DLP Week-2 Graded Assignment-2

Download the BookCorpus dataset. Take every 7th sample (the indices are multiple of 7:[0,7,14,21,..]) from the entire dataset. This will result in a dataset with 10 million samples (exactly, 10,572,033). Use these samples to build a tokenizer with the BPE tokenization algorithm by varying the vocabulary size.

Normalizer: LowerCase
PreTokenizer: WhiteSpace
Model: BPE
Special tokens: [GO],[UNK],[PAD],[EOS]
PostProcessing: None

Tokenize the input text: "SEBI study finds 93% of individual F&O traders made losses between FY22 and FY24." using the following configurations.

1) Keep the vocabulary size at 5000 and tokenize the input text using the learned vocabulary. Choose the number of tokens returned by the tokenizer. **1 point**

- 18
- 22
- 25
- 28
- 32
- 60

2) Increase the vocabulary size to 10K, 15K and 32K. For each case, tokenize the same input with the newly learned vocabulary. Choose all the correct statements **2 points**

- [x] Increasing vocab size from 5K to 10K decreases the number of tokens
- [ ] Increasing vocab size from 5K to 10K increases the number of tokens
- [ ] Increasing vocab size from 10K to 15K decreases the number of tokens
- [x] Increasing vocab size from 10K to 15K does not change the number of tokens
- [x] Increasing vocab size from 15K to 32K further decreases the number of tokens

Yes, the answer is correct.
Score: 2

Accepted Answers:

Increasing vocab size from 5K to 10K decreases the number of tokens
Increasing vocab size from 10K to 15K does not change the number of tokens
Increasing vocab size from 15K to 32K further decreases the number of tokens

3) Download the pre-trained tokenizer file "hopper.json" used in the lecture, from here . The tokenizer was trained on all 70 million samples in the BookCorpus dataset. Tokenize the same input text using this "hopper" tokenizer. How many tokens are there?

[After finding the answer, take a moment to compare the hopper tokenizer with the previous one]

25

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 25

1 point

4) Suppose we know that the acronym "FY" will likely appear very frequently in most of the input text (assume the text comes from the financial domain). Therefore, we hope that adding it manually to the vocabulary might help. Add the token "FY" to the vocabulary and tokenize (use the Hopper tokenizer) the input text. Enter the number of tokens produced.

[Question to ponder: Does reducing the number of tokens helpful?]

22

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 22

1 point

5) Load the "bert-base-uncased" and "gpt2" tokenizers (use **AutoTokenizer** function from *transformers*). Which of the following special tokens are used in these tokenizers? **1 point**

- [ ] [GO]
- [x] [CLS]
- [ ] [BOS]
- [x] [SEP]
- [x] [<|endoftext|>]

6) By now, we have four tokenizers.

1. Custom tokenizer (vocab size 32K, trained on 10 million samples)
2. bert-base-uncased
3. gpt2
4. hopper

Use these four tokenizers to count the number of tokens for the entire "imdb" dataset (drop the "unsupervised" part of the dataset). Enter the tokenizers in order such that the size of the dataset (measured in tokens) as returned by the tokenizers is in ascending order. For example, if the first tokenizer yields the smallest number of tokens and the fourth tokenizer yields the largest, you would enter 1234 (without any spaces)."

4312

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Numeric) 4231

2 points

7) The statement that the special tokens and their respective token ids are model-specific (model here refers to a language model) is **1 point**

- True
- False

8) Suppose that the context length of the model is 128. Assume that a mini-batch of size 8 samples is passed to a tokenizer that corresponds to a model from hub. After tokenization, the maximum length of sample in the batch is 64. The statement that zero is appended to the "input ids" of the remaining samples to make the length 64 is **1 point**

- True
- False
- Insufficient information to conclude

given :

context_len = 128

batch_size = 8

max_len = 64 (after tokenizⁿ)

padding is generally added to models when learning to deal with batches of unequal sequence lengths.

It's not clear ⇒

(i) padding is applied to match the longest seq. of 64.

(ii) or it's applied to match the content len. of 128.