tf.keras.preprocessing.sequence.make_sampling_table

```
make_sampling_table(
    size,
    sampling_factor=1e-05
)
```

Defined in `tensorflow/python/keras/_impl/keras/preprocessing/sequence.py`.

Generates a word rank-based probabilistic sampling table.

This generates an array where the ith element is the probability that a word of rank i would be sampled, according to the sampling distribution used in word2vec.

The word2vec formula is: p(word) = min(1, sqrt(word.frequency/sampling_factor) / (word.frequency/sampling_factor))

We assume that the word frequencies follow Zipf's law (s=1) to derive a numerical approximation of frequency(rank): frequency(rank) ~ 1/(rank * (log(rank) + gamma) + 1/2 - 1/(12*rank)) where gamma is the Euler-Mascheroni constant.

Arguments:

- `size`: int, number of possible words to sample.
- `sampling_factor`: the sampling factor in the word2vec formula.

Returns:

A 1D Numpy array of length `size` where the ith entry is the probability that a word of rank i should be sampled.

Last updated November 2, 2017.

English

Terms | Privacy