Alumne: Oriol Roqué Paniagua

Tipologia i cicle de vida de les dades – PRAC 1

Current News

Context

En aquesta pràctica de Web Scrapping s'ha optat per realitzar una extracció del portal digital de notícies 324, propietat del CCMA, que es recollirà en un dataset anomenat **CurrentNews**. L'objectiu és recollir la informació més rellevant de cada notícia que estigui activa en el moment de l'execució del script.

Prèviament s'havia valorat extreure informació d'un portal de venda d'instruments musicals i altres objectes relacionats amb la música, però després de fer una lectura completa dels termes i condicions, era necessari una confirmació per escrit del propietari per a poder realitzar web scrapping. Per aquest fet, s'ha optat per fer una extracció d'un portal de notícies. S'ha escollit el de 324 perquè en comparació a altres portals, la portada conté en general el títol i la url de la notícia en qüestió (a excepció de les primeres notícies). Aquest fet implica una complexitat afegida, que és la de accedir notícia a notícia per tal d'extreure la informació.

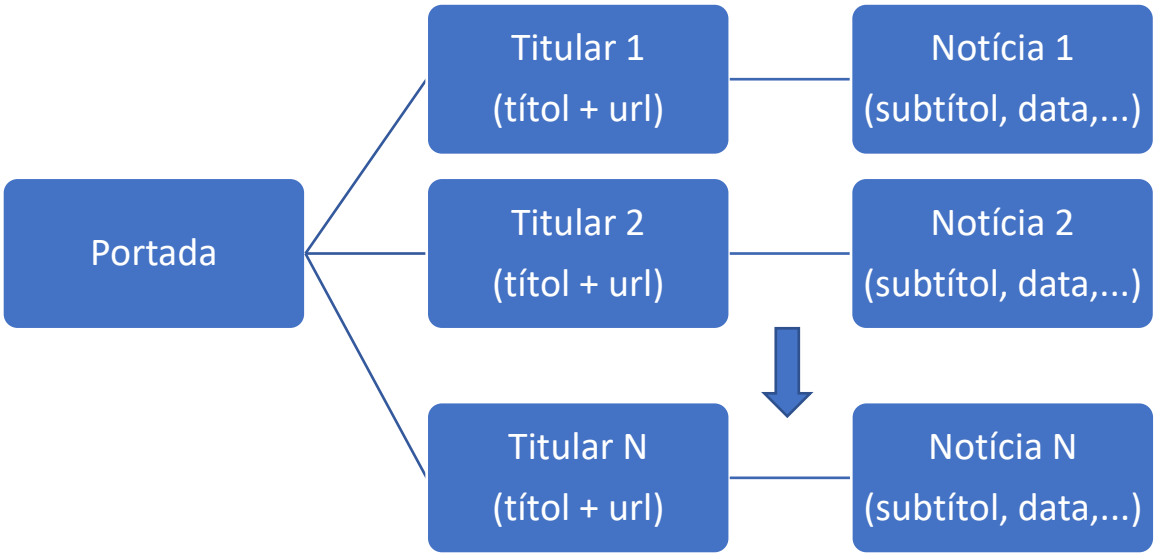
Definició de títol

El dataset resultant s'anomena CurrentNews, degut a què conté la informació bàsica de les notícies que presenta el portal en el moment d'execució.

Descripció del dataset

El dataset conté 11 variables amb la informació rellevant de cada notícia que aparegui en la portada principal de 324 en el moment de l'extracció. Aquestes variables són la data d'extracció, la url, el títol, el subtítol, la localització, la data i hora de redacció, la data i hora d'actualització, les etiquetes i el nombre d'etiquetes.

Representació gràfica



Contingut

L'estructura del dataset Current News és la següent:

Nom	Tipus	Descripció	Pot ser nul?
extraction_date	Datetime	Metadada que recull el moment en el temps de la recol·lecció de les dades de la notícia	No
url	String	url amb l'enllaç a la notícia	No
title	String	Títol de la notícia	No
subtitle	String	Subtítol que apareix sota la capçalera de la notícia	No
location	String	Si apareix, lloc on ocorre el fet esmentat en la notícia	Sí
publish_date	Date (dd/MM/yyyy)	Data de publicació o redacció de la notícia	No
publish_hour	Hour (hh:mm)	Hora de publicació o redacció de la notícia	No
last_update_date	Date (dd/MM/yyyy)	Si apareix, data d'actualització de la notícia	Sí
last_update_hour	Hour (hh:mm)	Si apareix, hora d'actualització de la notícia	Sí
tag	String	Si apareix, conjunt d'etiquetes vinculades a la notícia	Sí
num_tags	Integer	Nombre d'etiquetes vinculades a la notícia	No

L'extracció comença per accedir a la pàgina web https://www.ccma.cat/324/ i realitzar una primera lectura de la portada, per determinar els títols i les url de les notícies a analitzar. Seguidament, accedeix a la url de cada notícia i extreu les dades principals per tal de completar el dataset. Malgrat que l'estructura de les notícies és similar, hi ha elements que poden variar: per exemple, una notícia sempre serà creada en una data concreta, però la data d'actualització dependrà de que la notícia sigui modificada.

La volumetria del dataset resultant dependrà del nombre de notícies actives en el moment d'execució, però sol ser de l'ordre d'una vintena o trentena de registres. En el cas concret presentat, correspon a l'execució de les 8 del vespre del divendres 08/11/2019. En la següent imatge es presenta un exemple de les dades recollides d'una notícia en concret:

extraction_date	2019-11-08 20:04:49.655492
url	https://www.ccma.cat/324/el-tractament-per-prevenir-el-vih-ja-esta-disponible-a-catalunya-gratuitament/noticia/2961912/
title	El tractament per prevenir el VIH ja està disponible a Catalunya gratuïtament
subtitle	Es dispensarà en les 19 unitats de sida i en dos centres comunitaris de detecció de malalties de transmissió sexual
location	Barcelona
publish_date	08/11/2019
publish_hour	15.18
last_update_date	08/11/2019
last_update_hour	18.48
tag	Salut
num_tags	1

Agraïments

El propietari del conjunt de dades és la Corporació Catalana de Mitjans Audiovisuals (CCMA), atès la procedència de l'extracció. L'extracció es realitza seguint els criteris de l'avís legal del web. En concret, s'ha extret en relació als principis següents:

"Coses que et deixem fer amb el que publiquem:

Podràs consultar la informació en qualsevol moment i descarregar-la als teus dispositius si es compleixen les condicions següents:

1. Que l'ús que en facis sigui compatible amb els principis que et comentem en aquest avís i la resta de serveis digitals de la CCMA .
2. Que es faci només per obtenir la informació per a l'ús personal i privat. Que no la utilitzis amb finalitats comercials o per a la seva distribució, comunicació pública, transformació o descompilació.
3. Que no modifiquis cap dels continguts.
4. Que no copiïs o distribueixis separadament del text o de la resta de les imatges que l'acompanyin cap gràfic, icona o imatge de la web, aplicacions mòbils i de televisió connectada."

De la mateixa manera, l'extracció del conjunt de dades complirà la legalitat sempre que:

"A títol enunciatiu, el material que trobaràs en el portal no el podràs utilitzar per:

1. Incórrer en activitats il·lícites, il·legals o contràries a la bona fe i l'ordre públic.
2. Difondre continguts o propaganda de caràcter racista, xenòfob, pornogràfic, d'apologia del terrorisme o que atempti contra els drets humans.
3. Provocar danys en els sistemes físics i informàtics de la CCMA o dels seus proveïdors de serveis.
4. Intentar accedir a comptes d'altres usuaris o manipular les identitats de tercers."

El conjunt de drets i obligacions es pot llegir íntegrament al web https://www.ccma.cat/avis-legal/.

Inspiració

Una aplicació que podria tenir aquest procés de web scrapping seria generar un històric de notícies mitjançant una sèrie d'execucions programades. Si es tractessin els possibles casos de duplicats, es podria acabar generant un dataset acumulat, que resultaria en un conjunt de dades interessant per analitzar. Aquest es podria utilitzar per exemple per determinar la quantitat o distribució de notícies per zona geogràfica o tag, o si s'anotés la posició de la notícia respecte la portada, intentar determinar quins tipus de tags solen estar relacionats amb quines posicions. També es podria determinar el temps mig que duren certes tipologies de notícies en actiu. En cas que el propietari volgués utilitzar aquest tipus d'informació, podria generar recomanacions personalitzades en funció dels tags o localitzacions de les notícies que visita un usuari en concret.

Llicència

La llicència escollida pel dataset, d'acord amb l'avís legal del CCMA, correspon a la llicència Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0). En concret, la llicència obliga a:

- Donar crèdit de la procedència de les dades, que correspon al CCMA.
- No es pot utilitzar amb finalitats comercials.
- Per restringir de cares a les obligacions d'usuari del CCMA, no es permet realitzar modificacions sobre el propi dataset.

El detall de la llicència es pot trobar a https://creativecommons.org/licenses/by-nc-nd/4.0/.

Contribucions	Signa
Recerca prèvia	ORP
Redacció de les respostes	ORP
Desenvolupament de les respostes	ORP