

# PRA 1

Tipologia i cicle de vida de les dades

**FORMAR**  
**TRANS-**  
**FORMAR**



Universitat  
Oberta  
de Catalunya



La UOC posa a disposició la present plantilla corporativa per a l'elaboració de documents i presentacions docents, acadèmiques i administratives d'ús exclusiu per part dels usuaris autoritzats per la seva vinculació amb la UOC: estudiants, professors, personal docent col·laborador, així com personal de gestió i de recerca de la UOC.

Queda prohibida la divulgació, difusió, distribució, còpia o utilització de qualsevol classe d'aquesta plantilla corporativa o del seu contingut fora de l'àmbit esmentat sense la prèvia autorització expressa de la UOC.

La UOC pone a disposición la presente plantilla corporativa para la elaboración de documentos y presentaciones docentes, académicas y administrativas para uso exclusivo por parte de los usuarios autorizados por su vinculación con la UOC: estudiantes, profesores, personal docente colaborador, así como personal de gestión y de investigación de la UOC.

Queda prohibida la divulgación, difusión, distribución, copia o utilización de cualquier clase de la misma o de su contenido fuera del ámbito mencionado sin la previa autorización expresa de la UOC.

*The UOC provides this corporate template for teaching, academic and administrative documents and presentations. It is exclusively for authorized users: UOC students, faculty, course instructors, tutors or administrative and research staff.*

*The template and its contents must not be used, copied or distributed beyond this ambit without the express prior authorization of the UOC.*



## Context

En aquest treball s'ha realitzat web scraping a la pàgina web <https://pitchfork.com/best/>. Aquesta pàgina conté un recull dels millors àlbums de música, cançons i reedicions nous i per cada un d'ells conté una revisió.

En aquest Web Scraping s'ha volgut emmagatzemar en un fitxer CSV els noms dels àlbums, cançons i re-edicions juntament amb els noms dels artistes, el gènere i les URL de les revisions de cada una de les cançons i de la imatge de la portada.

Per fer-ho hem utilitzat les llibreries Requests, BeautifulSoup, Pandas i Urllib de Python.

## Descripció del dataset

Com s'ha comentat en l'apartat anterior, en aquest conjunt de dades hem volgut recollir tots els noms dels artistes, cançons o àlbums, gènere i url de les revisions de les diferents cançons i àlbums considerades com les millors. Per fer-ho hem obtingut sis atributs diferents. Els comentem a continuació:

- 'artist': on trobem els noms dels artistes en format String.
- 'title': on trobem els noms dels àlbums i cançons en format String.
- 'genre': on trobem el gènere de l'àlbum o cançó en format String.
- 'url': on trobem les urls de les revisions dels àlbums o cançons en format String.
- 'type': on trobem si es tracta d'un àlbum o d'una cançó en format String.
- 'image': on trobem la url de la imatge en format String.

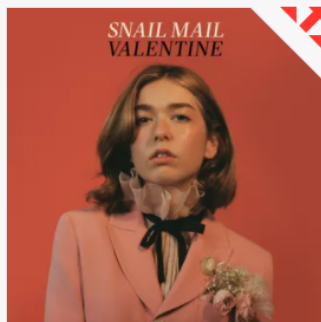
## Representació gràfica

A continuació mostren una representació gràfica de la informació que s'ha pogut extreure de la pàgina web.

# Best New Music

Launched in 2003, Best New Music is Pitchfork's way of highlighting the finest music of the current moment.

BEST NEW ALBUM



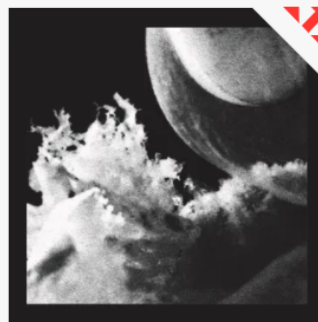
**Snail Mail**

Valentine

ROCK

BY: OLIVIA HORN  
NOVEMBER 4 2021

BEST NEW TRACK



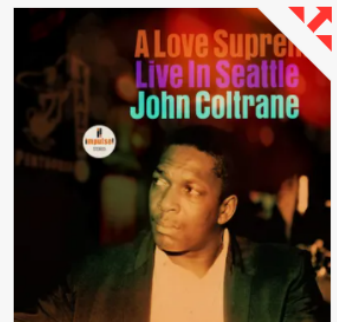
**Jlin**

"Embryo"

ELECTRONIC

BY: MATTHEW ISMAEL RUIZ  
OCTOBER 21 2021

BEST NEW REISSUE



**John Coltrane**

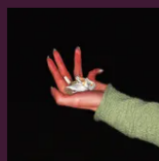
A Love Supreme: Live in Seattle

JAZZ

BY: MARK RICHARDSON  
OCTOBER 27 2021

## Best New Albums

ALL BEST NEW ALBUMS



**aya**

im hole

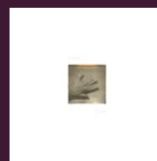
ELECTRONIC



**The War on Drugs**

I Don't Live Here  
Anymore

ROCK



**Grouper**

Shade

EXPERIMENTAL



**Xenia Rubinos**

Una Rosa

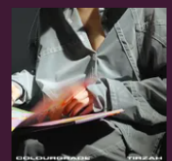
ROCK



**Wiki**

Half God

RAP




**Tirzah**

Colourgrade

EXPERIMENTAL


## Best New Tracks

ALL BEST NEW TRACKS




**The War on Drugs**  
"I Don't Live Here Anymore"

ROCK




**Snail Mail**  
"Valentine"

ROCK




**Water From Your Eyes**  
""Quotations""

ROCK




**Big Thief**  
"Little Things"

ROCK



**Indigo De Souza**  
"Hold U"

ROCK

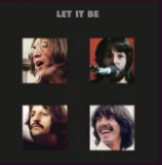


**Low**  
"Days Like These"

ROCK


## Best New Reissues

ALL BEST NEW REISSUES




**The Beatles**  
*Let It Be (Super Deluxe)*

ROCK




**Pastor T. L. Barrett and the Youth for Christ Choir**  
*I Shall Wear a Crown*

ROCK




**Trashcan Sinatras**  
*I've Seen Everything*

ROCK




**Laura Nyro**  
*American Dreamer*

ROCK



**Alice Coltrane**  
*Kirtan: Turiya Sings*

JAZZ / EXPERIMENTAL



**"Blue" Gene Tyranny**  
*Degrees of Freedom Found*

EXPERIMENTAL

Al nostre conjunt de dades o veiem de la següent manera:

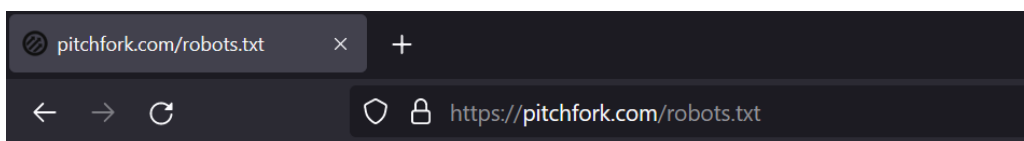
artist	title	genre	url	type	image
Snail Mail	Valentine	Rock	<a href="https://pitchfork.com/reviews/albums/snail-mail-valentine/">https://pitchfork.com/reviews/albums/snail-mail-valentine/</a>	album	<a href="https://media.pitchfork.com/photos/61412ae9d54eda5d82ec71">https://media.pitchfork.com/photos/61412ae9d54eda5d82ec71</a>
Jlin	Embryo	Electronic	<a href="https://pitchfork.com/reviews/tracks/jlin-embryo/">https://pitchfork.com/reviews/tracks/jlin-embryo/</a>	track	<a href="https://media.pitchfork.com/photos/6171861448d55524f73517">https://media.pitchfork.com/photos/6171861448d55524f73517</a>
John Coltrane	A Love Supreme	Jazz	<a href="https://pitchfork.com/reviews/albums/john-coltrane-a-love-supreme/">https://pitchfork.com/reviews/albums/john-coltrane-a-love-supreme/</a>	album	<a href="https://media.pitchfork.com/photos/612795465537dce55d2ccd">https://media.pitchfork.com/photos/612795465537dce55d2ccd</a>
Aya	Im Hole	Electronic	<a href="https://pitchfork.com/reviews/albums/aya-im-hole/">https://pitchfork.com/reviews/albums/aya-im-hole/</a>	album	<a href="https://media.pitchfork.com/photos/61761cfff0e49067279321">https://media.pitchfork.com/photos/61761cfff0e49067279321</a>
The War on Drugs	I Don't Live Here Anymore	Rock	<a href="https://pitchfork.com/reviews/albums/the-war-on-drugs-i-dont-live-here-anymore/">https://pitchfork.com/reviews/albums/the-war-on-drugs-i-dont-live-here-anymore/</a>	album	<a href="https://media.pitchfork.com/photos/60f1a8e8c64eabe66d59c1">https://media.pitchfork.com/photos/60f1a8e8c64eabe66d59c1</a>
Grouper	Shade	Experimental	<a href="https://pitchfork.com/reviews/albums/grouper-shade/">https://pitchfork.com/reviews/albums/grouper-shade/</a>	album	<a href="https://media.pitchfork.com/photos/60fb1c780f763fef3c3c6bf">https://media.pitchfork.com/photos/60fb1c780f763fef3c3c6bf</a>
Xenia Rubinos	Una Rosa	Rock	<a href="https://pitchfork.com/reviews/albums/xenia-rubinos-una-rosa/">https://pitchfork.com/reviews/albums/xenia-rubinos-una-rosa/</a>	album	<a href="https://media.pitchfork.com/photos/60eda941f58acdeaaad0e4">https://media.pitchfork.com/photos/60eda941f58acdeaaad0e4</a>
Wiki	Half God	Rap	<a href="https://pitchfork.com/reviews/albums/wiki-half-god/">https://pitchfork.com/reviews/albums/wiki-half-god/</a>	album	<a href="https://media.pitchfork.com/photos/612d9d52f64ed03c58cd3b">https://media.pitchfork.com/photos/612d9d52f64ed03c58cd3b</a>
Tirzah	Colourgrade	Experimental	<a href="https://pitchfork.com/reviews/albums/tirzah-colourgrade/">https://pitchfork.com/reviews/albums/tirzah-colourgrade/</a>	album	<a href="https://media.pitchfork.com/photos/60d48cea1a8b4ef7b3aafc">https://media.pitchfork.com/photos/60d48cea1a8b4ef7b3aafc</a>
The War on Drugs	I Don't Live Here Anymore	Rock	<a href="https://pitchfork.com/reviews/tracks/the-war-on-drugs-i-dont-live-here-anymore/">https://pitchfork.com/reviews/tracks/the-war-on-drugs-i-dont-live-here-anymore/</a>	track	<a href="https://media.pitchfork.com/photos/60f1a8e8c64eabe66d59c1">https://media.pitchfork.com/photos/60f1a8e8c64eabe66d59c1</a>
Snail Mail	Valentine	Rock	<a href="https://pitchfork.com/reviews/tracks/snail-mail-valentine/">https://pitchfork.com/reviews/tracks/snail-mail-valentine/</a>	track	<a href="https://media.pitchfork.com/photos/61424f6a4f9c18c6e9e77fc">https://media.pitchfork.com/photos/61424f6a4f9c18c6e9e77fc</a>
Water From Your Eyes	Quotations	Rock	<a href="https://pitchfork.com/reviews/tracks/water-from-your-eye-quotations/">https://pitchfork.com/reviews/tracks/water-from-your-eye-quotations/</a>	track	<a href="https://media.pitchfork.com/photos/611577e37961053f3f0ebd">https://media.pitchfork.com/photos/611577e37961053f3f0ebd</a>
Big Thief	Little Things	Rock	<a href="https://pitchfork.com/reviews/tracks/big-thief-little-things/">https://pitchfork.com/reviews/tracks/big-thief-little-things/</a>	track	<a href="https://media.pitchfork.com/photos/6112e67fe948362f2f4af5e">https://media.pitchfork.com/photos/6112e67fe948362f2f4af5e</a>
Indigo De Souza	Hold U	Rock	<a href="https://pitchfork.com/reviews/tracks/indigo-de-souza-hold-u/">https://pitchfork.com/reviews/tracks/indigo-de-souza-hold-u/</a>	track	<a href="https://media.pitchfork.com/photos/60fb265dd79d8ab7a7bb1">https://media.pitchfork.com/photos/60fb265dd79d8ab7a7bb1</a>
Low	Days Like These	Rock	<a href="https://pitchfork.com/reviews/tracks/low-days-like-these/">https://pitchfork.com/reviews/tracks/low-days-like-these/</a>	track	<a href="https://media.pitchfork.com/photos/60d246fb4d2d13ef18d3c7">https://media.pitchfork.com/photos/60d246fb4d2d13ef18d3c7</a>
The Beatles	Let It Be (Super Deluxe)	Rock	<a href="https://pitchfork.com/reviews/albums/the-beatles-let-it-be-super-deluxe/">https://pitchfork.com/reviews/albums/the-beatles-let-it-be-super-deluxe/</a>	album	<a href="https://media.pitchfork.com/photos/61674c78c35c407fc7c297c">https://media.pitchfork.com/photos/61674c78c35c407fc7c297c</a>
Pastor T. L. Barrett and the Youth for Christ Choir	I Shall Wear a Crown	Rock	<a href="https://pitchfork.com/reviews/albums/pastor-t-l-barrett-and-the-youth-for-christ-choir-i-shall-wear-a-crown/">https://pitchfork.com/reviews/albums/pastor-t-l-barrett-and-the-youth-for-christ-choir-i-shall-wear-a-crown/</a>	album	<a href="https://media.pitchfork.com/photos/6148e3560fc9ab36796880">https://media.pitchfork.com/photos/6148e3560fc9ab36796880</a>
Trashcan Sinatras	I've Seen Everything	Rock	<a href="https://pitchfork.com/reviews/albums/trashcan-sinatras-i-ve-seen-everything/">https://pitchfork.com/reviews/albums/trashcan-sinatras-i-ve-seen-everything/</a>	album	<a href="https://media.pitchfork.com/photos/614e42d2d3a861d44aafa2">https://media.pitchfork.com/photos/614e42d2d3a861d44aafa2</a>
Laura Nyro	American Dreamer	Rock	<a href="https://pitchfork.com/reviews/albums/laura-nyro-american-dreamer/">https://pitchfork.com/reviews/albums/laura-nyro-american-dreamer/</a>	album	<a href="https://media.pitchfork.com/photos/6143ac690fc9ab36796880">https://media.pitchfork.com/photos/6143ac690fc9ab36796880</a>
Alice Coltrane	Kirtan: Turiya Sings	Jazz/Experimental	<a href="https://pitchfork.com/reviews/albums/alice-coltrane-kirtan-turiya-sings/">https://pitchfork.com/reviews/albums/alice-coltrane-kirtan-turiya-sings/</a>	album	<a href="https://media.pitchfork.com/photos/60b7851c28fd170dc6ddcc">https://media.pitchfork.com/photos/60b7851c28fd170dc6ddcc</a>
Blue Gene Tyranny	Degrees of Freedom Found	Experimental	<a href="https://pitchfork.com/reviews/albums/blue-gene-tyranny-degrees-of-freedom-found/">https://pitchfork.com/reviews/albums/blue-gene-tyranny-degrees-of-freedom-found/</a>	album	<a href="https://media.pitchfork.com/photos/60c3787542a3ecb7ba1011">https://media.pitchfork.com/photos/60c3787542a3ecb7ba1011</a>

## Contingut

Per poder construir el nostre fitxer CSV amb les dades obtingudes de la pàgina web <https://pitchfork.com/best/> fent web scraping hem de seguir els passos que es comenten a continuació.

Primer de tot, revisem les limitacions de la web a l'arxiu robots.txt

<https://pitchfork.com/robots.txt>



```
User-agent: *
Allow: /
Disallow: /search/
Disallow: /search
Disallow: /*?redirectURL=
Disallow: /*{{url}}
Disallow: /*%7B%7Burl%7D%7D

Sitemap: https://pitchfork.com/sitemap.xml
Sitemap: https://pitchfork.com/branded-sitemap.xml
Sitemap: https://pitchfork.com/feed/google-latest-news/sitemap-google-news
Sitemap: https://pitchfork.com/feed/sitemap
```

Com podem observar, es permeten tots els user agents. També es permet fer crawl de les pàgines sota el root directory ("/"). En general, no tenim cap limitació a priori. En alguns casos les webs limiten el nombre de requests a fer en una finestra de temps, però no és el nostre cas.

Aquest tipus de comprovacions també les podem fer fent ús de l'eina robotparser.



```
if __name__ == '__main__':

    url_scrap = 'https://pitchfork.com/best/'
    url_robots = 'https://pitchfork.com/robots.txt'

    # Check de l'arxiu robots.txt
    rp = robotparser.RobotFileParser()
    rp.set_url(url_robots)
    rp.read()
    print(rp.crawl_delay("*"))
    print(rp.can_fetch("*", url_scrap))
```



```
None
True
```

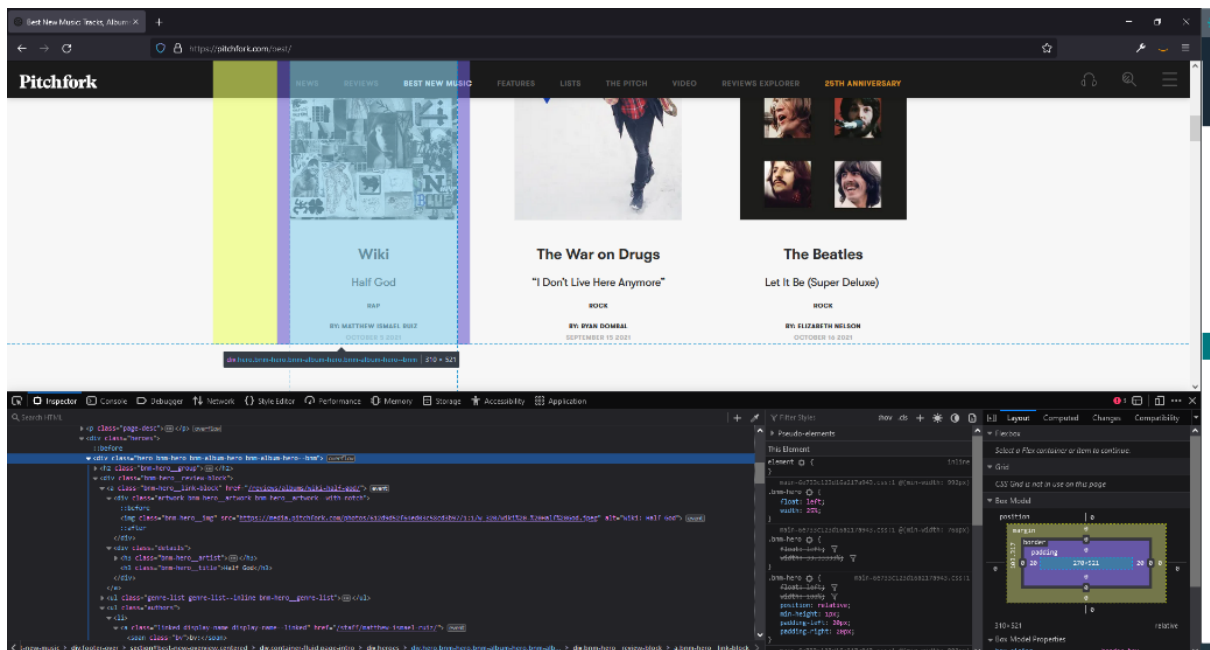
A la documentació de [robot.txt](#) trobarem el significat d'aquests arxius.

Un cop veiem que no hi ha problema en fer crawl a aquella pàgina, fem la request de la pàgina desitjada. Si obtenim el codi 200, és que la request s'ha pogut fer sense problemes.

Per tal de poder veure el contingut de la request de forma més user friendly, farem servir la llibreria BeautifulSoup.

Amb BeautifulSoup no només podem llegir el contingut de manera més amigable, sinó que també podem navegar pels objectes de l'arbre generat per la llibreria. D'aquesta manera podrem obtenir els objectes que ens interessin.

Per tal de veure quins són els objectes d'interès, podem inspeccionar la pàgina directament des del nostre navegador.



A la pàgina que volem explorar es mostra informació sobre la millor nova música. Volem extreure els següents camps per cada cançó:

- Artista
- Títol
- Link a la review
- Àlbum/cançó
- Gènere
- Portada

Hem de tenir en compte que cadascun dels elements que volem obtenir es troba encapsulat entre dues tags del HTML obtingut amb la request.

Veiem que els elements a obtenir estan encapsulats en un div, amb tota la informació que necessitem.

També observem 3 seccions les quals es corresponen amb una classe diferent del CSS, el qual ens ajudarà a obtenir les dades de l'objecte BeautifulSoup.

### 1. Heroes: bnm-hero\_\_review-block

```

▼ <div class="heroes">
  ::before
  ▼ <div class="hero bnm-hero bnm-album-hero bnm-album-hero--bnm"> overflow
    ▶ <h2 class="bnm-hero__group"> ... </h2>
    ▶ <div class="bnm-hero__review-block"> ... </div>
  </div>
  ▼ <div class="hero bnm-hero bnm-track-hero"> overflow
    ▶ <h2 class="bnm-hero__group"> ... </h2>
    ▶ <div class="bnm-hero__review-block"> ... </div>
  </div>
  ▼ <div class="hero bnm-hero bnm-album-hero bnm-album-hero--bnr"> overflow
    ▶ <h2 class="bnm-hero__group"> ... </h2>
    ▶ <div class="bnm-hero__review-block"> ... </div>
  </div>
  ::after
</div>

```

### 2. Best New Albums: bnm-small album-small

```

▼ <section id="best-new-albums" class="bnm-type">
  ▼ <div class="container-fluid">
    ::before
    ▼ <div class="section-detail">
      ::before
      ▶ <svg class="bnm-arrows" version="1.1" x="0px" y="0px" viewBox="0 0 80 40" enable-background="new 0 0 80 40"> ... </svg> overflow
      <h1 class="section-title">Best New Albums</h1> overflow
      ▶ <a class="see-all" href="/reviews/best/albums/"> ... </a> event
      ::after
    </div>
    ▼ <ul class="clearfix square-list">
      ::before
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
      ▼ <li> overflow
        ▶ <div class="bnm-small album-small"> ... </div>
      </li>
    </ul>
  </div>

```

### 3. Best New Albums: bnm-small album-small



```

<section id="best-new-tracks" class="bnm-type">
  <div class="container-fluid">
    ::before
    <div class="section-detail">...</div>
    <ul class="clearfix square-list">
      ::before
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small track-small">...</div>
      </li>
      ::after
    </ul>
    ::after
  </div>

```

#### 4. Best New Reissues: bnm-small album-small

```

<section id="best-new-reissues" class="bnm-type">
  <div class="container-fluid">
    ::before
    <div class="section-detail">...</div>
    <ul class="clearfix square-list">
      ::before
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      <li>overflow
        <div class="bnm-small album-small">...</div>
      </li>
      ::after
    </ul>
  </div>

```

Mitjançant el mètode `find_all()` de la llibreria BeautifulSoup obtenim les reviews, indicant les tags i les classes en que estem interessats per fer les particions de la informació. Observem com és la informació que ens arriba de cada review.

```

<html>
  <body>
    <div class="bnm-hero__review-block">
      <a class="bnm-hero__link-block" href="/reviews/albums/wiki-half-god/">

```

```

<div class="artwork bnm-hero__artwork bnm-hero__artwork--with-notch"></div>
<div class="details">
  <h3 class="bnm-hero__artist">
    <ul class="artist-list">
      <li>Wiki</li>
    </ul>
  </h3>
  <h3 class="bnm-hero__title">Half God</h3>
</div>
</a>
<ul class="genre-list genre-list--inline bnm-hero__genre-list">
  <li class="genre-list__item"><a class="genre-list__link"
href="/reviews/albums/?genre=rap">Rap</a></li>
</ul>
<ul class="authors">
  <li><a class="linked display-name display-name--linked"
href="/staff/matthew-ismael-ruiz/"><span class="by">by: </span>Matthew Ismael Ruiz</a></li>
</ul>
<p class="bnm-hero__date"><time class="pub-date" datetime="2021-10-05T04:00:00"
title="Tue, 05 Oct 2021 04:00:00 GMT">October 5 2021</time></p>
</div>
</body>
</html>

```

Aquest és el primer track obtingut de la llista. Podem veure que obtenim la informació dels atributs desitjats.

- Artista:

```

<h3 class="bnm-hero__artist">
  <ul class="artist-list">
    <li>Wiki</li>
  </ul>
</h3>

```

- Títol:

```

<h3 class="bnm-hero__title">Half God</h3>

```

- Link a la review:

```

<a class="bnm-hero__link-block" href="/reviews/albums/wiki-half-god/">

```

- Àlbum/cançó: es pot obtenir del link

```

<a class="bnm-hero__link-block" href="/reviews/albums/wiki-half-god/">

```

- Gènere:

```
<li class="genre-list__item"><a class="genre-list__link"
href="/reviews/albums/?genre=rap">Rap</a></li>
```

- Portada:

```

```

Pels següents tracks hem seguit el mateix procediment i hem anat guardant tota la informació en un diccionari que posteriorment hem convertit en un dataframe i l'hem guardat en un fitxer csv.

## Agraïments

El conjunt de dades creat en aquest treball s'ha recollit de la pàgina web Pitchfork el dia 7 de novembre de 2021. Per fer-ho s'ha utilitzat Web Scraping amb el llenguatge de programació Python juntament amb les llibreries requests, beautifulsoup, pandas i urllib.

També s'ha revisat l'arxiu robots.txt per saber quines restriccions hem hagut de tenir en compte a l'hora de rastrejar el lloc web.

## Inspiració

Aquest conjunt de dades s'ha creat amb la finalitat de recollir les millors cançons i àlbums segons la pàgina Pitchfork, i guardar les URLs de cada una de les revisions.

Aquesta base de dades la podem anar actualitzant periòdicament per així acabar obtenint una base de dades molt més completa. Per fer-ho ens podem fixar en el sitemap <https://pitchfork.com/feed/google-latest-news/sitemap-google-news> per veure les actualitzacions que s'han anat realitzant.

Com que la pàgina que hem rastrejat a mesura que van introduint noves revisions també en van traient, amb el nostre fitxer CSV podem emmagatzemar totes aquestes revisions sense perdre-les en un futur, i així tenir-les sempre guardades per poder-les analitzar quan ho necessitem.

## Llicència

La llicència que utilitzarem és la Released Under CC BY-NC-SA 4.0 License.

Aquesta llicència permet fer servir el conjunt de dades sempre que se'n citi l'autor, d'aquesta manera assegurem el nostre reconeixement; s'utilitzi només per projectes no comercials i es publiqui amb la mateixa llicència, així garantim que es segueixi distribuint aquesta base de dades amb els mateixos permisos que hem establert.

## Codi

En [aquest repositori GitHub](#) podem trobar el codi que permet realitzar el Web Scraping amb el nom de main.py.

Per executar aquest script és necessari instal·lar les llibreries següents:

```
pip install pandas
pip install requests
pip install beautifulsoup4
pip install urllib
```

El script s'ha d'executar amb la següent comanda:

```
python main.py
```

## Dataset

El dataset generat es pot trobar en [aquest link de Zenodo](#).

## Contribucions

Contribucions	Signatura
Investigació prèvia	MSM, OBR
Redacció de les respostes	MSM, OBR
Desenvolupament del codi	MSM, OBR

## Recursos

- Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

- <https://medium.com/@stephaniecaress/scraping-pitchforks-best-new-music-be563d18ea4f>