Package: wordpiece.data (via r-universe)

October 29, 2024

Title Data for Wordpiece-Style Tokenization

Version 2.0.0

Description Provides data to be used by the wordpiece algorithm in
order to tokenize text into somewhat meaningful chunks.
Included vocabularies were retrieved from
<https://huggingface.co/bert-base-cased/resolve/main/vocab.txt>
and
<https://huggingface.co/bert-base-uncased/resolve/main/vocab.txt>
and parsed into an R-friendly format.

License Apache License (>= 2)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

URL https://github.com/macmillancontentscience/wordpiece.data

BugReports https://github.com/macmillancontentscience/wordpiece.data/issues

Depends R (>= 3.5.0)

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

Repository https://jonthegeek.r-universe.dev

RemoteUrl https://github.com/macmillancontentscience/wordpiece.data

RemoteRef HEAD

RemoteSha f893df5061be8f53fd586b142274b7ed669112c9

Contents

wordpiece_vocab *Load a wordpiece Vocabulary*

Description

A wordpiece vocabulary is a named integer vector with class "wordpiece_vocabulary". The names of the vector are the tokens, and the values are the integer identifiers of those tokens. The vocabulary is 0-indexed for compatibility with Python implementations.

Usage

```
wordpiece_vocab(cased = FALSE)
```

Arguments

cased	Logical; load the uncased vocabulary, or the cased vocabulary?

Value

A wordpiece_vocabulary.

Examples

```
head(wordpiece_vocab())
head(wordpiece_vocab(cased = TRUE))
```

Index