

# M2.851 Tipología y ciclo de vida de los datos: PRA1

2021-22-Sem.1

**Estudiante: Neil Cotie, trabajo individual**

**Correo: ncotie@uoc.edu**

## Contexto

He deseado crear un raspador web para recoger datos bursátiles, para conectar con un interés personal del tema. El concepto básico ha sido que el raspador sea utilizado regularmente, para recoger datos a lo largo del tiempo, cada día.

Investigando cuál mercado usar y cuál sitio web, he encontrado muchos que no permitían el raspado, o por sus condiciones de uso, o por lo estipulado en el robots.txt. Finalmente encontré el mercado suizo y el sitio oficial suyo, <https://www.six-group.com/en/products-services/the-swiss-stock-exchange.html>, donde se permitía el raspado.

El grupo SIX “*operates the infrastructure for the financial centers in Switzerland and Spain*” (SIX Group s.f.), entonces es el pozo oficial de datos de los índices bursátiles en cuyos países.

## Titulo

Aunque no es un dataset estático, sino destinado a actualizarse diariamente, estaría apropiado llamarlo “Swiss Equity Index Component Daily Data”, por ejemplo. En Zenodo lo he llamado “Stock Data Scraper Output Example”, como que, en principio, se puede extender la implementación actual a más índices y más países.

## Descripción del dataset

En la implementación actual, los datos recogidos y almacenados por el raspador vienen de la página del SIX para cada acción componente del índice especificado por el usuario en la línea del comando.

Un ejemplo estaría la pagina de ABB, <https://www.six-group.com/en/products-services/the-swiss-stock-exchange/market-data/shares/share-explorer/share-details.CH0012221716CHF4.html#/>

Recogemos los datos que se encuentren en los “tabs” *News & Data* y *Share Details*, las secciones *Key Data* y *Performance* del primero, y *Key Data* y *Profile* del segundo. Con más implementación, se podría recoger aun más de los datos disponibles, pero con esas secciones, se puede apreciar no solo medidas de rendimiento bursátil típicas como

- ultimo precio, precios mayor-menor del día y cuando han ocurrido, precio del cierre ayer
- volúmenes de acciones negociados en el día,
- cambio de precio en los últimos 52 semanas
- volume-weighted average price (VWAP)

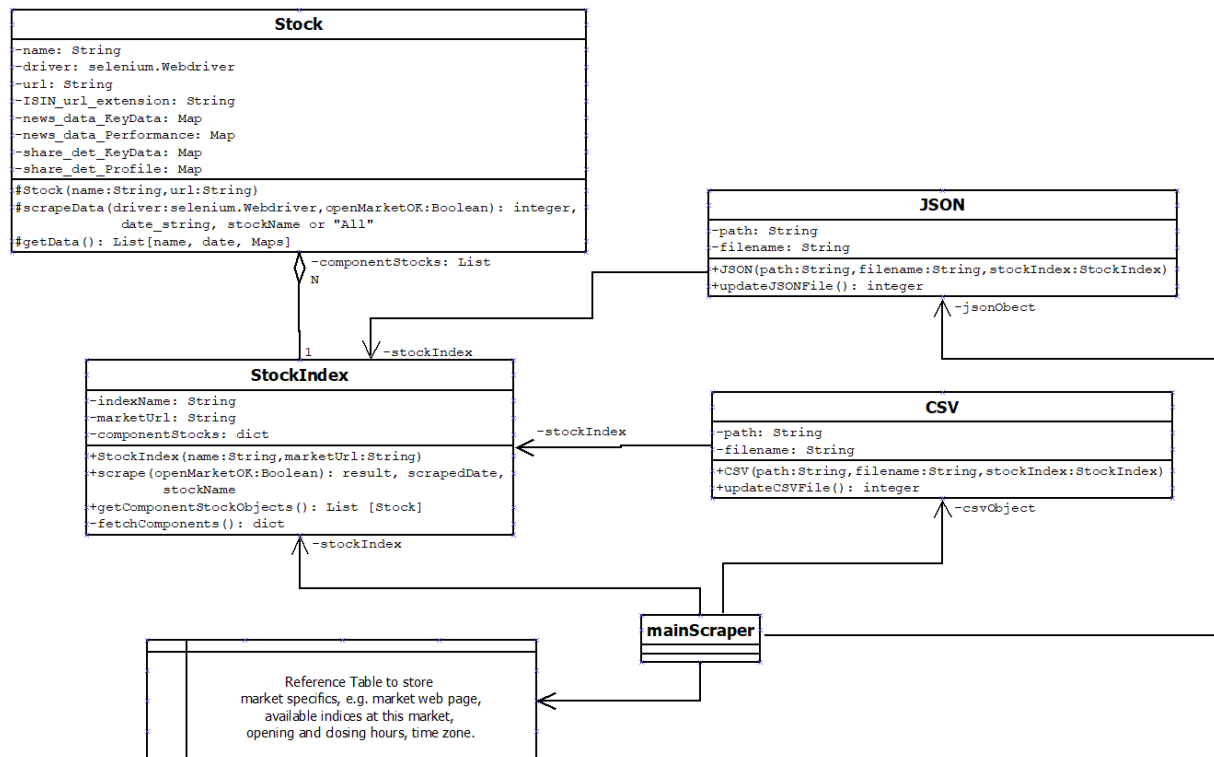
pero también datos más administrativos como:

- divisa de negociación
- número de acciones existentes
- el estándar regulatorio
- si dan un dividendo o no

Todos los datos recogidos por el raspador quedan conectados al nombre del dato, tal y como se aparecen en la página web, y el código añade la fecha de los datos y el nombre de la acción para cada una.

## Representación grafica

Incluyo como grafica el diagrama UML de la arquitectura del código, mostrando su orientación a objetos.



## Contenido

Como descrito arriba, el dataset no es estático, sino destinado a ser expandido día tras día con ejecución regular. La muestra cubre solo un par de días.

El raspador está codificado para recoger todos los campos de datos en las secciones descrito arriba, y, por tanto, si el grupo SIX añade o quita elementos de esas secciones, el cambio debería estar reflejado en los ficheros. No obstante, el código da un mensaje de error si se detecta que hay un cambio en los campos o en sus nombres, cada vez que se guarda en el fichero CSV, cuyo motivo es hacer que el usuario evalúe el impacto del cambio detectado.

La recogida de datos está hecha para generar, cada vez, unos diccionarios con parejas clave:valor. Eso es natural dada la forma de presentación de los datos en las páginas de SIX. Cuando se crea o actualiza el fichero CSV, no obstante, hay que “aplastar” esa estructura para qué acabe plana.

De momento, tenemos los siguientes campos:

- "Date": la fecha de los datos
- "Stock name": el nombre de la acción
- "Last trade / volume": el ultimo precio negociado y el volumen de acciones en esa transacción
- "Bid / ask price": los precios ofrecidos y pedidos al momento (o al cierre)
- "Bid / ask volume": las cantidades de acciones deseados al precio mas alto ofrecido, y al precio más bajo pedido
- "Last close": el precio del cierre anterior
- "Open": el precio de la primera transacción de la sesión
- "Volume": el volumen de acciones negociado hoy
- "Daily low / time": el precio más alto de la sesión y su hora
- "Daily high / time": el precio más bajo de la sesión y su hora
- "Spread absolute / in %": un valor usado en bonos
- "Opting up / Opting out": un atributo regulatorio
- "Management Transactions": una redirección a informaciones sobre transacciones hechas por gerentes del negocio
- "Significant Shareholders": una redirección a informaciones sobre los mayores dueños de acciones de esa empresa
- "52 week change": el cambio de precio en los últimos 52 semanas
- "52 week high / date": el precio más alto de los últimos 52 semanas y su fecha
- "52 week low / date": el precio más bajo de los últimos 52 semanas y su fecha
- "Year to date change": el cambio de precio desde el inicio del año
- "VWAP (60 days)": “Volume-weighted average price”, una medida de precio medio, ponderado por volumen a cada precio
- "Download": dos redirecciones a datos históricos de precios
- "Valor symbol": símbolo (“ticker”) usado en las transacciones
- "Valor number": número del símbolo (“ticker”)
- "ISIN": el número de identificación oficial de la acción
- "Trading currency": divisa de transacciones
- "Product type": tipo de producto/acción
- "Trading": fecha de inicio de uso
- "Security type": tipo de acción
- "Smallest tradable unit": unidad mínima de transacción
- "Security segment": segmento del mercado
- "Primary listed": si el mercado primario de acciones de esa empresa es el mercado suizo
- "Issued by": la empresa de cuál es la acción
- "Number in issue": cantidad existente de acciones
- "Nominal value": valor nominal de la acción
- "Dividend entitlement": si da dividende
- "Regulatory standard”: estándar regulatorio

Algunos datos cambiarían cada día (y más, cuando el mercado esté abierto) y otros no por mucho tiempo.

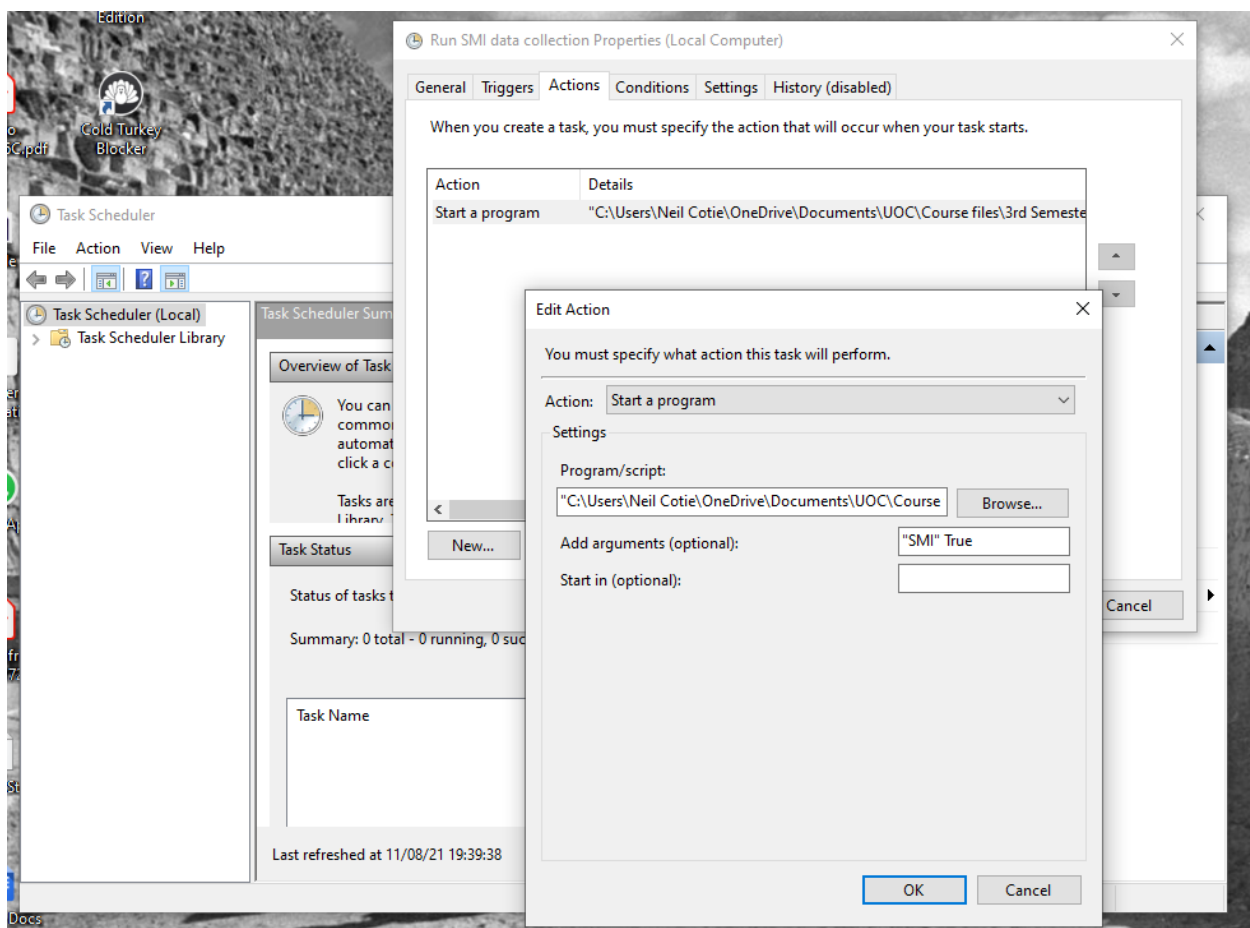
Los datos han sido recogidos usando detección de los “tags” HTML adecuados, tras inspeccionar la estructura de las páginas de SIX, implementado en rutinas de Python, del paquete *Selenium*.

El raspador debería estar gestionado en el *crontab* o *scheduled tasks* del sistema operativo en cuestión, por ejemplo, añadiendo:

```
23 59 * * 1-5 python3 {path apropiado}main.py "SMI" True
```

para ejecutar a 23:59 cada día de lunes a viernes.

o usando en Windows el *Task Scheduler*



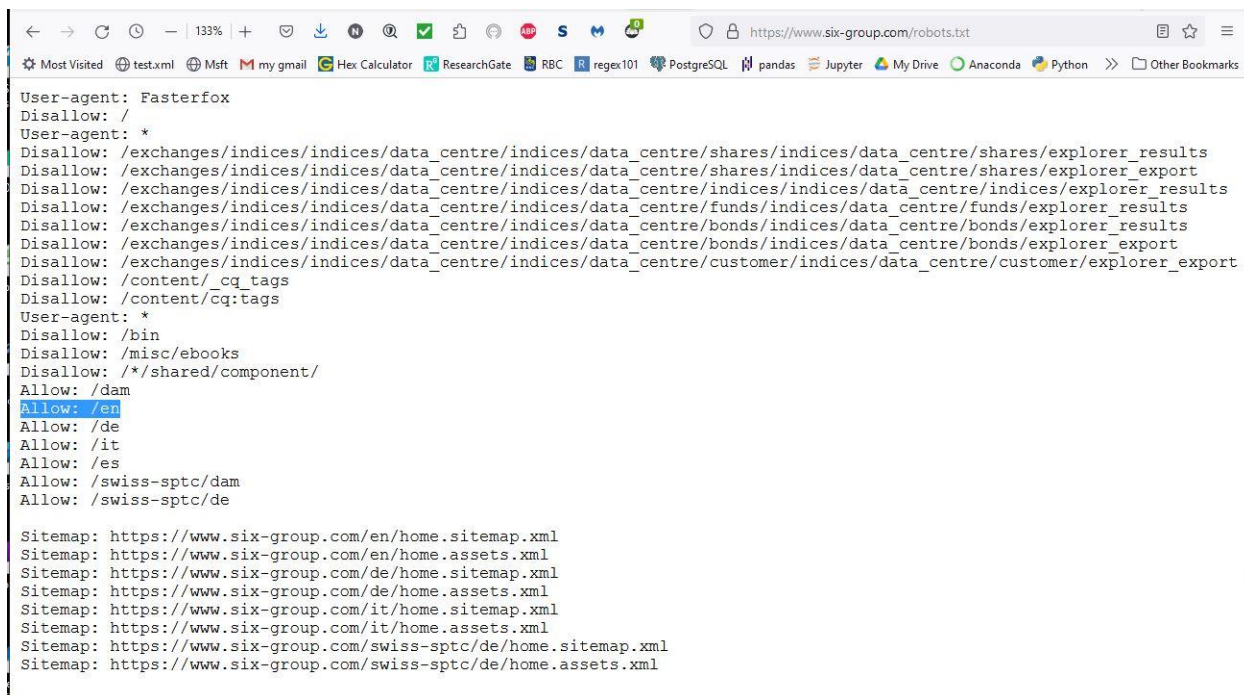
## Agradecimientos

Los datos son de origen del grupo SIX, <https://www.six-group.com/en/company.html>.

Dado a la naturaleza del raspador de datos web, no hay ningún proyecto o análisis anterior, que ha sido utilizado para crearlo, aunque seguramente existen muchos raspadores parecidos en uso privado por profesionales o semi-profesionales bursátiles.

Para estar seguro de la legalidad del raspado en cuestión:

- El fichero <https://www.six-group.com/robots.txt> incluye permisos para raspear las carpetas /en, de donde vienen todos esos datos.



```
User-agent: Fasterfox
Disallow: /
User-agent: *
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/shares/indices/data_centre/shares/explorer_results
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/shares/indices/data_centre/shares/explorer_export
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/indices/indices/data_centre/indices/explorer_results
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/funds/indices/data_centre/funds/explorer_results
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/bonds/indices/data_centre/bonds/explorer_results
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/bonds/indices/data_centre/bonds/explorer_export
Disallow: /exchanges/indices/indices/data_centre/indices/data_centre/customer/indices/data_centre/customer/explorer_export
Disallow: /content/_cq_tags
Disallow: /content/cq:tags
User-agent: *
Disallow: /bin
Disallow: /misc/ebooks
Disallow: /*/shared/component/
Allow: /dam
Allow: /en
Allow: /de
Allow: /it
Allow: /es
Allow: /swiss-sptc/dam
Allow: /swiss-sptc/de

Sitemap: https://www.six-group.com/en/home.sitemap.xml
Sitemap: https://www.six-group.com/en/home.assets.xml
Sitemap: https://www.six-group.com/de/home.sitemap.xml
Sitemap: https://www.six-group.com/de/home.assets.xml
Sitemap: https://www.six-group.com/it/home.sitemap.xml
Sitemap: https://www.six-group.com/it/home.assets.xml
Sitemap: https://www.six-group.com/swiss-sptc/de/home.sitemap.xml
Sitemap: https://www.six-group.com/swiss-sptc/de/home.assets.xml
```

- Los términos de uso del sitio web, <https://www.six-group.com/en/services/legal/terms-of-use.html>, no prohíben raspadores explícitamente, en contra de muchos sitios web americanos, por ejemplo. Solo prohíben “abusos” de las informaciones contenidos en el sitio web y/o sus sistemas TI. Para asegurar no infringir esas prohibiciones, he incluido muchas pausas entre interacciones con los servidores de SIX, para reducir al mínimo la carga necesaria para responder a peticiones del raspador. Incluso, cuando se cambia de “tab” para coger la segunda parte de los datos, en lugar de cargar la página de nuevo cambiando la dirección URL, como se podía hacer, el raspador navega y hace clic al botón adecuado, solamente iniciando cambios por JavaScript, evitando así carga a los servidores que produciría una carga completa.

## Inspiración

El uso de raspadores de ese tipo me hubiera ayudado en labores anteriores de trading, cuando la existencia y procuración de datos históricos con suficiente envergadura siempre ha sido problemática. En ese trabajo anterior, he luchado mantener datos históricos dentro de una plataforma propietaria, donde se podía fácilmente corromper los datos y/o perder trozos, entonces sé por experiencia personal que es un tema importante.

Además, datos históricos del tipo adecuado puede permitir búsquedas de patrones que puede ayudar a tomar decisiones de trading en el presente, así que, en ese contexto, “más siempre es mejor”.

Como descrito arriba, por el contexto del raspador, no hay conexiones concretas con análisis existentes de terceros.

## Licencia

Las opciones dadas en el enunciado son los siguientes. Haciendo referencia a Wikipedia (Wikipedia n.d.) y a choosealicense.com (Github n.d.):

- CC0: Public Domain License.
  - o Totalmente libre, considerado en dominio publico
- CC BY-NC-SA 4.0 License.
  - o « Attribution-NonCommercial-ShareAlike”
  - o Requiere atribución del autor, para usos no comerciales, y repartido más allá solo con licencia parecida y no más restrictiva.
  - o No incluido en las licencias ofrecido al crear un repositorio en GitHub.
- CC BY-SA 4.0 License
  - o Attribution-ShareAlike
  - o Lo mismo que BY-NC-SA, con la restricción de usos no comerciales quitado
  - o En (Github n.d.) dice que “Not recommended for software”.
- Database released under Open Database License, individual contents under Database Contents License.
  - o Eso es aplicable por bases de datos, cosa que no es el punto principal aquí, donde el código es el objeto destacado, y no un conjunto de datos ya producido.

Como que el proyecto aquí no es más que una utilidad para recopilar datos ya disponibles, y dada la baja complejidad, no parece apropiado elegir otro que Creative Commons Public Domain, en GitHub llamado “Creative Commons Zero v1.0 Universal”.

## Código

El repositorio GitHub se encuentra en el <https://github.com/ncotie/Share-data-scraper>.

Notas:

- BeautifulSoup no ha sido utilizable por su falta de manejo de JavaScript, entonces Selenium ha sido utilizado en lugar.
- La ambición inicial era de producir ambos tipos de ficheros de datos, CSV y JSON, pero limitaciones de tiempo han hecho imposible terminar la codificación de la parte JSON. En principio debería haber estado más sencillo que la parte CSV. La estructura JSON se adaptaría perfectamente a la forma de los datos recogidos, en diccionarios, entonces el trabajo incremental para terminarlo no debería estar mucho.
- Hay extensiones potenciales no explotados en el código, por ejemplos:
  - o Crear un fichero de configuración externo donde un usuario puede anotar mercados bursátiles, URL para datos, índices en esos mercados, etc. Eso facilitaría el uso del raspador, para evitar cualquier cambio de códigos solo para temas de configuración y/o URL.
  - o Con los índices disponibles en SIX, son de momento restringido a índices conteniendo solo el número de componentes que los diseñadores del web SIX han incluido en una

página de listado. Eso crea un tope, y por ejemplo no podemos raspear un índice de 100 componentes allí, por no haber desarrollado todavía el mecanismo para seguir a siguientes páginas del listado.

- He elegido usar `print()` para dar mensajes al usuario, pero otro mecanismo podría ser también el *logging*.
- Se podría extender los datos de configuración a incluir verificación de un horario de apertura del mercado y gestionar el cambio de zona horario, en lugar del mecanismo más sencillo actual.
- La fecha de los datos es tal cual como puesto en las páginas SIX. Un preprocesado de datos ante un análisis posterior podría cambiar el formato como necesario.
- El manejo de los campos de datos, para inyectarles en el formato CSV, no toma en cuenta el posible cambio de orden de esos campos en las páginas SIX en futuros modificaciones, y tampoco cambios pequeños en los nombres de los campos.
- Se podría haber extendido el rango de datos recogidos para tomar en cuenta más secciones de las páginas.

## Dataset

Una muestra de dataset formato CSV se encuentra en el <https://doi.org/10.5281/zenodo.5654517>.

Nota: Dado que ha sido un trabajo individual, no han sido incluidos cualquier firma/ iniciales como fuera necesario por un trabajo en pareja.

## Referencias

Github. n.d. "Choose an open source license." *https://choosealicense.com*. <https://choosealicense.com>.

Marzagao, Thiago. 2013. "webscraping with Selenium." *thiagomarzagao.com*. 11 12. Accessed 2021.  
<http://thiagomarzagao.com/2013/11/12/webscraping-with-selenium-part-1/>.

Muthukadan, Baiju. n.d. "Selenium with Python." <https://selenium-python.readthedocs.io/index.html>.

Selenium project. n.d. "Selenium." *selenium.dev*. <https://www.selenium.dev/documentation/overview/>.

SIX Group. n.d. "Overview of SIX." *six-group.com*. <https://www.six-group.com/en/company.html>.

Wikipedia. n.d. "Creative Commons license." *Wikipedia.org*.  
[https://en.wikipedia.org/wiki/Creative\\_Commons\\_license](https://en.wikipedia.org/wiki/Creative_Commons_license).

—. n.d. "List of stock exchanges." *Wikipedia*. [https://en.wikipedia.org/wiki/List\\_of\\_stock\\_exchanges](https://en.wikipedia.org/wiki/List_of_stock_exchanges).