TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐHQG TPHCM

BÁO CÁO ĐỒ ÁN

PROJECT 1: DATA COLLECTION

Môn: Nhập môn Khoa Học Dữ Liệu

Giảng viên: Lê Ngọc Thành

Thông tin nhóm:

Nhóm	MSSV	Họ và tên
	19127645	Bùi Đăng Khoa
	19127037	Võ Bách Khôi
	19127562	Chung Thế Thọ
	19127360	Dương Thị Xuân Diệu

Mục lục:

1. Tổng quan và mục tiêu của đồ án

○ Tổng quan

Trang SoundCloud là một trang web cho phép người dùng upload và chia sẻ các bài hát. Trong đồ án này, sinh viên sẽ thu thập thông tin về các bài hát, playlist và người dùng trên trang SoundCloud.

a. Làm quen và biết cách thu thập dữ liệu của một trang web bằng cách parse HTML và API.
b. Tự thiết kế và tổ chức các file notebook chứa mã nguồn, sao cho gọn gàng và dễ theo dõi.

2. Ý tưởng thực hiện đồ án

○ Bộ dữ liệu cần thu thập

Vì các file dữ liệu có dung lượng khá lớn, không thể nộp lên moodle, nên chúng tôi lưu tại link Google Drive.

■ User

- id: ID user.
- username: Tên/Bí danh của user.
- city: Nơi user sống/Nơi tạo tài khoản.
- country_code: Viết tắt chữ cái đầu của các nước.
- first_name + last_name: Họ + Tên của user.
- create_at: Thời gian tạo account.
- followers_count: Số người follow user.
- followings_count: Số người user follow.
- playlist_count: Số playlists đã tạo.
- track_count: Số track đã đăng tải.
- verified: User có tích xanh từ SoundCloud.
- likes_count: Lượt likes tổng tất cả các bài hát của user.
- comments_count: Tổng lượng comment trên tất cả bài hát của user.
- playlist_likes_count: Tổng lượng likes tất cả các playlist user đã tạo.
- last_modified: Lần cuối chỉnh sửa profile.

- avatar_url: Link url hình avatar của user.
- permalink_url: Link url dẫn đến account của user.

■ Playlist

- id: ID của playlist.
- title: Tên playlist.
- genre: Thể loại.
- duration: Thời lượng tổng các bài hát trong playlist
- likes_count: Lượt like của playlist.
- reposts_count: Lượng account repost lại playlist.
- is_album: Check nếu playlist là album
- release_date: ngày cụ thể khi track được phát hành đến công chúng
- created_at: Tạo vào ngày mấy.
- last_modified: Lần cuối chỉnh sửa
- label_name: Tên label (hãng) phân phối nhạc.
- purchase_title: tên của playlist khi bán.
- purchase_url: đường dẫn tới đến nơi để mua bán playlist
- license: Bản quyền.
- tag_list: Các tags liên quan đến playlist.
- user_id: ID user tạo playlist
- track_count: Số track trong playlist.
- published_at: Ngày giờ cụ thể khi playlist được phát hành đến công chúng
- display_date: Ngày giờ xuất hiện trên web.
- artwork_url: Link artwork.
- permalink_url: Link dẫn đến bài hát.

■ Track

- id: ID của track.
- title: Tên track.
- full_duration: Tổng thời lượng bài hát (tính theo ms).
- duration: Cũng là tổng thời lượng bài hát (tính theo ms). Thuộc tính này khi thu thập đa số bằng giá trị với full_duration. Nhưng khi một

track có duration khác full_duration, giá trị của duration luôn là 30000 và có một đặc điểm là các track này ban IP khu vực mà ta thu thập dữ liệu.

- playback_count: Số lần phát lại bài hát (số lượt nghe).
- likes_count: Lượt like của bài hát.
- reposts_count: Lượng account repost lại bài hát.
- comment_count: Lượng comment của bài hát.
- download_count: Lượng download bài hát
- downloadable: Có thể download được hay không
- commentable: Có thể comment vào bài hát được hay không
- has_download_left: hiện tại còn download được hay không. Có thể trước đó có thể download nhưng bây giờ thì không (downloadable = true, has_download_left = false)
- state: Trạng thái của track (demo hoặc finished)
- genre: Thể loại bài hát.
- streamable: Có thể set on streaming bài hát không.
- purchase_title: Địa chỉ mua bài hát.
- purchase_url: đường dẫn tới đến nơi để mua bán track
- tag_list: Các tags có liên quan đến bài hát.
- visuals: Ảnh đại diện của bài hát.
- user_id: ID của user sở hữu bài hát.
- release_date: Ngày cụ thể khi track được phát hành đến công chúng
- created_at: Ngày phát hành bài hát.
- last_modified: Lần cuối chỉnh sửa.
- license: Bản quyền.
- policy: Kiểm tra policy.
- waveform_url: Link waveform (đuôi .json).
- artwork_url: Link artwork.
- permalink_url: Link dẫn đến bài hát.

○ Thu thập dữ liệu bằng HTML

■ Ý tưởng thực hiện:

Để các dữ liệu có liên quan với nhau, chúng tôi quyết định lấy dữ liệu dựa theo thứ tự: User's Link -> User -> Playlist -> Track.

- **Thu thập User's link**: Được thực hiện theo 2 cách:
  - **Cách 1**: Dựa vào bảng xếp hạng "Top 50" và "New & hot" của Soundcloud. Cụ thể hơn, bằng sự hỗ trợ của thư viện Selenium, chúng tôi duyệt qua tất cả các thể loại nhạc (music genre) và thể loại âm thanh (audio genre), và ở mỗi bảng xếp hạng của các thể loại, chúng tôi lấy toàn bộ link của các user có trên bảng xếp hạng.
  - **Cách 2**: Trước tiên, chúng tôi tìm kiếm link các user thông qua chức năng search trên Soundcloud. Search trên Soundcloud sẽ tìm những người liên quan đến ký tự mình điền vào search, tuy nhiên, chúng tôi tìm được một ký tự Soundcloud không cho phép đặt tên chưa ký tự đó. Ký tự tôi muốn nhắc đến là " * ". Khi search bằng ký tự "*", kết quả trả về khá là ngẫu nhiên và những user nổi tiếng được đưa lên trên nhất. Vì số lượng dữ liệu khá lớn, nên tôi quyết định chỉ lấy những user được cho là có giá trị (good_user).
    "good_user" được cho là user có trên 10.000 follower hoặc được xác thực (verified)
- **Thu thập thông tin User**: Dựa trên "user link" đã lấy về bên trên, chúng tôi vào link các user đó để lấy thông tin trên đó.
- **Thu thập thông tin Playlist và Track**: Dựa trên "user link" đã lấy về bên trên, chúng tôi vào link và vào phần playlist của họ để lấy data. Trong playlist cũng có data về track nên chúng tôi tận dụng điều đó. Trong trường hợp User không có track, chúng tôi sẽ qua phần track.
- **Làm sạch dữ liệu**: Ở đây chúng tôi sẽ xử lý dữ liệu bị trùng lặp để khiến cho dữ liệu "sạch sẽ" hơn.
- **Đánh giá**:

- **Cách 1**:
  - Điểm mạnh: Do toàn bộ các user đều nằm trên bảng xếp hạng, nên đa phần các user chất lượng với số lượng track và cả lượng người theo dõi đều tương đối cao.
  - Điểm yếu: Do là dynamic page nên việc crawl dữ liệu theo cách này khá mất thời gian. Bên cạnh đó, do chỉ phụ thuộc vào các bảng xếp hạng, sự đa dạng và số lượng các user thu thập được theo cách này khá là hạn chế.
- **Cách 2**:
  - Điểm mạnh: User đa phần đều là user chất lượng. Khi đã có nhiều người theo dõi thì khả năng cao sẽ có nhiều playlist và track. Mặc dù vậy, vẫn tồn tại Good user nhưng không có track và playlist.
  - Điểm yếu: Tuy nhiên, sau khi tìm hiểu, tôi đã thấy parse theo cách này sẽ có những tình huống sau như, user có rất nhiều playlist, nhưng đều là playlist rỗng. Các bài hát trong playlist, có một số bài hát dường như không có thông tin. Vì vậy, có rất nhiều bài có trong track của playlist nhưng không có trong file track.csv.

■ Các bước thực hiện

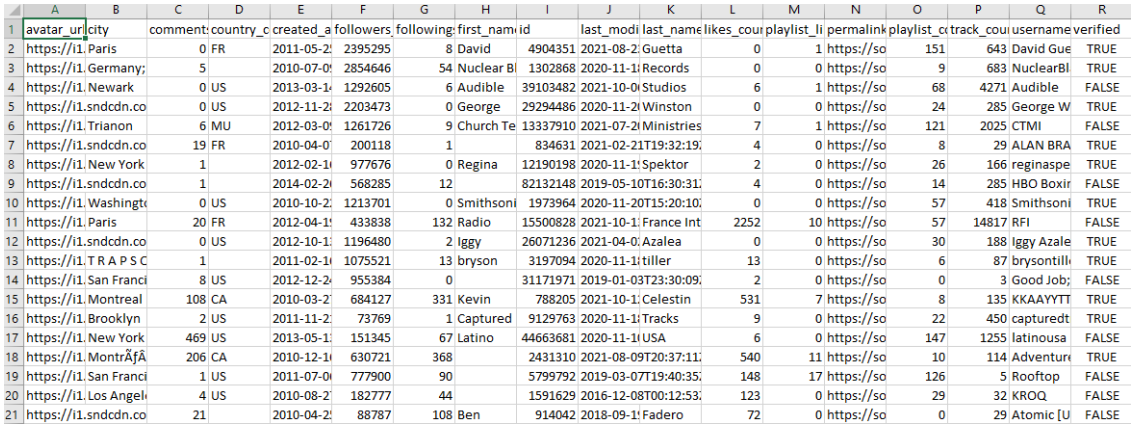
Thứ tự thực hiện các notebook:

1. **Cách 1:**
   1.1. get_urls.ipynb (kết quả "urls.csv")
   1.2. main.ipynb (kết quả "user.csv", "playlist.csv" và "track.csv")
   1.3. preprocessing.ipynb (kết quả "user.csv", "playlist.csv" và "track.csv")
2. **Cách 2**:
   2.1. Crawl_userslink.ipynb (kết quả "good_user.txt")
   2.2. Parse user.ipynb (kết quả "user.csv")
   2.3. Parse-playlist-track.ipynb (kết quả "track.csv" và "playlist.csv")

*Các bước thực hiện được chú thích rõ ràng, đầy đủ trong các file .ipynb.

■ Kết quả sau khi thực hiện

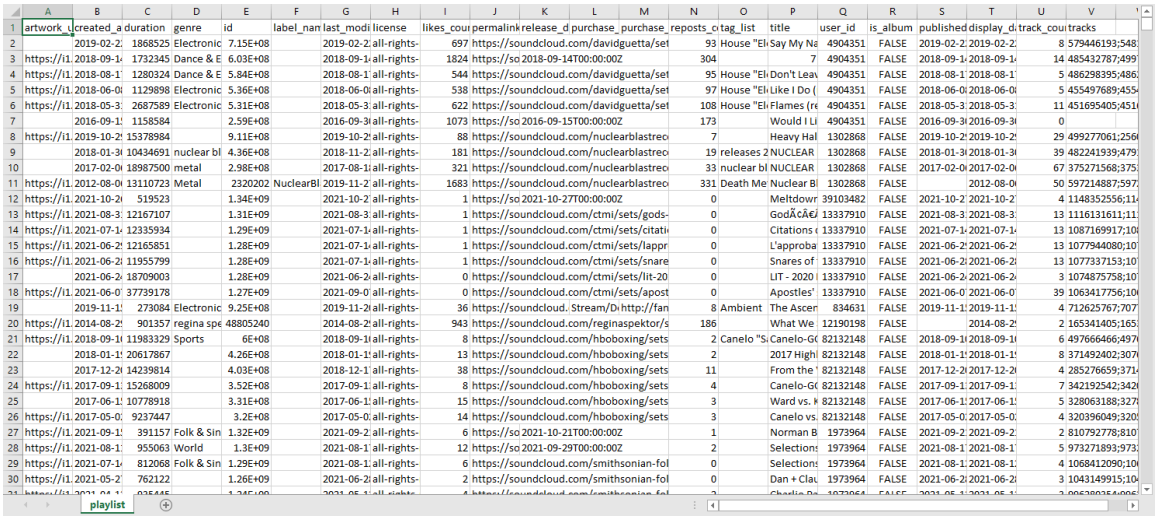
"User.csv": 2028 (Cách 1) và 2569 records (Cách 2)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	avatar_url	city	comments	country_c	created_a	followers	following	first_nam	id	last_modi	last_name	likes_cou	playlist_li	permalink	playlist_c	track_cou	username	verified
2	https://i1.	Paris	0	FR	2011-05-2	2395295	8	David	4904351	2021-08-2	Guetta	0	1	https://so	151	643	David Gue	TRUE
3	https://i1.	Germany;	5		2010-07-0	2854646	54	Nuclear Bl	1302868	2020-11-1	Records	0	0	https://so	9	683	NuclearBl	TRUE
4	https://i1.	Newark	0	US	2013-03-1	1292605	6	Audible	39103482	2021-10-0	Studios	6	1	https://so	68	4271	Audible	FALSE
5	https://i1.sndcdn.co		0	US	2012-11-2	2203473	0	George	29294486	2020-11-2	Winston	0	0	https://so	24	285	George W	FALSE
6	https://i1.	Trianon	6	MU	2012-03-0	1261726	9	Church Te	13337910	2021-07-2	Ministries	7	1	https://so	121	2025	CTMI	FALSE
7	https://i1.sndcdn.co		19	FR	2010-04-0	200118	1		834631	2021-02-21	T19:32:19	4	0	https://so	8	29	ALAN BRA	TRUE
8	https://i1.	New York	1		2012-02-1	977676	0	Regina	12190198	2020-11-1	Spektor	2	0	https://so	26	166	reginaspe	TRUE
9	https://i1.sndcdn.co		1		2014-02-2	568285	12		82132148	2019-05-10	T16:30:31	4	0	https://so	14	285	HBO Boxir	FALSE
10	https://i1.	Washingt	0	US	2010-10-2	1213701	0	Smithsoni	1973964	2020-11-20	T15:20:10	0	0	https://so	57	418	Smithsoni	TRUE
11	https://i1.	Paris	20	FR	2012-04-1	433838	132	Radio	15500828	2021-10-1	France Int	2252	10	https://so	57	14817	RFI	FALSE
12	https://i1.sndcdn.co		0	US	2012-10-1	1196480	2	Iggy	26071236	2021-04-0	Azalea	0	0	https://so	30	188	Iggy Azale	TRUE
13	https://i1.	T R A P S C	1		2011-02-1	1075521	13	bryson	3197094	2020-11-1	tiller	13	0	https://so	6	87	brysontill	TRUE
14	https://i1.	San Franci	8	US	2012-12-2	955384	0		31171971	2019-01-03	T23:30:09	2	0	https://so	0	3	Good Job;	FALSE
15	https://i1.	Montreal	108	CA	2010-03-2	684127	331	Kevin	788205	2021-10-1	Celestin	531	7	https://so	8	135	KKAAYYTT	TRUE
16	https://i1.	Brooklyn	2	US	2011-11-2	73769	1	Captured	9129763	2020-11-1	Tracks	9	0	https://so	22	450	capturedt	TRUE
17	https://i1.	New York	469	US	2013-05-1	151345	67	Latino	44663681	2020-11-1	USA	6	0	https://so	147	1255	latinousa	FALSE
18	https://i1.	MontrÃƒÂ	206	CA	2010-12-1	630721	368		2431310	2021-08-09	T20:37:11	540	11	https://so	10	114	Adventure	TRUE
19	https://i1.	San Franci	1	US	2011-07-0	777900	90		5799792	2019-03-07	T19:40:35	148	17	https://so	126	5	Rooftop	FALSE
20	https://i1.	Los Angel	4	US	2010-08-2	182777	44		1591629	2016-12-08	T00:12:53	123	0	https://so	29	32	KROQ	FALSE
21	https://i1.sndcdn.co		21		2010-04-2	88787	108	Ben	914042	2018-09-1	Fadero	72	0	https://so	0	29	Atomic [U	FALSE

Mặc dù chúng tôi vẫn có thể thu thập nhiều trường dữ liệu hơn, nhưng đây là các trường chúng tôi nghĩ là có giá trị về mặt phân tích cũng như lưu trữ. Những trường không có giá trị, chúng tôi bỏ qua và không lưu.

- "playlist.csv": 3764 records (Cách 1) và 5309 records (Cách 2)

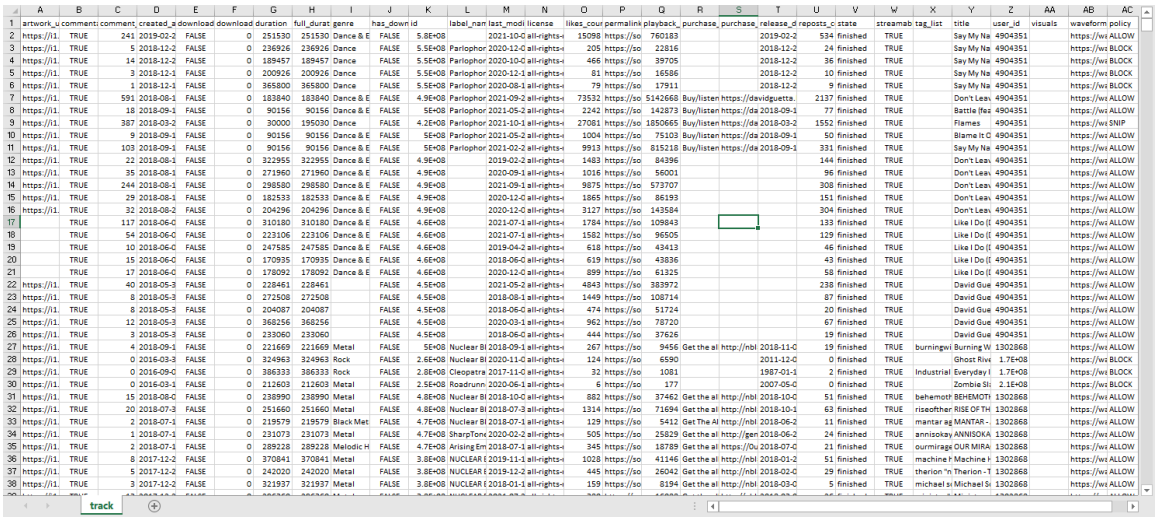


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	artwork_u	created_a	duration	genre	id	label_nam	last_modi	license	likes_cou	permalink	release_d	purchase_	purchase_	reposts_c	tag_list	title	user_id	is_album	published	display_d	track_cou	tracks
2		2019-02-2	1868525	Electronic	7.15E+08		2019-02-2	all-rights-	697	https://soundcloud.com/davidguetta/set				93	House "El	Say My Na	4904351	FALSE	2019-02-2	2019-02-2	8	579446193;548
3	https://i1.	2018-09-1	1732345	Dance & E	6.03E+08		2018-09-1	all-rights-	1824	https://so	2018-09-14	T00:00:00Z		304		7	4904351	FALSE	2018-09-1	2018-09-1	14	485432787;499
4	https://i1.	2018-08-1	1280324	Dance & E	5.84E+08		2018-08-1	all-rights-	544	https://soundcloud.com/davidguetta/set				95	House "El	Don't Leav	4904351	FALSE	2018-08-1	2018-08-1	5	486298395;486
5	https://i1.	2018-06-0	1129898	Electronic	5.36E+08		2018-06-0	all-rights-	538	https://soundcloud.com/davidguetta/set				97	House "El	Like I Do (	4904351	FALSE	2018-06-0	2018-06-0	5	455497689;455
6	https://i1.	2018-05-3	2687589	Electronic	5.31E+08		2018-05-3	all-rights-	622	https://soundcloud.com/davidguetta/set				108	House "El	Flames (re	4904351	FALSE	2018-05-3	2018-05-3	11	451695405;451
7		2016-09-1	1158584		2.59E+08		2016-09-3	all-rights-	1073	https://so	2016-09-15	T00:00:00Z		173		Would I Li	4904351	FALSE	2016-09-3	2016-09-3	0	
8	https://i1.	2019-10-2	15378984		9.11E+08		2019-10-2	all-rights-	88	https://soundcloud.com/nuclearblastrec				7		Heavy Hal	1302868	FALSE	2019-10-2	2019-10-2	29	499277061;256
9		2018-01-3	10434691	nuclear bl	4.36E+08		2018-11-2	all-rights-	181	https://soundcloud.com/nuclearblastrec				19	releases 2	NUCLEAR	1302868	FALSE	2018-01-3	2018-01-3	39	482241939;479
10		2017-02-0	18987500	metal	2.98E+08		2017-08-1	all-rights-	321	https://soundcloud.com/nuclearblastrec				33	nuclear bl	NUCLEAR	1302868	FALSE	2017-02-0	2017-02-0	67	375271568;375
11	https://i1.	2012-08-0	13110723	Metal	2320202	NuclearBl	2019-11-2	all-rights-	1683	https://soundcloud.com/nuclearblastrec				331	Death Me	Nuclear Bl	1302868	FALSE		2012-08-0	50	597214887;597
12	https://i1.	2021-10-2	519523		1.34E+09		2021-10-2	all-rights-	1	https://so	2021-10-27	T00:00:00Z		0		Meltdowr	39103482	FALSE	2021-10-2	2021-10-2	4	1148352556;114
13	https://i1.	2021-08-3	12167107		1.31E+09		2021-08-3	all-rights-	1	https://soundcloud.com/ctmi/sets/gods-				0		GodÃ¢Â€Â	13337910	FALSE	2021-08-3	2021-08-3	13	1116131611;11
14	https://i1.	2021-07-1	12335934		1.29E+09		2021-07-1	all-rights-	1	https://soundcloud.com/ctmi/sets/citati				0		Citations (	13337910	FALSE	2021-07-1	2021-07-1	13	1087169917;10
15	https://i1.	2021-06-2	12165851		1.28E+09		2021-07-1	all-rights-	1	https://soundcloud.com/ctmi/sets/lappr				0		L'approba	13337910	FALSE	2021-06-2	2021-06-2	13	1077944080;10
16	https://i1.	2021-06-2	11955799		1.28E+09		2021-07-1	all-rights-	1	https://soundcloud.com/ctmi/sets/snare				0		Snares of	13337910	FALSE	2021-06-2	2021-06-2	13	1077337153;10
17		2021-06-2	18709003		1.28E+09		2021-06-2	all-rights-	0	https://soundcloud.com/ctmi/sets/lit-20				0		LIT - 2020	13337910	FALSE	2021-06-2	2021-06-2	3	1074875758;10
18	https://i1.	2021-06-0	37739178		1.27E+09		2021-09-0	all-rights-	0	https://soundcloud.com/ctmi/sets/apost				0		Apostles'	13337910	FALSE	2021-06-0	2021-06-0	39	1063417756;10
19		2019-11-1	273084	Electronic	9.25E+08		2019-11-2	all-rights-	36	https://soundcloud.	Stream/D	http://fan		8	Ambient	The Ascer	834631	FALSE	2019-11-1	2019-11-1	4	712625767;707
20	https://i1.	2014-08-2	901357	regina spe	48805240		2014-08-2	all-rights-	943	https://soundcloud.com/reginaspektor/s				186		What We	12190198	FALSE		2014-08-2	2	165341405;165
21	https://i1.	2018-09-1	11983329	Sports	6E+08		2018-09-1	all-rights-	8	https://soundcloud.com/hboboxing/sets				2	Canelo "S	Canelo-G(	82132148	FALSE	2018-09-1	2018-09-1	6	497666466;497
22		2018-01-1	20617867		4.26E+08		2018-01-1	all-rights-	13	https://soundcloud.com/hboboxing/sets				2		2017 High	82132148	FALSE	2018-01-1	2018-01-1	8	371492402;307
23		2017-12-2	14239814		4.03E+08		2018-12-1	all-rights-	38	https://soundcloud.com/hboboxing/sets				11		From the '	82132148	FALSE	2017-12-2	2017-12-2	4	285276659;371
24	https://i1.	2017-09-1	15268009		3.52E+08		2017-09-1	all-rights-	8	https://soundcloud.com/hboboxing/sets				4		Canelo-G(	82132148	FALSE	2017-09-1	2017-09-1	7	342192542;342
25		2017-06-1	10778918		3.31E+08		2017-06-1	all-rights-	15	https://soundcloud.com/hboboxing/sets				3		Ward vs. K	82132148	FALSE	2017-06-1	2017-06-1	5	328063188;327
26	https://i1.	2017-05-0	9237447		3.2E+08		2017-05-0	all-rights-	14	https://soundcloud.com/hboboxing/sets				3		Canelo vs.	82132148	FALSE	2017-05-0	2017-05-0	4	320396049;320
27	https://i1.	2021-09-1	391157	Folk & Sin	1.32E+09		2021-09-2	all-rights-	6	https://so	2021-10-21	T00:00:00Z		1		Norman B	1973964	FALSE	2021-09-2	2021-09-2	2	810792778;810
28	https://i1.	2021-08-1	955063	World	1.3E+09		2021-08-1	all-rights-	12	https://so	2021-09-29	T00:00:00Z		2		Selections	1973964	FALSE	2021-08-1	2021-08-1	5	973271893;973
29	https://i1.	2021-07-1	812068	Folk & Sin	1.29E+09		2021-08-1	all-rights-	6	https://soundcloud.com/smithsonian-fol				0		Selections	1973964	FALSE	2021-08-1	2021-08-1	4	1068412090;10
30	https://i1.	2021-05-2	762122		1.26E+09		2021-06-2	all-rights-	2	https://soundcloud.com/smithsonian-fol				0		Dan + Clau	1973964	FALSE	2021-06-2	2021-06-2	3	1043149915;10
31	https://i1.	2021-04-1	935445		1.24E+09		2021-05-1	all-rights-	4	https://soundcloud.com/smithsonian-fol				2		Charlie Pa	1973964	FALSE	2021-05-1	2021-05-1	3	996280954;996

playlist

Tương tự như “user.csv”, chúng tôi chỉ lưu những trường có giá trị. Mỗi user chúng tôi lấy tối đa 6 playlists. Ở trường track, mỗi track được ngăn cách bởi dấu “;”

- “track.csv”: 13767 records (Cách 1) và 5309 records (Cách 2)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	artwork_u	comments	comment	created_a	download	download	duration	full_durat	genre	has_down	id	label_nam	last_modi	license	likes_cour	permalink	playback_	purchase_	purchase_	release_d	reposts_c	state	streamab	tag_list	title	user_id	visuals	waveform	policy
2	https://i1.	TRUE	241	2019-02-2	FALSE	0	251530	251530	Dance & E	FALSE	5.8E+08		2021-10-0	all-rights-r	15098	https://so	760183			2019-02-2	534	finished	TRUE		Say My Na	4904351		https://wa	ALLOW
3	https://i1.	TRUE	5	2018-12-2	FALSE	0	236926	236926	Dance	FALSE	5.5E+08	Parlophor	2020-12-0	all-rights-r	205	https://so	22816			2018-12-2	24	finished	TRUE		Say My Na	4904351		https://wa	BLOCK
4	https://i1.	TRUE	14	2018-12-2	FALSE	0	189457	189457	Dance	FALSE	5.5E+08	Parlophor	2020-10-0	all-rights-r	466	https://so	39705			2018-12-2	36	finished	TRUE		Say My Na	4904351		https://wa	BLOCK
5	https://i1.	TRUE	3	2018-12-1	FALSE	0	200926	200926	Dance	FALSE	5.5E+08	Parlophor	2020-12-1	all-rights-r	81	https://so	16586			2018-12-2	10	finished	TRUE		Say My Na	4904351		https://wa	BLOCK
6	https://i1.	TRUE	1	2018-12-1	FALSE	0	365800	365800	Dance	FALSE	5.5E+08	Parlophor	2020-08-1	all-rights-r	79	https://so	17911			2018-12-2	9	finished	TRUE		Say My Na	4904351		https://wa	BLOCK
7	https://i1.	TRUE	591	2018-08-1	FALSE	0	183840	183840	Dance & E	FALSE	4.9E+08	Parlophor	2021-09-2	all-rights-r	73532	https://so	5142668	Buy/listen	https://davidguetta.		2137	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
8	https://i1.	TRUE	18	2018-09-1	FALSE	0	90156	90156	Dance & E	FALSE	5E+08	Parlophor	2021-05-2	all-rights-r	2242	https://so	142873	Buy/listen	https://da	2018-09-1	77	finished	TRUE		Battle (fea	4904351		https://wa	ALLOW
9	https://i1.	TRUE	387	2018-03-2	FALSE	0	30000	195030	Dance	FALSE	4.2E+08	Parlophor	2021-10-1	all-rights-r	27081	https://so	1850665	Buy/listen	https://da	2018-03-2	1552	finished	TRUE		Flames	4904351		https://wa	SNIP
10	https://i1.	TRUE	9	2018-09-1	FALSE	0	90156	90156	Dance & E	FALSE	5E+08	Parlophor	2021-05-2	all-rights-r	1004	https://so	75103	Buy/listen	https://da	2018-09-1	50	finished	TRUE		Blame it C	4904351		https://wa	ALLOW
11	https://i1.	TRUE	103	2018-09-1	FALSE	0	90156	90156	Dance & E	FALSE	5E+08	Parlophor	2021-02-2	all-rights-r	9913	https://so	815218	Buy/listen	https://da	2018-09-1	331	finished	TRUE		Say My Na	4904351		https://wa	ALLOW
12	https://i1.	TRUE	22	2018-08-1	FALSE	0	322955	322955	Dance & E	FALSE	4.9E+08		2019-02-2	all-rights-r	1483	https://so	84396				144	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
13	https://i1.	TRUE	35	2018-08-1	FALSE	0	271960	271960	Dance & E	FALSE	4.9E+08		2020-09-1	all-rights-r	1016	https://so	56001				96	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
14	https://i1.	TRUE	244	2018-08-1	FALSE	0	298580	298580	Dance & E	FALSE	4.9E+08		2021-09-1	all-rights-r	9875	https://so	573707				308	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
15	https://i1.	TRUE	29	2018-08-1	FALSE	0	182533	182533	Dance & E	FALSE	4.9E+08		2020-12-0	all-rights-r	1865	https://so	86193				151	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
16	https://i1.	TRUE	32	2018-08-2	FALSE	0	204296	204296	Dance & E	FALSE	4.9E+08		2020-12-0	all-rights-r	3127	https://so	143584				304	finished	TRUE		Don't Lea	4904351		https://wa	ALLOW
17		TRUE	117	2018-06-0	FALSE	0	310180	310180	Dance & E	FALSE	4.6E+08		2021-07-1	all-rights-r	1784	https://so	109843				133	finished	TRUE		Like I Do ([	4904351		https://wa	ALLOW
18		TRUE	54	2018-06-0	FALSE	0	223106	223106	Dance & E	FALSE	4.6E+08		2021-07-1	all-rights-r	1582	https://so	96505				129	finished	TRUE		Like I Do ([	4904351		https://wa	ALLOW
19		TRUE	10	2018-06-0	FALSE	0	247585	247585	Dance & E	FALSE	4.6E+08		2019-04-2	all-rights-r	618	https://so	43413				46	finished	TRUE		Like I Do ([	4904351		https://wa	ALLOW
20		TRUE	15	2018-06-0	FALSE	0	170935	170935	Dance & E	FALSE	4.6E+08		2018-06-0	all-rights-r	619	https://so	43836				43	finished	TRUE		Like I Do ([	4904351		https://wa	ALLOW
21		TRUE	17	2018-06-0	FALSE	0	178092	178092	Dance & E	FALSE	4.6E+08		2020-12-0	all-rights-r	899	https://so	61325				58	finished	TRUE		Like I Do ([	4904351		https://wa	ALLOW
22	https://i1.	TRUE	40	2018-05-3	FALSE	0	228461	228461		FALSE	4.5E+08		2021-05-2	all-rights-r	4843	https://so	383972				238	finished	TRUE		David Gue	4904351		https://wa	ALLOW
23	https://i1.	TRUE	8	2018-05-3	FALSE	0	272508	272508		FALSE	4.5E+08		2018-08-1	all-rights-r	1449	https://so	108714				87	finished	TRUE		David Gue	4904351		https://wa	ALLOW
24	https://i1.	TRUE	8	2018-05-3	FALSE	0	204087	204087		FALSE	4.5E+08		2018-06-0	all-rights-r	474	https://so	51724				20	finished	TRUE		David Gue	4904351		https://wa	ALLOW
25	https://i1.	TRUE	12	2018-05-3	FALSE	0	368256	368256		FALSE	4.5E+08		2020-03-1	all-rights-r	962	https://so	78720				67	finished	TRUE		David Gue	4904351		https://wa	ALLOW
26	https://i1.	TRUE	3	2018-05-3	FALSE	0	233060	233060		FALSE	4.5E+08		2018-06-0	all-rights-r	444	https://so	37626				19	finished	TRUE		David Gue	4904351		https://wa	ALLOW
27	https://i1.	TRUE	4	2018-09-1	FALSE	0	221669	221669	Metal	FALSE	5E+08	Nuclear Bl	2018-09-1	all-rights-r	267	https://so	9456	Get the al	http://nbl	2018-11-0	19	finished	TRUE	burningwi	Burning W	1302868		https://wa	ALLOW
28	https://i1.	TRUE	0	2016-03-3	FALSE	0	324963	324963	Rock	FALSE	2.6E+08	Nuclear Bl	2020-11-0	all-rights-r	124	https://so	6590			2011-12-0	0	finished	TRUE		Ghost Rive	1.7E+08		https://wa	BLOCK
29	https://i1.	TRUE	0	2016-09-0	FALSE	0	386333	386333	Rock	FALSE	2.8E+08	Cleopatra	2017-11-0	all-rights-r	32	https://so	1081			1987-01-1	2	finished	TRUE	Industrial	Everyday I	1.7E+08		https://wa	BLOCK
30	https://i1.	TRUE	0	2016-03-1	FALSE	0	212603	212603	Metal	FALSE	2.5E+08	Roadrunn	2020-06-1	all-rights-r	6	https://so	177			2007-05-0	0	finished	TRUE		Zombie Sli	2.1E+08		https://wa	BLOCK
31	https://i1.	TRUE	15	2018-08-0	FALSE	0	238990	238990	Metal	FALSE	4.8E+08	Nuclear Bl	2018-10-0	all-rights-r	882	https://so	37462	Get the al	http://nbl	2018-10-0	51	finished	TRUE	behemoth	BEHEMOTH	1302868		https://wa	ALLOW
32	https://i1.	TRUE	20	2018-07-3	FALSE	0	251660	251660	Metal	FALSE	4.8E+08	Nuclear Bl	2018-07-3	all-rights-r	1314	https://so	71694	Get the al	http://nbl	2018-10-1	63	finished	TRUE	riseofther	RISE OF TH	1302868		https://wa	ALLOW
33	https://i1.	TRUE	2	2018-07-1	FALSE	0	219579	219579	Black Met	FALSE	4.7E+08	Nuclear Bl	2018-07-1	all-rights-r	129	https://so	5412	Get The Al	http://nbl	2018-06-2	11	finished	TRUE	mantar ag	MANTAR -	1302868		https://wa	ALLOW
34	https://i1.	TRUE	1	2018-07-1	FALSE	0	231073	231073	Metal	FALSE	4.7E+08	SharpTone	2020-02-2	all-rights-r	505	https://so	25829	Get the al	http://gen	2018-06-2	24	finished	TRUE	annisokay	ANNISOKA	1302868		https://wa	ALLOW
35	https://i1.	TRUE	2	2018-07-1	FALSE	0	289228	289228	Melodic H	FALSE	4.7E+08	Arising Em	2018-07-1	all-rights-r	345	https://so	18789	Get the al	https://ou	2018-07-0	21	finished	TRUE	ourmirage	OUR MIRA	1302868		https://wa	ALLOW
36	https://i1.	TRUE	8	2017-12-2	FALSE	0	370841	370841	Metal	FALSE	3.8E+08	NUCLEAR E	2019-11-1	all-rights-r	1028	https://so	41146	Get the al	http://nbl	2018-01-2	51	finished	TRUE	machine h	Machine H	1302868		https://wa	ALLOW
37	https://i1.	TRUE	5	2017-12-2	FALSE	0	242020	242020	Metal	FALSE	3.8E+08	NUCLEAR E	2019-12-2	all-rights-r	445	https://so	26042	Get the al	http://nbl	2018-02-0	29	finished	TRUE	therion ^n	Therion - T	1302868		https://wa	ALLOW
38	https://i1.	TRUE	3	2017-12-2	FALSE	0	321937	321937	Metal	FALSE	3.8E+08	NUCLEAR E	2018-01-1	all-rights-r	159	https://so	8194	Get the al	http://nbl	2018-03-0	5	finished	TRUE	michael si	Michael Si	1302868		https://wa	ALLOW

track

Tương tự như “user.csv” và playlist, chúng tôi chỉ lưu những trường hợp có giá trị. Các track nằm trong playlist sẽ chưa chắc nằm trong track.csv.

■ Đánh giá

Nhìn chung, 2 cách đều lấy đều mang lại dữ liệu khá tốt.

Tuy nhiên, ở cách 2, có thể bị bias vì lấy người ở thành phố London khá nhiều. Còn cách 1 thì lại không lấy được đa dạng dữ liệu bằng và phụ thuộc vào bản xếp hạng.

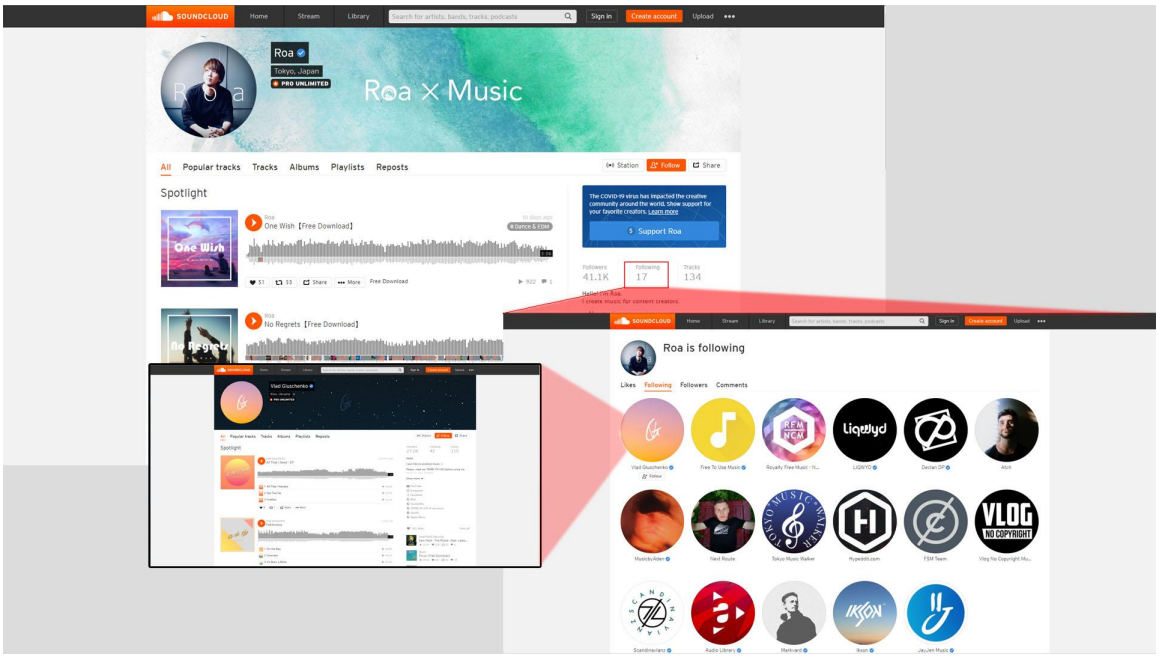
Điểm yếu chung là track của playlist nhưng không có trong file track.csv.

Vì để thầy dễ theo dõi nên chúng tôi chưa merge các file user.csv, track.csv và playlist.csv lại với nhau.

○ Thu thập dữ liệu bằng API SoundCloud

■ Ý tưởng thực hiện

Để tận dụng các tính năng API SoundCloud cung cấp, chúng tôi tiến hành thu thập dữ liệu theo cơ chế lan truyền từ một user này sang các user khác. Trong quá trình lan truyền này, chúng tôi chỉ quan tâm ID của user. Sau khi có được danh sách ID của các user, chúng tôi tiến hành lấy thông tin, các track và playlist của user đó dựa vào các api tương ứng.



API của SoundCloud có dạng tổng quát như sau:

https://api-v2.soundcloud.com/<entity>/<entity_id>?client_id=<client_id>

Trong đó:

- **entity:** loại object cần lấy thông tin. SoundCloud dùng danh từ số nhiều để định danh các object này như users, tracks, playlists.
- **entity_id:** là ID của object tương ứng
- **client_id:** là ID SoundCloud cung cấp cho phép ta khai thác API. Nhưng hiện nay việc đăng ký lấy client_id này đã ngừng hoạt động. Tuy nhiên ta có thể dùng thủ thuật để lấy client_id từ browser qua

inspect. Ví dụ ở Chrome ta ấn chuột phải inspect hoặc ấn F12, rồi ấn sang tab networks rồi ta download về. Ta mở file vừa tải về bằng bất kỳ text editor nào rồi dùng cơ chế tìm kiếm của text editor để tìm cụm từ "client_id=", khi đó ta có thể tìm được một kết quả có dạng như sau:
https://api.soundcloud.com/tracks/123456789/download?client_id=2t9loNQH90kzJcsFCODdigxfp325aq4z&oauth_token=2-274121-85658-y9KQYyZ6qG9oT2uvPq thì client_id ta có thể dùng là nằm ở sau "client_id=" và phía trước "&".

Cách lấy user_ID: Chúng tôi lấy các userID bằng cách tìm các followings của 1 user nhất định, link url API có dạng như sau:
https://api-v2.soundcloud.com/users/{list_user_id[index]}/followings?client_id={client_id}
Parse users: Dựa trên "user-id" đã lấy về bên trên, vào link url API như sau để lấy thông tin của các user đó:
https://api-v2.soundcloud.com/users/{user_id}?client_id={client_id
Parse playlists: Dựa trên "user-id" đã lấy về bên trên, vào link url API của playlists để lấy dữ liệu về playlists như sau:
https://api-v2.soundcloud.com/users/{user_id}/playlists?client_id={client_id}&limit=100

Chú ý: ta có thêm tham số limit=100 để báo API rằng ta chỉ lấy 100 request mỗi lần request.

Parse tracks: Dựa trên "user-id" đã lấy về bên trên, vào link url API của tracks dữ liệu về tracks như sau:
https://api-v2.soundcloud.com/users/{user_id}/tracks?client_id={client_id}&limit=100

Tương tự playlists, ở track ta cũng thêm tham số limit.

Đánh giá:

- **Điểm mạnh:** Tập dữ liệu user, track và playlist liên quan với nhau. Ngoài ra có thể tìm được khá nhiều "user-id" (10,000 Users) bằng cách tìm từ followings lần lượt của từng users. Các users này có rất nhiều playlists và tracks, dữ liệu về số playlist lên đến hơn 117,885 records, dữ liệu về tracks lên đến hơn 619,248 records.
- **Điểm yếu:** Tập dữ liệu bị bias hay không sẽ phụ thuộc vào user ban đầu cũng như số lượng user lấy được (càng ít càng dễ bias). Ngoài ra còn một điểm trừ khác là các user có lượng follow bằng không sẽ không được khai thác.

■ Các bước thực hiện

Thứ tự parse:

- get_userID.ipynb (kết quả user_id.txt").
- get_user.ipynb (kết quả "user.csv").
- get_playlist.ipynb (kết quả "playlist.csv").
- get_track.ipynb (kết quả "track.csv").

Các bước thực hiện cụ thể đã được ghi rõ trong từng file .ipynb

■ Kết quả sau khi thực hiện

"user.csv": 10,000 records.

	Unnamed: 0	id	username	city	country_code	first_name	last_name	created_at	followers_count	followings_count	playlist_count	tracl
0	0	630253005	Roa	Tokyo	JP	NaN	NaN	2019-05-02T08:00:39Z	40963	17	0	
1	1	217441590	Free To Use Music	NaN	NaN	NaN	NaN	2016-04-06T15:07:33Z	14000	201	5	
2	2	324531068	Royalty Free Music - No Copyright Music	NaN	NaN	NaN	NaN	2017-08-06T09:37:27Z	12509	212	9	
3	3	525378972	LiQWYD	NaN	NaN	NaN	NaN	2018-10-17T17:18:35Z	122746	16	7	
4	4	9855085	Declan DP	NaN	NaN	NaN	NaN	2011-12-16T18:35:32Z	59057	9	4	
...	...	...	...	...	...	...	...	...	...	...	...	
9995	9995	139675025	warez	worldwide	NaN	NaN	NaN	2015-02-21T07:38:54Z	8564	356	13	
9996	9996	77907316	Flakzz	NaN	NaN	Instagram :	@flakzzmusic	2014-01-30T18:20:35Z	14153	70	21	
9997	9997	15015	Thys	NaN	NaN	NaN	NaN	2008-09-26T03:36:51Z	19672	84	12	
9998	9998	85639485	Will Not Fear	Seoul, Korea	NaN	NaN	NaN	2014-03-21T13:02:05Z	3576	74	14	
9999	9999	26502362	nocolor	New York	US	Joseph	Workman	2012-10-19T22:46:21Z	1929	606	0	

10000 rows × 19 columns

“track.csv”: 619,248 records.

	Unnamed: 0	Unnamed: 0.1	id	title	duration	full_duration	playback_count	likes_count	reposts_count	comment_count	...	tag_list	visua
0	0	0	1150090951	One Wish 【Free Download】	186593	186593	830.0	49.0	31	1.0	...	Roa "Vlog Music" "Free to use" "Royalty Free M...	Na
1	1	1	1142178148	Daydream 【Free Download】	214779	214779	1901.0	87.0	51	1.0	...	Roa "Vlog Music" "Free Music" "Royalty Free Mu...	Na
2	2	2	1130272750	Cozy Fall 【Free Download】	186253	186253	1859.0	129.0	89	2.0	...	Roa "Free Music" "Free to use" "Royalty Free M...	Na
3	3	3	1126200826	No Regrets 【Free Download】	280633	280633	4233.0	380.0	323	204.0	...	Roa "Free to use" "Vlog Music" "Royalty Free M...	Na
4	4	4	1118068150	Endless Summer 【Free Download】	249704	249704	4353.0	356.0	311	207.0	...	Roa "Vlog Music" "Free to use" "Royalty free m...	Na
...	...	...	...	...	...	...	...	...	...	...	...	...	
619243	619243	167430	147282637	Nimbus (Original Mix) [click "buy" for free DL]	234783	234783	10290.0	258.0	55	16.0	...	Trap Chords Swag Swag Yolo Wow PLUR	Na
619244	619244	167431	141712915	8Er$ X Lowend - Bubble Wrap [Click "Buy" for f...	204112	204112	12615.0	322.0	73	20.0	...	NaN	Na
619245	619245	167432	139105364	Space Race X Lowend - Coma [Click "Buy" for fr...	195752	195752	3159.0	83.0	16	14.0	...	Race Lowend	Na
619246	619246	167433	130381511	Spider Grind (Original Mix)	162652	162652	2887.0	120.0	18	14.0	...	NaN	Na
619247	619247	167434	122974658	Eclipse (Original mix) [Free Download]	297352	297352	5373.0	168.0	28	24.0	...	NaN	Na

619248 rows × 29 columns

“playlist.csv”: 117,869 records.

	Unnamed: 0	id	title	genre	duration	likes_count	reposts_count	is_album	release_date	created_at	...	
0	0	1215929743	N3X - D R E A M S	Pop / Tropical House	1100486	11	0	True	2021-03-05T00:00:00Z	2021-02-23T13:45:56Z	...	https://www.toneden.io/fre
1	1	1211160604	Royalty Free Music	Royalty Free Music	9529286	6	0	False	NaN	2021-02-14T09:35:26Z	...	
2	2	1211139496	Copyright Free Music	Copyright Free Music	9529286	5	0	False	NaN	2021-02-14T08:19:07Z	...	
3	3	1211133652	Background Music For Videos	Background Music For Videos	9529286	6	1	False	NaN	2021-02-14T07:57:14Z	...	
4	4	1211125735	No Copyright Music	No Copyright Music	9529286	6	0	False	NaN	2021-02-14T07:29:15Z	...	
...	...	...	...	...	...	...	...	...	...	...	...	
117864	117864	594318714	00000 (with ACACY, Khundi Panda)	WNF	369402	9	0	True	2018-09-01T00:00:00Z	2018-09-01T14:44:03Z	...	
117865	117865	591039000	Brain Killer (with Minit, AIRAIR)	Electronic	182900	32	13	True	2018-08-27T00:00:00Z	2018-08-27T15:02:47Z	...	
117866	117866	554254233	Charged (with Sutt.)	NaN	228800	25	1	True	2018-07-04T00:00:00Z	2018-07-04T03:54:59Z	...	https://theartistu
117867	117867	504589122	Sympathy (with AIRAIR)	Electronic	234079	18	2	True	2018-04-23T00:00:00Z	2018-04-23T19:27:21Z	...	https://www.toneden.io/s
117868	117868	333988587	@weareredz	NaN	1343520	32	1	False	NaN	2017-06-29T01:09:03Z	...	

117869 rows × 23 columns

3. Đánh giá hoàn thành công việc

Họ Tên & MSSV	Công việc	Đánh giá hoàn thành (%)
Bùi Đăng Khoa - 19127645	- Code tất cả HTML cách 2 - Treo máy crawl dữ liệu theo cách 2 - Viết báo cáo cách 2 - Phân tích đánh giá 2 cách HTML	100%
Võ Bách Khôi - 19127037	- Code tất cả HTML cách 1 - Treo máy crawl dữ liệu theo cách 1 - Viết báo cáo cách 1 - Phân tích đánh giá 2 cách HTML	100%
Chung Thế Thọ - 19127562	- Code API lấy user ID - Code API lấy dữ liệu track - Treo máy crawl dữ liệu - Chỉnh sửa báo cáo phần API	100%
Dương Thị Xuân Diệu - 19127360	- Code API lấy dữ liệu user - Code API lấy dữ liệu playlist - Viết báo cáo phần API	100%

4. Nguồn tham khảo:

https://developers.soundcloud.com/

Scrape Soundcloud data using Selenium from scratch và Slide data science - Thầy Lê Ngọc Thành