Piloto webscraping para pasajes aéreos en el IPC chileno

Elementos de su desarrollo y aprendizajes

6 de septiembre 2023

ine.gob.cl

Web Scraping para precios de pasajes aéreos

¿Qué Es el *Web Scraping*?

La técnica de *Web Scraping* se refiere al proceso de extracción de datos (texto, imágenes, audios, y en general cualquier información presentada en el ciberespacio) directamente desde un navegador web.

Su ejecución puede ser realizada a través de un usuario de manera manual o automatizada, lo cual es realizado por un robot a través de un software de programación. Sin embargo, lo más frecuente es que sea de manera automatizada.

Tiene variadas aplicaciones y es posible implementarlo en distintas plataformas digitales, y con distintos procesos, es decir, no hay un método único de realizar *Web Scraping*.

Un ejemplo de *Web Scraping* es el sitio web Google, el cual rutinariamente scrapea o «rastrea» internet para indexar sitios web.

¿Qué nos motivó a usar *Web Scraping* para pasajes aéreos?

- El producto pasaje aéreo ha presentado una alta volatilidad desde la apertura de los *lockdowns* implementados durante la pandemia COVID.
- Esta volatilidad generó cuestionamientos entre los usuarios sobre si era una condición de este mercado particular o un sesgo debido a nuestra metodología de recolección de precios.
- Lo anterior nos motivó a desarrollar un trabajo paralelo de recolección alternativa de precios que aumentara significativamente el número de observaciones mensuales.
- Hay un gran beneficio en términos de costo/eficiencia, pues se logra tener un mayor volumen de datos a un menor costo y con la posibilidad de capturar precios fuera de horas laborales.

Objetivo

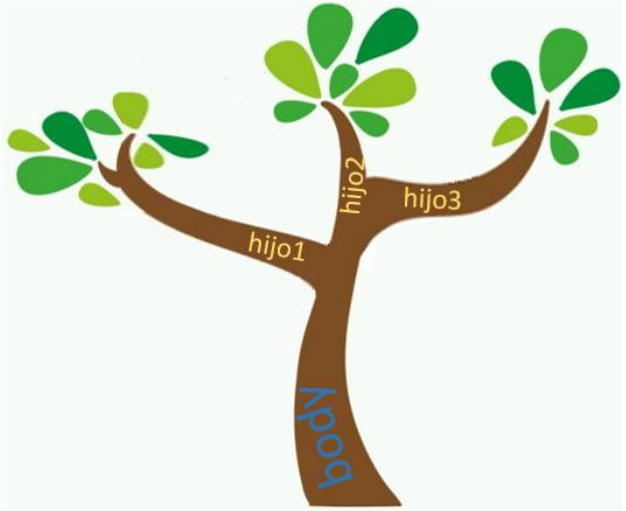
Realizar *Web Scraping* a través de la programación de un algoritmo en lenguaje automatizado en Python, con el fin de extraer un mayor número de observaciones de precios de vuelos aéreos internacionales y nacionales.

Puntos a presentar

- **Lenguaje**
- **Software**
- **Librerías**
- **Algoritmo**
- **Programación**
- **Resultados**
- **Problemáticas**
- **Soluciones**
- **Pasos a seguir**

Lenguaje HTML

HTML es usado recurrentemente para desarrollar páginas web y se basa en una estructura de árbol, donde los elementos están ordenados de manera estructurada y jerárquica.
El tronco principal es designado como body y cada rama corresponde a un hijo de este body llamado etiqueta html o tag html.



WEB SCRAPING

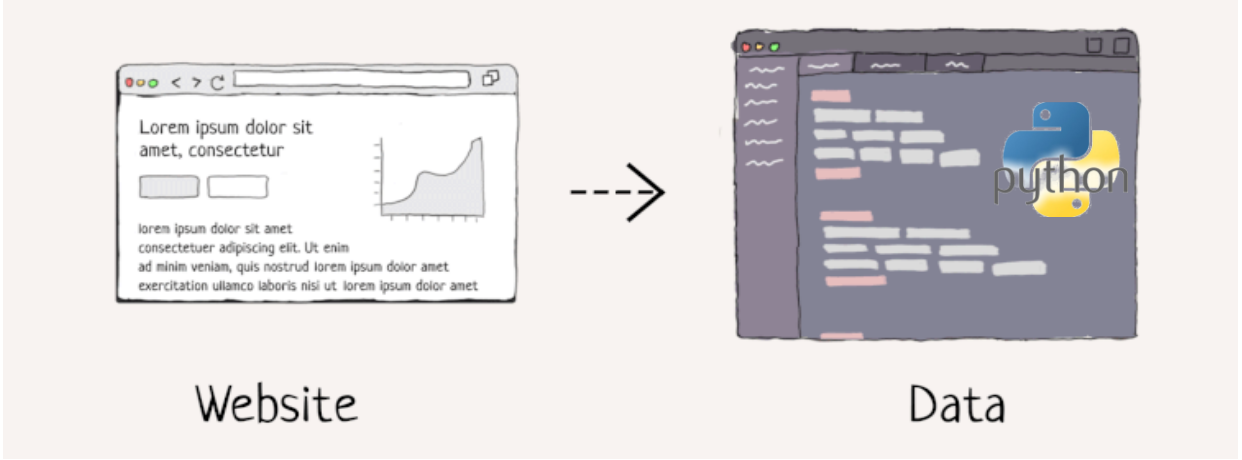
- **Lenguaje**
- **Software**
- **Librerías**
- **Algoritmo**
- **Programación**
- **Resultados**
- **Problemáticas**
- **Soluciones**
- **Pasos a Seguir**

Software Python y Librerías

Python es un lenguaje de programación independiente de plataforma, orientado a objetos y preparado para realizar cualquier tipo de programa, tales como aplicaciones, páginas web, análisis de datos, web scraping, entre otros. Además, contiene gran cantidad de librerías que incorporan funcionalidad en el lenguaje, de las cuales varias se destacan exclusivamente para la realización del Web Scraping. Sin embargo, para la ejecución de extracción de datos de vuelos aéreos se utilizó sólo Selenium.

Selenium

Esta librería tiene múltiples funciones en cuanto a la extracción de data en sitios de internet como, por ejemplo, abrir una página desde el compilador, acceder a datos, imágenes o archivos específicos, llenar formularios como loggin y password, hacer click, cambiar de pestaña, entre otras.



WEB SCRAPING

- **Lenguaje**
- **Software**
- **Librerías**
- **Algoritmo**
- **Programación**
- **Resultados**
- **Problemáticas**
- **Soluciones**
- **Pasos a Seguir**

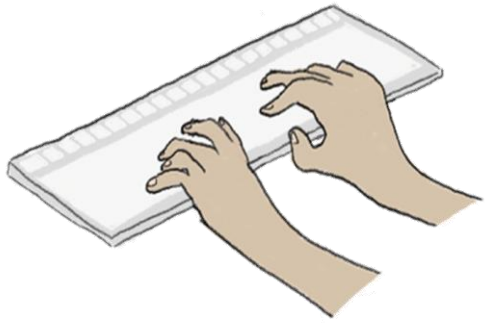
Algoritmo del Web Scraping y Programación para precios de pasajes aéreos

Algoritmo del Web Scraping

- **Abrir la página web a través del URL.**
- **Extraer automáticamente los datos estructurados a partir de los patrones.**
- **Resumir, almacenar, evaluar y/o combinar los datos extraídos.**

Programación del Web Scraping para precios de pasajes aéreos

- **El código se estructuró por separado, pero con la misma lógica para vuelos aéreos internacionales y nacionales, los cuales no requieren modificaciones, por ejemplo de fechas de consulta, y se ejecutan de forma automatizada diariamente.**
- **Para vuelos internacionales y nacionales se consultan diferentes fechas y destinos.**
- **Los scripts tomaron un tiempo de 2 meses aproximadamente su creación y correcta ejecución.**



Instituto Nacional de Estadísticas · Chile

WEB SCRAPING

- **Lenguaje**
- **Software**
- **Librerías**
- **Algoritmo**
- **Programación**
- **Resultados**
- **Problemáticas**
- **Soluciones**
- **Pasos a Seguir**

ine.gob.cl

Resultados del Web Scraping

- **Dependiendo de la fecha consultada y la frecuencia de recolección, se pueden obtener sobre 200 mil observaciones mensuales de precios para vuelos internacionales y nacionales.**
- **Los scripts se ejecutaban inicialmente durante la madrugada, demorando en promedio 2 horas para vuelos internacionales y 1 hora para vuelos nacionales.**
- **Los archivos resultantes se exportan en formato .csv, y se estructuran de la siguiente manera:**

Destino Lata	Destino	Fecha	Tipo	Aeropuerto	Horario	Escalas	Duracion	Clase	Premium	N de Vuelo	N de Clase	Precio	T Captura
SAO	Sao Paulo	09-07-2023	Ida	GRU	6:00	Directo	3 h 50 min	basic		0	0	404004	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	6:00	Directo	3 h 50 min	light		0	1	498324	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	6:00	Directo	3 h 50 min	plus		0	2	568278	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	6:00	Directo	3 h 50 min	top	Premium Ecc	0	3	764778	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	9:55	Directo	3 h 50 min	basic		1	0	275100	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	9:55	Directo	3 h 50 min	light		1	1	353700	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	9:55	Directo	3 h 50 min	plus		1	2	416580	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	9:55	Directo	3 h 50 min	top	Premium Ecc	1	3	591072	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	11:30	Directo	3 h 50 min	basic		2	0	404004	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	11:30	Directo	3 h 50 min	light		2	1	498324	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	11:30	Directo	3 h 50 min	plus		2	2	568278	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	11:30	Directo	3 h 50 min	top	Premium Ecc	2	3	764778	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	14:40	Directo	3 h 50 min	basic		3	0	404004	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	14:40	Directo	3 h 50 min	light		3	1	498324	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	14:40	Directo	3 h 50 min	plus		3	2	568278	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	14:40	Directo	3 h 50 min	top	Premium Ecc	3	3	764778	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	16:25	Directo	3 h 40 min	basic		4	0	326190	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	16:25	Directo	3 h 40 min	light		4	1	412650	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	16:25	Directo	3 h 40 min	plus		4	2	475530	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	16:25	Directo	3 h 40 min	plus	Premium Bu	4	3	491250	17-05-2023 21:50
SAO	Sao Paulo	09-07-2023	Ida	GRU	18:00	Directo	3 h 50 min	basic		5	0	539982	17-05-2023 21:50

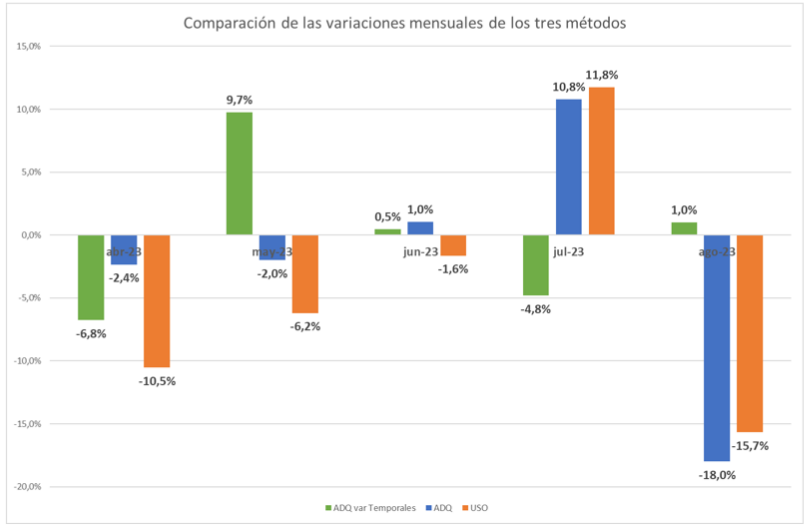
Resultados del Web Scraping

- **La densidad de registros de precios obtenidos nos ha permitido realizar ejercicios probando diferentes enfoques y opciones metodológicas, algo que no era posible con el método habitual de recolección.**
- **Estos ejercicios nos han permitido identificar diferentes fenómenos particulares a este mercado, lo que nos permitirá introducir mejoras en el cálculo del índice.**
- **Algo que cabe destacar es que la volatilidad observada en el índice oficial es propia de la dinámica de precios de pasajes aéreos.**

Ejercicio con enfoque de adquisiciones y más variedades temporales

ine.gob.cl

- Al comparar este método con el actual podemos apreciar que las variaciones mensuales disminuyen su dispersión al incorporar diferentes meses de partida del vuelo.



Desviación estándar enfoque de uso : 0,104
Desviación estándar método actual : 0,104
Desviación estándar mayor número de variedades : 0,064

WEB SCRAPING

- Lenguaje
- Software
- Librerías
- Algoritmo
- Programación
- Resultados
- Problemáticas
- Soluciones
- Pasos a Seguir

ine.gob.cl

Problemas que han surgido

- Los scripts pueden fallar por problemas de internet o conexión a la electricidad.
- Los scripts pueden fallar por problemas en la memoria RAM del computador.
- Los scripts pueden fallar por cambios inesperados en la estructura de la página web.
- Bloqueos al IP que está realizando la ejecución del Web Scraping.

Soluciones

- Se debe implementar continuamente la eliminación de las cookies de los sitios web, así como procesos de limpieza a los sistemas de almacenamiento de la información.
- Se deben revisar diariamente las salidas en .csv de los scripts, con el fin de confirmar que no haya habido cambios en la estructura de las páginas web consultadas.
- Para el bloqueo de la IP del computador, que es el problema más grave, se implementó una solución momnetánea que consiste en simular el Web Scraping como si fuese manual, es decir, se aumentan los tiempos de recolección y se simulan actualizaciones a las páginas consultadas. Esto significó que los scripts que anteriormente tomaban un promedio de 2 horas en levantarse, actualmente lo hacen en 18 horas aproximadamente, logrando que no bloqueen el IP del equipo. Sin embargo, lo más eficiente sería la utilización de API's de las propias empresas para realizar las consultas. Actualmente estamos trabajando en obtener estos permisos.

Pasos a seguir

- Ajustar los algoritmos de cálculo y probar diferentes opciones metodológicas.
- Continuar con el trabajo de relacionamiento con las empresas para lograr acceder a sus API’s.
- Comparar el comportamiento de la serie oficial con los resultados de la información proveniente de web scraping.
- Desarrollar protocolos operativos para cuando se decida su paso a la recolección mensual del IPC.

GRACIAS

ine.gob.cl