glove

# 라이브러리 링크

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

# 기초 설명

글로브(GloVe, Global Vectors for Word Representation)는 스탠퍼드에서 2014년 개발한 워드 임베딩 방법론입니다. 글로브는 카운트 기반의 LSA와 예측 기반의 Word2Vec의 문제점을 개선하기 위해 탄생했습니다.

LSA는 문서 전체의 통계적인 정보를 활용한다는 장점이 있지만, 단어간 유사도를 측정하기 어렵다는 단점이 있습니다. 반면 Word2Vec은 단어 간 유사도를 측정할 수 있다는 장점이 있지만, 사용자가 지정한 위도우 내에서만, 즉 주변 단어(맥락 단어) 몇 개만 활용하여 결과가 도출되기 때문에 문서 전체의 단어 정보가 반영되기 힘들다는 단점이 있습니다.

글로브는 LSA와 Word2Vec의 장점을 합친 워드 임베딩 방법론으로 단어 간 유사도도 측정할 수 있고 문서 전체의 통계적인 정보도 활용할 수 있습니다.

# 버전 정보

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

# 데이터셋 설명 및 출처

* xml 문법으로 작성되어 있어 자연어를 얻기 위해서는 전처리가 필요함. 얻고자 하는 실질적 데이터는 영어 문장으로만 구성된 내용을 담고 있는 <content>와 </content> 사이의 내용입니다.

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

# 코드 설명

* 훈련 데이터 셋을 사용하여 GlovVe 모델을 훈련시킨 후 유사한 단어를 리턴하는 코드.

# 검증 방법

* x