Jordi Serra
Paula Sobrevals

# PRACTICA1: WEB SCRAPING

**Link al GitHub:** https://github.com/PSobrevals/PAC1_WebScraping
**Document PDF amb les preguntes i respostes a respondre de la PRAC1:**

1. **Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.**

   La informació utilitzada per a fer el treball s'ha obtingut de la següent pàgina web: investing.com. Aquesta pàgina web està considerada una de les 3 millors pàgines webs financeres del món (segons SimilarWeb y Alexa). Investing.com està enfocat als mercats financiers i proporciona dades a temps real sobre 250 mercats d'arreu del món, així com cotitzacions, índexs, prediccions futures, anàlisis i notícies del sector.
   Hem escollit investing ja que ens proporciona accés il·limitat a temps real sobre els indexs bursàtils, és a dir, les cotitzacions mundials, tant principals com secundàries, d'arreu del món de manera gratuita. A més a més, aquesta informació es troba organitzada en taules segons el país de l'índex corresponent.

2. **Definir un títol pel dataset. Triar un títol que sigui descriptiu.**

   Cotitzacions mundials de borsa

3. **Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).**

   En el dataset podem observar les cotitzacions bursàtils de tot el món, organitzades per taules segons el país dels índexs en qüestió. Aquestes taules estan ordenades alfabèticament segons el país. Cada taula compta amb el nom del país el qual l'índex pertany, així com, el nom de l'índex, l'últim valor observat de cada un, el corresponent màxim i mínim observat, el canvi (tant numèric com amb percentatge) i per últim l'hora de l'observació. Al final del dataset hi trobem la data i hora en la qual s'han obtingut les dades.

Jordi Serra
Paula Sobrevals

4. **Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment**


Considerem que aquesta imatge descriu el dataset de manera visual, ja que, hi podem observar les cotitzacions bursàtils de diversos llocs del món (Nova York, Tokyo, Hong Kong i Gran Bretanya), així com el seu índex bursàtil. A més podem veure el nombre de les cotitzacions i el canvi del seu valor, tant si és positiu com negatiu, exactament com en el dataset.

5. **Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.**

El dataset conté la següent informació extreta directament de la web on s'ha fet el scraping. És a dir, que no s'edita en cap moment el contingut del fitxer CSV que representa al repositori de dades.
El dataset inclou diverses taules agrupades per països, és a dir, els indexs bursàtils (que contenen la informació explicada anteriorment a l'apartat 3), es troben a la taula del país al qual corresponen. Per tant, de manera rápida es pot

extreure informació localitzada per un país o continent ràpidament a partir del CSV directament.
El període de temps de les dades, és anotat al final del dataset, en el qual hi trobem la data i l'hora en la qual aquestes s'han extret. Com hem comentat, les dades varien molt ràpidament, així que els valors observats són vàlids en aquella data determinada.
Les dades d'han recollit mitjançant Web Scraping, concretament utilitzant la llibreria Selenium versió 3.141.0 de Python 3.6.0. El script utilitzat es pot trobar al *main.py* del GitHub.

6. **Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).**

El propietari de la web és © 2007-2020 Fusion Media Limited (Forexpros)[1]. Aquesta companyia fundada al 2007, proporciona eines i informació relacionada amb el mercat financer, com per exemple cotitzacions a temps real, gràfics en *streaming*, notícies financeres actualitzades, anàlisis tècniques, directori, llistats d'agents i un calendari econòmic. A més, com hem comentat anteriorment, proporciona informació detallada sobre divises, cotitzacions, índexs, accions i prediu els índexs futurs.[2]

1. Investing.com 2007, *About Investing.com* <https://www.investing.com/about-us/>
2. Crunchbase 2007, *Fusion Media* <https://www.crunchbase.com/organization/fusion-media-limited>

7. **Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.**

Per tal de trobar una web que canvia molt durant el dia, inclús varia en qüestió de minuts vam pensar en les possibilitats que poden tenir o bé les cotitzacions de les monedes o de la borsa de valors. En aquest cas, per tal de poder fer més accesible la variació dels valors al llarg del temps i poder preveure un posible canvi de tendència es va escollir fer servir una web de cotitzacions bursatils a nivell mundial.

8. **Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:**

    - **Released Under CC BY-NC-SA 4.0 License**

    La llicència creative commons permet a terceres persones poder fer servir el codi lliurament, basant-se en la compartició de coneixement lliure.
    En aquest cas s'escull la versió (BY-NC-SA) per a preservar aquesta compartició del codi i del dataset que hem creat. En el cas que algú faci una obra derivada haurà de citar la font i l'haurà de compartir en format identic, a la vegada que no permetem que es faci negoci amb el codi lliure que hem proporcionat a la comunitat.

9. **Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.**

    S'ha fet el codi en Python, fent servir la llibreria SELENIUM.

    Detallem molt breument els punts més importants:

    Instal·lem el driver del navegador Chrome per tal de visitar la pàgina web i poder obtenir les dades que després es pasaran al repositori en CSV. Ho fem amb el següent codi.

```
driver = webdriver.Chrome(ChromeDriverManager().install())
```

    En el codi següent s'accepta el diàleg per a les cookies que surt a l'obrir la pàgina web. Únicament s'accepta la política de cookies tal i com està.

```
driver.find_elements_by_xpath('//*[@id="onetrust-accept-btn-handler"]')[0].click()
```

    Per posteriorment obtenir les dades de la web mitjançant el següent codi de SELENIUM

```
# XPath del nom dels paisos: '//*[@id="leftColumn"]//h2'
paisos = driver.find_elements_by_xpath('//*[@id="leftColumn"]//h2')
```

```
# XPath del titol de les taules:'//*[contains(@id,"indice_table_")]/thead//tr/th'
rows_title = driver.find_elements_by_xpath('//*[@id="indice_table_28"]/thead/tr/th')
```

El codi complert i funcional en python és al GitHub amb el nom: *main.py>*

https://github.com/PSobrevals/PAC1_WebScraping

10. **Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.**

Allà El DOI per accedir al CSV penjat al Zenodo és el següent, clicant-hi a sobre podem accedir a la pàgina en qüestió: `10.5281/zenodo.4264782`.
Allà hi podem veure una breu descripció del dataset, una la previsualització d'aquest (en la qual només hi apareix la primera columna, però no s'hi pot veure tota la taula, un cop descarregat es té accés a tot el dataset) i accés per a descarregar-lo.
Exemple dels tres primers països:

Cotitzacions mundials

| **Alemania** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Índice** | Último | Máximo | Mínimo | Var. | % Var. | Hora | |
| **DAX** | 13.172,35 | 13.296,15 | 12.667,85 | 692,33 | 5,55% | 16:30:52 | |
| **Euro Stoxx 50** | 3.420,16 | 3.442,12 | 3.237,50 | 216,11 | 6,74% | 16:30:46 | |
| **Classic All Share** | 8.808,91 | 8.913,35 | 8.395,98 | 429,63 | 5,13% | 16:15:00 | |
| **Midcap** | 28.016,99 | 28.520,79 | 27.671,88 | 670,70 | 2,45% | 16:15:00 | |
| **Technology All Share** | 4.111,38 | 4.297,71 | 4.064,92 | -88,71 | -2,11% | 16:15:00 | |
| **HDAX** | 7.247,89 | 7.305,27 | 6.995,12 | 358,84 | 5,21% | 16:15:00 | |
| **Prime All Share** | 5.356,88 | 5.403,17 | 5.104,02 | 256,78 | 5,03% | 16:15:00 | |
| **SDAX** | 12.641,83 | 12.861,91 | 12.447,27 | 330,62 | 2,69% | 16:15:00 | |
| **TecDAX** | 2.980,29 | 3.084,25 | 2.953,63 | -33,82 | -1,12% | 16:15:00 | |
| **XETRA DAX Price** | 13.219,17 | 13.297,05 | 12.670,58 | 739,15 | 5,92% | 16:15:00 | |
| | | | | | | | |
| **Arabia Saudí** | | | | | | | |
| **Índice** | Último | Máximo | Mínimo | Var. | % Var. | Hora | |
| **MSCI TADAWUL 30** | 1.086,11 | 1.086,11 | 1.077,04 | 3,72 | 0,34% | 05/11 | |
| **Tadawul All Share** | 8.366,46 | 8.366,46 | 8.178,33 | 206,11 | 2,53% | 13:19:00 | |
| **NOMU Parallel Market Capped** | 17.654,95 | 17.831,92 | 16.606,56 | 755,84 | 4,47% | 13:19:00 | |
| | | | | | | | |
| **Argentina** | | | | | | | |
| **Índice** | Último | Máximo | Mínimo | Var. | % Var. | Hora | |
| **S&P Merval** | 49.916,44 | 50.586,65 | 48.492,54 | 1423,90 | 2,94% | 16:10:00 | |
| **S&P/BYMA Argentina General** | 2.102.229 | 2.127.134 | 2.041.535 | 60694 | 2,97% | 16:10:00 | |
| | | | | | | | |

Jordi Serra
Paula Sobrevals

**Taula de contribucions al treball:**

| **Contribucions** | **Signatura** | |
|---|---|---|
| | **Jordi** | **Paula** |
| Recerca Prèvia | ✔ | ✔ |
| Redacció de les respostes | ✔ | ✔ |
| Desenvolupament codi | ✔ | ✔ |