Trabajo Final de BootCamp

Estudiante: Sergio Alejandro Alvarado Parada

1. Introducción

El proyecto se ha pensado inicialmente para solucionar la problemática que enfrentan las personas de recursos humanos a la hora de encontrar un candidato que cumpla con los requisitos solicitados en los pocos días que les suelen dar para encontrar dichos candidatos, usualmente tres días. Partiendo de este principio, con este proyecto se busca reducir el tiempo que emplean en los departamentos de recursos humanos para mirar ciertos aspecto fundamentales como:

- Años de experiencia en puestos anteriores.
- Habilidades solicitadas.
- Referencias profesionales.
- Información de contacto.

Para dar una pequeña introducción al contexto del problema que se quiere solucionar, he podido conversar con diferentes profesionales del área de reclutamiento y mencionan el trabajo repetitivo que deben realizar para inspeccionar cientos Curriculum Vitae (u hojas de vida como se conocen en Colombia). Si bien en un principio el tiempo de inspección de una hoja de vida no tardará más de un minuto, este tiempo puede incrementarse rápidamente al tener que inspeccionar cientos de hojas de vida.

Siendo este un trabajo mecánico, se considera que se puede aprovechar las ventajas y capacidades de los LLM (Large Language Models) para buscar esta información en los documentos recibidos. A continuación se presenta un esquema con la solución general del proyecto.

2. Acceso a Datos

Como se ha mencionado en la introducción, los datos iniciales y fundamentales para esta solución son hojas de vida de diferentes candidatos. Debido a que estos documentos suelen tener información que puede considerarse como “sensible” se utilizará una hoja de vida real (la del estudiante) y otras con información inventada o generada con modelos LLM como GPT 3.5 o GPT 4o (dado su reciente lanzamiento). Esta generación de datos implica adaptarlos a un formato comúnmente utilizado en hojas de vida para hacer el experimento lo más real posible, es decir, un formato .pdf o .docx.

3. Diseño de flujo de trabajo de cada una de las etapas

En la Figura 1 se presenta un breve esquema del flujo de trabajo que debe realizar la solución implementada en este proyecto:

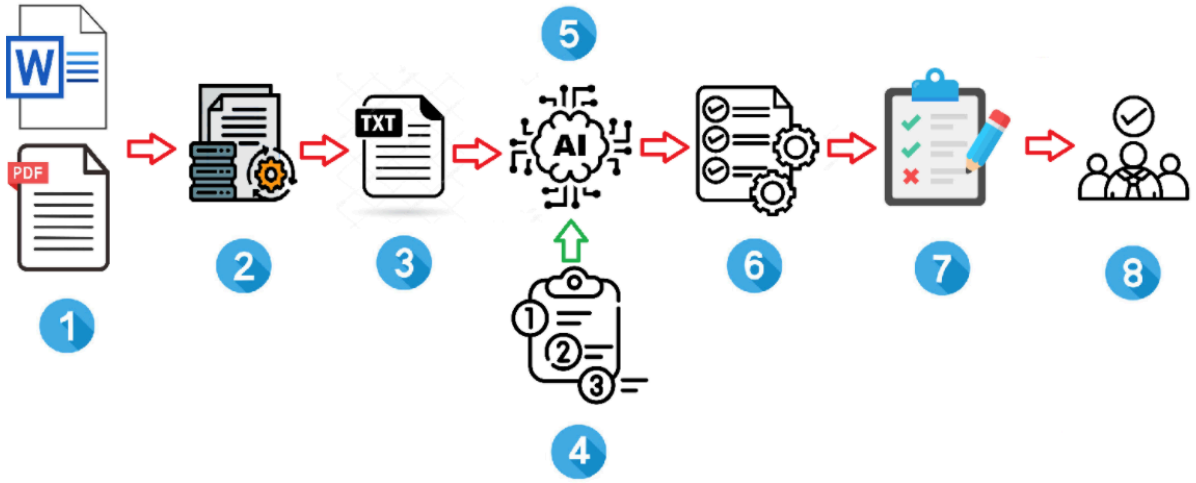


Figura 1. Flujo de trabajo
Fuente: Elaboración propia.

Cada una de las etapas del flujo de trabajo se especifican a continuación:

1-Documentos base: La solución recibirá los documentos (hojas de vida) en formato pdf o docx.

2-Preprocesamiento: Una vez se reciben los documentos en el formato indicado, se procede a extraer la información de ellos para almacenar el texto en un string de Python.

3-Salida del preprocesamiento: El texto contenido en la hoja de vida pasará a estar almacenada temporalmente en un archivo de Python.

4-Procesamiento 1: Una vez se obtiene el string de Python este pasa como input al modelo seleccionado para realizar la extracción de la información deseada.

5-Procesamiento 2: En este paso también se recibirán los requerimientos para la vacante por parte de la persona encargada del proceso de selección.

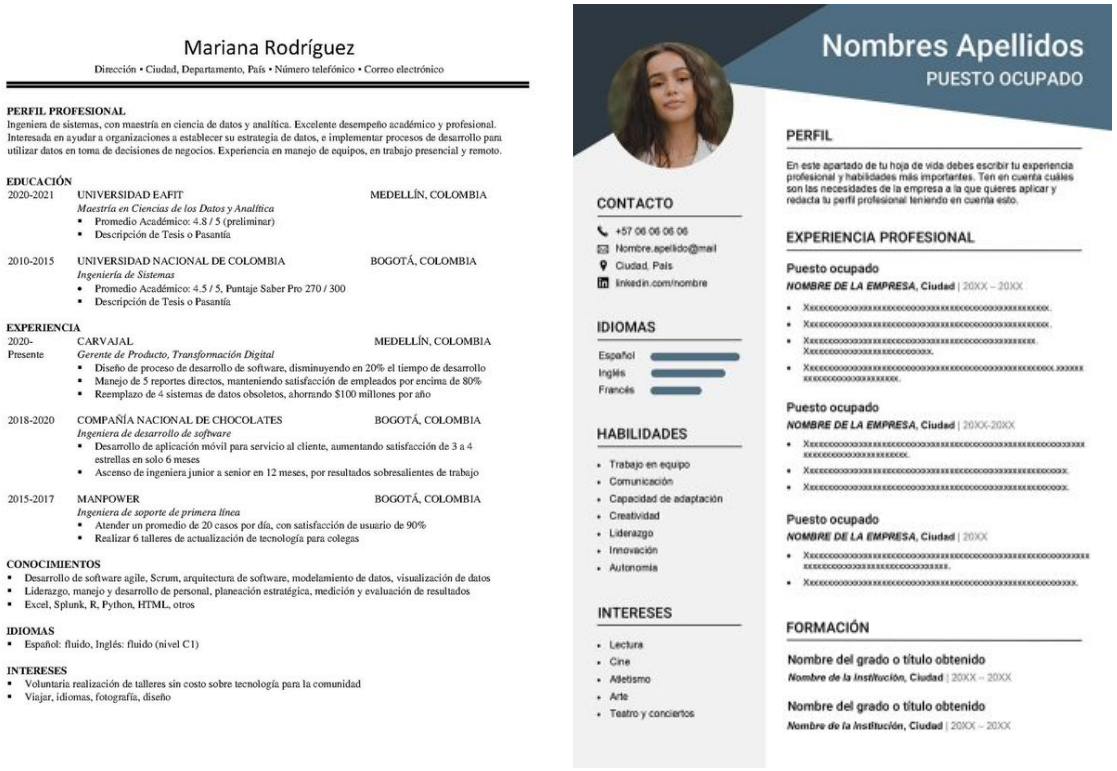
6-Posprocesamiento: La salida del modelo debería ser un formato JSON el cual contenga la información deseada para cada dato buscando en caso de encontrarse.

7-Verificación: En la verificación se identificará si el candidato cumple con los requisitos mínimos para la vacante ofertada.

8-Selección: Al ser planteada como una solución que recibe muchas hojas de vida, la solución presentará únicamente aquellos candidatos que cumplen con los requerimientos.

4. Análisis de datos y conclusiones preliminares

En esta sección no hay mucho que comentar, sin embargo, un detalle importante a tener en cuenta en la gran cantidad de formatos de hojas de vida presentes. Si bien existe una preferencia por las presentaciones sencillas y a veces sin foto, muchas personas suelen decorar mucho sus hojas de vida. En la Figura 2 se presenta a la izquierda se presenta una hoja de vida sencilla y a la derecha una hoja de vida llamativa estéticamente pero con decoraciones que pueden dificultar la extracción de la información.



Mariana Rodríguez

Dirección • Ciudad, Departamento, País • Número telefónico • Correo electrónico

PERFIL PROFESIONAL

Ingeniera de sistemas, con maestría en ciencia de datos y analítica. Excelente desempeño académico y profesional. Interesada en ayudar a organizaciones a establecer su estrategia de datos, e implementar procesos de desarrollo para utilizar datos en toma de decisiones de negocios. Experiencia en manejo de equipos, en trabajo presencial y remoto.

EDUCACIÓN

2020-2021 UNIVERSIDAD EAFIT — MEDELLÍN, COLOMBIA
Maestría en Ciencias de los Datos y Analítica
- Promedio Académico: 4.8 / 5 (preliminar)
- Descripción de Tesis o Pasantía

2010-2015 UNIVERSIDAD NACIONAL DE COLOMBIA — BOGOTÁ, COLOMBIA
Ingeniería de Sistemas
- Promedio Académico: 4.5 / 5, Puntaje Saber Pro 270 / 300
- Descripción de Tesis o Pasantía

EXPERIENCIA

2020-Presente CARVAJAL — MEDELLÍN, COLOMBIA
Gerente de Producto, Transformación Digital
- Diseño de proceso de desarrollo de software, disminuyendo en 20% el tiempo de desarrollo
- Manejo de 5 reportes directos, manteniendo satisfacción de empleados por encima de 80%
- Reemplazo de 4 sistemas de datos obsoletos, ahorrando $100 millones por año

2018-2020 COMPAÑÍA NACIONAL DE CHOCOLATES — BOGOTÁ, COLOMBIA
Ingeniera de desarrollo de software
- Desarrollo de aplicación móvil para servicio al cliente, aumentando satisfacción de 3 a 4 estrellas en solo 6 meses
- Ascenso de ingeniera junior a senior en 12 meses, por resultados sobresalientes de trabajo

2015-2017 MANPOWER — BOGOTÁ, COLOMBIA
Ingeniera de soporte de primera línea
- Atender un promedio de 20 casos por día, con satisfacción de usuario de 90%
- Realizar 6 talleres de actualización de tecnología para colegas

CONOCIMIENTOS
- Desarrollo de software agile, Scrum, arquitectura de software, modelamiento de datos, visualización de datos
- Liderazgo, manejo y desarrollo de personal, planeación estratégica, medición y evaluación de resultados
- Excel, Splunk, R, Python, HTML, otros

IDIOMAS
- Español: fluido, Inglés: fluido (nivel C1)

INTERESES
- Voluntaria realización de talleres sin costo sobre tecnología para la comunidad
- Viajar, idiomas, fotografía, diseño

Nombres Apellidos

PUESTO OCUPADO

CONTACTO
+57 00 00 00 00
Nombre.apellido@mail
Ciudad, País
linkedin.com/nombre

IDIOMAS
Español
Inglés
Francés

HABILIDADES
- Trabajo en equipo
- Comunicación
- Capacidad de adaptación
- Creatividad
- Liderazgo
- Innovación
- Autonomía

INTERESES
- Lectura
- Cine
- Atletismo
- Arte
- Teatro y conciertos

PERFIL

En este apartado de tu hoja de vida debes escribir tu experiencia profesional y habilidades más importantes. Ten en cuenta cuáles son las necesidades de la empresa a la que quieres aplicar y redacta tu perfil profesional teniendo en cuenta esto.

EXPERIENCIA PROFESIONAL

Puesto ocupado
NOMBRE DE LA EMPRESA, Ciudad | 20XX – 20XX
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx. Xxxxxxxxxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx. xxxxxx xxxxxxxxxxxxxxxxxxxx.

Puesto ocupado
NOMBRE DE LA EMPRESA, Ciudad | 20XX-20XX
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.

Puesto ocupado
NOMBRE DE LA EMPRESA, Ciudad | 20XX
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxxxx.
- Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.

FORMACIÓN

Nombre del grado o título obtenido
Nombre de la Institución, Ciudad | 20XX – 20XX

Nombre del grado o título obtenido
Nombre de la Institución, Ciudad | 20XX – 20XX

Figura 2. Tipos de hojas de vida.
Fuente 1: https://es.wikipedia.org/wiki/Curriculum_vitae
Fuente 2: https://www.hoja-de-vida.co

5. Selección de métricas técnicas y de negocio

Este es un problema interesante ya que encaja como un problema de clasificación como tal, ya que se extraer información correctamente de una fuente de origen. Sin embargo sí se puede aplicar una métrica propia de problemas de clasificación en esta solución la cual es la matriz de confusión (Figura 3).

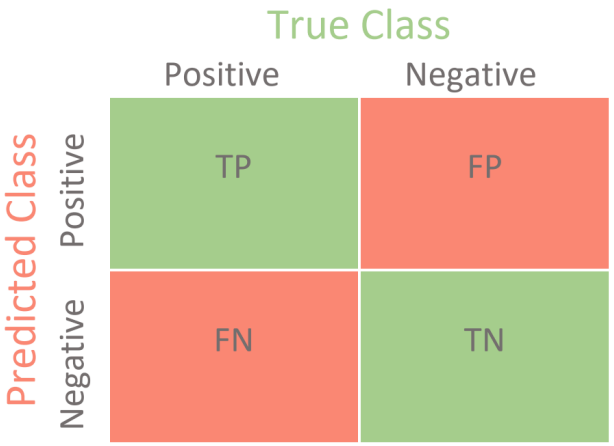


Figura 3. Matriz de Confusión
Fuente: towardsdatascience.com

La forma de evaluar el desempeño del modelo se realizará de la siguiente manera:

- Extracción correcta de la información (número de acierto)
- Extracción incorrecta de la información (número de no aciertos).
- Extracción de información cuando no está presente (falso positivo)
- No extracción de información cuando está presente (falso negativo)

Estas serán las métricas para evaluar el desempeño del modelo.

6. Preprocesamiento de datos

El preprocesamiento de datos, en este caso, archivos pdf o docx, se realiza en el fichero **`extraccion.py`** el cual contiene dos funciones: la primera llamada **`extraer_texto_word`** extrae el texto de documentos con formato docx y la segunda llamada **`extraer_texto_pdf`** extrae el texto de documentos pdf. El procesamiento en esta solución no requiere mayor dedicación, al menos para archivos sencillos, como los mostrados en la Figura 2.

Figura 4. PDF o DOCX a texto
Fuente: https://www.pdfmate.com/pdf-to-text.html

7. Selección de la arquitectura de modelo

Dado a que fue un modelo utilizado dentro del bootcamp y en general obtuvo buenos resultados, Gwen 1.5 1.8B fue el seleccionado para ser el corazón de la solución. Además se ha elegido porque es una opción open source y es relativamente pequeño para ejecutarse en un computador “normal” de forma local o también se puede ejecutar en Google Colab (que para la demostración del funcionamiento se adjuntará un notebook con los módulos de la solución).

Dentro del directorio del proyecto se encuentran los ficheros correspondientes a los procesos llevados a cabo durante la ejecución del programa. El fichero main.py es el único que debe ejecutarse, dentro de este fichero se importan los siguientes módulos:

- extraccion.py: Contiene las dos funciones que extraen información de los pdf o docx.
- model.py: Este fichero contiene la definición e inicialización del modelo a utilizar.
- prompts.py: Aquí se encuentran los prompts que dan las instrucciones de ejecución al modelo LLM.
- posprocesado.py: contiene las funciones que convierten la salida del modelo en un objeto de Python para su manipulación y su respectivo almacenamiento.

8. Entrenamiento del modelo

Debido al comportamiento del modelo la calibración del mismo se enfocó fuertemente en la elaboración de prompts. También se ajustó la temperatura a cero para tratar de obtener siempre el mismo resultado teniendo en cuenta el mismo input. Una de las particularidades que se evidenciaron al momento de crear los prompts es que cuanto más sencillo sea el prompt, mejor suelen ser los resultados.

Teniendo en cuenta que a menor prompt mejores resultados, se dividió el trabajo del modelo en tres llamados al modelo LLM. Uno que hace la extracción de la información básica y de contacto, otro que extrae la experiencia profesional junto con sus fechas y finalmente un tercer prompt que procesa la salida del segundo prompt.

Junto con el ajuste de la temperatura del modelo, otro parámetro que se ajustó bastante durante el “el entrenamiento” fue la salida máxima del modelo. Quizá esta es una de las limitaciones más grandes ya que si la salida del modelo es muy grande, este tiende a seguir generando texto que no se le ha solicitado en el prompt, evento que podríamos catalogar dentro de las alucinaciones del modelo.

9. Evaluación del modelo

Debido a la dificultad para conseguir información sobre hojas de vida en internet, se ha optado por utilizar un modelo generativo (en este caso GPT) para generar hojas de vida artificiales con información inventada (eso espero) para probar las capacidades de extracción del modelo.

En la Tabla 1 se presentan los registros utilizados en el modelo y se ha realizado la extracción de la información de forma manual para poder comparar y evaluar el modelo más adelante. Adicionalmente se incluyen estos documentos en formato pdf en los anexos de este proyecto.

Tabla 1. Registros base para evaluación del modelo

CV	Nombre Candidato	Teléfono	Correo	Años Mayor Experiencia	Puesto Mayor Experiencia
CV _1 GPT	Carlos Alberto Martínez Gómez	+34 612 345 678	carlos.martinez@gmail.com	4	Desarrollador de Software Senior
CV_2 GPT	Laura Fernández López	+34 678 901 234	laura.fernandez@gmail.com	5	Psicóloga Clínica
CV_3 GPT	Juan Carlos Rodríguez García	+34 645 123 456	juan.rodriguez@gmail.com	5	Ingeniero Civil Senior
CV_4 GPT	María Fernanda Gómez Martínez	+52 55 1234 5678	maria.gomez@gmail.com	6	Ingeniera Industrial
CV_5 GPT	Ana María Torres Rodríguez	+57 310 234 5678	ana.torres@gmail.com	0	NA
CV_6 GPT	Sebastián Valenzuela Ríos	+56 9 8765 4321	sebastian.valenzuela@gmail.com	6	Arquitecto de Software
CV_7 GPT	Mariana Silva Pereira	+55 11 9876 5432	mariana.silva@gmail.com	5	Analista de Datos
CV_8 GPT	Ekaterina Ivanova Petrova	+7 901 234 5678	ekaterina.petrova@gmail.com	6	Abogada Asociada
CV_9 GPT	Ana Catarina Ferreira	+351 91 784 6000	ana.ferreira@gmail.com	1	Técnica de Laboratorio Clínico
CV_10 GPT	-	-	-	1	Técnico de Laboratorio Clínico
CV_11 Real	Sergio Alejandro Alvarado Parada	57 3017394784	sergio.ing92@gmail.com	4	Ingeniero Civil
CV_12 GPT	Marcelo Gómez	+598 99 123 456	marcelo.gomez@gmail.com	6	Gerente de Servicio al Cliente
CV_13 GPT	Emily Johnson	+1 416 555 1234	emily.johnson@gmail.com	15	Gerente de Proyectos
CV_14 GPT	Aiko Tanaka	+81 90 1234 5678	aiko.tanaka@gmail.com	0.5	Asistente de Investigación

Fuente: Elaboración propia.

Cada uno de los registros de la Tabla 1 fue analizado por el modelo y los resultados se presentan en la Tabla 2.

Tabla 2. Resultados del modelo

CV	Nombre Candidato	Teléfono	Correo	Años Mayor Experiencia	Puesto Mayor Experiencia
CV _1 GPT	Carlos Alberto Martínez Gómez	+34 612 345 678	carlos.martinez@gmail.com	4	-
CV_2 GPT	Laura Fernández López	+34 678 901 234	laureafernandez@gmail.com	6	-
CV_3 GPT	Juan Carlos Rodríguez García	+34 645 123 456	juan.rodriguez@gmail.com	2	-
CV_4 GPT	María Fernanda Gómez Martínez	+52 55 1234 5678	maria.gomez@gmail.com	5	Ingeniera Industrial
CV_5 GPT	Ana María Torres Rodríguez	+57 310 234 5678	ana.torres@gmail.com	-	Inglés
CV_6 GPT	Sebastián Valenzuela Ríos	+56 9 8765 4321	sebastián.valenzuela@gmail.com	5	Innovatech Solutions Arquitecto de Software
CV_7 GPT	Mariana Silva Pereira	+55 11 9876 5432	mariana.silva@gmail.com	5	Analista de Datos
CV_8 GPT	Ekaterina Ivanova Petrova	+7 901 234 5678	ekaterina.petrova@gmail.com	2	Firma Legal Smirnov & Asociados
CV_9 GPT	Ana Catarina Ferreira	+351 91 784 6000	ana.ferreira@gmail.com	-	Firma Legal Smirnov & Asociados
CV_10 GPT	Juan Luis	+54 915-6789 01	jluiversant@gmail.com	8	Maestría en Ciencias Biomédicas
CV_11 Real	Sergio Alejandro Alvarado Parada	+57 301 379 4784	sergio.ing92@gmail.com	4	Ingeniero Civil
CV_12 GPT	Marcelo Gómez	+598 99 123 456	marcelo.gomez@gmail.com	5	Banco Santander, Montevideo
CV_13 GPT	Emily Johnson	+1 416 555 1234	emily.johnson@gmail.com	15	Tech Solutions Inc., Toronto, ON
CV_14 GPT	Aiko Tanaka	+81 90 1234 5678	aiko.tanaka@gmail.com	1	Bioingeniería

Fuente: Elaboración propia.

Analizando los resultados obtenidos podemos organizarlos en matrices de confusión para cada una de las cinco categorías presentes en la extracción. De la Tabla 3 a la Tabla 7 se presentan las matrices de confusión mencionadas.

Tabla 3. Matriz de confusión para Nombre, Teléfono y Correo Electrónico

		Valores Verdaderos	
		Verdaderos	Falsos
Valores Predichos	Verdaderos	13	1
	Falsos	0	0

Fuente: Elaboración propia.

Tabla 4. Matriz de confusión para Años de experiencia

		Valores Verdaderos	
		Verdaderos	Falsos
Valores Predichos	Verdaderos	4	0
	Falsos	10	0

Fuente: Elaboración propia.

Tabla 5. Matriz de confusión para Años de experiencia

		Valores Verdaderos	
		Verdaderos	Falsos
Valores Predichos	Verdaderos	2	0
	Falsos	12	0

Fuente: Elaboración propia.

En la extracción de información básica del candidato el modelo tuvo un muy buen desempeño, es capaz de identificar el nombre del candidato sin problemas, el correo también y el número de teléfono con su respectivo identificador de país. Solo cometió un error al crear información que no existía en la hoja de vida, generando información falsa, por esto se catalogó como un falso positivo.

Sin embargo, el modelo ha fallado en determinar la cantidad de años de experiencia y los puestos de experiencia de la mayoría de candidatos dentro de la evaluación. Analizando los resultados obtenidos tanto para años de experiencia como puesto de trabajo, se identificó que el modelo suele confundir con mucha frecuencia un puesto de trabajo con un estudio de formación académica. Este inconveniente se intentó solucionar con diferentes estrategias de prompting, sin embargo los resultados no mejoraron, desde intentar un chain of thoughts hasta una cadena de prompts individuales (que resultó ser la estrategia que “mejores” resultados arrojó). Dentro de la sección pasos a seguir se propondrán algunas ideas para implementar más adelante y poder mejorar estos resultados.

10. Pasos a seguir

A partir de los resultados obtenidos y considerando que no fueron los mejores, para futuras modificaciones del proyecto se incluyen los siguientes cambios:

- Realizar un fine tuning y brindar ejemplos al modelo para poder aumentar el número de aciertos.
- Buscar modelos quizás más capaces, quizá con un mayor número de parámetros para mejorar la detección de información y también reducir falsos negativos. Intentando siempre utilizar modelos open source, podría implementarse un modelo de pago a través de una API, como GPT.
- Implementar características de formateo para evitar incluir el formato de salida dentro del prompt con librerías como LangChain.
- Adicionar una interfaz gráfica para facilitar la adición de una gran cantidad de documentos para que se analicen de forma secuencial (en un principio).
- Incluir funciones de web scraping para ampliar la búsqueda de candidatos en plataformas de empleo.
- Implementar funciones CRUD para poder almacenar la información recopilada en una base de datos más allá de conservar la información en archivos de excel.
- Ampliar más el manejo de excepciones en casos de alucinaciones del modelo.
- Una vez se tenga implementado un mejor modelo será posible implementar la función de agregar los requerimientos para automatizar aún más el proceso de selección.

11. Conclusiones

- Las alucinaciones del modelo en este caso de uso particular no son un problema mayor, ya que en el peor de los casos el modelo inventará la información y no permitiría contactar al candidato, que sería ya de por sí raro que una persona enviara su hoja de vida sin incluir los datos de contacto pero podría pasar.
- El número de registros utilizados es muy bajito para dar un porcentaje de evaluación más acertado y confiable.