptación del sistema para otros idiomas y una mayor variedad de actividades recreativas.

7 Referencias

Navarro Bellido, J. (2023). Modelos de aprendizaje automático en análisis de sentimiento: Comparativa de rendimiento (Trabajo de fin de grado, p. 37). Universitat Politècnica de València, Escuela Técnica Superior de Ingeniería Informática. Recuperado de: https://m.riunet.upv.es/bitstream/handle/10251/198111/Navarro%20-%20Modelos%20de%20aprendizaje%20automatico%20en%20analisis%20de%20sentimiento%20comparativa%20de%20rendimiento.pdf?sequence=1&isAllowed=y