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.

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

"Dades diàries d'audiència de televisió a Espanya"

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.

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" 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, hem creat el dataset per a un any sencer.

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 (string de text)

Data: 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, pàgina editada per l'empresa Noxvo SL, però les dades que publiquen són elaborades por 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.

- main.ipynb: codi principal que efectua la generació del dataset entre unes dates concretes. Per obtenir les dades de cada dia, es fa una crida al segon codi: scraper.ipynb
- scraper.ipynb: codi que efectua el web scraping per a una data concreta

10 DATASET

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

S'accedeix a https://zenodo.org usant el mateix login i pwd que usem a Github i es publica el dataset.

INSERTAR CAPTURA PANTALLA

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".

DOI= es el codi que ens donarà el ZENODO al pujar el dataset

12 CONTRIBUCIONS

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