Tipologia i cicle de vida de les dades
Pràctica 1: Web scraping
Memòria

Nota: Les respostes a les diferents preguntes estan escrites amb blau per tal de ser fàcilment identificables

Components de l'equip (usuaris GitHub): Andrés Laverde Marín (alaverma)
Josep Mª Espasa Verdés (ilergeta)
Enllaç GitHub: https://github.com/alaverma/web-scraping-uoc

1. Context

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

L'objectiu principal d'aquesta pràctica ha estat recol·lectar informació de la cotització borsària històrica de les empreses que cotitzen a la Borsa de Madrid, principal mercat de valors d'Espanya, integrada en la Sociedad de Bolsas y Mercados Españoles (BME).

El lloc web www.bolsamadrid.es proporciona aquesta informació, ja que és el lloc web de la societat rectora de la borsa de valors de Madrid, propietària de d'aquestes dades.

2. Títol del dataset

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

Les dades recol·lectades es recullen en un dataset emmagatzemat en un fitxer de format CSV amb el nom corresponent al *ticker* de l'empresa cotitzada de la que s'està obtenint les dades històriques de cotització. Recordar que el *ticker* no deixa de ser un identificador de les empreses que cotitzen en un mercat borsari i que es correspon a una abreviació que pot contenir lletres i/o nombres.

Vàries empreses / fitxers??

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).

TODO

4. Representació gràfica

Presentar una imatge o esquema que identifiqui el dataset visualment

TODO

5. Contingut

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

TODO

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 les dades és: **Sociedad Rectora de la Bolsa de Valores de Madrid, S.A. Sociedad Unipersonal**

D'altra banda, per la realització d'aquesta pràctica, ja sigui, pròpiament, aquest document de memòria o el codi font, s'ha consultat la següent bibliografia:

Col·laboradors de Viquipèdia. 'Borsa de Madrid', *Viquipèdia, l'Enciclopèdia Lliure.* [Data de consulta: 10 de novembre de 2018]. <https://ca.wikipedia.org/wiki/Borsa_de_Madrid>

Col·laboradors de Viquipèdia. 'Ticker symbol', *Viquipèdia, l'Enciclopèdia Lliure.* [Data de consulta: 10 de novembre de 2019]. <https://en.wikipedia.org/wiki/Ticker_symbol>

Lawson, R. (2015). *Web Scraping with Python* (1a ed.). Birmingham: Packt Publishing Ltd.

Afegir més biblio emprada?!?

7. Inspiració

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

Si es planteja la inversió en renda variable des del punt de vista analític, conèixer la cotització històrica dels diferents valors és de vital importància per tal de poder aplicar les eines de modelatge/predicció que es considerin més adequades, així com per poder comprovar la seva efectivitat. En aquest sentit, l'obtenció de l'històric de cotitzacions borsàries en un format CSV que posteriorment sigui fàcilment exportable a qualsevol eina que es vulgui emprar, és clau per tal de poder realitzar aquesta anàlisi analítica i, d'aquesta manera, poder prendre les decisions adequades per tal d'intentar maximitzar els beneficis obtinguts de les inversions realitzades.

Des d'aquest punt de vista analític, mitjançant l'anàlisi del conjunt de dades històriques de determinats valors borsaris, emprant les eines i models que es considerin més adequats, es pretén donar resposta a la pregunta *clau* que tot inversor en borsa es fa: És un bon moment per comprar/vendre l'acció de l'empresa X?

8. Llicència

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

Tant el codi font com els datasets obtinguts estan sota la llicència **CC BY Reconeixement 4.0 Internacional** , mitjançant la qual, qualsevol persona pot mesclar, adaptar i crear a partir d'aquesta informació facilitada, fins i tot amb una finalitat comercial, sempre que reconegui l'autoria de la creació original.

El principal motiu per triar aquesta llicència és doble. Així, d'una banda, per tal d'aconseguir una màxima difusió de la informació aconseguida i de les possibilitats de les tasques de web scraping i, de l'altra banda, per ajudar a la comunitat aportant una nova implementació.

9. Codi

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

El codi desenvolupat per realitzar aquesta pràctica es pot trobar en l'enllaç del GitHub facilitat a l'inici del present document. Concretament, tot el codi es troba dins la carpeta `src`, agrupat en dos fitxers de la següent manera:

- **main.py:** fitxer principal d'execució del programa i que inicia el procés de scraping amb la crida a la classe BolsaScraper.
- **bolsaScraper.py:** conté la implementació de la classe BolsaScraper que, mitjançant els seus mètodes, genera la informació borsària sol·licitada i l'escriu al directori `data`.

Finalment, comentar que s'ha emprat el llenguatge de programació Python.

10. Dataset

Presentar el dataset en format CSV.

Un exemple de dataset en format CSV es pot trobar en el directori `data` de l'enllaç del GitHub facilitat en l'inici del present document, concretament, s'adjunten els fitxers de sortida per la cerca de les empreses XX, obtinguts mitjançant la comanda `python main XXX  XX XX XX`.

Igualment, com a part addicional de les dades recollides en el procés de web scraping, en el directori `images` del mateix enllaç facilitat, es poden trobar les imatges dels logos de les empreses cercades anteriorment, emmagatzemades com fitxers .gif, amb el nom del *ticker* identificatiu de l'empresa que, de fet, coincideix amb el nom del dataset anterior.

Contribucions	Signa
Recerca prèvia	A.L.M. i JM.E.V.
Redacció de les respostes	A.L.M. i JM.E.V.
Desenvolupament codi	A.L.M. i JM.E.V.