



Formação Inteligência Artificial



Processamento de Linguagem Natural

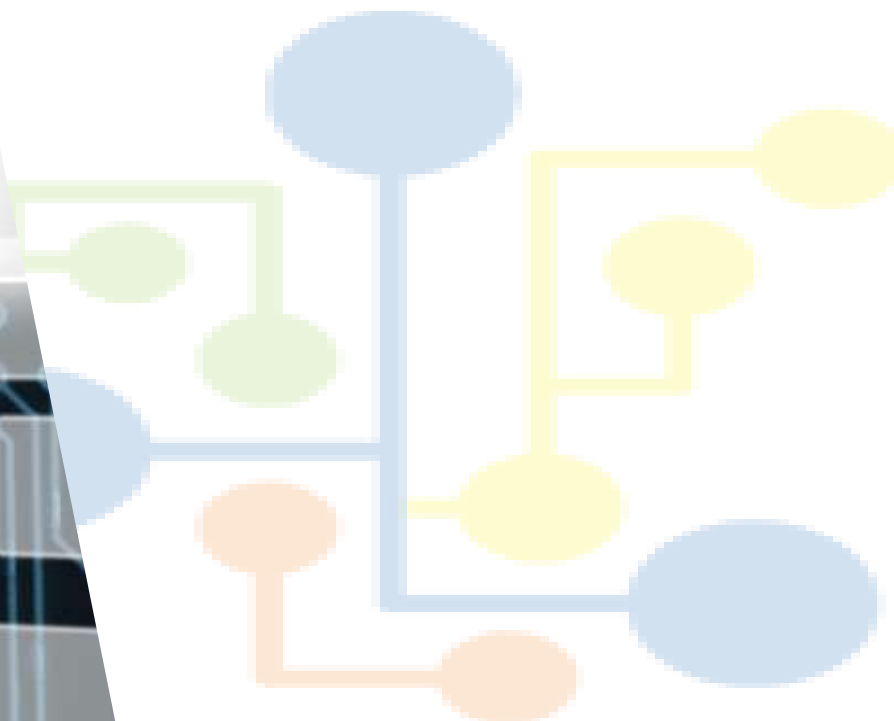




Data Science
Academy

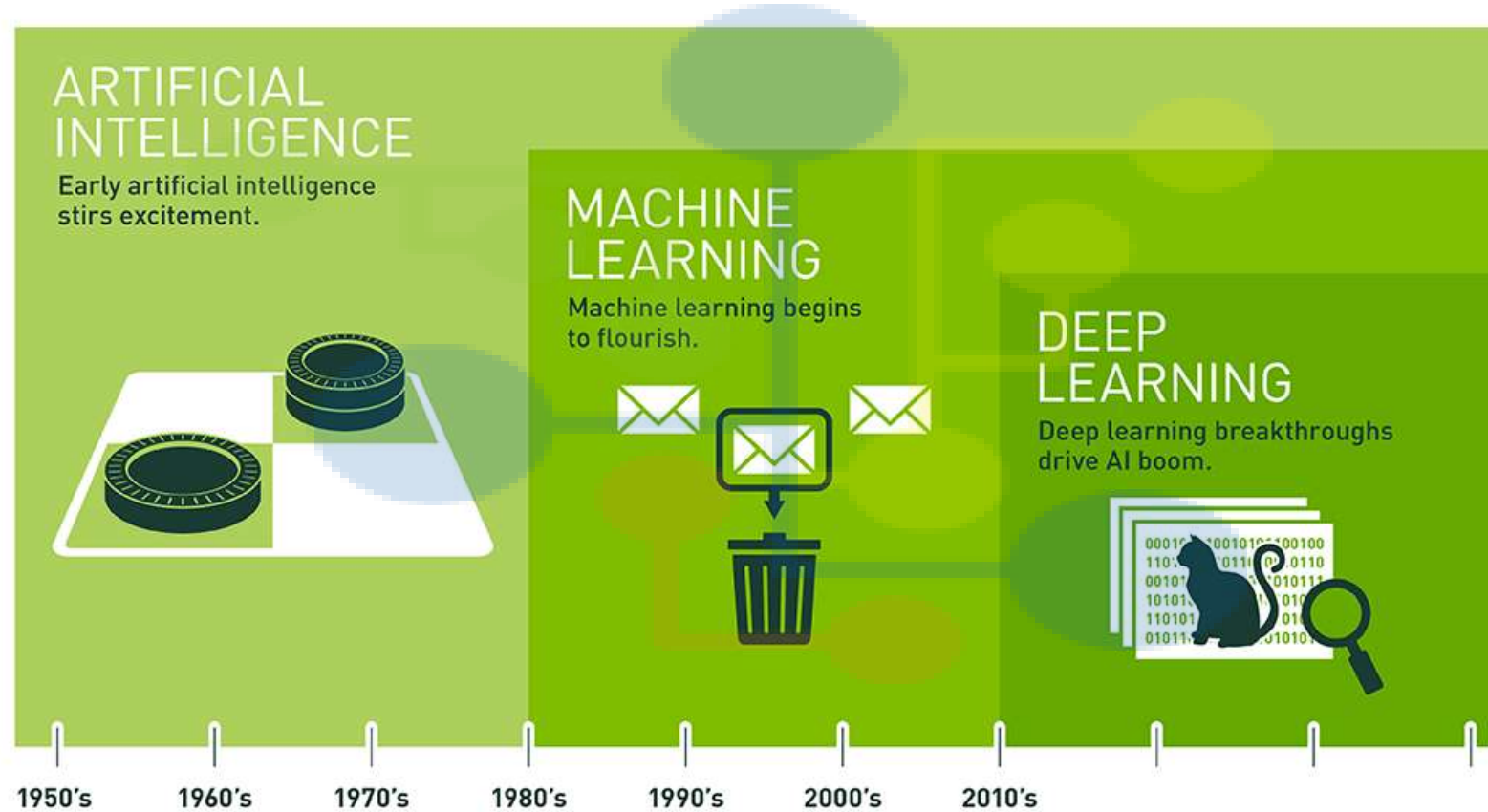
Data Science Academy felipe.oliveiras2000@gmail.com 5f8a0b3ee32fc37d576ba60d

Deep Learning em Processamento de Linguagem Natural





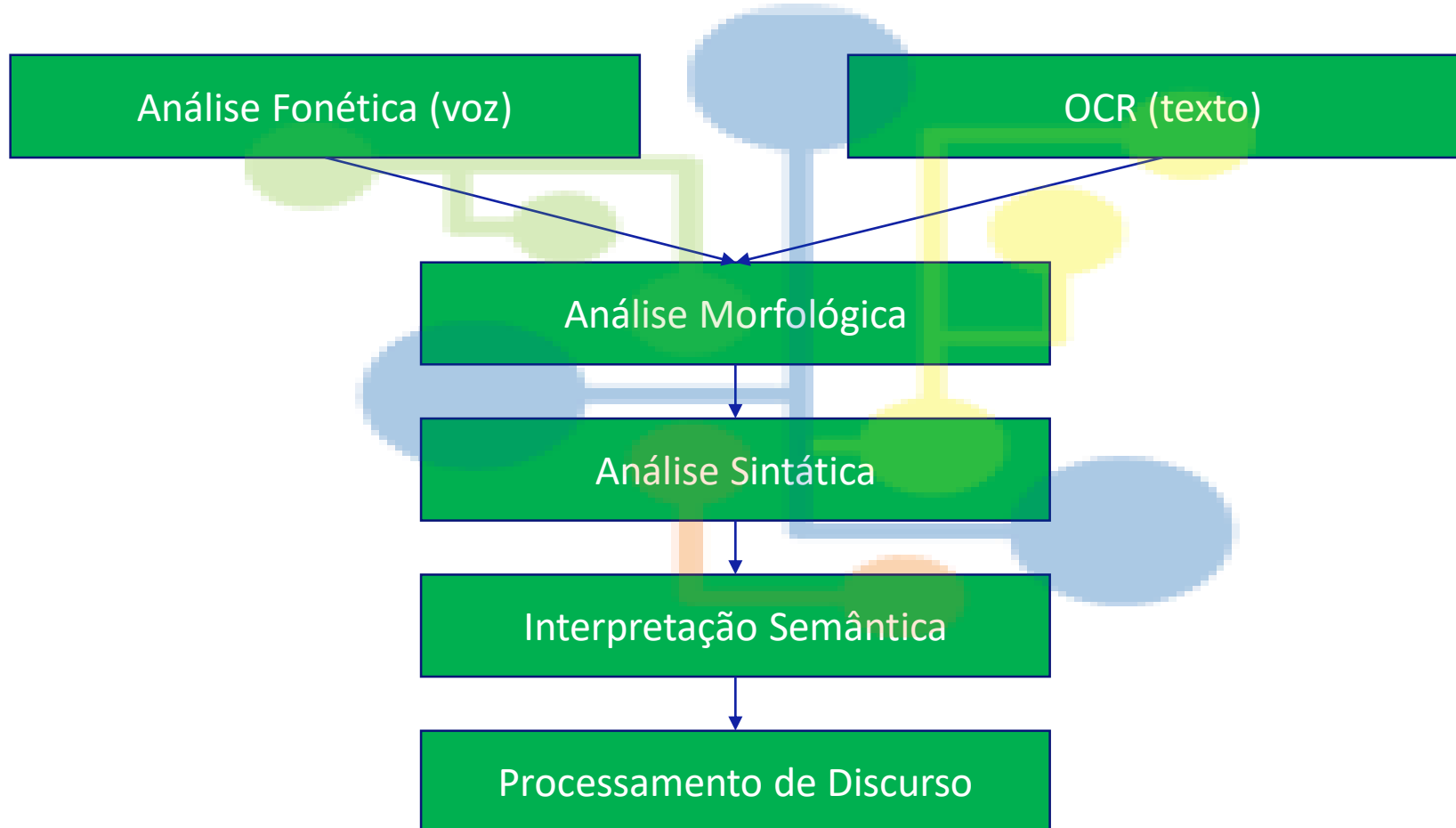
Deep Learning em Processamento de Linguagem Natural



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

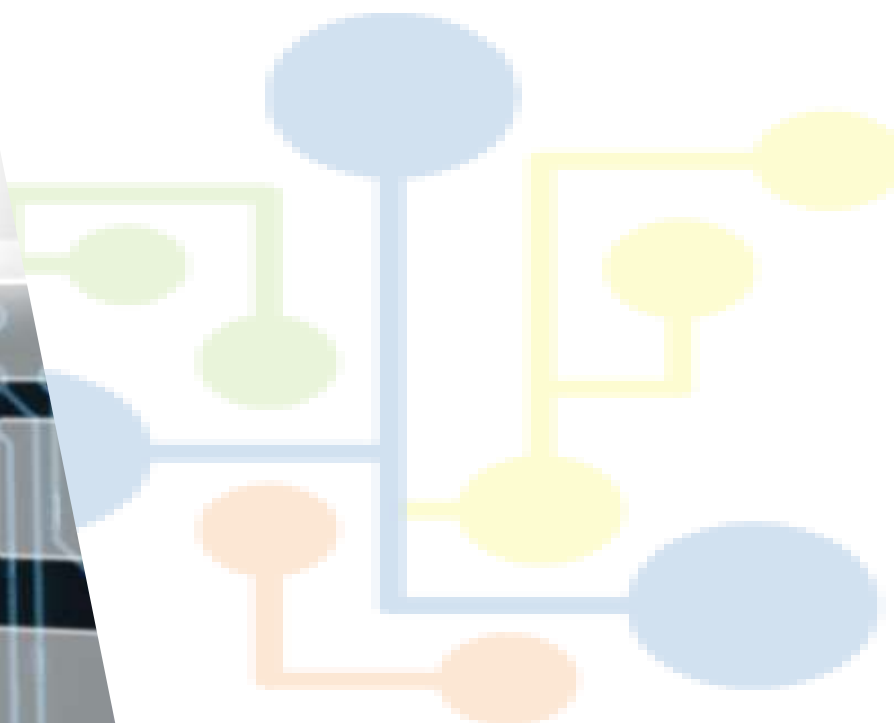


Deep Learning em Processamento de Linguagem Natural





Por que Deep Learning é o Estado da Arte em PLN?





Por que Deep Learning é o Estado da Arte em PLN?

A linguagem é um conjunto de símbolos que podem ser codificados como um “sinal”, que será usado para a comunicação em diferentes formatos:

- Som
- Gestos
- Imagens (texto)

Os símbolos (linguagem) são invariantes em todas as codificações.

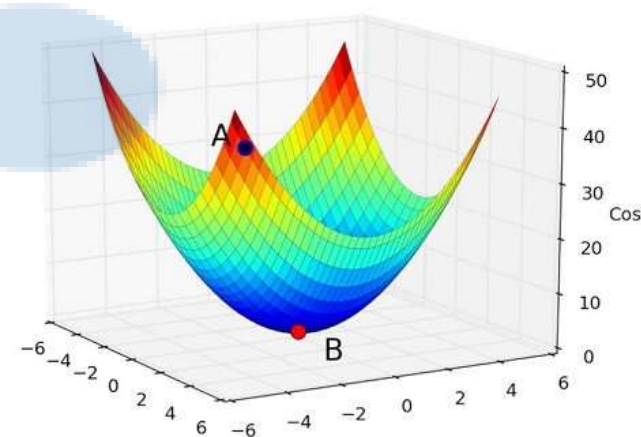




Por que Deep Learning é o Estado da Arte em PLN?

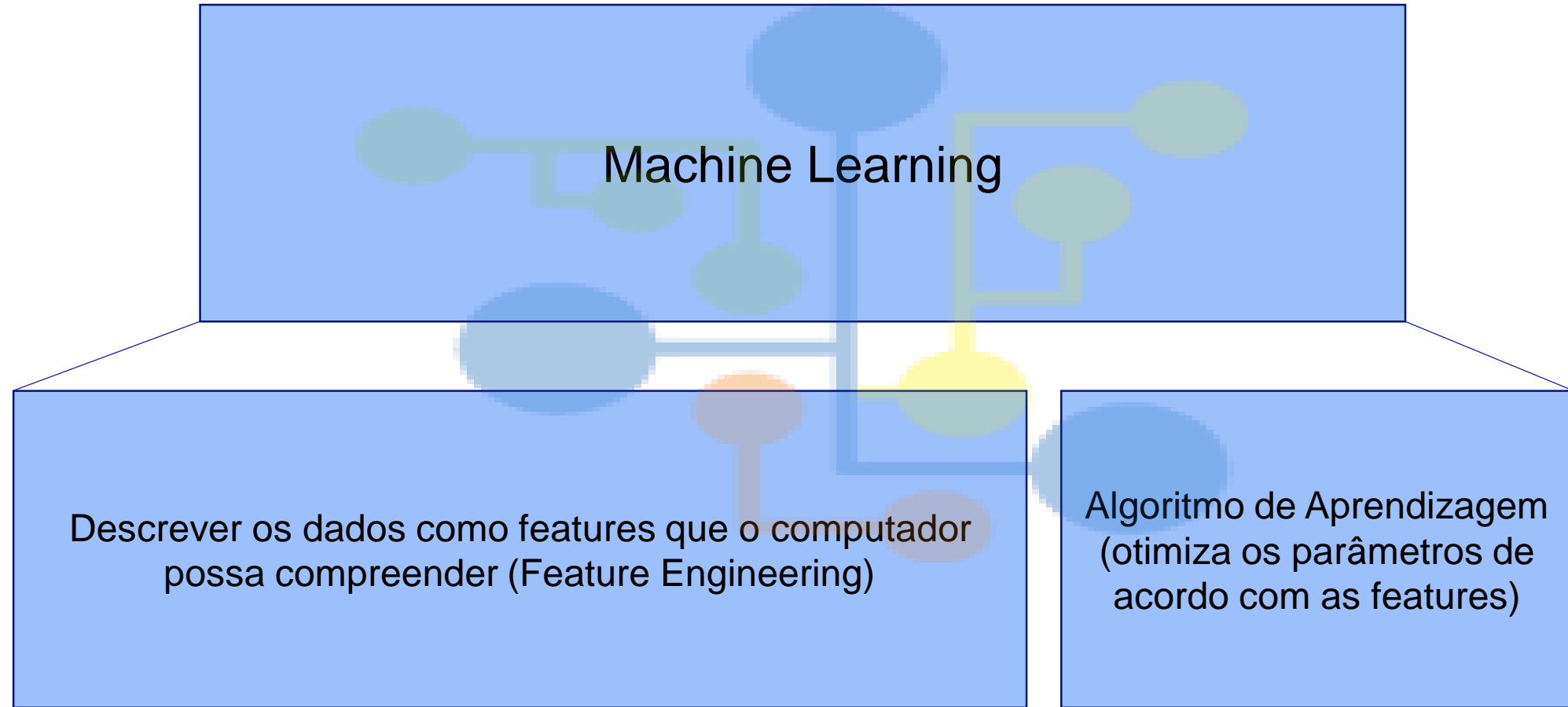
Muitos algoritmos de Machine Learning funcionam bem, porque Cientistas de Dados criam representações das features de entrada, que alimentam os algoritmos.

Machine Learning é, portanto, nada mais do que otimização de parâmetros, que melhor fazem as previsões.





Por que Deep Learning é o Estado da Arte em PLN?

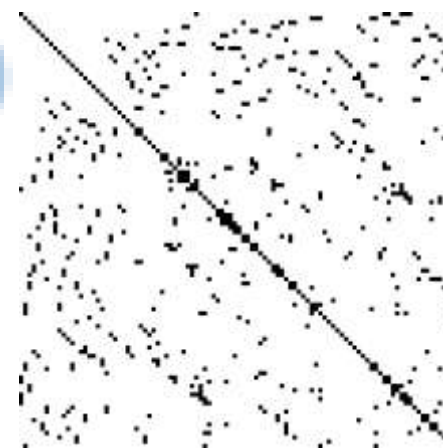




Por que Deep Learning é o Estado da Arte em PLN?

A linguagem humana é um sistema de símbolos, mas a forma como o cérebro funciona, demonstra um padrão contínuo de ativação e os símbolos são transmitidos como sinais contínuos de som/visão.

Quanto maior o vocabulário, a codificação simbólica das palavras cria um problema para Machine Learning: **esparsidade**.



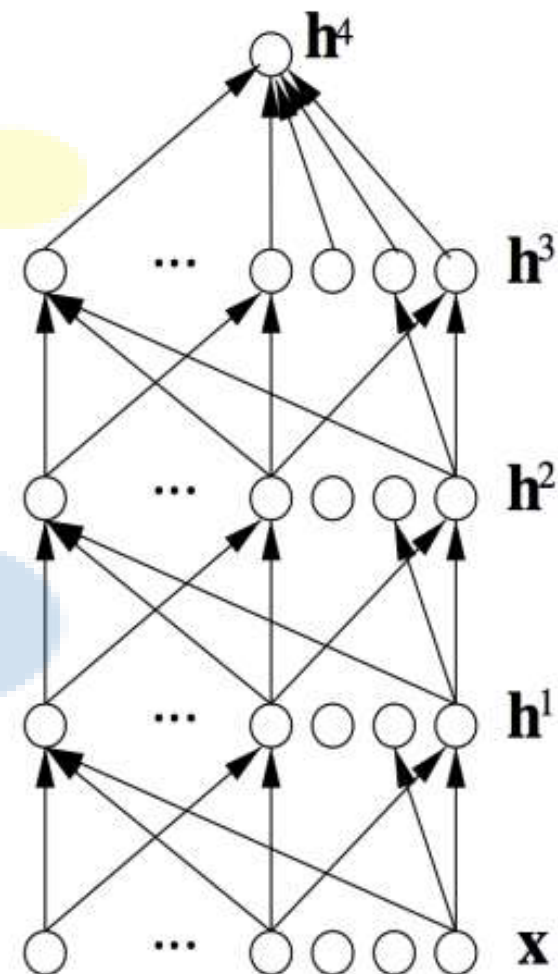


Por que Deep Learning é o Estado da Arte em PLN?

Em contraste a outros métodos de Machine Learning, Deep Learning tenta automaticamente aprender as melhores features.

Algoritmos de Deep Learning tentam aprender múltiplos níveis de representação (h_1 , h_2 , h_3) e a saída h_4 , como no diagrama ao lado.

E isso pode ser feito a partir de som, pixels, caracteres ou palavras.





Por que Deep Learning é o Estado da Arte em PLN?

- Descrever as features manualmente é um processo difícil e que requer muito tempo. Deep Learning resolve este problema aprendendo as features automaticamente.
- O aprendizado automático de features é normalmente mais rápido que o mesmo processo sendo feito manualmente.
- Deep Learning é flexível e pode ser usado para representar o mundo com informação visual e linguística.
- Deep Learning pode aprender de forma supervisionada ou não supervisionada.



Por que Deep Learning é o Estado da Arte em PLN?

A decorative background graphic consisting of a central vertical blue line with several horizontal and diagonal branches in blue, green, and yellow. At the end of these branches are circular nodes in the same color palette. Two large, dark blue rounded rectangular boxes are positioned on either side of the central graphic, containing white text.

Deep Learning requer
muitos dados (Big Data)

O processamento pode
ser longo e o uso de
GPU é quase uma
necessidade.

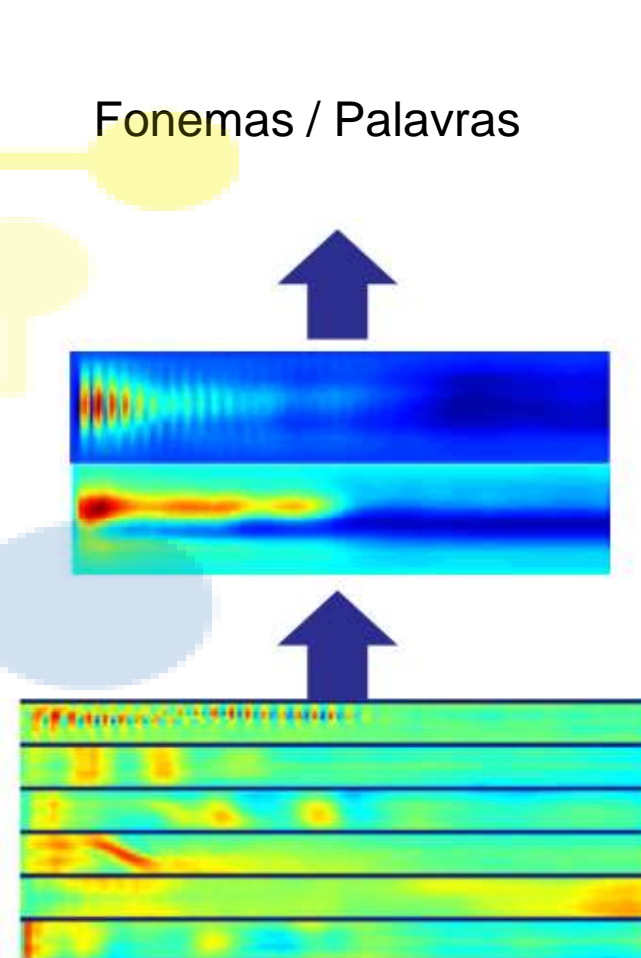


Por que Deep Learning é o Estado da Arte em PLN?

O primeiro grande resultado do Deep Learning em PLN ocorreu no reconhecimento de voz em 2010

Acoustic model and WER	RT03S FSH	Hub5 SWB
Traditional features	27.4	23.6
Deep Learning	18.5 (-33%)	16.1 (-32%)

Fonemas / Palavras

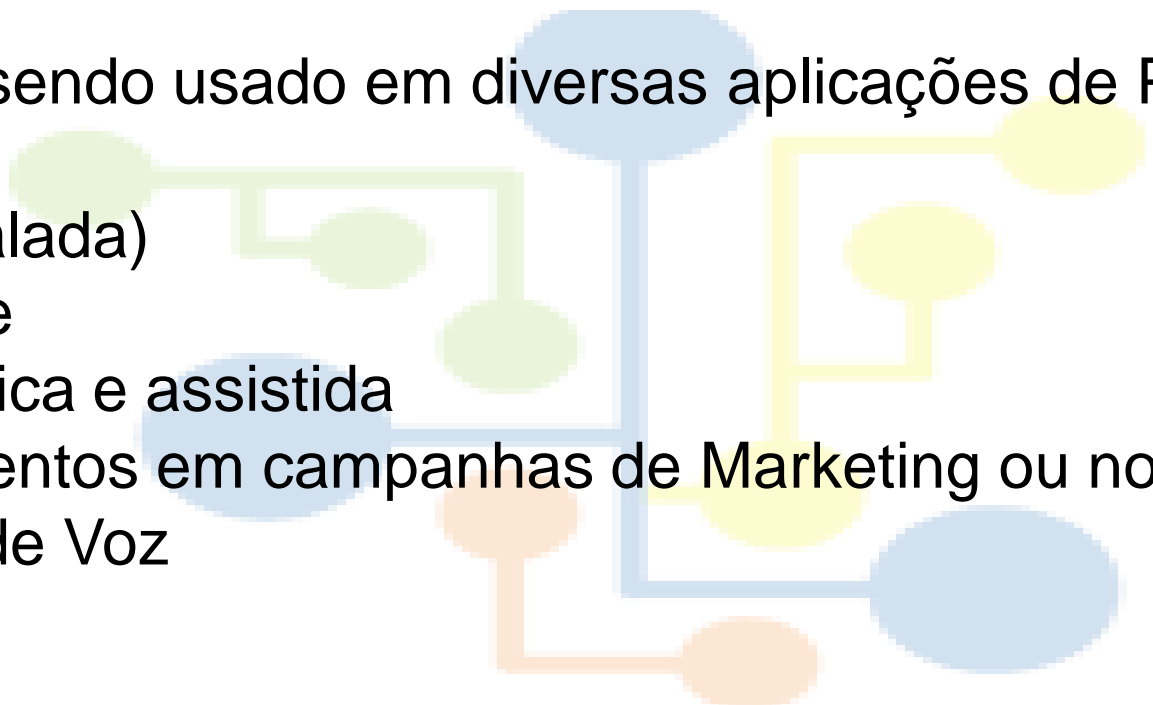




Por que Deep Learning é o Estado da Arte em PLN?

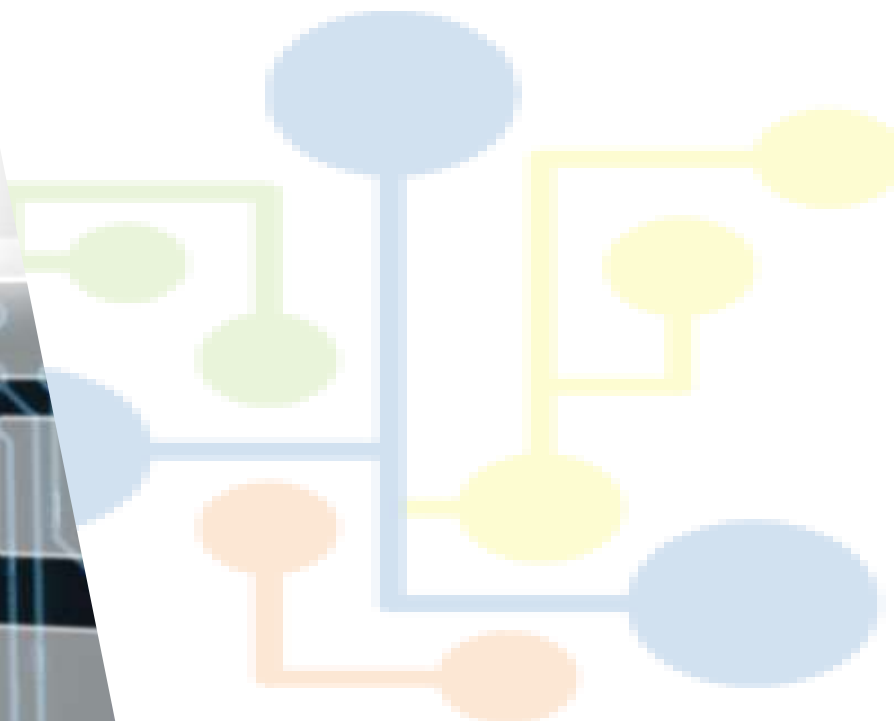
Deep Learning vem sendo usado em diversas aplicações de PLN, tais como:

- Busca (escrita e falada)
- Propaganda online
- Tradução automática e assistida
- Análise de sentimentos em campanhas de Marketing ou no Mercado Financeiro
- Reconhecimento de Voz
- Chatbots



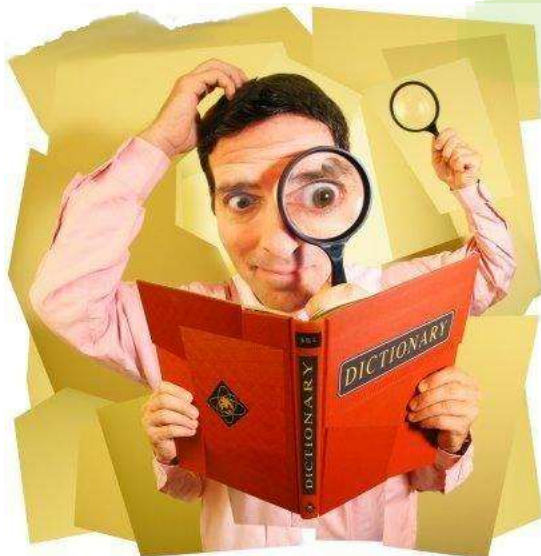


Como Representar o Significado de Uma Palavra?





Como Representar o Significado de Uma Palavra?

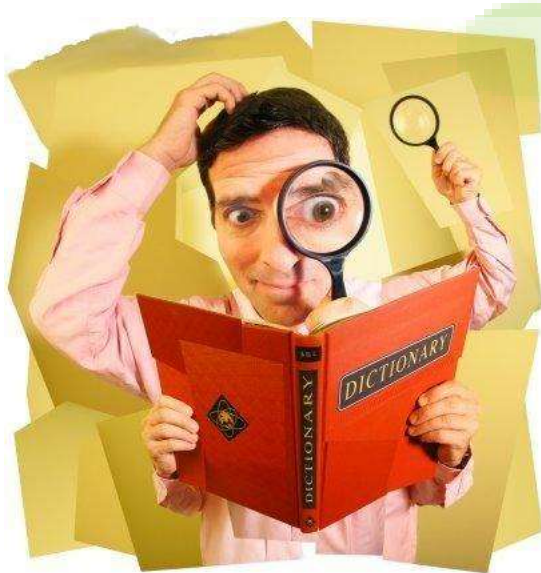


O Que é Significado?

- Definição atribuída a um termo, palavra, frase, texto; aquilo que alguma coisa quer dizer; sentido.
- Relevância que se dá a algo: sua participação teve muito significado.
- [Linguística] Significação; forma representativa e mental que se relaciona com a forma linguística; o que o signo quer significar; a parte do signo linguístico definida pelo conceito.



Como Representar o Significado de Uma Palavra?



O Que é Significado?

- Ideia que é representada por palavra, frase, etc..
- Ideia que uma pessoa quer expressar usando palavras, sinais, etc...
- Ideia que é expressada em um trabalho de escrita, de arte, etc...



Como Representar o Significado de Uma Palavra?

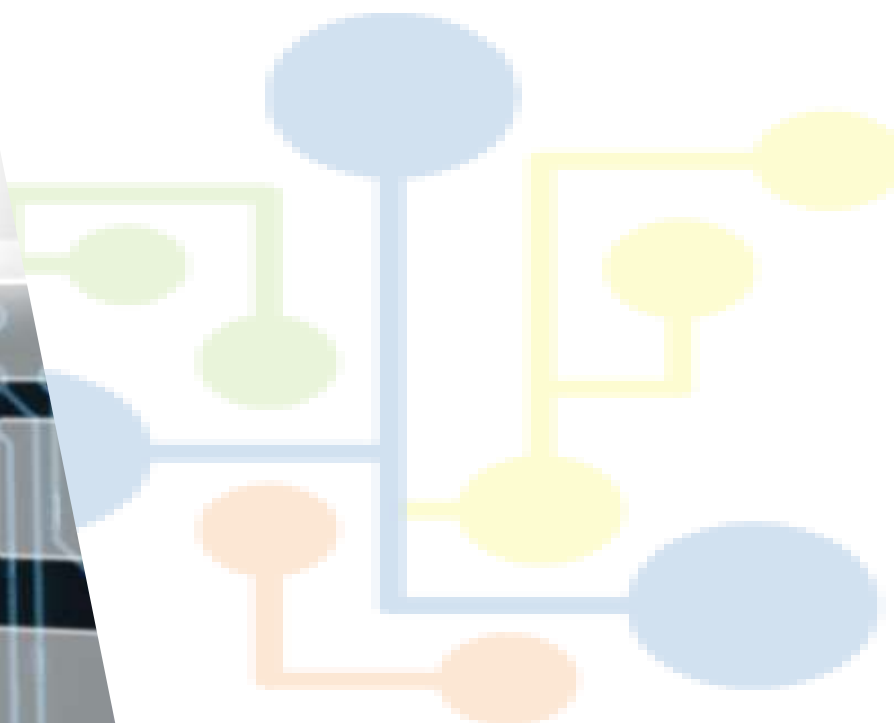
Wordnets

Símbolos Discretos
(One-Hot Encoding)

Contexto
(Word2vec)



Significado de Palavras com Wordnets





Significado de Palavras com Wordnets

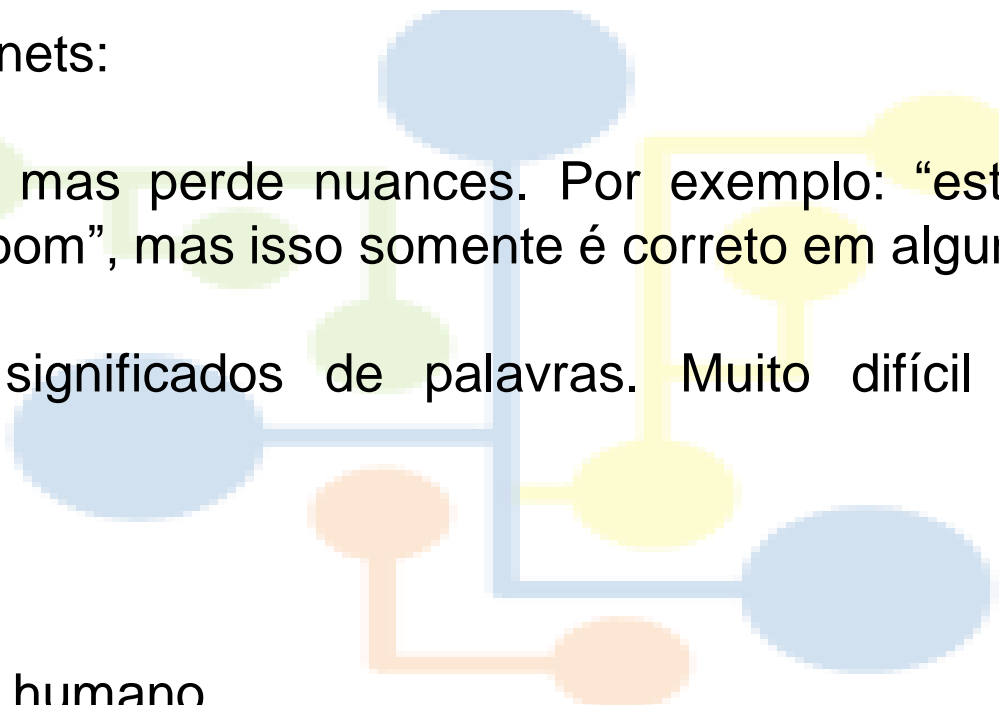
Uma wordnet pode ser entendida como uma base de dados que sistematiza o conjunto dos verbos, substantivos, adjetivos e advérbios de um dado idioma em termos de uma rede de quatro relações: sinonímia, antonímia, hiponímia/hiperonímia e meronímia/holonímia.



Significado de Palavras com Wordnets

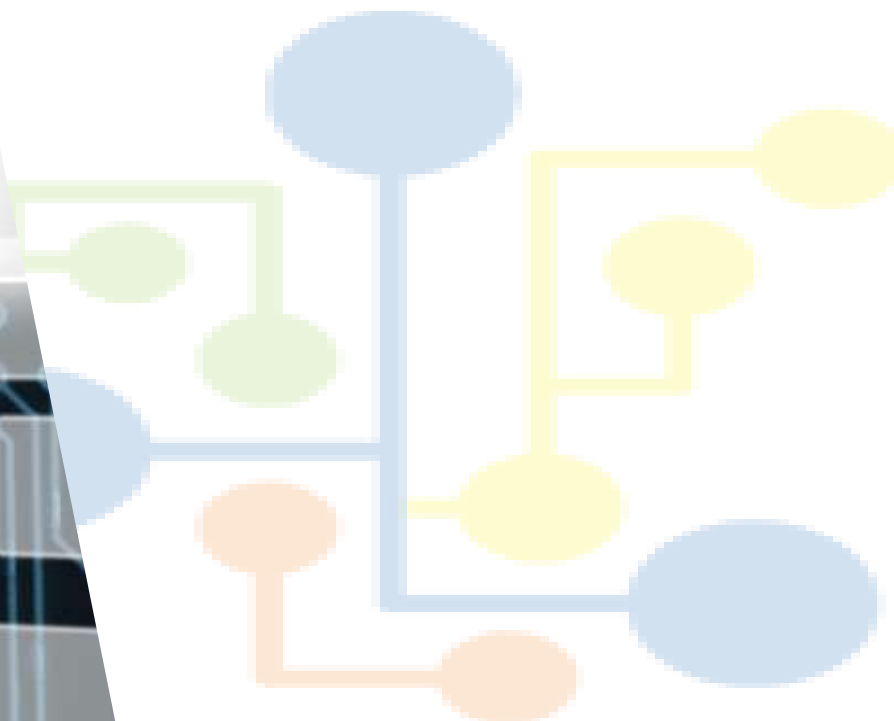
Problemas com as Wordnets:

- Ótimo como recurso, mas perde nuances. Por exemplo: “estimável” é listado como sinônimo para “bom”, mas isso somente é correto em alguns contextos.
- Não contém novos significados de palavras. Muito difícil de ser mantido atualizado.
- Subjetivo
- Requer muito trabalho humano.
- Difícil de computar a acurácia na similaridade entre as palavras.



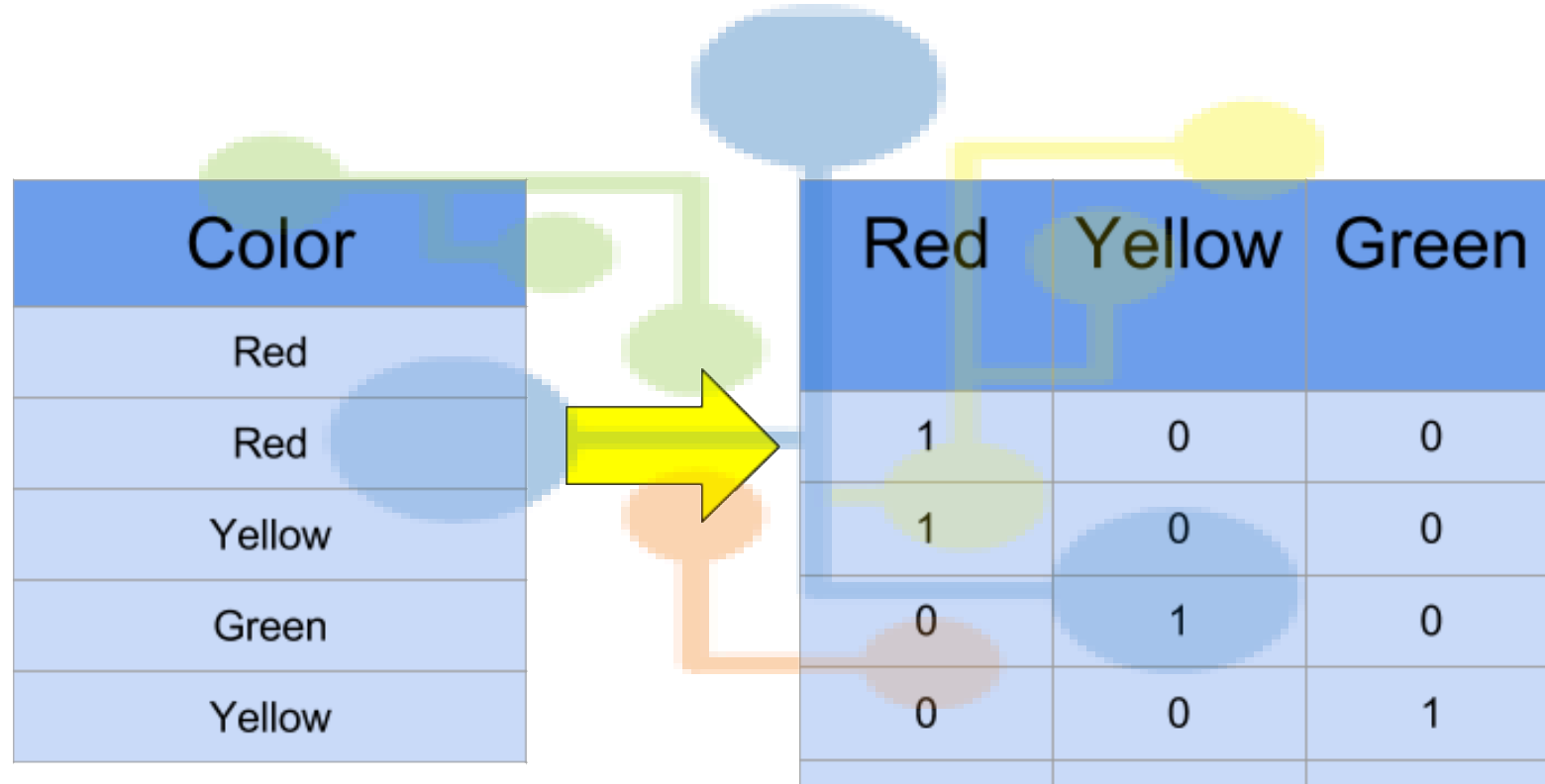


Significado de Palavras com One-Hot Encoding



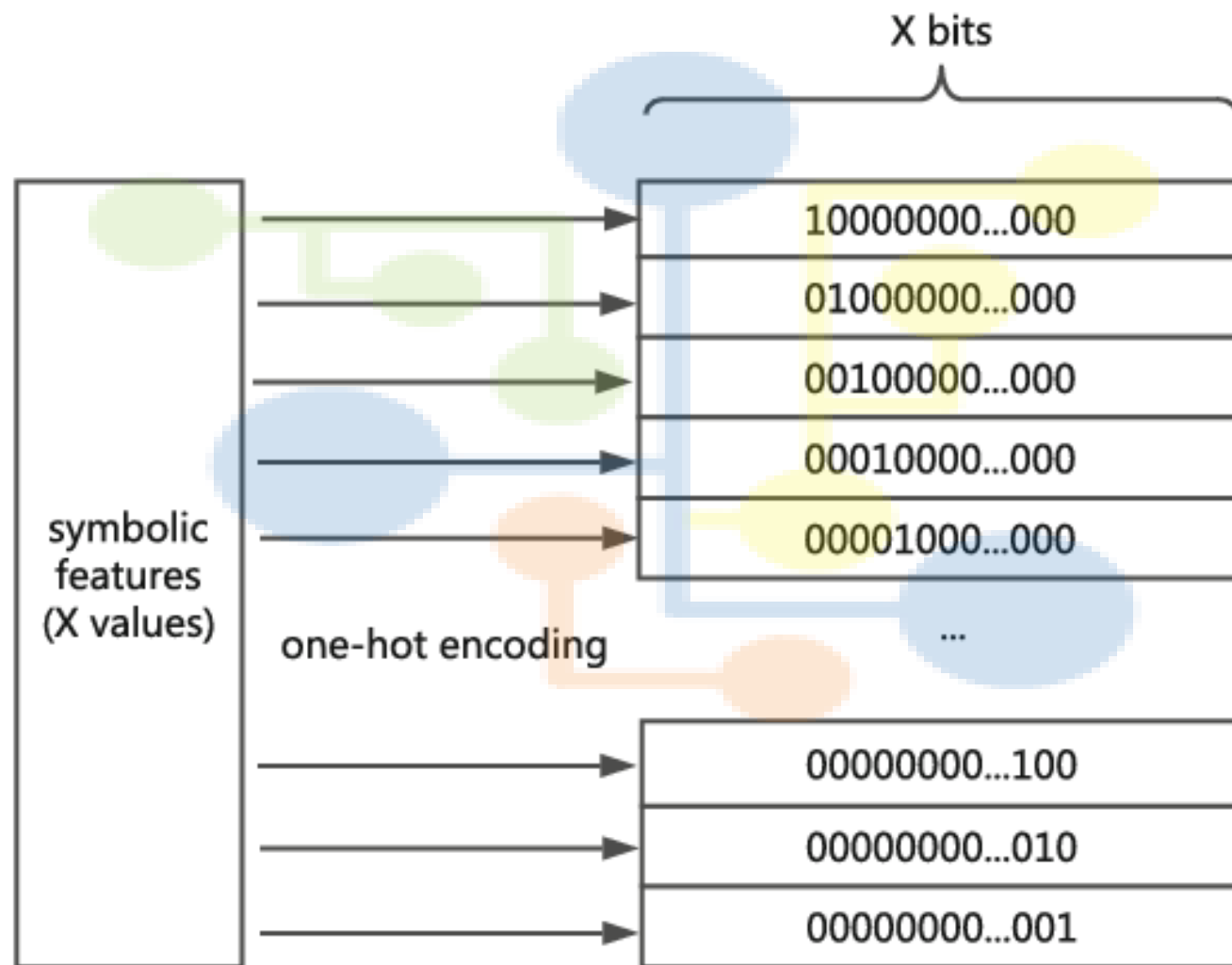


Significado de Palavras com One-Hot Encoding





Significado de Palavras com One-Hot Encoding





Significado de Palavras com One-Hot Encoding

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Palavras podem ser representadas por vetores one-hot, com o valor 1 representando a palavra e o restante composto de 0.

Dimensão do vetor = número de palavras em um vocabulário (~ 500.000)



Significado de Palavras com One-Hot Encoding

motel = [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

Se criamos um mecanismo de busca, queremos que sempre que o usuário buscar por “Copacabana Motel” a busca também retorne “Copacabana Hotel”. Isso é possível usando vetores one-hot?

Não, porque os vetores são ortogonais e não há similaridade natural entre vetores one-hot!!!



Significado de Palavras com One-Hot Encoding

motel = [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]
hotel = [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0]

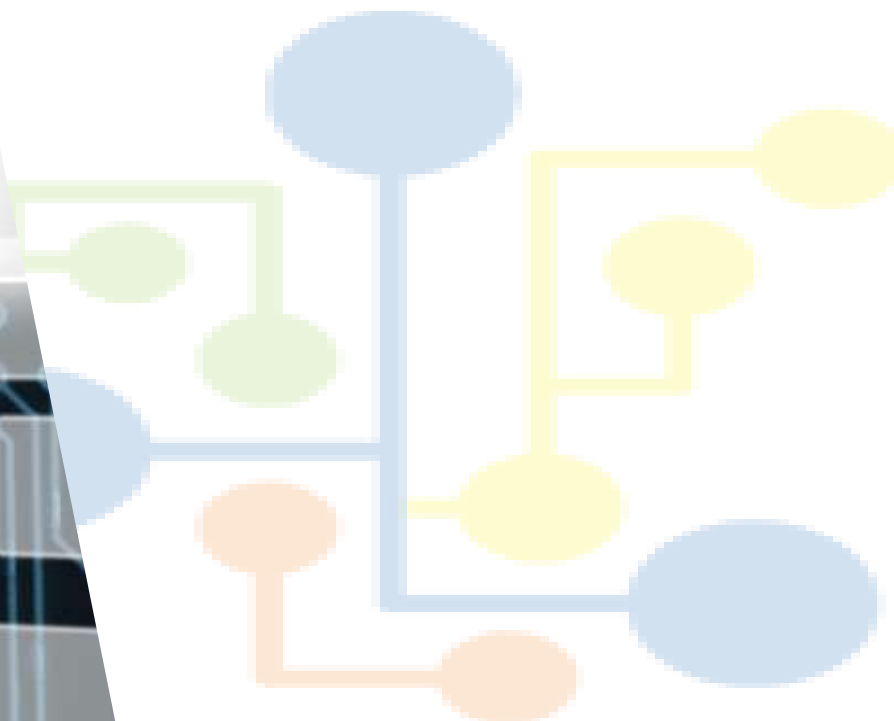
Solução?

Wordnets

Colocar o Encoding da similaridade nos vetores



Significado de Palavras em Contexto (Word Vectors)





Significado de Palavras em Contexto (Word Vectors)

O significado de uma palavra é dado pelas palavras que frequentemente estão próximas.

“You shall know a word by the company it keeps” (J. R. Firth 1957)

Esta é uma das ideias de maior sucesso em PLN Estatístico moderno.



Significado de Palavras em Contexto (Word Vectors)

Quando uma palavra w aparece em um texto, seu significado é representado pelo conjunto de palavras mais próximas (em uma janela fixa).

...problemas de débitos do governo causaram uma crise **bancária** que aconteceu em 2015...
...a comunidade Européia deve unificar a regulação **bancária** para substituir as leis atuais...
...China deu um tiro no pé em seu sistema **bancário** com as novas medidas...

O setor bancário terá significado a partir das palavras antes e depois em uma janela fixa. Isso permite representar o significado de uma palavra em PLN.



Significado de Palavras em Contexto (Word Vectors)

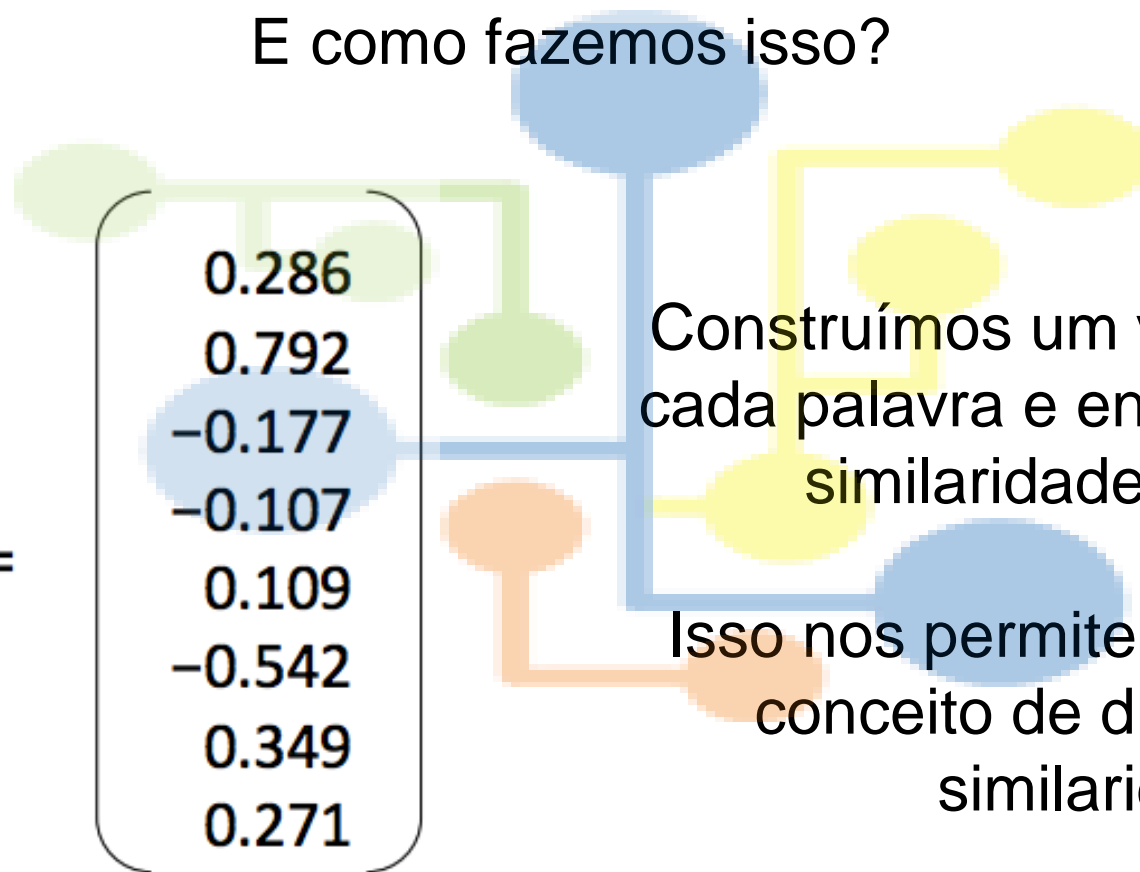
E como fazemos isso?

Linguística =

0.286
0.792
-0.177
-0.107
0.109
-0.542
0.349
0.271

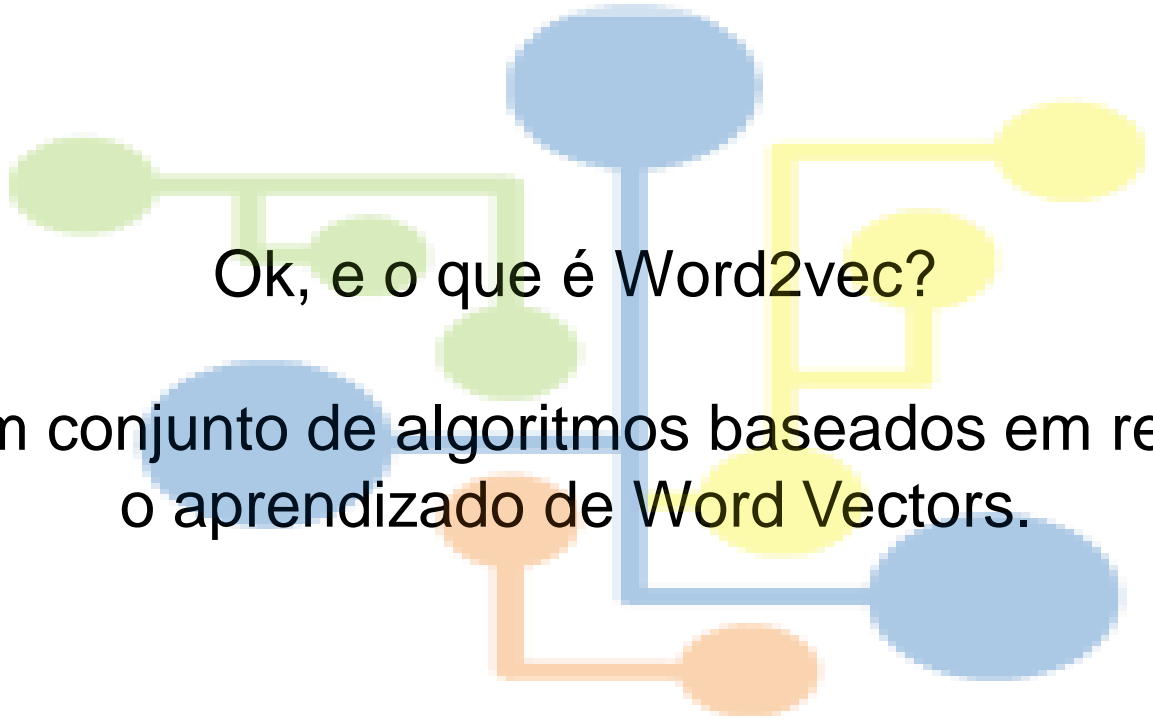
Construímos um vetor denso para cada palavra e então calculamos a similaridade entre eles.

Isso nos permite trabalhar com o conceito de distribuição de similaridades.





Significado de Palavras em Contexto (Word Vectors)

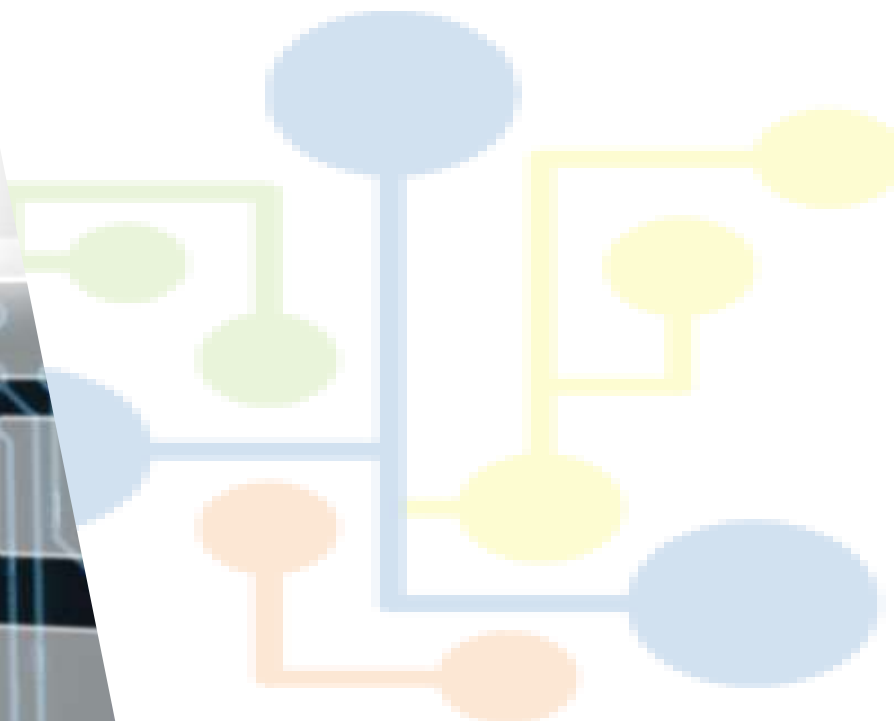
An abstract background diagram consisting of several colored circles (blue, green, yellow, orange) connected by lines, forming a network-like structure. The circles are of varying sizes and are distributed across the slide, with some lines connecting them in a hierarchical or branching manner.

Ok, e o que é Word2vec?

Um framework (um conjunto de algoritmos baseados em redes neurais) para o aprendizado de Word Vectors.



CBOW (Continuous Bag of Words)





CBOW (Continuous Bag of Words)

O CBOW funciona prevendo a probabilidade de uma palavra dada a um contexto. Um contexto pode ser uma única palavra ou um grupo de palavras.

this is sample corpus using only one co

	Hey	This	is	sample	corpus	using
nt 1	1	0	0	0	0	0
nt 2	0	1	0	0	0	0
nt 3	0	0	1	0	0	0
nt 4	0	0	1	0	0	0
nt 5	0	0	0	1	0	0
nt 6	0	0	0	1	0	0
nt 7	0	0	0	0	1	0
nt 8	0	0	0	0	1	0
nt 9	0	0	0	0	0	1
nt 10	0	0	0	0	0	1
nt 11	0	0	0	0	0	0

[illegible]



CBOW (Continuous Bag of Words)

C = “Hey, this is sample corpus using only one context word.”

Input	Output		Hey	This	is	sample	corpus	using	only	one	context	word
Hey	this	Datapoint 1	1	0	0	0	0	0	0	0	0	0
this	hey	Datapoint 2	0	1	0	0	0	0	0	0	0	0
is	this	Datapoint 3	0	0	1	0	0	0	0	0	0	0
is	sample	Datapoint 4	0	0	1	0	0	0	0	0	0	0
sample	is	Datapoint 5	0	0	0	1	0	0	0	0	0	0
sample	corpus	Datapoint 6	0	0	0	1	0	0	0	0	0	0
corpus	sample	Datapoint 7	0	0	0	0	1	0	0	0	0	0
corpus	using	Datapoint 8	0	0	0	0	1	0	0	0	0	0
using	corpus	Datapoint 9	0	0	0	0	0	1	0	0	0	0
using	only	Datapoint 10	0	0	0	0	0	1	0	0	0	0
only	using	Datapoint 11	0	0	0	0	0	0	1	0	0	0
only	one	Datapoint 12	0	0	0	0	0	0	1	0	0	0
one	only	Datapoint 13	0	0	0	0	0	0	0	1	0	0
one	context	Datapoint 14	0	0	0	0	0	0	0	1	0	0
context	one	Datapoint 15	0	0	0	0	0	0	0	0	1	0
context	word	Datapoint 16	0	0	0	0	0	0	0	0	1	0
word	context	Datapoint 17	0	0	0	0	0	0	0	0	0	1

Hey	this	is	sample	corpus	using	only	one	context	word
0	0	0	1	0	0	0	0	0	0

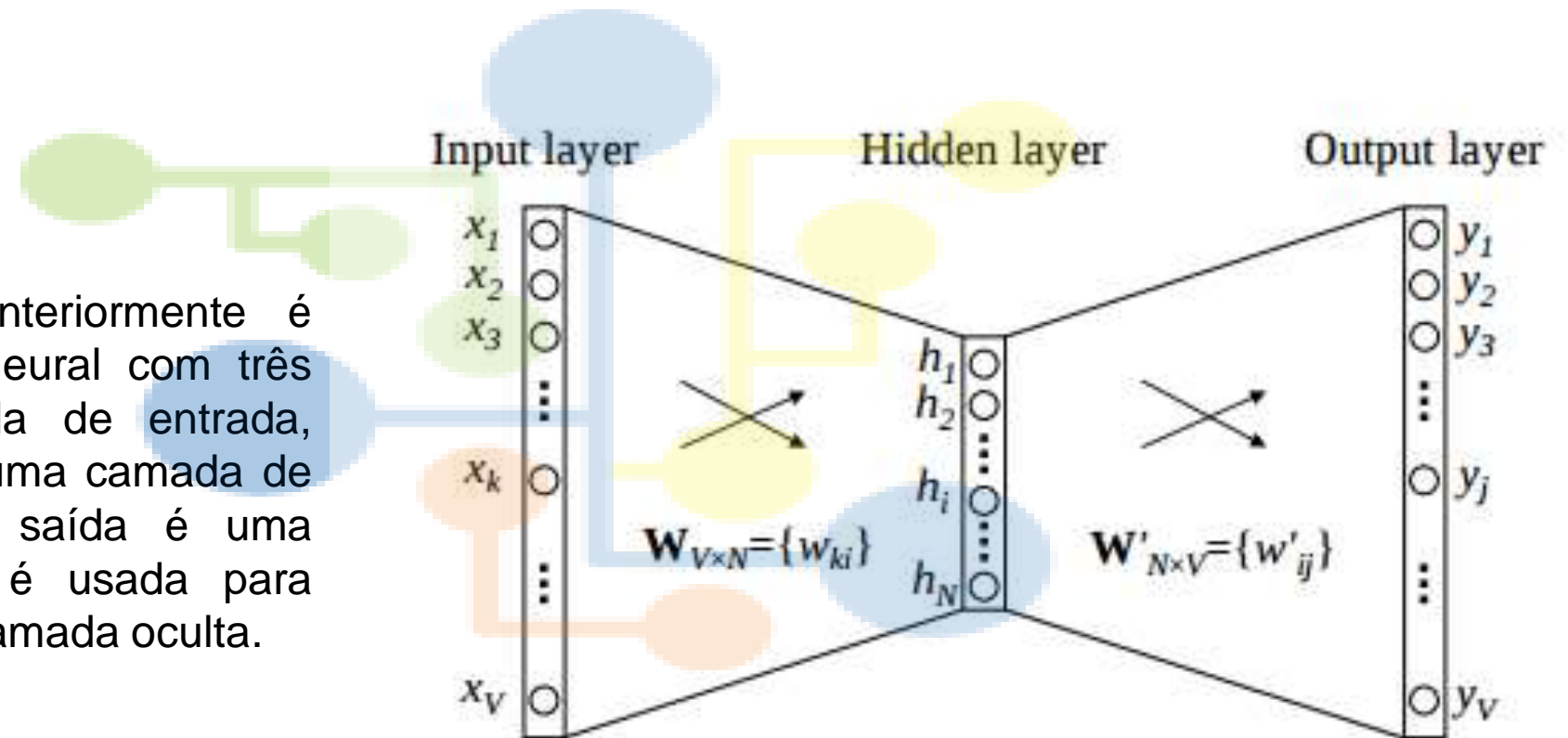


Vetor do Datapoint 4



CBOW (Continuous Bag of Words)

A matriz mostrada anteriormente é enviada a uma rede neural com três camadas: uma camada de entrada, uma camada oculta e uma camada de saída. A camada de saída é uma camada softmax que é usada para calcular a ativação da camada oculta.

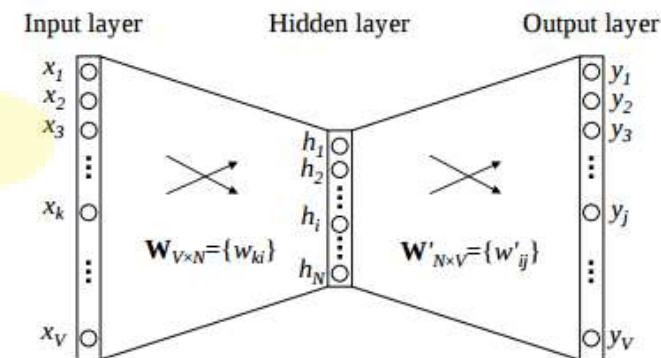




CBOW (Continuous Bag of Words)

O que acontece no treinamento da rede?

1. A camada de entrada e a de saída são “one-hot encoded” de tamanho $[1 \times V]$ e nesse caso $V = 10$.
2. Temos 2 conjuntos de pesos, um entre a camada de entrada e a camada oculta e outro entre a camada oculta e a saída. A camada Entrada/Oculta é uma matriz $[V \times N]$ enquanto a camada Oculta/Saída é uma matriz $[N \times V]$, sendo V o número de dimensões que nós escolhemos para representar a palavra (este é um hiperparâmetro da rede). N é o número de neurônios na rede, nesse caso 4.
3. Não há ativação linear entre as camadas.
4. As entradas são multiplicadas pelos pesos da camada Entrada/Oculta. As saídas da camada oculta são então multiplicadas pelos pesos da camada Oculta/Saída e temos os outputs.
5. O erro da rede é calculado comparando a saída da rede com a saída esperada e então retropropagado de volta na rede para atualizar os pesos na próxima passada.



O peso entre a camada oculta e a camada de saída é considerado como a representação vetorial da palavra ou word vector da palavra.



CBOW (Continuous Bag of Words)

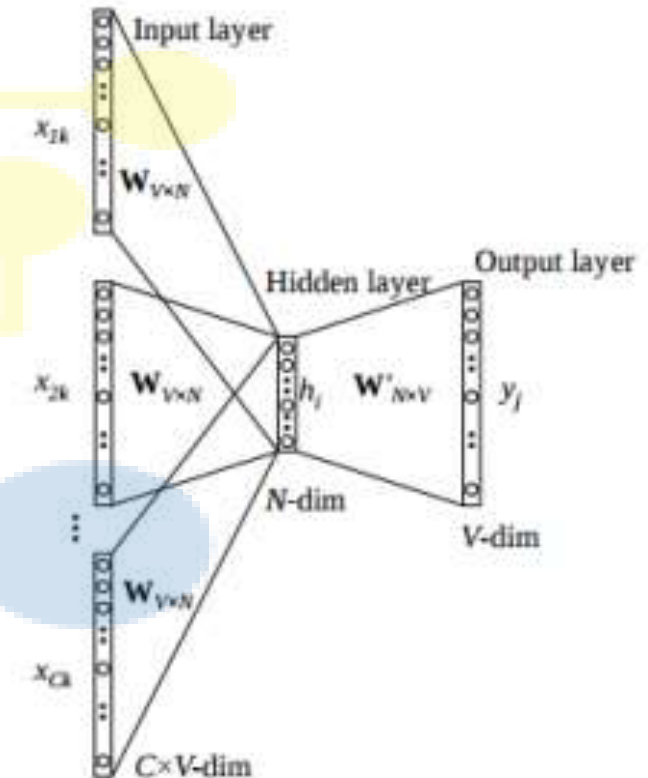
O exemplo anterior foi para apenas um contexto. Mas e quando temos múltiplos contextos para uma mesma palavra?



CBOW (Continuous Bag of Words)

Input	Output		Hey	This	is	sample	corpus	using	only	one	context	word
Hey	this	Datapoint 1	1	0	0	0	0	0	0	0	0	0
this	hey	Datapoint 2	0	1	0	0	0	0	0	0	0	0
is	this	Datapoint 3	0	0	1	0	0	0	0	0	0	0
is	sample	Datapoint 4	0	0	1	0	0	0	0	0	0	0
sample	is	Datapoint 5	0	0	0	1	0	0	0	0	0	0
sample	corpus	Datapoint 6	0	0	0	1	0	0	0	0	0	0
corpus	sample	Datapoint 7	0	0	0	0	1	0	0	0	0	0
corpus	using	Datapoint 8	0	0	0	0	1	0	0	0	0	0
using	corpus	Datapoint 9	0	0	0	0	0	1	0	0	0	0
using	only	Datapoint 10	0	0	0	0	0	1	0	0	0	0
only	using	Datapoint 11	0	0	0	0	0	0	1	0	0	0
only	one	Datapoint 12	0	0	0	0	0	0	1	0	0	0
one	only	Datapoint 13	0	0	0	0	0	0	0	1	0	0
one	context	Datapoint 14	0	0	0	0	0	0	0	0	1	0
context	one	Datapoint 15	0	0	0	0	0	0	0	0	1	0
context	word	Datapoint 16	0	0	0	0	0	0	0	0	1	0
word	context	Datapoint 17	0	0	0	0	0	0	0	0	0	1

Para múltiplas palavras movemos a “janela” pela matriz para obter diferentes contextos. Na figura ao lado, consideramos 3 palavras para prever o contexto de uma única palavra.





CBOW (Continuous Bag of Words)

A fórmula abaixo resume o que é feito no treinamento de um modelo CBOW.

$$p(w_O|w_I) = \frac{\exp(v'_{w_O} \top v_{w_I})}{\sum_{w=1}^W \exp(v'_w \top v_{w_I})}$$

w_O : output word
 w_I : context words



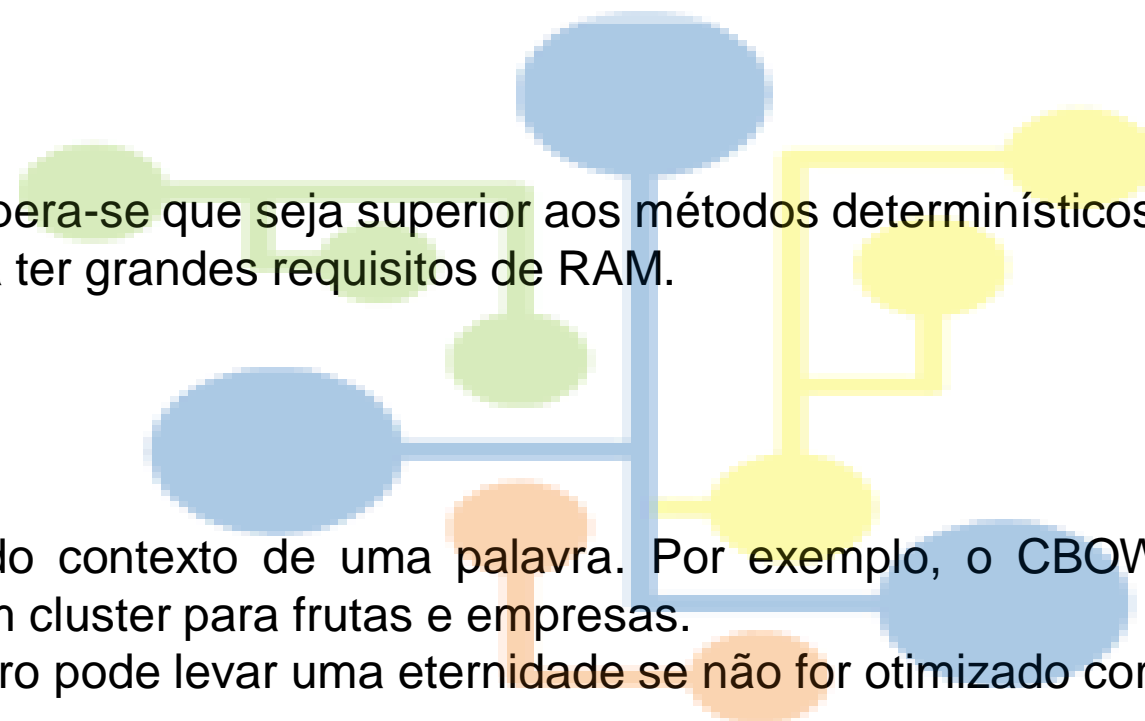
CBOW (Continuous Bag of Words)

Vantagens do CBOW:

- Sendo probabilístico, espera-se que seja superior aos métodos determinísticos (geralmente).
- É lento, mas não precisa ter grandes requisitos de RAM.

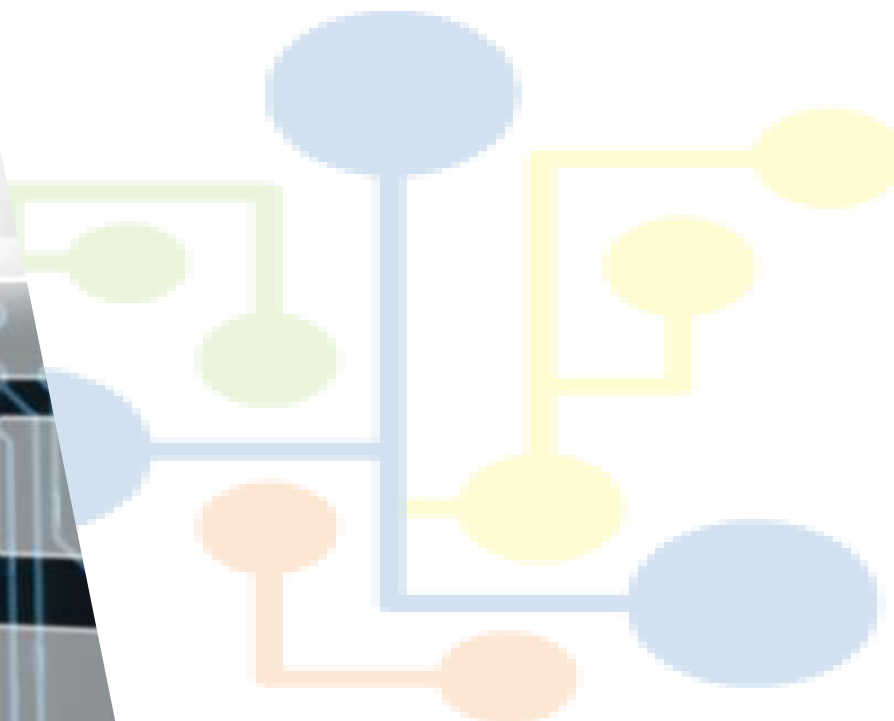
Desvantagens do CBOW:

- CBOW pega a média do contexto de uma palavra. Por exemplo, o CBOW usa uma média de contextos e locais em um cluster para frutas e empresas.
- Treinar um CBOW do zero pode levar uma eternidade se não for otimizado corretamente.





Skip – Gram Model





Skip – Gram Model

O objetivo do Skip-Gram é predizer o contexto, assim como o CBOW. O Skip-Gram segue a mesma topologia do CBOW, mas é inverso a ele, usando palavras de ambos os lados de cada palavra, para representar o contexto.

Vamos pegar o mesmo corpus que usamos para construir nosso modelo CBOW nas aulas anteriores.



Skip – Gram Model

C="Hey, this is sample corpus using only one context word."

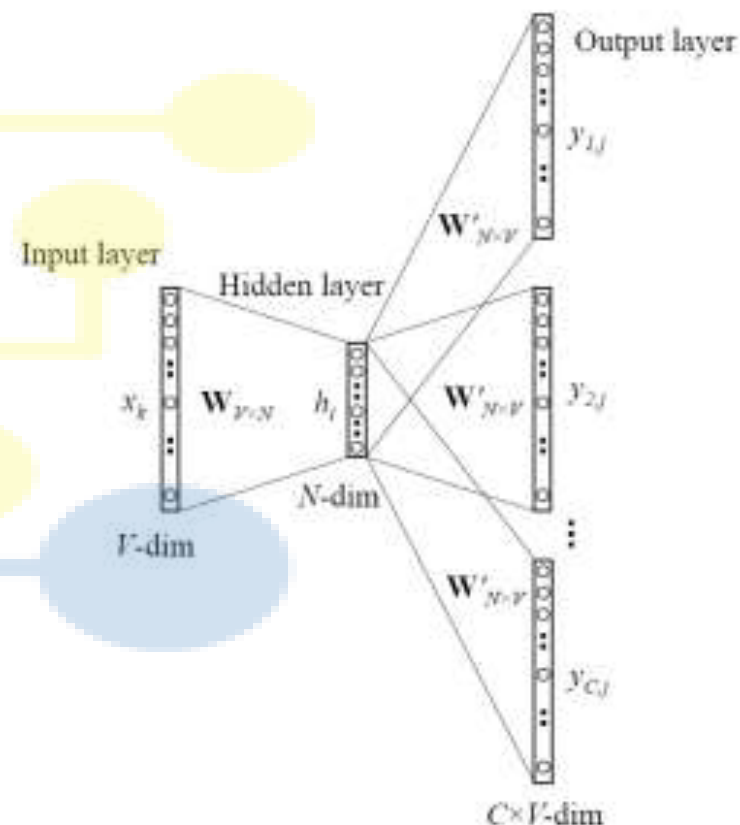
Input	Output(Context1)	Output(Context2)
Hey	this	<padding>
this	Hey	is
is	this	sample
sample	is	corpus
corpus	sample	corpus
using	corpus	only
only	using	one
one	only	context
context	one	word
word	context	<padding>



Skip – Gram Model

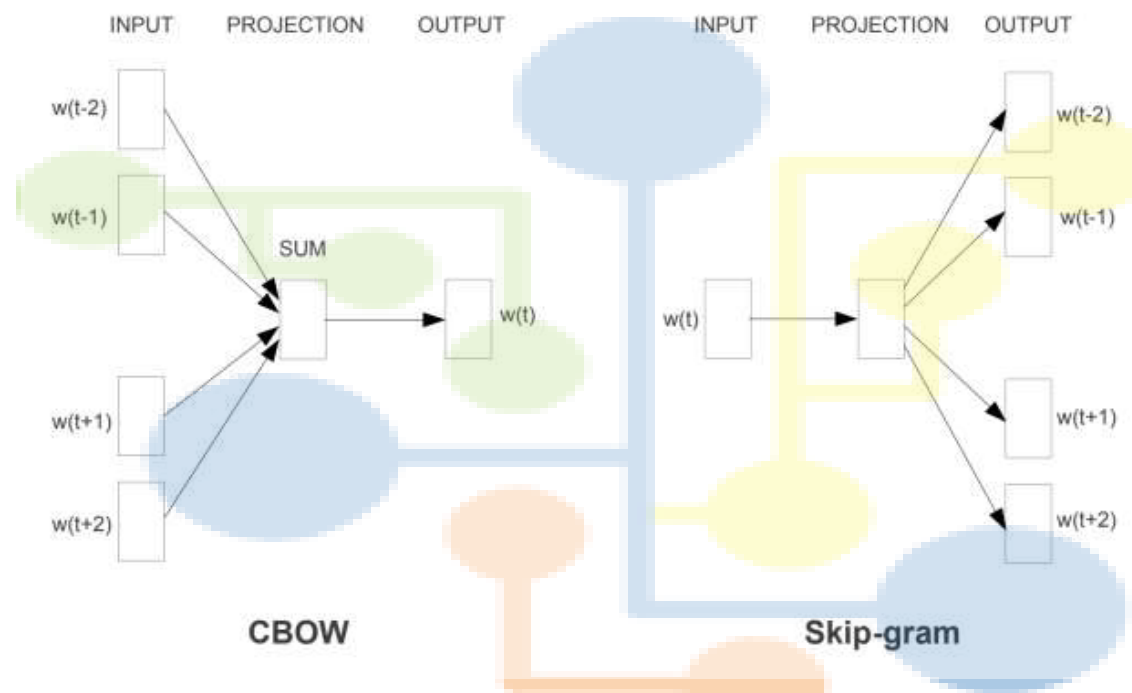
Dois erros separados são calculados em relação às duas variáveis-alvo.

As palavras entre a entrada e a camada oculta são tomadas como a word vector da palavra após o treinamento. A função de perda é a mesma do modelo CBOW.





Skip – Gram Model



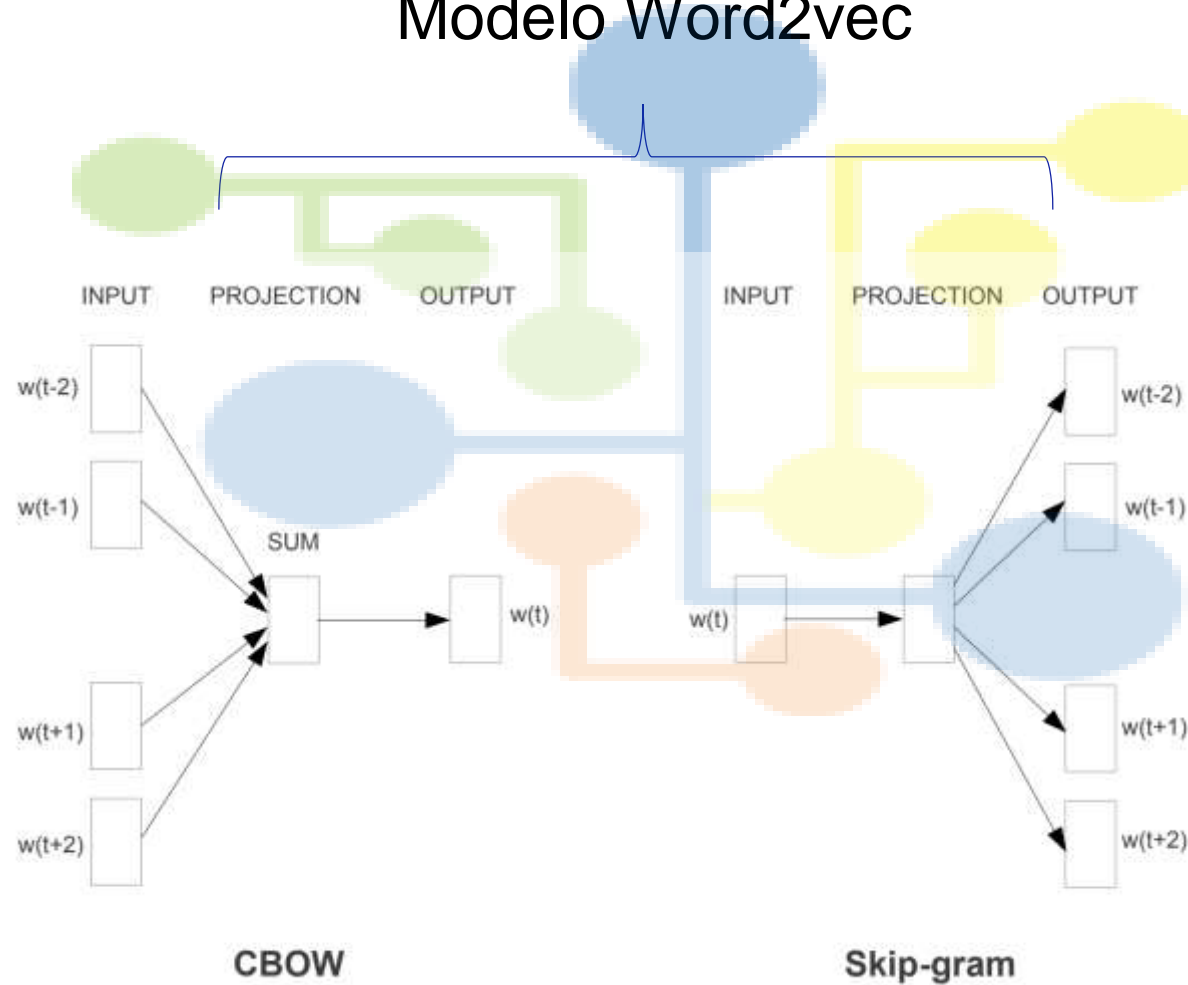
As palavras de contexto formam a camada de entrada. Cada palavra é codificada em uma forma one-hot, portanto, se o tamanho do vocabulário for V teremos vetores de V -dimensões com apenas um dos elementos definidos como 1 e o restante como 0s. Existe uma única camada oculta e uma camada de saída.

O vetor de entrada para Skip-Gram será um modelo CBOW de 1 contexto. Assim, os cálculos até as ativações de camadas ocultas serão os mesmos. A diferença está na variável de saída. Como definimos uma janela de contexto de 1 em ambos os lados, teremos 2 vetores one-hot na variável de saída e, logo, 2 saídas correspondentes.



Skip – Gram Model

Modelo Word2vec





Data Science
Academy

Data Science Academy felipe.oliveiras2000@gmail.com 5f8a0b3ee32fc37d576ba60d

Obrigado