CIÊNCIA DE DADOS (BIG DATA PROCESSING AND ANALYTICS)

RECUPERAÇÃO DA INFORMAÇÃO NA WEB E EM REDES SOCIAIS

Professor curador
Luciano Moreira Camilo e Silva

TRILHA 5
PROCESSAMENTO DE LINGUAGEM NATURAL
IDENTIFICAÇÃO DE PADRÕES DE TEXTO EM DOCUMENTOS

Processamento de Linguagem Natural (PLN) – Palavras

Palavra

•

Morfologia

por exemplo: radical, lema, afixos

•

Flexões

por exemplo: gênero, número, grau

•

Combinações

por exemplo: palavras compostos

Wikcionário em outras línguas

Mais de 1.000.000 de entradas

inglês (*English*) • francês (*français*) • malgaxe (*malagasy*)

Mais de 500.000 entradas

alemão (*Deutsch*) • chinês (中文) • espanhol (*español*) • holandês (*nederlands*) • lituano (*lietuvių*) • polonês / polaco (*polski*) • russo (*русский*) • servo-croata (*srpskohrvatski / српскохрватски*) • sueco (*svenska*)

Mais de 200.000 entradas

canarês (*ಕನ್ನಡ*) • catalão (*català*) • coreano (*한국어*) • curdo (*kurdî / ‌*) • (*کوردی*)finlandês (*suomi*) • grego (*ελληνικά*) • húngaro (*magyar*) • ido • italiano • português • tâmil (*தமிழ்*) • turco (*türkçe*) • vietnamita (*tiếng việt*)

Fonte: Wikcionário – CC BY-SA 3.0

Pré-tratamento – Lematização, steming e outras reduções

- **Lematização** consiste em reduzir a palavra para sua ideia principal.
  - Conhecimento de muitas palavras.
  - Complexo e demorado.

- Martin Porter publicou algoritmo de ***steming*** em 1980.
  - Baseado em regras.
  - Rápido e com bons resultados.

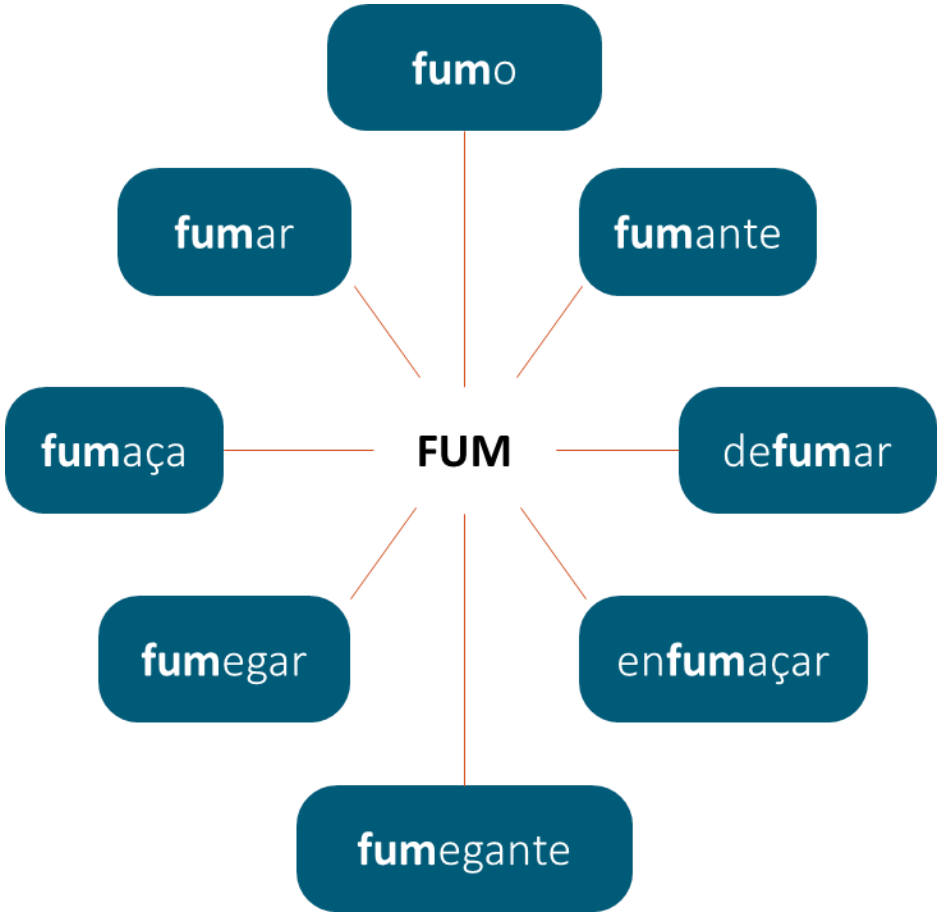


Ilustração do professor, baseada na apostila português IV da fundação CECIERJ.

Saco de Palavras

1. João adora as aulas de recuperação de informação. Maria adora também.

2. Maria também adora as aulas de *machine learning*.

Saco de Palavras

1. João adora as aulas de recuperação de informação. Maria adora também.

2. Maria também adora as aulas de *machine learning*.

	joao	adora	aulas	recuperacao	informacao	maria	tambem	*machine*	*learning*
1									
2									

Saco de Palavras

1. **João adora as aulas de recuperação de informação. Maria adora também.**

{joao: 1, adora: 2, aulas:1, recuperacao: 1, informacao: 1, maria: 1, tambem: 1, machine: 0, learning: 0}

2. **Maria também adora as aulas de *machine learning*.**

	joao	adora	aulas	recuperacao	informacao	maria	tambem	*machine*	*learning*
1	1	2	1	1	1	1	1	0	0
2									

Saco de Palavras

1. **João adora as aulas de recuperação de informação. Maria adora também.**

{joao: 1, adora: 2, aulas:1, recuperacao: 1, informacao: 1, maria: 1, tambem: 1, machine: 0, learning: 0}

2. **Maria também adora as aulas de *machine learning*.**

{joao: 0, adora: 1, aulas:0, recuperacao: 0, informacao: 0, maria: 1, tambem: 1, machine: 1, learning: 1}

	joao	adora	aulas	recuperacao	informacao	maria	tambem	*machine*	*learning*
1	1	2	1	1	1	1	1	0	0
2	0	1	0	0	0	1	1	1	1
T	1	3	1	1	1	2	2	1	1

TF-IDF

- Contagem de palavras ignora **tamanho** do corpus e relevância da palavra no **contexto** (corpora).

- ***Text Frequency (TF)*** mensura a importância relativa da palavra em um documento.

$$tf(t_n, d) = \frac{|\{t_n \in d\}|}{|\{t \in d\}|}$$

- ***Inverse Document Frequency (IDF)*** mensura a importância relativa da palavra em um documento.

$$idf(t_n, D) = log \frac{|\{d \in D\}|}{|\{d \in D : t_n \in d\}|}$$

Lorem Ipsum

"Neque porro quisquam est qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit..."

"Não há quem goste de dor, que a procure e a queira ter, simplesmente porque é dor..."

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec ipsum magna, efficitur vitae scelerisque at, finibus ut metus. Nulla fermentum erat felis, in condimentum purus varius vel. Quisque fringilla, mauris in pretium finibus, ex nisl lobortis velit, vel molestie velit diam ac turpis. Nunc eu sodales dolor. Maecenas vel enim imperdiet, malesuada massa et, placerat ex. In et sapien sed sapien pharetra tempor. Integer commodo lacus eget nibh ultrices vehicula. Mauris nec scelerisque tellus. Vivamus et urna venenatis, lacinia turpis in, blandit turpis. Ut interdum luctus vestibulum. Nulla ac porta nibh. Cras faucibus pellentesque quam a scelerisque.

Duis mattis eu tellus sit amet vehicula. Praesent pretium magna id nisl consectetur scelerisque. Cras volutpat consectetur mi eu commodo. Interdum et malesuada fames ac ante ipsum primis in faucibus. Maecenas efficitur eu nibh ac tristique. Sed efficitur posuere feugiat. Morbi id auctor sapien, a molestie velit. Aliquam vel turpis egestas, fringilla lacus quis, ultrices nibh. Morbi eget mauris semper metus vehicula rhoncus sit amet et massa. Donec eu magna lectus.

Donec suscipit fermentum tempor. Praesent vitae massa odio. Sed sed fermentum ex, in aliquam nibh. Nulla sed lectus lectus. Aliquam quis dictum mi. Etiam semper tellus a efficitur facilisis. Cras vitae nisl vitae est pharetra consectetur. Ut ornare quis metus sit amet rhoncus.

Quisque facilisis malesuada ornare. Donec sit amet vehicula magna. Donec id vulputate ligula. Nunc sollicitudin neque eu sem porttitor, quis porta libero mattis. Donec ultricies, dolor mollis ultrices convallis, sem turpis eleifend elit, eu viverra nibh massa vel odio. Sed non lobortis turpis. Quisque iaculis, ligula non elementum luctus, ligula purus luctus tortor, sed congue felis metus ac ipsum. Sed efficitur cursus mi vitae condimentum. Nullam sit amet arcu hendrerit, sagittis libero ac, iaculis orci. Ut hendrerit dolor ac posuere pulvinar. Phasellus vel fermentum leo, a eleifend dolor. Morbi id augue sit amet tortor mollis sagittis sit amet sed sem.

Vivamus sollicitudin elit lorem, eget ultricies odio pretium a. Sed ornare, tellus ultrices iaculis maximus, nisi erat eleifend diam, sed dictum est libero vitae purus. Vivamus metus nisi, consectetur vel tincidunt convallis, semper eu libero. Etiam nec pellentesque sem. Suspendisse scelerisque lacinia dui in varius. Aliquam turpis felis, aliquet et tincidunt sit amet, tincidunt ac elit. Integer nec tristique nibh. Nunc fringilla nisl enim, eget faucibus nisi pellentesque nec. In eget semper tortor. Mauris gravida dui in ante efficitur dictum id ac velit. Sed aliquam mollis risus eget efficitur. Vivamus leo ligula, consequat at ullamcorper at, ultrices ut nibh.

Interdum et malesuada fames ac ante ipsum primis in faucibus. Nulla dictum bibendum felis eget semper. Sed ipsum ante, tempor volutpat interdum faucibus, aliquet scelerisque neque. Quisque vehicula, nunc a consectetur facilisis, massa metus maximus metus, non consequat felis nibh dapibus nunc. Suspendisse euismod risus eget justo lobortis, id ultricies turpis posuere. Suspendisse faucibus est vel gravida posuere. Sed vitae fermentum est, sed maximus risus. Etiam suscipit, enim ac porta ultricies, lectus mauris facilisis purus, eget imperdiet felis purus eget metus. Curabitur quis augue odio. Sed pulvinar, odio a consectetur cursus, metus augue aliquam turpis, quis posuere leo nisl sed turpis. Duis iaculis lacus vel turpis varius, sit amet cursus libero bibendum. Vestibulum ultricies est sit amet cursus facilisis. Proin auctor purus a convallis blandit. Nulla hendrerit tincidunt ipsum ac vestibulum. Nulla ut tempor tellus, at rutrum nunc. Nunc quis magna dolor.

Ut vulputate nisi in turpis egestas consequat. Donec at iaculis dui. Proin lacinia nibh ac auctor posuere. Nullam ultricies, dui ac imperdiet mollis, purus felis aliquam nunc, a pellentesque libero dolor ac nibh. Nullam ac libero lacinia risus faucibus mollis. Praesent at felis sodales, vulputate mauris et, viverra est. Nulla mattis ut augue vitae porttitor. Donec consequat dolor et nibh efficitur, nec gravida dui tincidunt. Sed malesuada dui eu est sagittis eleifend. Proin porttitor massa et porta laoreet. Phasellus ut congue orci. Donec aliquam metus sollicitudin nisi interdum convallis. Pellentesque in eros sagittis, fringilla diam tincidunt, rhoncus libero. Nullam ac lacus enim. Cras sagittis purus eget quam imperdiet, vitae porttitor leo fermentum. Curabitur sed ornare nisi.

Suspendisse ex ante, consequat sit amet sem non, bibendum mollis nisl. Aliquam elementum velit tortor, vitae tincidunt mi interdum ac. In tristique arcu id nulla porta, eget gravida purus vestibulum. Phasellus vehicula aliquam bibendum. Sed ac risus placerat, molestie odio egestas, convallis mi. Fusce rhoncus, justo at malesuada scelerisque, nunc mauris fringilla metus, in feugiat nisi felis ac ex. Nulla scelerisque elit eget ipsum facilisis, nec condimentum ipsum feugiat. Maecenas dignissim risus arcu, nec imperdiet elit sodales non.

Fonte: <https://pt.lipsum.com>.

TF-IDF

1. João adora as aulas de recuperação de informação. Maria adora também.
2. Maria também adora as aulas de *machine learning*.

BoW

	1	2
joao	1	0
adora	2	1
aulas	1	0
recuperacao	1	0
informacao	1	0
maria	1	1
tambem	1	1
machine	0	1
learning	0	1

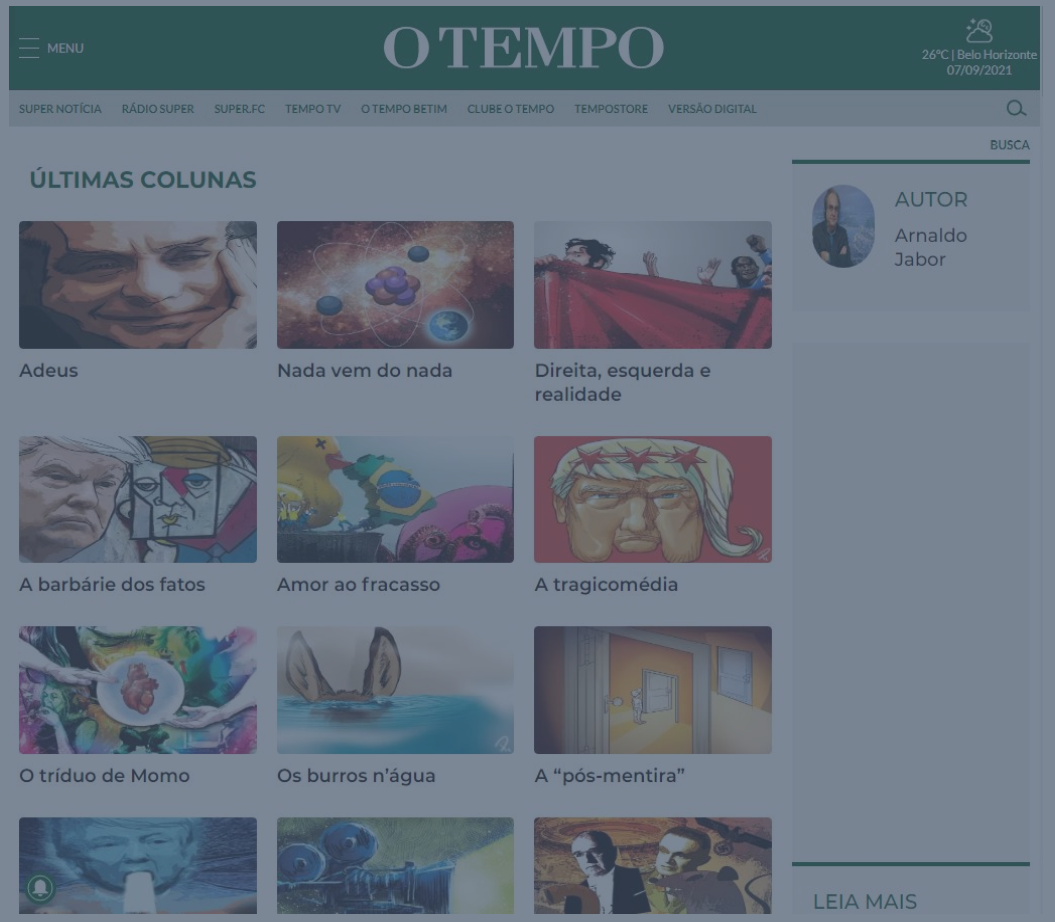
Exercício Prático

TRILHA 3
Recuperação de informação por raspagem – robôs

Arnaldo Jabor é um cineasta, roteirista, diretor de cinema e TV, produtor cinematográfico, dramaturgo, crítico, jornalista e escritor brasileiro.

Durante 26 anos escreveu mais de 1500 artigos para jornais e revistas. E teve seu nome associado a muito mais textos vistos na internet!

O exercício consiste em puxar e formatar todos os artigos de Arnaldo Jabor, disponível no site do jornal O Tempo.



Fonte: https://www.otempo.com.br/opiniao/arnaldo-jabor

Textos atribuídos a Arnaldo Jabor

Apesar de ter mais estudo do que qualquer professor de humanas da geração Paulo Freire e mais inteligência emocional do que qualquer outro político brasileiro, não há polidez em suas palavras e tão pouco elegância em seu comportamento.

Trecho do texto atribuído a Arnaldo Jabor

Brasileiro é um povo alegre. Mentira. Brasileiro é bobalhão. Fazer piadinha com as imundices que acompanhamos todo dia é o mesmo que tomar bofetada na cara e dar risada. Depois de um massacre que durou quatro dias em São Paulo.

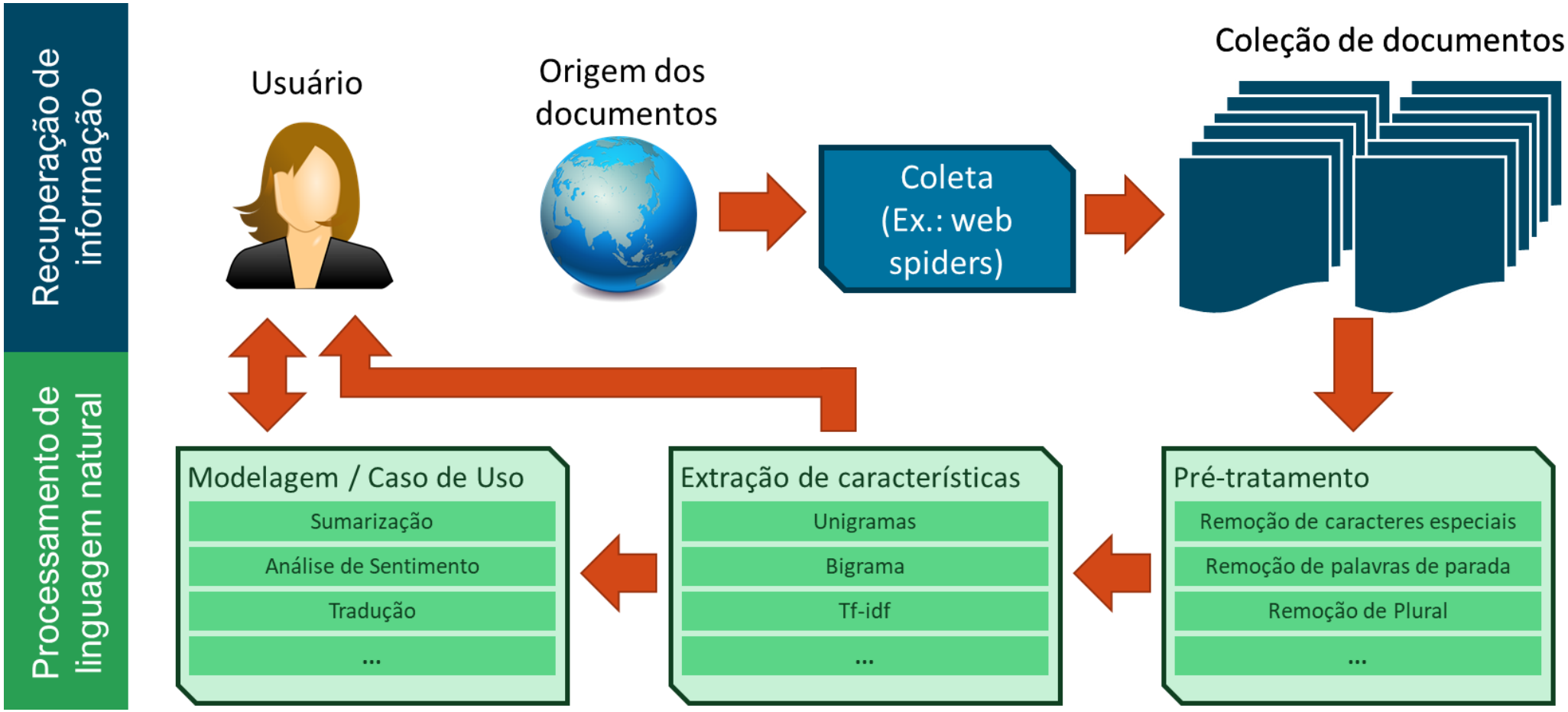
Trecho do texto atribuído a Arnaldo Jabor

Eu vi também a entrevista que a Marina Silva deu para outro grupo de jornalistas no dia seguinte. Meu Deus! Que coisa mais frágil! Que coisa mais pobre de ideias. Que coisa triste ver aquela boa mulher de ótimo caráter sem dúvida, mas sentadinha ali com sua vozinha falando em ética.

Trecho da transcrição da coluna à rádio CBN em 2 ago. 2018

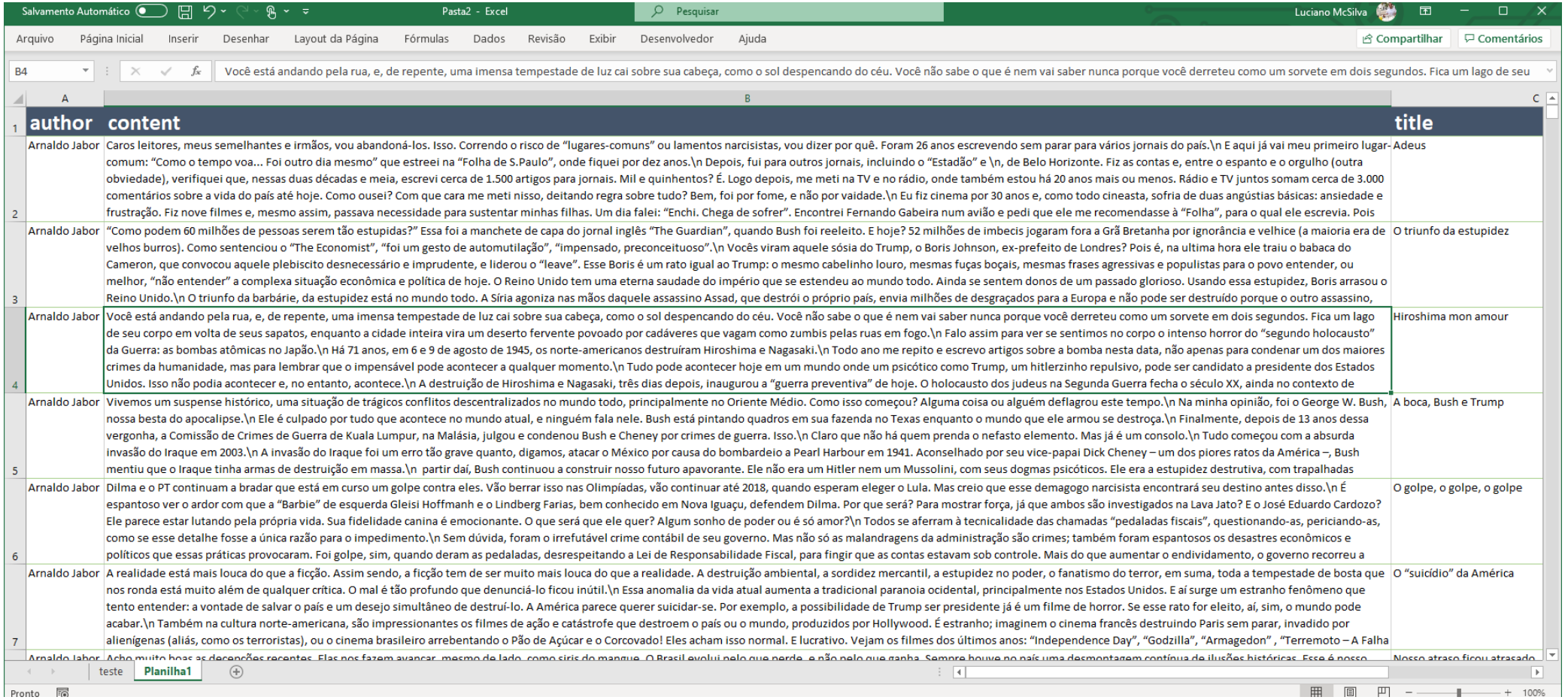
Fonte: Wikipédia

Fluxograma geral de PLN



Fonte: Elaborada pelo autor.

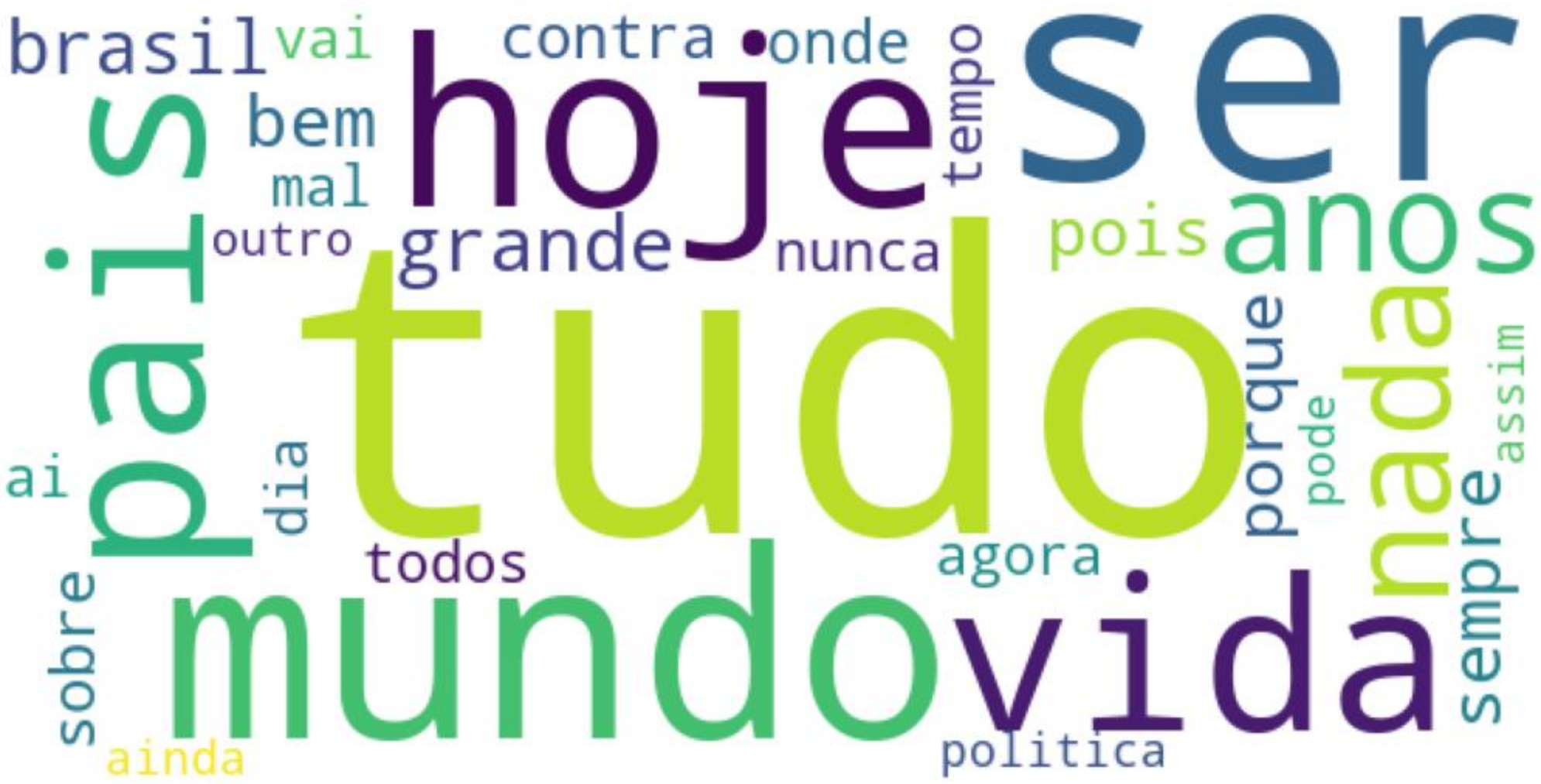
Arquivo com o conteúdo



	author	content	title
2	Arnaldo Jabor	Caros leitores, meus semelhantes e irmãos, vou abandoná-los. Isso. Correndo o risco de "lugares-comuns" ou lamentos narcisistas, vou dizer por quê. Foram 26 anos escrevendo sem parar para vários jornais do país.\n E aqui já vai meu primeiro lugar-comum: "Como o tempo voa... Foi outro dia mesmo" que estreei na "Folha de S.Paulo", onde fiquei por dez anos.\n Depois, fui para outros jornais, incluindo o "Estadão" e \n, de Belo Horizonte. Fiz as contas e, entre o espanto e o orgulho (outra obviedade), verifiquei que, nessas duas décadas e meia, escrevi cerca de 1.500 artigos para jornais. Mil e quinhentos? É. Logo depois, me meti na TV e no rádio, onde também estou há 20 anos mais ou menos. Rádio e TV juntos somam cerca de 3.000 comentários sobre a vida do país até hoje. Como ousei? Com que cara me meti nisso, deitando regra sobre tudo? Bem, foi por fome, e não por vaidade.\n Eu fiz cinema por 30 anos e, como todo cineasta, sofria de duas angústias básicas: ansiedade e frustração. Fiz nove filmes e, mesmo assim, passava necessidade para sustentar minhas filhas. Um dia falei: "Enchi. Chega de sofrer". Encontrei Fernando Gabeira num avião e pedi que ele me recomendasse à "Folha", para o qual ele escrevia. Pois	Adeus
3	Arnaldo Jabor	"Como podem 60 milhões de pessoas serem tão estupidas?" Essa foi a manchete de capa do jornal inglês "The Guardian", quando Bush foi reeleito. E hoje? 52 milhões de imbecis jogaram fora a Grã Bretanha por ignorância e velhice (a maioria era de velhos burros). Como sentenciou o "The Economist", "foi um gesto de automutilação", "impensado, preconceituoso".\n Vocês viram aquele sósia do Trump, o Boris Johnson, ex-prefeito de Londres? Pois é, na ultima hora ele traiu o babaca do Cameron, que convocou aquele plebiscito desnecessário e imprudente, e liderou o "leave". Esse Boris é um rato igual ao Trump: o mesmo cabelinho louro, mesmas fuças boçais, mesmas frases agressivas e populistas para o povo entender, ou melhor, "não entender" a complexa situação econômica e política de hoje. O Reino Unido tem uma eterna saudade do império que se estendeu ao mundo todo. Ainda se sentem donos de um passado glorioso. Usando essa estupidez, Boris arrasou o Reino Unido.\n O triunfo da barbárie, da estupidez está no mundo todo. A Síria agoniza nas mãos daquele assassino Assad, que destrói o próprio país, envia milhões de desgraçados para a Europa e não pode ser destruído porque o outro assassino,	O triunfo da estupidez
4	Arnaldo Jabor	Você está andando pela rua, e, de repente, uma imensa tempestade de luz cai sobre sua cabeça, como o sol despencando do céu. Você não sabe o que é nem vai saber nunca porque você derreteu como um sorvete em dois segundos. Fica um lago de seu corpo em volta de seus sapatos, enquanto a cidade inteira vira um deserto fervente povoado por cadáveres que vagam como zumbis pelas ruas em fogo.\n Falo assim para ver se sentimos no corpo o intenso horror do "segundo holocausto" da Guerra: as bombas atômicas no Japão.\n Há 71 anos, em 6 e 9 de agosto de 1945, os norte-americanos destruíram Hiroshima e Nagasaki.\n Todo ano me repito e escrevo artigos sobre a bomba nesta data, não apenas para condenar um dos maiores crimes da humanidade, mas para lembrar que o impensável pode acontecer a qualquer momento.\n Tudo pode acontecer hoje em um mundo onde um psicótico como Trump, um hitlerzinho repulsivo, pode ser candidato a presidente dos Estados Unidos. Isso não podia acontecer e, no entanto, acontece.\n A destruição de Hiroshima e Nagasaki, três dias depois, inaugurou a "guerra preventiva" de hoje. O holocausto dos judeus na Segunda Guerra fecha o século XX, ainda no contexto de	Hiroshima mon amour
5	Arnaldo Jabor	Vivemos um suspense histórico, uma situação de trágicos conflitos descentralizados no mundo todo, principalmente no Oriente Médio. Como isso começou? Alguma coisa ou alguém deflagrou este tempo.\n Na minha opinião, foi o George W. Bush, nossa besta do apocalipse.\n Ele é culpado por tudo que acontece no mundo atual, e ninguém fala nele. Bush está pintando quadros em sua fazenda no Texas enquanto o mundo que ele armou se destroça.\n Finalmente, depois de 13 anos dessa vergonha, a Comissão de Crimes de Guerra de Kuala Lumpur, na Malásia, julgou e condenou Bush e Cheney por crimes de guerra. Isso.\n Claro que não há quem prenda o nefasto elemento. Mas já é um consolo.\n Tudo começou com a absurda invasão do Iraque em 2003.\n A invasão do Iraque foi um erro tão grave quanto, digamos, atacar o México por causa do bombardeio a Pearl Harbour em 1941. Aconselhado por seu vice-papai Dick Cheney – um dos piores ratos da América –, Bush mentiu que o Iraque tinha armas de destruição em massa.\n partir daí, Bush continuou a construir nosso futuro apavorante. Ele não era um Hitler nem um Mussolini, com seus dogmas psicóticos. Ele era a estupidez destrutiva, com trapalhadas	A boca, Bush e Trump
6	Arnaldo Jabor	Dilma e o PT continuam a bradar que está em curso um golpe contra eles. Vão berrar isso nas Olimpíadas, vão continuar até 2018, quando esperam eleger o Lula. Mas creio que esse demagogo narcisista encontrará seu destino antes disso.\n É espantoso ver o ardor com que a "Barbie" de esquerda Gleisi Hoffmanh e o Lindberg Farias, bem conhecido em Nova Iguaçu, defendem Dilma. Por que será? Para mostrar força, já que ambos são investigados na Lava Jato? E o José Eduardo Cardozo? Ele parece estar lutando pela própria vida. Sua fidelidade canina é emocionante. O que será que ele quer? Algum sonho de poder ou é só amor?\n Todos se aferram à tecnicalidade das chamadas "pedaladas fiscais", questionando-as, periciando-as, como se esse detalhe fosse a única razão para o impedimento.\n Sem dúvida, foram o irrefutável crime contábil de seu governo. Mas não só as malandragens da administração são crimes; também foram espantosos os desastres econômicos e políticos que essas práticas provocaram. Foi golpe, sim, quando deram as pedaladas, desrespeitando a Lei de Responsabilidade Fiscal, para fingir que as contas estavam sob controle. Mais do que aumentar o endividamento, o governo recorreu a	O golpe, o golpe, o golpe
7	Arnaldo Jabor	A realidade está mais louca do que a ficção. Assim sendo, a ficção tem de ser muito mais louca do que a realidade. A destruição ambiental, a sordidez mercantil, a estupidez no poder, o fanatismo do terror, em suma, toda a tempestade de bosta que nos ronda está muito além de qualquer crítica. O mal é tão profundo que denunciá-lo ficou inútil.\n Essa anomalia da vida atual aumenta a tradicional paranoia ocidental, principalmente nos Estados Unidos. E aí surge um estranho fenômeno que tento entender: a vontade de salvar o país e um desejo simultâneo de destruí-lo. A América parece querer suicidar-se. Por exemplo, a possibilidade de Trump ser presidente já é um filme de horror. Se esse rato for eleito, aí, sim, o mundo pode acabar.\n Também na cultura norte-americana, são impressionantes os filmes de ação e catástrofe que destroem o país ou o mundo, produzidos por Hollywood. É estranho; imaginem o cinema francês destruindo Paris sem parar, invadido por alienígenas (aliás, como os terroristas), ou o cinema brasileiro arrebentando o Pão de Açúcar e o Corcovado! Eles acham isso normal. E lucrativo. Vejam os filmes dos últimos anos: "Independence Day", "Godzilla", "Armagedon" , "Terremoto – A Falha	O "suicídio" da América
	Arnaldo Jabor	Acho muito boas as decepções recentes. Elas nos fazem avançar, mesmo de lado, como siris do mangue. O Brasil evolui pelo que perde, e não pelo que ganha. Sempre houve no país uma desmontagem contínua de ilusões históricas. Esse é nosso	Nosso atraso ficou atrasado

Fonte: Elaborada pelo autor.

Nuvem de palavras Arnaldo Jabor (pós-tratamento)

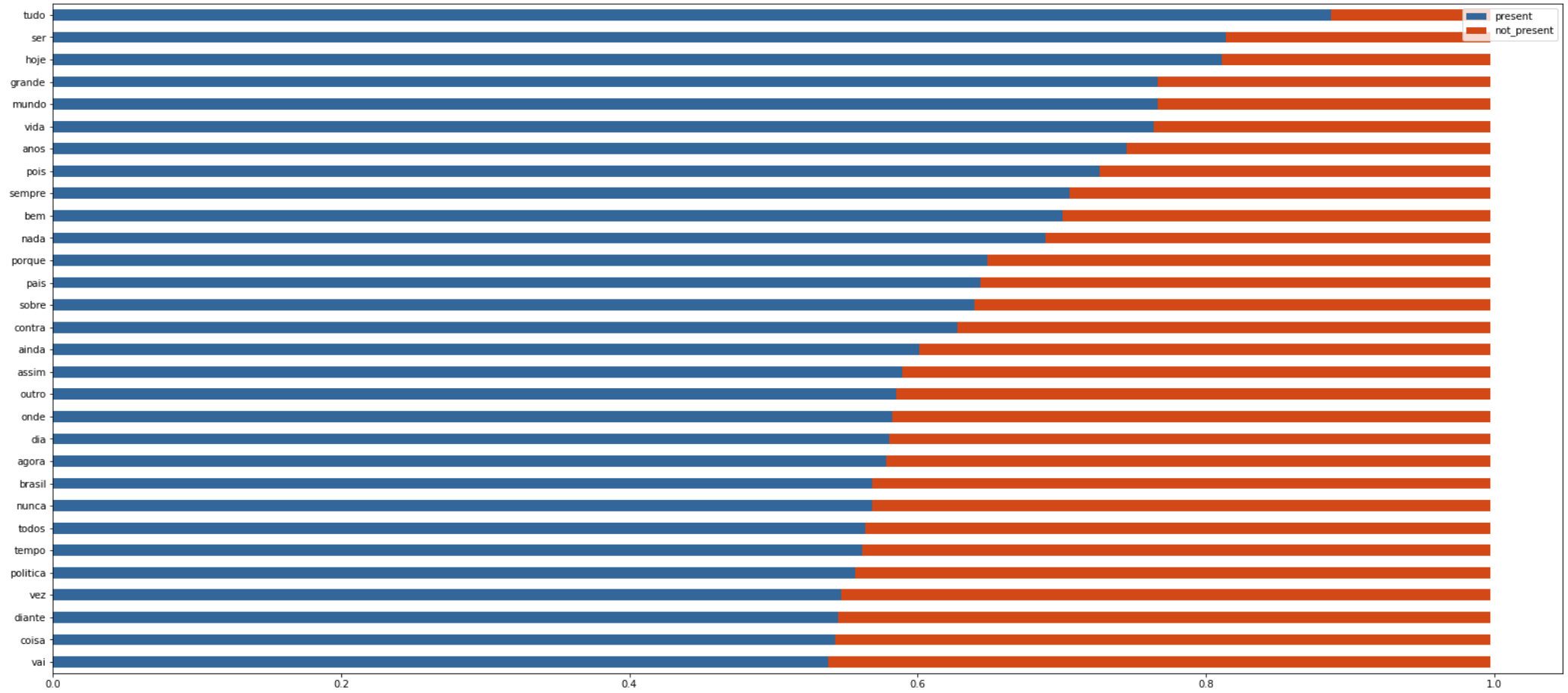


Saco de Palavras – Arnaldo Jabor

	tudo	ser	hoje	grande	brasileiro	mundo	vida	morte	morrer	otario	ahh	anos	nada	futuro
0	2	3	4	0	0	1	0	1	0	0	0	2	0	0
1	4	0	3	1	0	0	2	0	0	0	0	5	3	1
2	2	4	6	2	0	10	2	0	0	0	0	0	1	0
3	1	3	2	2	1	1	4	0	0	0	0	6	1	0
4	2	4	3	2	0	7	1	0	0	0	1	1	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
419	3	0	4	2	0	10	4	1	0	0	0	3	2	2
420	4	2	0	4	2	1	4	1	0	0	0	0	4	1
421	0	6	1	1	0	5	0	0	0	0	0	0	2	3
422	4	2	2	0	0	0	1	0	1	0	0	5	11	0
423	4	5	3	1	0	8	0	1	0	0	0	1	0	1

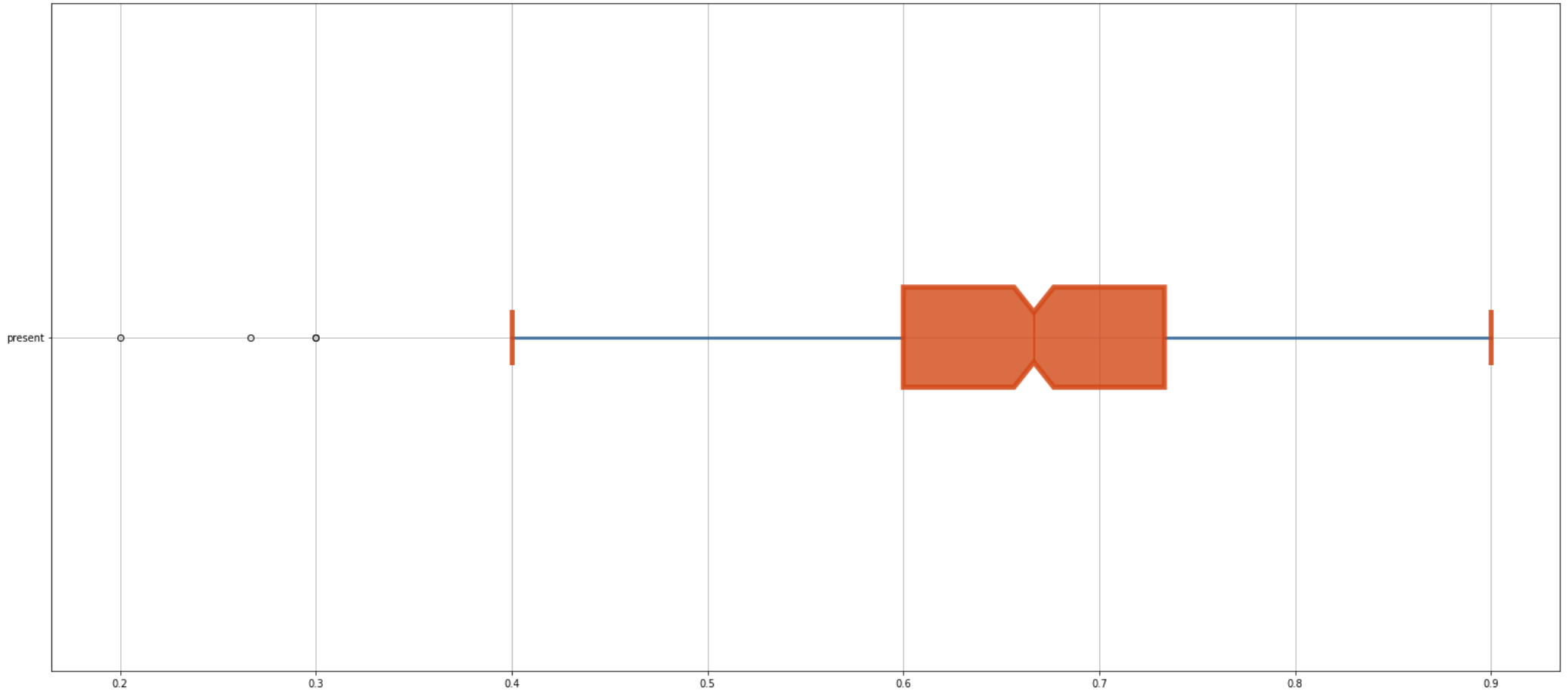
424 rows × 28080 columns

Frequência das 30 palavras mais utilizadas – Arnaldo Jabor



Fonte: Elaborada pelo autor.

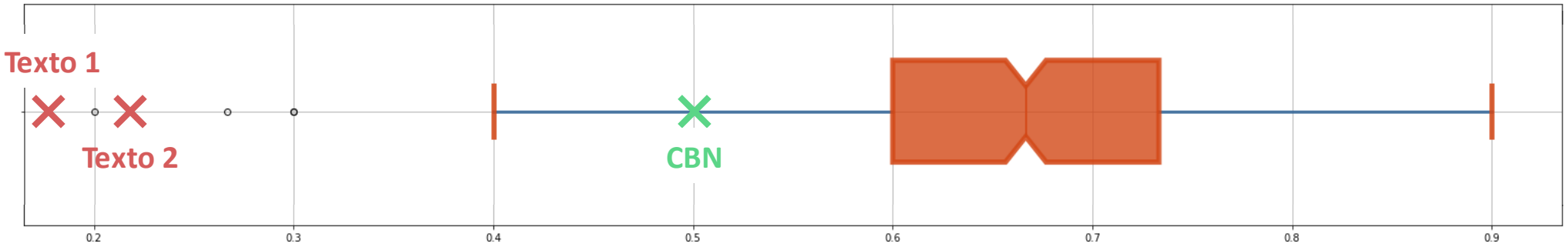
Presença das 30 palavras mais utilizadas – Arnaldo Jabor



Fonte: Elaborada pelo autor.

Presença das 30 palavras mais utilizadas – Arnaldo Jabor

	tudo	ser	hoje	grande	mundo	vida	anos	pois	sempre	bem	nada	porque	pais	sobre	contra	ainda	assim	outro	onde	dia	agora	brasil	nunca	todos	tempo	politica	vez	diante	coisa	vai	% uso
Texto 1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	**17%**
Texto 2	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0	2	0	**23%**
CBN	0	1	0	0	1	0	0	2	0	1	1	0	2	3	1	0	0	1	0	1	0	4	0	1	1	0	0	0	3	2	**50%**
Esperado	89%	81%	81%	77%	77%	76%	75%	73%	71%	70%	69%	65%	64%	64%	63%	60%	59%	58%	58%	58%	58%	57%	57%	56%	56%	56%	55%	54%	54%	54%	**62%**



Fonte: Elaborada pelo autor.

Universidade Presbiteriana
Mackenzie