

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 *i*th 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: $p(\text{word}) = \min(1, \sqrt{\text{word.frequency}/\text{sampling_factor}} / (\text{word.frequency}/\text{sampling_factor}))$

We assume that the word frequencies follow Zipf's law ($s=1$) to derive a numerical approximation of frequency(rank): $\text{frequency}(\text{rank}) \sim 1/(\text{rank} * (\log(\text{rank}) + \gamma) + 1/2 - 1/(12*\text{rank}))$ where γ 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 *i*th entry is the probability that a word of rank *i* should be sampled.

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 3.0 License](#), and code samples are licensed under the [Apache 2.0 License](#). For details, see our [Site Policies](#). Java is a registered trademark of Oracle and/or its affiliates.

Last updated November 2, 2017.

Stay Connected

[Blog](#)

[GitHub](#)

[Twitter](#)

Support

[Issue Tracker](#)

[Release Notes](#)

[Stack Overflow](#)

