UOC, Tipologia i cicle de vida de les dades

PRAC1:"WebScraping" Borsa Barcelona

Descripció de l'activitat

L'objectiu d'aquesta pràctica es l'obtenció de les dades contingudes de la Borsa de Barcelona que fan referència al'Ibex 35 (https://www.borsabcn.es/).

1. Context

La recol·lecta de dades s'ha fet en el sector financer. La borsa es una entitat privada on l'accés a les seves dades està restringit als brokers i mitjans de comunicació amb els que tenen acords.

La gent que es dedica a la borsa normalment ha d'adquirir un software específic per tal d'accedir des de casa a les dades diàries del mercat borsari. En cas de voler accedir a aquestes dades, existeixen tres tipus de plans:

1. Plans de difusió per demanda on cada petició efectuada té un import acordat.
2. Plans de difusió per retard amb una latència de 15 minuts.
3. Plansde difusió en temps real.

En el nostre cas, en adquirir les dades directament de la pàgina web, tenen una latència de 15 minuts.

El motiu principal que ens ha portat a triar la Borsa de Barcelona es la seva complexitat intermèdia, la llibertat de poder utilitzar mètodes de web scraping sense establir cap sessió d'usuari ique el control de la navegació s'obté fàcilment acceptant les "cookies" del lloc web.

2. Definició de títol pel dataset

El títol assignat al dataset ha estat "*Preus de les empreses de l'Ibex 35 amb latència de 15 minuts*" ja que reflecteix amb claredat i concisió el contingut de les dades que contindrà. No s'especifiquen els índexs ni el temps al que fa referència el dataset perquè les empreses que formen part de l'Ibex 35 poden canviar en qualsevol moment i el temps al que fan referència les dades és el que engloba els 15 minuts.

3. Descripció del dataset

Tal com suggereix el títol, les dades que formen el dataset fan referència a la cotització en borsa de les empreses més importants de l'Ibex 35. Són dades diàries de dilluns a divendres que es generen cada 15 minutsde 9:00 a 17:30.

D'aquesta manera, com les dades canvien cada 15 minuts, per fer un bon estudi de les dades de la Borsa de Barcelona que fan referència a l'Ibex 35 per a un dia concret s'ha de generar més d'un dataset, concretament n'hi haurà 30 referents als canvis d'aquests índexs durant un mateix dia (09/04/2021) i separats entre si 15 minuts. També hi ha el fitxer "finalDataset.csv", aquest, s'ha creat amb les dades dels 30 CSVs per a poder fer bones representacions gràfiques de les dades.

A data 08/04/2021 les empreses formant part de l'Ibex 35 i, per tant, les empreses de les quals tenim informació a tractar són:

```
['ACCIONA', 'ACERINOX', 'ACS', 'AENA', 'ALMIRALL', 'AMADEUS', 'ARCELORMIT.',
'B.SANTANDER', 'BA.SABADELL', 'BANKINTER', 'BBVA', 'CAIXABANK', 'CELLNEX', 'CIE AUTOMOT.',
'ENAGAS', 'ENDESA', 'FERROVIAL', 'FLUIDRA', 'GRIFOLS CL.A', 'IAG', 'IBERDROLA', 'INDITEX',
'INDRA A', 'INM.COLONIAL', 'MAPFRE', 'MELIA HOTELS', 'MERLIN', 'NATURGY', 'PHARMA MAR',
'R.E.C.', 'REPSOL', 'SIEMENS GAME', 'SOLARIA', 'TELEFONICA', 'VISCOFAN']
```

Aquest llistat d'empreses s'ha obtingut amb la funció "getIndexs()" del fitxer "ScrapBorsaBarc.ipynb" la qual serveix per obtenir els noms de les empreses de les quals s'estan adquirint dades.

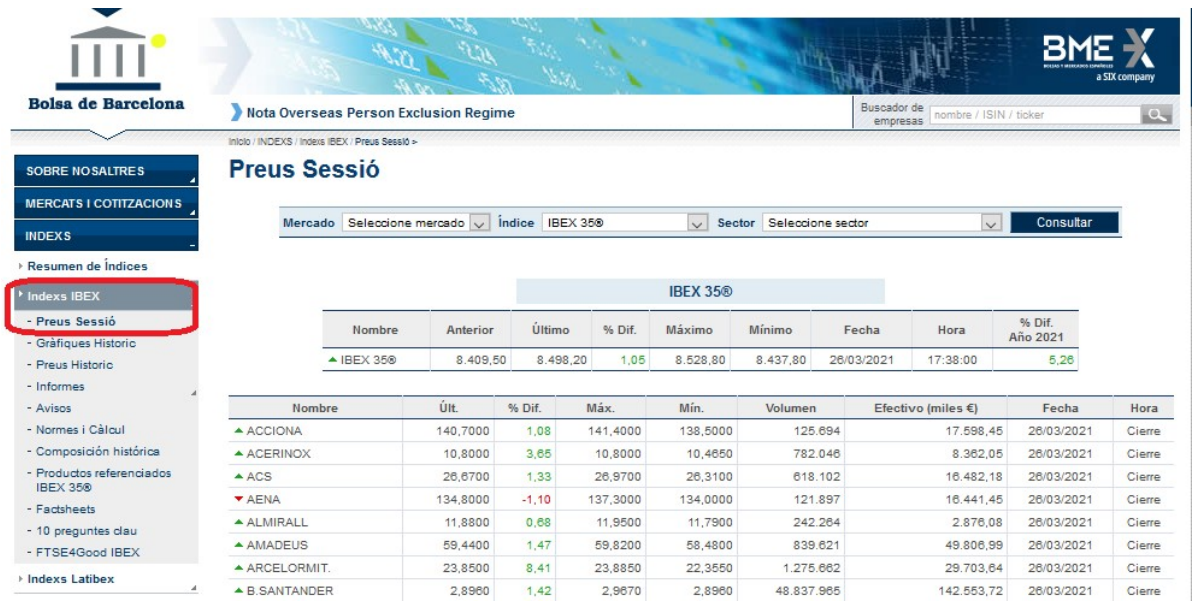
També interessa saber quin tipus d'informació s'obtindrà de cada una d'aquestes empreses, per fer-ho, s'ha generat la funció "getColNames()" la qual retorna el nom de les columnes que tindran els datasets generats:

```
['Nombre', 'Últ.', '% Dif.', 'Máx.', 'Mín.', 'Volumen', 'Efectivo (miles €)', 'Fecha', 'Hora', 'Capitalización']
```

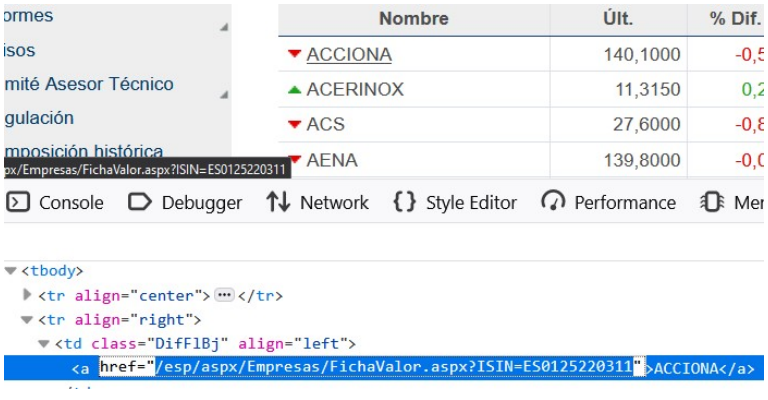
Tal i com s'observa, els datasets tindran informació de 10 atributs per a cada empresa de l'Ibex 35.

4. Representació gràfica

Les dades a recollir es troben al menú desplegable d'Índexs IBEX, a l'opció "Preus de Sessió" tal i com s'observa en la següent imatge:



Aquestes dades es complementen amb les dades de cada una de les empreses de l'Ibex, a aquesta opció s'accedeix clicant el nom de l'empresa del llistat al contenir aquest un link:



Tal i com s'observa cada una de les empreses té un link associat al seu nom que ens redirigeix a una ruta on es mostra més informació sobre l'empresa, incloent-hi la capitalització. Per tant, amb aquests links es pot extreure el valor de capitalització(en milers d'euros):

ACCIONA,S.A.

Domicilio	CL AVENIDA DE EUROPA 18, 28943 ALCOBENDAS (MADRID)
Capital Social Admitido	54.856.653,00 Euros

Ficha | Ampliaciones | Splits | Ref. Nominal | Pagos Especiales | Dividendos | Inf. Financiera | I.Priv./O.I.Relev. | Inf. Histórica | Negociaciones

ACCIONA

ISIN	ES0125220311	Ticker	ANA	Nominal	1,00	Mercado	Mercado Continuo	Capital Admitido	54.856.653,00 Euros

Últimos precios

Fecha	Hora	Cierre	Ref.	% Dif.	Últ.	Máx.	Mín.	Medio	Volumen	Efectivo
26/03/2021	Cierre	140,7000	139,2000	1,08	140,7000	141,4000	138,5000	140,0617	125.694	17.598.452,2

1 año | 5 años

Último pago

Tipo	Fecha Descuento	Fecha Abono	Concepto
Dividendo	30/06/2020	02/07/2020	Único 2019

Último split

Fecha	Equivalencia
30/11/1998	6 x 1

Última refundición de nominal (contrasplit)

Fecha	Equivalencia

Última ampliación

Periodo	Proporción	Tipo
30/11/1998	1 x 500	Par liberada

	2021 hasta el 26/03	2020	2019	2018	2017
Capitalización * *(Miles de Euros)*	7.718.331	6.401.771	5.145.554	4.231.481	3.895.94
Nº de acciones					

Per fer una representació gràfica, hem generat fitxers CSV per a un dia complet de mercat, es poden trobar els 30 fitxers a la carpeta “csv” del projecte, i els hem adjuntat en el CSV “finalDataset.csv”. Aquest CSV l’hem convertit un dataframe (combined_csv) de la llibreria pandas i n’hem extret un resum estadístic dels valors que contenen:

```
combined_csv.describe()
```

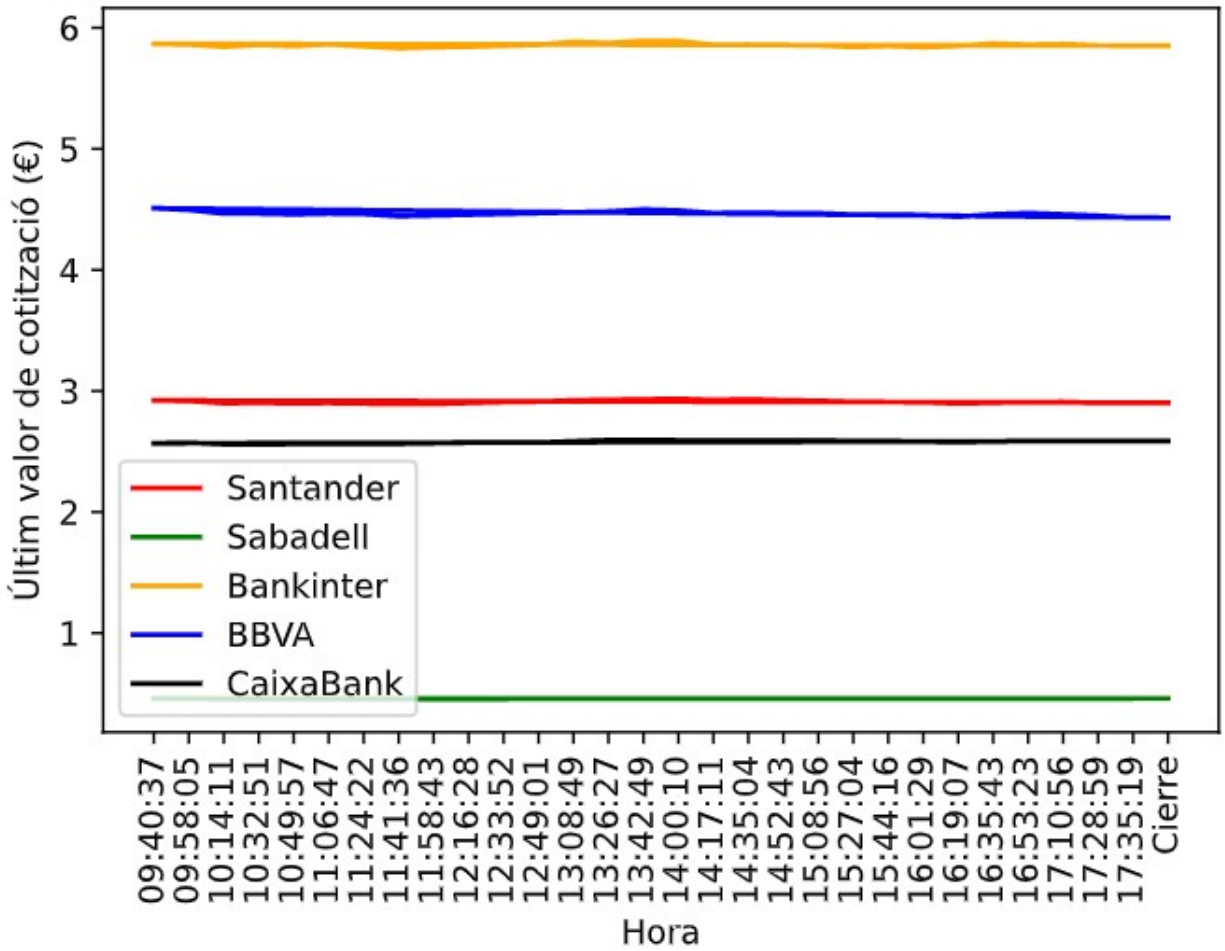
	Últ.	% Dif.	Máx.	Mín.	Efectivo (miles €)	Capitalización	Volumem
count	17850.000000	17850.000000	17850.000000	17850.000000	17850.000000	1.783300e+04	1.785000e+04
mean	26.764558	-0.629933	27.056480	26.628415	10259.870619	1.640049e+07	1.824402e+06
std	33.805975	0.985767	34.147178	33.685068	15787.391538	1.924144e+07	4.338276e+06
min	0.453400	-4.300000	0.464400	0.451200	173.970000	1.316075e+06	3.121000e+03
25%	6.512000	-1.000000	6.596000	6.500000	2191.200000	3.054467e+06	9.555900e+04
50%	17.917500	-0.570000	18.635000	17.765000	4861.500000	8.645796e+06	2.798530e+05
75%	27.600000	-0.060000	27.780000	27.540000	10890.890000	2.109991e+07	1.186616e+06
max	141.000000	2.190000	141.200000	139.600000	120117.270000	8.698576e+07	3.570338e+07

Tal i com s’observa hi ha una gran completesa de dades, de les 17850 dades extretes per a cada una de les 10 columnes només falten 17 dades per a la columna de capitalització. També es poden extreure conclusions interessants si s’analitza bé ja que tenim dades de la mitja, mediana, mínim, màxim, desviació estàndard i quartils d’interès. Si es vol continuar amb la anàlisis estadística es pot agrupar el dataset anterior pel valor de la columna Nombre per a, d’aquesta manera, tenir resums estadístics agrupats per empresa:

```
groupedDataframe = combined_csv.groupby(['Nombre'])
groupedDataframe.describe().transpose()
```

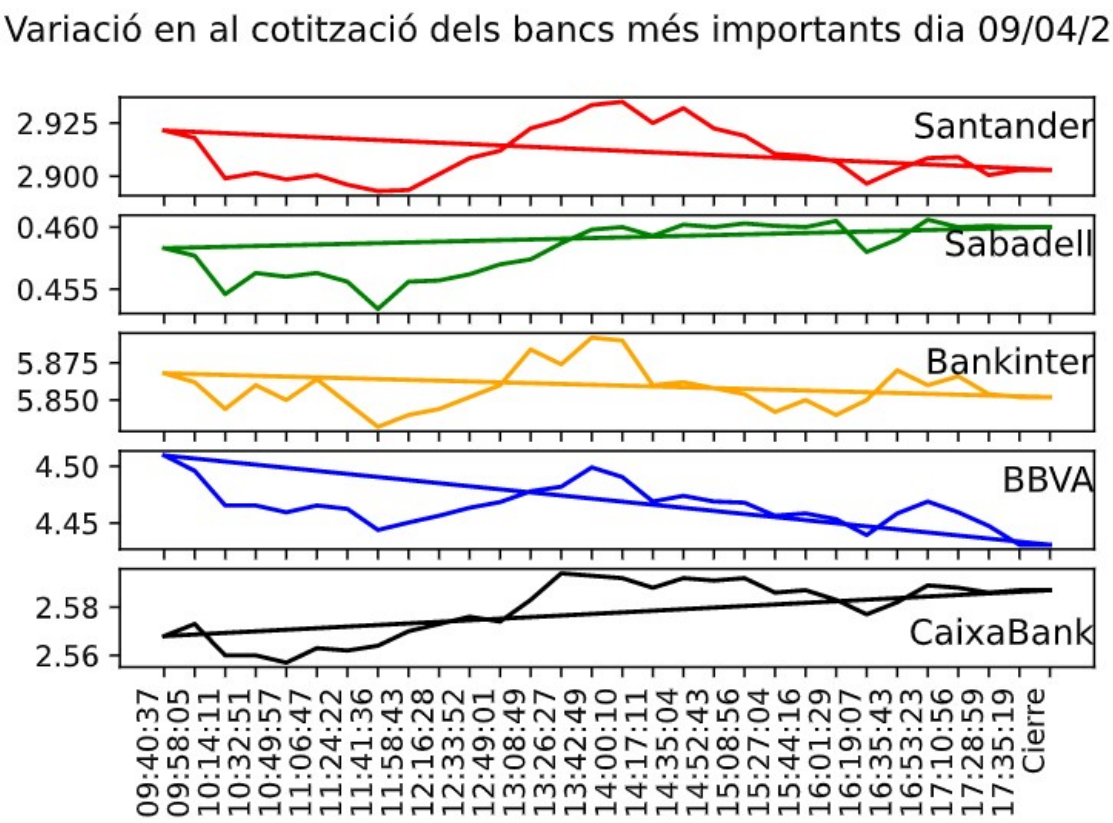
Nombre		ACCIONA	ACERINOX	ACS	AENA	ALMIRALL
Últ.	count	5.100000e+02	5.100000e+02	5.100000e+02	5.100000e+02	5.100000e+02
	mean	1.400833e+02	1.141100e+01	2.765433e+01	1.394933e+02	1.252467e+01
	std	3.628319e-01	4.089351e-02	6.807085e-02	6.361990e-01	1.912082e-02
	min	1.395000e+02	1.127500e+01	2.755000e+01	1.383000e+02	1.250000e+01
	25%	1.398000e+02	1.139500e+01	2.760000e+01	1.388000e+02	1.251000e+01
	50%	1.400000e+02	1.142000e+01	2.765000e+01	1.398000e+02	1.252000e+01
	75%	1.403000e+02	1.144000e+01	2.772000e+01	1.400000e+02	1.255000e+01
	max	1.410000e+02	1.146000e+01	2.776000e+01	1.402500e+02	1.256000e+01
% Dif.	count	5.100000e+02	5.100000e+02	5.100000e+02	5.100000e+02	5.100000e+02
	mean	-5.093333e-01	1.071333e+00	-6.316667e-01	-2.906667e-01	-6.753333e-01
	std	2.572339e-01	3.614615e-01	2.443173e-01	4.541041e-01	1.489128e-01

Finalment, deixant de banda l'anàlisi d'aquesta última taula amb els estadístics d'interès per a cada un dels atributs de cada una de les empreses que formen part de l'Ibex 35, com el dataset obtingut és molt gran i amb molta informació variada, es poden fer representacions gràfiques de mil tipus segons l'objectiu de l'anàlisi. En el nostre cas, com no és l'objectiu principal de la pràctica hem decidit comparar els principals bancs d'Espanya per veure les seves diferències pel que fa a la variació diària de la seva cotització un dia concret:



En aquest gràfic s'observa una clara diferència de cotització en els bancs més importants d'Espanya sent Bankinter el que té la taxa de cotització més elevada. A simple vista no s'observa cap mena de relació entre el valor de cotització d'un banc i l'hora del dia encara que per poder afirmar-ho amb seguretat es

necessitarien dades de molts més dies que la cotització i, a més a més, potser si que n'hi ha però a petita escala. Per comprovar que no hi ha variacions diàries fem el gràfic dels bancs per separat:



En fer aquest gràfic apilant la variació en la cotització diària dels 5 bancs un a dalt de l'altre amb la seva pròpia escala vertical si que s'observen variacions diàries on, aquestes, són bastant diferents entre els bancs. S'observa com per al dia 9 d'Abril de 2021 les cotitzacions dels bancs Sabadell i CaixaBank van anar augmentant mentre que les dels bancs Sabadell, Bankinter i BBVA van anar disminuint.

Com hem dit, amb aquest gran dataset es podrien fer mil tipus de gràfiques i inclús aplicar algun model de predicció de mineria de dades però com no és l'objectiu de l'assignatura ni de la pràctica donem per finalitzada la descripció del dataset.

5. Contingut

Per fer l'anàlisi inicial de la web el primer pas ha estat explorar la web amb el navegador web FireFox v85.0 per trobar els enllaços necessaris per tal d'anar a les taules on es troben les dades d'interès.

S'ha usat FireFox perquè permet veure de manera simultània el codi HTML i els components es que visualitzen a la pantalla;aquesta opció es troba al menú de configuració a la secció "desenvolupador web >Inspector". Encara que també es pot accedir a l'estructura HTML dels components d'interès fent clic dret sobre ells i clicant "Alt + q" o l'opció "Inspect Element".

Amb aquesta opció només cal situar-se en el punt exacte on es troben les taules per tal de trobar el codi HTML involucrat en la generació d'aquestes taules.

També s'ha utilitzat la comanda Python "print(soup.prettify())" durant el desenvolupament del codi per analitzar l'estructura "imbricada"; amb la informació que retorna aquesta comanda s'ha pogut analitzar el codi HTML dels diversos components dintre de l'estructura (taules, frames,tags imbricats, etc).

La taula que conté els índexs es troba a un pàgina que fa una càrrega dinàmica d'un frame amb una crida de tipus POST; es aquest "frame" carregat dinàmicament el que conté la taula. La pàgina inicial es carrega amb mètode GET i aquesta fa una càrrega d'un frame amb un mètode POST.

Per tal de combinar el mètode GET i POST de manera simultània s'ha usat la comanda "requests" de Python; amb l'adreça de la pàgina web s'incorporen els paràmetres de GET i amb l'atribut"soup.find" els paràmetres de POST.

Un altre repte ha estat com capturar la informació extra de les empreses, per aconseguir-ho,s'ha extret l'enllaç de cada una de les empreses de la taula de l'Ibex 35 i s'ha utilitzat per accedir a la taula particular de cada una d'elles i, un cop allí, s'ha tractat aquesta taula per extreure el valor de la seva capitalització.

En fer aquesta petició de la capitalització, com es fa per cada una de les empreses de les que s'està extraient informació, es pot generar un problema de peticions al servidor fent que pugui arribar a saturar-se. Per tal d'evitar aquest problema de saturació hem incorporat un retard de temps entre cada lectura proporcional al temps que el servidor triga en tornar una resposta.

De cara al futur pot ser interessant que l'accés a la web sembli feta per un navegador normal; es per això que a les peticions s'ha incorporat el "header" que envia el navegador FireFox. També cal dir que totes les excepcions que es produeixen en fer peticions de càrrega de les webs estan controlades, al igual que els temps de respostes els quals incorporen un "timeout" de 10 segons per evitar esperes molt llargues.

Finalment, els camps recollits per cada una de les empreses són:

a) **Nombre**: Nom de l'empresa.
b) **Últ.**: Darrera dada de cotització.
c) **%Dif.**: Variació del seu valor actual respecte al darrer tancament de la borsa.
d) **Máx.**: Valor màxim assolit.
e) **Mín.**: valor mínim assolit.
f) **Volumen**: Quantitat de títols negociats.
g) **Efectivo(miles€)**: Valor efectiu dels títols negociats en milers de euros.
h) **Fecha**: Data de la cotització.
i) **Hora**: Hora de la cotització.
j) **Capitalización**: Capitalització de l'empresa en milers d'euros.

El període de temps de les dades estan enregistrat als camps "Fecha" i "Hora" on, quan el valor del camp "Hora" pren el valor "Cierre" ja s'ha tancat la borsa (més tard de les 17:30h).

Tal i com hem descrit en apartats anteriors tenim un total de 31 CSVs però el format del dataset d'interès és el que s'extreu del fitxer CSV "finalDataset.csv" que ja ah estat creat per a que tingui una fàcil visualització i tractament.

Finalment, també caldestacar que a la web no existeix cap fitxer anomenat "robots.txt"; en intentar accedir a l'enllaç "https://www.borsabcn.es/robots.txt" no s'obté cap resultat (404 + redirecció interna).

6. Agraïments

Les borses Espanyoles pertanyen al grup SIX Group AG (SIX), aquest, ofereix serveis de productes i sistemes avançats de negociació i accés a mercats globals a emissors, intermediaris i inversors de dins i fora d'Espanya.

Aquest grup té els drets adquirits d'aquestes dades del mercat de valors a Espanya. Les dades adquirides les generen els inversors i empresesque, després, les posen a disposició d'aquestes entitats. Les dades introduïdes són processades i permeten realitzar l'operativa diària de la borsa.

Els nostres agraïments serien per aquest grup que posa a disposició de la gent les dades diàries de les sessions de les borses a Espanya.

7. Inspiració

L'anàlisi de les dades de la borsa amb nous tractaments de IA es un terreny molt nou; actualment n'hi ha molts algoritmes que es dediquen a fer compres i vendes. Degut al seu creixent interès i a la forta competència que hi ha,aquest procés s'ha anant automatitzant en bona part. Malauradament les dades estan restringides a un públic minoritari molt especialitzat.

Amb aquestes dades es pretén respondre a les següents preguntes:

- Quin es el grau de correlació en el comportament de les empreses de l'Ibex 35?

- Es pot determinar el comportament del valor d'una empresa si d'altres varien?.
- Es poden trobar mes variables predictives?.
- Quin es el grau màxim de predicció que es pot fer?
- Com poden estar relacionades les diferents borses locals i les que es troben a l'estranger?
- N'hi ha algú que controla les borses?
- Es poden detectar algoritmes de pressa de decisions a les borses?.

Aquestes preguntes tenen difícil resposta degut a la gran quantitat de variables i factors que els afecten, per tant, amb aquestes dades inicialment es pretén analitzar la completesa de les dades, les seves tendències i, crear una base per arribar a saber com estan correlacionades les diferents empreses a la borsa a l'Ibex per arribar a explicar com el comportament d'una afecta a la resta i, en els millors dels escenaris, predir aquest comportament. D'aquesta manera, donat que al sector borsari hi ha moltes més variables, aquest dataset seria un punt de partida per trobar totes aquestes variables que poden afectar.

A Internet hi ha exemples d'extraccions de dades d'altres borses que no són de la de Barcelona però no aporten gaire informació; en aquests exemples falta la complementació les dades de l'índex amb dades pròpies de les empreses.

Amb aquest WebScraping, a banda de la informació que es presenta a la taula de l'Ibex 35 de la Borsa de Barcelona,s'extreuen dades interessants de les empreses de la taula utilitzant els links extrets de cada una de les empreses per obtenir el portal web d'on extreure aquesta informació (columna Capitalización).

8. Llicència

"CreativeCommons" es una organització sense ànim de lucre que ajuda a superar obstacles derivats de la compartició de coneixement i creativitat per abordar els desafiaments del mon.De les llicències que tenen publicades la que mes s'adaptaria al dataset generat seria la"ReleasedUnder CC0: PublicDomainLicense"; és la menys restrictiva en permetre que altres puguin barrejar, adaptar i construir nous datasets amb qualsevol finalitat sempre que les noves creacions tinguin llicencies amb les mateixes condicions.

Les dades i l'ús que es pugui fer d'aquest dataset són propietat de les entitats borsàries; si aquestes les fan públiques ningú té cap dret sobre aquestes.

9. Publicació a Zenodo

Aquest dataset s'ha publicat en format CSV a Zenodo; el seu identificador DOI es el següent:

http://doi.org/10.5281/zenodo.4681539

Contribucions	Signa
Recerca prèvia	Sergio García Pérez
Redacció de les respostes	Sergio García Pérez, Andreu Fornós
Desenvolupament del Codi	Sergio García Pérez, Andreu Fornós