Atividade 02 - Análise exploratória

Equipe

Nome: Detetives Virtuais
Repositório GitLab da equipe: https://gitlab.com/jr_waine/csb-51
Membros:

- Waine Barbosa de Oliveira Junior, 1905120,
  jr_waine, waine@alunos.utfpr.edu.br, Eng. Comp, UTFPR
- Thiago Schinda Bubniak, 1540778,
  thiagobubniak, thiagobubniak@alunos.utfpr.edu.br, Eng. Comp, UTFPR

Tema

Compreender quais os temas de fake news mais relevantes em determinado período e comparar com eventos relacionados no período

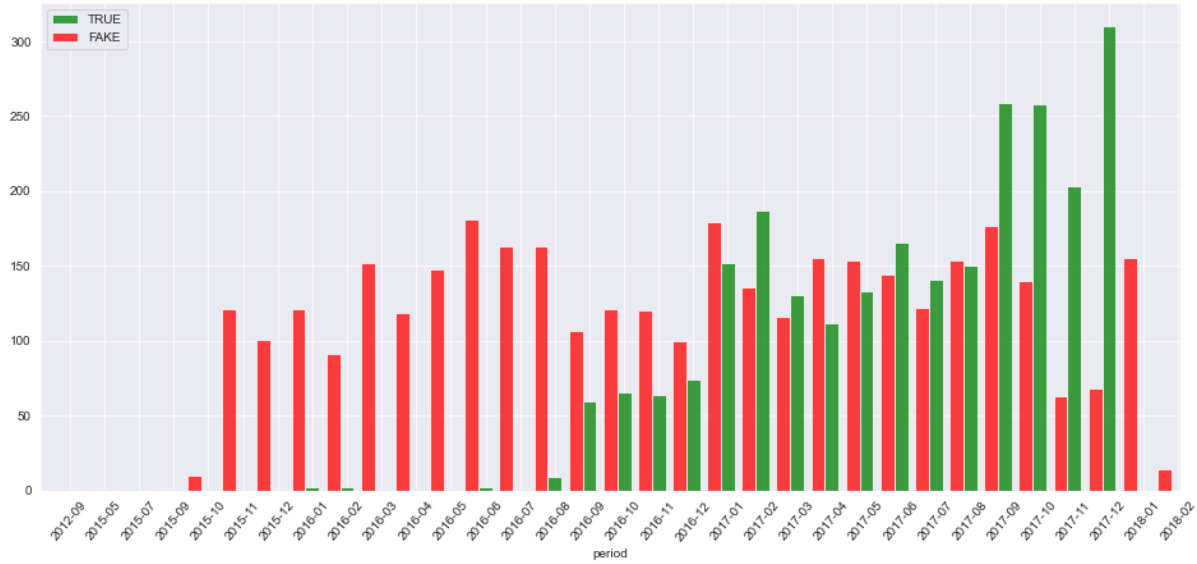
Obtenção e processamento de dados

Os dados foram obtidos a partir do Fake corpusBR.
Para o tratamento do conteúdo dos textos, eles foram processados, retirando acentos, pontuações, stopwords, tornando todos caracteres caixa baixa e também aplicando a normalização NFKD.
Com relação aos metadados, as datas eram escritas de vários formatos, o que necessitou de tratamento. Com relação ao número de substantivos, adjetivos, etc., estes estavam com um valor absoluto. Para análise, esses valores foram normalizados com relação ao número total de palavras, isto é, transformados em porcentagem.
Além disso, certas informações pré-processadas do corpus estavam erradas, o que necessitou que nós fizéssemos o pré-processamento.

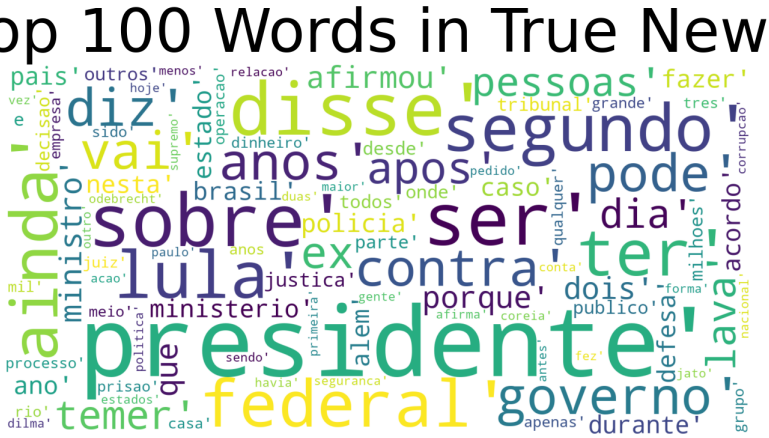
Cobertura e distribuição dos dados

Distribuição de fake news

A quantidade de notícias verdadeiras é falsa é a mesma, 3600, totalizando um total de 7200 notícias

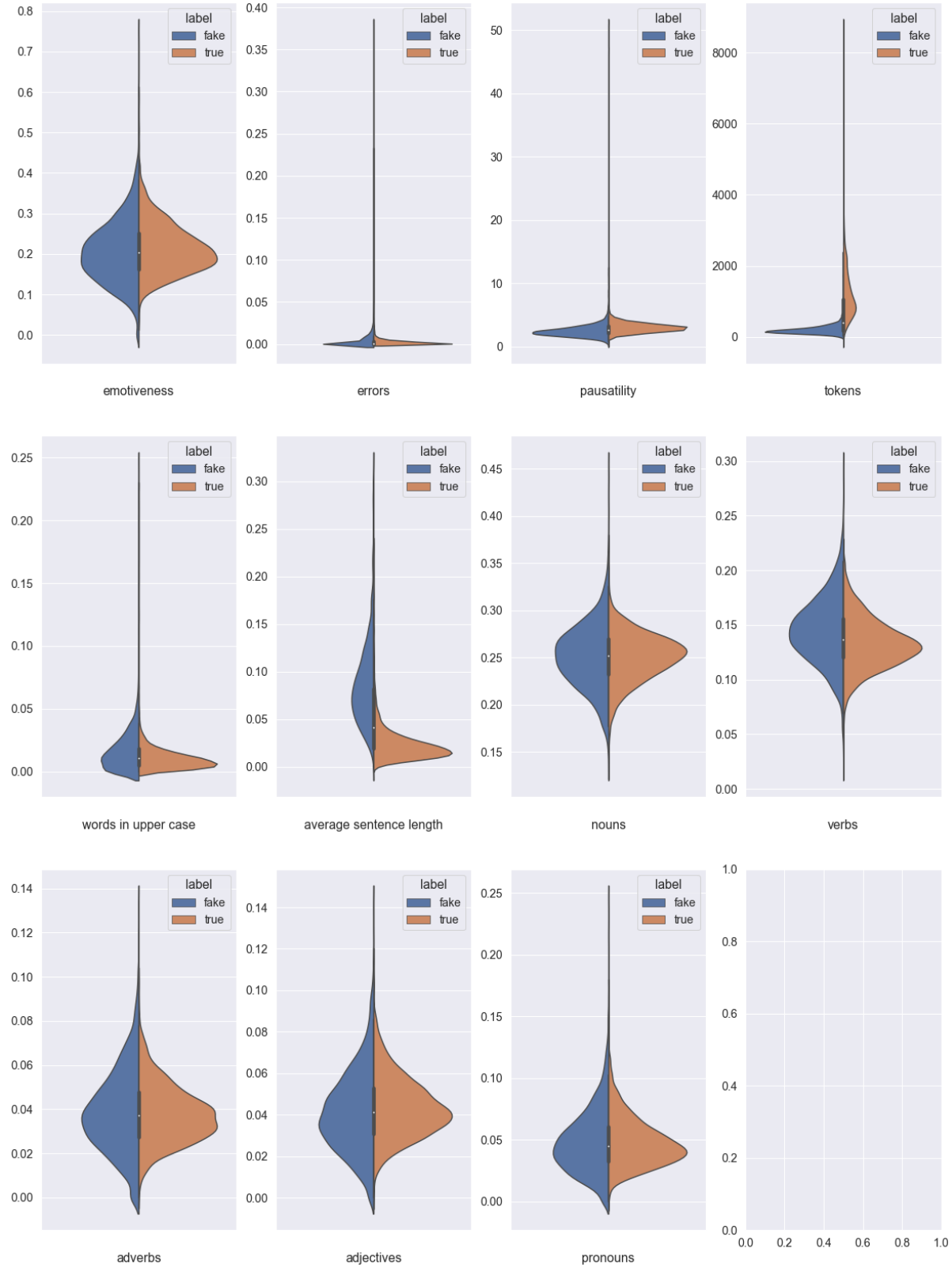


O intervalo coberto pelas notícias do corpus vai de 09/2015 até 02/2018.
É possível perceber que as notícias falsas tem uma distribuição relativamente uniforme com relação a todo esse período. Já as notícias verdadeiras são em maior parte do ano de 2017, principalmente no final deste ano.



Acima é possível ver também uma distribuição das palavras mais frequentes em cada tipo de notícia, percebe-se que o conteúdo mais comum das notícias são relacionados à política, com “Dilma” aparecendo com mais frequência em fake news quando comparado às notícias verdadeiras.

Características do corpus



Para visualização dos metadados, foram utilizados gráficos violinos para cada categoria, notícias verdadeiras e falsas.

As maiores diferenças nas características estão presentes no tamanho médio das sentenças (average sentence length) e no número de palavras (tokens). É possível perceber que há uma grande concentração de notícias falsas com poucas palavras, enquanto notícias verdadeiras são muito melhor distribuídas. O contrário ocorre para o tamanho médio das sentenças, enquanto as notícias falsas tem uma distribuição mais suave, as verdadeiras tem uma grande concentração de frases curtas.

Sites

link	label	
diariodobrasil.org	fake	3337
g1.globo.com	true	2300
politica.estadao.com.br	true	753
afolhabrasil.com.br	fake	174
internacional.estadao.com.br	true	114
cultura.estadao.com.br	true	95
www1.folha.uol.com.br	true	91
thejornalbrasil.com.br	fake	66
economia.estadao.com.br	true	51
brasil.estadao.com.br	true	46
esportes.estadao.com.br	true	30
alias.estadao.com.br	true	29
ceticismopolitico.com	fake	16
sao-paulo.estadao.com.br	true	16
saude.estadao.com.br	true	15
sustentabilidade.estadao.com.br	true	13
opiniao.estadao.com.br	true	11
topfivetv.com	fake	7
estadao.com.br	true	7
educacao.estadao.com.br	true	7
ciencia.estadao.com.br	true	5
viagem.estadao.com.br	true	5
emais.estadao.com.br	true	3
link.estadao.com.br	true	2
f5.folha.uol.com.br	true	2
territorioeldorado.limao.com.br	true	2
datafolha.folha.uol.com.br	true	1
blogdofred.blogfolha.uol.com.br	true	1
acervo.estadao.com.br	true	1

É possível ver que a maior parte das notícias vem de sites como diariodobrasil.org, g1.globo e estadao.com.br.

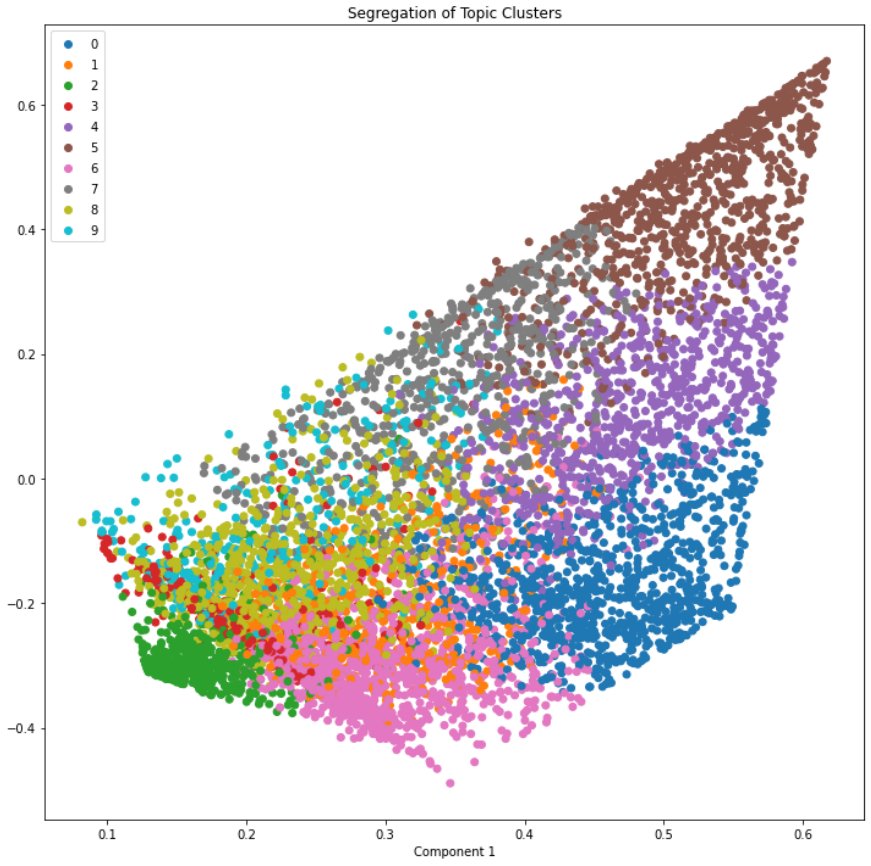
Dentre os sites utilizados, pode-se perceber que apenas os sites diariodobrasil.org, afolhabrasil.com.br, thejornalbrasil.com.br, ceticismopolitico.com.br e topfivetv.com.br foram utilizados como fontes de Fakenews.

Nenhum site teve disseminação de ambas as notícias (verdadeiras e falsas), ou um, ou outro.

Análise Exploratória

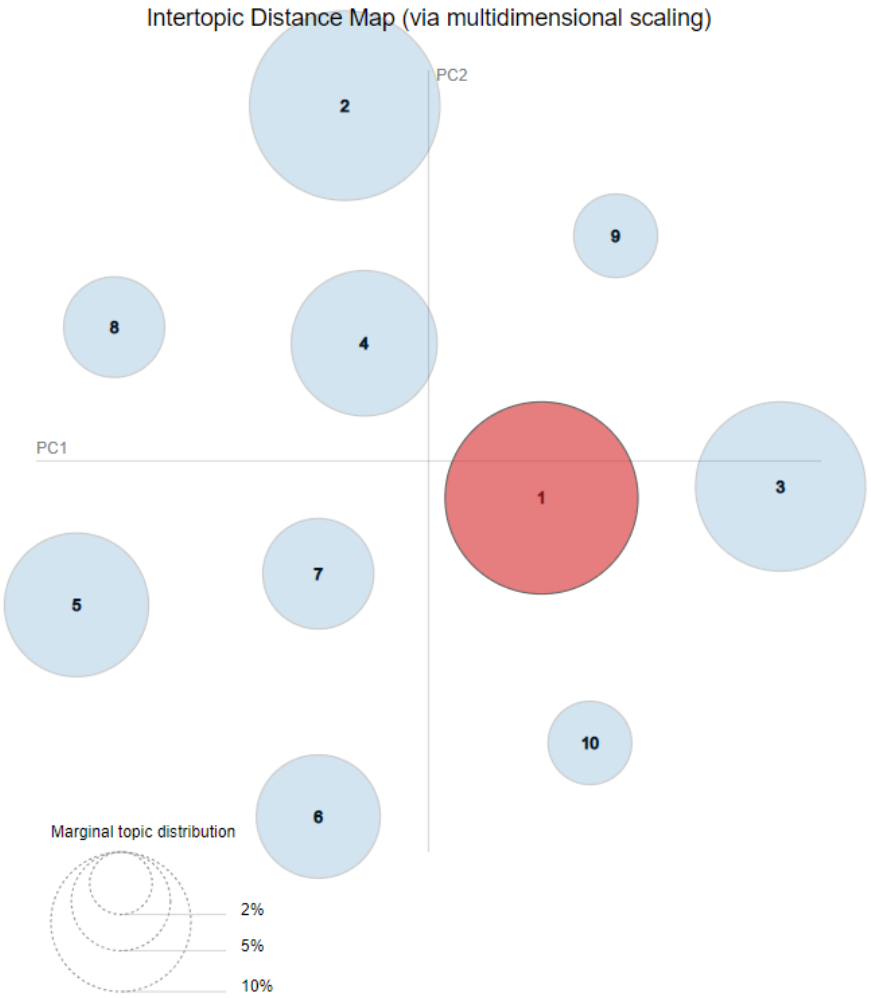
	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 0	pessoa	pai	politicar	vidar	querer	passar	formar	brasileiro	mundo	ficar	direito
Topic 1	militar	casar	pessoa	publicar	ficar	pai	policiar	gente	acontecer	video	passar
Topic 2	cidade	policiar	rir	janeiro	pessoa	saude	morto	governar	escola	morte	prefeitura
Topic 3	presidente	temer	deputar	governar	ministrar	camara	senador	pmdb	senado	politicar	afirmar
Topic 4	aumentar	mercar	empresar	produto	resultar	aguar	saude	brasileiro	medir	programar	analisar
Topic 5	falar	globo	jornalista	musicar	artista	filmar	cantor	show	amigo	tv	cinema
Topic 6	coreia	norte	trump	americano	nuclear	pai	armar	guerra	presidente	coreano	sul
Topic 7	empresar	milhoes	dinheiro	odebrecht	contar	receber	pagamento	propinar	campanha	pagar	comprar
Topic 8	policiar	apo	preso	hospital	familia	medicar	informar	filho	mulher	feirar	vitimar
Topic 9	presidente	federal	lavar	jato	pedir	defeso	juiz	publicar	advogar	processar	prisao

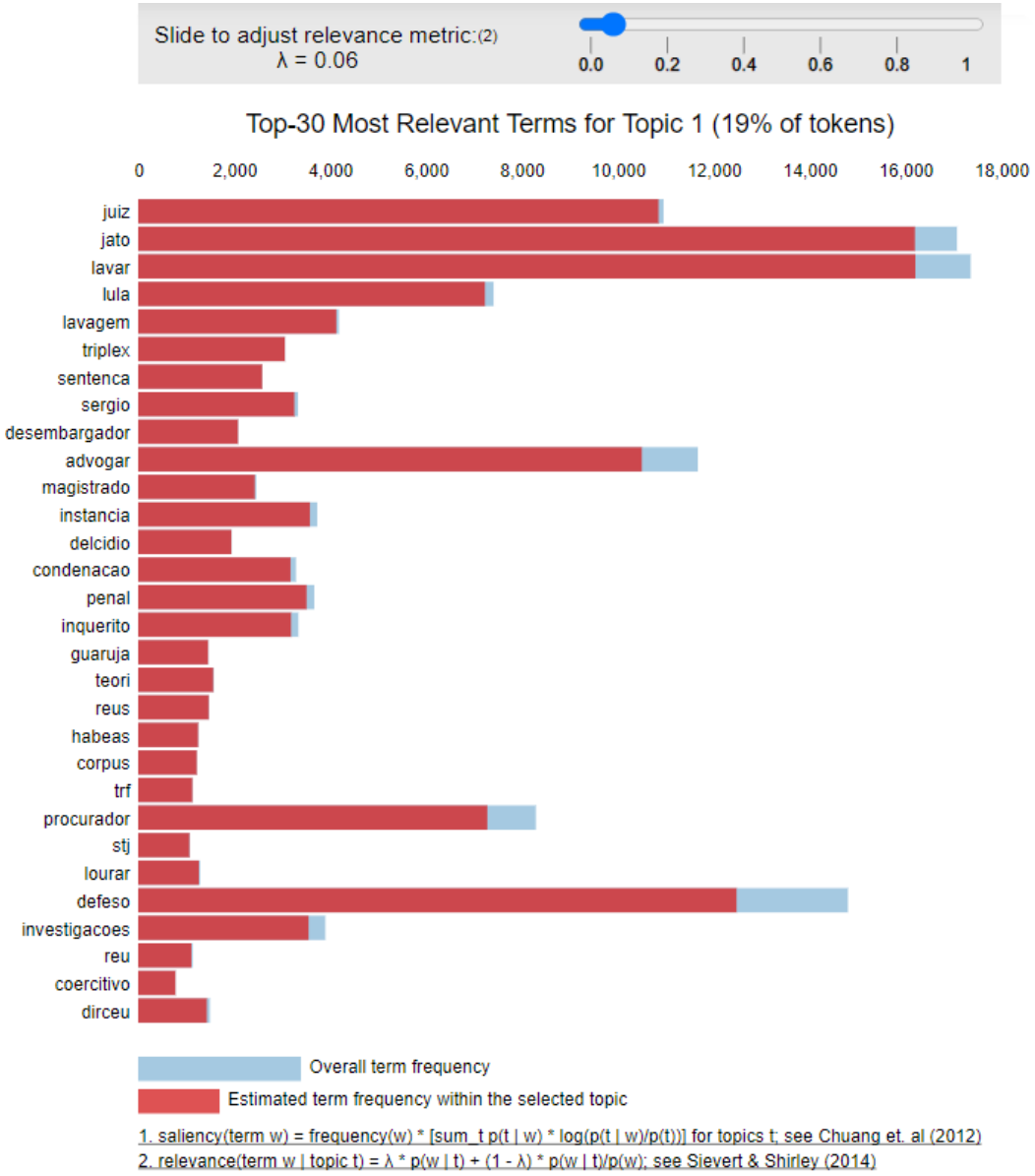
Os textos pós-tratados foram utilizados no algoritimo de LDA (Latent Dirichlet Allocation), a fim de descobrir os tópicos mais recorrentes considerando todos os documentos existentes. Acima, estão as palavras que mais representam cada um dos tópicos descobertos, diminuindo da esquerda para a direita a importância dela para o tópico.



Os textos então foram classificados de acordo com suas distâncias para cada um dos tópicos descobertos pelo LDA, atribuindo assim um tópico para cada notícia. A imagem acima mostra a clusterização dos diversos documentos para cada tópico. Foi então realizada a redução de dimensionalidade capturando a maior quantidade de informação das diversas dimensões e resumindo em duas, assim tornando possível realizar esta visualização da distribuição dos documentos em clusters de tópicos. Idealmente, um tópico ótimo possuiria uma distribuição sem sobreposição de outros.

Alternativamente foi possível visualizar as informações acima utilizando-se da ferramenta interativa pyLDAvis. No caso abaixo, foi selecionado o tópico 1, cujo possui uma grande quantidade de documentos classificados neste grupo e escolhido uma relevância de 0.06 dando uma maior importância a termos que aparecem apenas neste documento, e menos a termos mais comuns presentes em vários documentos. No tópico exemplo escolhido, é possível observar a grande quantidade de palavras relacionadas à operação lava jato no caso do triplex do ex-presidente Lula, estes termos, basicamente não encontrados quando se olha os demais tópicos.





Perguntas de pesquisa e explorações iniciais

Relembrando os objetivos definidos inicialmente:

- Obtenção e curadoria de dados sobre fake news a partir do corpus Fake.br
- Clusterização das notícias (falsas e verdadeiras) em temas
- Identificação de entidades presentes nas notícias
- Criar modelo para análise temporal dos dados obtidos
- Relacionar resultados obtidos com eventos ocorridos

A obtenção e curadoria dos dados pode ser considerada um sucesso. A clusterização dos dados também foi feita a partir da aplicação do LDA no corpus, a qual também ajudará nos últimos dois pontos.

Discussão e próximos passos

Os resultados obtidos da análise dos dados são considerados satisfatórios, elucidando características que devemos considerar nas próximas etapas, como a distribuição temporal e parâmetros importantes para notícias verdadeiras e falsas, tópicos chave de cada notícia e principais fontes de Fake news.

Acreditamos que com os resultados obtidos nessa análise, temos as informações necessárias para dar continuidade nos próximos passos de maneira mais focada e eficiente a fim de atingir os objetivos definidos e responder às perguntas propostas aplicando outros modelos de processamento de linguagem pertinentes quando necessários para obter o resultado desejado.