Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

CKP9011 – Introdução à Ciência de Dados
CK0223 - Mineração de Dados
2025.1

Lista 2

Exercício: Tratamento de Dados
Objetivos: Exercitar os conceitos referente à manipulação, tratamento e limpeza de dados.
Data da Entrega: 13/05/2025

1. Tarefa

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

a) Ler o dataset fakeTelegram.BR_2022.csv, o qual está disponível no link a seguir: https://drive.google.com/file/d/1c_hLzk85pYw-huHSnFYZM_gn-dUsYRDm/view?usp=drive_link
b) Remova os trava-zaps.
c) Exportar os dados para um arquivo Parquet.
d) Exportar os dados para o DuckDB.
e) Utlizando o DuckDB recupere:
    1. A quantidade de mensagens;
    2. A quantidade de usuários;
    3. A quantidade de grupos;
    4. Quantidade de mensagens que possuem apenas texto;
    5. Quantidade de mensagens contendo mídias;
    6. Quantidade de mensagens por tipo de mídia (jpg, mp4 etc);
    7. Quantidade de mensagens por estado;
    8. Quantidade de usuários por estado;
    9. Relação quantidade de usuários por quantidade de mensagens por estado;
    10. Quantidade de mensagens por país;
    11. Quantidade de mensagens Brasil X Países Estrangeiros;
    12. As 30 URLs que mais se repetem (mais compartilhadas);
    13. Os 30 domínios que mais se repetem (mais compartilhados);
    14. Os 30 usuários mais ativos;
    15. Os 30 usuários que mais compartilharam texto;
    16. Os 30 usuários que mais compartilharam mídias;
    17. As 30 mensagens mais compartilhadas;
    18. As 30 mensagens mais compartilhadas em grupos diferentes;
    19. Mensagens idênticas compartilhadas pelo mesmo usuário (e suas quantidades);
    20. Mensagens idênticas compartilhadas pelo mesmo usuário em grupos distintos (e suas quantidades);

21. Os 30 unigramas, bigramas e trigramas mais compartilhados;
22. As 30 mensagens mais positivas (distintas);
23. As 30 mensagens mais negativas (distintas);
24. O usuário mais otimista;
25. O usuário mais pessimista;
26. As 30 maiores mensagens;
27. As 30 menores mensagens;
28. O dia em que foi publicado a maior quantidade de mensagens;
29. As mensagens que possuem as palavras “FACÇÃO” e “CRIMINOSA”;
30. As mensagens que possuem a palavra “SEGURANÇA”.

2. Avaliação

Espera-se com a realização deste trabalho que cada estudante elabore e entregue (de forma digital) os seguintes documentos:

- Jupyter Notebook contendo o código utilizado na implementação das tarefas.
- Vídeo (disponibilizado no Youtube) apresentando e descrevendo as atividades desenvolvidas.

A avaliação deste trabalho se dará em duas etapas:

1ª. Vídeo de Apresentação do Dataset: Cada estudante irá disponibilizar um vídeo (no Youtube) apresentando o código desenvolvido para implementação das tarefas. O estudante pode utilizar slides e notebooks.

2ª. Avaliação do Notebook: O professor da disciplina irá avaliar a qualidade do notebook gerado pelo estudante, bem como dos códigos implementados e análises realizadas.

A avaliação do trabalho irá envolver os seguintes quesitos:

- Abrangência e Organização do Notebook
- Qualidade dos Códigos Utilizados
- Clareza do Texto Utilizado para Descrever as Atividades Realizadas e os Resultados Obtidos
- Domínio do Tema

3. Data da Entrega: 13/05/2025

- PS. O trabalho é individual.
- PS. Não serão aceitos trabalhos que não forem apresentados (por meio de vídeo disponibilizado no Youtube).
- PS. Cada estudante será responsável pela disponibilização do ambiente (software e hardware) necessário para a gravação da apresentação do seu trabalho.
- Os Notebooks deverão ser disponibilizados, em formato .ZIP, no SIGAA ou em um repositório público (GitHub ou GitLab).

“A Educação, qualquer que seja ela, é sempre uma teoria do conhecimento posta em prática”.
Paulo Freire