

## 9. 임베딩 벡터의 시각화

구글은 임베딩 프로젝터(embedding projector)라는 데이터 시각화 도구를 지원한다. 이것을 이용하여 학습한 임베딩 벡터들을 시각화 해보자. (코랩에서 실행함)


임베딩 프로젝터 논문 : <https://arxiv.org/pdf/1611.05469v1.pdf>


### 워드 임베딩 모델부터 2개 tsv파일 생성

우리는 기존에 실습했었던 영어 Word2Vec 모델인 eng\_w2v를 사용하도록 한다. 코랩에서 다음 커맨드를 실행시킨다.

```
!python -m gensim.scripts.word2vec2tensor --input eng_w2v --output eng_w2v
```

그러면 다음 2개의 tsv파일이 생성된다.

 eng\_w2v\_metadata.tsv

 eng\_w2v\_tensor.tsv

위 두 파일이 임베딩 벡터 시각화를 위해 사용할 파일이다. 위 두파일을 윈도우 로컬에 다운로드하도록 한다.

### 임베딩 프로젝터를 사용하여 시각화하기

먼저 아래의 링크에 접속한다.

링크 : <https://projector.tensorflow.org/>

DATA

5 tensors found

Word2Vec 10K

Label by  
word

Color by  
No color map

Edit by  
word

Tag selection as

Load

Publish

Download

Label

☒ Sphereize data

Checkpoint: Demo datasets

Metadata: oss\_data/word2vec\_10000\_200d\_labels.tsv

여기서 Load버튼을 클릭한다.

Load data from your computer

Step 1: Load a TSV file of vectors.

Example of 3 vectors with dimension 4:

0.1 0.2 0.5 0.9  
0.2 0.1 0.5 0.2  
0.4 0.1 0.7 0.8

Choose file

Step 2 (optional): Load a TSV file of metadata.

Example of 3 data points and 2 columns.

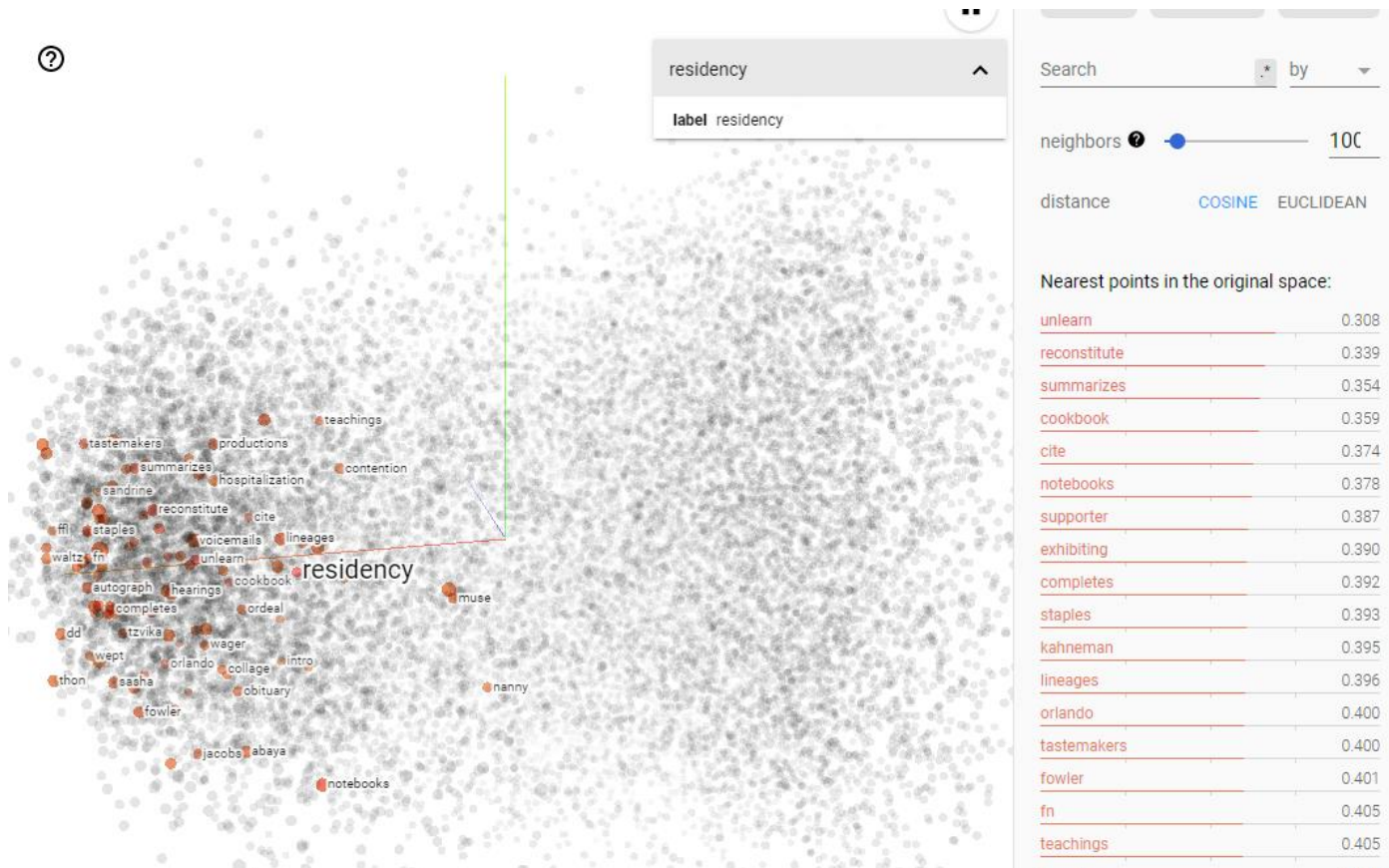
Note: If there is more than one column, the first row will be parsed as column labels.

Pokémon Species  
Wartort le Turtle  
Venusaur Seed  
Charmeleon Flame

Choose file

Click outside to dismiss.

위에 있는 Choose file 버튼을 누르고 eng\_w2v\_tensor.tsv 파일을 업로드하고, 아래에 있는 Choose file 버튼을 누르고 eng\_w2v\_metadata.tsv 파일을 업로드한다. 그 이후에는 학습했던 워드 임베딩 모델이 프로젝터에 시각화 된다.



위 사진은 residency라는 단어를 선택하여 유사도를 코사인 유사도로 기준을 잡고 가장 유사한 10개 벡터들을 표시한 것이다. 그리고 데이터 차원을 축소하여 시각화할 수 있도록 도와주는 PCA, t-SNE 등을 제공하기도 한다.