Bootcamp: Arquiteto(a) de Big Data

Módulo 5: Desafio Final

Para este desafio, vamos exercitar todas as etapas do processo de Big Data. Para isso, vamos utilizar datasets distintos. O primeiro enunciado visa coletar, processar, analisar e aplicar um algoritmo de Machine Learning no dataset de clientes de uma rede de supermercados cujo principal objetivo é conhecer o perfil do cliente. No segundo enunciado vamos coletar, tratar, analisar e aplicar o algoritmo de regras de associação nos datasets de movies e ratings. O objetivo desta atividade é tentar encontrar padrões de preferências de filmes por usuários e realizar uma indicação.

Objetivos geral do desafio

Exercitar os seguintes conceitos trabalhados no curso:

1. Coleta de dados estruturados.
2. Coleta de dados na Web.
3. Criação de estrutura de armazenamento em banco de dados.
4. Tratamento, limpeza e processamento de dados.
5. Análise de dados.
6. Visualização de dados.
7. Desenvolvimento de algoritmos de *Machine Learning;*
    a. K-means;
    b. Regras de associação.
8. Práticas de manipulação de dados.

9. Exercitar comandos Python, SQL.

Enunciado I

Uma rede de supermercados viu a necessidade de criar uma maneira de entender e conhecer melhor o seu público-alvo. Diante desse desafio, a rede precisa criar um processo de Big Data para auxiliar essa análise. A rede quer conhecer seu cliente como um todo, das compras que foram realizadas aos produtos mais vendidos e dessa forma criar uma estrutura que permita tomar decisões mais assertivas.

Os analistas de Big Data da rede identificaram que é necessário desenvolver um processo bem elaborado para transformar dados variados em informações úteis. Para isso é necessário:

1. Coletar dados em diversas fontes;
2. Armazenar os dados em um repositório;
3. Realizar análises de dados coletados;
4. Criar modelo analítico de *Machine Learning*;
5. Criar visualizações para os dados processados;

Para esse primeiro momento, vamos analisar os dados e realizar um agrupamento dos clientes baseados em algumas características que eles possuem.

As compras foram separadas por usuário. Deste modo, cada compra necessita possuir cliente, produto, quantidade de produtos, valor unitário e valor total da compra.

ATENÇÃO

Informação importante: Não existe compra sem produto ou sem cliente.

Os dados de clientes e compras é um dado fictício utilizado para o desenvolvimento das atividades a serem realizadas neste trabalho. Deste modo, os dados foram criados de forma aleatória e não possui nenhuma relação com dados no mundo real.

Atividades do enunciado I

Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados das seguintes fontes de dados.
    a. compras.xls
        i. Contém dados das compras realizadas por cliente;
    b. clientes.json
        i. Contém dados de clientes (análise de perfil);
    c. estados.txt
        i. Contém dados de estados dos clientes;
    d. O link: https://profleandrolessa.wordpress.com/exercicio-de-coleta-de-dados/
        i. Contém dados de produtos;
2. Criar estrutura de armazenamento;
3. Avaliar dados ausentes das colunas e corrigi-los;
4. Criar algoritmo de clusterização k-means;
5. Responder às questões de 1 a 10 práticas do desafio.

Informações de identificação de códigos base de clientes:

código	estado civil
0	solteiro(a)
1	casado(a)
2	viuvo(a)
3	divorciado(a)

código	indicador
0	não hipertenso
1	hipertenso
0	não diabetes
1	diabetes

código	sexo
0	feminino
1	masculino

Enunciado II

Uma empresa de entretenimento coletou dados de clientes e suas avaliações de filmes disponibilizados na sua plataforma on-line. A ideia da empresa é conhecer o perfil de preferência dos seus usuários e consequentemente poder recomendar outros itens de seu catálogo. Diante desse desafio, fez-se necessário criar um algoritmo que possa auxiliar e indicar outros itens do catálogo analisando os padrões e relações entre os filmes vistos. Os analistas de Big Data decidiram desenvolver uma POC com os dados de filmes e usuários para encontrar a relação dos filmes entre os usuários.

Atividades do enunciado II

Para esta atividade, não vamos levar em consideração os ratings enviado pelos usuários. Vamos apenas ver a relação quem assistiu os filmes.

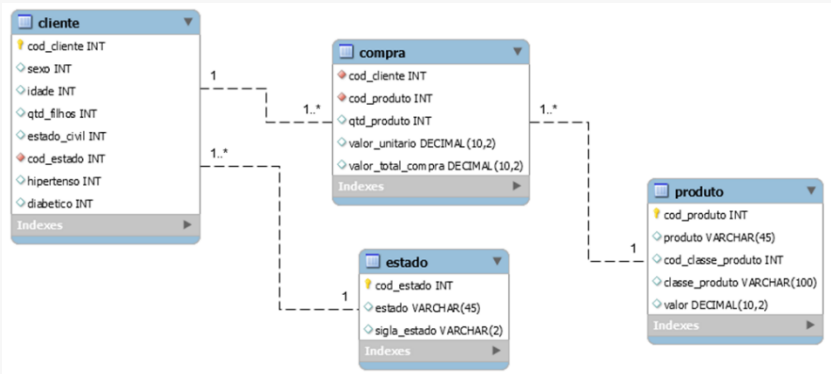
Os alunos deverão desempenhar as seguintes atividades:

1. Coletar dados dos datasets:
    a. movies.csv
        i. Contém dados dos filmes do catalogo.
    b. rating.csv
        i. Contém os dados de usuários e suas avaliações.

2. Analisar os dados coletados;
3. Tratar os dados coletados;
4. Avaliar dados ausentes nas colunas;
5. Criar algoritmo de regras de associação;
6. Identificar os itens frequentes;
7. Criar regras de associação;
8. Responder às questões de 11 a 15 práticas do desafio;

Dicas do professor I

1. Para armazenar os dados, pode utilizar banco de dados relacional MYSQL ou banco de dados não relacional Mongodb;
2. Para as atividades relacionados a clusterização realizada pelo k-means, crie um novo dataframe (compras_clientes) com os dados obtidos entre a relação das tabelas cliente, estado, produto e compras;
3. Muita atenção para realizar inserts nas tabelas para o banco de dados relacional. Verifique a hierarquia dos dados nas tabelas;
4. Segue uma sugestão para modelagem de dados.



5. Ao coletar e armazenar os dados de compras no formato xls, pode ser necessário instalar um complemento. Caso isso aconteça, utilize: !pip install xlrd;
6. Muita atenção para as informações dos códigos 0 e 1 da tabela de clientes;
7. Analisem bem os dados ausentes e utilizem o bom senso para corrigir os dados. Analisem bem os dados e suas relações;
8. Crie o agrupamento com n_clusters = 4 e random_state = 0;
9. Não existem pegadinhas nas questões. Tenham atenção e sigam as instruções;
10. Os datasets estão disponíveis no link:
    a. https://github.com/ProfLeandroLessa/desafio-final-ABD

Dicas do professor II

Segue algumas orientações para auxiliar no tratamento dos dados para as questões do enunciado II.

1. Para este desafio, não vamos levar em consideração os ratings avaliados pelos usuários.
2. Analisem bem os dados e criem uma estrutura que contenha usuário nas linhas e os filmes vistos nas colunas.
    a. Existem várias maneiras de fazer essa tarefa, no entanto sugiro utilizar o PIVOT TABLE ou PIVOT do pandas.
    b. Link da documentação:
        i. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.pivot.html
3. Crie uma coluna nova com o nome de “visto” e atribua o valor 1.

a. Essa coluna será utilizada para indicar os filmes que os usuários assistiram.

4. Para os filmes que os usuários não assistiram, atribua o valor 0.

5. Em resumo: O usuário viu o filme (possui rating) recebe 1. Usuário que não viu o filme (não possui rating) recebe 0.

6. Se utilizarem os PIVOT, siga as dicas:

   a. Utilize o usuário como index, os títulos do filme para colunas e a coluna “visto” para valores.

   b. Elimine o index gerado no pivot.

7. Exemplo do tratamento dos dados realizados:

ALGUNS EXEMPLOS DE FILMES ASSISTIDOS POR USUÁRIOS

titulo	'71 (2014)	'Hellboy': The Seeds of Creation (2004)	'Round Midnight (1986)	'Salem's Lot (2004)	'Til There Was You (1997)	'Tis the Season for Love (2015)	'burbs, The (1989)	'night Mother (1986)	(500) Days of Summer (2009)	*batteries not included (1987)	...	Zulu (2013)
11	0	0	0	0	0	0	1	0	0	0	...	0
12	0	0	0	0	0	0	0	0	0	0	...	0
13	0	0	0	0	0	0	0	0	0	0	...	0
14	0	0	0	0	0	0	0	0	1	0	...	0
15	0	0	0	0	0	0	0	0	0	0	...	0
16	0	0	0	0	0	0	0	0	0	0	...	0
17	0	0	0	0	0	0	0	0	1	0	...	0
18	0	0	0	0	0	0	1	0	0	0	...	0
19	0	0	0	0	0	0	0	0	0	0	...	0
20	0	0	0	0	0	1	0	0	0	0	...	0
21	0	0	0	0	0	0	0	0	1	0	...	0

8. Faz parte do desafio o aluno pesquisar a solução das atividades propostas.