Caso práctico III (Bloque III — RAG) — El asistente que contesta con los documentos de vuestro sector

Asignatura: Deep Learning para Business Analytics — Comillas (ICADE). **Profesor:** Eduardo C. Garrido-Merchán · `ecgarrido@comillas.edu`. **Curso:** 2026–2027. **Bloque:** III — Sistemas sobre el agente (Tema B, semanas 13–14). **Modalidad: entrega en grupo, defensa junto al pitch final.**

1. Qué se construye

Un asistente que responde preguntas de un empleado o de un cliente **leyendo los documentos de vuestro sector**, con la cita del fragmento del que sale cada cifra. El corpus lo montáis vosotros con documentos públicos y reales: normativa, memorias anuales, pliegos, manuales de producto, tarifas, condiciones generales. Este RAG es el que entra en el prototipo del proyecto final, así que montadlo sobre el sector que vais a defender en el pitch.

La práctica del Tema B era un RAG de juguete: convenio ficticio, TF-IDF y vectores que cabían en pantalla. Aquí el corpus es real, el recuperador es denso y el sistema tiene que sobrevivir a preguntas que nadie redactó pensando en él.

Tres exigencias. Toda respuesta **cita** el fragmento que la sostiene, con su documento y su fecha. El sistema **sabe callarse**: ante una pregunta cuya respuesta no está en el corpus, contesta que no está, y eso se mide. Y toda mejora que afirméis va con **su número antes y después**, sobre el mismo conjunto de evaluación. Nada de esto pide clave de API ni GPU.

2. El corpus y el conjunto de evaluación

Construir el conjunto de evaluación es la mitad del trabajo y la mitad de la nota: un RAG sin preguntas anotadas no se puede medir, y lo que no se mide no se vende.

Pieza	Mínimo	Qué se comprueba
Corpus	20 documentos, 150 fragmentos tras el troceado	`manifiesto.csv` con título, URL, fecha de descarga y licencia de cada documento. Sin datos personales reales.
Preguntas respondibles	20	Redactadas como las escribiría un usuario, no como las escribiría el documento. Cada una con su **fragmento de oro** anotado.
De esas, difíciles	5	Coloquiales, con sinónimo del término del documento, o que necesitan **dos** fragmentos.
Preguntas sin respuesta	5	Verosímiles y del mismo dominio, pero su respuesta no está en el corpus. El sistema debe rechazarlas.

3. Qué se mide

Cuatro números, los cuatro sobre el mismo conjunto y con la semilla escrita. **Acierto@**k para $k \in \{1, 3, 5, 10\}$: con qué frecuencia el fragmento de oro está entre los k recuperados. **Fidelidad**: de las respuestas en las que el fragmento de oro *sí* se recuperó, cuántas lo usan de verdad en lugar de inventar o de coger la cifra del fragmento de al lado. **Tasa de abstención** sobre las cinco preguntas sin respuesta. Y el **coste**, separado en lo que se paga una vez al indexar y lo que se paga en cada consulta.

Cada fallo se clasifica en uno de los cuatro sitios donde un RAG se rompe: **no indexado** (el troceado partió el dato), **no recuperado** (el encoder no lo ve), **recuperado pero enterrado** (estaba en el puesto siete y k era tres) o **recuperado y mal respondido** (el generador lo tuvo delante y contestó otra cosa). Cada uno se arregla en un sitio distinto, y confundirlos es gastar dinero donde no está el problema.

4. El sistema de referencia: `demo_rapida/01_rag.ipynb`

El cuaderno de clase, sobre la normativa inventada del Banco Iberia (6 documentos, 16 fragmentos), tiene los cuatro números medidos y sirve de listón. Con `intfloat/multilingual-e5-small` da acierto@1 del 83 % y acierto@3 del 100 %; cambiando a `paraphrase-multilingual-MiniLM-L12-v2`, de igual tamaño de vector y velocidad, el acierto@1 cae al 17 % y el @3 se queda en el 100 %: **cambiar de encoder es la optimización más barata de un RAG**. Sin contexto, el modelo se inventa una comisión de 0,50 € por sacar dinero en un cajero ajeno; con los fragmentos delante lee los 1,95 € del catálogo. La pregunta por los fondos indexados recupera bien el fragmento bueno y contesta con los 350 € de la tasación de otro documento, que es el fallo de fidelidad en estado puro. Y el umbral de similitud, la defensa que todo el mundo propone primero contra las preguntas sin respuesta, **no separa**: las respondibles puntúan entre 0,834 y 0,893 y una de las que no se puede contestar puntúa 0,880. Tendréis que abstener con otra cosa y explicar con cuál.

5. Qué se entrega

Una carpeta comprimida `grupoNN_caso_3.zip`, por debajo de 20 MB, con exactamente esto. El corpus va dentro si cabe; si no cabe, va el manifiesto y el script que lo descarga.

#	Fichero	Qué es
1	`rag/`	El sistema entero reproducible: troceado, índice, recuperación, generación con cita y la regla de abstención. Semilla fijada y un `README.md` de arranque de cinco líneas.
2	`evaluacion/preguntas.json`	Las 25 preguntas del §2 con su fragmento de oro anotado y la marca de cuáles no tienen respuesta.
3	`evaluacion/resultados.md`	Dos páginas: la tabla de acierto@k, fidelidad, abstención y coste; la clasificación de **cada** fallo en uno de los cuatro sitios; y los **dos arreglos** que probasteis, con el antes y el después de cada uno y la comprobación de que no rompieron nada que ya iba bien.

No hay memoria, ni informe, ni prompts que entregar. Lo que sí hay es un sistema que tiene que arrancar: en la defensa se os puede pedir que lo abráis y le hagáis una pregunta nueva.

6. Rúbrica

Componente	Peso	Qué se valora
Corpus y conjunto de evaluación	25 %	Documentos reales con manifiesto completo; 20 preguntas con oro anotado; las 5 difíciles, difíciles de verdad; las 5 sin respuesta, verosímiles.
Medida	25 %	Los cuatro números, bien calculados y sobre el conjunto de prueba. Barrido en k con la tabla completa.
Diagnóstico	20 %	Cada fallo en su casilla de las cuatro, justificado mirando lo que el sistema sí recuperó.
Arreglo medido	20 %	Dos intervenciones con antes y después y comprobación de regresión.
Abstención y cita	10 %	Toda respuesta citada; regla de abstención que funcione y que no sea un umbral de similitud a secas.

Medir no es arreglar, y afirmar no es medir. Decir que algo «mejora» sin número cuesta -1 por caso. **Suspende el caso** que una cifra de `resultados.md` no se pueda reproducir con el código entregado, que el sistema conteste sin citar, o que un arreglo se reporte sin comprobar si rompió alguna de las preguntas que antes acertaba.

Bonus, hasta $+1$ **punto**, sólo si lo obligatorio está completo: comparar dos encoders sobre vuestro corpus con la tabla de los dos (+0,5); un reordenador y la medida de si mueve algo o no (+0,25); troceado por secciones del documento en vez de por ventana fija, con el acierto@1 de los dos (+0,25).

7. La defensa sin IA

En la ronda de preguntas del pitch, ordenadores cerrados y sin asistente, con una página de notas de autoría propia. Cualquier miembro del grupo responde a: **(i)** enseñad una pregunta que falla y decid en cuál de los cuatro sitios falla y por qué; **(ii)** justificad el k que elegisteis con vuestra propia tabla, y decid qué ganáis y qué pagáis al subirlo; **(iii)** explicad qué hace vuestro sistema cuando la respuesta no está en el corpus, y por qué un umbral de similitud no bastaba. Incapacidad de defenderlo sin IA = cero en el caso (regla C, Día 01).

8. Fechas, política de IA y ayudas

Publicación en la semana 13 y entrega en Moodle en la semana 14, **antes del pitch**; la defensa es la ronda de preguntas del pitch, el mismo día; notas con las del proyecto final. Los grupos son los del proyecto final, y el RAG que entreguéis aquí es el que se espera ver en el prototipo del §2 de la memoria. La IA está permitida y fomentada para construirlo y prohibida en la defensa del §7: la responsabilidad de lo que el agente escriba es vuestra, y las cifras de `resultados.md` las firmáis vosotros. Para dudas, el foro primero y el profesor después. El andamiaje está entero en `demo_rapida/01_rag.ipynb`, que trae el troceado, el encoder, el índice y la abstención resueltos sobre el Banco Iberia, y en `nuevo_material/tema_B_sistemas/practica/`, que trae el barrido en k y la clasificación de los cuatro fallos. Copiadlos y cambiadles el corpus.