OFERTA DE INFORMACIÓN BENEFICIOSA PARA LOS ESTUDIANTES DE LA UNIVERSIDAD AUTÓNOMA DE OCCIDENTE REFERENTE A VIVIENDAS CERCANAS AL CAMPUS A UN COSTO ACCESIBLE

ESTUDIANTES

Camila Andrea Cardona Alzate

Gustavo Adolfo Chipantiza Manyoma

Maria Angelica Portocarrero Quintero

Xilena Atenea Rojas Salazar

DOCENTE

Jack Daniels Marquez Franco

UNIVERSIDAD AUTÓNOMA DE OCCIDENTE

INGENIERÍA DE DATOS E INTELIGENCIA ARTIFICIAL

Santiago de Cali

10 de Febrero del 2022

ÍNDICE

1.OBJETIVOS DEL PROYECTO:

Objetivo principal:

- Ofrecer información beneficiosa para los estudiantes de la Universidad Autónoma de Occidente, referente a viviendas cercanas al campus a un costo accesible.

Objetivos específicos:

- Evaluar el sitio web de la página escogida (fincaraiz.com).
- Recolectar información necesaria de apartamentos de Bochalema, como cantidad de habitaciones y precios por medio de técnicas de web scraping.
- Exponer la información recolectada sobre la url a través de la wiki del portal web Github.
- Implementar un código para el preprocesamiento de los datos para su posterior análisis y manejo.
- Analizar la información recolectada para la limpieza de los datos y su posterior manejo.
- Inferir conclusiones del proyecto a través de los resultados obtenidos.

2. EVALUACIÓN INICIAL DEL SITIO WEB:

La siguiente información también se encuentra en la Wiki del siguiente repositorio de Github: https://github.com/YOANGELICA/PROYECTO-PROGRAMACION

Análisis del archivo robots.txt:

En el archivo robots.txt de la página de Finca Raíz se pueden encontrar varios robots de búsqueda o *user agents*. Estos son los canales habilitados para que los usuarios entren y consulten la página web. Después de esos elementos se encuentran los directorios o archivos a los que los user agents no pueden acceder, como lo son */App_Modules/ , /comparar_inmuebles.aspx , /ClientAdmin/ ,* entre otros. Finalmente, en la parte de abajo están las urls donde se puede encontrar el mapa del sitio web en formato HTML.

Análisis del mapa del sitio web:

Los sitemaps del sitio web se encuentran en los siguientes enlaces:

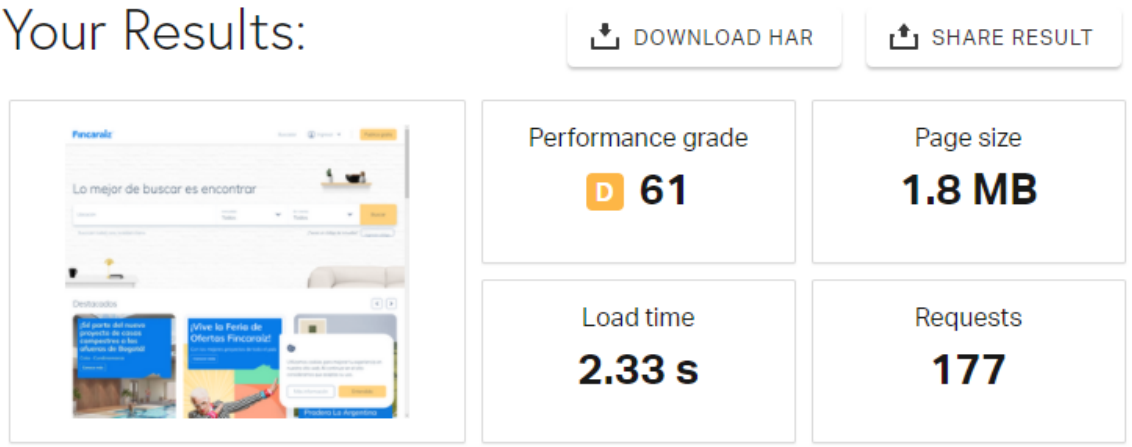
- http://www.fincaraiz.com.co/sitemaps/SiteMap_Index.xml
- http://www.fincaraiz.com.co/sitemaps/SiteMapFCCOL_Clients_Index.xml
- http://www.fincaraiz.com.co/sitemaps/image_sitemap.xml

En estos se encuentran los códigos estructurados en XML, y permiten la ejecución de una página dinámica, que incluye la descripción de la página web, y listas de

elementos que sirven como índice para navegar más fácilmente por la página. Estos índices se encuentran en la página principal y sirven para redireccionar a otras partes de la página como lo puede ser “Oferta de Finca Raíz en Colombia”, “Nuestros Clientes”, y “Servicio al cliente”. En el primer sitemap están en los links de los inmuebles publicados, junto a la fecha de la última modificación, todo dentro de nodos <url>. El segundo no tiene un link para descargar y visualizar el sitemap, abre una página web con una dirección web (https://www.sitemaps.org/schemas/sitemap/0.9/) que no contiene mucha información, mientras que el tercero no existe en el dominio de Fincaraiz.

Tamaño del sitio web:

Por medio de la herramienta *tools.pingdom.com* se encontró el tamaño de la página de Finca Raíz, el cual es 1.8 MB:



Whois:

La URL de fincaraiz.com.com fue creada el 18 de mayo del 2004 y actualizada por última vez el 23 de julio del 2021. La licencia del dominio o el registro de este tiene vigencia hasta el 27 de mayo del 2024. Tiene registrado como país Colombia, y de forma más específica, tiene registrado como “Estado” Bogotá. La demás información que se podría extraer referente al dueño de la página está completamente oculta ante este método de extracción de información a través de los dominios por motivos de privacidad.

Built with:

La página web fue construída con ASP.NET, utilizando el lenguaje de programación ASP. También se utilizó el lenguaje JavaScript, teniendo como herramienta Next.js.

3. PREPROCESAMIENTO DE LOS DATOS:

Para cumplir con los objetivos de este proyecto se creó un código estructurado para brindar al usuario que desee vivir cerca de la universidad (Específicamente en el barrio Bochalema) un filtro en el que este escoja la cantidad de personas que van a vivir en el apartamento, y se le dé como resultado las urls de los apartamentos indicados para la persona (tomando la cantidad de habitaciones como la cantidad de personas con las que viviría), junto a la cantidad de dinero que tendría que pagar de arriendo si viviese con la cantidad de personas especificada. El código se encuentra en los siguientes links:

https://github.com/YOANGELICA/PROYECTO-PROGRAMACION/blob/main/codigo-proyecto.py

https://colab.research.google.com/drive/1LRnCA09C5hETFc3hhptnJAVEmlcNmJmz?usp=sharing

Se utilizó Selenium para hacer el web scraping de una página dinámica, y se hizo la captura de la información hasta la cuarta página de los resultados, al ser la última página donde se encuentra información de Bochalema, Cali.

BIBLIOGRAFÍA:

-*Resultados de la búsqueda de WHOIS*. (2021). Godaddy.com. https://co.godaddy.com/whois/results.aspx?checkAvail=1&domain=fincaraiz.com.co

-*Bot Check*. (2022). Builtwith.com. https://builtwith.com/fincaraiz.com.co

-Pingdom Tools. (2022). *Pingdom Tools*. Pingdom.com. https://tools.pingdom.com/

-*Crear y enviar un archivo robots.txt | Centro de la Búsqueda de Google | Documentación | Google Developers*. (2022). Google Developers. https://developers.google.com/search/docs/advanced/robots/create-robots-txt?hl=es

-(2022). Fincaraiz.com.co. https://fincaraiz.com.co/robots.txt

-*Web Scraping interactivo con Selenium | Web Scraping PARTE 5*. (2020, 26 febrero). [Video]. YouTube. https://www.youtube.com/watch?v=fM_Os976HsQ