Pràctica 1: Web scraping
Document de respostes

L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realització, s'han de complir els següents punts:

1 CONTEXT

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

La viabilitat econòmica dels mitjans audiovisuals televisius depèn de forma gairebé absoluta de la inversió publicitària que són capaços de captar.

Les agències de mitjans són les empreses encarregades d'obtenir els millors rendiments de la inversió publicitària que fan els seus clients (els anunciants).

A Espanya, de forma general, es fa servir l'audiència com a referència per a determinar la forma de distribuir el budget entre els diferents mitjans.

Per tant, a més audiència, més ingressos potencials pot arribar a aconseguir un canal de televisió. Sovint, l'existència de grups mediàtics que concentren més d'un canal de televisió i que ofereixen als seus anunciants paquets de publicitat com a 'mix' de publicitat en diferents canals, distorsiona la relació entre audiència i ingressos publicitaris, però segueix sent un paràmetre bàsic.

Paradoxalment, la inversió publicitària es fa de forma anticipada, mentre que el valor d'audiència es coneix a posteriori (normalment l'endemà).

Resulta, per tant, important disposar d'informació sobre l'audiència per canals de forma diària per poder aplicar-hi després metodologies i algoritmes de manipulació de dades i poder, d'aquesta manera generar models que ens permetin preveure les audiències que s'obtindran i així, millorar en la optimització de la inversió publicitària.

La web *www.formulatv.com* recull informació sobre canals de televisió: oferta de programació, detalls de

series o presentadors,... i també disposa d'un apartat específic dedicat a les audiències.

Tot i que disposa d'informació sobre televisió de pagament i també als USA, ens centrarem només en la audiència de la televisió en obert (TDT).

2 DEFINIR UN TÍTOL PEL DATA SET

Triar un títol que sigui descriptiu.

3 DESCRIPCIÓ DEL DATA SET

Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

4 REPRESENTACIÓ GRÀFICA

Presentar una imatge o esquema que identifiqui el dataset visualment

5 CONTINGUT

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

6 AGRAÏMENTS

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

7 INSPIRACIÓ

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

8 LLICÈNCIA

Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció:

Released Under CC0: Public Domain License

Released Under CC BY-NC-SA 4.0 License

9 CODI

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

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

El dataset es troba disponible a Github, al repositori:

https://github.com/mboschga/PR1

El fitxer està disponible amb el nom “xxxxxxx.csv”.