

---

# Текстовые эмбединги простые и сложные

Штех Геннадий  
NAUMEN  
27.04.2019

**#DATASTART2019**

---



[https://github.com/ShT3ch/public\\_workshop](https://github.com/ShT3ch/public_workshop)

# О чем поговорим

---

- 1 Об эмбедингах в целом
  - 2 Историческая справка
  - 3 Эмбединги слов и их получение
  - 4 Способы применения к задачам
  - 5 Немного об актуальных инструментах
-

# Что такое эмбеддинги

# Эмбединги бывают разные

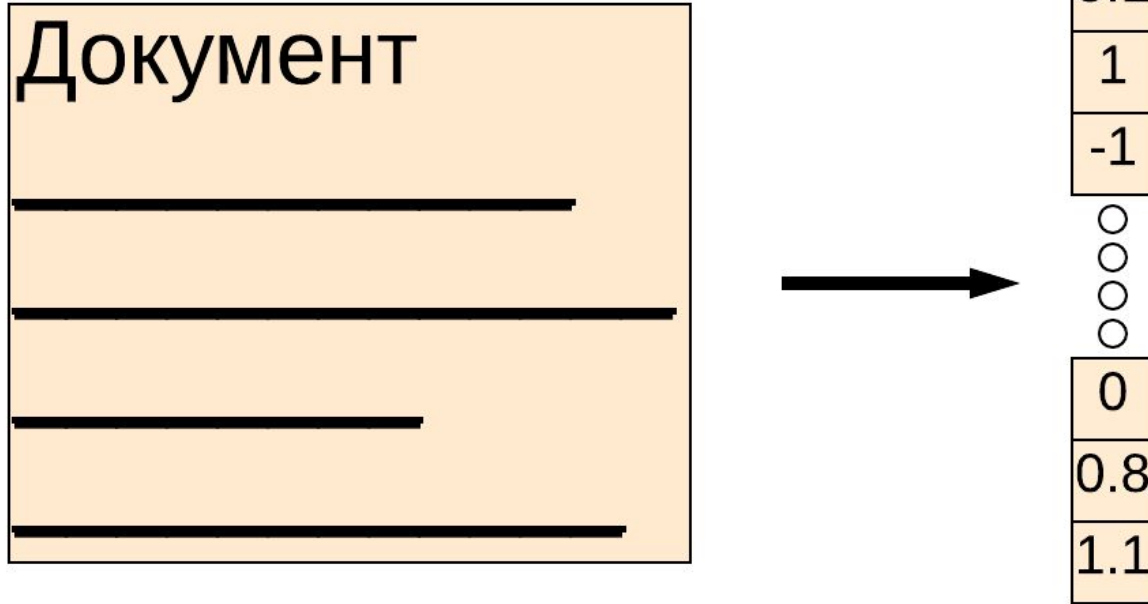
---

СЛОВО



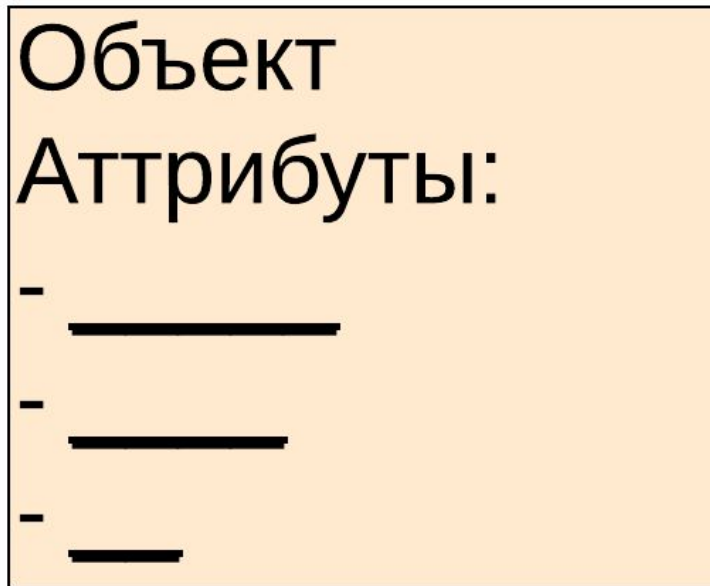
# Эмбединги бывают разные

---



# Эмбединги бывають разные

---



# Что было до эмбеддингов

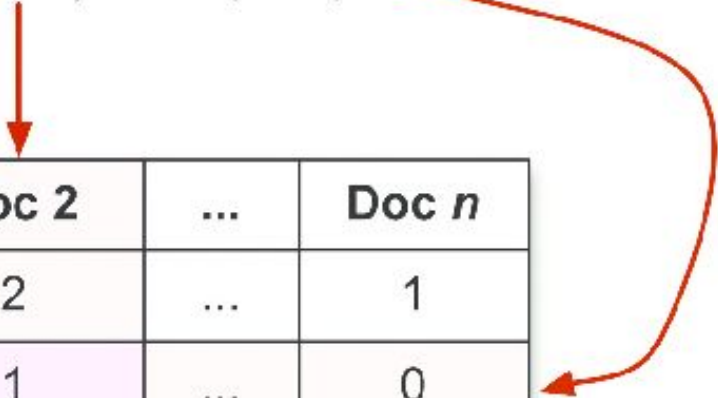
- Счетчики
- TF-IDF
- Факторизации и тематические модели



# Счетчики

---

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$



	Doc 1	Doc 2	...	Doc $n$
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	...	...	...	
Term(s) $n$	0	6	...	3

# TF-IDF

---

$$w_{x,y} = \text{tf}_{x,y} \times \log \left( \frac{N}{\text{df}_x} \right)$$

## TF-IDF

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$

$\text{df}_x$  = number of documents containing  $x$

$N$  = total number of documents

# Тематическое моделирование

## Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

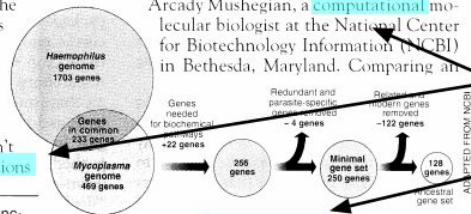
## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **trivial numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

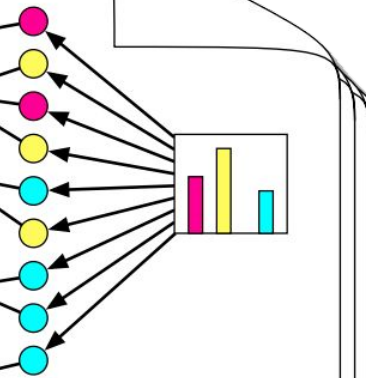


\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions & assignments



# Пример тематической модели

---

#106: приложение + реклама + сервис + продукт + пользователь + платформа + ...

#107: проект + рамка + мрф + реализовать + кц + решение + данный + филиал + ...

#108: работа + затрата + качество + время + количество + сотрудник + расход + ...

#109: олег + александр + сергей + спасибо + тема + согласный + комментарий + ...

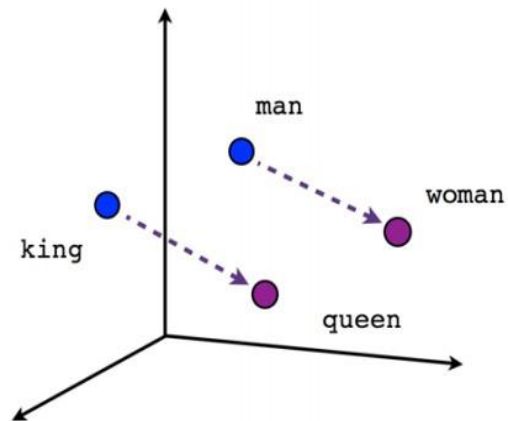
#110: приставка + компьютер + купить + пк + поставить + телевизор + питание + ...

#111: система + объект + управление + время + контроль + группа + прибор + ...

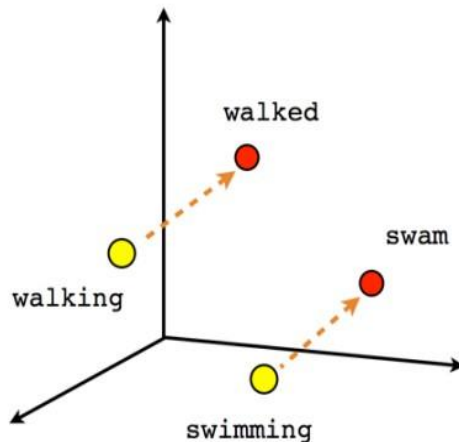
# Какими свойствами обладают эмбединги

- “Понимание” аналогий
- “Понимание” синтаксиса
- Память некоторых фактов
- Задача поиска синонимов

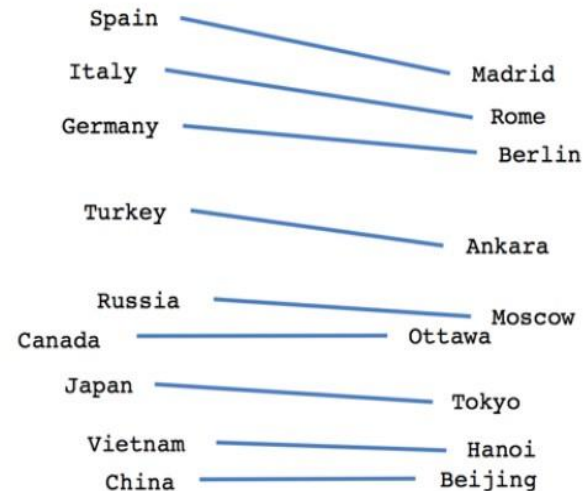
# Эмбеддинги слов



Male-Female



Verb tense



Country-Capital

# Эмбединги слов, близость

---

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

# Как получают эмбеддинги



# Skip-gram

---

Source Text	Training Samples					
<table><tr><td>The</td><td>quick</td><td>brown</td></tr></table> fox jumps over the lazy dog. ➡	The	quick	brown	(the, quick) (the, brown)		
The	quick	brown				
The <table><tr><td>quick</td><td>brown</td><td>fox</td></tr></table> jumps over the lazy dog. ➡	quick	brown	fox	(quick, the) (quick, brown) (quick, fox)		
quick	brown	fox				
The quick <table><tr><td>brown</td><td>fox</td><td>jumps</td></tr></table> over the lazy dog. ➡	brown	fox	jumps	(brown, the) (brown, quick) (brown, fox) (brown, jumps)		
brown	fox	jumps				
The <table><tr><td>quick</td><td>brown</td><td>fox</td><td>jumps</td><td>over</td></tr></table> the lazy dog. ➡	quick	brown	fox	jumps	over	(fox, quick) (fox, brown) (fox, jumps) (fox, over)
quick	brown	fox	jumps	over		

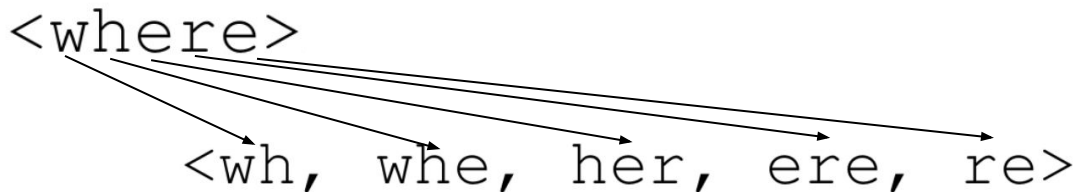
# Skip-gram

---

Source Text	Training Samples						
<table><tr><td>The</td><td>quick</td><td>brown</td></tr></table> fox jumps over the lazy dog. ➡	The	quick	brown	(the, quick) (the, brown)			
The	quick	brown					
<table><tr><td>The</td><td>quick</td><td>brown</td><td>fox</td></tr></table> jumps over the lazy dog. ➡	The	quick	brown	fox	(quick, the) (quick, brown) (quick, fox)		
The	quick	brown	fox				
<table><tr><td>The</td><td>quick</td><td>brown</td><td>fox</td><td>jumps</td></tr></table> over the lazy dog. ➡	The	quick	brown	fox	jumps	(brown, the) (brown, quick) (brown, fox) (brown, jumps)	
The	quick	brown	fox	jumps			
<table><tr><td>The</td><td>quick</td><td>brown</td><td>fox</td><td>jumps</td><td>over</td></tr></table> the lazy dog. ➡	The	quick	brown	fox	jumps	over	(fox, quick) (fox, brown) (fox, jumps) (fox, over)
The	quick	brown	fox	jumps	over		

# Проблема Out-Of-Vocabulary (OOV)

- Char-Ngramm



- Byte Pair Encoding

*Dictionary*

5 low  
2 lower  
6 new est  
3 widest

*Vocabulary*

l, o, w, e, r, n, w, s, t, i, d, es, est

# Проблема ограниченной выразительности

WORD	NEAREST NEIGHBOURS
python	java, php, shell, PHP, server, HTML plugin, zip, javascript
apple	iphone, android, mac, microsoft samsung, phone, galaxy, touch
date	registration, join, location, from changed, list, event, hours, festival
bow	gun, fire, shot, deep, down, snow head, ride, ball, dead
mass	energy, effect, impact, movement potential, military, weight, society exercise, lower

WORD	$p(z)$	NEAREST NEIGHBOURS
python	0.33 0.42 0.25	monty, spamalot, cantsin perl, php, java, c++ molurus, pythons
apple	0.34 0.66	almond, cherry, plum macintosh, iifx, iigs
date	0.10 0.28 0.31 0.31	unknown, birth, birthdate dating, dates, dated to-date, stateside deadline, expiry, dates
bow	0.46 0.38 0.16	stern, amidships, bowsprit spear, bows, wow, sword teign, coxs, evenlode
mass	0.22 0.42 0.36	vespers, masses, liturgy energy, density, particle wholesale, widespread

# Способы применения

0. [Скачать]
1. [Предобработать]
2. Усреднить и затем любой ML
3. Усреднить по TF-IDF
4. Усреднить по Arora et al 2017
5. Одномерные свёртки
6. ...

# Скачать

---

1 Word2Vec

2 GLoVe

3 Fasttext



download word vectors

# Скачать

Размер корпуса ▲▼	Объём словаря ▲▼	Частотный порог ▲▼	Target ▲▼	Алгоритм ▲▼	Размерность вектора ▲▼	Размер окна ▲▼
270 миллионов слов	189 193	5 (потолок словаря 250K)	<a href="#">Universal Tags</a>	Continuous Bag-of- Words	300	20
788 миллионов слов	248 978	5 (потолок словаря 250K)	<a href="#">Universal Tags</a>	Continuous Skipgram	300	2
почти 5 миллиардов слов	249 565	5 (потолок словаря 250K)	<a href="#">Universal Tags</a>	Continuous Skipgram	300	2
почти 5 миллиардов слов	192 415	5 (потолок словаря 250K)	Нет	fastText CBOW (3..5- граммы)	300	10

Скачать

---

**\*.bin** *vs* **\*.vec**

---



# Предобработка

---

## 1. Парсинг

"Нет возможности сделать  
корректировку минусовых  
остатков"



"ПОС 2 расхождение на 880р  
недосдача"



"Завис ПОС№1 на "Итого".  
Завис с 01:05"



"Кофе машина выдает ошибку  
№186 при промывке."



## 2. Нормализация текстов

['Нет', 'возможности', 'сделать',  
'корректировку', 'минусовых',  
'остатков']

['ПОС', '2', 'расхождение', 'на', '880р',  
'недосдача']

['Завис', 'ПОС', '1', 'на', 'Итого',  
'Завис', 'с', '01', '05']

['Кофе', 'машина', 'выдает',  
'ошибку', '186', 'при', 'промывке']

# Предобработка

---

## 1. —→ 2. Нормализация текстов

['Нет', 'возможности', 'сделать',  
'корректировку', 'минусовых',  
'остатков']



['ПОС', '2', 'расхождение', 'на',  
'880р', 'недосдача']



['Завис', 'ПОС', '1', 'на', 'Итого',  
'Завис', 'с', '01', '05']



['Кофе', 'машина', 'выдает',  
'ошибку', '186', 'при', 'промывке']



## 3. Формирование словаря

['нет', 'возможность', 'сделать',  
'корректировка', 'минусовый',  
'остаток']

['пос', '2', 'расхождение', 'на',  
'880р', 'недосдача']

['зависнуть', 'пос', '1', 'на',  
'итого', 'зависнуть', 'с', '01', '05']

['кофе', 'машина', 'выдавать',  
'ошибка', '186', 'при',  
'промывка']

# Предобработка

---

## 2. —→ 3. Формирование словаря

['нет', 'возможность', 'сделать',  
'корректировка', 'минусовый',  
'остаток']

['пос', '2', 'расхождение', 'на', '880р',  
'недосдача']

['зависнуть', 'пос', '1', 'на', 'итого',  
'зависнуть', 'с', '01', '05']

['кофе', 'машина', 'выдавать', 'ошибка',  
'186', 'при', 'промывка']

# Предобработка

---

## 2. —→ 3. Формирование словаря

- "онлайн" -> 137
- "перечеркнуть" -> 138
- "трм" -> 139
- "вод" -> 140
- "зал" -> 141
- "клиентский" -> 142
- "пол" -> 143
- "потечь" -> 144
- "протекать" -> 145
- "торговый" -> 146
- "туалет" -> 147
- "выйти" -> 148
- "выполнить" -> 149

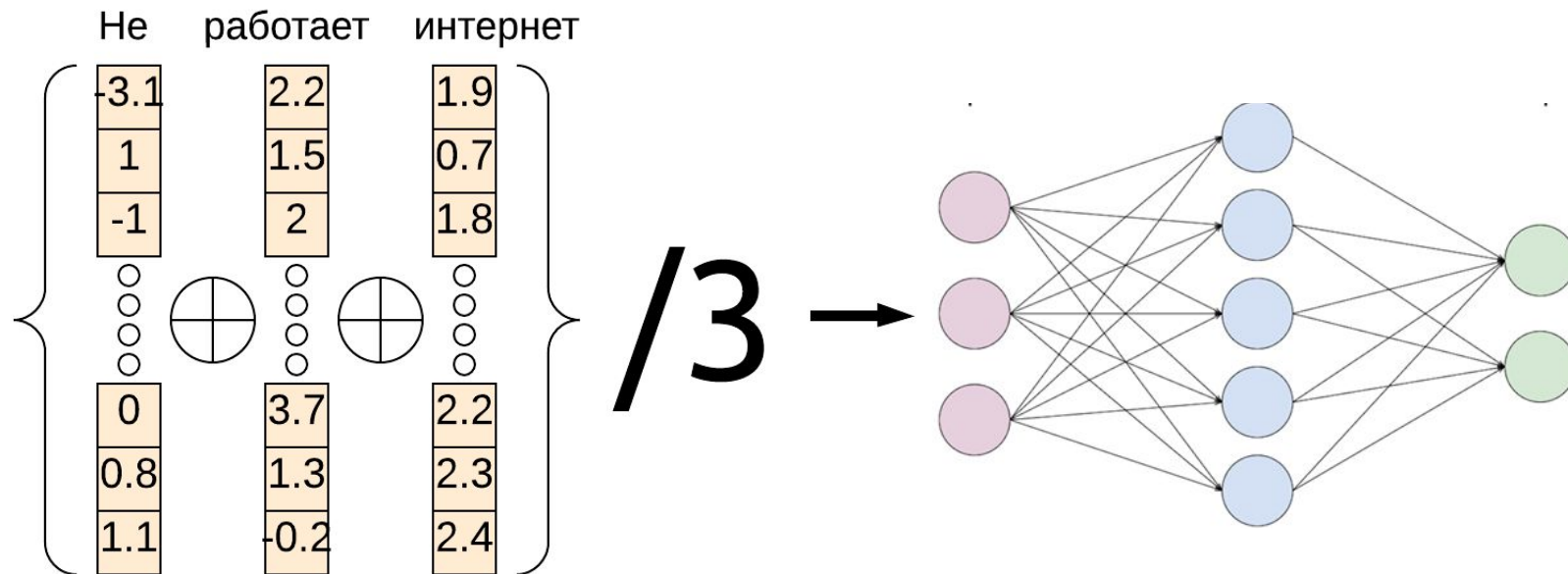
# Предобработка

---

## 3. —→ 4. Фильтрация стоп-слов

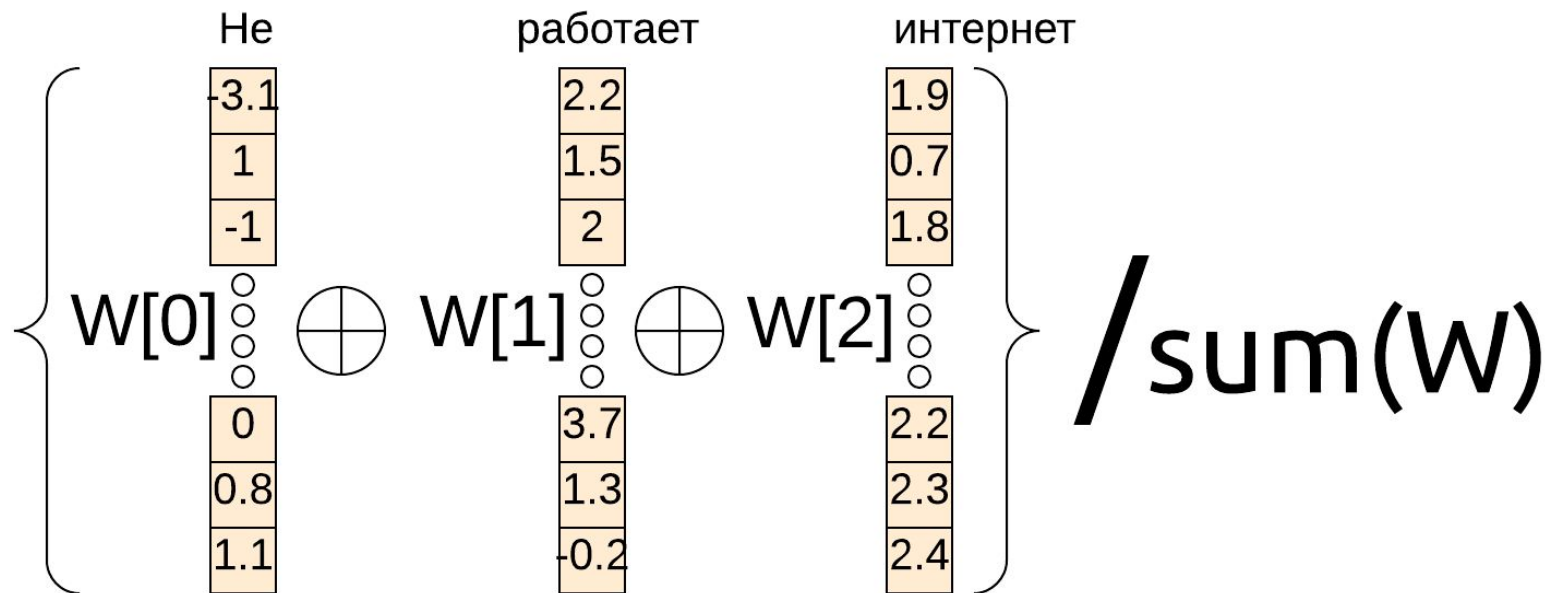
"нет", "быть", "к", "за", "после", "при", "как",  
"так", "между", "более", "до", "если", "здесь",  
"из", "можно", "о", "они", "перед", "сам", "то",  
"тот", "что", "вы", "или", "чем"

# Feature extraction. Усреднение векторов + ML



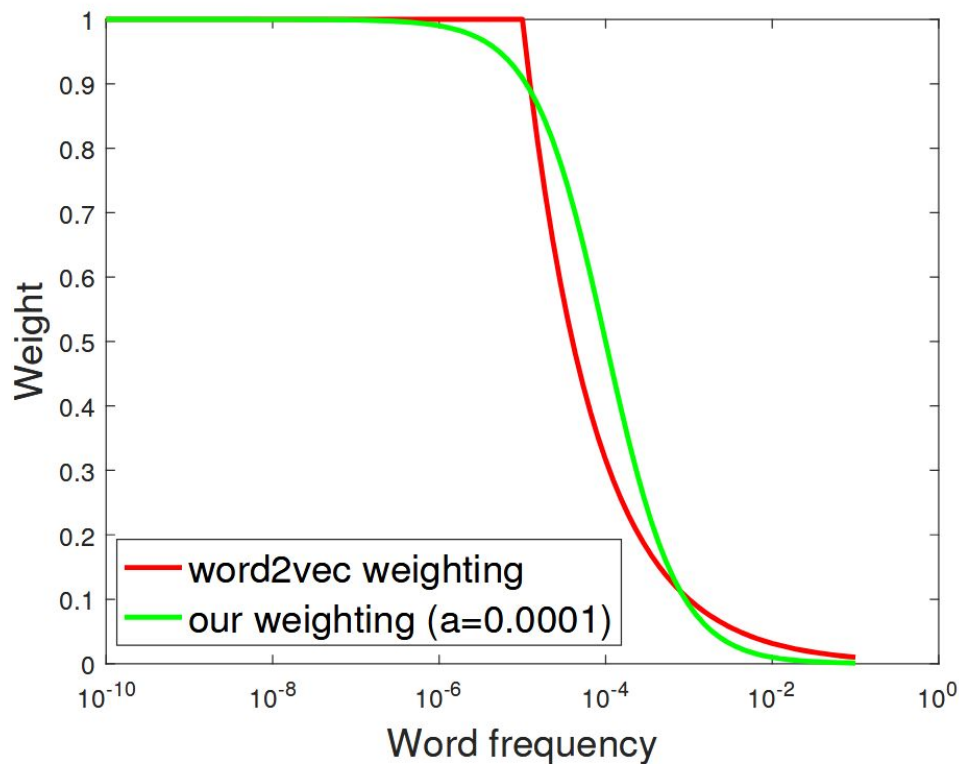
## Feature extraction. Усреднение по TF-IDF + ML

---



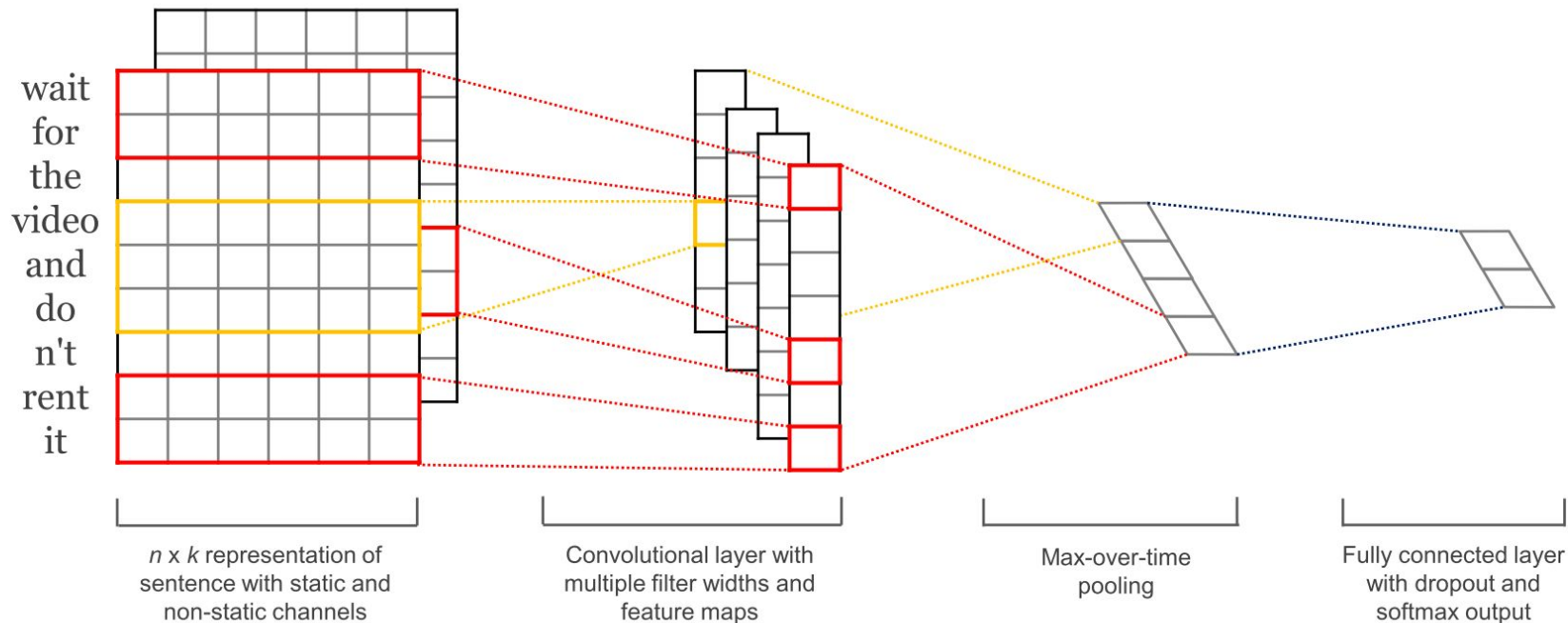
## Feature extraction. Усреднение по Arora et al 2017 + ML

---



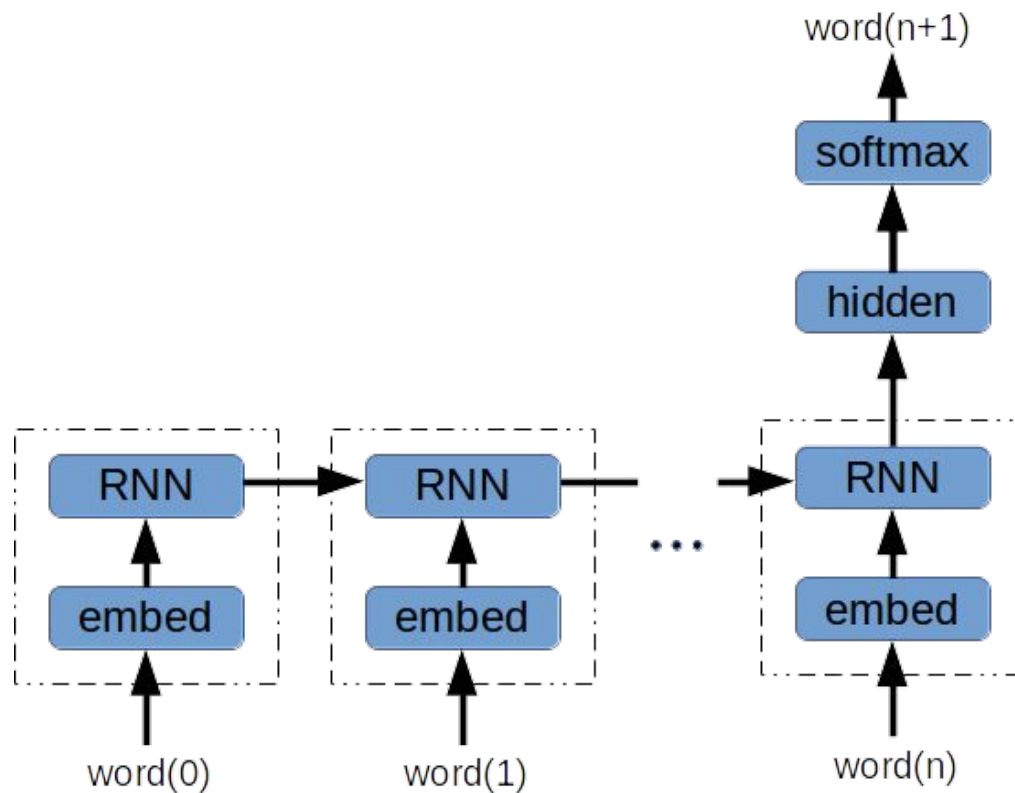


# Feature extraction. Свёрточные нейронные сети (CNN)



## Feature extraction. Рекуррентные нейронные сети (RNN)

---

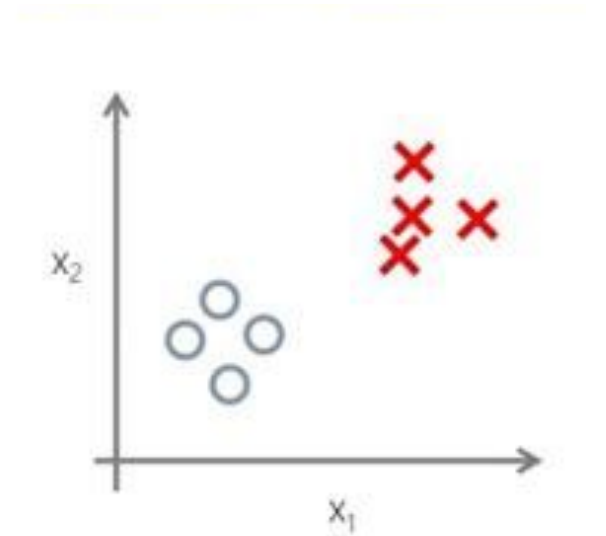
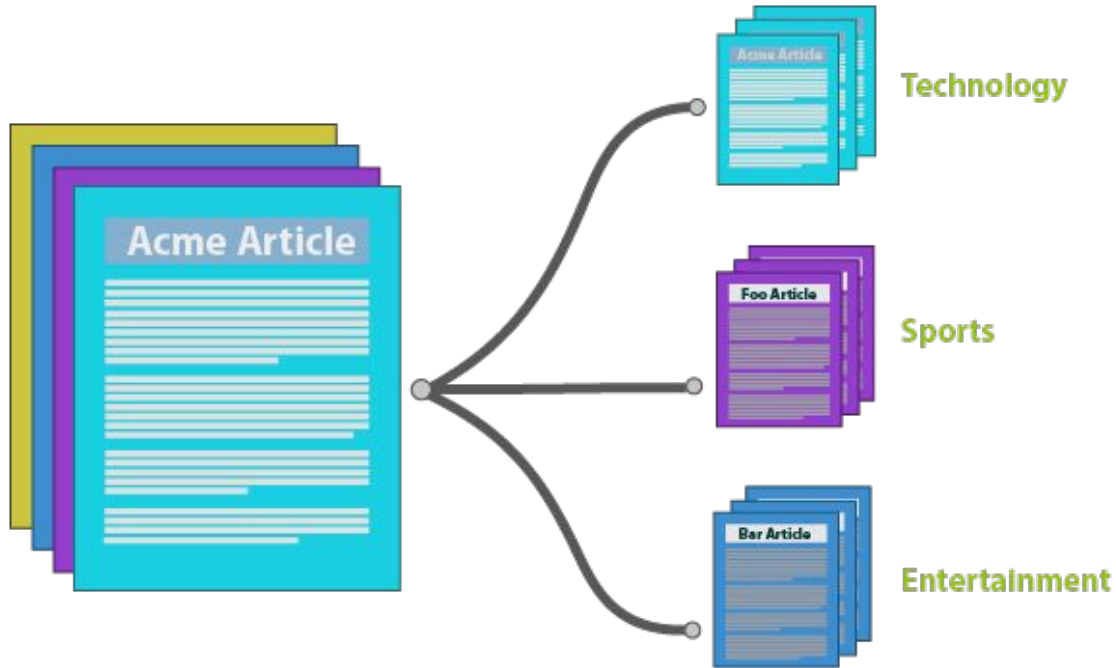


# Примеры применения

- Кластеризация
- Классификация
- Исправление опечаток
- Поиск

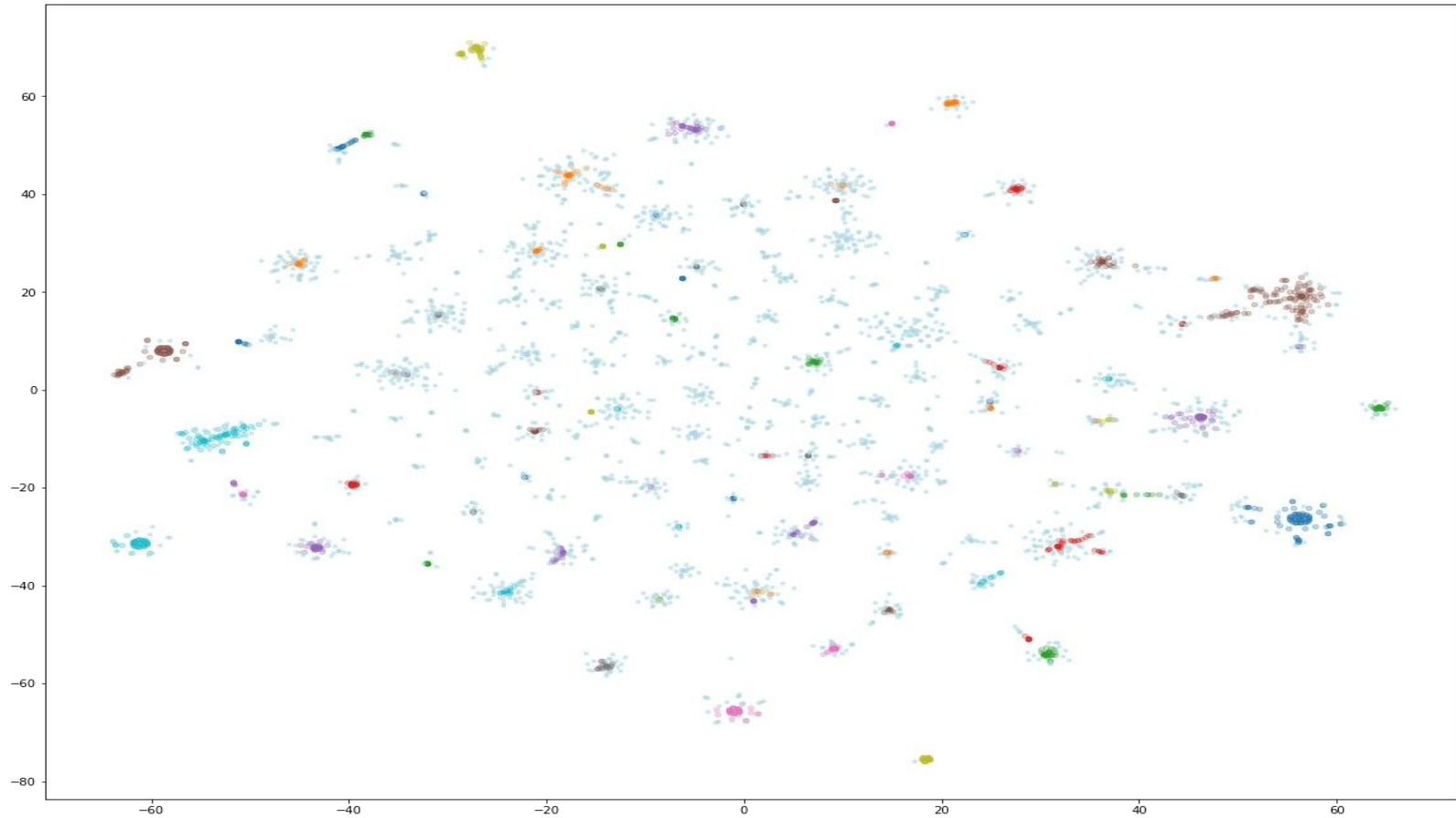
# Классификация

---



# Кластеризация

---



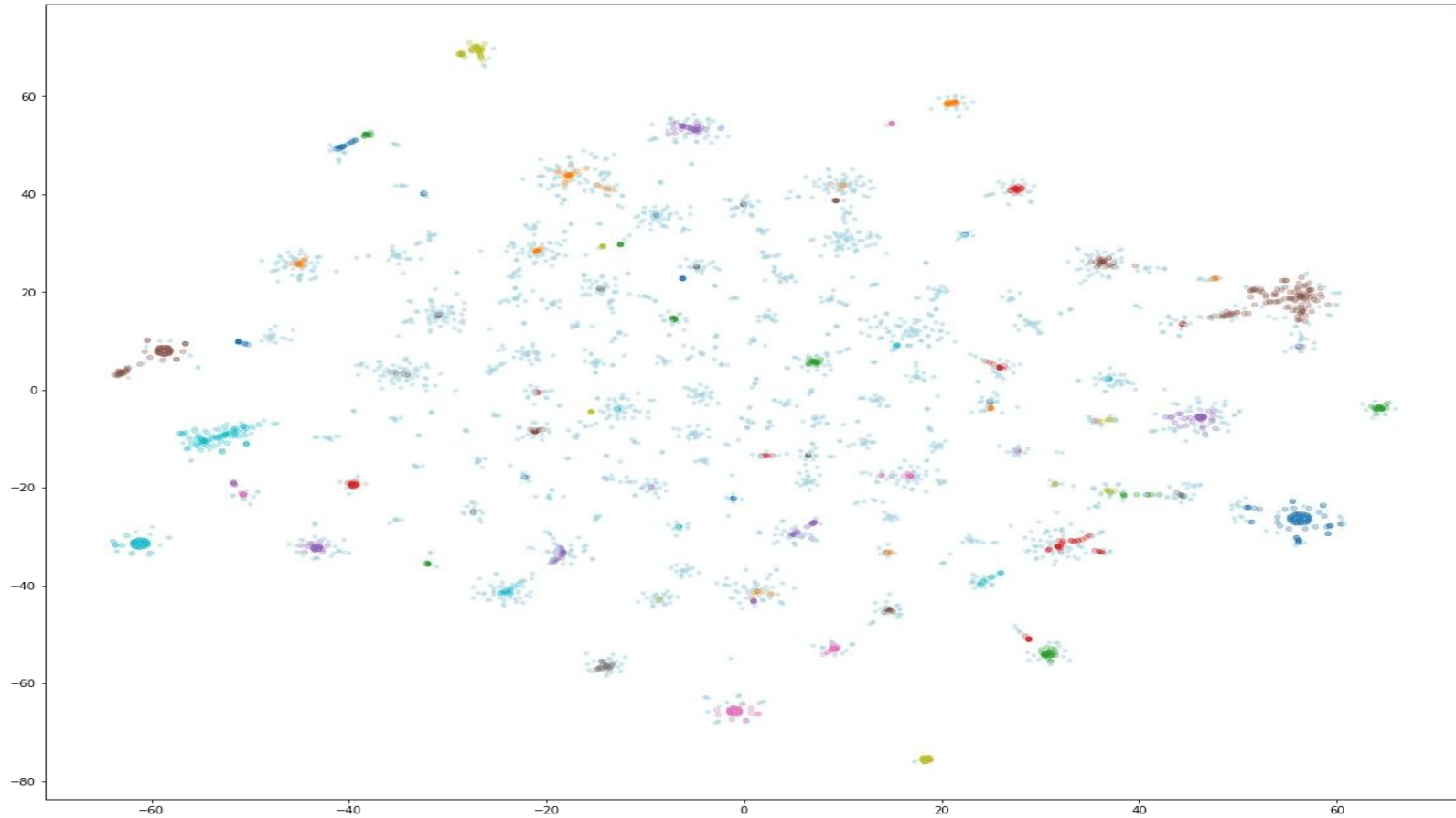
# Исправление опечаток (fasttext)

---

- |                                     |   |        |
|-------------------------------------|---|--------|
| ● wann, wanto, wanr, wany           | → | want   |
| ● havea, havr                       | → | have   |
| ● thiss, thise                      | → | this   |
| ● pleasee, pleasr, pleasw, pleaseee | → | please |
| ● numbe, numbet, numbee, numbr      | → | number |
| ● calll                             | → | call   |
| ● willl, wiill                      | → | will   |

# Поиск

---



---

# Актуальные алгоритмы

## Представление

- Tf-idf, nPMI, hashing trick, BPE

## Поиски

- BM25, HNSW, LSH

## Факторизация (декомпозиция)

- PCA, LSI-LSA, pLSA, nNMF

## Эмбединги

- word2vec, glove, doc2vec, fasttext, poincaré, ELMO

## Тематическое моделирование

- pLSA, LDA, HDP, ARTM

## Нейросетевые подходы

- LSTM, GRU, TCN, Attention, siamese network, similarity learning, Transformer, Augmented RNN



---

# Полезный NLP-софт

## Предобработка

**текста** (нормализация,  
токенизация)

- pymorphy2(ru), snowball stemmer(en), Stanford NLP(en)

## Фреймворки

- sklearn, NLTK, gensim, spaCy

## Узкоспециализированные фреймворки

- BigARTM, Vowpal Wabbit, Fasttext, faiss, annoy, NMSLib, lucene, sphinx, elastic

## Нейросетевые фреймворки

- Pytorch, HuggingFace, AllenNLP, torchtext

---

# Контакты

**Штех Геннадий \***  
**@ NAUMEN**  
gshtekh@naumen.ru

**Gennady Shtekh**  
shtechgen@gmail.com  
t.me/sht3ch  
github.com/ShT3cH

\*R&D Data Usage Department Executive

---

# Подходы и данные для тестирования моделей

---

- <https://github.com/facebookresearch/SentEval>
- <https://arxiv.org/pdf/1707.05589.pdf>
- <https://arxiv.org/pdf/1806.06259.pdf>
- <https://aclweb.org/anthology/D18-1009>
- <https://arxiv.org/pdf/1702.02170.pdf>
- <https://arxiv.org/pdf/1903.09442.pdf>
  
- <https://leaderboard.allenai.org/swag/submissions/public>
- <https://gluebenchmark.com/leaderboard>

# О прогрессе в НЛП

---

- <https://nlpoverview.com/#3>
- <https://arxiv.org/pdf/1708.02709.pdf>
- [http://nlpprogress.com/english/language\\_modeling.html](http://nlpprogress.com/english/language_modeling.html)
- <https://github.com/Separius/awesome-sentence-embedding>

**Посмотрим на будущее**

Появятся совсем простые  
фреймворки для  
использования глубоких  
предобученных сетей

Появятся фреймворки  
для семантического  
поиска документов

Разовьётся подход к  
генерации контента на  
основе RL



Скорее всего сети на  
гиперболических  
пространствах взорвут

BERT “облегчат”



[https://github.com/ShT3ch/public\\_workshop](https://github.com/ShT3ch/public_workshop)

# Хроника появления решений

---

- | Методы работы с текстами на LSTM
  - | Методы работы с текстами на GRU, CNN
  - | Attention и дополненные LSTM/GRU
  - | Transformer
  - | Transfer Learning
  - | Контекстно-зависимые эмбединги
  - | BERT
-

# Подходы к решению OOV

---

## Char-level Convolution

