***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.

Les dades recol·lectades s’ha recol·lectat en el sector financer. La borsa es una entitat privada i l’accés a les dades d’aquesta està restringit a brokers i mitjans de comunicació que tenen acords amb la borsa.

La gent que es dedica a la borsa normalment ha d’adquirir un software específic per tal d’accedir-hi des de casa a les dades diàries del mercat borsari.

Existeixen plans de difusió per demanda (preu per cada petició), difusió per retard amb una latència de 15 minuts i finalment la difusió en temps real.

Les dades que es mostren a la web de la borsa de Barcelona tenen una latència de 15 minuts.

La raó de triar la Borsa de Barcelona es que no es una web complexa, no es requereix establir cap sessió (sense identificació d’usuari) i només requereix “cookies” pel control de la navegació.

1. Definició de títol pel dataset.

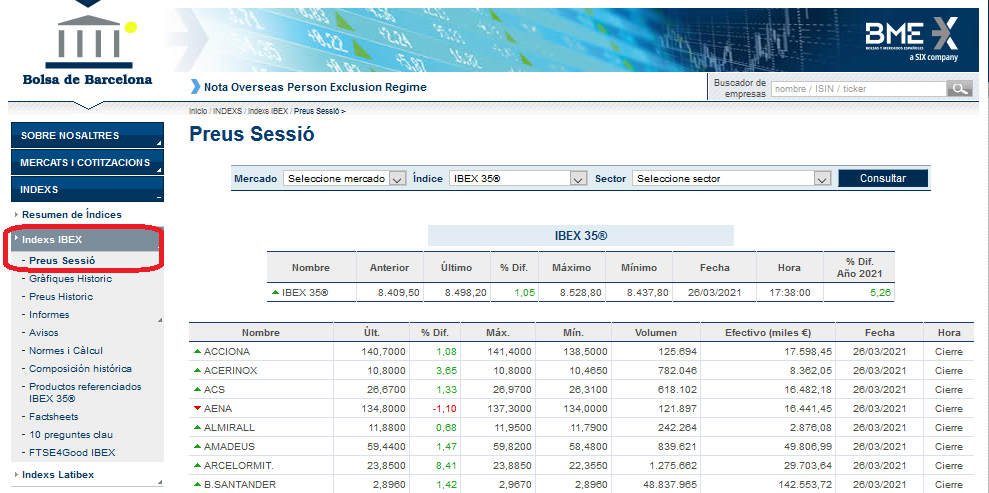
El títol del dataset seria “Indexs del Ibex amb latència de 15 minuts”; aquest reflecteix el contingut de l’ informació que es vol extreure.

1. Descripció del dataset.

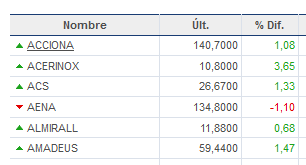
Les dades que tenim en aquest data set fan referència a la cotització en borsa de les empreses mes importants de l’Ibex 35. Les dades que es llegeixen son diàries de dilluns a divendres en horari de borsa i es fan cada 15 minuts.

1. Representació gràfica.

Les dades a recollir es troben al menú desplegable de Índexs IBEX, a l’opció “Preus de Sessió” segons es pot veure al dibuix:



Aquestes dades es complementen am les dades de les empreses de l’Ibex, a aquesta opció s’accedeix a través de link que n’hi ha a les empreses de l’anterior llistat.



I la dada complementaria que s’extreu es la capitalització(en milers d’euros):



1. Contingut. Explicació dels camps que inclou el dataset, del període de temps de les dades i de com s'han recollit.

Els camps que es recullen son els següents:

1. “Nombre” : nom de l’empresa
2. “Últ.”: utima cotització.
3. “%Dif”: variació que ha tingut respecte al darrer tancament de la borsa.
4. “Màx”: valor màxim que ha obtingut.
5. “Min”: valor mínim que ha obtingut.
6. “Volumen” : quantitat de títols negociats.
7. “Efectivo(miles€)”: valor efectiu dels títols negociats en milers de euros.
8. “Fecha”: Data de la cotització.
9. “Hora”: Hora de la cotització.
10. “Capitalitzacion” : aquest camp dona informació de la capitalització de l’empresa en milers d’euros.

El període de temps de les dades estan enregistrat als camps “Fecha” i “hora”.

El darrer camp pot prendre el valor “Cierre”; això vol dir que es correspon a les 17:30 h de la tarda (hora tancament de la borsa).

Les dades van canviat a lo llarg del dia i hora durant la sessió de la borsa. En aquests cas s’ha agafat com a exemple una hora aleatòria del dia.

1. 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 a Espanya i fora d’Espanya.

En principi aquest grup té els drets adquirits d’aquestes dades del mercat de valors a Espanya; aquestes dades les generen els inversors i les empreses i les posen a disposició d’aquestes entitats. Les dades introduïdes es processen i permeten realitzar l’operativa diària de la borsa.

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

A la web n’hi han extraccions d’altres borses que no son de la de Barcelona i no aporten gaire informació; falta complementar les dades amb dades de les empreses que componen l’Ibex .

1. 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. Aquest procés s’ha anant automatitzant en bona part; tenim que n’hi ha una forta competició. Malauradament les dades estan restringides a un públic minoritari molt especialitzat.

[És necessari comparar amb les anàlisis anteriors presentades a l’apartat 6.]

Aquest webscraping a banda de l’informació de l’Ibex extreu dades interessantes de les empreses de l’Ibex. Aquesta informació es troba als links de les empreses.

1. Llicència.

“Creative Commons” 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 meu dataset seria la

“Released Under CC0: Public Domain License”; aquesta es la menys restrictiva.

Permet que d’altres puguin barrejar, adaptar i construir nous datsets 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” son propietat de les entitats borsàries; si aquestes les fan públiques ningú té cap dret sobre aquestes.

1. Publicació a Zenodo:

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

|  |  |
| --- | --- |
| Contribucions | Signa |
| Recerca prèvia | Sergio García Pérez |
| Redacció de les respostes | Sergio García Pérez |
| |  | | --- | | Desenvolupament del Codi | |  | | Sergio García Pérez |