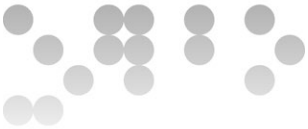
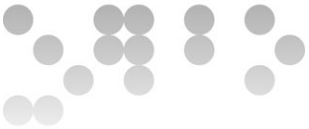


Taula de contingut



Presentació

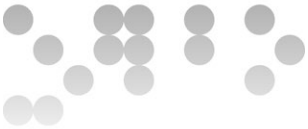
Aquest document forma part la pràctica de *web scraping* de l´assignatura **"Tipología i cicle de vida de les dades"** del **Màster en Ciència de Dades de la UOC**.

L´objectiu principal d´aquesta pràctica és elaborar un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d´extracció de dades.

En particular, s´ha realitzat *web scraping* a la pàgina **Huber Timing**[1] en la que es poden consultar les classificacions i els temps de curses de *running*. En aquest sentit s´han obtingut les dades de la cursa 10K solidària **Hope For Salem 2018**[2].

[1] **http://www.hubertiming.com/**

[2] **https://ugmsalem.org/walkforhope/**



Format d´entrega

La documentació generada en la realització de la pràctica es troba allotjada en GitHub al següent repositori:

- https://github.com/rsanchezs/web-scraping

En aquest repositori es poden trobar els següents fitxers

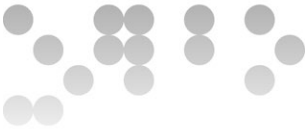
- **src/scraper.py:** conté el *scraper* que obté el conjunt de dades.
- **data/ranking.csv:** arxiu CSV amb el conjunt de dades.

Per altra banda, també es pot visitar el següent **enllaç** que conté la mateixa documentació que aquest document.

Requisits

La implementació del *scraper* s´ha realitzat amb la versió 3.7.0 de Python. Si és necessari caldrà instal.lar els següents mòduls, per exemple s´utilitzem `pip3`:

```
pip3 install beautifulsoup
pip3 install requests
pip3 install html5lib
```



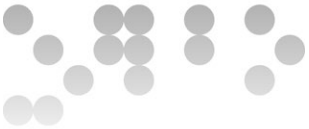
Informació del conjunt de dades

Aquest conjunt de dades consta de 15 variables i 129 observacions. Els tipus dels atributs són els següentes:

- Numèriques:
  - ✓ Discretes: `Chip Time`, `Chip Pace`, `Time to Start`, `Gun Time`
  - ✓ Continues: `Place`, `Bib`, `Age`, `Age Group`, `Age Group Place`
- Categòriques: `Gender`, `City, State`

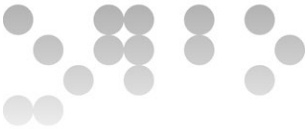
La variable continua `Gender Place` està agrupada en classes on la marca de classe es 9. Les variables `Gender Place` i `Age Group Place` són les posicions respecte el total i estàn representades per un *string*.

Per acabar, s´ha de tenir en compte que els atributs poden contindré valors desconeguts.



Informació dels atributs

Atribut	Descripció
Place	Classificació general
Bib	Dorsal del runner
Name	Nom i cognoms del runner
Gender	Gènere del runner
Age	Edat del runner
City	Ciutat del runner
State	Estat de la ciutat
Chip Time	Temps calculat pel chip
Chip Pace	Ritme enregistrat pel chip
Gender Place	Posició obtinguda respecte gènere
Age Group	Posició obtinguda respecte interval edat
Age Group Place	Posició obtinguda en la mateixa edad
Time to Start	Temps sortida carrera
Gun Time	Gun time = Chip time - time to start
Team	Equip al que pertany el runner



Origen de les dades

HUBR TIMING - Race Timing Servide

HUBER TIMING, LLC. Copyright 2009 - 2018

e-mail: timing@hubertiming.com

Llicència