tfds.features.text.SubwordTextEncoder

View source (https://github.com/tensorflow/datasets/blob/v1.3.0/tensorflow_datasets/core/features/t
on GitHub L385)

Class `SubwordTextEncoder`

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)

Used in the tutorials:

- Transformer model for language understanding (https://www.tensorflow.org/tutorials/text/transformer)

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:

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

ıd
ler = tfds.features.text.SubwordTextEncoder.load_from_file(vocab_filename)
```

```
: encoder.encode("hello world")
= encoder.decode([1, 2, 3, 4])
```

__init__

View source
(https://github.com/tensorflow/datasets/blob/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L65-L78)

```
.t__(vocab_list=None)
```

Constructs a SubwordTextEncoder from a vocabulary list.

To generate a vocabulary from a corpus, use
features.text.SubwordTextEncoder.build_from_corpus
://www.tensorflow.org/datasets/api_docs/python/tfds/features/text/SubwordTextEncoder#build_from_c

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.

Properties

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/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L260-L336)

```
smethod
_from_corpus(
ls,
orpus_generator,
arget_vocab_size,
ax_subword_length=20,
ax_corpus_chars=None,
eserved_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/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L90-L126)

```
le(ids)
```

Decodes a list of integers into text.

encode

View source (https://github.com/tensorflow/datasets/blob/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L80-L88)

```
le(s)
```

Encodes text into a list of integers.

load_from_file

View source (https://github.com/tensorflow/datasets/blob/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L251-L258)

```
smethod
from_file(
ls,
ilename_prefix
```

Extracts list of subwords from file.

save_to_file

View source (https://github.com/tensorflow/datasets/blob/v1.3.0/tensorflow_datasets/core/features/text/subword_text_encoder.py#L243-L249)

```
.to_file(filename_prefix)
```

Save the vocabulary to a file.