T5 + doc2query

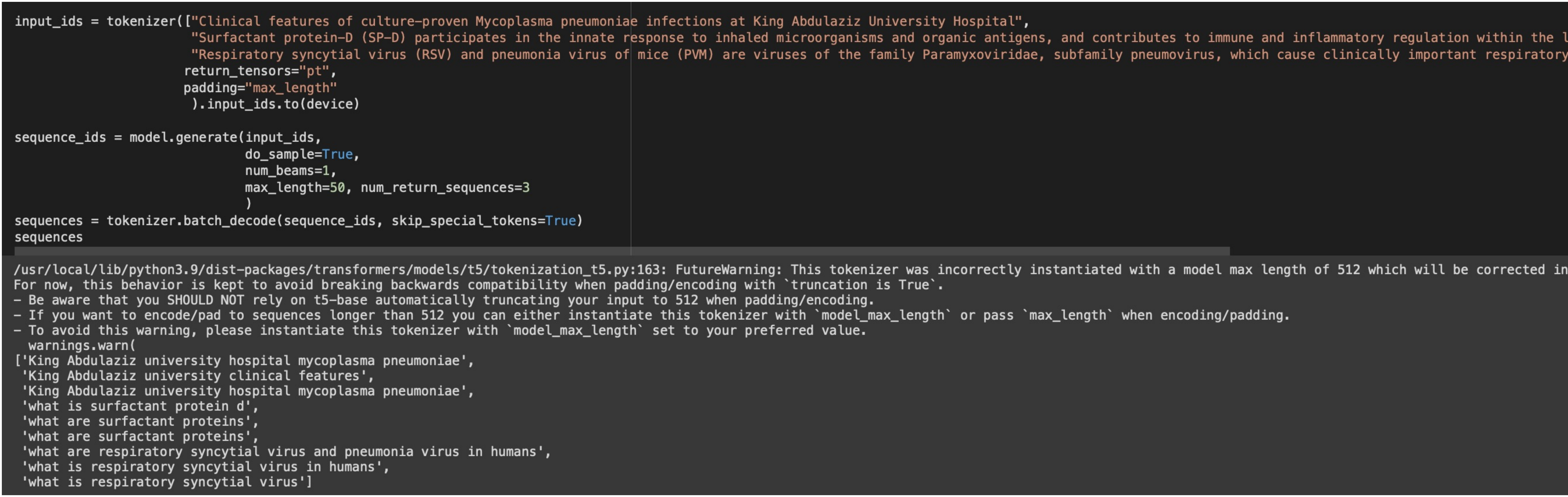
Monique Monteiro – moniquelouise@gmail.com

Conceitos importantes

- T5 (*__T__ext-__to__-__T__ext __T__ransfer __T__ransformer*) paper:
    - Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. arXiv:1910.10683 [cs.CL] (2019)

Técnica para garantir corretude

Inspeção visual das consultas geradas



```
input_ids = tokenizer(["Clinical features of culture-proven Mycoplasma pneumoniae infections at King Abdulaziz University Hospital",
                       "Surfactant protein-D (SP-D) participates in the innate response to inhaled microorganisms and organic antigens, and contributes to immune and inflammatory regulation within the l
                       "Respiratory syncytial virus (RSV) and pneumonia virus of mice (PVM) are viruses of the family Paramyxoviridae, subfamily pneumovirus, which cause clinically important respiratory
                      return_tensors="pt",
                      padding="max_length"
                       ).input_ids.to(device)

sequence_ids = model.generate(input_ids,
                              do_sample=True,
                              num_beams=1,
                              max_length=50, num_return_sequences=3
                              )
sequences = tokenizer.batch_decode(sequence_ids, skip_special_tokens=True)
sequences
```

```
/usr/local/lib/python3.9/dist-packages/transformers/models/t5/tokenization_t5.py:163: FutureWarning: This tokenizer was incorrectly instantiated with a model max length of 512 which will be corrected in
For now, this behavior is kept to avoid breaking backwards compatibility when padding/encoding with `truncation is True`.
- Be aware that you SHOULD NOT rely on t5-base automatically truncating your input to 512 when padding/encoding.
- If you want to encode/pad to sequences longer than 512 you can either instantiate this tokenizer with `model_max_length` or pass `max_length` when encoding/padding.
- To avoid this warning, please instantiate this tokenizer with `model_max_length` set to your preferred value.
  warnings.warn(
['King Abdulaziz university hospital mycoplasma pneumoniae',
 'King Abdulaziz university clinical features',
 'King Abdulaziz university hospital mycoplasma pneumoniae',
 'what is surfactant protein d',
 'what are surfactant proteins',
 'what are surfactant proteins',
 'what are respiratory syncytial virus and pneumonia virus in humans',
 'what is respiratory syncytial virus in humans',
 'what is respiratory syncytial virus']
```

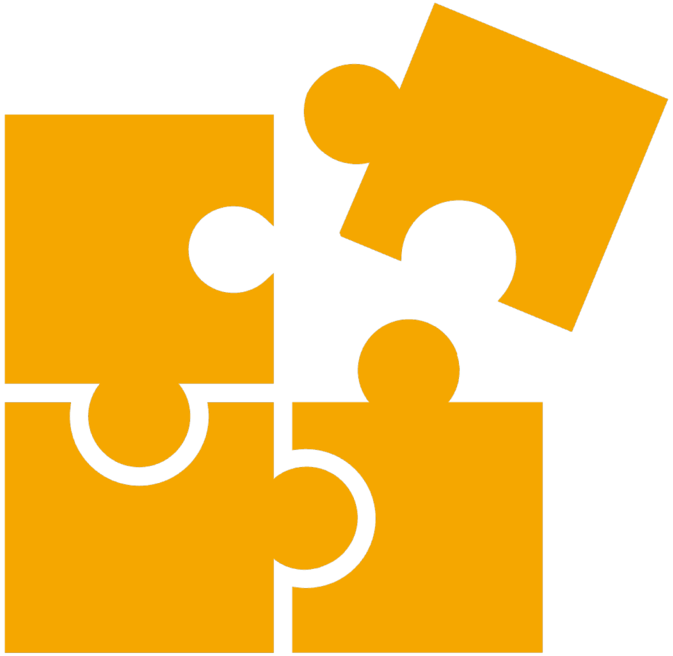
Hiperparâmetros - Finetuning do T5

Modelo	Otimizador	Batch size	Precisão mista	*Early stoping*	sacredBLEU
A	AdaFactor	32	Sim (bf16)[1]	Sim	19,92 (ép. 12)
B	AdaFactor	32	Não	Não	22,46
C	AdaFactor	32	Sim (fp16)	Não	Aprox. 11
D	AdamW	8 → 16 → 32	Sim (fp16)	Não	19,53

[1] https://huggingface.co/docs/transformers/v4.15.0/performance

Hiperparâmetros – Expansão dos documentos

- 5 consultas por passagem (1 “*greedy*”, 1 p/ *beam search* e 3 p/ amostragem)
    - Batch size = 8
- 18 consultas por amostragem (ideia inicial: 20 → 1 gs + 1 bs + 18 p/ amostragem)
    - Batch size = 32
- 3 consultas por amostragem
    - Batch size = 8



Resultados

Modelo	Número de consultas	NDCG@10
A	**18 (p/ amost.)**	**0,656**
A	3 (p/ amost.)	0,654
A	5 (1 gs + 1 bs + 3 p/ amost.)	0,651
B	5 (1 gs + 1 bs + 3 p/ amost.)	0,647

Impacto do processamento de arquivos sem texto

- Ao remover expansões de *queries* de arquivos sem texto:
  - NDCG@10 para 18 consultas: **0,654 → 0,656**
  - NDCG@10 para 3 consultas: **0,654 → 0,644 (!)**
- Na versão atual, apenas o conteúdo é expandido
- Conteúdo vazio gera muitas *queries* sem sentido (ex.: what/who, etc)
- Pendente de avaliação:
  - Expansão de títulos
  - Expansão por “janelas deslizantes”
    - https://github.com/castorini/docTTTTTquery/blob/master/convert_msmarco_doc_to_t5_format.py

Resultado interessante

- Crescimento do erro de validação (*overfitting*) com aumento do BLEU
  - O modelo aprendeu a gerar sentenças fluentes porém incorretas com alto grau de "*n-gram overlap*" para sacreBLEU
  - Conjunto de validação pequeno ou pouco representativo
  - Pequena redução na métrica-alvo (NDCG@10)