

UNIVERSIDADE FEDERAL FLUMINENSE

VITOR SANTOS DE ARAUJO

**Entendimento do algoritmo de criação de playlists
do spotify**

Niterói

2019

Sumário

1	Introdução	2
2	Dataset	3
3	Desiderata	5
3.1	What?	5
3.2	Why?	5
3.3	How?	5
4	Sistema	6
5	Casos de Uso	10
6	Conclusão	11
	Referências	12

Capítulo 1

Introdução

A idéia do trabalho consiste em utilizar alguma das ferramentas de visualização de dados em uma parcela de informações de playlists do spotify visando o entendimento do seu algoritmo de criação.

O dataset utilizado é apenas uma pequena parte da imensa quantidade de informação que o spotify possui, ele consiste do conteúdo de algumas playlist com informações detalhadas das músicas.

Com base no dataset utilização, foi desenvolvido uma aplicação web para a visualização dos dados [3], ela foi criada com o intuito de analisar esses dados da melhor forma possível.

Capítulo 2

O Dataset

O dataset [5] contém informação sobre 18835 músicas, sendo 14615 músicas diferentes, distribuídas em 300 playlists diferentes. Antes da análise dos dados foi realizado um pré-processamento dos dados onde foi realizada a sua contagem preparando eles para a análise e a renomeação de algumas colunas.

Os dados foram divididos entre dois arquivos, o primeiro continha as informações de cada música e o segundo a associação entre as músicas e as playlist [6]. As informações das músicas são divididas entre 15 atributos, sendo um deles o nome da música.

- `song_name`: Nome da música.
- `song_popularity`: A popularidade da música. É um valor numérico de 0-100.
- `song_duration_ms`: Duração da música em milissegundos.
- `acousticness`: É atributo usado pelo spotify com a ideia de prever se é uma música acústica. É um valor numérico de 0-1.
- `danceability`: É atributo usado pelo spotify que caracteriza se é uma música boa para dançar. É um valor numérico de 0-1.
- `energy`: É um atributo usado para medir a 'intensidade' da música. É um valor numérico de 0-1.
- `instrumentalness`: É atributo usado pelo spotify com a ideia de prever se é uma música instrumental. É um valor numérico de 0-1.
- `key`: É atributo usado pelo spotify para identificar a tonalidade da música. É um valor numérico discretizado.

- **liveness**: É atributo usado pelo spotify para identificar se a música foi gravada ao vivo. É um valor numérico de 0-1.
- **loudness**: É atributo usado pelo spotify para identificar o ganho de som da música em decibéis. É um valor numérico de -60 à 0.
- **audio_mode**: É um atributo usado para identificar a 'escala' da música. É um valor numérico binário, 0 e 1, onde 1 é maior e 0 menor.
- **speechiness**: É atributo usado pelo spotify para dizer o quão é 'falada' a música. É um valor numérico de 0-1.
- **tempo**: É um atributo usado para identificar o número de batidas por minuto da música, BPM. É um valor numérico positivo.
- **time_signature**: É um atributo usado para identificar o 'compasso' da música. É um valor numérico positivo.
- **audio_valence**: É atributo usado pelo spotify para avaliar a emoção transmitida pela música. É um valor numérico de 0-1.

O segundo dataset possui informações adicionais da música e a ligação da música com uma playlist.

- **song_name**: Nome da música.
- **artist_name**: O artista que perfoma essa música.
- **album_names(album_name)**: Nome do album em que essa música foi lançada.
- **playlist(playlist_name)**: Uma playlist que essa música esta associada.

Capítulo 3

Desiderata

3.1 What?

- Um amostral de músicas do spotify;
- Detalhes específico de cada música;
- Possível ligação entre todos os dado.

3.2 Why?

- Buscar entender o algoritmo do spotify, que define por exemplo uma sequência de músicas automaticas;
- Como é definido o gênero das músicas através desses atributos;
- Qual a influência desses atributos na popularidade da música.

3.3 How?

- Principalmente através de gráficos de barra e de linha;
- Buscando sempre visualização bidimensional.

Capítulo 4

Sistema criado

Após a análise dos dados, foi desenvolvido um sistema web [2] que gera uma visualização dos dados através de uma matriz de 'heatmap', utilizando os 7 atributos(audio_valence, speechiness, liveness, instrumentalness, energy, danceability, acousticness) que variam de zero até um discretizados.

Devido a grande quantidade de dados, teve a necessidade de se utilizar de muitos filtros no sistema. O primeiro 4.1 foi selecionar um conjunto com as 10 playlists mais interessantes, que é principal filtro, sendo utilizado na página principal. O segundo filtro 4.2, utilizado quando está selecionado para mostrar todos os dados, esse filtro remove os dados que são considerados interessantes, mostrando os dados restantes. O terceiro filtro 4.3, consiste em selecionar uma única playlist para estudar melhor as suas características.

O uso principal do heatmap é buscar clusters de dados em relação a um atributo, para o estudo, o atributo mais utilizado foi a playlist, porém dados interessantes como: Quais atributos são determinantes para uma música ser popular?

O sistema possui uma estrutura bem simplificada, sendo criado utilizando as ferramentas Vega-Light [1], utilizada para a construção dos gráficos, e jQuery [4], utilizado para facilitar em algumas manipulações da página. Ele possui apenas uma página na qual é possível através dos 'Radio Buttons' selecionar qual o filtro deseja ser utilizado.

Selecionando 'Interessante', será utilizado o filtro de interessantes 4.1. Selecionando 'Todos', será utilizado o filtro que irá remover as playlists interessantes 4.2, porém é possível mostrar todos os dados desmarcando a opção 'filtrar'. Selecionando 'Único', o sistema irá filtrar pela playlist que for definida na caixa de texto 4.3.



Figura 4.1: Playlist Interessantes.

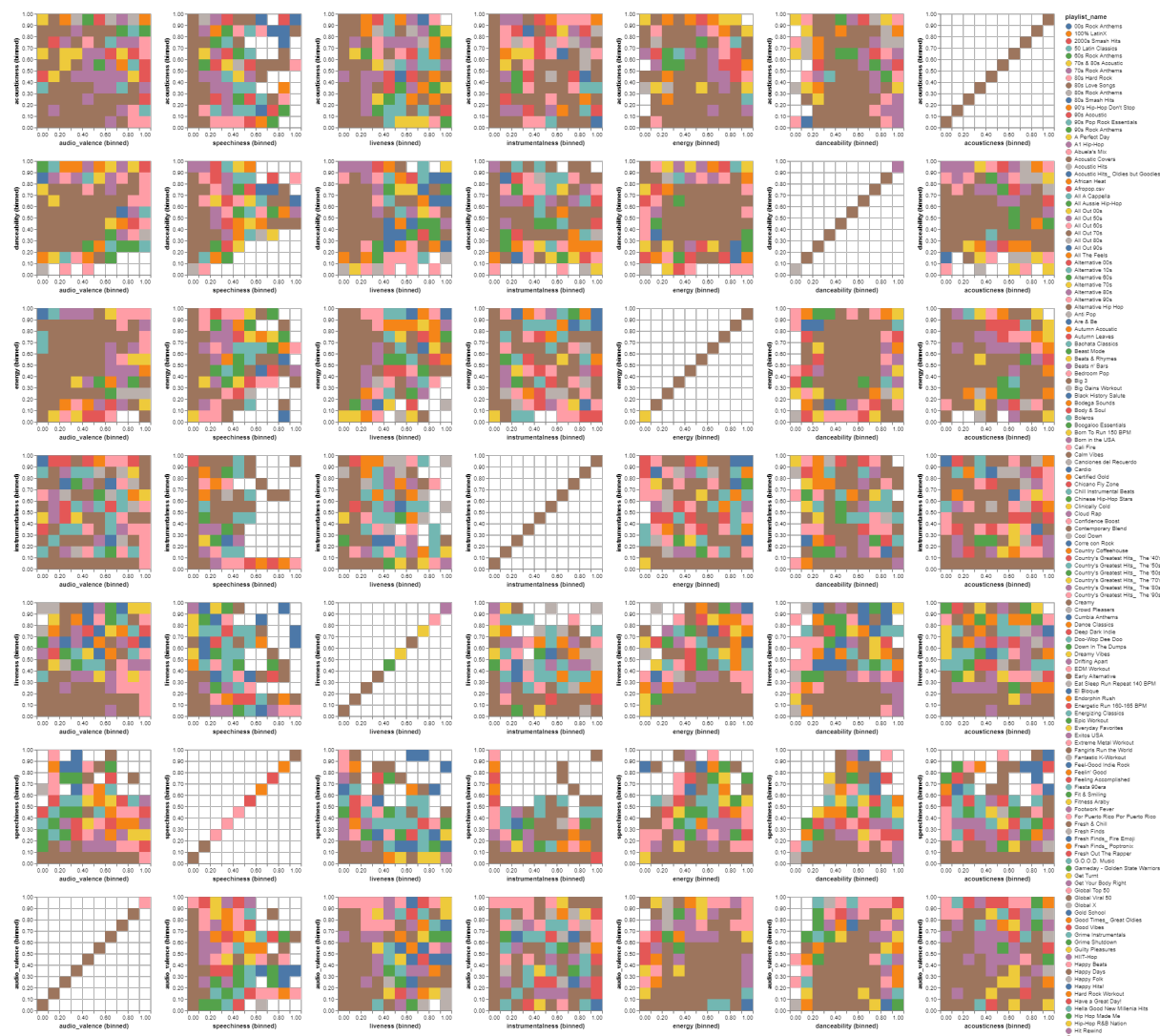


Figura 4.2: Todos os dados, filtrando as interessantes.

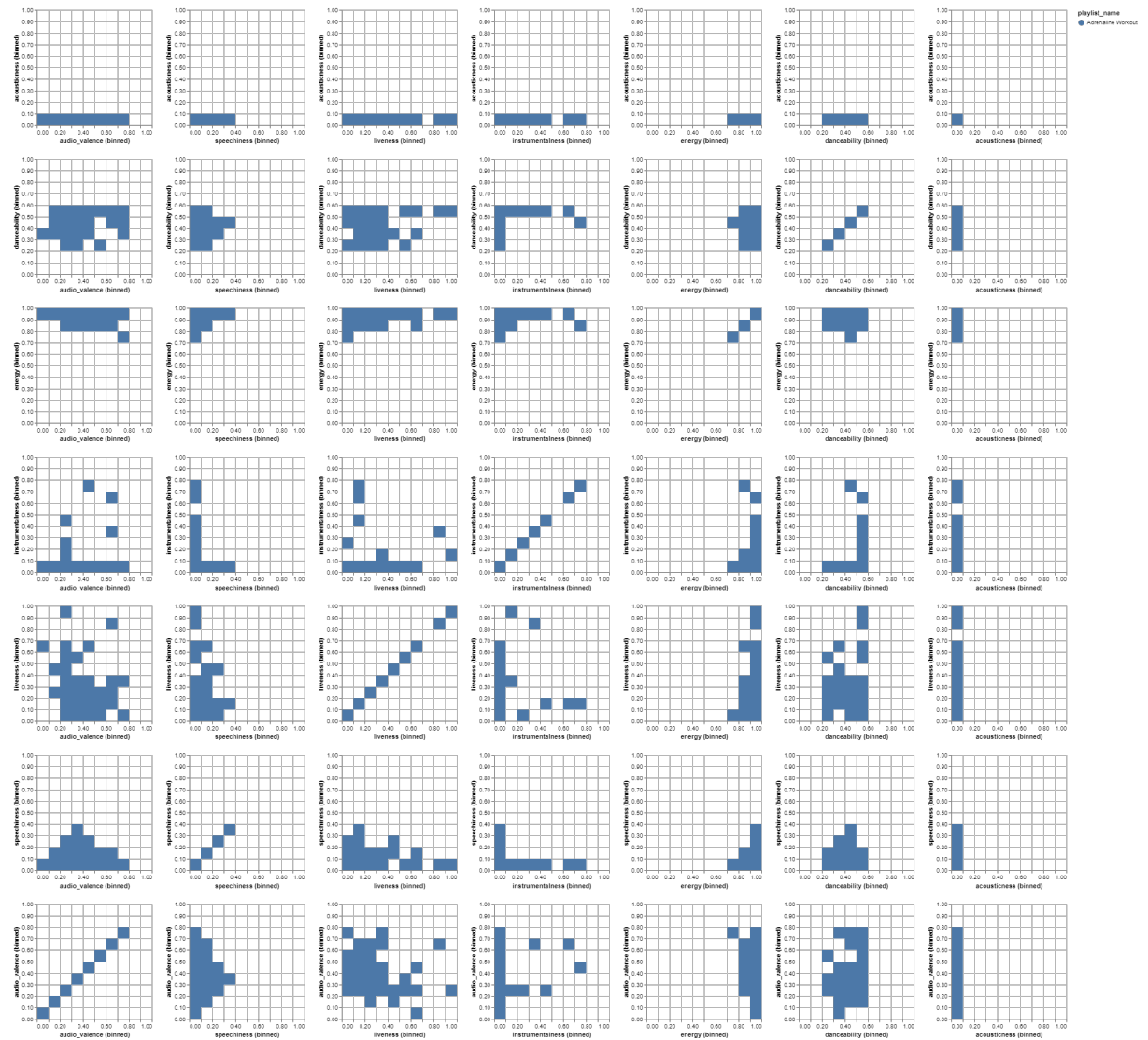


Figura 4.3: Playlist única, 'Adrenaline Workout'.

Capítulo 5

Casos de Uso

Com o sistema foi possível ter um entendimento das playlists do spotify, uma das conclusões encontradas é que uma playlist é montada em cima de uma ideia principal. Por exemplo, para se montar a playlist 'Adrenaline Workout', ele foi focada em músicas muito energéticas e com uma popularidade neutra. Baseado nisso, pelo sistema foi possível notar que ela não possui músicas acústicas pois elas são no geral pouco energéticas.

Com ele é possível fazer uma análise sobre os dados, como por exemplo, avaliar a popularidades das músicas, observar a distribuição dos dados em relação aos atributos, buscar correlações entre os atributos, além de validar ouvindo a playlist tais conclusões.

Capítulo 6

Conclusão

Através do uso do sistema [2] foi possível entender o funcionamento do processo de criações das playlists do spotify, além de adquirir melhor entendimento sobre o algoritmo do spotify.

As playlists são criadas com o intuito de que suas músicas seja similares, ou seja, quando analisada uma playlist sozinha, por exemplo, a playlist 'Your Favorite Coffeehouse' é possível pela análise dos dados concluir que essa playlist é prioritariamente de músicas acústicas e com uma certa capacidade de dançar, e tal conclusão pode ser confirmada após ouvir a mesma playlist no spotify.

Outro quesito a se levantar é responder perguntas, como por exemplo, se os atributos afetam a popularidade de uma música, pela análise usando o sistema, não é possível adquirir uma resposta muito forte para essa pergunta, visto que não se formam nenhum cluster de dados, porém olhando pontualmente é possível perceber que músicas dancantes tem a tendência de serem mais populares.

O sistema desenvolvido tem algumas limitações como por exemplo as 'tooltips' dos gráficos não possuem descrições detalhadas do quadrado, ou seja, não se sabe o número de dados da playlist que estão naquele quadrado. Uma grande melhoria para o sistema seria agrupar os dados por playlist de uma maneira mais inteligente, deixando de usar o agrupamento nativo da biblioteca do Vega [1].

Referências

- [1] *Vega-Lite – A Grammar of Interactive Graphics*. <https://vega.github.io/vega-lite/>.
- [2] DE ARAUJO, V. S. *VisUFF*. <https://vitora29.github.io/visuff-songs/>, 2019.
- [3] DE ARAUJO, V. S. *VisUFF - Github*. <https://github.com/VitorA29/visuff-songs>, 2019.
- [4] JQUERY FOUNDATION, T. *jQuery JavaScript Library*. <https://jquery.com/>.
- [5] RAMIREZ, E. *19,000 Spotify Songs*. <https://www.kaggle.com/edalrami/19000-spotify-songs>, 2018.
- [6] TAMER, N. *Top Spotify Tracks of 2017*. <https://www.kaggle.com/nadintamer/top-tracks-of-2017>, 2017.