Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

# CKP9011 e CK0268 – Introdução à Ciência de Dados

**2024.1**

**Lista 2**

Exercício: Tratamento de Dados
Objetivos: Exercitar os conceitos referente à manipulação, tratamento e limpeza de dados.
Data da Entrega: 25/03/2024

OBS 1: Exercício Individual.
OBS 2: A entrega da lista deverá ser executada utilizando-se o SIGAA.

## Questão 1

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

a) Realizar o "restore" do arquivo (dump) denominado fd_whatsapp_0911_2023.zip no PostgreSQL. Esse arquivo está disponível no link a seguir: https://drive.google.com/drive/folders/1kEEnmZUVJEgYTynZjU6qMICbEVd8wKca?usp=sharing
b) Remova os trava-zaps.
c) Exportar os dados para CSV.
d) Exportar os dados para um arquivo Parquet.
e) Exportar os dados para o DuckDB.
f) 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 “INTERVENÇÃO” e “MILITAR”;
30. As mensagens que possuem a palavra “STF”.

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