Nasdaq Website Scraper

Tipología y ciclo de vida de los datos, Práctica 1

Javier Advani (jadvani@uoc.edu)
Github: https://github.com/jadvani/NasdaqScraper

Zenodo: https://zenodo.org/record/6368832

Contexto

NASDAQ (National Association of Securities Dealers Automated Quotation) es el segundo mercado de valores y **bolsa de valores** automatizada y electrónica más grande de los Estados Unidos, siendo la **primera la Bolsa de Nueva York**, con más de 8000 compañías y corporaciones. Tiene más volumen de intercambio por hora que cualquier otra bolsa de valores en el mundo. Más de 7000 acciones de pequeña y mediana capitalización cotizan en la NASDAQ. Se caracteriza por comprender las empresas de alta tecnología en electrónica, informática, telecomunicaciones, biotecnología, y muchas otras más.

Descripción y Objetivos

Conocer los valores de acciones a tiempo real puede resultar crucial para decidir qué hacer con el capital invertido, por lo que la extracción automática del valor de la(s) empresa(s), sin necesidad de acudir al sitio web, así como la posterior manipulación para arrojar predicciones, son herramientas complementarias que pueden ayudar al inversor.
DISCLAIMER: Este trabajo se trata de una actividad académica para la asignatura de Topología y Ciclo de Vida de los Datos, para el master de Ciencias de Datos de la UOC.

El conjunto de funciones presentadas en este repositorio, elaboradas con Python 3.10, tienen como objetivo dos puntos principales:

- Obtener el listado actualizado de símbolos (identificador unívoco) de empresas.
- A partir del símbolo o listado de símbolos, obtener distintos parámetros relevantes para determinar el estado de la empresa.

El listado de símbolos de empresas completo a 13 de marzo de 2022 se ha guardado en `execution_results/symbols.txt`.

Los detalles de cada empresa del listado que se desee ejecutar se guardan en un CSV con el timestamp de ejecución, dentro del mismo directorio `execution_results/`

Contenido del dataset *Nasdaq Companies*

A partir del listado de símbolos, la utilidad de scraping accede a cada site asociado a la compañía para extraer los detalles de ésta. Los campos más interesantes son los recogidos en las siguientes 15 columnas del fichero CSV de salida:

- **symbol**: El símbolo de la compañía.
- **name**: El nombre completo.
- **price**: Precio de cada acción en el momento de ejecución.
- **pricing_changes**: Variación en dólares del precio en las últimas 24h.
- **pricing_percentage_changes**: Variación en % del precio en las últimas 24h.
- **sector**: Sector al que pertenece (Ejemplos: Technology, Consumer Services, Finance, ...)
- **industry**: Industria del sector (Ejemplo: Computer Manufacturing)
- **market_cap**: Capitalización total de la empresa, su valor total, en dólares.
- **share_volume**: Volumen de acciones. Son la cantidad de acciones que se han negociado en las últimas 24 horas.
- **earnings_per_share**: El beneficio por acción, lo que ha aportado en el periodo de un año.
- **annualized_dividend**: Utiliza el último dividendo pagado multiplicado por la frecuencia. Es el importe de un dividendo pagado a los accionistas en cuatro trimestres.
- **dividend_pay_date**: Fecha de pago de los últimos dividendos.
- **symbol_yield**: El rendimiento de la acción es la apreciación del precio de la acción más los dividendos pagados, dividido por el precio original de la acción.
- **beta**: La beta es una forma de medir la volatilidad de una acción en comparación con la volatilidad del mercado en general. El mercado en su conjunto tiene una beta de 1. Los valores con un valor superior a 1 son más volátiles que el mercado (lo que significa que generalmente subirán más de lo que sube el mercado y bajarán más de lo que baja el mercado)
- **errors**: Si se han detectado errores durante el escrapeo de la empresa, este campo se pone a True. Significa que no tenemos garantías de que todos los campos hayan sido correctamente obtenidos (especialmente los key data). Los del banner de la cabecera sí se obtienen correctamente, por eso los mantenemos. Quizá resulta de interés para la persona que analice y explote los datos corregirlos revisándolos manualmente, o simplemente los filtra y elimina.

Para estudiar y realizar predicciones de los precios de acciones, resultará interesante asociar la información al momento en que se toman los datos. Por esto mismo, se añade un timestamp al fichero que indica la fecha y hora de extracción. Podría añadirse una **columna adicional** que contuviese este **timestamp**, para así tener registradas varias entradas de la misma empresa en distintos momentos y aplicar técnicas tales como series temporales.

Agradecimientos

La propiedad de los datos tomados a tiempo real es íntegramente de Nasdaq, aunque se ofrece al público interesado gratuitamente, a través de la web y otros medios. Existen ya soluciones API de terceros, como esta de aquí, que permiten efectuar lo que se persigue con el script entregado. Como contrapartida, todas las APIs encontradas por la red limitan su uso gratuito a un número de llamadas: a partir de cierto volumen de llamadas, es necesario pagar una cuota.

El escrapeo de datos se efectuó con Python 3.x y Selenium. Utilizar otras herramientras de web scraping como Beautiful Soup resulta inviable porque buena parte de la web tiene contenido hecho en Javascript, lo que obliga a ir aplicando secuencias que emulen la interacción humana con la web.

Licencia

Se optó por una Licencia Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) para compartir el código y datos obtenidos de la web oficial de Nasdaq.
Aunque puede leerse el desglose de condiciones en el fichero de Licencia adjunto en este repositorio, hay dos motivos principales que su elección:

1. Se permite compartir, copiar y redistribuir el material en cualquier medio o formato.
2. Se permite adaptar, remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercial.

Bajo estas condiciones,

Atribución - Debe darse el crédito apropiado, proporcionar un enlace a la licencia e indicar si se hicieron cambios. Puede hacerse de cualquier manera razonable, pero no de ninguna forma que sugiera que el licenciante lo respalda.

ShareAlike - Si el material es remezclado, transformado, o se construye algo nuevo sobre éste, el material debe distribuirse bajo la misma licencia que el original.

Para más info, puede visitarse la web de Creative Commons.

Inspiración

Mi interés por las inversiones y una mejor comprensión del mundo financiero, dieron pie a la búsqueda de algo más de información acerca del tema. La excesiva cantidad de testimonios de inversores que han llegado a perderlo todo por una mala decisión, unido a las habilidades que estoy desarrollando al aprender nuevas técnicas de computación, han incentivado la creación de esta herramienta, que confío pueda ser útil para todos aquellos que también quieran empezar a indagar más en el mercado de valores.

Como posibles líneas de trabajo adicionales, y a fin de hacer una mejor explotación de los datos, consideraría muy interesante la elaboración de un bot en Telegram: alojado en algún tipo de dispositivo (como una Raspberry Pi) y ejecutándose periódicamente, enviamos un aviso en caso de que las acciones hayan variado considerablemente.

Otra opción aún más interesante, es la elaboración de un *roboadvisor*, un script que, a través de una API se comunicase con la entidad bancaria o la plataforma que facilite la compraventa de acciones, efectuando así la compraventa por sí solo, a partir de las predicciones que hiciera en el histórico de datos.

El código y dataset completos pueden descargarse desde el repositorio creado en GitHub. La carpeta src contiene el código, y export los resultados de ejecución: un fichero CSV con el resultado, y un txt “errors”, indicando las empresas que no han podido ser correctamente extraídas.

Bibliografía y recursos consultados

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.

Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC

R. Martin, and J. Coplien (2009) Clean code. Prentice Hall

https://selenium-python.readthedocs.io/