Trabalho Prático 1 - Computação em Nuvem

Aluno: Jean George Alves Evangelista

Matrícula: 2024661178

Introdução

Este documento apresenta um relatório do Trabalho Prático 1 da disciplina de Computação em Nuvem, cursada em 2024/02. Conforme sugerido, o desenvolvimento foi realizada utilizando Pyspark e o Jupyter Notebook instalado na máquina virtual do trabalho.

Após a leitura da introdução e descrição do conjunto de dados, o desenvolvimento ocorreu de maneira sequencial por tarefa. A abordagem empregada em cada tarefa está descrita na seção de desenvolvimento, onde também é feita uma análise e discussão acerca dos resultados obtidos.

Desenvolvimento

Após realizar a conexão via ssh na máquina virtual e realizar o tunelamento de portas, foi possível acessar o ambiente de desenvolvimento do Jupyter Notebook e iniciar o desenvolvimento. A primeira etapa do desenvolvimento envolveu:

1. Importação de bibliotecas necessárias
2. Obter conexão com o Spark
3. Carregar os dataframes de playlists e músicas

Em seguida, foi possível iniciar o desenvolvimento de cada uma das três tarefas definidas na descrição do trabalho.

A abordagem para cada tarefa foi a mesma: leitura inicial da descrição da tarefa; pesquisa e consulta a fontes para fortalecer o entendimento, quando necessário; escrita de código com Pyspark para implementação da tarefa, consultando a documentação oficial do Pyspark quando necessário.

Tarefa 1

Desenvolvimento

A tarefa 1 envolveu a identificação de outliers no conjunto de dados das músicas. Utilizando as funções nativas da biblioteca do Spark foi possível identificar a duração mínima, média e máximo das músicas no conjunto de dados:

Minimum	Average	Maximum
0	234408.54976216817	10435467

Tabela 1 - Duração mínima, média e máxima, em milissegundos, das músicas no conjunto de dados

A tabela 1 indica a presença de outliers no conjunto de dados, uma vez que existe pelo menos uma música com duração de 0 milissegundos e pelo menos uma música com 10435467 milissegundos (aproximadamente 174 minutos), enquanto a média é de duração é de cerca 234409 milissegundos (aproximadamente 3,9 minutos).

É interessante identificar e remover os outliers da base de dados, e isso foi feito calculando o primeiro quartil (Q1) e o terceiro quartil (Q3). Também foi calculado o intervalo intervalo interquartil (IRQ):

Q1	Q3	IRQ
198026.0	258133.0	60107.0

Tabela 2 - Primeiro quartil (Q1), terceiro quartil (Q3), intervalo interquartil (IRQ) da duração das músicas

Utilizando esses valores e aplicando a fómula definida na metodologia IQR, foi possível filtrar as músicas e remover os outliers. No total, 570909 músicas foram consideradas outliers. A tabela 3 abaixo mostra a duração mínima, média e máxima das músicas, após remoção dos outilers:

Minimum	Average	Maximum
107866	226795.8593433425	348293

Tabela 3 - Duração mínima, média e máxima, em milissegundos, das músicas no conjunto de dados após remoção de outliers de acordo com o método IQR

Discussão

Identificação de outliers é algo extremamente importante na análise de dados, tendo em vista que outliers podem representar valores incorretos ou irrelevantes. Utilizando o método IQR, foram identificados 570909 músicas cuja a duração pode ser considerada outlier, isto é, um valor muito pequeno ou muito grande. Esse valor representa cerca 5.3% do total de músicas do conjunto de dados, um número que pode ser considerável a depender da análise de dados que está sendo realizada.

Tarefa 2

Desenvolvimento

A tarefa 2 envolveu encontrar os 5 artistas mais populares no conjunto de dados. Por “artistas mais populares”, entende-se os artistas que estão presentes em um maior número de playlists. Para esses 5 artistas mais populares, foi necessário desenhar um gráfico demonstrando como a popularidade de cada artista se deu ao longo dos anos. Em outras palavras, o gráfico deveria mostrar a evolução do número de playlists, por artista mais popular, ao longo dos anos.

Para identificar os 5 artistas mais populares, primeiramente foi necessário realizar uma junção entre as músicas e as playlists, para garantir que a análise seria feita em músicas que necessariamente possuem uma playlist válida. Em seguida, foi necessário agrupar por artista e contar as ocorrências de playlists únicas. Após ordenar pelo número de playlists em ordem decrescente, a tabela abaixo mostra os 5 artistas mais populares encontrados:

artist_uri	artist_name	playlist_count
spotify:artist:3T...	Drake	32258
spotify:artist:5p...	Rihanna	23963
spotify:artist:5K...	Kanye West	22464
spotify:artist:1X...	The Weeknd	20046
spotify:artist:2Y...	Kendrick Lamar	19159

Tabela 4 - 5 artistas mais populares, por número de playlists

A partir dos dados acima e por meio das funções de agregação, contagem e cálculo de ano do Spark foi possível obter a tabela abaixo:

artist_name	year	playlist_number
Kanye West	2011	4
Rihanna	2011	1
Drake	2012	24
Kanye West	2012	25
Kendrick Lamar	2012	5
Rihanna	2012	31
The Weeknd	2012	2
Drake	2013	523
Kanye West	2013	463
Kendrick Lamar	2013	279
Rihanna	2013	506
The Weeknd	2013	99
Drake	2014	880
Kanye West	2014	820
Kendrick Lamar	2014	506
Rihanna	2014	795
The Weeknd	2014	212
Drake	2015	2517
Kanye West	2015	1880
Kendrick Lamar	2015	1304
Rihanna	2015	2049
The Weeknd	2015	2248
Drake	2016	7844
Kanye West	2016	5121
Kendrick Lamar	2016	2695
Rihanna	2016	6005
The Weeknd	2016	4429
Drake	2017	20470
Kanye West	2017	14151
Kendrick Lamar	2017	14370
Rihanna	2017	14576
The Weeknd	2017	13056

Tabela 5 - Número de playlists para cada artista mais popular, por ano

Por fim, com foi possível desenhar o gráfico abaixo:

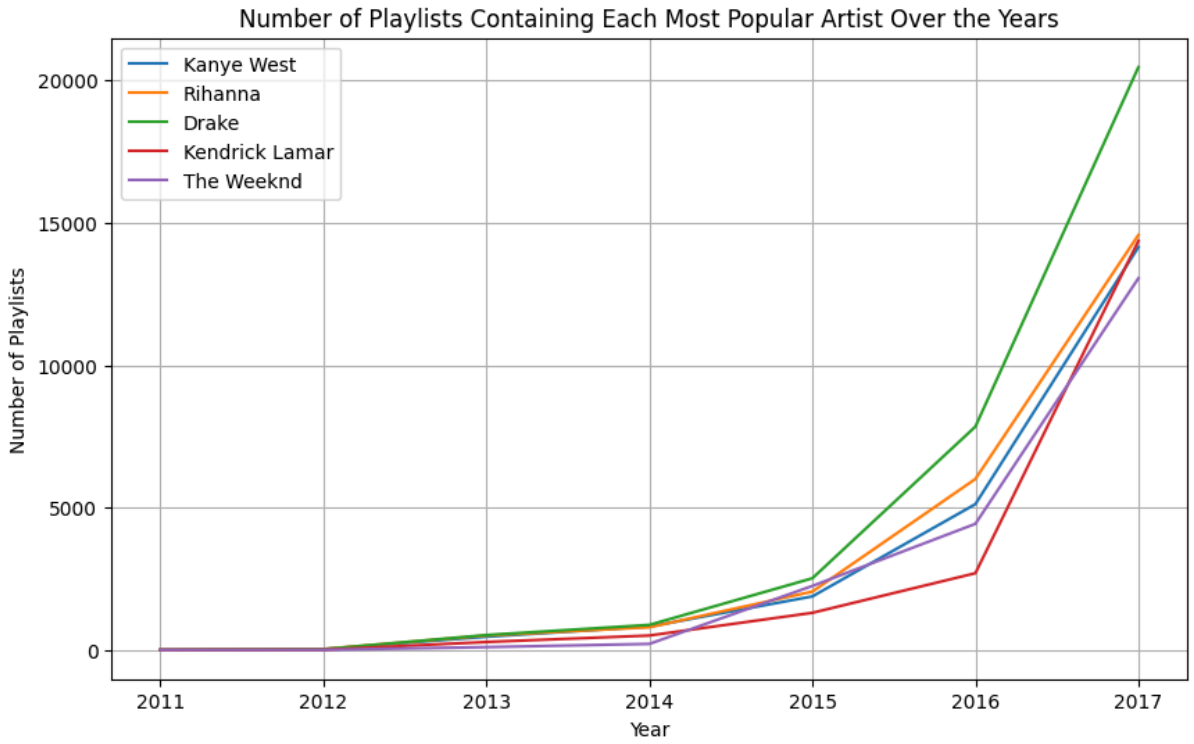

Gráfico 1 - Número de playlists contendo cada artista mais popular ao longo do tempo

Cabe citar que foi necessário realizar algumas manipulações nos dados para conseguir desenhar o gráfico acima, o que está explicitado no código.

Discussão

Os resultados obtidos nos mostram algumas informações interessantes. No geral, todos os artistas tiveram um grande aumento de popularidade entre 2015 e 2016 e, principalmente, entre 2016 e 2017.

É possível notar também uma cerca disparidadade do artista mais popular (Drake), para os demais, principalmente a partir de 2016.

A diferença de grandeza na quantidade de dados ao decorrer do tempo também chama atenção. Uma possível explicação seria o significativo aumento de utilização de serviço de streaming de música ao decorrer dos anos.

Tarefa 3

Desenvolvimento

A tarefa 3 envolveu a análise de como as playlists são criadas. Para identificar se é mais comum playlists com mais músicas do mesmo artista ou playlists com músicas mais diversas, foi calculada a prevalência do artista mais frequente em cada playlist. Por fim, calculou-se a Função de Distribuição Cumulativa, contendo a prevalência do artista em todas as playlists. Um gráfico com a CDF foi desenhado.

Para chegar no resultado desejado, os seguintes passos foram seguidos:

1. Agrupar por playlist e artistas, contar o número de música por artista
2. Encontrar o número máximo de músicas para o artista mais frequente em cada playlist
3. Calcular o número total de músicas para cada playlist
4. Juntar o resultado de 1 com 3, por id de playlist e calcular a prevalência
5. Computar o valor da CDF
6. Plotar o gráfico

Os passos 1-6 estão bem separados no código. Como resultado, obteve-se a seguinte CDF:

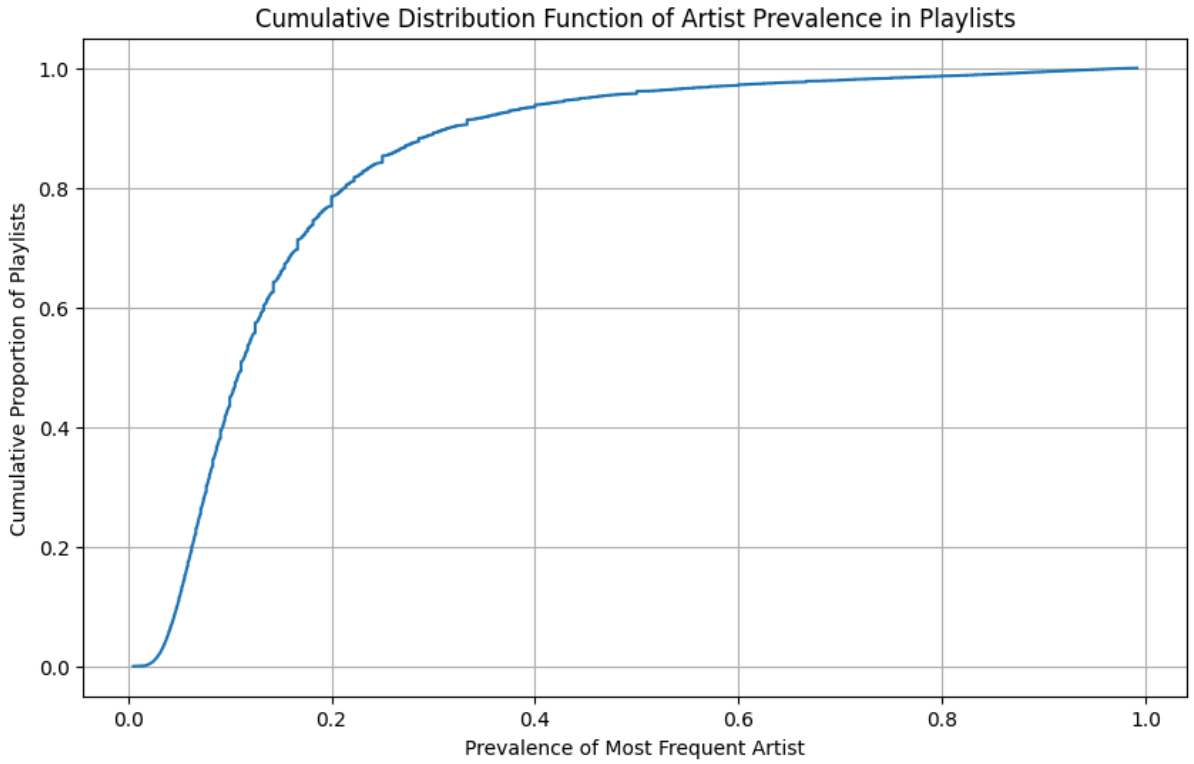

Gráfico 2 - CDF para a prevalência de artistas nas playlists

Discussão

A partir do gráfico 2 é possível obter algumas conclusões interessante sobre a criação das playlists.

A curva apresenta um crescimento rápido entre x=0 e x=0.2, alcançando y=0.8. Isso indica que 80% das playlists possuem prevalência de artista de no máximo 20%. Ou seja, nessas playlists, o artista mais frequente representa até 20% das músicas. Isso nos mostra claramente que a maioria das playlists é diversificada, contendo músicas de vários artistas.

Nota-se um crescimento lento entre x=0.2 e x=0.4, com y=0.95 (aproximadamente). Isso nos mostra que cerca de 15% das playlists possuem um artistas que representa entre 20% e 40% das músicas.

Por fim, é possível observar que apenas cerca de 5% das playlists tem prevalências de artista maiores que 40%, tendo em vista que o crescimento a partir de 0.4 é lento.

É possível concluir, portanto, que a grande maioria das playlists é divesificada, contendo músicas de vários artistas diferentes. Todavia, ainda existem casos, embora em menor quantidade, que a prevalência de artista é elevada, isso pode representar usuários que são extremamente fãs de um determinado artista.

Conclusão

Com o desenvolvimento do trabalho foi possível colocar em prática alguns conceitos de computação distribuída utilizando Spark.

No que se refere a análise dos dados, foi possível aplicar conceitos interessantes para a identificação de outliers, identificação de artistas mais populares e como sua popularidade se deu ao longo do tempo, e identificar o comportamento de criação das playlists.

Referências

INTERQUARTILE range. Wikipedia. Disponível em: https://en.wikipedia.org/wiki/Interquartile_range. Acesso em: 19 nov. 2024.

GREGOIRE, C. **What Is the Interquartile Range Rule?**. ThoughtCo. Disponível em: https://www.thoughtco.com/what-is-the-interquartile-range-rule-3126244. Acesso em: 19 nov. 2024.

CUMULATIVE distribution function. Wikipedia. Disponível em: https://en.wikipedia.org/wiki/Cumulative_distribution_function. Acesso em: 19 nov. 2024.

SPARK documentation. Apache Spark. Disponível em: https://spark.apache.org/docs/latest/api/python/index.html. Acesso em: 19 nov. 2024.