glove

# library link

<https://github.com/stanfordnlp/GloVe>

# basic description

Globe Vectors for word presentation is a word embedding methodology developed by Stanford in 2014. Globe was created to improve the problem of count-based LSA and predictive-based Word2Vec.

LSA has the advantage of utilizing statistical information throughout the document, but it has the disadvantage that it is difficult to measure the similarity between words. On the other hand, Word2Vec has the advantage of being able to measure the similarity between words, but has the disadvantage that it is difficult to relfect word information throughout the document because only a few surrounding words (contextual words) are used within the Window designated by the users.

Globe is a word embedding methodology that combines the advantages of LSA and Word2Vec that can measure similarities between words and utilize statistical information throughout the document.

# version

* Nltk >=3.4.5 (pip install nltk)
* Glove >= 0.2.0 (pip install glove\_python\_binary)
* colab

# dataset

* Dataset is written in xml grammar, so preprocessing is required to obtain natural language. The actual data you want to obtain is between <content> and </content>, which contain only English sentences.

<https://raw.githubusercontent.com/ukairia777/tensorflow-nlp-tutorial/main/09.%20Word%20Embedding/dataset/ted_en-20160408.xml>

# code description

* Code to train the Glove model using the training dataset and then return similar words.

# validation

* x