



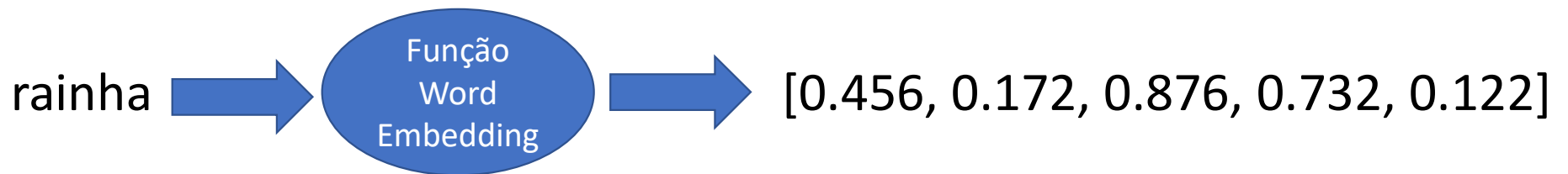
## Classificação com Word2vec

# 1. Conhecer a representação word embedding com Word2vec

## 1.1 - *Word embbeding*

## *Word embedding*

- Representação do significado de palavras.
- Diversos métodos diferentes para gerar (Word2vec, GloVe, FastText)
- Resultam em *word vectors*, que são vetores numéricos (não binários).



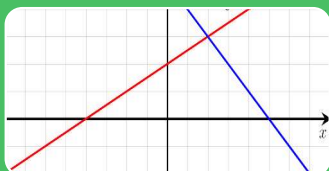


Podemos buscar pelo significado combinado de palavras.

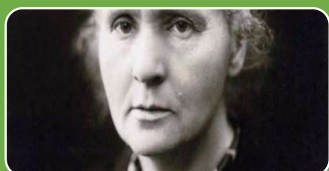


Queremos descobrir alguém que não lembramos o nome:

- Ela inventou algo a ver com a física na Europa no início do século XX.

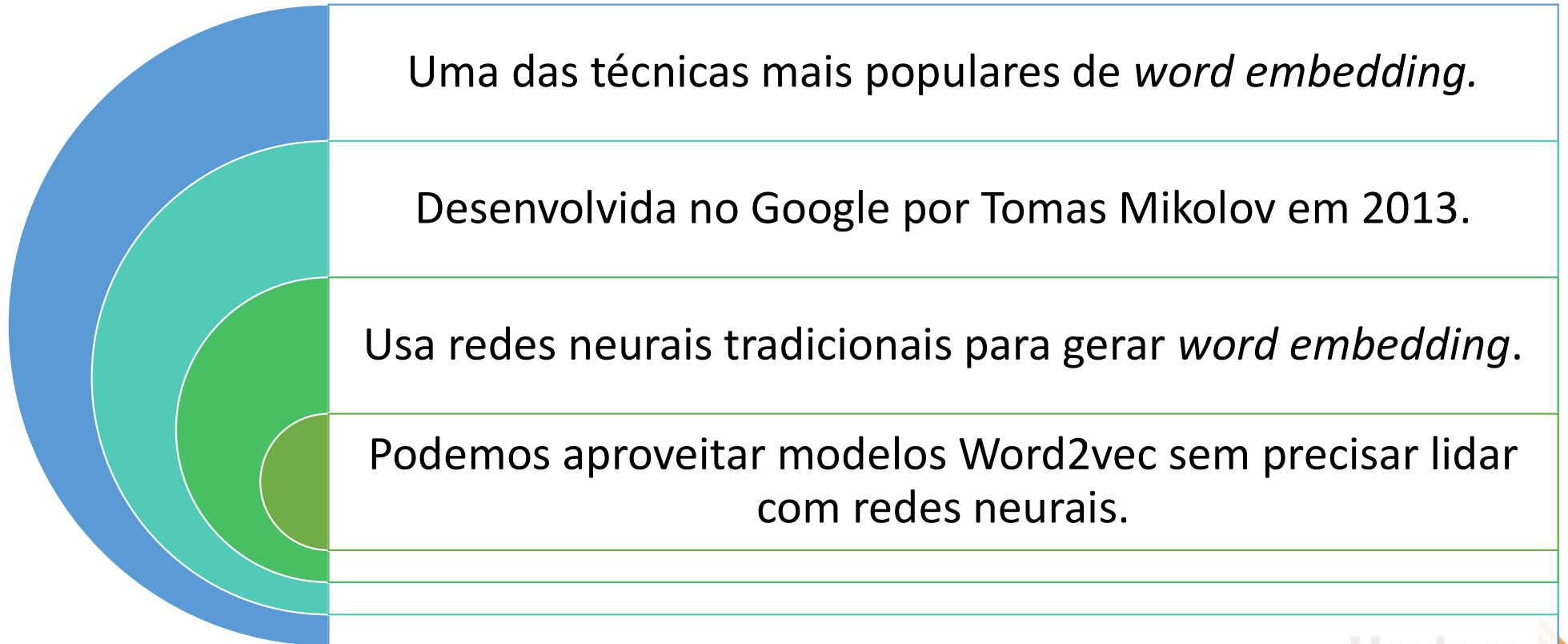


resposta = função("mulher") + função("física") +  
função("europa") + função("cientista") – 2\*função("esposa")



Descobriu?

# Word2vec



# Word2vec para classificar notícias

16 | Quarta-feira, 16 de outubro de 2019

Jornal do Comércio | Porto Alegre

**economia**

## Programa quer incentivar projetos de inovação

Financiamento é destinado a elevar competitividade no setor produtivo

/TECNOLOGIA

Adriana Lampert  
adriana@jornaldocomercio.com.br

Mecanismo adicional para as empresas habilitadas pela Lei de Informática cumprirem com os investimentos em pesquisa, desenvolvimento e inovação (PD&I), o Programa Prioritário (PP) em IoT/Manufatura 4.0, foi apresentado, nesta terça-feira, a empresários gaúchos, em evento ocorrido no Ritter Hotel, em Porto Alegre. Durante reunião-almoço promovida pela regional da Associação Brasileira da Indústria Elétrica e Eletrônica (Abinee), o diretor de operações Carlos Eduardo Pereira da Empresa Brasileira de Pesquisa e Inovação Industrial (Embrapii), explicou que o novo programa permite que as empresas aprofundem recursos da Lei de Informática (de incentivos fiscais) em um fundo para projetos de inovação para o setor.

Somando incentivos de R\$ 1,3 bilhão (de recursos públicos não reembolsáveis) destinados a 800 projetos desenvolvidos em cerca de quatro anos por quase 600 corporações em todo o País, a Embrapii é uma entidade privada sem fins lucrativos, que tem contrato de gestão com o Ministério da Ciência, Tecnologia, Inovações e Comunicações (MCTIC) para fomentar projetos de pesquisa aplicada em Internet das Coisas (IoT/Indústria 4.0) desenvolvidos em parceria com empresas e a academia. "Nosso objetivo principal é apoiar o avanço das novas tecnologias digitais e inovadoras do setor de TIC desenvolvidas para o setor produtivo", destacou Pereira.

Segundo o diretor da Embrapii, a contratação de projetos não está vinculada à realização de depósitos ao fundo, nem precisa de candidatura em edital com chamada pública. "Uma empresa que queria contratar projetos com uma das 42

instituições de ciência e tecnologia credenciadas como unidades da Embrapii, apenas deve observar algumas regras, a exemplo de buscar dois terços do total dos recursos em outras fontes para completar o orçamento." Isso porque a entidade oferece subsídio econômico de até um terço do valor total do projeto. "Para a aplicação dos recursos de PD&I neste Programa Prioritário, as empresas poderão escolher os diversos institutos de pesquisa, cadastrados como unidade Embrapii, para a realização de projetos conjuntos", ressaltou Pereira.

Ainda de acordo com o diretor de operações da Embrapii, não existe um valor mínimo ou máximo de financiamento dos projetos, mas a média tem ficado entre R\$ 200 mil para orçamentos menores e R\$ 20 milhões para os de maior porte. De acordo com Pereira, o diferencial dos programas de fomento que a entidade oferece é a "agilidade e flexibilidade na con-



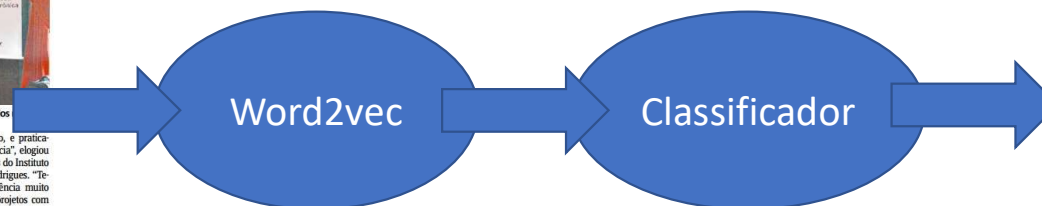
Pereira explicou como as empresas poderão fazer o aporte dos

recursos e na execução dos projetos". Ele também apontou que o Rio Grande do Sul é o estado com menor participação no escopo de fomento da entidade, apesar de apresentar um mercado com "grande potencial" para aproveitar o incentivo.

Destinado a projetos ligados a indústrias, agronegócios, cidades inteligentes e áreas da saúde, o programa atende cerca de 10 empresas no Estado, que trabalham em parceria com unidades da Embrapii - a exemplo do Instituto de Pesquisas Eldorado, que desenvolveu módulo de TV digital em parceria com a Motorola. "Em geral, o processo de liberação de recursos para os pro-

jetos é bastante rápido, e praticamente não há burocracia", elogiou o gerente de operações do Instituto Eldorado, Jefferson Rodrigues. "Temos tido uma experiência muito boa em desenvolver projetos com o fomento deste programa".

"Nosso setor está ligado diretamente com IoT e Indústria 4.0, o que nos faz ter todo o interesse de incentivar empresas a aderirem ao programa de fomento da entidade", destaca o presidente da Abinee, Regis Sell Haubert. "Além disso, precisamos agregar valor, para diminuir o tempo de desenvolvimento das pesquisas - o que ainda é um desafio muito grande", completa.



economia  
tecnologia  
política

## 1.2 - Word2vec vs. Bag of Words



## Revendo Bag of Words

- João gosta de assistir filmes.

João	gosta	de	assistir	filmes	Maria	também
1	1	1	1	1	0	0

- Maria também gosta de filmes.

João	gosta	de	assistir	filmes	Maria	também
0	1	1	0	1	1	1

# One-hot-encoding

- João.

João	gosta	de	assistir	filmes	Maria	também
1	0	0	0	0	0	0

- Bag-of-words: One Hot Encoding, Vetores de Frequência, TF/IDF.

# Word2vec vs. bag-of-words

- One-hot-encoding: João.

1	0	0	0	0	0	0
---	---	---	---	---	---	---

- Word2vec: João

0.133	0.231	0.009	0.878	0.764	0.189	0.320
-------	-------	-------	-------	-------	-------	-------

# Word2vec vs. bag-of-words

- One-hot-encoding: João.

João	gosta	de	assistir	filmes	Maria	também
1	0	0	0	0	0	0

- Word2vec: João

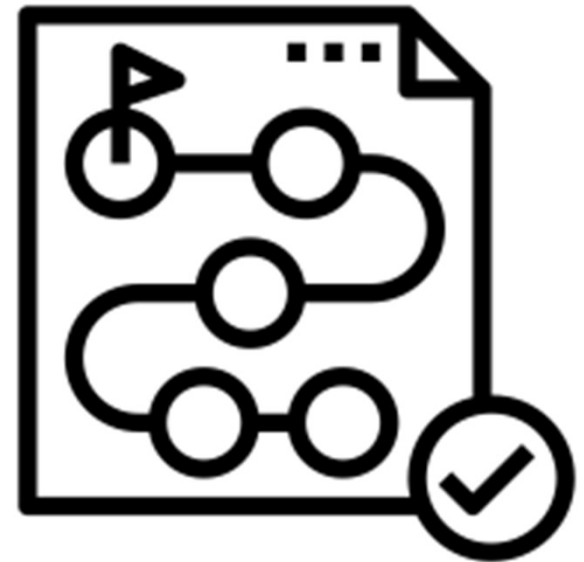
objeto	conceito	feminino	masculino	pessoal	sentimento	lugar
0.133	0.231	0.009	0.878	0.764	0.189	0.320

## 2. Computar vetores para comparar e encontrar conceitos

## 2.1 - Métodos de construção do Word2vec

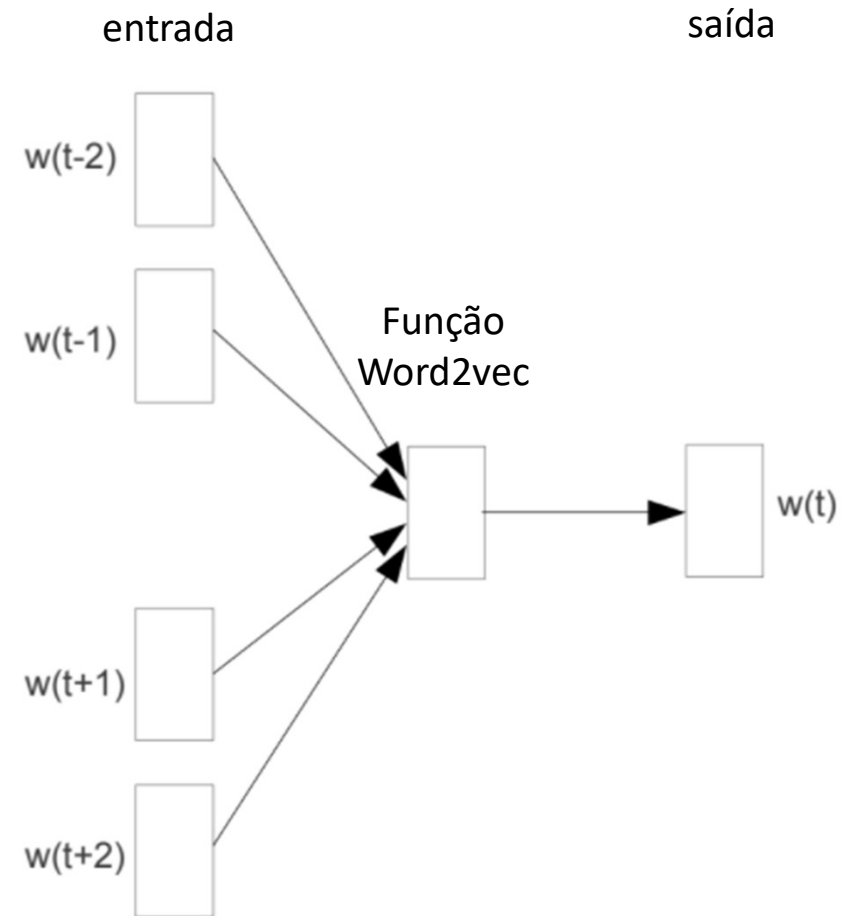
# Métodos da arquitetura

- Word2Vec é um método para construir a representação Word Embedding.
- Ele pode ser obtido usando dois métodos (ambos envolvendo Redes Neurais):
  - Continuous Bag Of Words (CBOW)
  - Skip Gram



# CBOW

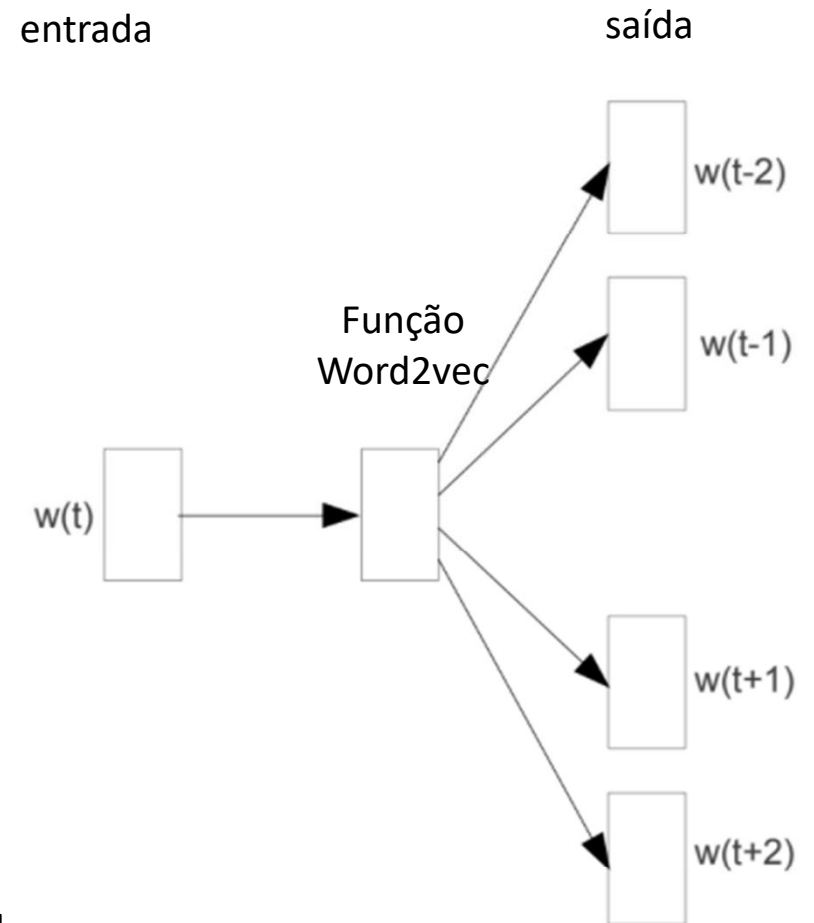
- Este método é utilizado para representar a palavra central de uma sentença.
- “O cachorro correu atrás do gato”
- ([palavras contexto], palavra alvo)
  - ([O, correu], cachorro)
  - ([cachorro, atrás], correu)
  - ([correu, do], atrás)
  - ([atrás, gato], do)
- Aprende a palavra a depender dos contextos





# Skip-gram

- Este método é utilizado para representar os contextos de uma palavra central de uma sentença.
- “O cachorro correu atrás do gato”.
- (palavra alvo, [palavras contexto])
  - (cachorro, [O, correu])
  - (correu, [cachorro, atrás])
  - (atrás, [correu, do])
  - (do, [atrás, gato])
- Aprende os contextos a depender da palavra.

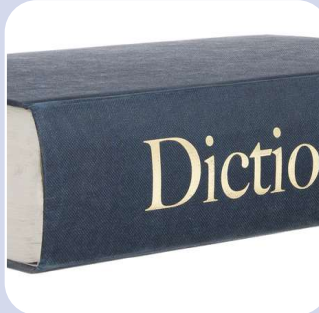


## 2.2 - Reusando modelos Word2vec

# Word vectors tem estrutura simples



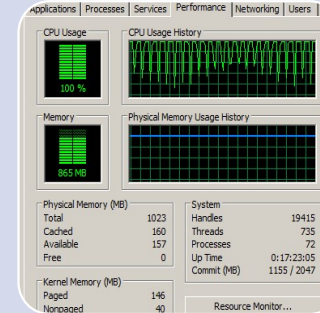
Word vectors  
podem ser  
usados  
independente  
do modelo de  
treinamento  
*word  
embedding*.



São vetores  
chaveados,  
com palavras  
como chaves  
que levam a  
vetores  
numéricos de  
muitas  
posições (100,  
300, 600...)



rainha =  
[0.456, 0.172,  
0.876, 0.732,  
0.122, 0.988,  
0.456, 0.701,  
...]



Usar o modelo  
completo de  
treinamento  
pode te  
permitir  
continuar  
treinando ele,  
mas isso exige  
mais  
computação.



Usar os  
vetores  
é muito  
mais  
prático.

Classificação Com Word2vec

# NILC

- Criado em 1993 para fomentar projetos de pesquisa e desenvolvimento em Linguística Computacional e Processamento de Linguagem Natural.
- Inclui cientistas da computação, linguistas e pesquisadores de diversas universidades e centros de pesquisa, como USP, UFSCar, UNESP, entre outras.
- Eles distribuem *word vectors* já treinados!



# Biblioteca Gensim

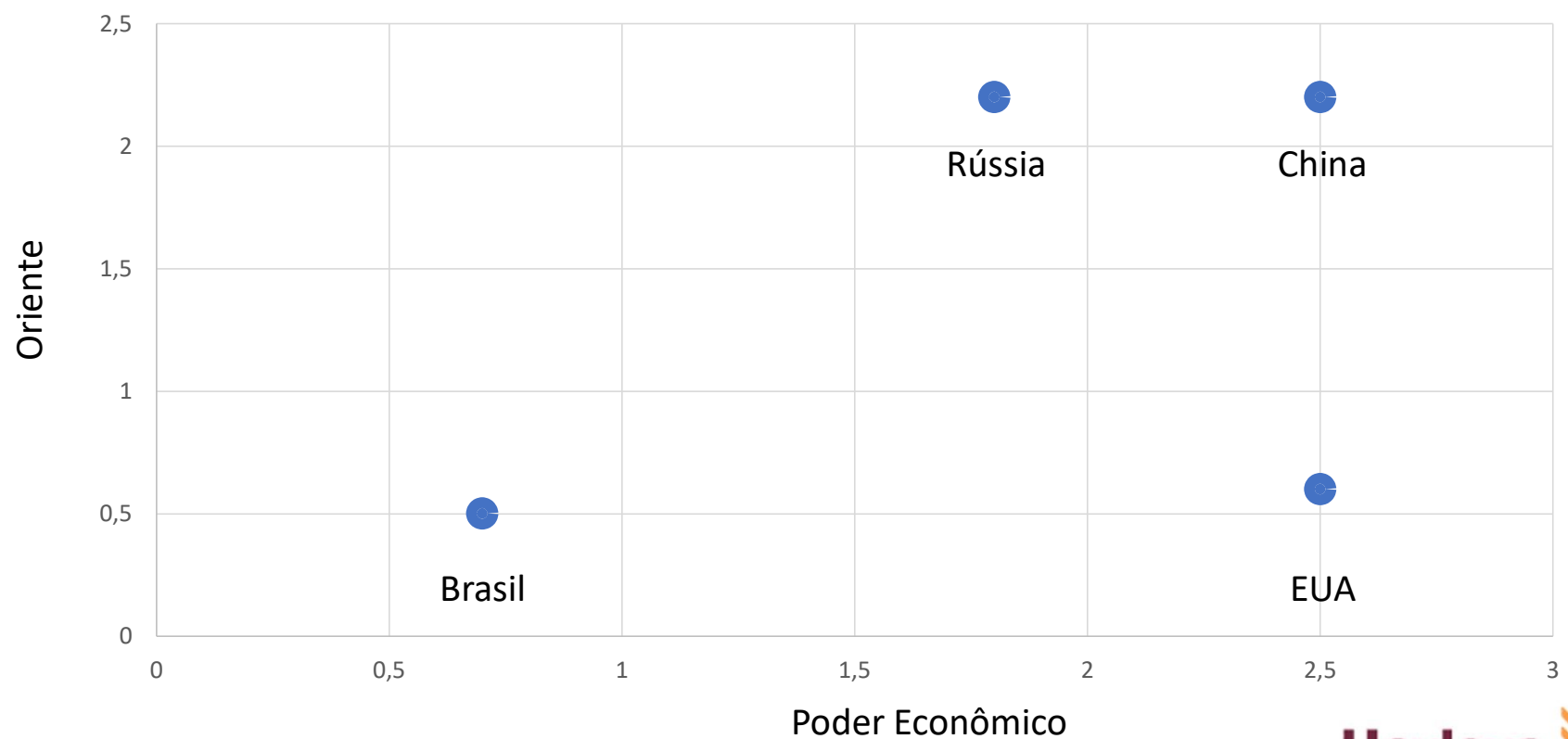
- Uma das bibliotecas mais rápidas para treinamento de *word embeddings*.
- A comunidade Gensim publica modelos pré-treinados para domínios específicos, como jurídico ou saúde, por meio do projeto Gensim-data.
- O NILC treina e publica seus *word vectors usando Gensim*.



### 3. Entender o pré-processamento para usar modelos de representação Word2vec

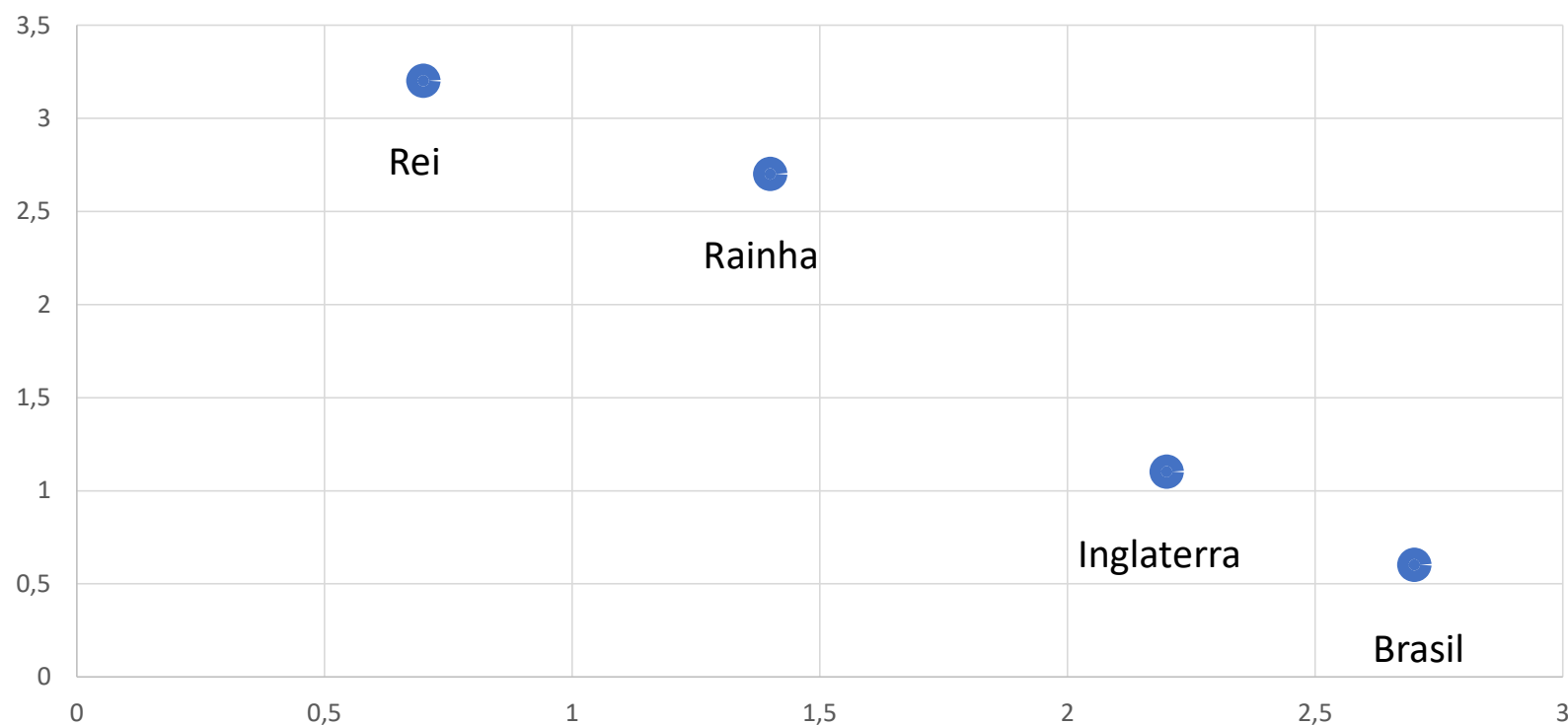
## 3.1 – As dimensões em Word2vec

# Interpretação da distribuição espacial



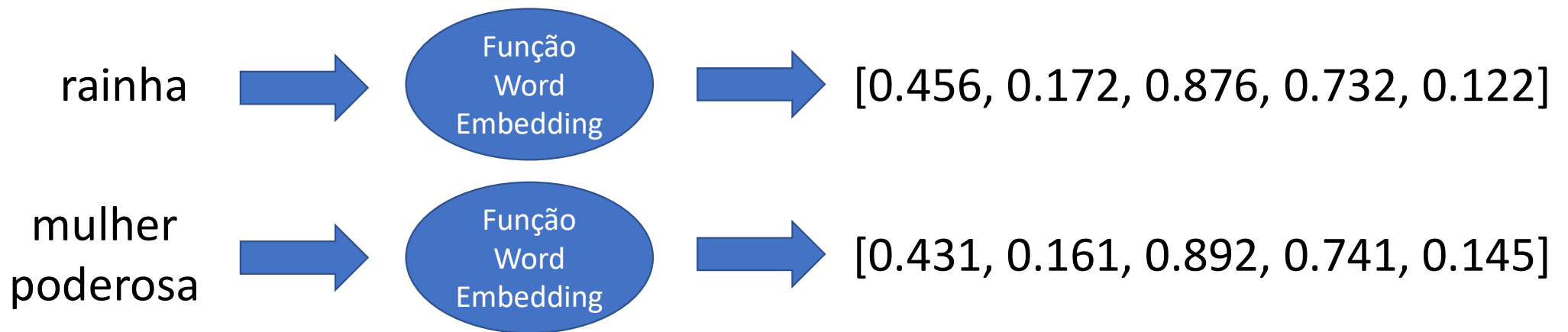


# Interpretação da distribuição espacial



## 3.3 - Combinando vetores de palavras

## Combinando palavras



# Encontrando similares

Brasil + Argentina



Word  
Embedding

1. Chile
2. Peru
3. Venezuela
4. Equador
5. Bolívia
6. Haiti
7. México
8. Paraguai

# Encontrando similares

nuvens + estrela  
- nuvem



Word  
Embedding

1. Estrelas
2. ...
3. ...

## 4. Aprender a classificar com Word2vec

## 4.3 - Comparando classificadores

# Precision e Recall

Valores verdadeiros





É gato

Não é gato

Valores previstos

É gato

Não é gato

 <p>3</p> <p>YOU ARE A CAT</p>	 <p>1</p> <p>YOU ARE A DOG</p>
 <p>2</p> <p>YOU ARE A CAT</p>	 <p>4</p> <p>YOU ARE NOT A CAT</p>



# Precision e Recall

Valores verdadeiros

É gato

Não é gato

Valores previstos

É gato

Não é gato

	É gato	Verdadeiro positivo 3 	Falso negativo 1 
		Falso positivo 2 	Verdadeiro negativo 4 

# Precision

$$\textit{Precisão} = \frac{\textit{Verdadeiro Positivo}}{\textit{Verdadeiro Positivo} + \textit{Falso positivo}}$$

## Recall

$$\text{Revogação} = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso negativo}}$$

# Precision e Recall

Valores verdadeiros

É gato

Não é gato

Valores previstos

É gato

Não é gato

	Valores previstos	
	É gato	Não é gato
Valores verdadeiros	<p>Verdadeiro positivo</p>  <p>3</p>	<p>Falso negativo</p>  <p>1</p>
	<p>Falso positivo</p>  <p>2</p>	<p>Verdadeiro negativo</p>  <p>4</p>

Precision

Recall



Obrigado!