**Tipología y ciclo de vida de los datos**

PRA 1

**Daniel Priego Barea - Raúl Martínez Ballarín**

Contenidos

[Contexto 3](#_Toc131490498)

[Título 3](#_Toc131490499)

[Descripción del dataset 3](#_Toc131490500)

[Representación gráfica 3](#_Toc131490501)

[Contenido 3](#_Toc131490502)

[Propietario 3](#_Toc131490503)

[Inspiración 3](#_Toc131490504)

[Licencia 3](#_Toc131490505)

[Código 3](#_Toc131490506)

# Contexto

Actualmente existen gran número de paginas web de bolsa y con mucha información económica y datos financieros de las empresas cotizadas en bolsa, algunas proporcionan algunos análisis de estas acciones basados en los mismos criterios o ratios más comunes usados por los analistas.

Este proceso de scraping iría dirigido hacia un analista de valores de bolsa independiente que quisiera obtener la información detallada sobre las cuentas de resultados, balance y flujo de caja de la empresa para realizar automáticamente sus propios análisis fundamentales detallados con gráficos a su propio criterio y comparaciones entre empresas.

La información para elaborar el conjunto de datos de partida para estos análisis se extrae de la web <https://finance.yahoo.com>. Esta web ofrece información financiera muy amplia como cotizaciones de bolsa, índices bursátiles, noticias financieras…además de herramientas para la gestión de finanzas personales.

El conjunto de datos obtenido a través del web scraping proporciona información del Estado de ingresos y Balance de cuentas sobre distintas compañías tecnológicas que servirá para realizar análisis y recomendaciones sobre las acciones de estas empresas.

# Título

Scraping financiero de empresas tecnológicas.

# Descripción del dataset

El dataset contiene los datos financieros de **Income Statement** y **Balance Sheet** de un grupo de empresas del sector tecnológico, pero puede adaptarse a cualquier otro tipo de empresa o incluso a todas las empresas de un determinado mercado o región. Por temas de optimización, consumo de memoria y tiempo de procesamiento se ha acotado el proceso a unas cuantas empresas tecnológicas.

# Representación gráfica

# Contenido

# Propietario

Un requisito básico que da sentido a las Bolsas es la transparencia y la limpieza de información de las compañías cotizadas. Los reguladores bursátiles, como la CNMV española, la FCA británica o la SEC americana, ponen mucha atención en velar por que las empresas cotizadas ofrezcan toda la información necesaria a los accionistas e inversores.

Un requisito de las compañías que cotizan en la mayoría de Bolsas es que deben presentar sus resultados trimestrales cuatro veces al año (tantas como trimestres tienen el ejercicio). De esa forma, los inversores se van haciendo una idea puntual de cómo están evolucionando los negocios de la compañía, pudiendo acceder a la información más completa y actualizada.

Por tanto, los datos recopilados por el scraper son de ámbito público y propiedad ultima de cada empresa, ya que son publicados y accesibles en la web de cada una de las diferentes empresas que cotizan en bolsa. Yahoo actúa como agrupador de estos datos en un formato web común, por ello se realiza en scrapeo de la información ahí, en vez de realizar consulta en cada una de las webs corporativas de las empresas que se quiera analizar.

# Inspiración

El motivo de la elección de este conjunto de datos es debido a un interés personal en este ámbito. Es cierto que hay diversas webs que proporcionan información similar (morningstar.es, investing.com…) sobre la “salud” financiera de las distintas empresas si bien se echa en falta la posibilidad de poder comparar de una manera ágil y sencilla la información entre las diferentes empresas.

Con el conjunto de datos obtenido se pretende poder realizar comparativas entre las diferentes empresas bien sea por las distintas métricas financieras obtenidas, así como su evolución a lo largo de los distintos ejercicios. Dicha información puede ser un factor a tener en cuenta en el caso de estar interesados en una posible inversión de capital en acciones de alguna empresa.

# Licencia

# Código

El código del scraper se ha realizado en Python en formato Jupyter Notebook.

Debido a la complejidad del sitio web elegido se ha tenido que implementar el control de un navegador web desde Python ya que el código HTML de la web contiene gran número de scripts que generan datos/links dinámicos.

Para ello se ha utilizado la librería Selenium:

<https://pythonbasics.org/selenium-firefox/>

con el webdriver para Mozila Firefox:

<https://github.com/mozilla/geckodriver/releases>

Con el objeto webdriver controlamos el navegador Firefox y realizamos las acciones de navegación y pulsación de botones desde el código Python.

Seguidamente una vez se ha navegado cada página HTML que contiene la información que nos interesa (Financials -> Income Statement / Balance Sheet) parseamos el código HTML con la librería BeautifulSoup seleccionando los datos de los <tags> que nos interesan.

<https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)>

Con respecto a las dificultades que presenta el sitio web, se destaca el hecho de tener que gestionar los pop-ups referentes tanto a la gestión de cookies, así como al inicio de sesión:

Text

Description automatically generated Diagram, application

Description automatically generated

Adicionalmente, la simulación de los clicks necesarios tanto para la navegación por la web así como para la recolección de los identificadores de acciones que requiere recorrer una tabla con paginación.

Table

Description automatically generated

Estos aspectos han sido resueltos utilizando la librería selenium en Python y el webDriver geckodriver para Firefox.

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa |  |
| Redacción de las respuestas |  |
| Desarrollo del código |  |
| Participación en el vídeo |  |