spotify_scraper

April 30, 2024

1 Spotify API Scraping for Creating Datasets

1.0.1 I used the code (but slightly modified it to my needs to include the `release_date`) from this lovely article!

[1]:
```
# pip install spotifyscraper
```

[2]:
```
# pip install pathlib
```

[3]:
```
# pip install ruamel-yaml
```

[4]:
```
# pip install spotipy
```

[5]:
```
import spotipy
from spotipy.oauth2 import SpotifyClientCredentials
from spotipy.oauth2 import SpotifyOAuth
import pandas as pd
import time
```

[7]:
```
cid = '6ff98eca336346ee942d607cc2d23879'
secret = '7d7292ad3cf0420e8f270e7d049f40ba'
client_credentials_manager = SpotifyClientCredentials(client_id=cid,␣
 ↪client_secret=secret)
sp = spotipy.Spotify(client_credentials_manager = client_credentials_manager)
```

[8]:
```
# Pagination (to extract more than 100 songs at a time)
def call_playlist(creator, playlist_id):
    # Step 1: Initialize DataFrame and other variables
    playlist_features_list = ["artist", "album", "track_name", "track_id",␣
 ↪"release_date", "danceability", "energy", "key", "loudness", "mode",␣
 ↪"speechiness", "instrumentalness", "liveness", "valence", "tempo",␣
 ↪"duration_ms", "time_signature"]
    playlist_df = pd.DataFrame(columns=playlist_features_list)
    offset = 0
    total_tracks = sp.user_playlist_tracks(creator, playlist_id)["total"]
    # Step 2: Fetch tracks with pagination
    while offset < total_tracks:
        playlist = sp.user_playlist_tracks(creator, playlist_id,␣
↪offset=offset)["items"]
        for track in playlist:
            playlist_features = {}
            playlist_features["artist"] =␣
↪track["track"]["album"]["artists"][0]["name"]
            playlist_features["album"] = track["track"]["album"]["name"]
            playlist_features["track_name"] = track["track"]["name"]
            playlist_features["track_id"] = track["track"]["id"]
            playlist_features["release_date"] =␣
↪track["track"]["album"]["release_date"]
            audio_features = sp.audio_features(playlist_features["track_id"])[0]
            for feature in playlist_features_list[5:]:
                playlist_features[feature] = audio_features[feature]
            track_df = pd.DataFrame(playlist_features, index=[0])
            playlist_df = pd.concat([playlist_df, track_df], ignore_index=True)
        offset += 100
    # Step 3: Return DataFrame
    return playlist_df
```

[9]:

```
# Function to fetch audio features with retry logic
def fetch_audio_features(track_id):
    retries = 10  # Maximum number of retry attempts
    for _ in range(retries):
        try:
            return sp.audio_features(track_id)[0]
        except spotipy.SpotifyException as e:
            if e.http_status == 429:
                # Retry after a fixed delay
                retry_after = int(e.headers.get('Retry-After', 10))  # Default␣
↪to 10 seconds if no Retry-After header
                print(f"Rate limited. Retrying after {retry_after} seconds...")
                time.sleep(retry_after)
            else:
                raise  # Re-raise the exception if it's not a 429 error
    raise Exception("Max retries reached, unable to fetch audio features")
```

[10]:

```
# https://open.spotify.com/playlist/5yAPuepGnApi5yc4QoZMDl
# "old" Playlist compiled by "emmabittinger" (this is a test)
old_playlist = call_playlist("spotify","5yAPuepGnApi5yc4QoZMDl")
```

[11]:

```
old_playlist.head()
```

[11]:

```
              artist                              album  \
0    Various Artists        Cars (VersiÃşn de ColecciÃşn)
1      Anthem Lights                     Covers Part IV
2  Bruce Springsteen                 Born In The U.S.A.
3     Lynyrd Skynyrd  Second Helping (Expanded Edition)
4        Rick Astley         Whenever You Need Somebody

                                          track_name                track_id  \
0  Life Is A Highway - From "Cars"/Soundtrack Ver...  1QezVl06xBzPfgJ2HXST5d
1                               Don't Stop Believing  0wBqAqxUygzHrUgw0MTJ6J
2                                 Born in the U.S.A.  0d0g1ySSI7NkpAe89Zo0b9
3                                 Sweet Home Alabama  7e89621JPkKaeDSTQ3avtg
4                            Never Gonna Give You Up  7GhIk7Il098yCjg4BQjzvb

  release_date  danceability  energy  key  loudness  mode  speechiness  \
0   2006-06-06         0.561   0.932    5    -5.475     1       0.0584
1   2015-07-17         0.516   0.391   10    -7.319     1       0.0315
2   1984-06-04         0.398   0.952    4    -6.042     1       0.0610
3   1974-04-15         0.596   0.605    7   -12.145     1       0.0255
4   1987-12-08         0.727   0.939    8   -11.855     1       0.0369

  instrumentalness  liveness  valence    tempo  duration_ms  time_signature
0                0    0.1810    0.670  103.062       275640               4
1                0    0.1440    0.395  117.873       218644               4
2         0.000077    0.1000    0.584  122.093       278680               4
3         0.000331    0.0863    0.886   97.798       283800               4
4         0.000044    0.1510    0.916  113.330       212827               4
```

[12]:

```
# https://open.spotify.com/playlist/66kbLWdmxWuMYeByFkqADT
# "throwbaccc" Playlist compiled by "emmabittinger"
throbaccc_playlist = call_playlist("spotify","66kbLWdmxWuMYeByFkqADT")
```

[14]:

```
throbaccc_playlist.to_csv("throbaccc_playlist.csv")
```

[15]:

```
# https://open.spotify.com/playlist/7dBWDKw7I8kZy0td1VYFIY
# "Songs Everyone Knows the Words To" Playlist compiled by "Ava Montgomery"
long_playlist = call_playlist("spotify","7dBWDKw7I8kZy0td1VYFIY")
```

[16]:

```
long_playlist.to_csv("long_playlist.csv")
```

[17]:

```
throbaccc_playlist.head()
```

[17]:

```
          artist                                album           track_name  \
0    Miley Cyrus             The Time Of Our Lives  Party In The U.S.A.
1        Rihanna                                 Loud      What's My Name?
2          Train                     Hey, Soul Sister     Hey, Soul Sister
3  Justin Bieber                             My World             One Time
```

```
4      Taio Cruz  Rokstarr (International Version)             Dynamite

                 track_id release_date  danceability  energy  key  loudness  \
0  5Q0Nhxo0l2bP3pNjpGJwV1   2009-01-01         0.652   0.698   10    -4.667
1  5FTCKvxzqy72ceS4Ujux4N   2010-11-16         0.692   0.786    2    -2.959
2  0KpfYajJVVGgQ32Dby7e9i   2009-08-06         0.675   0.885    1    -4.432
3  6eDApnV9Jdb1nYah0lbbUh   2009-01-01         0.691   0.853    1    -2.528
4  0bg6otrW5gxNnlCqrCrXyd   2010-05-28         0.754   0.804    4    -3.177

  mode  speechiness  instrumentalness  liveness  valence    tempo  duration_ms  \
0    0       0.0420          0.000115    0.0886    0.470   96.021       202067
1    1       0.0690          0.000000    0.0797    0.583  100.025       263173
2    0       0.0436          0.000000    0.0860    0.768   97.030       216667
3    0       0.0372          0.000071    0.0820    0.762  145.999       215867
4    1       0.0853          0.000000    0.0329    0.818  119.968       203867

  time_signature
0               4
1               4
2               4
3               4
4               4
```

```
[18]: throbaccc_playlist.describe()
```

```
[18]:        danceability      energy    loudness  speechiness  instrumentalness  \
      count    330.000000  330.000000  330.000000   330.000000        330.000000
      mean       0.647394    0.738092   -5.048933     0.075087          0.005875
      std        0.117276    0.164847    1.792816     0.061893          0.056663
      min        0.327000    0.056500  -15.099000     0.025400          0.000000
      25%        0.578500    0.676250   -5.889250     0.037700          0.000000
      50%        0.658500    0.773000   -4.823500     0.051100          0.000000
      75%        0.724750    0.857000   -3.848500     0.088050          0.000004
      max        0.979000    0.981000   -1.644000     0.449000          0.871000

               liveness     valence       tempo
      count  330.000000  330.000000  330.000000
      mean     0.177032    0.587910  121.181464
      std      0.128664    0.206436   25.593651
      min      0.019300    0.076500   65.043000
      25%      0.088675    0.428000  101.734500
      50%      0.127000    0.619000  122.624000
      75%      0.248750    0.748000  132.023000
      max      0.758000    0.965000  194.077000
```

```
[19]: long_playlist.head()
```

[19]:

```
             artist                                         album  \
0          Coldplay     Viva La Vida or Death and All His Friends
1           Rihanna                            Good Girl Gone Bad
2            *NSYNC                           No Strings Attached
3             Train  Save Me, San Francisco (Golden Gate Edition)
4  Carrie Underwood                                   Some Hearts

         track_name                track_id release_date  danceability  \
0      Viva La Vida  1mea3bSkSGXuIRvnydlB5b   2008-05-26         0.486
1          Umbrella  2yPoXCs7BSIUrucMdK5PzV   2007-01-01         0.583
2       Bye Bye Bye  62bOmKYxYg7dhrC6gH9vFn   2000-03-21         0.610
3  Hey, Soul Sister  4HlFJV71xXKIGcU3kRyttv   2010-12-01         0.673
4  Before He Cheats  0ZUo4YjG4saFnEJhdWp9Bt   2005-11-14         0.519

   energy key  loudness mode  speechiness  instrumentalness  liveness  \
0   0.617   5   -7.115     0       0.0287          0.000003    0.1090
1   0.829   1   -4.603     1       0.1340          0.000000    0.0426
2   0.926   8   -4.843     0       0.0479          0.001200    0.0821
3   0.886   1   -4.440     0       0.0431          0.000000    0.0826
4   0.749   6   -3.318     0       0.0405          0.000000    0.1190

   valence    tempo duration_ms time_signature
0    0.417  138.015      242373              4
1    0.575  174.028      275987              4
2    0.861  172.638      200400              4
3    0.795   97.012      216773              4
4    0.290  147.905      199947              4
```

[20]:

```
long_playlist.describe()
```

[20]:

```
       danceability      energy    loudness  speechiness  instrumentalness  \
count    345.000000  345.000000  345.000000   345.000000        345.000000
mean       0.614093    0.704965   -5.682510     0.079451          0.007984
std        0.139569    0.171480    2.219701     0.069564          0.049434
min        0.209000    0.111000  -18.064000     0.024900          0.000000
25%        0.522000    0.594000   -6.682000     0.038200          0.000000
50%        0.614000    0.728000   -5.223000     0.054500          0.000000
75%        0.717000    0.839000   -4.165000     0.085300          0.000090
max        0.967000    0.986000   -1.848000     0.449000          0.616000

        liveness     valence       tempo
count  345.00000  345.000000  345.000000
mean     0.17513    0.508641  125.870136
std      0.13550    0.228687   26.866936
min      0.02100    0.038500   65.997000
25%      0.09010    0.336000  106.970000
50%      0.12000    0.506000  125.072000
```

```
75%      0.23700    0.691000  142.673000
max      0.77000    0.969000  199.935000
```

```
[ ]:
```