Universitat Oberta de Catalunya

uoc.edu

PRA 1 Web Scrapping

Asignatura: Tipología y ciclo de vida de los datos

Autores: Paula Muñoz Lago
Abel Romero Búrdalo

Profesora: Mireia Calvo González

Fecha: 09/11/2020

ÍNDICE

1. Contexto

El desarrollo de esta práctica se encuentra en el marco de la asignatura "Tipología y Ciclo de Vida de los Datos" del Master Universitario de Ciencia de Datos de la Universitat Oberta de Catalunya.

El problema elegido para la realización de la práctica ha sido la evolución de las empresas de automoción en bolsa a lo largo de este último año, pudiendo estudiar cómo la pandemia del COVID-19 ha afectado a dicho sector. Además, ofrecemos la posibilidad de adaptar el análisis a las necesidades del usuario pudiendo elegir las fechas de inicio y fin que se quieren observar.

Para obtener el resultado objetivo se ha utilizado la página web YahooFinance, que permite buscar el nombre de cada empresa según su Ticker (siglas asociadas a las empresas que cotizan en bolsa para una rápida identificación). De cada empresa podemos obtener una gran cantidad de información, aunque nos hemos centrado en la pestaña 'historical data' para obtener el valor de cierre de la acción a lo largo de los días. Además, utilizando la api yfinance obtenemos datos sobre dividendos y stock splits (división del valor de una acción) que no podíamos obtener mediante el web scraping. Es por ello que se han integrado ambas técnicas.

2. Título dataset

El titulo escogido para el juego de datos extraído ha sido AccionesSector-Automocion.xlsx. En cuanto al titulo escogido para el proyecto realizado ha sido: Evolución de las acciones para las empresas con mayor cotización del sector de la automoción durante la pandemia de COVID-19.

3. Descripción del dataset

Los datos han sido recolectados para cuatro compañías del sector de la automoción:

- Tesla
- Porsche
- Nio
- Ferrari

Dichas compañías pertenecen al sector de la automoción, donde las cuatro cotizan en bolsa de valores de Nueva York. Esto quiere decir que se pueden comprar sus acciones, las cuales tienen un valor que fluctúa en función de

distintos parámetros como las ganancias de las empresas, la oferta y demanda de las acciones, etc.

El juego de datos extraído refleja los datos relativos a los valores de dichas acciones para las cuatro compañías mencionadas anteriormente, donde se reflejan los siguientes datos:

- Valor de la acción en la apertura del mercado
- Valor máximo adquirido a lo largo del día
- Valor mínimo adquirido a lo largo del día
- Valor de la acción en el cierre de mercado
- Valor estimado de la acción en el cierre de mercado
- Volumen de acciones compradas
- Dividendos recibidos por los accionistas
- Stock splits ese día

Dichos datos se han extraído para un intervalo de tiempo de un año donde para cada día se reflejan los valores de las acciones descritos anteriormente. Asimismo, cada resultado se ha almacenado en una hoja de Excel indicando el nombre de la compañía.

Está información permite comprender cual ha sido la evolución de las cuatro compañías en el marco de la pandemia, donde al tratarse del sector de la automoción permitiría extrapolar su comportamiento a otras compañías. Asimismo, se puede realizar una comparativa entre ellas y sacar conclusiones como si los mayores inversores han perdido confianza en estas compañías o si ha existido una mayor o menor demanda de las acciones, donde se reflejaría en si el valor accionarial ha incrementado o disminuido.

Para la extracción de los datos se ha tenido que examinar el fichero “robots.txt”, con la finalidad de saber que accesos se permiten a robots y que información se permite extraer de YahooFinance. En este caso el sitio web permite el acceso de todos los robots y se excluyen los siguientes directorios

- /m/
- /r/
- /__rapidworker-1.2.js
- /__blank
- /_td_api
- /_remote

Se adjunta el fichero “robot.txt” en el directorio “/Robots” del repositorio generado para el proyecto.

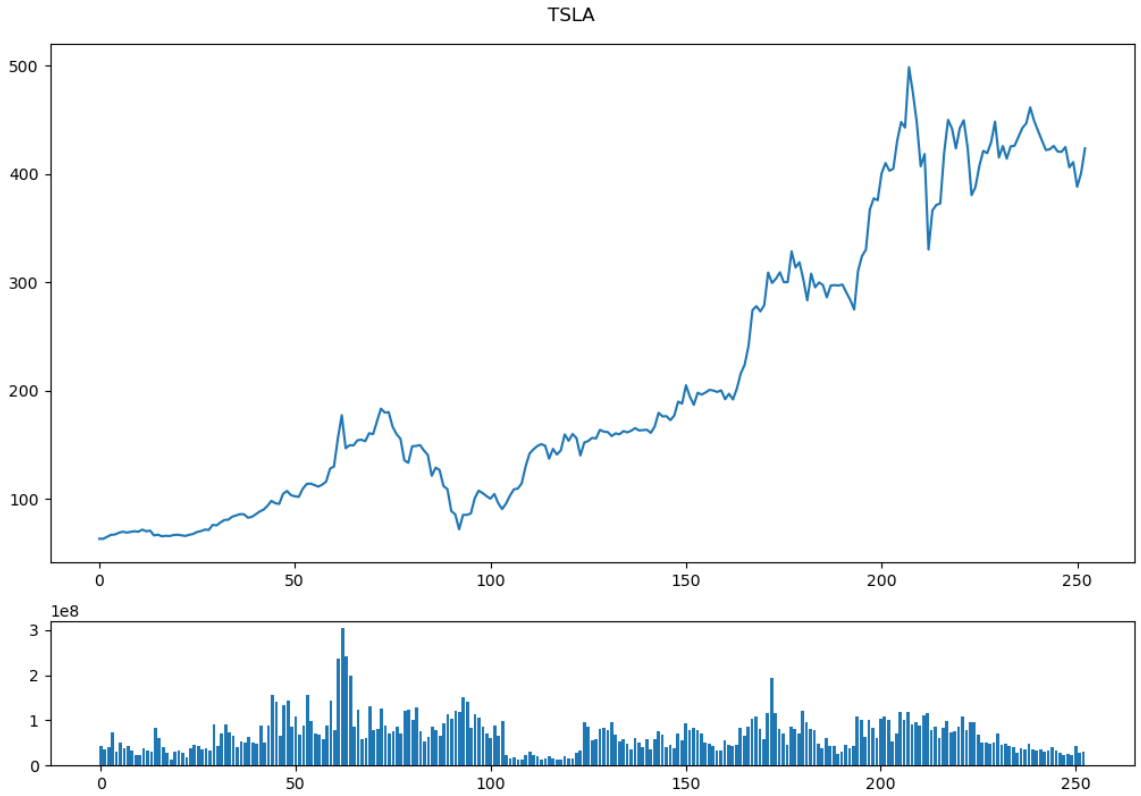
4. Representación gráfica

El Dataset generado se identifica de la siguiente forma:

Fecha	Abrir	Máx.	Mín.	Cierre*	Cierre ajus.**	Volumen	Dividends	Stock Splits
04/11/201	62,96	64,39	61,85	63,49	63,49	43935000	0	0
05/11/201	63,92	64,7	63,22	63,44	63,44	34717000	0	0
06/11/201	63,6	65,34	62,9	65,32	65,32	39704500	0	0
07/11/201	65,83	68,3	65,6	67,11	67,11	72336500	0	0
08/11/201	66,9	67,49	66,5	67,43	67,43	30346000	0	0
11/11/201	68,79	69,84	68,4	69,02	69,02	49933500	0	0
12/11/201	69,38	70,07	68,81	69,99	69,99	36797000	0	0
13/11/201	71	71,27	69,04	69,22	69,22	42100500	0	0
14/11/201	69,22	70,77	68,58	69,87	69,87	32324500	0	0
15/11/201	70,13	70,56	69,67	70,43	70,43	24045000	0	0
18/11/201	70,58	70,63	69,22	70	70	22002000	0	0
19/11/201	70,35	72	69,56	71,9	71,9	38624000	0	0
20/11/201	72	72,24	69,91	70,44	70,44	33625500	0	0
21/11/201	70,9	72,17	70,8	70,97	70,97	30550000	0	0
22/11/201	68,03	68,2	66	66,61	66,61	84353000	0	0
25/11/201	68,86	68,91	66,89	67,27	67,27	61697500	0	0
26/11/201	67,05	67,1	65,42	65,78	65,78	39737000	0	0
27/11/201	66,22	66,79	65,71	66,26	66,26	27778000	0	0
29/11/201	66,22	66,25	65,5	65,99	65,99	12328000	0	0
02/12/201	65,88	67,28	65,74	66,97	66,97	30372500	0	0
03/12/201	66,52	67,58	66,44	67,24	67,24	32868500	0	0
04/12/201	67,55	67,57	66,57	66,61	66,61	27665000	0	0
05/12/201	66,57	66,88	65,45	66,07	66,07	18623000	0	0
06/12/201	67	67,77	66,95	67,18	67,18	38062000	0	0
09/12/201	67,32	68,89	67,02	67,91	67,91	45115500	0	0
10/12/201	67,99	70,15	67,86	69,77	69,77	44141500	0	0
11/12/201	70,38	71,44	70,22	70,54	70,54	34489000	0	0
12/12/201	70,98	72,55	70,65	71,94	71,94	38819500	0	0
13/12/201	72,21	73,04	70,93	71,68	71,68	32854500	0	0
16/12/201	72,51	76,72	72,5	76,3	76,3	90871000	0	0
17/12/201	75,8	77,1	75,18	75,8	75,8	42484000	0	0
18/12/201	76,13	79,04	76,12	78,63	78,63	70605000	0	0
19/12/201	79,46	81,37	79,3	80,81	80,81	90535500	0	0
20/12/201	82,06	82,6	80,04	81,12	81,12	73763500	0	0
23/12/201	82,36	84,4	82	83,84	83,84	66598000	0	0
24/12/201	83,67	85,09	82,54	85,05	85,05	40273500	0	0
26/12/201	85,58	86,7	85,27	86,19	86,19	53169500	0	0
27/12/201	87	87,06	85,22	86,08	86,08	49728500	0	0
30/12/201	85,76	85,8	81,85	82,94	82,94	62932000	0	0
31/12/201	81	84,26	80,42	83,67	83,67	51428500	0	0
02/01/202	84,9	86,14	84,34	86,05	86,05	47660500	0	0

TSLA | PAH3.DE | NIO | RACE

Además, junto al archivo Excel, también se genera una visualización de los datos más representativos de cada empresa estudiada. Estos son, el valor de cierre por cada día y el volumen. El siguiente sería el ejemplo de TESLA.



En la parte superior del gráfico se observa la evolución de la cotización de la acción, esto es, el precio de cierre. Mientras que en la parte inferior se observa la evolución del volumen de operaciones realizadas por los accionistas sobre esta empresa.

5. Contenido

El conjunto de datos referentes al valor accionarial de las compañías Tesla, Ferrari, Nio y Porsche desde el 8 de noviembre de 2019 hasta el 8 de noviembre de 2020 extraídos en el fichero AccionesSectorAutomocion.xlsx son:

- **Fecha**: fecha en formato dd/MM/yyyy
- **Abrir**: valor de la acción en la apertura del mercado expresado en dólares estadounidenses (USD)
- **Máx**: valor máximo de la acción a lo largo del día expresado en USD
- **Cierre*:** valor de la acción en el cierre del mercado expresado en USD
- **Cierre ajus.***: valor de la acción estimado en el cierre del mercado expresado en USD.
- **Volumen:** la cantidad de un activo concreto en el que se invierte durante un día.
- **Dividends:** dinero recibido por los accionistas en forma de dividendos ese día.
- **Stock Splits:** Si se ha realizado una operación de división de acciones o no ese día.

6. Agradecimientos

La extracción de datos sobre el valor de las acciones de las empresas seleccionadas que cotizan en bolsa ha sido gracias al sitio web “Yahoo Finance”, cuyos datos están alojados en: https://es.finance.yahoo.com/ y son de acceso público y gratuito.

7. Inspiración

La pandemia en la que nos vemos inmersos ha afectado a numerosos campos, desde el más claro, la salud, hasta el económico, haciendo que muchas personas pierdan sus trabajos o cierren sus negocios. Esta situación también ha afectado a la bolsa de valores, dado que, en marzo, cuando se declaró el estado de alarma, las bolsas de todo el mundo cayeron en picado haciendo que millones de personas perdiesen una gran proporción de sus ahorros. Esto ha generado que muchos inversionistas viesen esta situación como un momento idóneo para comprar acciones, lo que ha producido que se esté generando una ‘burbuja bursátil’. Esto es, empresas como Amazon, Facebook o Tesla, estudiada en esta práctica, están en su cotización máxima histórica.

Es por la gravedad de este asunto y las consecuencias que pueden traer consigo que dicha ‘burbuja’ se ‘explote’, que hemos visto interesante obtener los datos de varias empresas de un sector concreto, el automovilístico, para comprobar qué supuso la pandemia a cada empresa y cómo se han recuperado (o no) de las caídas de marzo.

8. Licencia

La licencia escogida ha sido ***Released Under CCO: Public Domain License*** donde se renuncian a todos los derechos de la obra bajo las leyes de derechos autorales en todo el mundo. El juego de datos creado se puede copiar, distribuir, modificar e interpretar incluso para propósitos comerciales sin pedir permiso.

Se ha escogido la licencia explicada anteriormente debido a que los datos son públicos y se han obtenido de manera gratuita de YahooFinance. Asimismo, la finalidad de este proyecto es puramente académica y sin ánimo de lucro, por lo que no se pretende obtener ningún tipo de beneficio del trabajo realizado. Por último, se ha considerado que está licencia es adecuada para que terceras personas puedan utilizar el juego de datos con distintos fines como podrían ser académicos, introducción al mundo de las inversiones o estudiar la evolución accionarial del sector de la automoción durante la pandemia de COVID-19.

9. **DOI**

Para consultar el dataset y citarlo mediante DOI se puede acceder al siguiente enlace:
https://zenodo.org/record/4263399#.X6hpSWhKg2w

10. Código fuente

El código fuente creado para generar este dataset se encuentra publicado en el repositorio de GitHub https://github.com/paulamlago/Financial-Web-Scrapping/tree/main/Code

11. Contribuciones

Contribuciones	**Firma**
Investigación previa	ARB, PML
Redacción de las respuestas	ARB, PML
Desarrollo código	ARB, PML

REFERENCIAS

[1] Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC

[2] Masip, D. (2010). El lenguaje Python. Editorial UOC.

[3] Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

[4] Yahoo, Yahoo Finance, Noviembre 2020.https://finance.yahoo.com/

[5] PyPi, , Octubre 2020 https://pypi.org/project/beautifulsoup4/

[6] B. Muthukadan, "Selenium with Python", Noviembre 2020. https://selenium-python.readthedocs.io/

[7] M. Breuss, " Beautiful Soup: Build a Web Scraper With Python", Real Python, Diciembre 2019. https://realpython.com/beautiful-soup-web-scraper-python/

[8] C. OKeefe, "Modern Web Automation With Python and Selenium", Real Python, Noviembre 2020. https://realpython.com/modern-web-automation-with-python-and-selenium/