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 per cadascuna de les empreses sol·licitades es recullen en un dataset emmagatzemat en un fitxer de format CSV amb el nom corresponent al *ticker* de cadascuna de les empreses cercades de les que s'estan 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.

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).

El conjunt de dades extret, tal com s'ha comentat anteriorment, s'emmagatzema en fitxers de format CSV per cada empresa cercada. L'estructura d'aquest fitxer, està formada per una primera part a mode d'encapçalament, les tres primeres línies, i una segona part on s'adjunten el conjunt de dades de cotitzacions històriques borsàries sol·licitades, a mode de taula.

4. Representació gràfica

Presentar una imatge o esquema que identifiqui el dataset visualment

	A	B	C	D	E	F	G	H	I
1	AENA								
2	Image	url:	http://www.bolsamadrid.es/imagescomun/logosEmisoras/05046.gif						
3	Image	folder:	AENA.gif						
4	Fecha	Cierre	Referencia	Volumen	Efectivo	Último	Máximo	Mínimo	Medio
5	02/04/2019	161,65	160,7	372269	59969075,65	161,65	162,3	159,45	161,3003
6	03/04/2019	163,75	161,65	150812	24652421,45	163,75	164	161,15	163,4568
7	04/04/2019	162,15	163,75	671663	109887120,25	162,15	164,55	162,1	162,4527
8	05/04/2019	162,4	162,15	155221	25216630,55	162,4	163,15	161,95	162,4563
9	08/04/2019	159	162,4	659464	105909708,5	159	161,1	158,25	159,3445
10	09/04/2019	159	159	888904	141606617,3	159	160,3	158,55	159,0591
11	10/04/2019	162,35	159	1671098	269923941,55	162,35	162,7	160,95	162,2493
12	11/04/2019	163,5	162,35	1191386	193894183,45	163,5	163,75	161,35	162,889
13	12/04/2019	163,4	163,5	199845	32650754,9	163,4	163,95	162,25	163,3804
14	15/04/2019	165	163,4	600738	99110492,05	165	165,8	163,8	164,932
15	16/04/2019	164	165	170039	27916582,17	164	165,5	163,6	164,1343
16	17/04/2019	158,4	157,1	140672	22235336,75	158,4	158,95	156,15	158,0651
17	18/04/2019	158,9	158,4	659724	105269057,75	158,9	160,8	158,4	159,1436
18	23/04/2019	159	158,9	1812587	288670976,48	159	159,5	157,1	158,7215
19	24/04/2019	160,25	159	1739137	277314684,3	160,25	160,3	158,65	160,1022
20	25/04/2019	160,8	160,25	1254047	201286412,9	160,8	161,35	159,4	160,7115
21	26/04/2019	162,75	160,8	206255	33441464,65	162,75	163,35	160,65	162,4015
22	29/04/2019	162,8	162,75	628133	102198649,97	162,8	163,15	161,8	162,719
23	30/04/2019	165,35	162,8	226068	37316756,1	165,35	166,2	162,5	165,1709
24	02/05/2019	165,75	165,35	1067402	177325412,88	165,75	167,55	165,25	165,8847
25	03/05/2019	165,25	165,75	317858	52775008,52	165,25	166,65	165,25	165,5125
26	06/05/2019	164,9	165,25	718895	117377464,4	164,9	164,9	162,75	164,263
27	07/05/2019	164,4	164,9	214618	35340112,15	164,4	166,1	163,65	164,5262
28	08/05/2019	164,55	164,4	165068	27142098,73	164,55	165,1	163,4	164,4407
29	09/05/2019	165,1	164,55	259567	42751261,5	165,1	165,2	162,95	164,8332
30	10/05/2019	163,85	165,1	171328	28143675,15	163,85	165,2	163,55	164,0693
31	13/05/2019	163,7	163,85	302389	49636852,1	163,7	164,75	162,7	163,6291
32	14/05/2019	165,55	163,7	173901	28662677,4	165,55	165,55	163,9	165,2788
33	15/05/2019	167	165,55	396170	65767264,78	167	167,45	165,05	166,6926
34	16/05/2019	168,1	167	130613	21914652,47	168,1	168,1	166,4	167,7988
35	17/05/2019	166,75	168,1	139763	23106611,1	166,75	168,1	165,45	166,6144
36	20/05/2019	165	166,75	147864	24399612,45	165	166,55	164,05	164,9954
37	21/05/2019	167,75	165	113372	18988675,6	167,75	168,1	164,95	167,49
38	22/05/2019	168	167,75	197730	33113229,66	168	168,45	166,5	167,7441
39	23/05/2019	165,75	168	318190	53032198,05	165,75	167,6	165,55	165,9741
40	24/05/2019	167,1	165,75	133738	22336099,9	167,1	167,5	166,1	167,015

Figura 1. Captura de pantalla de les dades extretes de l'empresa AENA del 02/04/2019 fins 02/04/2019.

5. Contingut

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

Tal com s'ha comentat anteriorment, l'estructura dels fitxers de sortida està formada per dues parts diferenciades: l'encapçalament i la taula de dades pròpiament.

Respecte a l'encapçalament, indicar que està format per les primeres 3 línies del fitxer, on s'inclouen les següents dades:
línia 1 → Nom de l'empresa cotitzada de la que s'adjunta la informació històrica
línia 2 → Enllaç web on es troba allotjat el logo de l'empresa
línia 3 → Nom del fitxer del logo de l'empresa que s'ha emmagatzemat en local, concretament en la carpeta `images` i que coincideix amb el ticker de l'empresa amb format gif

D'altra banda, la taula de dades, que està formada per les línies 4 i posteriors del fitxers CSV, té la següent estructura:
línia 4 → Títol dels diferents camps dels que s'ha recollit informació
línies 5 i posteriors → informació borsària obtinguda per l'empresa sol·licitada amb els següents camps:

- *Fecha:* Data de la que s'adjunta la informació, en format dia/mes/any (p.e.: 01/04/2019).
- *Cierre:* Valor de l'acció, en euros, de l'empresa cotitzada en el moment de tancament de la sessió en la data indicada en el camp *Fecha*.
- *Referencia:* Valor de l'acció, en euros, de l'empresa cotitzada en el tancament de la sessió en la data indicada en el camp *Fecha* i que es pren com a referència en altres mercats com els futurs. Aquest valor s'obté mitjançant un algoritme per evitar que moviments especulatius en el tancament puguin influir en aquests derivats.
- *Volumen:* Quantitat d'accions de l'empresa cotitzada negociades (comprades i venudes) durant la sessió de la data indicada en el camp *Fecha*.
- *Efectivo:* Quantitat d'euros que implica la negociació de l'empresa cotitzada durant la sessió de la data indicada en el camp *Fecha*.
- *Último:* Valor de l'acció, en euros, en el darrer moviment de la sessió en la data indicada en el camp *Fecha*.
- *Màximo:* Valor de l'acció màxim, en euros, que ha assolit durant la sessió en la data indicada en el camp *Fecha*.
- *Mínimo:* Valor de l'acció mínim, en euros, que ha assolit durant la sessió en la data indicada en el camp *Fecha*.
- *Medio:* Valor de l'acció mig, en euros, que ha assolit durant la sessió en la data indicada en el camp *Fecha*.

Cal remarcar que totes les xifres tenen la coma com separador decimal i el punt com separador de milers.

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. 'IBEX 35', *Viquipèdia, l'Enciclopèdia Lliure.* [Data de consulta: 10 de novembre de 2019].

<https://es.wikipedia.org/wiki/IBEX_35>

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.

Finalment, indicar que s'ha consultat la documentació de les llibreries emprades `requests`, `csv`, `psutil`, `platform`, `os`, `datetime`, `sys`, `pickle`, `selenium` i `BeautifulSoup`.

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. 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 BANCO DE SABADELL i CAIXABANK, obtinguts mitjançant la comanda `python main.py 'sabadell' 'caixabank' --start 20/08/2019 -end 09/11/2019`.

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.