M2.951 Tipologia i cicle de vida de les dades

Nuria Aguilera Sánchez
Joan Antoni Reina i Romero

PRA1: WEB SCRAPING

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

Tant des de la vessant d'ésser un instrument de pagament com la d'un valor per invertir i obtenir una rendibilitat, el mercat de les monedes digitals es cada cop més important. I tant pel que fa a les transaccions comercials realitzades amb aquestes (cada vegada més freqüents entre empreses i particulars), com pel volum que es negocia en el mercat de criptodivises.

Es per això que l'interès per saber què es i com s'opera en aquest mercat es cada vegada més manifest entre el públic en general. D'aquí l'interès en conèixer més en detall aquesta possibilitat d'obtenir un guany econòmic pels inversors. De forma que resulta interessant conèixer les principals dades: volums negociats, rendibilitats, divises més atractives, capitalitzacions, etc.

Una de les moltes pàgines *web* especialitzada en reportar informació actual sobre les divises digitals en el mercat, on ofereixen informació online i històrics de les cotitzacions de les criptomonedes, es la que hem triat pel nostre treball:

https://coinmarketcap.com/.

On l'elecció es basa per una banda en què es una de les fonts més confiables per trobar dades estadístiques sobre criptodivises en circulació, a més es una de les més visitades pels fanàtics i inversors en el terreny de les criptomonedes digitals. On accedeixen per trobar el *ranking*, els seus preus i les capitalitzacions de mercat. Aquesta *web* recopila les dades reportades per les plataformes d'intercanvi més important de tot el món i les mostra de forma molt senzilla als inversors.

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

El títol triat per al data set es:

Comparativa de l'evolució de les dades històriques diàries del ranking de les 100 criptomonedes amb més capitalització de mercat acumulat.

Amb aquest títol es vol fer referencia a l'objecte que s'ha triat, les criptomonedes, i a la dimensió temporal de les dades que arrepleguem. On cada vegada que s'executi la generació del fitxer, s'obtindrà la informació de l'últim dia de tancament.

Només se seleccionaran les primeres 100 criptomonedes considerades les més importants segons la seva capitalització de mercat acumulat.

Cada dia es crearà un fitxer amb la informació del dia anterior tancat, de forma que amb aquests fitxers es podrà fer una comparativa de l'evolució temporal. Si només n'escollim un, es veurà la comparativa per un dia de les diferents criptomonedes.

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).

Amb la finalitat de poder fer una comparativa temporal de l'evolució de les diferents criptomonedes que ajudaran en la inversió, s'ha considerat que una informació molt interessant es la de conèixer les dades de les primeres 100 criptomonedes del *ranking* de *CoinMarketCap*. El *web* obté aquesta classificació a partir de la capitalització de mercat acumulat per cada criptomoneda, la qual cosa dona una visió més objectiva de la importància de cadascuna en termes de preu i adopció.

La informació seleccionada pel *dataset* correspon a les dades de tancament del mercat del dia anterior. I permet tenir informació per identificar cada criptomoneda, el seu preu d'obertura de mercat, el valor més alt del dia, el valor més baix, el valor de tancament, el volum negociat en un dia i la seva capitalització de mercat.

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

El *dataset* que s'ha creat està representat pel següent pictograma, que fa referència a les diferents criptomonedes i a la informació de cadascuna d'elles que permetran realitzar anàlisis i estudis per obtenir beneficis.

Per transmetre aquesta idea, s'ha recorregut a una imatge on apareixen diferents criptomonedes amb dades, percentatges d'increments positius o negatius dels seus valors i diferents símbols que representen l'anàlisi d'aquesta informació. Destacar que

al mig de la imatge mostra la criptomoneda més important i més coneguda per tothom, el ***Bitcoin***.

Com a segona més important hi es ***Ethereum***, que també es mostra a la imatge. La tercera a destacar correspondria al projecte ***Ripple*** amb la seva criptomoneda ***XRP***, pensada per servir de mitjà de pagament i ser més ràpida que ***Bitcoin*** i ***Ethereum*** en aquest sentit.

La quarta que destaca es ***Litecoin***, que es com el germà petit de ***Bitcoin***, ja que funciona de forma similar però pensat per ser un complement lleuger.

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

El *dataset* proporcionat consta de 100 files de informació més una de capçalera i 7 columnes per cada una de les variables amb la informació relativa a cada criptomoneda. Totes les dades de cada fitxer corresponen a informació d'un únic dia. La informació del dia que es tracta forma part del nom del fitxer amb format *AnyMesDia*, per exemple, ***'Criptomonedes_Historical_Data_20210325.csv'***

- **Coin**: Nom identificador de la criptomoneda
- **Open**: Preu d'obertura de mercat de la criptomoneda, expressat en dòlars nord-americans ($)
- **High**: Valor més alt del dia de la criptomoneda, expressat en dòlars nord-americans ($)
- **Low**: Valor més baix del dia per de la criptomoneda, expressat en dòlars nord-americans ($)
- **Close**: Valor de tancament de la criptomoneda, expressat en dòlars nord-americans ($)
- **Volum**: Correspon al volum negociat en un dia de la criptomoneda, expressat en dòlars nord-americans ($)
- **Market_cap**[1]: Correspon a la capitalització de mercat de la criptomoneda (valor total de mercat d'una moneda), expressat en dòlars nord-americans ($)

A l'executar-lo genera la informació relativa als camps descrits només per al dia anterior i per les cent primeres monedes del *ranking* de *CointMarketCap*, que corresponen a les més importants segons el seu valor de mercat, ja que es troben ordenades per la variable ***Market Cap*** (o el valor total del mercat).

Les dades utilitzades en la creació del *dataset* no s'hi troben disposades de la forma en què apareixen al *web*. Sinó que s'han hagut d'agafar des de la pàgina *Historical Data*, on s'ha accedit prement cadascuna de les criptomonedes des de la pàgina principal. Amb la qual cosa, s'ha d'accedir moneda a moneda agafant la informació relativa al dia anterior. Que es la primera línia de la taula d'aquesta pàgina.

Per exemple, per la primera divisa, ***Bitcoin***, la informació que es pren es la que apareix a la primera línia de la taula d'aquesta adreça:

https://*CointMarketCap*.com/currencies/bitcoin/historical-data/

[1] Aquest valor es el resultat de multiplicar el preu de la moneda per la seva oferta circulant:
Market Cap = Current Price x Circulating Supply

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.

Les dades han sigut recollides des de la base de dades online de *CointMarketCap*. Per això, s'ha fet ús del llenguatge de programació *Python* així com de les tècniques de *Web Scraping* per extreure la informació allotjada en les diferents pàgines.

S'ha d'agrair el *web CointMarketCap* per la seva dedicació en la tasca de recollir la informació, integrar les dades i fer-les accessibles al públic en general.

S'han trobat diferents anàlisis anteriors referents a aquesta *web* que han servit de base per la realització d'aquest treball. Es el cas d'aquestes referències que hem considerat més destacades:

- https://medium.com/coinmonks/downloading-historical-data-from-*CointMarketCap*-41a2b0111baf. Aquí es comprova que l'accés a les dades històriques via API no aporta gratuïtament la informació requerida. A banda de veure aspectes importants com: accedir a una taula de dades històriques per una única moneda, convertir una llista en un *dataframe* i netejar les dades obtingudes.

- https://morioh.com/p/429f9d8ec035. Aquí també mostren com obtenir dades històriques però de dues formes diferents. Primer amb la llibreria *BeautifulSoup* agafant les dades de la pàgina *Historical Data* columna a columna. I en segon lloc fent ús de la llibreria *Selenium* i la funció *xpath*.

- https://towardsdatascience.com/web-scraping-crypto-prices-with-python-41072ea5b5bf. En aquesta referència s'observa com construir una llista de monedes (des de la pàgina principal) per obtenir dades històriques (en la pàgina *Historical Data*). La qual cosa es fa mitjançant una funció iterativa a l'adreça per accedir-hi a la pàgina de dades històriques de cadascuna de les criptomonedes. També es mostra la utilització de les llibreries *json* i *time*.

- https://towardsdatascience.com/scrape-tabular-data-with-python-b1dd1aeadfad. Tot i que aquesta referència no fa referència a *CointMarketCap*, es interessant perquè permet fer *scraping* de *datasets* tabulars, es a dir els elements que tenen etiqueta *table*. Per a la qual cosa es fa ús de *Selenium* i del *chromedriver*.

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.

Avui dia no es pot negar que les criptomonedes s'estan tornant cada vegada més populars i acceptades, i que tenen un gran potencial al món de les finances i la

tecnologia. Havent usuaris que les fa servir a diari per realitzar pagaments, transferir diners, invertir en criptomònades, activitats humanitàries i d'ajuda al desenvolupament, i programes de préstecs. A mesura que l'adopció de les monedes digitals augmenti també evolucionaran les seves funcions, expandint-se i arribant a altres àmbits del sector públic i privat.

Un pas fonamental per les criptomonedes es passar de ser un actiu només digital al seu ús com a moneda real i poder fer-les servir com a mètode de cobrament i pagament.

Un exemple es *Eurocoinpay*, una *startup* espanyola que disposa d'una plataforma de pagament *online*, *app* inclosa, per cobrar i pagar en comerços (la compra del supermercat, el compte del restaurant, etc).

El futur de la majoria de les criptomonedes es incert, ja que no compten amb l'acceptació esperada, però altres es podran fer servir a gran escala, gracies a que cada dia es desenvolupen tecnologies que permeten incloure-les en el món quotidià, industrial i/o financer.

Es important indicar que el *dataset* podria ser de gran utilitat en el camp de la mineria de dades, per elaborar models predictius que permetin prendre decisions sobre quines son les millors criptomonedes on invertir, també l'import o quantitat que seria adequat i quin moment pot resultar més convenient.

Per la realització d'aquest treball s'ha triat el *web* https://*CointMarketCap*.com/ per diferents raons. En primer lloc per les prestacions que ofereix, com la facilitat per accedir-hi (no cal enregistrar-se) i les escasses dificultats per realitzar *web scraping*, també per la gran quantitat d'informació que conté (incloses les dades històriques), la ràpida actualització, i per la confiança i seguretat que dona que aquesta *web* tingui més de cent milions de visites mensuals.

Si comparem aquest treball respecte als que hem fet referència a l'apartat anterior, s'observa que l'objecte del *scraping* també son les dades històriques de les diferents criptomonedes (una necessitat recurrent a l'hora d'elaborar anàlisi d'inversions).

Però aquí s'ha tractat de donar un enfocament diferent, mentre que els treballs consultats agafen la totalitat de les dades històriques existents (bé per una o per totes les monedes), aquí s'han obtingut les dades per una sola data concreta, com es la del darrer dia de cotització, i per a les principals cent criptomonedes.

De manera que per una banda es disposi d'informació ràpida de la posició de les monedes més importants al tancament del dia anterior i, per una altra, la possibilitat de generar un històric per veure'n evolució temporal i fer comparació entre les diferents divises.

8. Llicència. Seleccionar una d'aquestes llicències pel *dataset* resultant i explicar 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

Seguint les indicacions de les llicencies de Creative Commons, s'ha optat per la publicació d'aquest conjunt de dades per la **CC BY-NC-SA 4.0 License**.

L'elecció d'aquesta llicencia té a veure amb la idoneïtat de les clàusules que presenta en relació al treball que s'ha dut a terme.

- El beneficiari de la llicència ha de proporcionar el nom del creador del conjunt de dades generat, indicant els canvis que s'han realitzat. Així, es reconeix el treball aliè i en quina mesura s'han realitzat aportacions en relació amb el treball original.
- No es permet un ús comercial. Per evitar que es pugui treure profit econòmic d'un treball que s'ha elaborat a través de l'ús gratuït de dades acumulades per un tercer. En canvi si es permetrà la màxima difusió del treball en l'àmbit acadèmic.
- Les contribucions realitzades a posteriori sobre el treball publicat amb aquesta llicencia s'haurà de distribuir sota la mateixa. Així, permetrà que el treball de l'autor original es continuï distribuint acomplint els termes que ell mateix va contemplar.

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

A continuació s'inclou el codi que s'ha emprat per a la realització de *Web Scraping*.

```
# -------- Importació de llibreries --------------------
import requests
from bs4 import BeautifulSoup
from selenium import webdriver
import pandas as pd
import json
import time
import os
from datetime import datetime

# --------- WEB SCRAPING -----------------------------
## Obtencio de la llista de crytomonedes ##
# Seleccionem la pàgina
url_site = 'https://coinmarketcap.com/'

# Baixem el codi de la pàgina
# Utilitzem la funció get de la llibreria request i l'objecte BeautifulSoup per manegar
el codi html
page = requests.get(url_site)
soup_list = BeautifulSoup(page.content, 'html.parser')

data = soup_list.find('script', id="__NEXT_DATA__",type="application/json")
coins = {}

coin_data = json.loads(data.contents[0])
listings = coin_data['props']['initialState']['cryptocurrency']['listingLatest']['data']

## Obtenció de les dades històriques per cada criptomoneda ##
for i in listings:
    coins[str(i['id'])] = i['slug']

# Declarar un dataframe buit
df = pd.DataFrame(columns=('Coin', 'Open($)', 'High($)', 'Low($)', 'Close($)', 'Volum($)',
'Market_cap($)'))

# Fem seguiment de les diferents criptomonedes
for i in coins:
    url_web = f'https://coinmarketcap.com/currencies/{coins[i]}/historical-data'
    driver = webdriver.Chrome("C:\chromedriver")
    driver.get(url_web)
```

```
    time.sleep(3)
    soup = BeautifulSoup(driver.page_source,'html')
    driver.quit()
    # Només hi apareix una taula, que serà l'índex [0]
    tables = soup.find_all('table')

    table = tables[0]
    # Agafem tots els "th" ("head") i "td" ("data")
    row_header = table.find_all(["tr","th","td"])
    date rows = row header[9].text
    date_row = datetime.strptime(date_rows, '%b %d, %Y').date()
    open_row = pd.to_numeric(row_header[10].text.replace('$', '').replace(',',''))
    high_row = pd.to_numeric(row_header[11].text.replace('$', '').replace(',',''))
    low_row = pd.to_numeric(row_header[12].text.replace('$', '').replace(',',''))
    close_row = pd.to_numeric(row_header[13].text.replace('$', '').replace(',',''))
    volume row = pd.to numeric(row header[14].text.replace('$', '').replace(',',''))
    market_cap_row = pd.to_numeric(row_header[15].text.replace('$', '').replace(',',''))
    coin = coins[i]

    # Amb pandas bolquem les dades a un "dataframe", a partir del diccionari creat
    df = df.append({'Coin': coin, 'Open($)': open_row, 'High($)': high_row, 'Low($)':
low_row, 'Close($)': close_row, 'Volum($)': volume_row, 'Market_cap($)': market_cap_row},
ignore_index=True)

# I el "dataframe" es grava a un fitxer "csv"
fileName = 'Criptomonedes_Historical_Data_' + date_row.strftime('%Y%m%d') + '.csv'
pathName = r"C:\python\\" + fileName
# Es comprova si no existeix abans de crear-lo
if not os.path.exists(pathName):
    df.to_csv(pathName, index = False)
```

Aquesta mateix codi s'ha adjuntat en format *py*.

PRA1_Criptomonedes_NA_JAR_Web_Scraping.py

10. *Dataset*. Publicar el *dataset* en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.

S'ha deixat un *dataset* corresponen a **Criptomonedes_Historical_Data_20210325.csv** al repositori *Zenodo* amb el *Digital Object Identifier* (DOI), en el següent *link*:

https://zenodo.org/record/4641835#.YF8ap69KiUl

Contribució	Signatura
Disseny i concepció de la pràctica	Núria Aguilera i Joan Antoni Reina
Recerca prèvia de la temàtica	Núria Aguilera i Joan Antoni Reina
Redacció i elaboració de respostes	Núria Aguilera i Joan Antoni Reina
Disseny de la imatge representativa	Núria Aguilera i Joan Antoni Reina
Elaboració i desenvolupament del codi	Núria Aguilera i Joan Antoni Reina
Execucions i discussions sobre la pràctica	Núria Aguilera i Joan Antoni Reina