tf.keras.datasets.imdb.load_data

```
load_data(
    path='imdb.npz',
    num_words=None,
    skip_top=0,
    maxlen=None,
    seed=113,
    start_char=1,
    oov_char=2,
    index_from=3
)
```

Defined in `tensorflow/python/keras/_impl/keras/datasets/imdb.py` .

Loads the IMDB dataset.

Arguments:

- `path` : where to cache the data (relative to `~/.keras/dataset`).
- `num_words` : max number of words to include. Words are ranked by how often they occur (in the training set) and only the most frequent words are kept
- `skip_top` : skip the top N most frequently occurring words (which may not be informative).
- `maxlen` : truncate sequences after this length.
- `seed` : random seed for sample shuffling.
- `start_char` : The start of a sequence will be marked with this character. Set to 1 because 0 is usually the padding character.
- `oov_char` : words that were cut out because of the `num_words` or `skip_top` limit will be replaced with this character.
- `index_from` : index actual words with this index and higher.

Returns:

Tuple of Numpy arrays: `(x_train, y_train), (x_test, y_test)` .

Raises:

- `ValueError` : in case `maxlen` is so low that no input sequence could be kept.

Note that the 'out of vocabulary' character is only used for words that were present in the training set but are not included because they're not making the `num_words` cut here. Words that were not seen in the training set but are in the test set have simply been skipped.

Last updated November 2, 2017.

Stay Connected

Blog

GitHub

Twitter

Support

Issue Tracker

Release Notes

Stack Overflow

English

Terms | **Privacy**