N-gram

N개의 단어를 가지고 바로 다음에 위치할 단어를 예측하는 문제

Target 단어 다음의 단어들은 참조하지 않음.

I am a college (student.)

→'I', 'am', 'a' 라는 **세 개의 단어**를 보고 네 번째에 위치한 'student' 예측

Tri-gram

CBOW(Continuous Bag-Of-Word)

Target 단어 양 옆의 N개 단어를 보고 예측하는 문제

즉, 주변 단어(Context)를 가지고 중심 단어(Center) 예측

Word2Vec(Skip-gram)

CBOW의 반대 작용.

중심 단어(Center)를 가지고 주변 단어(Context)를 예측하는 문제.

중심 단어가 update될 기회가 많지만, 그만큼 계산량이 늘어남.

사용한 데이터셋

ABC 뉴스 기사의 헤드라인을 모아놓은 csv 파일

Pandas 라이브러리 이용 읽어오고, 셔플 후 5000개만 잘라서 tokenize ->재사용 위해 pkl

Hyper parameters:

Embedding dim= 10

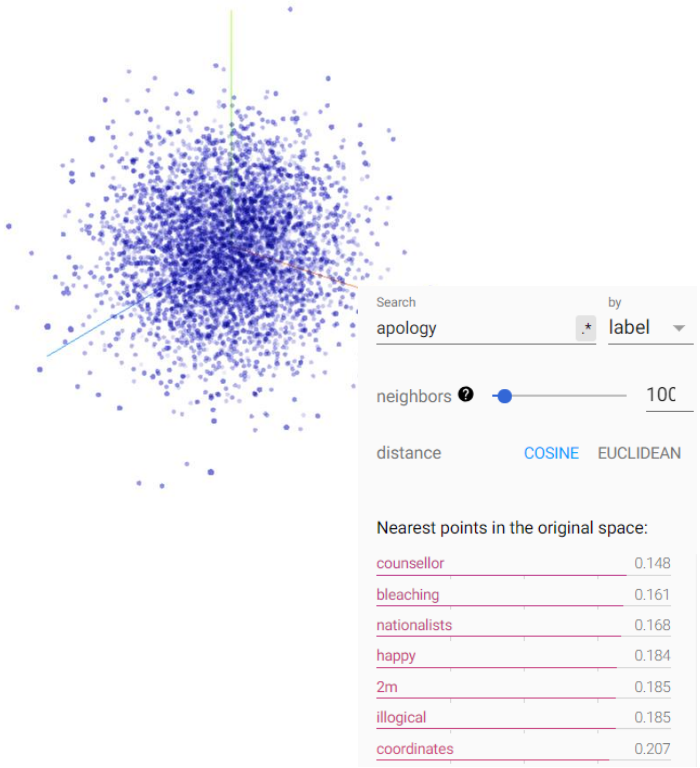
Num_epochs = 10

Learning rate = 0.05

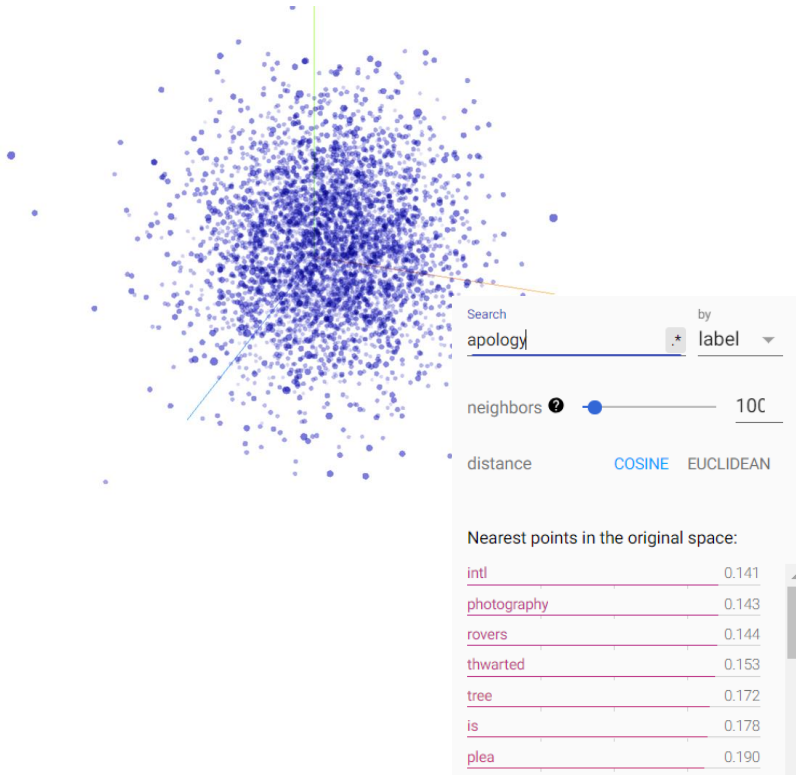
+ Skipgram에서의 subsampling 빈도수 threshold=2, negative sampling num=10

TensorboardX 이용 임베딩 결과 확인

N-grams



CBOW



Skip-gram(NS, SS 포함)