Máster universitario en Ciencia de datos
(Data science)

Práctica 1: Web Scraping

M2.851 - Tipología y ciclo de vida de los datos

Autor:
Yosry Elsayed
Laura Pastor

Profesor/a:
Mireia Calvo González

Índice

1. Introducción

El mercado financiero está en cambio constante, por lo que resulta complicado intentar analizarlo a tiempo real extrayendo el dato mediante la descarga de ficheros tipo Excel de manera continua. Por ello, es interesante extraer los diferentes valores del mercado mediante técnicas de web scraping para así conseguir crear un histórico a tiempo real que pueda alimentar un modelo predicativo cuyo objetivo es optimizar la cantidad de dinero a invertir y cuando, y visualizar dichos valores.

Los gráficos de la propia web no tienen cierta flexibilidad u opciones para modificar sus ejes y parámetros (como podemos observar en la siguiente imagen), algo que sí que podemos manipular y diseñar según nuestras necesidades en Power BI.

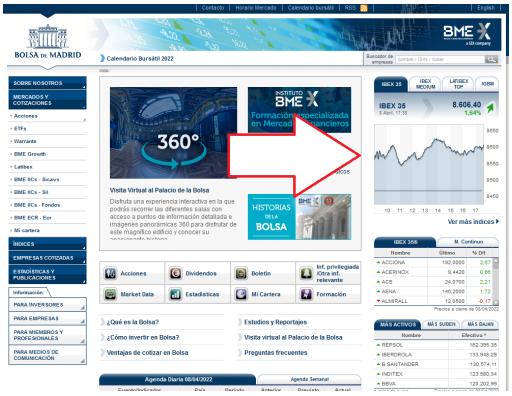

Figura 1. Gráfico ejemplo de la página web de Bolsa de Madrid

Resumidamente, en esta práctica solo nos centraremos en una extracción puntual del histórico de los diferentes valores financieros de las empresas que componen el índice bursátil IBEX35 desde la página web oficial de la Bolsa de Madrid, y presentarlos en un report de visualizaciones usando Power BI.

La Bolsa de Madrid es el principal mercado de valores de España, cuyo propietario es el BME que es el operador de todos los mercados de valores y sistemas financieros en España. El presidente actual del BME es Johannes Bernardus Dijsselhof.

El titular y prestador de la página web en cuestión se presenta como "Sociedad Rectora de la Bolsa de Valores de Madrid, S.A. Sociedad Unipersonal", como se puede ver en el apartado de "Aviso Legal".

En el mismo apartado de "Aviso Legal" de la página web nos encontramos con el subapartado "Uso de la Página Web" donde se señala que los contenidos de la página pueden ser descargados, copiados e impresos para uso personal y no comercial. Por lo que es legal la extracción y análisis de los diferentes valores de la página usando técnicas de Web Scraping e usarlos con fines estudiantiles.

Por el otro lado, El IBEX 35 es el principal índice bursátil de la bolsa española elaborado por BME. Está formado por un total de 35 empresas las cual extraeremos un histórico sus valores de precio de cierre, e información descriptiva propia de cada empresa (por ejemplo, dirección, teléfono, logo, etc.).

2. Algoritmo

Presentaremos como funciona el algoritmo, creado en Python, y su navegación en la página web para extraer los datos correspondientes a cada miembro que forma parte del IBEX35.

1. Entra en la página web `https://www.bolsamadrid.es` y extrae todos sus datos en formato HTML.

Figura 2. Interfaz de la página web de Bolsa de Madrid

Figura 3. Información en formato HTML de la página de Bolsa de Madrid

2. Extrae toda la información que contiene la tabla señalada, mediante el id de la tabla que se encuentra en el HTML.

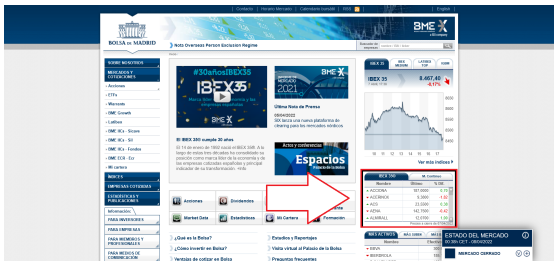

Figura 4. La tabla de resumen del índice bursátil del IBEX35 que nos interesa para llegar al detalle de cada empresa que lo forma

3. Datasets

La información extraída se dividirá en dos tablas (ficheros CSV que tienen como separador la comma ','), y una carpeta que contendrá el logo de cada empresa (formato gif y jpg).

3.1. DIM_Miembro.CSV

Este fichero contiene información descriptiva de los diferentes miembros/empresas que forman parte del IBEX35. Es una tabla de 35 observaciones con los siguientes campos:

- **miembro:** El nombre de la empresa.
- **dirección:** La dirección de la central de la empresa.
- **postal:** El código postal junto a la ciudad donde se encuentra la central.
- **teléfono:** Teléfono de contacto.
- **logo:** La carpeta y el nombre del logo de cada empresa.

Si pensamos en la estructura de un modelo de datos, esta tabla se considera como una dimensión dado que contiene datos descriptivos de cada empresa y no incluye ninguna medida numérica a analizar.

3.2. FACT_IBEX35.CSV

Este fichero contiene un histórico los valores financieros diarios de cada miembros/empresas que forman parte del IBEX35 por día. Es una tabla de 805 observaciones con los siguientes campos:

- **Id:** Identificador o clave primaria de cada observación.
- **fecha:** La fecha del registro/observación.

- **cierre:** El precio de cierra de cada empresa para cada día.
- **Referencia:** Código identificativo de un valor en el sistema de representación de Bolsa de Madrid.
- **volumen:** El número de títulos o contratos que se negocian de un activo financiero en un mercado.
- **Efectivo:** El efectivo negociado de una operación es el resultado de multiplicar el número de acciones negociado por el precio de cada acción.
- **logo:** La carpeta y el nombre del logo de cada empresa.
- **último:** Último precio de acción, que coincide con el precio de cierre.
- **máximo:** Máximo precio alcanzado.
- **mínimo:** Mínimo precio alcanzado.
- **medio:** Precio medio de acción a lo largo del día.
- **activo:** Nombre de la empresa/miembro del IBEX35.

Si pensamos en la estructura de un modelo de datos, esta tabla se considera como una tabla de hechos dado que contiene las medidas numéricas que nos interesa analizar y no incluye información descriptiva en exceso.

Por lo tanto, nuestro modelo de datos se construirá por una tabla de hechos y una dimensión.

3.3. Carpeta: miembros

Se trata de una carpeta que incluye iconos o logos de cada empresa cuyos formatos varían entre gif y jpg. Se puede saber el nombre del icono correspondiente a cada empresa mediante la columna 'logo' en el fichero 'DIM_Miembro.CSV'.

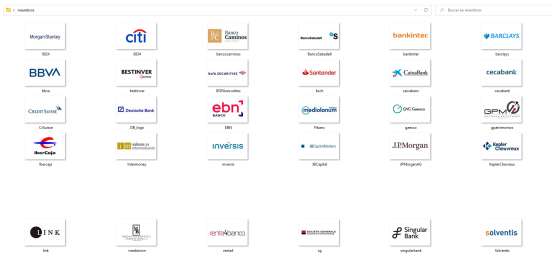

Figura 5. Los logos de los miembros del IBEX35

Nuestros datasets tendría una licencia ODbL. Una licencia Open Database License permite a los usuarios compartir, modificar y usar libremente las bases

de datos en cuestión, sin temor a los derechos de autor o cuestiones de propiedad del dato. Este tipo de licencia encaja con nuestros datasets dado que se tratan de datos públicos informativos sobre los cual no tenemos ninguna propiedad, y se pueden usar libremente por cualquier persona ya que nuestro repositorio será público.

Los ficheros CSV se encuentran tanto en el repositorio GitHub (el repositorio es Yoyazoooo20/DBT-Trainning) como en Zenodo (el DOI de Zenodo es 10.5281/zenodo.6428376).

4. Visualización

Para visualizar y analizar los datos se usará la herramienta Power BI. Se trata de un servicio de análisis de datos de Microsoft orientado a proporcionar visualizaciones interactivas, la cual se puede conectar a cualquier base de datos, página web o API. En nuestro caso Power BI se conectará directamente con los archivos CSV que guardaremos en un directorio GitHub público, así mismo cualquier persona que descargue el fichero de visualizaciones '.pbix' puede interactuar con los datos sin la necesidad de tener los ficheros CSV.

Una vez conectados ambos ficheros CSV a Power BI, construimos el siguiente modelo de datos:

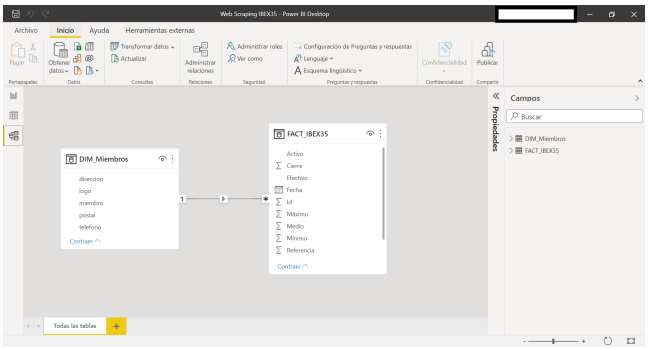

Figura 6. Modelo de los datos

Se tiene una relación 1:N entre la tabla dimensión y la tabla de hechos. Es decir, cada registro de la tabla de hechos está relacionada únicamente con un registro de la tabla de dimensión, pero cada registro de la tabla de dimensión puede estar relacionada con más de un registro de la tabla de hechos.

Las visualización que nos encontramos en el informe son:

1. Un gráfico clásico de velas del índice bursátil del IBEX35 (se usan las distintas métricas de cierre, medio, mínimo y máximo de precio por activo).

2. Una tabla de información descriptiva de cada miembro del IBEX35.

3. un gráficos de rectángulos de los miembros del IBEX35, donde el tamaño de cada rectángulo/empresa va relacionado con su precio de cierre. Es decir, cuanto más grande es el rectángulo de una empresa, mayor es su precio de cierre a comparación del resto de empresas.

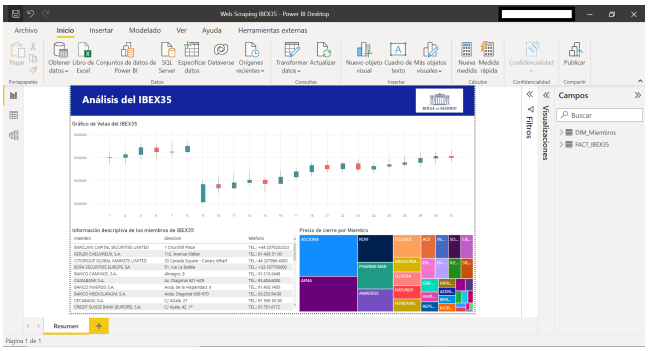

Figura 7. Report con las distintas visualizaciones descritas anteriormente

Como se puede observar, a partir de los datos extraídos se pueden personalizar visualizaciones interactivas dependiendo de nuestras necesidades. Otra ventaja con Power BI es que pueden publicar los reports de manera que sean fáciles de visualizar, analizar y compartir en distintos dispositivos (tablet, móvil, etc). A partir de estas visualizaciones podemos ver la tendencia del IBEX35, por lo que si se observa una tendencia alcista podemos comprar acciones suponiendo que esa tendencia se mantendrá. También podemos ver las empresas con mayor precio de cierre de acciones dentro del IBEX35 en todo momento (se puede aplicar filtros por día, mes y/o año). Por último, también se pueden consultar información propia de cada empresa.

5. Aspectos claves a mejorar

Una de las mejoras que se podrían aplicar en este trabajo es aplicar el proceso de Web Scraping directamente en la interfaz de Power BI, así evitamos la necesidad de extraer el dato en un archivo CSV, y guardarlo en un repositorio GitHub para que posteriormente sea ingestado dentro de visualizaciones. Power BI actualmente tiene la posibilidad de incorporar código de Python o R, pero todavía esta en fase de mejora por lo que todavía no es óptimo ni fácil de usar.

6. Contribución al trabajo

Una tabla donde se muestra que cada uno de los autores/integrantes han participado en cada uno de los apartados del trabajo presentado.

Contribuciones	Firma
Investigación previa	Laura y Yosry
Redacción de las respuestas	Laura y Yosry
Desarrollo del código	Laura y Yosry