
Data Intelligence
INFORMACIÓN PARA EL DESARROLLO

Lo más buscado en Chile en 2019

Consulta las tendencias del 2019 - Chile

Tendencias 2019

1. Copa América
2. Javiera Suárez
3. Bono Marzo 2019
4. Toque de queda
5. Cameron Boyce

Quién es

1. Quién es el Vicepresidente de Chile 2019
2. Quién es Greta Thunberg
3. Quién es Pareman
4. Quién es Bessy Gallardo
5. Quién es Chimuelo

Qué es

1. Qué es una asamblea constituyente
2. Qué es la Constitución
3. Qué es TPP-11
4. Qué es Estado de Excepción
5. Qué es Estado de Emergencia

Acontecimientos

1. Copa América
2. Mundial Femenino
3. La caída del Muro de Berlín
4. Eclipse solar
5. Día del Profesor

Cómo

1. Cómo inscribir a las mascotas
2. Cómo ver el eclipse
3. Cómo saber si soy reservista
4. Cómo hacer lentes para eclipse
5. Cómo ver el eclipse sin lentes

Deportes

1. Copa América
2. Mundial Femenino
3. Inter de Milán
4. Christian Garín
5. Chile vs. Argentina

Fuente:
Tendencias Google
https://trends.google.com/trends/yis/2019/CL/

Principales Fuentes de datos oficiales en Chile

http://www.ide.cl/	IDE Chile	Infraestuctura de datos Geoespaciales	Genera iniciativas para promover que la información geográfica de carácter público este accesible a organismos de la administración del Estado, entidades privadas y ciudadanía en general
http://datosabiertos.chilecompra.cl/	Chile Compra	Analisis de Compras del Estado	Datos Abiertos es la plataforma de análisis de compras del Estado a través de la Dirección de Compra y Contratación Pública
https://www.ine.cl/	INE	Instituto Nacional de Estadisticas de Chile	Es encargada de las estadísticas y censos oficiales de la República, estas tareas la realiza a través de la elaboración y difusión de información confiable, oportuna, accesible, de relevancia y comparable a nivel nacional e internacional
https://www.bcentral.cl/	Banco Central	Banco Central de Chile	Permiten crear un entorno predecible para la toma de decisiones, contribuyendo a suavizar los ciclos económicos y sentando las bases para un crecimiento sostenido del país.
http://www.deis.cl/category/bloques_home/	DEIS	Departamento de Estadísticas e Información de Salud	información estadística pertinente, confiable y oportuna, dentro del marco definido por la Autoridad Sanitaria, participando en el diseño y en la implantación de mecanismos de control y evaluación,
http://www.cmfchile.cl/	CMF	Comisión para el Estado Financiero	Principal entidad supervisora de los mercados financieros en Chile, tras integrarse con la Superintendencia de Bancos e Instituciones Financieras

El proceso de Extraer Datos

Web scraping se podría definir como la técnica por la que un equipo de desarrolladores es capaz de rascar, *escrapear* o liberar datos de páginas web de gobiernos, instituciones públicas u organizaciones para acceder a datos privados o públicos que puedan ser publicados o distribuidos en formato abierto. El problema es que la mayoría de los datos de interés están en formatos no reutilizables y poco transparentes como un PDF, por ejemplo.

Para acceder y distribuir este tipo de información existe una gran cantidad de herramientas o procesos mediante el uso de lenguajes de programación. Esta es una guía de uso de los principales métodos de extracción de datos.

Técnicas de Extracción de Datos Web

- Google Spreadsheets
- Google Chrome
- Script Python
- Web Scrapping
- Web Harvesting
- Softwares especializados

Google Spreadsheets

¿Cómo Funciona?

Técnica que se realiza teniendo como destino una Hoja de Cálculo mediante el uso de fórmulas como ImportFeed, ImportHTML e ImportXML.
Se utiliza para extraer datos de tablas o listados de forma ordenada desde cualquier página web. Dependiendo de si es una tabla o una lista, el tipo de fórmula varía.

Ventajas y Desventajas

Se depende de la Suite Google y su capacidad de almacenamiento. Puede ser lenta en caso de encontrar archivos de más de 5Mb. Es recomendable en casos de tener la url del sitio pero no la url exacta de descarga, o tener tablas y datos en medio de la web, no como un descargable. Es muy fácil de implementar y automatizar, y su costo es cero.

Google Chrome

¿Cómo Funciona?

A través del navegador Chrome, se puede instalar una extensión denominada Table Capture, que proporciona a un usuario los datos de una web de forma sencilla y práctica. Obtiene la información contenida en una tabla en HTML de una página web a cualquier formato de tratamiento de datos como Google Spreadsheet, Excel o CSV. Es muy similar a la fórmula ImportHTML.

Ventajas y Desventajas

Se depende del Navegador Google Chrome y su capacidad de procesamiento. Es muy útil porque pueden descargar imágenes e incluso almacenar capturas de pantalla.

Script de Python

¿Cómo Funciona?

A través de un script elaborado en Python y la URL de una web, se puede hacer extracción de datos de una pagina web. Es muy útil cuando una pagina tiene varios ítems, como un catalogo de retail o una pagina de clasificados. Se puede redactar el código para que extraiga los datos de esa pagina y las siguientes que se incrementen a través de la URL de la web.

Ventajas y Desventajas

Es muy útil y fácil de implementar. Se debe programar una malla o un automatizador que los ejecute cada cierto tiempo porque este tipo de información cambia con mucha frecuencia. Una de sus funcionalidades más interesantes y profesionales es la extracción de datos no estructurados. En Python existen numerosas librerías para el acceso a datos.

Web Harvesting

¿Cómo Funciona?

Es el conjunto de todas las opciones que existen para extraer información de una página web, cuando generalmente existe una API. Puede recopilar informes XML, RSS o JSON.

Ventajas y Desventajas

Es uno de los métodos más eficaces. Su rapidez dependerá del servicio API que disponga la web, lo que a su ve nos condiciona a la existencia de ese código API.

Web Scrapping

¿Cómo Funciona?

Web Scraping forma parte de todo el conjunto Harvesting, pero utiliza métodos más específicos. Hace referencia principalmente al rastreo HTML a través de un servidor estático. Es muy útil cuando no hay una API disponible.

Ventajas y Desventajas

Normalmente se utilizan programas informáticos diseñados para extraer información de los sitios. La ventaja es que extrae datos del sitio web y con código HTML y de la base de datos. Existen negocios en la red cuya finalidad es recopilar este tipo de información, y hay una diversidad de software que podemos implementar para ello.

Software especializados

¿Cómo Funcionan?

Existen diferentes plataformas de software con licencias gratuitas inclusive que pueden facilitar enormemente las tareas de Scrapping de una web. Se instalan, se conectan a nuestros servidor de Datos o la fuente destino y se puede acompañar de un programa que orqueste la actualización y las consultas.

A continuación enumeramos algunas de ellas con sus ventajas y desventajas.

- **Tabula**

  Es una herramienta gratuita y fácil de usar. Su principal habilidad es la capacidad de extraer información ordenada y estructurada de archivos PDF que estén basados en texto (para los documentos escaneados se necesita otro programa). Soporta grandes volúmenes de información, y es compatible con Microsoft, Mac OS y Linux.

- **Web Scraper**

  Web Scraper es un software confiable para extraer datos de una página web. Actualmente está disponible para usuarios de Google Chrome y puede realizar una variedad de tareas de raspado de datos en pocos minutos. El raspador web puede extraer información de múltiples páginas al mismo tiempo y tiene capacidades de extracción de datos dinámicas incomparables. También puede manejar páginas con AJAX, cookies, redirecciones y Javascript.

Software automatizados

- **Spinn3r**

Spinn3r es adecuado para programadores, desarrolladores y startups. Puede extraer datos de un sitio web completo y se dirige principalmente a sitios de noticias, fuentes RSS, sitios de redes sociales y portales de viajes. Spinn3r usa API y gestiona hasta el 90% de los proyectos de rastreo y extracción de datos en Internet. Su sistema de rastreo web es similar a Google y Spinn3r guarda sus datos en formatos CSV y JSON. Esta herramienta escanea continuamente páginas web y obtiene los resultados deseados en cuestión de minutos.

- **Fminer**

Fminer es un raspador visual de datos que combina características de primera categoría. Con Fminer, puede realizar múltiples tareas de raspado web simultáneamente y así ahorrar tiempo y energía. También puede manejar sitios con AJAX y cookies. Fminer es perfecto para webmasters y startups y no les cuesta nada. Recoge datos de los medios de comunicación y garantiza la protección contra correo no deseado en Internet.

- **Octoparse**

Octoparse primero identifica sus datos, los raspa instantáneamente y guarda la información extraída en su disco duro. Navega a través de múltiples sitios y recopila contenido útil para usted. Octoparse es una buena opción para programadores y analistas de datos. Es mejor conocido por su tecnología de aprendizaje automático y exporta sus datos a formatos HTML, Excel, CSV y TXT.

Software automatizados

- **Dexi.io**

Dexi.io es uno de los mejores y más confiables software para raspar datos en Internet. No necesita descargar esta herramienta; de hecho, solo necesita abrir su sitio web y obtener sus datos al instante. Es una herramienta basada en navegador que viene con muchas capacidades y características únicas. Dexi.io exporta sus datos a archivos JSON y CSV o los guarda en Google Drive y Box.net.

- **Import.io**

Import.io ofrece un constructor para formar sus propios conjuntos de datos simplemente importando los datos de una página web en particular y exportando los datos a CSV. Puede scrapear fácilmente miles de páginas web en minutos sin escribir una sola línea de código.
Import.io utiliza tecnología de vanguardia para obtener millones de datos todos los días, que las empresas pueden aprovechar a cambio de pequeñas tarifas. Junto con la herramienta web, también ofrece aplicaciones gratuitas para Windows, Mac OS X y Linux para construir extractores de datos y rastreadores, descargar datos y sincronizarlos con la cuenta en línea.

Software automatizados

- **Scrapinghub**

Scrapinghub es una herramienta de extracción de datos basada en la nube que ayuda a miles de desarrolladores a obtener datos valiosos. Scrapinghub usa Crawlera, un rotador de proxy inteligente que admite omitir las contramedidas de los bots para rastrear fácilmente sitios enormes o protegidos contra bots.
Scrapinghub convierte toda la página web en contenido organizado . Su equipo de expertos está disponible para ayudarlo en caso de que su creador de rastreo no pueda satisfacer sus necesidades.

ParseHub

ParseHub está diseñado para rastrear sitios web únicos y múltiples con soporte para JavaScript, AJAX, sesiones, cookies y redirecciones. La aplicación utiliza la tecnología de aprendizaje automático para reconocer los documentos más complicados de la web y genera el archivo de salida en función del formato de datos requerido.

ParseHub, además de la aplicación web, también está disponible como una aplicación de escritorio gratuita para Windows, Mac OS X y Linux que ofrece un plan básico gratuito que cubre 5 proyectos de rastreo.