Hackerspace - Mariana Mercucci

Natural Language Processing (NLP)

serasa experian

DATA LAB

Mariana Mercucci

Cientista de Dados Pleno - Experian Datalab

Linkedin: https://www.linkedin.com/in/mariana-mercucci/

De onde veio?
Goiânia

O que estudou?
Bacharela em Física – USP

Tópicos de interesse em DS?
Dados não estruturados, optimização de modelos e fairness.

Hobbies?
Dança, Circo, Filmes...

DATA LAB

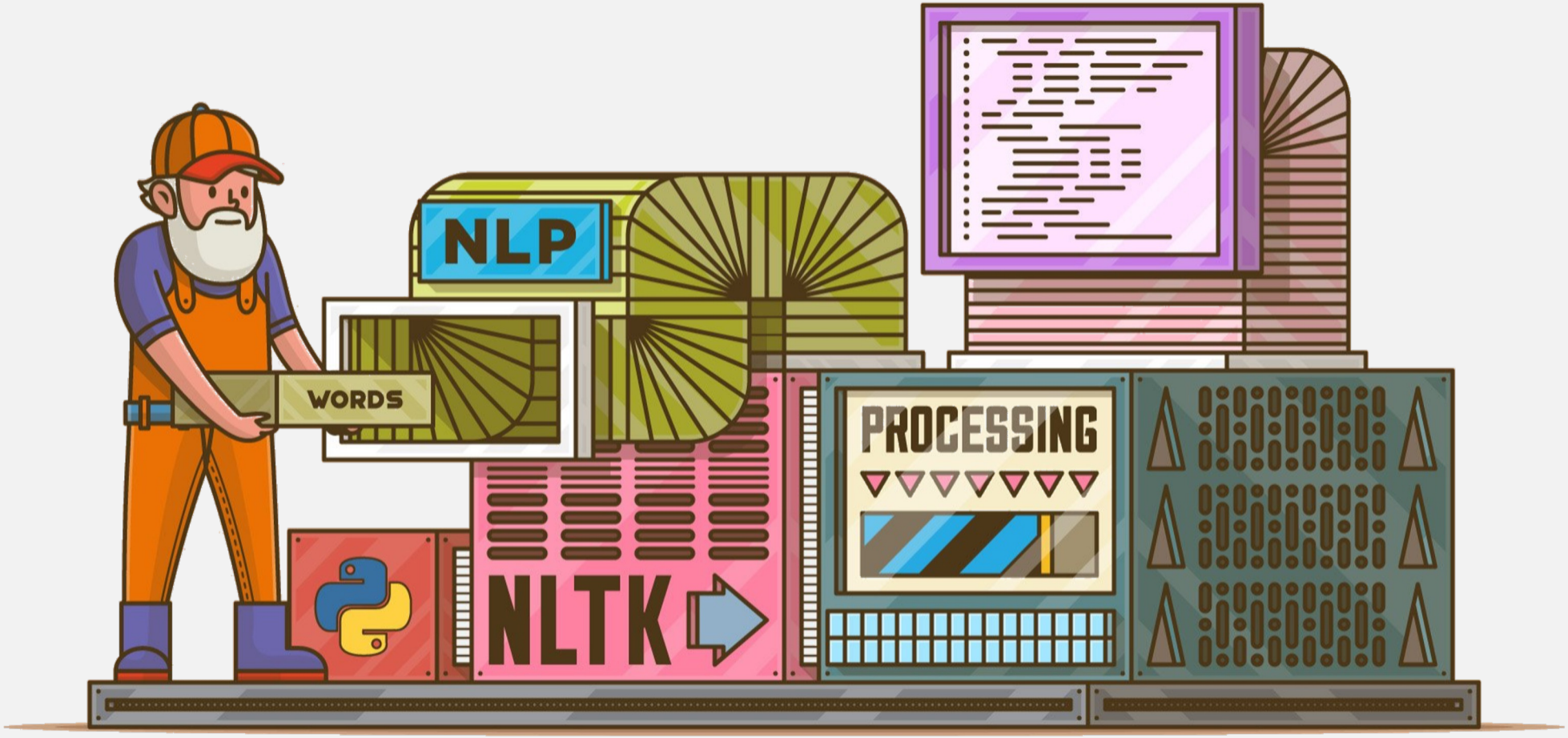
O que é NLP?

Segundo a IBM...

> Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the *same way human beings can*.

DATA LAB

O processamento de linguagem natural tem como input os textos. E o output?



DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

Reconhecimento da fala

Reconhecimento de entidades

Resolução de referência

Análise de Sentimentos

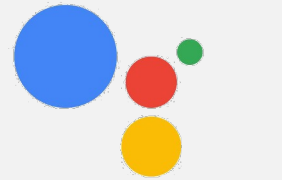
Geração de texto natural

Tradução de textos

DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

Reconhecimento da fala



DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

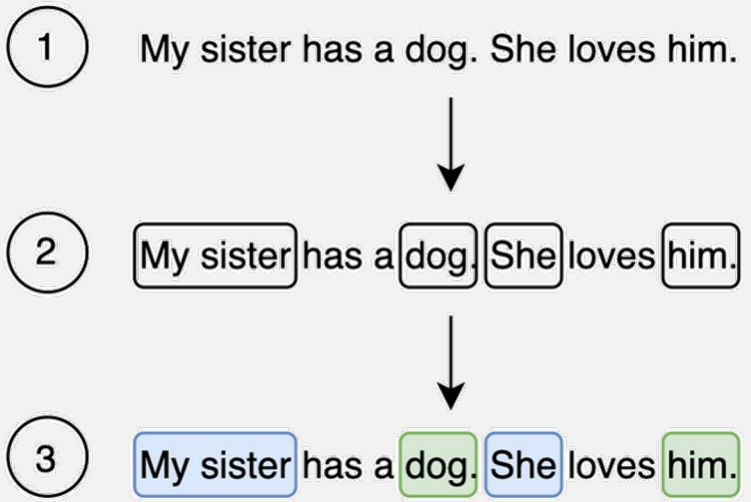
Reconhecimento de entidades

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

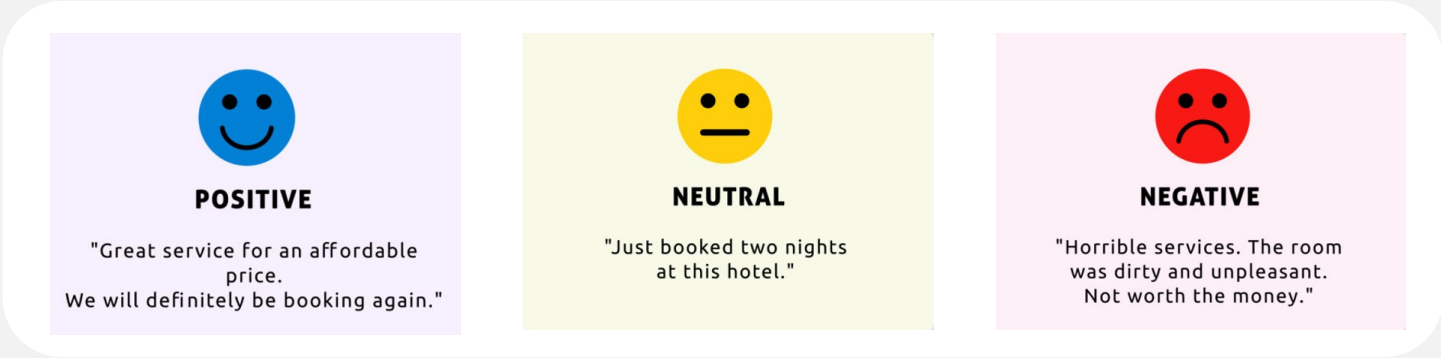
Resolução de referência



DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

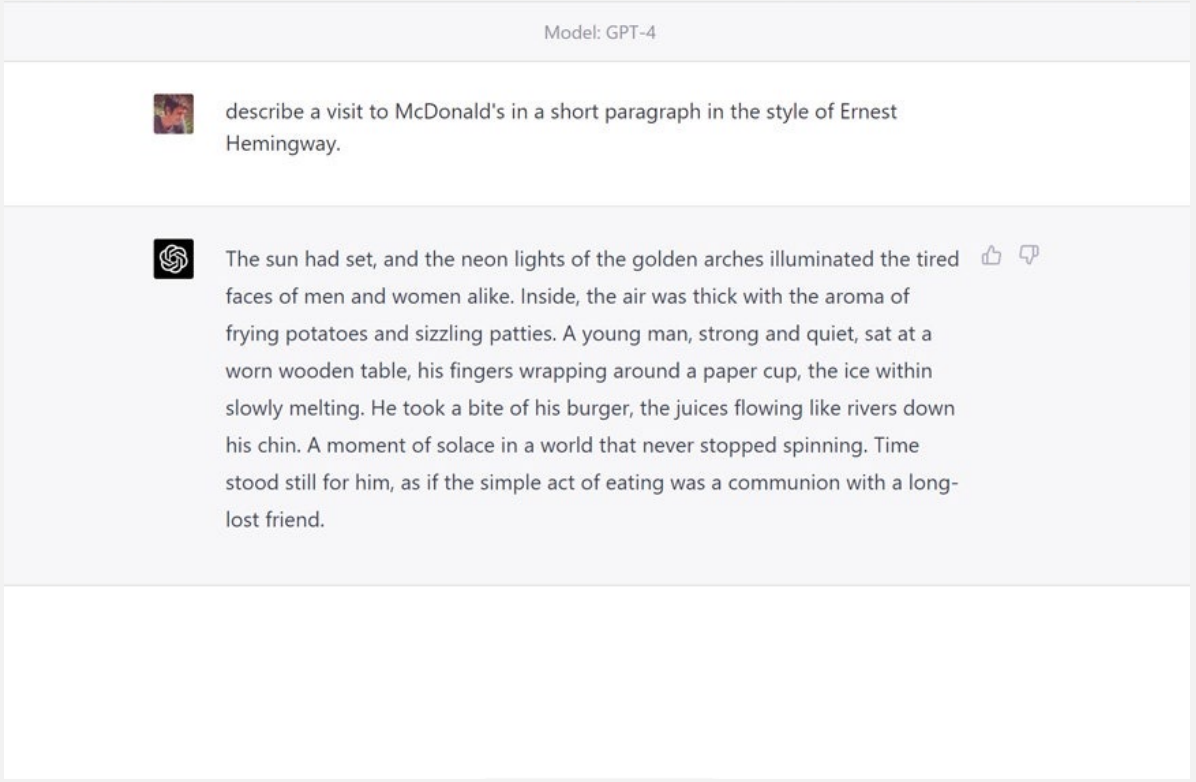
Análise de Sentimentos



DATA LAB

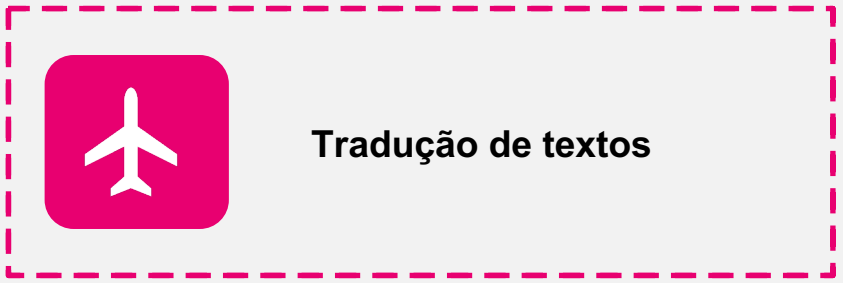
As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:

Geração de texto natural



DATA LAB

As tarefas do processamento de linguagem natural são muitas, algumas das mais famosas são:



DeepL

Google Tradutor

DATA LAB

Tokenização: transformar o texto em unidades menores

Como transformar texto em tokens?

Primeira estratégia:

Transformar caracteres em tokens

R → 1
O → 2
M → 3
A → 4

Qual é o problema dessa estratégia?

Anagramas!!!

A → 4
M → 3
O → 2
R → 1

DATA LAB

Tokenização: transformar o texto em unidades menores

Como transformar texto em tokens?

Segunda estratégia:

Transformar palavras em tokens

Eu	→	1
amo	→	2
Machine	→	3
Learning	→	4

Aqui temos vantagens e desvantagens...

Eu	→	1
odeio	→	7
Machine	→	3
Learning	→	4

Eu	→	1
amo	→	2
Redes	→	5
Neurais	→	6

DATA LAB

Com esse tipo de estratégia já conseguimos resolver as primeiras tarefas de NLP

Como classificar quais são as palavras mais importantes de Memórias Póstumas?

CAPÍTULO PRIMEIRO / *ÓBITO DO AUTOR*

Algum tempo hesitei se devia abrir estas memórias pelo princípio ou pelo fim, isto é, se poria em primeiro lugar o meu nascimento ou a minha morte. Suposto o uso vulgar seja começar pelo nascimento, duas considerações me levaram a adotar diferente método: a primeira é que eu não sou propriamente um autor defunto, mas um defunto autor, para quem a campa foi outro berço; a segunda é que o escrito ficaria assim mais galante e mais novo. Moisés, que também contou a sua morte, não a pôs no intróito, mas no cabo: diferença radical entre este livro e o Pentateuco.

DATA LAB

O primeiro passo é o pré-processamento dos dados

Como padronizar meu texto?

Padronizar as letras em **minúsculas** (função nativa do python)

```
# define um texto
text = 'Algum tempo hesitei se devia abrir estas memórias pelo princípio ou pelo fim

# padroniza as letras em minúsculas
text_min = text.lower()
text_min
```

✓ 0.0s

```
'algum tempo hesitei se devia abrir estas memórias pelo princípio ou pelo fim, isto é,
```

DATA LAB

O primeiro passo é o pré-processamento dos dados

Como padronizar meu texto?

Separar apenas palavras com **regex** (utilizando a biblioteca **re** do python)

```
# Separar apenas letras
text_letras = re.findall(r'[a-zéóáêâãõç]+', text_min)
text_letras
```

✓ 0.0s

```
['algum',
 'tempo',
 'hesitei',
 'se',
 'devia',
 'abrir',
 'estas',
 'memórias',
 'pelo',
 'princ',
 'pio',
 'ou',
 'pelo',
 'fim',
 'isto',
 'é',
 'se',
 'poria',
```

DATA LAB

O primeiro passo é o pré-processamento dos dados

Como padronizar meu texto?

Remover **stop words** (dependendo do algoritimo)

```
new_test = " ".join(text_letras)
tokens = word_tokenize(new_test)

# Remove stop words
portuguese_stopwords = stopwords.words('portuguese')
tokens_wo_stopwords = [t for t in tokens if t not in portuguese_stopwords]
tokens_wo_stopwords
```

✓ 0.0s

```
['algum',
 'tempo',
 'hesitei',
 'devia',
 'abrir',
 'memórias',
 'princ',
 'pio',
 'fim',
 'poria',
 'primeiro',
 'lugar',
 'nascimento',
 'morte',
```

DATA LAB

BoW

Bag of Words

---

*A very common feature extraction procedures for sentences and documents is the bag-of-words approach (BOW). In this approach, we look at the **histogram of the words** within the text, i.e. considering each word count as a feature.*

Page 69, Neural Network Methods in Natural Language Processing, 2017.

É uma medida de **frequência**, a ordem **não** importa

Vantagens:

- Funciona muito bem para entender tópico mais falado (ex: comentários, resenhas...)

Desvantagens:

- Não captura significado (desconsidera ordem)
- Vetores muito esparsos
- Tamanho do vocabulário reduzido
- Pré tratamento de dados cauteloso

Conseguiríamos classificar as palavras mais **recorrentes** no texto do Machado

DATA LAB

TF-IDF

Term Frequency Inverse Document Frequency

Objetivo:

Identificar termos raros e importantes em uma coleção de textos

Definição

TF-IDF = Term Frequency * Inverse Document Frequency

Term Frequency: tf(t) (frequência de um termo **t**)

Inverse Document Frequency: ln[(1+n)/(1+df(t))], sendo **n** o tamanho do corpus.

Para evitar divisões com zero

Vantagens:

- Funciona muito bem para classificação de tipos de documentos e análise de resenhas de produtos
- Não precisa remover stop words
- Pré-tratamento de dados muito mais direto

Desvantagens:

- Não captura significado
- Computacionalmente caro

Conseguiríamos entender **relevância** das palavras no texto do Machado

DATA LAB

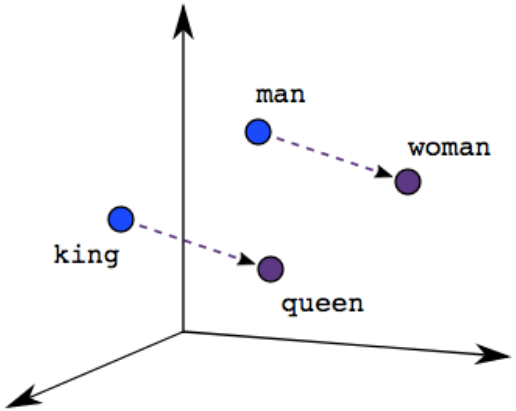
O que mais gostaríamos de ensinar para um modelo?

REI − HOMEM + MULHER = RAINHA

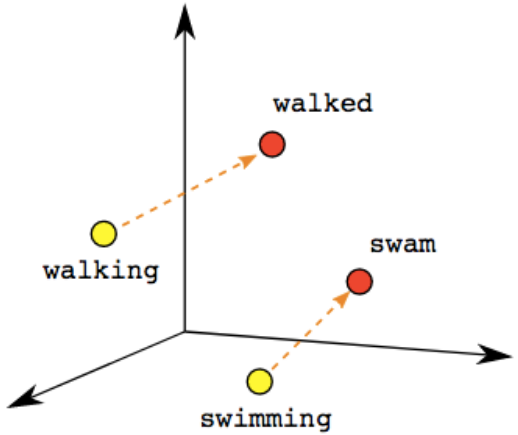
Relações (operações) de sentido entre os tokens (palavras)

DATA LAB

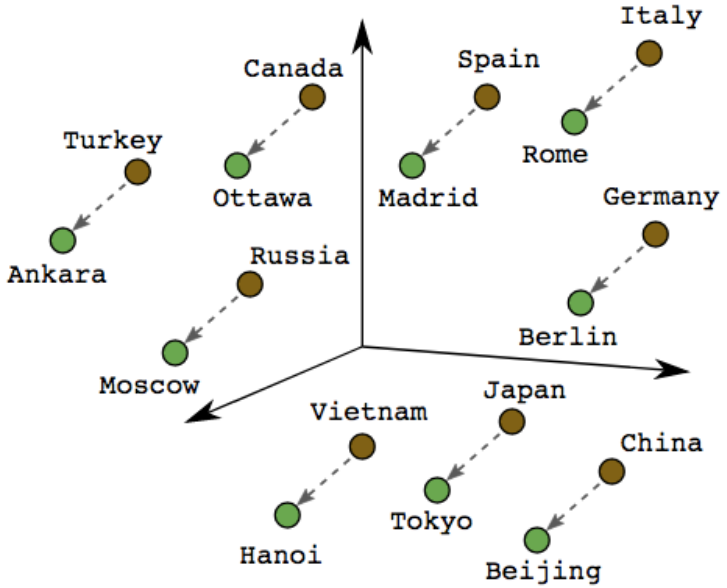
Para essas relações serem possíveis precisamos construir um espaço vetorial em que observemos relações semânticas



Male-Female



Verb Tense



Country-Capital

DATA LAB

WORD2VEC

Parameter Learning Explained

word2vec Parameter Learning Explained

Xin Rong
ronxin@umich.edu

Abstract

The word2vec model and application by Mikolov et al. have attracted a great amount of attention in recent two years. The vector representations of words learned by word2vec models have been shown to carry semantic meanings and are useful in various NLP tasks. As an increasing number of researchers would like to experiment with word2vec or similar techniques, I notice that there lacks a material that comprehensively explains the parameter learning process of word embedding models in details, thus preventing researchers that are non-experts in neural networks from understanding the working mechanism of such models.

This note provides detailed derivations and explanations of the parameter update equations of the word2vec models, including the original continuous bag-of-word (CBOW) and skip-gram (SG) models, as well as advanced optimization techniques, including hierarchical softmax and negative sampling. Intuitive interpretations of the gradient equations are also provided alongside mathematical derivations.

In the appendix, a review on the basics of neuron networks and backpropagation is provided. I also created an interactive demo, wevi, to facilitate the intuitive understanding of the model.[1]

DATA LAB

WORD2VEC

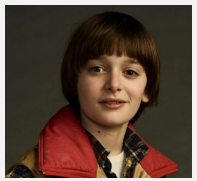
Adentrando na arquitetura do modelo

> You shall know a word by the company it keeps.
>
> J. R. FIRTH

DATA LAB

WORD2VEC-SKIP GRAM

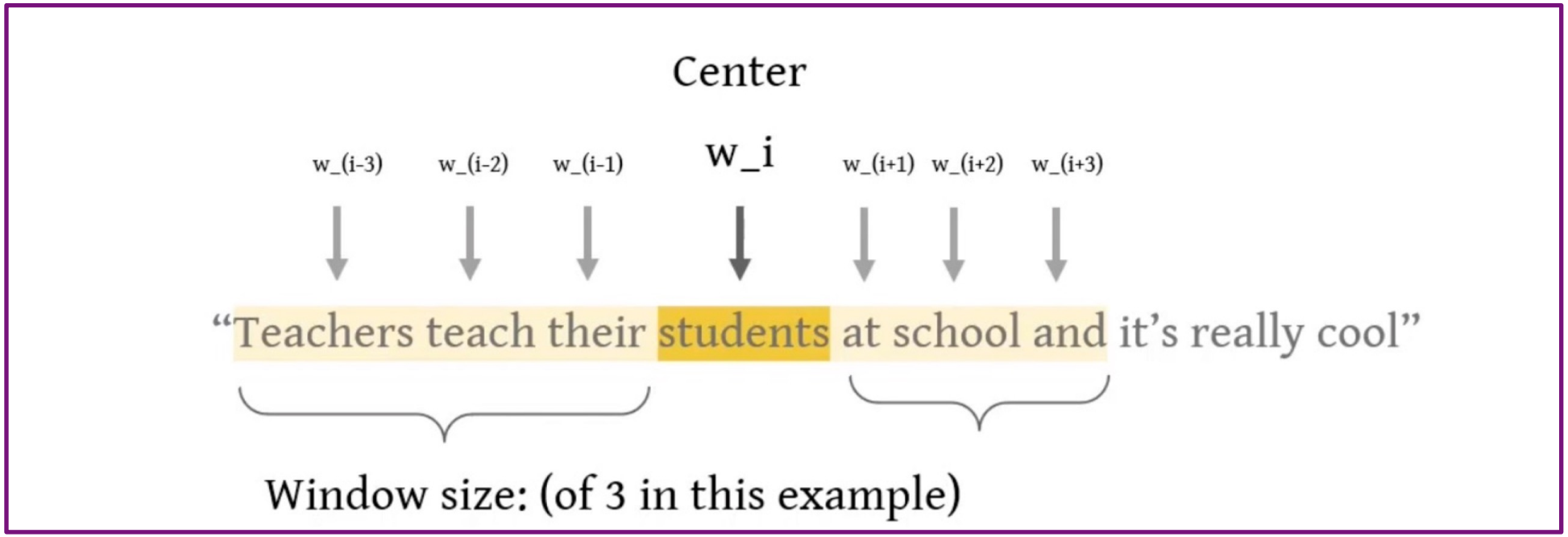
One hot encoding- input do modelo

	1	0	0	0	0
	0	1	0	0	0
	0	0	1	0	0
	0	0	0	1	0
	0	0	0	0	1

DATA LAB

WORD2VEC-SKIP GRAM

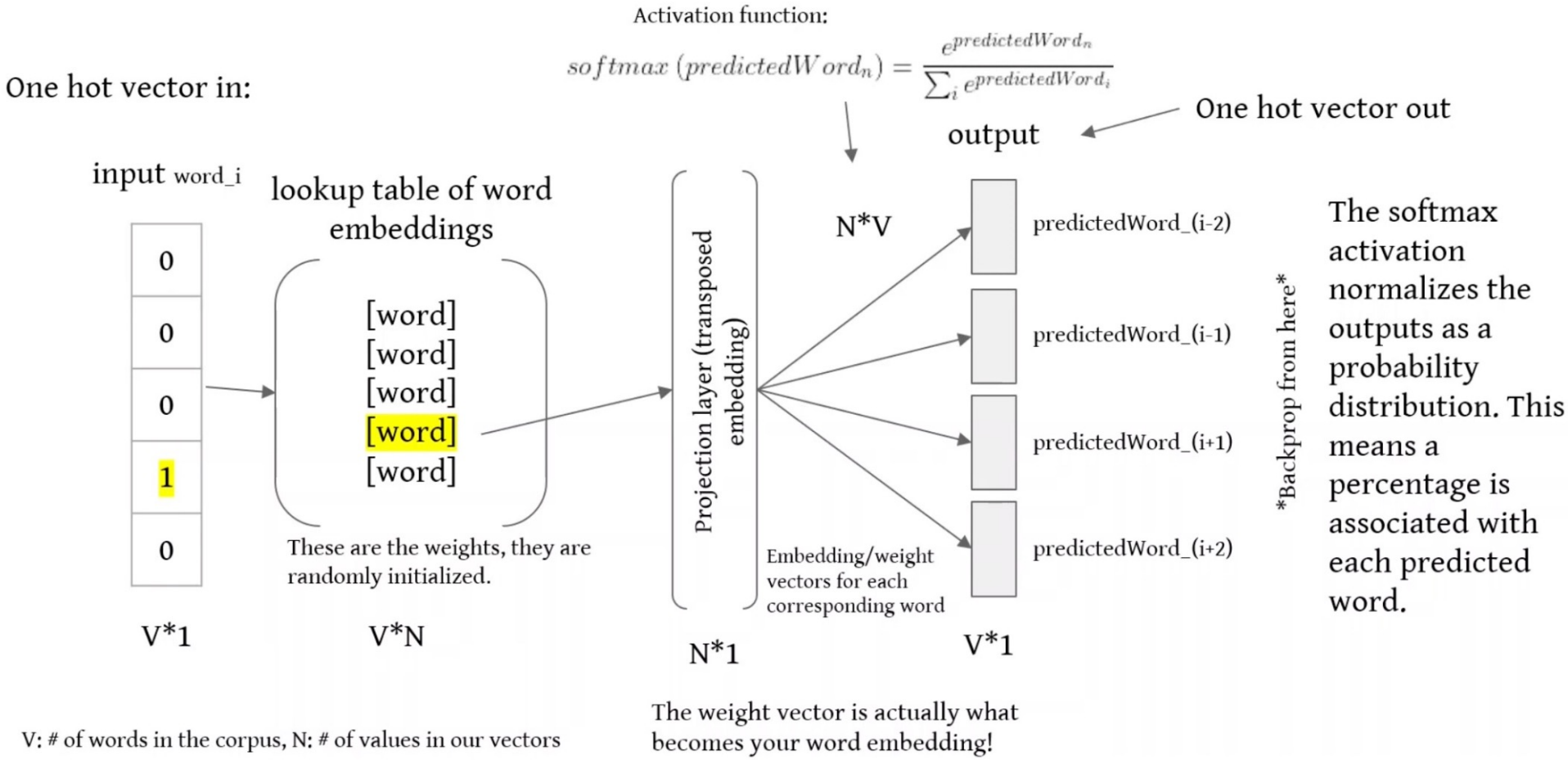
Prediz as palavras de contexto dado uma palavra central



DATA LAB

WORD2VEC-SKIP GRAM- Dentro da arquitetura

Prediz as palavras de contexto dado uma palavra central



DATA LAB

WORD2VEC- BAG OF WORDS

Prediz a palavra central dado as palavras de contexto

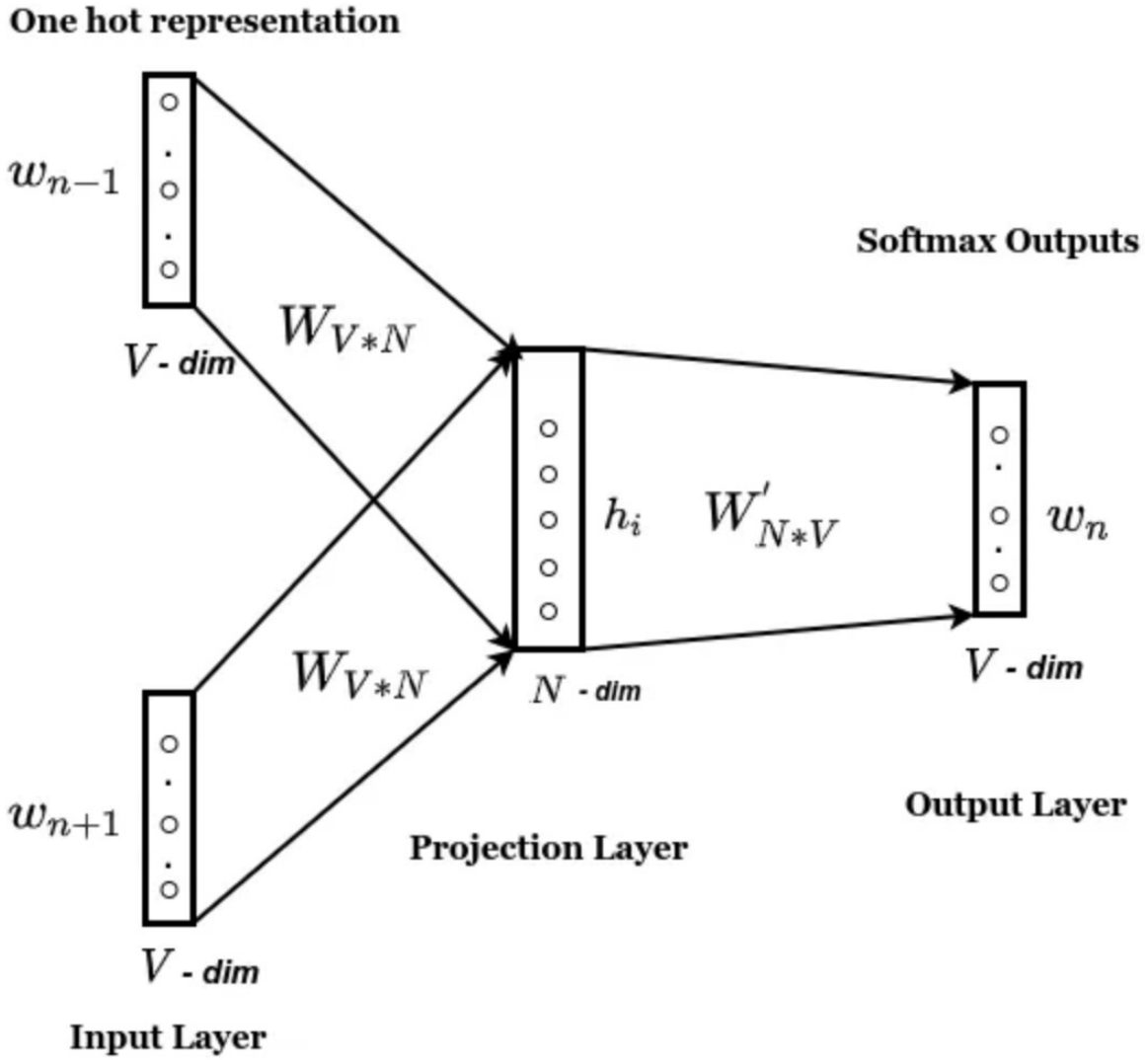
Diz	que	é	?	que	tem	saudade

verdade

DATA LAB

WORD2VEC- BAG OF WORDS- Dentro da arquitetura

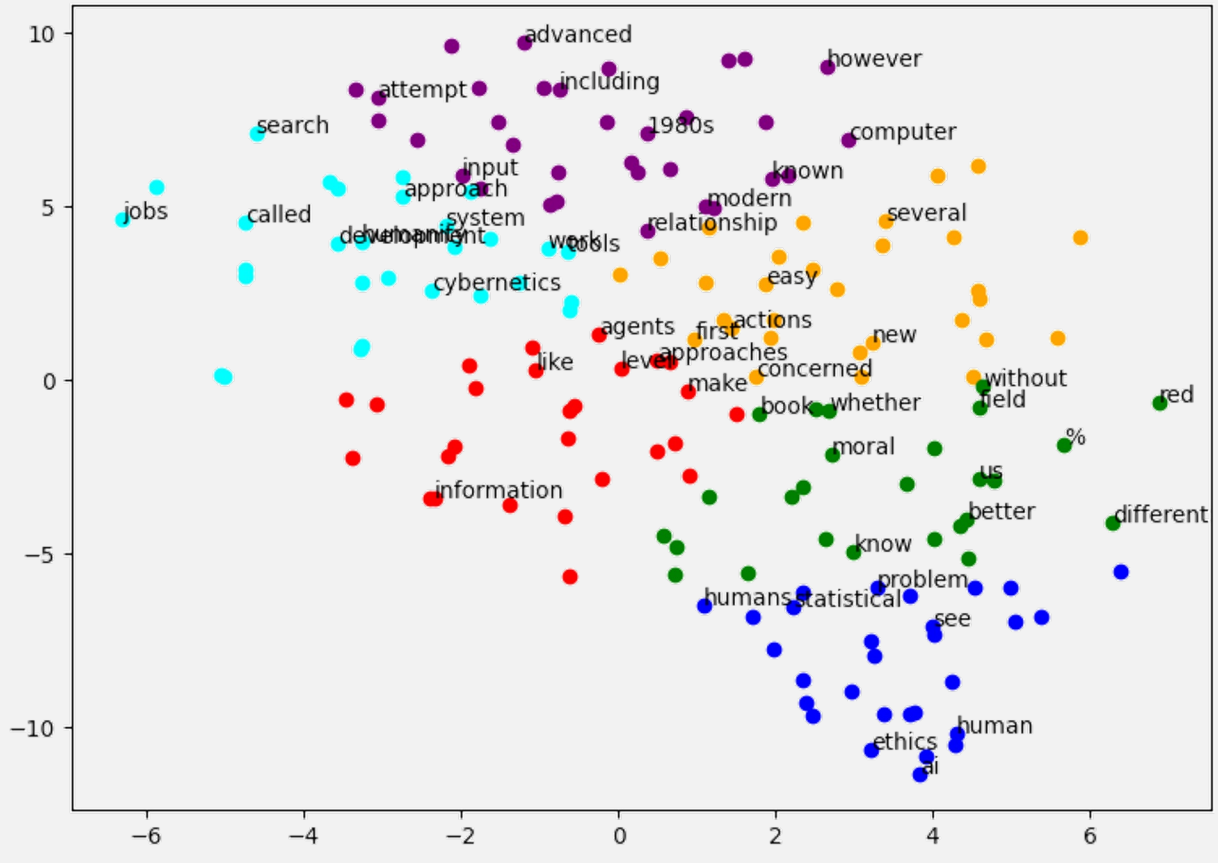
Prediz a palavra central dado as palavras de contexto



DATA LAB

WORD2VEC

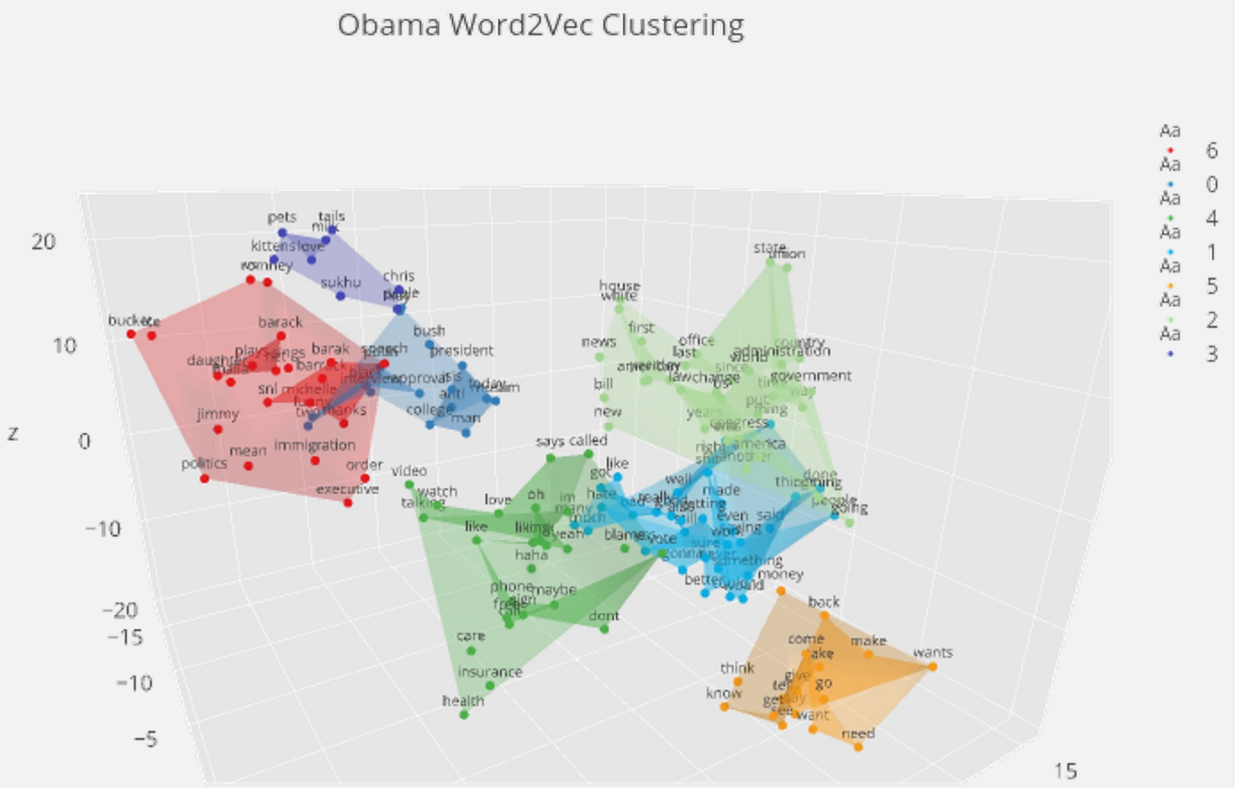
Clusterização examples



DATA LAB

WORD2VEC

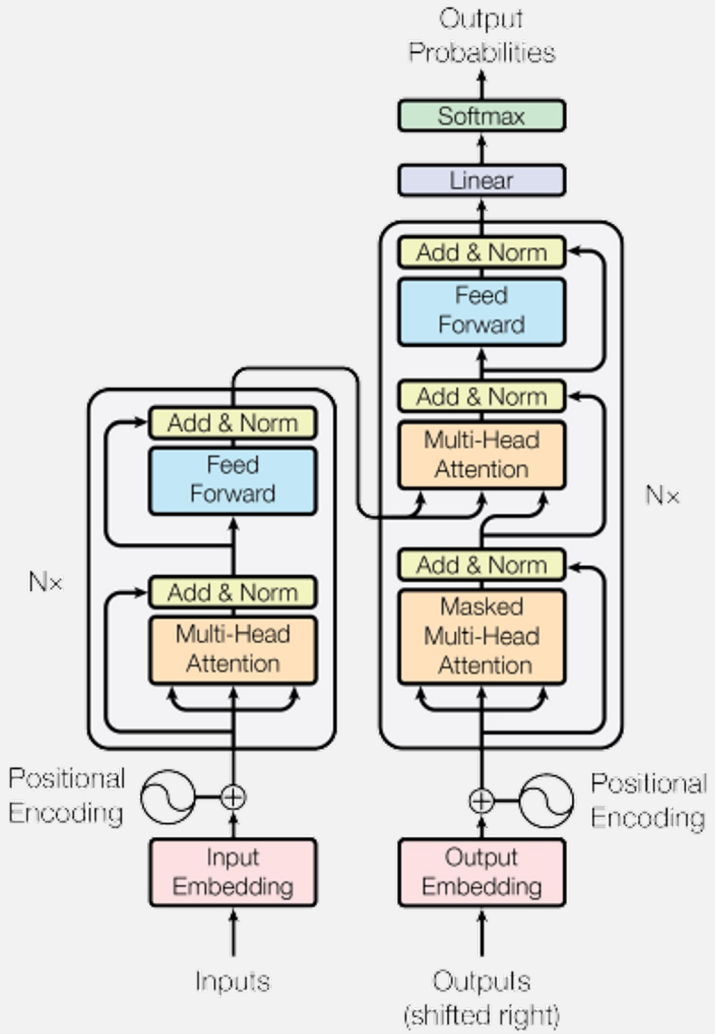
Clusterização examples



DATA LAB

BERT

Bidirectional Encoder Representation from Transformers



- Treinado em todo o corpos da Wikipedia e BookCorpus
- Treinado com 1 milhão de passos
- Foi treinado com múltiplas tarefas
- Consegue representar tanto palavras quanto sentenças
- Pode ser optimizado para uma tarefa específica
- Entende importância da ordem

DATA LAB

BERT

Como é usado?

Respondendo questões

What is ChatGPT?

ChatGPT is an app built by OpenAI. Using the GPT language models, it can answer your questions, write copy, draft emails, hold a conversation, explain code in different programming languages, translate natural language to code, and more—or at least try to—all based on the natural language prompts you feed it. It's a chatbot, but a really, really good one.

Sugerindo respostas

[Mensagem cortada] Exibir toda a mensagem

Boa tarde! | Bom dia! | Obrigada!

Entendendo similaridade de textos

ZDNet
https://www.zdnet.com › ... › Artificial Intelligence

What is ChatGPT and why does it matter? Here's what you ...

15 de set. de 2023 — **ChatGPT is a natural language processing tool driven by AI technology** that allows you to have human-like conversations and much more with the ...

TechRadar
https://www.techradar.com › ... › Artificial Intelligence

everything you need to know about the AI chatbot

27 de nov. de 2023 — ChatGPT stands **for "Chat Generative Pre-trained Transformer"**, which is a bit of a mouthful. Let's take a look at each of those words in turn.

What exactly is it? · What does ChatGPT stand for? · How much does it cost?

Business Insider
https://www.businessinsider.com › ... · Traduzir esta página

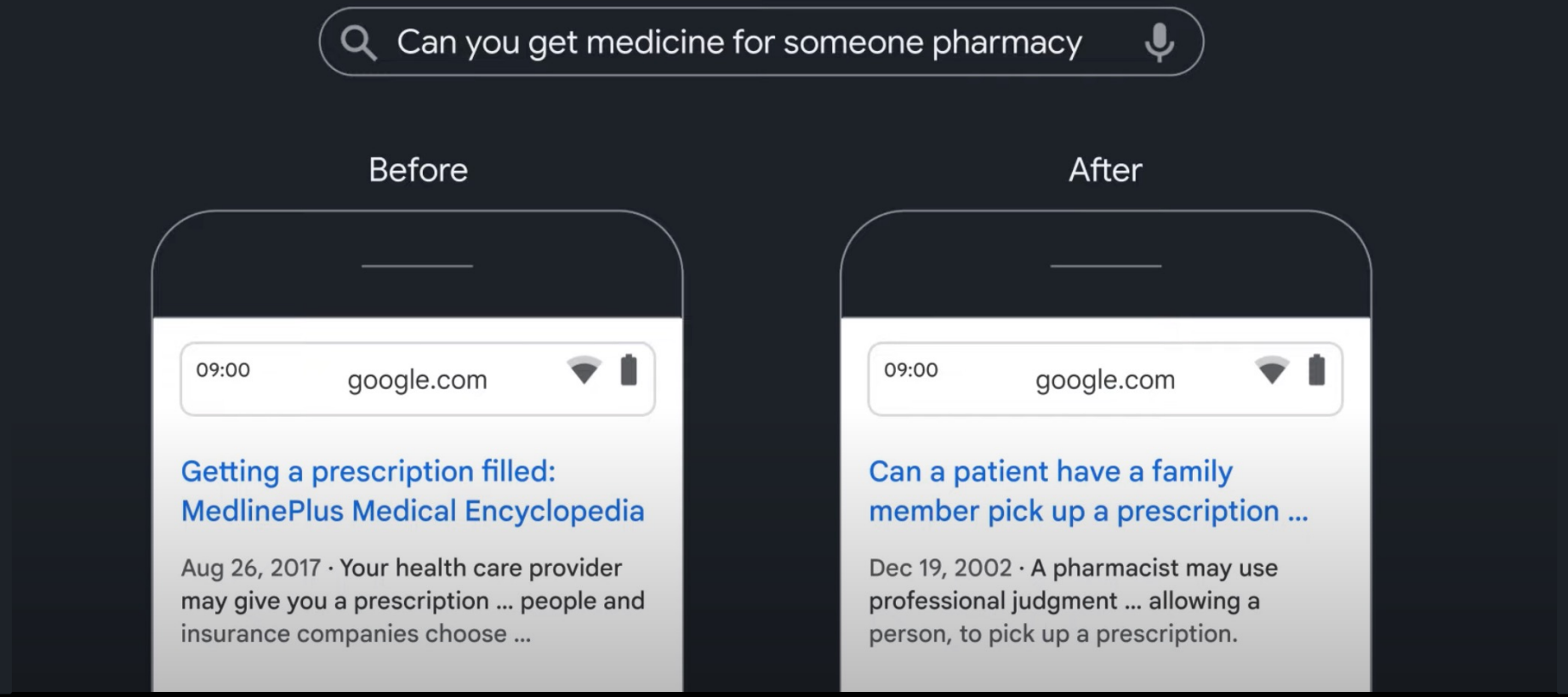
What Is ChatGPT? Everything You Need to Know About the ...

21 de ago. de 2023 — Chatbots like ChatGPT are **powered by large amounts of data and computing techniques to make predictions to string** words together in a meaningful ...

DATA LAB

BERT

Como esse modelo impactou o sistema de busca do google?



DATA LAB

BERT

Modelo baseado na arquitetura de Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

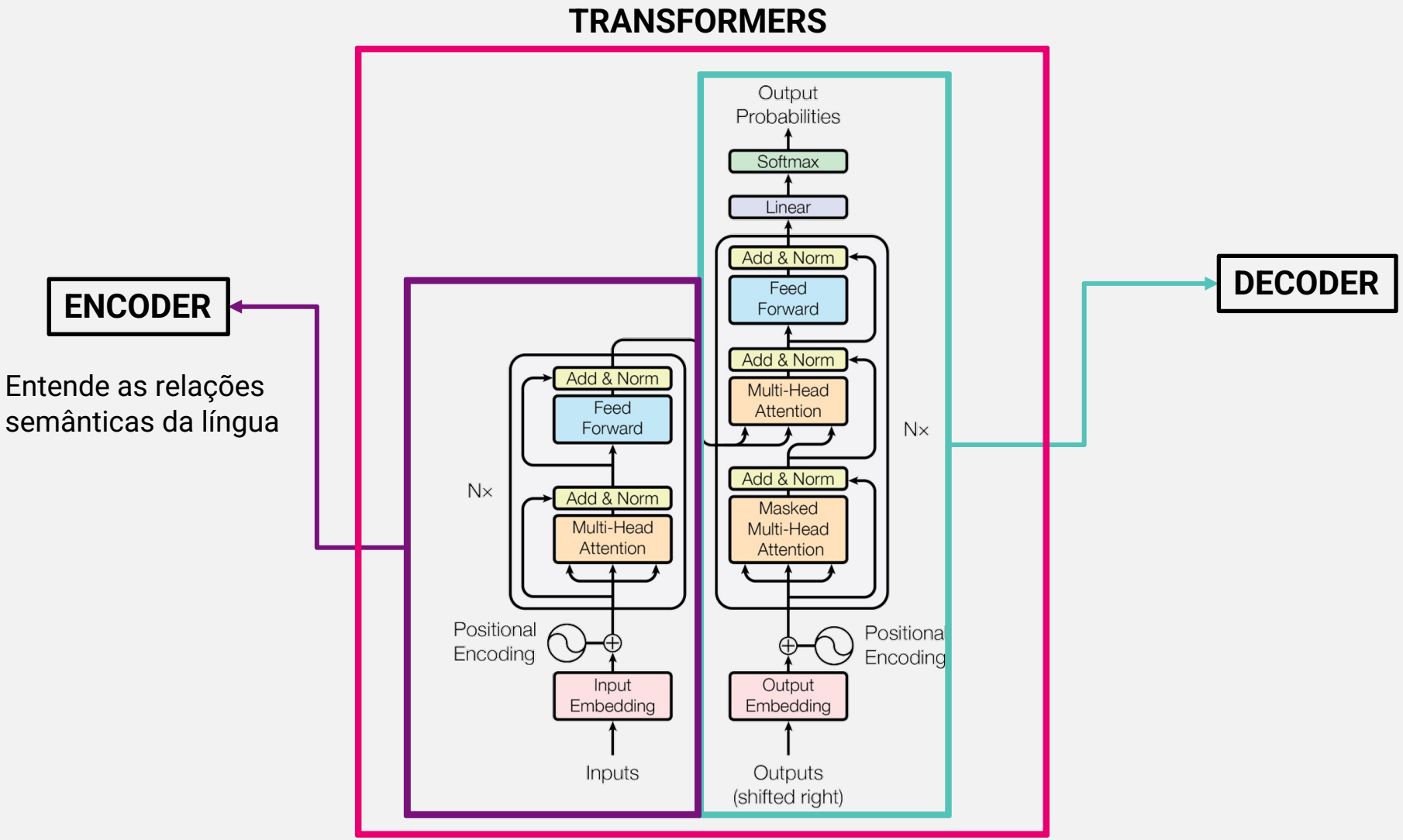
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

DATA LAB

BERT

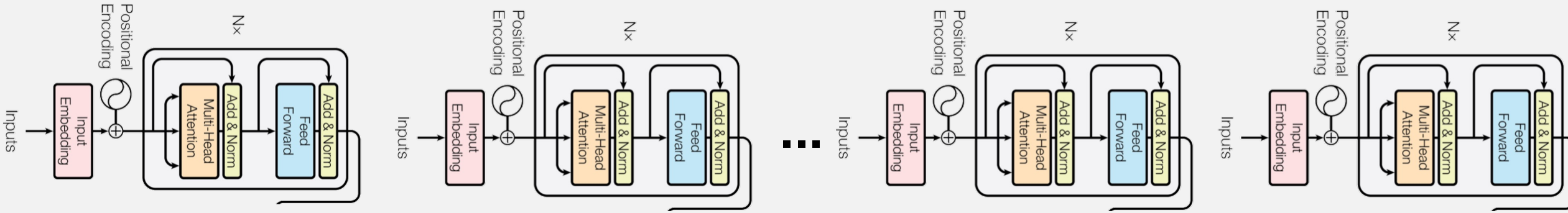
Modelo baseado na arquitetura de Transformers



DATA LAB

BERT

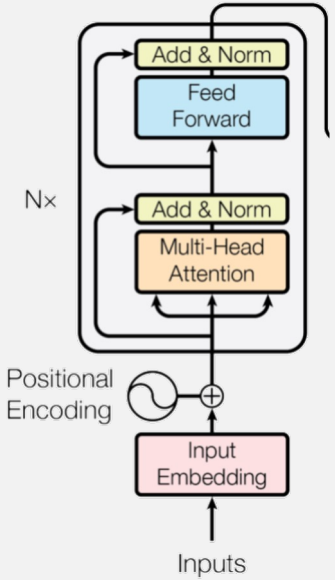
Bidirectional **Encoder Representation** from **Transformers**



DATA LAB

BERT

Como é feito o seu pré-treinamento?



Masked Language Model (MLM)

Prevê as palavras mascaradas

No meio do caminho tinha uma [mask1] tinha uma pedra no meio do [mask2]

Next Sentence Prediction (NSP)

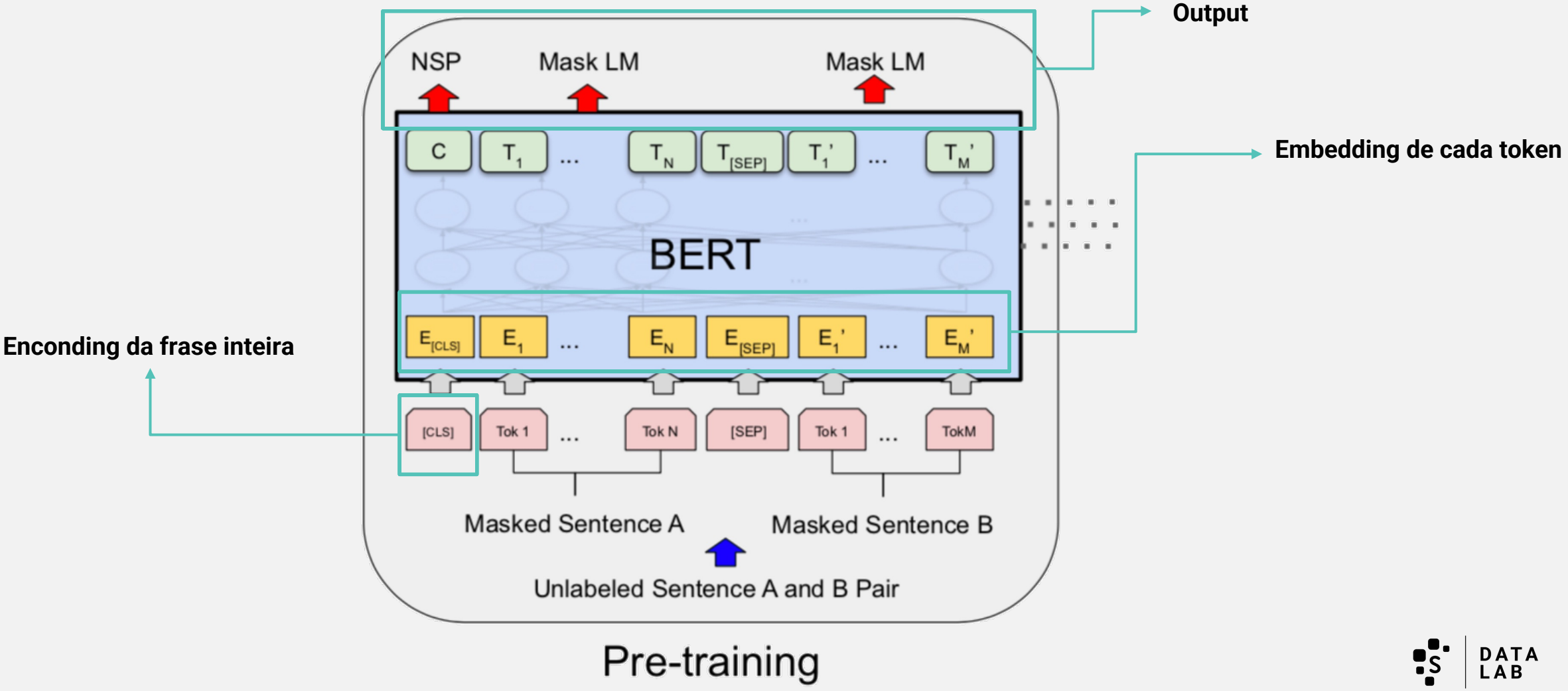
Prevê se a segunda frase segue a primeira

Nunca me esquecerei desse acontecimento
na vida de minhas retinas tão fatigadas.
Nunca me esquecerei que no meio do caminho
tinha uma pedra
tinha uma pedra no meio do caminho
no meio do caminho tinha uma pedra.

DATA LAB

BERT

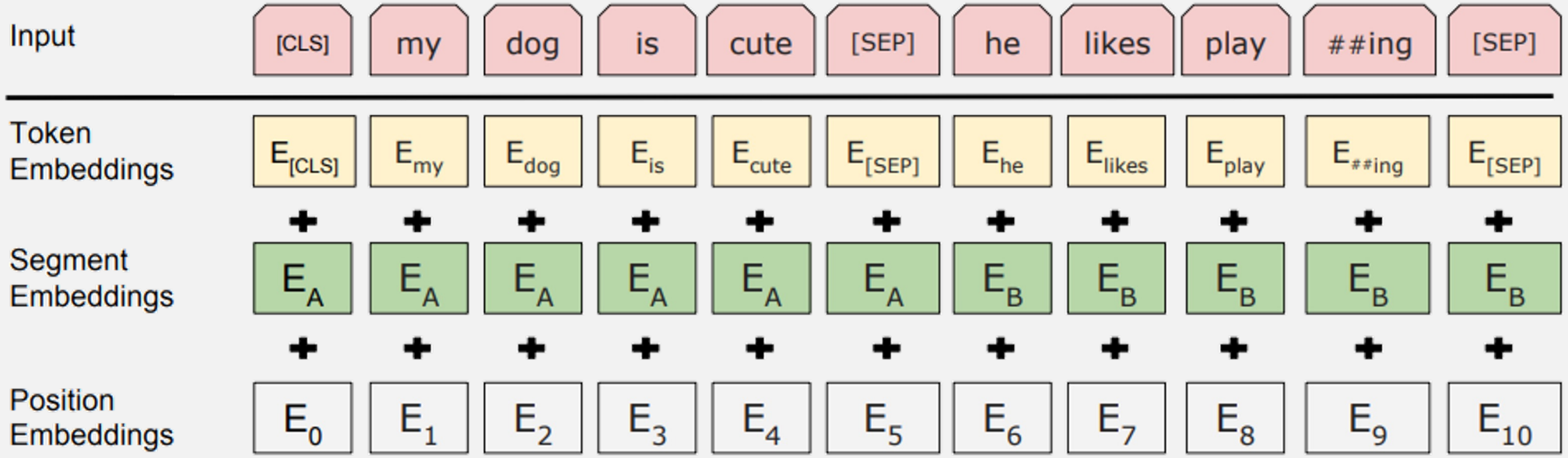
MLM e NSP são feitas ao mesmo tempo!!



DATA LAB

BERT

Como os embeddings são gerados?

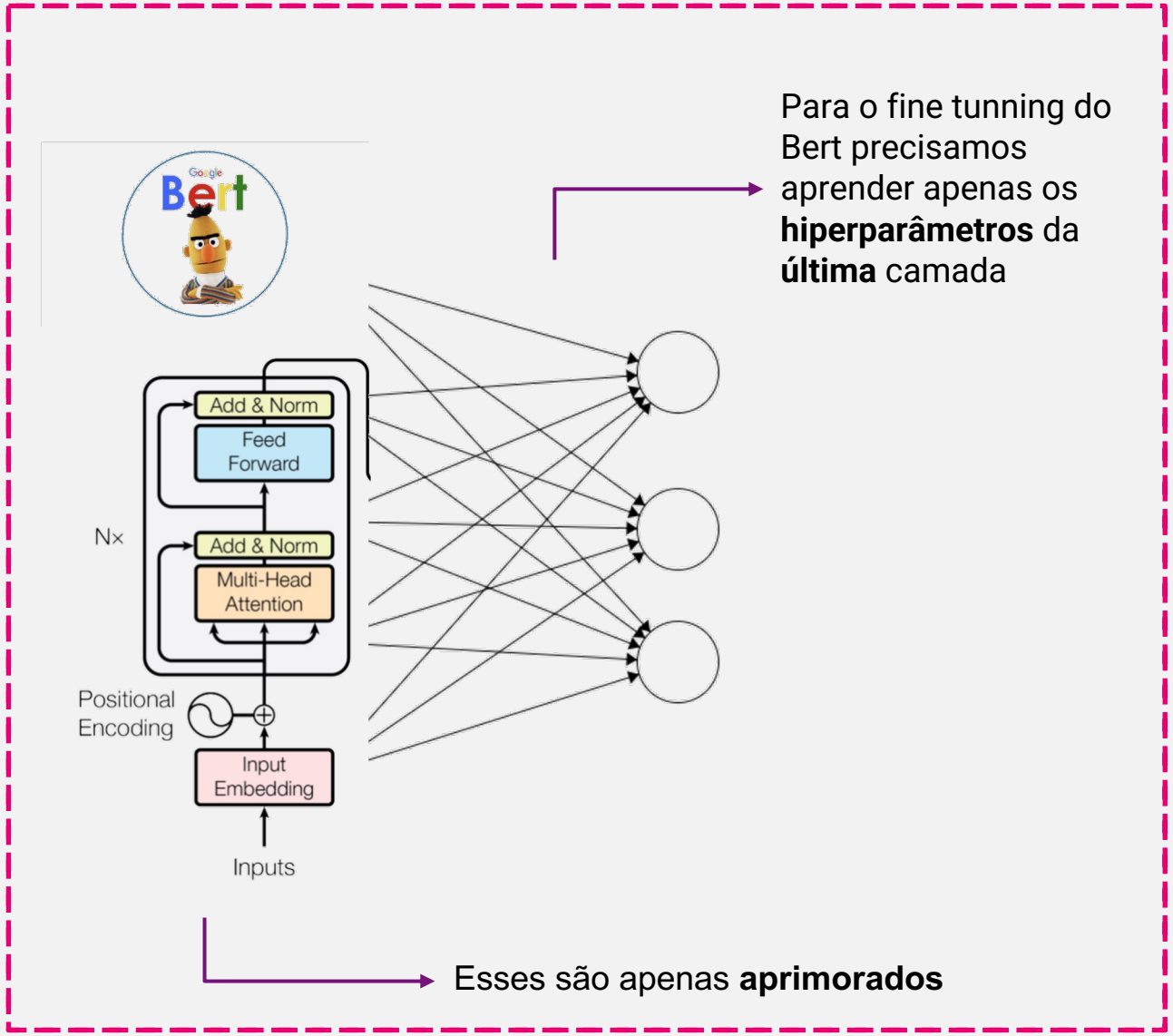


DATA LAB

BERT

Fine tunning

Para o fine tunning do Bert performamos um aprendizado supervisionado com uma nova camada de output



DATA LAB