

TRABAJO PRACTICO FINAL

Ulises Nemeth

Procesamiento de Lenguaje Natural.

Profesores:

Juan Pablo Manson

Alan Geary

Constantino Ferrucci

Dolores Sollberger

Fecha de entrega: 18/12/24

Indice

1. [Resumen General](#)
2. [Ejercicio 1](#)
 - 2.1. [Resumen](#)
 - 2.2. [Introducción](#)
 - 2.3. [Metodología](#)
 - 2.4. [Desarrollo e Implementación](#)
 - 2.5. [Resultados](#)
 - 2.6. [Conclusiones](#)
3. [Ejercicio 2](#)
 - 3.1. [Resumen](#)
 - 3.2. [Introducción](#)
 - 3.3. [Metodología](#)
 - 3.4. [Desarrollo e Implementación](#)
 - 3.5. [Resultados](#)
 - 3.6. [Conclusiones](#)
4. [Referencias](#)
5. [Anexos](#)

1. Resumen General

En este trabajo se presenta la tarea (y los conflictos) de aplicar los conocimientos adquiridos al intentar desarrollar un chatbot y un Agente basados en la técnica **RAG (Retrieval Augmented Generation)** para el juego de mesa "The White Castle". Hay tres fuentes principales de datos: documentos de texto, datos tabulares y bases de datos de grafos. A través de una combinación de herramientas y modelos de PLN, se implementó un clasificador que permite determinar la fuente adecuada para generar respuestas precisas. Se compararon dos versiones del clasificador (basado en LLM y embeddings), optimizando la eficiencia y relevancia en la recuperación de información. Además, se diseñó un agente ReAct para combinar las capacidades del sistema. Al final, los resultados obtenidos, una interpretación acerca de los logros en contraste con la consigna y las limitaciones identificadas.

2. Ejercicio 1

2.1 Resumen

En este apartado se va a desarrollar el webscrapping en dos partes ya que el contenido se extrajo de páginas web y videos. Luego, en tres partes, la utilización de los archivos creados para crear y llenar las fuentes de datos y las problemáticas que surgieron. Al final los clasificadores, retrievers y ReRank y una demostración del chatbot utilizando todos los recursos.

2.2 Introducción

Este primer ejercicio abarca casi la totalidad del trabajo, esto se va a ver reflejado en este informe. Si bien mientras el trabajo estaba en desarrollo no seguía un orden, mientras lo estaba terminando acomodé las cosas para que esta parte sea más ordenada.

Contexto y justificación: Los chatbots con RAG mejoran la generación de respuestas al integrar diversas fuentes de datos. Se busca crear un chatbot que responda preguntas sobre reglas, estrategias y estadísticas del juego The White Castle.

Objetivos específicos:

1. Implementar RAG con múltiples fuentes de datos.
2. Desarrollar dos clasificadores y comparar sus resultados.
3. Optimizar la búsqueda de información mediante queries dinámicas.

Estructura del informe: Se describen los métodos utilizados, el proceso de implementación y los resultados obtenidos.

Metodologia

Pasos seguidos:

WebScrapping:

La obtencion de texto y datos de forma automatizada aprovechando los recursos publicados online es una forma poco costosa de acceder a grandes volúmenes de informacion.

En este trabajo se ve un ejemplo de como en segundos se pueden obtener los subtitulos de algun video de youtube que tenga una gran profundidad sobre un tema especifico. Gracias al PLN se puede conseguir procesar estos datos y darles un proposito sin necesidad de consumirlo por completo de forma tradicional.

Tambien en el desarrollo se consigue el texto de parrafos en paginas web gracias al HTML, mirando el codigo fuente de sitios relacionados al juego The White Castle, se pueden obtener todo el contenido necesario omitiendo partes no tan relevantes.

Obtener datos de forma estructurada (clave valor) para llenar una fuente de datos tabular SQLite.

Obtener datos de forma estructurada para generar luego las tríadas e insertarlas en la base de datos de grafos

Fuentes de datos:

Estan categorizadas por su contenido, de esta forma se pueden filtrar mejor los datos para dar una respuesta precisa a la hora de interactuar con el usuario.

- **Documentos:** Reglas del juego, reseñas y tutoriales.
- **Tablas:** Estadísticas del juego y datos relevantes (SQLite).
- **Grafos:** Relaciones entre diseñadores y elementos del juego.

Selección y Análisis de Fuentes

- **Documentos Textuales:** Reseñas y tutoriales desde youtube y paginas web ([ver referencias](#))
- **Datos Tabulares:** Estadísticas de bgg.
- **Base de Datos de Grafos:** Relación entre diseñadores, mecánicas y artistas modelada con **Neo4j**.

Procesamiento de Datos

- **Segmentación de Textos:** Uso de **LangChain** para dividir textos en chunks.
- **Limpieza de Datos:** Eliminación de redundancias y normalización de texto.
- **Embeddings:** Generación de vectores con OpenAI.
- **Datos Tabulares y Grafos:** Consultas dinámicas utilizando SQL y Cypher.

Herramientas Utilizadas

- **LangChain** para textos.
- **ChromaDB** para almacenamiento vectorial.
- **Neo4j** para grafos.
- **Pandas/SQL** para tablas.

Desarrollo/Implementación

Instalación de librerías

Webscrapping

Creacion de fuentes de datos

Creacion de clasificadores

Queries dinamicas

Busqueda hibrida

Chatbot

Resultados

1. Presentación de resultados:

- Ejemplos de consultas que requieren más de una fuente (documento + tabla).
- **Gráficos y tablas:** Comparación de tiempo de respuesta y precisión entre clasificadores.

2. Errores comunes:

- Ejemplos de preguntas donde el agente falló, como inconsistencias en queries de grafos.

3. Análisis:

Por la poca cantidad de tiempo disponible para el trabajo, no conseguí hacer pruebas exhaustivas, pero sin dudas el objetivo fue cumplido

Ejemplos de Consultas:

Pregunta: ¿Quién diseñó *The White Castle*?

Respuesta: Basado en el contexto proporcionado, hay dos diseñadores mencionados:

1. Isra C.
2. Shei S.

La pregunta "¿Quién es el diseñador?" es un poco ambigua ya que hay dos diseñadores en la lista. Si necesitas información específica sobre uno de ellos, necesitaría que me indiques cuál es el diseñador que te interesa. ¿Podrías aclarar tu pregunta?

Casos de Error:

Cuando el modelo tiene un contexto sin muchos datos, por ejemplo, con resultados a la pregunta anterior, que solamente le llega la query y dos nombres, suele indicar que no está seguro de que responder, ya que el contexto queda muy acotado por la naturaleza de los resultados de fuentes de datos tabulares.

Conclusiones

1. Resumen de logros:

Implementación exitosa del chatbot con RAG.

2. Evaluación del cumplimiento de objetivos.

No conseguí comparar los clasificadores como para mostrar un benchmark, termine utilizando el LLM porque conseguí que se adapte mas facilmente a las fuentes de datos, ya que desde que hice los clasificadores hasta la entrega, estas cambiaron mucho.

3. Recomendaciones:

Dedicarle mas tiempo al desarrollo del trabajo.

Ampliar las fuentes de datos para enriquecer las respuestas e implementar mejor las herramientas y recursos creados.

3. Ejercicio 2

3.1 Resumen

Se extendió el chatbot a un **agente interactivo** basado en ReAct utilizando **Llama-Index**. El agente combina tres herramientas (`doc_search`, `graph_search`, `table_search`) para responder consultas complejas.

Esta parte del trabajo quedo incompleta para esta primer entrega.

3.2 Introducción

Objetivo del Trabajo

Desarrollar un agente que razone y actúe utilizando múltiples herramientas dinámicamente.

Justificación

La combinación de herramientas mejora la capacidad del sistema para resolver consultas más complejas.

3.3 Metodología

3.4 Desarrollo e Implementación

3.5 Resultados

3.6 Conclusiones

- Logros: -
- Limitaciones: -
- Futuras Mejoras: -

4. Referencias.

<https://misutmeeple.com/2023/11/resena-the-white-castle/> *Reseña: The White Castle.*

<https://pandas.pydata.org/> *Pandas.*

<https://boardgamegeek.com/boardgame/371942/the-white-castle/> *Boardgamegeek.*

<https://youtu.be/UDupg6a1GRM> *How to play The White Castle - Devir.* Devir Games channel

<https://www.langchain.com> *LangChain. LangChain Documentation.*

<https://chromadb.com> *ChromaDB. ChromaDB Documentation.*

<https://neo4j.com/> *Neo4j. Graph Database Technology.*

<https://huggingface.co/Qwen> *Hugging Face. Qwen Models*

[How to play The White Castle boardgame - Full teach + Visuals - Peaky Boardgamer](#) *How to play The White Castle boardgame - Full teach + Visuals - Peaky Boardgamer. Peaky Boardgamer channel*

5. Anexos

1. Descripción del Eurogame: *The White Castle*

Este juego combina mecánicas de colocación de trabajadores y manejo de recursos, ambientado en la construcción y administración de un castillo japonés. Los jugadores compiten por optimizar su estrategia y obtener la mayor cantidad de puntos de victoria.

Recursos sobre el juego:

- **BoardGameGeek:** Plataforma principal con información detallada sobre el juego, estadísticas, reglas, reseñas y foros de discusión.
Enlace: <https://boardgamegeek.com/boardgame/>
Secciones clave:
 - **Files:** Archivos con manuales, hojas de ayuda y resúmenes de reglas.
 - **Forum:** Discusión de reglas, estrategias y aclaraciones de dudas.
 - **Stats:** Estadísticas que pueden ser usadas para análisis tabular.
- **Análisis en blogs especializados:** Se utilizaron entradas de blogs con reseñas y análisis sobre *The White Castle*.

- **Videos de reseñas:** Plataforma como YouTube ofrece revisiones del juego, disponibles en varios idiomas. Se consiguió subtítulos de varios videos para enriquecer las fuentes textuales.
-