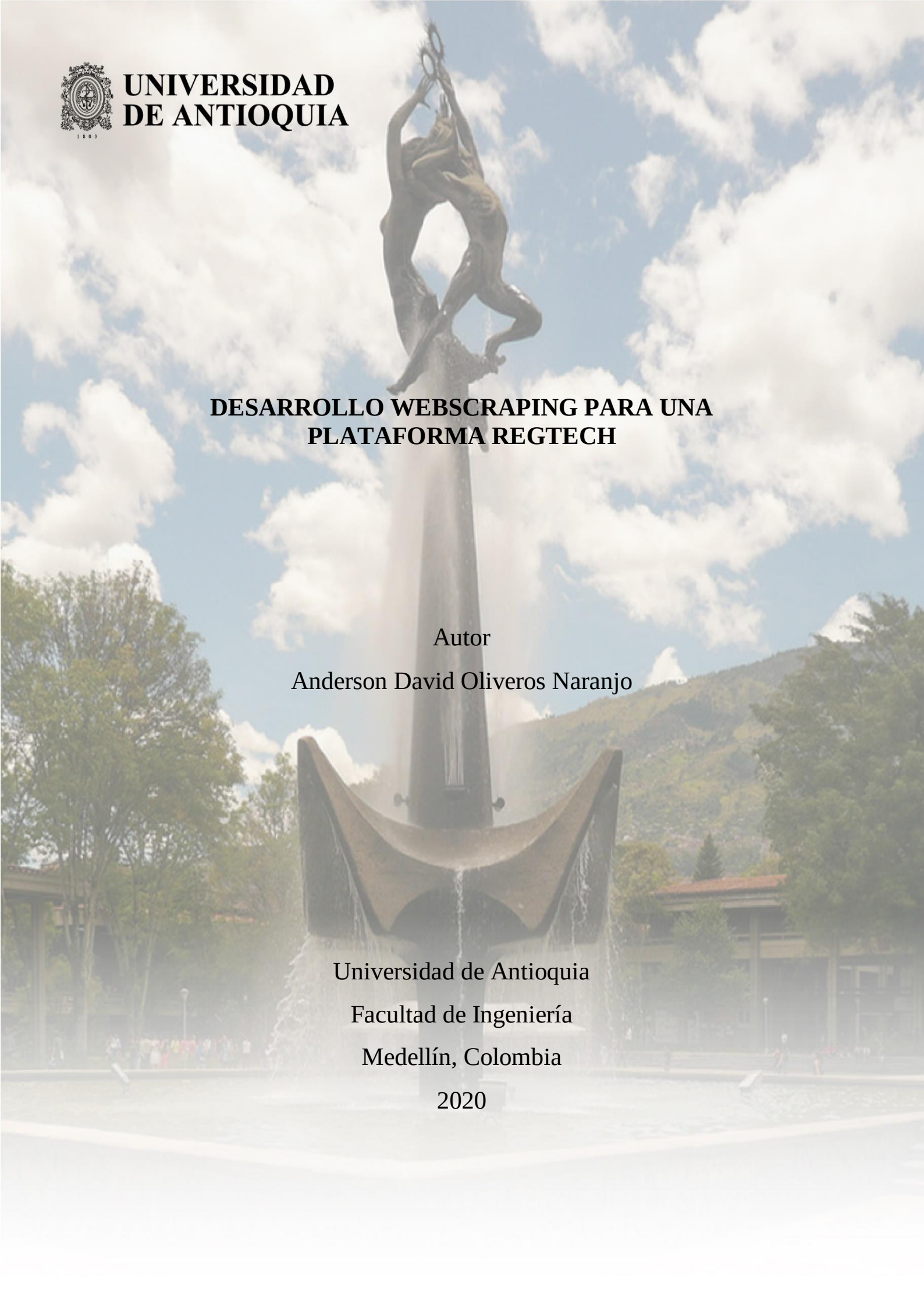

UNIVERSIDAD DE ANTIOQUIA

DESARROLLO WEBSCRAPING PARA UNA PLATAFORMA REGTECH

Autor

Anderson David Oliveros Naranjo

Universidad de Antioquia

Facultad de Ingeniería

Medellín, Colombia

2020

DESARROLLO WEBSCRAPING PARA UNA PLATAFORMA REGTECH DE BANCOLOMBIA

Anderson David Oliveros Naranjo

Practica empresarial como requisito parcial para optar al título de:

Ingeniería de Sistemas

Asesor interno:

Jaime Fonseca Espinal

Universidad de Antioquia

Facultad de Ingeniería

Medellín, Colombia

2020

Contenido

Resumen

La gerencia Investigación, Desarrollo e Innovación (I+D+I) perteneciente a la Vicepresidencia de Auditoria Interna del Grupo Bancolombia propuso desarrollar una iniciativa RegTech la cual permita a los clientes prevenir riesgos y adaptarse a los cambios. Por dicha razón se hizo un apoyo en el desarrollo de todo el proyecto asociado a la técnica webscraping en diferentes listas de control propuestas y definidas por el Banco. Para este desarrollo, fue necesario consultar la documentación sobre Django [2] que fue el framework propuesto para el desarrollo y estudiar sobre la técnica webscraping [3] la cual se definió para extraer la información de las páginas web, además a lo largo de esta práctica también se definió una metodología con la que se llevaran a cabo los objetivos. Por último, se hizo un análisis de los resultados y se obtuvieron unas conclusiones del proyecto realizado en esta práctica académica.

Introducción

El Grupo Bancolombia es una entidad financiera que día a día evoluciona al ritmo de las nuevas tecnologías y se expande conforme avanza el desarrollo de nueva información de sus clientes. En la gestión empresarial de Bancolombia, disponer de información actualizada y de calidad agrega gran valor a la organización, mejorando los servicios ofrecidos por el Banco, ya que actualmente con la implementación de la analítica de datos se puede entender mejor las necesidades del cliente y del mercado.

Cada una de las áreas del Banco entiende la importancia de aplicar analítica de datos en sus proyectos a desarrollar, y es aquí donde la Vicepresidencia de Auditoría se encarga de velar por el cumplimiento de los objetivos estratégicos de la organización por medio de las revisiones y acompañamiento continuo.

La Vicepresidencia de Auditoría Interna (VAI), adaptándose a las estrategias definidas por el Banco ha comenzado a implementar en sus auditorías soluciones innovadoras donde la Gerencia Investigación, Desarrollo e Innovación (I+D+I) se propone el desarrollo de una iniciativa *RegTech*.

El objetivo principal de esta práctica empresarial es desarrollar una solución tecnológica para el cumplimiento de requerimientos regulatorios de los clientes, ampliar el esquema de

prevención de riesgos, el cual se trabaja bajo el estándar internacional del modelo de las tres líneas de defensa y donde se busca desarrollar una línea de defensa cero con el fin de que los clientes puedan prevenir riesgos, sean más sostenibles y tengan mayor adaptación a los cambios.

Objetivo General

Apoyar al desarrollo de la plataforma REGTECH mediante la descarga, integración y automatización de información de clientes de diversas fuentes públicas que permitan profundizar el conocimiento del cliente de manera integral.

Objetivos Específicos

- Analizar listas de control públicas o privadas para consultar la información disponible de un cliente.
- Desarrollar el código en el lenguaje de programación Python para la extracción de la información consultada en la lista de control.
- Integrar el código desarrollado a la plataforma creada para el proyecto.
- Realizar automatización y pruebas del código integrado a la plataforma.
- Planear las tareas a realizar cada semana como participación en equipo y con formas ágiles de trabajo.

Marco Teórico

Las *RegTech* son soluciones tecnológicas a los procesos regulatorios donde se comprende el uso de las tecnologías de la información en el contexto de la supervisión, la presentación de informes y el cumplimiento de la normativa [1]. Algunos de los servicios que se pretenden abordar son:

- Verificación y controles de identidad.
- Monitoreo de transacciones y auditoría.
- Detección de normas de cumplimiento de archivos y automatización de procesos de gestión.
- Verificar datos de usuario antes de formar una relación.
- Crear perfiles de riesgo LAFT (Lavado de Activos y Financiación de Terrorismo) para reducir falsos positivos y minimizar operatividad.
- Proyección escenario futuros.
- Cálculo exposición al riesgo
- Reporte de riesgos.

Django es un *framework* de aplicaciones web gratuito y de código abierto escrito en Python el cual contiene un conjunto de componentes que facilitan el desarrollo de sitios web, cuenta con ventajas como escalabilidad, seguridad, versatilidad, usabilidad y por estar escrito en Python se puede ejecutar en todas las plataformas sin problemas. [2]

El *Web Scraping* consiste en navegar automáticamente una web y extraer información de dicha página, esto se hace mediante un *Bot* de software, programado para ir a cualquier página web y obtener información mediante la estructura *HTML* del sitio web. [3]

Git es un sistema de control de versiones distribuido, gratuito, rápido y eficiente el cual permite versionar el código con el fin de guardar cambios efectuados para luego ver el historial de cambios o regresar a versiones anteriores, es compatible con cualquier plataforma web y además incluye un sistema de seguimiento de incidencias para atender todo tipo de problemas. [4]

El acompañamiento en este desarrollo estará compuesto por las siguientes cuatro etapas:

La primera etapa, consiste en verificar que la lista tenga un sitio web con información pública disponible para ser consultada por la plataforma, además se debe verificar que dicha información se pueda extraer mediante *Web Scraping.*

Luego de verificar las condiciones de la primera etapa se procede con el desarrollo del código en Django, en el cual se utilizarán librerías para el *Web Scraping* y dependiendo de la estructura del sitio web se hará el desarrollo.

Para la integración del desarrollo del código con el proyecto, se hará una sincronización con *Git* para que los demás desarrolladores tengan disponible el nuevo código y poder realizar después el ensamble con el apoyo de un desarrollador encargado del *Backend* de la plataforma.

Por último, se realizan las pruebas por medio de consultas para verificar que la plataforma sí obtenga la información correcta del sitio al que se le aplicó *Web Scraping.*

Metodología

El desarrollo de los objetivos anteriores se llevó a cabo de la siguiente manera:

1. Investigación y propuesta del proyecto:

    a) Inducción al desarrollo del proyecto por parte de dos auditores para el contexto y el desarrollo de cada componente.

    b) Consulta e investigación sobre las librerías en Python para hacer Web Scraping y cómo aplicarlas.

2. Desarrollo del proyecto:

a) Análisis y evaluación para cada una de las listas de control asignada.

b) Crear el código de Web Scraping para cada una de las listas que sean factibles.

c) Integrar el código creado a todo el proyecto trabajando en conjunto con un desarrollador encargado del Backend del proyecto.

d) Realizar las pruebas y hacer modificaciones de ser necesario con el fin de verificar el correcto funcionamiento de la integración del código.

3. Realimentación y mejoras:

a) Planear las tareas a realizar cada sprint con el fin de proponer objetivos y ver el avance en cuanto al cumplimiento de estos.

b) Asistir a los Daily virtuales para que todo el equipo tenga conocimiento de los avances del día anterior, los objetivos del día y los obstáculos que se presenten para continuar con el logro de estos.

Resultados y análisis

En el proyecto se recibió la asignación de cincuenta y cinco listas de control de las cuales a cuarenta se les aplicó un mismo proceso siendo guardadas con éxito; las restantes no fue posible hacerles webscraping debido a que tenían captcha o contaban con algún tipo de seguridad para proteger los datos.

El proceso que se les aplicó a las listas que fueron guardadas con éxito es el siguiente:

1. Análisis de página web y componentes HTML para conocer la estructura del código fuente.

2. Desarrollo en Python para extraer la información de la página web.

3. Integrar el desarrollo creado en la plataforma del proyecto.

4. Realizar pruebas y modificaciones en el desarrollo ya integrado.

Los resultados se presentarán con el desarrollo webscraping de una de las cuarenta listas de control guardadas con éxito, la cual es una página de la policía nacional de dominio y datos públicos que servirá de ejemplo para mostrar el proceso que se le aplicó a cada una de las listas, el cual fue el siguiente:

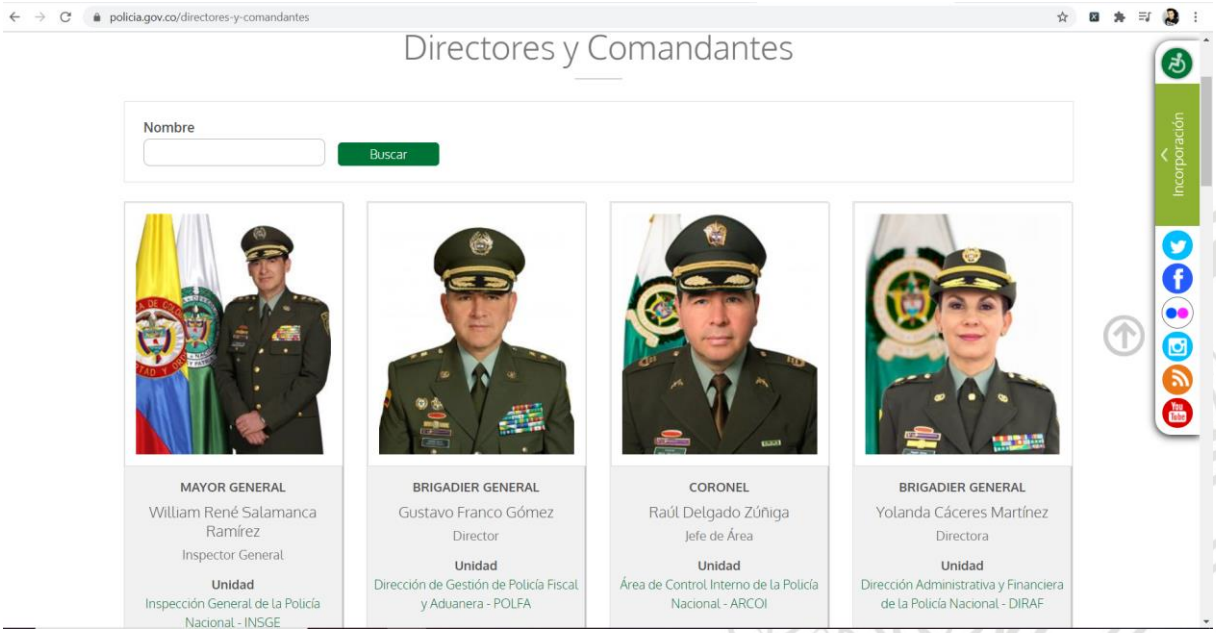

Figura 1: Pagina web lista de control

En la figura 1 se muestra la página web asociada a la lista de control de la policía nacional de Colombia; donde se observa con claridad una lista de los directores y comandantes de la institución la cual muestra el rango, nombre y la unidad a la que pertenece. Para esto se hará uso de la herramienta de desarrollador de google y la extensión XPath Helper para verificar los elementos de la página web.

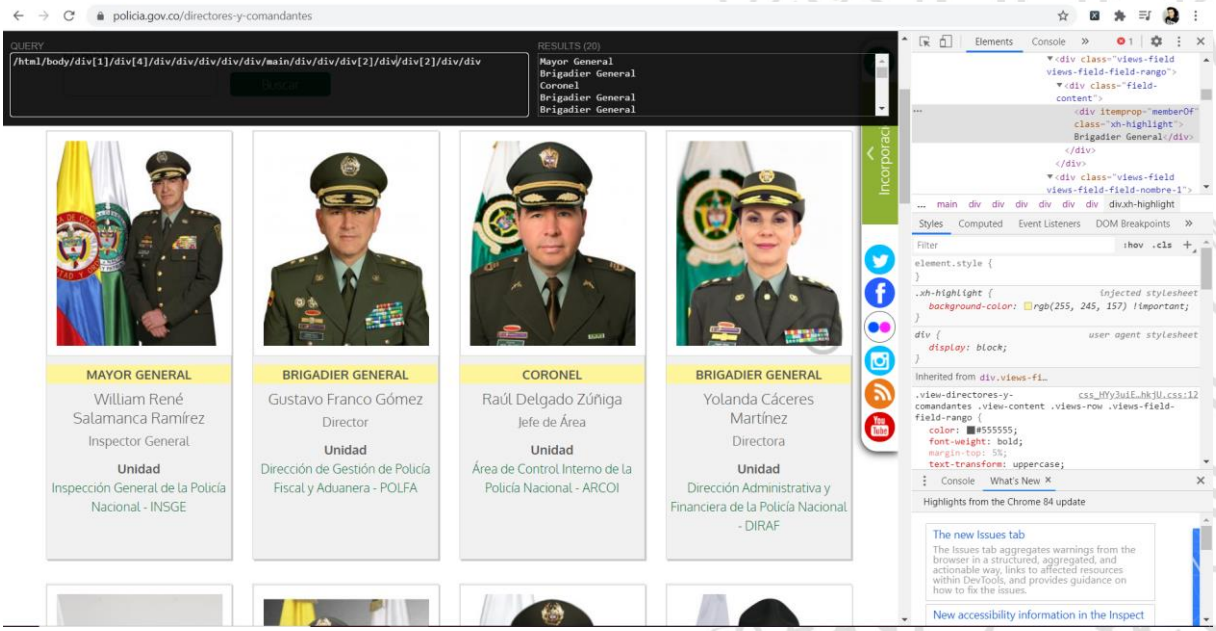

Figura 2: Lista rangos con elementos XPath de HTML

A continuación, se presenta la identificación por medio de la herramienta XPath Helper de cada elemento listado donde se puede ver el rango de todos los funcionarios de la policía que

aparecen en esta página, esto se hace por medio del XPath y la estructura HTML el cual permite obtener el texto asociado a cada uno de los elementos que en este caso es el rango.

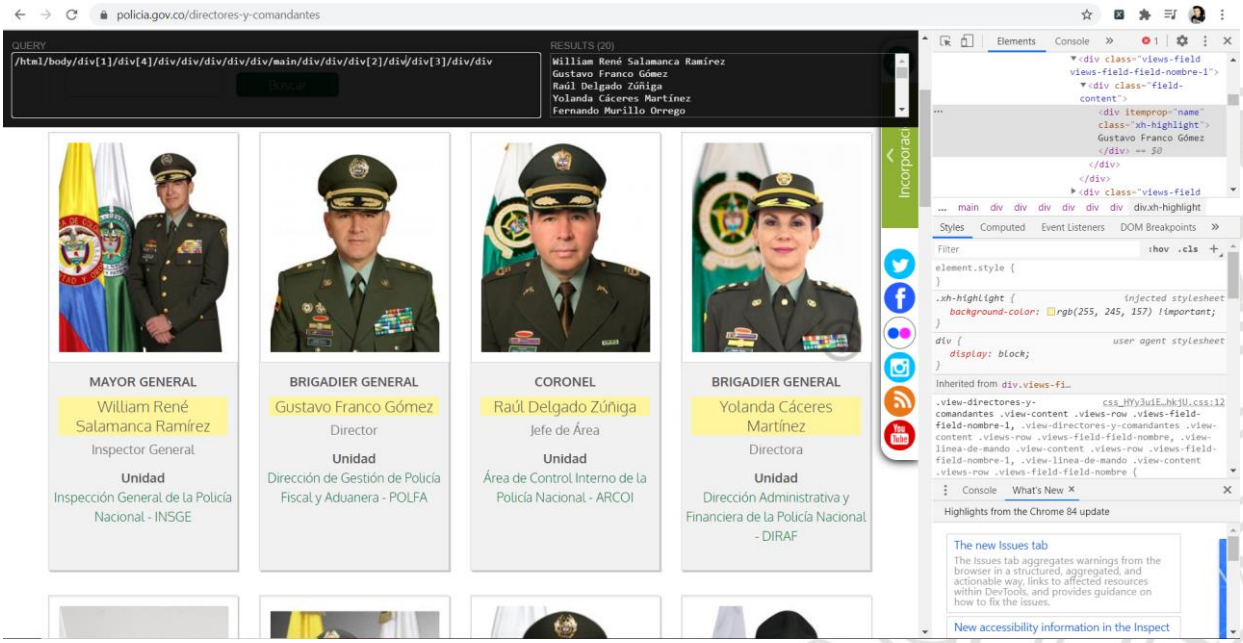

Figura 3: Lista nombre con elementos XPath de HTML

Después se presenta la identificación por nombre de cada elemento listado donde se puede ver dicho texto asociado a todos los funcionarios de la policía que aparecen en esta página.

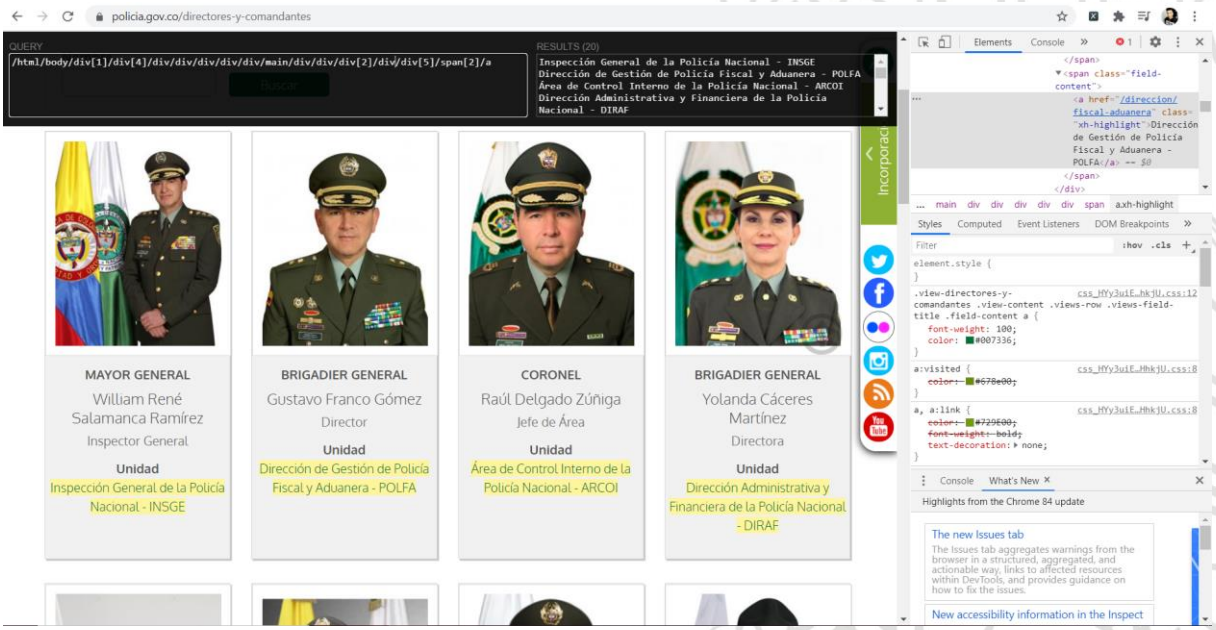

Figura 4: Lista unidad con elementos XPath de HTML

Luego con el mismo análisis de figura 2 y 3, se procede a obtener la unidad a la que pertenece cada funcionario de la policía donde con la estructura XPath se obtiene una lista con el nombre de todas las unidades correspondientes a cada miembro de la policía.

```
Lista_Rangos=driver.find_elements_by_xpath('/html/body/div[1]/div[4]/div/div/div/div/div/main/div/div/div[2]/div/div[2]/div/div')

Lista_Nombre=driver.find_elements_by_xpath('/html/body/div[1]/div[4]/div/div/div/div/div/main/div/div/div[2]/div/div[3]/div/div')

Lista_Unidad =driver.find_elements_by_xpath('/html/body/div[1]/div[4]/div/div/div/div/div/main/div/div/div[2]/div/div[5]/span[2]/a')
```

Figura 5: Búsqueda de elementos XPath en Python

En base a las librerías investigadas al inicio de la práctica, se utiliza el método 'find_elements_by_xpath' [5], donde se le da como parámetro el XPath identificado en las figuras anteriores para que el script encuentre los elementos de la página web y puedan ser guardados en un vector.

```
                Rango              Nombre_Funcionario                                Unidad a la que pertenece
0        MAYOR GENERAL   William René Salamanca Ramírez  Inspección General de la Policía Nacional - INSGE
1    BRIGADIER GENERAL             Gustavo Franco Gómez  Dirección de Gestión de Policía Fiscal y Aduan...
2              CORONEL               Raúl Delgado Zúñiga  Área de Control Interno de la Policía Nacional...
3    BRIGADIER GENERAL          Yolanda Cáceres Martínez  Dirección Administrativa y Financiera de la Po...
4    BRIGADIER GENERAL           Fernando Murillo Orrego  Dirección Antisecuestro y Antiextorsión de la ...
..                 ...                               ...                                                ...
96             CORONEL          Edilberto Garcia Guauta            Departamento de Policía Vichada - DEVIC
97             CORONEL  Juan Carlos Castellanos Álvarez      Oficina de Comunicaciones Estratégicas - COEST
98    TENIENTE CORONEL   Yasid Alberto Montaño Granados  Escuela Antidrogas "Mayor Wilson Quintero Mart...
99             CORONEL     Alba Patricia Lancheros Silva       Unidad Policial para la Edificación de la Paz
100   TENIENTE CORONEL       Carlos Antonio Ardila Rocha  Área de Relaciones y Cooperación Internacional...

[101 rows x 3 columns]
```

Figura 6: Resultado webscraping visualizado en Python

Por último, se presenta el resultado en Python de los datos obtenidos a la página web los cuales se guardarán en una base de datos para después ser integrados a la plataforma.

Dificultades

Una de las mayores dificultades fue la situación que se presentó debido a la pandemia causada por el Covid-19, el cual nos obligó a todos por seguridad trabajar desde la casa y que también ocasionó el cambio en el proyecto de mi práctica académica que originalmente era un tablero de control; Este cambio fue un gran reto para mí ya que tuve que acoplarme rápidamente al nuevo proyecto asignado, a nuevos compañeros de trabajo y a una metodología diferente a la que me habían explicado al principio, pero a pesar de todos los cambios asumí de buena manera el reto y utilicé todos los conocimientos adquiridos en la universidad para cumplir los objetivos del nuevo proyecto asignado el cual finalizó con éxito y satisfacción.

Ya en el contexto del proyecto, la mayor dificultad fueron las estructuras de los HTML y las malas prácticas de las páginas web a las que se les hacia webscraping, esto producía una mayor dificultad ya que había que buscar alternativas desde la lógica para obtener los datos sin errores.

Conclusiones

- Los conocimientos adquiridos en la universidad durante la carrera fueron una base fundamental para haber logrado todos los retos afrontados en esta práctica y cumplir con todos los objetivos propuestos por la empresa.

- La técnica de webscraping es un insumo muy valioso para las empresas que en sus proyectos buscan extraer información de sitios web, por lo que el conocimiento adquirido en esta práctica sobre la técnica webscraping fue una experiencia muy valiosa para la vida profesional de cualquier ingeniero de sistemas.

- La práctica académica es una oportunidad enorme para afrontar la transición de estudiante a ingeniero, ya que une los aspectos académicos e investigativos de la universidad con los requerimientos y necesidades del mundo laboral.

Bibliografía

[1] A. Douglas, J. Barberis, R. Buckley. *FinTech, RegTech, and the Reconceptualization of Financial Regulation*. Ed. 37. Nw. J. Int'l L. & Bus. 2017. [Internet] Disponible en: http://scholarlycommons.law.northwestern.edu /njilb/vol37/iss3/2

[2] Jeff Forcier, Paul Bissex, Wesley J Chun. Python Web Development with Django. 2008. [Internet] Disponible en: https://books.google.com.co/books?hl=es&lr=&id=M2D5nnYlmZoC&oi=fnd&pg=PT31&dq=django+python+framework&ots=vY_KCx8RPO&sig=Pwx_aJTFLEVJuCZvbzGQYClipMI&redir_esc=y#v=onepage&q=django%20python%20framework&f=false

[3] R. Mitchell. *Web Scraping with Python* . Ed. 2. EEUU. O' Reilly Media Inc. 2018. [Internet] Disponible en: https://books.google.com.co/books?hl=es&lr=&id=TYtSDwAAQBAJ&oi=fnd&pg=PT8&dq=web+scraping&ots=y0B3rIjofm&sig=btNFmGhnW4woMyKdeB615Cie9k&redir_esc=y#v=onepage&q=web%20scraping&f=false

[4] Kalliamvakou, E., Damian, D., Blincoe, K., Singer, L. and German, D., 2015. Open Source-Style Collaborative Development Practices in Commercial Projects Using GitHub. 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering.

[5] Selenium-python.readthedocs.io. 2020. Selenium With Python — Selenium Python Bindings 2 Documentation. [online] Available at: <https://selenium-python.readthedocs.io/> [Accessed 6 August 2020].