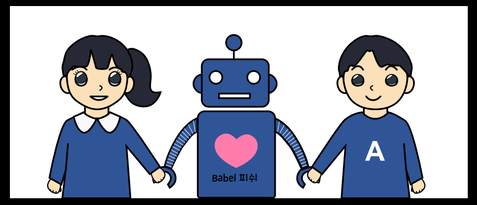
한국어 임베딩

Korean Embedding

nlp4kor

http://github.com/bage79/nlp4kor

박혜웅



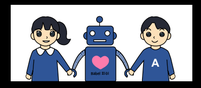
바벨피쉬

What is Word Embedding?

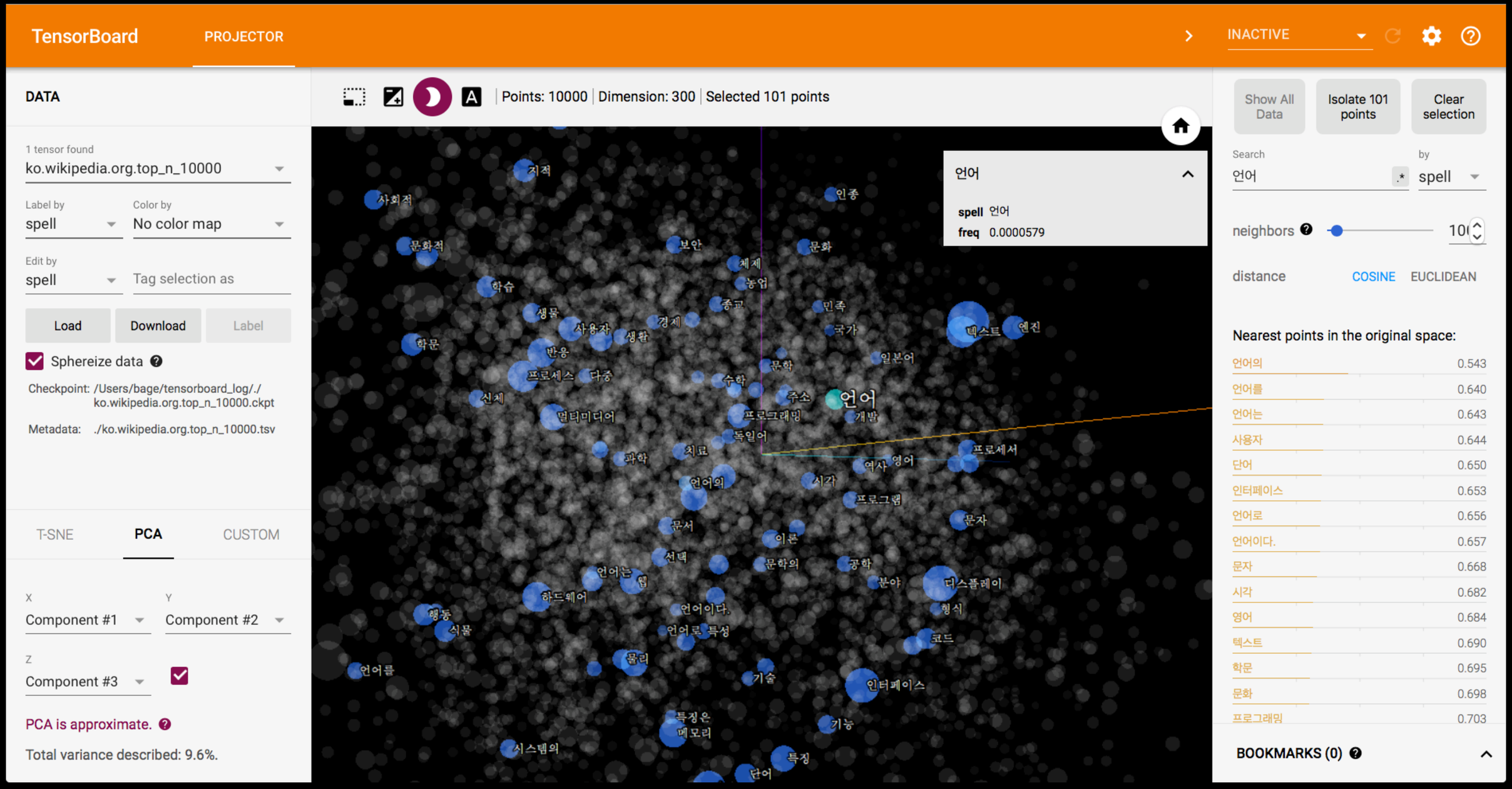
em·bed [1] | imbéd |

동사 타동사 (embed·ded, embed·ding)

1.《보통 수동태》〈물건을〉 (…에) 박아 넣다, 〈꽃 등을〉 (…에) 심다(*in* …);



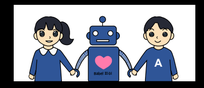
Word Embedding



Various Similarity

Semantic & Syntatic

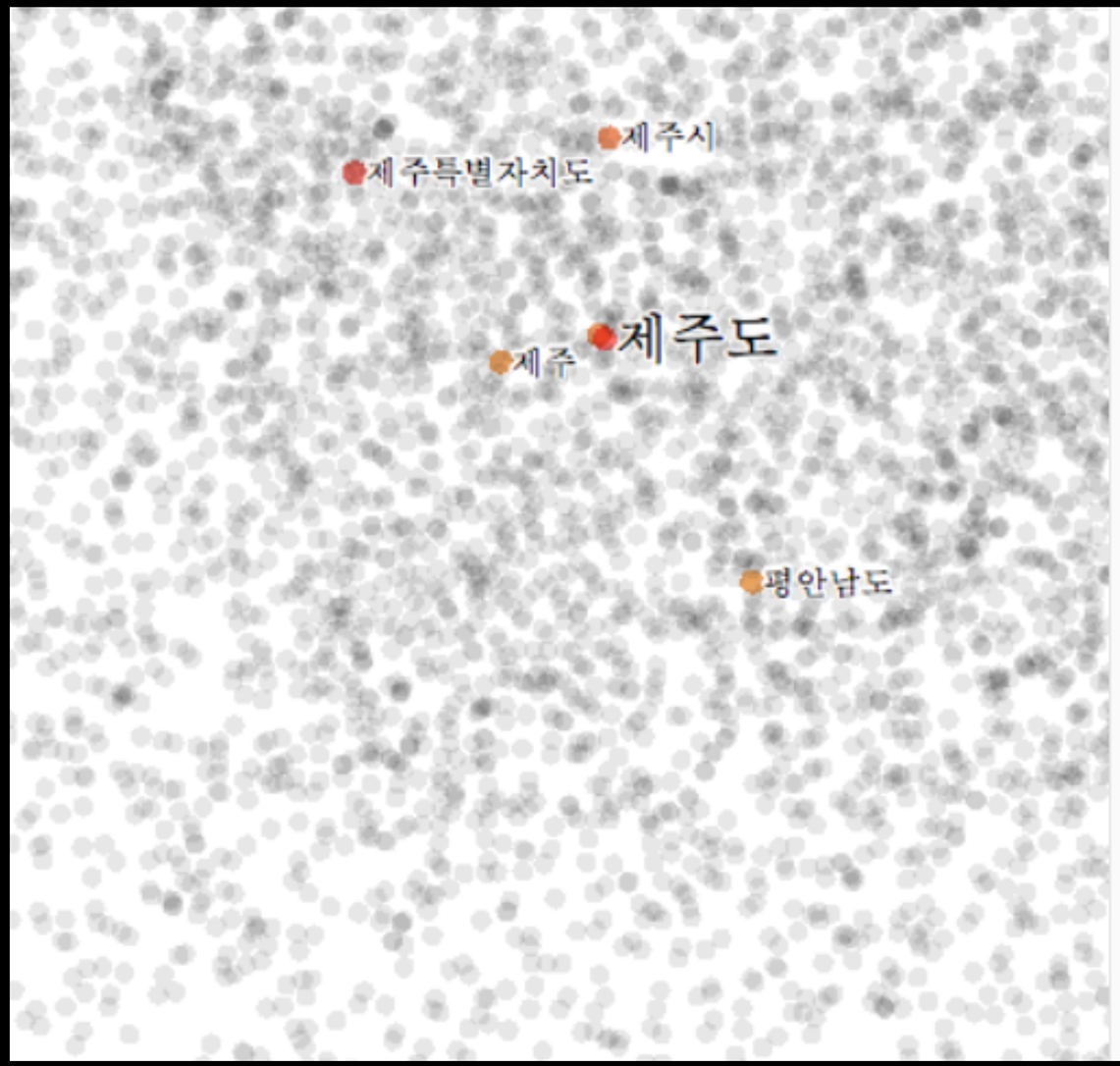
nlp4kor



What are Similar two words?

EBS 다큐프라임 <명사로 세상을 보는 서양인, 동사로 세상을 보는 동양인>

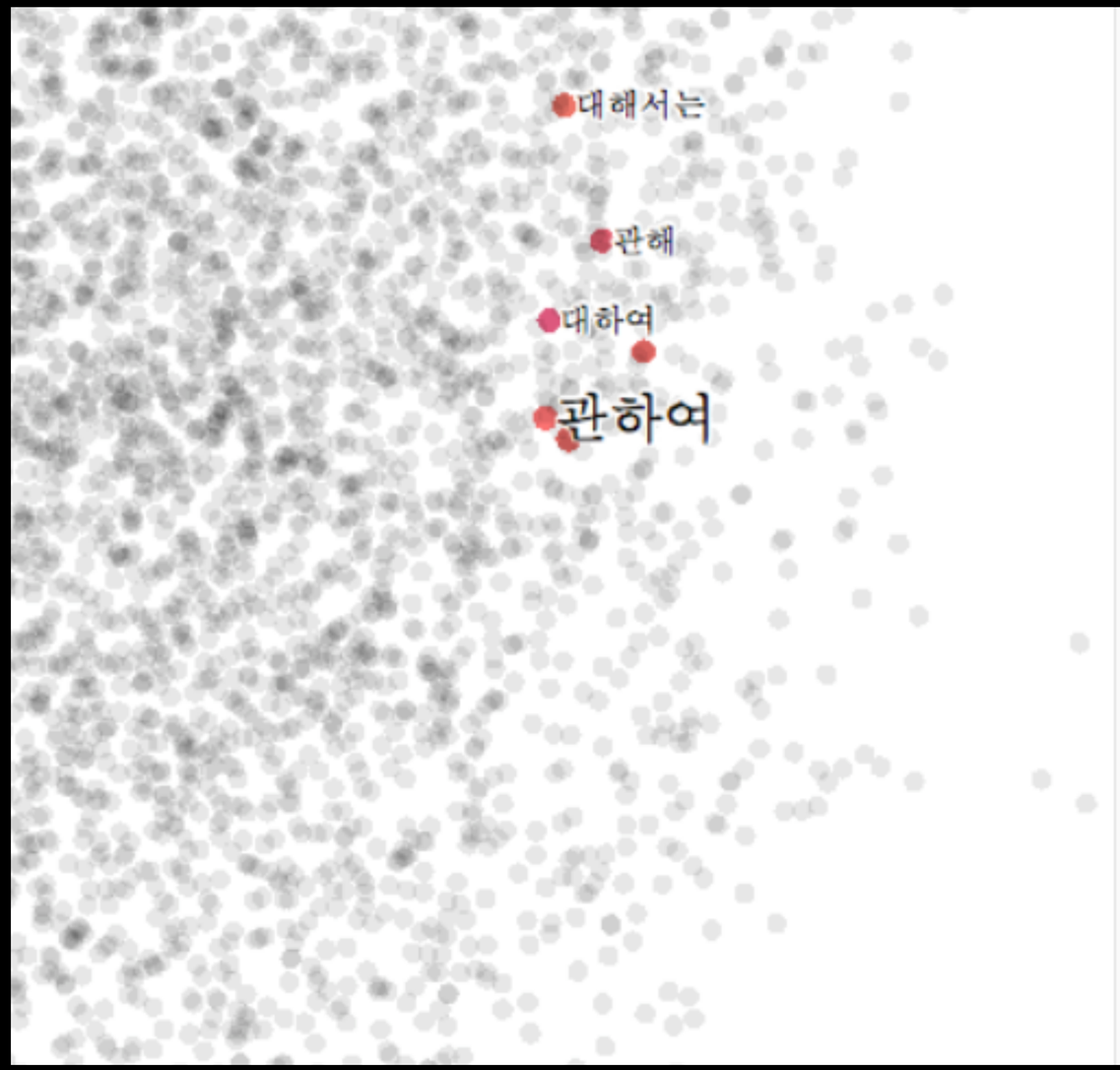
Synonym



Nearest points in the original space:

제주특별자치도	0.325
제주시	0.428
서귀포시	0.472
제주	0.502
평안남도	0.509

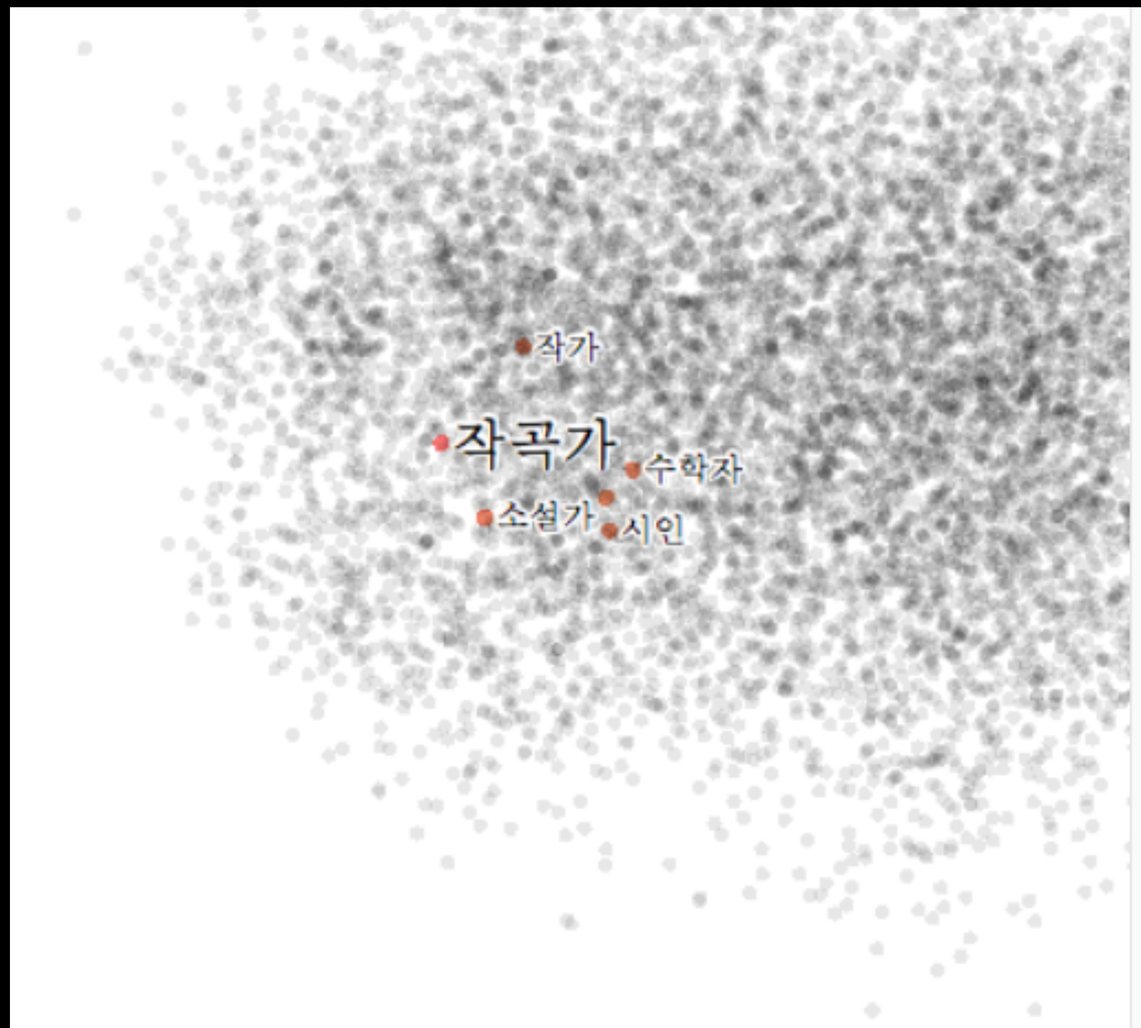
Synonym



Nearest points in the original space:

대하여	0.233
관해	0.267
대해서	0.314
대해	0.316
대해서는	0.340

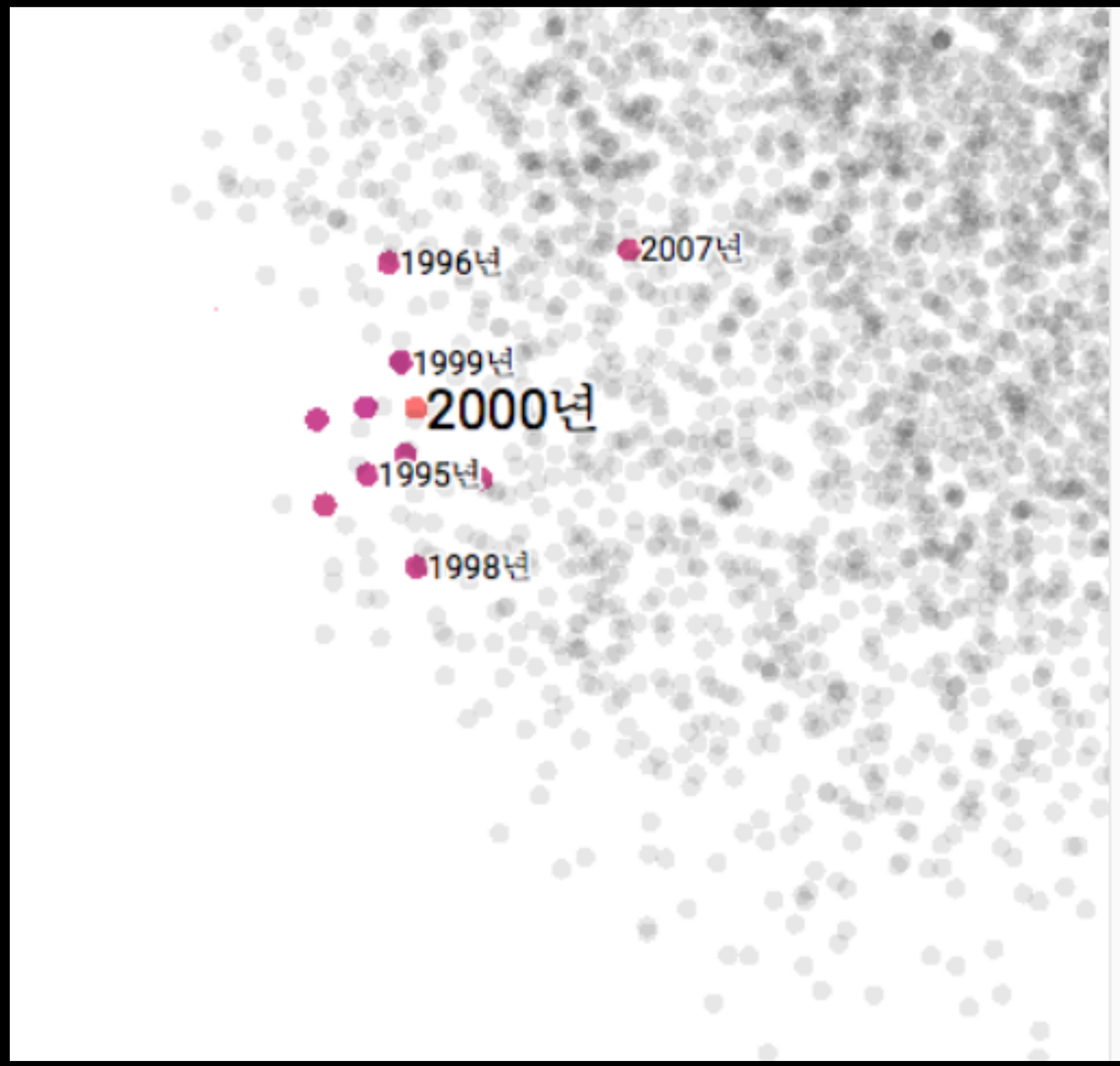
Similiar Types



Nearest points in the original space:

소설가	0.382
시인	0.403
수학자	0.404
작가	0.407
정치인	0.422

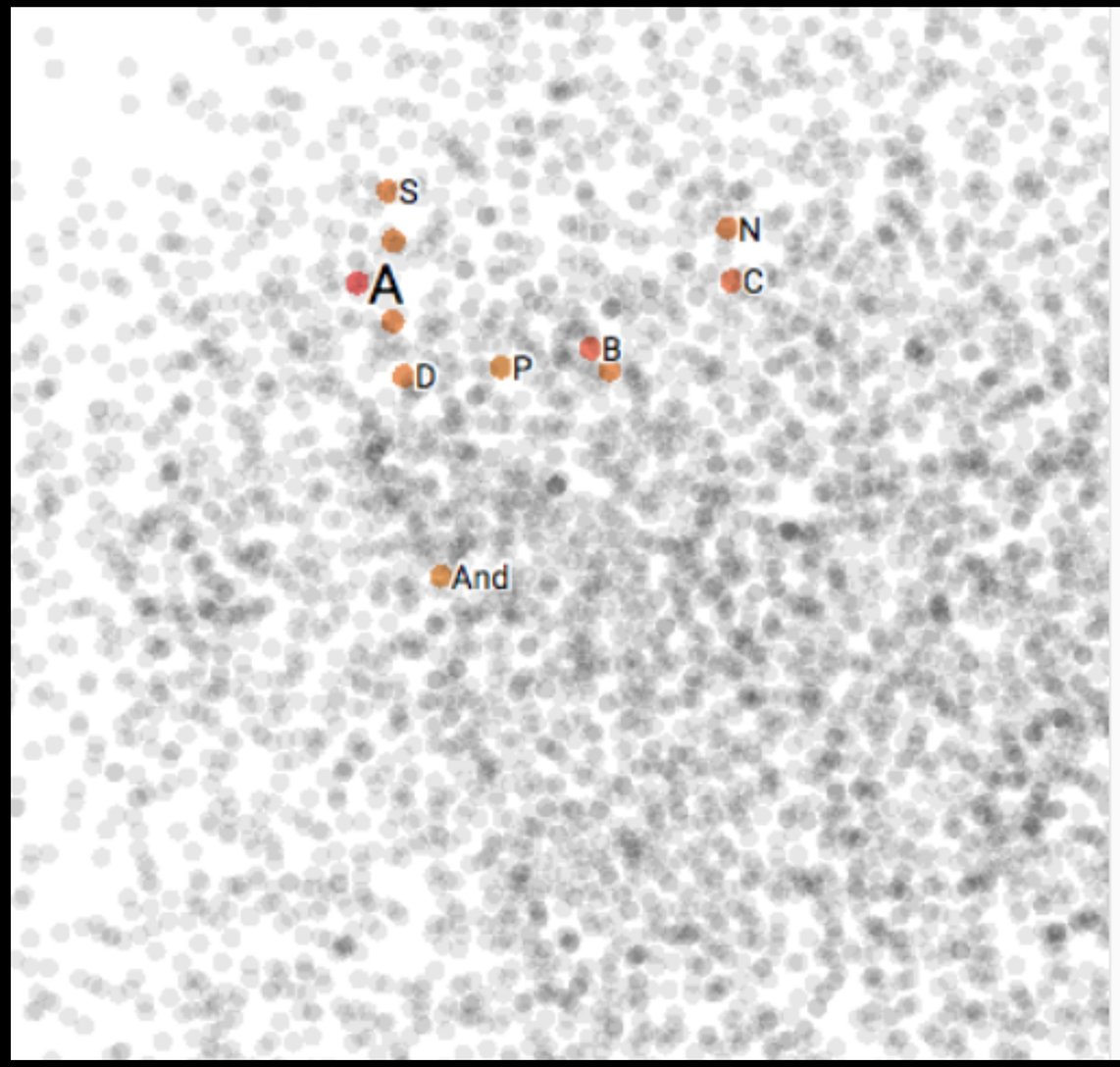
Similiar Types



Nearest points in the original space:

2001년	0.158
1999년	0.161
2003년	0.166
1995년	0.170
2004년	0.171
1998년	0.178
1996년	0.181
1997년	0.187
1993년	0.190
2007년	0.197

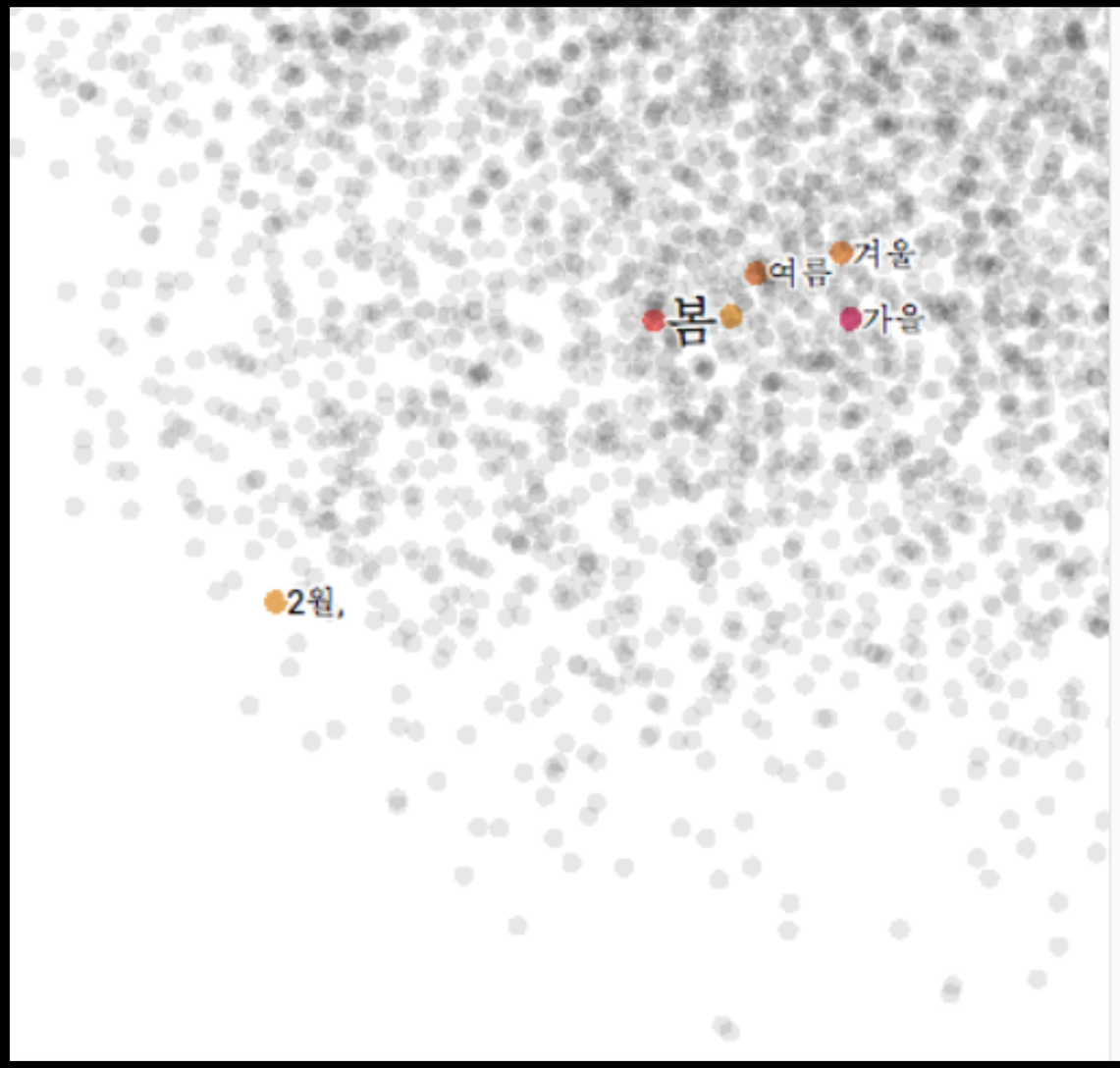
Similiar Types



Nearest points in the original space:

B	0.367
C	0.413
N	0.459
E	0.463
D	0.465
S	0.468
H	0.470
F	0.476
P	0.507
And	0.508

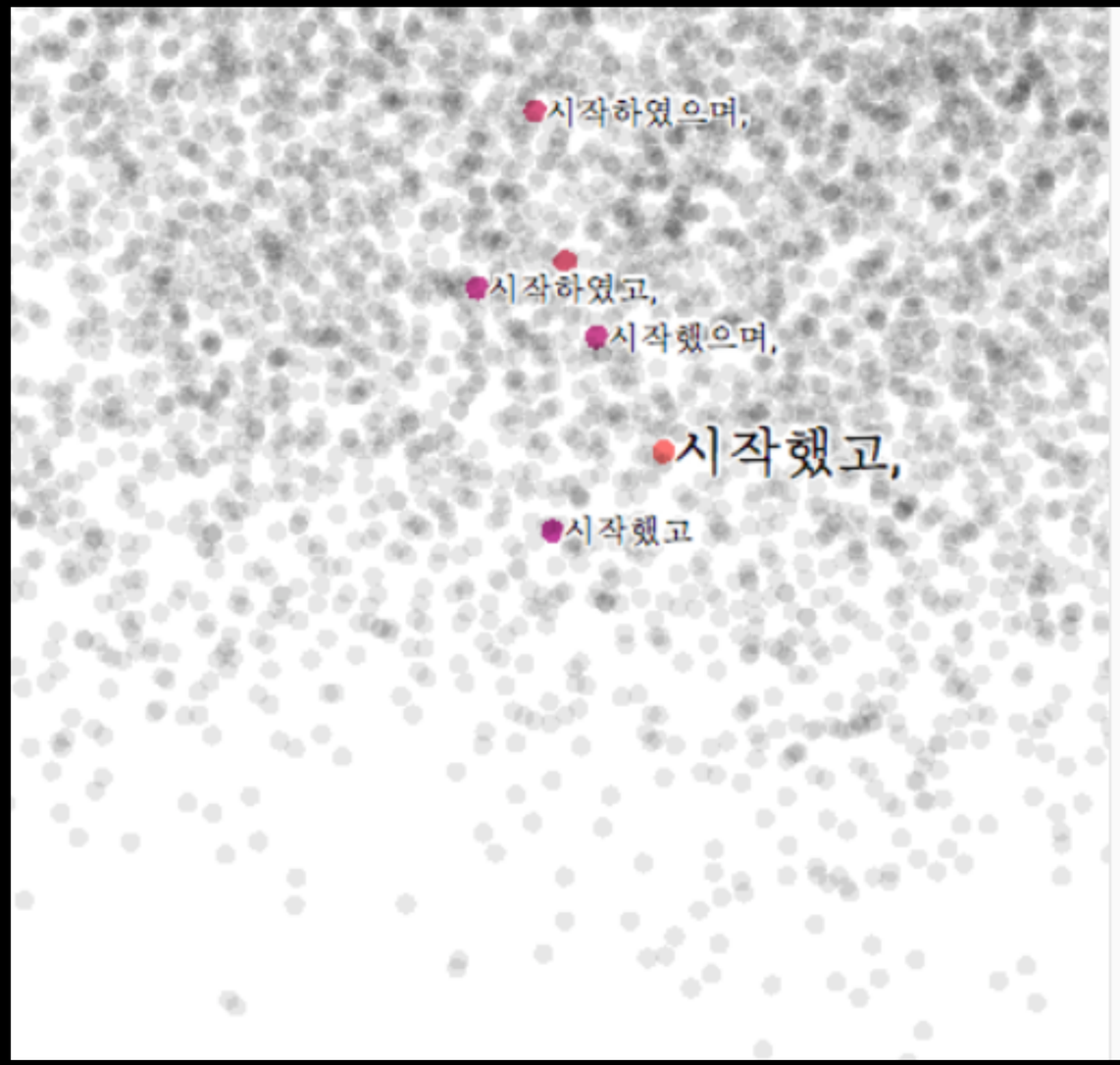
Similiar Types



Nearest points in the original space:

가을	0.222
여름	0.449
겨울	0.485
7월부터	0.533
2월,	0.539

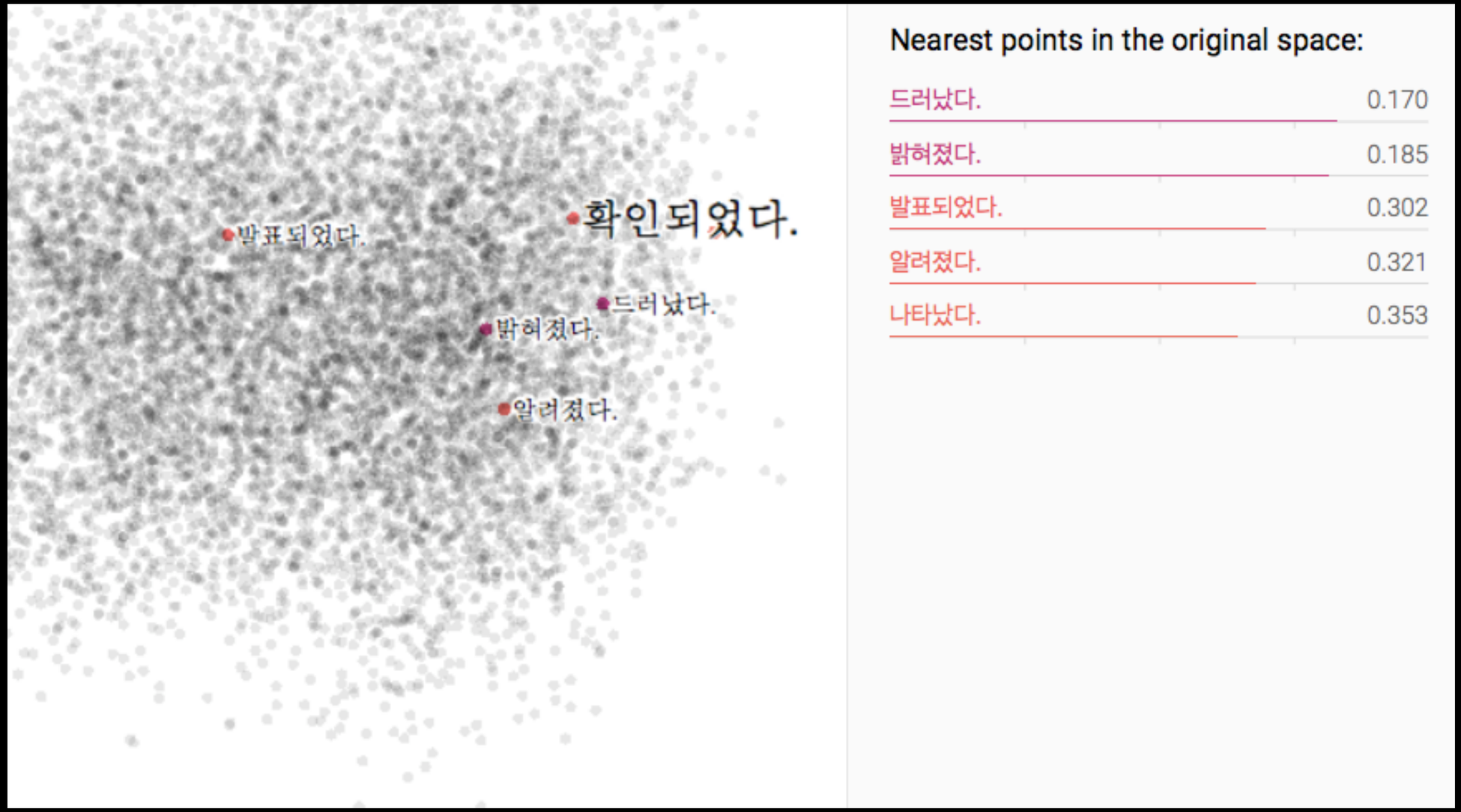
Derivateve words



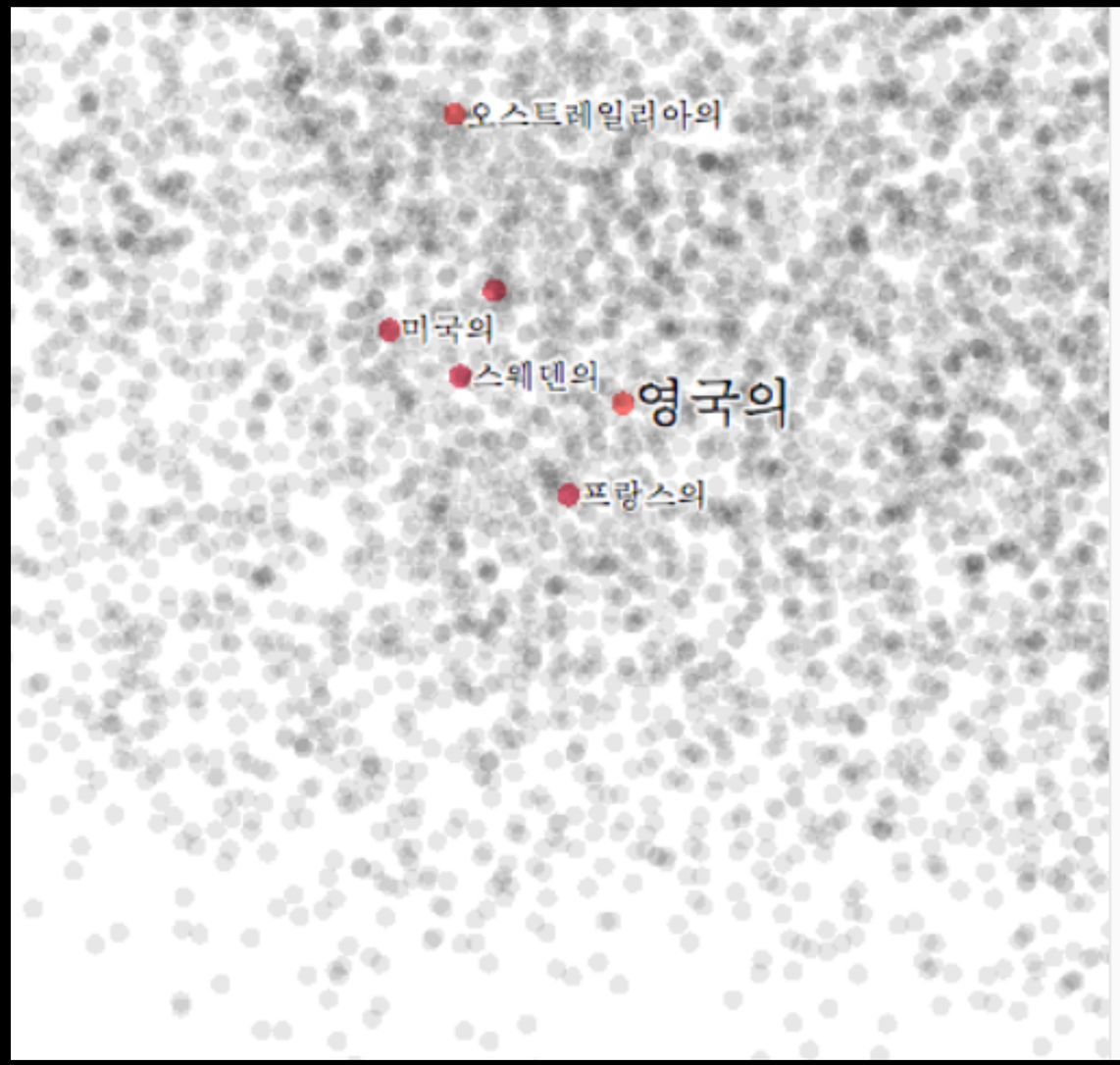
Nearest points in the original space:

시작했고	0.134
시작하였고,	0.161
시작했으며,	0.169
시작하였으며,	0.221
시작해	0.251

Synonym + Derivateve words



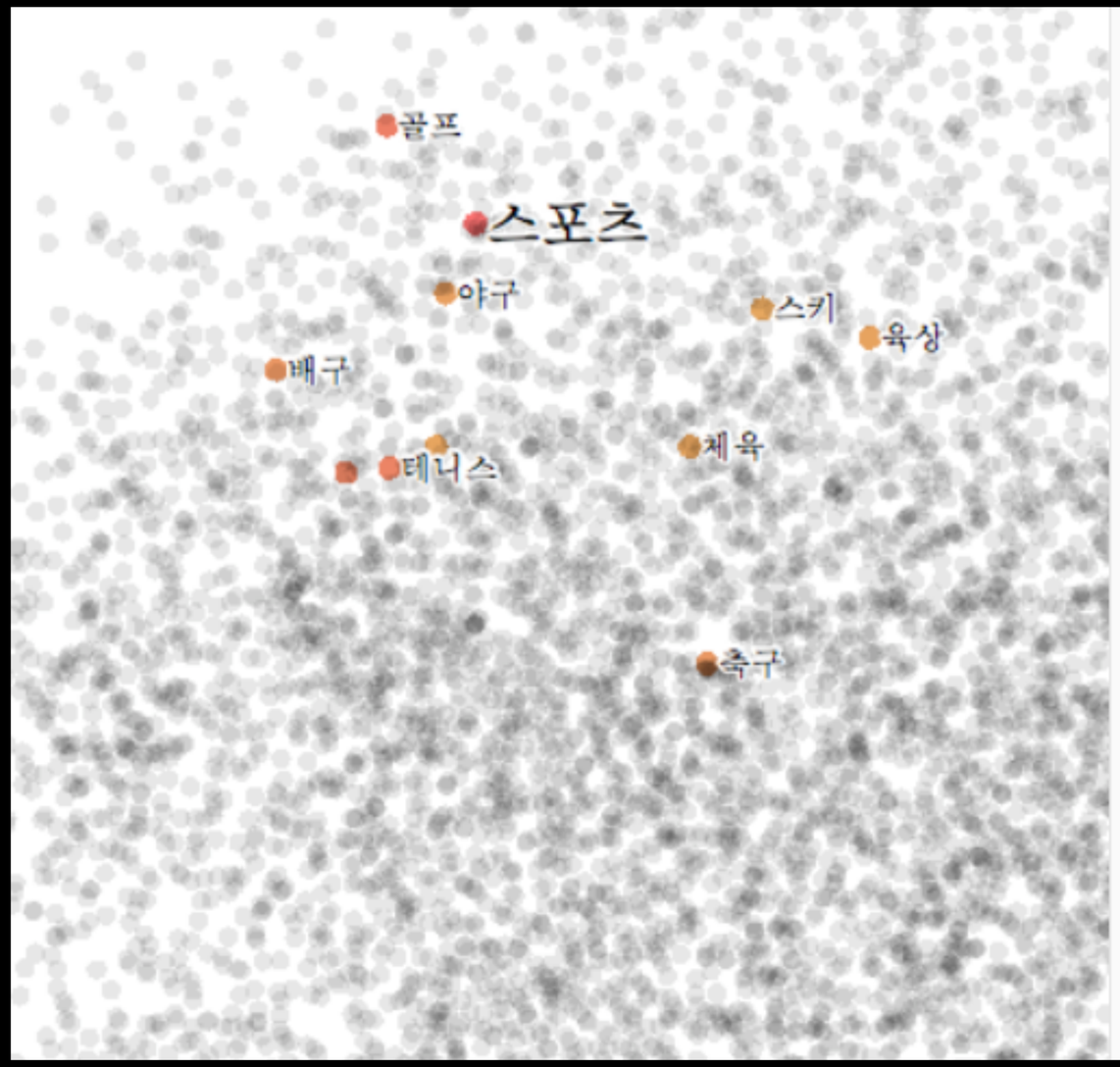
Similiar Types + Derivateve words



Nearest points in the original space:

프랑스의	0.254
스웨덴의	0.255
미국의	0.261
캐나다의	0.266
오스트레일리아의	0.295

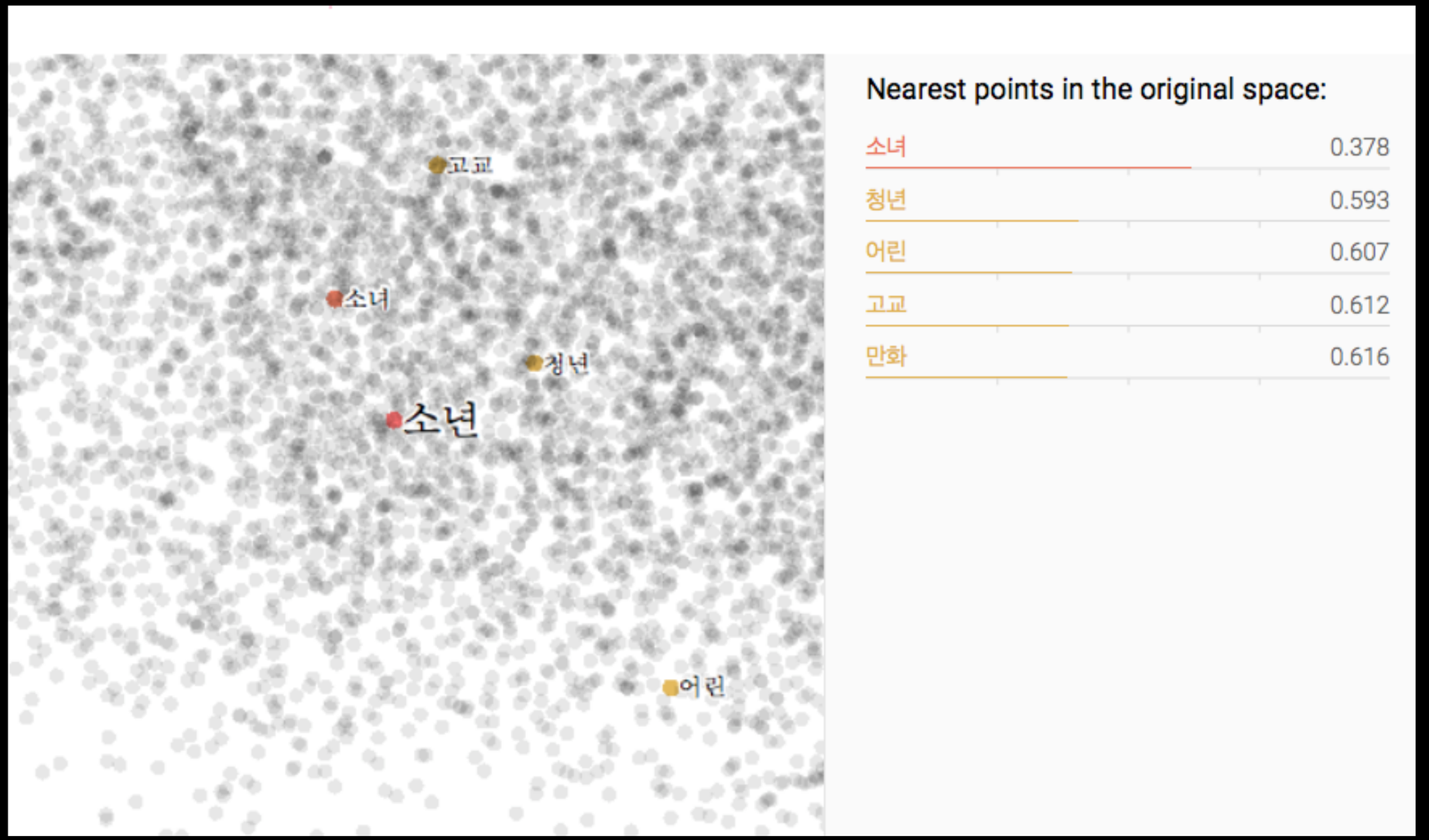
a Part of



Nearest points in the original space:

골프	0.376
테니스	0.392
농구	0.399
배구	0.461
축구	0.466
야구	0.497
육상	0.509
수영	0.520
체육	0.522
스키	0.528

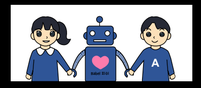
Antonym


고교
소녀
청년
소년
어린
Nearest points in the original space:
소녀 0.378
청년 0.593
어린 0.607
고교 0.612
만화 0.616

Word Representaion in NLP

Natural Language Processing

nlp4kor



Dictionary vs Embedding

남자[2] (男子) ★★★

〔명사〕

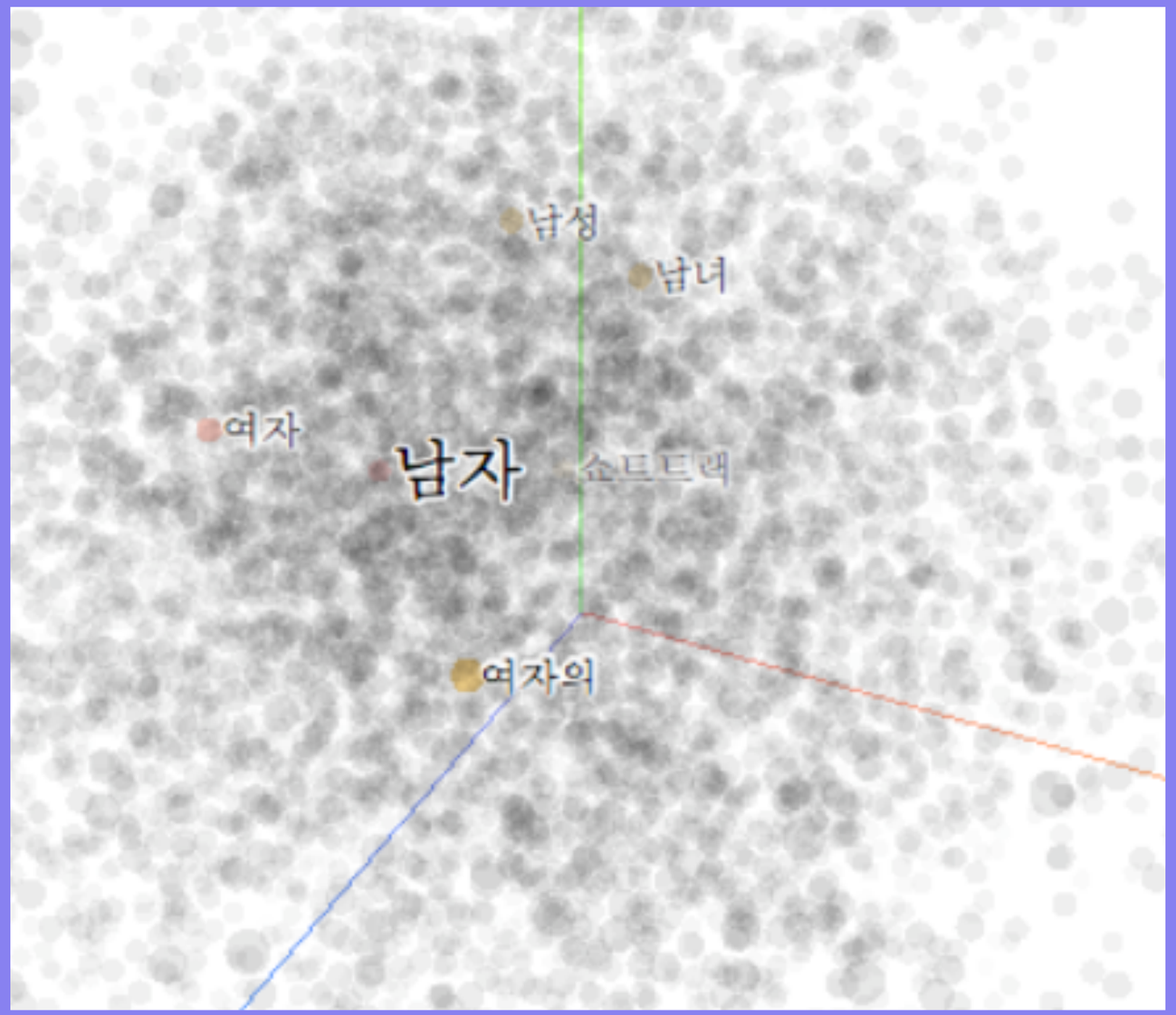
1. 남성으로 태어난 사람.
2. 사내다운 사내.
3. 한 여자의 남편이나 애인을 이르는 말.

여자[2] (女子) ★★★

〔명사〕

1. 여성으로 태어난 사람.
2. **여자**다운 **여자**.
3. 한 남자의 아내나 애인을 이르는 말.

유의어 : 부녀자, 부녀[2], 아낙



Dictionary vs Embedding

남자[2] (男子) ★★★

〔명사〕

1. 남성으로 태어난 사람.
2. 사내다운 사내.
3. 한 여자의 남편이나 애인을 이르는 말.

늑대 〔늑때〕 ★

〔명사〕

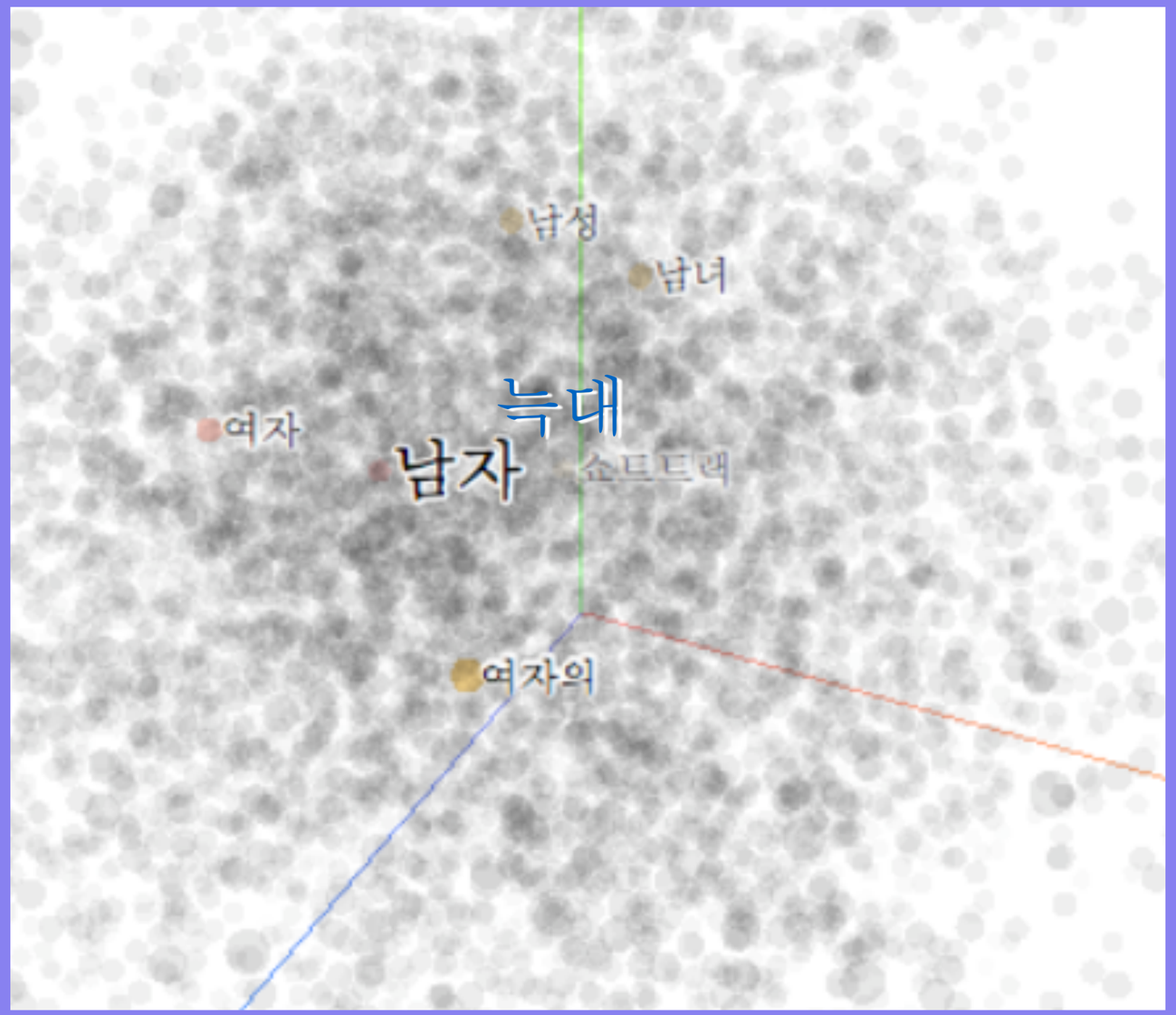
1. 〈동물〉갯과의 포유류. 몸의 길이는 120cm, 꼬리는 35cm,
2. 여자에게 음흉한 마음을 품은 남자를 비유적으로 이르는 말.

여자[2] (女子) ★★★

〔명사〕

1. 여성으로 태어난 사람.
2. **여자**다운 **여자**.
3. 한 남자의 아내나 애인을 이르는 말.

유의어 : 부녀자, 부녀[2], 아낙



Dictrionary based Representation

남자[2] (男子) ★★★

〔명사〕

1. 남성으로 태어난 사람.
2. 사내다운 사내.
3. 한 여자의 남편이나 애인을 이르는 말.

늑대 〔늑때〕 ★

〔명사〕

1. 〈동물〉갯과의 포유류. 몸의 길이는 120cm, 꼬리는 35cm,
2. 여자에게 음흉한 마음을 품은 남자를 비유적으로 이르는 말.

여자[2] (女子) ★★★

〔명사〕

1. 여성으로 태어난 사람.
2. **여자**다운 **여자**.
3. 한 남자의 아내나 애인을 이르는 말.

유의어 : 부녀자, 부녀[2], 아낙

	남성	사람	사내	여자	남편	애인	여성	남자	아내	애인	포유류	몸	음흉	마음
“남자”	O	O	O	O	O	O								
“늑대”								O			O	O	O	O
“여자”		O		O			O	O	O	O				

Character based Representation

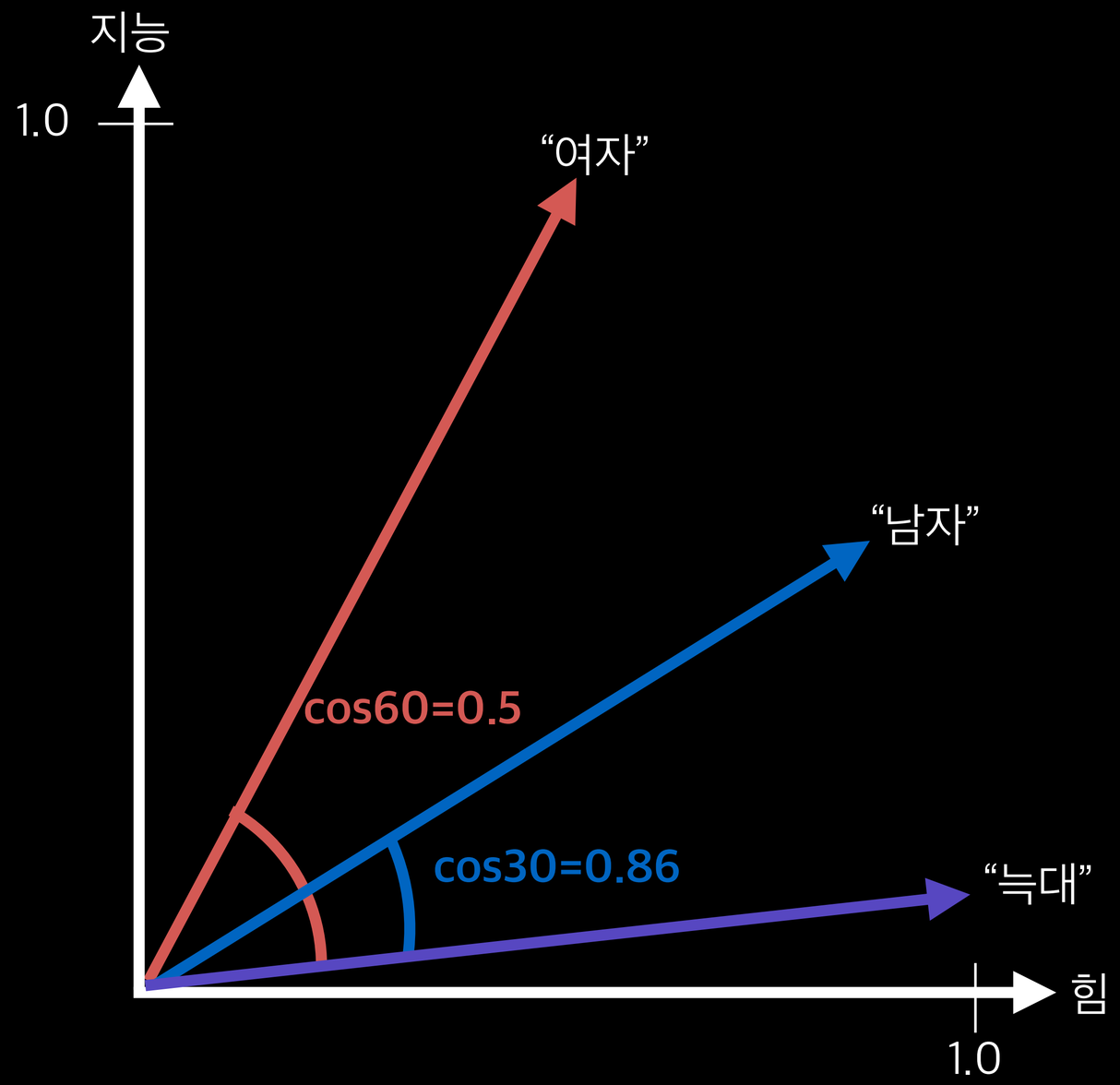
	남	여	자	늑	대
“남자”	O		O		
“늑대”				O	O
“여자”		O	O		

Vector based Representaion=Embedding

	힘	지능	덩치	고음
“남자”	0.9	0.5	0.8	0.2
“늑대”	1.0	0.1	0.9	0.7
“여자”	0.4	0.9	0.5	1.0

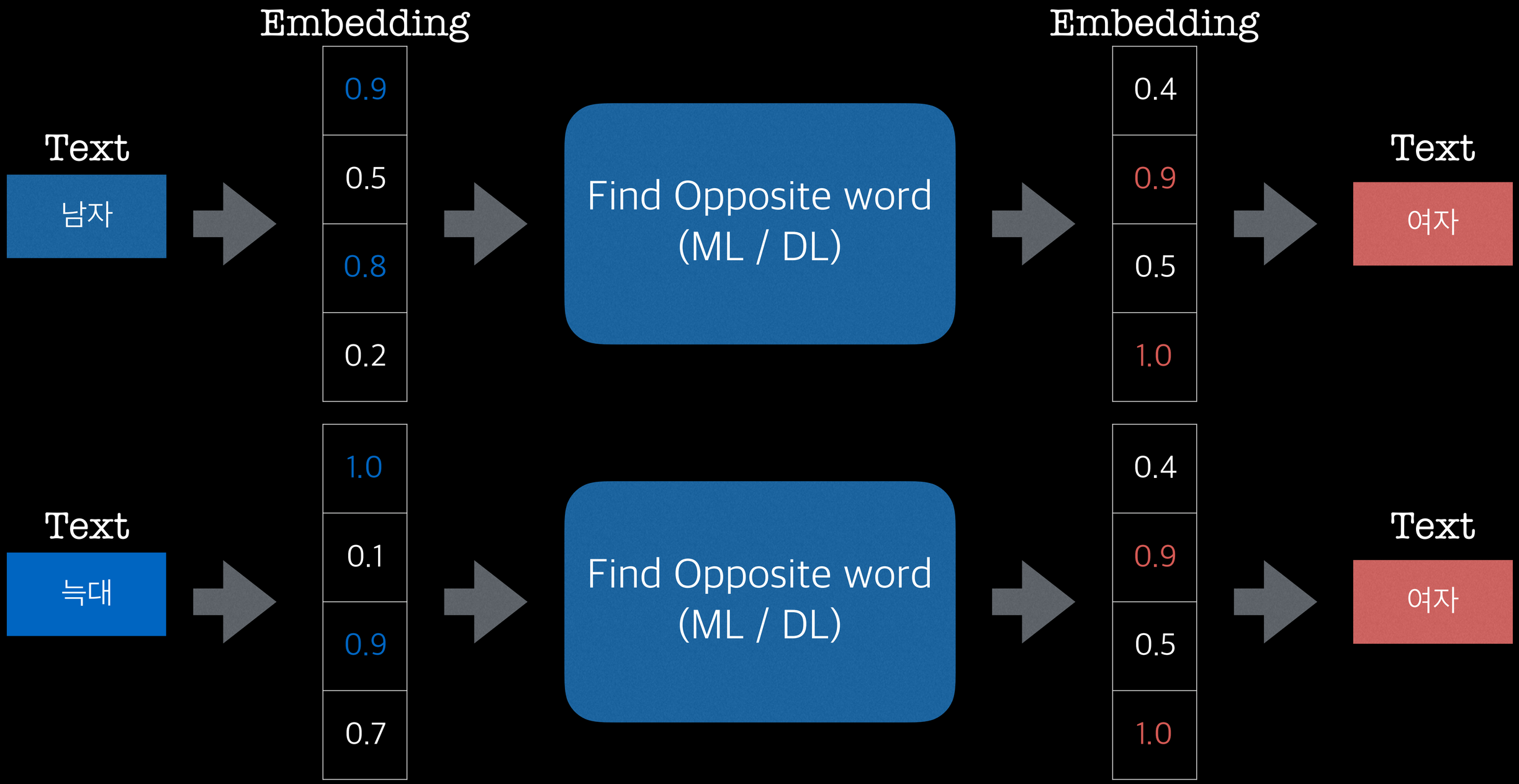
Word Similarity

	힘	지능
“남자”	0.9	0.5
“늑대”	1.0	0.1
“여자”	0.4	0.9



sim(“남자”, “늑대”) > sim(“여자”, “늑대”)

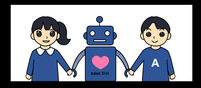
남자 ≈ 늑대


Embedding
Embedding
Text
남자
0.9
0.5
0.8
0.2
Find Opposite word
(ML / DL)
0.4
0.9
0.5
1.0
Text
여자
Text
늑대
1.0
0.1
0.9
0.7
Find Opposite word
(ML / DL)
0.4
0.9
0.5
1.0
Text
여자

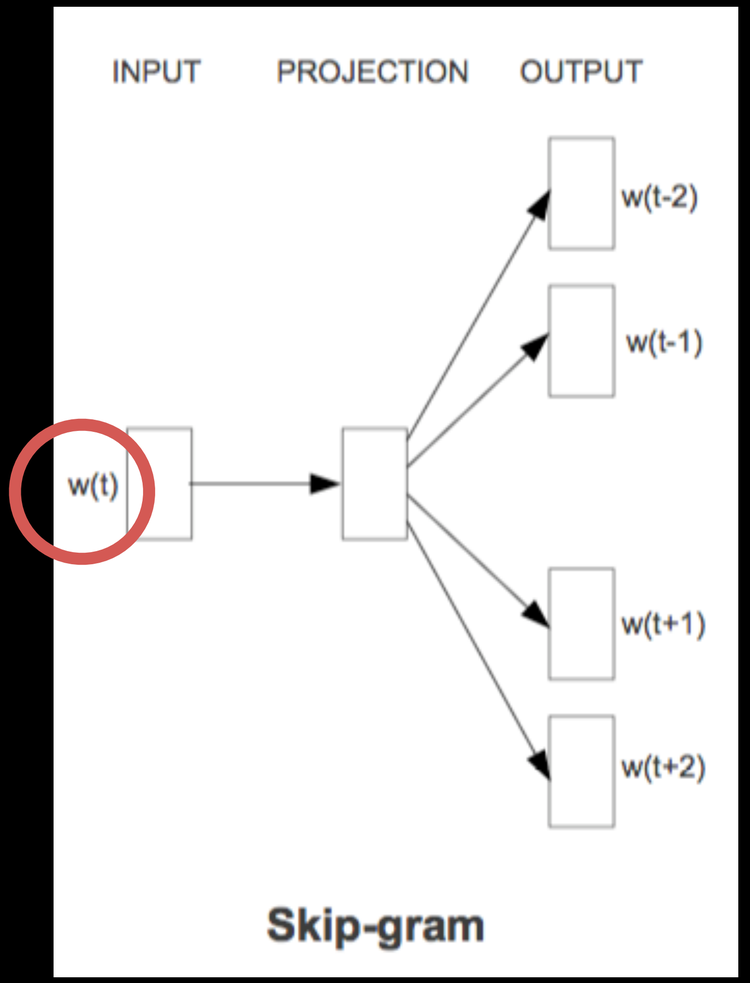
Language Models

N-Gram, Word2vec, Fasttext, Glove...

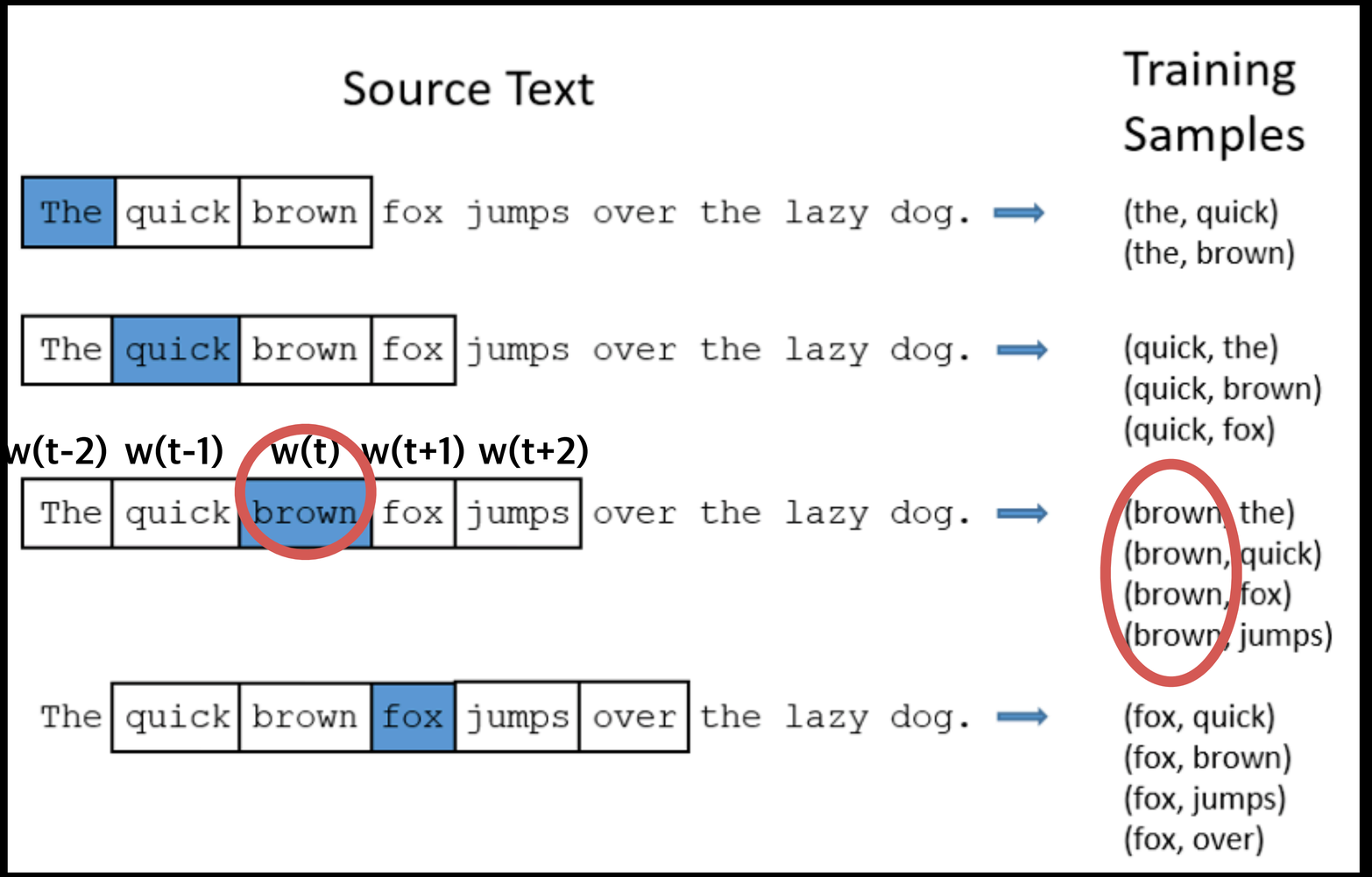
nlp4kor



Word2vec Skip-Gram Model



https://arxiv.org/pdf/1301.3781v3.pdf

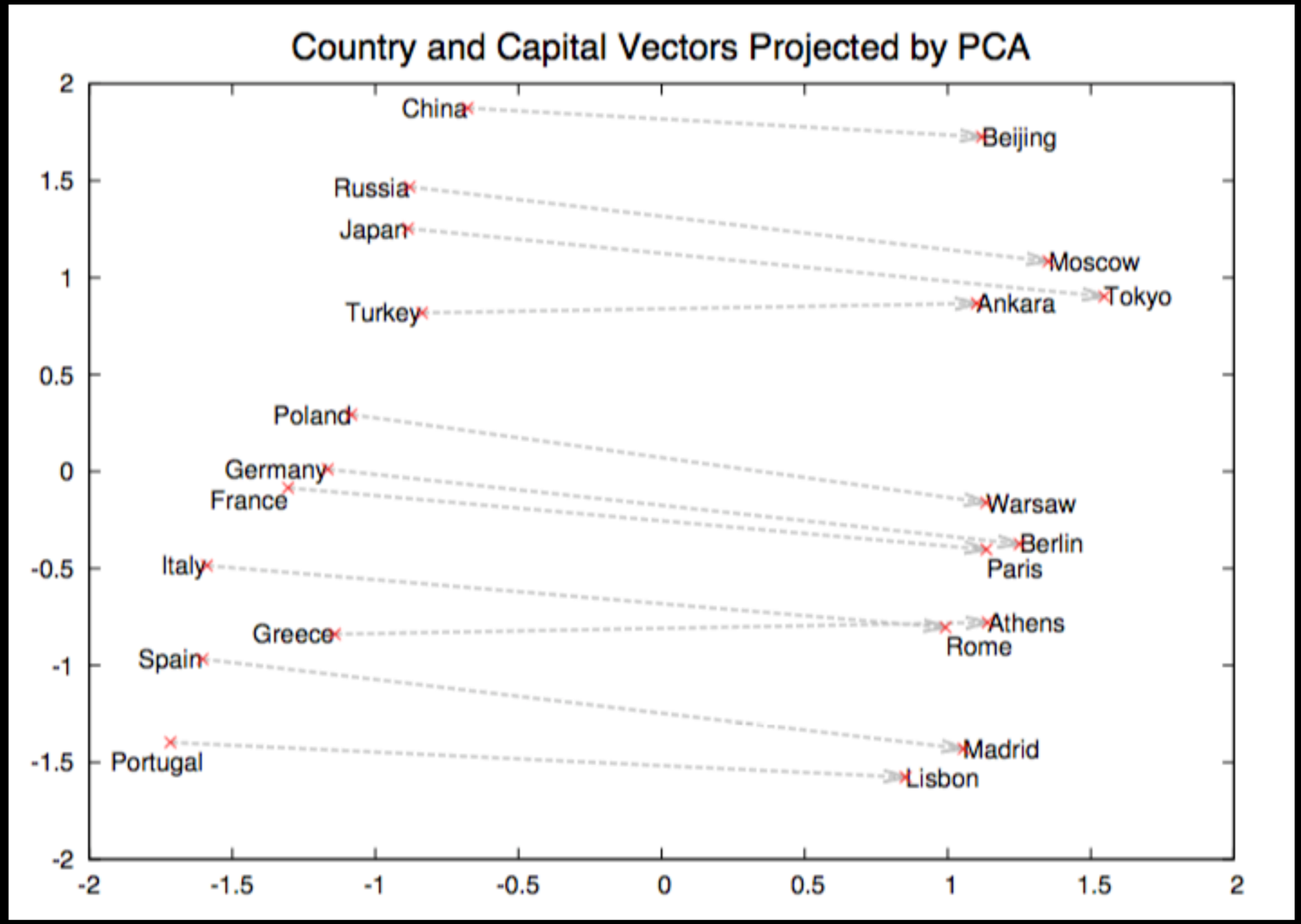


http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Word Similiarity in word2vec

30 page in https://www.slideshare.net/lucypark/nltk-gensim

Word Relationship in word2vec

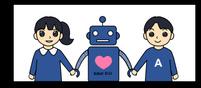


https://arxiv.org/pdf/1310.4546.pdf

Create your own word embedding

with open-sources

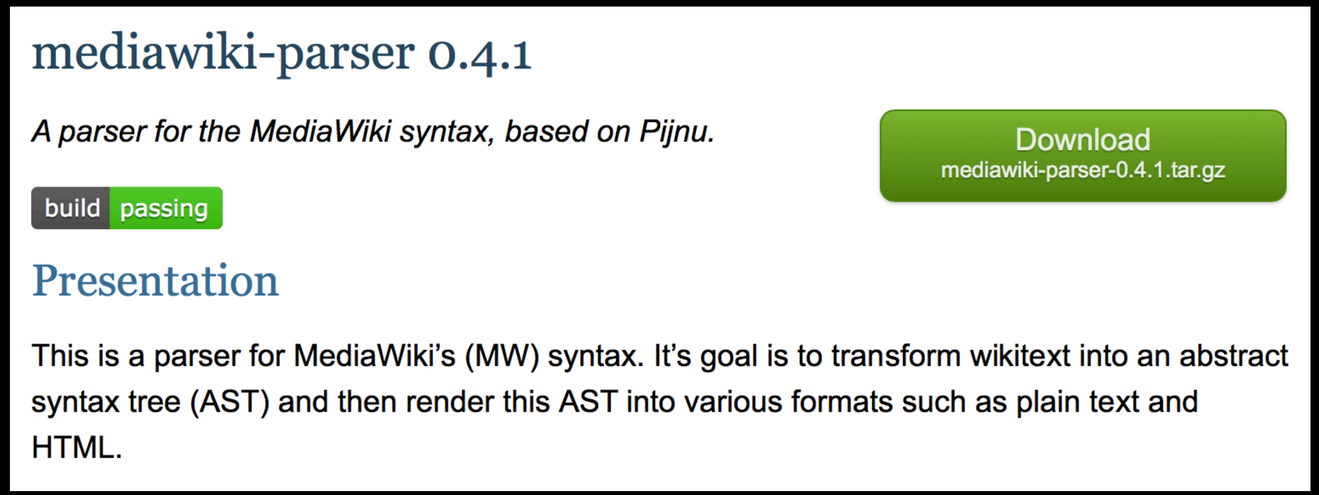
nlp4kor

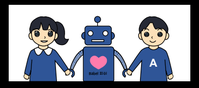


Prepare Corpus

- **Download Korean Wikipedia dump file**
  - https://dumps.wikimedia.org/kowiki/20180220/

- **Parse dump file(mediawiki format) to text file**
  - https://pypi.python.org/pypi/mediawiki-parser/





Word2vec open source

- https://github.com/theeluwin/pytorch-sgns
  - Fast training Word2vec Skip-gram with Pytorch
  - Fully class based source codes

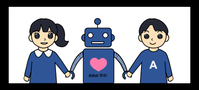
PyTorch SGNS

Word2Vec's **SkipGramNegativeSampling** in Python.

Yet another but quite general negative sampling loss implemented in PyTorch.

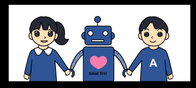
It can be used with ANY embedding scheme! Pretty fast, I bet.

```
vocab_size = 20000
word2vec = Word2Vec(vocab_size=vocab_size, embedding_size=300)
sgns = SGNS(embedding=word2vec, vocab_size=vocab_size, n_negs=20)
optim = Adam(sgns.parameters())
for batch, (iword, owords) in enumerate(dataloader):
    loss = sgns(iword, owords)
    optim.zero_grad()
    loss.backward()
    optim.step()
```

Korean Embedding on Github

- https://github.com/bage79/word2vec4kor
- **File format for Tensorboard Embedding(Projector)**
- **Corpus: https://ko.wikipedia.org**
  - Total sentences: about 3,115,431
  - Total Unique words: 10,000
  - Tokenized: white-space
  - Embedding Dimension: 300
  - Skip-Gram + Negative Sampling + Subsampling

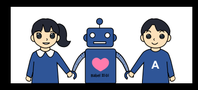


Q & A

[바벨피쉬 모임] Word2vec Embedding 실습

https://www.facebook.com/events/529492287451286/?notif_t=event_description_mention&notif_id=1519395371494159

nlp4kor



Tensorboard Demo