XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
14 d'Abril de 2020

M2.951 - TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1 – WEB SCRAPING

DATASET WORLD HEALTH INDICATORS

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

En un context de pandèmia per la Covid-19, existeix una especial preocupació a escala internacional per la situació de feblesa dels sistemes sanitaris dels països menys desenvolupats i/o altament poblats.
El dataset World Health Indicators ha estat elaborat a partir de dades proporcionades per l'Organització de les Nacions Unides i pel Banc Mundial a través dels seus llocs web. Ambdues són organitzacions que treballen pel desenvolupament humà, la seguretat i l'erradicació de la pobresa.

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

El dataset que es presenta s'ha titulat ***World Health Indicators***, perquè integra indicadors de salut pública i socioeconòmics pels diferents països del món.

3. Descripció del dataset.

El dataset World Health Indicators està format per 681 files i 22 columnes. Conté dades relacionades amb l'àmbit de la salut pública i dades socioeconòmiques referents a 227 països i regions del món.

4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment.

World Health Indicators Dataset image.

XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
14 d'Abril de 2020

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

Captura:

El procés de captura de dades s'ha dut a terme conforme a les següents accions:

1. Web scraping de diverses URL del lloc web de les Nacions Unides que contenen indicadors socioeconòmics i de salut pública pels diferents països del món. L'índex d'URLs es pot trobar al següent link: http://data.un.org/en/index.html. S'utilitza la llibreria Selenium de Python.

2. S'obtenen indicadors addicionals pels diferents països del món del lloc web del Banc Mundial a través d'API. L'índex dels diferents indicadors existents es pot trobar a: https://data.worldbank.org/indicator.

3. S'obté la nomenclatura ISO pels països a través d'un procés de web scraping mitjançant la llibreria BeautifulSoup de Python.

4. Es crea un únic dataframe que aglutina els valors de diversos indicadors donats per les dues institucions i es guarda en un arxiu en format .csv.

Rang temporal de les dades:

Els valors dels indicadors socioeconòmics i de salut pública s'han capturat pels anys 2005, 2010 i 2019. Per les dades del Banc Mundial, s'ha hagut d'assimilar les dades de 2018 a les de 2019 a l'espera de que es publiquin les dades de 2019.

Camps del dataset:

country_code – Codi ISO3 de país

country – Nom del país

year – Any

population – Població total

population_grow – Creixement de la població (% anual)

population_under_14 – Població d'edats entre els 0 – 14 anys (% de la població total)

population_above_65 – Població d'edats superior als 64 anys (% de la població total)

gdp - Producte Interior Brut (milions de dòlars actuals) – GDP per les seves sigles en anglès.

gdp_growth_rate – Taxa de creixement del Producte Interior Brut (% anual a preus constants de 2010)

unemployment – Atur (% de població en edat laboral)

education_gov_expenditure – Despesa governamental en educació (% GDP)

XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
14 d'Abril de 2020

health_expenditure – Despesa en salut (% GDP)

life_expectancy_fem – Esperança de vida, dones (anys)

life_expectancy_male – Esperança de vida, homes (anys)

non_commun_disease_death – Morts per malalties no transmissibles (% del total)

commun_disease_death – Morts per malalties transmissibles i condicions d'embaràs, prenatals i nutricionals (% del total)

tuberculosis – Incidència de tuberculosi (per cada 100.000 habitants)

hiv – Prevalença de VIH (en % de població d'edats entre els 15 i 49 anys)

infant_mortality – Mortalitat infantil (per cada 1000 naixements amb vida)

undernourishment – Prevalença de desnutrició (% població)

hospital_beds – Llits d'hospital (per cada 1000 habitants)

physicians – metges (per cada 1000 habitants)

6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Les dades socioeconòmiques i sanitàries s'han extret en part del lloc web de l'**Organització de les Nacions Unides,** ONU (UN per les seves sigles en anglès), en concret del link: http://data.un.org/en/index.html. Es tracta de dades proporcionades per la seva Divisió d'Estadística. L'ONU és una organització internacional encarregada de mantenir la pau i seguretat internacionals, fomentar les aliances entre nacions i posar solució a problemes globals a través de la cooperació internacional.
Terms of use: https://data.un.org/Host.aspx?Content=UNdataUse.

D'altra banda, d'altres dades d'indicadors socioeconòmics i de salut pública s'han extret del lloc web del Banc Mundial, en concret de The World Bank Indicators, amb link: https://data.worldbank.org/indicator. Es tracta d'una organització que dona assistència financera i tècnica als països en vies de desenvolupament per tal de reduir l'escletxa de pobresa.
Terms of use: https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets.

Les dades obtingudes i que conformen el dataset no han patit cap modificació respecte a les fonts mencionades.

7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

En el context actual de pandèmia per la Covid-19 en el que ens trobem, els sistemes sanitaris d'arreu del món es troben sota una forta pressió. La manca de mitjans materials i personals fa que les xifres oficials d'infectats no puguin reflectir la situació real de cada territori. Aquest dataset proporciona informació general sobre l'estat sanitari pre-pandemia en el que es troben els diferents països i es podria utilitzar per preveure quines regions parteixen d'una situació sociosanitària més precària, i per tant podrien ser més vulnerables enfront de la pandèmia. És possible estudiar a partir d'ell quins factors socioeconòmics tenen un major impacte en la salut pública, quins territoris pateixen malalties transmissibles preexistents, i es podria fer servir en un futur per calcular l'impacte real de la pandèmia un cop controlada. Aquest és tan sols un

XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
14 d'Abril de 2020

suggeriment, però les aplicacions del dataset poden ser molt diverses i de ben segur no estan limitades a l'impacte de la Covid-19.

8. Llicència.

Tenint en consideració les llicències de les fonts de dades originals, la llicència que ens sembla més adient pel dataset World Health Indicators i el material que l'acompanya és la CC BY-SA 4.0, ja que permet el següent:

- Copiar i redistribuir el material per qualsevol mitjà i en qualsevol format.
- Transformar, fusionar i elaborar nous materials per a qualsevol propòsit, també per usos comercials.

S'exigeix que:

- Quedi reconeguda l'atribució als autors del material, aportant un enllaç a la llicència i indicant si s'han dut a terme modificacions. L'atribució s'ha de realitzar sense suggerir que el nou ús té el support del llicenciant.
- Les transformacions o fusions del material han de ser distribuïdes sota la mateixa llicència original.

D'aquesta manera, qualsevol persona o entitat podria treure profit d'aquest material per nous usos tant comercials com no comercials.

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

El codi està compost per tres arxius Python, que es troben al següent enllaç de Github: https://github.com/adtsune/tipologia_cicle_dada

1. *practica1.py* – Arxiu principal, que crida a diverses funcions dels dos arxius següents, i obté el dataset i el guarda en un arxiu tipus .csv.

2. *practica1selenium.py* – Arxiu que conté les funcions que realitzen el web scraping i la manipulació de dades del lloc web de les Nacions Unides utilitzant la llibreria Selenium de Python.

3. *practica1funcions.py* – Arxiu que conté les funcions que obtenen les dades del lloc web del Banc Mundial a través d'API, i la nomenclatura ISO pels països amb web scraping utilitzant la llibreria BeautifulSoup de Python.

10. Dataset. Publicar el dataset en format CSV a Zenodo.

11. Lliurar. Presentar el treball amb el DOI del dataset a Github

Tots els fitxers relacionats amb aquesta pràctica es troben al següent enllaç de Github:
https://github.com/adtsune/tipologia_cicle_dada

XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
14 d'Abril de 2020

Taula de contribucions al treball

Contribucions	**Signa**
Recerca prèvia	XVD, ADS
Redacció de les respostes	XVD, ADS
Desenvolupament codi	XVD, ADS