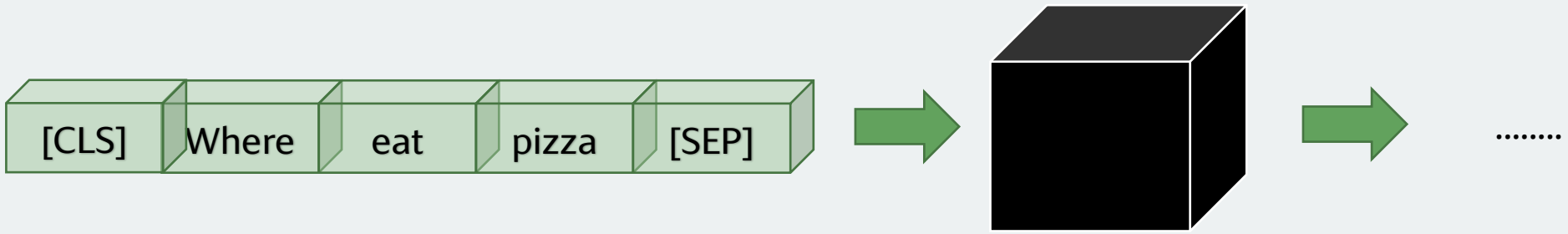
Notebook – Splade

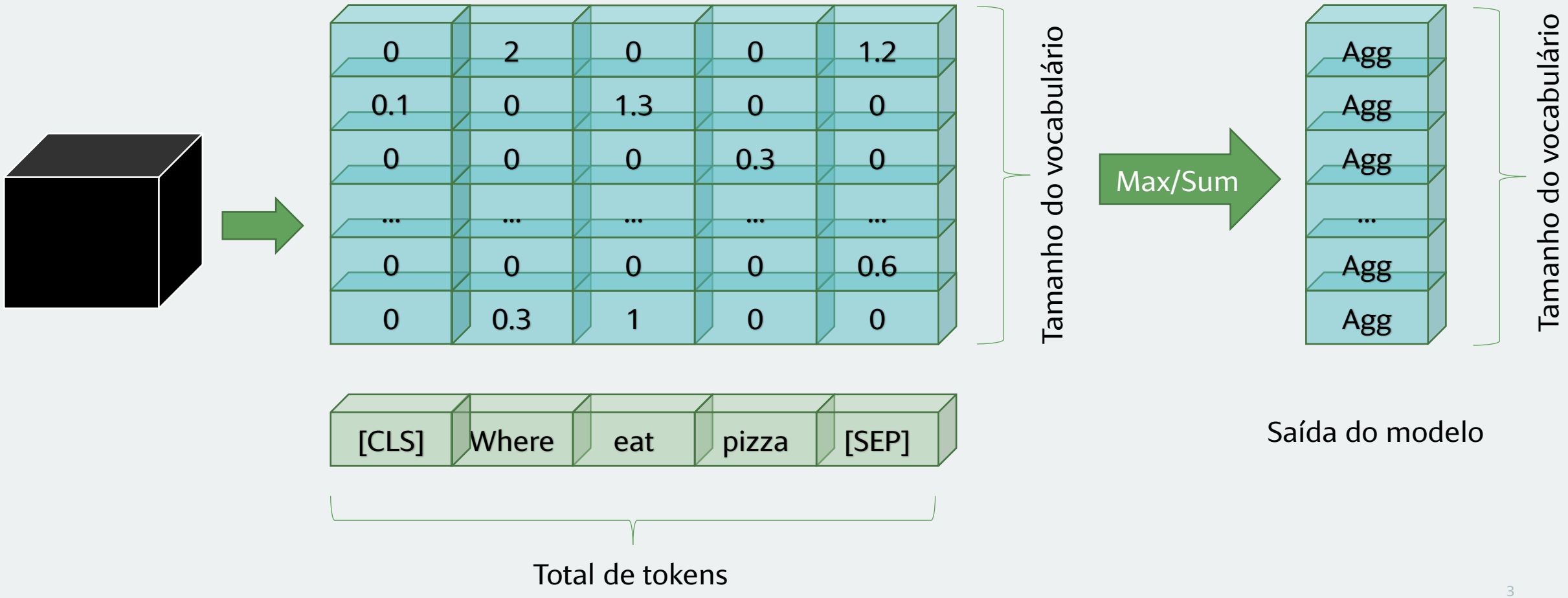
Leandro Carísio

Conceitos do exercício

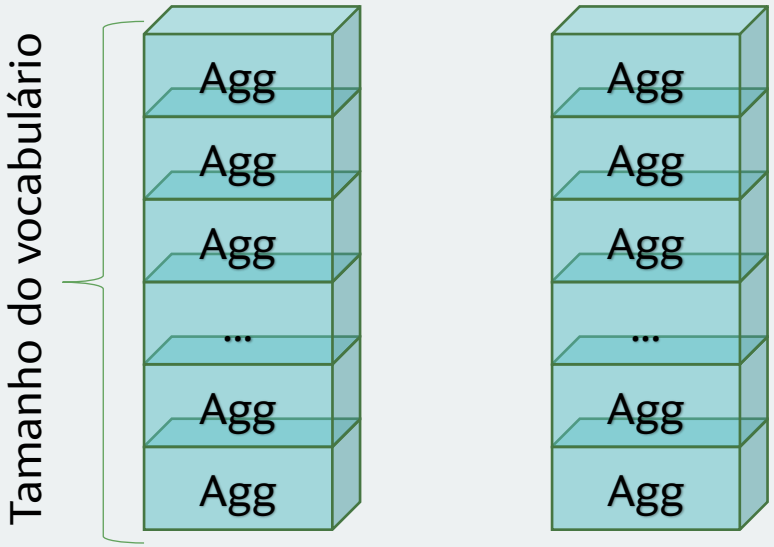


Texto de entrada

Conceitos do exercício



Conceitos do exercício



Score(d, q) = produto escalar entre d e q

No SpladeV2 os elementos do vetor da query são {0, 1}.

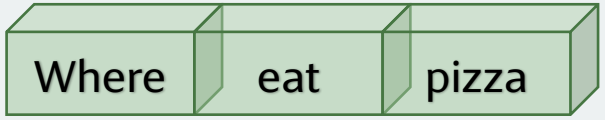
⇒ Se o corpus for pequeno, dá pra guardar a matriz de documento e resolver com uma multiplicação de matrizes.

⇒ Se o corpus for grande, dá pra resolver com um índice invertido guardando como chave o token (ou token_id) e a lista de documento e score do documento

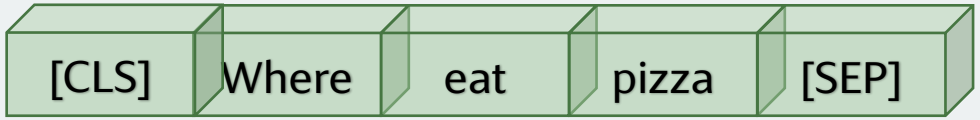
Problemas e soluções no desenvolvimento

1. Resultado da primeira simulação era nDCG = 0.0000

Solução: Eu estava gerando os textos sem os tokens [CLS] e [SEP]



' , - . / a cheese dave dom food fred is meat pie pizza restaurant smith the

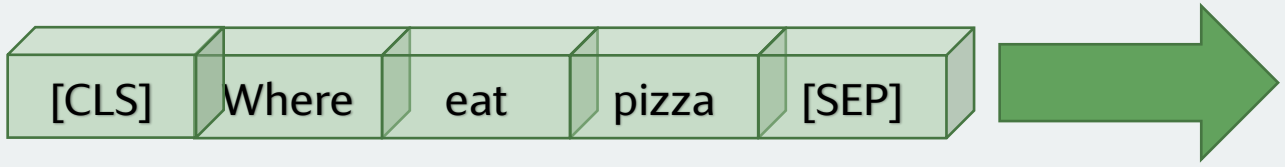


anywhere best cafe country culture eat eating famous farm favorite food habitat headquarters hotel location locations murphy pie pizza place places restaurant restaurants shop that venue visit where

Problemas e soluções no desenvolvimento

2. naver/splade_v2_distil estava gerando resultado nada a ver:

Solução após ver caderno do Pedro Gabriel: Usar DistilBertTokenizer e DistilBertForMaskedLM (ou usar os AutoModels...)



BertTokenizer
BertForMaskedLM

('ব', tensor(0.0004)), ('ে', tensor(0.4884)), ('ভ', tensor(0.5233)),
('শ', tensor(0.1958)), ('ক', tensor(0.4602)), ('ઢ', tensor(0.8806)),
('ষ', tensor(0.6900)), ('ச', tensor(0.5440)), ('ଚ', tensor(0.4128)),
('স', tensor(0.1393)), ('ட', tensor(0.1255)), ('ఒ', tensor(0.2532)),
('ো', tensor(0.7554)), ('த', tensor(0.0173)), ('ಳ', tensor(0.3629)),
('ি', tensor(0.5658)), ('ந', tensor(0.1367)), ('ಕ', tensor(0.8841)),
('ী', tensor(0.3767)), ('ண', tensor(0.2874)), ('ಇ', tensor(0.5743)),
('ৌ', tensor(0.7418)), ('ப', tensor(0.4996)), ('ಪ', tensor(0.1146)),
('।', tensor(0.1557)), ('ம', tensor(0.8778)), ('ಸ', tensor(0.1561)),
('ং', tensor(0.6785)), ('ய', tensor(0.0894)), ('ಜ', tensor(0.0676)),
('অ', tensor(0.2900)), ('வ', tensor(0.3731)), ('ಷ', tensor(0.1247)),
('আ', tensor(0.3225)), ('ా', tensor(0.2420)), ('ಅ', tensor(0.3174)),
('ই', tensor(0.6525)), ('ీ', tensor(0.3318)), ('ಎ', tensor(0.2280)),
('উ', tensor(0.2975)), ('ూ', tensor(0.4002)), ('ಠ', tensor(0.1788)),
('ক', tensor(0.2984)), ('ಆ', tensor(0.2981)),

Problemas e soluções no desenvolvimento

3. Lentidão na etapa de inferência

Solução: Ordenar o dataset antes de indexar e usar fp16 (ideia de Marcos Piau) diminui o tempo de processamento

Solução: Máscara estava na cpu (e não na gpu) (tks Borela)

```
mask_tokens_validos = 1 - inputs['special_tokens_mask']
mask = mask_tokens_validos.squeeze().unsqueeze(-1).expand(logits.size()).to(device)
```

CPU GPU

```
mask_tokens_validos = 1 - inputs['special_tokens_mask'].to(device)
mask = mask_tokens_validos.squeeze().unsqueeze(-1).expand(logits.size())
```

GPU

Resultados interessantes/inesperados

A ordem dos fatores altera o resultado:

where eat pizza: anywhere best cafe country culture eat eating famous farm favorite food habitat headquarters hotel location locations murphy pie pizza place places restaurant restaurants shop that venue visit where

pizza where eat: cafe country culture eat eating famous food habitat headquarters hotel italy location locations menu murphy pie pizza place places position restaurant restaurants venue where

Resultados interessantes/inesperados

Muitos tokens são extraídos do CLS/SEP (where eat pizza)

Com contribuição do CLS/SEP: anywhere baker best booth bowl cafe church city country culture dave diner dish domino don eat eating eden famous farm favorite find food foods garden habitat headquarters hotel hut italy joe kitchen location locations meat murphy pie pizza place places restaurant restaurants sandwich shop store that variety venue visit website where york

Sem contribuição do CLS/SEP: anywhere best cafe country culture eat eating famous farm favorite food habitat headquarters hotel location locations murphy pie pizza place places restaurant restaurants shop that venue visit where

Resultados interessantes/inesperados

Parece que os tokens extraídos do CLS/SEP não são muito relevantes

naver/splade-cocondenser-ensembledistil	**nDCG@10**
Multiplicação matricial, fp32	De 0,7290 para 0,7276
Índice, fp16	De 0,7269 para 0,7242

=> Apesar de reduzir o tamanho do índice, é uma redução singela (de 242,1 MB para 235,6 MB). Espera-se uma redução na latência.

Resultados

Método	nDCG@10
BM25 (Aula 5)	0,5956
BM25 doc. original + expansão (Aula 5)	0,6719
Implementação busca densa (Aula6)	0,3322
Resultados desta aula - naver/splade-cocondenser-ensembledistil	
max, f16, índice	0,7269
max, f16, índice, sem contrib. SEP e CLS	0,7242
max, f32, matriz	0,7290
max, f32, índice	0,7282
sum, f32, matriz	0,1921
sum, f16, índice	0,1931

Resultados

Modelo (max, f16, índice, com contrib. SEP e CLS)	nDCG@10
baseplate/splade-cocondenser-selfdistil	0,7348
naver/splade-cocondenser-selfdistil	0,7348
naver/splade-cocondenser-ensembledistil	0,7269
naver/splade_v2_distil	0,7103

Tópico para discussão

1. A implementação usando a função SUM como agregação deu um resultado muito ruim.

2. Calcular o score conforme proposto no SPLADEv2 piorou os resultados.

3. Parece que a melhor combinação foi a função agregação MAX (SPLADEv2) com o score calculado conforme SPLADEv1.

Faz sentido isso? Ou foi algum bug no desenvolvimento?

Obrigado

Leandro Carísio

carisio@gmail.com