

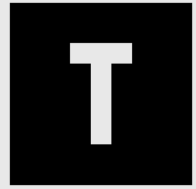
Tera



03/dezembro/2019

Processamento de Linguagem Natural

Cinthia M. Tanaka



Cinthia

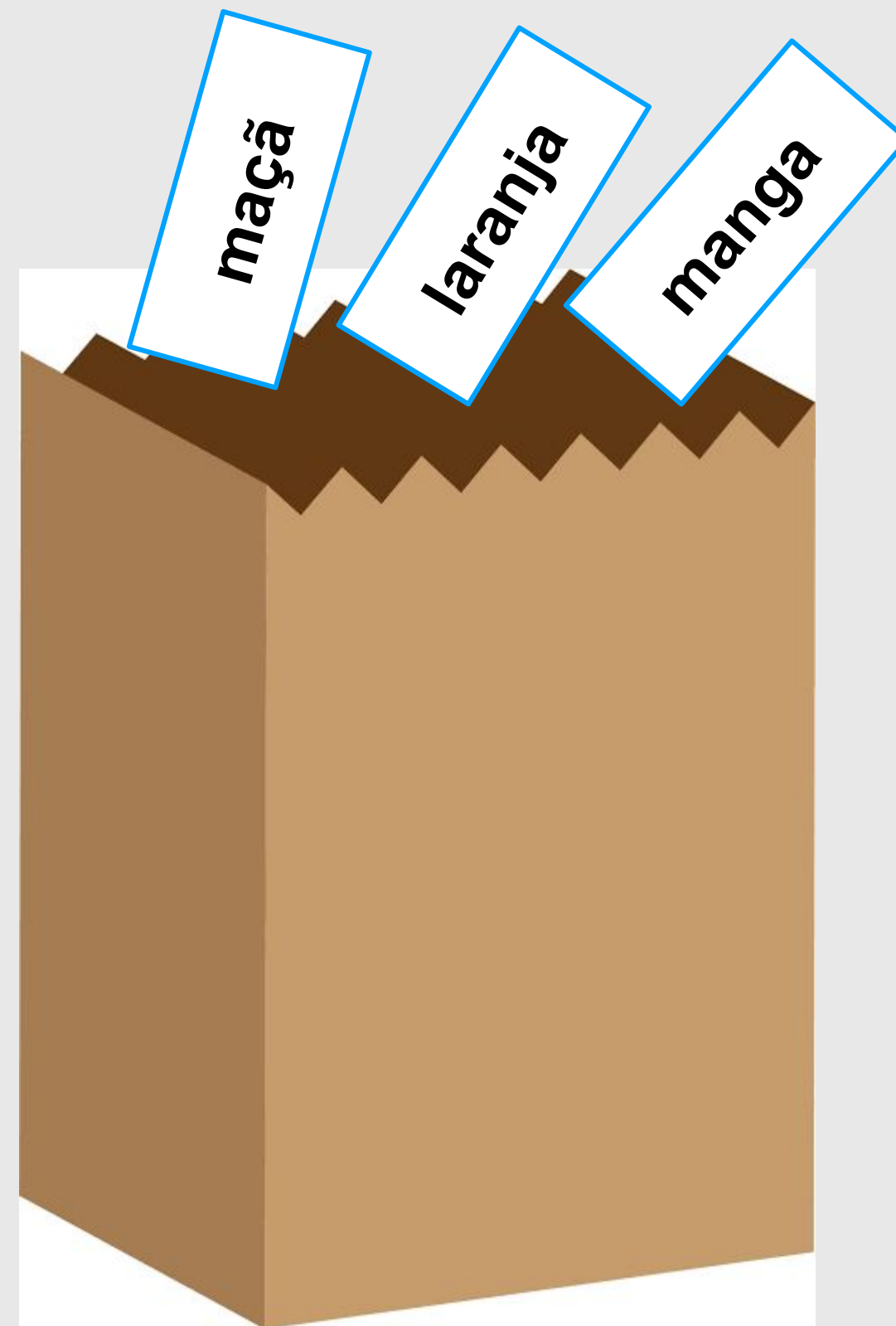
Data scientist no **Elo7**

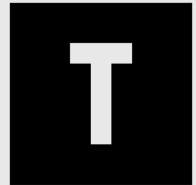
 @cimarieta

 cinthia-tanaka



Bag of words





Bag of words

	maçã	manga	laranja
Uma manga e uma maçã			
Uma maçã e uma manga			

Bag of words

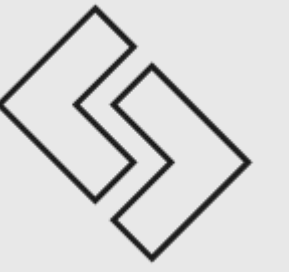
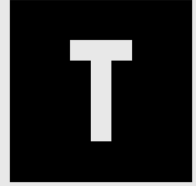
	maçã	manga	laranja
Uma manga e uma maçã	1	1	0
Uma maçã e uma manga	1	1	0

Bag of words

	maçã	manga	laranja
Uma manga e uma maçã	1	1	0
Uma árvore sem frutas			

Bag of words

	maçã	manga	laranja
Uma manga e uma maçã	1	1	0
Uma árvore sem frutas	0	0	0

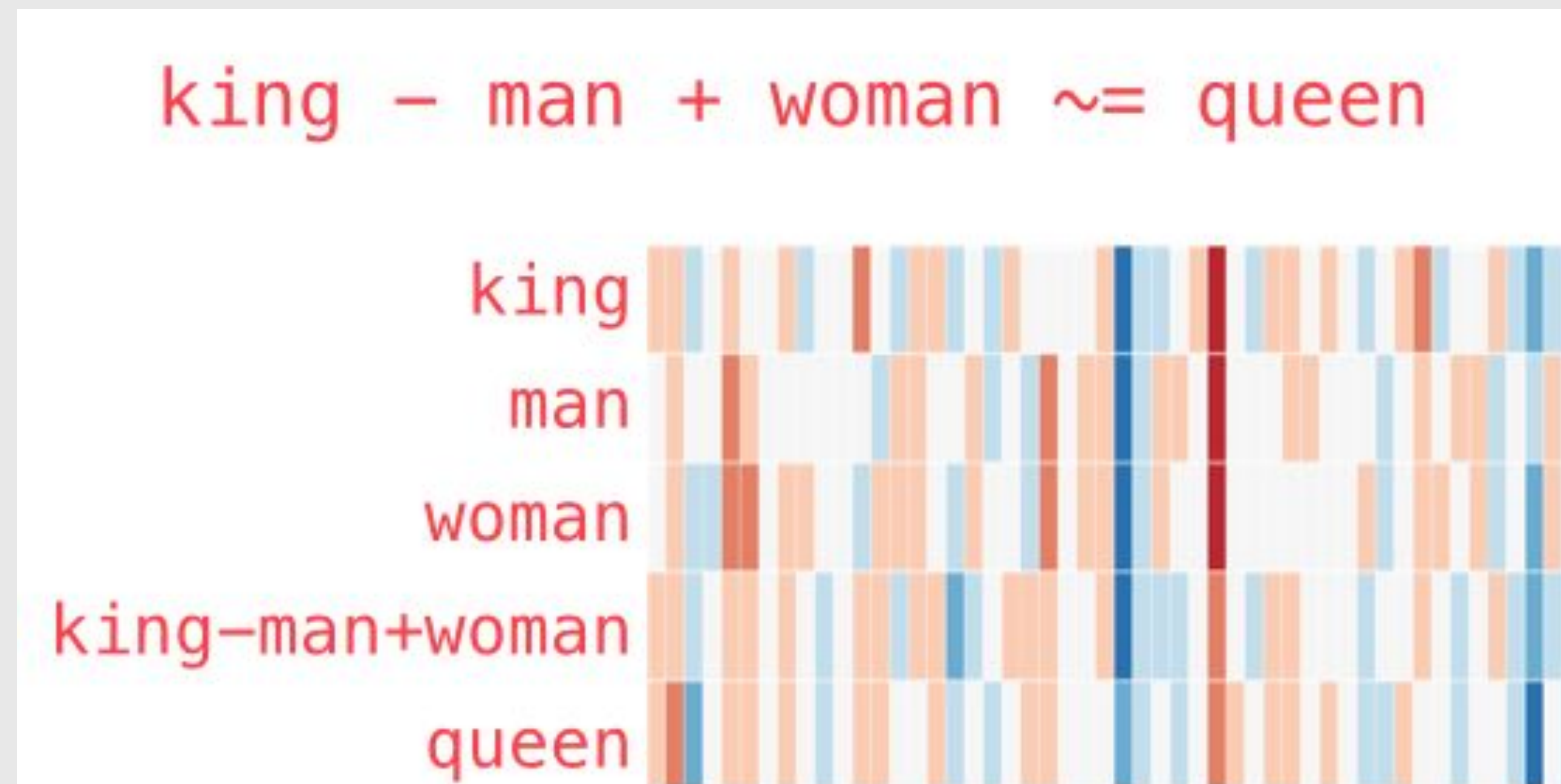


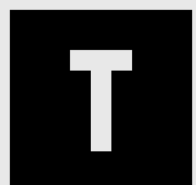
Word2Vec

- Incluindo o contexto em nossas representações de cada palavra

Word2Vec

- Incluindo o contexto em nossas representações de cada palavra





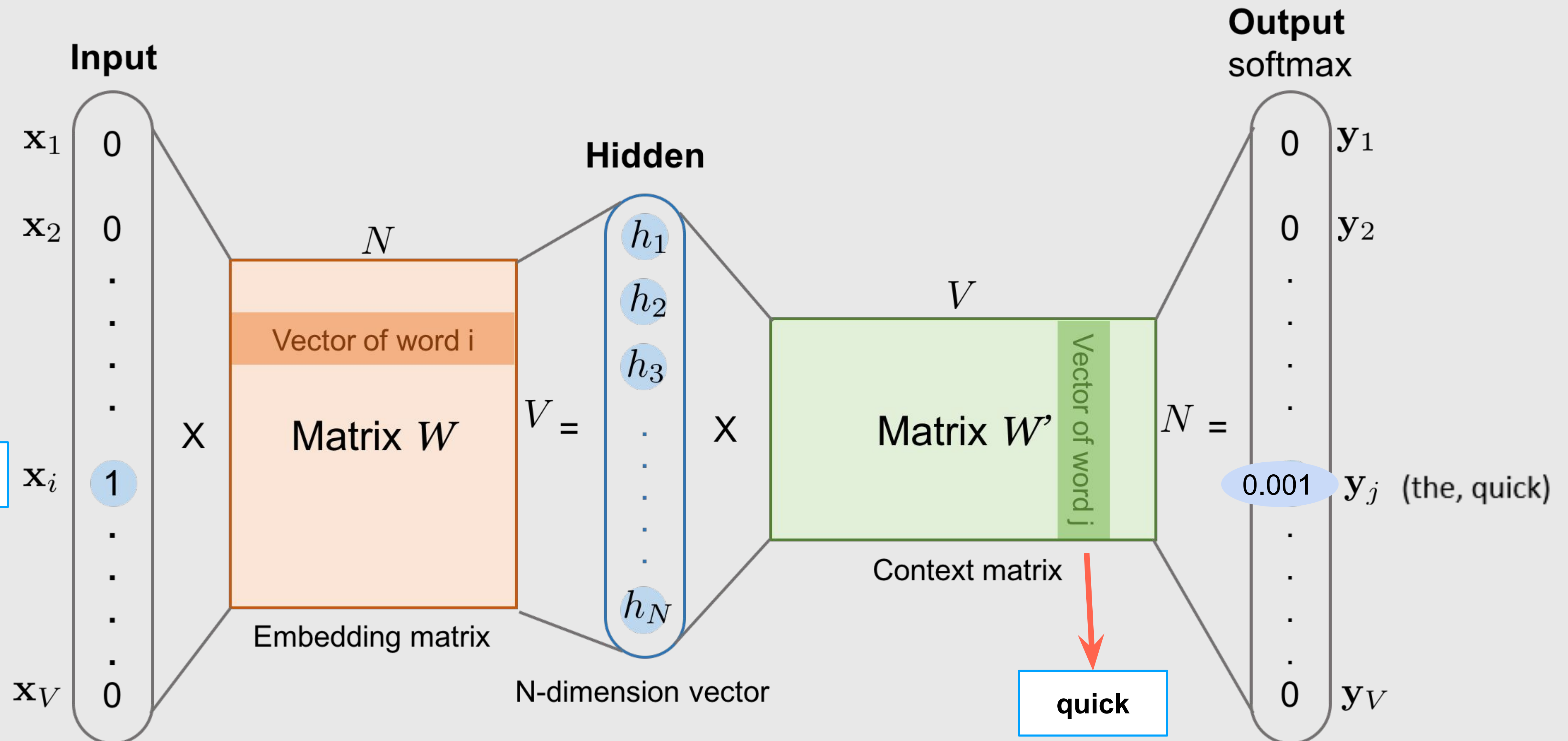
Word2Vec - Skipgram

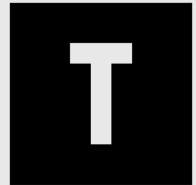
Source Text

Training
Samples

The quick brown fox jumps over the lazy dog. →	(the, quick) (the, brown)
The quick brown fox jumps over the lazy dog. →	(quick, the) (quick, brown) (quick, fox)
The quick brown fox jumps over the lazy dog. →	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown fox jumps over the lazy dog. →	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

Training Samples (the, quick)





Training Samples (the, quick)

Actual Target		Model Prediction
0		0 aardvark
0		0 aarhus
0		0.001 aaron
...		...
0		0.4 taco
1	-	0.001 quick
...		...
0		0.0001 zyzzzyva

Training
Samples (the, quick)

Actual
Target

0
0
0
...
0
1
...
0

-

Model
Prediction

0	aardvark
0	aarhus
0.001	aaron
...	
0.4	taco
0.001	quick
...	
0.0001	zyzzyva

=

Error

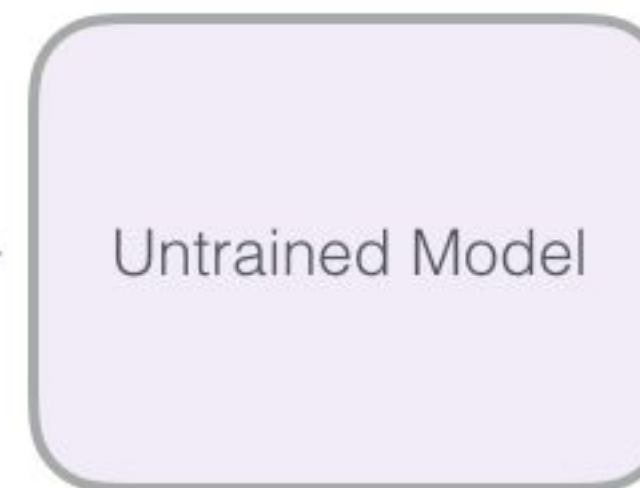
0
0
-0.001
...
-0.4
0.999
...
-0.0001

Training Samples (the, quick)

Actual
Target

0
0
0
...
0
1
...
0

not



Model
Prediction

0	aardvark
0	aarhus
0.001	aaron
...	...
0.4	taco
0.001	quick
...	...
0.0001	zyzzyva

Error

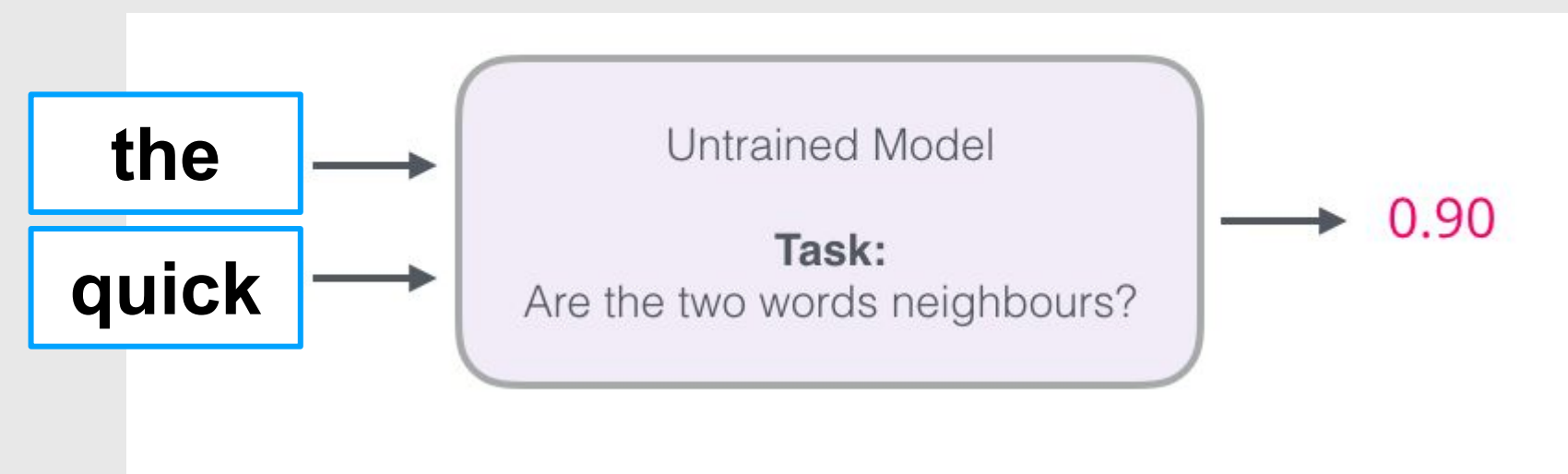
0
0
-0.001
...
-0.4
0.999
...
-0.0001

Update
Model
Parameters





Word2Vec - o treino “eficiente”



Word2Vec - o treino “eficiente”



input word	output word	target
the	quick	1
the		0
the		0
the	fox	1

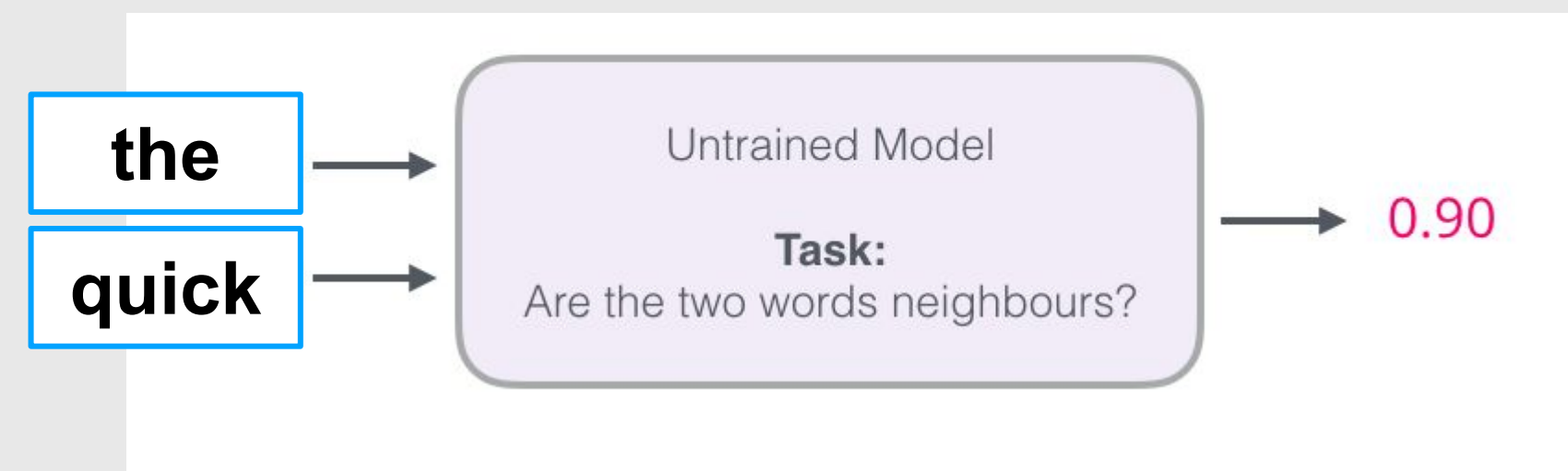
➤ Negative examples

input word	output word	target
the	quick	1
the	aaron	0
the	taco	0
the	fox	1

Pick randomly from vocabulary
(random sampling)

Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

Word2Vec - o treino “eficiente”



input word	output word	target
the	quick	1
the		0
the		0
the	fox	1

➤ Negative examples


input word	output word	target
the	quick	1
the	aaron	0
the	taco	0
the	fox	1

Pick randomly from vocabulary
(random sampling)

Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

E mais truques...

Limitações

- Não lida com palavras fora do vocabulário
- Palavras polissêmicas
 - Moranguinho:
 - Morango pequeno 
 - Personagem Moranguinho



Outros *word vector models*

- Um único vetor por palavra
 - [Glove](#) (da Stanford University)
 - [FastText](#) (do mesmo criador do Word2Vec)
 - Têm representações para palavras fora do vocabulário
- Tem mais de um vetor por palavra (tratam o problema da polissemia):
 - [ELMO](#) (da AllenNLP)
 - [BERT](#) (da Google)

Analogias podem ter vieses

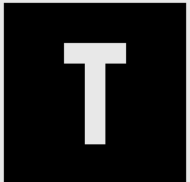
Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairstylist-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Referência: Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016. [\[link\]](#)



Aplicações de NLP no elo7

elo7

Produtos ▼ tapete urso

CEP para frete + barato

Cadastrar

Entrar

Categorias

Acessórios

Aniversário e Festas

Bebê

Bijuterias

Bolsas e Carteiras

Casa

Casamento

Convites

Decoração

Doces

Eco

Infantil

Jogos e Brinquedos

Jóias

Lembrancinhas

Papel e Cia

Pets

Religiosos

Roupas

Saúde e Beleza

Técnicas de Artesanato

Tapete Urso

1942 PRODUTOS ENCONTRADOS

Buscas relacionadas: tapete quarto menino tapete de barbante infantil tapete bebê tapete de ursinho tapete de elefante

Preço R\$ até R\$ Cidade Digite uma cidade Filtrar Filtrar por Todos produtos Ordenar por Relevância

Tapete Urso Théo

Tia Baby

R\$ 197,70 12x R\$ 16,48 sem juros

Tapete Urso Théo com Coroa

Tia Baby

R\$ 197,70 12x R\$ 16,48 sem juros

Tapete Urso Principe

Tia Baby

R\$ 169,40 12x R\$ 14,12 sem juros

Tapete Urso

Café Com Leite Artesanatos

R\$ 151,10 12x R\$ 12,59 sem juros

Tapete Urso Majestoso Bege

Tia Baby

R\$ 169,40 12x R\$ 14,12 sem juros

Tapete urso

Studio Mandarin

R\$ 255,40 12x R\$ 21,28 sem juros

Tapete Urso Navy

Tia Baby

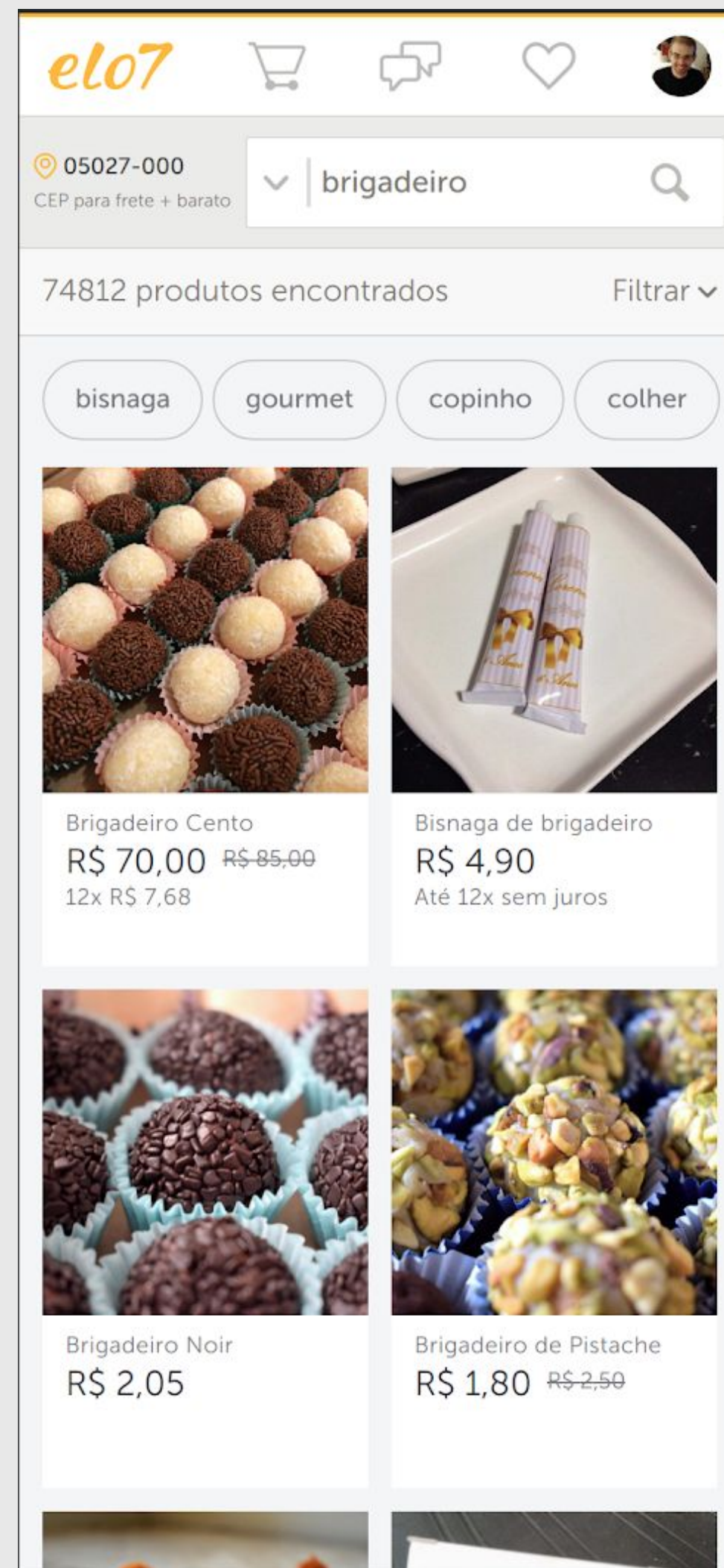
R\$ 169,40 12x R\$ 14,12 sem juros

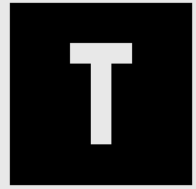
Tapete de urso

Ateliê croche e arte

R\$ 330,00 12x R\$ 27,50 sem juros

Aplicações de NLP no elo7





NLP é uma área em progresso



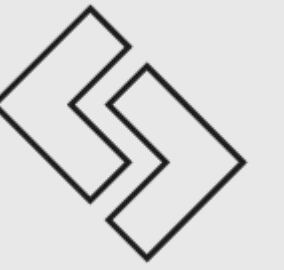
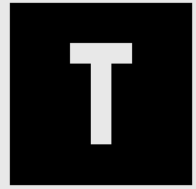
Jack Hessel
@jmhessel



Me, an NLP researcher: it's amazing how much language technologies have improved! The field is doing great!!

Me, a human, interacting with a customer service chat bot: OPERATOR OPERATOR OPERATOR

5:12 PM · Nov 30, 2019 · [Twitter for Android](#)



DÚVIDAS?!

