Нейронные сети и обработка текстов

лекция 4. Векторное представление слова. word2vec

Иван Андреевич Ташков, 2025

Векторное представление слов

Тезаурус

Хотим научиться понимать связи между словами.
Синонимы, антонимы, гиперонимы и т.п.

Должны быть похожи:
Отель и гостиница
Университет и институт
Голкипер и вратарь

Пример: WordNet

two cute handsome twins playing football and wearing goalkeeper gloves

Stanford. Лекция о word2vec

DLS. Лекция word2vec

статья на habre про word2vec

Векторное представление слов

Статьи

A neural probabilistic language model (Bengio et al. 2003)

A unified architecture for natural language processing: Deep neural networks with multitask learning (Collobert & Weston, 2008)

Efficient Estimation of Word Representations in Vector Space (Mikolov et. al 2013) — word2vec

https://code.google.com/archive/p/word2vec

Векторное представление слов
Дистрибутивная семантика

"You shall know a word by the company it keeps" (J. R. Firth 1957:11)

"По окружению слова вы узнаете само слово"

«Бесцветные зеленые идеи яростно спят» (Хомский)

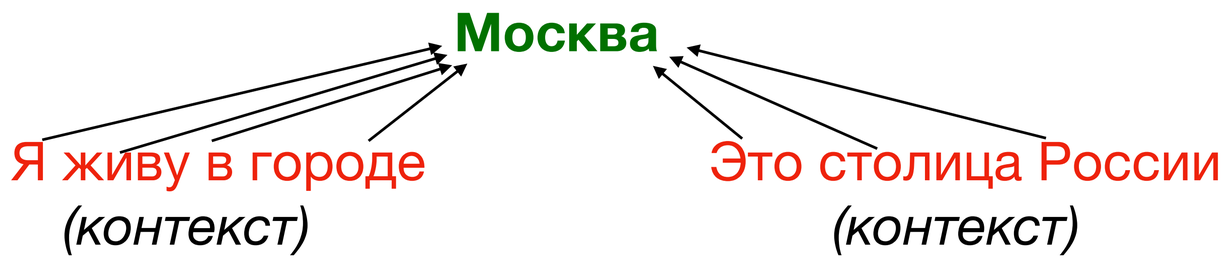
Александр Пиперски —
лексическая сочетаемость

Векторное представление слов

Дистрибутивная семантика

Скажи-ка, дядя, ведь не даром **Москва**, спаленная пожаром, Французу отдана?

Как полагают исследователи, слово «**Москва**» принадлежало ранее к древнерусскому типу склонения на *-ū-, именительный падеж которого заканчивался на -ы



Векторное представление слов

Записываем частоту встречаемости слов в контексте

Теперь мы можем сравнивать слова по схожести (cos, dist)

Понижаем размерность (например, методом главных компонент). С векторами становится удобнее работать

Автомобиль -> (0.15, -0.15, 0.2, 0.1, 0.99)

	я	**люблю**	**хлопья**	**тебя**	**автом обиль**	**куплю**
я		100	5	80	10	80
люблю	100		10	80	10	0
хлопья	5	10		2	1	12
тебя	80	80	2		4	5
автом обиль	10	10	1	4		23
куплю	80	0	12	5	23	

Векторное представление слов
word2vec

Вектора хотелось бы обучать!

Везде далее векторные представления слов будем называть ***эмбэддингами***

CBOW (continuous bag of words) — по контексту предсказываем слово
Skip-gram — по слову предсказываем слова из контекста

окно размера 5

Векторное представление слов

skip-gram

Задача классификации. Количество классов — размер словаря N

На вход слово one hot encoding = *(0, 0, 0, 0, ..., 0, 0, 1, 0, 0, ..., 0, 0, 0)*

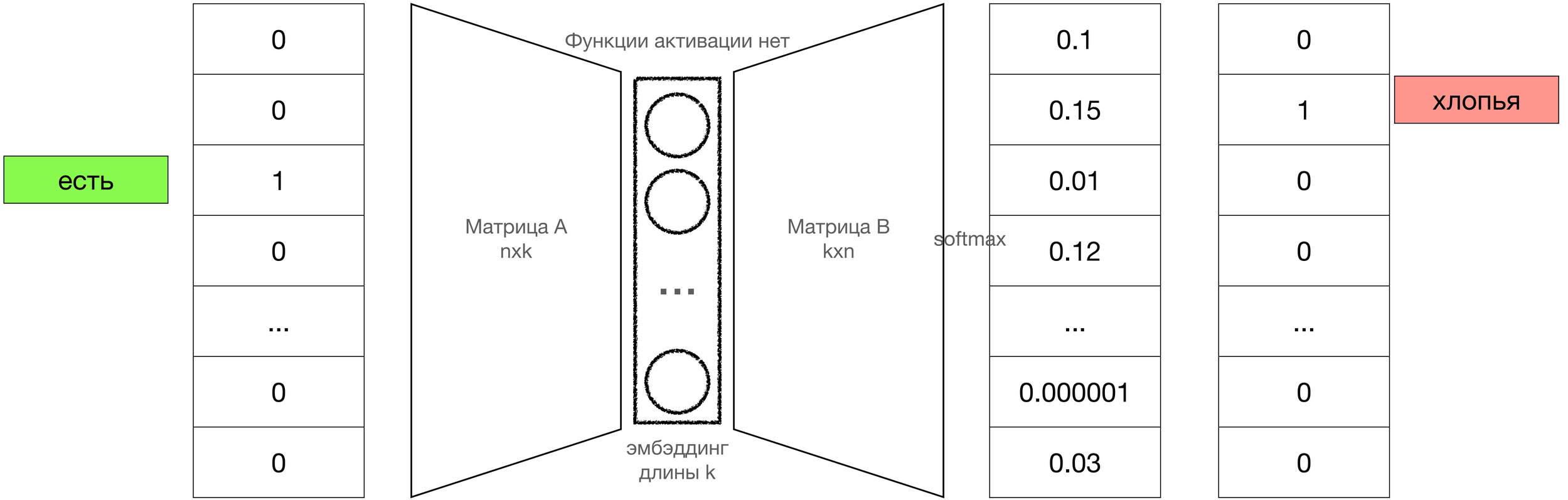
На выходе — вероятность всех слов из словаря

Loss — кроссэнтропия

окно размера 5

Векторное представление слов

skip-gram



Векторное представление слов

Матрица A

я	0.5	0.3	...	0.2
овсяные	0.1	-1.5	...	0.2
...				
есть	0.15	-0.1	...	0.9

Матрица B

я	хлопья	шерстяные	...	баклажаны
0	0.21	-1.5	...	0.2
-0.5	-0.15	1.0	...	-0.15
...	...	...	...	...
0.15	0.88	-0.1	...	0.1

word2vec

Обучение. Метрики

После обучения получим вектора размерности k для всех слов в словаре

k (размер эмбэддинга) можно менять: 50, 100, 300, 600, 1000

Эмбэддинги слов сравнивать (например, считать между ними косинус)

Обучаем на Google News 6B dataset, adagrad (модификация стох. град. спуска)

Table 6: *Comparison of models trained using the DistBelief distributed framework. Note that training of NNLM with 1000-dimensional vectors would take too long to complete.*

Model	Vector Dimensionality	Training words	Accuracy [%]			Training time [days x CPU cores]
			Semantic	Syntactic	Total	
NNLM	100	6B	34.2	64.5	50.8	14 x 180
CBOW	1000	6B	57.3	68.9	63.7	2 x 140
Skip-gram	1000	6B	66.1	65.1	65.6	2.5 x 125

word2vec

Арифеметические операции

Вектора можно складывать!

v(king) - v(man) + v(woman) ≈ v(queen)

Table 1: *Examples of five types of semantic and nine types of syntactic questions in the Semantic-Syntactic Word Relationship test set.*

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

примеры из датасетов, на которых тестировали обученные вектора

Table 8: *Examples of the word pair relationships, using the best word vectors from Table 4 (Skip-gram model trained on 783M words with 300 dimensionality).*

Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

примеры работы модели

Векторное представление слов

word2vec

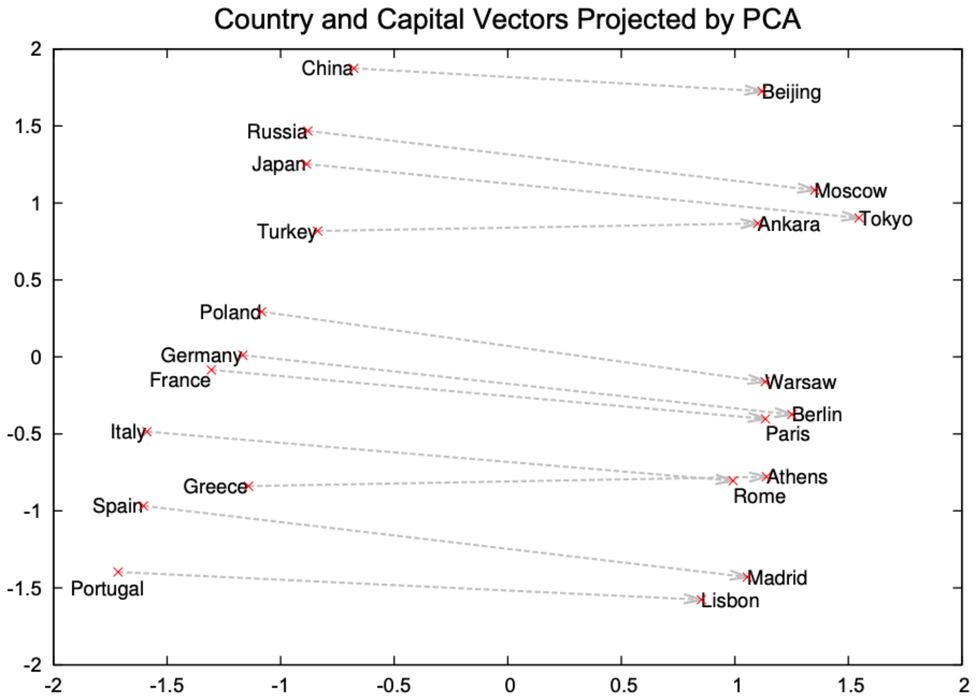


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Векторное представление слов

Преимущества:

- векторы отражают смысл слов
- маленькая размерность (по сравнению с размером словаря)
- можно дообучать вектора на новых текстах

Проблемы:

- не умеем работать с новыми словами
- плохо обрабатываем редкие слова
- похожие слова (**корол**ь, **корол**ю) считаются АБСОЛЮТНО разными словами
- не умеем обрабатывать пары слов: Соединённое Королевство, New York

Векторное представление слов

Негативное сэмплирование

$$softmax(z_1, \ldots, z_j, \ldots, z_V)_j = \frac{e^{z_j}}{\sum_k^V e^{z_k}} = \frac{e^{<x,w_j>}}{\sum_k^V e^{<x,w_k>}}$$

$$\log \sigma({v'_{w_O}}^\top v_{w_I}) + \sum_{i=1}^{k} \mathbb{E}_{w_i \sim P_n(w)} \left[\log \sigma(-{v'_{w_i}}^\top v_{w_I}) \right]$$

loss, negative sampling

V — размер словаря. Может быть огромным!

Negative sampling — давайте в качестве негативов возьмём 2-20 случайных элементов из словаря. Их и будем использовать в качестве негативов.
Частотные слова реже попадают в качестве негативов по сравнению с редкими

$U^{3/4}$, $P(w_i) = \frac{cnt(w_i)^{3/4}}{\sum cnt(w_j)^{3/4}}$, где $cnt(w_i)$ — количество вхождений слова w_i

Векторное представление слов

Частотные слова

Сэмплирование частотных слов

Постоянное обновление популярного слова даёт меньше информации, чем обновление редких слов.

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

$f(w_i)$ — частота слова. С вероятностью P(w) будем оставлять слово из таргета и контекста

В оригинальной статье t = 10^{-5} (подобран эмпирически)

Векторное представление слов

Method	Time [min]	Syntactic [%]	Semantic [%]	Total accuracy [%]
NEG-5	38	63	54	59
NEG-15	97	63	58	**61**
HS-Huffman	41	53	40	47
NCE-5	38	60	45	53
The following results use 10^{-5} subsampling				
NEG-5	14	61	58	60
NEG-15	36	61	61	**61**
HS-Huffman	21	52	59	55

Table 1: Accuracy of various Skip-gram 300-dimensional models on the analogical reasoning task as defined in [8]. NEG-k stands for Negative Sampling with k negative samples for each positive sample; NCE stands for Noise Contrastive Estimation and HS-Huffman stands for the Hierarchical Softmax with the frequency-based Huffman codes.

Векторное представление слов

Словосочетания

$$\text{score}(w_i, w_j) = \frac{\text{count}(w_i w_j) - \delta}{\text{count}(w_i) \times \text{count}(w_j)}.$$

Спасаем New York!

если score > порога, то оставляем пару слов

Идея: если пара слов встречается рядом друг с другом, чем по отдельности — добавим их как новое "слово" тоже (на самом деле новое слово состоит из двух слов)!

1 млн. слов -> 3 млн. слов

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Table 2: Examples of the analogical reasoning task for phrases (the full test set has 3218 examples). The goal is to compute the fourth phrase using the first three. Our best model achieved an accuracy of 72% on this dataset.

Векторное представление слов

Словосочетания

Method	Dimensionality	No subsampling [%]	10^{-5} subsampling [%]
NEG-5	300	24	27
NEG-15	300	27	42
HS-Huffman	300	19	**47**

Table 3: Accuracies of the Skip-gram models on the phrase analogy dataset. The models were trained on approximately one billion words from the news dataset.

Больше данных, больше векторы -> лучше качество (72%)

Векторное представление слов

Словосочетания

	NEG-15 with 10^{-5} subsampling	HS with 10^{-5} subsampling
Vasco de Gama	Lingsugur	Italian explorer
Lake Baikal	Great Rift Valley	Aral Sea
Alan Bean	Rebbeca Naomi	moonwalker
Ionian Sea	Ruegen	Ionian Islands
chess master	chess grandmaster	Garry Kasparov

Table 4: Examples of the closest entities to the given short phrases, using two different models.

Czech + currency	Vietnam + capital	German + airlines	Russian + river	French + actress
koruna	Hanoi	airline Lufthansa	Moscow	Juliette Binoche
Check crown	Ho Chi Minh City	carrier Lufthansa	Volga River	Vanessa Paradis
Polish zolty	Viet Nam	flag carrier Lufthansa	upriver	Charlotte Gainsbourg
CTK	Vietnamese	Lufthansa	Russia	Cecile De

Table 5: Vector compositionality using element-wise addition. Four closest tokens to the sum of two vectors are shown, using the best Skip-gram model.

Векторное представление

fasttext (2016)

Fasttext — https://arxiv.org/pdf/1607.04606

Идея — word2vec, в котором слова строим для n-грам букв, из которых он состоит

Вектор слова — сумма векторов его n-gram

яблоко = [<яб, бло, лок, око, ко>, <яблоко>] (< и > — специальный символ начала или конца слова)

Обучали на мультиязычной википедии

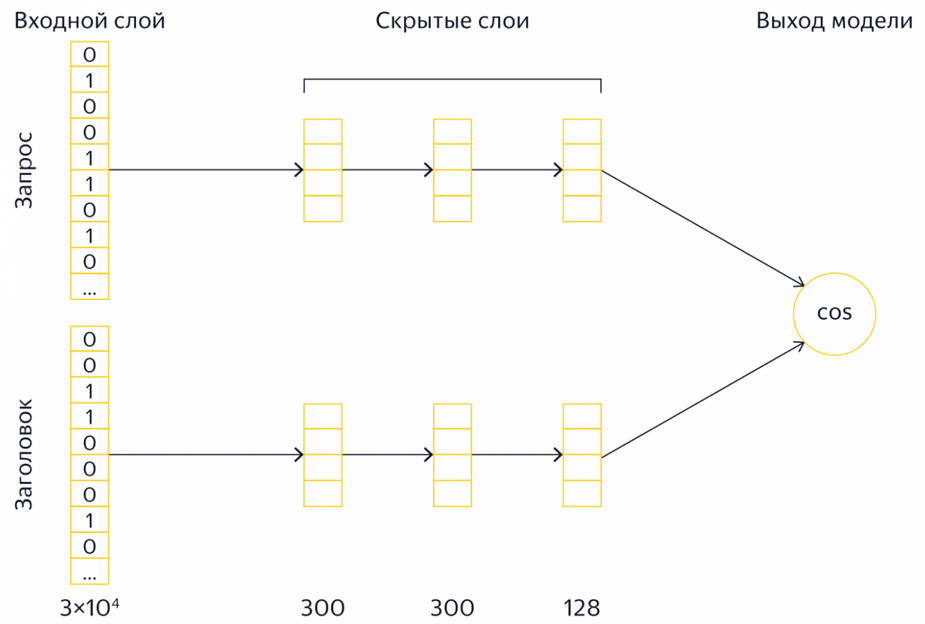
Векторное представление

dssm (2013)

Deep Structured Semantic Model (microsoft)

На вход триграммы слов из запроса и заголовка документа

Клик в поиске bing — положительный пример, отсутствие клика — отрицательный



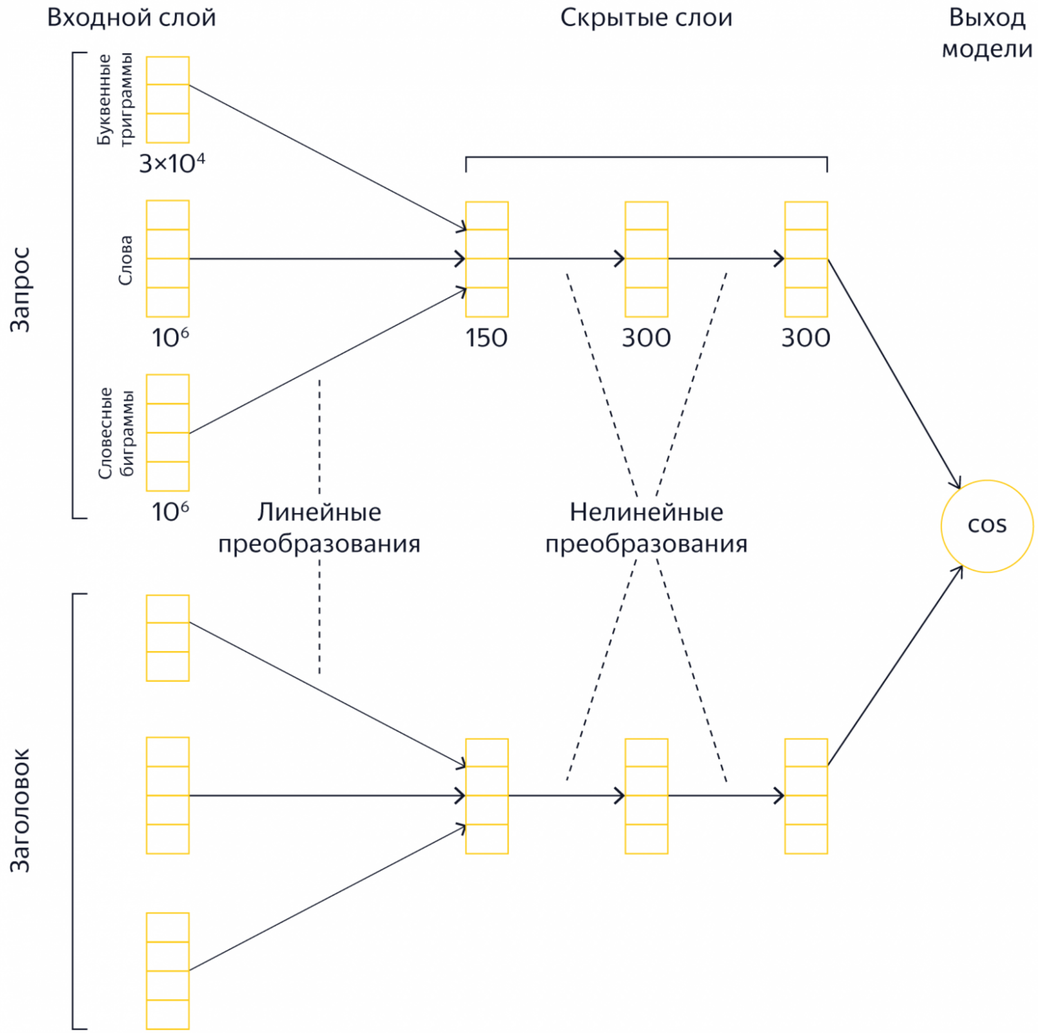
Векторное представление

Модификация dssm

Искусственный интеллект в поиске. Как Яндекс научился применять нейронные сети, чтобы искать по смыслу, а не по словам — 2016

Обучение:

1. Кликнутый документ против полностью случайного
2. Кликнутый документ против случайного, который содержит хотя бы 1 слово из запроса
3. Кликнутый документ против hard негатива



Векторное представление

Dssm в поиске Яндекса (2016)

В качестве примера возьмем запрос [келлская книга] и посмотрим, какое значение принимают факторы на разных заголовках. Для контроля добавим в список заголовков явно нерелевантный результат.

Заголовок страницы	BM25	Нейронная модель
келлская книга википедия	0.91	0.92
ученые исследуют келлскую книгу вокруг света	0.88	0.85
book of kells wikipedia	0	0.81
ирландские иллюстрированные евангелия vii viii вв	0	0.58
икеа гипермаркеты товаров для дома и офиса ikea	0	0.09

Теперь давайте посмотрим, как будут себя вести наши факторы, если мы переформулируем запрос, не меняя его смысла: [евангелие из келлса].

Заголовок страницы	BM25	Нейронная модель
келлская книга википедия	0	0.85
ученые исследуют келлскую книгу вокруг света	0	0.78
book of kells wikipedia	0	0.71
ирландские иллюстрированные евангелия vii viii вв	0.33	0.84
икеа гипермаркеты товаров для дома и офиса ikea	0	0.10

Как обучить модель

Как получить эмбэддинг предложения?

Простой способ — усреднить все слова (мешок слов)

Над эмбэддингами текста можно докинуть несколько слоёв нейросети или подать их на вход град. бустингу (или хоть даже лог. регрессии).

Векторное представление всего

site2vec

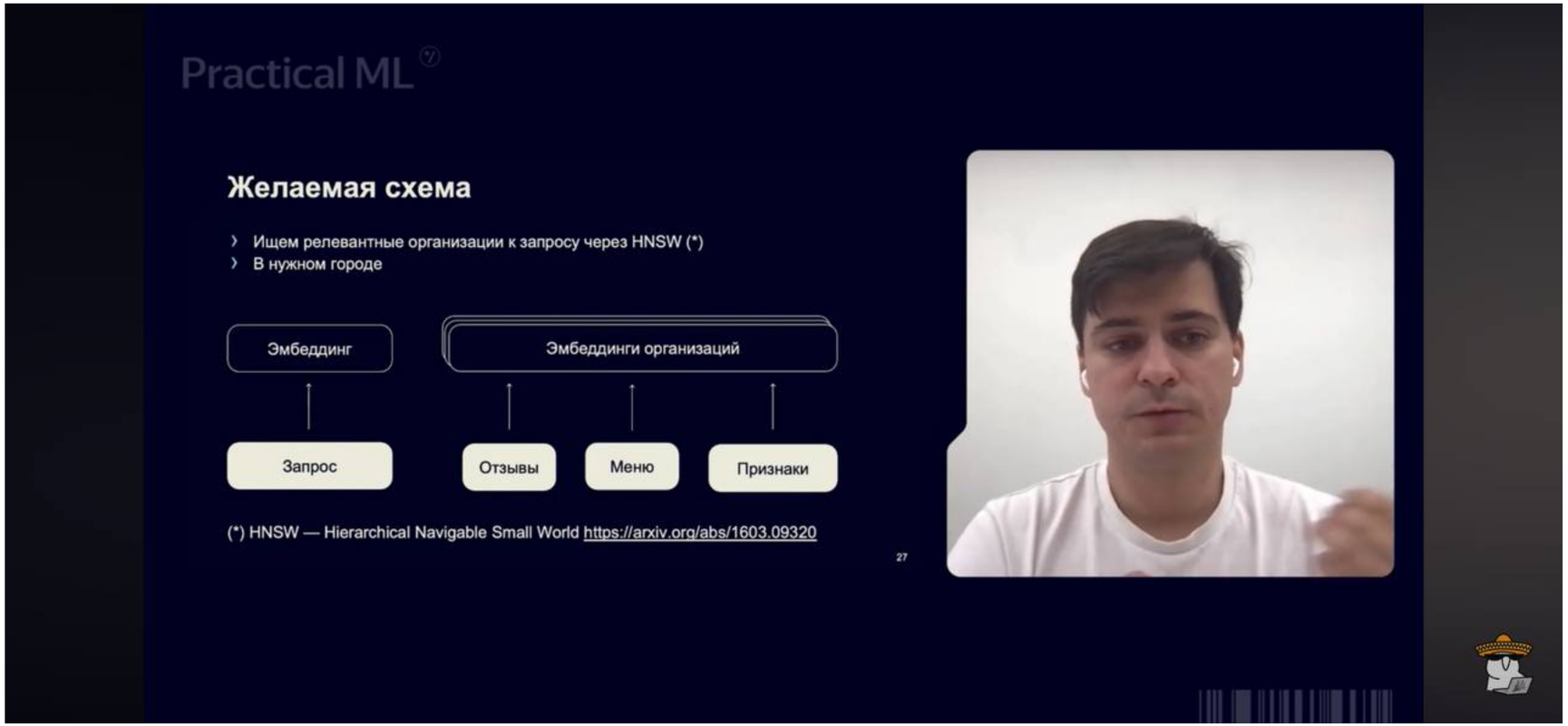
Давайте посчитаем эмбэддинги сайтов!

Предположим, мы работаем в большой корпорации и у нас есть история посещения пользователя

В качестве слова используем url сайта (хост).
В качестве текста — историю посещения пользователя

Векторное представление всего

Поиск по организациям



Андрей Данильченко, Никита Киселев | Поиск организаций по смыслу

Спасибо!