TIPOLOGIA I CICLE DE VIDA DE LES DADES

Pràctica 1
Web scraping

Enric Pou Robert
Màster en Data Science
UOC - 2019/2020

Dataset: Lletra de totes les cançons del grup "Metalica"

Descripció del dataset

El conjunt de dades generat com a resultat d'aquesta pràctica és el conjunt de totes les lletres de les cançons del conegut grup de música "*Metallica*" al llarg de la seva trajectòria. En aquest dataset s'aporta informació, a més, de l'artista, l'àlbum al qual pertany la cançó i sobre cadascuna de les cançons d'aquest grup en concret. Vull dir que, s'ha escollit aquest grup pel fet de ser un grup famós, però que el codi desenvolupat no és estanc a aquest grup, sinó que permet a l'usuari obtenir els resultats fruit de la cerca de qualsevol grup, àlbum o cançó.

Imatge descriptiva

Context

Tal com s'ha explicat anteriorment, el dataset proporcionat és el recull de totes les cançons del grup "*Metallica*" al llarg de la seva trajectòria. És per això, que s'ha buscat una web de referència en el que si pogués trobar una gran ventall de lletres de cançons de molts grups. En aquest cas es tracta de la web https://www.azlyrics.com. Aquesta web, proporciona informació bàsica sobre l'artista, l'album i cadascuna de les cançons de que es troben en la seva base de dades.

Contingut

Per a cadascuna de les cançons resultants, es recullen les següents característiques:

- ***artist:*** El nom de l'artista.
- ***album_name:*** El nom de l'album el qual pertany la cançó.
- ***album_year:*** El nom de l'any el qual va ser publicat l'album
- ***album_category:*** La categoria de l'album
- ***song_name:*** El nom de la canço
- ***song_genre:*** El genere de la cançó
- ***song_lyrics:*** La lletra de la canço, després de ser netejada.

Cal dir que la web de *azlyircs* té recopilada tota la discografia de cadascun dels artistes que tenen a la seva base de dades. Per tant, el contingut del dataset serà totes les cançons de la banda.

Com ja s'ha comentat anteriorment, el codi desenvolupat permet a l'usuari fer qualsevol tipus de consulta. En concret, per extreure aquest dataset s'ha executat:

```
python main.py —-search_by 'arist' —-search 'metallica' -o metallica_lyrics.csv
```

El que fa el codi és buscar el nom de la banda a la web, i itera per a cadascun dels resultats i n'extreu (entre d'altres paràmetres) la lletra de les cançons.

Vull comentar que, *azlyrics*, té implementat un sistema de bloqueig de IP que detecta i bloqueja diferents casos: tant en el cas de fer consultes des de maquines virtuals al *cloud* (tals com AWS, DigitalOcean, etc) com en el cas si es detecten moltes consultes consecutives des de la mateixa IP. Per això s'ha desenvolupat un sistema de consultes en *batch*, per a fer pauses (paràmetres configurables) entre un finestra de consultes consecutives i evitar esser bloquejat. Tanmateix, que l'usuari executi el codi sota el seu risc de esser bloquejat.

Agraïments

Les dades han estat recol·lectades des de la web pública de *azlyrics*. Per a exteure la informació s'ha utilitzat tècniques de web scraping per a extreure la informació dels resultats en HTML (codi HTML o JS) de la seva web.

Inspiració

La música sempre m'ha apassionat. És per això que treballo en el sector des de fa més de 3 anys i m'he donat compte que es té informació sobre gran quantitat d'aspectes de la música: bpm, genere, any, charts, posició a rankings, etc. Tot i això, es té poca informació sobre l'intenció emocional o de significat de cada cançó, i alhora grup, que fa que transmeti quelcom únic.

Amb el conjunt de dades proporcionats es pot fer anàlisis de sentiments mitjançant algoritmes de processament de llenguatge natural basant-se en la lletra de les cançons. Es podria extreure el sentiments principals de cada cançó analitzant la seva lletra. Es per això que s'ha facilitat la feina mitjançant un procés de neteja de la lletra de les cançons. A partir d'aquests anàlisis, i tinguen la lletra de cadascun dels albums, a més es podria saber la evolució de les emocions de cadascuna de les etapes de la banda.

Vull remarcar, que aquest dataset només aporta la lletra de les cançons d'una banda en concret, però el codi dona la llibertat d'obtenir la lletra de les cançons a partir de diferents paràmetres de recerca. Així aquest estudi no només es podria centrar en una banda en concret sinó en un conjunt de bandes diferents que desitgés l'usuari, per més tard extreure evolucions de sentiments i extreure conclusions. Una pràctica interessant sería la de buscar els referents musicals reivindicatius d'una època i extreure els sentiments conjunts de les cançons. Així, podríem saber l'evolució de les lluites ideològiques de cada època.

Llicència

La llicència que he escollit és la **CC BY-NC-SA 4.0 License**. Principalment, he escollit aquesta llicència ja que no permet l'ús comercial de les dades. Clarament, es pot observar a la web de *azlyrics* el següent comentari: "*Usage of azlyrics.com content by any third-party lyrcs provider is prohibited by our licensing agreement.*". Per tal de respectar el propietari de les dades, i evitar que aquest dataset acabi en mans d'una empresa proveïdora de lletres de cançons, aquesta llicència permet que els datasets resultants dels meus script no podran ser usat amb fins comercial.

A part d'això, garanteix en les seves clàusules la:

- *Llibertat per a compartir el material en qualsevol format i mitja.*
- *Llibertat de transformar, i construir sobre el material.*

Ara bé, sota els següents termes:

- *No permet que el material s'utilitzi amb fins comercials.* Per tant ens assegurem que les lletres de les cançons no acabin en mans de empreses, possibles empreses de lletres de cançons.
- *En cas de transformació o desenvolupament s'ha de continuar usant la mateixa llicència.* Per tant es respecta la decisió de la llicència escollida.

Codi i dataset

Trobareu tota la informació sobre el codi i el dataset resultat de la pràctica al repositori:

https://github.com/epou/lyrics_scraper

Recursos

1. *"Beautiful Soup Documentation - Beautiful Soup 4.4.0.", https://www.crummy.com/software/BeautifulSoup/bs4/doc/*
2. *Lawson, R (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.*
3. *Al Sweigart (2015). Automate the boring stuff with Python. Chapter 11. Web Scraping.*