Práctica 1: Web Scraping

Acciones más activas de España y su histórico desde enero 2019

Realizado por: María Dolores Higuera González y Andrea González Vicario.

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. (0.25 puntos)**

La página elegida para poder realizar esta práctica, ha sido Investing.com, ya que nos proporciona datos en tiempo real sobre la variación de los mercados financieros. En nuestro caso, hemos visto interesante conocer las acciones de España que están más activas a día de hoy, así como el histórico de cada una de estas empresas, desde enero de 2019, hasta noviembre de 2020.

2. **Definir un título para el dataset. Elegir un título que sea descriptivo. (0.25 puntos)**

Acciones más activas de España y su histórico desde enero 2019

3. **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). (0.25 puntos)**

Para poder extraer los datos necesarios de la web de Investing, para conocer las acciones más activas de nuestro país en tiempo real, hemos utilizado la librería Beautiful Soup, la cual nos ha permitido extraer la información necesaria en formato HTML, utilizando el parser de lxml.

Hemos navegado a través de la web, hasta que hemos conseguido nuestro objetivo, conseguir el nombre de las empresas que tienen más acciones activas a día de hoy en España. Según vayan variando estos datos en la tabla de origen, también variarán en nuestro csv resultante.

Además, hemos creado un histórico de datos, donde tenemos los valores de:

Fecha	Último valor	Valor apertura	Máx. valor	Min. valor	Volumen	% Variación

De cada una de las 50 empresas que se encuentran en la lista de acciones de España más activas en tiempo real proporcionada por Investing.com, desde enero de 2019 hasta la fecha actual.

4. **Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente. (0.25 puntos)**

En nuestro caso, hemos obtenido dos tipos de CSVs:

- El primero guarda las 50 acciones más activas de España en el día actual, obtenemos el Nombre, último valor, valor máximo, mínimo, variación, el % de variación, el volumen y la hora.

```
        # Creamos las entreadas del diccionario y lo convertimos en un dataframe
        dict_data['Nombre'] = list_name
        dict_data['Último valor'] = list_last
        dict_data['Máx. valor'] = list_high
        dict_data['Min. valor'] = list_low
        dict_data['Variación'] = list_var
        dict_data['% Variación'] = list_por_var
        dict_data['Volumen'] = list_vol
        dict_data['Hora'] = list_time

        df = pd.DataFrame(dict_data, columns = ['Nombre', 'Último valor', 'Máx. valor', 'Min. valor', 'Variación', '% Variació
n', 'Volumen', 'Hora'])

        # Descargamos un archivo .csv
        final_datafile = df.to_csv('WebScraping_SP_stocks.csv', sep=',')
```

Mostramos un ejemplo de los datos que encontramos en el archivo csv obtenido:

	Nombre	Último valor	Máx. valor	Min. valor	Variación	% Variación	Volumen	Hora
0	Iberdrola	11,127	11,152	11,030	+0,152	+1,38%	2,17M	11:47:52
1	Santander	1,831	1,835	1,790	+0,044	+2,46%	11,11M	11:49:25
2	BBVA	2,582	2,582	2,514	+0,096	+3,86%	6,59M	11:49:09
3	Telefónica	2,905	2,905	2,820	+0,111	+3,97%	5,08M	11:49:20

- Además, obtenemos 1 csv por cada acción, donde encontramos el histórico de los datos de cada una de estas acciones. En este caso, los datos que encontramos son: la fecha, el último valor, el valor de apertura, el máximo, el mínimo, el volumen y el % de variación.

```
                # Creamos las entreadas del diccionario y lo convertimos en un dataframe
                dict_data_hist["Fecha"] = list_date_hist
                dict_data_hist["Último valor"] = list_price_hist
                dict_data_hist["Valor apertura"] = list_open_hist
                dict_data_hist["Máx. valor"] = list_high_hist
                dict_data_hist["Min. valor"] = list_low_hist
                dict_data_hist["Volumen"] = list_vol_hist
                dict_data_hist["% Variación"] = list_var_hist

                df_hist = pd.DataFrame(dict_data_hist, columns = ['Fecha', 'Último valor', 'Valor apertura', 'Máx. valor', 'Mi
n. valor', 'Volumen', '% Variación'])

                # Descargamos un archivo .csv
                final_datafile_hist = df_hist.to_csv('WebScraping_SP_stock_'+ web_name + '.csv', sep=',')
```

Por ejemplo, en la siguiente imagen podemos encontrar el histórico del grupo Acciona SA:

	Fecha	Último valor	Valor apertura	Máx. valor	Min. valor	Volumen	% Variación
0	Nov 2020	92,375	86,925	95,600	85,750	161,18K	6,55%
1	Oct 2020	86,700	93,500	98,775	85,150	137,17K	-6,57%
2	Sep 2020	92,800	100,200	105,000	91,200	105,47K	-7,39%

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. (1 punto)

Como comentábamos anteriormente, tenemos 2 tipos de archivos csv. El primero de ellos, es en el que encontramos los datos que nos proporciona la página Investing.com a día de hoy, los cuales se van actualizando cada día. Por este motivo, el periodo de los datos obtenido es el día actual, ya que es la propia web la que nos los va actualizando. Los campos que incluye este primer dataset son:

- *Nombre:* de la empresa
- *Último valor:* consiste en el precio de cierre, es decir, el ultimo valor antes del cierre de la bolsa
- *Valor máximo:* es el valor máximo alcanzado durante la sesión diaria
- *Valor mínimo:* es el valor mínimo alcanzado durante la sesión diaria
- *La variación:* consiste en la diferencia entre el valor de apertura y de cierre
- *El % de variación*
- *El volumen:* el cambio de valor del total de acciones de la empresa con respecto al cierre del día anterior
- *La hora*: momento en el que se ha registrado la actualización de la bolsa

Por otro lado, como también se ha comentado anteriormente, hemos creado un csv con el histórico de cada una de estas acciones. Estos datos los hemos conseguido utilizando la librería Selenium, junto con el controlador de Chrome (chromedriver) y seleccionamos los datos mensuales que queremos obtener.

```
# Elegimos el desplgable y la opción que queremos mostrar
select = Select(driver.find_element_by_id('data_interval'))
select.select_by_visible_text('Mensual')
```

Los campos que incluye este dataset, son muy similares a los del csv general, variando los siguientes campos:

- *Fecha:* Encontramos la fecha en formato Mes Año de los datos obtenidos.
- *Valor de apertura:* la media de los valores de apertura de todos los días del mes
- *Max. Valor*: obtenido en los últimos 30 días
- *Min. Valor:* obtenido en los últimos 30 días

En este caso, el periodo de tiempo de los datos es desde enero de 2019, hasta noviembre 2020.

	Fecha	Último valor	Valor apertura	Máx. valor	Min. valor	Volumen	% Variación
0	Nov 2020	92,375	86,925	95,600	85,750	161,18K	6,55%
1	Oct 2020	86,700	93,500	98,775	85,150	137,17K	-6,57%
2	Sep 2020	92,800	100,200	105,000	91,200	105,47K	-7,39%

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay). (1 punto)

En primer lugar, queríamos agradecer a *Investing.com* porque gracias a esta plataforma, hemos podido obtener todos los datos en tiempo real que necesitábamos para poder realizar esta práctica. Esta plataforma fue creada en 2007, y actualmente cuenta con 250 empleados que se encargan del mantenimiento de la página, y gracias a los cuales hemos podido obtener los datos necesarios.

Por otro lado, queríamos dar las gracias también a Daniel Romero Pérez, ya que nos ha guiado durante el proceso de la práctica, dándonos las directrices necesarias para poder cumplir con éxito nuestros objetivos.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. (1 punto)

Esta página nos ha parecido muy interesante porque podemos obtener una gran variedad de datos, que se van actualizando día, tras día, en tiempo real, por lo que cada día obtendremos nuevos valores, y cada mes vamos guardando los valores más significativos en un historial.

Con nuestro código, hemos conseguido responder a las siguientes preguntas:

- *¿Cuáles son las acciones más activas en España?*
- *¿Cuáles es el historial de dichas acciones?*
- *¿Cuál fue el valor máximo del Banco Santander en diciembre de 2019?*

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: (1 punto)**
    - 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

Después de realizar un estudio sobre las diferentes licencias disponibles para nuestro Dataset, creemos que la que mejor se adapta a nuestro programa es Released Under CC BY-NC-SA 4.0 License, ya que, al subirlo al Github, queremos que pueda servir de ayuda para futuros estudiantes que quieran aprender a realizar Web Scraping sobre cualquier Web utilizando BeautifulSoup.

Esta licencia, nos da la posibilidad de poder redistribuir nuestro código, así como readaptarlo por futuros usuarios. Éste se puede utilizar para fines lúdicos, pero en ningún caso para un fin comercial. Además, otro aspecto importante es que, siempre que se comparta, deberá realizarse bajo la misma licencia.

Recursos utilizados para escoger la licencia que mejor se adapta a nuestro Proyecto:

https://creativecommons.org/licenses/?lang=es_ES

https://creativecommons.org/choose/non-web-popup?license_code=by-nc-sa&jurisdiction=&version=4.0&lang=en

https://creativecommons.org/licenses/by-nc-sa/4.0/

https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

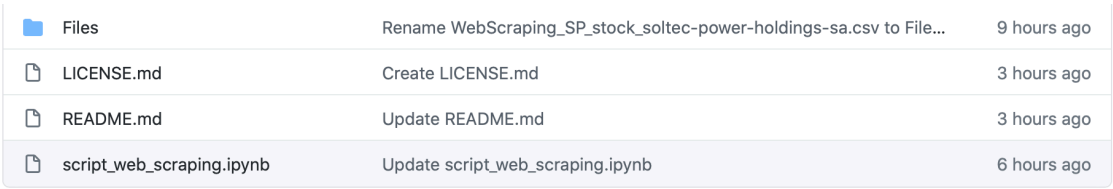
9. **Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. (2.5 puntos)**

Puede encontrar nuestro código en Python en el siguiente enlace de Github:

https://github.com/andreagonvic/Web-Scraping

En nuestro repositorio de GitHub podemos encontrar los siguientes documentos:

- FILES -> en esta carpeta tenemos todos los archivos csv generados
- LICENSE.md -> nos indica la licencia seleccionada
- README.md -> Donde encontraremos las directrices que debemos de seguir para poder ejecutar nuestro script final
- Script_web_scraping.ipynb -> Es el código que nos genera los distinos csv



10. **Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción. (2.5 puntos)**

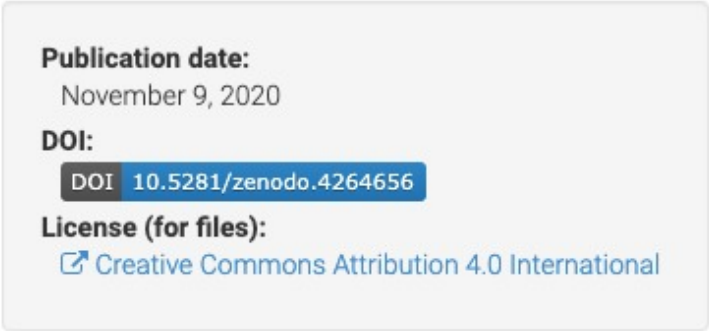
Hemos publicado nuestro dataset Acciones más activas de España y su histórico desde Enero 2019 en Zenovo. Lo podemos comprobar en el siguiente enlace:

https://zenodo.org/record/4264656#.X6ldv5NKhQI

Cite as:

Mª Dolores, & Andrea. (2020). Acciones más activas de España y su histórico desde Enero 2019 (Version 1.0) [Data set]. Zenodo.

http://doi.org/10.5281/zenodo.4264656



November 9, 2020

Dataset Open Access

Acciones más activas de España y su histórico desde Enero 2019

Mª Dolores; Andrea

Este script recoge los datos de las acciones de las cincuenta empresas más activas en la bolsa española, así como el histórico de cada una de ellas desde Enero de 2019.

Preview

	Fecha	Último valor	Valor apertura	Máx. valor	Min. valor	Volumen	% Variación
0	Nov 2020	92,375	86,925	95,600	85,750	161,18K	6,55%
1	Oct 2020	86,700	93,500	98,775	85,150	137,17K	-6,57%
2	Sep 2020	92,800	100,200	105,000	91,200	105,47K	-7,39%
3	Ago 2020	100,200	93,750	103,650	91,875	67,57K	6,82%
4	Jul 2020	93,800	87,200	101,000	84,875	204,99K	7,63%
5	Jun 2020	87,150	89,750	98,000	86,050	124,24K	-2,95%
6	May 2020	89,800	89,950	89,950	76,300	155,74K	-0,66%
7	Abr 2020	90,400	93,150	99,500	85,000	76,84K	-7,14%
8	Mar 2020	97,350	116,000	126,700	76,850	140,20K	-14,90%
9	Feb 2020	114,400	102,500	118,700	100,000	2,30M	11,72%

TABLA DE CONTRIBUCIONES AL TRABAJO:

Contribuciones	Firma
Investigación previa	AGV, MHG
Redacción de las respuestas	AGV, MHG
Desarrollo código	AGV, MHG