Tipologia i cicle de vida de les dades · Pràctica 1

Alejandro Tortosa Molla i Sergi Poy Garcia

1. Context

El projecte pren com a base el Big Mac Index, que serveix per mesurar la paritat del poder adquisitiu entre països, a més de donar un tipus implícit de canvi de divisa que pugui ser comparable amb el tipus de canvi real de mercat, amb un objecte de consum tan conegut com el Big Mac.

En el nostre cas hem considerat oportú i segons "scrappejar" el web d'Apple i treure no només els preus dels iPhones per diferents països sinó també el del Macbook i de l'iPad. No som els primers que han volgut fer un "iPhone Index" com podem veure aquí, però sí que volem oferir de cara a la propera pràctica un context molt més ampli on no només prenguem de referència aquest producte icònic d'Apple, sinó també d'altres.

A més, però, és necessari poder tenir dades per a comparar el preu dels dispositius. Aquestes les extraurem d'una web on ofereixen un canvi de divises mundial amb actualització diària, i també del salari mitjà mensual de diferents països.

Per això mateix les dades han estat extretes de les següents webs:
https://www.apple.com/
https://www.numbeo.com/common/currency_settings.jsp
https://www.numbeo.com/cost-of-living/prices_by_country.jsp?displayCurrency=EUR&itemId=105

2. Títol

iPhone, Mac and Macbook Index.

Per la naturalesa i per la inspiració del projecte creiem que aquest títol s'escau en la descripció.

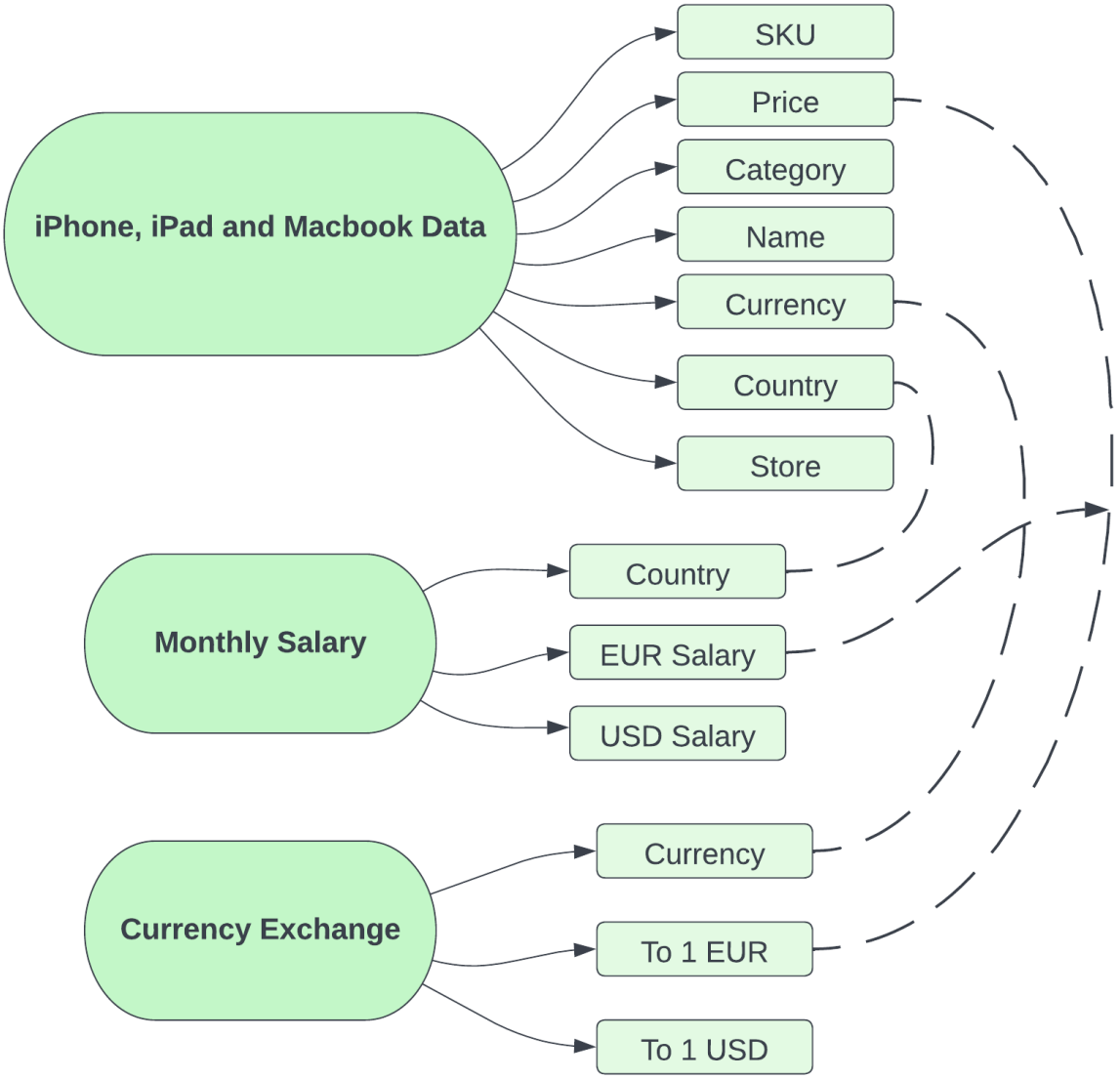
3. Descripció del Dataset

El que obtenim són tres conjunts de dades, que de cara a la propera pràctica consolidarem en un de sol.
Aquests datasets són:

- Extractes d'una llista on figuren els models d'Iphone, Mac i Ipad. El corresponent preu i model.
- Extracte de països amb el corresponent salari mensual mitjà després d'impostos.
- Nom de les diferents divises mundials i el seu tipus de canvi a temps real (diari).

Totes aquestes dades per separat poden ser interessants per comprar diferents tipus d'informació (per exemple models disponibles i variacions de preu dels dispositius).
Però el que realment fa interessant aquesta pràctica és l'interès que pot suscitar la combinació dels datasets, per així obtenir una informació molt més rellevant i descriptiva del món actual.

4. Representació Gràfica



5. Contingut

El dataset inclou tres conjunts de dades, que consolidarem per a treballar més fàcilment a la propera pràctica.

- Al primer conjunt trobem dades de preu, país i tenda als que es comercialitzen, producte, nom del model del producte, i una clau única del producte amb que ens descarreguem les dades. Les dades descarregades venen de la pàgina web d'Apple, i consistiran en diversos models de Mac, iPad i iPhone, de les pàgines web de tots els països als que es comercialitzen els productes.
- El segon conjunt de dades contindrà una llista de països amb el salari mensual mitjà després de taxes.
- Finalment el tercer conjunt de dades contindrà una llista amb quasi totes les divises del món i el seu canvi a relació d'1 EUR o 1 USD al temps de la descàrrega, ja que s'actualitza diàriament.

Respecte al període de temps, les dades corresponen al 20/11, i són les que trobem a la web en aquell mateix moment. En aquest cas treballem en unes dades estàtiques corresponents al moment esmentat (encara que executant el codi en altres moments podem donar-li altra perspectiva al treball).

6. Propietari

Les dades provenen de la pàgina web d'Apple.com per al primer DataSet, i de numbeo.com per al segon i tercer. Per tant, per una banda, el propietari seria Apple, i per l'altra el lloc web Numbeo, que es dedica a publicar investigacions i dades i que aparentment no estan sota la influència de cap govern.

Com hem dit en apartats anteriors, la idea prové de l'existència de l'Index Big Mac, no ve inspirada de cap estudi conegut per nosaltres, per tant, citaria aquí https://www.economist.com/big-mac-index com a anàlisi similar respecte a concepte, però no com a font de dades.

Pel que fa als aspectes legals, abans de fer cap descàrrega, hem examinat l'arxiu "robots.txt" de cadascun dels arxius i ens n'hem assegurat que les pàgines a les que anàvem a fer l'scrape, no estaven a la llista de "disallowed" per l'usuari "*".

A més si busquem els propietaris per consultes "whois" online podem trobar:

Per a apple:

```
Domain Name: apple.com
Registry Domain ID: 1225976_DOMAIN_COM-VRSN
```

```
Registrar WHOIS Server: whois.corporatedomains.com
Registrar URL: www.cscprotectsbrands.com
        http://cscdbs.com
Updated Date: 2022-02-16T01:15:06+00:00
        2022-02-16
Creation Date: 1987-02-19T00:00:00+00:00
        1987-02-19
Registrar Registration Expiration Date: 2023-02-20T05:00:00+00:00
        2023-02-20
Registrar: CSC CORPORATE DOMAINS, INC.
        CSC Corporate Domains, Inc.
Sponsoring Registrar IANA ID: 299
Registrar Abuse Contact Email: domainabuse@cscglobal.com
Registrar Abuse Contact Phone: 18887802723
Status:
        clientTransferProhibited
        serverDeleteProhibited
        serverTransferProhibited
        serverUpdateProhibited
Registry Registrant ID:
Registrant Name: REDACTED FOR PRIVACY (DT)
Registrant Organization: Apple Inc.
Registrant Street: One Apple Park Way
Registrant City: Cupertino
Registrant State/Province: CA
Registrant Postal Code: 95014
Registrant Country: us
Registrant Phone: 14089961010
```

Per a numbeo:

```
Domain Name: numbeo.com
Registry Domain ID: 1553640512_DOMAIN_COM-VRSN
Registrar WHOIS Server: WHOIS.DREAMHOST.COM
Registrar URL: WWW.DREAMHOST.COM
        http://www.DreamHost.com
Updated Date: 2022-03-27T07:44:48+00:00
        2022-03-27
Creation Date: 2009-04-27T06:38:47+00:00
        2009-04-27
Registrar Registration Expiration Date: 2023-04-27T13:38:47+00:00
        2023-04-27
Registrar: DREAMHOST
        DreamHost, LLC
Sponsoring Registrar IANA ID: 431
Registrar Abuse Contact Email: domain-abuse@dreamhost.com
Registrar Abuse Contact Phone: 12132719359
Status:
        ok
Registry Registrant ID:
Registrant Name: Proxy Protection LLC
```

```
Registrant Organization: Proxy Protection LLC
Registrant Street: 417 Associated Rd #324
        C/O numbeo.com
Registrant City: Brea
Registrant State/Province: CA
Registrant Postal Code: 92821
Registrant Country: us
Registrant Phone: 17147064182
```

7. Inspiració

El conjunt de dades resulta interessant per la mateixa naturalesa de la companyia de la qual venen. Igual que McDonalds, Apple és una companyia global que ven telèfons, tauletes tàctils i ordinadors arreu del món, per tant ens resultarà útil per a fer una comparació, a tots els països dels quals tinguem dades, dels preus a cadascun dels països i del cost relatiu de fer aquesta compra a partir del salari mitjà de cada lloc. Hem de pensar que la companyia inverteix grans esforços en fer estudis de mercat, igual que McDonalds, per tant, les dades són per si mateixes molt valuoses en aquest sentit.

A més, resulta interessant veure el tipus de canvi de divisa implícit a partir dels preus dels aparells en diferents països (Preu en divisa estrangera/Preu en divisa local) i comparar-la amb el preu real de la divisa del mercat de FX, i saber així si el tipus de divisa de mercat està sobrevalorat o infravalorat respecte d'aquest tipus de canvi implícit que hem calculat. Un altre resultat interessant que podria veure's, és l'evolució de les dades al llarg del temps, si anàrem executant el codi i extraient Datasets en diferents moments.

Per tant, la pròpia naturalesa de les dades i d'on provenen, així com el potencial que tenim per treballar amb elles ens han inspirat a fer aquest treball.

8. Llicència

En el nostre cas veiem convenient utilitzar la CC BY-NC-SA 4.0 o bé la "*Database released under Open Database License, individual contents under Database Contents License*", ja que permet que altres usuaris la puguin utilitzar sense cost, i les modificacions que en facin es comparteixen sota el mateix pretext i sense pretensió comercial.
Fer-ho de domini públic (CC0) és arriscat en el sentit que ni tan sols cal que una segona persona faci referència al nostre projecte i el pugui fer servir com seu sense cap mena de problema.
Finalment CC BY-SA 4.0 seria una opció quan no ens importés que algú en fes un ús comercial, però d'entre totes CC BY-NC-SA 4.0 ens sembla la més adient.

9. Codi

El trobem al GitHub a la carpeta arrel:

https://github.com/SergiPoy/iPhone-iPad-and-Macbook-Index

A l'hora de treure les dades del Website de Apple, hem trobat diverses complicacions.

Una d'elles és que es tracta d'una pàgina molt dinàmica en la que vas canviant de pantalles no només amb links, sinó que també amb formularis. La traducció d'aquest fet a llenguatge HTML fa que les etiquetes no siguin directes, i molts cops que no es pugui accedir a certs valors amb Beautifulsoup, ja que sota certes etiquetes sembla que no trobem visibilitat per a poder accedir els camps buscats. A més de la visibilitat d'aquests valors, també ens trobàvem en què molts valors que necessitàvem estaven en idiomes locals de cadascun dels països, la qual cosa feia complicat homogeneïtzar camps que trèiem de tots els països. Aquest problema l'hem resolt després d'estudiar la pàgina a fons, i adonar-nos que existeix un JSON (amb etiqueta HTML// <![CDATA[) dins de la web a la qual estan totes les dades (i que entenem que podria ser d'on es serveixen els formularis per a recuperar les dades). A més, aquestes dades són consistents a tots els websites de tots els països independentment de l'idioma.

Una altra dificultat va sorgir en el moment d'intentar recuperar tots els productes que existeixen, i un enllaç per a accedir i poder fer scrap de les seves dades. Finalment, després de provar molt, no vàrem trobar cap etiqueta a la qual poguéssim accedir i ens tornés tots els productes, i vàrem optar per fer una llista dels productes als quals estàvem interessats a recuperar els preus.

A partir d'aquí, el que fa el codi que fa l'scrap d'Apple és, per a la llista de productes predefinida, primerament mitjançant una funció, recuperar tots els enllaços de tots els països als quals es comercialitza el producte, i després, amb una altra funció, per a cada país i cada producte, treure els preus de tots els models de cada producte (per exemple, per a tots els països traiem el preu de l'Iphone 14 Pro en totes les combinacions de colors i capacitats que existeixen). Totes les dades que anem recuperant, les guardem i acumulem en un DataFrame de pandas, i el darrer DataFrame el transformem a CSV.

Pel que fa als dataset de la web Numbeo la major dificultat no ha estat en la quantitat de dades, ja que era realment petita en comparació amb altres webs, tot i això per aquesta mateixa raó gairebé no tenia tags ("divs"," td" amb clases determinades, etc...) amb el que l'extracció ha estat més aviat de força bruta amb Beautifulsoup i "for loops". Tot i això ha estat molt més senzilla.

10. Dataset

L'enllaç és el següent:

https://doi.org/10.5281/zenodo.7344067

11. Vídeo

L'enllaç és el següent:

https://drive.google.com/file/d/1CKC9qnsd_HVBO43CtLwLBh-HLg_cYpbv/view?usp=sharing

Contribucions	Signatura
Investigació prèvia	Sergi Poy, Alejandro Tortosa
Redacció de les respostes	Sergi Poy, Alejandro Tortosa
Desenvolupament del codi	Sergi Poy, Alejandro Tortosa
Participació al vídeo	Sergi Poy, Alejandro Tortosa