WORD2VEC (PART 2)

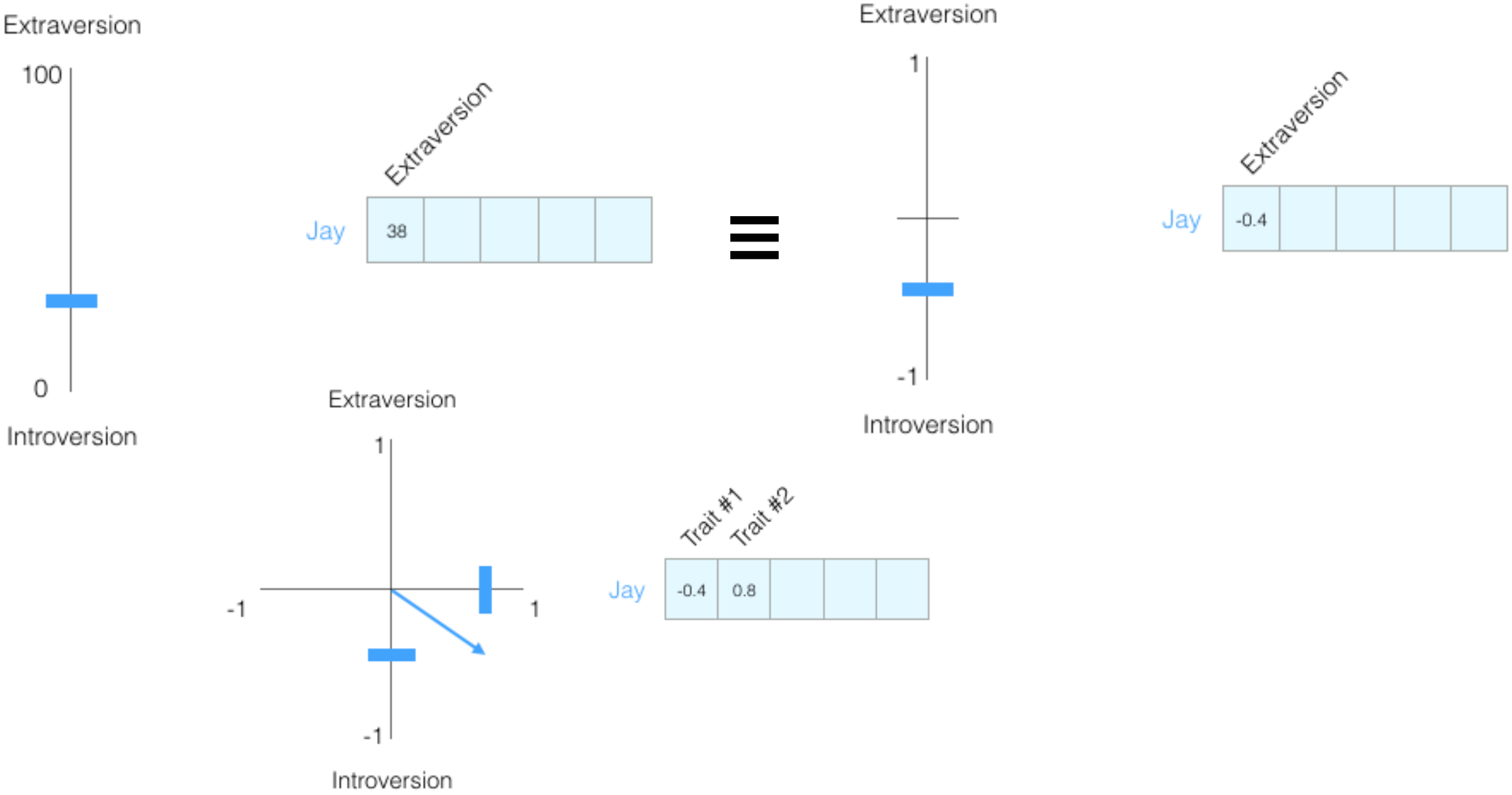
~~Совершенно секретно~~

ВСТРАИВАНИЕ ЛИЧНОСТЕЙ

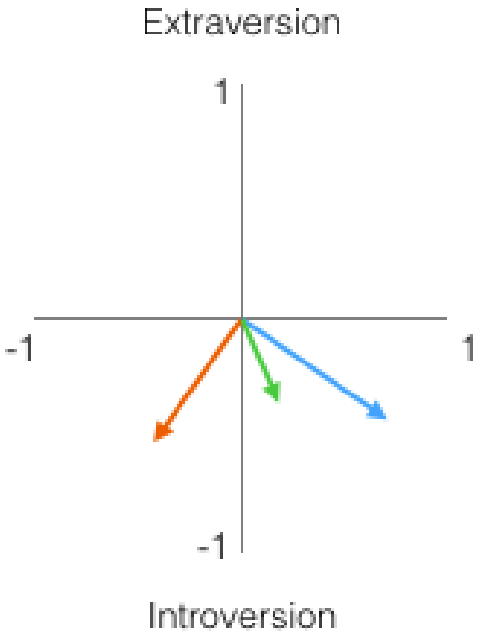
Openness to experience	79 out of 100
Agreeableness	75 out of 100
Conscientiousness	42 out of 100
Negative emotionality	50 out of 100
Extraversion	58 out of 100

https://projects.fivethirtyeight.com/personality-quiz/

ВСТРАИВАНИЕ ЛИЧНОСТЕЙ



ВСТРАИВАНИЕ ЛИЧНОСТЕЙ



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

	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	-0.2	0.7	-0.1

```
cosine_similarity(
```
Jay: [-0.4, 0.8] , Person #1: [-0.3, 0.2]) = 0.87 ✓

```
cosine_similarity(
```
Jay: [-0.4, 0.8] , Person #2: [-0.5, -0.4]) = -0.20

ВСТРАИВАНИЕ ЛИЧНОСТЕЙ

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

1- We can represent things (and people) as vectors of numbers
(Which is great for machines!)

Jay	-0.4	0.8	0.5	-0.2	0.3

2- We can easily calculate how similar vectors are to each other

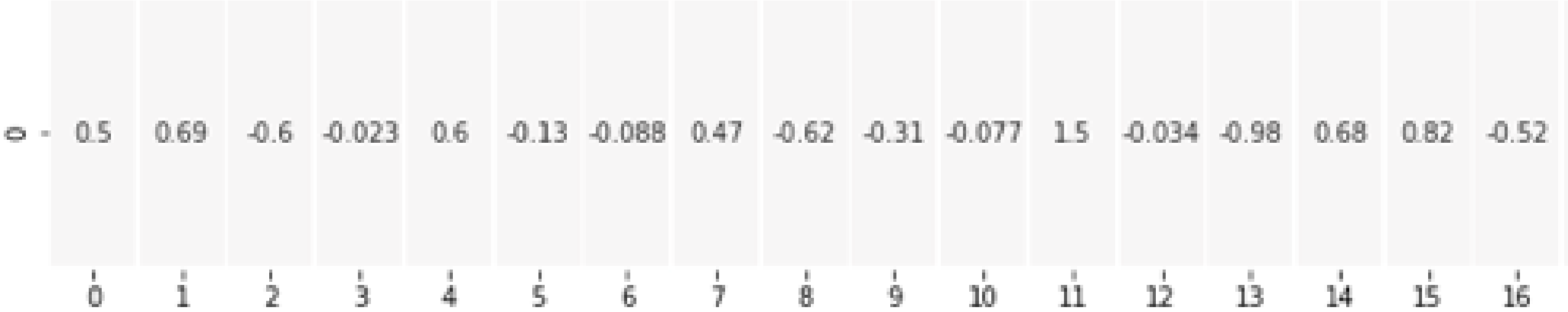
The people most similar to Jay are:

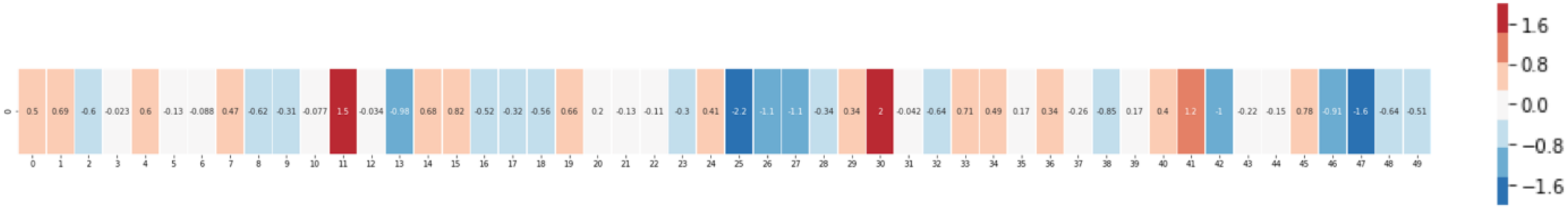
	cosine_similarity ▼
Person #1	0.86
Person #2	0.5
Person #3	-0.20

ВСТРАИВАНИЕ СЛОВ

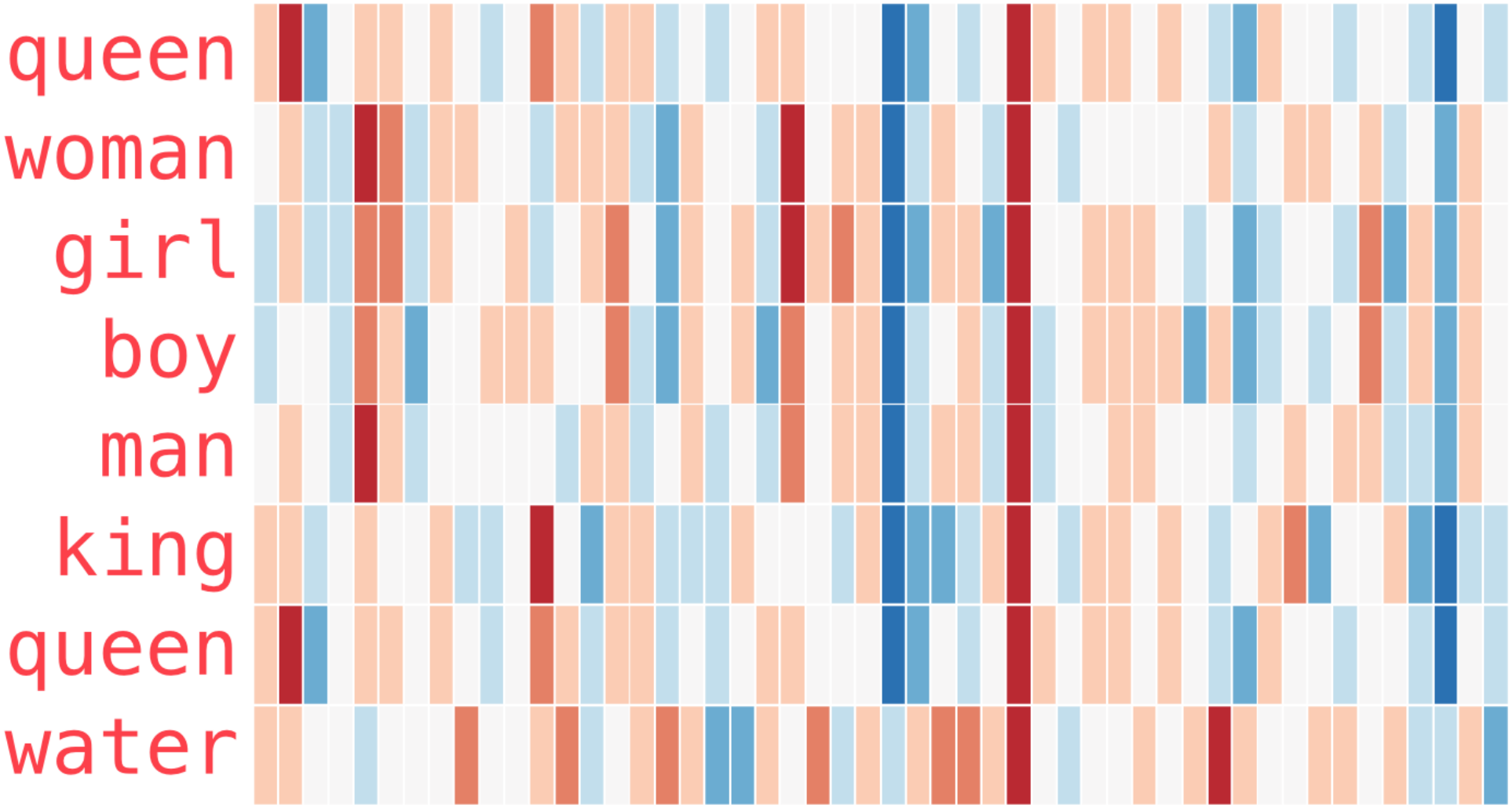
GloVe

[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





ВСТРАИВАНИЕ СЛОВ



АНАЛОГИИ

king – man + woman ~= queen

НЕЙРОННАЯ МОДЕЛЬ ЯЗЫКА

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

Thou shalt _____

Input

Features

Output

Prediction

Thou →

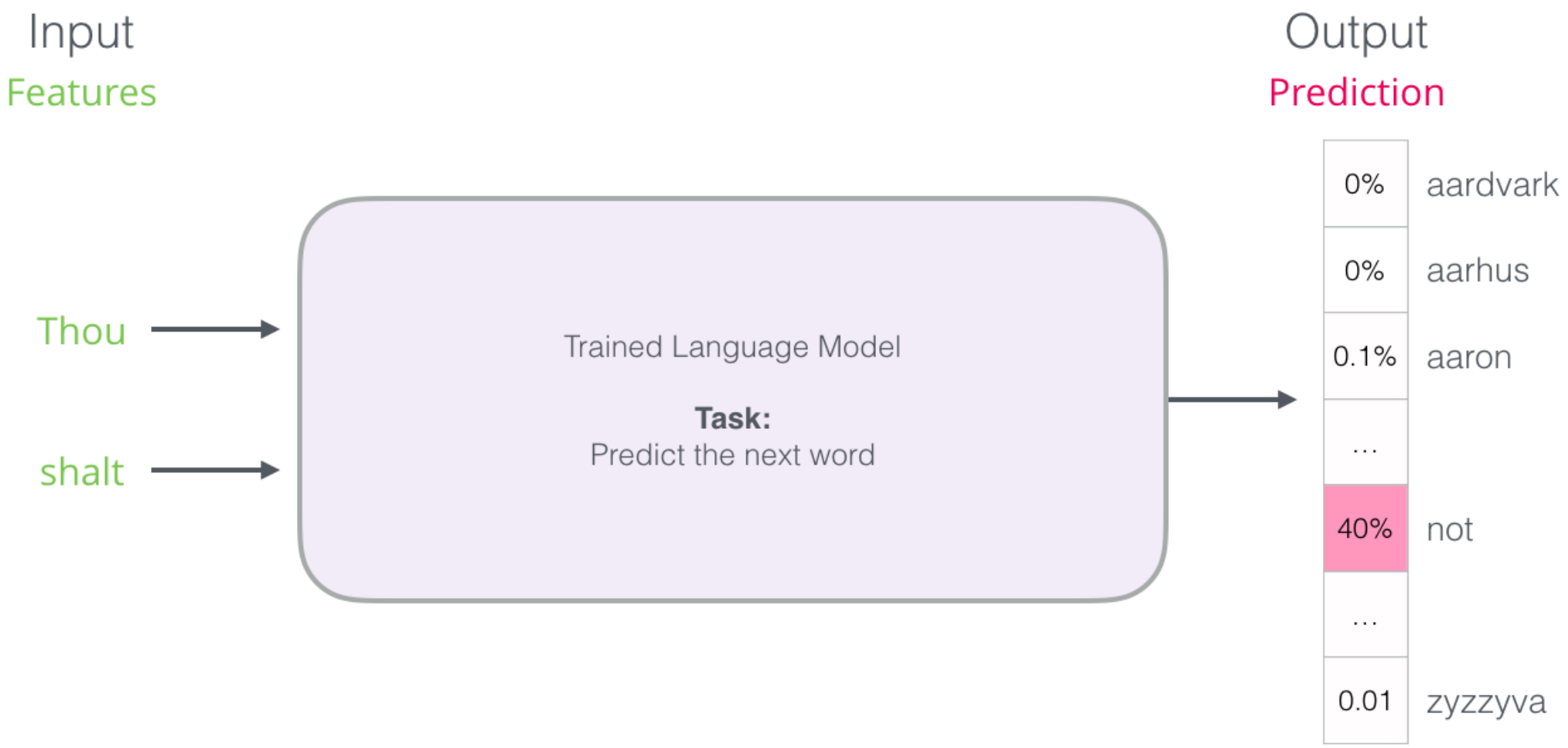
shalt →

Trained Language Model

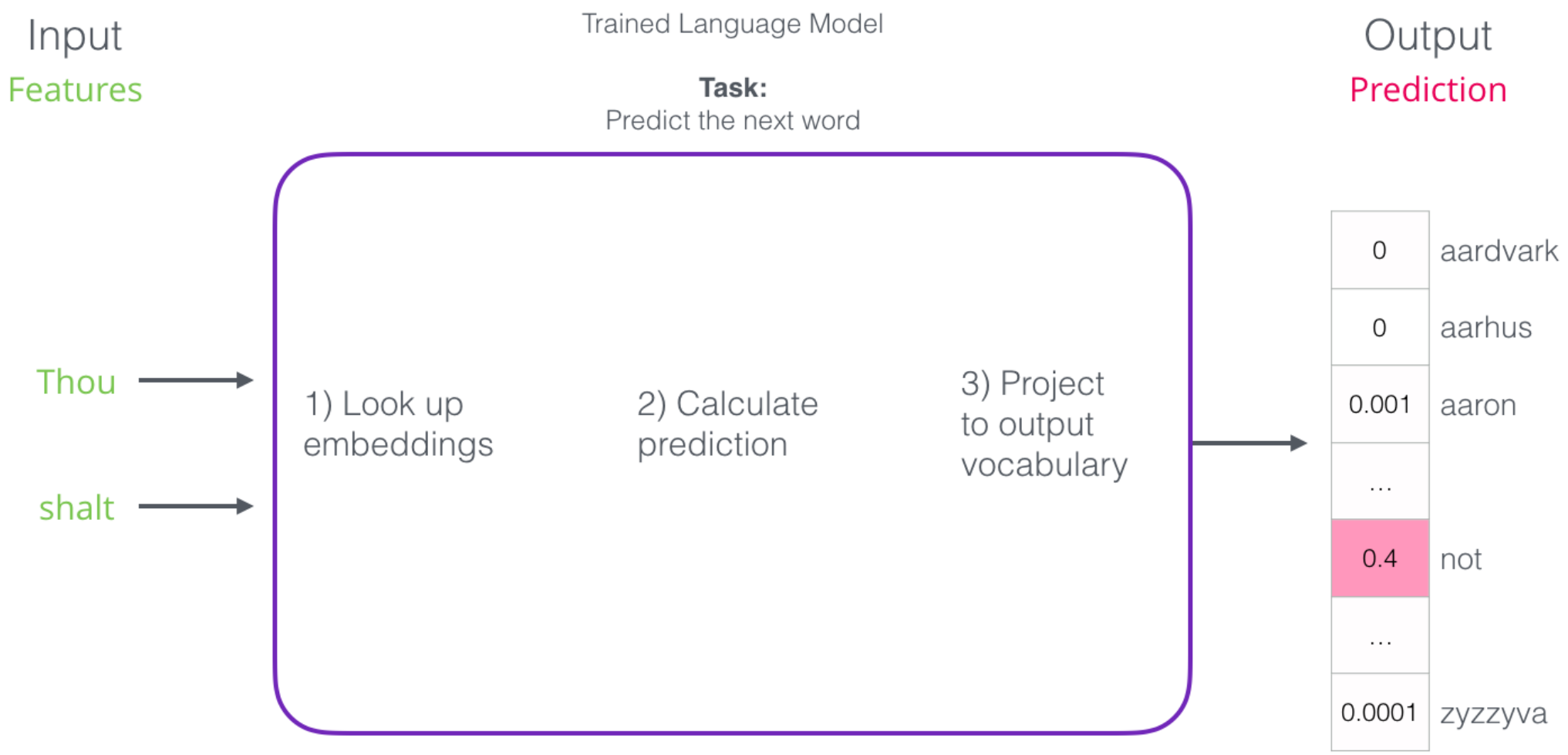
Task:
Predict the next word

→ not

НЕЙРОННАЯ МОДЕЛЬ ЯЗЫКА

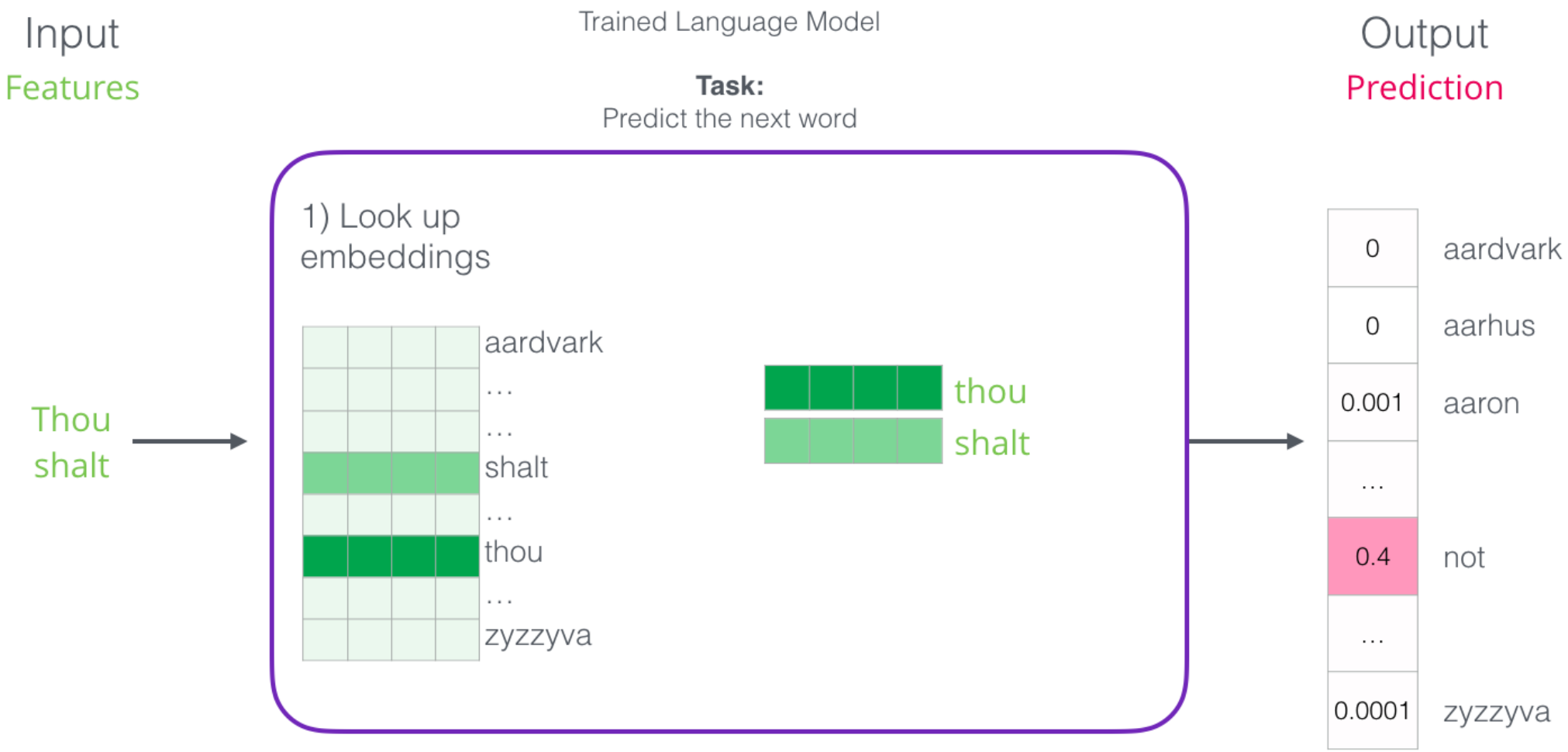


НЕЙРОННАЯ МОДЕЛЬ ЯЗЫКА



НЕЙРОННАЯ МОДЕЛЬ ЯЗЫКА

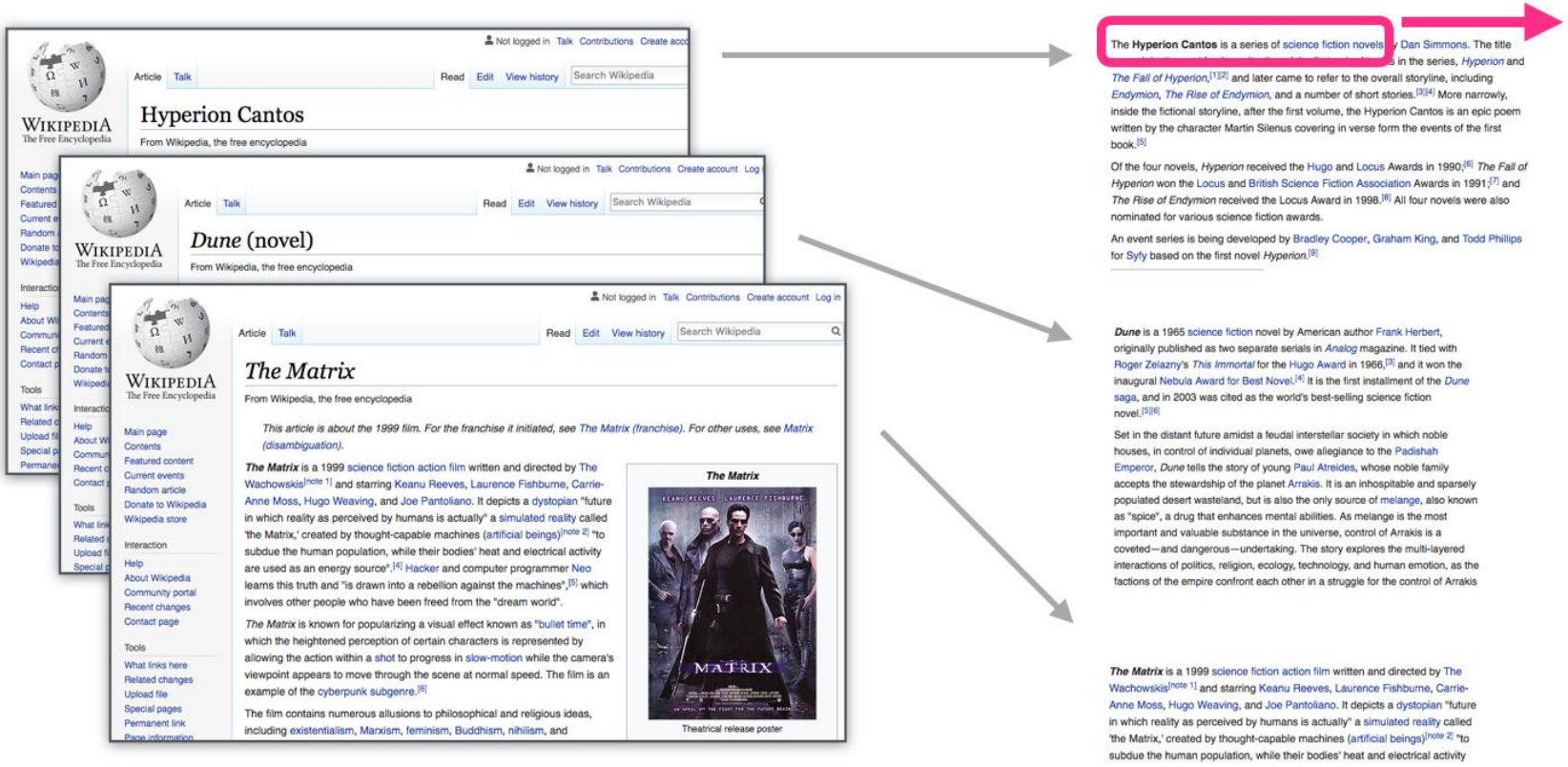
В результате обучения создаётся матрица с вложениями всех слов нашего словаря.



ОБУЧЕНИЕ МОДЕЛИ ЯЗЫКА

Как создаётся матрица вложений?

1. Получаем много текстовых данных (скажем, все статьи Википедии)
2. Устанавливаем окно (например, из трёх слов), которое скользит по всему тексту.
3. Скользящее окно генерирует образцы для обучения нашей модели



ОБУЧЕНИЕ МОДЕЛИ ЯЗЫКА

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

Первые два слова принимаем за признаки, а третье слово — за метку

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 ЯЗЫКА

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

На практике модели обычно обучаются непосредственно в процессе движения скользящего окна. Но логически фаза «генерации набора данных» отделена от фазы обучения.

ОБУЧЕНИЕ МОДЕЛИ ЯЗЫКА

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

CBOW

Jay was hit by a _____ → Jay was hit by a _____ bus

Jay was hit by a _____ bus in...

by	a	red	bus	in

набор данных для обучения модели:

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

SKIP-GRAM

В зелёном слоте — входное слово, а каждое розовое поле представляет возможный выход

input	output
red	by
red	a
red	bus
red	in

SKIP-GRAM

Визуализируем скользящее окно:

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

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

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

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	…

thou	shalt	not	make	a	machine	in	the	…

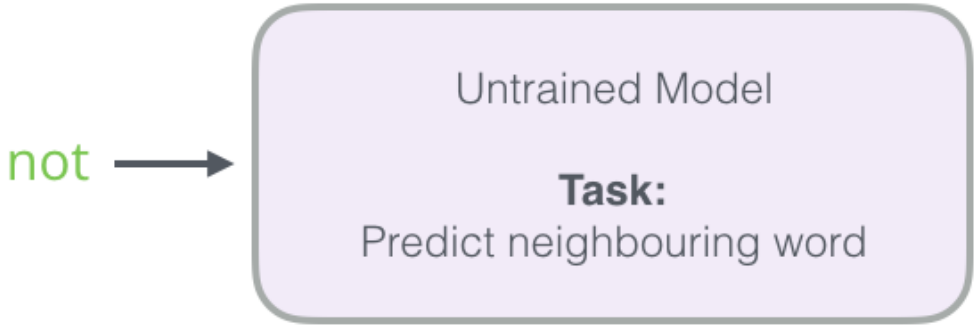
thou	shalt	not	make	a	machine	in	the	…

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

ПЕРЕСМОТР ПРОЦЕССА ОБУЧЕНИЯ

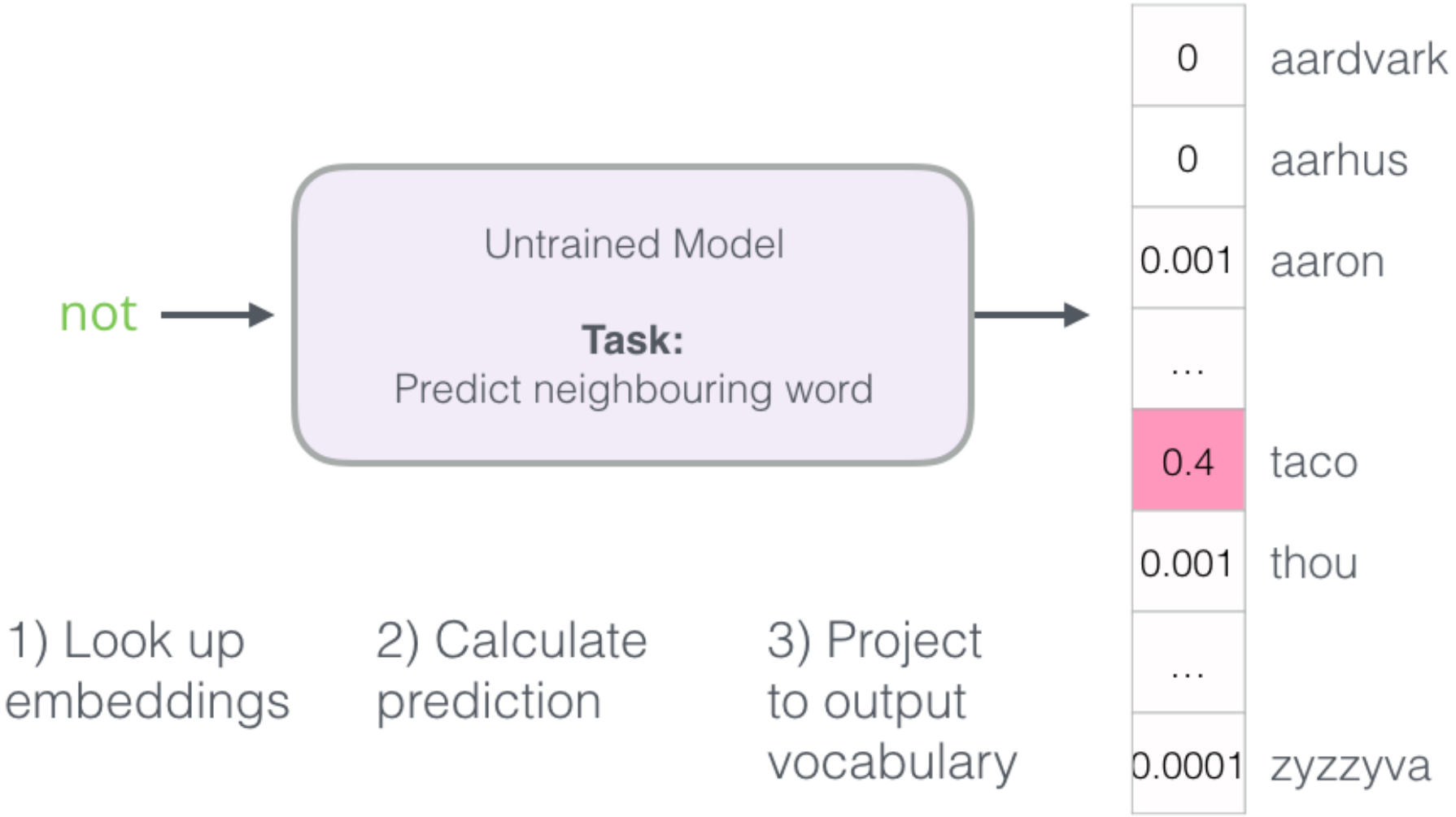
input word	target word
not	thou
not	shalt
not	make
not	a
make	shalt
make	not
make	a
make	machine
a	not
a	make
a	machine
a	in
machine	make
machine	a
machine	in
machine	the
in	a
in	machine
in	the
in	likeness



Предсказание соседнего слова

ПЕРЕСМОТР ПРОЦЕССА ОБУЧЕНИЯ

модель не обучена, на данном этапе её прогноз наверняка неправильный



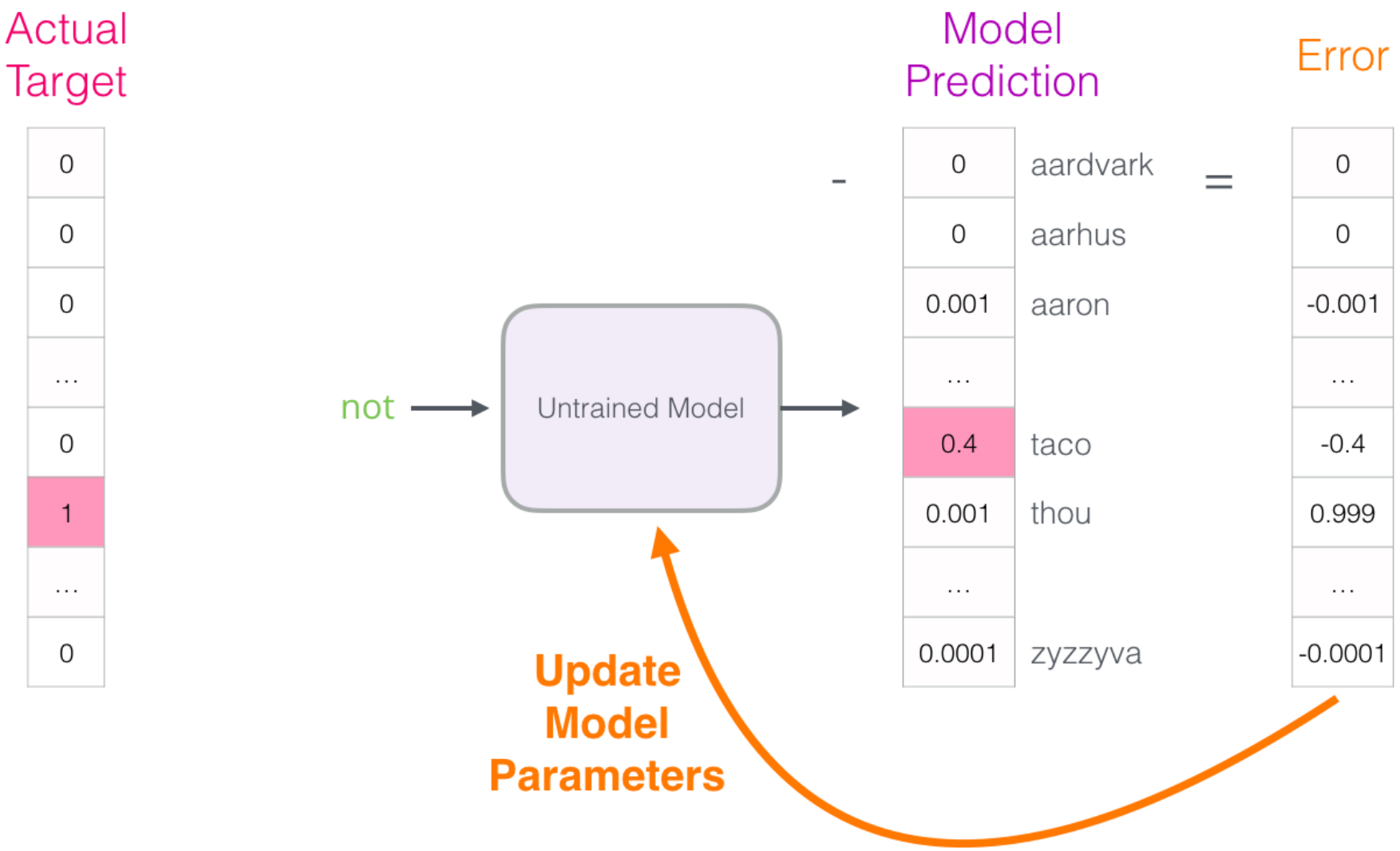
ПЕРЕСМОТР ПРОЦЕССА ОБУЧЕНИЯ

«Целевой вектор» — тот, в котором у целевого слова вероятность 1, а у всех остальных слов вероятность 0

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

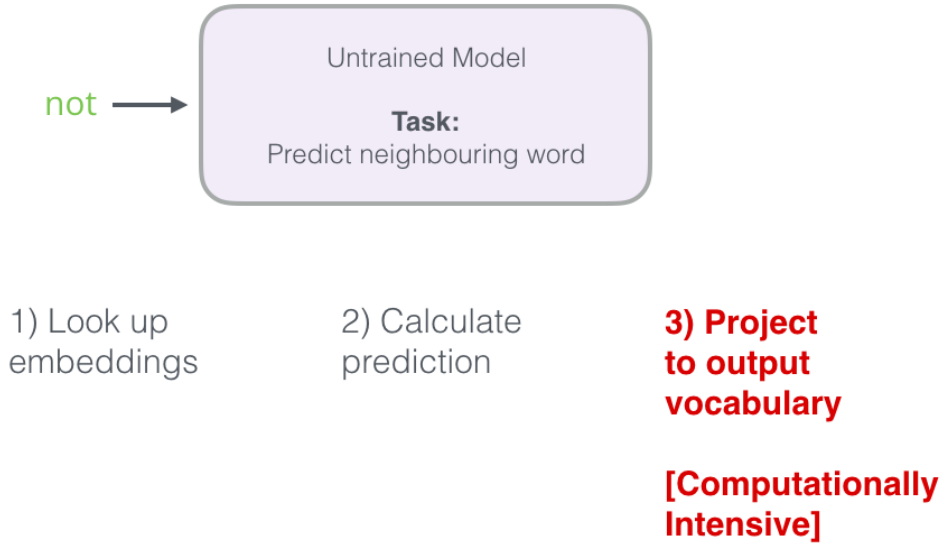
Насколько ошиблась модель?
Вычитаем вектор прогноза из целевого и получаем вектор ошибки

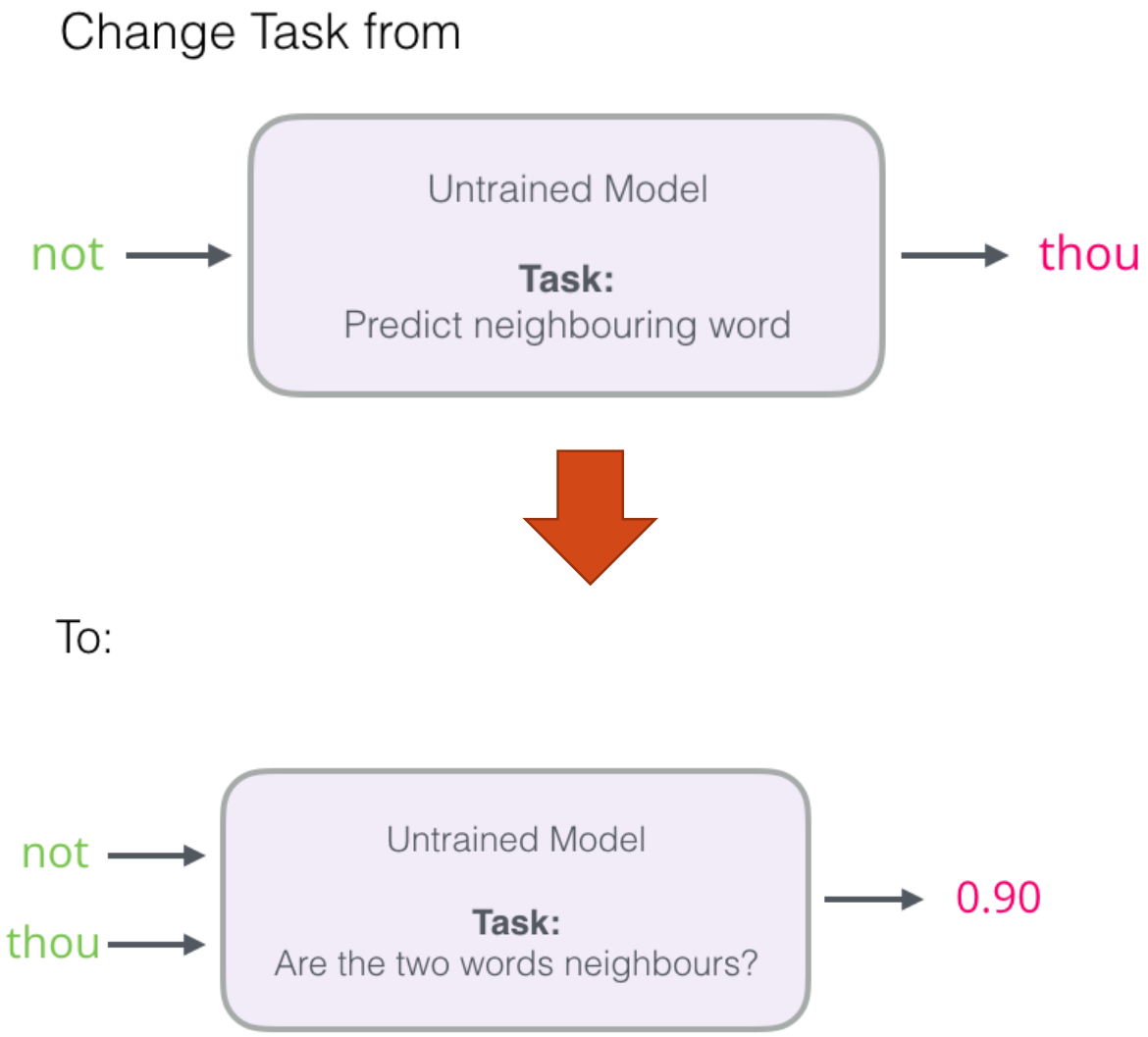
ПЕРЕСМОТР ПРОЦЕССА ОБУЧЕНИЯ



Повторяем всё снова и снова в течение нескольких эпох, и в итоге получаем обученную модель: из неё можно извлечь матрицу вложений и использовать в любых приложениях.

ОТРИЦАТЕЛЬНЫЙ ОТБОР





Разделим 3й этап:

1. Создать высококачественные вложения слов (без прогноза следующего слова).
2. Использовать эти высококачественные вложения для обучения языковой модели (для прогнозирования).

модель логистической регрессии

ОТРИЦАТЕЛЬНЫЙ ОТБОР

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

Такая модель вычисляется с невероятной скоростью: миллионы образцов за считанные минуты.

ОТРИЦАТЕЛЬНЫЙ ОТБОР

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

ОТРИЦАТЕЛЬНЫЙ ОТБОР

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		

SKIP-GRAM С ОТРИЦАТЕЛЬНОЙ ВЫБОРКОЙ (SGNS)

Skipgram

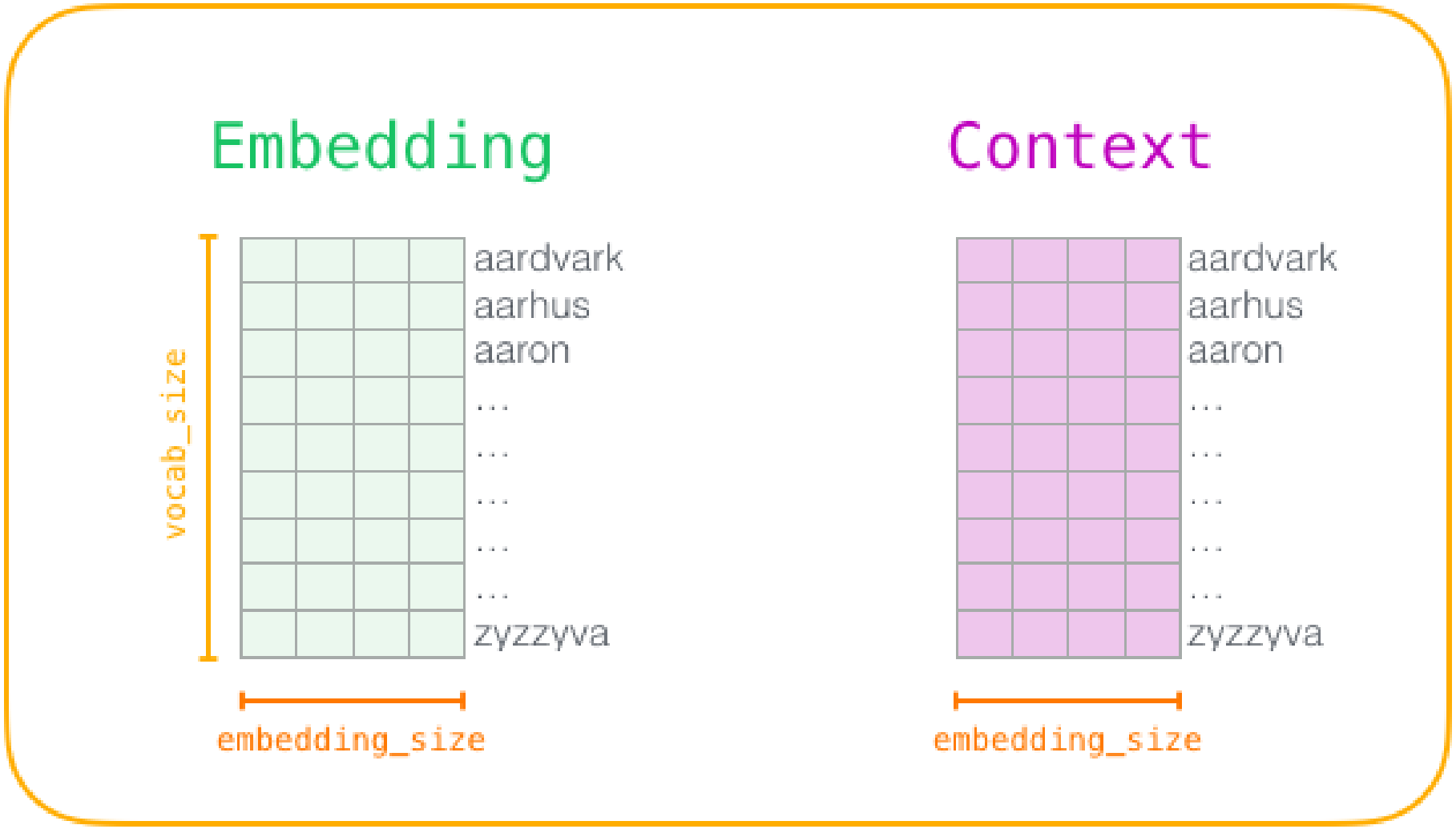
shalt	not	make	a	machine

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

ОБУЧЕНИЕ WORD2VEC

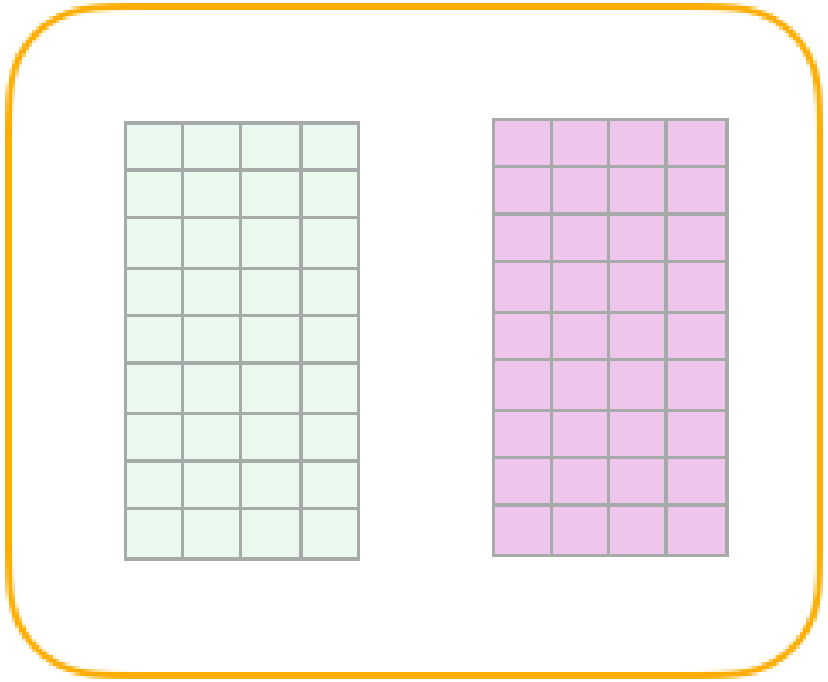


ОБУЧЕНИЕ WORD2VEC

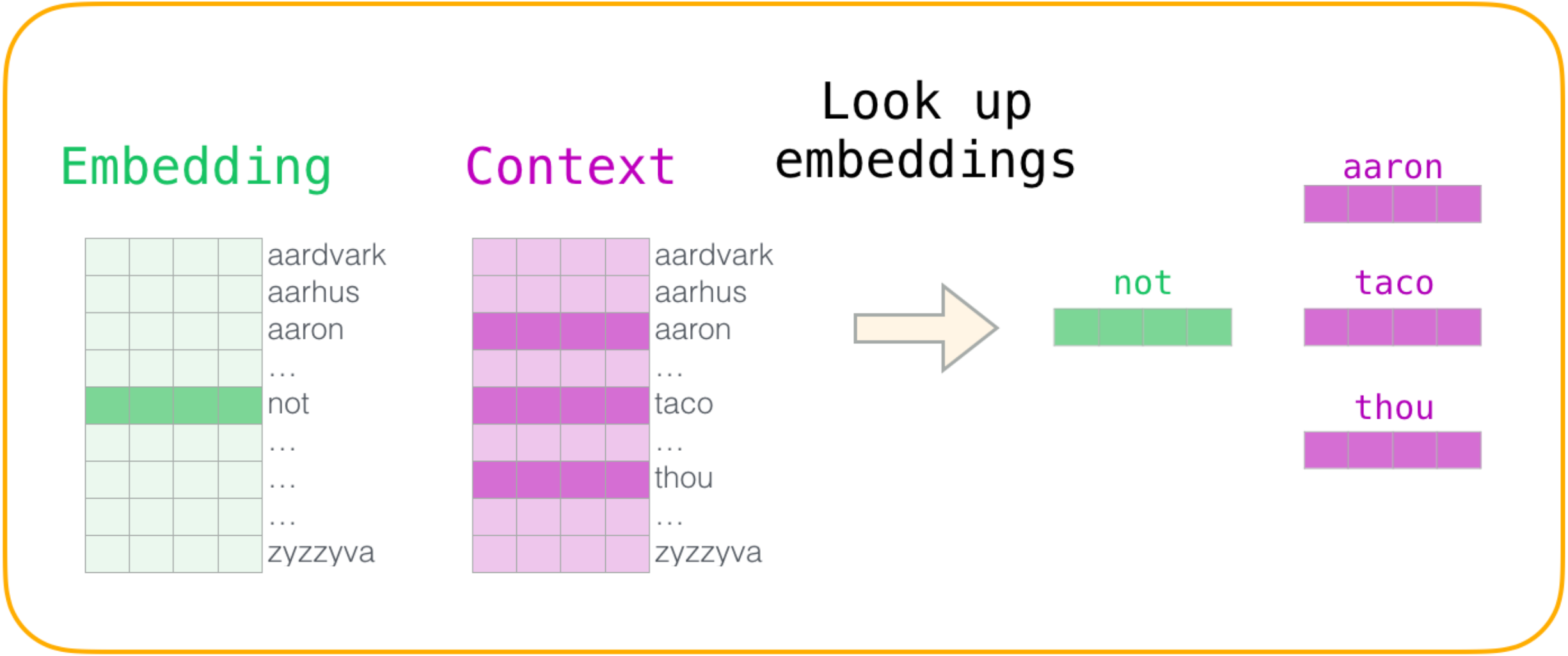
dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

model



ОБУЧЕНИЕ WORD2VEC



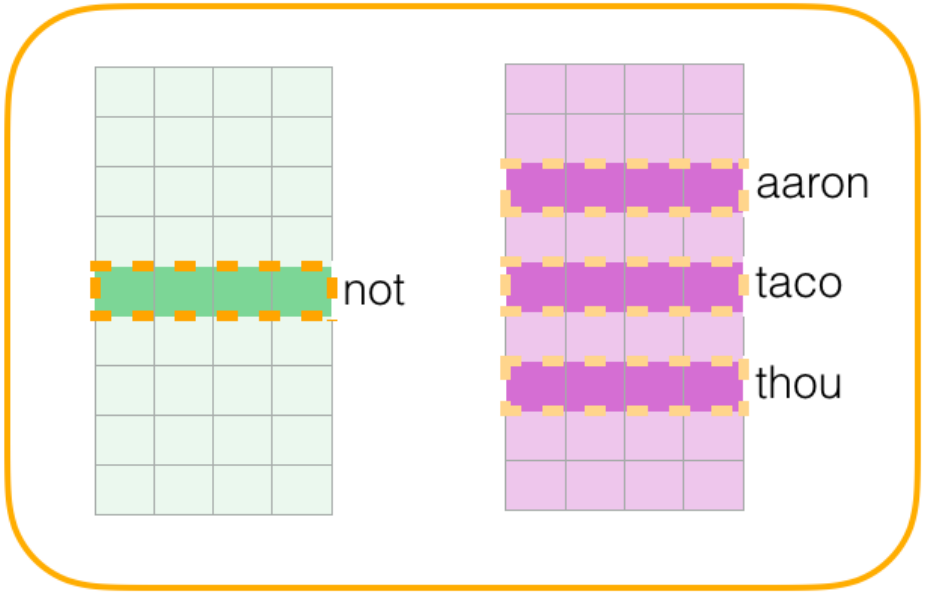
input word	output word	target	input • output
not	thou	1	0.2
not	aaron	0	-1.11
not	taco	0	0.74

input word	output word	target	input • output	sigmoid()
not	thou	1	0.2	0.55
not	aaron	0	-1.11	0.25
not	taco	0	0.74	0.68

input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68

ОБУЧЕНИЕ WORD2VEC

input word	output word	target	input • output	sigmoid()	Error
not	thou	1	0.2	0.55	0.45
not	aaron	0	-1.11	0.25	-0.25
not	taco	0	0.74	0.68	-0.68



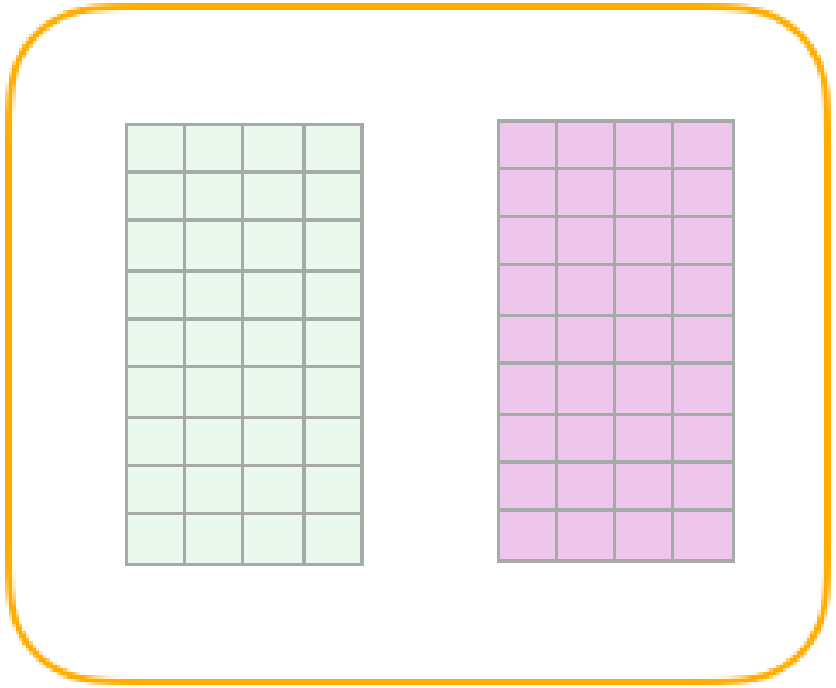
Update Model Parameters

ОБУЧЕНИЕ WORD2VEC

dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

model



РАЗМЕР ОКНА И КОЛИЧЕСТВО ОТРИЦАТЕЛЬНЫХ ОБРАЗЦОВ

Window size: 5

Window size: 15

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

ПРИМЕР

https://github.com/thushv89/exercises_thushv_dot_com/blob/master/word2vec_light_on_math_ml/word2vec_light_on_math_ml.ipynb