Chatbot para a documentación e normativa da UDC

Marcelo Ferreiro Sánchez
Marcos Grobas Martínez
José Romero Conde

10 de decembro do 2025

Resumo

O obxectivo do proxecto é desenvolver un chatbot capaz de solventar dúbidas acerca do funcionamiento dos procesos burocráticos e de documentación da Universidade da Coruña (UDC). Para conseguilo optamos por unha arquitectura xenerativa aumentada por recuperación (RAG), técnica moi empleada para mellorar o desempeño de chatbots baseados en LLM cando buscan información específica dun dominio sobre o que o modelo de linguaxe orixinal non foi entrenado.

Índice xeral

1. Introdución

A burocracia de calquera campo pode chegar a ser moi complexa e pode chegar a consumir moito tempo e recursos ás persoas que teñen que lidiar con ela. Os procesos universitarios non son unha excepción e moitas das persoas involucradas neles (tanto alumnos como profesores e persoal administrativo) os poder atopar inabarcables ou imposibles de navegar sen axuda.
Neste contexto, apreciouse como podería ser de gran utilidade o desenvolvemento dun chatbot capaz de responder a preguntas relacionadas coa documentación e normativa da Universidade seguindo a gran tendencia da actualidade de empregar chatbots para diversas tarefas.
O obxectivo deste proxecto é desenvolver un chatbot que poida simplificar e explicar **referindo sempre ás fontes burocráticas oficiais** os procesos burocráticos e de documentación da Universidade da Coruña (UDC).
O sistema é resultado da unión de compoñentes e técnicas xa ben coñecidas, sempre tendo en mente o obxectivo a cumprir. Polo tanto, tratouse máis ven dunha tarefa de aplicación e adaptación de técnicas e ferramentas xa existentes nun caso particular, antes que de deseño ou resolución de novos problemas. Se ben este traballo está circunscrito no contexto da asignatura de Técnicas Avanzadas de Procesamiento de Linguaxe Natural (TAPLN), debido á súa natureza e obxectivo, gran parte do tempo invertido no proxecto dedicouse á tarefa de recolección de información para o RAG.
Ao non existir unha 'base de datos' oficial da UDC sobre a que un usuario poda descargar toda a documentación relativa ao centro, senón que atópase repartida nas páxinas web dos seus diferentes centros, foi necesario deseñar unha solución que nos permita extraela a partir dos seus portais oficiais. Este tipo de tarefas non son novas no mundo da informática, de feito pertencen a unha área máis que consolidada chamada Recuperación de Información (IR) e tales métodos que buscan, extraen e organizan información disposta en webs html son coñecidos como *crawlers*.

2. Solución proposta

ostrarase nesta sección a arquitectura xeral do sistema e posteriormente describirase cada un dos seus compoñentes, sendo estos principalmente o *crawler* e o *RAGsystem*

2.1 A Arquitectura

2.2 O *Crawler*

2.3 O *tf-idf*

2.4 Os modelos

2.5 Ferramentas usadas

3. Instalación e uso

4. Resultados

 esults presentation here.

5. Conclusions


ola castro