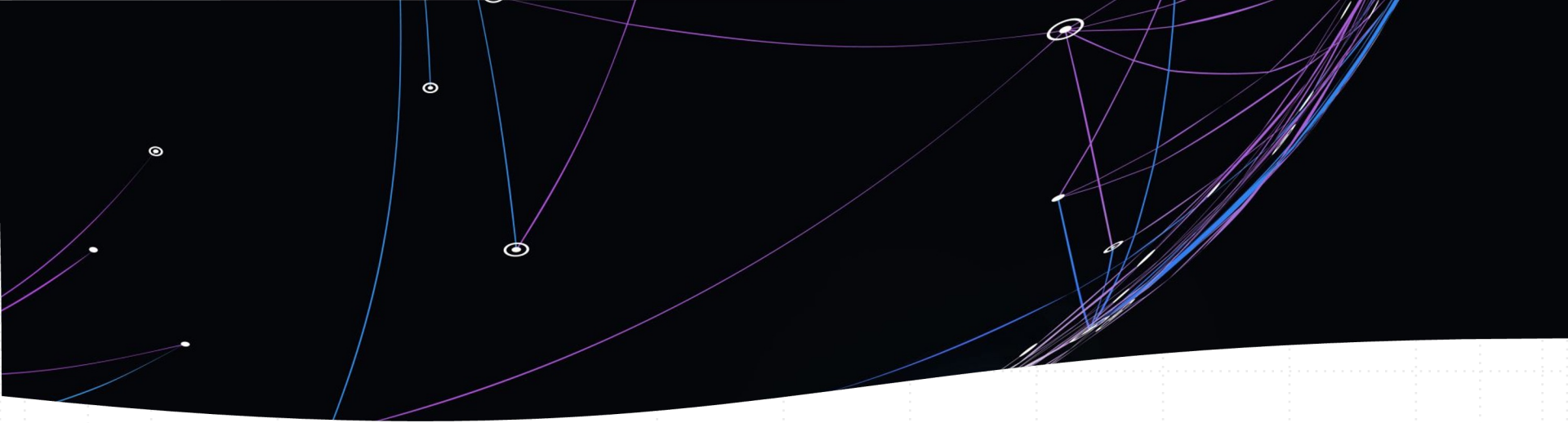


SPLADE-V2 para Português

Leonardo Bernardi de Avila – leo_avila92@msn.com

Monique Louise Monteiro – moniquelouise@gmail.com

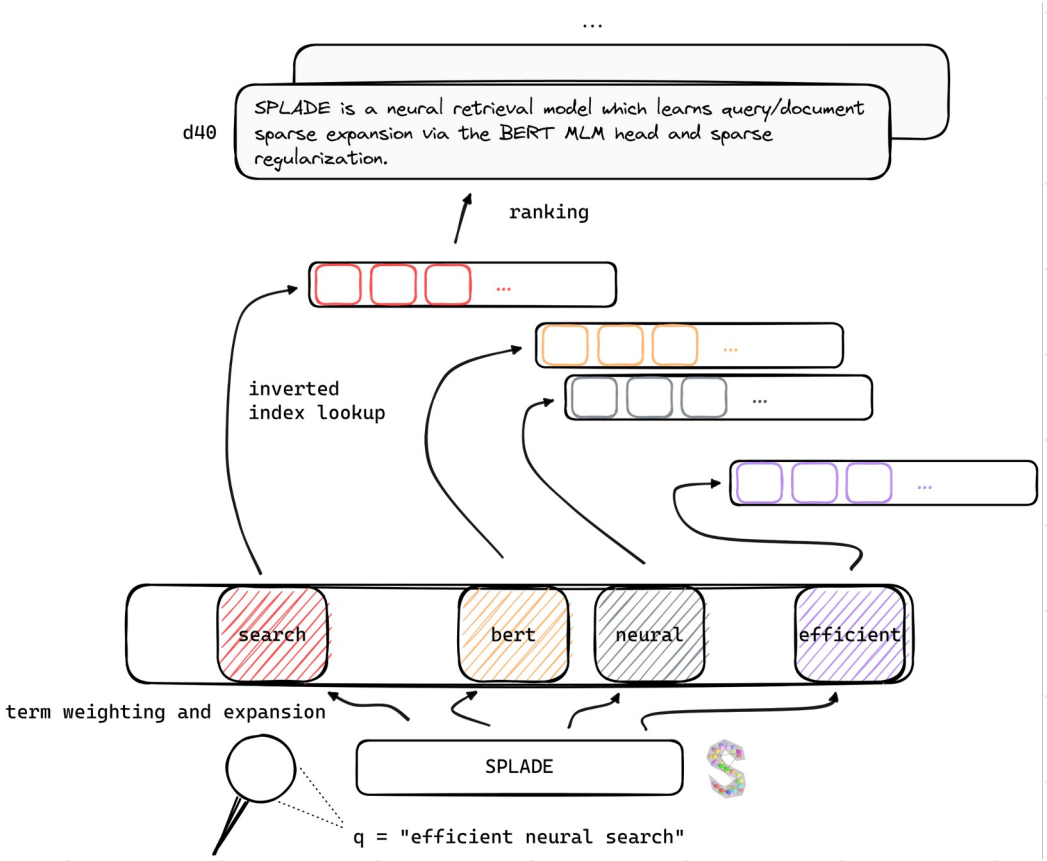
Descrição do projeto

Treinar e validar SPLADE-v2 para Português

Hipóteses testadas

- Uso do BERTimbau como modelo-base
- Uso do T5 Encoder como modelo-base

Descrição do projeto



Metodologia empregada

Reutilização do código do SPLADE-v2

Uso de *forks* para experimentos específicos

ex.: dataloader ids/corpus, ptt5

Datasets

- Treinamento
  - mMARCO
- Teste
  - mRobust

mMARCO

triplets:

39.780.811 (query, pos_doc, neg_doc)

triplets_ids:

39.780.811 (query_id, pos_doc_id, neg_doc_id)

corpus:

Mapeamento ids → textos (ex.: dataset de validação)

queries:

Mapeamento ids → textos (ex.: dataset de validação)

mRobust

característica: bastante ruidoso

queries: 250 queries

```
301     Identifique as organizações que participam da atividade criminosa internacional, a atividade e, se possível, as organizações colaboradoras e os países envolvidos.
302     A doença da poliomielite (poliomielite) está sob controle no mundo?
303     Identifique realizações positivas do telescópio Hubble desde que foi lançado em 1991.
304     Faça uma lista de mamíferos considerados em perigo, identifique seu habitat e, se possível, especifique o que os ameaça.
305     Quais são os veículos de passageiros mais resistentes a choques e menos resistentes a choques?
306     Quantos civis não combatentes foram mortos nas várias guerras civis na África?
307     Identifique projetos hidrelétricos propostos ou em construção por país e localização. É desejável uma descrição detalhada da natureza, extensão, propósito, problemas e consequências.
308     Quais são as vantagens e / ou desvantagens dos implantes dentários?
```

corpus: 528.032 docs

```
LA021690-0184   <P> 16 de fevereiro de 1990, sexta-feira, P.M. Final </P>
LA021690-0185   <P> 16 de fevereiro de 1990, sexta-feira, P.M. Final </P> <P> VENCEDOR DO CAMPO DE DIXIE BELLE TOPS HOT SPRINGS </P> <P> Duas potras invictas e Breezy Macbree, vencedora da Dixie Belle, lidera um field de 1
LA021690-0186   <P> 16 de fevereiro de 1990, sexta-feira, P.M. Final </P>
LA021690-0187   <P> 16 de fevereiro de 1990, sexta-feira, P.M. Final </P>
LA021690-0188   <P> 16 de fevereiro de 1990, sexta-feira, Orange County Edition </P> <P> CRIAÇÃO DE AVES: WESTERN FLYCATCHER </P> <P> WESTERN FLYCATCHER </P> <P> (Empidonax difficilus Baird) </P> <P> Descrição: Verde acast
LA021690-0189   <P> 16 de fevereiro de 1990, sexta-feira, edição de Orange County </P> <P> NA VIDA COMO NA ESCRITA, SUSAN GRIFFIN ENDEREÇA GRANDES TEMAS </P> <P> A voz de Susan Griffin é clara, mas suave o suficiente para
LA021690-0190   <P> 16 de fevereiro de 1990, sexta-feira, edição doméstica </P> <P> O FABRICANTE DE CINEMA ITALIANO FAZ NOME PARA SI MESMO; </P> <P> FILMES: 'CINEMA PARADISO' DE GIUSEPPE TORNATORE RECEBEU UMA NOMINAÇÃO DE
LA021690-0191   <P> 16 de fevereiro de 1990, sexta-feira, edição doméstica </P> <P> DO SHIELD CARRIER PARA OS PRINCIPAIS PAPÉIS CIRÚRGICOS; </P> <P> ÓPERA: RODNEY GILFRY ACHOU QUE SERIA PROFESSOR DE MÚSICA. EM VEZ, ESTÁ CA
LA021690-0192   <P> 16 de fevereiro de 1990, sexta-feira, edição doméstica </P> <P> SEM CERIMÔNIA, ELLIOTT ESTÁ SUPERANDO; </P> <P> PISTA E CAMPO: ELE É O HERDADEIRO APARENTE PARA A TRADIÇÃO BRITÂNICA DE MÉDIA DISTÂNCIA. <
```

Estratégia de avaliação

- Durante o treinamento:
  - Validação segundo a loss (split da base de treinamento)
  - Validação de métricas relativas a indexação/recuperação (base definida pelo naver/splade e "traduzida" na nossa versão)
- Avaliação no mRobust

Métricas

- As mesmas utilizadas pela versão original do SPLADE-V2
  - nDCG@10
  - nDCG@20
  - MRR@10
  - R@1000

Resultados Esperados

	en	*fr*	*pt*	*it*	*id*	*ru*	*es*	*de*	*zh*	**avg**
					nDCG@20					
BM25	0.389	0.389	0.389	0.387	0.383	0.372	0.364	0.333	0.289	0.367
mT5	0.466	0.376	0.391	0.384	0.374	0.372	0.402	0.375	0.358	0.389
mColBERT	0.362	0.302	0.323	0.305	0.287	0.265	0.309	0.280	0.262	0.300
					nDCG'@20					
BM25	0.394	0.418	0.409	0.411	0.407	0.403	0.394	0.372	0.349	0.396
mT5	0.486	0.429	0.439	0.436	0.432	0.431	0.454	0.435	0.418	0.440
mColBERT	0.414	0.383	0.401	0.379	0.367	0.348	0.389	0.361	0.345	0.377
					R@1000					
BM25	0.649	0.655	0.657	0.628	0.649	0.627	0.640	0.514	0.517	0.616
mColBERT	0.597	0.526	0.549	0.525	0.510	0.475	0.547	0.503	0.423	0.518

Table 2: Main results in the mRobust04 dataset. MT5 and MCOLBERT were finetuned on mMARCO.

Fonte: https://arxiv.org/pdf/2209.13738.pdf

Treinamento SPLADE-v2

Ranking loss. Let $s(q, d)$ denote the ranking score obtained via dot product between q and d representations from Eq. (2). Given a query q_i in a batch, a positive document d_i^+, a (hard) negative document d_i^- (e.g. coming from BM25 sampling), and a set of negative documents in the batch (positive documents from other queries) $\{d_{i,j}^-\}_j$, we consider a constrastive loss, which can be interpreted as the maximization of the probability of the document d_i^+ being relevant among the documents d_i^+, d_i^- and $\{d_{i,j}^-\}$:

$$\mathcal{L}_{rank-IBN} = -\log \frac{e^{s(q_i,d_i^+)}}{e^{s(q_i,d_i^+)} + e^{s(q_i,d_i^-)} + \sum_j e^{s(q_i,d_{i,j}^-)}} \quad (3)$$

The *in-batch negatives* (IBN) sampling strategy is widely used for training image retrieval models, and has shown to be effective in learning first-stage rankers [13, 16, 24].

Learning sparse representations. The idea of learning sparse representations for first-stage retrieval dates back to SNRM [32], via ℓ_1 regularization. Later, [22] pointed-out that minimizing the ℓ_1 norm of representations does not result in the most efficient index, as nothing ensures that posting lists are evenly distributed. Note that this is even more true for standard indexes due to the Zipfian nature of the term frequency distribution. To obtain a well-balanced index, *Paria et al.* [22] introduce the FLOPS regularizer, a smooth relaxation of the average number of floating-point operations necessary to compute the score between a query and a document, and hence directly related to the retrieval time. It is defined using a_j as a continuous relaxation of the activation (i.e. the term has a non-zero weight) probability p_j for token j, and estimated for documents d in a batch of size N by $\bar{a}_j = \frac{1}{N}\sum_{i=1}^{N} w_j^{(d_i)}$. This gives the following regularization loss

$$\ell_{\text{FLOPS}} = \sum_{j \in V} \bar{a}_j^2 = \sum_{j \in V} \left(\frac{1}{N} \sum_{i=1}^{N} w_j^{(d_i)} \right)^2 \quad (4)$$

Overall loss. By jointly optimizing the model in Eq. (2) with ranking and regularization losses, SPLADE combines the best of both worlds for end-to-end training of sparse, expansion-aware representations of documents and queries:

$$\mathcal{L} = \mathcal{L}_{rank-IBN} + \lambda_q \mathcal{L}_{\text{reg}}^q + \lambda_d \mathcal{L}_{\text{reg}}^d \quad (5)$$

where $\mathcal{L}_{\text{reg}}$ is the sparse FLOPS regularization from Eq. 4. We use two distinct regularization weights (λ_d and λ_q) for queries and documents – allowing to put more pressure on the sparsity for queries, which is critical for fast retrieval.

Fonte: https://arxiv.org/abs/2109.10086

Etapas de execução

1. (Re)leitura de referências bibliográficas
2. Avaliação do repositório naver/splade
3. Organização dos datasets (mMARCO, mRobust)
4. Finetuning
    - Cerca de 9 experimentos
    - Variações:
        i. tamanho do dataset
        ii. carregamento dos dados
        iii. encoder
        iv. Hiperparâmetros (max length, taxa de aprendizagem, regularização)
        v. Amostras negativas
        vi. Novo branch "hf" (ainda muito instável)

Dificuldades

- triplets com ~39 milhões de linhas + código do naver/splade estoura memória do colab (A100 80GB)
  - **Solução:** fork no github do naver/splade para alterar o dataloder para rodar com os triplets ids.

- custos colab
  - **Solução Parcial:** com o fork, houve redução de memória e foi possível rodar parte dos treinamentos na V100

- impossibilidade de realizar muitos testes (modelos, configurações de FLOPS, número iterações)
  - custos do colab
  - demora na execução
    - treinamento: ~20 horas cada configuração de experimento (ex.: A100)
    - indexação mRobust04: ~0,5 hora
    - retrieval: ~15 minutos

Experimentos - BERTimbau

Código naver/splade:

10 MM de triplets:

1) max_length: 256, triplets negatives, FLOPS: (λ_q: 5e-4, λ_d: 3e-4, T=3)

Fork naver/splade com ajuste no dataloader (triplets ids e carregamento dinâmico):

39 MM de triplets ids:

2) max_length: 256, triplets negatives, FLOPS: (λ_q: 5e-4, λ_d: 3e-4, T=3)

3) max_length: 384, triplets negatives, FLOPS: (λ_q: 5e-4, λ_d: 3e-4, T=3)

4) configs splade-max:

max_length: 384, triplets + in-batch negatives, FLOPS: (λ_q:3e-4 , λ_d:1e-4, T=50.000)

Experimentos - T5 Encoder

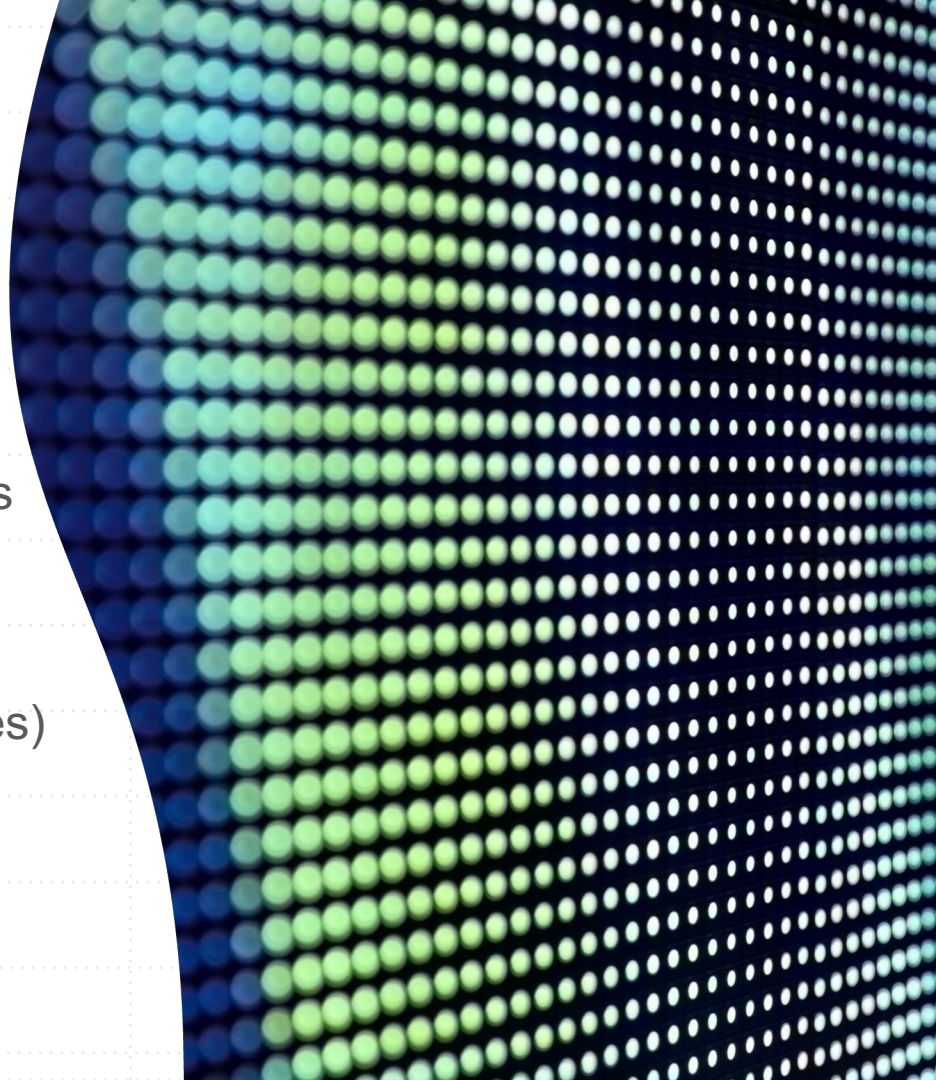
- Uso do PTT5v2
- Dificuldade de convergência do modelo
  - Diferentes taxas de aprendizagem foram testadas
- Problemas:
  - MRR@10 igual a zero nas primeiras milhares de iterações
  - MRR@10 diminuindo muito lentamente
  - Instabilidade nos valores da *loss*

Resultados - mRobust

modelo	triplets	settings	ndcg@10	ncdg@20	MRR@10	recall@1000
bm25	-	k1 = 0.82, b = 0.68	0.4098	0.3893	0.6690	0.6572
neuralmind/bert-base-portuguese-cased	10 MM	triplet negatives, max length = 256	0.2797	0.2581	0.5183	0.4006
neuralmind/bert-base-portuguese-cased	39 MM	triplet negatives, max length = 256	0.3051	0.2740	0.5529	0.4070
neuralmind/bert-base-portuguese-cased	39 MM	triplet negatives, max length = 384	0.3149	0.2862	0.5885	0.4636
neuralmind/bert-base-portuguese-cased	39 MM	in-batch, max length = 384	0.3295	0.2977	0.5929	0.4572

Trabalhos futuros

- Quebra e janelamento (sliding) dos textos longos em passagens menores
- Finetuning com 512 tokens
- Treinar por mais iterações
- Avaliação no mMARCO dev (6980 queries)
- Destilação
- Comparação com OpenAI Embeddings
- Investigações adicionais com PTT5-v2



Referências

- SPLADE
- SPLADE v2
- Repositório SPLADE V2
- Albertina PT-BR
  - Repositório: https://huggingface.co/PORTULAN/albertina-ptbr
  - Paper: https://arxiv.org/abs/2305.06721
- BERTimbau
- mMarco
- mRobust
  - Repositório: https://huggingface.co/datasets/unicamp-dl/mrobust
  - Paper: https://arxiv.org/pdf/2209.13738.pdf
- Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation