justingrisanti / dsc-capstone-project Public

Code Issues Pull requests Actions Projects Wiki Security

main 1 branch 0 tags Go to file Add file Code

About

No description, website, or topics provided.

Readme
0 stars
1 watching
0 forks

346a2d0 1 minute ago 10 commits

Visu...	Project Update...	1 hour ago
.DS_...	Project Update...	1 hour ago
.gita...	Final Project U...	1 minute ago
.giti...	Project Update...	1 hour ago
Cap...	Project Update...	1 hour ago
Jupy...	Project Update...	1 hour ago
Proj...	Project Update...	1 hour ago
Proj...	Project Update...	1 hour ago
Proj...	Project Update...	1 hour ago
REA...	Final Project U...	1 minute ago

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Jupyter Notebook 100.0%

README.md

Spotify Recommendation 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

Step 1: Business Understanding

The purpose of this section is to define the business problem and understand the stakeholders for the work that I am performing. Spotify is an audio streaming and media services platfrom, 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.

What I aim to perform is to create a recommendation system from scratch that can reperform the functionality of Enhance, which is to obtain a selection of songs 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 playlsit 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)

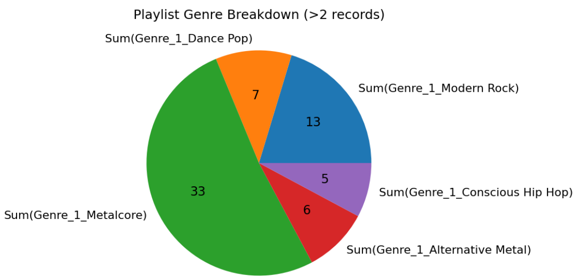
Reasonableness Test: 102585

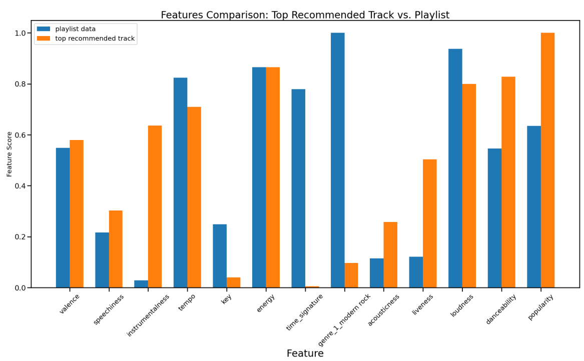
Thank you for inputting your playlist. Please see playlist tracks below:

pos	track_uri	track_name	artist_name	album_name
0	67WTwafOMgegV6ABnBQxcE	Some Nights	fun.	Some Nights
1	6Ep6BzIOB9tz3P4sWqiiAB	Radioactive	Imagine Dragons	Night Visions
2	4wCmqSrbyCgxEXROQE6vtV	Somebody That I Used To Know	Gotye	Making Mirrors
3	5j9iuo3tMmQIfnEEQOOjxh	Best Day Of My Life	American Authors	Oh, What A Life
4	6cpk00i5TxCqSeqNi2Hule	One More Night	Maroon 5	Overexposed Track By Track
5	5edBgVtRD0fvWk140SI21T	Counting Stars	OneRepublic	Native
6	1jdNcAD8Ir58RlsdGjJJdx	Ho Hey	The Lumineers	The Lumineers
7	7zWj09xkFgA9tcV6YhfU6q	Drive By	Train	California 37
8	6g1NlCpW7fgqDnWbCCDrHl	Wake Me Up - Radio Edit	Avicii	True
9	5rgy6ghBq1eRApCkeUdJXf	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
0	5zT5cMnMKoyruPj13TQXGx	I Found	Amber Run	5AM (Deluxe)
1	4GITtbZtRCQXhWLMXrWXHt	Roots	Imagine Dragons	Roots
2	673WCjn0SxKJD4qRKczaCk	She Had The World	Panic! At The Disco	Pretty. Odd.
3	0z8yrlXSjnl29Rv30RssNl	Shots - Broiler Remix	Imagine Dragons	Smoke + Mirrors
4	2AObzKd3JYlWQqQ067Z0Yl	Release	Imagine Dragons	Smoke + Mirrors





As we can see above, the sample variance between our recommended song and our playlist average seems to be very low, driven mainly by the genre data. 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 most of the features are comparable, except for time signature, acousticness, liveness, and key. As a musician, I would argue that key and time signature does not matter too much in terms my enjoyment for music, so at first glance, I would not say that our recommendation is that far off.

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 nostaglic sound to them.

For our recommendations, we have 3 songs by Imagine Dragons, 1 song by Panic at the Disco and our top song was Amber Run. I already know all of the songs except for the one made by Amber Run, and I would say that it is reasonable that these songs would be suggested, seeing the selection of songs on the playlist. I had to listen to the Amber Run song, and it definitely is a slower song that is lesser known to me than the others. It is definitely more acoustic and I wouldn't necessarily classify it as modern rock.

Overall, I would deem that these 5 recommended songs pass the reasonableness test.

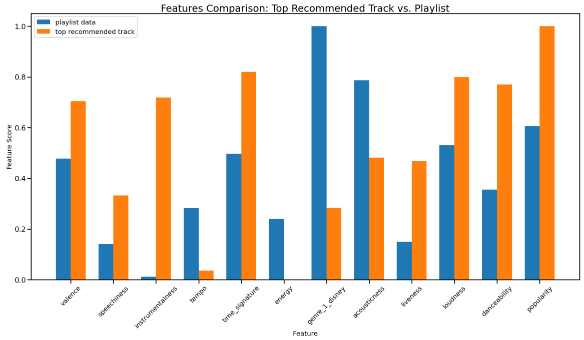
Reasonableness Test: 765

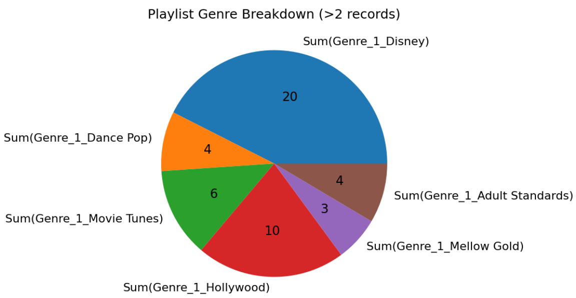
Thank you for inputting your playlist. Please see playlist tracks below:

pos	track_uri	track_name	artist_name	album_name
0	03xWMkKEbeO4SnylA53ipj	When Will My Life Begin - From "Tangled"/Sound...	Mandy Moore	Tangled
1	1lOSxJNCLvWm2bYaTcTSmK	Mother Knows Best - From "Tangled"/Soundtrack ...	Donna Murphy	Tangled
2	0TCt7OFRdD8PQ6vTRQxNgQ	I've Got a Dream - From "Tangled"/Soundtrack V...	Mandy Moore	Tangled
3	6klpXs2uAjagnZMFkt4qkl	I See the Light - From "Tangled"/Soundtrack Ve...	Mandy Moore	Tangled
4	75VVIB2x1h6BfxD2PqOO57	Healing Incantation - From "Tangled"/Soundtrac...	Mandy Moore	Tangled
5	0loKEMa9DP6mLibdFMe6Uv	You'll Be In My Heart	Phil Collins	Tarzan Original Soundtrack
6	28UMEtwyUUy5u0UWOVHwil	I'll Make a Man Out of You - From "Mulan"/Soun...	Donny Osmond	Mulan
7	2rm72l5pjfby1cWvPMUojl	Go the Distance - From "Hercules"/Soundtrack	Roger Bart	Hercules
8	7jZcWttgE7HF35YeB9dzq8	The Bells Of Notre Dame	Tony Jay	The Hunchback of Notre Dame Original Soundtrack
9	1OYOLWqKmhkFlx2KC9ek1a	Colors Of The Wind	Judy Kuhn	Pocahontas

Here are 10 tracks that might fit this playlist:

	track_uri	track_name	artist_name	album_name
0	2stkLJ0JNcXkIRDNF3ld6c	You've Got A Friend In Me - From "Toy Story"/ ...	Randy Newman	Toy Story
1	7G061Oqw7NXFr1NDTpXol4	Happy Working Song - From "Enchanted"/Soundtra...	Amy Adams	Enchanted
2	7wjiMdiSgsL9Vkrwb10Num	On My Way	Phil Collins	Brother Bear
3	6mDxu0xwhv5tn1oMTNUypu	Something There	Robby Benson	Beauty and the Beast
4	6ZgigeSB0XUMqc0jjzaq6d	You're A Mean One, Mr. Grinch	Thurl Ravenscroft	How The Grinch Stole Christmas
5	2dlxN435ZY9ruxGYND2Hq0	Almost There	Anika Noni Rose	The Princess and the Frog
6	2s6wCS3vDZFPY9NOTlPXJZ	Take Me Home - 2016 Remastered	Phil Collins	No Jacket Required (Remastered)
7	0jkGkwy510cvhy0jYPFme4	Kingdom Dance - From "Tangled"/Score	Alan Menken	Tangled
8	4qKDjmz094Bu2wMepNuwVN	Main Title / Once Upon A Dream / Prologue - Fr...	Chorus - Sleeping Beauty	Sleeping Beauty
9	0QKHM0vKGZcgRpryeqtYkG	The Chipmunk Song	Alvin & The Chipmunks	Alvin & The Chipmunks / OST





As we can see above, the sample variance between our recommended song and our playlist average is very low, driven mainly by the genre data.

Looking at our visualization, for the most part it appears most of the features are comparable, except for the instrumentalness and energy. Seeing that the general theme of this playlist seems to be soundtrack for childrens movies, again, I believe that genre is the most important driver of the recommender here. Our most common genre, disney, seems a little off, but that can be explained because not all childrens movies are necessarily Disney movies.

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 movie is the main category for this playlist, I will not review the other audio features.

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

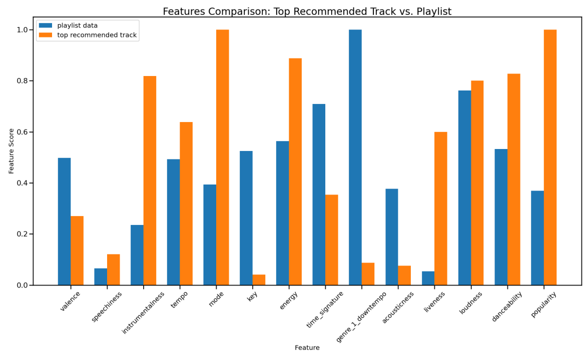
Reasonableness Test: 160101

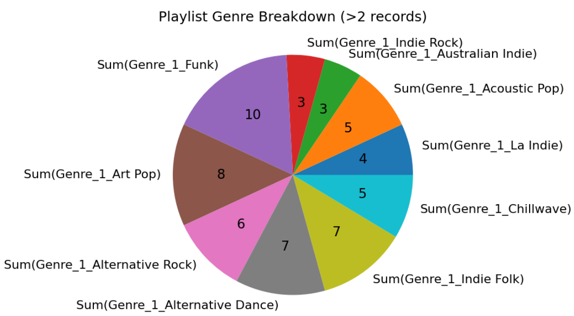
Thank you for inputting your playlist. Please see playlist tracks below:

pos	track_uri	track_name	artist_name	album_name
0	4w3dm0pGQ9otu7cG5uWy88	Not Just a Girl	She Wants Revenge	Valleyheart
1	37r6i0GTqgR05rGe5wNhmp	When They Fight, They Fight	Generationals	Con Law
2	0grFc6klR3hxoHLcgCYsF4	Howlin' For You	The Black Keys	Brothers
3	6M23RkYPbVR91c4iWVNkcl	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	4xifZ2mYHM6iMHJ7e83eyz	Grey Lynn Park	The Veils	Troubles of the Brain - EP
7	2eiuumi6SdF4CvGCfEt8V0	You're A Wolf	Sea Wolf	Leaves In The River
8	04xYrikbiBH39wHO7Sjae0	Greenback Boogie	Ima Robot	Greenback Boogie - Single
9	5Fli1xRi01bvCjsZvKWro0	Houdini	Foster The People	Torches

Here are 5 tracks that might fit this playlist:

	track_uri	track_name	artist_name	album_name
0	54KFQB6N4pn926IUUYZGzK	To Build A Home	The Cinematic Orchestra	Ma Fleur
1	13PUJCvdTSCT1dn70tlGdm	Welcome Home, Son	Radical Face	Ghost
2	2gZNwCpx5Twi1fQO94AY3G	The Lakes	James Vincent McMorrow	Post Tropical
3	5dulWCZyRqio1YhzwCc4P4	King Of Spain	The Tallest Man On Earth	The Wild Hunt
4	3iTi975Q6qnoRKrBL1FNsl	Gold	Matt Hartke	Gold





As we can see above, the sample variance between our recommended song and our playlist average is very low, driven mainly by the genre data.

According to our visualization, our top song features appear to be a little more variable, with most feautres having a large distance. I think this once again proves that our specific genre data is the driver for our model.

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. To Build a Home is definitely a very slow song, and I would argue it sounds more like pop than alt-rock, but if our main genre for our playlist is "Downtempo," I would say it fits. Welcome Home, Son by Radical Face sounds like chill folk-rock, comparable to the Lumineers. Based on the type of playlist, I think this is a decent recommendation. King of Spain also sounds like folk, similar to Welcome Home, Son. The Lakes and Gold give me a similar feeling to To Build a Home, they are very slow and sound more like chill pop.

This playlist seems like there is more variety between the genres, and as such, my recommendations also seem varied. The thing in common does appear to be the style of songs being acoustic and instrumental and downtempo overall. I would say that these recommendations are adequate.

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 beleive 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 usng 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.