Pràctica 1 - Tipologia i cicle de vida de les dades

Noms i cognoms: Judith Cid i Jordi Tormo

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

   Després de la pandèmia del coronavirus, ens trobem en un procés inflacionari que afecta sobretot els preus de les matèries primeres o productes bàsics, els quals reflecteixen l'economia global, altament interconnectada. Especialment l'energia, es tracta d'una gran arma geopolítica que en moments d'inestabilitat pot ser, i està sent, molt rellevant.
   En aquest context, i amb la intenció d'estudiar indirectament l'economia mundial, ens proposem recollir dades dels preus de les matèries primeres en el rang de període dels últims vint anys.
   Per fer-ho, s'ha escollit la web IndexMundi, en concret per les matèries primeres https://www.indexmundi.com/commodities/, que es tracta d'un portal de dades que reuneix fets i estadístiques de múltiples fonts.

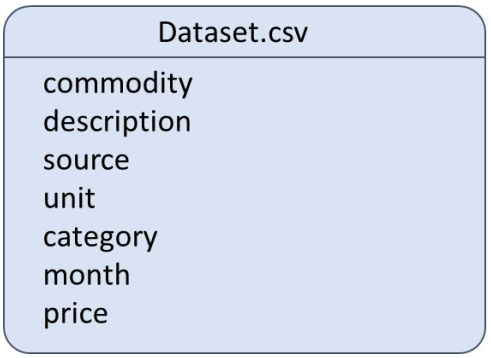
2. **Títol**. Definir un títol que sigui descriptiu pel dataset.

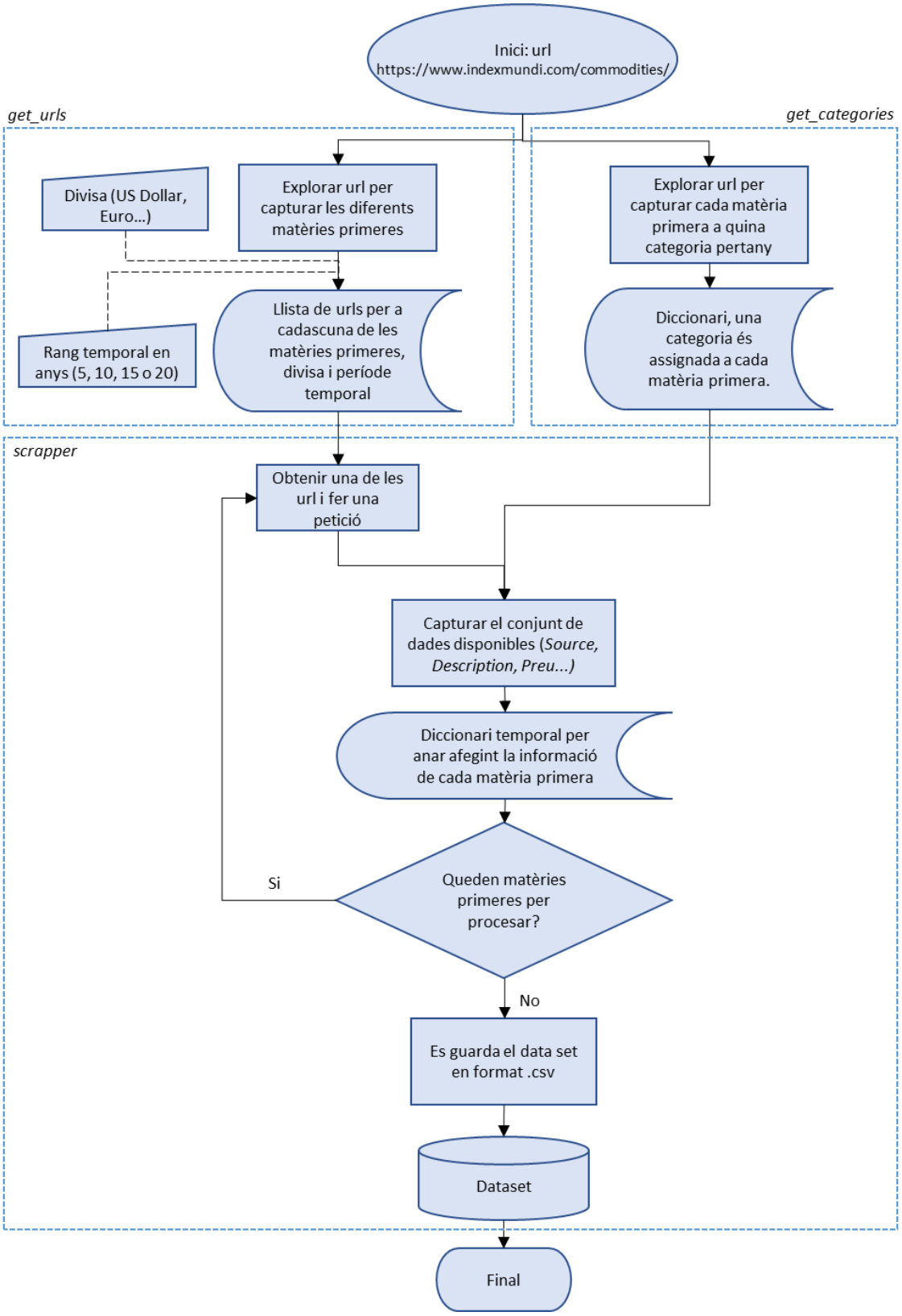
   Evolució del preu de les matèries primeres

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.

   S'ha extret dades del preu mensual de totes les matèries primeres disponibles a la web IndexMundi amb dades dels últims vint anys disponibles. En concret s'ha extret el nom de la matèria primera (*commodity*), juntament amb la seva categoria, una descripció, la font de les dades, el preu per mes i la moneda (*currency*).

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




Inici: url
https://www.indexmundi.com/commodities/
get_urls
get_categories
Divisa (US Dollar, Euro...)
Explorar url per capturar les diferents matèries primeres
Explorar url per capturar cada matèria primera a quina categoria pertany
Rang temporal en anys (5, 10, 15 o 20)
Llista de urls per a cadascuna de les matèries primeres, divisa i període temporal
Diccionari, una categoria és assignada a cada matèria primera.
scrapper
Obtenir una de les url i fer una petició
Capturar el conjunt de dades disponibles (Source, Description, Preu...)
Diccionari temporal per anar afegint la informació de cada matèria primera
Si
Queden matèries primeres per procesar?
No
Es guarda el data set en format .csv
Dataset
Final

5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de les dades.

   El període de temps de les dades és de les dades mensuals dels últims 20 anys. Cal tenir en compte que hi ha un retard d'uns dos mesos perquè fins que no acaba un mes complet el portal no pot començar a consolidar les dades. Així mateix, si d'alguna matèria primera no es disposa de la informació de tot el període, es reunirà tota la informació disponible.

   Els camps inclosos en el data set són els següents:

   - *Commodity:* Nom de la matèria primera, en anglès.
     Exemple: Gold, Aluminum, Rice…
   - *Description:* Descripció de quina referència de preu s'utilitza per a cadascuna de les matèries primeres, en anglès.
     Exemple: *Gold (UK), 99.5% fine, London afternoon fixing, average of daily rates*
   - *Source:* Font d'on el portal ha extret la informació.
     Exemple: World Bank
   - *Unit:* Unitat en la que s'expressa el preu, tant la divisa com per unitat de massa/volum
     Exemples: *US Dollars per Gallon, US Dollars per Kilogram…*
   - *Category:* Les matèries primeres es poden agrupar en categories ja siguin metalls, cereals, energies… etc. En anglès
     Exemple: Fruits, Meat, Metals, Seafood
   - *Month:* Mes i any del preu de la matèria primera [Mmm YYYY], el nom del mes abreujat a tres dígits i en anglès.
     Exemple: Apr 2004
   - *Price:* Preu de la matèria primera segons descripció i enunciat. Valor numèric.

6. **Propietari.** 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.

   IndexMundi és el propietari de tot el contingut de la seva web, tal i com s'especifica en l'apartat "Intellectual Property Rights" dels "terms of use".

   Aquest tema està molt estudiat, especialment actualment, degut a la crisis de matèries primeres en la que ens trobem. Dos exemples podrien ser aquest article de Caixa Bank o aquest altre del World Bank.

   Per actuar d'acord amb els principis ètics i legals s'ha comprovat l'arxiu robots.txt, per saber si era possible l'*scrapping*. El presentem a continuació:

   ```
   Sitemap: http://www.indexmundi.com/sitemap.xml

   User-agent: *
   Disallow: /api/v2/
   ```

User-agent: compatible; attributor
Disallow: /

User-agent: attributor
Disallow: /

User-agent: kalooga
Disallow: /

User-agent: Mp3Bot
Disallow: /

User-agent: WebAlta Crawler
Disallow: /

User-agent: TurnitinBot
Disallow: /

User-agent: GbPlugin
Disallow: /

User-agent: Domain Re-Animator Bot
Disallow: /

User-agent: trendkite-akashic-crawler
Disallow: /

A més, s'ha inclòs un retard entre peticions consecutives per evitar saturar el servidor web.

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.

Aquest conjunt de dades es creu interessant especialment pel context en què ens trobem, en una recessió global degut a diverses causes com podrien ser la pandèmia per coronavirus o la guerra a Ucraïna. A conseqüència de la pandèmia, hi ha hagut una caiguda del subministrament de les matèries primeres i això ha fet que els seus preus augmentin, i amb això també els preus als supermercats, la benzina o qualsevol altre producte derivat de les matèries primeres.

Evidentment, per poder confirmar aquests fets i estudiar en profunditat la correlació que poden tenir amb la geopolítica actual, s'han d'estudiar en detall les dades de l'evolució del preu de les primeres matèries, especialment en els últims anys, i això és justament el que es proposa en aquest treball.

Un exercici interessant que pot ajudar a comprendre millor l'actualitat, i a trobar possibles respostes, pot ser comparar les dades actuals amb dades d'altres crisis històriques, com per exemple la crisi econòmica del 2008.

Al disposar de les *commodities* classificades per categories, també es poden fer estudis per categories i focalitzar en els grups més interessants, com per exemple l'energia. Aquesta anàlisi és atractiva especialment a l'Europa actual, on per exemple el proveïment de gas s'ha vist afectat pel fet que en gran majoria provenia de Rússia i aquest es troba en guerra amb Ucraïna.

Justament l'article del World Bank presentat en l'exercici 6 se centra en l'estudi de l'energia. També utilitza el recurs de comparar les dades actuals amb dades històriques, per exemple , es representa l'evolució del preu de l'oli i s'explica que amb la pujada de l'oli a 1970 es va respondre amb polítiques per augmentar l'eficiència energètica i així tornar a disminuir el preu. Tot i això, l'article després conclou que aquest enfocament no seria viable avui en dia.

També es poden comparar els increments dels preus entre categories, estudi que es fa en l'article de CaixaBank presentat en l'exercici anterior. Per exemple, l'article conclou que l'escalada de preus més gran és la de l'energia, seguit dels béns agrícoles i dels metalls industrials.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar el motiu de la seva selecció. Exemples de llicències que poden considerar-se:
    - 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.
    - Altres (especificar quina).

S'ha escollit la llicència "*Creative Commons Attribution Non Commercial ShareAlike 4.0 International*" (CC BY-NC-SA 4.0).

Aquesta llicència permet copiar i redistribuir el material en qualsevol medi o format. A més, les dades es poden remesclar, transformar i crear a partir del material. Ara bé, les dades no es poden utilitzar per a finalitats comercials.

S'ha escollit aquesta llicència perquè als terms del portal web diu:

- *Users may use this Site solely for their own personal purposes. Each User agrees that it shall not copy, reproduce or download any information, text, images, video clips, directories, files, databases or listings available on or through the Site for the purpose of re-selling, re-distributing or re-publishing its content for mass mailing (via emails, wireless text messages, physical mail or otherwise), operating a business that competes with IndexMundi.com, or otherwise commercially exploiting the Site content.*

9. **Codi.** Codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.
    - El codi haurà de situar-se a la carpeta /source del repositori.
    - S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python poden obtenir-se mitjançant la comanda `pip3 freeze > requirements.txt`
    - Al document PDF s'han de comentar els aspectes més rellevants sobre com el codi realitza el procés de recol·lecció de dades, quines dificultats presenta el lloc web triat, i com les heu resolt.

L'estructura del *crawler* està dividida en tres mètodes, tots ells part d'una mateixa classe `IndexMundiScrapper`, que s'executen de forma successiva des del main.

Per executar el codi, és necessari incloure els paràmetres *currency*, la moneda en la que es volen extreure els preus, i *years_range*, el rang de temps (en anys) en el qual es vol extreure la informació. Pel que fa al nostre *Dataset*, s'ha escollit el dòlar americà i un període de 20 anys.

Aquests dos paràmetres s'utilitzen com arguments per la primera funció que s'executa, `get_urls`. Aquesta part del codi, partint de la web principal "*https://www.indexmundi.com/commodities/*", extreu els enllaços de totes les matèries primeres, seleccionant la moneda i el rang d'anys desitjat.

En un inici, per aconseguir els enllaços de totes les matèries primeres, es va optar per processar el *sitemap* de la web, però es va comprovar que aquest es troba incomplet.

Finalment, es va decidir fer navegar de forma autònoma per la web principal amb la intenció de trobar els enllaços de totes les commodities, les quals es troben classificades per categories en el menú de la barra lateral esquerra de la web.

El primer pas és extreure tots els enllaços de la pàgina, després, es filtren només els que ens interessen (els de les commodities). Un cop trobats, es fa *scrapping* un per un, per trobar llavors l'enllaç amb el rang d'anys i la moneda escollits.

En segon lloc s'executa la funció `get_categories`, que té com a objectiu crear un diccionari per tal de poder relacionar totes i cadascuna de les diferents matèries primeres amb una categoria. Aquest diccionari s'utilitzarà posteriorment a l'hora de generar el dataset. S'ha optat per utilitzar la classificació utilitza el mateix portal web d'Indexmundi.

Per últim, s'executa la funció `scrapper`. Aquesta és la funció que s'encarrega de fer les peticions a totes les url generades en la primera fase, recollir i guardar la informació desitjada per a cada matèria primera. Un cop s'obté tot el codi corresponent a una matèria primera, en primer lloc s'extreuen aquells atributs comuns, per exemple la descripció, la font originària de les dades, les unitats de mesura i la categoria, que s'obté del diccionari generat al segon pas. Seguidament s'obté la taula que conté les dades del preu mensual i es recorre per recollir el preu per cada mes. Tota la informació de cada matèria primera es guarda a un diccionari i un cop s'ha finalitzat el recorregut es consolida a un Dataframe. Aquest Dataframe, que conté tots

els atributs i registres desitjats, és el que retorna la funció. No es fa cap processament de les dades perquè es deixa per al preprocessament de la Pràctica 2.

Finalment, la funció principal guarda el Dataframe generat al pas anterior com un csv a la carpeta `dataset`. El separador utilitzat és el tab (`\t`) perquè hi ha atributs que tenen comes o semicomes.

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset (https://doi.org/…). El dataset també s'haurà d'incloure a la carpeta /dataset del repositori.

https://doi.org/10.5281/zenodo.7336214

11. **Vídeo.** Realitzar un breu vídeo explicatiu de la pràctica (màxim 10 minuts), que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo s'haurà de realitzar una presentació del projecte, destacant els punts més rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure les dades. Indicar l'enllaç del vídeo (https://drive.google.com/…), que haurà d'estar al Google Drive de la UOC.

url del video

Contribucions

Investigació prèvia	Judith Cid | Jordi Tormo
Desenvolupament del codi	Judith Cid | Jordi Tormo
Redacció de la memòria de la pràctica	Judith Cid | Jordi Tormo
Preparació del vídeo	Judith Cid | Jordi Tormo