

# Лингвистика для математиков

Дистрибутивная семантика

# План на жизнь

На следующей паре тест!

Пользоваться можно будет всем кроме соцсетей и соседей



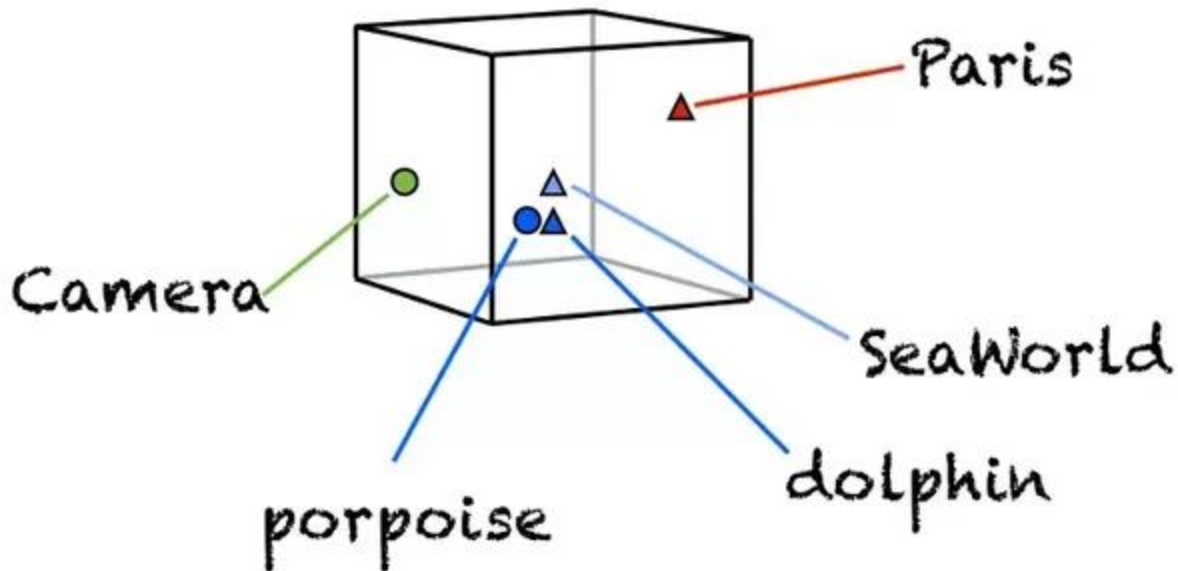
# Дистрибутивная семантика

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

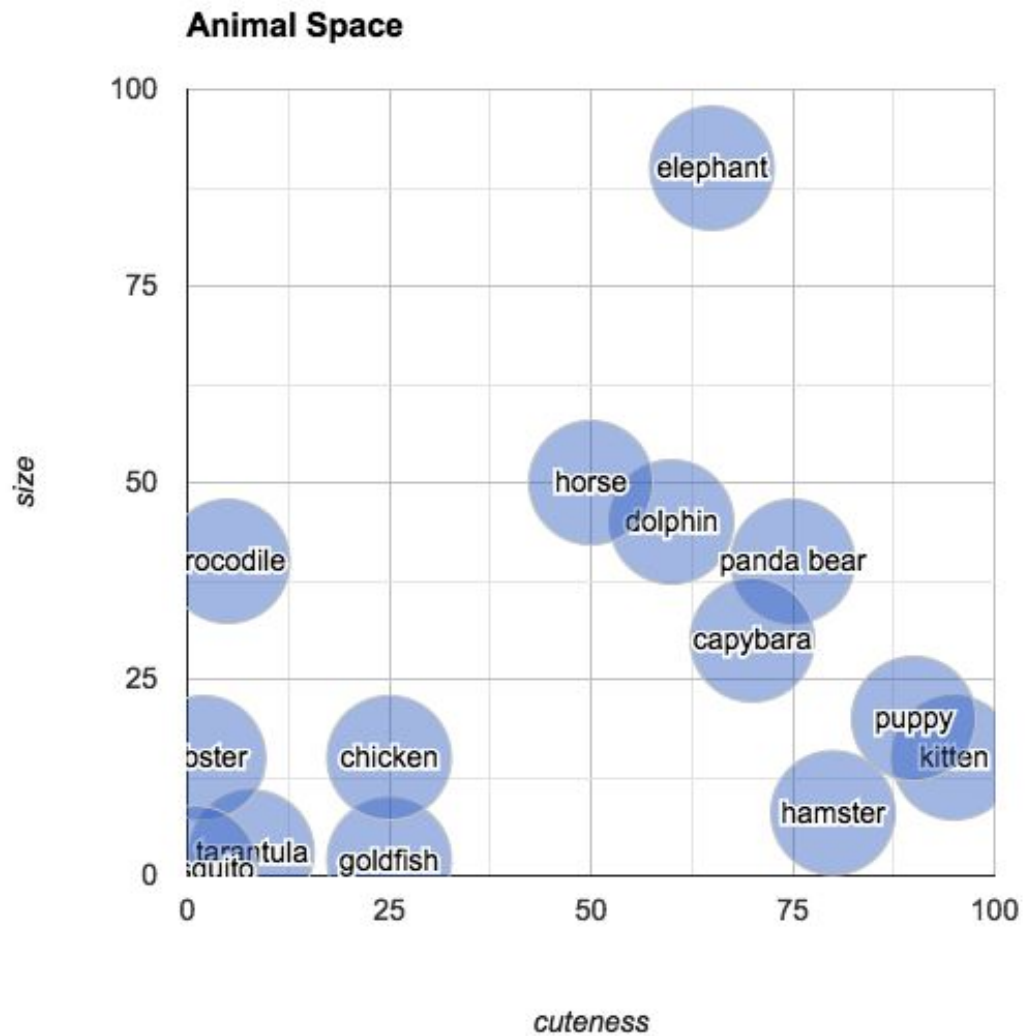
Иначе: word embeddings

# 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



# Аналогии

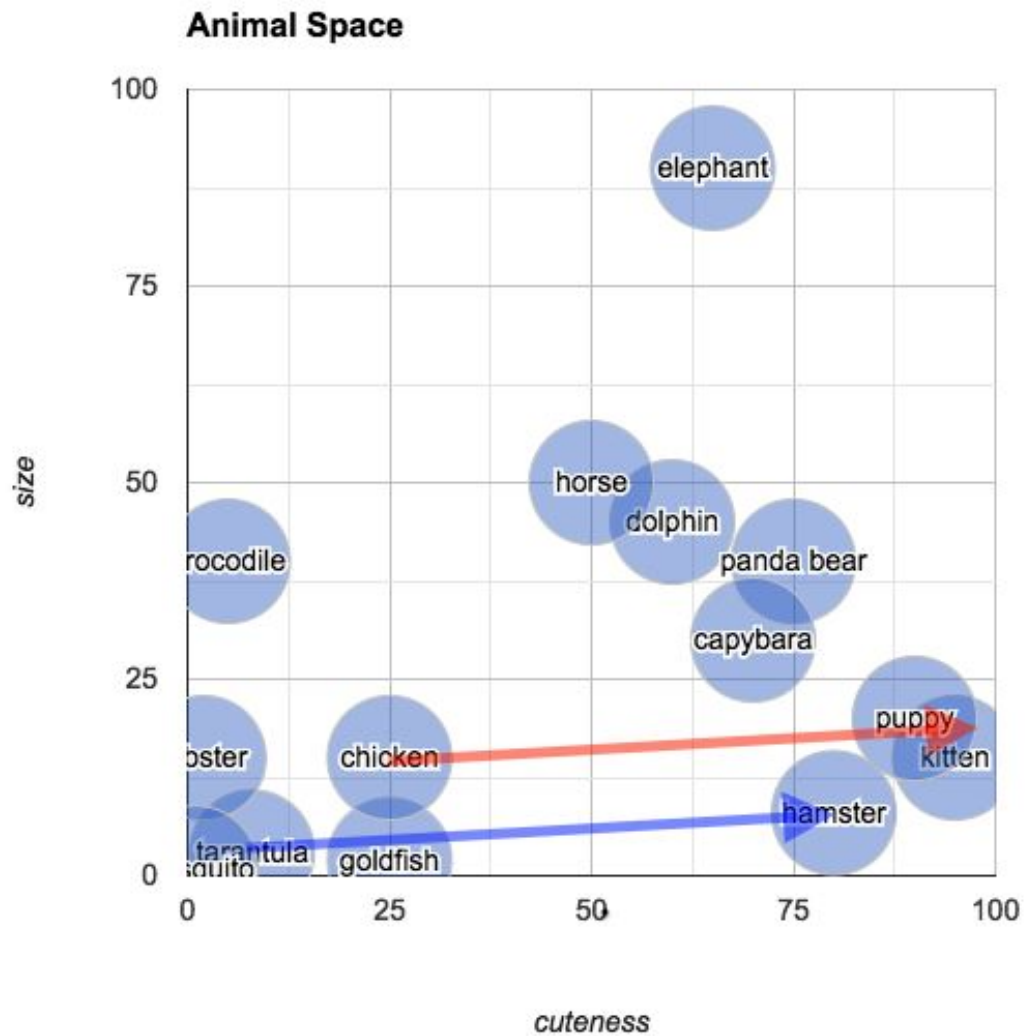
Следующий шаг - Расстояние!

С животными мы исследуем Евклидово расстояние:

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



# Аналогии



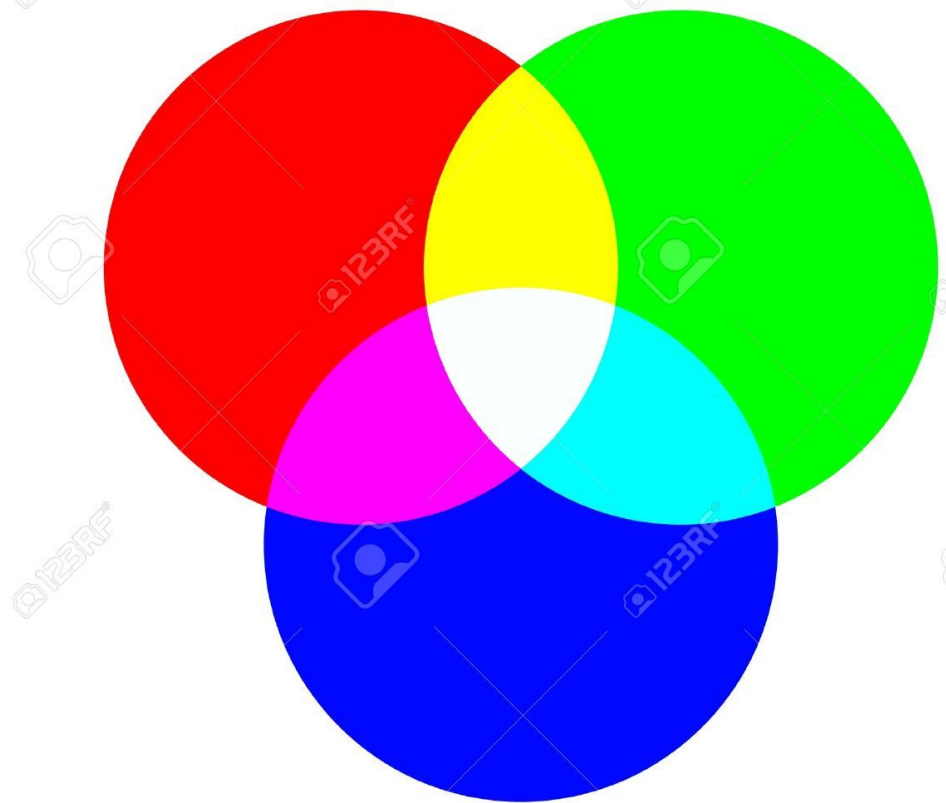
# Аналогии

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

Extraversion

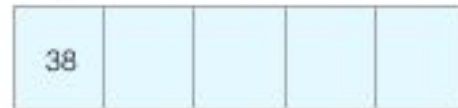
100

0

Introversion



Jay



# Аналогии

Extraversion

1

0

-1

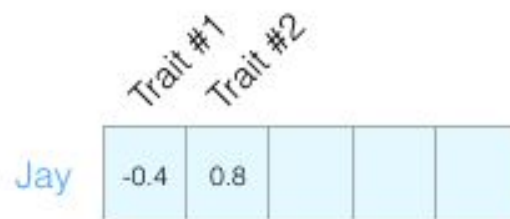
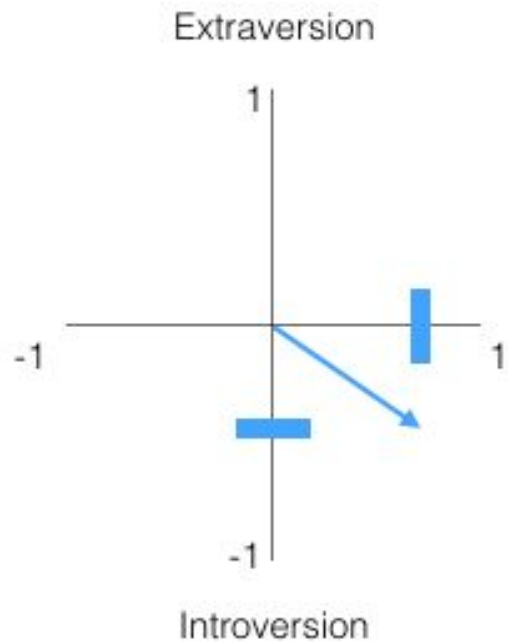
Introversion



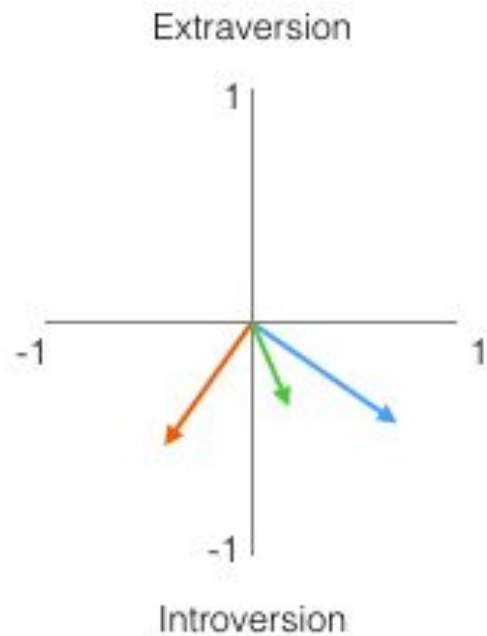
Jay



# Аналогии



# Аналогии



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

# Аналогии

$$\text{cosine\_similarity}\left(\begin{array}{c|c} \text{Jay} & \\ \hline -0.4 & 0.8 \end{array}, \begin{array}{c|c} \text{Person \#1} & \\ \hline -0.3 & 0.2 \end{array}\right) = 0.87 \quad \checkmark$$

$$\text{cosine\_similarity}\left(\begin{array}{c|c} \text{Jay} & \\ \hline -0.4 & 0.8 \end{array}, \begin{array}{c|c} \text{Person \#2} & \\ \hline -0.5 & -0.4 \end{array}\right) = -0.20$$



# Аналогии

Jay

Trait #1	Trait #2	Trait #3	Trait #4	Trait #5
-0.4	0.8	0.5	-0.2	0.3

Person #1

-0.3	0.2	0.3	-0.4	0.9
------	-----	-----	------	-----

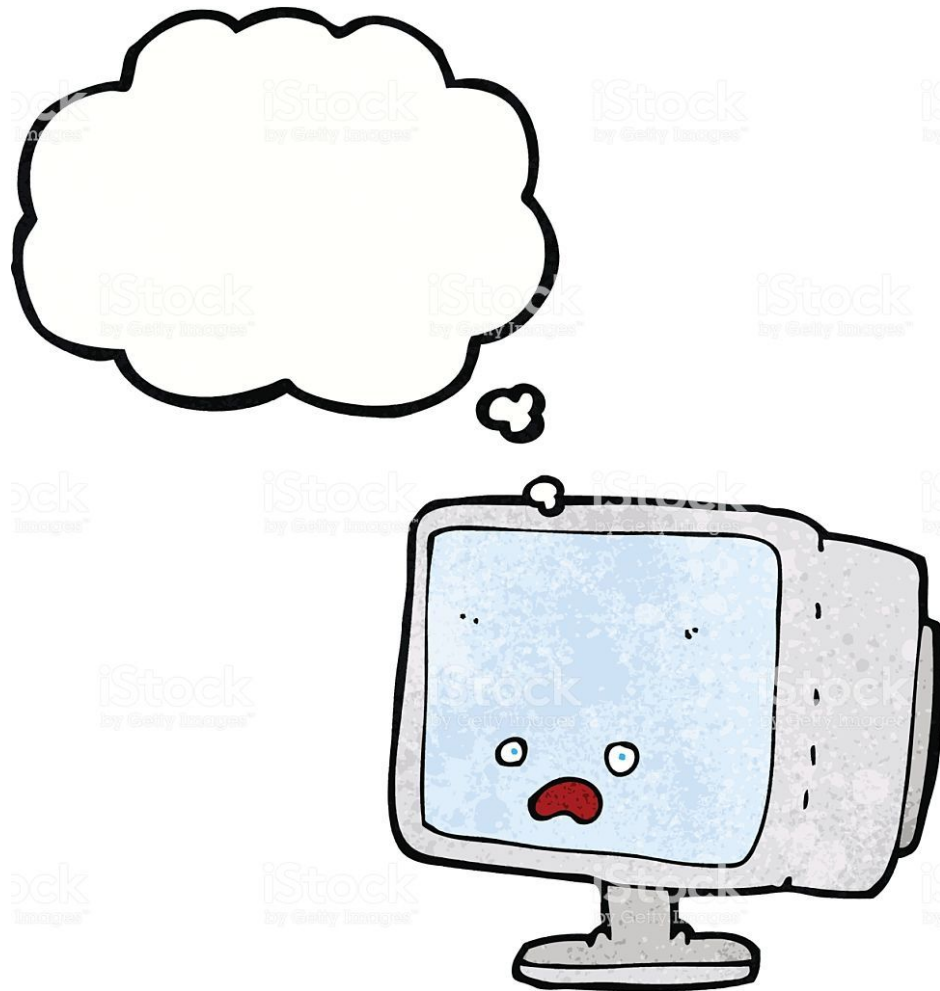
Person #2

-0.5	-0.4			
------	------	--	--	--

$$\text{cosine\_similarity}(\begin{matrix} \text{Jay} \\ \begin{bmatrix} -0.4 & 0.8 & 0.5 & -0.2 & 0.3 \end{bmatrix} \end{matrix}, \begin{matrix} \text{Person \#1} \\ \begin{bmatrix} -0.3 & 0.2 & 0.3 & -0.4 & 0.9 \end{bmatrix} \end{matrix}) = 0.66 \quad \checkmark$$

$$\text{cosine\_similarity}(\begin{matrix} \text{Jay} \\ \begin{bmatrix} -0.4 & 0.8 & 0.5 & -0.2 & 0.3 \end{bmatrix} \end{matrix}, \begin{matrix} \text{Person \#2} \\ \begin{bmatrix} -0.5 & -0.4 & -0.2 & 0.7 & -0.1 \end{bmatrix} \end{matrix}) = -0.37$$

Наивный подход



# Наивный подход

Разные виды близости слов: лексическая и семантическая

- петух
- курица
- цыпленок





# Векторное представление слов

One-hot encoding:

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND  
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

# 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  $\Rightarrow [0, 1, 0, 0, 1, 1, 0, 1]^T$

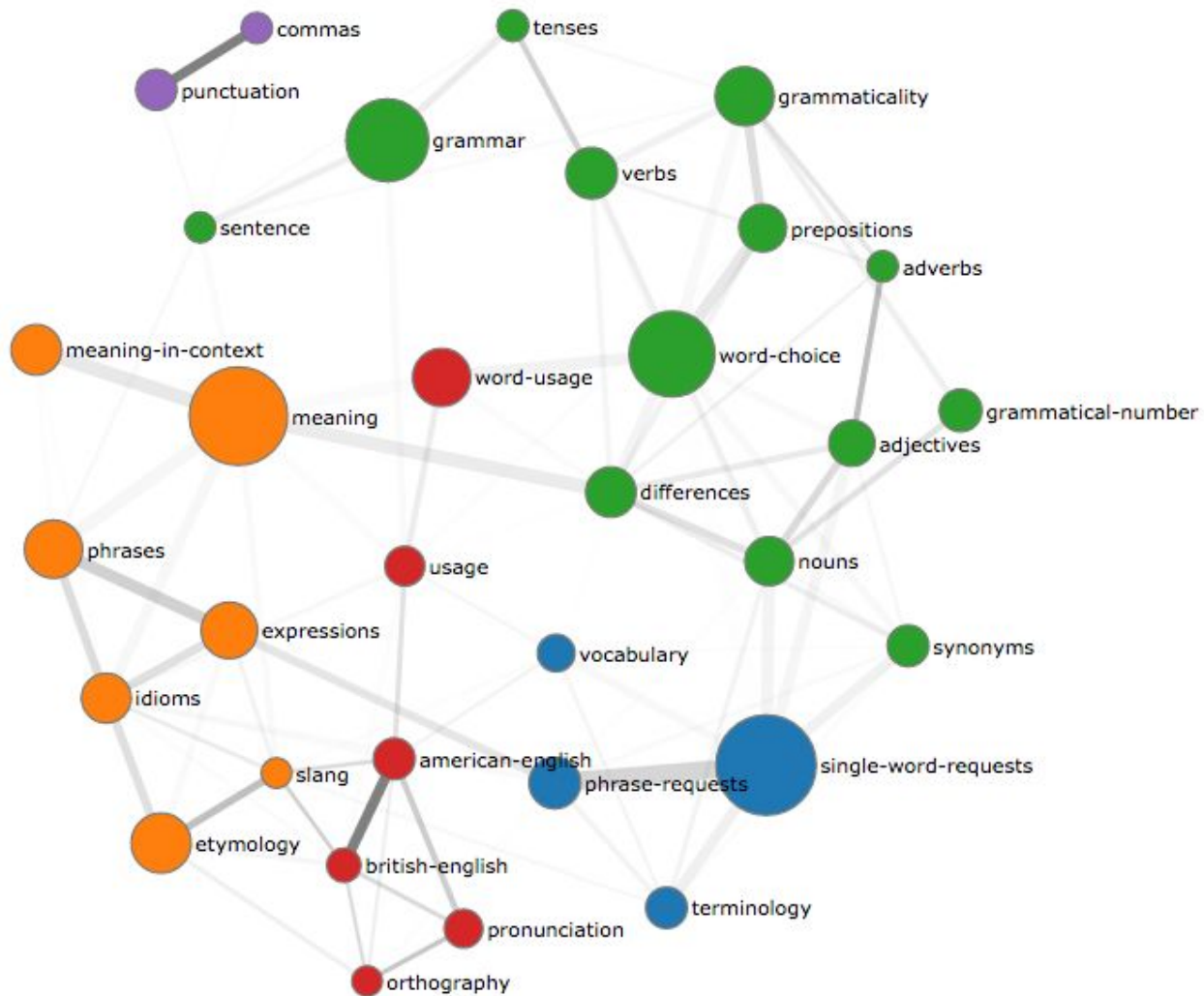
cat: cute, furry, miaowed  $\Rightarrow [0, 1, 1, 0, 1, 0, 0, 0]^T$

dog: loud, furry, ran, bit  $\Rightarrow [1, 0, 1, 1, 0, 0, 1, 0]^T$

$$\text{sim}(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{\|w_1\| \cdot \|w_2\|}$$

Как обучить  
компьютер  
отделять  
слова?

**word2vec**

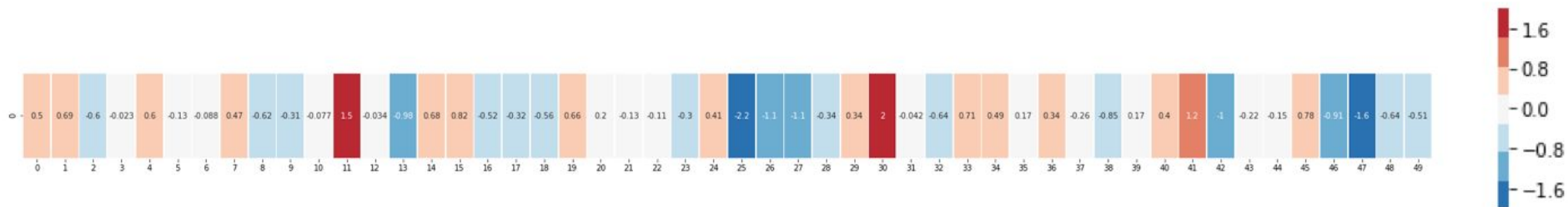


# Embeddings

```
[ 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

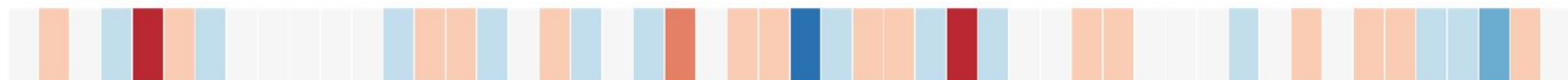


# Embeddings

“king”

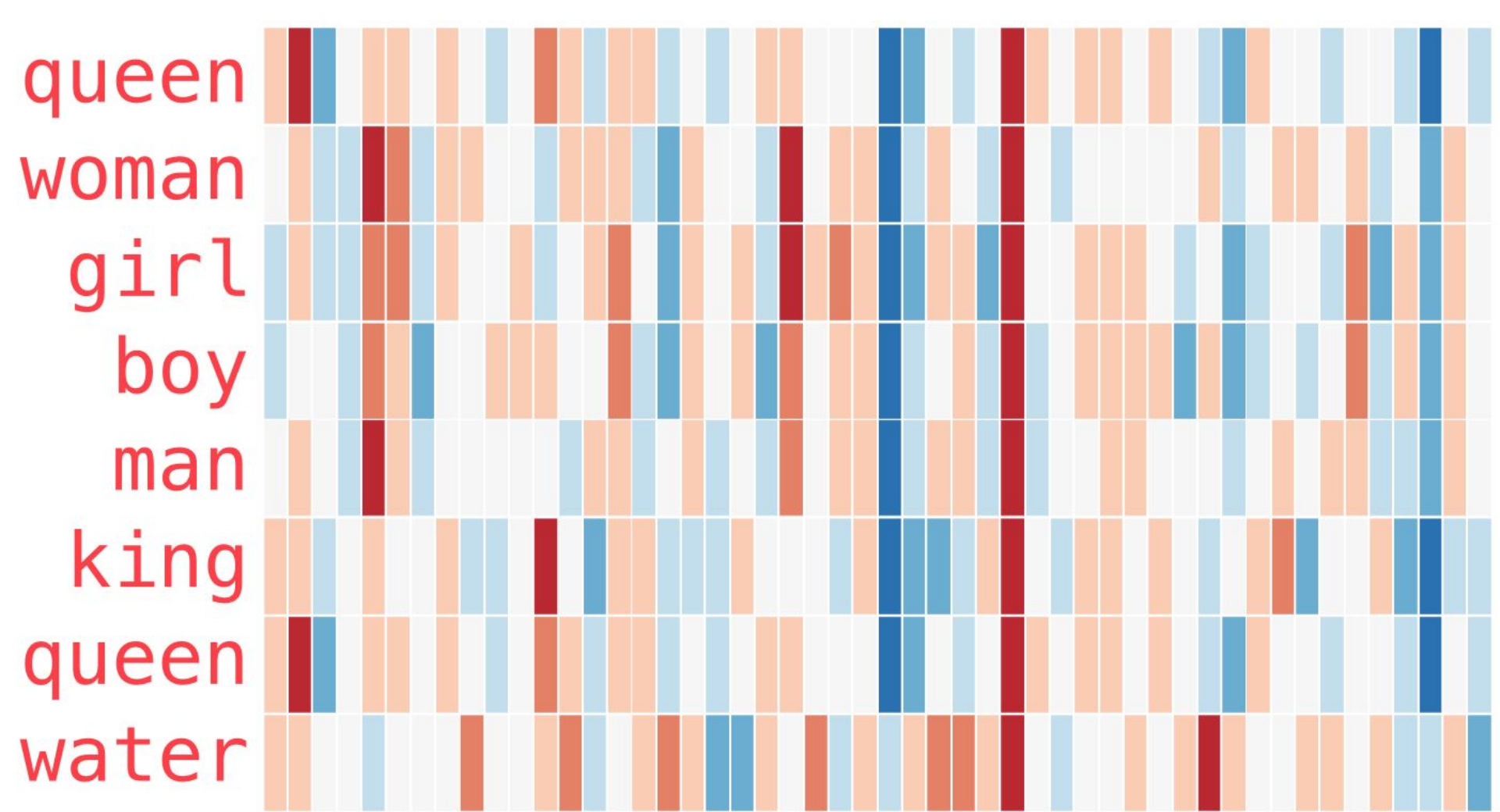


“Man”



“Woman”



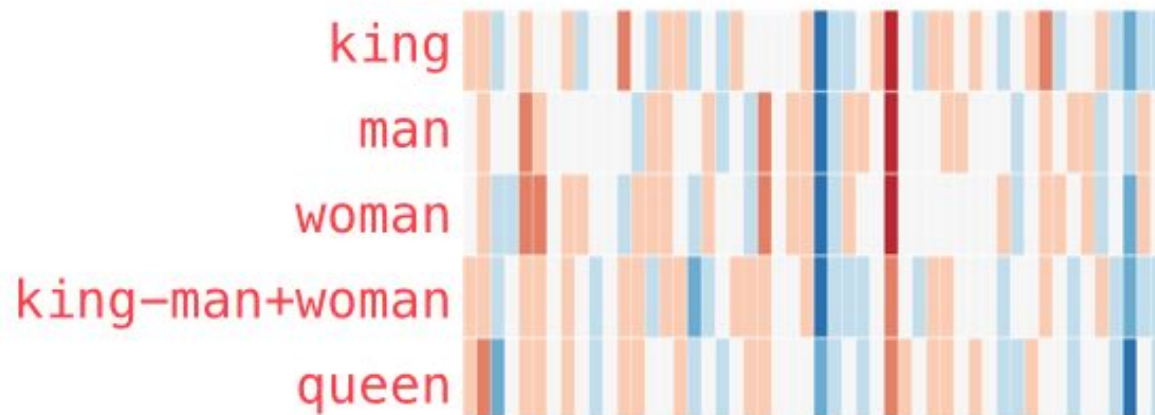


# Embeddings

Вновь про понятие аналогий:

Кот соотносится с котенком так же, как курица с цыпленком

king - man + woman  $\approx$  queen



# Нейронная модель языка



input/feature #1

input/feature #2

output/label

Thou shalt

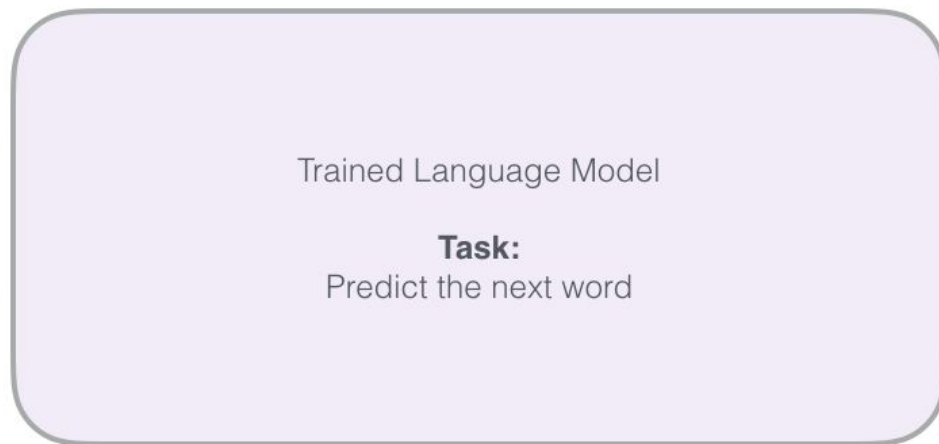


# Нейронная модель языка

Input  
Features

Thou →

shalt →



Output  
Prediction

0%	aardvark
0%	aarhus
0.1%	aaron
...	
40%	not
...	
0.01	zyzzyva

# Нейронная модель языка

Input  
Features

Trained Language Model

Output  
Prediction

**Task:**  
Predict the next word

Thou →

shalt →

1) Look up  
embeddings

2) Calculate  
prediction

3) Project  
to output  
vocabulary

0	aardvark
0	aarhus
0.001	aaron
...	...
0.4	not
...	...
0.0001	zyzzyva

# Нейронная модель языка

Input

Features

Thou  
shalt

Trained Language Model

**Task:**

Predict the next word

1) Look up embeddings

				aardvark
				...
				...
				shalt
				...
				thou
				...
				zyzzyva

				thou
				shalt

Output

Prediction

0	aardvark
0	aarhus
0.001	aaron
...	...
0.4	not
...	...
0.0001	zyzzyva



# 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...**



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

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

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

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

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

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

thou	shalt	not	make	a	machine	in	the	...
------	-------	-----	------	---	---------	----	-----	-----

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



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	...
------	-------	-----	------	---	---------	----	-----	-----

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

not →



Предсказываем слова

not



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

0
0
0
...
0
1
...
0

-

## Model Prediction

0	aardvark
0	aarhus
0.001	aaron
...	
0.4	taco
0.001	thou
...	
0.0001	zyzzyva

# Actual Target

0
0
0
...
0
1
...
0

-

# Model Prediction

0	aardvark
0	aarhus
0.001	aaron
...	
0.4	taco
0.001	thou
...	
0.0001	zyzzyva

=

# Error

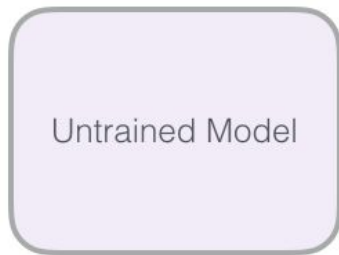
0
0
-0.001
...
-0.4
0.999
...
-0.0001



Actual  
Target

0
0
0
...
0
1
...
0

not



Model  
Prediction

0	aardvark	=	0
0	aarhus		0
0.001	aaron		-0.001
...			...
0.4	taco		-0.4
0.001	thou		0.999
...			...
0.0001	zyzzyva		-0.0001

Error

Update  
Model  
Parameters

not



1) Look up embeddings

2) Calculate prediction

**3) Project to output vocabulary**

**[Computationally Intensive]**

# Вместо этого....

Change Task from



ЭТО:

To:

not



thou



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	<b>1</b>
not	shalt	<b>1</b>
not	make	<b>1</b>
not	a	<b>1</b>
make	shalt	<b>1</b>
make	not	<b>1</b>
make	a	<b>1</b>
make	machine	<b>1</b>

# Проблемка

## Smartass Model

### Task:

Are the two words neighbours?

```
def model(in, out):  
    return 1.0
```

not →

thou →

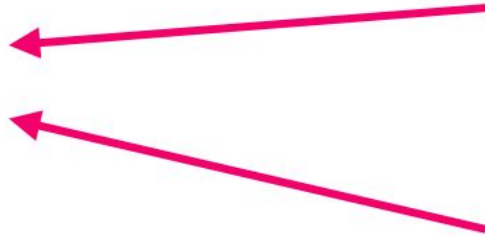
input word	output word	target
not	thou	<b>1</b>
not		<b>0</b>
not		<b>0</b>
not	shalt	<b>1</b>
not	make	<b>1</b>

 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	<b>1</b>
make	aaron	<b>0</b>
make	taco	<b>0</b>

# Параметры обучения

Window size: 5



Window size: 15



# Параметры обучения

Negative samples: 2

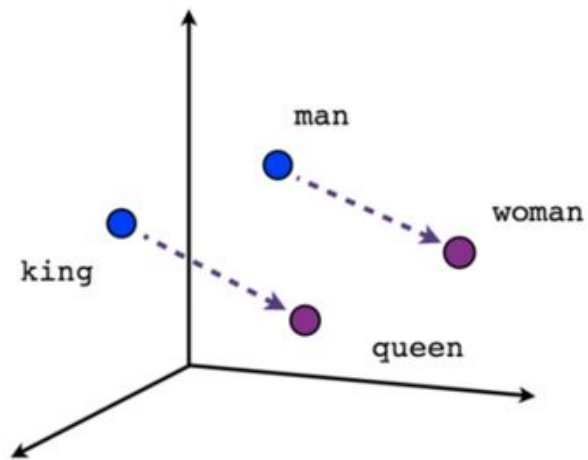
input word	output word	target
make	shalt	<b>1</b>
make	aaron	<b>0</b>
make	taco	<b>0</b>

Negative samples: 5

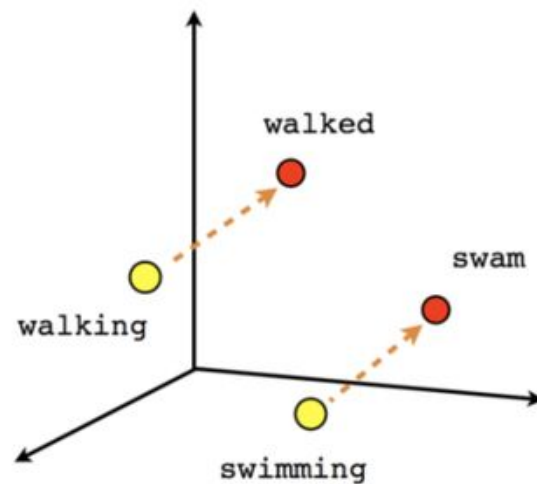
input word	output word	target
make	shalt	<b>1</b>
make	aaron	<b>0</b>
make	taco	<b>0</b>
make	finlonger	<b>0</b>
make	plumbus	<b>0</b>
make	mango	<b>0</b>

# RusVectores

<https://rusvectores.org/ru/>



Male-Female



Verb tense

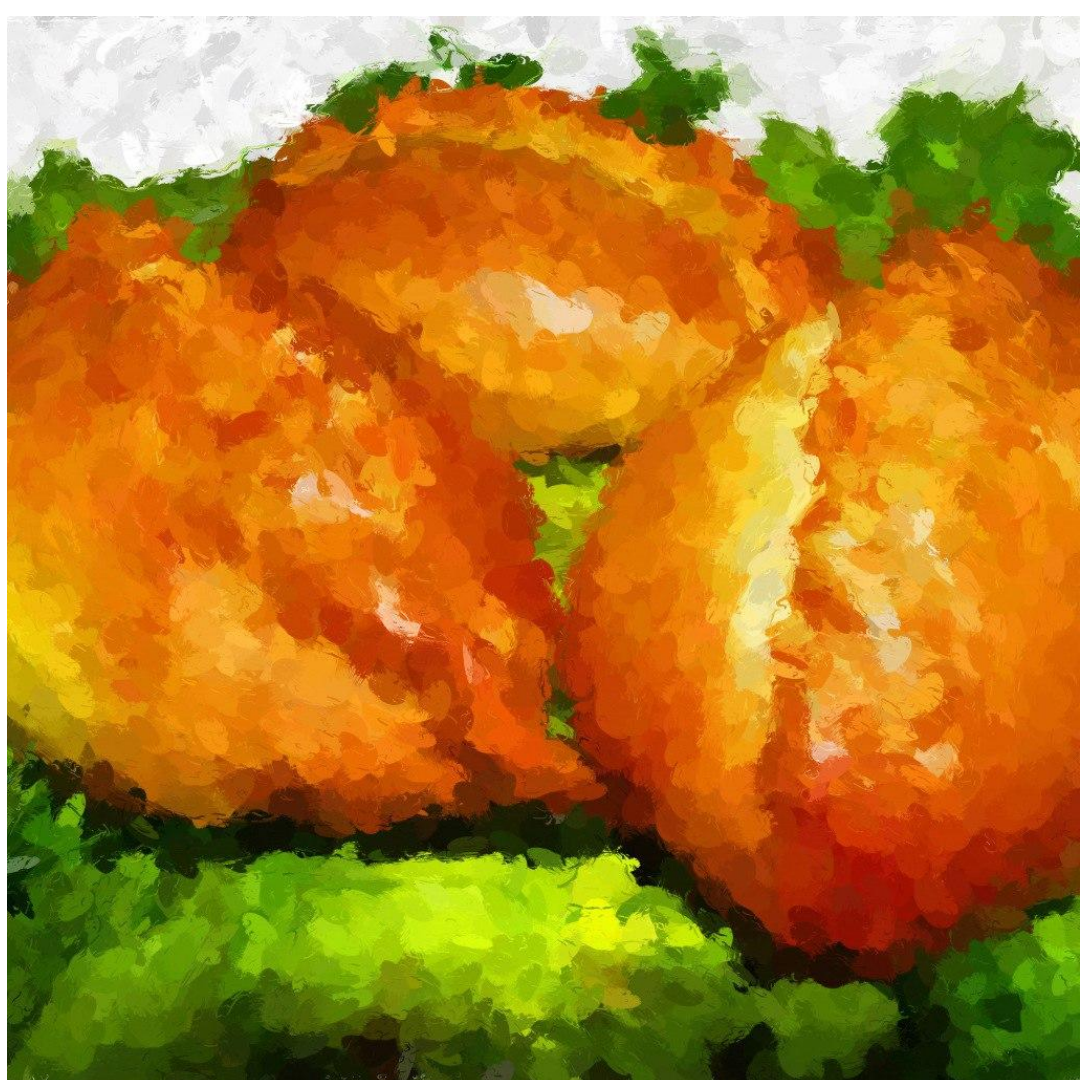
# Векторные романы

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/>