Taula de contingut

[Presentació 2](#_Toc529821196)

[Format d´entrega 3](#_Toc529821197)

[Requisits 3](#_Toc529821198)

[Informació del conjunt de dades 4](#_Toc529821199)

[Informació dels atributs 5](#_Toc529821200)

[Origen de les dades 6](#_Toc529821201)

[Llicència 6](#_Toc529821202)

## 

## 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](https://estudios.uoc.edu/es/masters-universitarios/data-science/presentacion).

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]](#footnote-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]](#footnote-2).

## 

## 

## 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](https://rsanchezs.github.io/web-scraping/) 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](http://www.hubertiming.com/home)

HUBER TIMING, LLC. Copyright 2009 - 2018

e-mail: [timing@hubertiming.com](mailto:timing@hubertiming.com)

## 

## Llicència



To the extent possible under law, Ruben SANCHEZ SANCHO has waived all copyright and related or neighboring rights to Ranking Hope for Salem 2018 | Pràctica web scraping assignatura Tipología i cicle de vida de les dades de la UOC. This work is published from: España.

1. <http://www.hubertiming.com/> [↑](#footnote-ref-1)
2. <https://ugmsalem.org/walkforhope/> [↑](#footnote-ref-2)