Bootcamp: Cientista de Dados

Trabalho Prático

Módulo 2	CDD – Coleta e Obtenção de Dados

Objetivos

Exercitar os seguintes conceitos trabalhados no Módulo:

- ✓ Realizar o cadastro no Twitter para obter a conta de desenvolvedor.
- ✓ Criar uma aplicação no Twitter para ser utilizada posteriormente.
- ✓ Realizar coleta de dados em arquivos utilizando uma das seguintes opções de tecnologia: Linguagem R ou Python, ou plataforma Knime analytics.
- ✓ Realizar operações de criação de bases de dados relacional utilizando uma das seguintes opções: Linguagem R ou Python, ou plataforma Knime analytics.
- ✓ Realizar carga e coleta de dados em SGBD relacional utilizando uma das seguintes opções de tecnologia: Linguagem R ou Python, ou plataforma Knime analytics.

Enunciado

Para esta atividade, o aluno deverá assistir atentamente as seguintes aulas, disponíveis no ambiente virtual de aprendizagem, **exatamente na ordem indicada nas atividades**.

Considere ainda os arquivos complementares anexo ao enunciado do trabalho prático:

- Script para criação do banco de dados: *script_BD_bootcamp.sql*
- Arquivos de carga para as tabelas do banco de dados:
  - *estado.xlsx*
  - *cidade.csv*
  - *caracteristicasgerais.csv*

  - *imoveis.csv*
  - *caracteristicageralimovel.txt*
  - *caracteristicageralimovel.csv*

- Arquivo com workflow Knime: *cargaSGBDMySQL.knwf*
- Arquivos com os notebooks R e Python:
  - coletaDadosMySQL_R.ipynb
  - coletaDadosMySQL_Python.ipynb

Atividades

Os alunos deverão desempenhar as seguintes atividades:

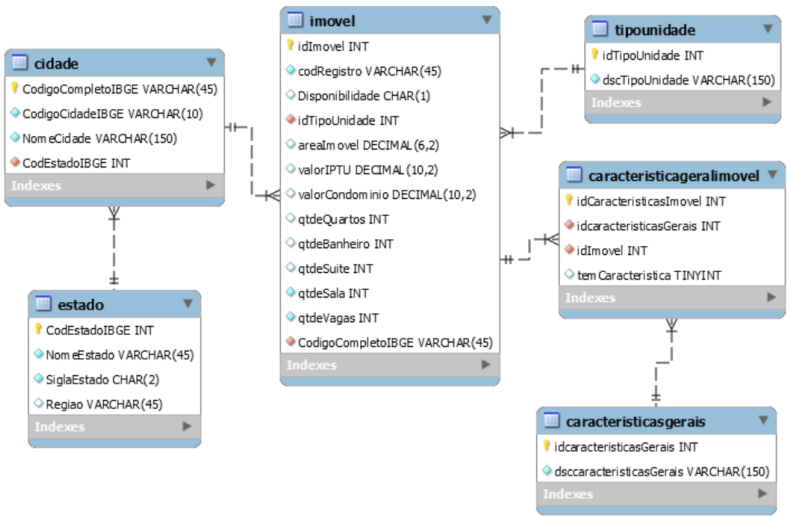
- Para realizar as atividades 1 e 2, deve-se assistir a videoaula denominada “APIs de coleta de dados”.

1. Realizar o cadastro no Twitter para desenvolvedor e obter as chaves necessárias para utilizar a API do Twitter.
   - Obs.: O link para realizar o cadastro é https://developer.twitter.com/en
2. Após receber o acesso de desenvolvedor na API do Twitter, crie sua aplicação chamada de “Coletor IGTI”, pois vamos utilizar esta aplicação no Desafio do módulo.
   - Obs.: No campo Website URL informe https://www.igti.com.br/
3. Assistir as videoaulas do tópico “Ferramentas usadas para coleta de dados”, e seguir o tutorial de instalação do ambiente Anaconda (*Aula Instalação e configuração do Framework Anaconda*).
4. Após a instalação do Anaconda, criar o ambiente denominado “*dev*” para trabalhar com a linguagem R e instalar o RStudio (*Aula Framework Anaconda: Ambiente R Studio*). Caso não tenha sido instalado na criação do ambiente, instale ainda o Prompt do Anaconda (CDM.exe Prompt), o Jpyter Notebook e o Spyder.

- Se achar necessário, para validar seu ambiente "*dev*" com a linguagem Python e R, replique os exemplos "Alô Mundo!" das respectivas aulas.

5. Assistir as videoaulas do tutorial MySQL Server e MySQL Workbench, e realizar a instalação das duas ferramentas em seu computador.

6. Criar o banco de dados "*bootcamp*" com as tabelas do esquema conforme o modelo a seguir. Para isso, utilize o arquivo "*bootcamp.sql*", disponível em conjunto com o enunciado deste trabalho. Durante a criação, analise o script gerado, pois você vai precisar dele para responder as questões.



7. Assistir as videoaulas do tópico "Coleta de dados estruturados: Exemplo utilizando a linguagem R" e, em seguida, executar o notebook "coletaDadosMySQL_R.ipynb", disponível em conjunto com o enunciado deste trabalho. Analise e compreenda o código do notebook e sua respectiva execução, pois você vai responder questões relacionadas a ele.

   - Para esta atividade, você vai precisar do arquivo "estados.xlsx", salvo na pasta "C:\Bootcamp\Datasets\XLS". Se deseja alterar a pasta, certifique-se de fazer a mesma alteração no notebook.

8. Alterar o notebook "coletaDadosMySQL_R.ipynb", utilizando a linguagem R, para inserir novos registros na tabela *caracteristicasgerais*. Para apoiar a execução desta atividade, foi fornecido o arquivo csv chamado "*caracteristicasgerais.csv*". Ao seu final, sua tabela deve ficar como a imagem abaixo.

idcaracteristicasGerais	dsccaracteristicasGerais
1	Portaria 24 horas
2	Elevador
3	Piscina
4	Salão de festas
5	Área gourmet
6	Água individual
7	Gás canalizado
8	Aquecimento solar
9	Vaga coberta
10	Vaga livre
11	Armários na cozinha
12	Closet
13	Armários no quarto

9. Assistir as videoaulas do tópico “Coleta de dados estruturados: Exemplo utilizando a linguagem Python” e, em seguida, executar o notebook “coletaDadosMySQL_Python.ipynb”, disponível em conjunto com o enunciado deste trabalho. Analise e compreenda o código do notebook e sua respectiva execução, pois você vai responder questões relacionadas a ele.

    – Para esta atividade você vai precisar do arquivo “cidades.csv” salvo em na pasta “C:\Bootcamp\Datasets\CSV”. Se deseja alterar esta pasta, certifique-se de fazer a mesma alteração no notebook.

10. Alterar o notebook “coletaDadosMySQL_Python.ipynb”, utilizando a linguagem Python, para inserir registros na tabela *caracteristicageralimovel*. Para apoiar a execução desta atividade, foi fornecido o arquivo txt chamado “*caracteristicageralimovel.txt*”. Ao final, sua tabela deve ficar com 101 registros.

    – O arquivo *caracteristicageralimovel.txt* possui apenas 3 colunas: *idcaracteristicasGerais, idImovel e temCaracteristica*. A coluna *idCaracteristicasImovel* é definida como auto incrementável, e por isso vai gerar seu valor automaticamente no momento da inserção de cada registro.

11. Assistir as videoaulas do tópico “Plataforma Knime Analytics” (parte 1 e 2), e seguir o tutorial de instalação da Plataforma Knime Analytics para instalar esta Plataforma em seu computador. Em seguida, assistir as videoaulas do tópico “Coleta de dados estruturados: Exemplo utilizando a Plataforma Knime (parte 1 e 2)”.

12. Utilizando a Plataforma Knime Analytics, importar o workflow "*cargaSGBDMySQL.knwf*" (*Aula Importar e Exportar workflows com a Plataforma Knime*). Este workflow está disponível em conjunto com o enunciado deste trabalho. Execute-o para realizar a carga na tabela de imóveis.

    – Para esta atividade você vai precisar do arquivo "imoveis.csv", salvo em na pasta "C:\Bootcamp\Datasets\CSV". Se deseja alterar esta pasta, certifique-se de fazer a mesma alteração no nó do workflow.

13. Alterar o workflow "*cargaSGBDMySQL*" para inserir mais registros na tabela *caracteristicageralimovel*. Para apoiar a execução desta atividade, foi fornecido o arquivo csv chamado "*caracteristicageralimovel.csv*". Ao final, sua tabela deve ficar com 233 registros.

Respostas Finais

Os alunos deverão desenvolver a prática e, depois, responder às seguintes questões objetivas: