Análisis de perfiles de candidatos para puestos de trabajo en el área de TI mediante Recuperación Aumentada por Generación (RAG)

Karen Riveros

Amanda Sosa

Jorge Benítez

Maestría en Ciencia de Datos | Universidad Comunera | Paraguay

Resúmen

La selección de personal en el área de Tecnología de la Información (TI) representa un desafío constante para empresas y consultoras, debido a la diversidad de perfiles y al gran volumen de currículums que deben analizarse. Frente a esta necesidad, el presente proyecto propone la implementación de un sistema de Recuperación Aumentada por Generación (RAG), capaz de realizar búsquedas semánticas e inteligentes sobre documentos de candidatos, facilitando así la identificación de talentos según criterios específicos.

Para ello, se recopiló un corpus de 77 currículums vitae en formato PDF, provenientes de candidatos del área tecnológica en su mayoría, y otras áreas, incluyendo también perfiles exportados de LinkedIn. Estos documentos, con estructura y calidad variadas, fueron procesados mediante el framework LangChain en un entorno colaborativo de Google Colab. El sistema RAG se construyó utilizando el modelo generativo **GPT-4o-mini** y embeddings **text-embedding-3-large** de OpenAI. A través de FAISS, se creó una base vectorial que permite recuperar los fragmentos más similares a cada consulta, sobre los cuales el LLM genera respuestas detalladas.

Se formularon 17 preguntas que simulan escenarios reales de reclutamiento, evaluando así la capacidad del sistema para extraer y sintetizar información como habilidades técnicas, formación académica, experiencia laboral, liderazgo, y certificaciones. El análisis mostró un **59%** (10 respuestas) de respuestas **completamente correctas**, **29%** (5 respuestas) de las respuestas **parcialmente correctas**, y **11%** (2 respuestas) de las respuestas **incorrectas,** destacando la influencia de la calidad del documento y la precisión del prompt sobre los resultados.

Planteamiento del problema

En el ámbito de Recursos Humanos, la búsqueda y selección de personal en el área de Tecnología de la Información (TI) representa un proceso continuo y desafiante. Las empresas y consultoras de RRHH deben gestionar grandes volúmenes de información sobre candidatos, como datos personales, antecedentes laborales, habilidades, experiencia y especializaciones, en perfiles tan diversos como desarrolladores, gestores de proyectos, expertos en ciberseguridad o analistas de datos.

Contar con bases de datos bien estructuradas es esencial, pero no siempre posible ni suficiente. La identificación rápida y precisa del talento adecuado requiere herramientas modernas que permitan explorar eficientemente esta información y facilitar la coincidencia entre los requerimientos del puesto y las competencias de los candidatos.

En este contexto, la inteligencia artificial, y en particular el enfoque de Retrieval-Augmented Generation (RAG), ofrece una solución innovadora. Mediante RAG, es posible afinar la búsqueda de perfiles al combinar recuperación de información relevante con generación de respuestas basadas en lenguaje natural. Esto permite interactuar con la documentación (currículums digitalizados) de manera más flexible e inteligente, filtrando y proponiendo candidatos que se ajusten a criterios específicos como nivel de experiencia, conocimientos técnicos o trayectoria profesional.

Implementar un sistema basado en RAG no solo optimiza el proceso de preselección, sino que también proporciona una ventaja competitiva para las empresas y consultoras, al facilitar la identificación de los perfiles más adecuados antes de iniciar el proceso formal de entrevistas y selección.

Descripción del corpus

El objetivo de este corpus es alimentar un sistema de recuperación aumentada (RAG) que permita realizar búsquedas semánticas y generar respuestas relevantes en procesos de prospección de perfiles y reclutamiento. Así, se podrán identificar de manera más precisa candidatos adecuados según criterios específicos definidos por reclutadores o empresas.

El corpus está conformado por documentos en formato PDF que contienen principalmente información de profesionales del ámbito tecnológico. Estos documentos provienen principalmente de dos fuentes:

- Currículums vitae (CVs): Archivos enviados directamente por candidatos a que se encuentran a la búsqueda de nuevas vacantes laborales.
- Perfiles de LinkedIn exportados: Información profesional estructurada o descargada desde perfiles públicos de LinkedIn, también en formato PDF.

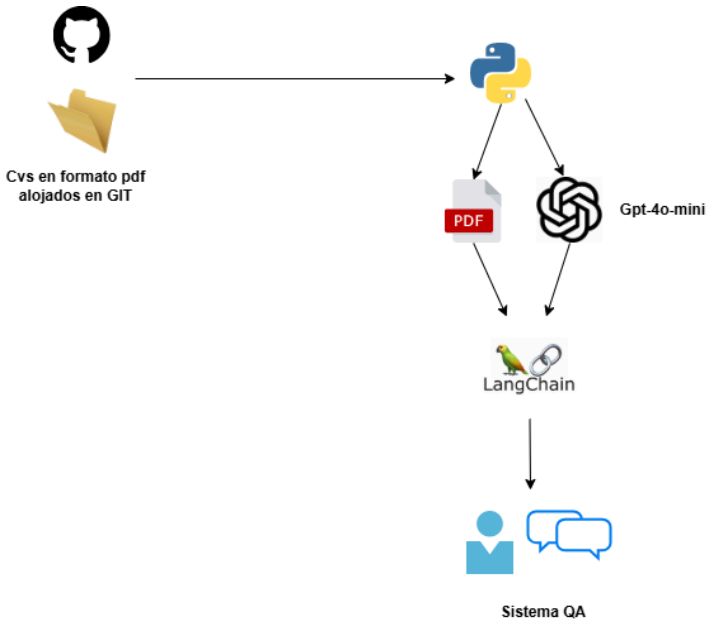
Cada documento incluye datos semi-estructurados o no estructurados relacionados con:

- Datos personales (nombre, ubicación, contacto)
- Experiencia laboral (puestos ocupados, duración, empresas)
- Formación académica (títulos, instituciones, fechas)
- Habilidades técnicas (lenguajes de programación, herramientas, frameworks)
- Certificaciones y logros
- Idiomas
- Resúmenes personales o extractos del perfil profesional

En cuanto al formato presenta las siguientes características:

- Tipo de archivo: PDF (Portable Document Format)
- Longitud promedio: 1 a 3 páginas por documento
- Idioma predominante: Español e inglés
- Estructura del texto: variable en contenido; algunos documentos están bien organizados, otros presentan formatos complejos (tablas, columnas, gráficos, logos)

Metodología (modelo(s) utilizado(s), procesamiento, herramientas, hiperparámetros)



En un repositorio de git almacenamos 77 currículums de profesionales del área de TI y unos pocos de profesionales de otras áreas (marketing, administración , etc). Los documentos están en formato .pdf.

En una notebook de google colab:

1- Se establece conexión con el repositorio Git donde están alojados los Cvs y se procede a la descarga.

2- Se importa el framework langchain que permite aplicar la técnica de recuperación aumentada (RAG) combinando un modelo generativo y documentos relacionados a una temática específica.

2.1 La partición de documentos en segmentos más pequeños, se realiza con el método RecursiveCharacterTextSplitter, con los parámetros:

- chunk_size: Se divide cada documento en fragmentos de 2.000 caracteres cada uno.
- chunk_overlap: Se genera una superposición de 200 caracteres entre un segmento y otro para no perder el contexto.
- separators: ["\n\n", "\n", ". ", " ", ""], La lista de caracteres a ser usados para dividir el texto.
- strip_whitespace: True, eliminar espacios innecesarios.

2.2 El modelo generativo usado es GPT-4o-mini y el modelo embeddings usado es: text-embedding-3-large de OpenIA.

2.3 La extracción del contenido textual de cada documento y la conversión de cada texto en un vector numérico, se hace con el método: FAISS.from_documents

2.4 Los parámetros para la búsqueda de los resultados de una consulta en la base de datos de vectores se especifican con el método: as_retriever()

Parámetros:

- search_type: Similarity (Por defecto). Calcula la similitud entre la consulta vectorizada y los vectores de la base de datos. Usa el método del coseno para el cálculo.
- k: 15, retorna los K primeros documentos de mayor similitud con la consulta efectuada.

Por último, con el método create_retrieval_chain, se asocia el objeto devuelto por as_retriever() con el LLM definido y un prompt a usar en cada consulta.

Evaluación de resultados

El proyecto se basó en un enfoque RAG (Recuperación Aumentada por Generación) sin etiquetas de verdad para cada consulta. Para llevar a cabo la evaluación se realizó un análisis cualitativo y observación de la precisión y relevancia de las respuestas generadas por el modelo frente a las expectativas.

Se formularon 17 preguntas que simulan necesidades reales de reclutadores, como:

- Listado de candidatos según habilidades técnicas, blandas o formación académica.
- Extracción de datos de contacto o experiencia profesional.
- Generación de resúmenes de perfiles.
- Obtención de antecedentes laborales de candidatos.

Los resultados obtenidos se pueden clasificar en 3 principales categorías:

- **Respuestas correctas:** en **10** de las 17 consultas, el sistema devolvió información precisa, detallada y directamente relacionada con los documentos fuente. Ejemplo:

```
[ ]  #5 Escribe la pregunta
     response = retrieval_chain.invoke({"input": "Listar el nombre de personas con formación académica de maestría en ciencia de da
     print(response["answer"])
```

```
Basado en el contexto proporcionado, las personas con formación académica de maestría en ciencia de datos son:

1. **Jorge Dagoberto Benítez Cabana**
   - Maestría en Ciencia de Datos (mar 2024 - presente) en Universidad Comunera, Asunción.

2. **Karen Riveros**
   - Maestría en Ciencia de Datos (2024 - 2025) en Universidad Comunera.

3. **Fernando Manuel Enciso**
   - Diplomado en Ciencias de Datos (2023) en Universidad Comunera. (Nota: No se menciona explícitamente una maestría, pero se

Estas son las personas mencionadas en el contexto que están relacionadas con la formación en ciencia de datos a nivel de maestr
```

```
[6]  agregar_fila(5, "Correcta")
```

RESPUESTA CORRECTA: Fernando Encisco tiene un diplomado en ciencia de datos; el LLM aclara que no tiene una maestría; pero si algunos conocimientos adquiridos en el diplomado

- **Respuestas parcialmente correctas:** en **5** de las 17 consultas, el sistema incluyó información extra o excluyó información relevante según lo preguntado. Esto revela una limitación en la capacidad del modelo. Ejemplo:

```
#10 Make a question
response = retrieval_chain.invoke({"input": " ¿Cuáles son las habilidades blandas del candidato Willian Paez?"})
print(response["answer"])
```

```
Las habilidades blandas del candidato Willian Paez, según el contexto proporcionado, son las siguientes:

1. **Proactividad**: Willian se describe como una persona proactiva, lo que indica que toma la iniciativa y busca soluciones antes de que se presenten problemas.

2. **Capacidad para trabajar en equipo**: Menciona que tiene predisposición y capacidad para trabajar en equipo, lo que sugiere que puede colaborar eficazmente con otros

3. **Adaptabilidad**: Se destaca su capacidad para adaptarse a distintos ambientes laborales, lo que es crucial en un campo como la tecnología, donde las condiciones y h

4. **Autodidacta**: Willian se considera autodidacta, lo que implica que tiene la habilidad de aprender de manera independiente y buscar nuevos conocimientos por su cuen

5. **Resolución de problemas**: Muestra habilidades en la resolución de problemas, lo que es esencial para enfrentar desafíos técnicos y encontrar soluciones efectivas e

6. **Orientación al detalle**: Se menciona su orientación al detalle, lo que sugiere que presta atención a los aspectos finos de su trabajo, lo cual es importante en el

7. **Habilidades organizativas y de gestión del tiempo**: Estas habilidades son fundamentales para manejar múltiples tareas y cumplir con plazos en un entorno laboral.

8. **Flexibilidad**: La flexibilidad es otra habilidad blanda que se destaca, lo que implica que puede adaptarse a cambios y nuevas situaciones sin dificultad.

9. **Habilidades de comunicación**: Aunque no se menciona explícitamente, la capacidad para trabajar en equipo y resolver conflictos sugiere que Willian tiene buenas hab

Estas habilidades blandas complementan su formación técnica y experiencia laboral, haciéndolo un candidato valioso en el campo de la tecnología y el desarrollo de softwa
```

PROPORCIONÓ INFORMACION ADICIONAL QUE NO ESTABA EN EL CV

- **Respuestas incorrectas:** en **2** de las 17 consultas, el sistema respondió incorrectamente a la pregunta realizada. Por ejemplo:

```
[ ]  # 16  Escribe la pregunta
     response = retrieval_chain.invoke(
         {"input": "Lista los nombres de las personas con experiencia laboral en la empresa GRUPO HORUS S.A."})
     print(response["answer"])
```

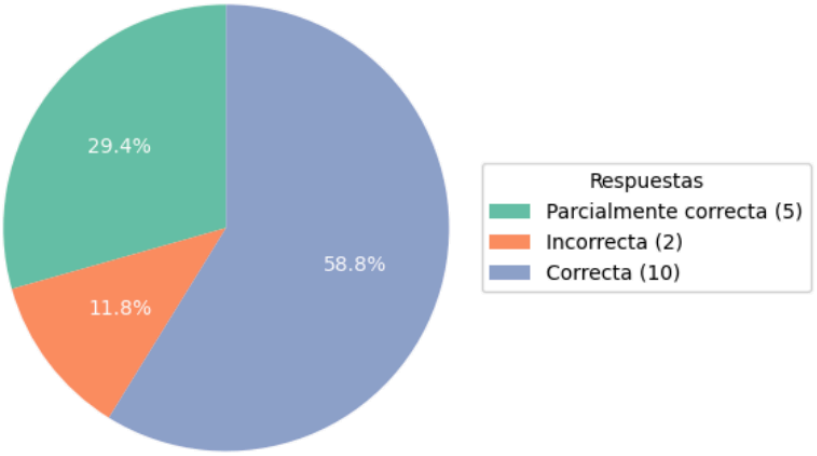
```
En el contexto proporcionado, no se menciona ninguna experiencia laboral específica en la empresa GRUPO HORUS S.A. Por lo tanto
```

```
[17]  agregar_fila(16, "Incorrecta")
```

RESPUESTA INCORRECTA: Edgar Missael Cabral Baez menciona a la empresa GRUPO HORUS S.A. como experiencia laboral

El detalle de los resultados de las consultas realizadas se puede ver en el archivo de la notebook ejecutada.

En el siguiente gráfico se pueden observar los **resultados obtenidos**.



El modelo fue capaz de identificar nombres, cargos, fechas, lenguajes de programación, lugares de estudio y habilidades blandas, lo que demuestra una buena capacidad de extracción semántica incluso en documentos con formatos desorganizados.

Es importante mencionar que para obtener buenos resultados, fue clave la selección de parámetros y la elaboración de buenos prompts que proporcionen el contexto adecuado para que el modelo pueda analizar los cvs y obtener la información relevante. Para optimizar aun más los resultados se sugiere realizar pruebas más exhaustivas con buena estructura de prompts, basada en un análisis de requerimientos con profesionales del área de recursos humanos.

Conclusiones y recomendaciones

La implementación del sistema basado en Recuperación Aumentada por Generación (RAG) ha demostrado ser una herramienta efectiva para la búsqueda semántica y la generación de respuestas en el contexto del análisis de perfiles laborales. La combinación del modelo generativo GPT-4o-mini con los embeddings text-embedding-3-large de OpenAI permitió procesar información no estructurada contenida en currículums vitae, generando respuestas coherentes, relevantes y detalladas ante consultas específicas.

Los resultados obtenidos evidencian que el sistema es capaz de identificar con precisión aspectos clave como experiencia laboral, formación académica, habilidades técnicas y datos de contacto. No obstante, se observó que la calidad y precisión de las respuestas están directamente influenciadas por dos factores principales:

- La estructura y legibilidad del documento fuente (CV).
- La claridad y especificidad de las preguntas formuladas al sistema.

El trabajo en colaboración con expertos del área de Recursos Humanos es relevante para optimizar los resultados. Con este trabajo en conjunto se podría diseñar prompts especializados, como plantillas predefinidas de consultas frecuentes (e.g., búsqueda por lenguaje de programación, experiencia mínima, certificaciones), que optimicen la interacción con el sistema y minimicen la ambigüedad.

Se propone experimentar el modelo en una etapa betha, siendo una herramienta fundamental para reclutadores, analizando las enormes cantidades de cvs recibidas y como mecanismo de sugerencias. Es importante llevar a producción el modelo para evaluar con más precisión el desempeño y pulir los resultados. Una vez que el modelo demuestre funcionar robustamente, esto mismo se puede implementar también a otros tipos de reclutamiento que no pertenezcan necesariamente al área de TI.