

# SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking

Monique Monteiro – [moniquelouise@gmail.com](mailto:moniquelouise@gmail.com)

# Conceitos Importantes e Principais Contribuições

## Representação esparsa

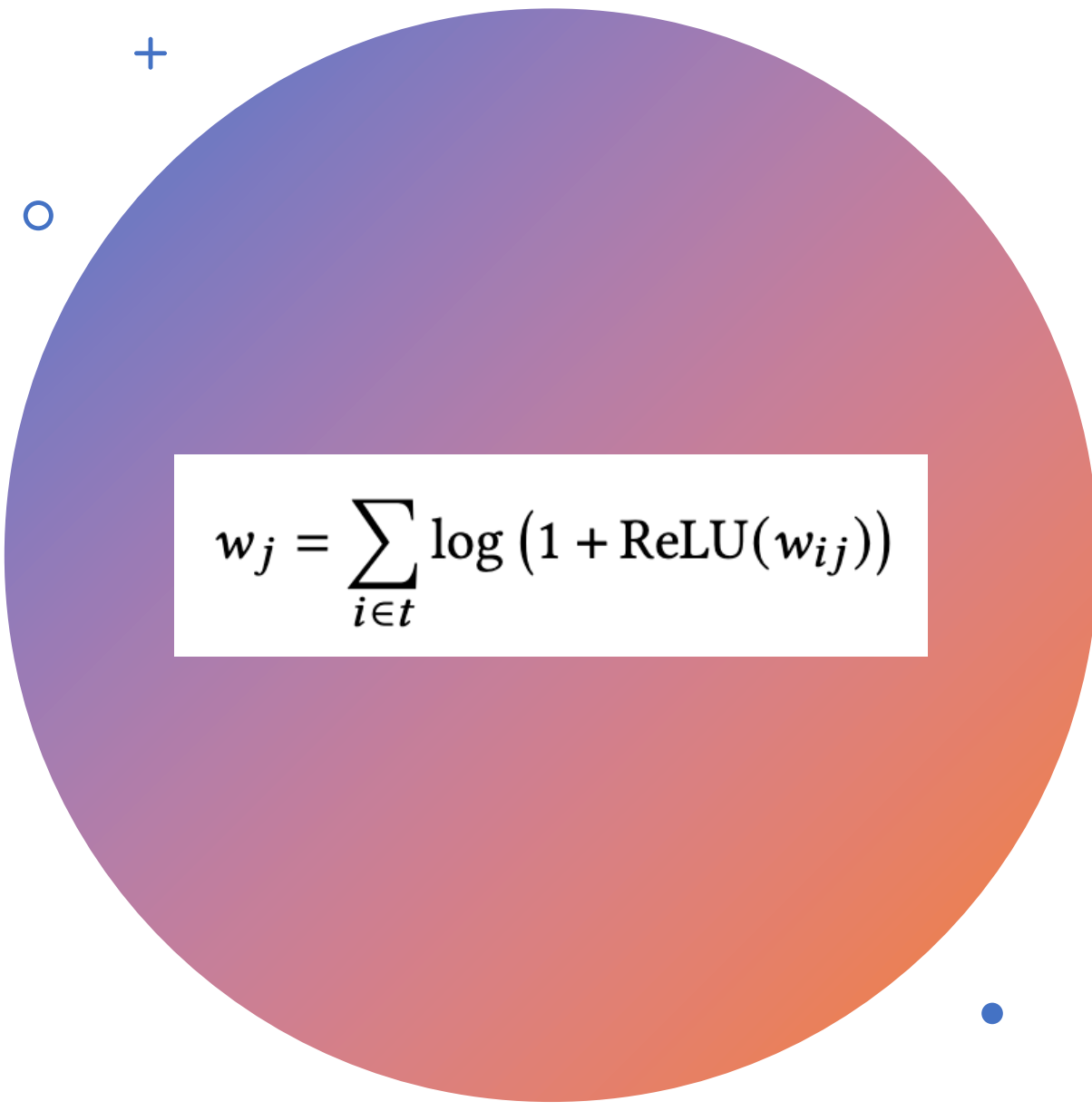
- Propriedades desejáveis de modelos *bag-of-words* (*exact matching*, eficiência de índices invertidos, interpretabilidade)
- Similaridade semântica (expansão de *queries* e documentos)

## Necessidade de métodos onde:

- A maior parte da computação possa ser feita offline
- Inferência online seja rápida

## Sparse Lexical AnD Expansion (SPLADE)

- Expansão de documentos

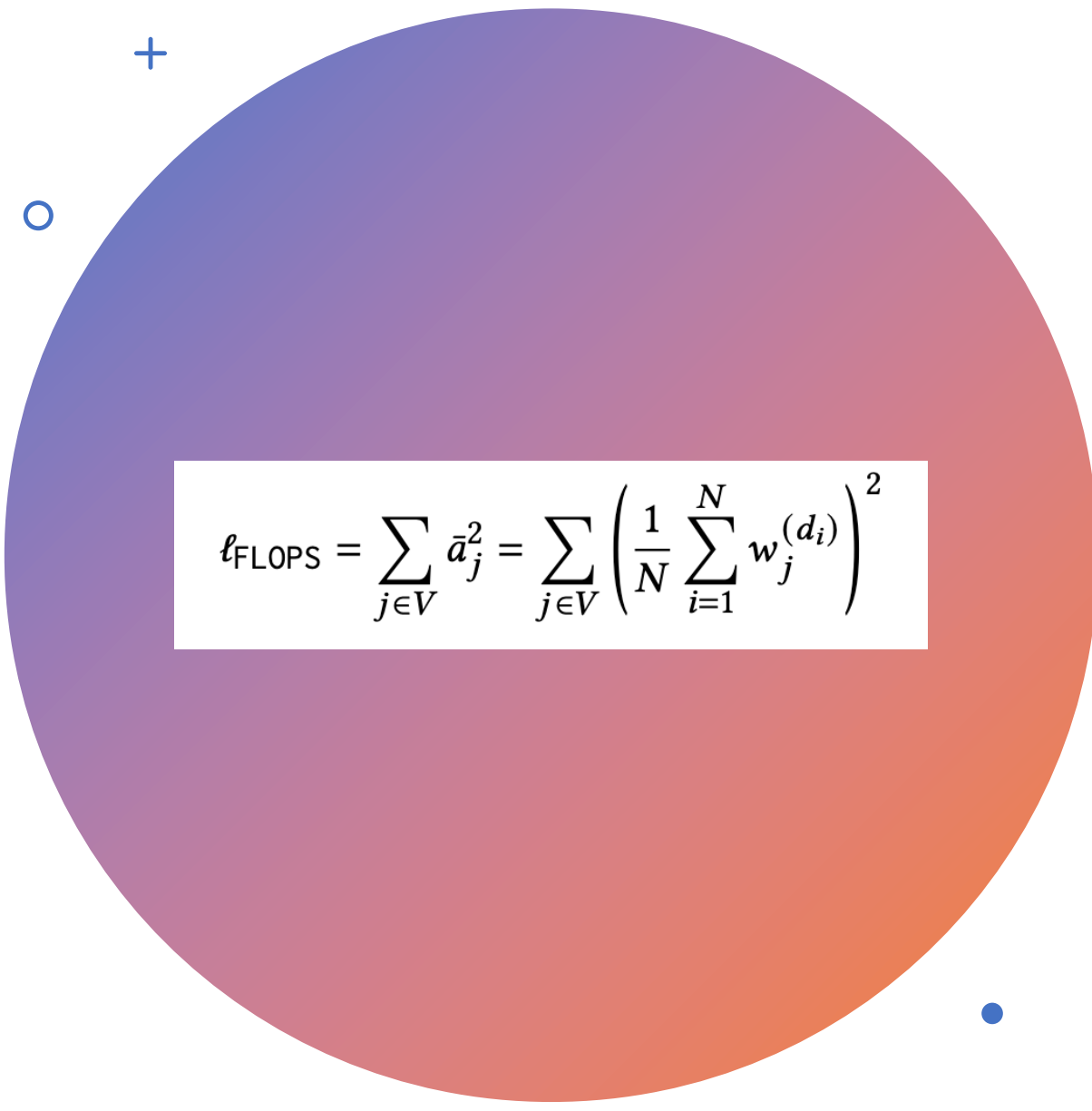

$$w_j = \sum_{i \in t} \log (1 + \text{ReLU}(w_{ij}))$$

# SPLADE

- $W_j$ : contribuição de cada termo do vocabulário
- Efeito de saturação que impede que alguns termos dominem:
  - Uso do logaritmo
- Esparsividade:
  - ReLU: se  $w_{ij} < 0 \rightarrow \text{ReLU}(w_{ij}) = 0 \rightarrow w_j = 0$

# Função de perda

$$\mathcal{L}_{rank-IBN} = -\log \frac{e^{s(q_i, d_i^+)}}{e^{s(q_i, d_i^+)} + e^{s(q_i, d_i^-)} + \sum_j e^{s(q_i, d_{i,j}^-)}}$$


$$\ell_{\text{FLOPS}} = \sum_{j \in V} \bar{a}_j^2 = \sum_{j \in V} \left( \frac{1}{N} \sum_{i=1}^N w_j^{(d_i)} \right)^2$$

# Dúvida básica

- Regularização baseada no tempo de computação da consulta
  - $d_i$  é binário?
- 