DATA ENGINEERING INDIVIDUAL COURSEWORK

SPOTIFY PLAYLIST DATABASE:

A SENTIMENT ANALYSIS

April 2022
Word Count:

TABLE OF CONTENTS

1.0 INTRODUCTION

“Music can heal the wounds which medicine cannot touch”, says Debasish Mridha. From many decades ago, music has already been recognised as an opportunity to address mental health challenges (Schriewer and Bulaj, 2016). Nowadays, audience can enjoy the music more conveniently via music streaming services, instead of downloading the original audio file of a song. As one of the biggest music streaming platforms, Spotify had over 365 million users by February 2022 (Caddy, 2022).

One important feature of Spotify is the editorial playlists. Spotify’s in-house teams curate these playlists by selecting collections of songs that somehow have some similarities so that meaningful playlists are created. Many playlists have emotion-related tags such as “happy”, “moody” and “chilling”, and each of them is designed to match the audience’s emotions.

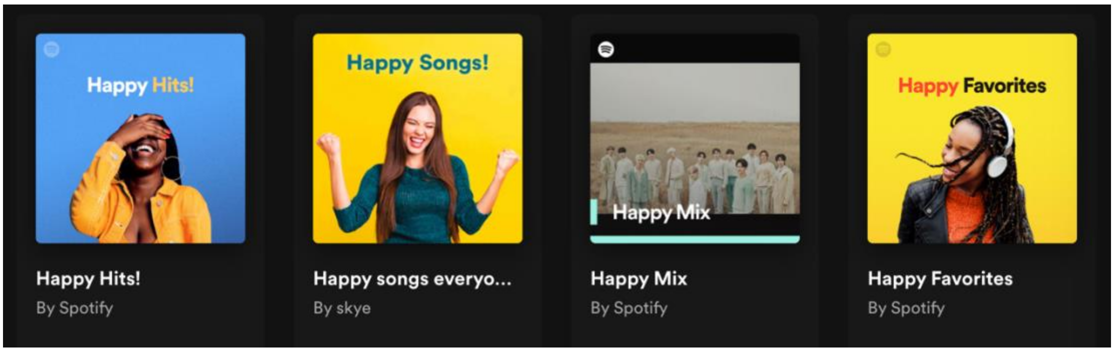


Figure 1: Spotify Curated Playlists Related to “Happy”

While the Spotify playlist is an important feature of the platform in helping the artists reach more of the target audiences, it is worthy of investigation that the exact criterion of how a song is featured on a certain playlist. This project aims to create a database of a specific Spotify playlist that contains a number of different attributes of the songs within that playlist and conduct further sentiment analyses of these songs based on the attributes.

In achieving the objective of the project, the following steps are executed and will be explained in detail throughout this report: First, real-time streaming data will be extracted through API scraping from Spotify and several other relevant platforms. The data will then be processed and stored in a suitable way so that further analyses are allowed. This project will then conduct sentiment analyses on these processed data so that we can evaluate whether the "emotion" of the songs in the playlist match with the playlist title. It is believed that successful execution of the project with provide precious guidelines for sentiment related research on a bigger database beyond the scope of this project.

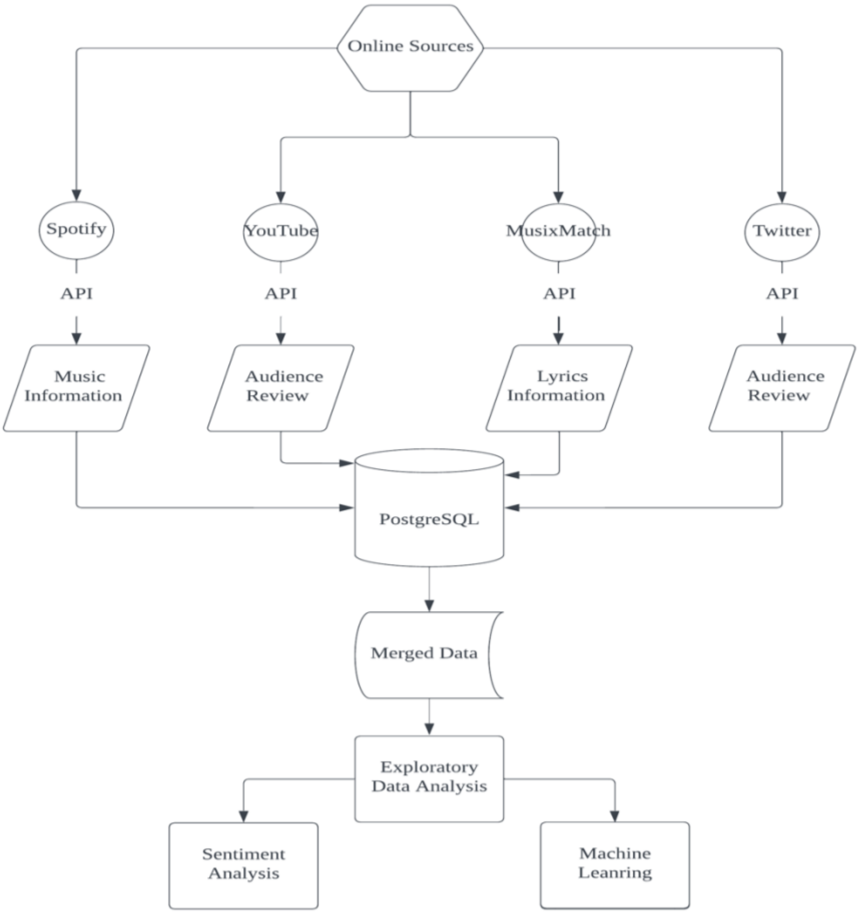


Figure 2: Project Workflow

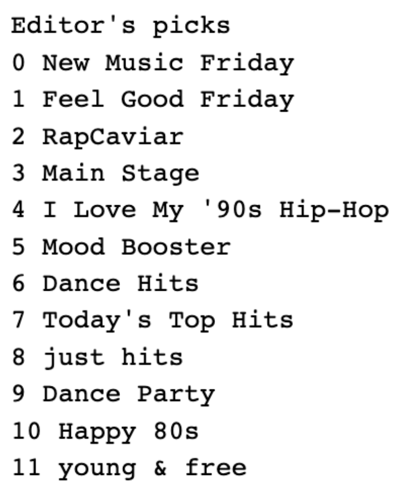
2.0 DATA COLLECTION

This project mainly used the Application Programming Interface (API) techniques in the Python language to extract data from four different platforms including Spotify, MusixMatch, YouTube and Twitter. The API is an important tool to obtain data from dynamic websites and allows some level of customisations (Medium,2022).

2.1 SPOTIFY DATA

As the main subject matter of this project, data from Spotify was scraped first. Spotify provides API services on its developer website and a number of functions can be achieved under the free user plan. On the developer website, a Python library called Spotipy (https://github.com/plamere/spotipy) is recommended for executing the API calls. This package was utilised by this project throughout the data extraction from Spotify.

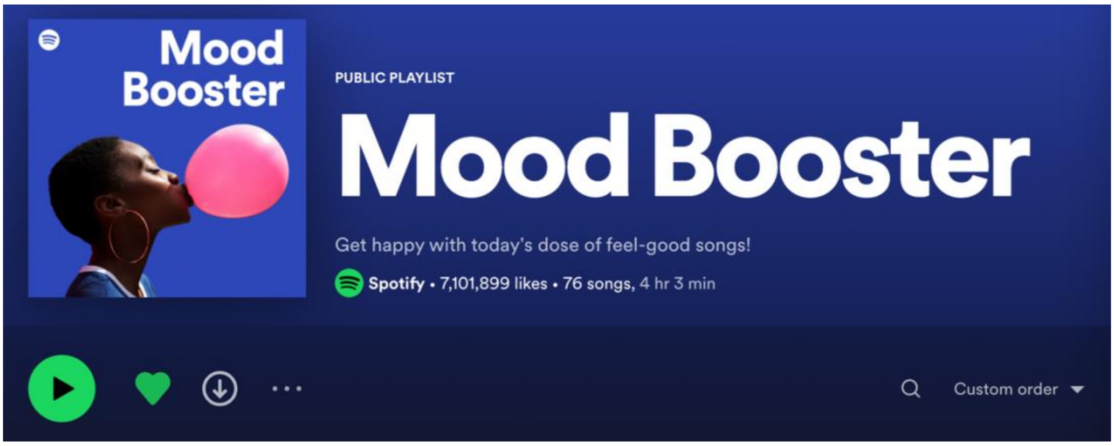
First of all, the project wanted to retrieve a list of featured playlists (Editor's Picks) on Spotify to decide which playlist to extract and evaluate. By using the *featured_playlists()* method, a list of playlists which were the Editor’s picks on the day when the data was scraped was returned:



```
Editor's picks
0 New Music Friday
1 Feel Good Friday
2 RapCaviar
3 Main Stage
4 I Love My '90s Hip-Hop
5 Mood Booster
6 Dance Hits
7 Today's Top Hits
8 just hits
9 Dance Party
10 Happy 80s
11 young & free
```

Figure 3: Featured Playlists on 8th April

The sixth result was a playlist named ***"Mood Booster"***, which was highly related to the research objective - to analyse the emotion of songs with sentimental analysis techniques and evaluate whether the emotion of the song match with the overall playlist genre.



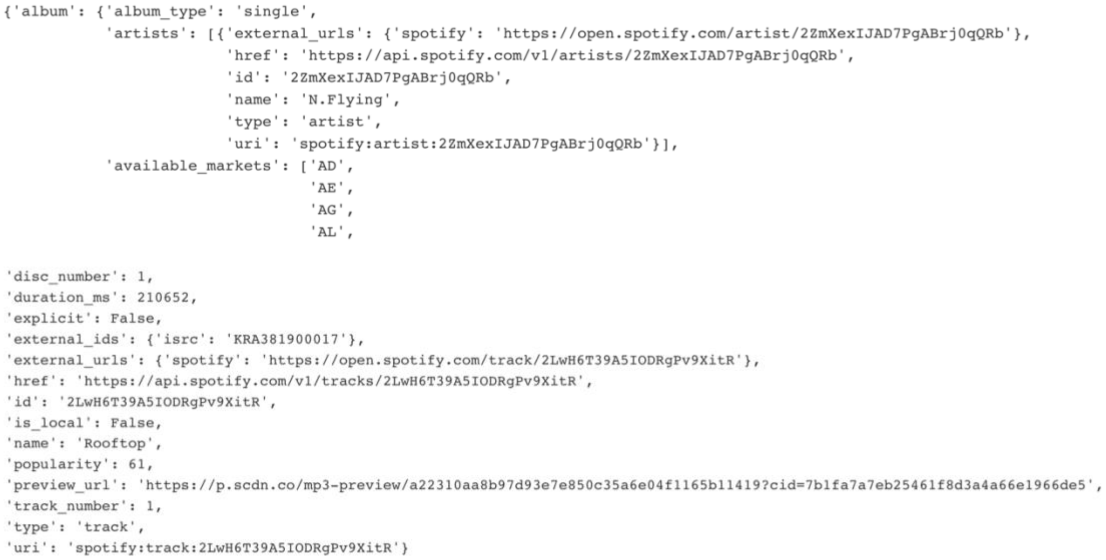
Get happy with today's dose of feel-good songs!

Spotify • 7,101,899 likes • 76 songs, 4 hr 3 min

Figure 4: The *Mood Booster* Playlist

According to Spotify’s description of this playlist, the songs in the playlist are supposed to make the audience “feel good” and “get happy”. This project is interested in the if the emotions of the songs in this playlist actually match that purpose. To dig deeper into the tracks in this playlist, the playlist's ID on Spotify needs to be known. Unfortunately, at the moment, the only feasible way to get a playlist's ID is through getting a user's current playlist. Therefore, this playlist was manually followed on Spotify and added to the profile using a personal Spotify account.

With the playlist id being known, the project was able to retrieve more information about the tracks in this the playlist. A list 76 song names were first retrieved via the *playlist_items()* method, and the 76 songs' corresponding Spotify IDs were retrieved via similar ways. The songs' Spotify IDs allowed the project to retrieve a number of attributes of the track by using the Spotify API's *track()* method, including the artist information. By calling the method and looping into the nested dictionary (see outcome example in Figure 5) returned by the method, this project was able to get the *artist's name, album, Spotify popularity, release date, duration* information of the track. The method could also tell whether a track contains explicit content or not: most music streaming platforms distinguish and differentiate between tracks that is suitable for mainstream consumption, and those songs that may contain a parental advisory or may be considered explicit content (Soundplate, 2022). On Spotify, a track with explicit content will have a "E" or "Explicit" symbol next to its title. With the API, boolean values of True or False was returned regarding the "explicit" attribute.



```
{'album': {'album_type': 'single',
           'artists': [{'external_urls': {'spotify': 'https://open.spotify.com/artist/2ZmXexIJAD7PgABrj0qQRb'},
                        'href': 'https://api.spotify.com/v1/artists/2ZmXexIJAD7PgABrj0qQRb',
                        'id': '2ZmXexIJAD7PgABrj0qQRb',
                        'name': 'N.Flying',
                        'type': 'artist',
                        'uri': 'spotify:artist:2ZmXexIJAD7PgABrj0qQRb'}],
           'available_markets': ['AD',
                                 'AE',
                                 'AG',
                                 'AL',

'disc_number': 1,
'duration_ms': 210652,
'explicit': False,
'external_ids': {'isrc': 'KRA381900017'},
'external_urls': {'spotify': 'https://open.spotify.com/track/2LwH6T39A5IODRgPv9XitR'},
'href': 'https://api.spotify.com/v1/tracks/2LwH6T39A5IODRgPv9XitR',
'id': '2LwH6T39A5IODRgPv9XitR',
'is_local': False,
'name': 'Rooftop',
'popularity': 61,
'preview_url': 'https://p.scdn.co/mp3-preview/a22310aa8b97d93e7e850c35a6e04f1165b11419?cid=7b1fa7a7eb25461f8d3a4a66e1966de5',
'track_number': 1,
'type': 'track',
'uri': 'spotify:track:2LwH6T39A5IODRgPv9XitR'}
```

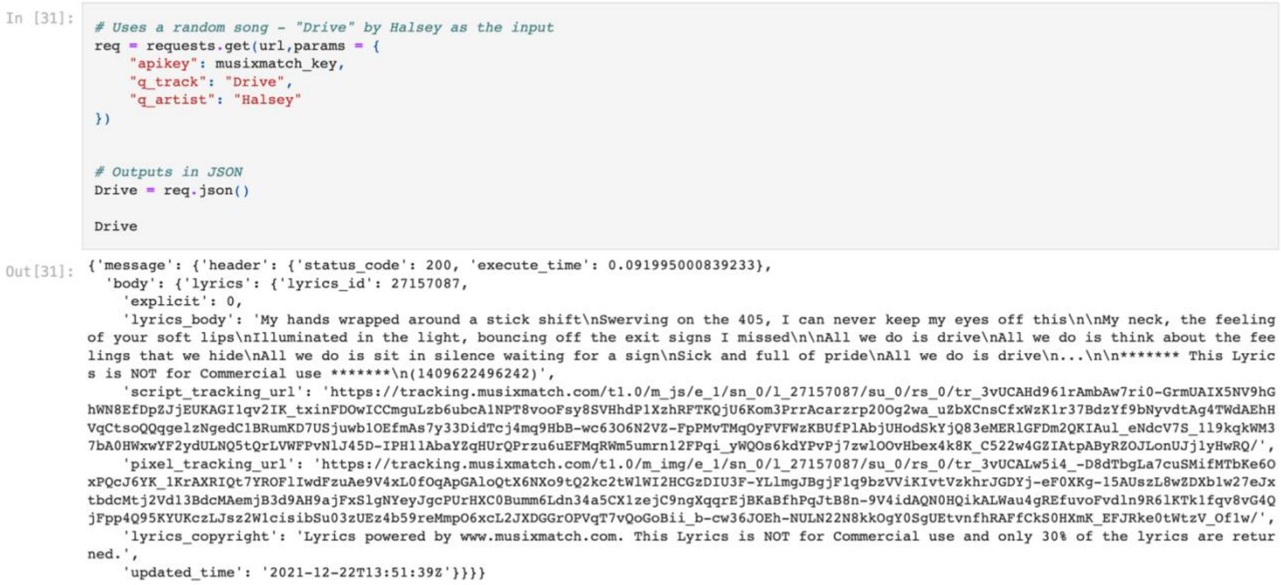
Figure 5: Output Example of the *track()* Method

2.2 MUSIXMATCH DATA

MusixMatch is an Italian music data company which has the world's largest database of 14 million lyrics items in various different languages (Baydeer, 2021). With the lyrics data provided by MusixMatch, this project would be able to conduct further sentiment analysis on the lyric strings.

With the free API plan provided by MusixMatch, the account created was limited to 2,000 API calls per day, and only 30% of the lyrics of a song was accessible. The project would have to assume most songs would have their emotions set in stone in the very first bit.

By inputting the title and the artist's name of the song as the parameters of the MusixMatch request call, JSON styled results were pulled. In Figure 6, this project used *Halsey*'s song "*Drive*" as an example to examine the output.



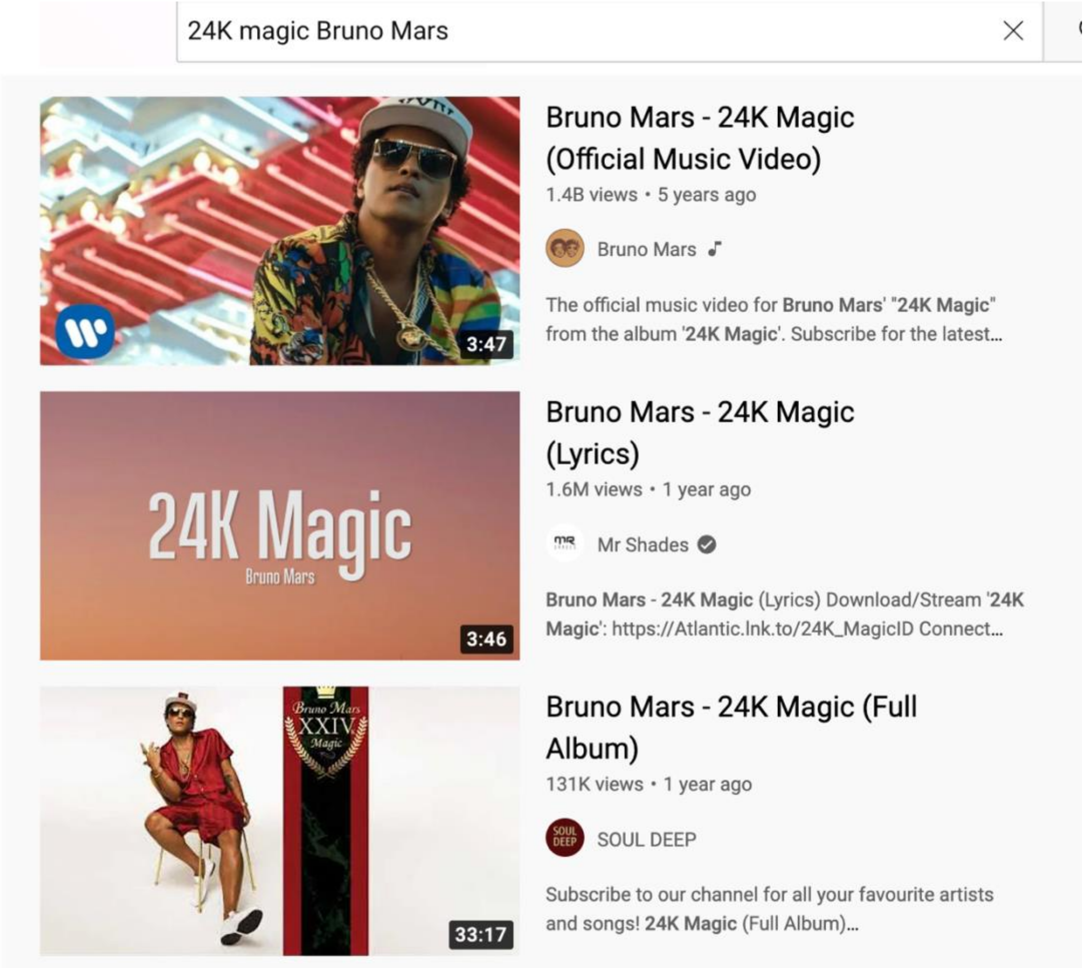
```
In [31]:  # Uses a random song - "Drive" by Halsey as the input
          req = requests.get(url,params = {
              "apikey": musixmatch_key,
              "q_track": "Drive",
              "q_artist": "Halsey"
          })


          # Outputs in JSON
          Drive = req.json()

          Drive
```

```
Out[31]: {'message': {'header': {'status_code': 200, 'execute_time': 0.091995000839233},
           'body': {'lyrics': {'lyrics_id': 27157087,
             'explicit': 0,
             'lyrics_body': 'My hands wrapped around a stick shift\nSwerving on the 405, I can never keep my eyes off this\n\nMy neck, the feeling
         of your soft lips\nIlluminated in the light, bouncing off the exit signs I missed\n\nAll we do is drive\nAll we do is think about the fee
         lings that we hide\nAll we do is sit in silence waiting for a sign\nSick and full of pride\nAll we do is drive\n...\n\n******* This Lyric
         s is NOT for Commercial use *******\n(1409622496242)',
             'script_tracking_url': 'https://tracking.musixmatch.com/t1.0/m_js/e_1/sn_0/l_27157087/su_0/rs_0/tr_3vUCAHd961rAmbAw7ri0-GrmUAIX5NV9hG
         hWN8EfDpZJjEUKAGI1qv2IK_txinFDOwICCmguLzb6ubcA1NPT8vooFsy8SVHhdP1XzhRFTKQjU6Kom3PrrAcarzrp20Og2wa_uZbXCnsCfxWzK1r37BdzYf9bNyvdtAg4TWdAEhH
         VqCtsoQQqgelzNgedC1BRumKD7USjuwb1OEfmAs7y33DidTcj4mq9HbB-wc63O6N2VZ-FpPMvTMqOyFVFWzKBUfPlAbjUHodSkYjQ83eMERlGFDm2QKIAul_eNdcV7S_1l9kqkWM3
         7bA0HWxwYF2ydULNQ5tQrLVWFPvNlJ45D-IPH11AbaYZqHUrQPrzu6uEFMqRWm5umrn12FPqi_yWQOs6kdYPvPj7zwlOOvHbex4k8K_C522w4GZIAtpAByRZOJLonUJj1yHwRQ/',
             'pixel_tracking_url': 'https://tracking.musixmatch.com/t1.0/m_img/e_1/sn_0/l_27157087/su_0/rs_0/tr_3vUCALw5i4_-D8dTbgLa7cuSMifMTbKe6O
         xPQcJ6YK_1KrAXRIQt7YROFlIwdFzuAe9V4xL0fOqApGAloQtX6NXo9tQ2kc2tWlWI2HCGzDIU3F-YLlmgJBgjF1q9bzVViKIvtVzkhrJGDYj-eF0XKg-15AUszL8wZDXblw27eJx
         tbdcMtj2Vdl3BdcMAemjB3d9AH9ajFxSlgNYeyJgcPUrHXC0Bumm6Ldn34a5CX1zejC9ngXqqrEjBKaBfhPqJtB8n-9V4idAQN0HQikALWau4gREfuvoFvdln9R6lKTk1fqv8vG4Q
         jFpp4Q95KYUKczLJsz2W1cisibSu03zUEz4b59reMmpO6xcL2JXDGGrOPVqT7vQoGoBii_b-cw36JOEh-NULN22N8kkOgY0SgUEtvnfhRAFfCkS0HXmK_EFJRke0tWtzV_Of1w/',
             'lyrics_copyright': 'Lyrics powered by www.musixmatch.com. This Lyrics is NOT for Commercial use and only 30% of the lyrics are retur
         ned.',
             'updated_time': '2021-12-22T13:51:39Z'}}}}
```

Figure 6: Output Example a Random Song

2.3 YOUTUBE DATA

Other than emotion-related attributes, this project is also interested in the popularity of the official music videos and user-uploaded lyrics videos of the tracks in the playlist on YouTube.



24K magic Bruno Mars

Bruno Mars - 24K Magic (Official Music Video)
1.4B views • 5 years ago
Bruno Mars
The official music video for Bruno Mars' "24K Magic" from the album '24K Magic'. Subscribe for the latest...
3:47

Bruno Mars - 24K Magic (Lyrics)
1.6M views • 1 year ago
Mr Shades
Bruno Mars - 24K Magic (Lyrics) Download/Stream '24K Magic': https://Atlantic.lnk.to/24K_MagicID Connect...
3:46

Bruno Mars - 24K Magic (Full Album)
131K views • 1 year ago
SOUL DEEP
Subscribe to our channel for all your favourite artists and songs! 24K Magic (Full Album)...
33:17

Figure 7: Top3 YouTube Results of Searching "24K Magic" + Bruno Mars

Figure 7 displays the top results of querying the keywords "*24K Magic*" (song title) and "*Bruno Mars*" (artist's name) on YouTube, which includes the official music video of the song uploaded by the artist's own channel as well as user-uploaded lyrics and album videos.

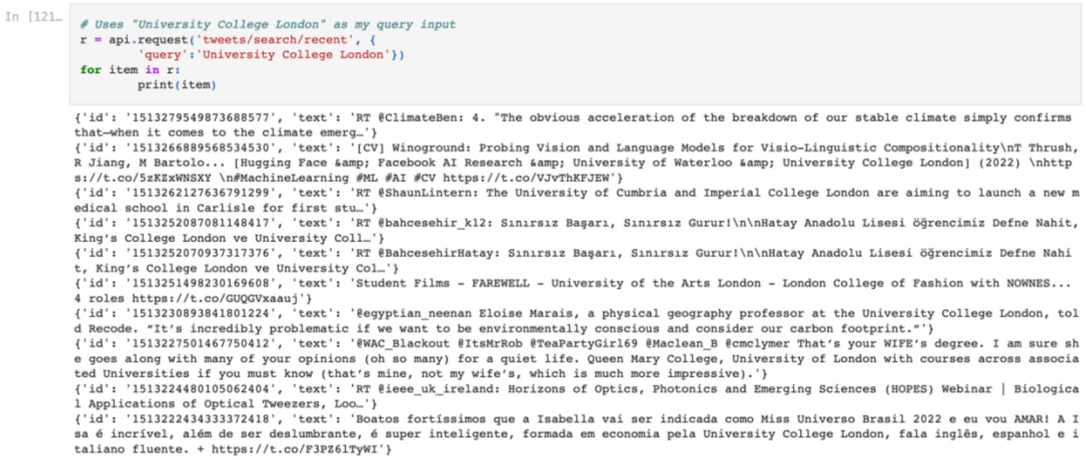
By using the YouTube V3 API's method *Search()*, this project first got the top 10 query results' YouTube video IDs for each of the 76 songs and stored them in a 76*10 dictionary. The number 10 was decided because they must be the most relevant videos to the song. To get the attributes like the view and like count for each video, the YouTube V3 *statistics()* method was used. This project retrieved a total of four statistics of each of the 760 videos: the view, like, favourite and comment count, and calculates the mean of each count for each of the 76 songs. The four attributes should reflect the popularity of the songs on YouTube.

2.4 TWITTER DATA

Twitter is one of the biggest social networking services in the world where users can post and interact with messages known as "tweets" (Igi-global, 2022). This project is also interested in the emotions associated with the tweets that discuss the tracks in the *Mood Booster* playlist.

To retrieve the tweet strings, the Twitter API was used, and a simple Python wrapper called *Twitter API* (https://github.com/geduldig/TwitterAPI) suggested by the Twitter developer's website was used for making the requests in Python.

The standard result of searching a random query ("University College London" in this case) is displayed in Figure 8.



```
In [121…
# Uses "University College London" as my query input
r = api.request('tweets/search/recent', {
        'query':'University College London'})
for item in r:
        print(item)
```

```
{'id': '1513279549873688577', 'text': 'RT @ClimateBen: 4. "The obvious acceleration of the breakdown of our stable climate simply confirms
that—when it comes to the climate emerg…'}
{'id': '1513266889568534530', 'text': '[CV] Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality\nT Thrush,
R Jiang, M Bartolo... [Hugging Face &amp; Facebook AI Research &amp; University of Waterloo &amp; University College London] (2022) \nhttp
s://t.co/5zKZxWNSXY \n#MachineLearning #ML #AI #CV https://t.co/VJvThKFJEW'}
{'id': '1513262127636791299', 'text': 'RT @ShaunLintern: The University of Cumbria and Imperial College London are aiming to launch a new m
edical school in Carlisle for first stu…'}
{'id': '1513252087081148417', 'text': 'RT @bahcesehir_k12: Sınırsız Başarı, Sınırsız Gurur!\n\nHatay Anadolu Lisesi öğrencimiz Defne Nahit,
King's College London ve University Coll…'}
{'id': '1513252070937317376', 'text': 'RT @BahcesehirHatay: Sınırsız Başarı, Sınırsız Gurur!\n\nHatay Anadolu Lisesi öğrencimiz Defne Nahi
t, King's College London ve University Col…'}
{'id': '1513251498230169608', 'text': 'Student Films - FAREWELL - University of the Arts London - London College of Fashion with NOWNES...
4 roles https://t.co/GUQGVxaauj'}
{'id': '1513230893841801224', 'text': '@egyptian_neenan Eloise Marais, a physical geography professor at the University College London, tol
d Recode. "It's incredibly problematic if we want to be environmentally conscious and consider our carbon footprint."'}
{'id': '1513227501467750412', 'text': '@WAC_Blackout @ItsMrRob @TeaPartyGirl69 @Maclean_B @cmclymer That's your WIFE's degree. I am sure sh
e goes along with many of your opinions (oh so many) for a quiet life. Queen Mary College, University of London with courses across associa
ted Universities if you must know (that's mine, not my wife's, which is much more impressive).'}
{'id': '1513224480105062404', 'text': 'RT @ieee_uk_ireland: Horizons of Optics, Photonics and Emerging Sciences (HOPES) Webinar | Biologica
l Applications of Optical Tweezers, Loo…'}
{'id': '1513222434333372418', 'text': 'Boatos fortíssimos que a Isabella vai ser indicada como Miss Universo Brasil 2022 e eu vou AMAR! A I
sa é incrível, além de ser deslumbrante, é super inteligente, formada em economia pela University College London, fala inglês, espanhol e i
taliano fluente. + https://t.co/F3PZ6lTyWI'}
```

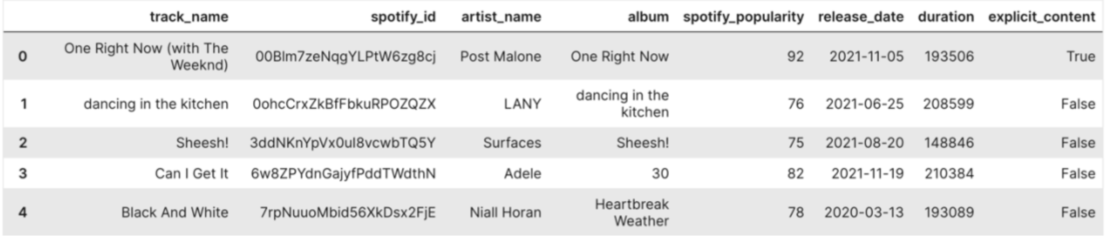
Figure 8: Output Example of the Twitter API

The project is only interested in the "text" section in the requesting result.

Similar to the previous two requests from MusixMatch and YouTube, this project also used the combination of the "song title" and the "artist's name" as the input parameter. By querying the 76 combinations and only retaining the "text" output, the last part of the database was successfully retrieved.

3.0
DATA PROCESSING

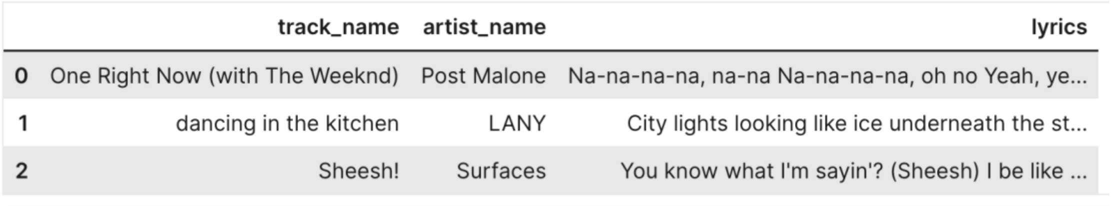
Most of the API result are in JSON formats or in a nested dictionary. By enumerating through these formats in Python, this project was able to generate four dataframes for four of the different platforms. Each dataframe has 76 rows and has the track name and artist's name columns as the foreign keys.



	track_name	spotify_id	artist_name	album	spotify_popularity	release_date	duration	explicit_content
0	One Right Now (with The Weeknd)	00Blm7zeNqgYLPtW6zg8cj	Post Malone	One Right Now	92	2021-11-05	193506	True
1	dancing in the kitchen	0ohcCrxZkBfFbkuRPOZQZX	LANY	dancing in the kitchen	76	2021-06-25	208599	False
2	Sheesh!	3ddNKnYpVx0ul8vcwbTQ5Y	Surfaces	Sheesh!	75	2021-08-20	148846	False
3	Can I Get It	6w8ZPYdnGajyfPddTWdthN	Adele	30	82	2021-11-19	210384	False
4	Black And White	7rpNuuoMbid56XkDsx2FjE	Niall Horan	Heartbreak Weather	78	2020-03-13	193089	False

Figure 9: *Spotify_df*

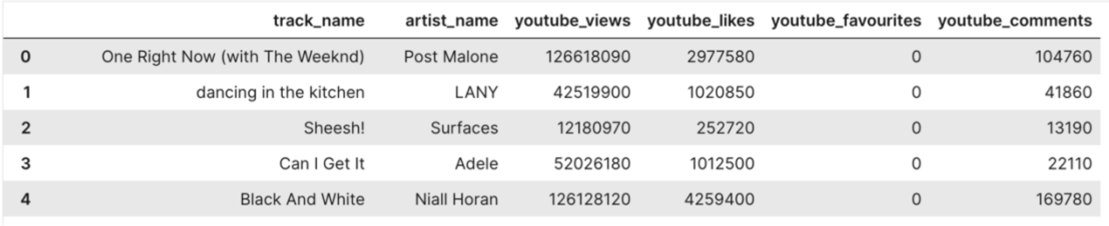
The main dataframe, *Spotify_df*, has 8 columns in total and 6 of them are unique Spotify attributes: the Spotify ID, album, Spotify popularity, release date, duration information of a song and whether it contains explicit content.



	track_name	artist_name	lyrics
0	One Right Now (with The Weeknd)	Post Malone	Na-na-na-na, na-na Na-na-na-na, oh no Yeah, ye...
1	dancing in the kitchen	LANY	City lights looking like ice underneath the st...
2	Sheesh!	Surfaces	You know what I'm sayin'? (Sheesh) I be like ...

Figure 10: *Musixmatch_df*

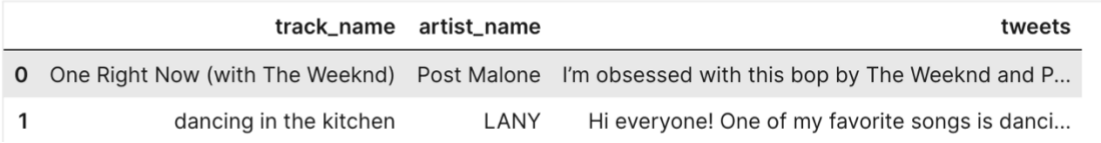
The *Musixmatch_df* dataframe has 3 columns and only the lyrics column is unique which contains the lyric string of each song.



	track_name	artist_name	youtube_views	youtube_likes	youtube_favourites	youtube_comments
0	One Right Now (with The Weeknd)	Post Malone	126618090	2977580	0	104760
1	dancing in the kitchen	LANY	42519900	1020850	0	41860
2	Sheesh!	Surfaces	12180970	252720	0	13190
3	Can I Get It	Adele	52026180	1012500	0	22110
4	Black And White	Niall Horan	126128120	4259400	0	169780

Figure 11: *Youtube_df*

The *Youtube_df* dataframe has 4 unique columns which are the average view, like, favourite and comment counts for the 10 most relevant videos of each of the 76 songs on YouTube.



	track_name	artist_name	tweets
0	One Right Now (with The Weeknd)	Post Malone	I'm obsessed with this bop by The Weeknd and P...
1	dancing in the kitchen	LANY	Hi everyone! One of my favorite songs is danci...

Figure 12: *Twitter_df*

Similar to the *Musixmatch_df* dataframe, the *Twitter_df* only contains one unique column, which is the tweet strings that discuss each of the 76 songs in the playlist.

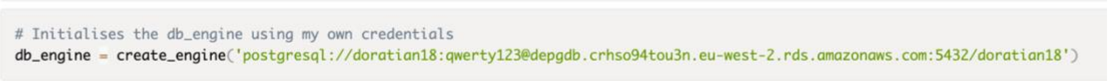
4.0
DATA STORAGE

4.1 LOCAL STORAGE

The four dataframes are first exported to 4 csv files for local storage. However, for more flexible, affordable, and scalable data management, the dataframes need to be stored in a more reliable cloud storage database.

4.2 CLOUD DATABASE

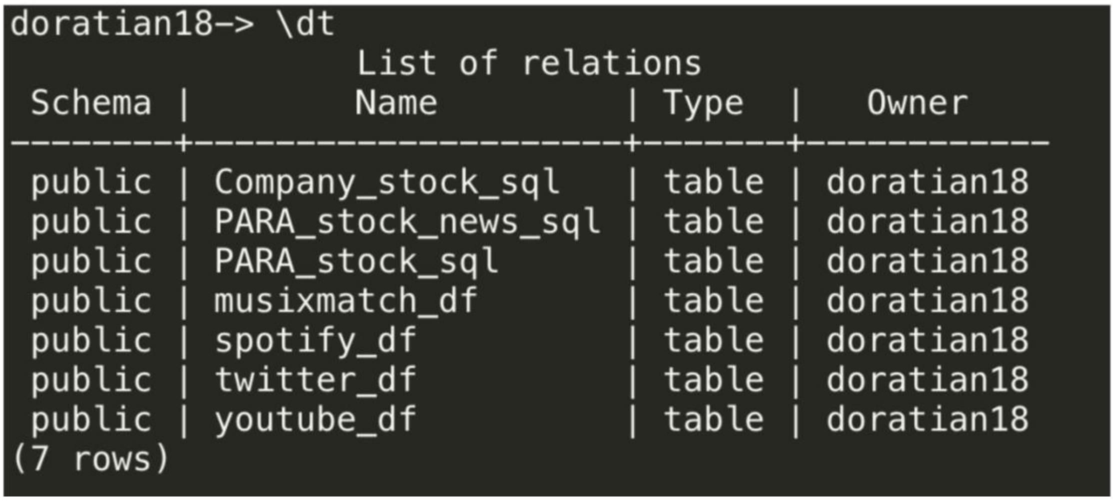
This project chose the PostgreSQL as the database management system. PostgreSQL is a powerful open-source object-relational database system with a solid reputation for active development and stability, functional robustness, and good performances for over 30 years (PostgreSQL, 2022). This system would also allow the project to conduct analyses via SQL queries via the Postgres connection and a relational database that had meaningful linkages could be created.



```
# Initialises the db_engine using my own credentials
db_engine = create_engine('postgresql://doratian18:qwerty123@depgdb.crhso94tou3n.eu-west-2.rds.amazonaws.com:5432/doratian18')
```

Figure 13: Initialising the Database Connection

With the user, host name and port number being initialised, the *db_engine* was created for future connections to this database.



```
doratian18-> \dt
                 List of relations
 Schema |         Name          | Type  |   Owner
--------+-----------------------+-------+------------
 public | Company_stock_sql     | table | doratian18
 public | PARA_stock_news_sql   | table | doratian18
 public | PARA_stock_sql        | table | doratian18
 public | musixmatch_df         | table | doratian18
 public | spotify_df            | table | doratian18
 public | twitter_df            | table | doratian18
 public | youtube_df            | table | doratian18
(7 rows)
```

Figure 14: Dataframes Stored in PostgreSQL

By connecting to the database in the terminal and using the command line prompts to check the tables, the last four rows in Figure 14 indicates that the four dataframes were successfully stored in the database and had the correct ownership.

5.0 RELATIONAL DATABASE

5.1 SCHEMA

Now that

5.2 SQL QUERIES

1234

6.0 EXPLORATORY DATA ANALYSIS

HEADING 2

View and edit this document in Word on your computer, tablet or phone. You can edit text, easily insert content such as pictures, shapes and tables, and seamlessly save the document to the cloud from Word on your Windows, Mac, Android or iOS device.

7.0 SENTIMENT ANALYSIS

7.1 WORD CLOUD

1234

7.2 NLTK SENTIMENT ANALYSIS

1234

8.0
REGRESSION MODEL

HEADING 2

View and edit this document in Word on your computer, tablet or phone. You can edit text, easily insert content such as pictures, shapes and tables, and seamlessly save the document to the cloud from Word on your Windows, Mac, Android or iOS device.

9.0 FUTURE OPPORTUNITIES

HEADING 2

View and edit this document in Word on your computer, tablet or phone. You can edit text, easily insert content such as pictures, shapes and tables, and seamlessly save the document to the cloud from Word on your Windows, Mac, Android or iOS device.

10.0
CONCLUSION

HEADING 2

View and edit this document in Word on your computer, tablet or phone. You can edit text, easily insert content such as pictures, shapes and tables, and seamlessly save the document to the cloud from Word on your Windows, Mac, Android or iOS device.

11.0 BIBLIOGRAPHIES

REPORT REFERENCES

Schriewer, K. and Bulaj, G., 2016. Music Streaming Services as Adjunct Therapies for Depression, Anxiety, and Bipolar Symptoms: Convergence of Digital Technologies, Mobile Apps, Emotions, and Global Mental Health. Frontiers in Public Health, 4.

Caddy, B., 2022. *The best music streaming services 2022: Spotify, Apple Music, Tidal and more*. [online] TechRadar. Available at: <https://www.techradar.com/best/the-best-music-streaming-services-2021> [Accessed 18 April 2022].

Medium. 2022. *Using an API for Web Scraping: A List of the Best Advantages*. [online] Available at: <https://medium.com/api-world/using-an-api-for-web-scraping-a-list-of-the-best-advantages-586e9fec2660> [Accessed 19 April 2022].

Soundplate. 2022. *What Does 'Explicit Content' Mean on Spotify, Apple Music & Other Music Streaming Platforms*. [online] Available at: <https://soundplate.com/what-does-explicit-content-mean-on-spotify-apple-music-other-music-streaming-platforms/> [Accessed 19 April 2022].

Baydeer, J., 2021. Let the Music Speak. [online] Medium. Available at: https://medium.com/swlh/let-the-music-speak-8c524ed45809 [Accessed 9 April 2022].

Igi-global.com. 2022. *What is Twitter | IGI Global*. [online] Available at: <https://www.igi-global.com/dictionary/i-found-myself-retweeting/30754> [Accessed 20 April 2022].

PostgreSQL. 2022. *PostgreSQL*. [online] Available at: <https://www.postgresql.org/> [Accessed 20 April 2022].

12.0 APPENDIX

HEADING 2

View and edit this document in Word on your computer, tablet or phone. You can edit text, easily insert content such as pictures, shapes and tables, and seamlessly save the document to the cloud from Word on your Windows, Mac, Android or iOS device.