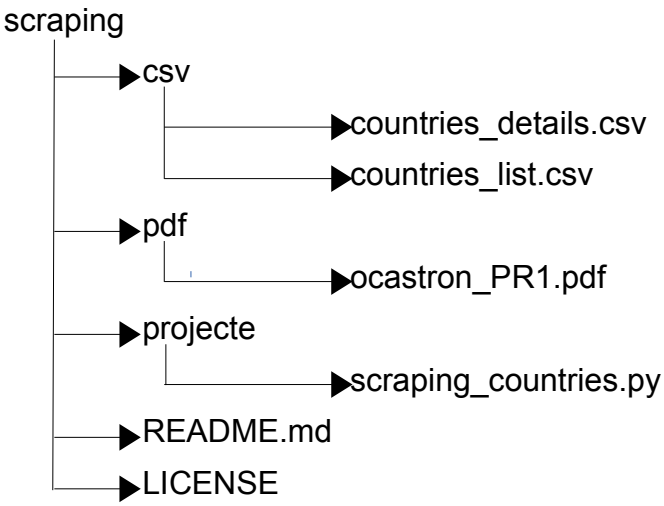
Pràctica1

Autor: **Octavi Castro Nuez**

Estructura dels fitxers a Github:

Carpeta/Fitxer	Contingut a	Descripció
scraping		Carpeta arrel. Conté la resta de carpetes.
csv	scraping	Carpeta amb els fitxers csv del projecte.
pdf	scraping	Carpeta que conté aquest fitxer pdf.
projecte	scraping	Carpeta amb el fitxer py que conté el codi Python del projecte
countries_details.csv	csv	Informació relativa als països africans
countries_list.csv	csv	Llistat dels països d'Àfrica.
LICENSE	scraping	Llicència.
ocastron_PR1.pdf	pdf	Aquest fitxer amb les respostes a les preguntes i el nom de l'autor.
README.md	scraping	Fitxer amb informació del projecte.
scraping_countries.py	projecte	Codi Python del projecte.



Universitat Oberta de Catalunya uoc.edu

Respostes de la Pràctica

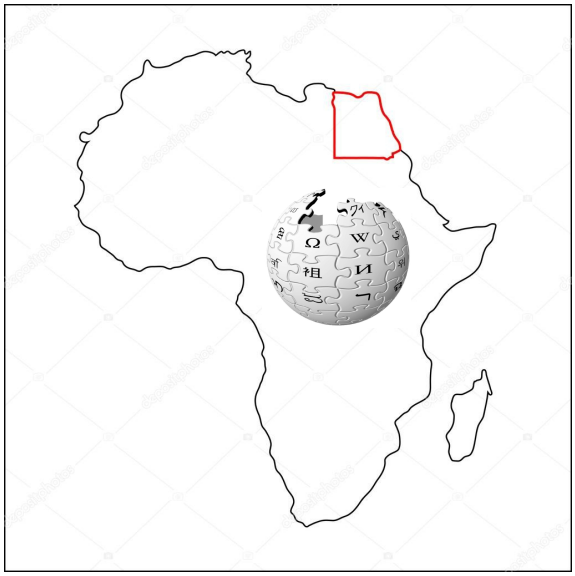
1. Títol del dataset. Cal que poseu un títol que sigui descriptiu.

   Països sobirans i territoris dependents d'Àfrica

2. Subtítol del dataset. Agregueu una descripció àgil del vostre conjunt de dades pel vostre subtítol.

   Llistat dels països sobirans i territoris dependents africans amb informació de cada un d'ells extreta de la wikipedia en anglés

3. Imatge. Agregueu una imatge que identifiqui el vostre dataset visualment



4. Context. Quina és la matèria del conjunt de dades?

   El context es centra en un dels cinc continents de la Terra, el continent africà, i la divisió política per països del mateix, que en aquest cas difereix de la divisió socio-cultural, i part de la informació sobre aquests que trobem a la web de la wikipedia de cadascun d'ells.

5. Contingut. Quins camps inclou?

   La informació obtinguda és divideix en dos datasets.

   El primer dataset és, bàsicament, per a la obtenció de les url dels països africans i conté la següent informació:

   Link: L'adreça web (url) del país.

   Country: nom del país.

   Capital: la ciutat del país que n'és la capital.

   Status: estat en que es troba el país respecte a la seva independència.

   El segon dels datasets conté els detalls de cada país següents:

   Country: nom del país.

   Capital: ciutat que és la capital del país i les seves coordenades.

Official language: l'idioma o idiomes oficials del país.

Government: tipus de govern del país.

Legislature: tipus de legislatura del país.

Gini: darrer coeficient Gini publicat del país i la seva valoració nominal (low, medium, high, very high).

HDI: darrer *Human Development Index* publicat del país, la seva valoració nominal (low, medium, high, very high) i la seva posició mundial.

Currency: moneda del país.

Time zone: ús horari del país.

Drives on the: si en el país. la conducció es fa per la dreta o per l'esquerra

Calling code: codi internacional telefònic.

ISO 3166 code: codi ISO 3166 del país.

Internet TLD: domini web del país.

Quin és el període de temps de les dades i com s'ha recollit?

Les dades s'han extret totes durant el mateix període de temps i al ser unes dades que varien poc en el temps només s'ha dut a terme una recollida de dades mitjançant *scraping web* a https://en.wikipedia.org/wiki/List_of_sovereign_states_and_dependent_territories_by_continent, pel que fa al primer dataset, i a les pàgines web de la wikipedia en anglés per al segon dataset.

6. Agraïments. Qui és propietari del conjunt de dades? Inclou cites de recerca o anàlisi anteriors.

Les dades són propietat de Wikimedia Foundation, https://wikimediafoundation.org/wiki/Home.

Pel que fa al coeficient Gini podem trobar informació a https://en.wikipedia.org/wiki/Gini_coefficient

I del *Human Development Index* podem veure la classificació a https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index.

Agraïments especials a Jimmy Wales, https://en.wikipedia.org/wiki/Jimmy_Wales, i Larry Sanger, https://en.wikipedia.org/wiki/Larry_Sanger, creadors de la wikipedia, i a tots els usuaris que la fan possible.

7. Inspiració.

Durant els darrers anys Europa ha sofert canvis importants en referència a la distribució del territori per països. I això em va fer mirar el continent africà, molt desconegut per a mi, i concretament els països que actualment el conformen i que històricament han pertangut a altres països d'altres continents.

Per què és interessant aquest conjunt de dades?

Podem tenir informació de diferents aspectes dels països africans, tant pel que fa a dades internacionals, com per exemple *Calling code*, com financera, *Gini*, o geogràfica, *Capital* entre d'altres

Quines preguntes li agradaria respondre la comunitat?

La comunitat pot trobar resposta directa, entre d'altres, a:

- Costat pel que es condueix en un país.
- Moneda en curs.
- Codi internacional per trucar-hi.
- Domini d'internet.
- Capital d'un país. i coordenades.
- Zona horària.

I, posterior a l'anàlisi de les dades, entre altres, a:

- Nombre de llengües que es parlen de forma oficial al país. i quina és la més parlada.
- Comparatives, entre països, a través del coeficient Gini o del HDI.
- Tipus de govern o legislació més comú .
- Quants països es troben en una zona horària concreta i quina és la més comú.

8. Llicència. Cal que seleccioneu una d'aquestes llicències i cal dir per què l'heu seleccionada:

   Sembla lògic pensar que tractant-se de dades extretes de la wikipedia la llicència seleccionada sigui la mateixa (CC BY-SA 3.0) o la més propera a aquesta de les que tenim al llistat, per la qual cosa la llicència per a les dades és

   CC BY-SA 4.0

   que és la darrera versió de la que fa anar wikipedia.

9. Codi: Cal adjuntar el codi amb el que heu generat el dataset, preferiblement amb R o Python, que us ha ajudat a generar el dataset

   El codi que s'ha utilitzat en aquest projecte es troba en Python i es troba en el Github dins de la carpeta projecte amb el nom *scraping_countries.py*.

10. Dataset: Dataset en format CSV

    El dataset del projecte el formen dos arxius CSV que es troben a la carpeta csv del Github.