Glove

# example code

|  |
| --- |
| import re  import urllib.request  import zipfile  from lxml import etree  from nltk.tokenize import word\_tokenize, sent\_tokenize  import nltk  nlkt.download(‘punkt’)  urllib.request.urlretrieve("https://raw.githubusercontent.com/ukairia777/tensorflow-nlp-tutorial/main/09.%20Word%20Embedding/dataset/ted\_en-20160408.xml", filename="ted\_en-20160408.xml")  targetXML = open('ted\_en-20160408.xml', 'r', encoding='UTF8')  target\_text = etree.parse(targetXML)  # xml 파일로부터 <content>와 </content> 사이의 내용만 가져온다.  parse\_text = '\n'.join(target\_text.xpath('//content/text()'))  # 정규 표현식의 sub 모듈을 통해 content 중간에 등장하는 (Audio), (Laughter) 등의 배경음 부분을 제거.  # 해당 코드는 괄호로 구성된 내용을 제거.  content\_text = re.sub(r'\([^)]\*\)', '', parse\_text)  # 입력 코퍼스에 대해서 NLTK를 이용하여 문장 토큰화를 수행.  sent\_text = sent\_tokenize(content\_text)  # 각 문장에 대해서 구두점을 제거하고, 대문자를 소문자로 변환.  normalized\_text = []  for string in sent\_text:       tokens = re.sub(r"[^a-z0-9]+", " ", string.lower())       normalized\_text.append(tokens)  # 각 문장에 대해서 NLTK를 이용하여 단어 토큰화를 수행.  result = [word\_tokenize(sentence) for sentence in normalized\_text]  from glove import Corpus, Glove  corpus = Corpus()  # 훈련 데이터로부터 GloVe에서 사용할 동시 등장 행렬 생성  corpus.fit(result, window=5)  glove = Glove(no\_components=100, learning\_rate=0.05)  # 학습에 이용할 쓰레드의 개수는 4로 설정, 에포크는 20.  glove.fit(corpus.matrix, epochs=20, no\_threads=4, verbose=True)  glove.add\_dictionary(corpus.dictionary)  print(glove.most\_similar("man"))  print(glove.most\_similar("boy"))  print(glove.most\_similar("university"))  print(glove.most\_similar("water"))  print(glove.most\_similar("physics"))  print(glove.most\_similar("muscle"))  print(glove.most\_similar("clean")) |

# testing result

|  |
| --- |
| [('woman', 0.9572215493215784), ('guy', 0.8789844818739406),  ('girl', 0.853532334763873), ('young', 0.8280370404227984)]  [('girl', 0.9345645231775356), ('kid', 0.8289540816541461),  ('woman', 0.8218024833004672), ('man', 0.8135440459915495)]  [('harvard', 0.8812833619052991), ('mit', 0.8527717147812021),  ('stanford', 0.8358544889267849), ('cambridge', 0.8338871501757076)] [('clean', 0.8416763648953807), ('air', 0.8303662013859833), ('fresh', 0.8295296399506084), ('food', 0.8101058146369846)]  [('chemistry', 0.8873987607832885), ('economics', 0.8843470585052818), ('beauty', 0.8740753299346331), ('mathematics', 0.871922781053764)] [('tissue', 0.8518306102851879), ('nerve', 0.8251863618967884), ('bone', 0.7643866682812236), ('channel', 0.7622640357975308)]  [('water', 0.8416763648953807), ('fresh', 0.8337178701696939), ('heat', 0.8154577327246422), ('wind', 0.7949307244160057)] |