XAVIER VENTURA DE LOS OJOS
ANNA DE LA TORRE SUÑE
BARCELONA
11 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.

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

El dataset que es presenta s'ha titulat ***world _health_indicators***, perqué integra indicadors de salut pública i socioeconòmics pels diferents països del món.

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret.

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
11 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 procès 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 econòmics i socials 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 diversos indicadors demogràfics i de salut pública dels 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 mitjançacnt 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 del 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 la espera de que es publiquin les dades de 2019.

Camps del dataset:

```
# Llista indicadors UN data
indicadors_UN = ["GDP: Gross domestic product (million current US$)", "GDP per capita (current US$)",
                 "GDP growth rate (annual %, const. 2010 prices)", "Unemployment (% of labour force)",
                 "Education: Government expenditure (% of GDP)", "Health: Current expenditure (% of GDP)",
                 "Infant mortality rate (per 1 000 live births)"]


# Llista d'indicadors World Bank

indicadors_BM = ["SP.POP.TOTL", "SP.POP.GROW", "SP.DYN.LE00.FE.IN", "SP.DYN.LE00.MA.IN",
                 "SP.POP.65UP.TO.ZS", "SH.DTH.COMM.ZS", "SN.ITK.DEFC.ZS", "SH.DYN.AIDS.ZS",
"SH.DTH.NCOM.ZS",
                 "SH.MED.BEDS.ZS", "SH.MED.PHYS.ZS", "SH.SGR.PROC.P5"]


# Population, total
#https://data.worldbank.org/indicator/SP.POP.TOTL?view=chart

#Population growth (annual %)
#https://data.worldbank.org/indicator/SP.POP.GROW?view=chart


# Life expectancy at birth, female (years)
#https://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN?view=chart

# Life expectancy at birth, male (years)
```

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

#https://data.worldbank.org/indicator/SP.DYN.LE00.MA.IN?view=chart

Population ages 65 and above (% of total population)
#https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS?view=chart

Prevalence of undernourishment (% of population)
#https://data.worldbank.org/indicator/SN.ITK.DEFC.ZS?view=chart

Prevalence of HIV, total (% of population ages 15-49)
#https://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?view=chart

Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total)
#https://data.worldbank.org/indicator/SH.DTH.COMM.ZS?view=chart

Cause of death, by non-communicable diseases (% of total)
#https://data.worldbank.org/indicator/SH.DTH.NCOM.ZS?view=chart

Hospital beds (per 1,000 people)
#https://data.worldbank.org/indicator/SH.MED.BEDS.ZS?view=chart

Physicians (per 1,000 people)
#https://data.worldbank.org/indicator/SH.MED.PHYS.ZS?view=chart

Number of surgical procedures (per 100,000 population)
#https://data.worldbank.org/indicator/SH.SGR.PROC.P5?view=chart

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 una part de les dades sanitàries s'han extret 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, les dades 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 técnica 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.

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

En el context actual de pandèmia pel Covid-19 en el que ens trobem, els sitemes 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 precaria, i per tant podrien ser més vulnerables enfront la pandèmia. És possible estudiar a partir d'ell quins factors socioeconòmics tenen un major impacte en la salut pública, quins territoris pateixen malaties transmisibles 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 sugeriment, però les aplicacions del dataset poden ser molt diverses i de ben segur no estan limitades a l'impacte del 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 la atribucó 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 sugerir que el nou ús té recolzament de l llicenciant.
- Les transformacions o fusions del material han de ser distribuides 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 amb una xicoteta descripció.

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

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

https://github.com/adtsune/tipologia_cicle_dada

Taula de contribucions al treball

, la qual ha de signar cada integrant del grup
amb les seves inicials. Les inicials representen la confirmació per part del grup que l'integrant ha participat en aquest apartat. Tots els integrants han de participar a cada apartat, per la qual cosa, idealment, els apartats haurien d'estar signats per tots els integrants.

Contribucions	**Signa**
Recerca prèvia	XAVIER VENTURA DE LOS OJOS, ANNA DE LA TORRE SUÑE
Redacció de les respostes	XAVIER VENTURA DE LOS OJOS, ANNA DE LA TORRE SUÑE
Desenvolupament codi	XAVIER VENTURA DE LOS OJOS, ANNA DE LA TORRE SUÑE