# Pràctica 1 - *Web Scraping*

## Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació. [0.25 punts]

El consum televisiu a Espanya es mesura des de fa més de 30 anys. Tot i que els hàbits dels espectadors han anat canviant, la clàssica graella de programes encara és molt rellevant. Com també sabem, la programació de TV d’una corporació competeix amb la resta d’operadors de mitjans. En aquest context és important, doncs, conèixer al detall la composició de de tota l’oferta televisiva. Una Base de Dades de Continguts unificada podria permetre analitzar i treure conclusions sobre quina estratègia utilitza cadascun dels competidors, i veure quines són les tendències a cada moment.

I a partit del context que hem descrit, podem saber que qui millor coneix la oferta de Programació televisiva són els mateixos propietaris de les cadenes. Així que probablement la informació més actualitzada la trobarem als portals propis de cada operador.

Volem que aquesta activitat serveixi de prova de concepte per demostrar el potencial dels *websites* de mitjans per obtenir les dades de la graella de TV i el detall dels continguts que la composen. En aquest cas concret, analitzem la companyia que manté el lideratge d’audiència a nivell estatal: [MEDIASET](https://www.mediaset.es/).

## Definir un títol pel *dataset*. Triar un títol que sigui descriptiu. [0.25 punts]

Oferta de programes de TV de diferents cadenes en l’àmbit de l’Estat espanyol.

## Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat). [0.25 punts]

*Explicar amb més detall el que es vol emmagatzemar, però sense arribar al nivell de la composició de camps. Es pot agafar com exemple:* [*https://www.kaggle.com/shivamb/netflix-shows*](https://www.kaggle.com/shivamb/netflix-shows)

## Representació gràfica. Presentar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit. [0.25 punts]

*A pensar i intentar trobar una imatge que sigui original i alhora significativa*.

## Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit. [1 punt]

*Cal preparar l’aplicació abans de respondre això, doncs llavors tindrem el detall dels camps que hem pogut capturar i el període que avarca.*

## Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars. [1.5 punts]

*Cal verificar el propietari de la pàgina web (tal com s’explica als apunts) i intentar esbrinar si algú més a “scrapejat” aquestes dades. Veure també al fòrum la pregunta (Dubte pregunta 6) i resposta que hi ha al respecte.*

## Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. És necessari comparar amb les anàlisis anteriors presentades a l’apartat 6. [1.5 punts]

*Aquí també es pot mirar un exemple como el de Netflix:* [*https://www.kaggle.com/shivamb/netflix-shows*](https://www.kaggle.com/shivamb/netflix-shows)*.*

## Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció: [1 punt]

### 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

*Pensar el que cal fer en aquest exercici.*

## Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R. [2 punts]

*Desenvolupar el notebook i intentar introduir una mica de complexitat per intentar aconseguir els 2 punts, alhora que fem la pràctica més interessant i profitosa. Una proposta és utilitzar scrapy, que se surt de la simplicitat del BeatifulSoup i permet fer crawling per inspeccionar vàries pàgines al mateix temps que es fa scraping.*

## Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció. [2 punts]

*Caldrà veure com es fa això.*

## CONTRIBUCIONS

|  |  |
| --- | --- |
| ACTIVITAT | SIGNATURA |
| Recerca prèvia | CM, JB |
| Redacció de les respostes | CM, JB |
| Desenvolupament codi | CM, JB |