Tipologia i cicle de vida de les dades

Pràctica 1

Web Scraping

Autors:

Antoni Sànchez Magraner
Joan Solà Porta

1 Context

La web **datosmacro.com** publica diversos tipus indicadors globals de tipus econòmic, demogràfic, sociològic i cultural, amb un abast mundial, desglossant-los per països i sèries temporals (en unitats mensuals, tot i que en molts casos són informats només un màxim de dos mesos per cada any). La web forma part del grup editorial del diari *Expansión*, especialitzat en informació empresarial, econòmica i financera. Concretament publica dades de població, index de natalitat i defuncions, immigració, índex de desenvolupament humà, de progrés social, risc de pobresa, religió, etc.

Des de la pàgina principals de la web es permet seleccionar un indicador i, des de la de l'indicador un país, d'entre els disponibles de la llista, i així accedir a la visualització de la taula evolutiva de l'indicador per al país seleccionat (codificada en html). La funcionalitat que ofereix a l'hora de realitzar consultes elaborades i complexes és molt limitada, atès que no permet la descàrrega directa de les diferents taules, aquest fet dificulta la comparació i anàlisi conjunt de diferents indicadors .

A partir de les diferents pàgines d'indicadors de cada país, que es troben desagregades entre sí, es proposa la creació d'un únic dataset que unifiqui tots els indicadors, implementant un procés d'extracció de dades ad hoc, emprant el llenguatge de programació python.
La pàgina principal des d'on es pot accedir a les demés pàgines d'indicadors és:
`https://datosmacro.expansion.com/demografia`

2 Títol

Sèries històriques d'indicadors socio-econòmics per països.

3 Descripció del dataset

A partir de les dades dels diferents indicadors de la web datosmacro, es proposa crear un conjunt de dades que reculli, de manera unificada, tots els indicadors de tots els països desglossats per sèries temporals. D'aquesta manera cada registre del dataset contindrà:

- dos camps, que posen en context temporal i geogràfic (país i mes-any), i
- la resta de camps del dataset, on figurarien els respectius valors dels indicadors.

Amb el dataset generat es podrien realitzar anàlisis, comparatives i estudis de correlació entre els diferents indicadors, que no seria possible realitzar a través de la web.

4 Representació gràfica

A la figura 1 s'hi mostra l'evolució d'una de les estadístiques recollides en el conjunt de dades per a 5 països diferents. En aquest cas s'ha escollit l'esperança de vida al néixer com a mètrica a representar. Tot i això, la figura ens dona una idea del potencial que té el conjunt de dades recollit.

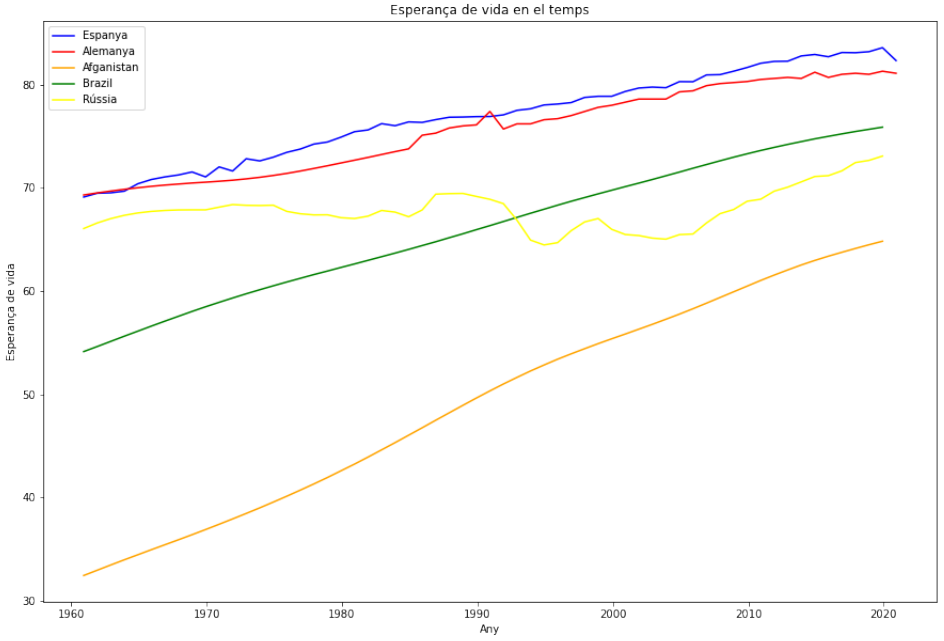

Figura 1: Representació de l'evolució de l'esperança de vida de 5 dels països del conjunt de dades.

5 Contingut

Com ja s'ha comentat amb anterioritat, les dades que s'inclouen al conjunt de dades s'han recollit a partir de les pàgines que publiquen estadístiques per a cada país. Per a fer això,

doncs, el codi segueix els següents passos:

1. Llistar les estadístiques disponibles en l'apartat de demografia.
2. Seleccionar una estadística en concret.
3. Seleccionar un país en concret.
4. Guardar les dades de l'estadística per a tot l'històric disponible.
5. Repetir els passos 3 i 4 per a tots els països disponibles.
6. Repetir els passos 2-5 per a totes les estadístiques de l'apartat de demografia.

Així doncs, el conjunt de dades recollit conté diverses dades sociodemogràfiques dels diversos països del món al llarg del temps. Els camps recollits són els següents:

- 'País': país.
- 'Fecha': data (en format 'mes-any').
- 'Ranking de la Brecha de Género': posició en índex de bretxa de gènere respecte a la resta de països.
- 'Índice de la Brecha de Género': índex de bretxa de gènere.
- 'Densidad': densitat de població.
- 'Hombres': població d'homes.
- 'Mujeres': població de dones.
- 'Población': població total.
- 'Inmigrantes hombres': nombre d'immigrants homes.
- 'Inmigrantes mujeres': nombre d'immigrants dones.
- 'Inmigrantes': nombre d'immigrants total.
- '% Inmigrantes': percentatge d'immigrants.
- 'Saldo remesas (M.$)': diferència entre les remeses enviades i rebudes
- 'Remesas recibidas (%PIB)': percentatge del PIB que representen les remeses rebudes.
- 'Remesas recibidas (M.$)': remeses rebudes.
- 'Remesas enviadas (%PIB)': percentatge del PIB que representen les remeses enviades.
- 'Remesas enviadas (M.$)': remeses enviades
- 'Emigrantes hombres': nombre d'emigrants homes.
- 'Emigrantes mujeres': nombre d'emigrants dones.
- 'Emigrantes': nombre d'emigrants total.

- '% Emigrantes': percentatge d'emigrants total.
- 'IDH': índex de desenvolupament humà.
- 'Ranking IDH': posició en IDH respecte a la resta de països.
- 'SPI': índex de progrés social.
- 'SPI_2': posició en SPI respecte a la resta de països.
- 'Necesidades Humanas Básicas': posició en cobertura de les necessitats humanes bàsiques.
- 'Fundamentos del Bienestar': posició en cobertura dels fonaments del benestar.
- 'Oportunidades': posició en l'oferta d'oportunitats.
- 'Índice de Paz Global': índex de pau global.
- 'Ranking Paz Global': posició en índex de pau global respecte a la resta de països.
- 'Ranking': posició en índex d'envelliment respecte a la resta de països.
- 'Índice de envejecimiento': índex d'envelliment.
- 'Entorno': índex que indica la llibertat d'elecció que tenen les persones grans per viure de manera autosuficient i independent.
- 'Competencias': índex que indica la inversió en ocupació i educació.
- 'Salud': índex que mesura el benestar físic dels habitants.
- 'Ingresos': índex que indica la facilitat d'accés a una pensió digna.
- 'Nacidos': nombre total de naixements.
- 'Nacidos Hombres': nombre de naixements d'homes.
- 'Nacidos Mujeres': nombre de naixements de dones.
- 'Tasa Natalidad': taxa de natalitat.
- 'Índice de Fecund.': índex de fecunditat.
- 'Muertes': morts.
- 'Muertes - Hombres': morts d'homes.
- 'Muertes - Mujeres': morts de dones.
- 'Tasa mortalidad': taxa de mortalitat.
- 'Esperanza de vida - Mujeres': esperança de vida de dones.
- 'Esperanza de vida - Hombres': esperança de vida d'homes.
- 'Esperanza de vida': esperança de vida.
- 'Matrimonios': nombre de matrimonis.

- 'Edad Media primer mat. mujeres': edat mitjana al primer matrimoni per a les dones.
- 'Edad Media primer mat. hombres': edat mitjana al primer matrimoni per als homes.
- 'Tasa de primer matrimonio - mujeres': taxa de primers matrimonis per a dones.
- 'Tasa de primer matrimonio - hombres': taxa de primer matrimoni per a homes.
- 'Tasa bruta de nupcialidad': taxa bruta de nupcialitat.
- 'Divorcios': nombre de divorcis.
- 'Tasa bruta de divorcios': taxa bruta de divorcis.
- 'Suicidios mujeres': nombre de suïcidis de dones.
- 'Suicidios hombres': nombre de suïcidis d'homes.
- 'Suicidios ': nombre de suïcidis total.
- 'Suicidios tasa femenina': nombre de suïcidis per cada 100.000 dones.
- 'Suicidios tasa masculina': nombre de suïcidis per cada 100.000 homes.
- 'Suicidios por 100.000': nombre de suïcidis per cada 100.000 habitants.
- 'Número de Homicidios': homicidis totals.
- 'Homicidios Mujeres': nombre d'homicidis de dones.
- 'Homicidios Hombres': nombre d'homicidis d'homes.
- 'Homicidios Mujeres por familiares': nombre d'homicidis de dones per familiars.
- 'Homicidios hombres por familiares': nombre d'homicidis d'homes per familiars.
- 'Homicidios por 100.000': homicidis totals per cada 100.000 habitants.
- 'Reclusos mujeres adultas': nombre de dones adultes recluses.
- 'Reclusos mujeres adultas / 100.000': dones adultes recluses per cada 100.000 dones.
- 'Reclusos varones adultos': nombre d'homes adults reclusos.
- 'Reclusos varones adultos / 100.000': homes adults reclusos per cada 100.000 homes.
- 'Reclusos jóvenes': nombre de reclusos joves.
- 'Reclusos jóvenes / 100.000': nombre de reclusos joves per cada 100.000 joves.
- 'Total reclusos': nombre total de reclusos.
- 'Reclusos por 100.000': nombre total de reclusos per cada 100.000 habitants.
- 'Personas en riesgo de pobreza': nombre de persones en risc de pobresa.
- 'Umbral persona': llindar de pobresa per persona.

- 'Umbral persona_3': posició del llindar de pobresa per persona respecte a la resta de països.
- 'Umbral hogar': llindar de pobresa per família.
- 'Umbral hogar_5': posició del llindar de pobresa per família respecte a la resta de països.
- '% Riesgo Pobreza': percentatge de la població en risc de pobresa.
- 'Índice de Gini': índex GINI.
- 'Ranking Felicidad': posició de líndex de felicitat respecte a la resta de països.
- 'Índice Felicidad': índex de felicitat.
- '0-14 años %': percentatge de la població amb 14 anys o menys.
- '15-64 años %': percentatge de la població entre 15 i 64 anys.
- '> 64 años %': percentatge de la població amb més de 64 anys.
- 'Tasa de alfabetización mujeres': taxa d'alfabetització de dones.
- 'Tasa de alfabetización hombres': taxa d'alfabetització d'homes.
- 'Tasa de alfabetización de adultos': taxa d'alfabetització d'adults.
- 'Tasa de alfabetización jovenes mujeres': taxa d'alfabetització de dones joves.
- 'Tasa de alfabetización jovenes hombres': taxa d'alfabetització d'homes joves.
- 'Tasa de alfabetización jovenes': taxa d'alfabetització de joves.

Les dades recollides es comprenen, des d'un punt de vista temporal, entre el desembre de 1920 i el desembre de 2021, estant agrupades mensualment.

Tot i això, no es tenen dades per a tots els mesos, motiu pel qual per a cada país només s'inclouen registres per als mesos dels quals es té informació. A més, per a un mes i país pot no existir informació de totes les estadístiques, de manera que no tots els camps d'un mateix registre tenen per què estar informats.

6 Agraïments

Tal com ja s'ha comentat amb anterioritat, el propietari de les dades és Datosmacro.com[1], un servei de recol·lecció i publicació de dades associat al diari Expansión[2], que s'especialitza en economia. Tal com s'indica en aquesta pròpia pàgina[3], aquesta pàgina té com a principal objectiu oferir al públic les principals variables econòmiques i sociodemogràfiques dels països per tal d'oferir una visió de la situació econòmica d'aquests. Tal com s'indica, en aquest mateix comunicat, les fonts utilitzades per a les dades recol·lectades són diversos

[1] Veure la pàgina web `https://datosmacro.expansion.com/`.
[2] Accés a través de l'adreça `https://www.expansion.com/`.
[3] Veure la nota inclosa a `https://datosmacro.expansion.com/legal/acerca-de`.

organismes oficials.

A l'hora de dur a terme l'anàlisi presentat en aquesta pràctica no se n'ha tingut en compte cap de previ que fes servir aquesta font. Tot i això, sí que s'ha pres com a inspiració la pàgina web publicada per l'Organització Mundial de la Salut de les Nacions Unides per donar informació sobre l'evolució de la pandèmia de la Covid-19 (veure la referència [1]). Això és degut a que, tant aquest projecte com el que presentem aquí tenen com a objectiu traslladar a la població en general coneixement sobre algun aspecte d'interès a partir de dades recollides d'organismes oficials. Tot i això, és cert que la referència citada va un pas més enllà i, a més de recollir les dades, presenta un dashboard amb què l'usuari pot interactuar i fer les seves exploracions.

Tal com es pot veure a l'avís legal de la pàgina web de Datosmacro.com[4], el propietari de les dades no posa en cap moment cap restricció a l'hora de fer ús d'aquestes (simplement imposa que no se'l pot fer responsable dels possibles errors que puguin contenir). A més, si es mira el fitxer de *robots.txt*[5], el propietari només demana que no es faci rastreig de les dades que publica per mitjà d'uns bots concrets que no hem fet servir en cap cas. Per tant, des d'un punt de vista ètic i legal l'únic que hem hagut de tenir en compte a l'hora de recopilar les dades ha sigut fer referència a Datosmacro.com com a propietari i recopilador d'aquestes i vigilar de no afectar el funcionament dels seus servidors a l'hora de fer el *web scraping*.

7 Inspiració

Tal com es pot veure a partir dels camps del conjunt de dades recollit (presentats a l'apartat 5), aquest té un gran potencial, ja que ens permet respondre preguntes socioeconòmiques de tot tipus respecte als països que s'hi inclouen. Com a petit recull, les següents preguntes es poden respondre a partir de les dades recollides (segurament se'n poden pensar moltes més que no s'han inclòs aquí):

- Població:
  - Com evoluciona en el temps la densitat de població dels països?
  - Quins són els països més/menys densament poblats al llarg del temps?
  - Com ha evolucionat al llarg del temps la població dels diversos països?
  - Quins són els països més/menys poblats al llarg del temps?
  - Quina és la proporció actual d'homes-dones per a cada país? I al llarg del temps?
  - Com ha anat variant la piràmide de població d'un determinat país al llarg del temps?
  - Quins són els països amb la població més envellida/jove?
- Immigració/emigració:
  - Com ha evolucionat la quantitat d'immigrants/emigrants d'un país concret? I per gènere?

[4]Veure `https://datosmacro.expansion.com/legal/terminos`.
[5]Es pot consultar a l'adreça `https://datosmacro.expansion.com/robots.txt`.

  - Quins són els països amb una taxa d'immigració/emigració més alta/baixa actualment? I històricament?
  - Quin impacte sobre el PIB d'un país concret tenen les remeses rebudes/enviades? Com ha evolucionat al llarg del temps?

- Natalitat/mortalitat:
  - Com ha evolucionat la taxa de natalitat dels països?
  - Quins són els països amb majors/menors taxes de natalitat?
  - Com ha evolucionat la taxa de mortalitat dels països?
  - Quins són els països amb majors/menors taxes de mortalitat?
  - Com ha evolucionat l'esperança de vida dels països al llarg del temps?
  - Quins són els països amb majors/menors esperances de vida?

- Matrimonis/divorcis:
  - Com ha evolucionat la taxa de matrimonis/divorcis al llarg del temps en cada país?
  - Quins són els països amb major/menor taxa de nupcialitat/divorcis? I històricament?

- Suïcidis/homicidis/reclusos:
  - Quina és l'evolució històrica de suïcidis en cada país?
  - Quins són els països amb major/menor taxa de suïcidis?
  - Quina evolució ha tingut la taxa d'homicidis en cada país?
  - Quins són els països amb major/menor taxa d'homicidis?
  - Quina proporció de gènere hi ha en les persones recluses de cada país?
  - Com ha anat evolucionant la taxa de reclusos de cada país al llarg del temps?

Si es comparen les possibilitats del dataset recollit en aquesta pràctica amb el que es presenta a la referència [1], es pot observar que presenten moltes similituds des d'un punt de vista del tipus de preguntes de cada temàtica que ens permeten respondre. Això és degut a que inclouen informació sobre molts aspectes diferents del tema d'estudi (la pandèmia en les dades de [1] i la informació sociodemogràfica en el cas de les nostres dades) per a cadascun dels diferents països. A més, per a cadascun dels països es presenta una evolució temporal de cadascun dels indicadors recollits (de manera que la *primary key* dels dos conjunts de dades és essencialment la mateixa: data-país). Per acabar, les dades dels dos conjunts de dades procedeixen de fonts oficials (en el cas de la referència [1] la OMS recull les dades que proporcionen les autoritats sanitàries de cada país).

8 Llicència

A l'hora de seleccionar la llicència d'ús del dataset resultant del webscraping, en primer lloc, s'han consultat les condicions d'ús de la font de dades original sobre les que s'ha realitzat l'extracció. Concretament datosmacro.com disposa d'una pàgina de *Termes i condicions d'ús*, en la que s'indica el titular del lloc web (Alldatanow S.L.) i la regulació de l'accés, navegació i utilització del lloc web: *l'accés i navegació del lloc web no requereixen registre i*

ambdós suposen que l'usuari accepta en la seva totalitat i s'obliga a complir les condicions de l'avís legal.

A l'apartat responsabilitats i garanties es remarca que el propietari de lloc web no es fa responsable de qualsevol conseqüència derivada de l'ús de les dades del lloc. Més enllà d'indicar que l'accés i navegació del lloc web no requereixen registre no es detalla cap tipus de llicència d'ús de les dades que conté, i no es fa cap tipus de referència a l'aplicació de pràctiques de web scraping.

Per altra banda, també s'ha consultat el fitxer robots.txt `https://datosmacro.expansion.com/robots.txt`, on es pot comprovar que es restringeix l'accés a diversos contextos del domini per qualsevol agent (*), a més de restringir el rastreig de tota la web per part de determinats agents concrets. Per tant, tot i que que el propietari de la web intenta evitar, o advertir en contra, del rastreig per part de determinats tercers o eines, no sembla que tenguin la intenció explícita d'excloure les pràctiques de web scraping en general, sinó algunes concretes realitzades per determinats agents (que pot ser tenguessin un impacte negatiu sobre els servidors), per tant es podria considerar que es permet un ús moderat del web scraping.

S'ha de tenir en compte que les dades que es publiquen a la web són estadístiques de diversos indicadors en un abast mundial que, segurament, no han estat elaborades pels propietaris de la web, sinó que simplement n'han fet una recopilació o compendi. Tenint en compte això, en aquest cas, podrien acollir-se a la Directiva de la Unió Europea sobre bases de dades de l'any 1996, que proporciona protecció jurídica als creadors de bases de dades que no estiguin coberts per drets de propietat intel·lectual.

Tot i això, en el cas que ens ocupa, tenint en compte el que s'indica als *Termes i condicions d'ús* (es permet l'ús de les dades de manera completament oberta, sense registre) i a la no exclusió de l'user agent emprat, *python-requests*, sembla que la generació del dataset seria admesa com una de les moltes finalitats per a les que es poden emprar les dades que es publiquen a la web datosmacro. També s'ha de tenir en compte que no s'imposa cap restriccions a la realització d'un ús comercial de les dades que es puguin consultar. Tot i això, a l'hora de realitzar el web scraping i fer un ús posterior del dataset resultant, seria recomanable:

- Evitar causar qualsevol tipus de dany i disrupció sobre el servidor.
- Sol·licitar permís a la web de la realització del web scraping i posterior ús de les dades (sobretot si se'n fa comptes fer un ús comercial).
- Mencionar la font original d'on s'han extret les dades, en cas de fer-ne un ús públic o cedir-les a tercers.

Per tant, una vegada s'ha determinat que amb la generació del dataset no s'ha incomplert cap norma ni infringit drets i que les dades que conté són públiques, el dataset que ha resultat del procés de web scraping es podria publicar emprant la llicència **CC BY-SA 4.0**, que permet copiar, redistribuir (en qualsevol mitjà o format), remesclar, transformar i crear nou contingut a partir d'ell (sempre i quan es reconegui l'autoria i sigui alliberat amb la mateixa llicència i condicions.
Es pot accedir a l'avís legal a: `https://datosmacro.expansion.com/legal/terminos`

9 Codi

Enllaç al repositori GitHub: `https://github.com/jsolatsanchez/Practica1Webscraping`.

10 Dataset

S'indica el DOI del dataset: 10.5281/zenodo.6426889

Referències

[1] World Health Organization (WHO). *WHO Coronavirus (COVID-19) Dashboard*[en línia]. Disponible a: `https://covid19.who.int/`.

Contribucions al treball

Contribucions	Signatura
Investigació prèvia	A.S.M, J.S.P
Redacció de les respostes	A.S.M, J.S.P
Desenvolupament del codi	A.S.M, J.S.P