Super Easy Way to Get Sentence Embedding using fastText in Python

sobamchan Follow
Dec 2, 2019 · 2 min read ★

Photo by Patrick Tomasso on Unsplash

Super easy way to get word embeddings by tofunlp/sister.

When you are working with applications that contain some NLP techniques, it is very common that you want word embeddings of your text data. So that you can do a variety of things, such as calculate the distance of sentences, classification or cool visualization.

But it often takes some time to download pre-trained word embeddings (e.g., word2vec, fastText) and load them somehow. And these procedures are not very fun (at least to me) and not easy to manage and keep these codes clean.

So in this article, I would like to introduce a little python library, named SISTER (SImple SenTence EmbeddeR). (Yes, I like to name my projects strange names). This pip-installable library allows you to do two things, 1) download pre-trained word embedding, 2) provide a simple interface to use it to embed your text. As an extra feature, since I wrote this library to be easy to extend so supporting new languages or algorithms to embed text should be simple and easy.

. . .

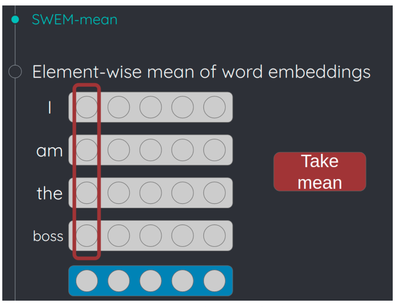
Now, let me show you how easy it is to use. It can be done in just 4 lines.

See? Very easy.

You do not need to do any downloading or building, when you use *MeanEmbedding* for the first time, it just downloads pre-trained facebookresearch/fastText vector automatically.

Currently, this library only supports English and Japanese. However, since the implementation is extensible, just adding URL to the pre-trained vector here, you can easily add your favorite language. Or you can create an issue for the request so I may add it.

And the way the *MeanEmbedding* creates sentence vector is illustrated below, this is a very simple method but very effective too (ref: one of my favorite paper).



How it gets sentence vector from sequence of words.

As you can see in the figure above, it first converts all the given words into word embeddings, then takes their mean in element-wise. So the sentence vector will have the same size as each word embeddings (300-dim in the previous example code).

In this article, we went through how to use word embeddings to obtain sentence embeddings. I hope you enjoyed it, embed a bunch of sentences by yourself!

Machine Learning NLP Python AI Data Science