Proposta de Trabalho Final
Processamento de Dados Massivos em Nuvem

Danilo Pimentel [1] **– 2016058077**
Gabriel Bastos [1] **– 2016058204**

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

1. Introdução

Dados gerados a partir da atividade de usuários em plataformas de música são um dos aspectos mais valiosos para este tipo de negócio. Tais dados constituem informações sobre as tendências e preferências de seus usuários, e portanto são o objeto principal de estudo para tais plataformas.

Além disso, tais dados são insumo para funcionalidades de recomendação, um dos diferenciais mais relevantes das plataformas modernas. Sistemas de recomendação não são uma novidade, tal que há uma considerável diversidade na literatura de pesquisa. Neste trabalho, buscamos aplicar metodologias já bem estabelecidas em uma base de dados dos hábitos musicais de usuários reais.

2. Base de Dados

A base de dados adotada para estudo provém da plataforma Last.fm[1], e contém os hábitos musicais de seus usuários no período de 2002 até 2009. Tais dados estão dispostos em duas tabelas:

1. Reproduções: id de usuário; timestamp; id e nome do artista; id e nome da faixa.
2. Usuários: id de usuário; gênero; idade; país; data de inscrição.

Em particular, a base apresenta 19.150.868 registros de reprodução de 992 usuários distintos. Tal volume de dados se mostra suficiente para a aplicação de metodologias de aprendizado de máquina em sistemas de recomendação, como por exemplo filtragem colaborativa[2].

3. Objetivos

Definimos objetivos em dois principais aspectos. O primeiro deles visa uma análise geral, com o objetivo de identificar a disposição dos dados. Tal objetivo é importante para identificar características desbalanceadas, valores inesperados, e obter uma intimidade geral com a base de dados. O segundo aspecto visa a elaboração de funcionalidades de recomendação para os usuários, baseada em tendências presentes nos dados.

Tais objetivos devem ser implementados sobre a plataforma Spark, de forma a explorar os conhecimentos adquiridos na disciplina. Baseado nas demais experiências obtidas na disciplina, acreditamos que o Spark é uma plataforma que nos empodera a realizar tais análises sem dificuldades adicionais.

3.1. Análise geral

Propomos os seguintes tópicos de análise geral:

1. Em qual faixa etária se encontram a maioria dos usuários? É proposta a construção de um histograma de idade dos usuários para entender melhor o público alvo do Last.fm.
2. Quais foram as faixas mais populares entre os usuários no mundo todo? Propomos a produção de um ranking e a análise de distribuição de popularidade das faixas.
3. No mundo todo, quais são os artistas mais populares entre os usuários? Será construído um ranking e analisada a distribuição de popularidade dos artistas com músicas na plataforma.
4. Quem são os *heavy users* do Last.fm? Deseja-se obter um ranking e analisar a distribuição de atividade dos usuários do sistema.
5. Onde estão os usuários do Last.fm? É proposta a análise da distribuição de localização dos usuários da plataforma no mundo.
6. Onde ocorrem a maioria das reproduções de faixas? Propõe-se a construção de um mapa de calor de reproduções por país.
7. Quais são os *hits* do momento? Será construído um ranking de faixas que tiveram os maiores surtos de reproduções no mundo.

3.2. Sistema de recomendação

Propomos a construção de um sistema de recomendação baseado em filtragem colaborativa. O sistema deve considerar as interações dos usuários como implícitas, visto que não há informação explícita sobre preferência na base de dados. A interação implícita se dá pela reprodução de uma faixa.

A filtragem colaborativa é uma técnica oportuna para este conjunto de dados, uma vez que não são disponibilizadas características das faixas em si. Sem informações inerentes às faixas, como o gênero musical, não acreditamos que a utilização de algoritmos baseados em proximidade da vizinhança seja eficaz.

4. Documentação

O produto final a ser produzido neste trabalho é a documentação, onde será descrita a metodologia elaborada, as principais técnicas empregadas e os principais resultados obtidos. A documentação cumpre o papel de relatório do trabalho, servindo como principal objeto de avaliação.

Referências

[1] LAST.FM. *The last.fm website*. 2010. `https://www.last.fm/`.

[2] SPARK. *Collaborative Filtering in Spark*. `https://spark.apache.org/docs/2.2.0/ml-collaborative-filtering.html`.