

 main


 1 branch













 0 tags


Go to file


Add file

Code







	d0efcfa	1 minute ago	🕒 16 commits
	Visu...	Final Project R...	5 minutes ago
	.DS_...	Final Project R...	1 minute ago
	.gita...	Final Project U...	4 days ago
	.giti...	Final Project U...	4 days ago
	Cap...	Final Project R...	5 minutes ago
	GitH...	Final Project U...	4 days ago
	Jup...	Final Project R...	5 minutes ago
	Proj...	Final Project R...	5 minutes ago
	Proj...	Final Project R...	5 minutes ago
	Proj...	Final Project R...	5 minutes ago
	REA...	Final Project R...	5 minutes ago

 README.md



Spotify Recommendation

No description, website, or topics provided.

-  Readme
-  0 stars
-  1 watching
-  0 forks

Releases

No releases published
[Create a new release](#)

Packages

No packages published
[Publish your first package](#)

Languages

-  Jupyter Notebook 100.0%

System

Preface:

Link to Presentation:

<https://github.com/justingrisanti/dsc-capstone-project/blob/main/Project%20Presentation.pdf>

Link to Visualization:

<https://github.com/justingrisanti/dsc-capstone-project/tree/main/Visualizations>

Link to Data:

<https://www.dropbox.com/s/m8yqo6wucwzizzf/Data.zip?dl=0>

Link to Jupyter Notebook:

<https://github.com/justingrisanti/dsc-capstone-project/blob/main/Capstone%20Project-Recommendation%20System.ipynb>

Link to Blog Post:

https://justingrisanti.github.io/spotify_recommendation_system

Business Understanding

Context:

Spotify is an audio streaming and media services platform, created in 2006. It is one of the largest music streaming service providers with over 406 million monthly active users, including 180 million paying subscribers, as of December 2021. Spotify offers digital copyright restricted recorded music and podcasts, including more than 82 million songs, from record labels and media companies. As a freemium service, basic features are free with advertisements and limited control, while additional features, such as offline listening and commercial-free listening, are offered via paid subscriptions. Spotify is currently available in 180+ countries as of October 2021. Users can search for music based on artist, album, or genre, and can create, edit, and share playlists. Two of the most important aspects of Spotify that has led to its popularity are its music discovery functionalities, and playlist creation fostering a new social aspect to music listening. In a 2021 How-To Geek article called "6 Awesome Spotify Features You Should Be Using," 3 of the 6 features are related to playlists, and one speaks about music discovery. One

of these features related to music discovery is called "Enhance."

Enhance allows you to discover new tracks that might best fit one of your existing playlists. For example, if you have a playlist of a collection of 80s rock songs, Enhance might suggest that you add "Eye of the Tiger" by Survivor.

Business Problem:

Music is now one of the easiest forms of media to consume, due to smartphone technology allowing users to access millions of songs at their fingertips. However, with all of these choices, it is hard to know where to begin when it comes to finding new music that matches someone's preferences, and users may become overwhelmed when trying to find music. The solution to this problem is to create a recommendation system from scratch that can reperform the functionality of Enhance, which is to use a selection of songs from a playlist and use content-based filtering to suggest a list of songs that are similar. The stakeholders of this project are Spotify, music-listeners, DJs, and other music-related occupations. The main purpose of this recommendation system is inferential, meaning that this model should be able infer information about songs from a given playlist and then to predict songs that a user will likely add to that same playlist.

Step 2: Data Understanding

We are sourcing our data using the Spotify Million Playlist Dataset, with the following metrics

- Number of playlists: *1,000,000*
- Number of tracks: *66,346,428*
- Number of unique tracks:
2,262,292
- Number of unique albums:
734,684
- Number of unique artists:
295,860
- Number of unique titles: *92,944*

Our raw playlist data has the following features:

Playlist Attributes

- Playlist Name
- Playlist Type
- Number of Tracks
- Number of Unique Albums
- Number of Followers
- Number of Edits
- Duration in Milliseconds
- Number of Artists

Song Attributes

- Artist Name
- Track URI
- Artist URI
- Track Name
- Album URI
- Duration in Milliseconds

- Album Name

As we see in the summary above, there are 1 million playlists with over 66 million songs. Within these playlists, there are 2.2 million unique songs, 734k unique albums, and 300k unique artists. There is a lot of data to work with here. The playlists are sorted into categories, with country and chill being the top, followed by rap and workout. Drake, Kanye West, and Kendrick Lamar are the 3 top artists. The next step in this process is to convert all of this json data to be compatible with python.

Note: For our data, we will select the top 100 playlists based on number of followers from 200k playlists due to runtime.

Step 3: Data Preparation

For our Content-Based model, we will compare 2 vectors using cosine similarity. The first vector will be our unique track data, with audio features, genre data, and popularity for each track, and the second vector will be a playlist average vector, with average audio features, average genre data and average popularity by track.

Popularity and genre are self explanatory, but please see below for a guide to our audio features:

- **Danceability:** describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- **Acousticness:** A measure from 0.0 to 1.0 of whether the track is acoustic.
- **Energy:** a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy.
- **Instrumentalness:** Predicts whether a track contains no vocals. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content.
- **Liveness:** Detects the presence of an audience in the recording. Higher liveness values represent an increased probability

that the track was performed live.

- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track. Values typical range between -60 and 0 db.
- Speechiness: detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.
- Tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).

Step 4: Modeling

Our model is a content-based recommendation system using cosine similarity. I created a function that allows you to put in a `playlist_id` and number of recommendations desired, and then it provides that number of recommended tracks.

See below the three playlists selected for our reasonableness tests (102585, 765, 160101)

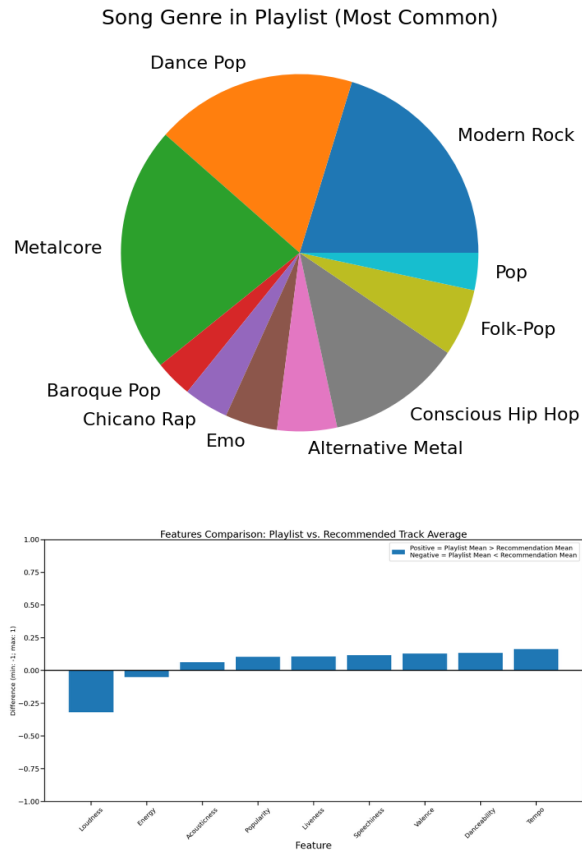
Reasonableness Test: 102585

Thank you for inputting your playlist. Please see playlist tracks below:

pos	track_uri	track_name	artist_name	album_name
0	67WtWafOMpgeV6ABrBQxcE	Some Nights	fun.	Some Nights
1	6Ep6BzICB9Lz3P4sWqIAAB	Radioactive	Imagine Dragons	Night Visions
2	4wCmqSrtYcGxEXROQE6vV	Somebody That I Used To Know	Gotye	Making Mirrors
3	59luo3tMmQlInEQOOjxh	Best Day Of My Life	American Authors	Oh, What A Life
4	6cpk005TXCqSeqN2Hule	One More Night	Maroon 5	Overexposed Track By Track
5	5edBgVVRD0nWk140S21T	Counting Stars	OneRepublic	Native
6	1jdNcAD8r58RsdQJJdx	Ho Hey	The Lumineers	The Lumineers
7	7zWj09xkFga9tcv6YhU6q	Drive By	Train	California 37
8	6g1NlCpW7fgDnWbCCDhI	Wake Me Up - Radio Edit	Avicii	True
9	5rgy6ghBq1eRApCkeUdX7	We Are Young (feat. Janelle Monáe) - feat. Jan...	fun.	Some Nights

Here are 5 tracks that might fit this playlist:

track_uri	track_name	artist_name	album_name
3s5ogvexUgA6Xqj37zpnP	Pliedriver waltz	Alex Turner	Submarine - Original Songs From The Film By Al...
57LTVY8oPSGy2a7c8SzJpD	What Goes Around...	Alesana	Punk Goes Pop, Vol. 2
0fB5cRA2ingpF06X1Ou4u	Bang Bang You're Dead	Dirty Pretty Things	Waterloo To Anywhere
5dHlBukd22hQH9zTcFv	Dirty Little Secret	The All-American Rejects	Move Along
0X0Lz7LwplWodGqWiaxD	Mess Around	Cage The Elephant	Tell Me I'm Pretty



As we can see above, the aggregate difference between our recommended songs and our playlist average is 14, which is driven mainly by the genre data. This metric is only useful when compared to other aggregate differences, so I will analyze when I review another playlist. While it can be argued that our genre data is so specific that it is dominating our model, I personally think the specificity of the genre helps us narrow our recommendations down to suggest more accurate songs.

Looking at our visualization, for the most part it appears that our features for our recommendations are comparable to the playlist, with Tempo driving the main difference.

Now for the sound test. I will document my subjective findings below: Some Nights, Radioactive, and Somebody That I Used to Know were all hits around 2013, and I personally have listened to these songs already. My description of Radioactive is an arena rock, bass thumping anthem that brings a high level of intensity. Some Nights and Somebody That I Used to Know are more chill and have a very nostalgic sound to them.

For our recommendations, What Goes Around..., Bang Bang You're Dead, Dirty Little Secret and Mess Around are all alternative/rock sounding tracks, whereas Piledriver Waltz is definitely more on the chill side.

Overall, based on our genre analysis and feature comparison, I would deem that these 5 recommended songs pass the reasonableness test.

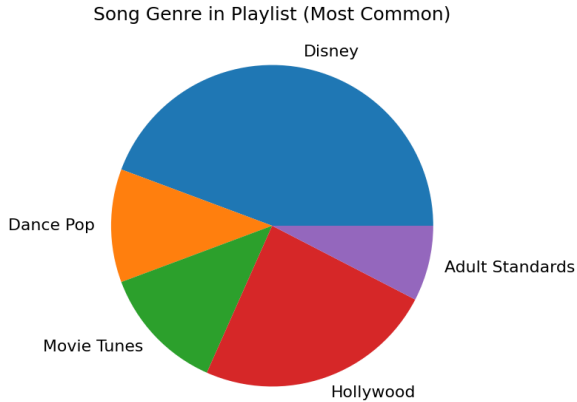
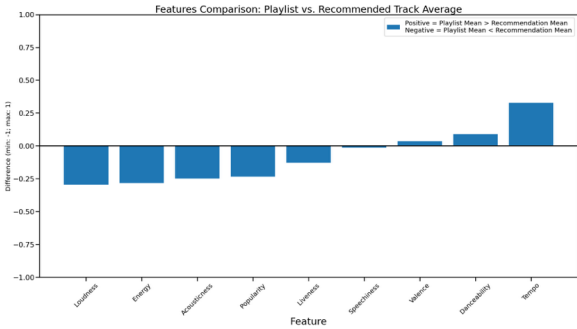
Reasonableness Test: 765

Thank you for inputting your playlist. Please see playlist tracks below:

	track_url	track_name	artist_name	album_name
pos				
0	03xWMKEbeO4SnyAS3pj	When Will My Life Begin - From "Tangled"/Sound...	Mandy Moore	Tangled
1	110SkxNCLWm2bY6t5SmK	Mother Knows Best - From "Tangled"/Soundtrack ...	Donna Murphy	Tangled
2	0T0Z0FRd8PQ6vTRQzNgQ	I've Got a Dream - From "Tangled"/Soundtrack V...	Mandy Moore	Tangled
3	6kspKx2uA9p2MFH4q6	I See the Light - From "Tangled"/Soundtrack Vb...	Mandy Moore	Tangled
4	75VfV82x1H8HnQ2PyQO57	Healing Incantation - From "Tangled"/Soundtrac...	Mandy Moore	Tangled
5	06oKEMa8OPemL8d9FMe6Jv	You'll Be In My Heart	Phil Collins	Tarzan Original Soundtrack
6	28UjMEhwYUy5uUWVOVHwJ	I'll Make a Man Out of You - From "Mulan"/Soun...	Donny Osmond	Mulan
7	2m7Z25pby1cVwPMAUqj	Go the Distance - From "Hercules"/Soundtrack	Roger Bart	Hercules
8	7jZCvMtGE7HF35W89zq2g	The Bells Of Notre Dame	Tony Jay	The Hunchback of Notre Dame Original Soundtrack
9	1OYOLWqmhKfKzKC3ek1a	Colors Of The Wind	Judy Kuhn	Pocahontas

Here are 10 tracks that might fit this playlist:

	track_url	track_name	artist_name	album_name
0	7wJMdSgSL9Vwvwb10Num	On My Way	Phil Collins	Brother Bear
1	23kLJ0JNcxNcXRDNF3d5c	You've Got A Friend In Me - From "Toy Story"/ ...	Randy Newman	Toy Story
2	0odIT9B88vOCxX50e4B5	Bette Davis Eyes	Kim Carnes	Mistaken Identity
3	2s6wCS3vGZFPY9NOTfPKJZ	Take Me Home - 2016 Remastered	Phil Collins	No Jacket Required (Remastered)
4	4LCyWjRtsZ5qJ2ck7w	Tornado	Lea Michele	Places
5	27m5YFYTi1K3eXMMWmZVZ	The Rendezvous (feat. Madi Diaz)	Rob Cantor	Not a Trampoline
6	0qxtQ28r3W1nd3Q2r9dH4	I Just Can't Wait to Be King - From "The Lion ...	Jason Weaver	The Lion King
7	7G061QqW7NKP1NDTpk04	Happy Working Song - From "Enchanted"/Soundtra...	Amy Adams	Enchanted
8	6mDsuQawhV5m1t0MTNUpv	Something There	Robby Benson	Beauty and the Beast
9	0QKHb0vKGZc9Ppyeq7HKG	The Chipmunk Song	Alvin & The Chipmunks	Alvin & The Chipmunks / OST



As we can see above, the aggregate difference between our recommended songs and our playlist average is 6, which is driven mainly by the genre data. Most of our songs here are Disney/Soundtrack, so there seems to be little variation here. This aggregate difference is better than our first playlist, although it seems our features may be offsetting some of that difference.

Looking at our visualization, for the most part it appears all of the features are comparable, with tempo, energy, popularity, loudness and danceability having the largest difference. Seeing that the general theme of this playlist seems to be soundtrack for childrens movies, I believe that genre is the most important driver of the recommender here, and it seems like our genre hit the mark.

Now for the sound test. I will document my subjective findings below: As we can see from our selection of songs, they all seem to be Disney songs from Moana and Tangled. I would expect that our recommendation system would suggest other kids songs accross childrens movies. After pulling more songs from the unwrapped track data, I can confirm that there are some songs that are not Disney.

For our recommendations, all 10 songs appear to be soundtrack songs, which is great, and they are all mainly from beloved kids movies. Because it seems like kids movies is the main category for this playlist, I will deem this recommendation adequate.

Overall, I would deem that these 10 recommended songs pass the reasonableness test.

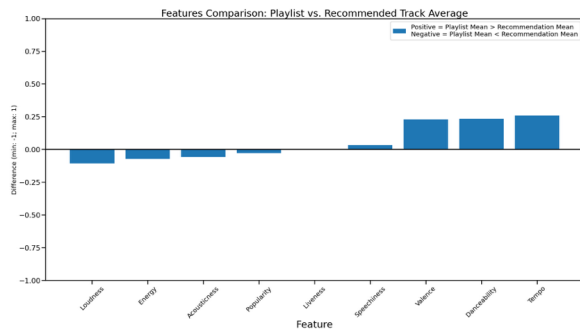
Reasonableness Test: 160101

Thank you for inputting your playlist. Please see playlist tracks below:

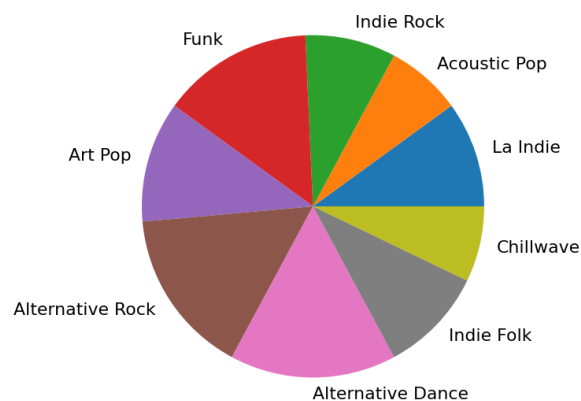
pos	track_url	track_name	artist_name	album_name
0	4w3dmOpGQ9otu7cG5uWy88	Not Just a Girl	She Wants Revenge	Valleyheart
1	37r6i0GTgR05rGe5wNhmP	When They Fight, They Fight	Generational	Con Law
2	0grF6kR3hoHLcgCYsF4	Howlin' For You	The Black Keys	Brothers
3	6M23RkYPdVR91c4iWVnKcl	Changing	The Airborne Toxic Event	All At Once
4	5nHRIKsXdwUpse9gZrAxLR	Oxford Comma	Vampire Weekend	Vampire Weekend
5	4u2KDCwcj8WGcFXDhy1aKL	The World (Is Going Up In Flames)	Charles Bradley	No Time For Dreaming
6	4xifZ2mYHm6MHJ7e83eyz	Grey Lynn Park	The Veils	Troubles of the Brain - EP
7	2eluumf6Sdf4CvGCFE18VO	You're A Wolf	Sea Wolf	Leaves In The River
8	04xYrikbIH39wHO7Sjae0	Greenback Boogie	Ima Robot	Greenback Boogie - Single
9	5FI11xRi01bvCjsZvKWro0	Houdini	Foster The People	Torches

Here are 5 tracks that might fit this playlist:

	track_url	track_name	artist_name	album_name
0	6mHw0o12JUfDodSxMwp8Tl	Caves	Haux	All We've Known
1	629Cjw0fUyZUMkBJnjtDR	Places	Shlohmo	Bad Vibes
2	2BvDTv6akKbIKLT6ol2NXp	Coast to Coast	Waxahatchee	Cerulean Salt
3	2ByM6ejjN2EERm2NmXlq5t	Emerge From Smoke	Shlohmo	Dark Red
4	2VvTTbJWcguP88dvnph7hl	Adult Diversion	Alvvays	Alvvays



Song Genre in Playlist (Most Common)



As we can see above, the aggregate difference between our recommendations and our playlist data is pretty large, at 44. I believe this is driven by our genre data, as it is very evenly spread out across multiple genres

According to our feature visualization, our differences are very minimal compared to the other playlists. The largest differences are Speechiness, Acousticness and Energy.

Now for the sound test. I will document my subjective findings below: From our playlist, I already recognize M83, the Black Keys and Vampire Weekend. Outro by M83 is very spacy and slow, whereas Howlin' For You is more uptempo and rock themed. Oxford Comma is also uptempo and kind of spaced out, but overall sounds chill. Not Just a Girl is slower paced, but also sounds like alt-rock overall. For our recommendations, I would expect chiller alt-rock that is slower in pace and has mostly fleshed out instrumentals.

For our recommendations, I recognize none of the songs. Caves, Coast to Coast and Diversion are very Indie sounding, somewhat Alternative. The two songs by Shlohmo are dance-like, and sound more electronic. They are also instrumentals. Outro by M83 fits a similar vibe to these songs.

This playlist seems like there is more variety between the genres, and as such, my recommendations also seem varied. However, due to the feature difference being relatively small, and due to the genres fitting the broad scope of the playlist, I deem these recommendations sufficient.

Step 5: Results

To summarize: using cosine similarity, I was able to successfully create a recommendation system using a selection of data from the Spotify Million Playlist Dataset. Overall, I have identified 3 potential drawbacks that might affect this model: 1. Suggesting songs from the same basket that the machine is learning from may create bias, 2. There are few metrics to assess our system, which makes it harder to determine the broader success of the model, 3. I analyzed the model using subjectivity. Tastes can be very different and subjectivity begets bias, as well. With these in mind, however, I believe that this recommendation system is adequate at suggesting similar songs based on our tests in the results section.

As shown in the results section, our music recommendations seem to be driven mainly by the genre.

Extracting Genre data from Spotipy provided us with very specific genre data that might've been able to isolate our list to a very select handful of songs. It then used the audio features and popularity vectors to differentiate the songs when it came to using cosine similarity.

I believe some good improvements to this project would be to source more general genre data if it is eventually made available as an export from Spotipy. As I had to manually export the data from Spotipy using different means, it is hard to tell how accurate the genre data actually is. I would also like to include information on the year released, in case a playlist is related to a certain year, for example, 2013s hits. I would also like user-review data that shows how much a user liked a song, so I could try a collaborative-based filtering approach. Another thing I would like to do is research other forms of content-based recommendation systems, perhaps some form of clustering. Lastly, if I had a faster computer I would include all of the playlists in my analysis to capture more songs for my recommendation system.