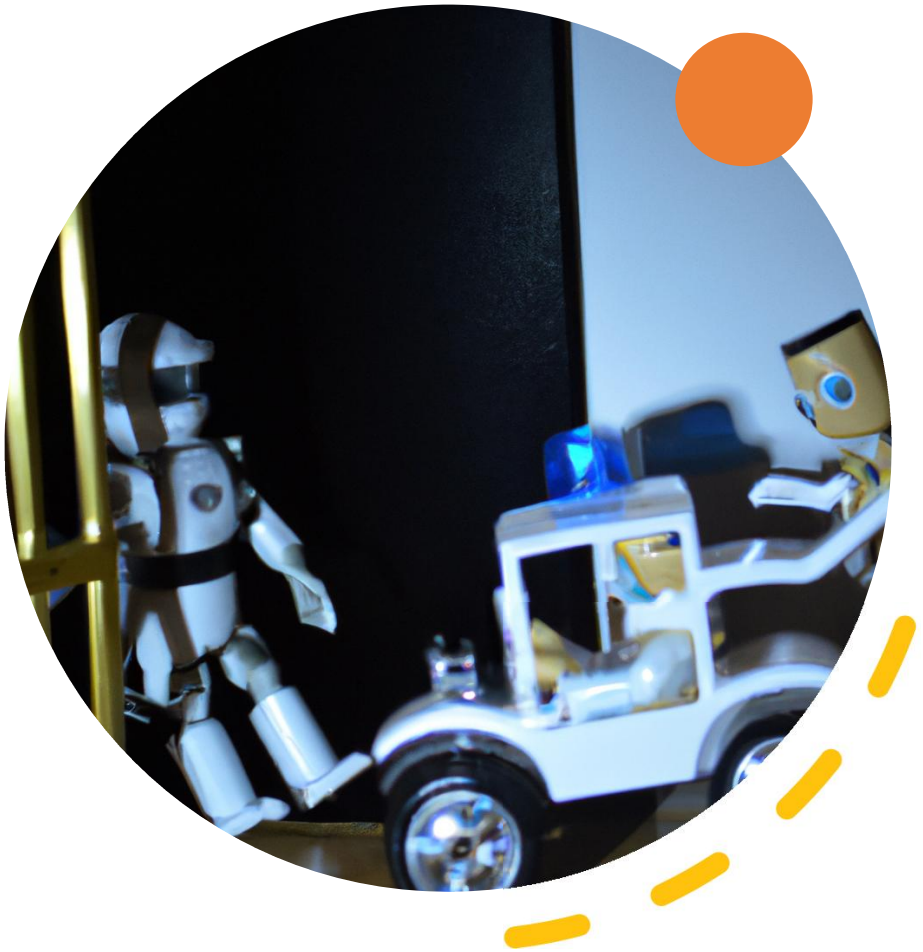
Computadores

- **São capazes apenas de processar números**
- **Não lidam muito bem com informação não estruturada!**



Computadores

- **É preciso criar formas de representação de texto as quais os computadores sejam capazes de processar!**
  - **Transformar texto em uma representação numérica**
  - **Transformar informação não estruturada em estruturada**

Word Embedding

- Representações computacionais de texto
  - Independentes de contexto
  - Dependentes de contexto

One Hot Encoding

- Nossa vida é controlada por algoritmos

	algoritmos	controlada	é	Nossa	por	vida
Nossa	0	0	0	1	0	0
vida	0	0	0	0	0	1
é	0	0	1	0	0	0
controlada	0	1	0	0	0	0
por	0	0	0	0	1	0
algoritmos	1	0	0	0	0	0

TF-IDF

- Processos Judiciais
  - Comum: Processo, lei, partes
  - Incomum: Ambiental

```
['ambiental' 'de' 'jurídico' 'lei' 'processo'
'segundo' 'trabalhista']
  (0, 3)  0.26469629755874713
  (0, 5)  0.4146978997095072
  (0, 0)  0.8293957994190144
  (0, 4)  0.26469629755874713
  (1, 2)  0.7772211620785797
  (1, 4)  0.6292275146695526
  (2, 1)  0.6406554311067799
  (2, 2)  0.5051001005334584
  (2, 3)  0.4089220628888078
  (2, 4)  0.4089220628888078
  (3, 6)  0.8429263481500496
  (3, 3)  0.5380289691033573
```

Word2Vec

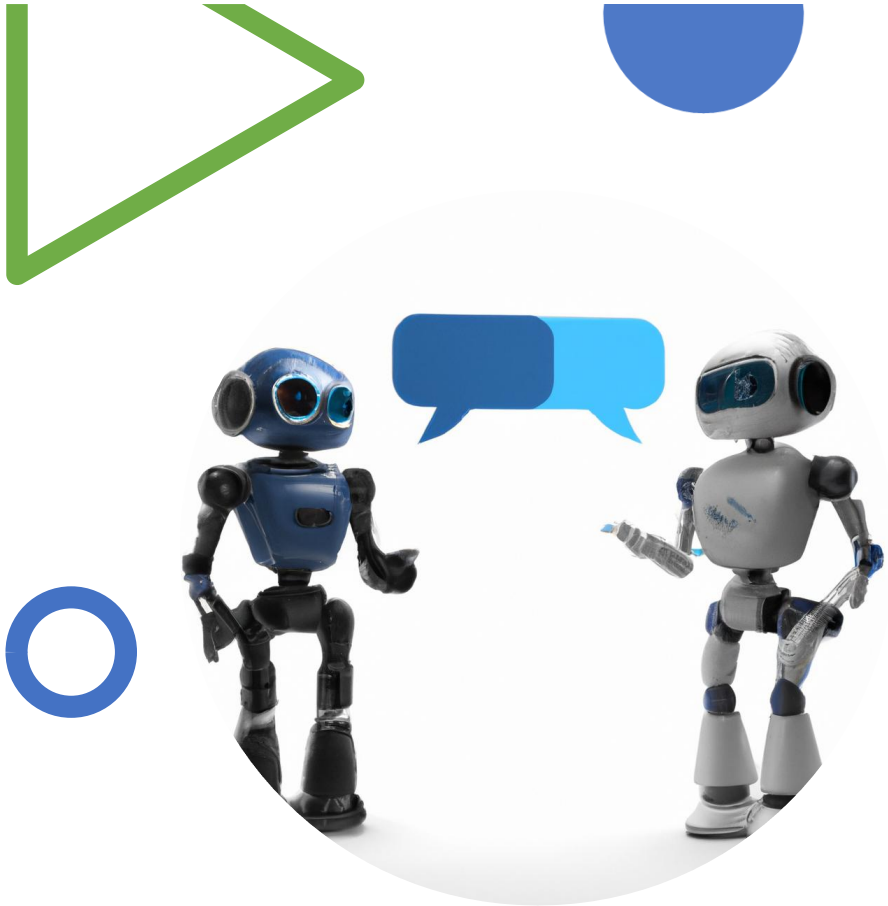
- Através de um processo de treinamento, produz um vetor demonstrando matematicamente a relação entre palavras
- Duas formas principais:
  - CBOW: busca prever uma palavra central no seu contexto
  - Skip-gram: busca prever o contexto a partir de uma palavra central

BANANA.VECTOR

```
array([2.02280000e-01,  -7.66180009e-02,   3.70319992e-01,
       3.28450017e-02,  -4.19569999e-01,   7.20689967e-02,
      -3.74760002e-01,   5.74599989e-02,  -1.24009997e-02,
       5.29489994e-01,  -5.23800015e-01,  -1.97710007e-01,
      -3.41470003e-01,   5.33169985e-01,  -2.53309999e-02,
       1.73800007e-01,   1.67720005e-01,   8.39839995e-01,
       5.51070012e-02,   1.05470002e-01,   3.78719985e-01,
       2.42750004e-01,   1.47449998e-02,   5.59509993e-01,
       1.25210002e-01,  -6.75960004e-01,   3.58420014e-01,
       # ... and so on ...
       3.66849989e-01,   2.52470002e-03,  -6.40089989e-01,
      -2.97650009e-01,   7.89430022e-01,   3.31680000e-01,
      -1.19659996e+00,  -4.71559986e-02,   5.31750023e-01], dtype=float32)
```

Outras Formas...

- *FastText*
- *Glove*
- *Bert*



Transformers

- Arquitetura de RNA utilizada para NLP
- Pode utilizar word embeddings como entrada para a RNA
- Transformers serão estudados em seção posterior