Avances reto estudios de seguridad y gestión documental

Planteamiento inicial del proyecto

El reto consiste en completar dos objetivos principales el primero es “desarrollar un modelo autónomo que se encargue de validar los estudios de seguridad y la búsqueda de antecedentes que determinan los posibles factores que intervienen en la selección de un candidato” y el segundo es “Implementar un modelo de análisis de información que se encargue de realizar de forma autónoma los procesos de validación documental para concretar la contratación de las personas.” Ambos objetivos se enfocan en construir un modelo de inteligencia artificial que se entrene por medio de un usuario capacitado para que este realice la búsqueda, el análisis, la validación, la transformación y la estructuración de los datos para que por medio de este se pueda determinar un score que permita calificar la compatibilidad de un candidato respecto a un cargo de acuerdo a la información encontrada en sus estudios de seguridad y su documentación.

Tratamiento de datos

Uno de los objetivos específicos del proyecto consiste en tomar los datos que nos proporciona Zoho y procesarlos para estructurar la información, para realizar este procedimiento se realizó la creación de varias vistas que permiten centralizar la información y de igual forma sectorizarla para poder analizarle de forma más sencilla. Las rutas de esos informes se encuentran de la siguiente manera:

Estudios de seguridad – Análisis reporte de antecedentes

Para el modelo de consulta de antecedentes se exporto la información de zoho, la cual proporciona los datos del candidatos con sus respectivos hallazgos, el resumen de consulta de antecedentes y cada uno de los antecedentes encontrados por página, esta información es la que se obtienen de los candidatos por medio de la plataforma de tus datos, el reporte de análisis permite obtener el json que proporciona esta página (Tus datos) y de esta forma evaluar cada una de las consultas, este formulario se importó como csv para realizar un análisis de datos en Python, para este proceso se crearon dos archivos el primero se emplea para analizar principalmente el JSON de tus datos y poder ver cada antecedentes en un dataframe, de esta forma se podrá ver cada antecedentes y realizar estudios de hallazgos, de datos demográficos, de afiliaciones, antecedentes judiciales, búsqueda en Google, entre otros. El segundo archivo se centra en validar la información de los antecedentes de los candidatos de forma individual, para esto se creó un formulario en zoho que permite visualizar cada antecedente de forma individual.

Vistas desarrollo - Análisis de datos gestión documental

Para el modelo de gestión documental se creó una vista en zoho que centraliza toda la documentación de los candidatos desde que ingresan en el sistema hasta que se contratan o se retiran, esta vista trae información de hoja de vida, de aplicar convocatorias, de documentos postulados, de la requisición, de contratación, de compensación e incluso de consulta de

antecedentes, aquí se puede ver cada aspecto del candidato referentes a su información personal y a cada documentación durante su proceso en el sistema. Esta información se exporta como csv a un script de Python para poder ser procesada, adicionalmente a la creación de la vista se realizó un flujo que permite actualizar la información de historia laboral, otros estudios, otros idiomas, referencias personales y núcleo familiar, estos campos son subformularios internos de hoja de vida por tal razón no se pueden traer por medio de lookups en este formulario, por ende el flujo se encarga de concatenar esta información para que sea incluida en la vista de análisis documental.

Nota: Este proceso de importación desde zoho debe realizarse por medio de un API para que se esté actualizando la información que se procesa en Python cada vez que ingresa un nuevo candidato

Limpieza de información

El tratamiento de datos tanto para estudios de seguridad como para gestión documental son incorporados en dos scripts de jupyter donde se realiza un preprocesamiento de la data, en ellos se busca realizar la limpieza de todos los atributos, variables nulas, outliers, datos vacíos , erróneos y cual información incongruente, cada programa se encarga de analizar la información faltante en cada una de las variables para poder transformar la base de datos para su consulta en los procesos que se requieren y así no se generen errores de funcionalidad por algún dato en mal estado.

Análisis exploratorio de datos

Finalmente, los mismos scripts de jupyter se emplean para realizar el análisis exploratorio de datos, es este se busca analizar cada aspecto de la información que nos proporciona las bases de datos traída de zoho, para consulta de antecedentes y gestión documental se busca sacar análisis de la información de progreso del candidato en el cargue de documentos, en los posibles antecedentes que puedan presentar un problema al cliente o en desarrollo de la ejecución de su trabajo, los atributos que no son relevantes, ya sea por la relevancia de la variable o por la falta cantidad de datos nulos o vacíos, las variables que están correlacionas, los análisis invariados y multivariados, las variables que deben corregirse de entrada para su análisis las variables que deben imputarse o cambiar su formato como es el tema de las fechas o las edades, entre muchos otros aspectos que se han venido construyendo en estos archivos empleados para la ejecución del EDA.

Estos scripts no solo se encargan de analizar la información si no que proporcionan un archivo csv con la data corregida para que esta ingrese dentro del proceso de consulta y validación que es donde se estructuraran los modelos, los cuales se explican en las siguientes secciones

Avances modelo estudios de seguridad

En la sección de estudios de seguridad se realizó la recopilación de cada uno de los antecedentes que se realizan como consulta por medio de tus datos y se implementó un código para realizar la búsqueda de antecedentes en diferentes plataformas. El proceso de web scraping permite obtener la información con más detalle y trabajar cada dato de manera puntual, es decir toda la información que se obtenga mediante una consulta automática podrá ser almacenado en un dataframe y permitirá ser analizada de forma independiente. Actualmente la plataforma de Tus Datos proporciona alrededor de 60 consultas las cuales muchas de ellas no son tomadas en cuenta y otras no son necesarias de acuerdo con esto se realizó una tabla donde se observa cada una de las paginas donde se debe realizar consulta de antecedentes y a cuáles de ellas se les puede hacer web scraping.

antecedente	link	captcha
Interpol	https://www.interpol.int/es/Como-trabajamos/Notificaciones/Ver-las-notificaciones-rojas	libre
OFAC	https://sanctionssearch.ofac.treas.gov/	libre
Lista ONU	https://www.un.org/securitycouncil/es	libre
Proveedores ficticios	https://www.dian.gov.co/Paginas/Resultados.aspx?k=proveedores+ficticios	libre
Postulados (Desmovilizados)	https://www.fiscalia.gov.co/colombia/justicia-transicional-2/consulta-postulados/	libre
Europol	https://eumostwanted.eu/es	libre
Rama unificada	https://consultaprocesos.ramajudicial.gov.co/Procesos/Index	libre
SIMIT	https://fcm.org.co/simit/#/home-public	libre
RUAF	https://www.sispro.gov.co/central-prestadores-de-servicios/Pages/RUAF-Registro-Unico-de-Afiliados.aspx	libre
PEPS	https://www.compliance.com.co/que-o-quienes-son-pep-que-es-una-lista-de-peps/	libre
Consulta de procesos judiciales	https://procesos.ramajudicial.gov.co/procesoscs/ConsultaJusticias21.aspx?EntryId=KhFNqaxOiCvDXco6pZpZH53UrL8%3d	libre
Rnmc (Medidas Correctivas)	https://srvcnpc.policia.gov.co/PSC/frm_cnp_consulta.aspx	libre
Libreta militar	https://www.libretamilitar.mil.co/	libre
Contadores Sancionados	https://www.jcc.gov.co/es/estadisticas-de-contadores	libre
Garantías mobiliarias	https://www.garantiasmobiliarias.com.co/	libre
Sisbén	https://www.sisben.gov.co/paginas/consulta-tu-grupo.aspx	libre
certificado registraduría	https://registraduria.gov.co/	tiene
Policía	https://www.policia.gov.co/	tiene
Procuraduría	https://www.procuraduria.gov.co/portal/	tiene
Contaduría	https://www.contaduria.gov.co/	tiene
Contraloría	https://www.contraloria.gov.co/	tiene
Concordato (SuperSociedades)	https://www.supersociedades.gov.co/imagenes/boletin/A%C3%91O%202012/CONCORDATOS/CONCORDATOS%20ACUMULADOS%20OCTUBRE%2031%20%20DE%202012.htm	libre
INPEC	https://www.inpec.gov.co/	tiene
personería Bogotá	https://www.personeriabogota.gov.co/	tiene

Registraduría Certificado	https://wsp.registraduria.gov.co/estado_docs/documento/consultar/	tiene
SECOP S	https://colombiacompra.gov.co/secop/consulte-en-el-secop-i	libre
Juzgados	https://www.ramajudicial.gov.co/juzgados-de-ejecucion-de-penas-y-medidas-de-seguridad	pendiente
Juzgados Tyba	https://procesojudicial.ramajudicial.gov.co/Justicia21/Administracion/Ciudadanos/frmConsulta	tiene
SECOP	https://www.contratos.gov.co/consultas/resultadoListadoProcesos.jsp#	libre
SECOP 2	https://colombiacompra.gov.co/secop-ii	tiene
PEPS	https://www.uiaf.gov.co/transparencia/informacion_interes/glosario/personas_politicamente_expuestas_7122	pendiente
SIMUR	https://www.movilidadbogota.gov.co/web/consulta_de_comparendos	tiene

Las casillas marcadas en verde son las páginas a las cuales ya se realizó el código de web scraping, las azules son las que están pendientes y las blancas son páginas a las cuales no se les puede realizar web scraping debido a que estas poseen un captcha que no permite la consulta automática.

Para el desarrollo de este primer modelo de consulta de antecedentes se implementó cada programa en Python donde se realiza cada paso necesario para hacer la consulta automática en cada página, esto fue necesario trabajarlo de forma independiente ya que cada una de las paginas presenta una particularidad en su arquitectura y para poder realizar la consulta es necesario seguir una serie pasos determinados, los cuales pueden ser más complejos o sencillos dependiendo de la entidad. Se realizaron 16 códigos individuales para realizaron consultas a las siguientes entidades:

- Interpol
- OFAC
- Lista ONU
- Proveedores ficticios
- Postulados (Desmovilizados)
- Europol
- Rama unificada
- SIMIT
- RUAF
- PEPS
- Consulta de procesos judiciales
- Rnmc (Medidas Correctivas)
- Libreta militar
- Contadores Sancionados
- Garantías mobiliarias
- Sisbén

Cabe aclar que todas estas entidades proporcionan información general del candidato referentes a sus antecedentes sin embargo las plataformas a que proporcionan información son la contraloría,

la procuraduría, el INPEC y la policía nacional, estas plataformas utilizan un catpcha para evitar la entrada de bots por tal razón para poder consultarlas debe realizarse por medio de un API, se realizó la consulta por medio de la web y se buscaron cotizaciones para los accesos a las bases de datos pero es necesario contactar a estas entidades directamente para obtener respuesta de la funcionalidad de sus APIs. Por medio de web scraping se realizó la consulta de la página de la policía nacional para medidas correctivas y la página de juzgados, tanto la unificada como la de general donde se puede validar si el candidato tiene un proceso judicial abierto o si ha estado en la cárcel, estas dos entidades son algunas de las páginas que proporcionan información más relevante de los antecedentes del candidato.

El modelo de programación que se utilizó sigue el siguiente flujo:

1. Importación de librerías para realizar web scraping y realizar procesos de modificación de datos (Selenium, json, pandas, time, os, etc.)
2. Creación de dataframe para el almacenamiento de los resultados de consultada
3. Lectura de un archivo csv que contiene la base de datos transformada y procesada
4. Creación de variables de consulta para el almacenamiento de la información
5. Creación de un ciclo de consulta
6. Parámetros de indexación de las consultas a las variables de consulta
7. Definir opciones de navegación
8. Inicialización de las pantallas y el navegador
9. Declaración de condiciones y pasos para realizar web scraping
10. Declarar tiempos de espera
11. Declarar parametrización de fallos
12. Ingresar información al dataframe de resultados
13. Tomar pantallazos de las consultas
14. Almacenar archivos pdf o imprimir las vistas en documentos de las consultas
15. Cierre de pantallas y del navegador
16. Generar archivo csv

De acuerdo a lo anterior se comenzó a estructurar un script en Python donde se unificarán todas las consultas realizadas por web scraping y las APIs de todas las plataformas que nos proporcionen datos sobre los antecedentes del candidato y sobre su comportamiento, gustos, intereses, hobbies, y demás, todo lo que pueda obtenerse de la web se esperar que se incorpore en el programa, de esta forma solo es ingresar algunos datos personales del candidatos como es la cedula, la fecha de expedición, la nacionalidad, el departamento y la ciudad de residencia, entre otros.

El archivo que se implementó maneja un flujo similar al de las consultas individuales de web scraping solo que la diferencia consiste en su estructura más robusta y en una funcionalidad ininterrumpida, la estructura de este condigo es la siguiente:

1. Importación de librerías
2. Creación de dataframe para el almacenamiento de los resultados de consultada (Con la información del candidato y con los campos de los resultados)
3. Importación de datos desde zoho por medio del API
4. Procesamiento de la información
5. Declaración de variables de consulta

6. creación de ciclos de consulta
7. Indexación de información en las variables de consulta
8. Definición de parámetros de configuración
9. Despliegue de consultas
10. creación de ciclos de consulta
11. Declaración de pasos para la consulta de páginas (web scraping)
12. Extracción de archivos (pdf, imágenes, pantallazos) e información
13. Ingresar resultados en el dataframe
14. Cierre de aplicaciones
15. Repetición de ciclos de consulta
16. Concatenación de las consultas al dataframe
17. Validación de errores
18. Construcción de modelo relacional de datos

Avances modelos validación documental

El proceso de gestión documental requiere inicialmente de un estrcuturamiento de la información donde se pueda visualizar cada uno de los archivos de un candidato desde que se postula hasta que se contrata o se retira tal como se explicó en la sección de tratamiento de datos, esto es necesario ya que la información debe descargarse de forma masiva de todos los candidatos que existan en zoho para poder conformar el modelo y entrenarlo.

Como modo de prueba se realizó la descarga de todos los archivos de un candidato y se estructuro la arquitectura del código para que se comiencen a validar cada documento, la mayoría de estos archivos vienen en PDF o imagen por esta razón se emplearon varias librerías de Python que permiten realizar estas validaciones.

Proyección Futura

El reto es un proceso que busca crear una inteligencia artificial que realice procesos autónomos dentro y fuera de zoho, por esta razón requiere del apoyo de varias herramientas tecnológicas que permitan realizar la construcción de los procesos esperados así como la integración de un ambiente donde la información se este procesando analizando, validando y buscando cada vez mas datos que alimenten el modelo y que permitan aplicar machine learning, en cada uno de los flujos donde se realicen las validaciones y el análisis de datos para calificar al candidato.

La primera etapa que se ha venido realizando esta enfoca fuertemente en la obtención y transformación de la información, la creación del modelo, su testeo y su entrenamiento requiere un modelo de datos relacional, es decir una gran base de datos consolidada con muchos datos que permita tomar como punto de referencia para la ejecución de los procesos, teniendo en cuenta lo anterior la continuidad de estas tareas se centran en las siguientes tareas:

- Unificar la consulta de antecedentes por medio de web scraping
- Realizar las consultas APIs a las entidades más importantes requeridas en la consulta de antecedentes
- Realizar web scraping a páginas que permitan deducir el comportamiento del usuario (Redes Sociales, Consultas en navegadores, etc.)

- Validación de información Laboral ingresada por el usuario
- Analizar la veracidad de la documentación subida por el candidato comparándola con las consultas realizadas por web scraping a las entidades que proporcionan información personal del candidato como es la registraduría, la RUAF (Afiliaciones), El address (EPS – caja de compensación) de igual forma con la misma información que proporciona el candidato.
- Validar más páginas que permitan comprobar la veracidad de la documentación del candidato
- creación de un flujo en zoho para la descarga masiva en zoho
- Análisis de documentación individual
- Análisis de documentación Masiva
- Creación de dataframe unificado con toda la información y compararlo con la data analizada en el EDA con la base de datos de zoho
- Validar la información obtenida con web scraping, generar insights y compararlo con el reporte y la información que proporciona tus datos para validar la efectividad de la implementación del proceso

Cronograma de tiempos

Tarea	TITULO ÁREA	TAREA DUEÑO	EMPEZAR FECHA	PENDIENTE FECHA	DURACIÓN EN DÍAS	PCT DE LA TAREA INTEGRO
1	Planeación y planteamiento del problema					
	Identificación del problema	DAVID	09/13/22	09/15/22	2	**100%**
	Estructuración el documento	DAVID	09/15/22	09/17/22	2	**100%**
	descripción del problema (hipotesis)	DAVID	09/15/22	09/17/22	2	**100%**
	Objetivos del proyecto	DAVID	09/15/22	09/17/22	2	**100%**
	Datos aspiracionales	DAVID	09/15/22	09/17/22	2	**100%**
2	recolección y extracción de datos					
	Búsqueda de información disponible	DAVID	09/19/22	09/21/22	2	**60%**
	Organización y estructuración de información en Zoho	DAVID	09/22/22	10/07/22	15	**40%**
	Preparación de datos	DAVID	09/22/22	10/07/22	15	**50%**
	Creación de entorno de Zoho para	DAVID	09/23/22	10/07/22	14	**50%**

	el análisis de información					
	Consolidación e integración de bases de datos en datos	DAVID	09/23/22	10/07/22	14	**30%**
	Web Scraping	DAVID	09/28/22	11/14/22	47	**50%**
	Definición de datos para las consultas	DAVID	10/01/22	10/30/22	29	**70%**
	Búsqueda y estructuración de los códigos para realizar la consulta	DAVID	10/01/22	10/30/22	29	**50%**
	Documentación	DAVID	10/01/22	10/30/22	29	**45%**
	Antecedentes	DAVID	10/01/22	10/30/22	29	**50%**
	Datos aspiracionales	DAVID	10/01/22	10/30/22	29	**30%**
	Creación de algoritmo para automatizar las consultas a cada pagina	DAVID	10/01/22	10/30/22	29	**50%**
	Agrupación de la información de cada una de las consultas (MODELO MDR)	DAVID	10/30/22	11/11/22	12	**10%**
	Sincronización con Zoho para la conformación de los datos a consultar	DAVID	11/01/22	11/12/22	11	**10%**
3	Análisis de datos					
	Estado y seguimiento	DAVID	11/12/22	12/12/22	30	**0%**
	Cargar bases de datos	DAVID	11/12/22	12/12/22	30	**0%**
	Análisis exploratorio de datos	DAVID	11/12/22	12/12/22	30	**0%**
	Distinción de atributos	DAVID	11/12/22	12/12/22	30	**0%**
	Análisis univariado	DAVID	11/12/22	12/12/22	30	**0%**
	Análisis multivariado	DAVID	11/12/22	12/12/22	30	**0%**
	Detectar valores perdidos o anormales	DAVID	11/12/22	12/12/22	30	**0%**
	Detectar outliers	DAVID	11/12/22	12/12/22	30	**0%**

	Rasgos de ingeniería	DAVID	11/12/22	12/12/22	30	**0%**
	Mostrar visión de negocio	DAVID	11/12/22	12/12/22	30	**0%**
4	Modelamiento					
	Transformación y selección de datos	DAVID	12/13/22	01/12/23	30	**0%**
	Selección de los modelos	DAVID	12/13/22	01/12/23	30	**0%**
	Construcción del modelos	DAVID	12/13/22	01/12/23	30	**0%**
	Validación del modelo	DAVID	12/13/22	01/12/23	30	**0%**
5	Evaluación					
	Despliegue del modelo	DAVID	01/13/23	01/28/23	15	**0%**
	Evaluación y monitores de resultados	DAVID	01/13/23	01/28/23	15	**0%**
6	Generación de insights y reportes					
	Estructuración de la información a presentar	DAVID	01/28/23	02/27/23	30	**0%**
	Selección de software para la presentación del reporte	DAVID	01/28/23	02/27/23	30	**0%**
	Definir el medio por el cual se va enviar la información	DAVID	01/28/23	02/27/23	30	**0%**
7	Despliegue					
	incorporar al sistema de zoho	DAVID	02/28/23	03/30/23	30	**0%**
	vincular las respectivas plataformas y desarrollos realizados	DAVID	02/28/23	03/30/23	30	
8	Toma de decisiones					
	Determinar los beneficios y el valor agregado con base en la información analizada	GENERENCIA	03/01/23	03/31/23	30	**0%**
	establecer las determinadas decisiones de acuerdo a las métricas y las perspectivas realizadas	GENERENCIA	03/01/23	03/31/23	30	**0%**

	Conclusiones	GENERENCIA	03/01/23	03/31/23	30	**0%**