word2vec in action

Ермаков Петр

Семантическая близость?

Светильник

Лампа

Кипятильник

WordNet

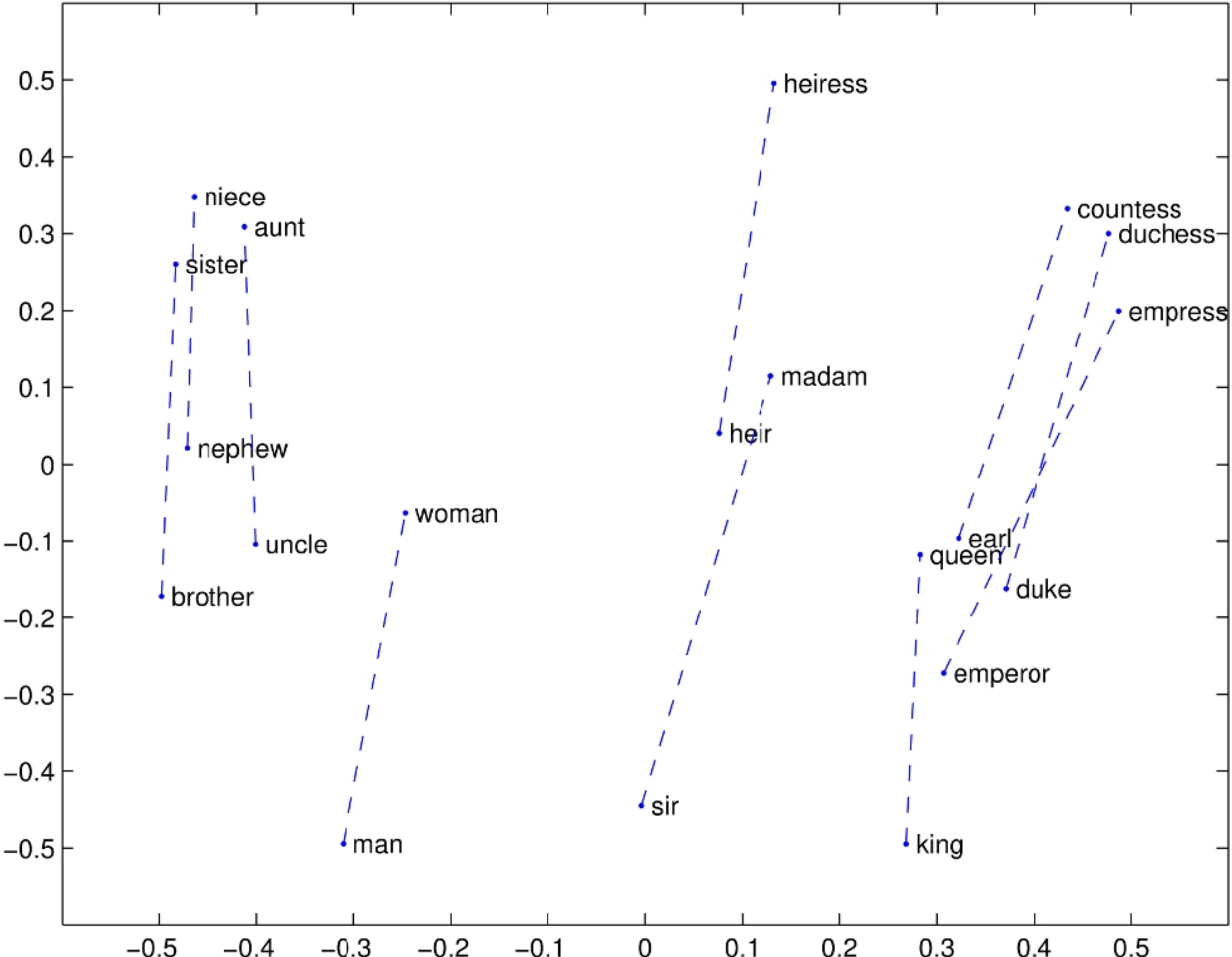
Но зачем?

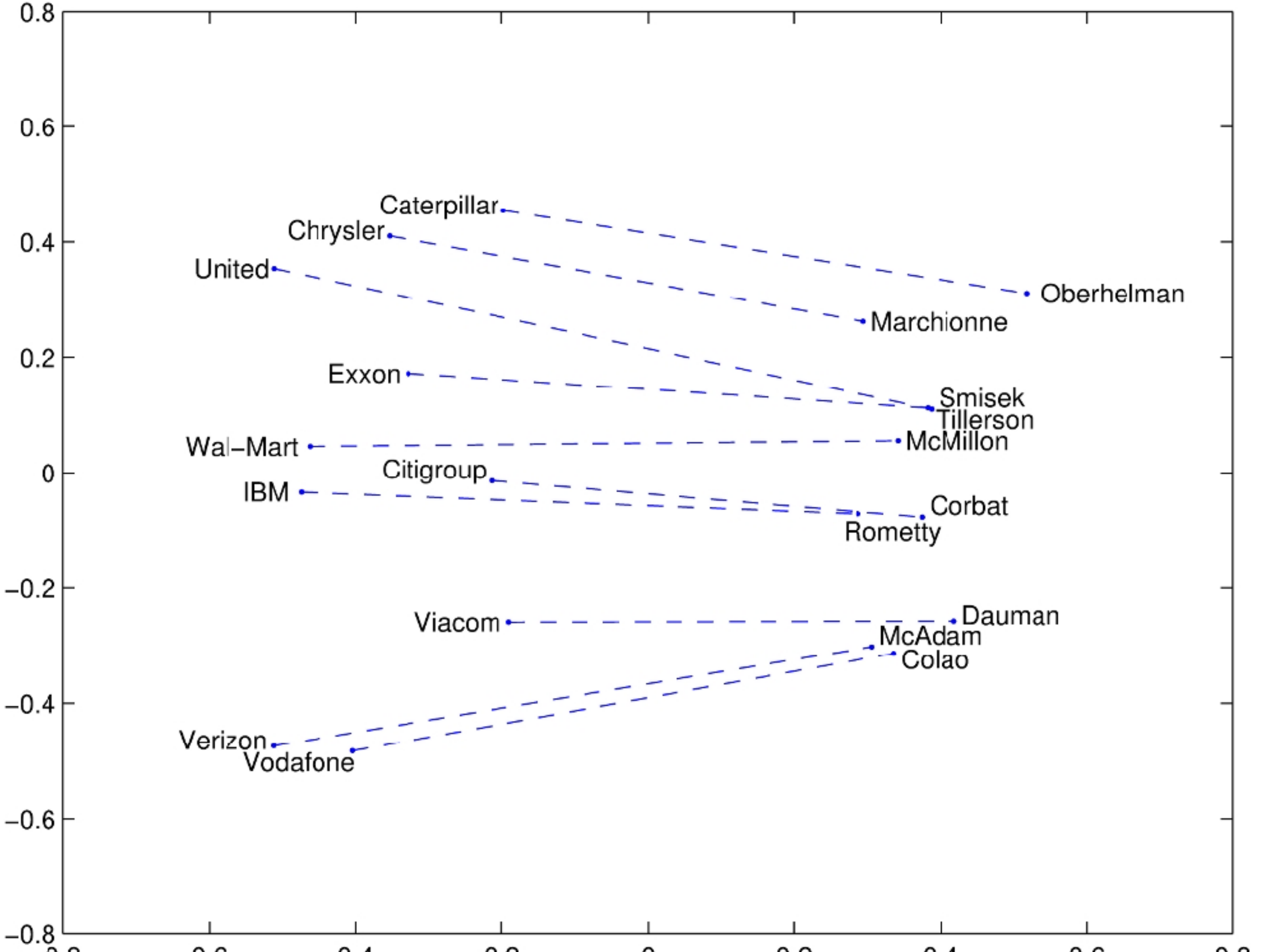
- Выделение синонимов и связанных слов
- Построение “семантической карты”

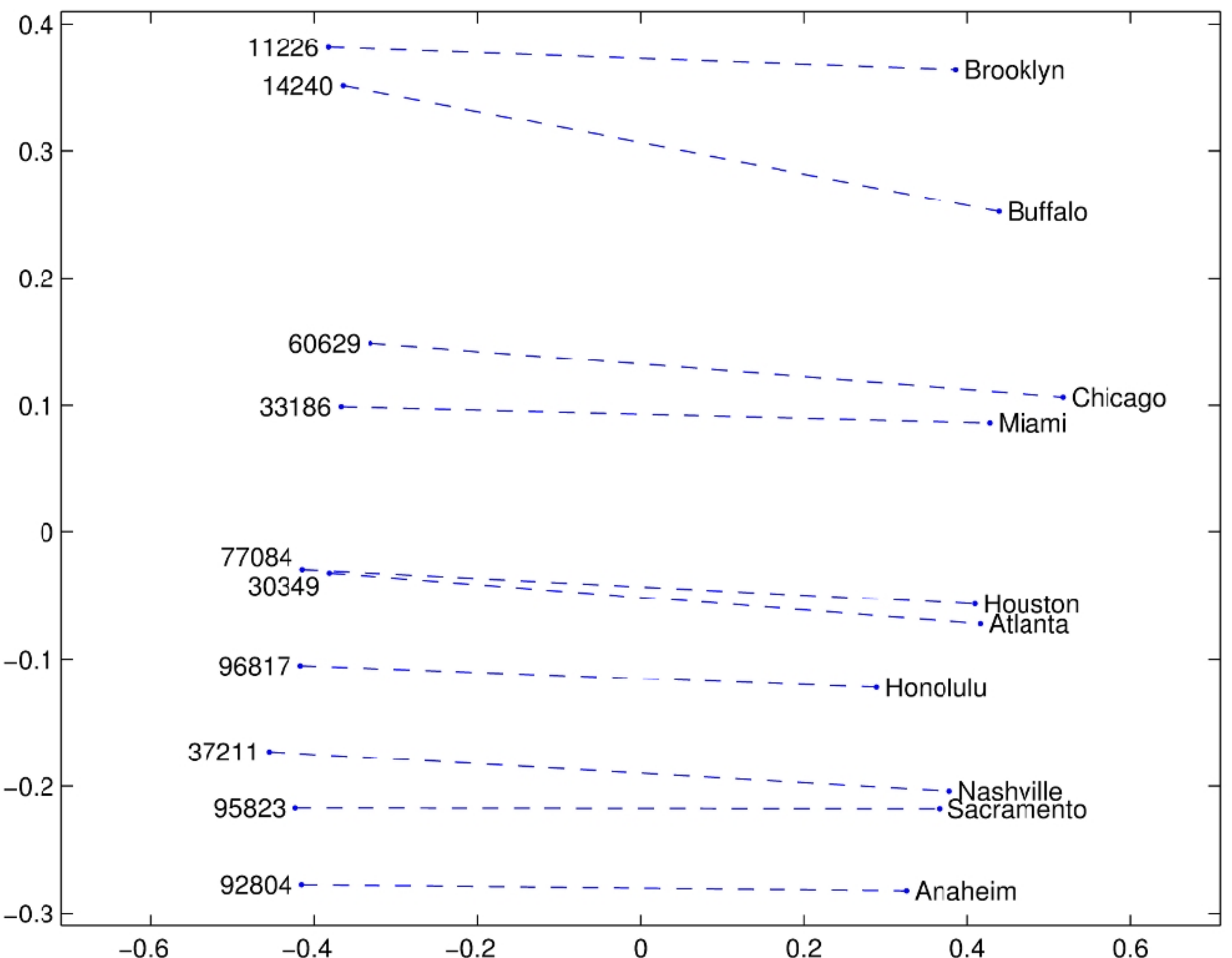
Семантика из паттернов употребления

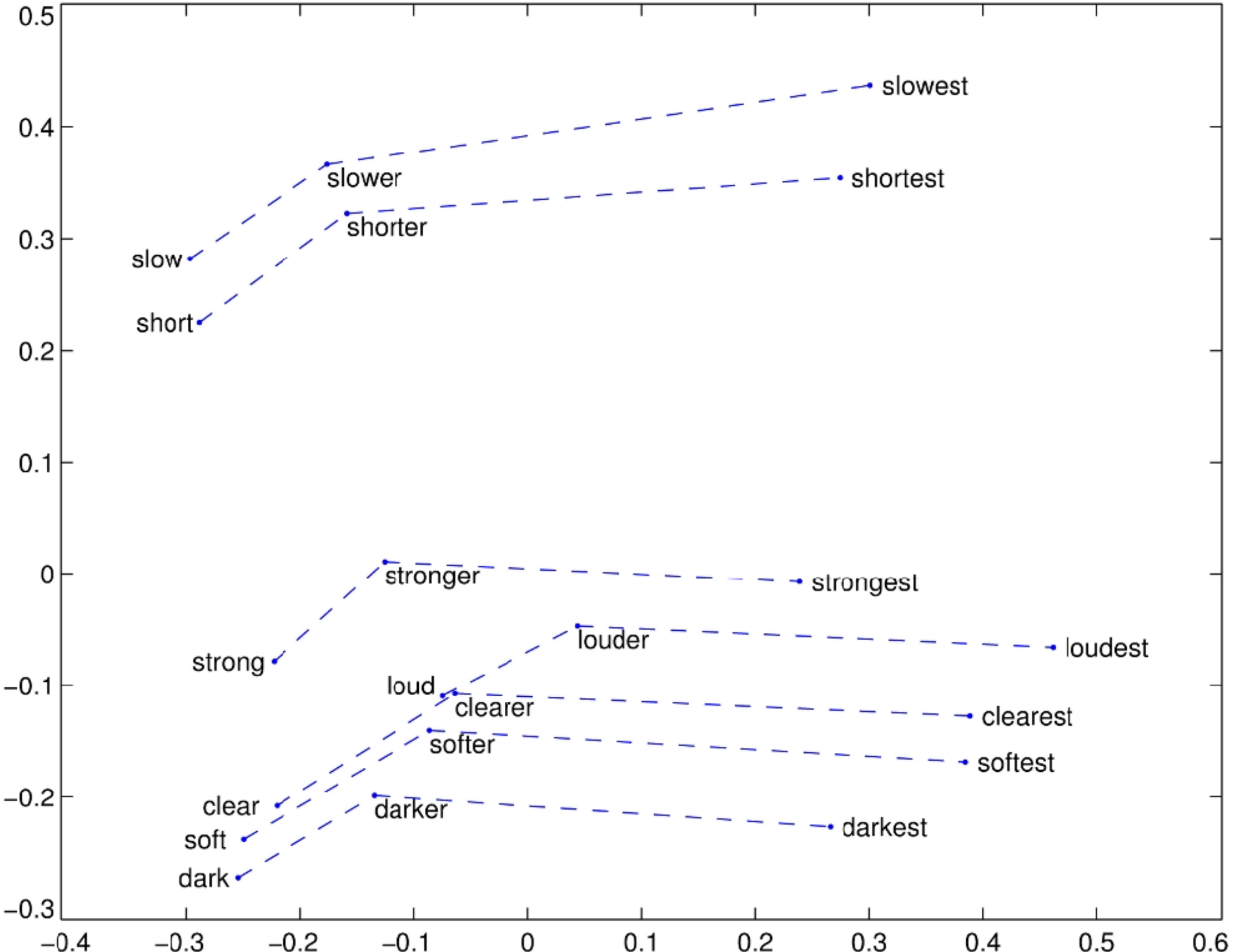
Счетные дистрибутивно-семантические модели

Count-based distributional semantic models


heiress
niece
aunt
countess
duchess
sister
empress
madam
heir
nephew
woman
uncle
earl
queen
duke
brother
emperor
man
sir
king
0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.5
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5


0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
Caterpillar
Chrysler
United
Exxon
Wal−Mart
Citigroup
IBM
Viacom
Verizon
Vodafone
Oberhelman
Marchionne
Smisek
Tillerson
McMillon
Corbat
Rometty
Dauman
McAdam
Colao


0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
11226
Brooklyn
14240
Buffalo
60629
Chicago
33186
Miami
77084
30349
Houston
Atlanta
96817
Honolulu
37211
Nashville
95823
Sacramento
92804
Anaheim

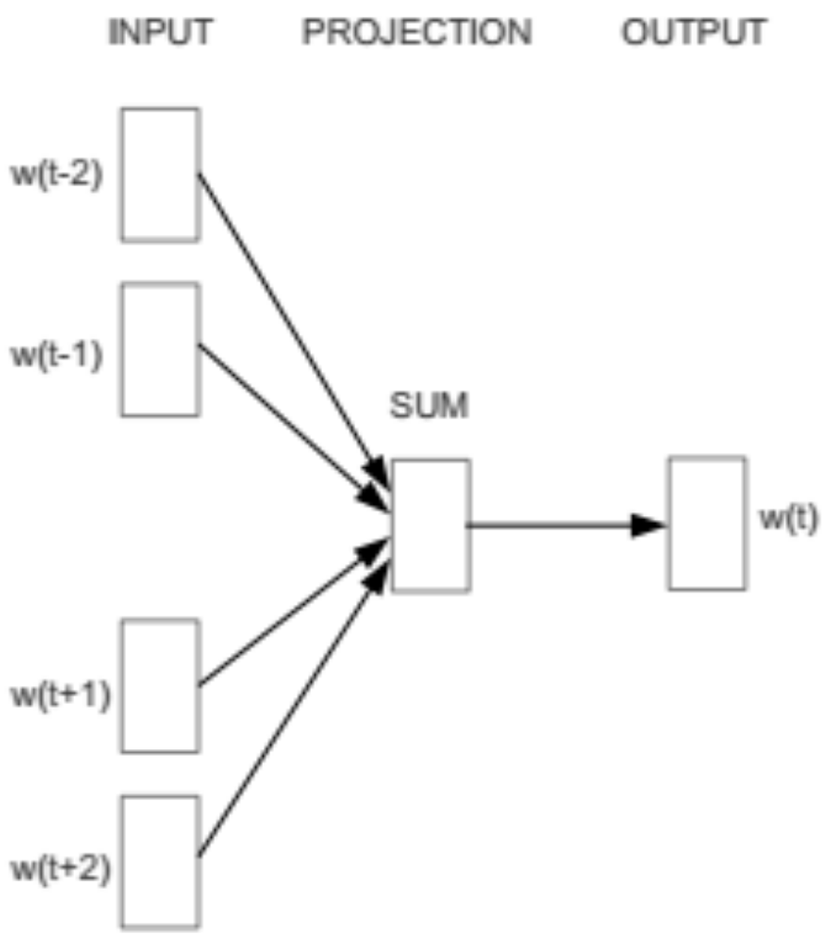

0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
slow
slower
slowest
short
shorter
shortest
strong
stronger
strongest
loud
louder
loudest
clear
clearer
clearest
soft
softer
softest
dark
darker
darkest

Neural Network time

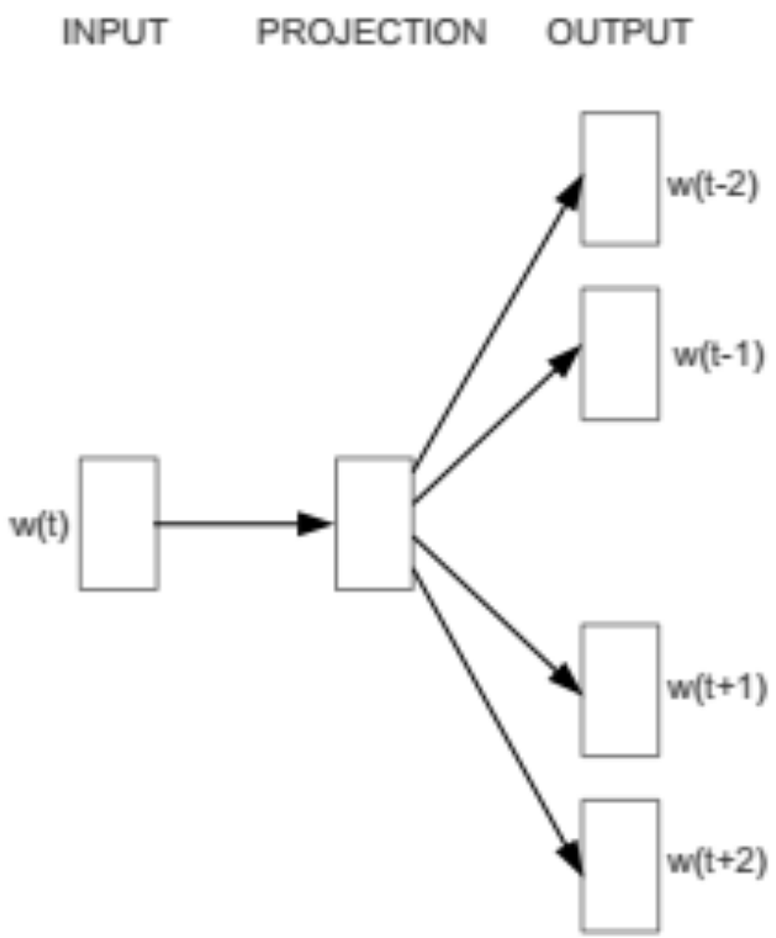
- Счетные модели

- Предсказательные модели

Модели



CBOW



Skip-gram

Параметры обучения

- Модель
- Размерность вектора
- Размер окна
- Порог частотности

Что же выбрать?

UEDA
@hurutoriya

Following

Japanese NLP Problem. #MLSSKYOTO

RETWEETS 2,152 FAVOURITES 1,375

8:24 a.m. - 27 Aug 2015

Japanese lost sight of "nation" - the essence of foreign carrot regime problem (Shogakukan Novel) (2000) ISBN: 4094050914 [Japanese Import] Paperback Bunko

Лойс

Лойс

- ГОДНО, ЛОЙС

Лойс

- годно, лойс
- лойс за песню
- из принципа не поставлю лойс
- взаимные лойсы
- лойс если не согласен

Кек

Кек

- кек, что ли?

Кек

- кек, что ли?
- кек))))))
- ну ты кек

In [61]:
```
m.most_similar(positive=['привет'], negative=[],
               topn=50)
```

Out[61]:
```
[('приветик', 0.9303760528564453),
 ('здравствуй', 0.8811429142951965),
 ('приветики', 0.857758760452270 5),
 ('привеет', 0.7910666465759277),
 ('здарова', 0.7907091379165649),
 ('здорова', 0.775124192237854),
 ('салют', 0.7467054128646851),
 ('привееет', 0.7370582818984985),
 ('здраствуй', 0.7104891538619995),
 ('прив', 0.7028274536132812),
 ('дарова', 0.6875522136688232),
 ('прива', 0.6685740351676941),
 ('привт', 0.659491240978241),
 ('саламалейкум', 0.6565588116645813),
 ('превет', 0.6515309810638428),
 ('пртвет', 0.6511324644088745),
 ('ривет', 0.6400219202041626),
 ('приветь', 0.639562368392944 3),
 ('салам', 0.6318571567535 4),
 ('привееееет', 0.6198863387107849),
 ('слушай', 0.6170339584350586),
 ('приветки', 0.6149711012840271),
 ('приветы', 0.6130839586257935),
```

In [59]:
```
m.most_similar(positive=['лойс'], negative=[],
               topn=50)
```

Out[59]:
```
[('луйс', 0.8383901119232178),
 ('лайк', 0.8372862339019775),
 ('камент', 0.824302077293396),
 ('взаимн', 0.8030086755752563),
 ('запесь', 0.796702980995178 2),
 ('комент', 0.7937247157096863),
 ('взаим', 0.790043294429779),
 ('лайни', 0.771076679229736 3),
 ('стену', 0.7693438529968262),
 ('позязя', 0.765210747718811),
```

```
m.most_similar(positive=['кек'], negative=[],
               topn=50)
```

```
[('азазат', 0.8450093865394592),
 ('лол', 0.6260291337966919),
 ('кека', 0.5775775909423828),
 ('кекво', 0.5586311817716919),
 ('мема', 0.5537031888961792),
```

music playlist

https://github.com/mattdennewitz/playlist-to-vec

social networks

http://arxiv.org/abs/1403.6652

user interests

http://cs.stanford.edu/~quocle/paragraph_vector.pdf

Что почитать?

- Mikolov, Tomas, et al. "E cient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

- Baroni, Marco, Georgiana Dinu, and Germ´an Kruszewski. "Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol. 1. 2014.

- Levy, Omer, and Yoav Goldberg. "Dependency-based word embeddings." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Vol. 2. 2014.

- Kutuzov, Andrey and Andreev, Igor. “Texts in, meaning out: neural language models in semantic similarity task for Russian.” Proceedings of the Dialog 2015 Conference, Moscow, Russia

“Type a quote here.”

–Johnny Appleseed