

CHATBOT PARA A DOCUMENTACIÓN E NORMATIVA DA UDC

Marcelo Ferreiro Sánchez
Marcos Grobas Martínez
José Romero Conde

19 de decembro do 2025

Resumo

O obxectivo do proxecto é desenvolver un chatbot capaz de solventar dúbihdas 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	2
2. Solución proposta	3
2.1. A Arquitectura	3
2.2. O <i>Crawler</i>	3
2.2.1. Selección de páxinas relevantes	3
2.2.2. Procesado dos contidos de cada páxina	4
2.2.3. Funcionamento interno do <i>Crawler</i>	6
2.3. Análise dos datos obtidos mediante o <i>Crawler</i>	7
2.4. O <i>RAGsystem</i>	9
2.4.1. Construcción da base de vectores	9
2.4.2. O sistema de recuperación	10
2.5. Os modelos	11
2.6. Ferramentas usadas	11
3. Instalación e uso	12
4. Resultados	13
5. Conclusions	14
A. Máis información acerca do <i>crawler</i>	16
A.1. <i>Keywords</i> utilizadas polo crawler	16
A.2. Saída OCR da táboa de exemplo	18
A.3. Progresión de páxinas <i>retrieveadas</i> ao longo do crawleo	18
B. Análise do Corpus	20
B.1. Distribución de tipos de ficheiros	20
B.2. Palabras más frecuentes	21
B.3. Palabras menos frecuentes	22
B.4. Resultados validación <i>RAGsystem</i>	22
B.4.1. <i>Heatmaps</i> de varias métricas	22
B.4.2. <i>Surface plots</i> de varias métricas	23
B.4.3. Comparación de métricas para diferentes valores de perso de BM25 co mellor <i>chunk size</i> (2048)	24

1. Introducción



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 (sexan, alumnos, profesores ou persoal administrativo) os poden atopar inabarcables ou imposibles de navegar sen axuda.

Neste contexto, aprecioouse 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 cumplir. Polo tanto, tratouse máis ben 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.

Non existindo unha 'base de datos' oficial da UDC sobre a que un usuario poda descargar toda a documentación relativa ao centro, senón que atópase esta 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 más que consolidada chamada Recuperación de Información (IR) e tales métodos que buscan, extraen e organizan información disposta en webs html son coñecidos como *crawlers*, parte esencial do traballo final pois para asegurar que o sistema sexa capaz de responder á maioría de dúbidas dos usuarios, é necesario ter un corpus de toda información mínimamente relevante acerca do funcionamento interno da universidade.

2. Solución proposta



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

2.1 A Arquitectura

Fundamentalmente e como indicouse antes, o sistema compõe de dúas partes relativamente independientes: *RAGsystem*, que será un modelo de linguaxe grande (LLM) xa adestrado que recibirá xunto con cada consulta de usuario, unha serie de documentos (ou fragmentos dos mesmos) para responder mellor á mesma, e por outra parte, un sistema de *crawling* que ocuparase de navegar as diversas páxinas e portais da UDC recolectando aquelas que considere que poden conter información relevante.

Nas siguientes subseccións explicaranse en detalle o funcionamento interno de cada parte e cómo interactúan entre elas. Optouse por unha orde de exposición que siga o fluxo de execución do sistema, é dicir, comezando polo *Crawler* e seguindo polo *RAGsystem*. Sopesouse seguir a orde cronolóxica do desenvolvemento do sistema mais concluíuse que introduciría demasiadas exégesis e reviraría demasiado e sen utilidade a narrativa desta memoria. De calquera forma, explicarase a evolución das partes que máis cambios sufrieron ao longo da súa implementación.

2.2 O *Crawler*

Considerouse que a posesión dun bo volume de datos sería crucial para resolver o problema a tratar, na maioría dos casos as dúbidas burocráticas resólvense encontrando o documento adecuado, e de non telo, o sistema veríase obrigado a comunicarlle ao usuario que non ten acceso á tal ou cal información, voltándose completamente inútil. É por iso que a Recuperación de Información xoga un papel crucial para o RAG.

2.2.1 Selección de páxinas relevantes

Sendo así, o primeiro problema a solucionar é o de deseñar un bo criterio do que é un **documento relevante** para a aplicación. Se ben tal tarefa pódese complicar *ad infinitum* e é un área de estudo activa aínda nestes días [Pezzuti, MacAvaney e Tonellotto 2025], neste caso optouse por un enfoque moito más sinxelo, baseándose primariamente na premisa de que esta tarefa non require de conxuntos masivos de datos. Só é necesario navegar un certo número de URLs (principalmente do directorio raíz de cada facultade),

non a totalidade da web, polo que pódese permitir descargar páxinas e documentos potencialmente pouco relevantes. A información a colleitar ten un límite de tamaño finito e razonable: nun escenario extremo, mesmo descargando todas as URLs da UDC, estaríase lonxe de requirir terabytes de almacenamento, como sí acontecería nun *crawler* de propósito xeral.

Deste xeito, a solución proposta sitúase como un método arcaico aínda que funcional, unha asignación de relevancia binaria (relevante ou non relevante) mediante *keywords*. Este enfoque, se ben simple, remite aos métodos de *focused crawling* pioneiros [Chakrabarti, Berg e Dom 1999] que sentaron as bases da recuperación de información temática na web.

As palabras clave utilizadas son fixas e propias do ámbito académico e do vocabulario burocrático, aparecendo tanto en galego, coma en castelán e inglés (véxase o Apéndice A.1 para a lista completa).

2.2.2 Procesado dos contidos de cada páxina

Unha vez unha páxina estímase como relevante, é necesario procesar seu contido, para este caso, procurar presentalo en texto plano, evitando ao máximo posible decoradores. Mais nunha páxina atópase máis que texto, especialmente neste caso de uso, moita información burocrática útil para un usuario final atoparse nun documento PDF ligado ao URL, é por iso que o *crawler* tamén os procesa e converte en texto plano listo para que posteriormente un LLM poda leelos sen maior complicación.

Non obstante, este enfoque presenta unha limitación importante: información que se atope únicamente en imaxes (capturas de pantalla, carteis dixitalizados ou documentos escaneados) permanece invisible para o *crawler*.

Tal debilidade do sistema fixose obvia nas primeiras fases de teste do *chatbot*, pois este non era capaz de responder a unha pregunta moi sinxela e básica para os estudiantes: "Cantos libros pódense emprestar?". Tras investigar a casuística e comprobarse que o URL onde figura tal información foi efectivamente analizada polo crawler, caéuse na conta de que, se ben a información figura nunha táboa dentro da páxina, esta está en formato imaxe [ver 2.1], polo que próbase que é necesario aproveitar mellor a información de cada páxina para cumplir os obxectivo establecido.

Tipos de usuarios	Nº de documentos en préstamo	Días de préstamo	Renovaciones	Reservas
GRUPO 1				
Estudiantado de Grado de centros propios y adscritos				
Estudiantado de programa de movilidad (Erasmus, Sigue-Séneca)				
Estudiantado de doble Grado, de simultaneidad				
Estudiantado de grados interuniversitarios				
Estudiantado da Universidad Sénior				
GRUPO 2				
Estudiantado de MÁSTERES e posgrados propios, incluyendo los de la Fundación Universidade da Coruña				
Estudiantado de Trabajos de fin de grado				
Personal de administración en servicio (PTGAS)				
GRUPO 3				
Estudiantado de Doctorado	35	30 días	Indefinidas	Límite 35 docs.
Personal investigador visitante: visitante senior/ visitante predocente o postdoctoral	35	30 días	Indefinidas	Límite 35 docs.
GRUPO 4				
Becarios/as de investigación				
Personal contratado investigador	35	Curso académico (cada biblioteca podrá excluir de este tipo de préstamo documentos por razón de uso y disponibilidad)	Indefinidas	Límite 35 docs.
GRUPO 5				
PDI da UDC (incluyendo centros adscritos), de la Fundación UDC				
Profesorado emérito, jubilado incentivado y honorario				
Profesorado visitante				
Lectores/as				
GRUPO 6				
Cualquier persona ajena a la comunidad universitaria de la UDC que sea autorizada por la Biblioteca Universitaria	6	10 días	Indefinidas	Límite 6 docs.

Figura 2.1 Táboa de préstamos

Procesado de imaxes mediante OCR

O recoñecemento óptico de caracteres (OCR polas súas siglas en inglés) é unha técnica que permite extraer texto a partir de imaxes dixitais, se ben o seu uso máis típico é a dixitalización de documentos físicos escaneados, serve para calquera imaxe que conteña caracteres, como é o caso.

Neste caso, implementouse mediante a librería *pytesseract* unha interface en *Python* do motor OCR Tesseract, creado inicialmente por HP e mantido actualmente por Google.

Procesar cada imaxe presente nun URL pode semexar moi custoso computacionalmente, mais a extracción de texto mediante OCR é verdadeiramente rápida, isto sumado a que a maioría de imaxes son irrelevantes e de moi pequeno tamaño, otorga a capacidade de poder procesar todas as presentes en calquera documento sen apenas engadir latencia ao sistema. Como exemplo, o tempo de procesado de OCR na imaxe da táboa de préstamos [2.1] foi de 0,6280 segundos.

Demostrado que é posible e relativamente sinxelo á vez que barato procesar arquivos de imaxe, é necesario agora demostrar que tal texto é fidedigno ao realmente existente na mesma.

Un exemplo que demostra á perfección un caso de uso real é precisamente a táboa mencionada anteriormente, o texto extráido en crú pódese ver no apéndice A.2. Revisándoo, o texto extráido non asemixa nada parecido á táboa orixinal, mais hai que ter en mente que non é necesario que sexa intelixible para o usuario final, senón para o LLM que o debe de analizar. Unha forma rápida de comprobar que o chatbot poida ser

capaz de entender o texto extraído (coa complicación de que neste caso está en formato táboa) é pedirlle mediante un prompt no portal online dalgún LLM comercial, que refaga a táboa por nós con esa información. Fíxose tal experimento usando ChatGPT-5 e a táboa que reconstruiu foi a seguinte:

Grupo	Tipo de usuarios	Nº docs	Días	Renov.
1	Estud. Grao (centros propios/adscritos), Mobilidade, Dobre Grao, Graos interuniv., Univ. Sénior	10	10 días	Indef. (Lím. 10)
2	Estud. másteres/posgraos propios (Fund. UDC), Estud. TFG, PTGAS	15	21 días	Indef. (Lím. 15)
3	Estud. Doutoramento, Inv. visitante/senior, Inv. predoc./posdoc	35	30 días	Indef. (Lím. 35)
4	Becarios inv., Pers. contr. inv. (<i>Exclusión posible</i>)	35	Curso acad.	Indef. (Lím. 35)
5	PDI UDC (inc. adscritos), Fund. UDC, Prof. emérito/xub./hon., Prof. visitante, Lectores (<i>Exclusión posible</i>)	100	Curso acad.	Indef. (Lím. 100)
6	Persoal externo á UDC, Pers. autoriz. Biblioteca	6	10 días	Indef. (Lím. 6)

Cadro 2.1 Táboa reconstruída mediante LLM (ChatGPT-5)

Se ben existe certa información que o LLM obviou por ser incompleta, a información importante está presente, demostrando que os LLMs son capaces de reconstruir e interpretar a información extraída mediante OCR, incluso cando esté representada en formatos mais complexos como o é unha táboa.

2.2.3 Funcionamento interno do *Crawler*

A implementación proposta usa da librería *Python BeautifulSoup* para navegar a rede e extraer información das páxinas. O proceso comeza cunha lista de URLs semente que representa os portais principais da universidade (concretamente a *homepage* de cada facultade). Para cada URL, o sistema descarga o contido HTML, extrae todos os enlaces e imaxes presentes na páxina, e identifica documentos PDF que conteñan termos relacionados coa burocracia universitaria segundo a lista de *keywords* antes exposta. Os recursos descargados almacénanse de forma organizada no sistema de ficheiros local, mentres que un sistema de metadatos rexistra información sobre cada recurso para optimizar futuras execucións, concretamente garda o *etag* e a data de última modificación de cada páxina, ademáis da data de descarga.

En canto ás imaxes, aplícase OCR a todas as presentes en cada documento e decídese se gardar o texto extraído ou non segundo se presenta máis dun mínimo de caracteres. Todo este proceso realiza ao final de crawlear cada páxina, os arquivos de imaxe vanse engadindo a unha pila, e esta é procesada cando xa extraéronse as hiperreferencias a outras páxinas e os documentos PDF.

Para acelerar o proceso de *crawling* aplicáronse técnicas de programación concurrente. É posible establecer un número de traballadores aos cales se lles asignará unha URL

como base desde onde iniciar o crawler, deste xeito, é posible ter múltiples crawlers simultáneamente navegando a rede da universidade. Para non xerar problemas de lectura/escritura nos ficheiros onde se escriben os metadatos ou o rexistro de páxinas visitadas, implementáronse *locks*, deste xeito dous *crawlers* non podrán escribir á vez nun mesmo documento. Implementar esta técnica foi de gran utilidade, pois a navegación e descarga de páxinas web vai fancéndose máis lenta segundo o tempo de *crawling*, pois é mais difícil atopar páxinas relevantes non visitadas, no apéndice A.1 pódense consultar gráficas acerca dun crawleo que o confirman.

2.3 Análise dos datos obtidos mediante o *Crawler*

Tras un facer un *crawl* sobre as *homepages* das facultades e escolas da UDC, obtiveronse un total de 4966 documentos, dos cales 2668 son páxinas *HTML* e 2298 son documentos *PDF* (véxase B.1).

Debido a que esta práctica está tan enlazada coas técnicas de Recuperación da Información clásicas, pareceu interesante realizar unha análise máis exploratoria típica no sector. En concreto estudar a distribución das palabras do vocabulario e lonxitude dos documentos:

Cadro 2.2 Estadísticas del corpus

Métrica	Valor
Arquivos procesados	4.966
Total de caracteres	75.088.461
Total de palabras	10.164.265
Tamaño del vocabulario	73.640
Palabras únicas (<i>hapax legomena</i>)	15.889

Atopamos que o *corpus* obtido alcanza os 10 millóns de palabras, cun vocabulario de 73.640 palabras ([2.2]). Chama a atención o elevado número de *hapax legomena*, estas son aquelas palabras cunha única instancia na colección completa, un 20 % do vocabulario pertence a esta clase, o cal pode dar a pensar que unha gran cantidade de elas son faltas de ortografía (xa sexan debidas ao erro humano de quen as escribiu ou na técnica de OCR).

Se ben podería ser así, un estudo posterior mostra o contrario, sen embargo, só 230 de tales palabras son faltas ortográficas, e 30 delas foron xeradas polo OCR.

O fenómeno de que unha porcentaxe tan elevada do vocabulario teña tan pouco uso non é exclusivo deste corpus, senón que é unha propiedade universal dos idiomas naturais descrita pola lei de Zipf. Esta lei establece que a frecuencia dunha palabra é inversamente proporcional ao seu rango na distribución de frecuencias. En outras palabras, un pequeno número de palabras moi frecuentes coexiste cunha longa cola de palabras raras, como se pode apreciar na Figura ?? e 2.3.

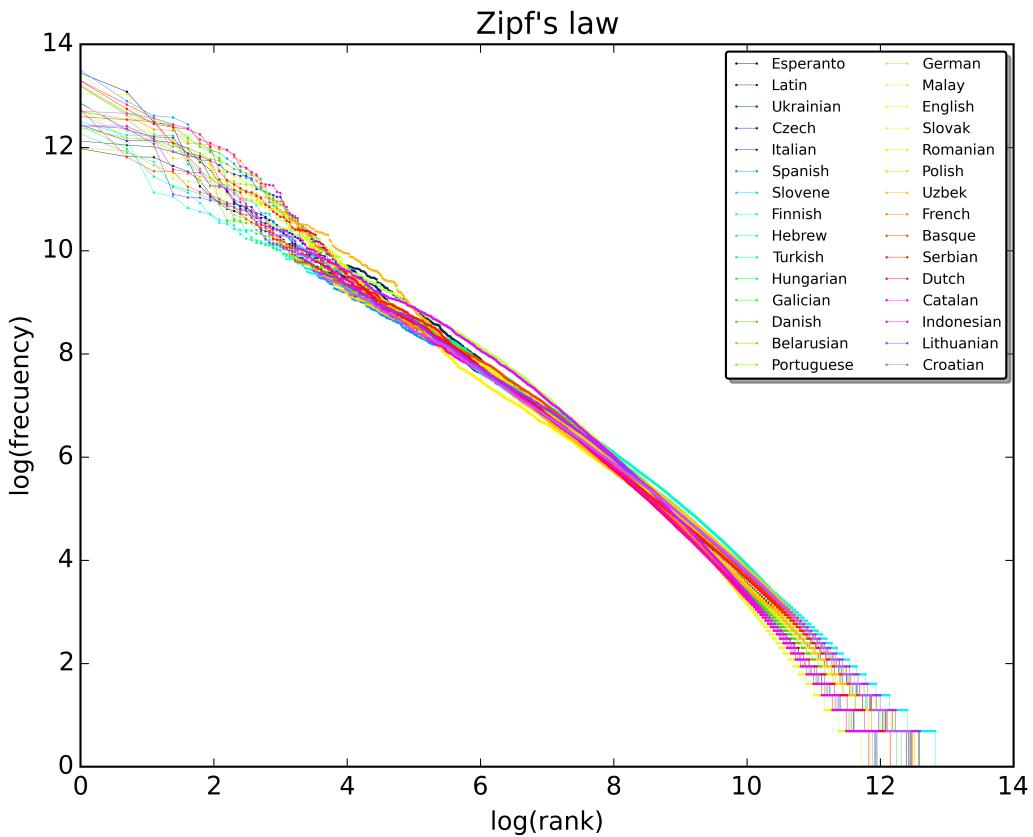


Figura 2.2 Frecuencia de termos en diversas lingüas [Wikipedia 2024]

Ademais, cúmprese o principio de Pareto: aproximadamente o 20 % das palabras mais frecuentes do vocabulario cobren o (aproximadamente) 80 % das ocorrencias totais, mentres que para alcanzar o 80 % das ocorrencias só se necesita o 20 % do vocabulario (Figura 2.3). No caso deste corpus este principio e aínda más esaxerado e estremo, chegaríamos a ese 80 % só co 2.1 % do vocabulario, e se o extenderamos ao 20% o *coverage* sería dun 97.5 %.

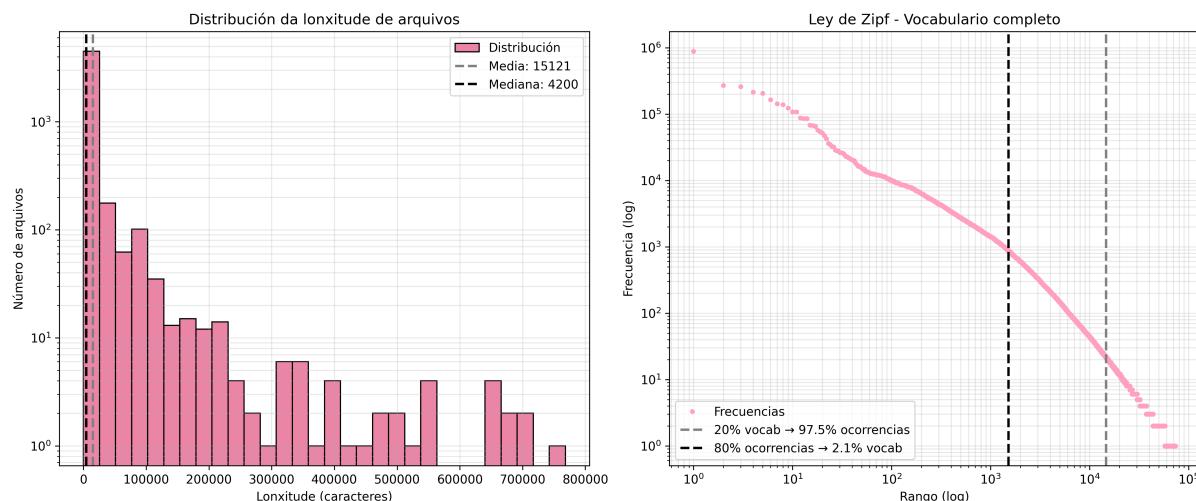


Figura 2.3 Gráficos de distribución de lonxitude de documentos e frecuencia de termos

A distribución das lonxitudes dos documentos (Figura 2.3, esquerda) revela unha gran cantidade de documentos moi curtos. A mediana sitúase en 4200 caracteres, mentres que a media alcanza os 15121 caracteres, o que indica unha distribución asimétrica con presenza de documentos extensos que elevan a media. Esta heteroxeneidade na lonxitude non é de extrañar pois existen moitas páxinas web cun contido moi limitado que só mostran unha serie de ligazóns ou documentos para descargar. De feito, algunas URLs teñen a única finalidade de mostrar unha ou varias imaxes sen texto asociado. No presente *crawleo* atopáronse 115 páxinas completamente baleiras (0 caracteres), ben por conter únicamente imaxes sen texto extraíble por OCR, ben por tratarse de páxinas de redirección ou navegación pura.

2.4 O *RAGsystem*

O sistema RAG (Retrieval-Augmented Generation) implementado constitúe o núcleo do chatbot, combinando técnicas de recuperación de información con modelos de linguaxe xerativos. O proceso de resposta a unha consulta desenvólvese en varias etapas: primeiro, a pregunta do usuario convértese nun vector mediante embeddings semánticos que capturan o seu significado; a continuación, este vector empregase para buscar no vectorstore os documentos más relevantes mediante similitude coseno, recuperando os k fragmentos de texto más próximos semanticamente á consulta. Para mellorar a calidade da recuperación, o sistema incorpora un compoñente de reranking baseado en TF-IDF que permite reordenar os documentos recuperados segundo a súa relevancia léxica, ou ben combinarse de forma híbrida coas puntuacións vectoriais mediante pesos configurables. Unha vez identificados os fragmentos más relevantes, estes incorpóranse como contexto nunha plantilla de prompt xunto co historial recente da conversación, permitindo que o modelo de linguaxe (Claude, ChatGPT ou calquera compatible con *Langchain*) xere unha resposta fundamentada na documentación oficial da universidade. O sistema mantén un historial de conversación de lonxitude configurable que permite responder a preguntas que fan referencia a interaccións previas, dotando ao chatbot de capacidade contextual e conversacional. Ademais, cada documento recuperado enriquecése con metadatos que inclúen a súa puntuación de relevancia tanto vectorial como TF-IDF, facilitando a trazabilidade e verificación das fontes empregadas para xerar cada resposta.

2.4.1 Construcción da base de vectores

[Lewis et al. 2020] A base de vectores (vectorstore) constrúese a partir dos documentos recollidos e previamente convertidos a texto plano (.txt) polo *crawler*. Cada documento é fragmentando despois en varios *chunks* de tamano configurable, á vez que se establece un *overlap* que especifica cantos tokens comparten os fragmentos consecutivos, súa utilidade é impedir que existan referencias cruzadas dentro dun texto non se vexan cortadas por terminarse o chunk, xa que nese caso o LLM non podería

entender o contexto completo.

Para o almacenamento e indexación eficiente dos embeddings, emprégase FAISS (*Facebook AI Similarity Search*) [Johnson, Douze e Jégou 2019], unha biblioteca optimizada para a busca de similaridade en espazos vectoriais de alta dimensionalidade. FAISS implementa algoritmos de busca aproximada de veciños más próximos (ANN, *Approximate Nearest Neighbors*), que permiten realizar consultas en millóns de vectores de forma eficiente. O proceso funciona do seguinte xeito: cada chunk de texto convértese nun vector denso (embedding) mediante un modelo de linguaxe, estes vectores almacénanse nunha estrutura de índice optimizada, e cando se realiza unha consulta, FAISS calcula rapidamente os vectores más similares utilizando medidas de distancia como a similaridade coseno ou a distancia euclidiana [Douze et al. 2024].

2.4.2 O sistema de recuperación

O sistema polo que o RAG recibe os documentos relevantes para cada consulta recibiu múltiples iteraciones e cambios tras varias fases de testeo e replanteamento teórico. Trátase dun compoñente clave do sistema e non ten unha solución trivial, pois é un área de investigación moi activa [Author e Author 2025], poderíase dicir que se ben a área de IR parecía estar algo estancada nos últimos anos, debido aos bons resultados que alcanzaron os buscadores web, agora volve a verse moi activa grazas ao xurdimento dos RAGsystems.

Comezou empregando un sistema de recuperación baseado exclusivamente en similitude coseno de vectores FAISS. Mais tras comprobar que tal *approach* podía non dar os resultados esperados, pois da a mesma importancia a todas as palabras da consulta, non ten en conta cal pode ser a palabra clave máis relevante. A maioría das consultas (especialmente burocráticas) van ter moitas palabras moi comúns e repetidas ao longo da colección de documentos (proceso, documento, normativa) que poden facer que o sistema recupere documentos pouco relevantes.

A primeira solución implementada foi aplicar MMR (Maximal Marginal Relevance) no canto da similaridade coseno. O fundamento detrás desta idea era que deste xeito non desenvolverianse documentos moi redundantes entre si, xa que no caso de existir varios documentos moi relevantes entre si, súa información útil para a consulta sería menor, mais durante a experimentación comprobouse que na realidade non se cumplía tal suposición.

É moi común ver diferentes documentos oficiais que repiten a mesma información ou varias versións e revisións dun proceso en diferentes documentos. En tal caso é probable que o sistema devolva un só deles, perdéndose información relevante ou directamente indicando un proceso obsoleto, xa que prioriza diversidade ante exhaustuvidade [Carbonell e Goldstein 1998]. Por outra banda, esta problemática acerca da vixencia temporal dos documentos é unha das maiores problemáticas neste proxecto (e en calquera RAG que manexe información cambiante ao longo do tempo [Grofsky 2025]) e tratarase máis adiante, xa que require modificación en diversas partes do sistema.

Sabendo que o MMR mom é axeitado ao noso dominio, decidiuse volver cara o enfoque orixinal, pero esta buscando a maneira de discriminar mellor entre os términos relevantes e irrelevantes, existe un método que consegue estes resultados desde fai tempo, o TF-IDF (Term Frequency - Inverse Document Frequency) [Spärck Jones 1972]. A forma de aplicalo inicialmente foi utilizalo para rerankear os documentos recuperados polo FAISS, é dicir, cos n documentos devoltos mediante a similaridade coseno, estes reordéanse segundo a ponderación tf-idf. Prontamente caeuse na conta de que deste xeito non obtense resultados demasiado óptimos pois simplemente cambia de orde os documentos retrieveados polo FAISS, mais non engade novos documentos que poidan ser relevantes e que a similaridade coseno obvia. Por iso, a solución final foi empregar un sistema híbrido, onde se combinan as puntuacións de similitude coseno e tf-idf mediante pesos configurables (50/50 no caso actual). Tal enfoque é moi común nos RAGsystems actuais como no caso de Author et al. 2025 e Sharma et al. 2024.

2.5 Os modelos

2.6 Ferramentas usadas

3. Instalación e uso

4. Resultados



esults presentation here.

5. Conclusions

 ola castro

Bibliografía

- Author, A. e B. Author (2025). "A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions". En: *arXiv preprint arXiv:2507.18910*. Consultado: Diciembre 2025.
- Author, A. et al. (2025). "A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts". En: *arXiv preprint arXiv:2502.16767*. Consultado: Diciembre 2025.
- Carbonell, Jaime e Jade Goldstein (1998). "The use of MMR, diversity-based reranking for reordering documents and producing summaries". En: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98*. New York, NY, USA: ACM, pp. 335–336.
- Chakrabarti, Soumen, Martin Van den Berg e Byron Dom (1999). "Focused crawling: a new approach to topic-specific Web resource discovery". En: *Computer Networks* 31.11-16, pp. 1623–1640. DOI: [10.1016/S1389-1286\(99\)00052-3](https://doi.org/10.1016/S1389-1286(99)00052-3).
- Douze, Matthijs et al. (2024). *The Faiss library*. <https://github.com/facebookresearch/faiss>. Accessed: 2024-12-15.
- Grofsky, Mark (2025). "Solving Freshness in RAG: A Simple Recency Prior and the Limits of Heuristic Trend Detection". En: *arXiv preprint arXiv:2509.19376*. Consultado: Diciembre 2025.
- Johnson, Jeff, Matthijs Douze e Hervé Jégou (2019). "Billion-scale similarity search with GPUs". En: *IEEE Transactions on Big Data* 7.3, pp. 535–547.
- Lewis, Patrick et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks". En: *Advances in Neural Information Processing Systems* 33, pp. 9459–9474.
- Pezzuti, Francesca, Sean MacAvaney e Nicola Tonello (xul. de 2025). "Neural Prioritisation for Web Crawling". En: *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*. ICTIR '25. ACM, 307–314. DOI: [10.1145/3731120.3744597](https://doi.org/10.1145/3731120.3744597). URL: <http://dx.doi.org/10.1145/3731120.3744597>.
- Sharma, A. et al. (2024). "Domain-specific Question Answering with Hybrid Search". En: *arXiv preprint arXiv:2412.03736*. Consultado: Diciembre 2025.
- Spärck Jones, Karen (1972). "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". En: *Journal of Documentation* 28.1, pp. 11–21. DOI: [10.1108/eb026526](https://doi.org/10.1108/eb026526).
- Wikipedia (2024). *Ley de Zipf*. Wikipedia, La enciclopedia libre. URL: https://es.wikipedia.org/wiki/Ley_de_Zipf (accedido en 15/12/2025).

A. Más información acerca do *crawler*

A.1 *Keywords* utilizadas polo *crawler*

A continuación móstrase o conxunto completo de palabras clave utilizadas polo *crawler* para determinar a relevancia das páxinas:

- | | | |
|------------------|-----------------|-----------------|
| ■ regulation | ■ calendario | ■ resolución |
| ■ reglamento | ■ syllabus | ■ circular |
| ■ normativa | ■ programa | ■ instruccións |
| ■ procedure | ■ requirements | ■ instrucciones |
| ■ procedimiento | ■ requisitos | ■ bases |
| ■ proceso | ■ regulamento | ■ anexo |
| ■ form | ■ regulación | ■ catalog |
| ■ formulario | ■ procedemento | ■ catalogo |
| ■ solicitud | ■ solicitude | ■ catálogo |
| ■ guideline | ■ guía | ■ library |
| ■ guia | ■ docente | ■ biblioteca |
| ■ manual | ■ asignatura | ■ collection |
| ■ policy | ■ política | ■ colección |
| ■ politica | ■ matrícula | ■ colección |
| ■ norma | ■ inscripción | ■ acquisition |
| ■ enrollment | ■ académico | ■ adquisicion |
| ■ matricula | ■ convocatoria | ■ adquisición |
| ■ inscripcion | ■ prazo | ■ loan |
| ■ administrative | ■ prazos | ■ prestamo |
| ■ administrativo | ■ documentación | ■ préstamo |
| ■ academic | ■ tramite | ■ reserve |
| ■ academico | ■ trámite | ■ reserva |
| ■ calendar | ■ ordenanza | ■ interlibrary |

- interbibliotecario ■ bibliographic ■ autorizacion
- reference ■ bibliografico ■ autorización
- referencia ■ bibliográfico ■ notification
- circulation ■ holdings ■ notificacion
- circulacion ■ fondos ■ notificación
- circulación ■ serials ■ registration
- periodical ■ publicacions seriadas ■ registro
- periodico ■ special collections ■ protocol
- periódico ■ coleccions especiais ■ protocolo
- journal ■ reading room ■ statute
- revista ■ sala de lectura ■ estatuto
- archive ■ stacks ■ ordinance
- archivo ■ depósito ■ decree
- arquivos ■ microfilm ■ decreto
- repository ■ microficha ■ resolution
- repositorio ■ digital library ■ resolucion
- classification ■ biblioteca dixital ■ official
- clasificacion ■ opac ■ oficial
- clasificación ■ marc ■ office
- indexing ■ application ■ oficina
- indexacion ■ deadline ■ department
- indexación ■ plazo ■ departamento
- cataloging ■ documentation ■ service
- catalogacion ■ documentacion ■ servicio
- catalogación ■ certification ■ servizo
- dewey ■ certificado ■ unit
- isbn ■ certificación ■ unidad
- issn ■ authorization ■ unidade

A.2 Saída OCR da táboa de exemplo

Texto OCR

[Tipos de usuarios
Nº de documentos en préstamo
Días de préstamo
Renovaciones
Reservas
GRUPO 1
Estudiantado de Grado de centros propios y adscritos
Estudiantado de programa de movilidad (Erasmus, Sicue-Séneca)
Estudiantado de doble Grado, de simultaneidad 10 10 días Indefinidas | Límite 10 docs.
Estudiantado de grados interuniversitarios Estudiantado da Universidad Séntor GRUPO 2
Estudiantado de MASTERES e posgrados propios, incluyendo los dela Fundación Universidade da Coruña . , 15 21 días Indefinidas | Límite 15 docs. Estudiantado de Trabajos de fin de grado. Personal de administración en servicio (PTGAS) GRUPO 3 Estudiantado de Doctorado 35 30 días Indefinidas | Límite 35 does. Personal investigador visitante; visitant tant 'ersonal investigador visitante: visitante senior] visitante 35 30 días indefinidas | Límite 38 docs predoctoral o postdactoral GRUPO 4 Becarios /as de investigación Curso académico Personal contratado investigador (cada biblioteca podrá excluir 35 de este tipo de préstamo Indefinidas | Límite 35 docs. documentos par razón de uso y disponibilidad) GRUPO 5 PDI da UDC (incluyendo centros adscritos), de la Fundación UDC Curso académico Profesorado emérito, jubilado incentivado y honorario (cada biblioteca podrá excluir Profesorado visitante 100 de este tipo de préstamo Indefiridas | Límite 100 docs. Lectores /as documentos par razón de uso y disponibilidad) GRUPO 6 Cual ! idad taria de la UDC cualquier persona ajena a la comunidad universitaria dela 6 10 días indefinidas | Límite 6 docs que sea autorizada por la Biblioteca Universitaria

A.3 Progresión de páxinas *retrieveadas* ao longo do crawleo

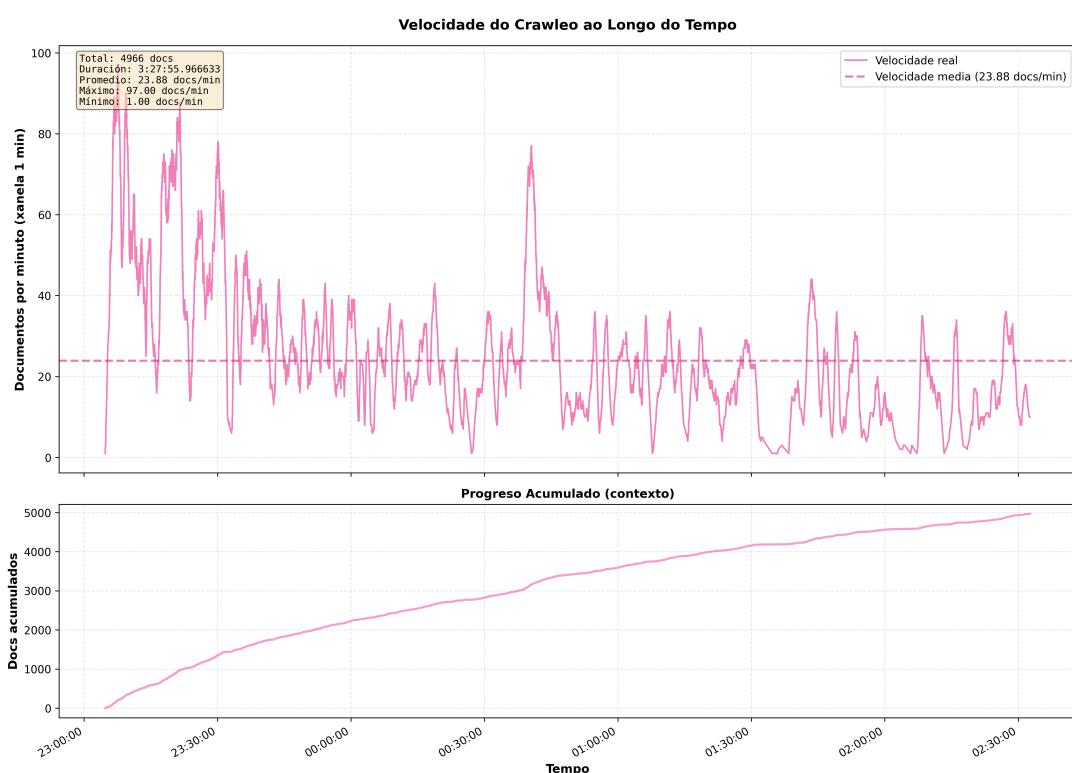


Figura A.1 Progresión da eficiencia do crawler ao longo do tempo

B. Análise do Corpus

B.1 Distribución de tipos de ficheiros

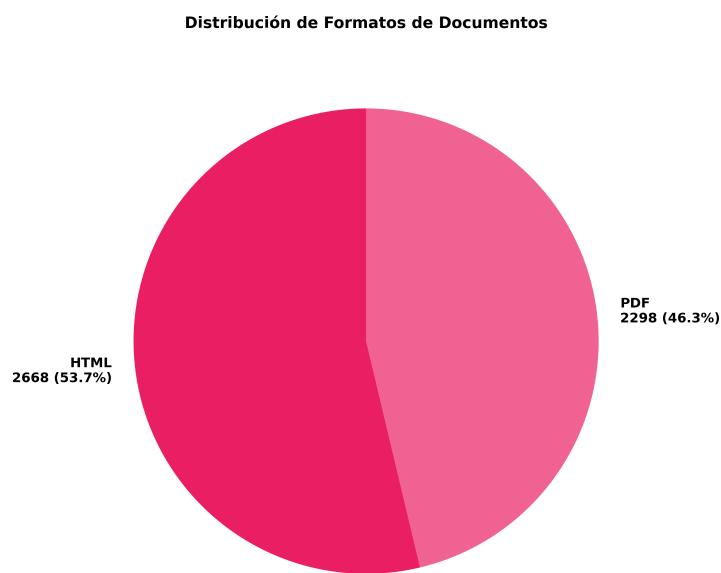


Figura B.1 Distribución de tipos de ficheiros no corpus

B.2 Palabras más frecuentes

Pos.	Palabra	Frecuencia
1	de	385,703
2	a	113,823
3	en	102,598
4	la	102,014
5	e	90,196
6	y	77,587
7	que	73,069
8	o	57,323
9	da	49,975
10	el	48,951
11	para	43,204
12	do	40,254
13	se	35,387
14	los	35,200
15	del	33,793
16	las	29,507
17	por	27,315
18	no	25,683
19	con	25,326
20	udc	25,182

Cadro B.1Top 20 palabras más frecuentes no corpus

B.3 Palabras menos frecuentes

Pos.	Palabra	Frecuencia
1	ricondo	- 1
2	acompañamiento	- 1
3	recomendaciéns	- 1
4	aproximacién	- 1
5	nosum	- 1
6	climent	- 1
7	vengut	- 1
8	empar	- 1
9	retransmision	- 1
10	toxicoloxia	- 1
11	landeira	- 1
12	angelines	- 1
13	psicoloxicos	- 1
14	mase	- 1
15	lameiras	- 1
16	cartelixornadasumisionquimicaevs	- 1
17	conciliacíons	- 1
18	conciliaciones	- 1
19	operatoria	- 1
20	extranjeos	- 1

Cadro B.2Top 20 palabras menos frecuentes no corpus

B.4 Resultados validación RAGsystem

B.4.1 Heatmaps de varias métricas

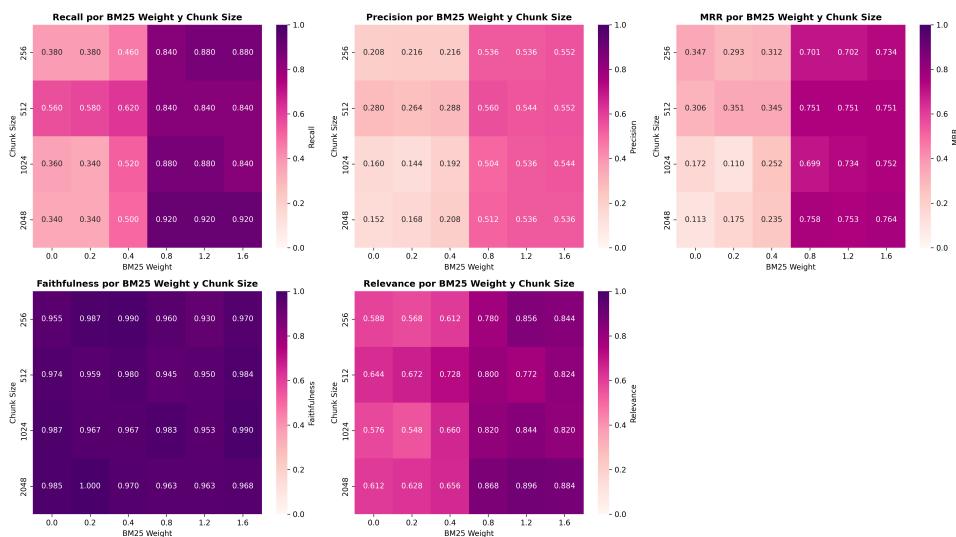


Figura B.2Heatmaps das métricas de validación

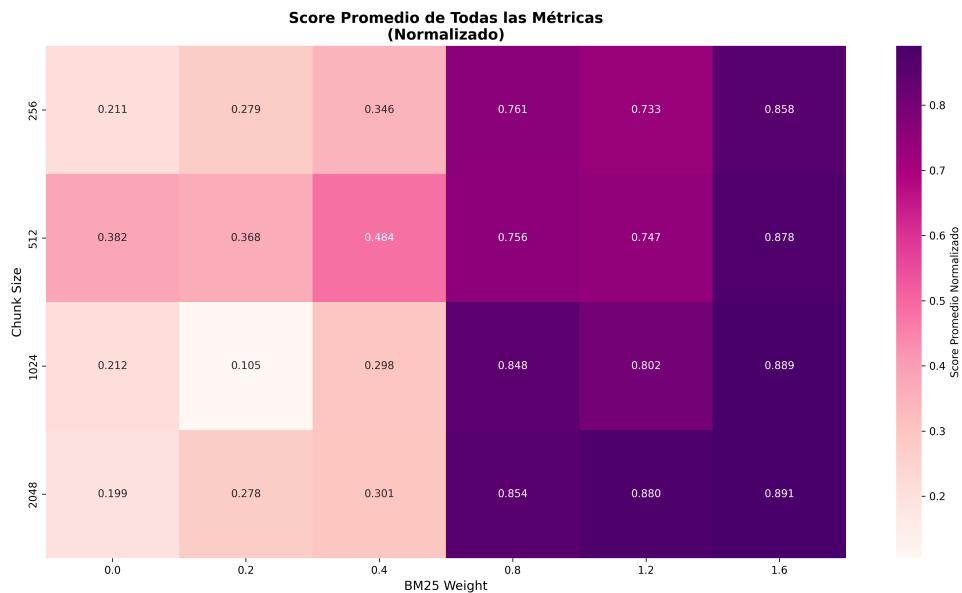


Figura B.3 Heatmap combinado de todas las métricas normalizadas

B.4.2 Surface plots de varias métricas

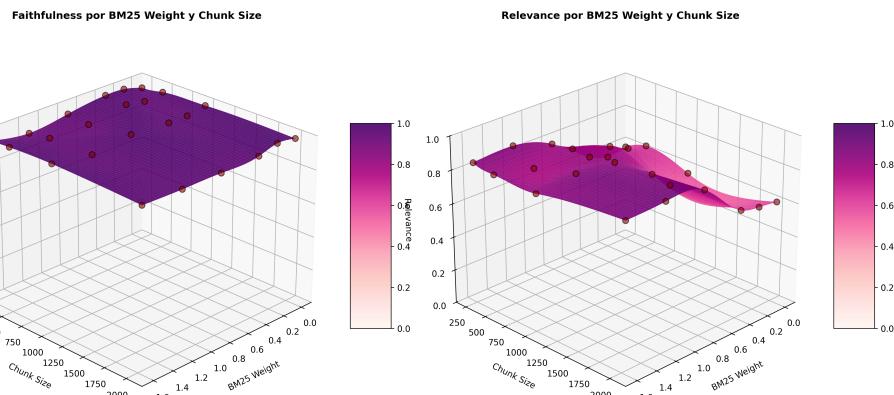


Figura B.4 Surface plots de Faithfulness y Relevance

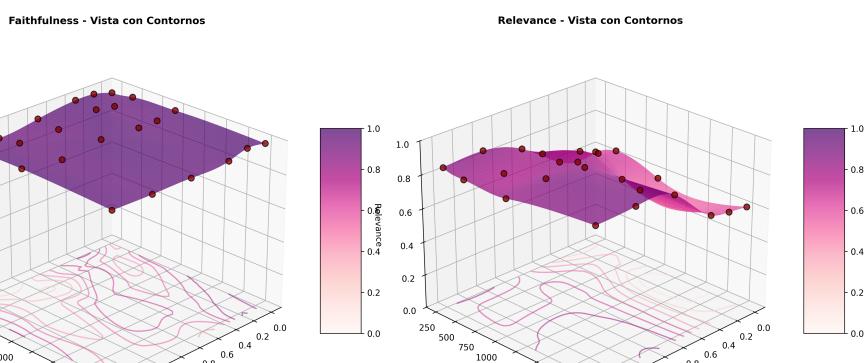


Figura B.5 Surface plots con contornos de Faithfulness y Relevance

B.4.3 Comparación de métricas para diferentes valores de peso de BM25 co mellor *chunk size* (2048)

Cadro B.3 Faithfulness según BM25 Weight (Chunk Size = 2048)

BM25 Weight	Faithfulness
0.0	0.985
0.2	1.000
0.4	0.970
0.8	0.963
1.2	0.963
1.6	0.968

Cadro B.4 Relevance según BM25 Weight (Chunk Size = 2048)

BM25 Weight	Relevance
0.0	0.612
0.2	0.628
0.4	0.656
0.8	0.868
1.2	0.896
1.6	0.884