tfds.features.text.SubwordTextEncoder

View source on GitHub (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L41-L386)

Invertible `TextEncoder` using word pieces with a byte-level fallback.

Inherits From: `TextEncoder` (https://www.tensorflow.org/datasets/api_docs/python/tfds/features/text/TextEncoder)

```
features.text.SubwordTextEncoder(
ocab_list=None
```

Encoding is fully invertible because all out-of-vocab wordpieces are byte-encoded.

The vocabulary is "trained" on a corpus and all wordpieces are stored in a vocabulary file. To generate a vocabulary from a corpus, use `tfds.features.text.SubwordTextEncoder.build_from_corpus` (https://www.tensorflow.org/datasets/api_docs/python/tfds/features/text/SubwordTextEncoder#build_from_corpus).

Typical usage:

```
ld
ler = tfds.features.text.SubwordTextEncoder.build_from_corpus(
orpus_generator, target_vocab_size=2**15)
ler.save_to_file(vocab_filename)

d
ler = tfds.features.text.SubwordTextEncoder.load_from_file(vocab_filename)
 encoder.encode("hello world")
 = encoder.decode([1, 2, 3, 4])
```

Args	
`vocab_list`	`list<str>`, list of subwords for the vocabulary. Note that an underscore at the end of a subword indicates the end of the word (i.e. a space will be inserted afterwards when decoding). Underscores in the interior of subwords are disallowed and should use the underscore escape sequence.

Attributes	
`subwords`	
`vocab_size`	Size of the vocabulary. Decode produces ints [1, vocab_size).

Methods

build_from_corpus

View source (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L261-L337)

```
smethod
_from_corpus(
orpus_generator, target_vocab_size, max_subword_length=20,
ax_corpus_chars=None, reserved_tokens=None
```

Builds a `SubwordTextEncoder` based on the `corpus_generator`.

Args	
`corpus_generator`	generator yielding `str`, from which subwords will be constructed.
`target_vocab_size`	`int`, approximate size of the vocabulary to create.
`max_subword_length`	`int`, maximum length of a subword. Note that memory and compute scale quadratically in the length of the longest token.
`max_corpus_chars`	`int`, the maximum number of characters to consume from `corpus_generator` for the purposes of building the subword vocabulary.
`reserved_tokens`	`list<str>`, list of tokens that will always be treated as whole tokens and not split up. Note that these must contain a mix of alphanumeric and non-alphanumeric characters (e.g. "") and not end in an underscore.

Returns
`SubwordTextEncoder`.

decode

View source (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L91-L127)

```
le(
ds
```

Decodes a list of integers into text.

encode

View source (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L81-L89)

```
le(
s
```

Encodes text into a list of integers.

load_from_file

View source (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L252-L259)

```
ssmethod
_from_file(
filename_prefix
```

Extracts list of subwords from file.

save_to_file

View source (https://github.com/tensorflow/datasets/blob/v3.2.1/tensorflow_datasets/core/features/text/subword_text_encoder.py#L244-L250)

```
_to_file(
filename_prefix
```

Save the vocabulary to a file.

Last updated 2020-09-11 UTC.