Лингвистика для математиков

Дистрибутивная семантика

План на жизнь

На следующей паре тест!

Пользоваться можно будет всем кроме соцсетей и соседей

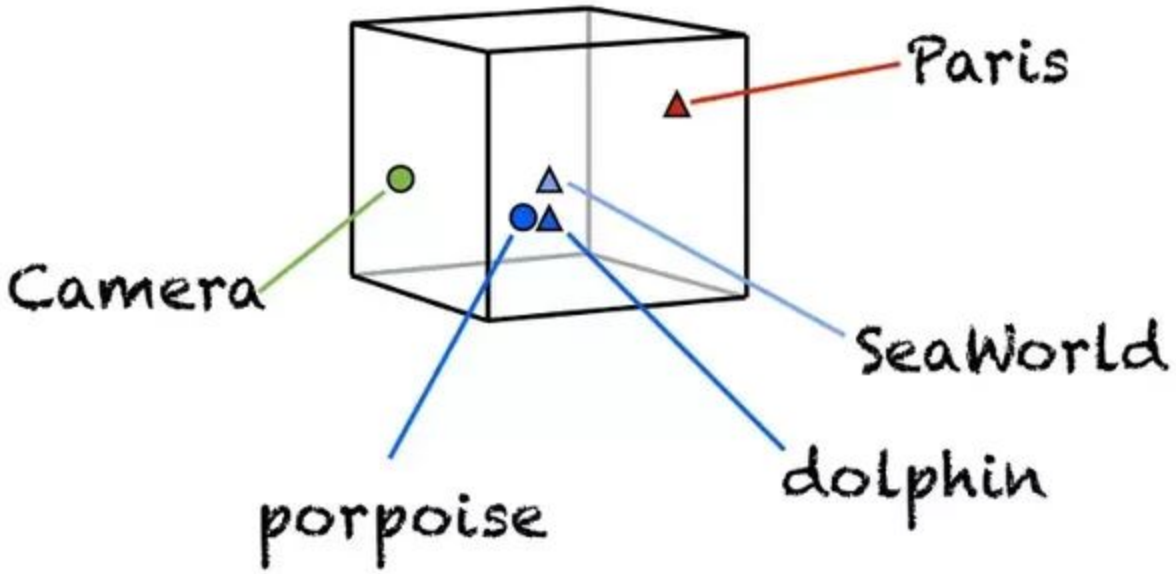
Дистрибутивная семантика

Дистрибутивная семантика — это область лингвистики, которая занимается вычислением степени семантической близости между лингвистическими единицами на основании их дистрибуционных признаков в больших массивах лингвистических данных.

Дистрибутивная семантика

Значение слова - это **сумма** всех его контекстов

Каждое слово\лексическая единица - вектор



Векторное представление слов

Каждое слово - это вектор

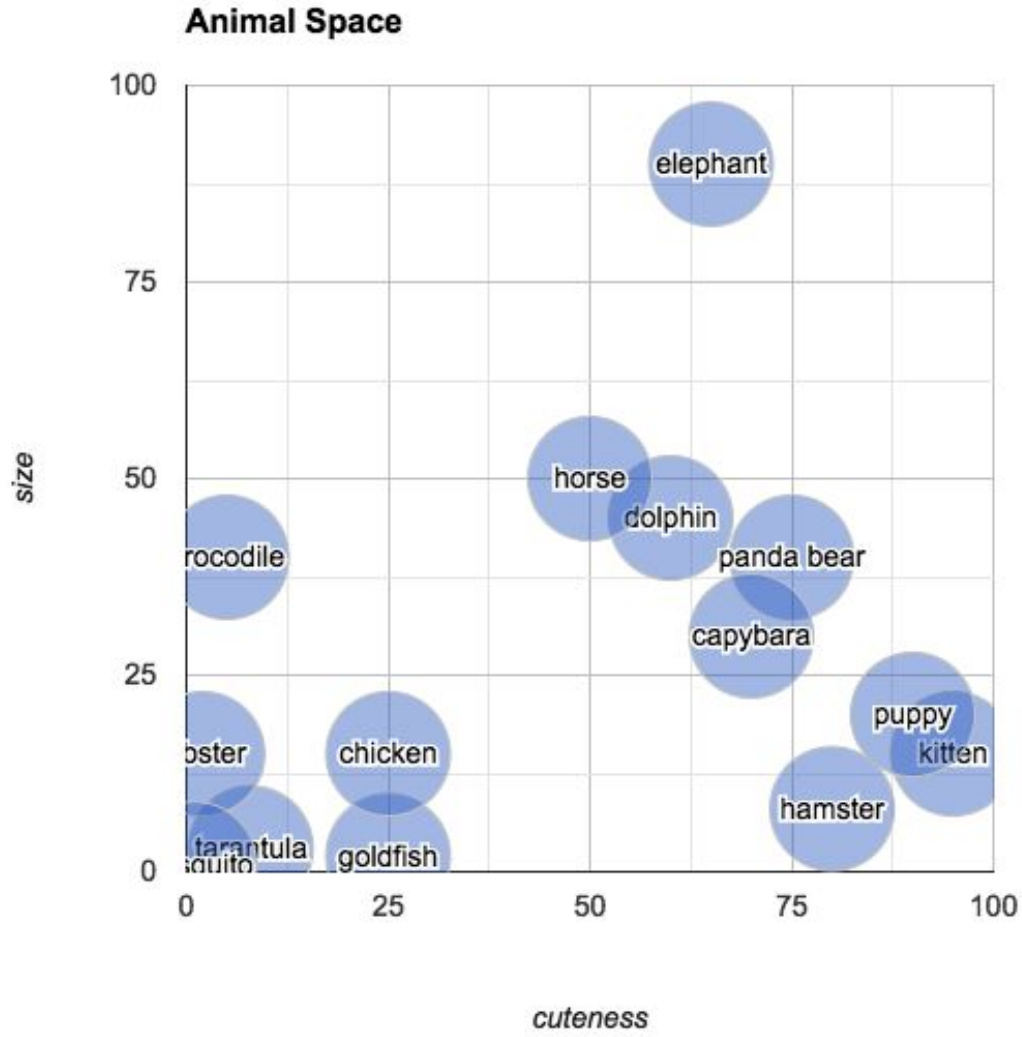
Иначе: word embeddings

Аналогии

Как вообще мы используем концепт векторного представления для сравнения каких-либо сущностей?

	cuteness (0–100)	size (0–100)
kitten	95	15
hamster	80	8
tarantula	8	3
puppy	90	20
crocodile	5	40
dolphin	60	45
panda bear	75	40
lobster	2	15
capybara	70	30
elephant	65	90
mosquito	1	1
goldfish	25	2
horse	50	50
chicken	25	15

Аналогии


Animal Space
100
75
50
25
0
size
elephant
horse
dolphin
panda bear
capybara
rocodile
puppy
kitten
chicken
hamster
goldfish
0
25
50
75
100
cuteness

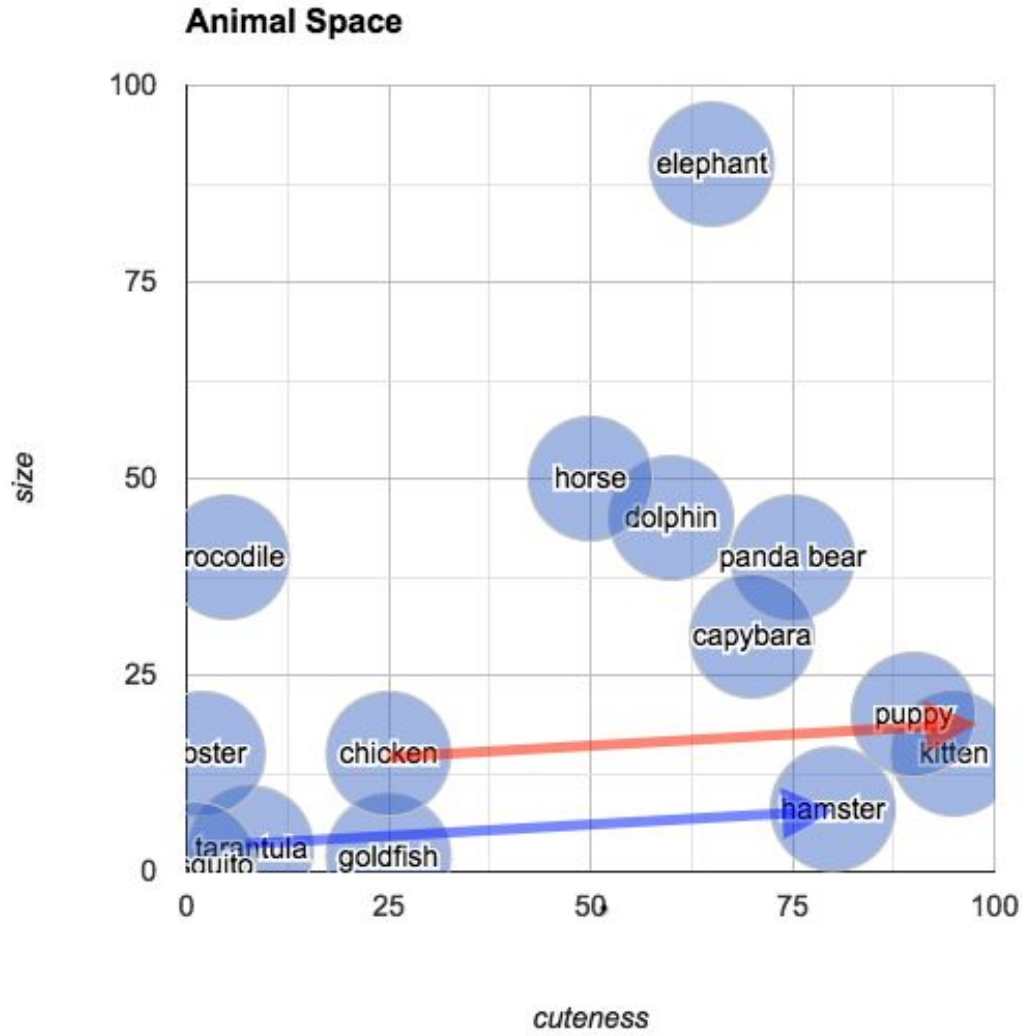
Аналогии

Следующий шаг - Расстояние!

С животными мы исследуем Евклидово расстояние:

$$P = \sqrt{\sum_{i=1}^{N} \left(A_i - B_i \right)^2}$$

Аналогии


Animal Space
100
75
50
25
0
size
elephant
horse
dolphin
panda bear
rocodile
capybara
puppy
kitten
chicken
hamster
goldfish
0
25
50
75
100
cuteness

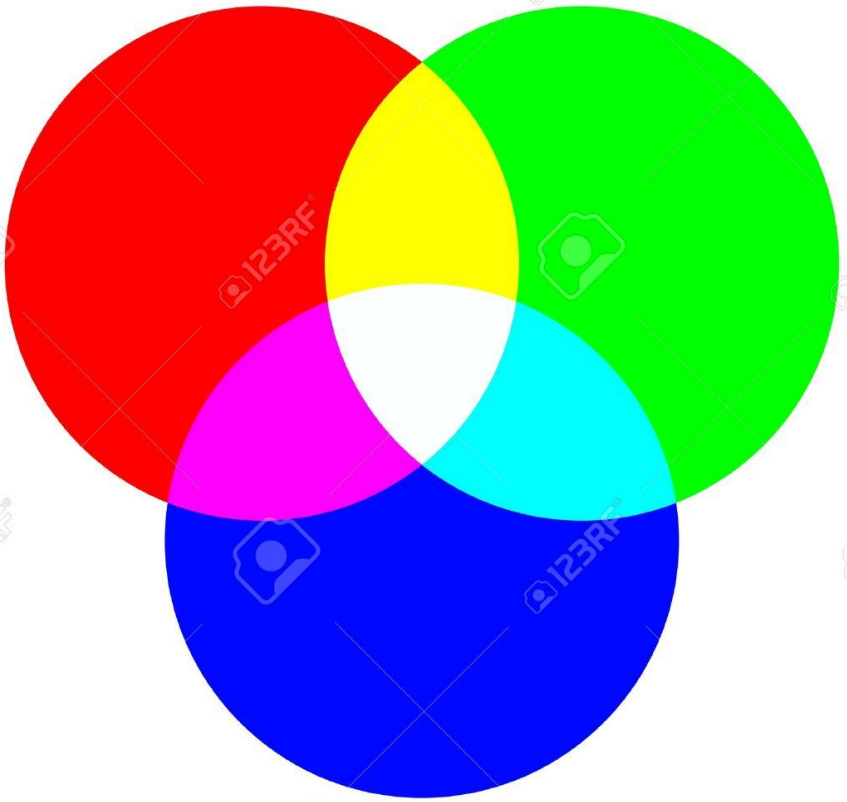
Аналогии

Цвета RGB = вектора в трехмерном пространстве

Красный - (229, 0, 0)

Черный - (0, 0, 0)

Оливковый - (110, 117, 14)

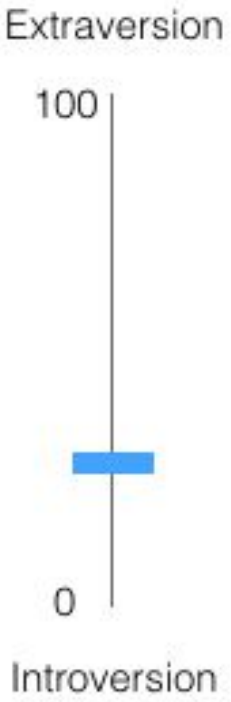


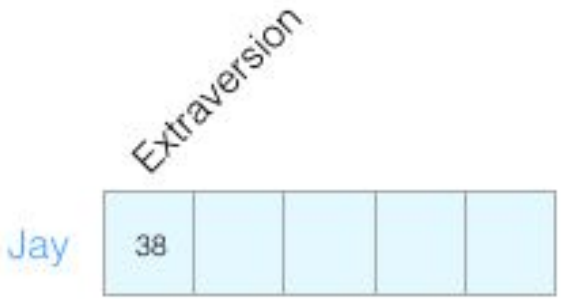
RGB

Аналогии

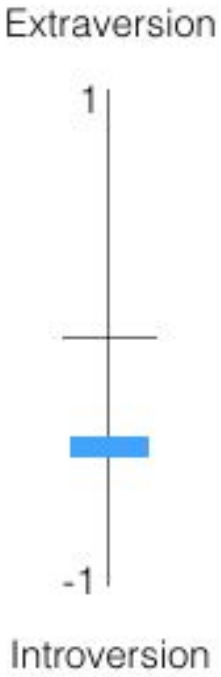
Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100

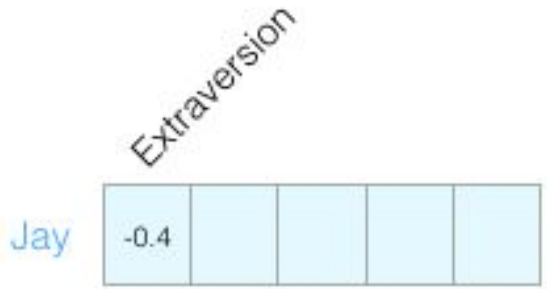
Аналогии



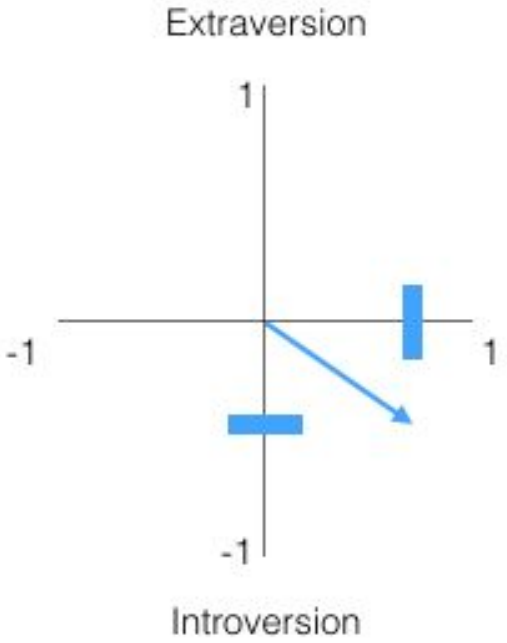


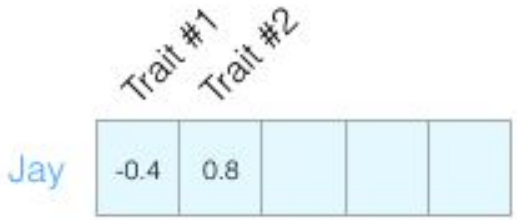
Аналогии



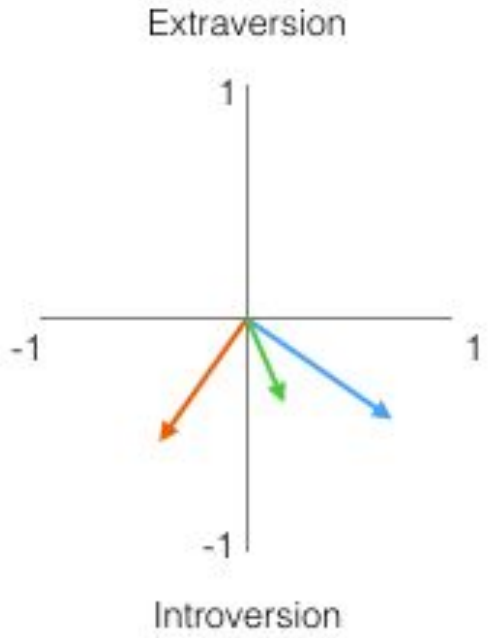


Аналогии



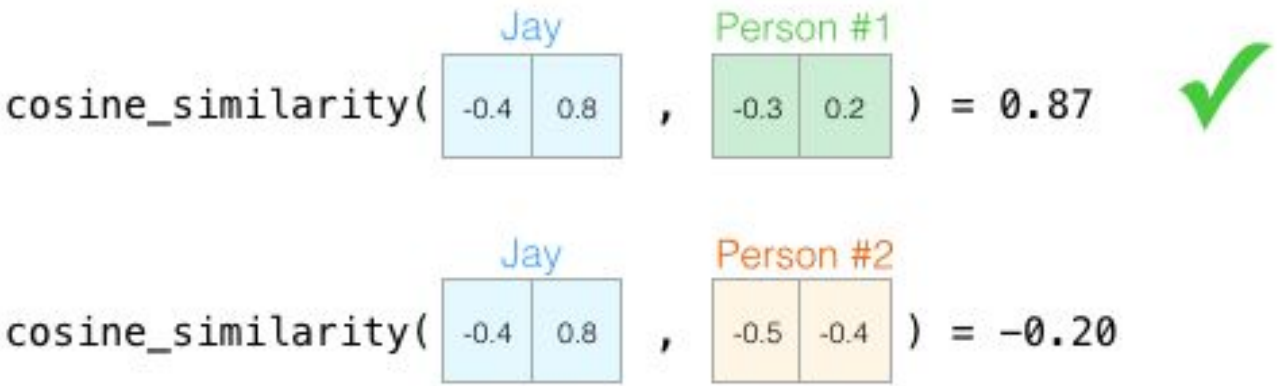


Аналогии

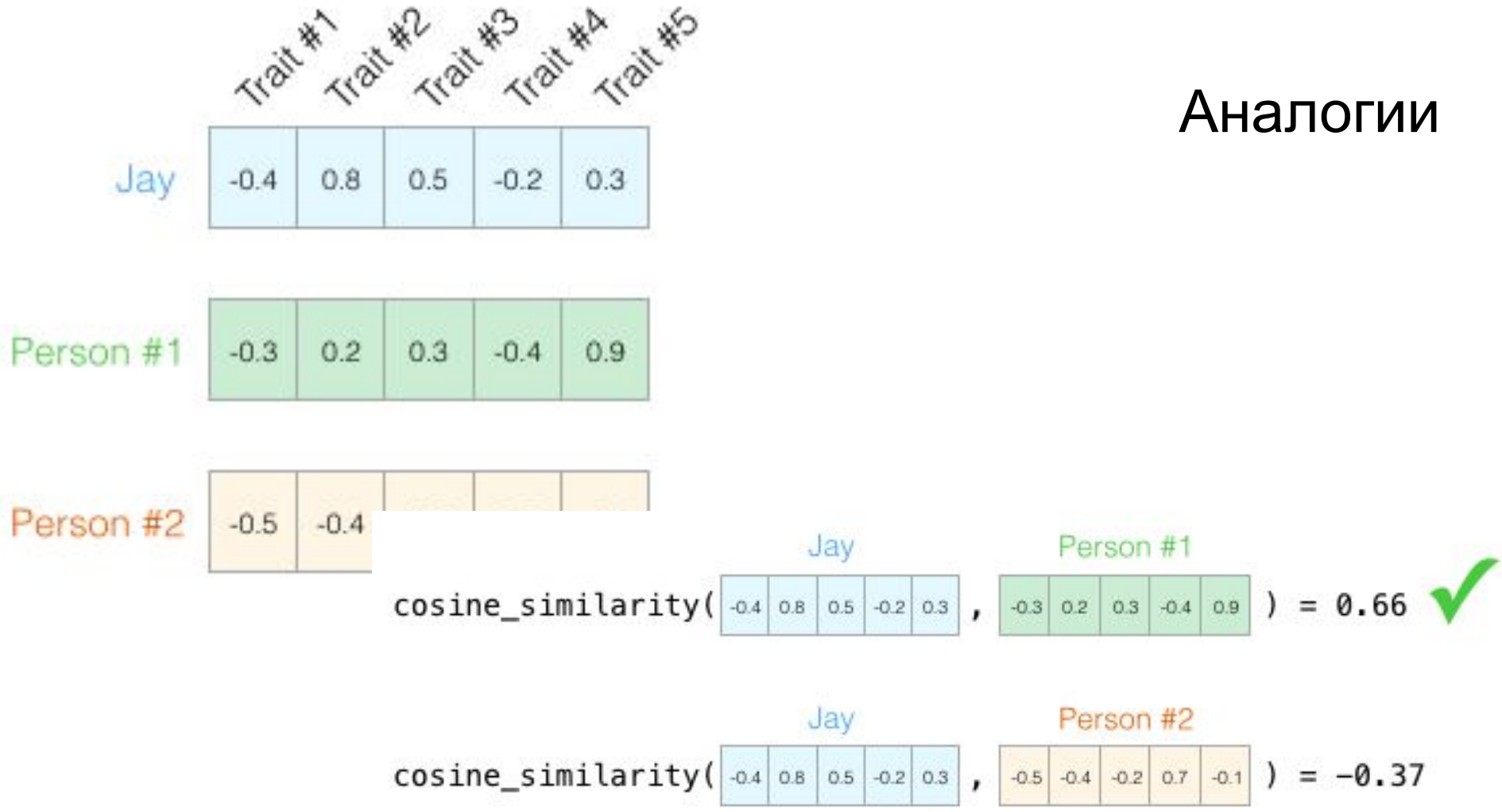

Extraversion
1
-1
1
-1
Introversion


Trait #1
Trait #2
Jay
-0.4
0.8
Person #1
-0.3
0.2
Person #2
-0.5
-0.4

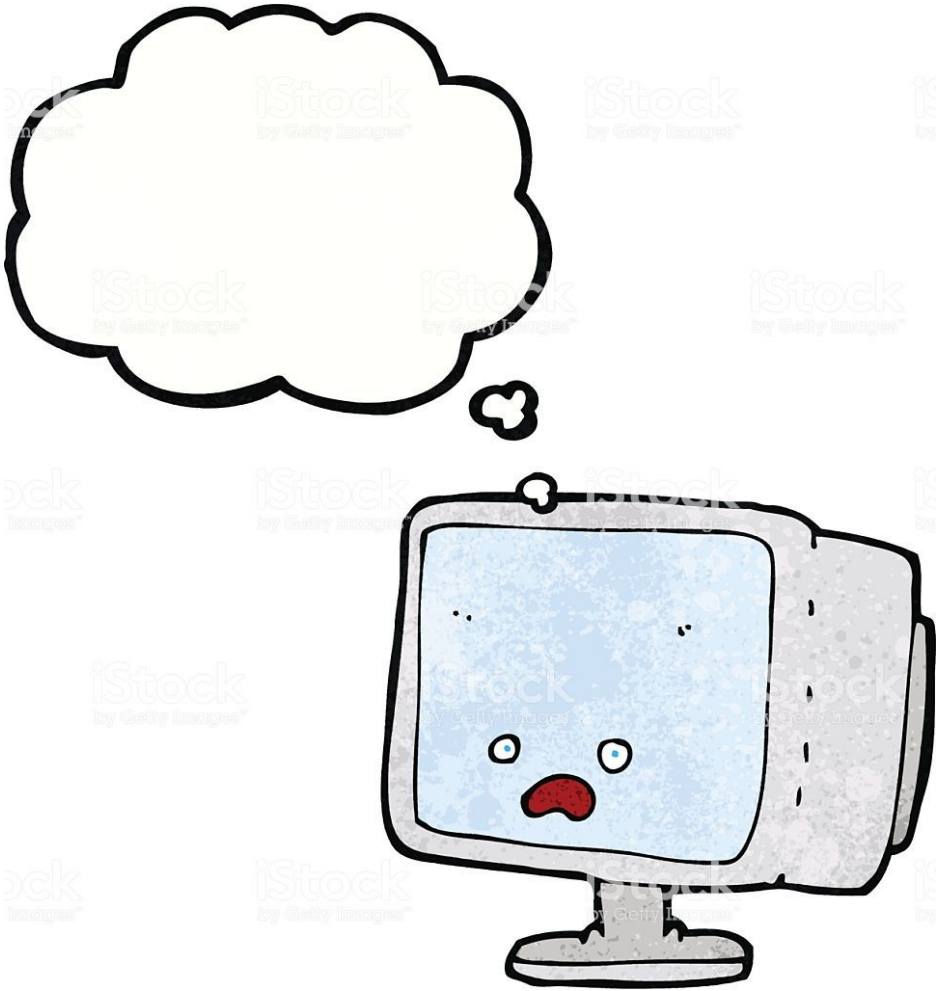
Аналогии


Jay
Person #1
cosine_similarity(-0.4 0.8 , -0.3 0.2) = 0.87
Jay
Person #2
cosine_similarity(-0.4 0.8 , -0.5 -0.4) = -0.20

Аналогии


Trait #1
Trait #2
Trait #3
Trait #4
Trait #5
Jay
-0.4
0.8
0.5
-0.2
0.3
Person #1
-0.3
0.2
0.3
-0.4
0.9
Person #2
-0.5
-0.4
Jay
Person #1
cosine_similarity(-0.4 0.8 0.5 -0.2 0.3 , -0.3 0.2 0.3 -0.4 0.9) = 0.66
Jay
Person #2
cosine_similarity(-0.4 0.8 0.5 -0.2 0.3 , -0.5 -0.4 -0.2 0.7 -0.1) = -0.37

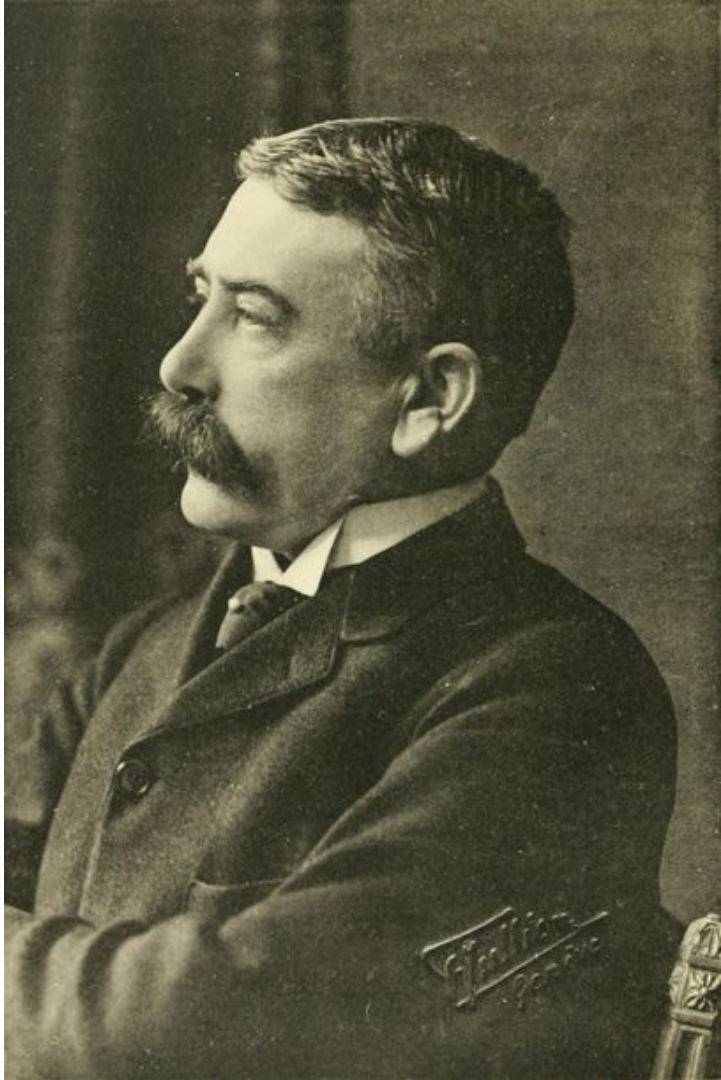
Наивный подход



Наивный подход

Разные виды близости слов: лексическая и семантическая

- петух
- курица
- цыпленок



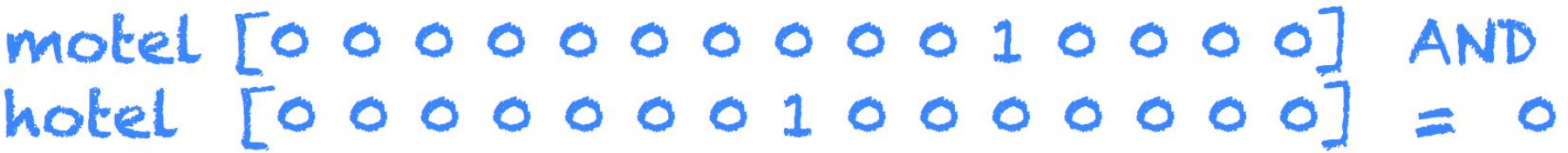
Векторное представление слов

One-hot encoding:

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Векторное представление слов

One-hot encoding:



Как закодировать слово

Счётчик:

... and the cute kitten purred and then ...
... the cute furry cat purred and miaowed ...
... that small kitten miaowed and she ...
... the loud furry dog ran and bit ...

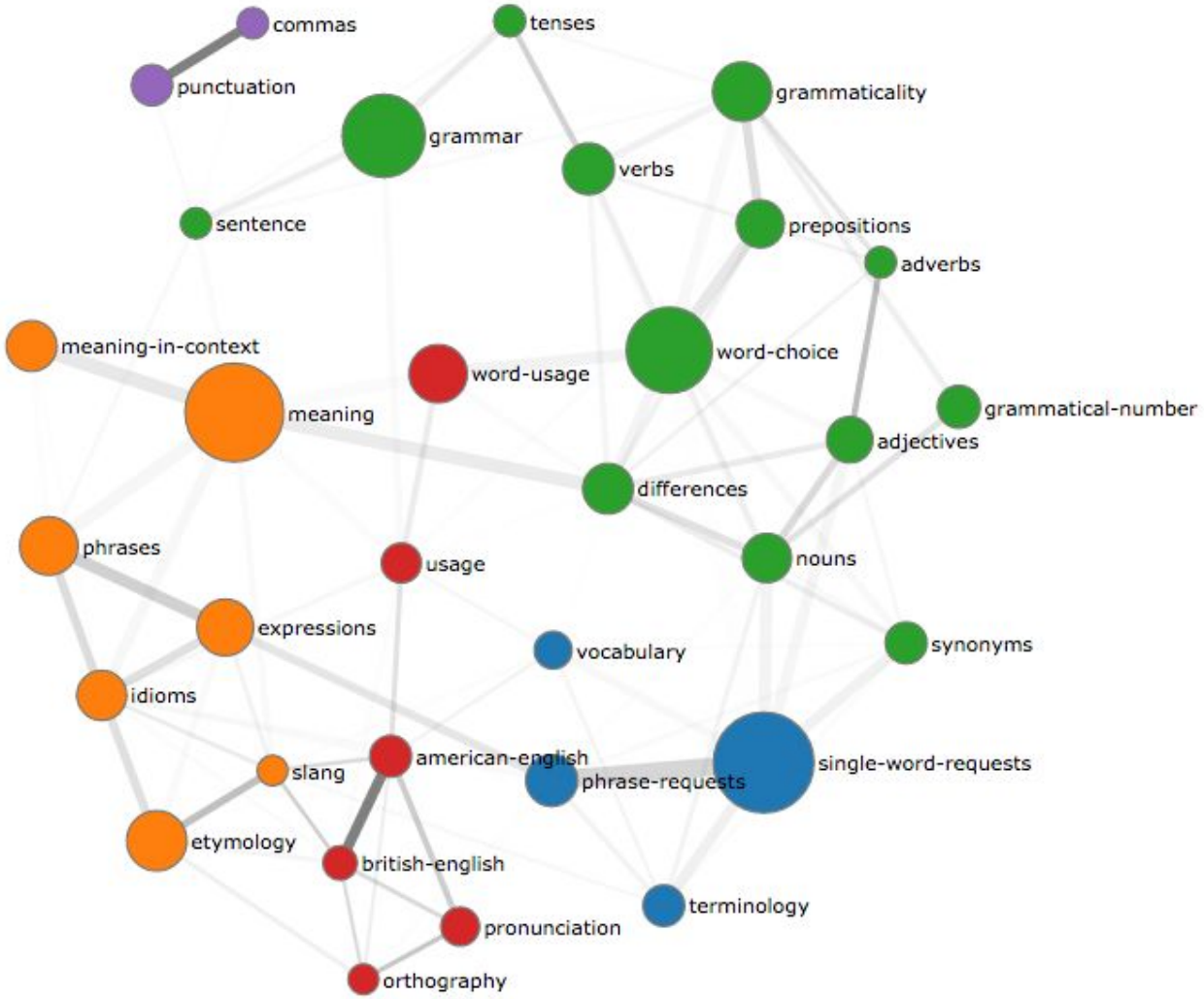
Словарь: bit, cute, furry, loud, miaowed, purred, ran, small

kitten: cute, purred, small, miaowed $\Rightarrow [0,1,0,0,1,1,0,1]^T$
cat: cute, furry, miaowed $\Rightarrow [0,1,1,0,1,0,0,0]^T$
dog: loud, furry, ran, bit $\Rightarrow [1,0,1,1,0,0,1,0]^T$

$$sim(w_1, w_2) = \frac{<w_1, w_2>}{||w_1|| \cdot ||w_2||}$$

Как обучить компьютер отделять слова?

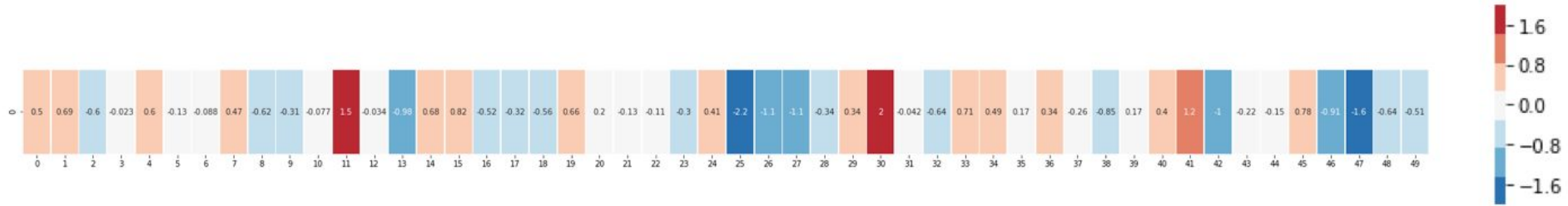
word2vec



Embeddings

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 ,
-0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 ,
-0.55809 , 0.66421 , 0.1961 , -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 ,
-0.34354 , 0.33505 , 1.9927 , -0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663
, -0.8523 , 0.1661 , 0.40102 , 1.1685 , -1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106
, -0.64426 , -0.51042 ]
```

Embeddings

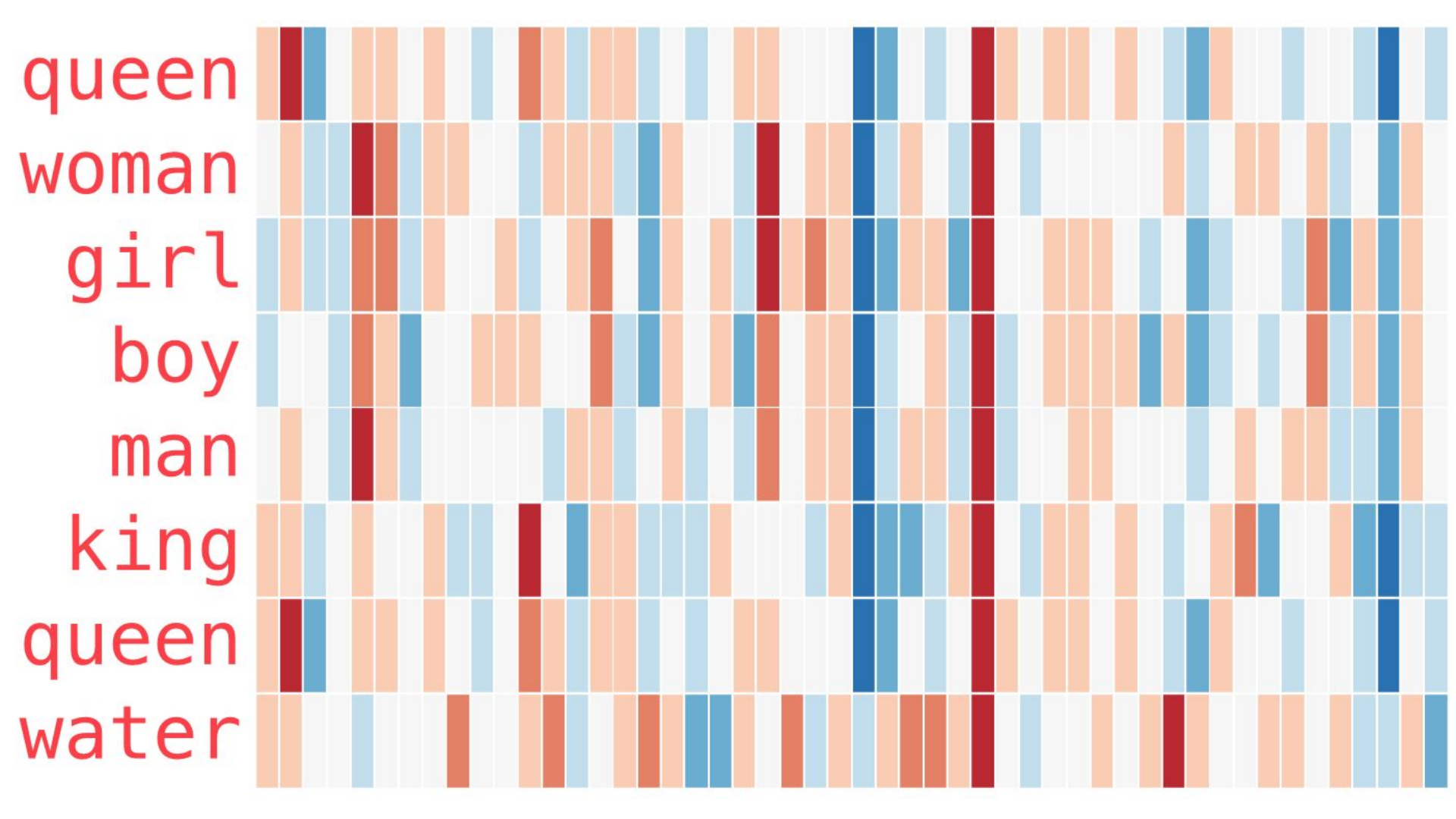

0
0.5 0.69 -0.6 -0.023 0.6 -0.13 -0.088 0.47 -0.62 -0.31 -0.077 1.5 -0.034 -0.98 0.68 0.82 -0.52 -0.32 -0.56 0.66 0.2 -0.13 -0.11 -0.3 0.41 -2.2 -1.1 -1.1 -0.34 0.34 2 -0.042 -0.64 0.71 0.49 0.17 0.34 -0.26 -0.85 0.17 0.4 1.2 -1 -0.22 -0.15 0.78 -0.91 -1.6 -0.64 -0.51
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
1.6
0.8
0.0
−0.8
−1.6

Embeddings

“king”

“Man”

“Woman”


queen
woman
girl
boy
man
king
queen
water

Embeddings

Вновь про понятие аналогий:

Кот соотносится с котенком так же, как курица с цыпленком

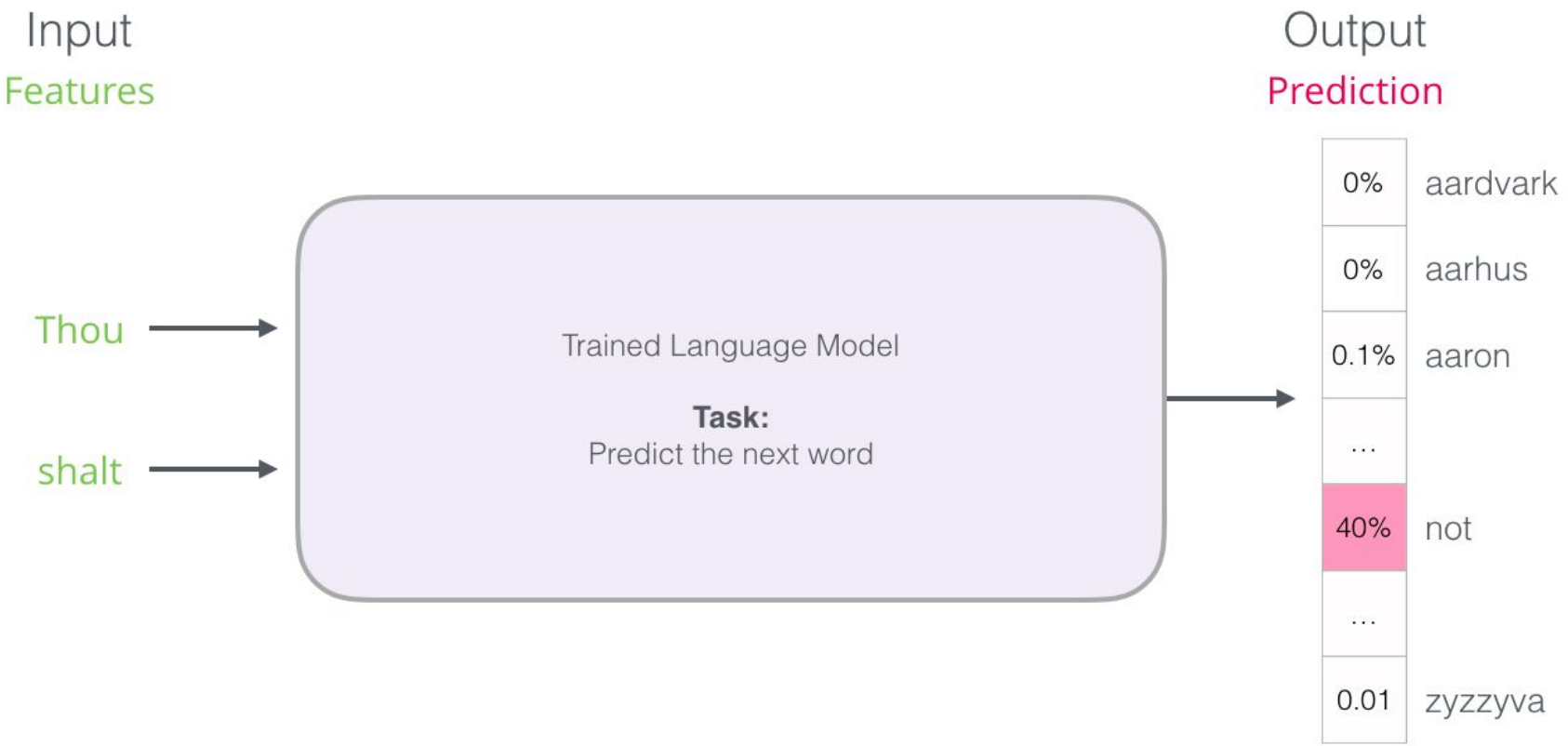
king − man + woman ~= queen

Нейронная модель языка

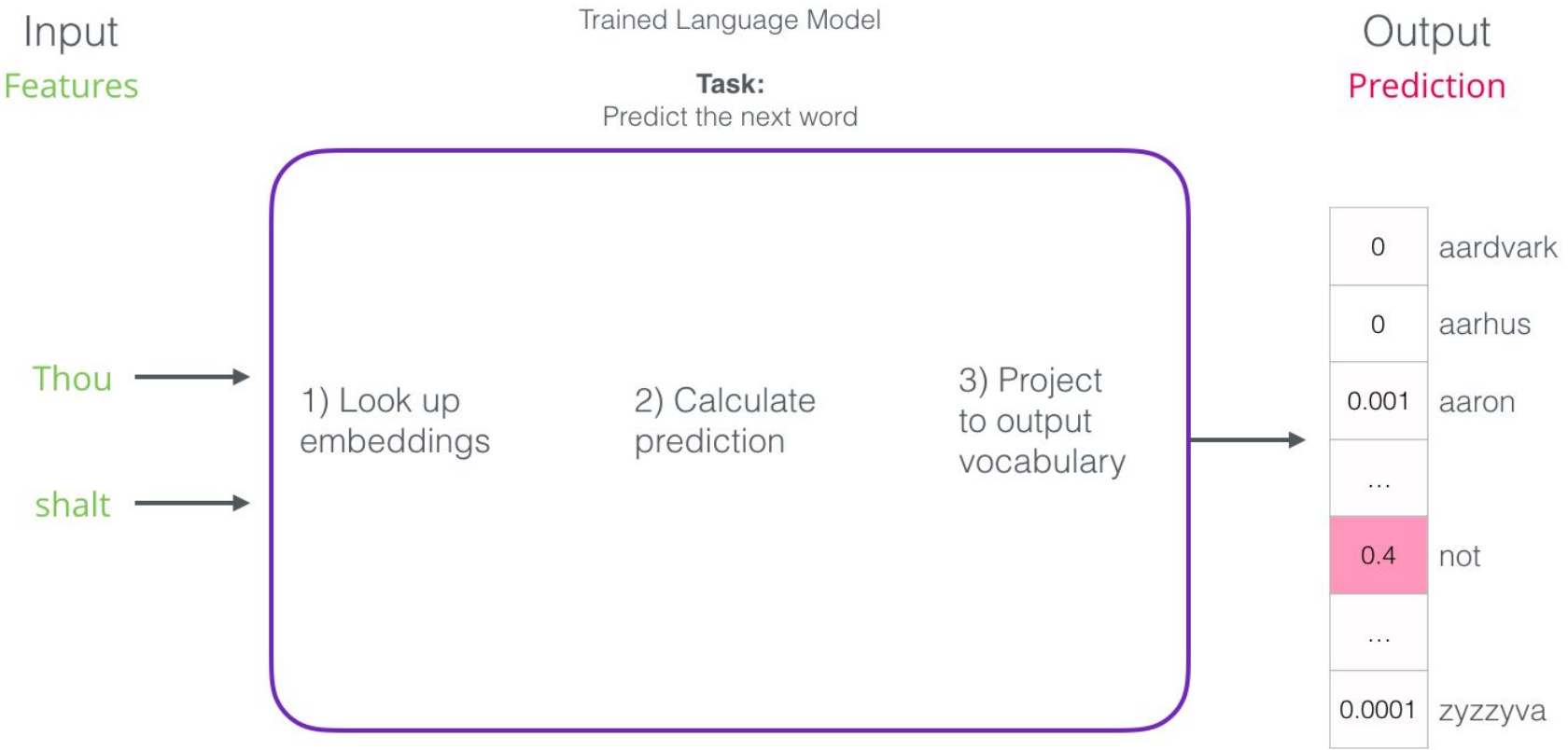
input/feature #1 input/feature #2 output/label

Thou shalt _____

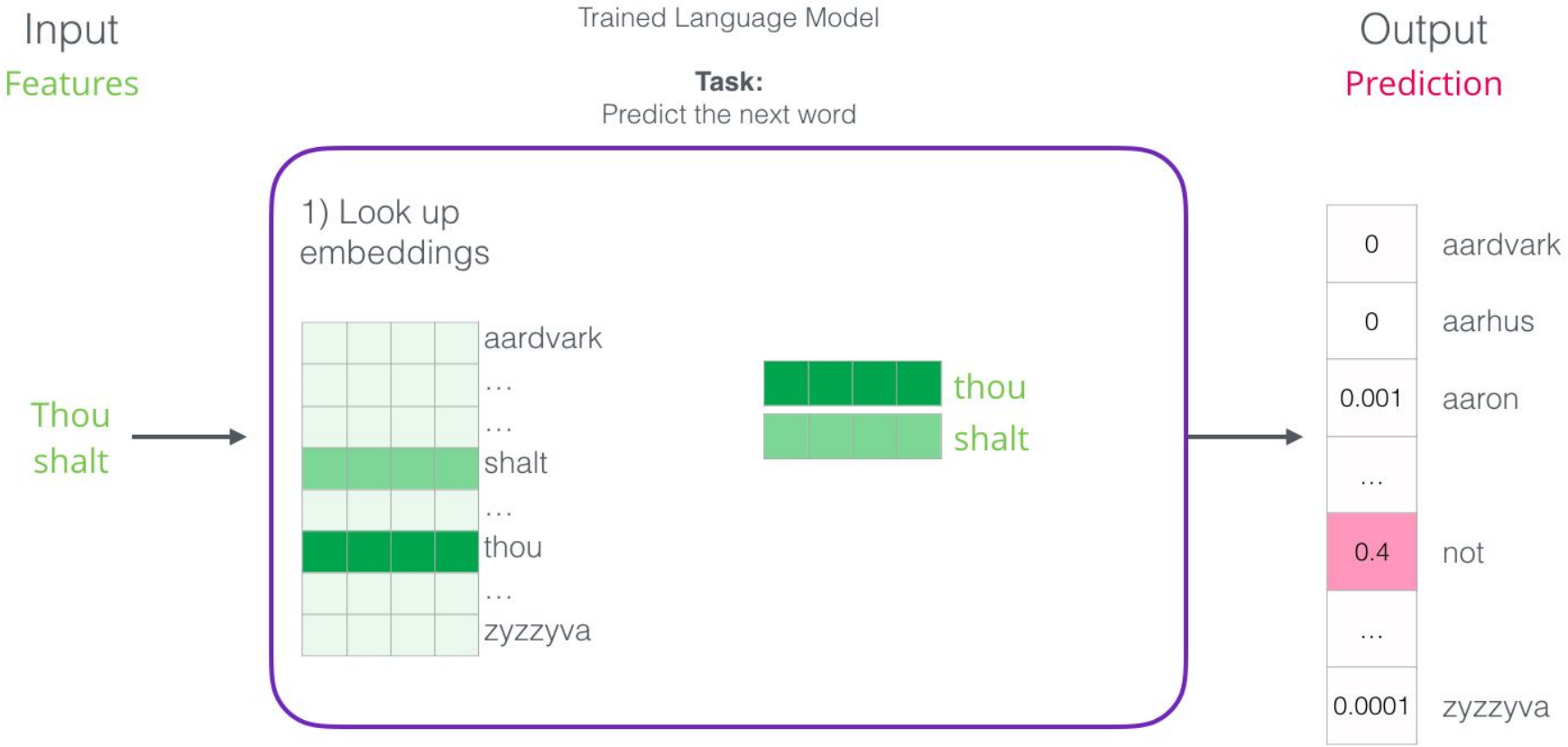
Нейронная модель языка



Нейронная модель языка



Нейронная модель языка



Обучение модели

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…

Dataset

input 1	input 2	output

Обучение модели

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…

Dataset

input 1	input 2	output
thou	shalt	not

Обучение модели

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make

Обучение модели

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make
not	make	a
make	a	machine
a	machine	in

Как работает word2vec?

Обучение модели более формально:

- **CBOW** предсказывает текущее слово, исходя из окружающего его контекста.
- **Skip-gram**, наоборот, использует текущее слово, чтобы предугадывать окружающие его слова.

Мешок слов

Jay was hit by a _____ bus in...

by	a	red	bus	in

input 1	input 2	input 3	input 4	output
by	a	bus	in	red

Skip-gram

Jay was hit by a red bus in...

Jay was hit by a red bus in...

by	a	red	bus	in

input	output
red	by
red	a
red	bus
red	in

Skip-gram

Thou shalt not make a machine in the likeness of a human mind

input word	target word

Skip-gram

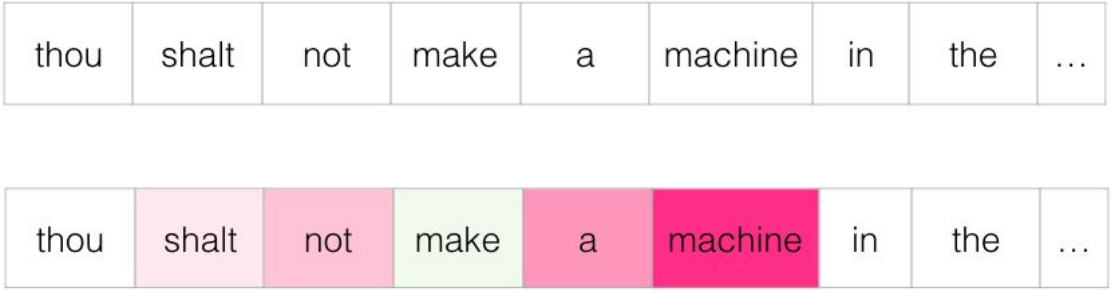
Thou shalt not make a machine in the likeness of a human mind

thou	shalt	not	make	a	machine	in	the	…

input word	target word
not	thou
not	shalt
not	make
not	a

Skip-gram

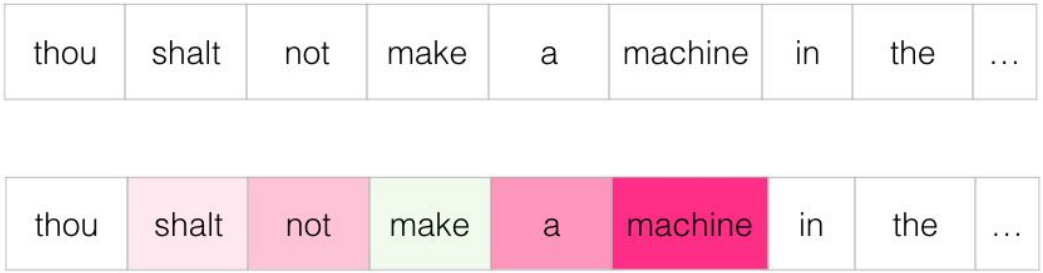
Thou shalt not make a machine in the likeness of a human mind



input word	target word
not	thou
not	shalt
not	make
not	a

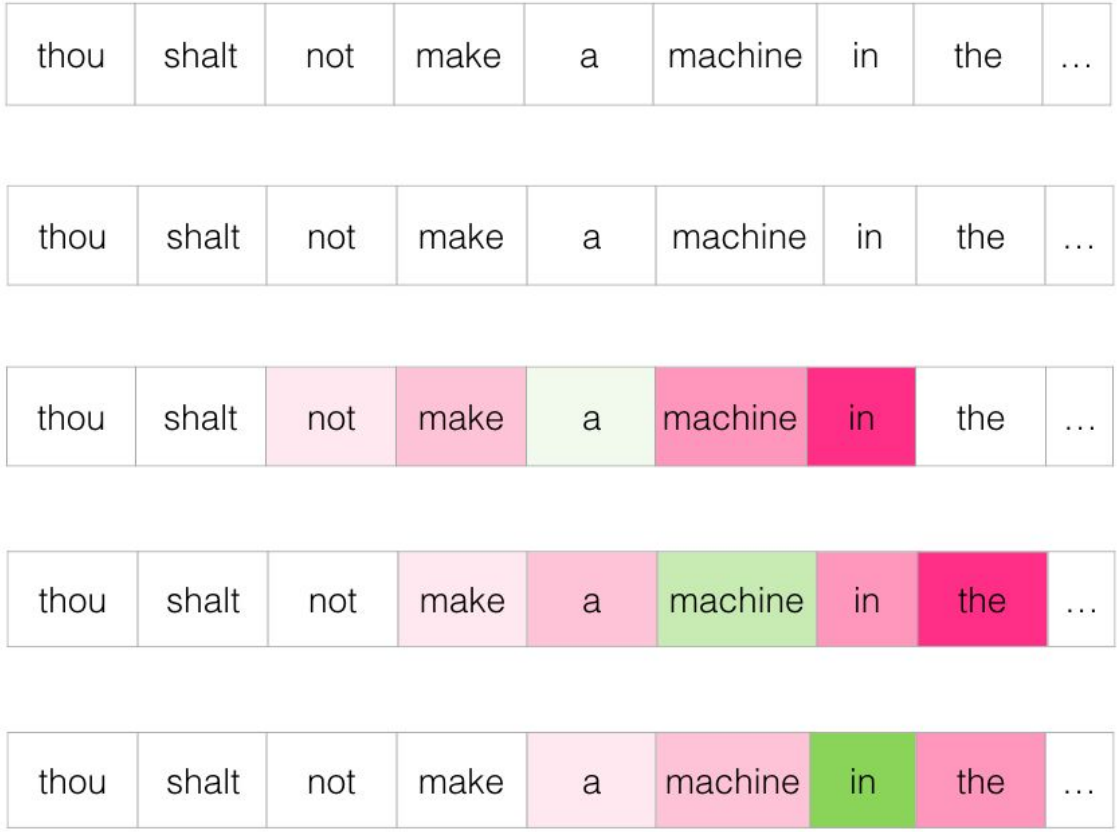
Skip-gram

Thou shalt not make a machine n the likeness of a human mind



input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

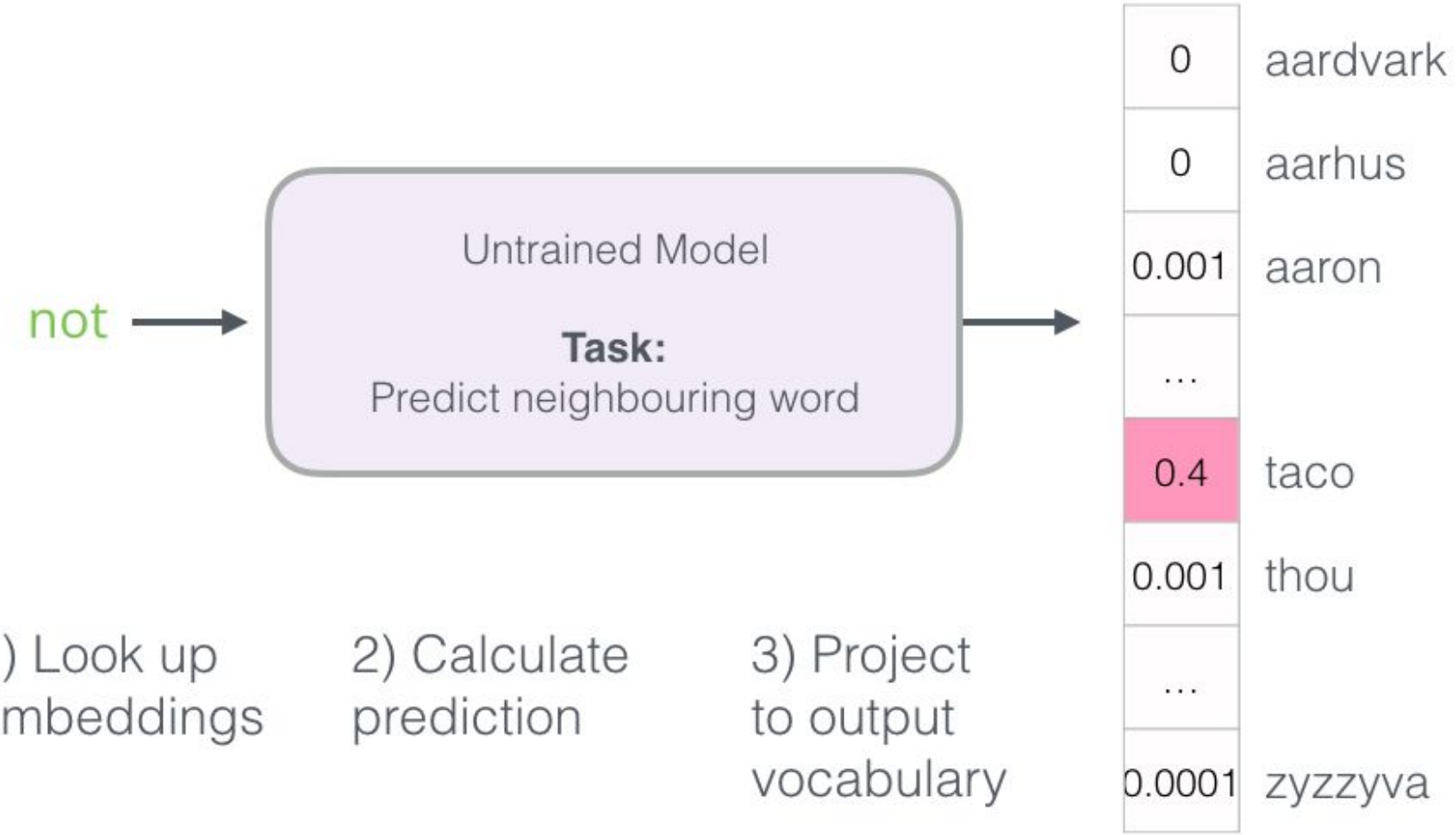
Thou shalt not make a machine in the likeness of a human mind

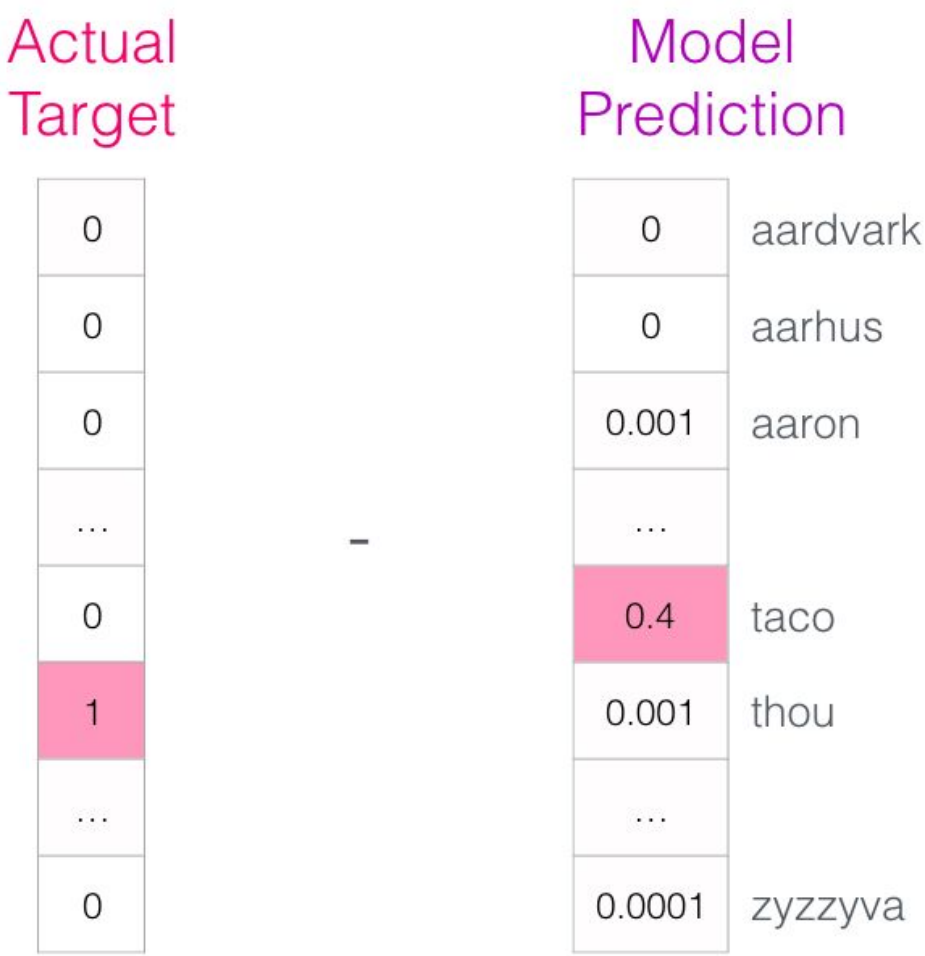


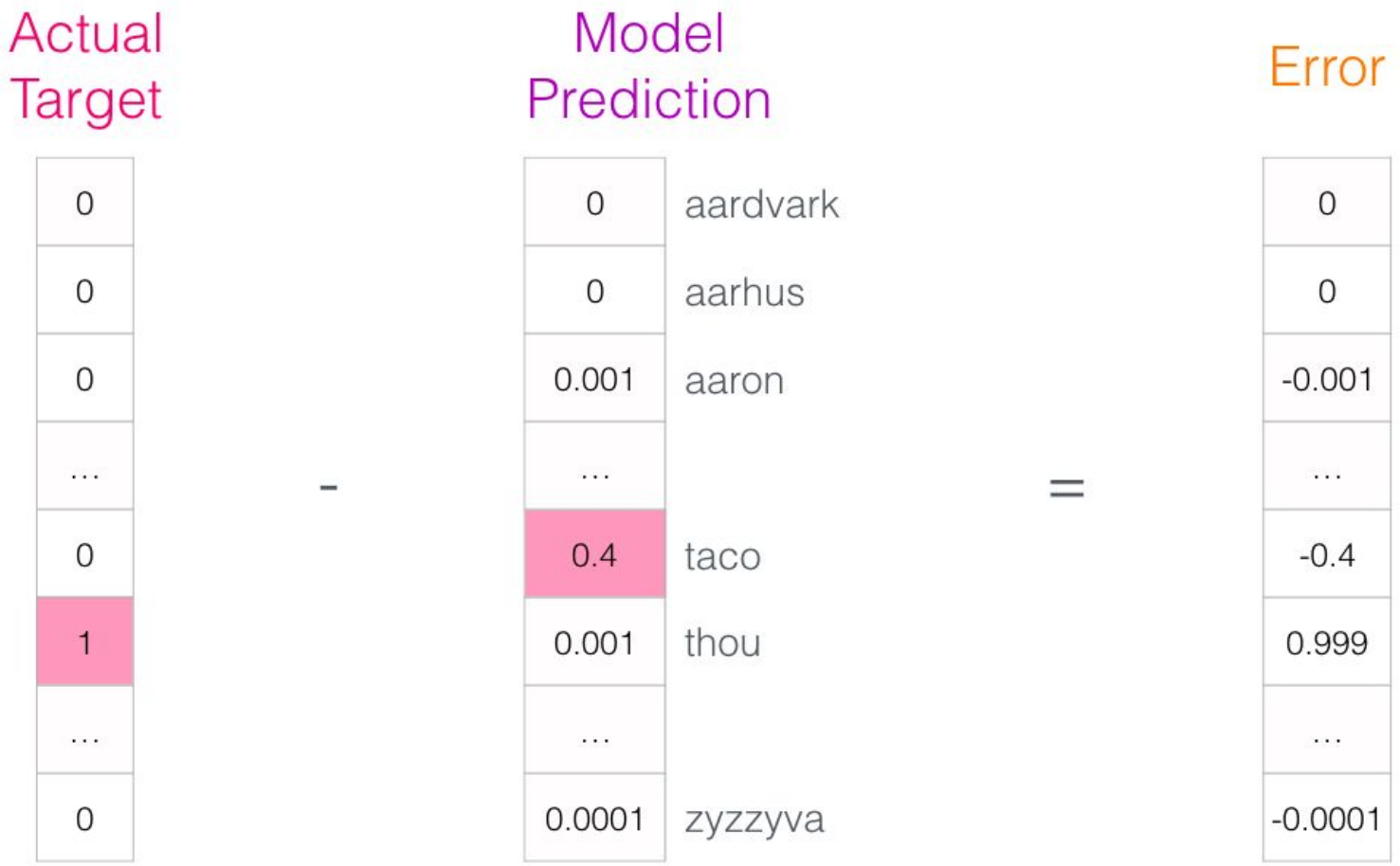
input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

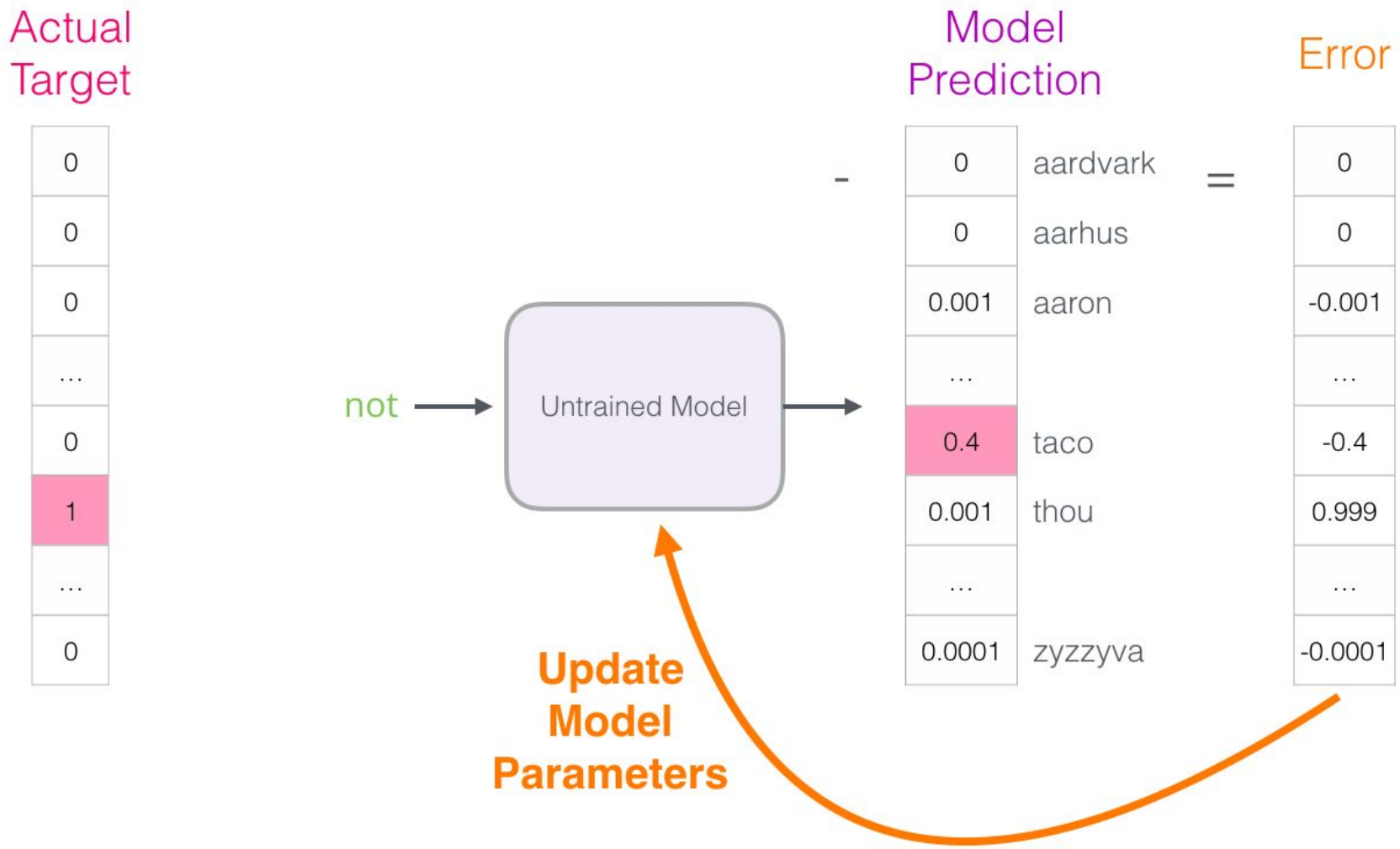
input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness

Предсказываем слова


not
Untrained Model
Task:
Predict neighbouring word
0
aardvark
0
aarhus
0.001
aaron
…
0.4
taco
0.001
thou
…
0.0001
zyzzyva
1) Look up embeddings
2) Calculate prediction
3) Project to output vocabulary


Actual Target
Model Prediction
0
0
0
…
0
1
…
0
-
0
aardvark
0
aarhus
0.001
aaron
…
0.4
taco
0.001
thou
…
0.0001
zyzzyva


Actual Target
Model Prediction
Error
0
0
0
…
0
1
…
0
-
0
aardvark
0
aarhus
0.001
aaron
…
0.4
taco
0.001
thou
…
0.0001
zyzzyva
=
0
0
-0.001
…
-0.4
0.999
…
-0.0001


Actual Target
Model Prediction
Error
0
0
0
…
0
1
…
0
-
0 aardvark
0 aarhus
0.001 aaron
…
0.4 taco
0.001 thou
…
0.0001 zyzzyva
=
0
0
-0.001
…
-0.4
0.999
…
-0.0001
not
Untrained Model
Update Model Parameters

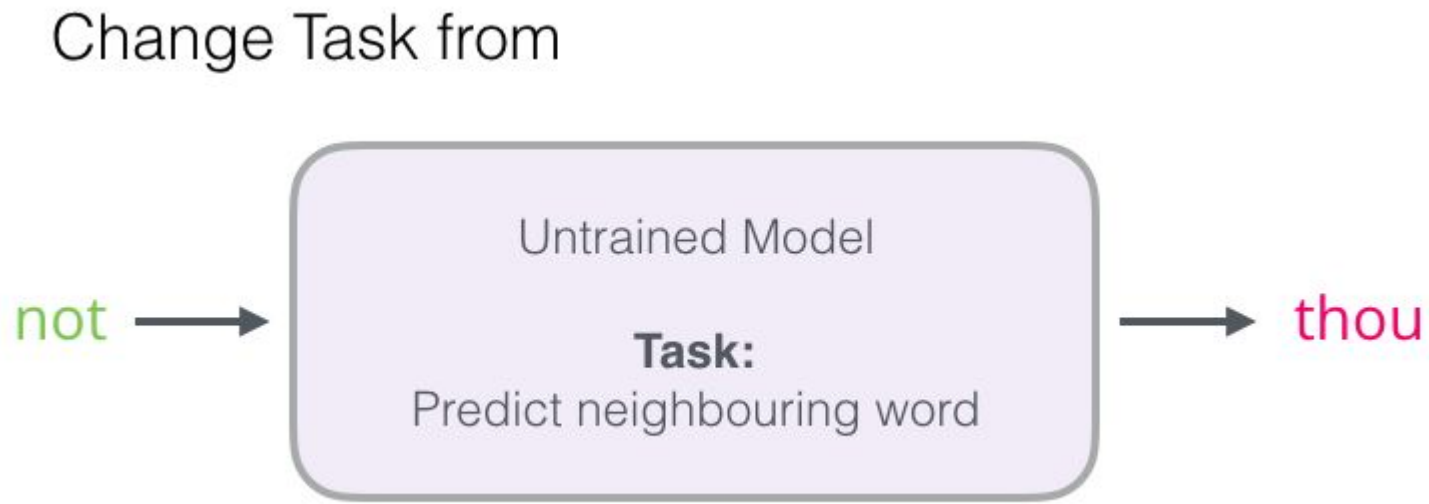
1) Look up embeddings

2) Calculate prediction

3) Project to output vocabulary

[Computationally Intensive]

Вместо этого....



Это:

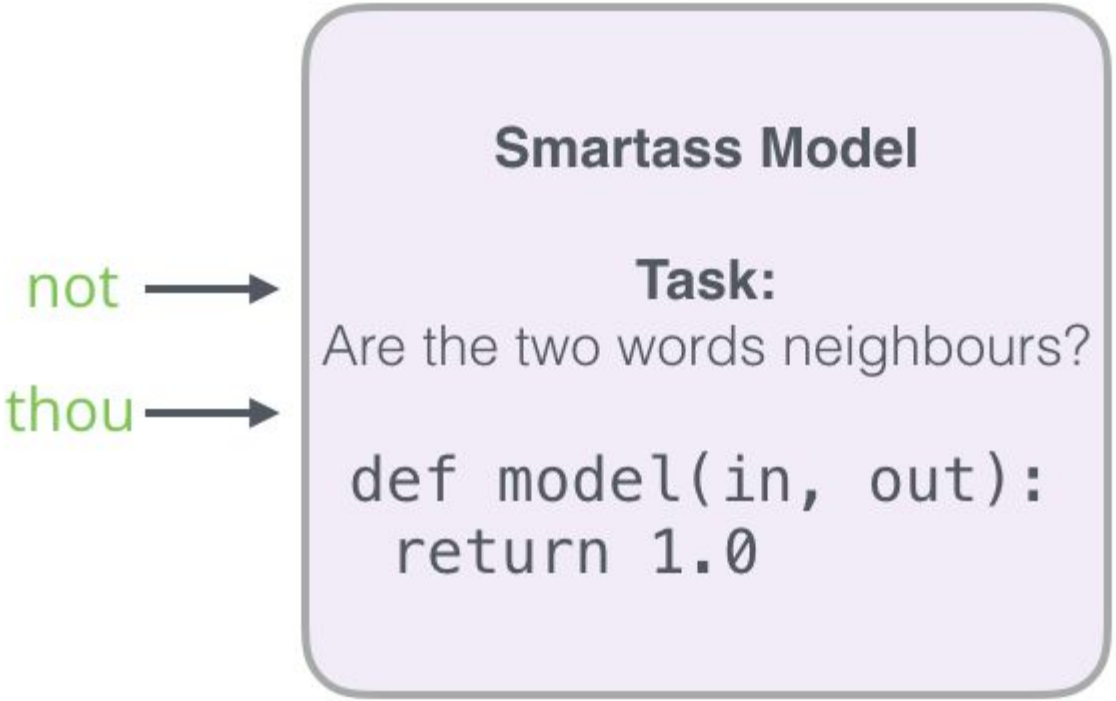
To:


not
thou
Untrained Model
Task:
Are the two words neighbours?
0.90

input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine

input word	output word	target
not	thou	1
not	shalt	1
not	make	1
not	a	1
make	shalt	1
make	not	1
make	a	1
make	machine	1

Проблемка

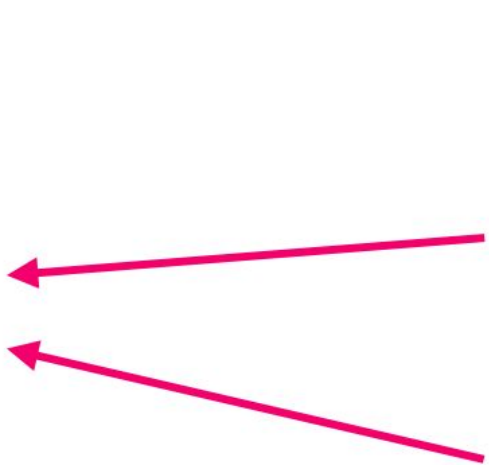

not
thou
Smartass Model
Task:
Are the two words neighbours?
def model(in, out):
return 1.0

input word	output word	target
not	thou	**1**
not		**0**
not		**0**
not	shalt	**1**
not	make	**1**

Negative examples

Pick randomly from vocabulary
(random sampling)

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	make	1



Word	Count	Probability
aardvark		
aarhus		
aaron		
taco		
thou		
zyzzyva		

Ключевые концепты еще раз

Skipgram

shalt	not	make	a	machine

input	output
make	shalt
make	not
make	a
make	machine

Negative Sampling

input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0

Параметры обучения

Параметры обучения

Negative samples: 2

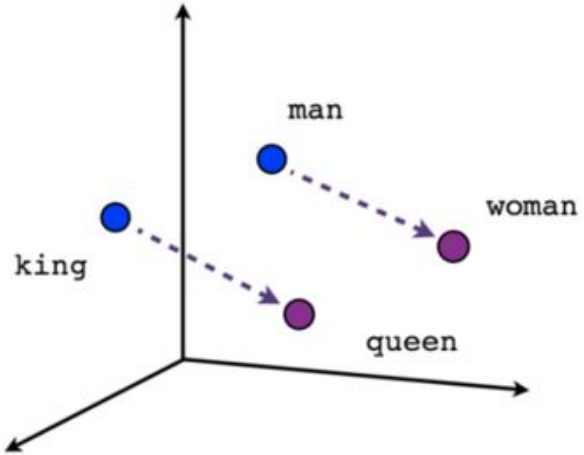
input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0

Negative samples: 5

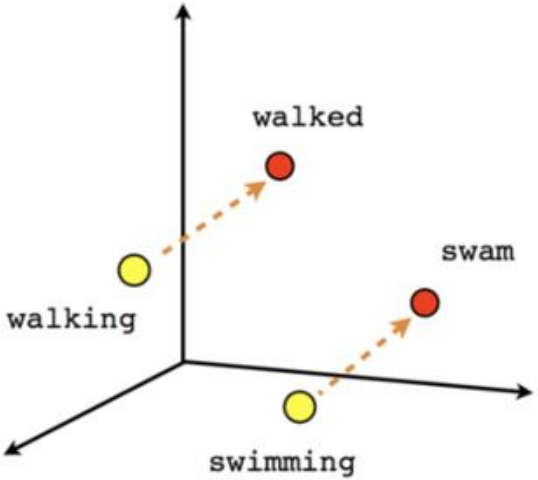
input word	output word	target
make	shalt	1
make	aaron	0
make	taco	0
make	finglonger	0
make	plumbus	0
make	mango	0

RusVectores

https://rusvectores.org/ru/



Male-Female



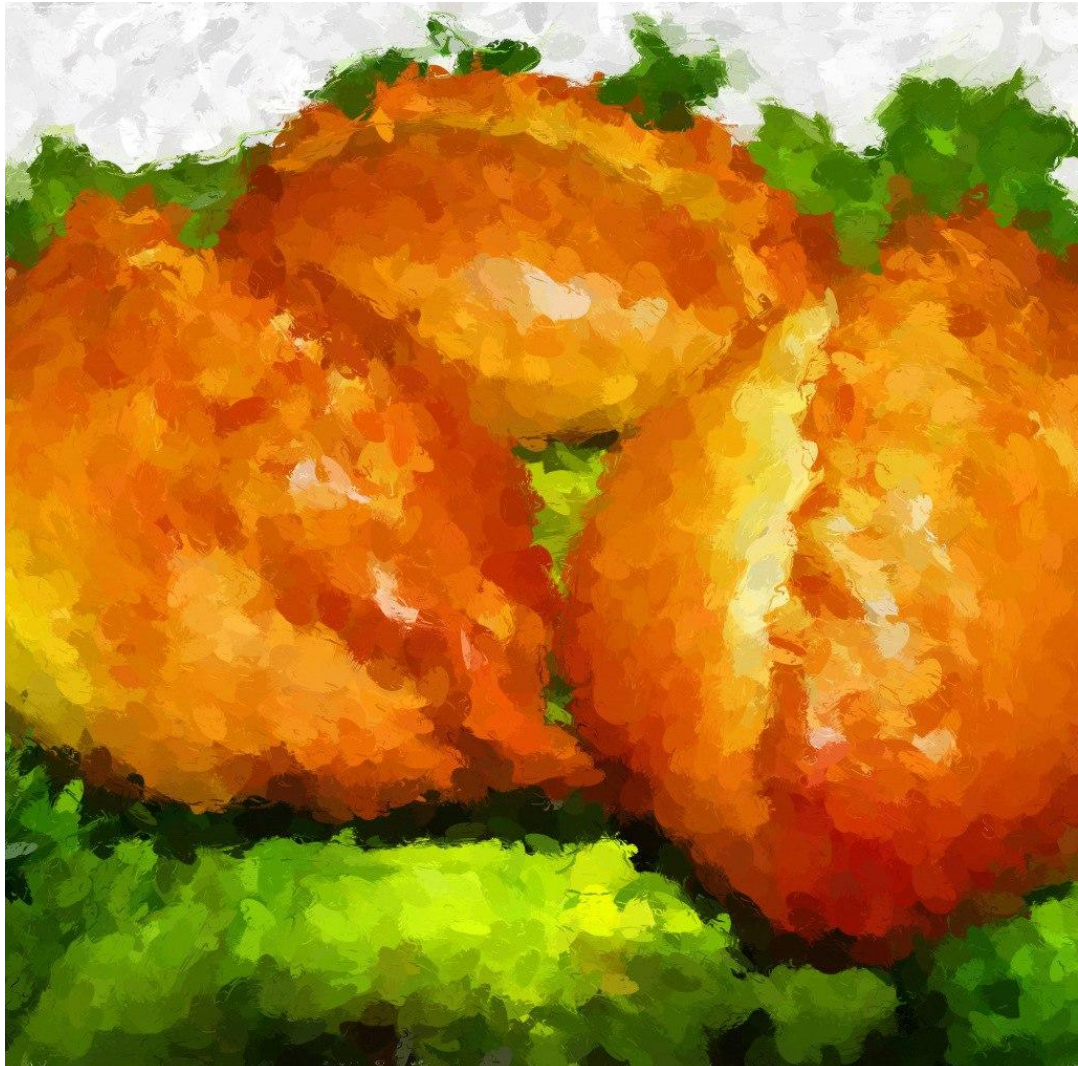
Verb tense

Векторные романы

Ещё давайте посмотрим на векторные романы
https://nevmenandr.github.io/novel2vec/

Пирожки в дистрибутивной семантике

https://habr.com/ru/post/275913/



Применение

- разрешение лексической неоднозначности
- информационный поиск
- кластеризация документов
- машинный перевод
- автоматическое формирование словарей (словарей семантических отношений, двуязычных словарей)
- создание семантических карт
- моделирование перифраз
- определение тематики документа
- определение тональности высказывания

Вопросы к тесту

Спасибо за внимание!

Литература

https://habr.com/ru/post/446530/