CIÊNCIA DE DADOS (BIG DATA PROCESSING AND ANALYTICS)

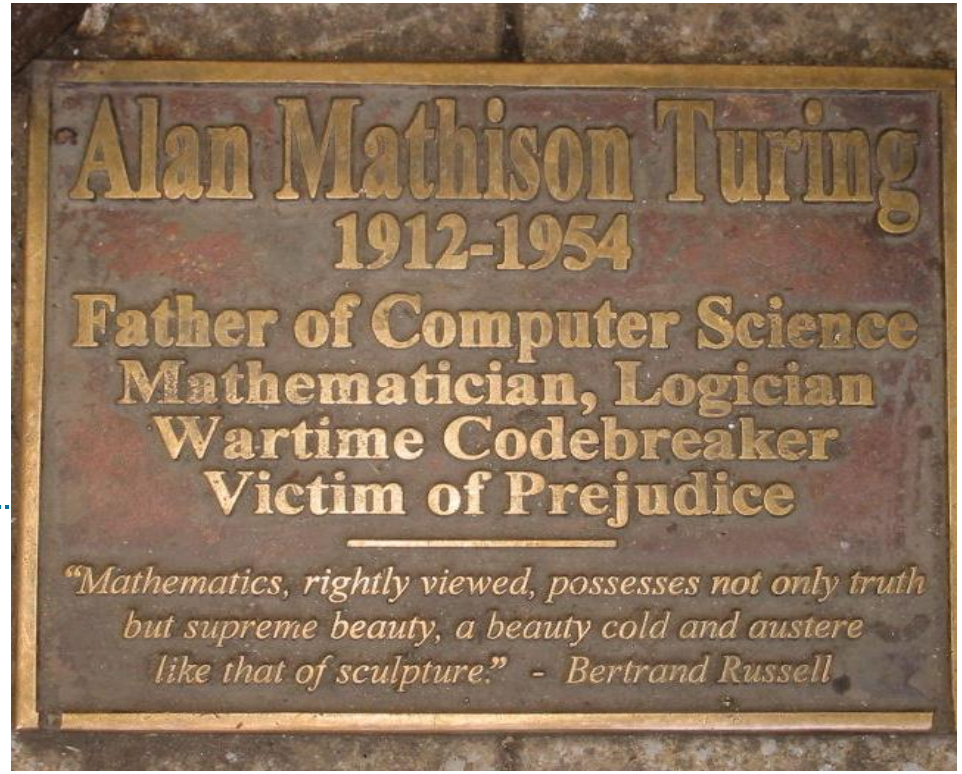
RECUPERAÇÃO DA INFORMAÇÃO NA WEB E EM REDES SOCIAIS

Professor curador
Luciano Moreira Camilo e Silva

TRILHA 4
PROCESSAMENTO DE LINGUAGEM NATURAL: EXTRAÇÃO DE PALAVRAS-CHAVE

Processamento de Linguagem Natural (PLN)

- Área da **ciência da computação**, em particular da **inteligência artificial**.
- Habilitar os computadores a compreender a **linguagem natural** dos humanos.
- Combinar **linguísticas** com regras, **estatísticas** e modelos de **aprendizado de máquinas**
- Em 1950, **Alan Turing** publicou as bases para avaliação da linguagem natural em seu artigo *"Computing Machinery and Intelligence".*
- **Rápida evolução a partir da década de 1980.** Progressos recentes culminaram na criação de assistentes virtuais como Waton, Siri e Alexa.



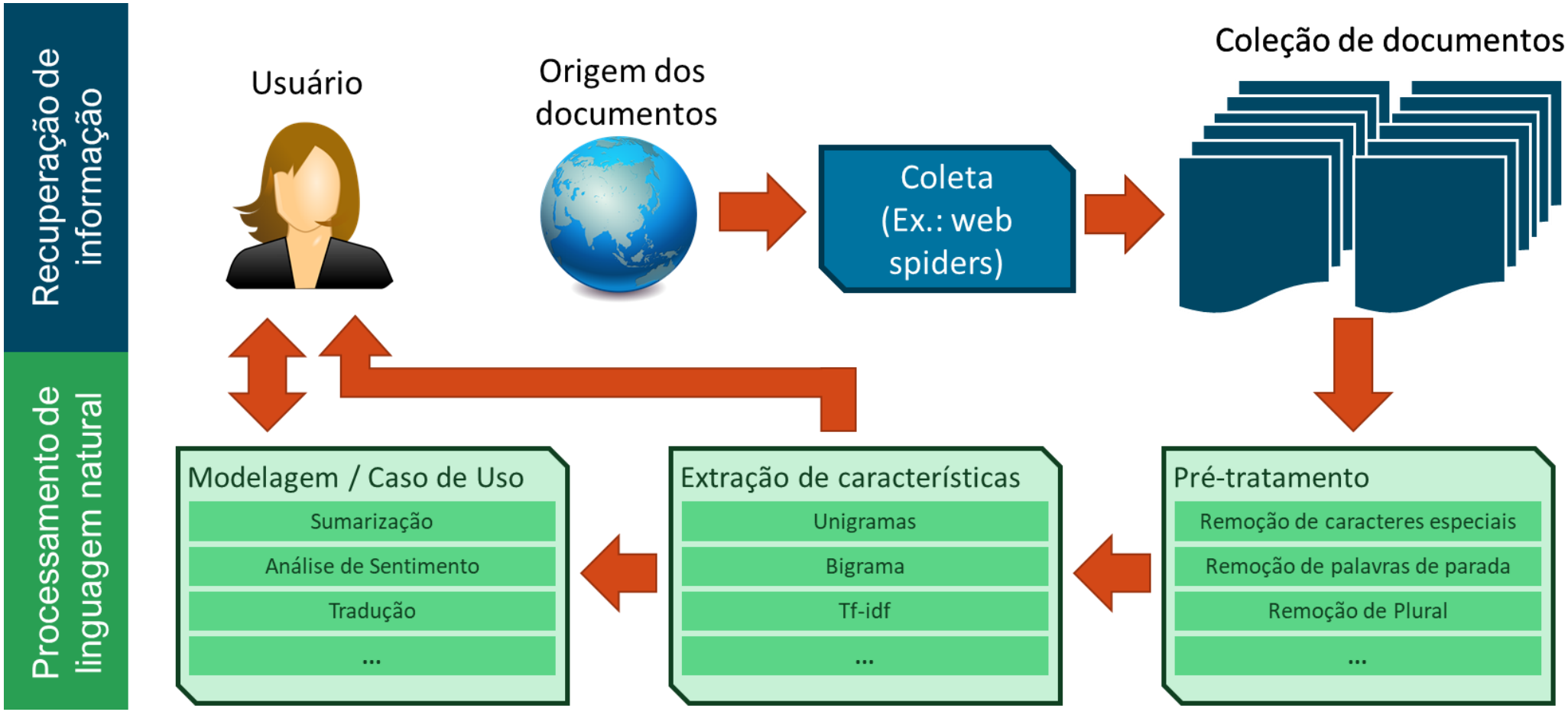
Fonte: Wikipédia.

Casos de uso em Processamento de Linguagem Natural (PLN)

- Auto completar de formulários
- Avaliação de CV (Recrutamento)
- Sumarização de textos
- Correção de textos
- Traduções
- Chatbots
- Análise de sentimentos
- **Detecção de SPAM**
- Buscas de documentos
- Geração de textos

Fonte: GettyImages.

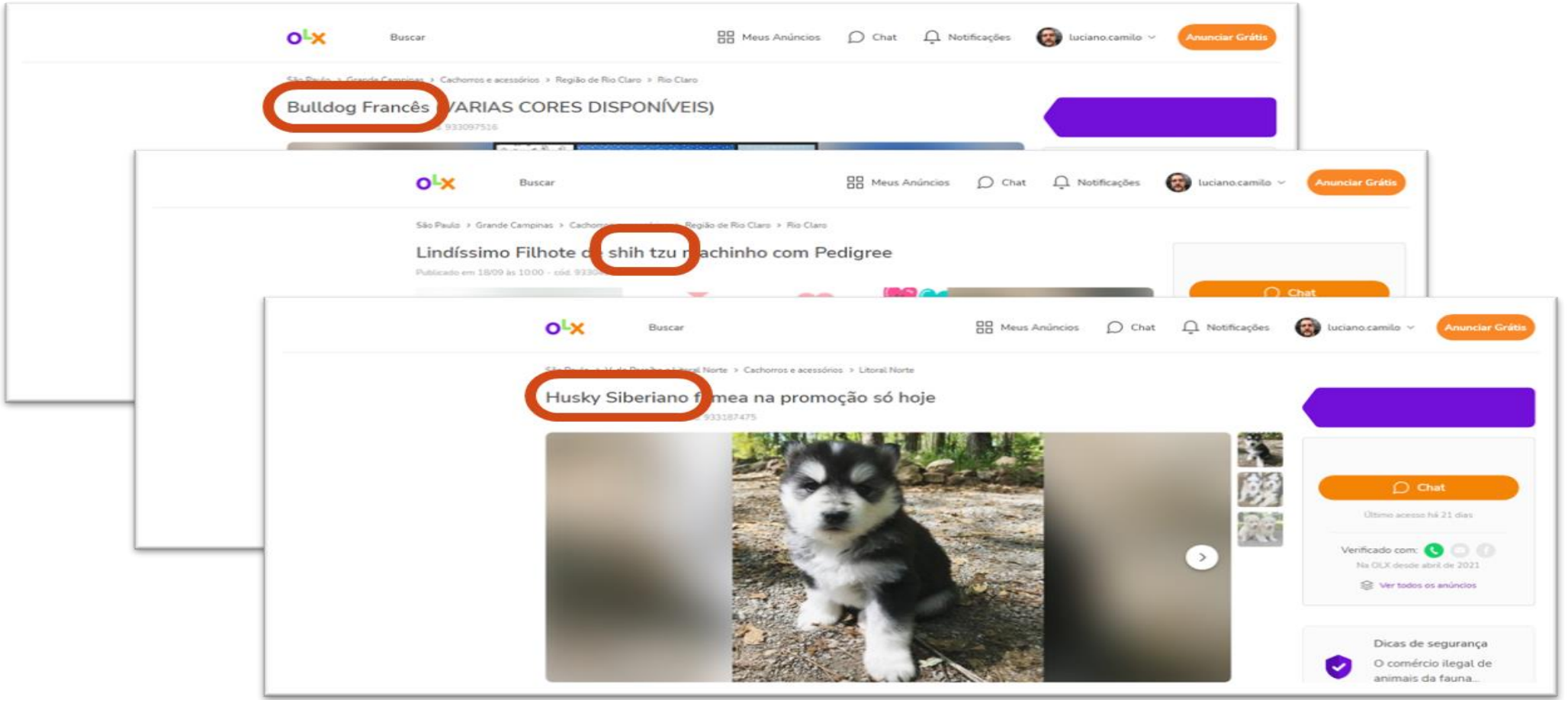
Fluxograma geral de PLN



Fonte: Elaborada pelo autor.

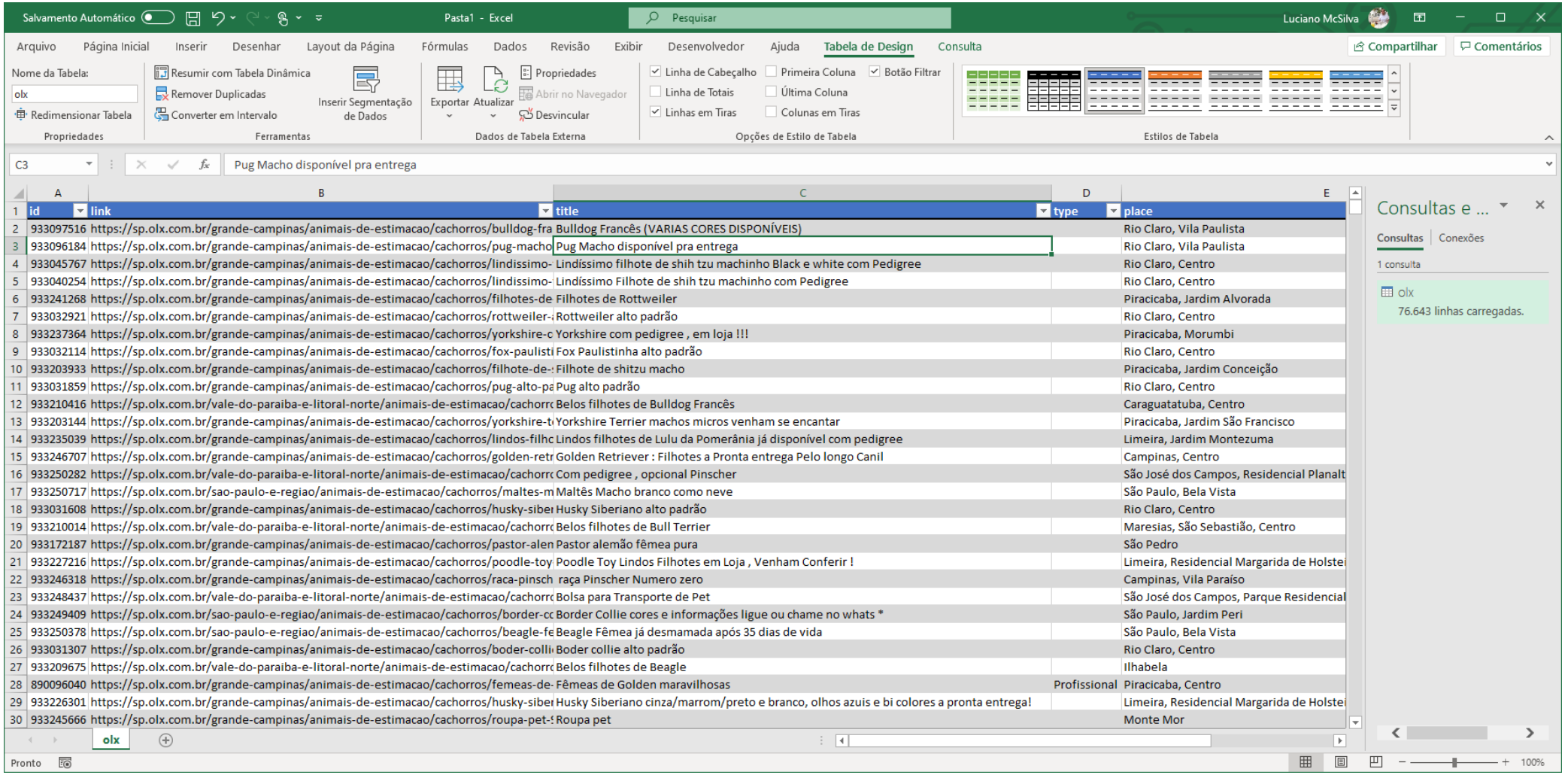
O caso de uso

O caso de uso



Fonte: http://www.olx.com.br. Acesso em: 29 set. 2021.

76.643 anúncios classificados organizados



Salvamento Automático — Pasta1 - Excel — Pesquisar — Luciano McSilva

Arquivo | Página Inicial | Inserir | Desenhar | Layout da Página | Fórmulas | Dados | Revisão | Exibir | Desenvolvedor | Ajuda | Tabela de Design | Consulta

Compartilhar | Comentários

Nome da Tabela: olx
Redimensionar Tabela
Propriedades
Resumir com Tabela Dinâmica
Remover Duplicadas
Converter em Intervalo
Inserir Segmentação de Dados
Ferramentas
Exportar | Atualizar | Propriedades | Abrir no Navegador | Desvincular
Dados de Tabela Externa
Linha de Cabeçalho | Primeira Coluna | Botão Filtrar
Linha de Totais | Última Coluna
Linhas em Tiras | Colunas em Tiras
Opções de Estilo de Tabela
Estilos de Tabela

C3: Pug Macho disponível pra entrega

	A	B	C	D	E
1	id	link	title	type	place
2	933097516	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/bulldog-fra	Bulldog Francês (VARIAS CORES DISPONÍVEIS)		Rio Claro, Vila Paulista
3	933096184	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/pug-macho	Pug Macho disponível pra entrega		Rio Claro, Vila Paulista
4	933045767	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/lindissimo-	Lindíssimo filhote de shih tzu machinho Black e white com Pedigree		Rio Claro, Centro
5	933040254	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/lindissimo-	Lindíssimo Filhote de shih tzu machinho com Pedigree		Rio Claro, Centro
6	933241268	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/filhotes-de	Filhotes de Rottweiler		Piracicaba, Jardim Alvorada
7	933032921	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/rottweiler-	Rottweiler alto padrão		Rio Claro, Centro
8	933237364	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/yorkshire-c	Yorkshire com pedigree , em loja !!!		Piracicaba, Morumbi
9	933032114	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/fox-paulisti	Fox Paulistinha alto padrão		Rio Claro, Centro
10	933203933	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/filhote-de-	Filhote de shitzu macho		Piracicaba, Jardim Conceição
11	933031859	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/pug-alto-pa	Pug alto padrão		Rio Claro, Centro
12	933210416	https://sp.olx.com.br/vale-do-paraiba-e-litoral-norte/animais-de-estimacao/cachorr	Belos filhotes de Bulldog Francês		Caraguatatuba, Centro
13	933203144	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/yorkshire-t	Yorkshire Terrier machos micros venham se encantar		Piracicaba, Jardim São Francisco
14	933235039	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/lindos-filhc	Lindos filhotes de Lulu da Pomerânia já disponível com pedigree		Limeira, Jardim Montezuma
15	933246707	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/golden-retr	Golden Retriever : Filhotes a Pronta entrega Pelo longo Canil		Campinas, Centro
16	933250282	https://sp.olx.com.br/vale-do-paraiba-e-litoral-norte/animais-de-estimacao/cachorr	Com pedigree , opcional Pinscher		São José dos Campos, Residencial Planalt
17	933250717	https://sp.olx.com.br/sao-paulo-e-regiao/animais-de-estimacao/cachorros/maltes-m	Maltês Macho branco como neve		São Paulo, Bela Vista
18	933031608	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/husky-siber	Husky Siberiano alto padrão		Rio Claro, Centro
19	933210014	https://sp.olx.com.br/vale-do-paraiba-e-litoral-norte/animais-de-estimacao/cachorr	Belos filhotes de Bull Terrier		Maresias, São Sebastião, Centro
20	933172187	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/pastor-alen	Pastor alemão fêmea pura		São Pedro
21	933227216	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/poodle-toy	Poodle Toy Lindos Filhotes em Loja , Venham Conferir !		Limeira, Residencial Margarida de Holste
22	933246318	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/raca-pinsch	raça Pinscher Numero zero		Campinas, Vila Paraíso
23	933248437	https://sp.olx.com.br/vale-do-paraiba-e-litoral-norte/animais-de-estimacao/cachorr	Bolsa para Transporte de Pet		São José dos Campos, Parque Residencial
24	933249409	https://sp.olx.com.br/sao-paulo-e-regiao/animais-de-estimacao/cachorros/border-cc	Border Collie cores e informações ligue ou chame no whats *		São Paulo, Jardim Peri
25	933250378	https://sp.olx.com.br/sao-paulo-e-regiao/animais-de-estimacao/cachorros/beagle-fe	Beagle Fêmea já desmamada após 35 dias de vida		São Paulo, Bela Vista
26	933031307	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/boder-colli	Boder collie alto padrão		Rio Claro, Centro
27	933209675	https://sp.olx.com.br/vale-do-paraiba-e-litoral-norte/animais-de-estimacao/cachorr	Belos filhotes de Beagle		Ilhabela
28	890096040	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/femeas-de-	Fêmeas de Golden maravilhosas	Profissional	Piracicaba, Centro
29	933226301	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/husky-siber	Husky Siberiano cinza/marrom/preto e branco, olhos azuis e bi colores a pronta entrega!		Limeira, Residencial Margarida de Holste
30	933245666	https://sp.olx.com.br/grande-campinas/animais-de-estimacao/cachorros/roupa-pet-s	Roupa pet		Monte Mor

Consultas e ...
Consultas | Conexões
1 consulta
olx
76.643 linhas carregadas.

olx

Pronto — 100%

Nuvem de palavras sem nenhum tratamento

Pré-tratamento

Pré-tratamento – Texto insensível e remoção de caracteres especiais

Fonte: GettyImages

Pré-tratamento – Remoção de plural

Palavras terminadas em	Regra	Exemplos
‘a’, ‘e’, ‘i’, ‘o’, ‘u’, ‘n’	Acrescenta-se a letra **‘s’**	carro – carros, cadeira – cadeiras, lúmen – lúmens
‘m’	Troca-se o **‘m’** por **‘ns’**	viagem – viagens, som – sons
‘ão’	Utiliza-se: **‘aõs’**, **‘ões’** ou **‘ães’**	pão – pães, cidadão – cidadãos, avião – aviões
‘r’, ‘s’ ou ‘z’	Acrescenta-se **‘es’**	cor – cores, holandês – holandeses, raiz – raízes
‘al’, ‘el’, ‘ol’ ou ‘ul’	Troca-se o **‘l’** por **‘is’**	farol – faróis, anzol – anzóis
‘il’	Troca-se o **‘il’** por **‘is’**	míssil – misseis, cantil – cantis
‘x’	Não se altera	durex – durex, xerox – xerox

Pré-tratamento – Palavras de Parada

Fonte: GettyImages

a	às	num	pelo	isto	nossa	tiver	seriam	tiverem	estivesse
o	tu	nem	pela	está	delas	terei	tinham	teremos	estiverem
e	te	meu	isso	haja	estes	terão	tivera	estivera	houvessem
é	há	nós	seus	eram	estas	teria	tenham	houvemos	houvermos
à	que	lhe	quem	fora	estou	quando	teriam	houveram	houveriam
de	com	vos	esse	seja	estão	também	aqueles	houvesse	tivéramos
do	não	teu	eles	será	houve	depois	aquelas	houverem	estivessem
da	uma	tua	você	tive	hajam	minhas	estamos	houverei	estivermos
em	por	hei	essa	teve	somos	nossos	estavam	houverão	houvéramos
um	dos	hão	suas	terá	fomos	nossas	estejam	houveria	houveremos
os	mas	sou	numa	muito	foram	aquele	estiver	fôssemos	tivéssemos
no	ele	são	elas	entre	sejam	aquela	havemos	seríamos	estivéramos
se	das	era	qual	mesmo	fosse	aquilo	houvera	tínhamos	houvéssemos
na	seu	fui	este	minha	forem	estive	hajamos	tenhamos	houveríamos
as	sua	foi	dele	pelos	serei	esteve	houverá	tivessem	estivéssemos
ao	nos	for	lhes	deles	serão	estava	fôramos	tivermos	
ou	até	tem	meus	essas	seria	esteja	sejamos	teríamos	
já	ela	tém	teus	esses	tenho	houver	seremos	estivemos	
eu	sem	para	tuas	pelas	temos	éramos	tivemos	estiveram	
só	aos	mais	dela	vocês	tinha	fossem	tiveram	estávamos	
me	nas	como	esta	nosso	tenha	formos	tivesse	estejamos	

Fonte: "Natural Language Toolkit" por NLTK team licenciado sob Apache 2.0

Pré-tratamento – Palavras a serem ignoradas (dado o contexto)

IGNORADAS

c	dia	hora	total	ideal	pronta	fisica	vendese	adquirir	incriveis	lindissimo	veterinaria
r	via	leia	todas	reais	pronto	cabeca	reserva	vacinada	garantias	disponivel	assistencia
x	vai	veja	todos	puros	lindas	adulto	reserve	adoravel	microchip	entregamos	vermifugado
ja	chip	veje	todas	sonho	lindos	vacina	pelagem	condicao	vacinados	entregamos	apartamento
so	hoje	casa	venha	preco	padrao	azuis]	tamanho	amorosos	saudaveis	disposicao	companheiro
sp	raca	apto	belos	unico	compro	prontos	conosco	imediata	vacinados	parcelamos	filhotinhos
hs	loja	belo	lojas	unica	compra	entrega	filhote	conhecer	adoraveis	belissimos	maravilhosos
cm	dias	azul	porte	sonho	chamar	suporte	amarelo	namorada	seguranca	felicidade	oportunidade
ate	voce	macho	amigo	otimo	melhor	confira	machinho	conferir	condicoes	transporte	parcelamento
pra	todo	femea	super	otima	cartao	adquira	pedigree	proprias	polegadas	oferecemos	vermifugados
vet	toda	lindo	hiper	chama	varias	tamanho	lindinho	servicos	perfeitos	brincalhao	olhocaramelo
vet	vida	linda	meses	cinza	visita	vacinas	lindinha	retirada	descricao	exclusivas	exclusividade
top	pura	fofos	vezes	preto	unicos	garanta	fofinhos	cachorro	alexandre	exclusivos	exclusividades
bem	casa	saude	docil	preta	unicas	amoroso	garantia	legitimo	companhia	beneficios	brancochocolate
vez	capa	racas	bebes	canil	otimos	contato	contrato	tamanhos	qualidade	filhotinho	vacinado
sim	amor	vendo	chame	olhos	otimas	contato	promocao	linhagem	companhia	merlepreto	maravilhoso
nao	face	vende	horas	white	gratis	procuro	whatsapp	filhotes	excelente	lindissimos	
hrs	aqui	venda	ainda	black	fofura	alegria	gratuito	vermelho	exclusiva	disponiveis	
ver	novo	juros	unica	merle	branco	procura	saudavel	merlered	exclusivo	veterinario	
lar	bebe	whats	feliz	machos	cabeca	visitar	clinicas	lindinhos	pagamento	informacoes	
apt	info	ligue	fotos	femeas	ultimo	conheca	melhores	lindinhas	chocolate	procedencia	

Pré-tratamento – Código utilizado

```
1.  import nltk
2.  from nltk.corpus import stopwords
3.  from unidecode import unidecode
4.  import string
5.
6.  # download das palavras de parada em português
7.  nltk.download('stopwords')
8.  stop = stopwords.words('portuguese')
9.
10. #Transforma em minúscula e remoção de acentos
11. titles['title_trd'] = titles['title'].str.lower().apply(lambda x: unidecode(x))
12.
13. # remoção de pontuação
14. titles['title_trd'] = titles['title_trd'].str.replace('[{}]'.format(string.punctuation), '')
15.
16. # remoção de números
17. titles['title_trd'] = titles['title_trd'].str.replace('[{}]'.format(string.digits), '')
18.
19. # remoção de palavras de parada
20. titles['title_trd'] = titles['title_trd'].str.apply(lambda x: ' '.join([w for w in x.split() if w not in (stop)]))
21.
22. # remoção plural
23. titles['title_trd'] = titles['title_trd'].apply(lambda x: ' '.join([singularizar(word) for word in x.split()]))
24.
25. # remoção palavras a serem ignoradas
26. titles['title_treated_ignored'] = titles['title_trd'].str.apply(lambda x: ' '.join([w for w in x.split() if w not in (ignorewords)]))
```

Extração de características

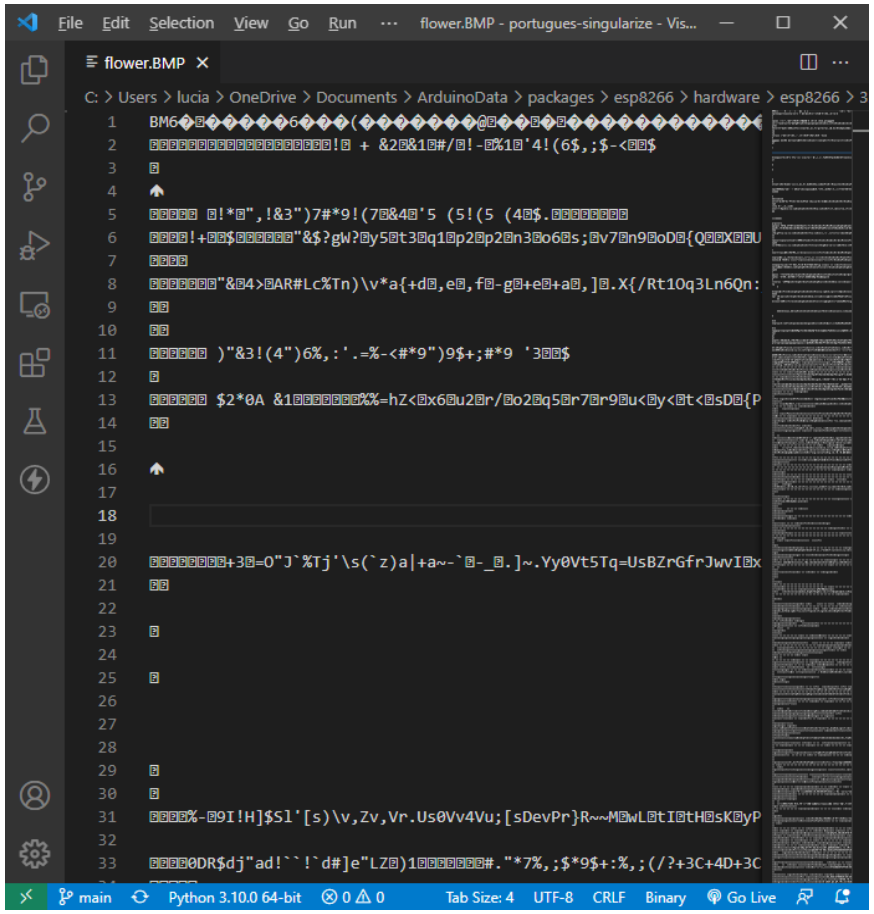
(falta de) Padrão

"O bom de usar padrões de mercado é que existem muitos diferentes para você escolher."

Andrew S. Tanenbaum

Fonte: Wikipédia

Arquivos



Fonte: Elaborada pelo autor.

- Arquivos computacionais são meras representações de 0s e 1s e **metadados**.
- Comumente separados em **Binários** e **Textos**.
- Arquivos textos não demandam metadados complexos para sua leitura, somente codificação dos bytes (***charset***).
- Há milhares de codificação possível para os bits de um texto. Os mais comuns em português são UTF-8, Windows 1250 e ISSO 8859-1.

Diversos **padrões** de codificação

	p	a	d	r	õ	e	s
UTF-8	01110000	01100001	01100100	01110010	11110101	01100101	01110011
ISO 8859-1	112	97	100	114	195	101	115
Windows 1250	112	97	100	114	213	101	115
Mac OS Roman	112	97	100	114	155	101	115

Codificação de palavras

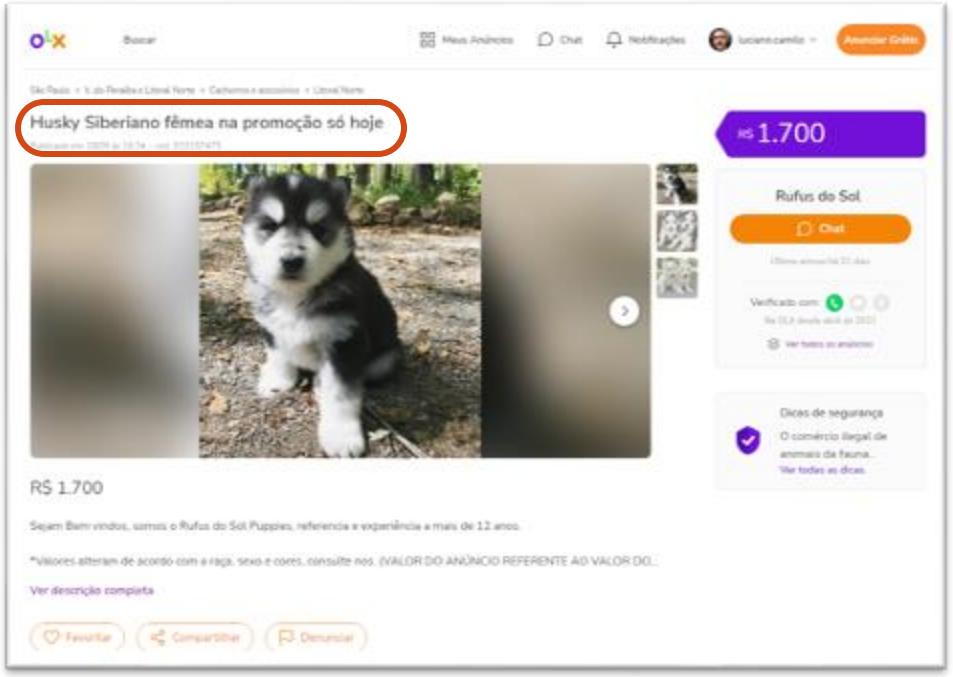
Diversos **padrões** de codificação

Padrões		Decodificação			
		UTF-8	**ISO 8859-1**	**Windows 1250**	**Mac OS Roman**
codificação	**UTF-8**	padrões	padrÃµes	padrÃµes	padrVµes
	ISO 8859-1	padr&#65533;es	padrões	padrőes	padrıes
	Windows 1250	padrões	padroes	padrões	padroes
	Mac OS Roman	padr&#65533;es	padre&#0;s	padr›es	padrões

Transformação feita pelo site https://string-functions.com/encodedecode.aspx

Tokenização de palavras

Husky Siberiano fêmea na promoção só hoje



Fonte: https://www.olx.com.br/

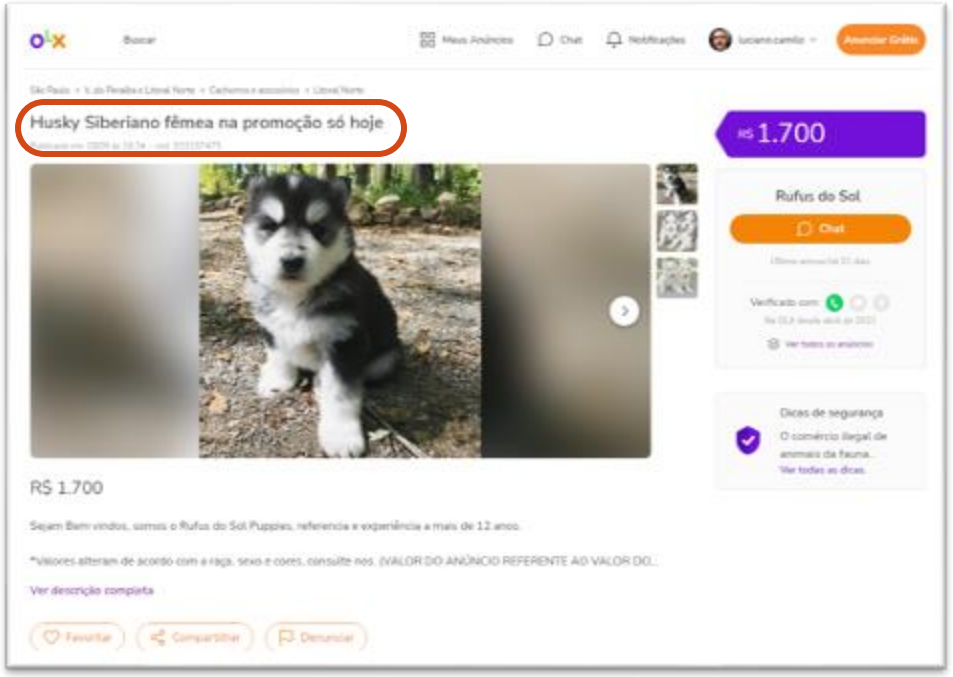
Separar cada palavra como um item próprio

Anúncio Classificado #933187475

- Husky
- Siberiano
- fêmea
- na
- promoção
- só
- hoje

Tokenização de palavras

husky siberiano fêmea na promoção só hoje



Fonte: https://www.olx.com.br/

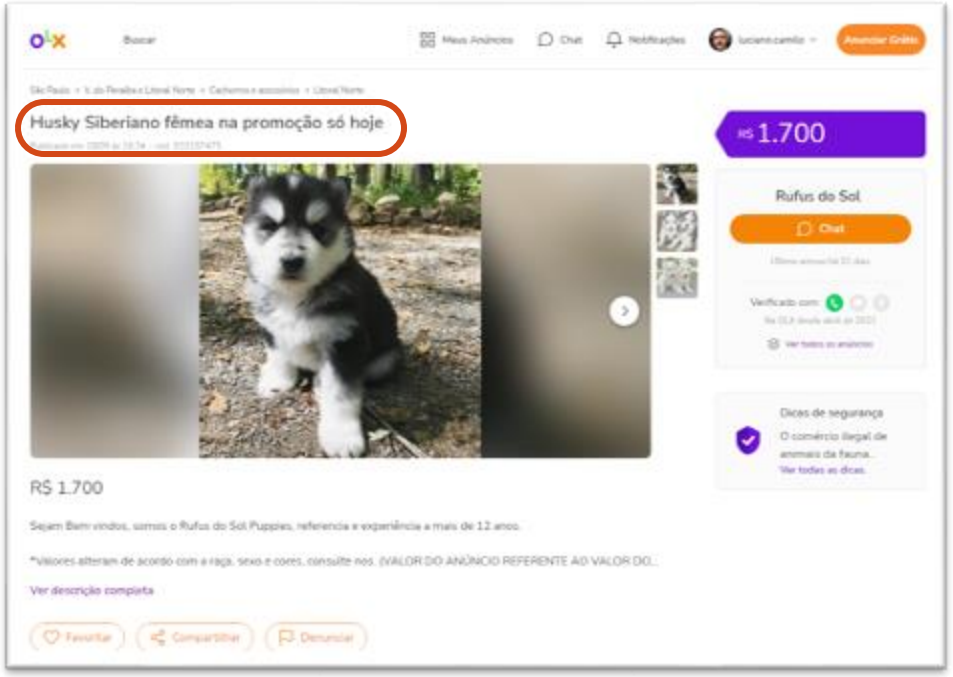
Todas em minúscula

Anúncio Classificado #933187475

- husky
- siberiano
- fêmea
- na
- promoção
- só
- hoje

Tokenização de palavras

husky siberiano femea na promocao so hoje



Fonte: https://www.olx.com.br/

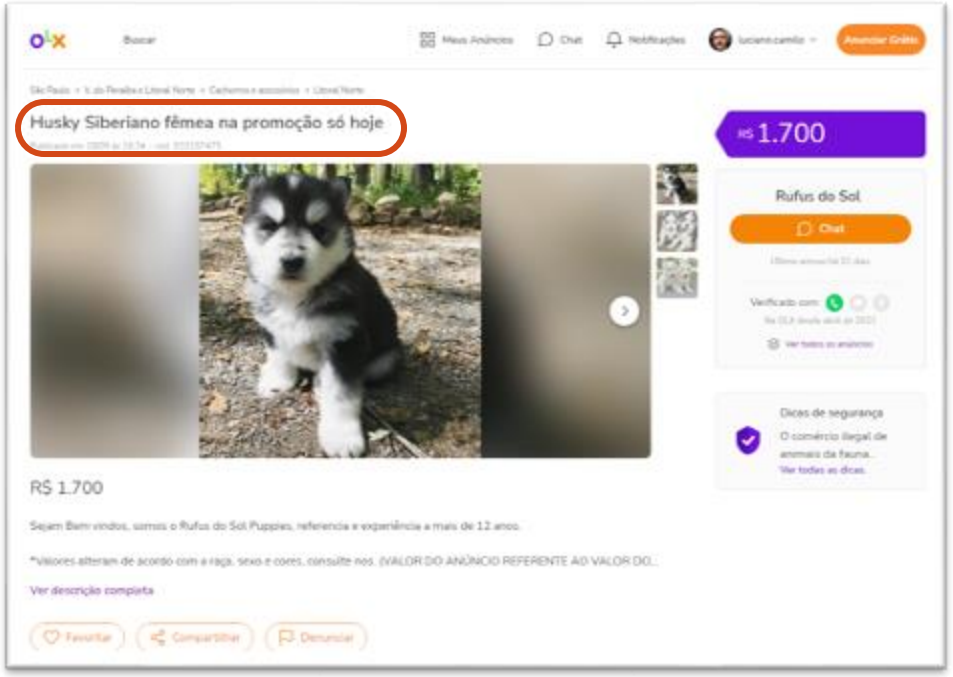
Sem caracteres especiais

Anúncio Classificado #933187475

- husky
- siberiano
- femea
- na
- promocao
- so
- hoje

Tokenização de palavras

husky siberiano femea ~~na~~ promocao so hoje



Fonte: https://www.olx.com.br/

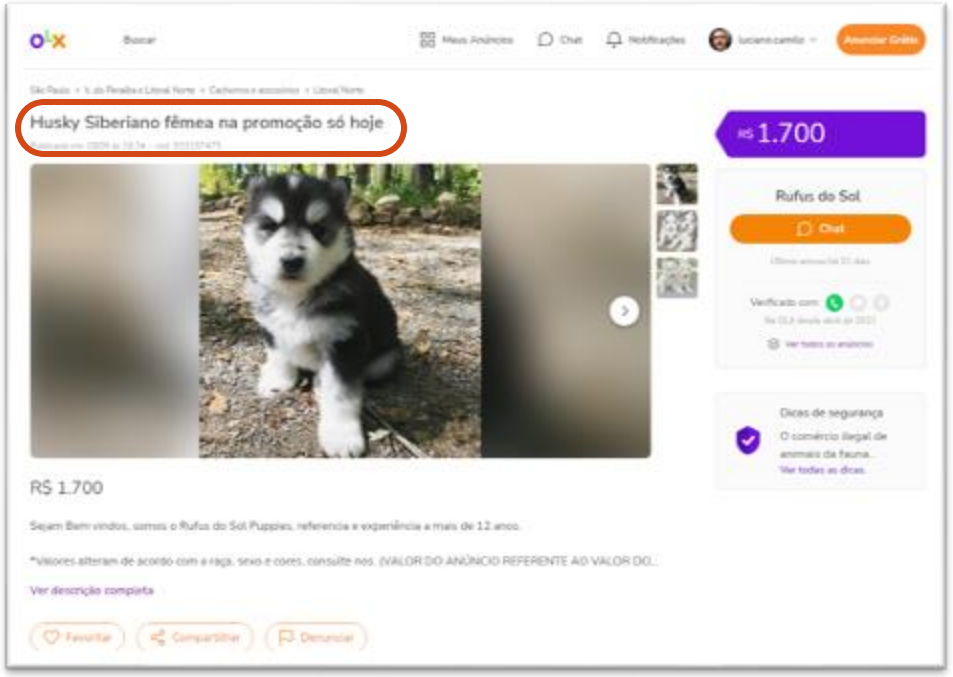
Removendo palavras de parada

Anúncio Classificado #933187475

- husky
- siberiano
- femea
- ~~na~~
- promocao
- so
- hoje

Tokenização de palavras

husky siberiano ~~femea na promocao so hoje~~



Fonte: https://www.olx.com.br/

Removendo palavras a serem ignoradas

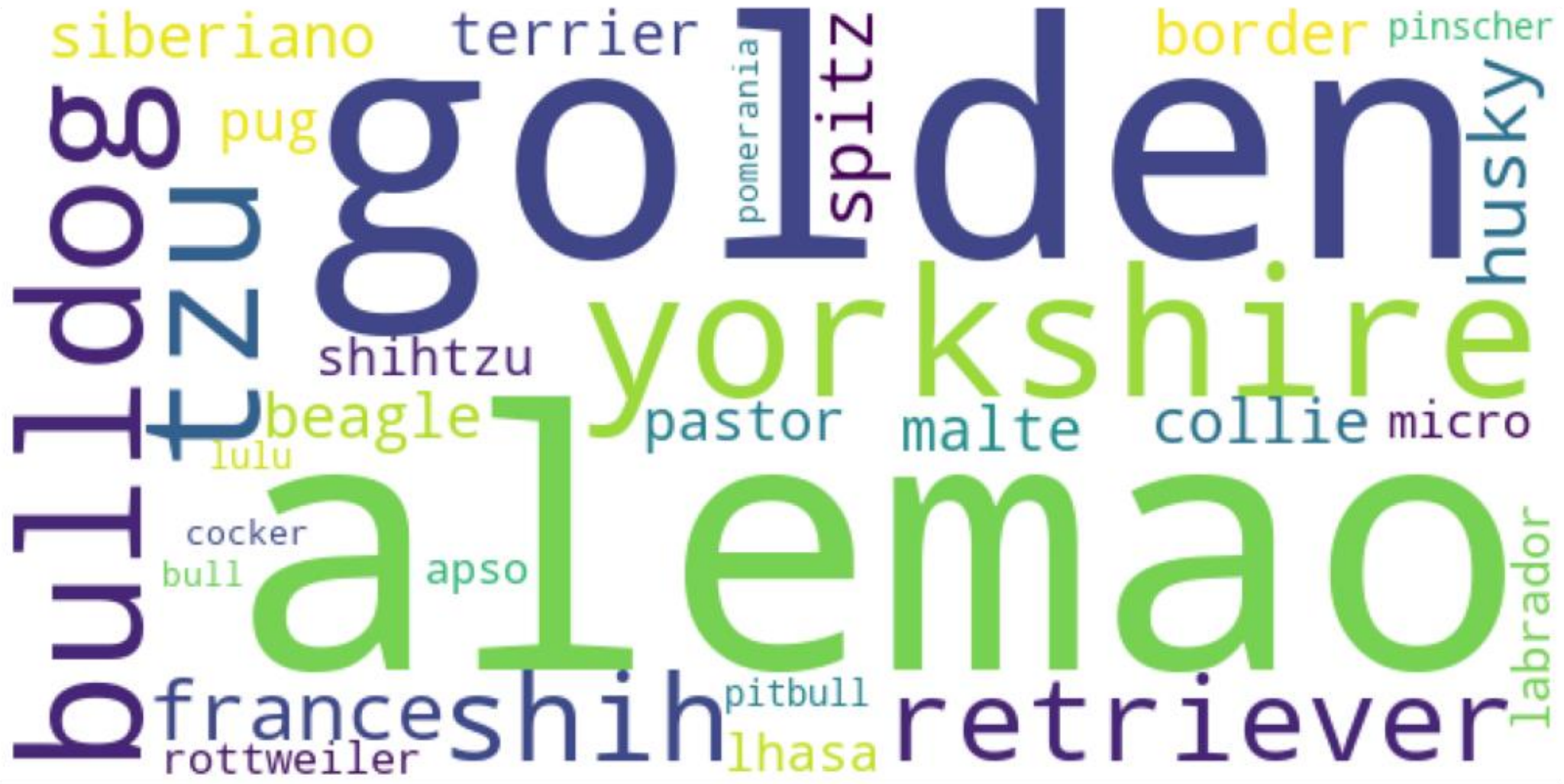
Anúncio Classificado #933187475

- husky
- siberiano
- ~~femea~~
- ~~na~~
- ~~promocao~~
- ~~so~~
- ~~hoje~~

Unigramas – Código utilizado

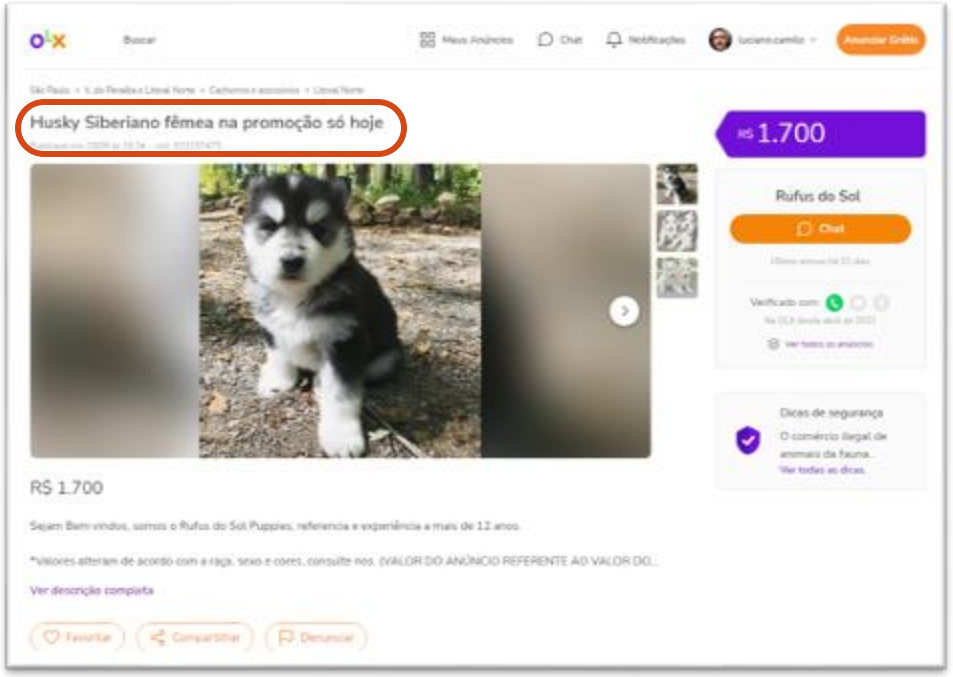
```
1.  from nltk.tokenize import word_tokenize
2.  from nltk import FreqDist
3.
4.  #cria uma lista com todos os anúncios concatenados
5.  treated_text = ' '.join(titles['title_treated_ignored'])
6.
7.  # cria o token quebrando por palavras
8.  tokens = word_tokenize(treated_text)
9.
10. # calcula a frequência, ou seja, quantas vezes cada palavra apareceu
11. unigram_fd = nltk.FreqDist(tokens)
12.
13. # ordena a frequência do maior para o menor
14. sorted_ugm = sorted(unigram_fd.items(), key=lambda x: x[1], reverse=True)
15.
16. # Carrega na biblioteca pandas para visualizar melhor
17. pd_unigram_df = pd.DataFrame.from_dict(sorted_ugm)
18.
19. # Apresenta as 30 palavras mais frequentes (desenhada na nuvem de palavras)
20. pd_unigram_df.head(30)
```

Nuvem de palavras após pré-tratamento (unigrama)



Tokenização de Palavras – Bigramas

Husky Siberiano fêmea na promoção só hoje



Fonte: https://www.olx.com.br/

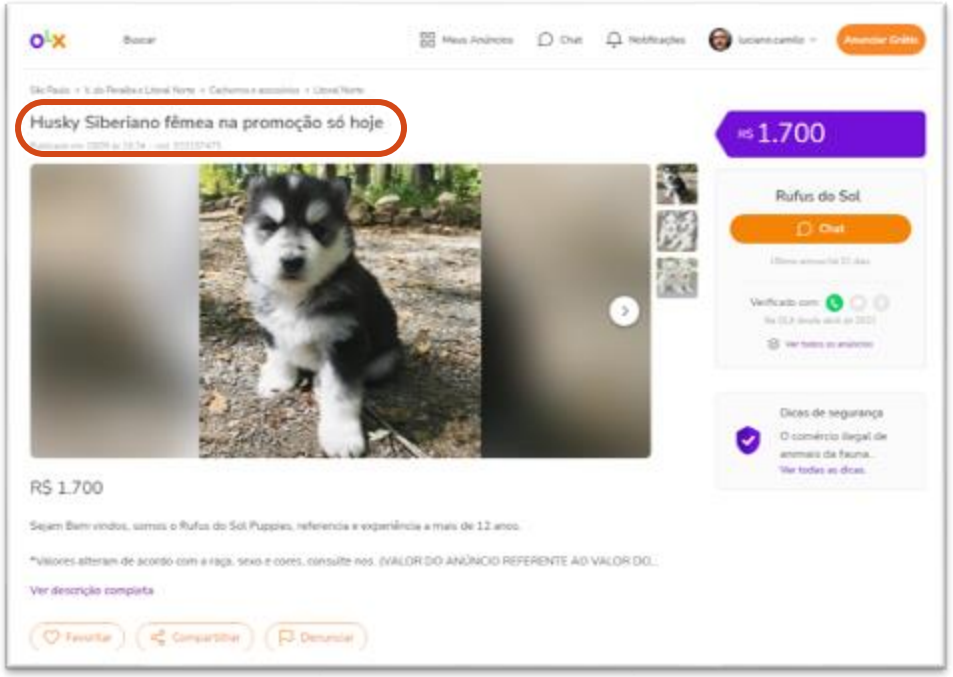
Separar cada palavra como um item próprio

Anúncio Classificado #933187475

- Husky Siberiano
- Siberiano fêmea
- fêmea na
- na promoção
- promoção só
- só hoje

Tokenização de Palavras – Bigramas

husky siberiano fêmea na promoção só hoje



Fonte: https://www.olx.com.br/

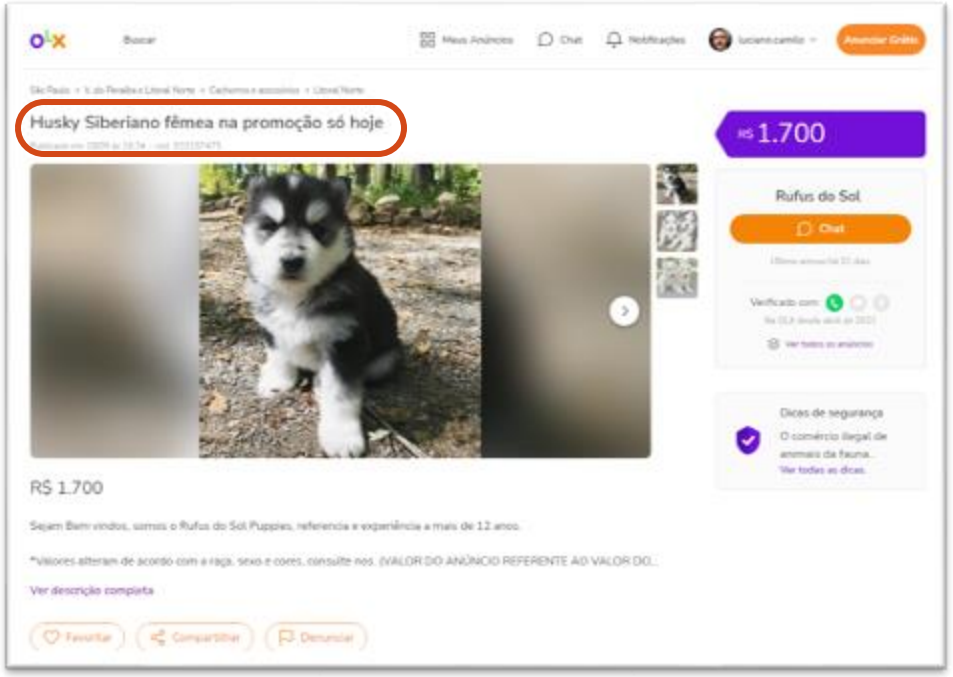
Todas em minúscula

Anúncio Classificado #933187475

- husky siberiano
- siberiano fêmea
- fêmea na
- na promoção
- promoção só
- só hoje

Tokenização de Palavras – Bigramas

husky siberiano femea na promocao so hoje



Fonte: https://www.olx.com.br/

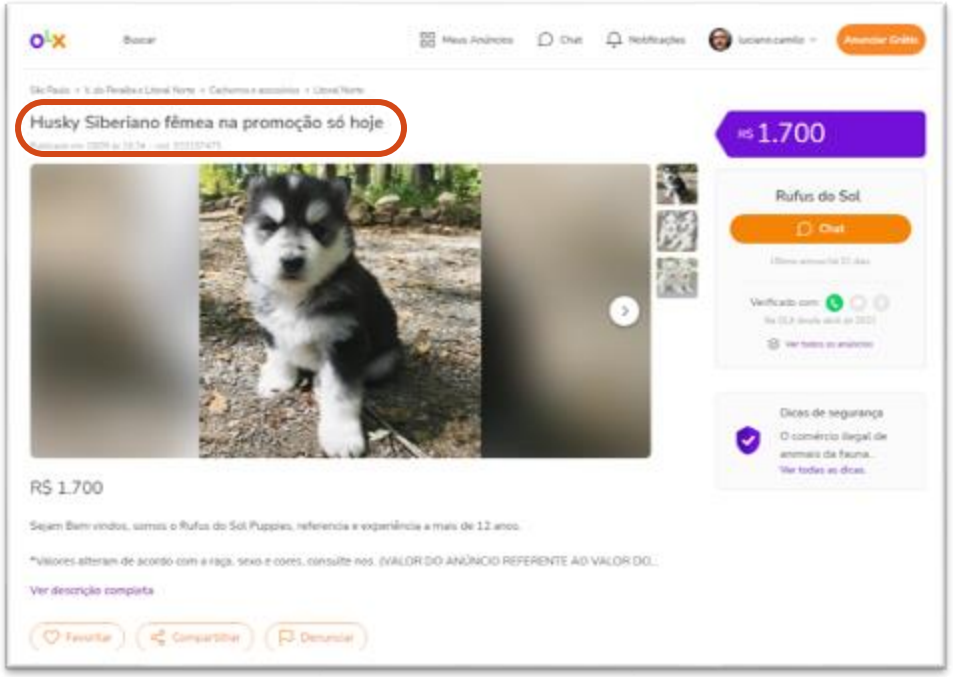
Sem caracteres especiais

Anúncio Classificado #933187475

- husky siberiano
- siberiano femea
- femea na
- na promocao
- promocao so
- so hoje

Tokenização de Palavras – Bigramas

husky siberiano femea ~~na~~ promocao so hoje



Fonte: https://www.olx.com.br/

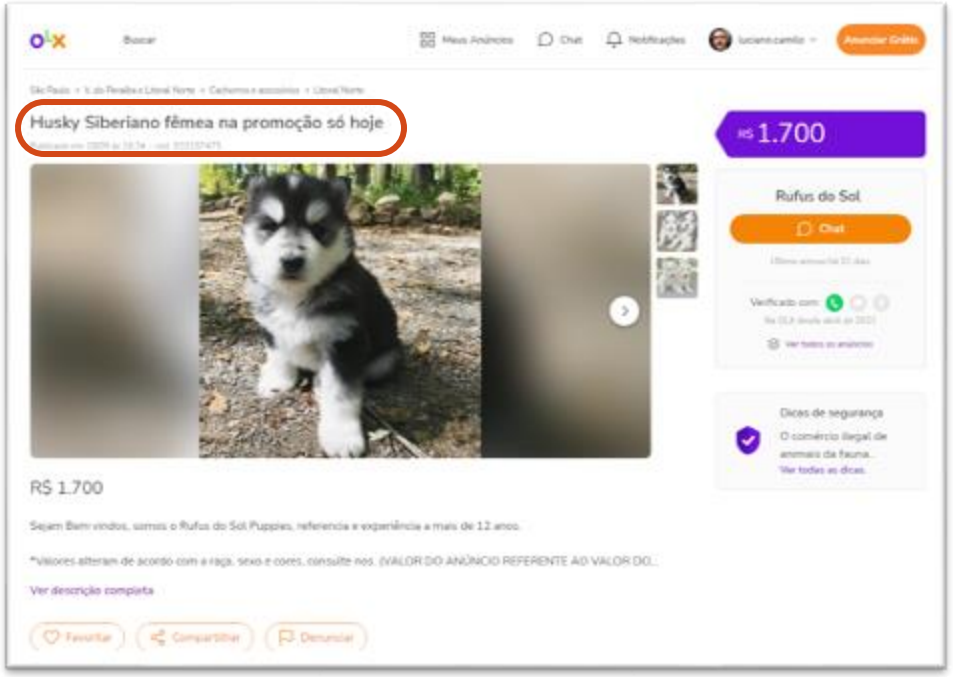
Removendo palavras de parada

Anúncio Classificado #933187475

- husky siberiano
- siberiano femea
- ~~femea na~~
- ~~na promocao~~
- femea promoção
- promoção so
- so hoje

Tokenização de Palavras – Bigramas

husky siberiano ~~femea na promocao so hoje~~



Fonte: https://www.olx.com.br/

Removendo palavras a serem ignoradas

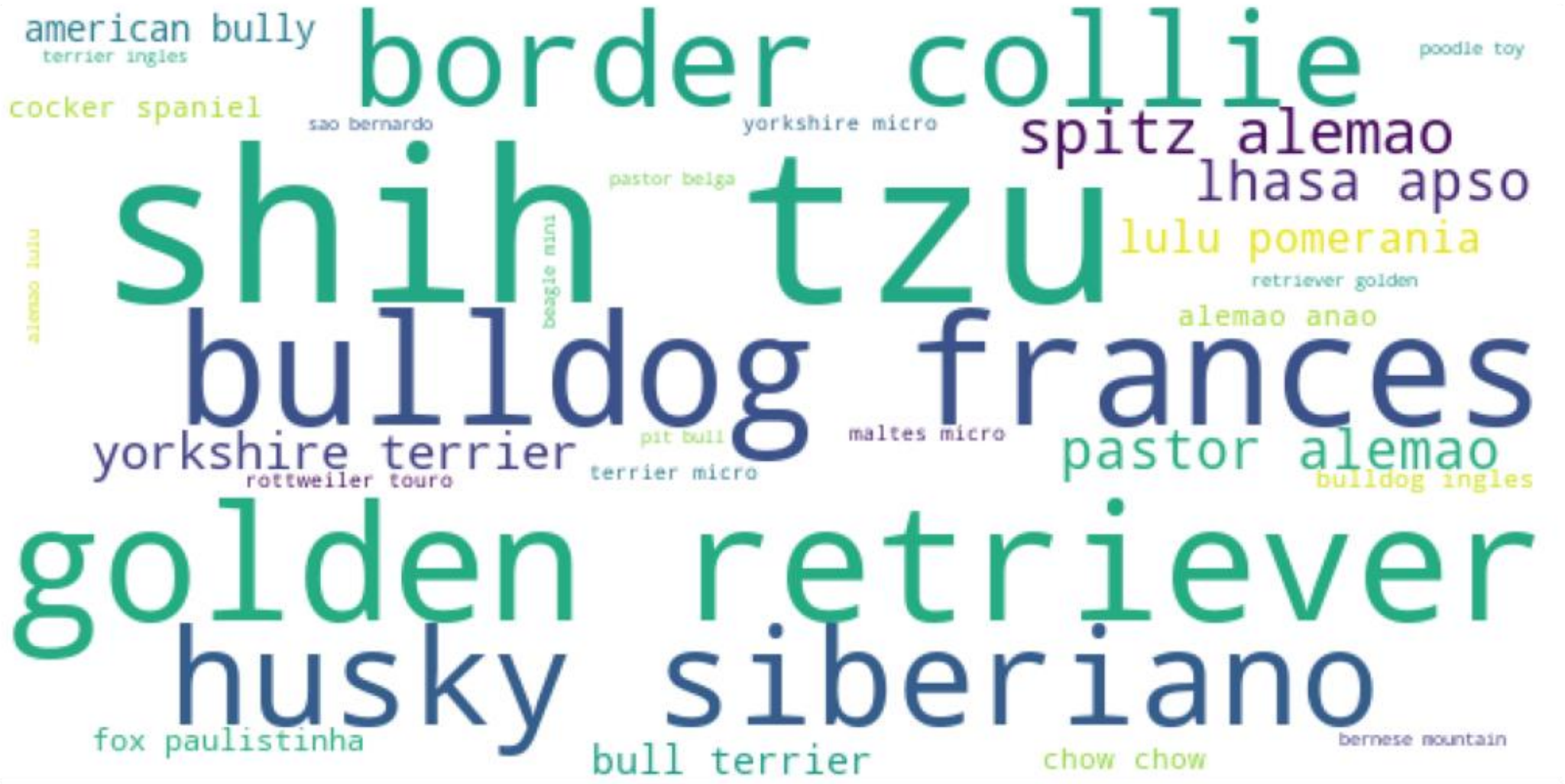
Anúncio Classificado #933187475

- husky siberiano
- ~~siberiano femea~~
- ~~femea na~~
- ~~na promocao~~
- ~~femea promoção~~
- ~~promoção so~~
- ~~so hoje~~

Bigramas – Código utilizado

```
1.  from nltk.tokenize import word_tokenize
2.  from nltk import FreqDist
3.  from nltk.util import ngrams
4.
5.  #cria uma lista com todos os anúncios concatenados
6.  treated_text = ' '.join(titles['title_treated_ignored'])
7.
8.  # cria o token quebrando por palavras
9.  tokens = word_tokenize(treated_text)
10.
11. # cria o token por bigrama
12. bigram_tokens = nltk.bigrams(tokens)
13.
14. # calcula a frequência, ou seja, quantas vezes cada palavra apareceu
15. bigram_fd = nltk.FreqDist(bigram_tokens)
16.
17. # ordena a frequência do maior para o menor
18. sorted_bgm = sorted(bigram_fd.items(), key=lambda x: x[1], reverse=True)
19.
20. # Carrega na biblioteca pandas para visualizar melhor
21. pd_bigram_df = pd.DataFrame.from_dict(sorted_bgm)
22.
23. # Apresenta as 30 palavras mais frequentes (desenhada na nuvem de palavras)
24. pd_bigram_df.head(30)
```

Nuvem de palavras após seleção dos 30 mais frequentes bigramas



Universidade Presbiteriana
Mackenzie