***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’han 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), plans de difusió per retard amb una latència de 15 minuts i finalment plans de 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.

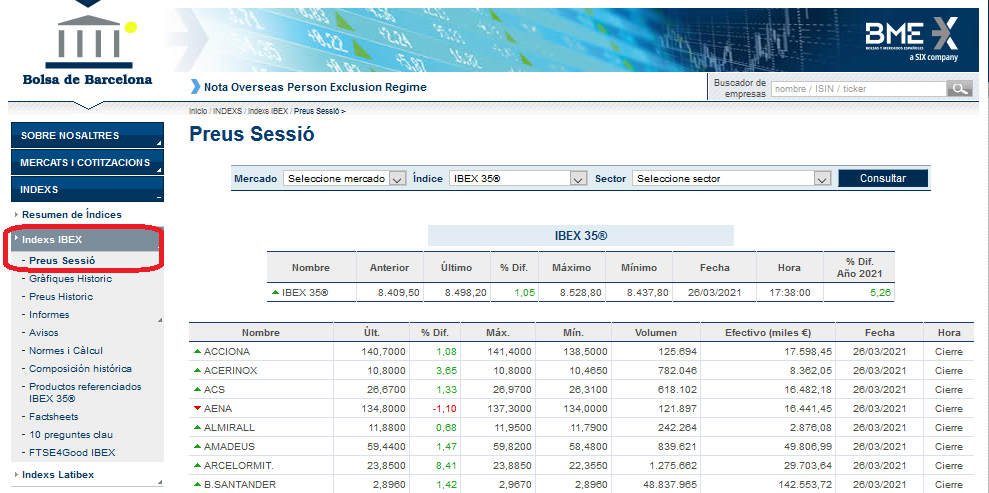
El títol del dataset seria “Indexs de l’Ibex amb latència de 15 minuts”; doncs aquest reflecteix el contingut de l’ informació que es vol extreure.

1. Descripció del dataset.

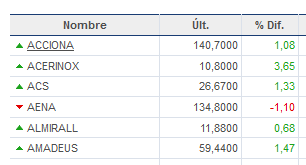
Les dades que tenim en aquest data set fan referència a la cotització en borsa de les empreses mes importants de l’Ibex 35. Les dades que es llegeixen son diàries de dilluns a divendres en horari de borsa i es fan cada 15 minuts.

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):



Fer una representació gràfica

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

Per fer l’anàlisi inicial de la web el primer que he fet es explorar la web amb el navegador web “FireFox” (versió 85.0). Amb aquest navegador he trobat els enllaços necessaris per tal d’anar a les taules on es trobaven les dades que m’interessaven.

Aquest navegador té una opció molt interessant que 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”.

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.

Altre mètode que he fet servir per analitzar l’estructura “imbricada” ha sigut executar la comanda Python “print(soup.prettify())” durant el desenvolupament del codi; amb aquest he pogut analitzar el codi HTML dels diversos components dintre del codi (taules, frames,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 poder combinar el mètode GET i POST de manera simultània he hagut de combinar tots dos mètodes amb la comanda “requests” de Python; amb l’adreça de la pàgina web s’incorporen els paràmetres de GET i amb el atribut de “soup.find” els paràmetres de POST.

Un altre qüestió que m’he plantejat ha sigut com capturar l’ informació extra de les empreses; per això m’he anat a l’enllaç d’aquestes que es troba a la primera columna de la taula.

Quan llegeixo les dades de la taula principal d’índexs també haig d’accedir a les dades de les empreses; això pot comportar un problema de peticions al servidor i el pot saturar. Per tal d’evitar aquest problema de saturació incorporo un retard de temps entre cada lectura que es 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 he incorporat a les peticions el “header” que envia el navegador “FireFox”.

Totes les excepcions que es produeixen en fer peticions de càrrega de les webs estan controlades i els temps de respostes també amb un “timeout” de 10 segons (per evitar esperes molt llargues).

Els camps que es recullen finalment son els següents:

1. “Nombre” : nom de l’empresa.
2. “Últ.”: darrera 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 canviant 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.

Haig de destacar també que a la web no existeix cap fitxer anomenat “robots.txt”; si poso al navegador web “FireFox” el següent l’enllaç “<https://www.borsabcn.es/robots.txt>” no surt res.

El dataset que es genera recull els índexs de les 36 de les primeres 36 empreses de l’Ibex 35.

En aquest dataset es presenta els valors de les cotitzacions(en euros) de les accions; també surten re reflectits els màxims, els mínims, la quantitats de títols venuts/comprats, la quantitat de negoci que s’ha fet, la data i l’hora de les dades.

Les magnituds dels camps son euros,milers d’euros, dia i hora. Excepte el camps relacionats amb l’empresa, el valor de la cotització i la data/hora la resta de camps son calculats.

Aquestes dades no tenen inconsistències i el format d’aquestes es el adequat amb una excepció:

el camp de l’hora pot contenir el text “Cierre”; aquest valor s’ha de substituir per l’hora del tancament de la borsa.

Aquest fitxer pràcticament es pot fer servir per ser processats una vegada s’han extret els camps calculats i s’han corregits els valors incorrectes del camp relacionat amb l’hora.

El format del dataset és un fitxer CSV que facilita la seva visualització i tractament.

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.

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?.

Amb aquestes dades inicialment es pretén analitzar com estan correlacionades les diferents empreses a la borsa a l’Ibex, es pretén explicar com el comportament d’una afecta a la resta i si es pot predir aquest comportament. Donat que al sector borsari n’hi han moltes mes variables aquest seria un punt de partida per trobar totes aquestes variables que poden afectar.

En el meu cas aquesta extracció es un punt de partida

A Internet n’hi han exemples d’extraccions a d’altres borses que no son de la de Barcelona però no aporten gaire informació; en aquests exemples falta complementar les dades de l’índex amb dades pròpies de les empreses.

Aquest webscraping a banda de l’informació de l’Ibex extreu dades interessants 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 |