



Universidade Federal do Ceará
Centro de Ciências
Departamento de Computação

CKP9011 e CK0268 – Introdução à Ciência de Dados 2024.1

Lista 4

Exercício: Análise Exploratória de Dados

Objetivos: Exercitar os conceitos referente à análise exploratória de dados.

Data da Entrega: 29/04/2024

OBS 1: Exercício Individual.

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

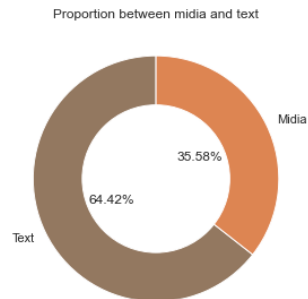
Questão 1

Crie um arquivo Jupyter Notebook e realize as seguintes operações:

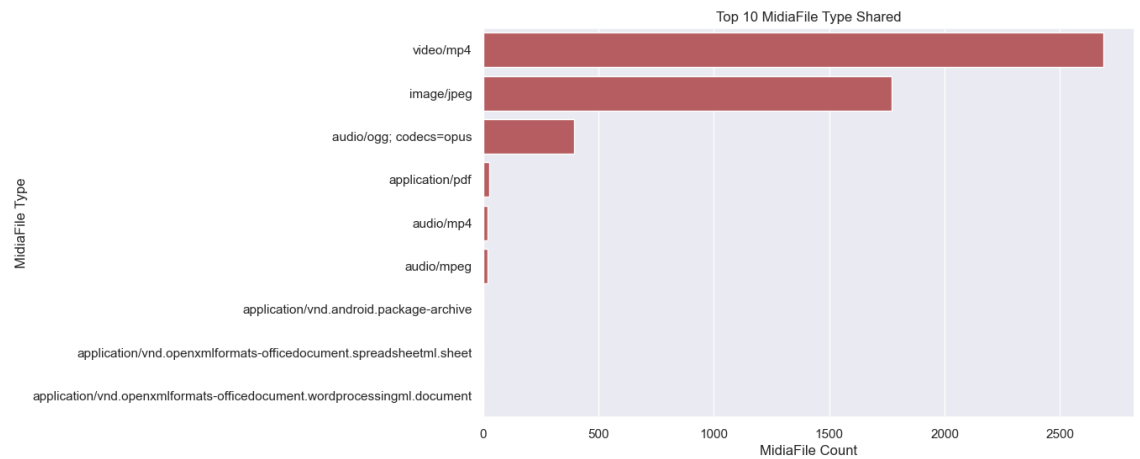
- a) Realizar o “restore” do arquivo (dump) denominado fd_whatsapp_0911_2023.zip no PostgreSQL. Esse arquivo está disponível no link a seguir:
<https://drive.google.com/drive/folders/1kEEnmZUVJEgYTynZjU6qMICbEVd8wKca?usp=sharing>
- b) Remova os trava-zaps.
- c) Remover as linhas repetidas (duplicadas).
- d) Remover textos com menos de 5 palavras.
- e) Para cada atributo numérico apresente:
 1. Medidas de Tendência Central
 2. Medidas de Variabilidade
 3. Tabela de Frequência e Histograma
 4. Boxplot
 5. QQ-Plot
 6. Teste de Normalidade
 7. Best Fit Distribution (usando diferentes métodos)
- f) Para cada par de atributos numéricos apresente:
 1. O Coeficiente de Correlação apropriado
 2. Um Gráfico de Dispersão
- g) Para cada par de atributos categóricos apresente:
 1. O resultado do método V de Cramer

h) Crie uma visualização (gráfico) para apresentar:

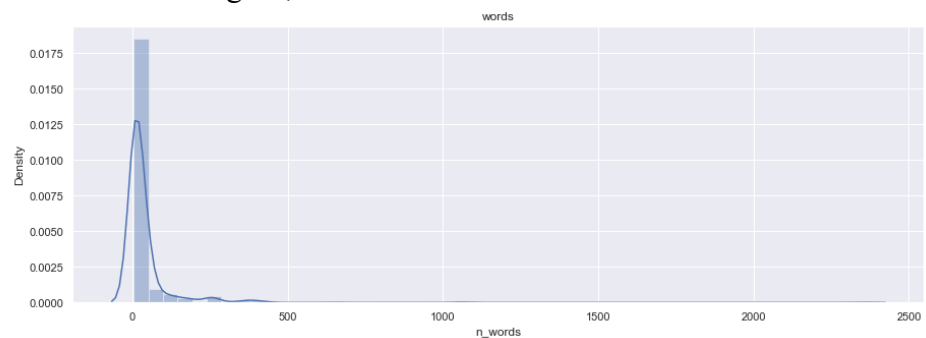
1. As quantidades de grupos, usuários e mensagens;
2. A quantidade de mensagens que possuem apenas texto X mídia;



3. Quantidade de mensagens por tipo de mídia (jpg, mp4 etc);



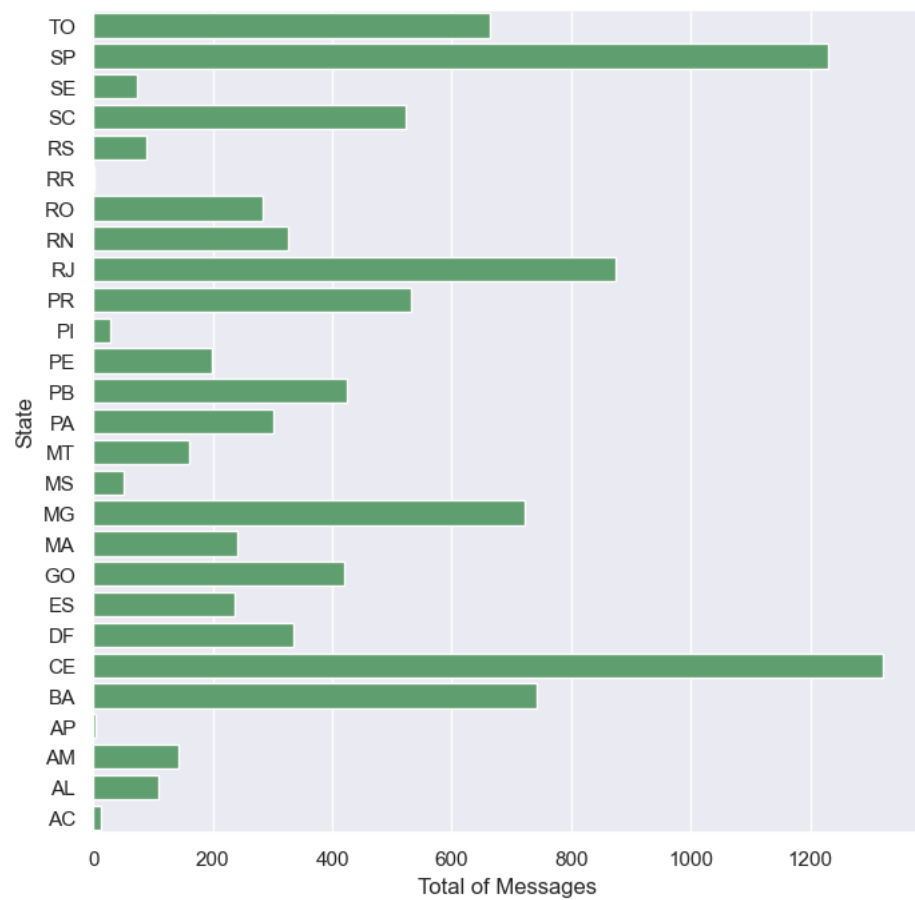
4. A relação entre a quantidade de mensagens e a quantidade de palavras presente nas mensagens;



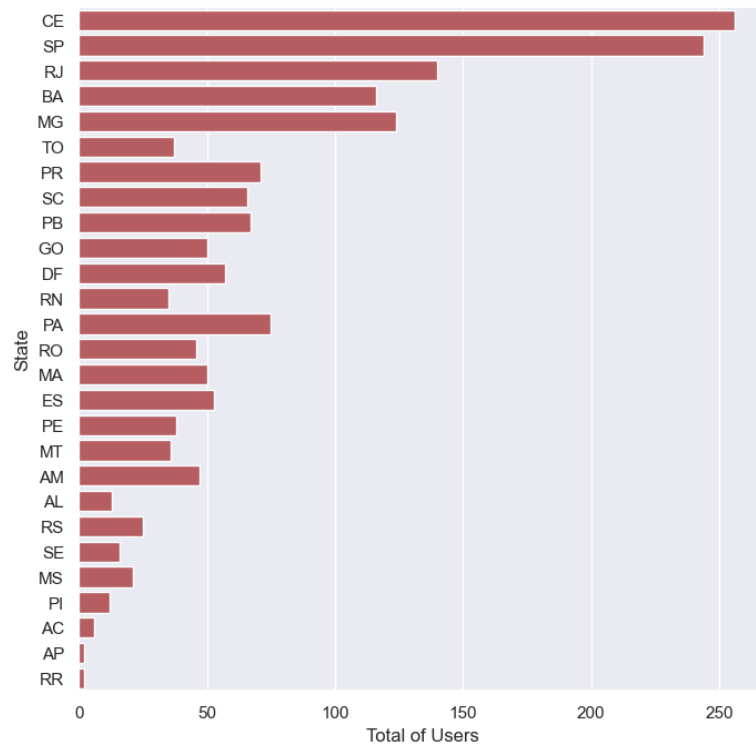
5. Quantidade de mensagens por estado;

Out[23]:

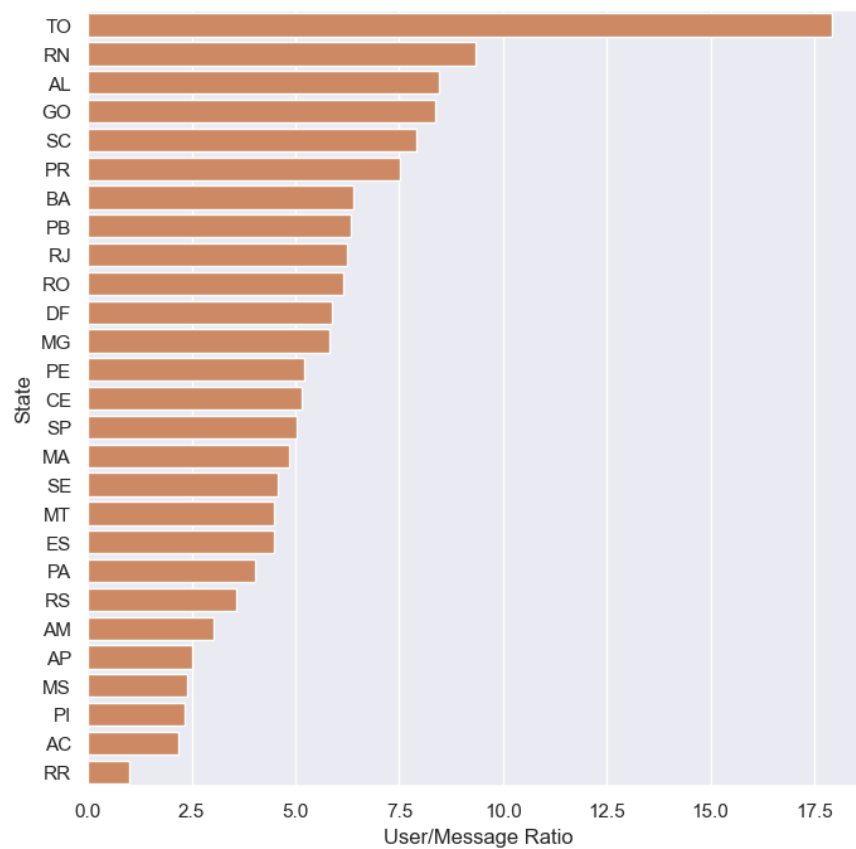
	state	total messages	total_users	messages_ratio
0	TO	229	23	9.956522
1	SP	1126	110	10.236364
2	SE	46	9	5.111111
3	SC	560	85	6.588235
4	RS	129	17	7.588235
5	RR	27	2	13.500000
6	RO	32	5	6.400000
7	RN	161	13	12.384615
8	RJ	487	48	10.145833
9	PR	536	46	11.652174
10	PI	47	9	5.222222
11	PE	250	41	6.097561
12	PB	473	36	13.138889
13	PA	425	84	5.059524
14	MT	191	19	10.052632
15	MS	21	6	3.500000
16	MG	750	68	11.029412
17	MA	366	51	7.176471



6. Quantidade de usuários por estado;



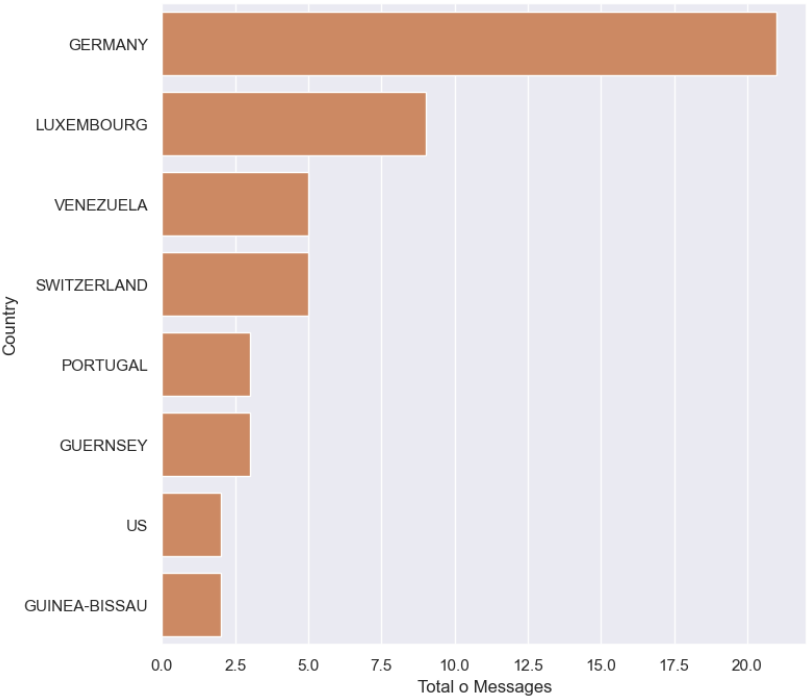
7. Relação quantidade de usuários por quantidade de mensagens por estado;



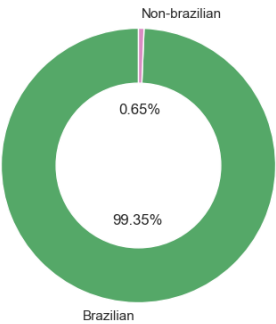
8. Quantidade de mensagens por país;

Out[31]:

	country	total messages	total unique users	message/user ratio
0	VENEZUELA	5	2	2.5
1	US	2	2	1.0
2	SWITZERLAND	5	2	2.5
3	PORTUGAL	3	2	1.5
4	LUXEMBOURG	9	1	9.0
5	GUINEA-BISSAU	2	1	2.0
6	GUERNSEY	3	1	3.0
7	GERMANY	21	2	10.5



9. Quantidade de mensagens Brasil X Países Estrangeiros;



10. As 30 URLs que mais se repetem (mais compartilhadas);



11. Os 30 domínios que mais se repetem (mais compartilhados);

Out [40]:

	URL Domain	URL count	Number of unique URLs	Number of users
2	youtube	602	385	602
8	terrabrasilnoticias	298	164	298
1	facebook	278	231	278
3	instagram	267	214	267
0	twitter	36	29	36
5	tiktok	24	23	24
7	jornaldacidade	12	9	12
9	g1.globo	10	8	10
13	gazetabrasil	7	4	7
4	drive	1	1	1
6	rumble	1	1	1
10	bit.ly	1	1	1
11	chat.whatsapp.com	0	0	0
12	t.me	0	0	0

12. Os 30 usuários mais ativos;

Out [35]:

	id	count	messages	texts	media	ddi	ddd	country	state
0	cebd2107ff2001db85851cc0e81e0667		55	55	0	55	83	BRAZIL	PB
1	ca191ceb779cd47c0519ca5208f3d13		50	41	9	55	92	BRAZIL	AM
2	dfa59e63e5046ab8786f8beffea4995		32	31	1	55	33	BRAZIL	MG
3	0e2bc3223037e8e43f6706aea65154af		29	29	0	55	11	BRAZIL	SP
4	c4c18dfc587b1d40a4c6bfc6d1cc8c89		27	27	0	55	77	BRAZIL	BA
5	53abfb5e4dd3e082443f3e921d442380		24	23	1	55	49	BRAZIL	SC
6	15d6e5ca7c6da2b8b7d00521bd17cac8		23	23	0	55	41	BRAZIL	PR
7	1e0b20142e2dcba76ca1e81a6487a50f		23	23	0	55	11	BRAZIL	SP
8	2b8b665928822d425cb00df9771cf455		19	19	0	55	51	BRAZIL	RS
9	acb33f60b30463b1af944e02dd473aef		17	17	0	55	11	BRAZIL	SP

13. Relação entre quantidade de mensagens contendo somente texto e mensagens com tendo mídia dos usuários mais ativos;

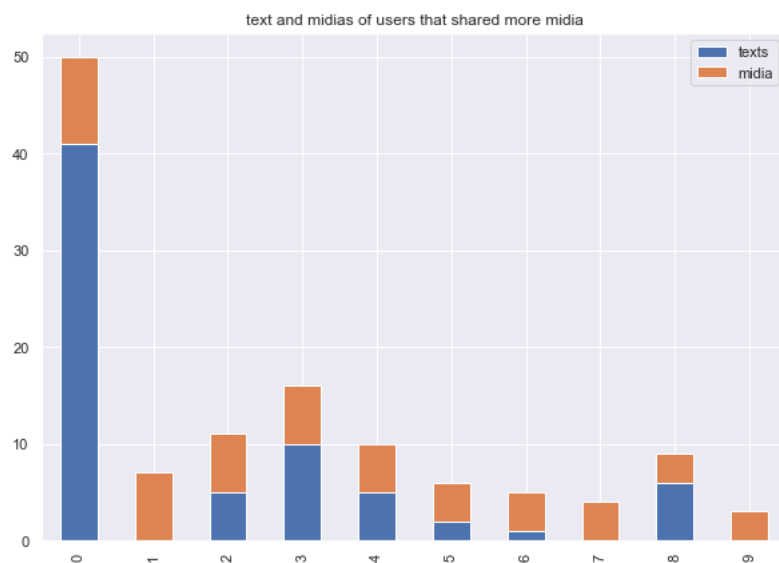


14. Os 30 usuários que mais compartilharam texto;

15. Os 30 usuários que mais compartilharam mídias;

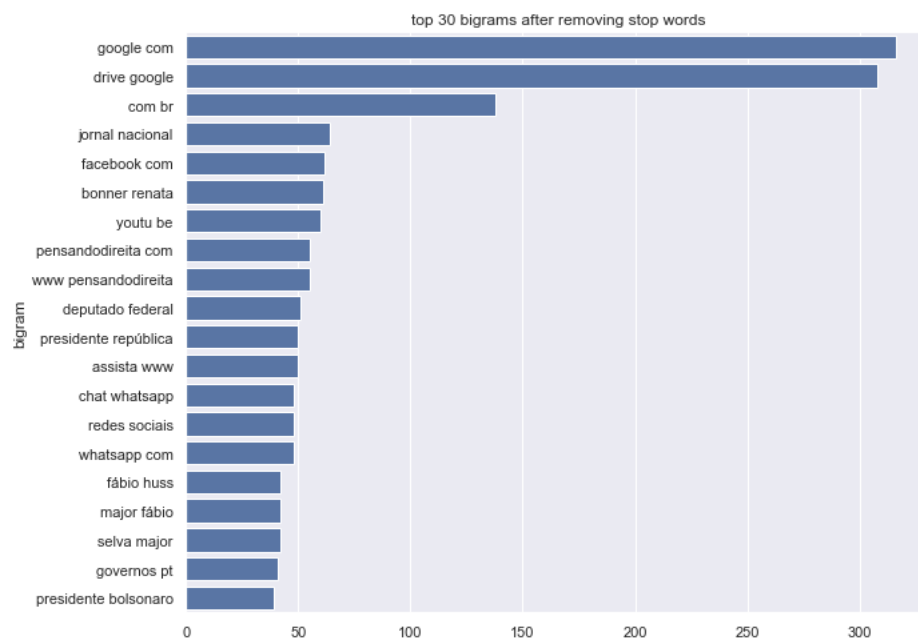
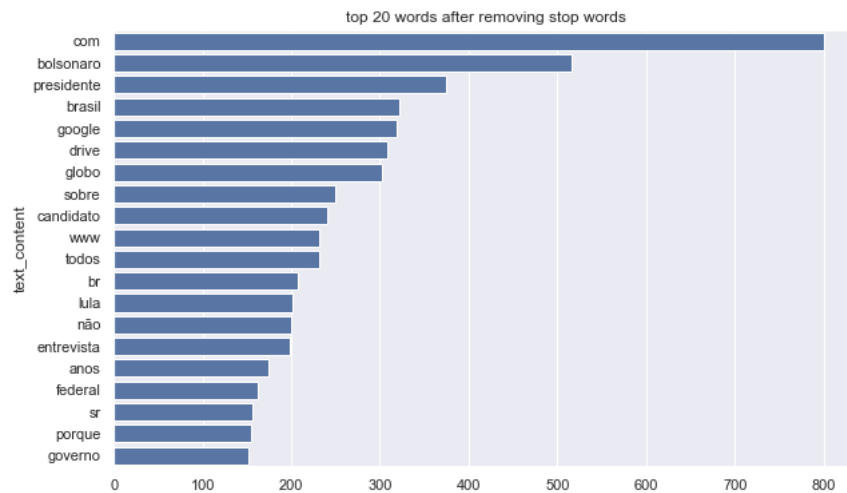
Out[37]:

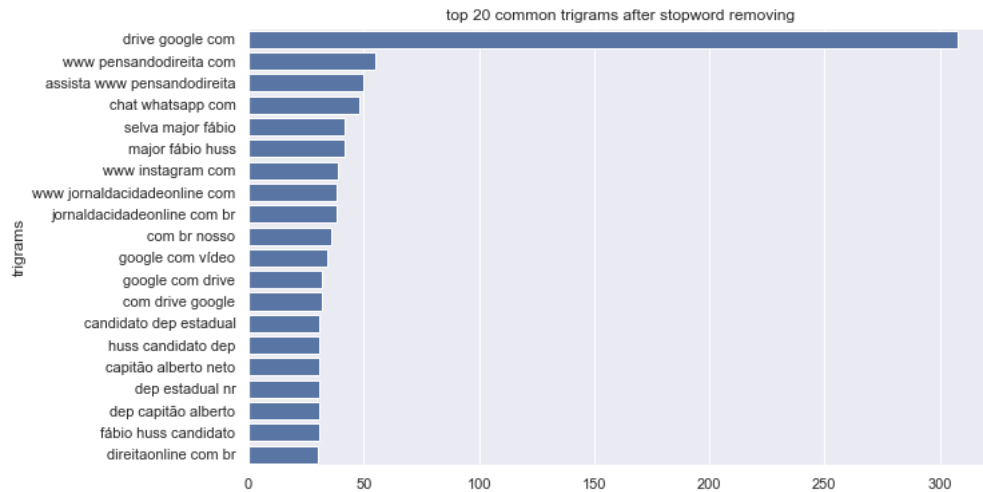
	id	count messages	texts	midia	ddi	ddd	country	state
0	ca191ceb779cd4f7c0519ca5208f3d13	9	41	9	55	92	BRAZIL	AM
1	71353b3768166691b5596f1426da8ce7	7	0	7	55	84	BRAZIL	RN
2	f55fc70c99788530140ce3ec0600bc41	6	5	6	55	61	BRAZIL	DF
3	9a69159052d4f67ff97cf02e2e0505ff	6	10	6	55	45	BRAZIL	PR
4	e41177a9f2a9b69507e0412dc6f2b505	5	5	5	55	21	BRAZIL	RJ
5	784ca076f2cd5634e123843b8095fbd4	4	2	4	55	98	BRAZIL	MA
6	6b4b08d74e4df1b4684ccc76b3cb570f	4	1	4	55	11	BRAZIL	SP
7	2868acfd7bcd31484732e38a65417571	4	0	4	55	64	BRAZIL	GO
8	211a2dbc44b6bd0742cda56af511c6e	3	6	3	55	92	BRAZIL	AM
9	9f7b2f9fa6828e155f9350ff48f4038a	3	0	3	55	92	BRAZIL	AM



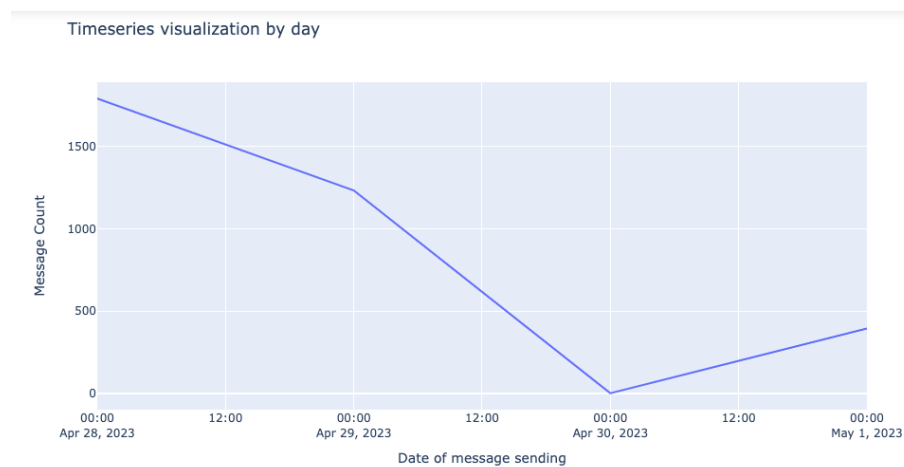
16. As 30 mensagens mais compartilhadas;

17. As 30 mensagens mais compartilhadas em grupos diferentes;
18. Mensagens idênticas compartilhadas pelo mesmo usuário (e suas quantidades);
19. Mensagens idênticas compartilhadas pelo mesmo usuário em grupos distintos (e suas quantidades);
20. Os 30 unigramas, bigramas e trigramas mais compartilhados (após a remoção de stop words);

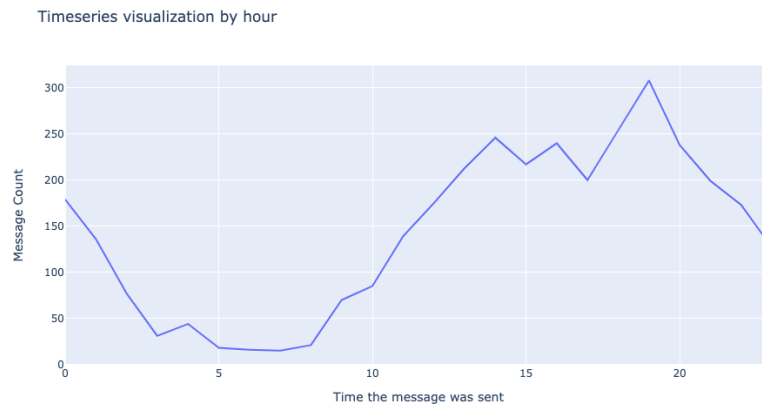




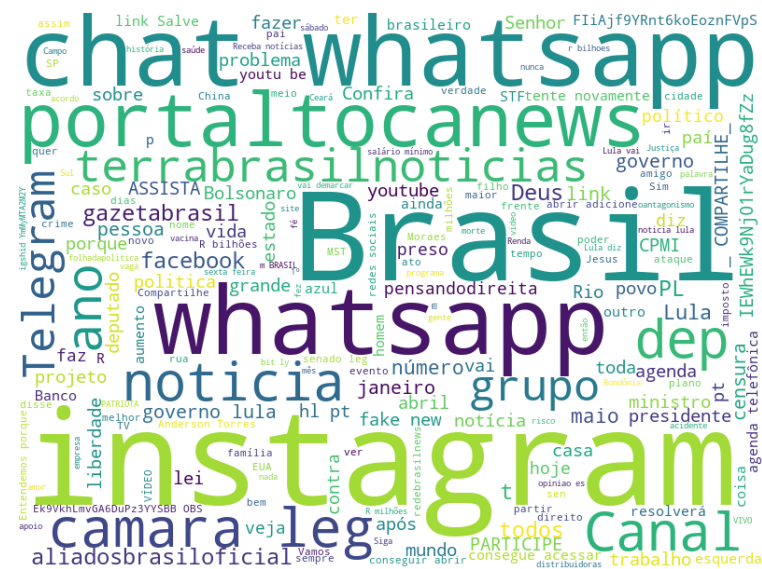
21. As 30 mensagens mais positivas (distintas);
22. As 30 mensagens mais negativas (distintas);
23. O usuário mais otimista;
24. O usuário mais pessimista;
25. As 30 maiores mensagens;
26. As 30 menores mensagens;
27. O dia em que foi publicado a maior quantidade de mensagens;
28. As mensagens que possuem as palavras “INTERVENÇÃO” e “MILITAR”;
29. Quantidade de mensagens por dia e hora;



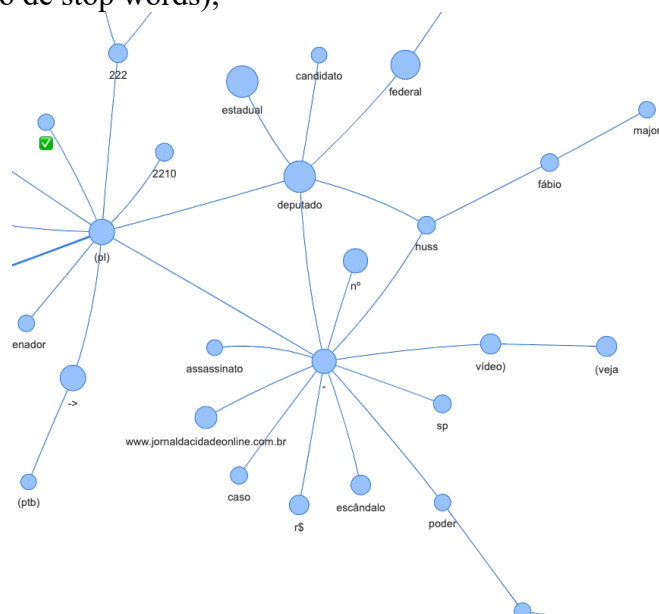
30. Quantidade de mensagens por hora;



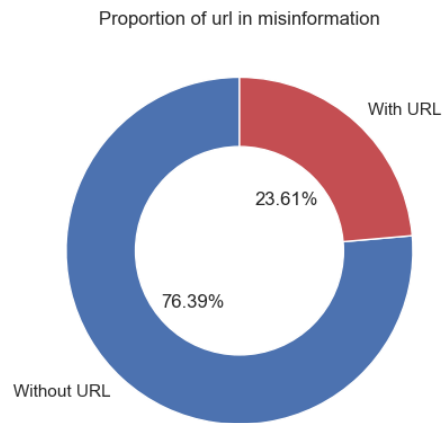
31. A nuvem de palavras referente às mensagens de texto (após a remoção de stop words);



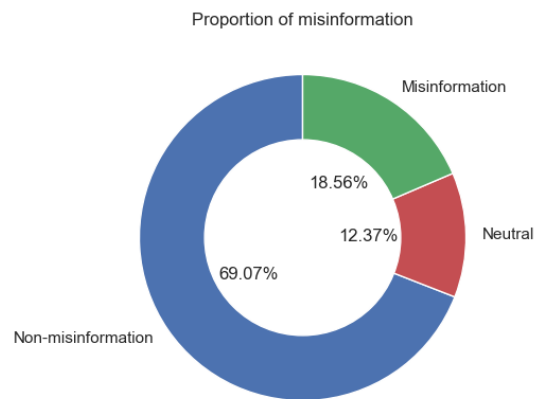
32. A rede interativa das palavras referente às mensagens de texto (após a remoção de stop words);



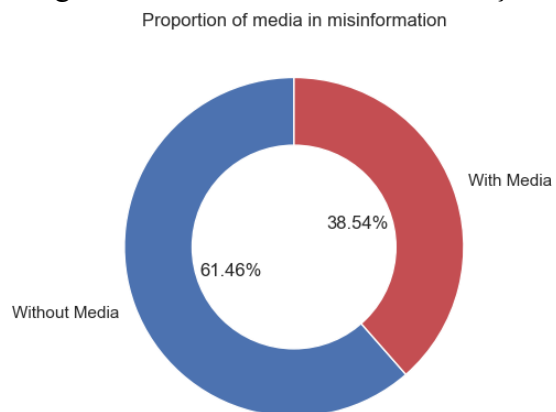
33. Proporção de mensagens com e sem URL;



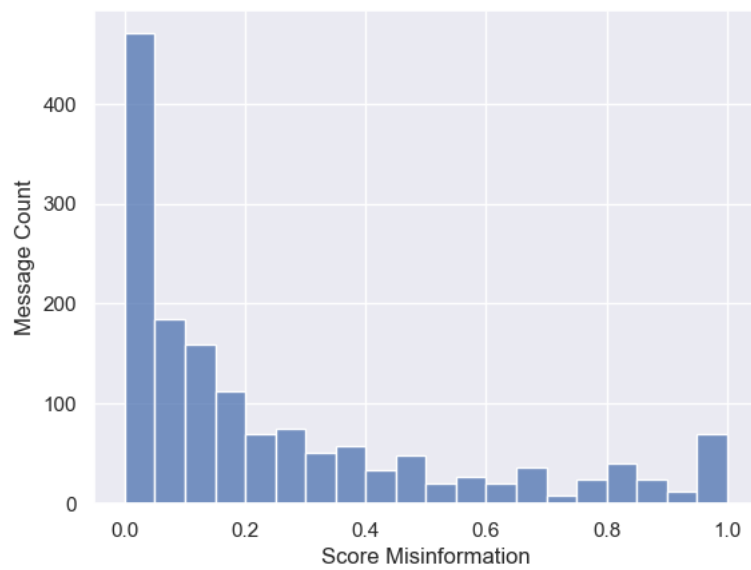
34. Proporção de desinformação;



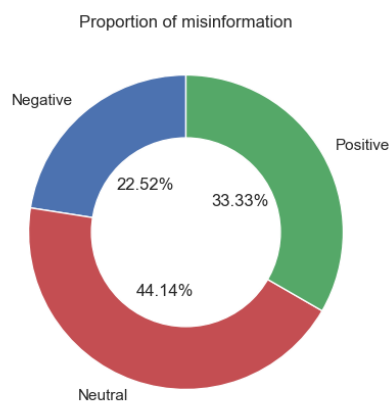
35. Proporção de mensagens contendo mídia e desinformação;



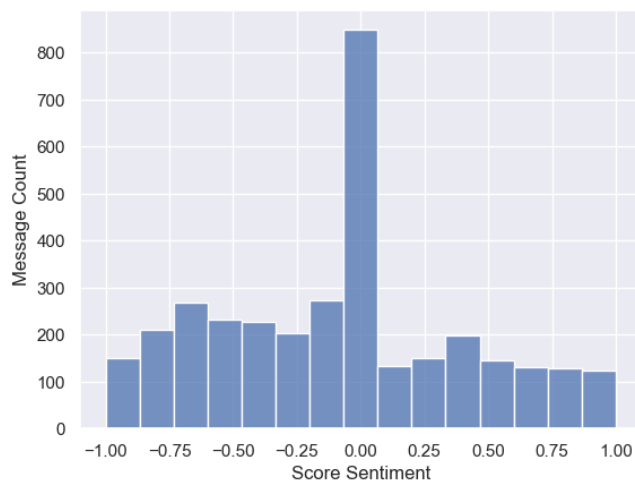
36. Distribuição de mensagens por score de desinformação;



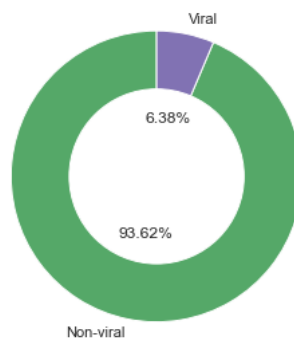
37. Proporção de sentimentos;



38. Distribuição de mensagens por score de sentimentos;



39. Proporção entre mensagens virais e não virais;



40. Algo que você julga importante e que ainda não foi solicitado;

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