

# 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](http://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.*

El nom que creiem, que li encaixa al dataset és:

“Dades diàries d’audiència de televisió a Espanya (1Q19)”

### 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).*

Recull de les audiències diàries obtingudes pels diferents canals de televisió en obert que emeten a Espanya amb cobertura estatal durant el primer quadrimestre del 2019.

### 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.*

El nostre dataset contindrà informació d'audiències de televisió en obert (TDT) a Espanya, per cadena i per dies, tot i que s'ha pogut observar que només hi apareixen els canals d'àmbit estatal, no hi ha canals

autonòmics ni locals.

Tal com s'ha exposat anteriorment, la web "[www.formulatv.com/audiencias](http://www.formulatv.com/audiencias)" conté informació de les audiències, però de forma no agrupada en una sola pàgina, sinó que les dades diàries estan presentades en pàgines web diferents, afegint a la ruta "/aaaa-mm-dd".

D'aquesta manera, per a accedir a les audiències del dia 25 de gener de 2020, cal accedir a <https://www.formulatv.com/audiencias/2020-01-25/>.

Per això, la tècnica de web scraping aplicada només a una pàgina ens hauria donat un data set de un sol dia de dades i per tant, amb només 2 columnes: "cadena" i "audiència", hauria estat suficient (o bé amb una sola fila i tantes columnes com cadenes)

Tenint en compte l'objectiu presentat en el context, té sentit que el dataset tingui dades de més d'un dia. Inicialment, volíem creat el dataset per a un any sencer, però la web no admetia tantes consultes. Malgrat haver aplicat un retard entre consultes, un any requeria una estratègia més avançada. És per això que, finalment, hem capturat les dades del primer quadrimestre (gener a abril de 2019).

S'ha automatitzat el procés i s'ha construït un dataset amb 3 columnes:

- Audiencia: Share d'audiencia en %
- Cadena: nom de la cadena de televisió
- Fecha: data en format aaaa-mm-dd

## 6 AGRAÏMENTS

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

Les dades les hem capturat de la web [www.formulatv.com](http://www.formulatv.com), pàgina editada per l'empresa Noxvo SL, però les dades que publiquen són elaborades per l'empresa Kantar Media.

A Espanya, l'empresa que té l'exclusiva d'explotació comercial de les dades que genera Kantar Media es Barlovento Comunicacion <https://www.barloventocomunicacion.es>

## 7 INSPIRACIÓ

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

L'any 2019 la televisió a Espanya va captar aproximadament 2000 milions d'euros de publicitat, dels gairebé 6000 milions d'euros que es van invertir a Espanya en publicitat en tots els mitjans. És a dir, gairebé un terç de la inversió en publicitat recau sobre un sol suport: la televisió.

La contractació de minuts publicitaris es fa de forma prèvia a la emissió, naturalment. I les dades d'audiència en televisió (a diferència del mitjà digital) no es coneix fins l'endemà.

Per tant, la decisió sobre quin import es destina al mitjà i també quin import es destina a cada cadena no és més que un exercici basat en l'experiència. L'experiència permet als professionals saber que a l'estiu es veu menys la televisió que a la tardor, primavera i hivern, que les nits de diumenge a dijous l'audiència decau més d'hora que les nits de dissabte i diumenge, ...

Seria molt interessant, però poder extreure patrons estacionals i arribar a modelar l'audiència per tal de poder conèixer a priori, sense un error molt elevat, quina audiència tindrà cada cadena l'endemà.

Aquest data set, és interessant perquè ens permet iniciar la busca de les estacionalitats, de moment en un rang diari i setmanal. Podem analitzar els comportaments per dia de la setmana o per mes o per setmana dins del mes.

Les preguntes que ens pot respondre són, de moment: quin dia de la setmana es mira més la tele, quina hora del dia es mira més la tele, quins dies de la setmana es mira més un canal en concret,....

## 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*

*Released Under CC BY-SA 4.0 License*

*Database released under Open Database License, individual contents under Database Contents License*

*Other (specified above)*

*Unknown License*

Considerem adequat que aquest joc de dades sigui publicat sota una llicència de tipus “Released Under CC BY-NC-SA 4.0 License”.

Hem escollit aquesta opció perquè:

- Obliga a declarar el nom del creador del joc de dades i indicar quins canvis s’hi ha fet. D’aquesta manera es reconeix la feina feta per tercers.
- Impedeix que se’n faci un ús comercial. Tot i que en principi és contrari a l’objectiu inicial amb el que s’ha considerat la creació d’aquest joc de dades, l’avís legal de la web de formulatv es prohibeix l’ús comercial de les dades que conté, per tant, hem de respectar aquesta limitació
- Obliga a que les futures contribucions mantinguin els mateixos criteris de publicació, de manera que la voluntat dels seus creadors no es perd.

## 9 CODI

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

S’ha optat per fer servir python com a llenguatge principalment per les llibreries existents per a efectuar web scraping.

S’han creat dos fitxer de codi diferents, ubicats a la carpeta /codi del repositori.

- scraper-1Q-year2019.ipynb: codi que efectua la generació del dataset entre unes dates concretes, en aquest cas de 1 de gener de 2019 a 30 d’abril de 2019.
- scraper-yesterday.ipynb: codi que efectua el web scraping per al dia anterior. És el codi bàsic que s’ha fet servir per analitzar el comportament i poder així generar el codi principal.

## 10 DATASET

*Publicar el dataset en format CSV a Zenodo amb una xicoteta descripció.*

El dataset en format csv està disponible al repositori, a la carpeta /fitxers. Seguint amb el que es requereix, s'accedeix a <https://zenodo.org> usant el mateix login i pwd que usem a Github i es publica el dataset.



April 4, 2020

Dataset Open Access

### Dades diàries d'audiència de televisió a Espanya (1Q19)

Marta Martínez; Meritxell Bosch

Data collector(s)

Meritxell Bosch; Marta Martínez

Recull de les audiències diàries obtingudes pels diferents canals de televisió en obert que emeten a Espanya amb cobertura estatal durant el primer quadrimestre del 2019

[Preview](#)

Files (1.1 MB)

Name

Size

export\_1Q\_year\_2019.csv

1.1 MB

[Preview](#)
[Download](#)

md5:eea9b7966b62fc69924be5facd9df34a

**Publication date:**

April 4, 2020

**DOI:**

DOI 10.5281/zenodo.3739731

**Keyword(s):**
[audiència](#) [TDT](#) [televisió](#) [Espanya](#) [2019](#)  
[audiències](#) [canals](#)
**License (for files):**
[Creative Commons Attribution 4.0 International](#)

## 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> i també a Zenodo, un cop l'hem pujat amb el DOI: 10.5281/zenodo.3739731

El fitxer està disponible amb el nom "export\_1Q\_year\_2019.csv".

### Versions

Version 1

Apr 4, 2020

10.5281/zenodo.3739731

**Cite all versions?** You can cite all versions by using the DOI 10.5281/zenodo.3739730. This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

### Share



### Cite as

Marta Martínez, & Meritxell Bosch. (2020). Dades diàries d'audiència de televisió a Espanya (1Q19) [Data set]. Zenodo. <http://doi.org/10.5281/zenodo.3739731>

Start typing a citation style...

## 12 CONTRIBUCIONS

CONTRIBUCIONS	INTEGRANT	SIGNATURA
<b>Recerca prèvia</b>	Meritxell Bosch	MB
	Marta Martínez	MM
<b>Redacció de les respostes</b>	Meritxell Bosch	MB
	Marta Martínez	MM
<b>Desenvolupament codi</b>	Meritxell Bosch	MB
	Marta Martínez	MM