

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

PRA 1

**Daniel Priego Barea - Raúl Martínez Ballarín**

## Contenidos

Contexto .....	3
Título .....	3
Descripción del dataset .....	3
Representación gráfica.....	3
Contenido .....	4
Propietario.....	7
Inspiración .....	8
Licencia.....	8
Código .....	8
Dataset .....	9
Video .....	10
Firmas.....	10

## 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.

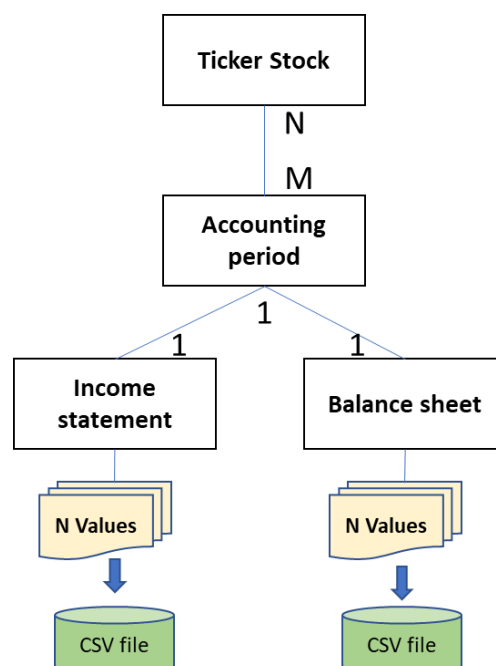
## Título

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

El dataset resultante consta de 2 dataframes que se traducen en dos ficheros csv generados. Por una parte, el primero de ellos "Income" hace referencia al estado de ingresos de las distintas compañías en los diferentes ejercicios. Por la otra, el dataset "Balance" el cual hace referencia al balance de las empresas.

Las dimensiones de cada uno de los datasets son:

Income: 45 columnas, 99 filas

Balance: 129 columnas, 79 filas

Estos dos datasets tienen en común los siguientes campos:

1. Company
2. Breakdown

Los campos específicos de cada dataset son:

Dataset Income:

1. Total Revenue
2. Operating Revenue
3. Cost of Revenue
4. Gross Profit
5. Operating Expense
6. Selling General and Administrative
7. Research & Development
8. Operating Income
9. Net Non Operating Interest Income Expense
10. Interest Income Non Operating
11. Interest Expense Non Operating
12. Other Income Expense
13. Other Non Operating Income Expenses
14. Pretax Income
15. Tax Provision
16. Net Income Common Stockholders
17. Net Income
18. Net Income Including Non-Controlling Interests
19. Net Income Continuous Operations
20. Diluted NI Available to Com Stockholders
21. Basic EPS
22. Diluted EPS
23. Basic Average Shares
24. Diluted Average Shares
25. Total Operating Income as Reported
26. Total Expenses
27. Net Income from Continuing & Discontinued Operation
28. Normalized Income
29. Interest Income
30. Interest Expense
31. Net Interest Income
32. EBIT
33. EBITDA

34. Reconciled Cost of Revenue
35. Reconciled Depreciation
36. Net Income from Continuing Operation Net Minority Interest
37. Normalized EBITDA
38. Tax Rate for Calcs
39. Tax Effect of Unusual Items
40. Total Unusual Items Excluding Goodwill
41. Total Unusual Items
42. Earnings from Equity Interest Net of Tax
43. Average Dilution Earnings

Dataset Balance:

1. Total Assets
2. Current Assets
3. Cash, Cash Equivalents & Short Term Investments
4. Cash And Cash Equivalents
5. Cash
6. Cash Equivalents
7. Other Short Term Investments
8. Receivables
9. Accounts receivable
10. Other Receivables
11. Inventory
12. Other Current Assets
13. Total non-current assets
14. Net PPE
15. Gross PPE
16. Properties
17. Land And Improvements
18. Machinery Furniture Equipment
19. Leases
20. Accumulated Depreciation
21. Investments And Advances
22. Investment in Financial Assets
23. Available for Sale Securities
24. Other Investments
25. Other Non Current Assets
26. Total Liabilities Net Minority Interest
27. Current Liabilities
28. Payables And Accrued Expenses
29. Payables
30. Accounts Payable
31. Current Debt And Capital Lease Obligation
32. Current Debt
33. Commercial Paper
34. Other Current Borrowings
35. Current Deferred Liabilities
36. Current Deferred Revenue
37. Other Current Liabilities
38. Total Non Current Liabilities Net Minority Interest
39. Long Term Debt And Capital Lease Obligation
40. Long Term Debt

41. Trade and Other Payables Non Current
42. Other Non Current Liabilities
43. Total Equity Gross Minority Interest
44. Stockholders' Equity
45. Capital Stock
46. Common Stock
47. Retained Earnings
48. Gains Losses Not Affecting Retained Earnings
49. Total Capitalization
50. Common Stock Equity
51. Net Tangible Assets
52. Working Capital
53. Invested Capital
54. Tangible Book Value
55. Total Debt
56. Net Debt
57. Share Issued
58. Ordinary Shares Number
59. Gross Accounts Receivable
60. Allowance For Doubtful Accounts Receivable
61. Raw Materials
62. Work in Process
63. Finished Goods
64. Hedging Assets Current
65. Buildings And Improvements
66. Other Properties
67. Goodwill And Other Intangible Assets
68. Goodwill
69. Other Intangible Assets
70. Long Term Equity Investment
71. Total Tax Payable
72. Income Tax Payable
73. Pension & Other Post Retirement Benefit Plans Current
74. Long Term Capital Lease Obligation
75. Non Current Deferred Liabilities
76. Non Current Deferred Taxes Liabilities
77. Non Current Deferred Revenue
78. Capital Lease Obligations
79. Prepaid Assets
80. Construction in Progress
81. Non Current Deferred Assets
82. Non Current Deferred Taxes Assets
83. Non Current Prepaid Assets
84. Other Payable
85. Current Accrued Expenses
86. Interest Payable
87. Current Capital Lease Obligation
88. Employee Benefits
89. Preferred Stock
90. Additional Paid in Capital
91. Treasury Stock
92. Treasury Shares Number

- 93. Notes Receivable
- 94. Due from Related Parties Current
- 95. Receivables Adjustments Allowances
- 96. Investments in Associates at Cost
- 97. Held To Maturity Securities
- 98. Dividends Payable
- 99. Due to Related Parties Current
- 100. Non Current Pension And Other Post-Retirement Benefit Plans
- 101. Other Equity Adjustments
- 102. Other Equity Interest
- 103. Minority Interest
- 104. Preferred Stock Equity
- 105. Preferred Shares Number
- 106. Taxes Receivable
- 107. Inventories Adjustments Allowances
- 108. Assets Held for Sale Current
- 109. Financial Assets
- 110. Non Current Accounts Receivable
- 111. Non Current Note Receivables
- 112. Defined Pension Benefit
- 113. Non Current Accrued Expenses
- 114. Derivative Product Liabilities
- 115. Loans Receivable
- 116. Other Inventories
- 117. Current Deferred Assets
- 118. Current Provisions
- 119. Long Term Provisions
- 120. Investments in Other Ventures Under Equity Method
- 121. Preferred Securities Outside Stock Equity
- 122. Liabilities Held for Sale Non Current
- 123. Restricted Cash
- 124. Line of Credit
- 125. Unrealized Gain Loss
- 126. Minimum Pension Liabilities
- 127. Foreign Currency Translation Adjustments

## 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 última 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. Adicionalmente, 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

En referencia a la licencia adecuada para el dataset se trataría de la licencia CC BY-NC-SA. La elección de esta se basa en los siguientes aspectos:

- Se permite compartir (copiar y redistribuir) el material en cualquier medio o formato.
- Se permite adaptar, es decir remezclar, transformar y construir a partir del material.
- Se debe dar crédito de manera adecuada e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.
- No se permite un uso comercial de la obra original ni de las posibles obras derivadas.
- La distribución de estas obras derivadas se debe hacer con una licencia igual a la que regula la obra original.

## 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 Mozilla 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\\_\(HTML\\_parser\)](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:



Cuando utilizas nuestros sitios y aplicaciones, usamos **cookies** para:

- proporcionarte nuestros sitios y aplicaciones;
- autenticar usuarios, aplicar medidas de seguridad y evitar el spam y los abusos, y
- medir el uso que haces de nuestros sitios y aplicaciones.

Si haces clic en «**Aceptar todo**», nosotros y **nuestros socios** también utilizaremos cookies y tus datos personales (como tu dirección IP, ubicación precisa y datos de navegación y búsqueda) para:

- mostrar anuncios y contenido personalizados basados en perfiles de interés;
- medir la efectividad de los anuncios y el contenido personalizados, y
- desarrollar y mejorar nuestros productos y servicios.


Si no quieres que nosotros ni nuestros socios utilicemos cookies y datos personales para estos propósitos adicionales, haz clic en «**Rechazar todo**».

Si quieres personalizar tus opciones, haz clic en «**Gestionar configuración de privacidad**».

Puedes cambiar tus opciones en cualquier momento haciendo clic en el enlace «Panel de control de privacidad» de nuestros sitios y aplicaciones. Para obtener más información sobre cómo utilizamos tus datos personales, consulta nuestra [Política de privacidad](#) y [Política de cookies](#).

Aceptar todoRechazar todoGestionar configuración de privacidad

**Sign in to save ANSS**



Follow ANSS and similar companies and stay on top of the trends

Sign In

**Maybe later**

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.

**Matching Stocks** 1-25 of 425 results [Add to Portfolio](#)

⚠ Results were generated a few mins ago. Pricing data is updated frequently

<input type="checkbox"/> Symbol	Name
<input type="checkbox"/> AAPL	Apple Inc.
<input type="checkbox"/> MSFT	Microsoft Corporation
<input type="checkbox"/> NVDA	NVIDIA Corporation
<input type="checkbox"/> TSM	Taiwan Semiconductor Manufacturing Company Limited
<input type="checkbox"/> ASML	ASML Holding N.V.
<input type="checkbox"/> AVGO	Broadcom Inc.
<input type="checkbox"/> ORCL	Oracle Corporation
<input type="checkbox"/> CSCO	Cisco Systems, Inc.

## Dataset

<https://doi.org/10.5281/zenodo.7825842>

## Video

[https://drive.google.com/file/d/1sIOF0u6GMN-ILCKmKIAalrtb0D6WwW8d/view?usp=share\\_link](https://drive.google.com/file/d/1sIOF0u6GMN-ILCKmKIAalrtb0D6WwW8d/view?usp=share_link)

## Firmas

<b>Contribuciones</b>	<b>Firma</b>
Investigación previa	Daniel Priego Barea, Raúl Martínez Ballarín
Redacción de las respuestas	Daniel Priego Barea, Raúl Martínez Ballarín
Desarrollo del código	Daniel Priego Barea, Raúl Martínez Ballarín
Participación en el vídeo	Daniel Priego Barea, Raúl Martínez Ballarín