# Representando Significado com Vetores

Bruno Ferrari Guide

Orientador: Marcos Lopes
Universidade de São Paulo

*bruno.guide@usp.br*

9 de novembro de 2018

# Tópicos

- Vetorizando a Língua

# Juntando e contando coisas - Corpora e estatística

```
#contando as palavras
def histograma(corpus):
    dic = dict()
    for palavra in corpus:
        if (palavra not in dic):
            dic[palavra] = 1
        else:
            dic[palavra] += 1
    return dic
```

Figura: Script em Python para contar palavras

# Frequência de termo (TF)

**Variants of term frequency (tf) weight**

| weighting scheme | tf weight |
|---|---|
| binary | $0, 1$ |
| raw count | $f_{t,d}$ |
| term frequency | $f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$ |
| log normalization | $\log(1 + f_{t,d})$ |
| double normalization 0.5 | $0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$ |
| double normalization K | $K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$ |

# Matriz de termos-documentos

| | romance/marm01.txt | romance/marm02.txt | romance/marm03.txt | romance/marm04.txt | romance/marm05.txt | romance/marm06.txt | romance/marm07.txt | romance/marm08.txt |
|---|---|---|---|---|---|---|---|---|
| 0 | romance, | romance, | romance, | romance, | romance, | romance, | romance, | romance, |
| 1 | ressurreição,1872 | a | helena,1876 | iaiá | memórias | casa | quincas | dom |
| 2 | ressurreição | mão | helena | garcia,1878 | póstumas | velha,1885 | borba,1891 | casmurro, |
| 3 | texto-fonte: | e | texto-fonte: | iaiá | de | casa | quincas | 1899 |
| 4 | obra | a | obra | garcia | brás | velha | borba | dom |
| 5 | completa, | luva,1874 | completa, | texto-fonte: | cubas, | texto-fonte: | texto-fonte: | casmurro |
| 6 | machado | a | de | obra | 1880 | obra | obra | texto |

Figura: Matriz termo-documento

# Frequência inversa em documentos (IDF)

$$\operatorname{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Figura: Inverse Document Frequency

# TF-IDF

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

| | As You Like It | Twelfth Night | Julius Caesar | Henry V |
|---|---|---|---|---|
| battle | 0.074 | 0 | 0.22 | 0.28 |
| good | 0 | 0 | 0 | 0 |
| fool | 0.019 | 0.021 | 0.0036 | 0.0083 |
| wit | 0.049 | 0.044 | 0.018 | 0.022 |

**Figure 6.8** A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.

Figura: TF-IDF com exemplo

- Vetores, Álgebra e Geometria

# Plano Cartesiano

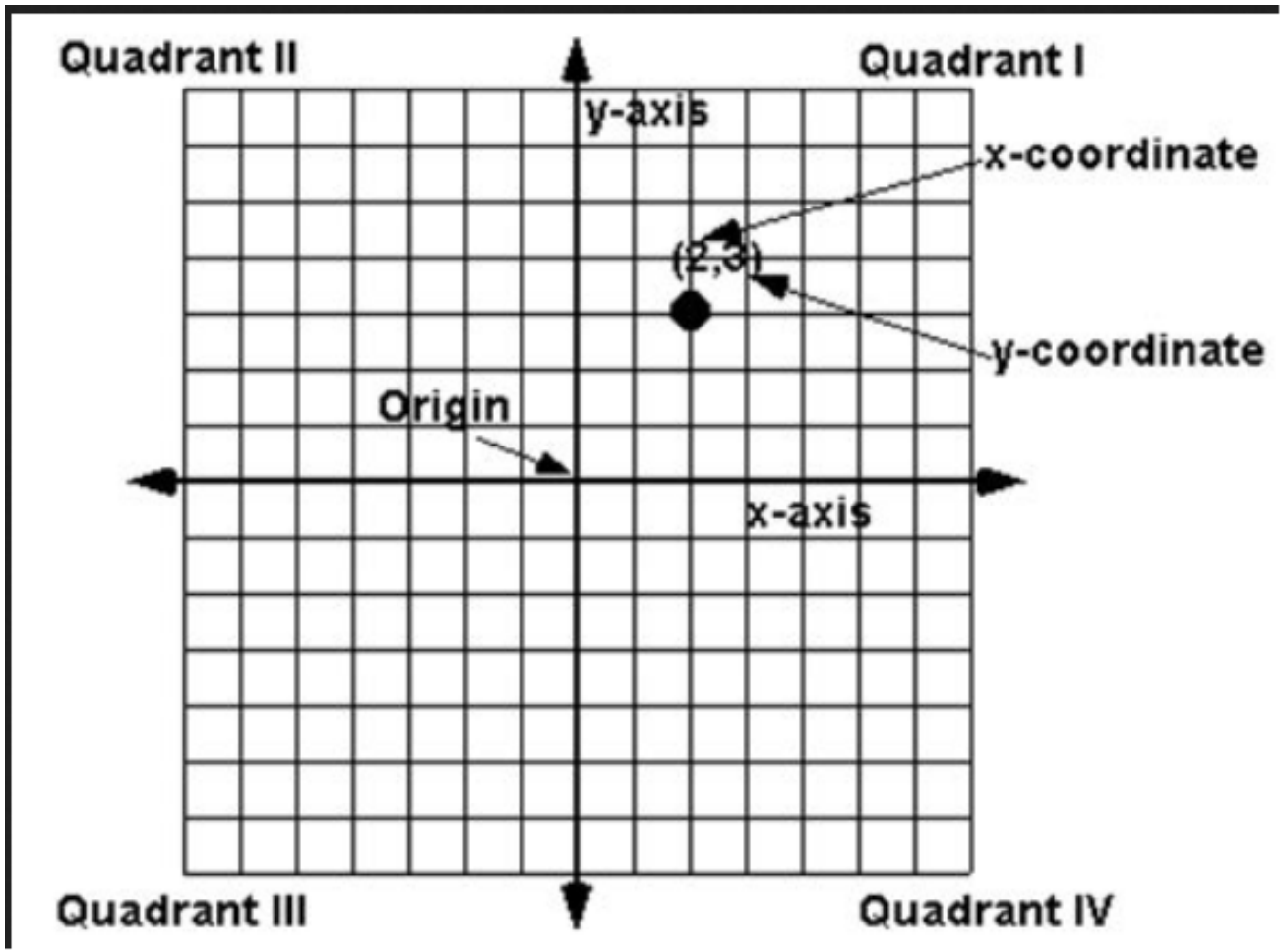
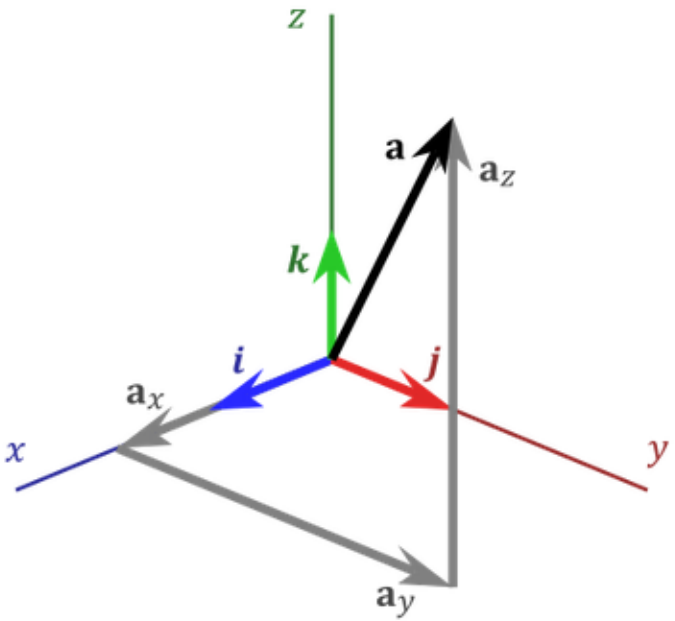



Figura: O plano Cartesiano

# Ponto e vetor




$$\mathbf{a} = (2, 3)$$

$$\mathbf{a} = (a_x, a_y, a_z)$$

$$\mathbf{a} = (a_1, a_2, a_3, \cdots, a_{n-1}, a_n)$$

Figura: Pontos, Vetores e dimensões

# Palavras são vetores

```
('helena', [0.28651808681165336, 0.7020584309878144, 0, 0.3414983950784797])
('casmurro', [0.6454947187463482, 0, 0, 0])
('casa', [0.009719545955246493, 0.008234123230390626, 0.00015107275510106155, 0.005648214929258 18])
('casamento', [0.6661184290694248, 0.7330428931966182, 0.4373260999084505, 0.8577315612168996])
('bom', [0.004945725328765154, 0.007919284835379035, 0.0005926042584393865, 0.007123514055950858])
('verdade', [0.12926613589474578, 0.9170286217031972, 0.16477416499879327, 0.9762289194905985])
('para', [0.0002519198505075053, 5.618377428077959e-05, 0.00013623218392854386, 0.000145825581112210 55])
```

Figura: Vetores de palavras

# Distância entre vetores

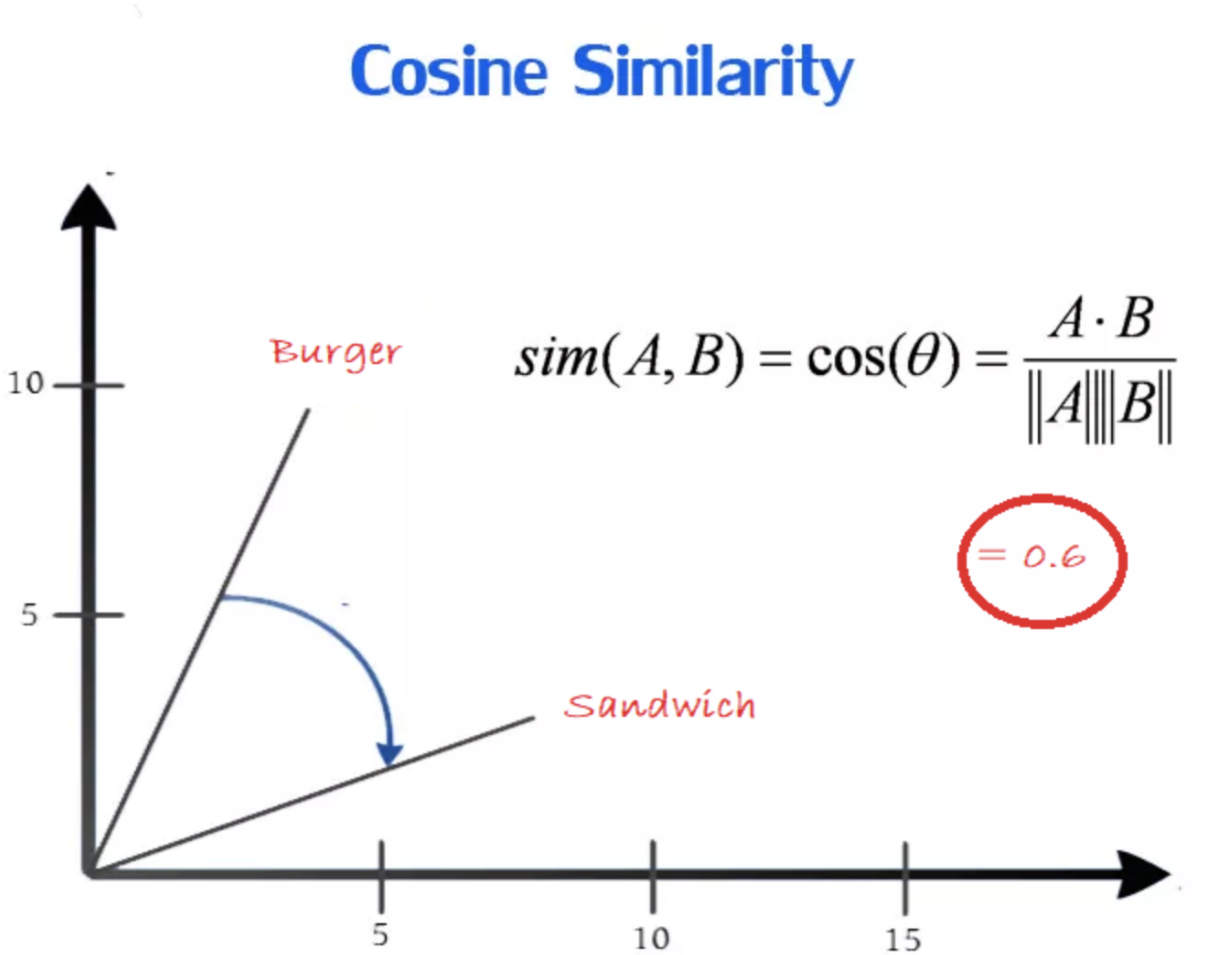



Figura: Similaridade de cosseno

- Vetores Esparsos e Densos

# A Maldição da Dimensionalidade

$$\begin{pmatrix} 1.0 & 0 & 5.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.0 & 0 & 0 & 0 & 0 & 11.0 & 0 \\ 0 & 0 & 0 & 0 & 9.0 & 0 & 0 & 0 \\ 0 & 0 & 6.0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 7.0 & 0 & 0 & 0 & 0 \\ 2.0 & 0 & 0 & 0 & 0 & 10.0 & 0 & 0 \\ 0 & 0 & 0 & 8.0 & 0 & 0 & 0 & 0 \\ 0 & 4.0 & 0 & 0 & 0 & 0 & 0 & 12.0 \end{pmatrix}$$

Figura: Coleção de Vetores (Matriz)

# Vetores Densos vs. Vetores Esparsos

| Tipo de Vetor | Interpretação | Exigência para processamento |
|---|---|---|
| Esparso | Transparente | Exigente |
| Denso | Opaco | Amigável |

Tabela: Comparação entre vetores densos e esparsos

# Adensando Vetores

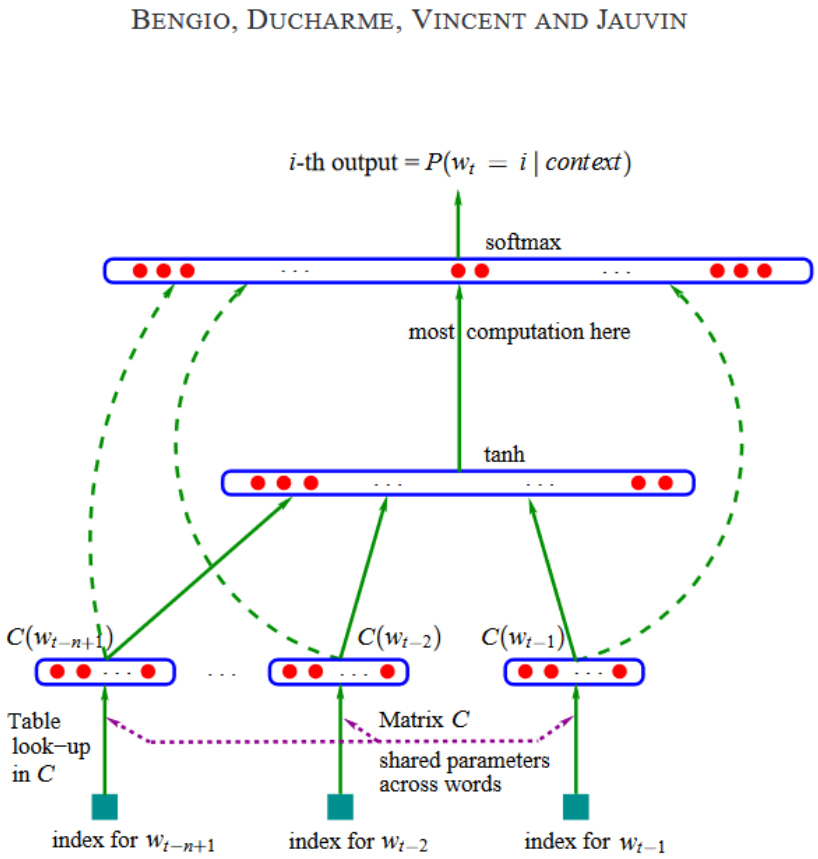



Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where $g$ is the neural network and $C(i)$ is the $i$-th word feature vector.

Figura: É possível, mas não cabe a discussão no momento

# Funciona?

- Muito!
- Hoje a semântica de vetores densos é dominante nos algoritmos mais bem sucedidos para interpretação de significado, como chatbots, tradução automática, busca.
- Os modelos principais não são iguais a LSA do slide anterior, mas o principio de adensar vetores é o mesmo.
- A seguir, explorarei os tipos de relação semântica que podemos explorar nos vetores gerados pelo modelo Word2Vec.

- Noções Linguísticas Subjacentes

# Operações com vetores - similaridade

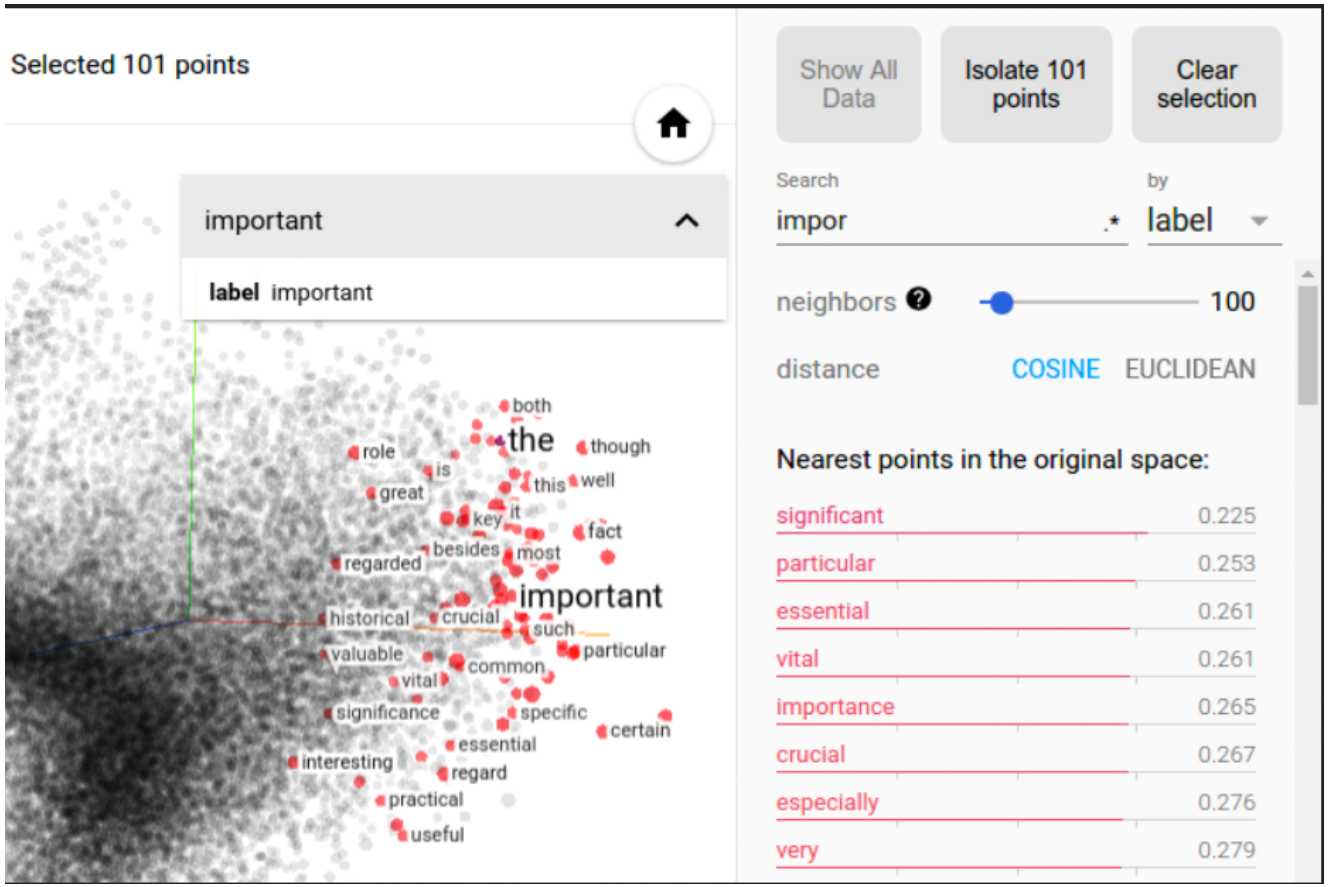



Figura: Similaridade com Word2Vec

# Operações com vetores - linearidades

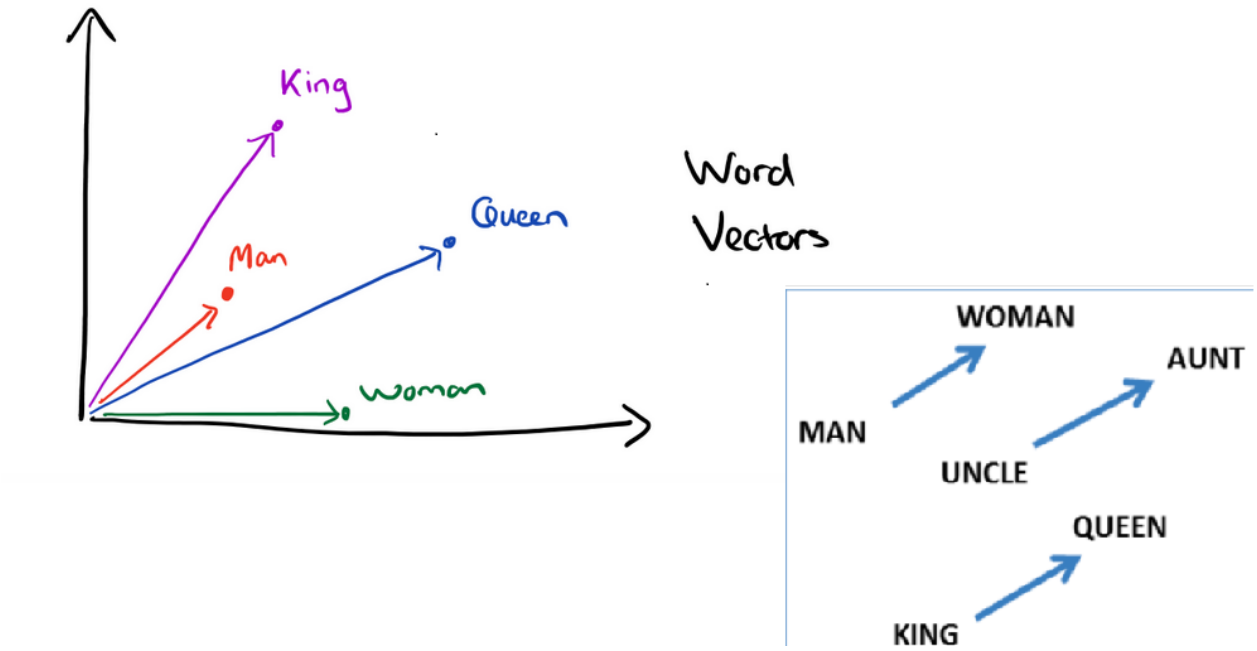
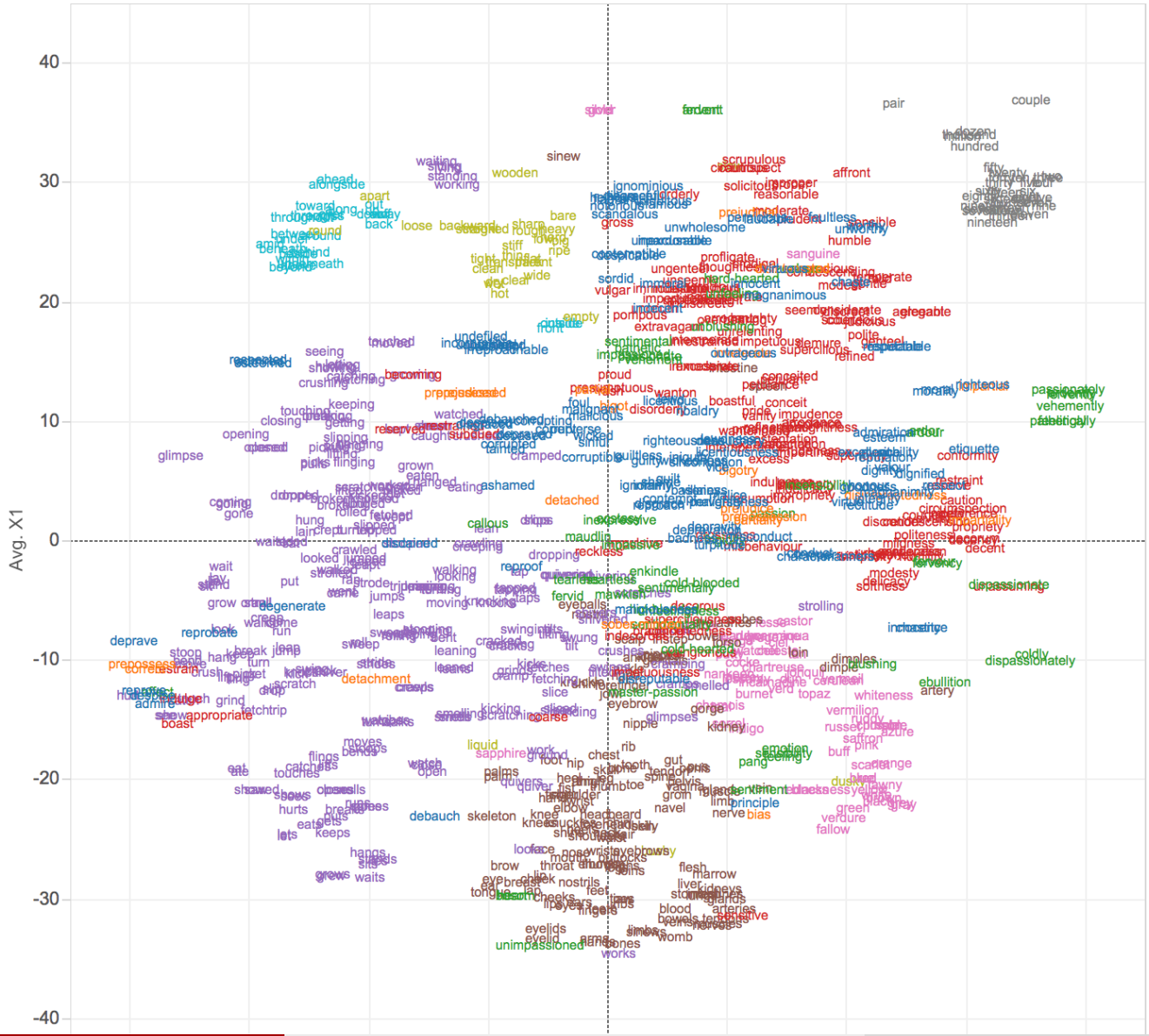



Figura: Relações lineares

# O que um vetor codifica?



- Limitações e questões

# Limitações e questões

- Sensibilidade ao corpus de treinamento
- Na prática, é um modelo muito bem sucedido para tarefas de NLP, mas a semantização do modelo é algo complexo.
- Modelo é treinado a partir de uma quantidade absurda de dados linguísticos. Não possui nenhuma expectativa de ter alguma veracidade neurolinguística.
- A noção de 'relação semântica' é intuitiva, mas formalmente vaga.

# Próximos passos

- Investigar um tipo específico de relação semântica: a sinonímia.
- Analisar o funcionamento da sinonímia nos modelos.
- Criar classificadores de sinônimos para esse tipo de modelo, usando abordagens matemáticas e/ou linguísticas.

# References


- Mikolov et al. (2013). Linguistic regularities in continuous space word representations.*Proceedings of NAACL-HLT 2013*.
- Jurafsky, D. & Martin, J.(2018). Speech and Language Processing.*3rd Edition draft, 2018*.
- Bengio, Y et al (2003). A Neural Probabilistic Language Model.*Journal of Machine Learning Research 3, 2003*.
- Manning, C. et al (2008). Introduction to information retrieval.*Cambridge University Press, 2018*.
- Widdows, D. (2004). Geometry and Meaning.*CSLI Publications, 2004*.
- Tensorflow, Vector Representations of words (2018).*https://www.tensorflow.org/tutorials/representation/word2vec*.
- Colyer, A. The amazing power of word vectors (2016).*https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/*.
- Deisenroth, M.P. et al (2018). Mathematics for Machine Learning.*Cambridge University Press, 2018*.


Figura: Obrigado!