Tipologia i Cicle de Vida de les Dades
PRAC 1: Web scraping.

Òscar del Álamo i Guaus

14 d'abril de 2020

Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'extracció de dades. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github (https://github.com) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius del vostre lliurament. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu revisar aquests exemples com a guia:

- Exemple: `https://github.com/rafoelhonrado/foodPriceScraper`
- Exemple complex: `https://github.com/tteguayco/Web-scraping`

Competències

En aquesta PAC es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-ho i resoldre-ho.
- Capacitat per aplicar les tècniques específiques de web scraping.

Objectius

Els objectius concrets d'aquesta Prova d'Avaluació Contínua són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants que el seu tractament aporta valor a una empresa i la identificació de nous projectes analítics.
- Saber identificar les dades rellevants per dur a terme un projecte analític.
- Capturar dades de diferents fonts de dades (tals com a xarxes socials, web de dades o repositoris) i mitjançant diferents mecanismes (tals com queries, API i scraping).
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.

- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la PAC a realitzar

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.

Avui en dia, hi ha molta oferta musical a una multitud diversa de plataformes. Un dels grans problemes que hi ha és descobrir contingut musical de qualitat, sigui antic o nou, ja que moltes vegades se'ns redueix el món a la nostra bombolla de cada plataforma o xarxa social i les recomanacions que ens fan els algoritmes deixen de ser interessants o simplement es tornen repetitives. Aquesta immensa oferta que hi ha avui en dia, pot tenir la desavantatge de què hi hagi molta música que es perdi entre terabytes d'informació.

Byte FM és una ràdio "d'internet"o online nascuda a Alemanya i fundada pel periodista musical Ruben Jonas Schnell. Aquesta ràdio online ha anat guanyant popularitat al llarg dels anys i, segons la seva presentació, tenen uns 850.000 oients mensuals. Una de les seves principals característiques pel qual molta gent la valora és perquè la seva programació musical no utilitza cap mena de software de rotació de música, sinó que és 100% creada per humans.

2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

El títol pel dataset seria: "ByteFM 2020 March Mornings", ja que recull totes les cançons que han sona al programa "ByteFM am Morgens"durant els matins de març del 2020.

3. 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).

El dataset resultant recull la informació de les cançons "scrapejades"que han sonat durant el març de 2020 a la secció "ByteFM am Morgens". És a dir, es pot saber en quin programa han sonat, quin dia han sonat i la informació disponible que hi ha a ByteFM per poder buscar la cançó i tornar-la a escoltar. Tot i això, el codi permet scrapejar molt més contingut que no pas un mes i un programa, sinó que es pot cercar per diferents períodes de temps i diferents programes o inclús tota la programació sense filtrar per programa.

4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment.

Figura 1: ByteFM 2020 March Mornings

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

El contingut scrapejat forma part del l'emissió del programa "ByteFM am Morgens"durant el més de març del 2020. El dataset està format per 6 columnes:

- program: el nom del programa
- date: la data d'emissió en format 'YYYY-MM-DD'
- artist: artista de la cançó
- title: títol de la cançó
- album: àlbum de la cançó. No sempre està disponible.
- label: el segell que produït la cançó. A vegades n'hi ha dos. Generalment un fa referència al segell musical que ha produït el vinil i l'altre al cd. No sempre hi ha aquesta informació disponible.

La informació s'obté a través d'scrapejar la informació de tots els programes que té ByteFM, disponible des del gener del 2008 fins al dia d'avui. Per tal d'extreure el dataset generat per la pràctica s'ha executat l'scraper de la següent manera:

A més, a nivell de codi, també s'ha intentat no carregar el servidor posant delays entre peticions i creant un cache amb mongodb per evitar de descarregar contingut que ja haguem descarregat abans.

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

Les dades han estat recollides de la web de ByteFM: https://www.byte.fm/. Així doncs, el propietari del conjunt de dades n'és la pròpia radio.

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

A nivell individual aquest dataset pot servir per descobrir i tornar a escoltar els temes que han sonat al programa de ràdio escollit.

A nivell pràctic, a partir de la informació recollida, es podria agafar més informació fent servir APIs públiques o "scrapejan"altres webs per ampliar el dataset. Alguns casos d'ús serien:

- Utilitzar la api d'spotify: https://developer.spotify.com/documentation/web-api/ per trobar característiques de les cançons i intentar fer un anàlisis utilitzant machine learning i descriptors de baix nivell de les cançons.
- Es podria utilitzar la cerca feta a través de la API d'spotify per crear llistes de reproducció a spotify amb les cançons trobades. En aquest article se'n pot veure un exemple: https://towardsdatascience.com/using-python-to-create-spotify-playlists-of-the-samples-on-an-album-e3f20187ee5e
- Utilitzar la api de discogs: https://www.discogs.com/developers/# per obtenir més informació de les cançons i saber on trobar botigues de proximitat on en venguin còpies en físic.
- Scrapejar webs que continguin informació de lletres de les cançons i fer anàlisis de sentiments de les cançons emeses per programes o períodes, tenint en compte que es podria ampliar el dataset amb molta més informació de ByteFM.

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)

La llicència seleccionada és la BY-NC-SA 4.0 License, ja que no permet un ús comercial de les dades. Els motius són diversos, però el principal és evitar que altres emisores o empreses relacionades del sector s'aprofitin de la feina humana que hi ha al darrera de cada programa de ByteFM.

Per contra, aquesta llicència permet compartir i transformar el material. D'aquesta manera, es podria seguir treballant amb el material des d'un punt de vista d'investigació o fer-ne ús particular.

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

La informació del codi amb el que s'ha generat al dataset es pot trobar a: https://github.com/poskinx/bytefm_scraper.

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

El link a Zenodo és el següent: https://zenodo.org/record/3751710.

11. Lliurar. Presentar el treball amb el DOI del dataset a Github.

El link de github és el següent: https://github.com/poskinx/bytefm_scraper.

Contribucions al treball

Òscar del Álamo i Guaus.

Bibliografia

Els següents recursos són d'utilitat per la realització de la PAC:

- Calvo, M., Pérez, D., Subirats, L. (2019). Introducció al cicle de vida de les dades. Editorial UOC.
- Subirats, L., Calvo, M. (2019). Web Scraping. Editorial UOC.
- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.