PR1 Tipologia i cicle de vida de les dades

Autors: Francesc Ballester Lecina i Oriol Raurell Gan

1. Context. Explicar en quin context s'ha recollit la informació. Explicar per què el lloc web triat proporciona aquesta informació.

Primer de tot, s'han proposat diferents webs per a la recollida d'informació (https://www.transfermarkt.es/ , https://datosmacro.expansion.com/otros/coronavirus , https://www.idealista.com , https://coinmarketcap.com) . Finalment, ens hem decantat per un tema d'actualitat com són les criptomonedes, ja que actualment es un dels principals actius d'inversió i ens seria de gran utilitat per conèixer i analitzar l'evolució de les principals criptomonedes existents en el mercat d'inversió de divises alternatives.

2. Títol. Definir un títol que sigui descriptiu pel dataset.

Top20 criptomonedes amb més valor de mercat

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 escollit.

El conjunt de dades generat un cop fet el web scraping, conté informació diària relativa a les 20 principals criptomonedes del mercat. En el conjunt de dades obtingut podem trobar informació sobre el nom de la criptomoneda, el seu símbol, el valor de capitalització de mercat, el preu actual de la moneda (en dòlars), el nombre actual de criptomonedes que hi ha en circulació d'aquella criptomoneda, el volum de dolares tractat en les últimes 24 hores, la variació del preu de la criptomoneda per hora, dia i setmana.

4. Representació gràfica. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

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

La pàgina web https://coinmarketcap.com es la que ens proporciona la informació del nostre conjunt de dades. El conjunt de dades conté informació diària relativa a les 20 principals criptomonedes del mercat. Les dades són actualitzades del moment en que s'extreuen.

Els atributs que apareixen en el conjunt de dades són:

- **Nom:** Nom de la criptomoneda.
- **Símbol**: Símbol de la criptomoneda.
- **Cap. de mercat** : Valor de capitalització de mercat.
- **Preu** : Preu de la criptomoneda en dòlars.
- **En circulació** : Nombre de criptomonedes que hi ha en circulació.
- **Volum(24h)** : Volum de dòlars tractat en les 24h.
- **%1h** : Variació del preu de la criptomoneda en 1h.
- **%24h** : Variació del preu de la criptomoneda en 24h.
- **%7d** : Variació del preu de la criptomoneda respecte la setmana anterior.
- **Dia**: Dia de la extracció.
- **Hora**: Hora de la extracció.

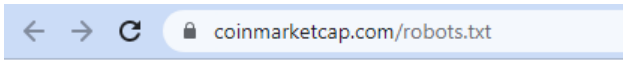
6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-n'hi, justificar aquesta cerca amb anàlisis similars. Justificar quins passos s'han seguit per actuar d'acord amb els principis ètics i legals en el context del projecte.

Pel que fa als principis ètics i legals s'han seguit des del primer moment, revisant les pàgines de termes i condicions generals i els arxius robot.txt de cada lloc. Inicialment la proposta més motivadora va ser la pàgina idealista, però després de revisar les condicions ens va fer retractar davant les dificultats que s'exposen a continuació:

> Acceder, controlar o copiar cualquier información incluida en esta Web y apps utilizando para ello cualquier tipo de robot, spider, scraper u otro medio automático o proceso manual para cualquier propósito, sin nuestro permiso expreso y por escrito.
>
> En concreto, no está permitido revender, realizar deep-links, utilizar, copiar, monitorizar (por ejemplo, spider, scrape), mostrar, descargar, guardar o reproducir el contenido, la información, el software, los productos o los servicios disponibles en nuestro sitio web para cualquier actividad comercial o competitiva sin autorización previa y por escrito por nuestra parte.

https://www.idealista.com/ayuda/articulos/terminos-y-condiciones-generales-de-idealista/

En canvi, la pàgina web escollida (www.coinmarketcap.com) en el seu arxiu robots.txt indicava l'exclusió únicament d'una sèrie de directoris sense afectar a les taules de les pàgines principals. D'aquesta manera, els propietaris del lloc web permeten accedir a la resta de directoris mitjançant rastrejadors i per tant es pot dur a terme la pràctica complint amb el codi ètic i legal sense problemes:




```
User-agent: *
Allow: /

Disallow: */currencies/*/social/$
Disallow: */currencies/*/onchain-analysis/$
Disallow: */currencies/*/ratings/$
Disallow: */currencies/*/price-estimates/$

Disallow: /*/headlines/*$
```

https://coinmarketcap.com/robots.txt

Per la creació del codi s'ha buscat en la xarxa diferents anàlisis anteriors i resolució de dubtes de web scraping com ara:

https://tommycc.medium.com/web-scraping-crypto-prices-with-python-41072ea5b5bf

https://tcoil.info/coinmarketcap-python-web-scraper/

https://stackoverflow.com/questions/49192522/web-scraping-coinmarketcap-com-with-python-requests-beautifulsoup

I mitjançant el manual de la UOC: Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

Per tant, s'ha rastrejat una informació pública sense extreure informació confidencial, privada o que el autor no desitja. Les dades extretes no causen cap dany i no seran utilitzades amb fins comercials sinó mes aviat amb un objectiu educatiu per part dels autors del robot.

Encara que existeix un API del lloc web, s'ha decidit no accedir degut a la manca d'històrics de dades i els preus i usos exigits pel propietari que fa més senzill realitzar la practica utilitzant el codi html.

Mitjançant el programa Colaboratory de Google s'ha pogut treballar des de qualsevol ordinador i provar amb diferents user agents, gràcies a que no cal instal·lar cap programari addicional, sinó només mitjançant la conta educativa de la UOC i Google Drive.

Pel que fa la qualitat i robustesa de les dades podem afirmar que encara que no tenen una completesa màxima, si que permeten obtenir les principals 20 criptomonedes i extreure conclusions de com es troba el mercat. Tampoc seran úniques, ja que diferents web reporten els mateixos valors i estem parlant de valors d'intercanvi internacionals. Les dades tindran validesa, puntualitat, exactitud i consistència sempre i quan s'executi de nou l'script i s'obtinguin les dades actualitzades al moment.

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l'apartat 6.

El 'boom' de les criptomonedes ha fet que cada cop més la societat tingui més interès en elles pel seu potencial gràcies a la tecnologia blockchain i per inversió, ja que es poden obtenir grans rendibilitats.

Mitjançant el conjunt de dades es pot obtenir informació actualitzada dels diferents preus de les criptomonedes, la seva evolució al llarg de la setmana, així com les 20 principals monedes del mercat. En aquest cas, podem saber quina ha sigut la criptomoneda que el seu valor ha crescut més en l'última setmana, quina moneda ha mogut més volum de dòlars en les transaccions, quina són les 10 millors criptomonedes, quina te millor capitalització entre d'altres.

Aquesta informació pot ser de gran utilitat per aquelles persones que volen invertir o bé per aquelles que vulguin obtenir informació sobre la situació actual de les principals criptomonedes.

8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

El tipus de llicencia escollit pel dataset a l'hora de publicar-se a Zenodo ha sigut la llicencia del tipus CC BY-SA 4.0 License bàsicament per les millors condicions i ajuda que permet aplicar aquest tipus de llicencia al dataset.

El tipus de llicencia CC BY-SA 4.0 License permet compartir i redistribuir, adaptar, transformar.. les dades existents en el dataset sigui quina sigui la seva finalitat, ja pot ser personal, educativa o comercial.

Les úniques condicions que exigeix són el reconeixement dels autors, proporcionar un enllaç a la llicencia i indicar si s'han realitzat canvis. Si es modifiquen les dades o s'utilitzen per crear un material nou s'haurà de comunicar les contribucions sota la llicencia original. No existiran mesures addicionals que restringeixin realitzar el que pel tipus de llicencia es permet.

Així que aquests tipus de llicencia partint d'un projecte educatiu sembla la millor opció per compartir el dataset obtingut.

9. Codi. Adjuntar al repositori Git el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

Per a l'extracció del conjunt de dades s'ha implementat una funció en Python que ens permet extreure la informació relativa a les criptomonedes de la web Cryptocurrency Market Capitalitzacions. El codi font de la implementació el podem trobar en el següent repositori de GitHub:

https://github.com/fballesterl/PR1-Tipologia-de-dades

10. Dataset. Publicar el dataset obtingut(*) en format CSV a Zenodo amb una breu descripció. Obtenir i adjuntar l'enllaç del DOI.

Enllaç del DOI: **10.5281/zenodo.5651172**
url d'accés: **https://doi.org/10.5281/zenodo.5651172**

TAULA DE CONTRIBUCIONS

Contribucions	Signatura
Investigació prèvia	**FBL, ORG**
Redacció de les respostes	**FBL, ORG**
Desenvolupament del codi	**FBL, ORG**