PRÁCTICA 1

Pablo Rivas Castellanos

Diego Alberto López Herrera

Tipología y ciclo de vida de los datos

Máster universitario de Ciencia de Datos

12/Mar/2021

Índice

1. Contexto

Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

Se ha escogido el sitio web https://www.bolsamadrid.es/ por aportar gran cantidad de información detallada sobre valores cotizados en la bolsa de Madrid. Este sitio web no establece ninguna limitación en cuanto a la descarga automática de datos mediante *web scraping* o técnicas similares (se ha realizado esta revisión a 12 de marzo de 2021).

Se pretende, por un lado, obtener información general sobre las empresas cotizadas: nombre, sector, mercado, índices. Por otra parte, se recopilarán datos detallados sobre la cotización a nivel diario: valores de cierre, referencia, último, máximo, mínimo y medio, así como el volumen y efectivo en cuanto a los títulos negociados.

Como posible mejora futura, se podrá establecer una actualización de los datos con la frecuencia conveniente para mantener un listado de empresas actualizado y un histórico completo de cotización diaria.

2. Título del dataset

Definir un título para el dataset. Elegir un título que sea descriptivo.

Bolsa de Madrid: empresas y detalles de cotización diaria

3. Descripción del dataset

Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Se dispone del listado de empresas cotizadas de la bolsa de Madrid, aportando detalles generales sobre ellas (sector, índice al que pertenecen, etc.) y el detalle de cotización diaria con varias métricas relevantes (valor de cierre, volumen, valor medio, etc.). El conjunto de datos se compone de dos tablas en las que el elemento común es el ISIN (https://www.cnmv.es/portal/ANCV/CodigoISIN.aspx), un identificador unívoco de los valores mobiliarios a nivel internacional, a través del cual se pueden relacionar los datos de ambas tablas.

En la siguiente tabla se incluye todo el detalle de los campos del dataset:

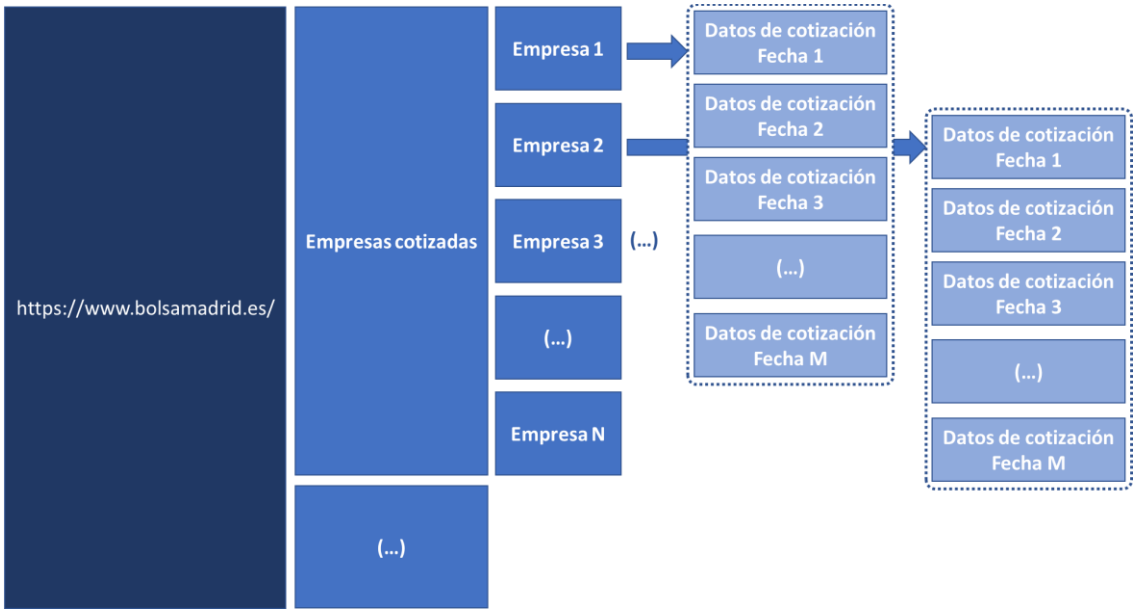
Tabla	Campo	Descripción	Tipo	Ejemplo
empresas	isin	ISIN. Identificador unívoco de los valores mobiliarios a nivel internacional.	Texto	*ES0125220311*
empresas	nombre	Nombre de la empresa	Texto	*ACCIONA, S.A.*

empresas	sector_subsector	Sector y subsector al que pertenece la empresa	Texto	*Mat.Basicos, Industria y Construcción - Construcción*
empresas	mercado	Mercado en el que cotiza la empresa	Texto	*Mercado Continuo*
empresas	indices	Índice bursátil al que pertenece la compañía	Texto	*IBEX 35®*
cotizaciones_diarias	isin	ISIN. Identificador unívoco de los valores mobiliarios a nivel internacional.	Texto	*ES0125220311*
cotizaciones_diarias	fecha	Fecha de la cotización	Fecha	*02/03/2021*
cotizaciones_diarias	valor_cierre	Valor de cierre de cotización	Numérico (decimal)	*133,1000*
cotizaciones_diarias	valor_referencia	Valor de referencia de cotización (cierre del día previo)	Numérico (decimal)	*135,0000*
cotizaciones_diarias	volumen_titulos_negociados	Volumen de títulos negociados	Numérico (entero)	*71.918*
cotizaciones_diarias	efectivo_titulos_negociados	Volumen de efectivo en títulos negociados	Numérico (decimal)	*9.616.442,00*
cotizaciones_diarias	valor_ultimo	Valor último de cotización	Numérico (decimal)	*133,1000*
cotizaciones_diarias	valor_maximo	Valor máximo de cotización	Numérico (decimal)	*135,3000*
cotizaciones_diarias	valor_minimo	Valor mínimo de cotización	Numérico (decimal)	*132,7000*
cotizaciones_diarias	valor_medio	Valor medio de cotización	Numérico (decimal)	*133,7140*

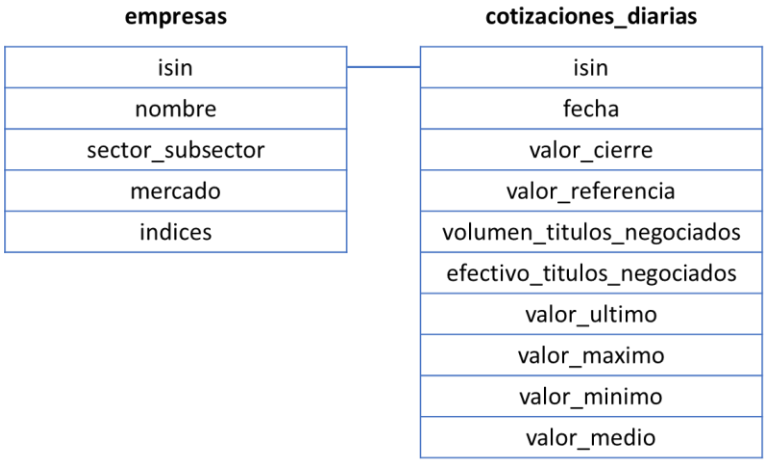
4. Representación gráfica

Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.

La información se recoge tal y como se muestra en el siguiente esquema:



En cuanto al dataset, en la siguiente figura se detallan las tablas y campos que los conforman, junto con las relaciones correspondientes:



5. Contenido

Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Se tienen los siguientes sitios webs de obtención de los datos para cada subconjunto de datos del *dataset*:

- Información general sobre empresas: https://www.bolsamadrid.es/esp/aspx/Empresas/Empresas.aspx . Requiere navegación vía *JavaScript* para visualizar todas las compañías en base a la letra inicial de su nombre (se agrupan alfabéticamente).

- Datos sobre cotizaciones a nivel diario: https://www.bolsamadrid.es/esp/aspx/Empresas/InfHistorica.aspx?ISIN=[ISIN] (donde *[ISIN]* es el ISIN de la empresa cotizada). La información que proporciona por defecto es la de los últimos 30 días, pero se puede incluir información sobre fechas de inicio y fin en los formularios para obtener datos con vigencia temporal de un máximo de un año.

Se presentan a continuación una serie de consideraciones relevantes sobre la recogida de datos y buenas prácticas de *web scraping* y diseño general de la solución que se han incluido en el desarrollo:

a) El programa es capaz de navegar por cualquiera de las páginas de las que se desea extraer datos, incluyendo:
   - Cumplimentación y envío de formularios.
   - Navegación por el contenido de tablas basadas en *JavaScript*, interactuando con botones y recogiendo los resultados (botón "siguiente").

b) Se evita colapsar al servidor a través de los siguientes mecanismos:
   - En cada uno de los hilos, las peticiones se envían secuencialmente, de forma que no se envíen peticiones adicionales hasta obtener las correspondientes respuestas (o se sobrepase el *timeout* establecido).
   - En caso de que las peticiones no generen respuesta, se ha diseñado un sistema de esperas exponenciales entre el reenvío de peticiones para permitir la potencial recuperación del servidor.
   - Se especifican un número máximo de intentos de acceso al recurso. Si al finalizar el último intento no se obtiene el recurso se notifica al usuario.

c) Enfoque basado en *threads*. Se diseña una subclase de Thread para extraer los datos con las siguientes funcionalidades:
   - Se crea un grupo de tareas encargadas de extraer la información.
   - El hilo principal añade las *URLs* a procesar a una cola.
   - Los hilos toman uno a uno elementos de la cola de entrada, los procesan y envían los resultados a una cola de salida.
   - El hilo principal obtiene los resultados de la cola y los guarda en ficheros, informando adicionalmente de los errores identificados (en caso de que los hubiese) y de los resultados.

6. Agradecimientos

Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

En primer lugar, se agradece a Bolsa de Madrid la puesta a disposición al público general de información de máxima calidad y actualizada diariamente. El objeto de este proyecto ha cubierto una mínima parte de todos los datos que podrían recopilarse de forma periódica para obtener un conocimiento profundo sobre las empresas cotizadas de la bolsa de Madrid, ya que adicionalmente se dispone en el sitio web de información relativa a dividendos, ampliaciones de capital, ofertas públicas de venta y suscripción y *splits*, entre muchos otros aspectos. Los autores esperamos que este trabajo sea de utilidad pública para cualquier empresa o particular que lo use bajo la licencia establecida y que pueda servir de punto de partida para la creación de proyectos de alcances mucho más amplios.

Por otra parte, nos gustaría agradecer también los contenidos de los siguientes sitios web que han servido de referencia e inspiración para el diseño e implementación de la solución:

1. Sitio web: https://developer.ibm.com/articles/au-threadingpython/ . En el apartado "Listing 4. Multiple queues data mining websites" hay un ejemplo de múltiples *worker threads* empleando dos colas con un objeto que hereda de *Thead*. En nuestro diseño se optó por una solución de corte similar para paralelizar el proceso, haciendo que la clase que extrae la información herede de *Thread*, reciba los trabajos de una cola y devuelva la tabla extraída a través de una segunda cola. No obstante, hay diferencias sustanciales de funcionamiento, como el envío de señales de salida a las tareas trabajadora, o el procesado por parte de *main* (escritura de resultados en fichero) de cada elemento de la cola inmediatamente después de su recepción (no se espera a que se completen todos los trabajos).
2. Sitio web: https://www.youtube.com/watch?v=Xjv1sY630Uc&list=PLzMcBGfZo4-n40rB1XaJ0ak1bemvlqumQ . Video-curso de *Selenium*, donde se explican sus funcionalidades más destacadas de manera clara y detallada. Nos ha resultado muy ilustrativo para el diseño de la solución y la implementación de algunas de las funcionalidades.

7. Inspiración

Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El conjunto de datos que extrae la solución ofrece un detalle muy completo sobre las empresas que cotizan en la bolsa de Madrid.

A continuación, se presentan algunas de las cuestiones que se pueden responder a través del análisis de este conjunto de datos:

- ¿Qué empresas cotizan en la bolsa de Madrid? ¿A qué mercados pertenecen (mercado continuo, parqué, Latibex, etc.)?
- ¿Qué sectores y subsectores presentan un mayor número de empresas en cada índice (IBEX 35, IBEX TOP Dividend, etc.)? ¿Qué peso relativo tienen las compañías de “Servicios de Consumo - Ocio, Turismo y Hostelería”? ¿Qué sectores y subsectores están presentando un mayor auge? ¿Y un mayor retroceso?
- ¿Qué evolución están presentando los valores de cotización de las acciones en cada sector? ¿Qué sector está presentando un mejor comportamiento? ¿Cuál está presentando el peor rendimiento?
- ¿Qué grado de volatilidad están presentando los valores de cotización en los últimos meses? ¿Hay algún índice / mercado / sector que presente un grado de volatilidad mayor? ¿Y alguno que presente una mayor estabilidad?
- ¿Cuál es el top 5 de acciones por volumen negociado en los últimos 10 días?

El conjunto de datos podría ser de utilidad para multitud de cometidos. Se exponen a continuación una serie de escenarios de uso en los que podría ser de utilidad:

- **Empresas de inversión e inversores particulares**: puede ser de utilidad en la toma de decisiones sobre inversión en ciertas compañías en base a su histórico, el rendimiento general del sector y

la consistencia del valor en el tiempo. También permite a los inversores intradía identificar oportunidades de inversión en valores que estén presentando una gran volatilidad en las fechas más recientes.

- **Medios de comunicación y empresas de estudio de mercado**: gracias a la categorización de las empresas por sector, índice, mercado, etc., podría permitir la elaboración estudios y artículos muy segmentados. También pueden identificarse patrones que puedan anticipar movimientos de fusión o adquisición relevantes, por ejemplo, si se identifican en un mismo rango temporal un aumento significativo en el volumen de títulos negociados y el valor de cotización en dos empresas específicas.
- **Gobierno y administraciones públicas**: puede servir de apoyo a la toma de decisiones sobre inversión o ayudas a compañías / sectores que estén en declive, anticipando problemas financieros futuros (ya que el comportamiento del mercado de valores anticipa con gran antelación potenciales situaciones de las compañías). Por otra parte, permite comprender mejor el interés de inversión que provocan ciertos sectores y compañías que cotizan en la bolsa de Madrid en cada momento. Adicionalmente, se pueden identificar potenciales escenarios de alta especulación (por ejemplo, identificando un alto volumen de títulos negociados) que permitan a las autoridades tomar medidas con la antelación suficiente.

8. Licencia

Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- Released Under CC0: Public Domain License
- Released Under CC BY-NC-SA 4.0 License
- Released Under CC BY-SA 4.0 License
- Database released under Open Database License, individual contents under Database Contents License
- Other (specified above)
- Unknown License

Se ha elegido la licencia “CC BY-NC-SA 4.0 License” para que se requiera citación de los autores, solo se permita el uso no comercial y todas las adaptaciones se distribuyan bajo los mismos términos. De esta forma, se pretende que quien use el material no persiga beneficios económicos y, en caso de que se enriquezca / adapte el material, pueda accesible por cualquier persona / compañía con las mismas condiciones que el material compartido.

9. Enlace github

El proyecto completo se puede encontrar en el siguiente enlace de *github*:

https://github.com/dlopezherr/bolsaMadridScraper/