

Informe Final: Asistente Conversacional RAG para Políticas Internas

Tema elegido para el asistente conversacional

El tema seleccionado para el desarrollo del asistente conversacional es el "Reglamento de Políticas Internas" de una firma de servicios profesionales. Se trata de un documento extenso que regula normas y procedimientos sobre licencias, beneficios, responsabilidades, horarios, entre otros aspectos laborales internos.

Justificación de la elección

La motivación para elegir este tema radica en la necesidad de mi empresa de desarrollar un sistema tipo chatbot para evacuar consultas de carácter normativo a la interna. Con este asistente que permite realizar consultas en lenguaje natural sobre este reglamento, se puede aumentar significativamente el acceso a la información por parte de los empleados, mejorar el cumplimiento de normas internas y reducir la carga de consultas al área encargada de estos temas.

Fuentes utilizadas para construir la base de conocimiento

La fuente principal fue el PDF original "Políticas Internas" provisto por la empresa. Este archivo fue almacenado en el repositorio bajo la carpeta data/, con el nombre politicas_internas.pdf. No se utilizaron fuentes externas ni bases de datos adicionales.

Preprocesamiento y parseo de datos

El preprocesamiento consistió en la limpieza manual de la documentación (anonimización), consolidación de distintos documentos (originalmente separados por temas de interés) y etiquetado por temáticas. El PDF fue procesado utilizando pdfminer.six para extraer el texto plano. Posteriormente, se aplicó un script de parsing personalizado en parser.py, a partir del cual se identificó artículos, títulos, capítulos y subcapítulos mediante expresiones regulares. Se estructuró el contenido jerárquicamente para preservar el contexto. Luego, los bloques de texto fueron convertidos en objetos Document con contenido y metadatos (número de artículo título, capítulo, subcapítulo y subtítulo) para su posterior indexación. Finalmente, cada documento se transformó en embeddings y los mismos se almacenaron en un índice vectorial (faiss_index) para lograr búsquedas eficientes por similitud. Se utilizó el modelo inffloat/e5-large-v2 para la generación de los embeddings. Este modelo fue seleccionado por su buen rendimiento en tareas de recuperación. A partir de esta indexación de los documentos se logra recuperar los documentos más relevantes dada una consulta.

Modelo LLM utilizado

Se utilizó el modelo google/flan-t5-large, descargado desde Hugging Face y ejecutado localmente mediante la librería transformers. Este modelo fue elegido dadas las restricciones de capacidad y memoria del equipo en el que se ejecutó por un lado, y por su capacidad para seguir instrucciones en lenguaje natural y generar respuestas coherentes y contextualmente relevantes. La inferencia se realiza a través de un pipeline de generación de texto integrado por LangChain utilizando HuggingFacePipeline con una cadena RAG personalizada.

Debido a resultados no tan buenos como los esperados, se desarrolló un sistema equivalente pero basado en “Mistralai/Mistral-7B-Instruct-v0.1” y otro basado en “meta-llama/Llama-3.1-8B-Instruct” ejecutados en Google Colab con GPU. De esta forma se pudo corroborar que la causa de las malas respuestas era la potencia del LLM utilizado, ya que con los otros dos modelos más potentes las respuestas mejoraron sustancialmente.

Diseño del sistema RAG

El sistema RAG implementado combina:

- **Recuperación:** Los documentos se indexan usando FAISS y se consultan con embeddings mediante métricas de similitud (se optó por los 3 documentos más similares ya que la documentación está bien segmentada y separada por capítulos y subcapítulos). De esta forma el sistema se basa en los documentos relevantes y no se mezcla con otros documentos).
- **Generación:** El modelo flan-t5 genera respuestas basadas exclusivamente en el contexto recuperado. El modelo se seteó para que no retorne todo el texto (simplemente debe dar la respuesta indicada en la cantidad de tokens que necesite), además de una temperatura de 0.1 para evitar alucinaciones.
- **Prompting:** Se diseñó un prompt personalizado que obliga al modelo a responder solo con el contexto provisto, indicando Título, Capítulo y Subcapítulo, y concluyendo con la recomendación institucional de contactar a Talento y Cultura si es necesario. En caso de no tener la respuesta en el contexto proporcionado debe contestar “No lo sé” y no inventar respuestas. Posteriormente, con StrOutParser() nos aseguramos de que la salida como respuesta sea un string limpio.

Funcionamiento del backend

El backend fue desarrollado con FastAPI. Al recibir una pregunta, se ejecuta el siguiente flujo:

1. La pregunta se transforma en embedding.
2. Se consulta FAISS para recuperar los documentos más relevantes.
3. El contenido se formatea como contexto.
4. Se envía al modelo LLM con el prompt estructurado.
5. La respuesta generada se devuelve a Streamlit.

El backend está completamente desacoplado del frontend. Ambos servicios corren en un contenedor Docker, expuestos en los puertos 8000 (API) y 8501 (Streamlit).

Guía para la ejecución del sistema

Se desarrollaron dos métodos para ejecutar el asistente: una modalidad local, pensada para entornos de desarrollo, y otra contenerizada mediante Docker, orientada a portabilidad y despliegue sin configuración adicional.

Ejecución local

Requiere tener instalado Python 3.9 o superior, pip y virtualenv. Los pasos para la ejecución local son:

1. Clonar el repositorio del asistente desde GitHub y acceder a la carpeta del proyecto.
2. Ejecutar `setup_api` para crear entorno virtual y descargar dependencias.
3. (Opcional) Descargar localmente los modelos de lenguaje y de embeddings desde Hugging Face para evitar su descarga automática en tiempo de ejecución.
4. Ejecutar el script `build_index.py` para procesar el PDF, generar los embeddings y construir el índice FAISS con los documentos vectorizados.
5. Lanzar el sistema con el script `run_all.sh`, que inicia tanto el backend con FastAPI como el frontend con Streamlit.

Una vez iniciado, la interfaz de usuario se encuentra disponible en <http://localhost:8501> y la documentación del backend (FastAPI) en <http://localhost:8000/docs>.

Ejecución con Docker

Requiere tener Docker Desktop instalado y ejecutándose. Los pasos para ejecutar el asistente en un contenedor Docker son:

1. Ejecutar `setup_api` para crear entorno virtual y descargar dependencias.
2. Ejecutar el script `build_index.py` para construir el índice FAISS (es un paso previo obligatorio a la construcción de la imagen).
3. Construir la imagen Docker utilizando el script `build_api.sh`.
4. Ejecutar el contenedor mediante el script `run_docker.sh`.

Una vez iniciado, el asistente se encuentra disponible en el puerto 8501 (Streamlit) y la API en el puerto 8000 (FastAPI).

Consideraciones sobre el uso de modelos

Si los modelos no están descargados localmente, el sistema intentará descargarlos automáticamente desde Hugging Face al momento de la ejecución, siempre que haya conexión a internet. En caso de querer trabajar offline, se recomienda ejecutar previamente `download_models.py` para obtener los modelos necesarios y almacenarlos en la carpeta `models/`.

Capturas de pantalla:

Por lo comentado previamente, las capturas tanto del frontend como del backend están vinculadas al modelo que obtuvo respuestas malas. Para constatar de que efectivamente quedó funcional se muestran algunas capturas como referencia. Debido a recursos limitados (explicación en video) este sistema quedó funcional en otra PC (de todas formas se muestra en video el sistema funcionando). De este modo, en esta sección se muestra el frontend y backend con respuestas insuficientes por un lado y las respuestas correctas que devuelve el modelo de Llama 3-8B ejecutado en Google Colab. Para que el sistema no devuelva solamente "No lo sé" y se vean respuestas más largas en Streamlit, se decidió agregarle un fallback que devuelve los documentos que trajo la API para generar la respuesta.

- Pantalla inicial del asistente.

Frontend

The screenshot shows a Streamlit application window. At the top, it says "Asistente de Políticas Internas" and "Consultá el reglamento interno de la empresa con ayuda de un modelo de lenguaje." Below this, there are two input fields: "Ingresá tu consulta sobre el reglamento interno:" and "Cuál es la duración del permiso por maternidad?". A "Consultar" button is located below the inputs. To the left of the main area, there is a sidebar titled "Información" which lists the system architecture: Backend: FastAPI + LangChain, Frontend: Streamlit, Embeddings: intfloat/e5-large-v2, Vector DB: FAISS, and LLM: google/flan-t5-large. Another section titled "Ejemplos de consulta" contains questions like "¿Qué hago ante una situación de acoso laboral?" and "¿Cuál es la política sobre diversidad?".

This screenshot shows the same Streamlit application as above. The main area now displays a yellow warning message: "⚠ La respuesta fue tomada directamente del reglamento, ya que el modelo no pudo generar una respuesta." Below this message, there is a detailed view of the "Respuesta" section, which includes the query, the title and chapter of the regulation, the fragment used, and the full text of the regulation fragment. The sidebar and other UI elements remain the same.

Backend

Asistente de Políticas Internas 0.1.0 OAS 3.1

/openapi.json

default

POST /query Query Rag

Parameters

No parameters

Request body required

application/json

```
{ "question": "Cuál es la duración del permiso por maternidad?" }
```

Curl

```
curl -X 'POST' \
'http://localhost:8000/query' \
-H 'accept: application/json' \
-H 'Content-type: application/json' \
-d '{
  "question": "Cuál es la duración del permiso por maternidad?"
}'
```

Request URL

http://localhost:8000/query

Server response

Code Details

200 Response body

```
{
  "answer": "TÍTULO: Política de maternidad y paternidad - CAPÍTULO: Permiso por maternidad:\n\n2. Capítulo: Permiso por maternidad: \n\nTodas las colaboradoras de LA FIRMA embarazadas tendrán un permiso por maternidad de 16 semanas. La colaboradora podrá elegir cuándo gozar de dicho permiso y podrá dividirlo (entre el final del embarazo y luego del nacimiento del o los bebés en la proporción que elija) o usar íntegramente las 16 semanas desde el nacimiento del o los bebés. En caso de nacimientos múltiples o si el peso del bebé al nacer no supera 1.5 kilogramos, el permiso por maternidad se extenderá hasta 18 semanas. Independientemente de la semana de gestación en que se produzca el nacimiento, cuando el o los recién nacidos presenten algún trastorno que implique riesgo o compromiso de vida, con internación o con tratamiento domiciliaria, el permiso por maternidad podrá extenderse hasta que el o los bebés de la colaboradora cumplan 6 meses de edad. El permiso por maternidad también podrá extenderse hasta que el o los bebés de la colaboradora cumplan 6 meses de edad cuando presenten algún afección que, sin implicar riesgo de vida, involucre discapacidades sensoriales, físicas o intelectuales, que a juicio del médico tratante se beneficie de los cuidados de la madre. Cuando el parto se produzca antes de la fecha probable, pero a partir de las 34 semanas de gestación inclusive, la colaboradora iniciará el permiso por maternidad de inmediato. El permiso por maternidad se verá prolongado hasta completar las 16 semanas o las 8 semanas posteriores a la fecha prevista como probable de parto, una vez este último periodo vence con posterioridad al parto. Cuando el parto se produzca antes de la fecha probable y hasta las 31 semanas de gestación (inclusive), la colaboradora iniciará el permiso por maternidad de parte inicialmente, si hasta este último periodo vence con posterioridad al primero.",

  "response_time": "22.94s",
  "context_docs": [
    {
      "articulo": "",
      "título": "Política de maternidad y paternidad",
      "capítulo": "Permiso por maternidad",
      "subcapítulo": "",
      "subtítulo": "",
      "contenido": "TÍTULO: Política de maternidad y paternidad - CAPÍTULO: Permiso por maternidad:\n\n2. Capítulo: Permiso por maternidad: \n\nTodas las colaboradoras de LA FIRMA embarazadas tendrán un permiso por maternidad de 16 semanas. La colaboradora podrá elegir cuándo gozar de dicho permiso y podrá dividirlo (entre el final del embarazo y luego del nacimiento del o los bebés en la proporción que elija) o usar íntegramente las 16 semanas desde el nacimiento del o los bebés. En caso de nacimientos múltiples o si el peso del bebé al nacer no supera 1.5 kilogramos, el permiso por maternidad se extenderá hasta 18 semanas. Independientemente de la semana de gestación en que se produzca el nacimiento, cuando el o los recién nacidos presenten algún trastorno que implique riesgo o compromiso de vida, con internación o con tratamiento domiciliaria, el permiso por maternidad podrá extenderse hasta que el o los bebés de la colaboradora cumplan 6 meses de edad. El permiso por maternidad también podrá extenderse hasta que el o los bebés de la colaboradora cumplan 6 meses de edad cuando presenten alguna afección que, sin implicar riesgo de vida, involucre discapacidades sensoriales, físicas o intelectuales, que a juicio del médico tratante se beneficie de los cuidados de la madre. Cuando el parto se produzca antes de la fecha probable, pero a partir de las 34 semanas de gestación inclusive, la colaboradora iniciará el permiso por maternidad de inmediato. El permiso por maternidad se verá prolongado hasta completar las 16 semanas o las 8 semanas posteriores a la fecha prevista como probable de parto, una vez este último periodo vence con posterioridad al parto. Cuando el parto se produzca antes de la fecha probable y hasta las 31 semanas de gestación (inclusive), la colaboradora iniciará el permiso por maternidad de parte inicialmente, si hasta este último periodo vence con posterioridad al primero."
    }
  ]
}
```

Response headers

- Ejemplo de consulta simple (con modelo Llama 3-8B). En video se muestra en vivo.

Bajo nivel de dificultad

Pregunta: **¿Cuál es la duración del permiso por maternidad?**

 Respuesta generada:

El permiso por maternidad dura 16 semanas, aunque puede extenderse hasta 18 semanas en casos específicos, como nacimientos múltiples o bebés con bajo peso al nacer. También puede extenderse hasta que el o los bebés cumplan 6 meses de edad en caso de trastornos o afecciones que requieran cuidados especiales.

Consultá con Talento y Cultura para más información.

Consulta simple con respuesta incorrecta:

Pregunta: ¿Qué tipo de residuos deben ir en el cesto marrón?

Lo siento, pero no hay información disponible en el contexto proporcionado sobre qué tipo de residuos deben ir en el cesto marrón. El capítulo mencionado solo lista algunos ejemplos de objetos que pueden ir en el cesto marrón, pero no especifica qué tipo de residuos son los adecuados.

Sin embargo, puedo sugerirte buscar en el capítulo mencionado (Capítulo: Cesto marrón COMPOSTABLES) para obtener más detalles sobre qué tipos de residuos se consideran adecuados para el cesto marrón.

Consultá con Talento y Cultura para más información.

Mayor nivel de dificultad:

Pregunta: ¿Qué comportamiento en redes sociales está prohibido para los colaboradores de la firma?

Respuesta generada:

Según el Artículo 9, Capítulo: Lineamientos sobre el comportamiento en redes sociales, subcapítulo 3, NO HACER | DONT'S:

- * No pubiques contenido confidencial de LA FIRMA, sus clientes, sus proveedores o cualquier otro tercero.
- * No proporciones asesoramiento ni envíes documentos relevantes a través de las redes mencionadas en la presente política.
- * No pubiques comentarios personales sobre cualquier tema que ponga en conflicto o comprometan la posición u opiniones de LA FIRMA.
- * No cites o hagas referencia a clientes, colaboradores o proveedores sin su aprobación.
- * No utilices insultos ni ninguna terminología que pueda ser considerada ofensiva para un grupo de personas o para una persona determinada.
- * No ingreses en discusiones a través de las redes sobre temas que puedan incidir negativamente en la imagen de LA FIRMA.
- * No subas fotografías de los clientes (tomadas por ellos o donde ellos salen fotografiados) sin su autorización expresa, a ninguna red social.
- * No subas fotografías de las instalaciones de LA FIRMA donde pueda verse comprometida información sensible de la organización, clientes o colaboradores.

Consultá con Talento y Cultura para más información.

Pregunta: ¿Cómo debo proceder si sufri acoso laboral?

- Respuesta generada:

No hay información específica en el contexto proporcionado sobre cómo proceder si sufrieste acoso laboral. Sin embargo, se menciona que LA FIRMA tiene tolerancia cero con el acoso y prohíbe expresamente cualquier forma de hostigamiento.

En el Artículo 2, Capítulo ALCANCE, Subcapítulo 3.7, se establece que LA FIRMA no tolerará ninguna forma de acoso y que se compromete a mantener un ambiente de trabajo productivo, libre de acosos y de actividades perturbadoras.

En el Artículo 17, Capítulo Introducción, se menciona que LA FIRMA cuenta con medidas tendientes a prevenir, desalentar y sancionar toda conducta de acoso de cualquier clase, y que la víctima de conductas de acoso debe denunciarlas ante la mínima señal.

Por lo tanto, si sufrieste acoso laboral, te recomiendo:

- * Denunciarlo ante tu supervisor o responsable inmediato.
- * Informar a la Dirección de Recursos Humanos o a quien corresponda en la empresa.
- * Consultar la Política contra el Acoso Laboral de LA FIRMA, que se proporciona a todos los Miembros.

Recuerda que la confidencialidad y la reserva son fundamentales en casos de acoso laboral.

Consultá con Talento y Cultura para más información.

- Ejemplo de consulta compleja que involucra deducción.

Pregunta: ¿Un miembro puede invertir en acciones de un cliente si tiene información confidencial que aún no es pública?

- Respuesta generada:

No, un miembro no puede invertir en acciones de un cliente si tiene información confidencial que aún no es pública, debido a la norma establecida en el Subcapítulo 6.1: Confidencialidad y secreto profesional, que indica que los Miembros no deben divulgar ni utilizar información confidencial sobre los Clientes para fines personales.

Consultá con Talento y Cultura para más información.

Pregunta: ¿Un miembro puede aceptar un regalo por parte de un cliente si es un espectáculo para fomentar vínculos comerciales?

- ✓ Respuesta generada:

Según el Artículo 2, Capítulo ALCANCE, Subcapítulo 3.1: Beneficios personales, los Miembros no pueden aceptar Beneficios Personales por parte de los clientes, pero hay algunas excepciones. Entre ellas se encuentra:

* Las cortesías comerciales de estilo que facilitan la discusión de los negocios de LA FIRMA o fomentan la amabilidad y las buenas relaciones comerciales y/o profesionales (almuerzos, espectáculos, etc.).

Por lo tanto, si un miembro puede aceptar un regalo por parte de un cliente si es un espectáculo para fomentar vínculos comerciales, la respuesta es afirmativa.

Consultá con Talento y Cultura para más información.

- Respuesta cuando la información no está presente:

Pregunta: ¿Cuáles son las políticas para el uso de vehículos corporativos?

- ✓ Respuesta generada:

Lo siento, pero no hay información disponible en el contexto proporcionado sobre las políticas para el uso de vehículos corporativos.

Consultá con Talento y Cultura para más información.

Pregunta: ¿Quién es el presidente de Uruguay?

- ✓ Respuesta generada:

Lo siento, pero no tengo información disponible sobre quién es el presidente de Uruguay en el contexto proporcionado. Consultá con Talento y Cultura para más información.

Conclusiones del trabajo

El desarrollo de este asistente conversacional permitió poner en práctica una arquitectura RAG completa, integrando procesamiento de lenguaje natural, recuperación semántica y una interfaz intuitiva para el usuario final. A través de un único documento estructurado e indexado, se logró implementar un sistema que responde consultas en lenguaje natural con precisión variable (según modelo LLM utilizado) y referencias explícitas al contenido normativo recuperado.

Aprendizajes

- Profundización en técnicas de procesamiento de texto y segmentación desde numerosos documentos a uno homogéneo y estructurado.

- Implementación del sistema funcional con utilización de LangChain, FAISS y Transformers en un flujo de preguntas y respuestas con Frontend intuitivo para el usuario final y backend con buenos resultados de recuperación.

Dificultades encontradas

- Bloqueo de puertos, capacidad y memoria muy limitadas para la elaboración de este sistema, lo que obligó a recurrir a otro equipo para elaborarlo.
- Restricciones de memoria que limitaron el uso de modelos más grandes.

Oportunidades de mejora y próximos pasos

- Contenerizar y poner en funcionamiento el sistema pero utilizando los modelos más potentes (los utilizados en Colab) con un equipo con mayores capacidades que lo soporten.
- Incluir autenticación por usuario para limitar el acceso a información sensible.
- Registrar logs de consultas y respuestas más detallados para evaluar desempeño y detectar mejoras.