TECH 1: SPOTIFY WRAPPED

Jacob Kaspar, Hande Huang, Hannah Sullivan,
Raksha Sridharan, and Jack Haggerty

THE QUESTION

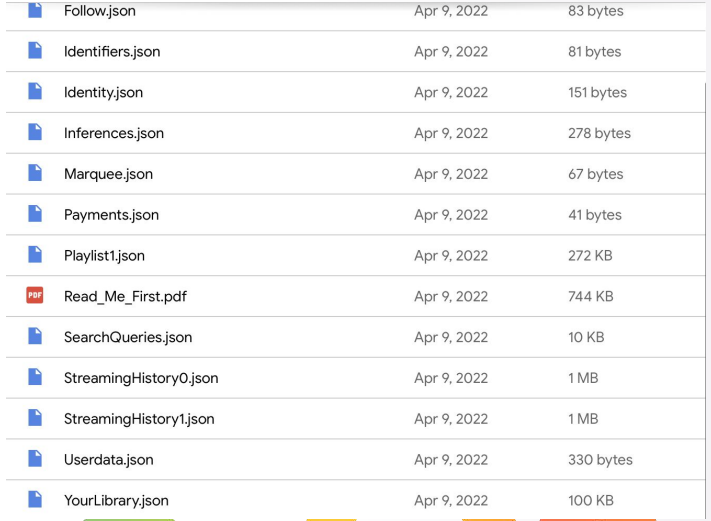
Can we use our CMSE 202 knowledge to model a recommendation feature using Spotify?

SPOTIPY

- Python library for the Spotify Web APi
- Provides full access to all music data from the Spotify platform (via Spotify Developer, as an independent developer with spotify data)
- Used to cross reference artists and genres for the recommendation feature

GATHERING DATA

- Using Spotify's "Download your data" tool, we requested our personal Spotify data to use.
- We were delivered a zip file containing .json files of playlists, the past year's streaming history, search results, and more.



Follow.json	Apr 9, 2022	83 bytes
Identifiers.json	Apr 9, 2022	81 bytes
Identity.json	Apr 9, 2022	151 bytes
Inferences.json	Apr 9, 2022	278 bytes
Marquee.json	Apr 9, 2022	67 bytes
Payments.json	Apr 9, 2022	41 bytes
Playlist1.json	Apr 9, 2022	272 KB
Read_Me_First.pdf	Apr 9, 2022	744 KB
SearchQueries.json	Apr 9, 2022	10 KB
StreamingHistory0.json	Apr 9, 2022	1 MB
StreamingHistory1.json	Apr 9, 2022	1 MB
Userdata.json	Apr 9, 2022	330 bytes
YourLibrary.json	Apr 9, 2022	100 KB

FORMATTING OUR DATA

- We created a function called get_100_favorite_artists() that took a streaming history file as the input.
- The output was a dataframe containing the 100 most listened to artists, their most listened to song, and milliseconds played

	endTime	artistName	trackName	msPlayed
2099	2021-05-21 13:48	Maribou State	Tongue	512575
6922	2021-07-26 20:49	Smino	Amphetamine	469830
961	2021-04-27 22:34	Cuco	Lover Is a Day	456000
2271	2021-05-25 04:51	Phantogram	Black Out Days	392805
1752	2021-05-17 01:02	Gorillaz	She's My Collar (feat. Kali Uchis)	388283
...	...	...	...	...
5212	2021-07-12 03:19	The Weeknd	The Morning	312426
4753	2021-07-05 21:00	Fetty Wap	Again	312240
5138	2021-07-11 19:39	Fetty Wap	Again	312240
5076	2021-07-10 23:48	Fetty Wap	Again	312240
5095	2021-07-11 00:21	Fetty Wap	Again	312240

100 rows × 4 columns

FINDING MOST LISTENED TO SONGS

- Next, we created a function: get_track_name_time_artist() that took the streaming history dataframe as the input
- The outputs were a dictionary with song names as the key and the artists + milliseconds played as the value, and a dataframe with song names as one column and artists + milliseconds as the other column

	track_names	time_artist
0	She's My Collar (feat. Kali Uchis)	[14392692, Gorillaz]
1	Pursuit Of Happiness (Nightmare)	[10297252, Kid Cudi]
2	Ms. Jackson	[8262151, Outkast]
3	Promiscuous	[7709360, Nelly Furtado]
4	YKWIM?	[7526871, Yot Club]
...	...	...
1759	Can't Do It Without You (Austin & Ally Main Ti...	[0, Austin Moon]
1760	Spice Girl	[0, Aminé]
1761	Potential Breakup Song	[0, Aly & AJ]
1762	Easier	[0, 5 Seconds of Summer]
1763	out for the night	[0, 21 Savage]

1764 rows × 2 columns

COMPARING SONGS

- We found a csv file containing over 600k songs from 1921-2020 that are on Spotify. We read this file in as a dataframe called “tracks”
- Using a mask, we created a dataframe of listened to songs from “tracks”. This new dataframe has 15 columns including danceability, energy, and loudness that correlate to number values

	name	popularity	artists	release_date	danceability	energy	key	loudness
56433	Ms. Jackson	82	['Outkast']	2000-10-31	0.843	0.806	4	-5.946
70153	YKWIM?	79	['Yot Club']	2019-05-31	0.599	0.537	9	-7.625
102515	Hotel Room Service	56	['Pitbull']	2009-10-23	0.851	0.634	1	-8.424
257487	Find Your Love	62	['Drake']	2010-01-01	0.627	0.614	6	-6.006
67227	New Person, Same Old Mistakes	74	['Tame Impala']	2015-07-17	0.470	0.809	8	-6.740
...	...	...	...	...	...	...	...	...
62320	Circus	74	['Britney Spears']	2008-12-02	0.791	0.733	6	-5.215
151086	Oops!...I Did It Again	54	['Britney Spears']	2004-11-08	0.755	0.832	11	-4.869
69176	Spice Girl	70	['Aminé']	2017-07-28	0.693	0.460	0	-7.912
412668	Potential Breakup Song	72	['Aly & AJ']	2020-12-29	0.554	0.899	6	-4.149
153497	Easier	74	['5 Seconds of Summer']	2020-03-27	0.562	0.460	5	-4.173

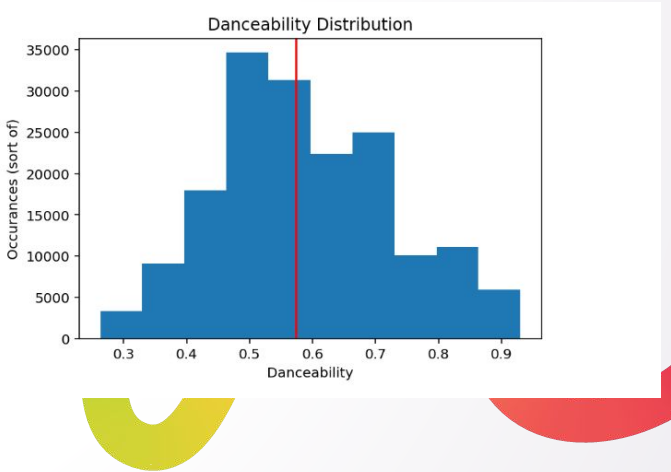
SONG FEATURES

- 9 of the 15 columns from the dataframe are song features like danceability, energy, and, loudness, as mentioned before
- We found the 50th percentile of each feature and put all those values into a list that we named Center
- We used Center to input into a K Nearest Neighbor Model that outputs a number of songs from “tracks” that have similar values for the 9 features and have not been listened to before

```python
#k nearest neighbors
def knn (Center,nineD,k=3):
    N = len(nineD)
    dist_list = []
    for i in range(N):
        dist = distance(Center,nineD[i])
        dist_list.append(dist)

    min_value = []
    min_index = []
    for j in range(k):
        min_val = sorted(dist_list)[j]
        min_value.append(min_val)

        min_ind = dist_list.index(min_val)
        min_index.append(min_ind)
    return min_value , min_index
```



RESULTS

Congratulations on 36646.21 minutes of listening!

This past year, your top 5 songs were:
She's My Collar (feat. Kali Uchis) by Gorillaz
Pursuit Of Happiness (Nightmare) by Kid Cudi
Voyager by Daft Punk
Aria Math by C418
Who's Been Sleeping In My Bed by Barry Manilow

This past year, your top 10 artists were:
Gorillaz who you listened to 692 times!
Daft Punk who you listened to 380 times!
Kanye West who you listened to 358 times!
The Neighbourhood who you listened to 331 times!
The Beatles who you listened to 277 times!
Pitbull who you listened to 238 times!
C418 who you listened to 211 times!
Greta Van Fleet who you listened to 207 times!
Arctic Monkeys who you listened to 194 times!
The Killers who you listened to 179 times!

This past year, your top genres were:
POP
ROCK
MODERN ROCK

Going into the next year, here's some music we think you'll love!

Song Title ['Band Name']

I Think I'm In Love ['Kat Dahlia']

Roscoe - Beyond the Wizard's Sleeve Remix ['Midlake', 'Beyond The Wizards Sleeve']

Back on My Feet Again ['The Babys']

Carry On (from the Original Motion Picture "POKÉMON Detective Pikachu") ['Kygo', 'Rita Ora']

Knocking At Your Back Door ['Deep Purple']

Alles auf Rot ['Capo']

Blue & Grey ['BTS']

Strange Days ['The Doors']

La déclaration d'amour - Remasterisé en 2004 ['France Gall']

Spirit in the Sky ['KEiiNO']

DIFFICULTIES/COMPLICATIONS

- There were issues with cross referencing our streaming history and tracks.csv, where not all our listened to songs were in tracks and so we lost some data.
- The process of going through the dataframes takes a long amount of time, depending on the size of the data inputted.
- Getting ML to work was quite hard, and in the end we didn't end up using ALL the data we collected, nor the initial methods we started with, as the knn proved difficult enough on it's own..
- Having issues pushing tracks.csv to the repository.


DEMO