Package 'wordpiece'

February 11, 2021

Type Package

Title R Implementation of Wordpiece Tokenization

Version 1.0.2

Description Apply 'Wordpiece' (<arXiv:1609.08144>) tokenization to input text, given an appropriate vocabulary. The 'BERT' (<arXiv:1810.04805>) tokenization conventions are used by default.

Encoding UTF-8

LazyData true

URL https://github.com/jonathanbratt/wordpiece

BugReports https://github.com/jonathanbratt/wordpiece/issues

Depends R (>= 3.3.0)

License Apache License (>= 2)

RoxygenNote 7.1.1

Imports digest (>= 0.6.5), purrr (>= 0.2.3), rappdirs (>= 0.3), stringi (>= 1.0)

Suggests testthat (>= 2.1.0), knitr, rmarkdown, covr

VignetteBuilder knitr

NeedsCompilation no

Author Jonathan Bratt [aut, cre] (<https://orcid.org/0000-0003-2859-0076>), Jon Harmon [aut] (<https://orcid.org/0000-0003-4781-4346>), Bedford Freeman & Worth Pub Grp LLC DBA Macmillan Learning [cph]

Maintainer Jonathan Bratt <jonathan.bratt@macmillan.com>

Repository CRAN

Date/Publication 2021-02-11 15:40:06 UTC

R topics documented:

`get_cache_dir` *Retrieve Directory for vocabulary Cache*

Description

Retrieve Directory for vocabulary Cache

Usage

```
get_cache_dir()
```

Value

A unique filename to use for cacheing the vocabulary.

`load_or_retrieve_vocab` *Load a vocabulary file, or retrieve from cache*

Description

Load a vocabulary file, or retrieve from cache

Usage

```
load_or_retrieve_vocab(
  vocab_file,
  use_cache = TRUE,
  cache_dir = get_cache_dir()
)
```

Arguments

`vocab_file`	path to vocabulary file. File is assumed to be a text file, with one token per line, with the line number corresponding to the index of that token in the vocabulary.
`use_cache`	Logical; if TRUE, will attempt to retrieve the vocabulary from the specified cache location, or, if not found there, will ask to save the vocabulary as an .rds file.
`cache_dir`	Character; the path to a cache directory (defaults to location returned by `get_cache_dir()`).

Value

The vocab as a named integer vector. Names are tokens in vocabulary, values are integer indices. The casedness of the vocabulary is inferred and attached as the "is_cased" attribute.

Note that from the perspective of a neural net, the numeric indices *are* the tokens, and the mapping from token to index is fixed. If we changed the indexing, it would break any pre-trained models. This is why the vocabulary is stored as a named integer vector, and why it starts with index zero.

Examples

```
# Get path to sample vocabulary included with package.
vocab_path <- system.file("extdata", "tiny_vocab.txt", package = "wordpiece")
vocab <- load_or_retrieve_vocab(vocab_file = vocab_path, use_cache = FALSE)
```

`load_vocab` *Load a vocabulary file*

Description

Load a vocabulary file

Usage

```
load_vocab(vocab_file)
```

Arguments

`vocab_file`	path to vocabulary file. File is assumed to be a text file, with one token per line, with the line number corresponding to the index of that token in the vocabulary.

Value

The vocab as a named integer vector. Names are tokens in vocabulary, values are integer indices. The casedness of the vocabulary is inferred and attached as the "is_cased" attribute.

Note that from the perspective of a neural net, the numeric indices *are* the tokens, and the mapping from token to index is fixed. If we changed the indexing, it would break any pre-trained models. This is why the vocabulary is stored as a named integer vector, and why it starts with index zero.

Examples

```
# Get path to sample vocabulary included with package.
vocab_path <- system.file("extdata", "tiny_vocab.txt", package = "wordpiece")
vocab <- load_vocab(vocab_file = vocab_path)
```

`wordpiece_tokenize` *Tokenize Sequence with Word Pieces*

Description

Given a single sequence of text and a wordpiece vocabulary, tokenizes the text.

Usage

```
wordpiece_tokenize(text, vocab, unk_token = "[UNK]", max_chars = 100)
```

Arguments

`text`	Character scalar; text to tokenize.
`vocab`	Named integer vector containing vocabulary words
`unk_token`	Token to represent unknown words.
`max_chars`	Maximum length of word recognized.

Value

A named integer vector, giving the tokenization of the input sequence. The integers values are the token ids, and the names are the tokens.

Examples

```
# Get path to sample vocabulary included with package.
vocab_path <- system.file("extdata", "tiny_vocab.txt", package = "wordpiece")
vocab <- load_or_retrieve_vocab(vocab_file = vocab_path, use_cache = FALSE)
tokens <- wordpiece_tokenize(
  text = "I love tacos!",
  vocab = vocab
)
```

Index