

# Глубинное обучение

Векторное представление текста

Михаил Гущин

[mhushchyn@hse.ru](mailto:mhushchyn@hse.ru)

НИУ ВШЭ, 2024



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# Генерация текста

A screenshot of a text generation interface. At the top, a user profile picture shows a person with short hair, and the text reads: "give some examples of the types of writing AI can do, and why that will transform businesses in a paragraph." To the right of the text is a small edit icon. Below this, an AI-generated response is shown, preceded by an AI icon (a green square with a white swirl). The response text is: "AI can be used to write a wide range of business-related documents, including reports, proposals, and marketing materials. By utilizing AI's ability to generate ideas and provide information, businesses can quickly and easily create high-quality written materials. This can save time and resources, allowing businesses to focus on other important tasks. Additionally, AI can help ensure accuracy and consistency in writing, improving the overall professionalism and effectiveness of business documents. The use of AI in writing will transform businesses by streamlining the writing process and improving the quality and accuracy of written materials." To the right of the response are a thumbs-up icon and a thumbs-down icon.

# Машинный перевод

Яндекс Браузер

Яндекс.Браузер обновился. Версия 21.8.2

1

Hi! I'm David and I lead the NLP team at Yandex

Закадровый перевод видео с английского

Нейросети Яндекса научились сами переводить и озвучивать видео на английском языке. Пока — не везде, но уже скоро любой ролик на английском можно будет смотреть на русском.

Сразу попробовать новую функцию можно [по ссылке](#).

Как включить перевод видео?

Подписывайтесь на новости Яндекс.Браузера:

Дзен ВКонтакте Твиттер Телеграм

Перевод не доступен для видео, у которых есть технические средства защиты авторских прав (DRM).

# Голосовые помощники

The screenshot shows the Yandex.Alice app's main screen. At the top center is the Alice logo, a stylized white egg inside a purple circle. Below it is a large, rounded white button containing the text "Привет, я Алиса!" (Hello, I'm Alice!). The background is a solid purple color. In the center, the text "Я готова помочь" (I'm ready to help) is displayed in white. Below this, there are six horizontal cards, each with an icon and text: 1. "Определить песню" (Identify song) with a musical note icon. 2. "Узнать, что на фото" (Know what's in the photo) with a camera icon. 3. "Включить сказку" (Turn on story) with a book icon. 4. "Одеться по погоде" (Dress according to the weather) with a person icon. 5. "Поиграть" (Play) with a game controller icon. 6. "Построить маршрут" (Build route) with a map pin icon. 7. "Вызвать такси" (Call a taxi) with a car icon. 8. "Найти нужное место" (Find the right place) with a location pin icon. 9. "Купить на Беру" (Buy on Beru) with a shopping box icon.

Привет, я Алиса!

Я готова помочь

- Определить песню >
- Узнать, что на фото >
- Включить сказку >
- Одеться по погоде >
- Поиграть >
- Построить маршрут >
- Вызвать такси >
- Найти нужное место >
- Купить на Беру >

# Векторизация текстов

# Задача классификации текстов

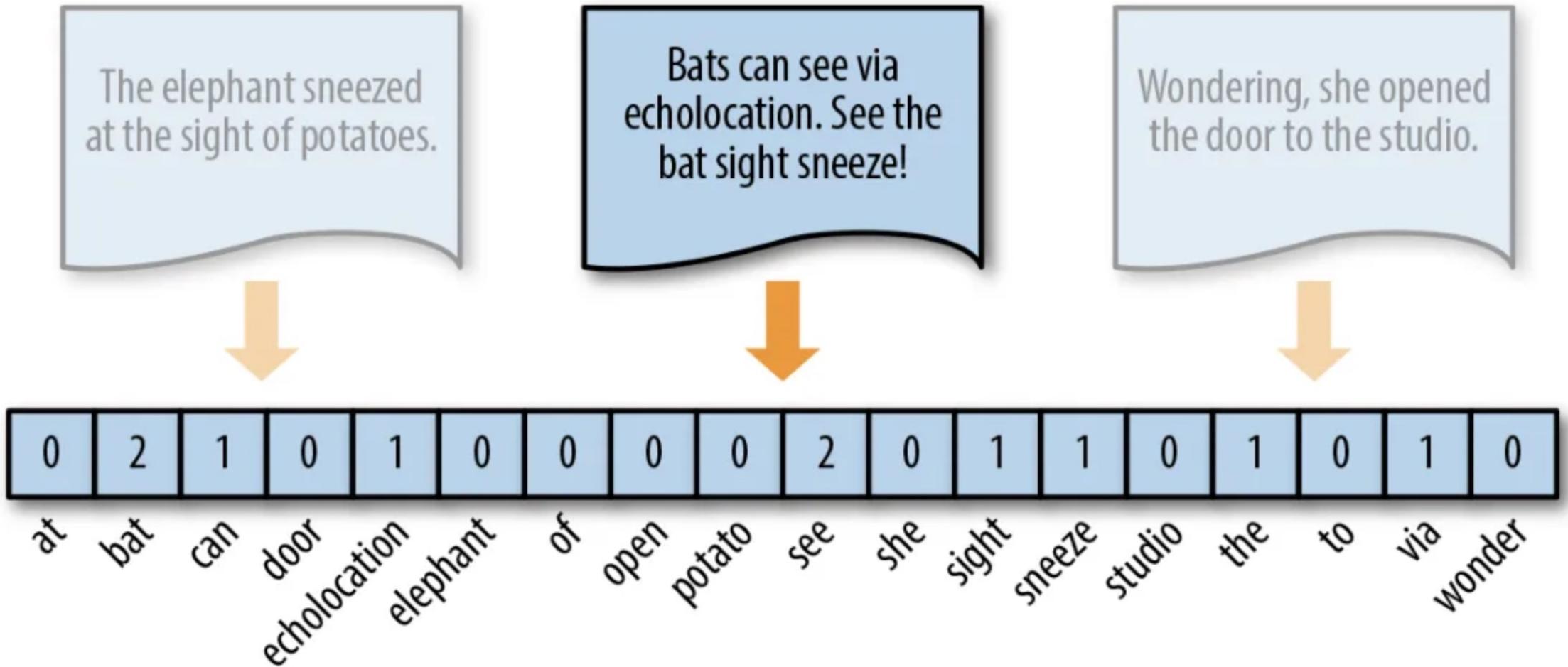
**"text":** "Двое налетчиков совершили нападение на охранника банка \"ЦентрКредит\" в Алматы и завладели его оружием, сообщает пресс-служба ДВД Алматы.\n\"Сегодня, в 08.10 в Центр оперативного управления ДВД города Алматы поступило сообщение о нападении на охранника одного из отделений банка \"Центркредит\" в Алмалинском районе Алматы, в результате чего двое неустановленных лиц завладели его оружием\", — говорится в сообщении.\nПо неподтвержденным данным, чрезвычайное происшествие случилось в отделении, которое находится на улице Маметовой, между улицами Сейфуллина и Дзержинского.\nКак информирует ведомство, в настоящее время по городу объявлен спецплан \"Сирена\". Отрабатывается комплекс оперативных мероприятий, которые направлены на розыск и задержание преступников.\n",  
**"sentiment":** "negative»

**"text":** "АСТАНА. КАЗИНФОРМ - Карим Масимов провел совещание по вопросам деятельности Актауского международного морского торгового порта. Read on the original site", "id": 2085,  
**"sentiment":** "positive"

# Мешок слов (Bag of Words)

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

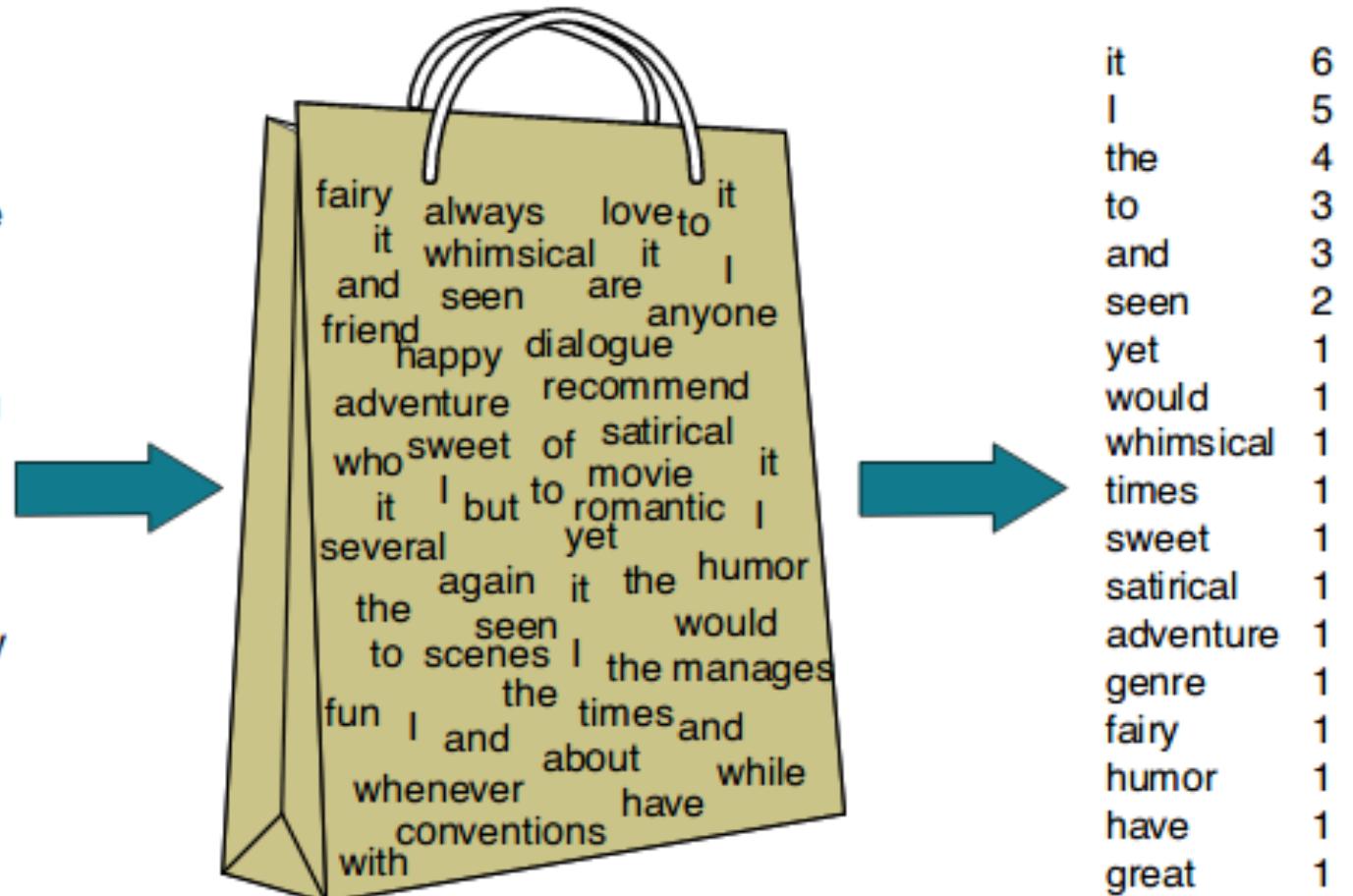
# Мешок слов (Bag of Words)



Источник: <https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5>

# Мешок слов (Bag of Words)

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



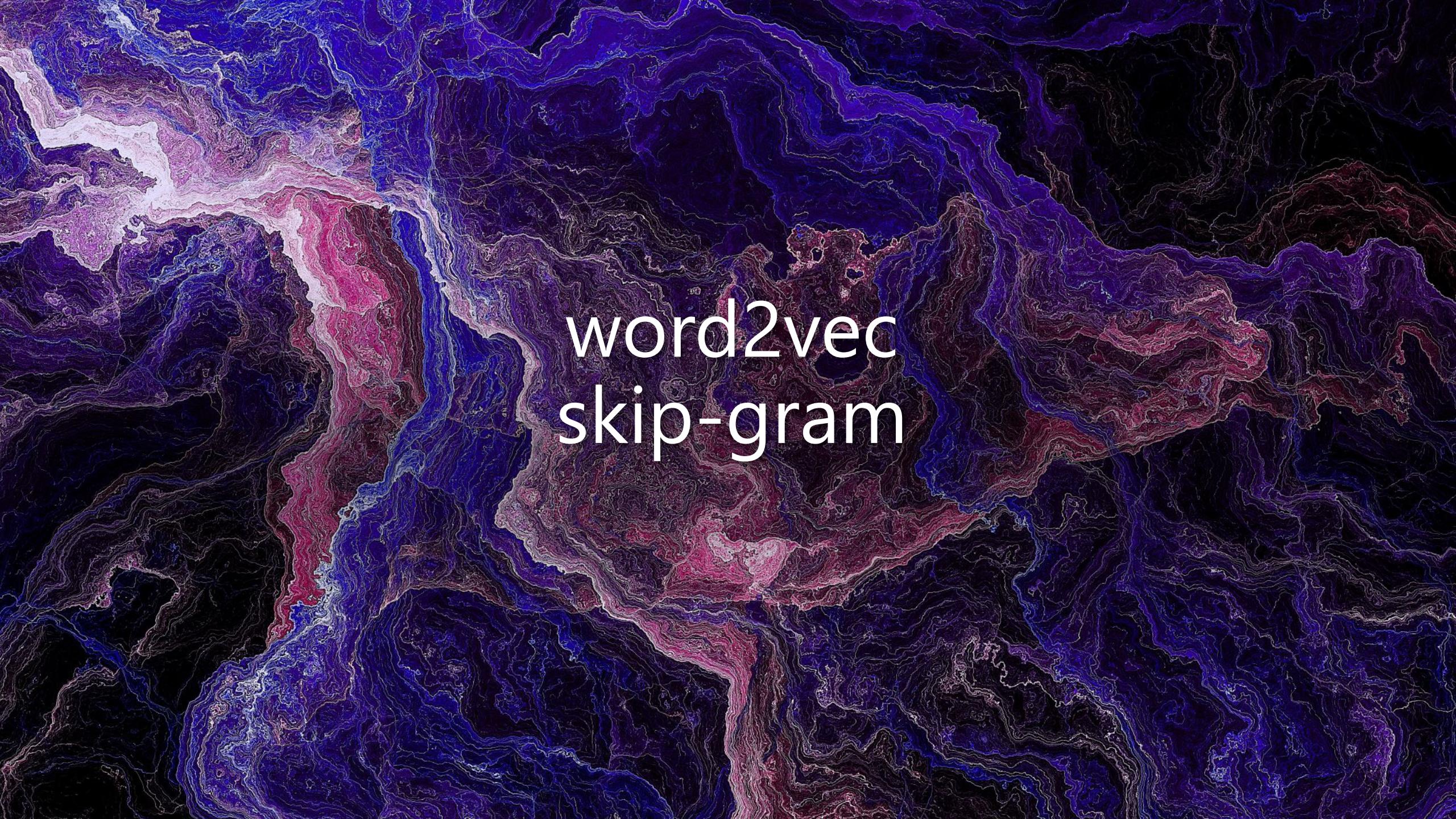
Источник: <https://dudeperf3ct.github.io/lstm/gru/nlp/2019/01/28/Force-of-LSTM-and-GRU/>

# Алгоритм решения

- ▶ Пусть есть набор текстов
- ▶ Делим каждый текст на токены и приводим слова к нормальной форме
- ▶ Используем Bag Of Words или TF-IDF, чтобы получить векторное представление каждого текста
- ▶ Используем эти вектора как вектора признаков
- ▶ Используем любые модели классификации и регрессии для анализа текстов

# Мешок слов (Bag of Words)

- ▶ Слишком много признаков
- ▶ Похожие по смыслу тексты могут иметь разные представления
- ▶ Можно ли иначе получить векторные представления слов?

The background of the image is a dark, abstract pattern resembling a topographic map or a complex liquid. It features intricate, swirling lines in shades of purple, blue, and white against a black background, creating a sense of depth and motion.

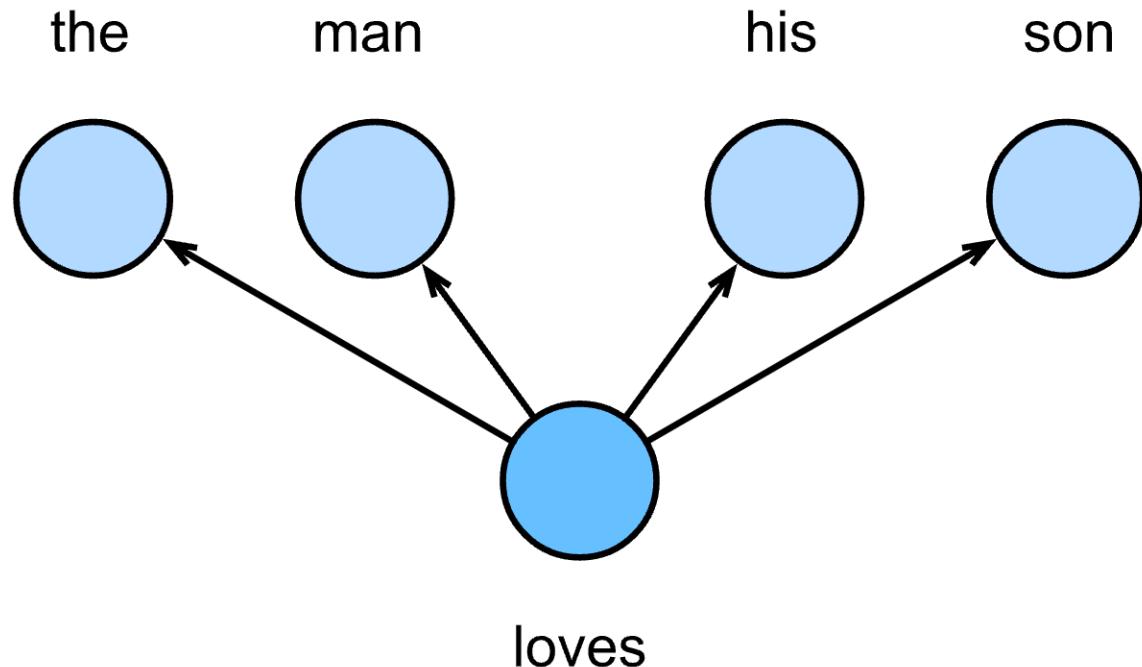
word2vec  
skip-gram

# Пример

The man **loves** his son

- ▶ Назовем "loves" – центральным словом
- ▶ "the", "man", "his", "son" – контекстом
- ▶ Word2vec предполагает, что вектора центрального слова и контекстных близки

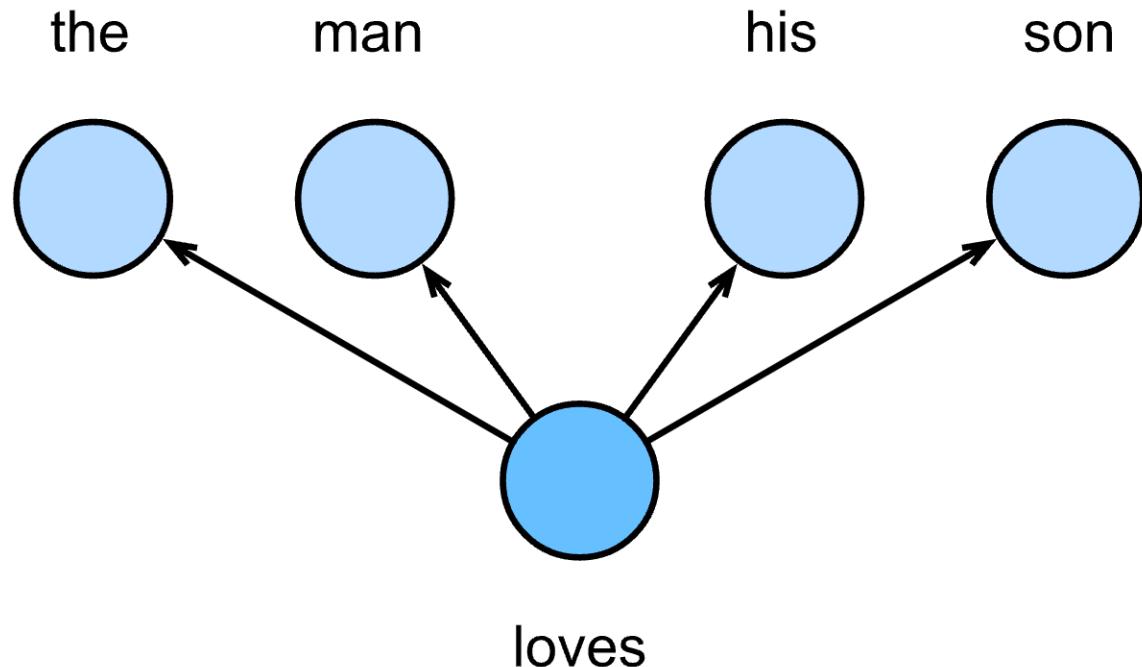
# Модель skip-gram



- ▶ Skip-gram модель предполагает, что контекст зависит от центрального слова:

$$P(\text{"the"}, \text{"man"}, \text{"his"}, \text{"son"} \mid \text{"loves"})$$

# Модель skip-gram



- ▶ Предположим, что распределение условно независимое:

$$\begin{aligned} P(\text{"the"}, \text{"man"}, \text{"his"}, \text{"son"} \mid \text{"loves"}) &= \\ P(\text{"the"} \mid \text{"loves"}) \times P(\text{"man"} \mid \text{"loves"}) \times P(\text{"his"} \mid \text{"loves"}) \times P(\text{"son"} \mid \text{"loves"}) \end{aligned}$$

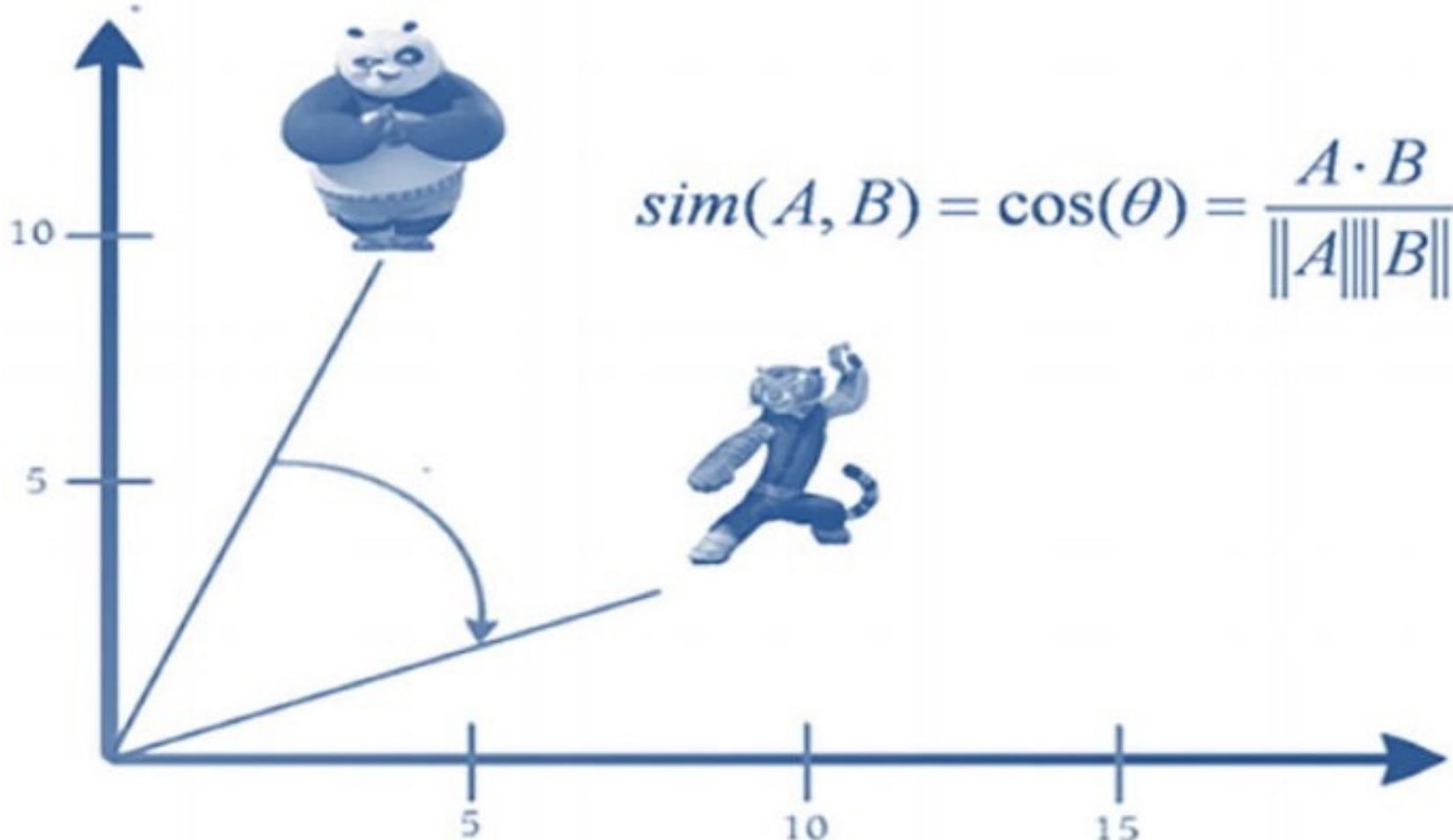
# Модель skip-gram

- ▶ Обозначим центральное слово ("loves") как  $w_c$
- ▶ Обозначим контекстное слово ("man") как  $w_o$
- ▶ Для слов будем искать векторные представления:  $u_o, v_c \in R^d$
- ▶ Тогда опишем вероятности как:

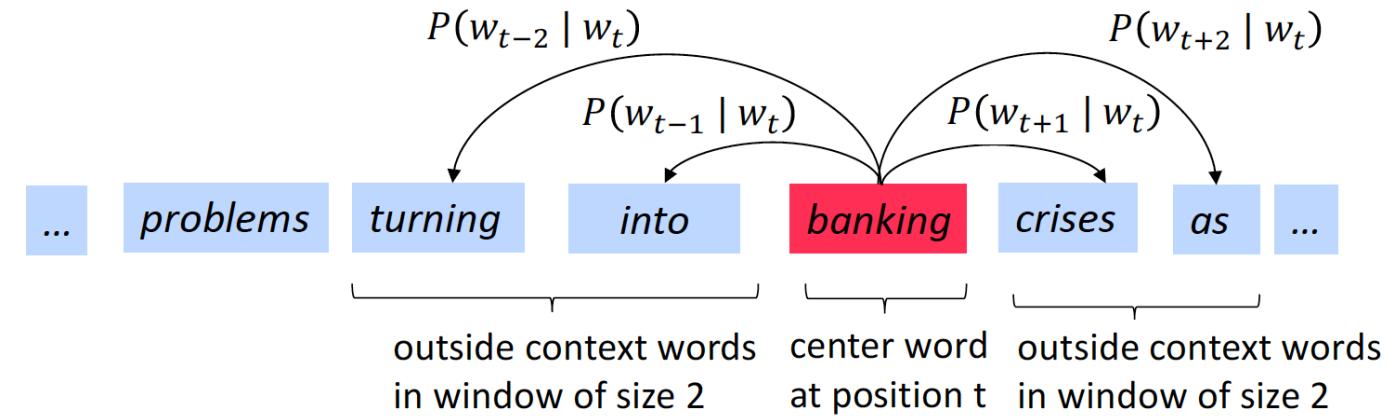
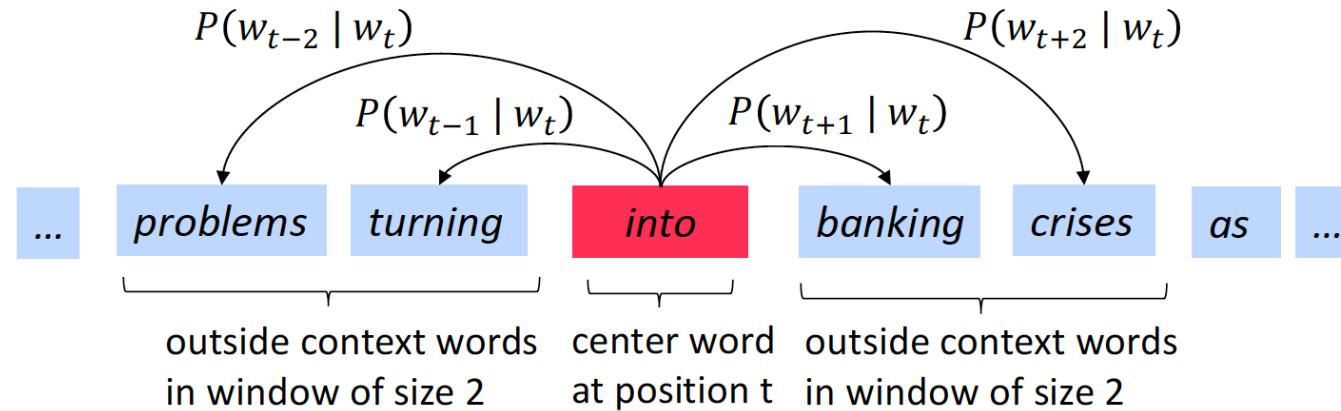
$$P(w_0|w_c) = \frac{\exp(u_o^T v_c)}{\sum_{i \in V} \exp(u_i^T v_c)}$$

Где  $V$  – текст длиной  $T$  слов

# Почему uv?



# Пример



# Модель skip-gram

- ▶ Тогда функция потерь word2vec модели запишем как:

$$L = - \sum_{t=1}^T \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \log P(w_{t+j} | w_t) \rightarrow \min$$

- ▶ Т.е. сумма логарифмов вероятностей для всех последовательностей слов длиной  $2m$
- ▶ Есть связь с функцией потерь для классификации? ☺

# Пример текста

## Text Corpus

Window Size = 2

The	quick	brown
-----	-------	-------

fox jumps over the red dog

The	quick	brown	fox
-----	-------	-------	-----

jumps over the red dog

The	quick	brown	fox	jumps
-----	-------	-------	-----	-------

over the red dog

The	quick	brown	fox	jumps	over
-----	-------	-------	-----	-------	------

the red dog

## Training Samples

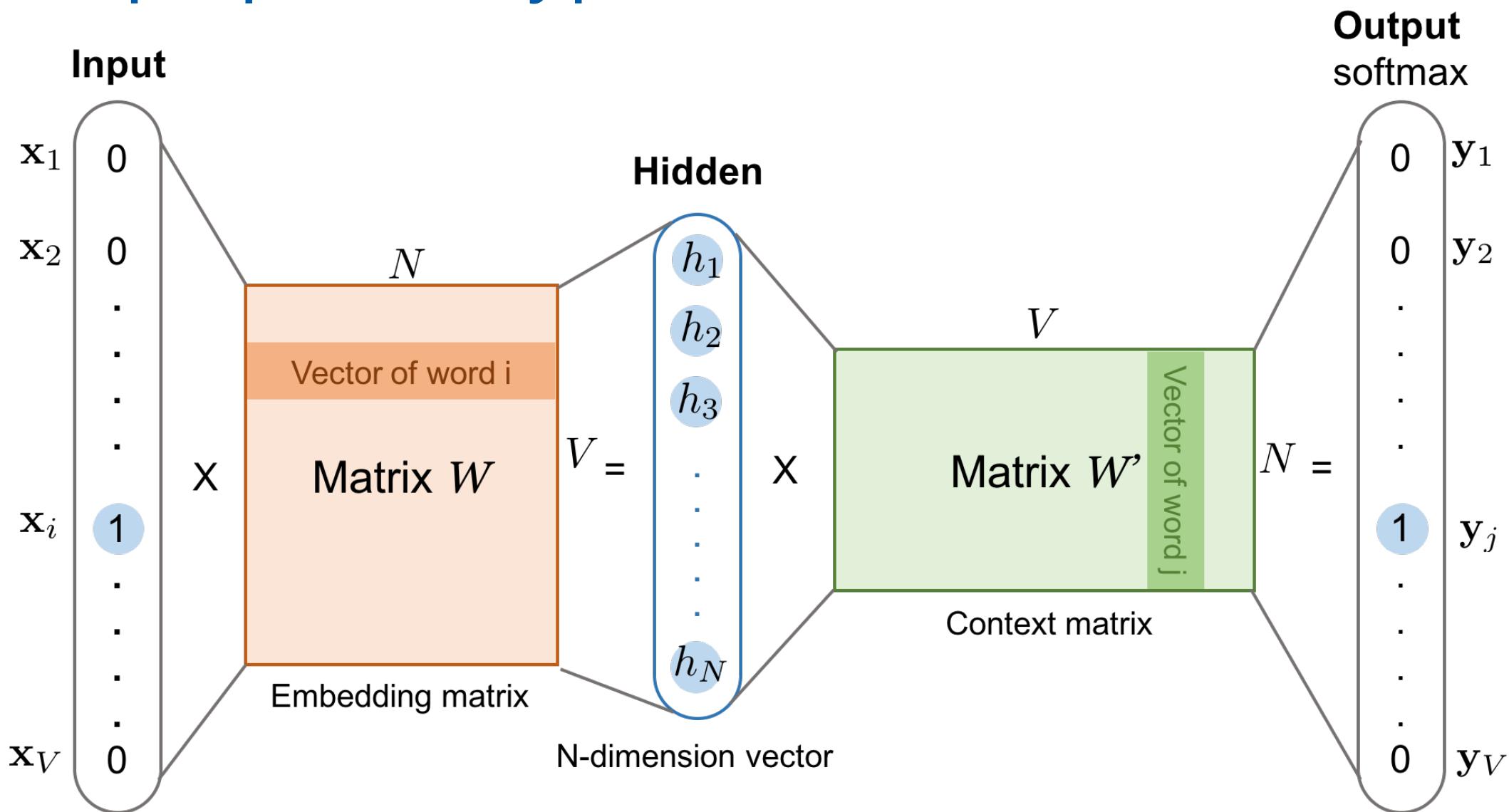
( The , quick )  
( The , brown )

( quick, the )  
( quick , brown )  
( quick, fox )

( brown , the )  
( brown , quick )  
( brown , fox )  
( brown , jumps )

( fox , quick )  
( fox , brown )  
( fox , jumps )  
( fox , over )

# Пример архитектуры



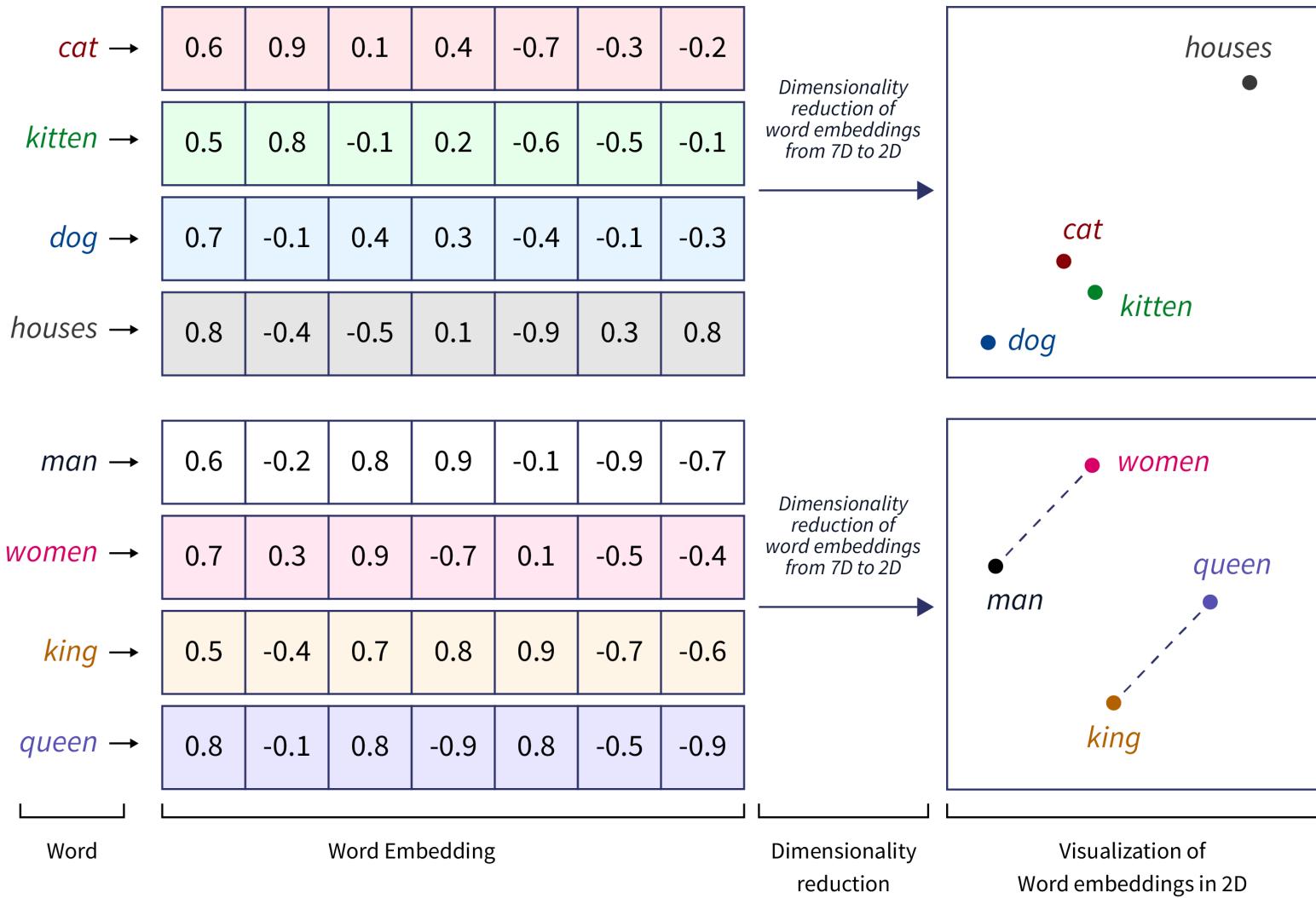
# Вопрос

- ▶ Матрицы  $W$  и  $W'$  - разные матрицы. Почему?

# Вопрос

- ▶ Матрицы  $W$  и  $W'$  - разные матрицы. Почему?
- ▶ Пример: The man **loves** his son
- ▶  $P(\text{loves}|\text{loves}) \sim \exp(u^T v)$ , где  $u \in W'$ ,  $v \in W$
- ▶ Если  $W = W'$ , то  $u = v$  и  $P(\text{loves}|\text{loves}) \sim \exp(v^T v)$
- ▶ Но в тексте редко встречается комбинация «**loves loves**», поэтому  $P(\text{loves}|\text{loves}) \approx 0$ , что невозможно

# Пример эмбеддингов (векторов)



# Ускорение обучения

- ▶ При подсчете условных вероятностей

$$P(w_0|w_c) = \frac{\exp(u_o^T v_c)}{\sum_{i \in V} \exp(u_i^T v_c)}$$

Требуется суммирование по всему словарю  $V$

- ▶ Это дорогая операция во время обучения
- ▶ Как ускорить обучение модели?

# Negative Sampling

- ▶ Рассмотрим  $\log P(w_0|w_c)$ :

$$\log P(w_0|w_c) = u_o^T v_c - \log \sum_{i \in V} \exp(u_i^T v_c)$$

- ▶ Сумму нужно как-то уменьшить 😊
- ▶ Воспользуемся идеей из бинарной классификации!

# Negative Sampling

- ▶ Определим вероятность того, что два слова встречаются вместе (класс 1):

$$P(D = 1 | w_c, w_o) = \sigma(u_o^T v_c)$$

- ▶ Тогда вероятность, что два слова не встречаются вместе (класс 0):

$$P(D = 0 | w_c, w_o) = 1 - \sigma(u_o^T v_c)$$

- ▶ Случайным образом выберем К негативных примеров (класс 0)
- ▶ Тогда переопределим  $\log P(w_0 | w_c)$ :

$$\log P(w_0 | w_c) = \log \sigma(u_o^T v_c) - \sum_{k=1}^K \log(1 - \sigma(u_k^T v_c))$$

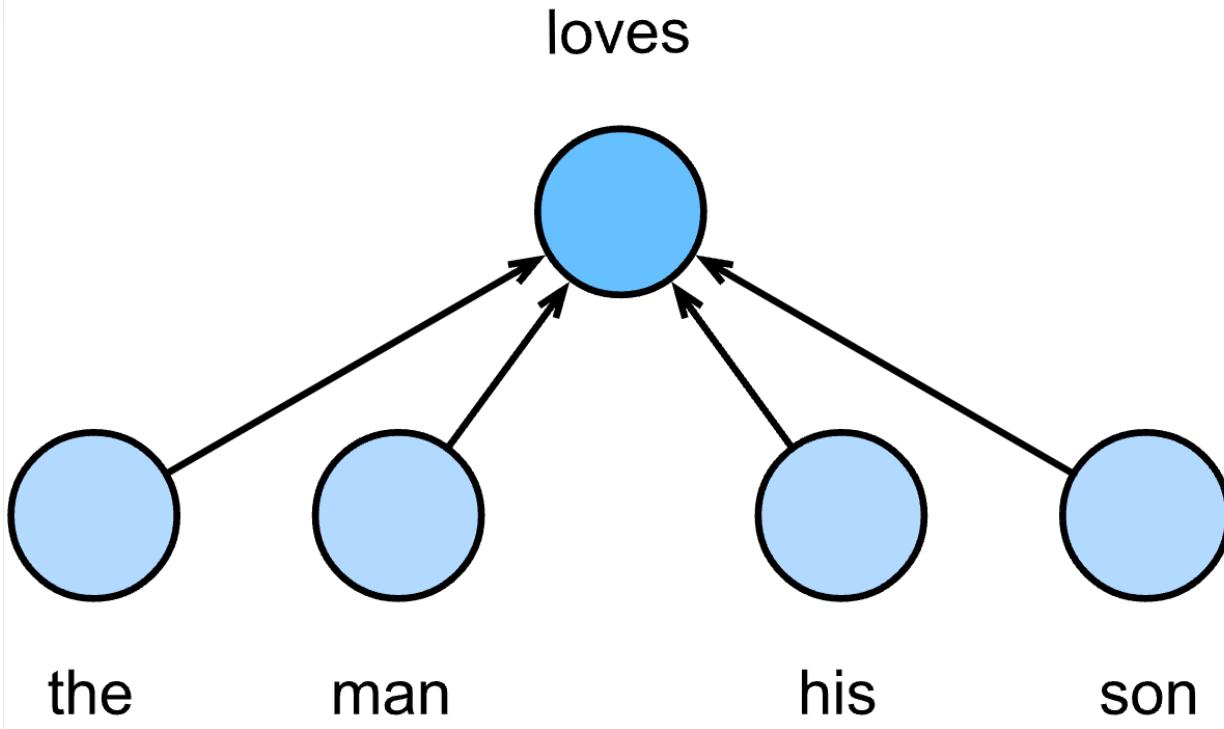
# Negative Sampling

- ▶ К – гиперпараметр, который мы выбираем сами
- ▶ Если К маленький – меньше вычислений для функции потерь, но и меньше штраф за негативные примеры



word2vec  
CBOW

# Модель continuous bag of words (CBOW)



- ▶ СВОУ модель предполагает, что центральное слово зависит от контекста:

$$P(\text{"loves"} \mid \text{"the"}, \text{"man"}, \text{"his"}, \text{"son"})$$

# Модель CBOW

- ▶ Обозначим центральное слово ("loves") как  $w_c$
- ▶ Обозначим контекстное слово ("man") как  $w_{oi}$
- ▶ Для слов будем искать векторные представления:  $u_{oi}, v_c \in R^d$
- ▶ Тогда опишем вероятности как:

$$P(w_c | w_{o1}, w_{o2}, \dots, w_{o2m}) = \frac{\exp(u_c^T \bar{v}_o)}{\sum_{i \in V} \exp(u_i^T \bar{v}_o)}$$

$$\bar{v}_o = \frac{1}{2m} (v_{o1} + v_{o2} + \dots + v_{o2m})$$

Где  $V$  – текст длиной  $T$  слов

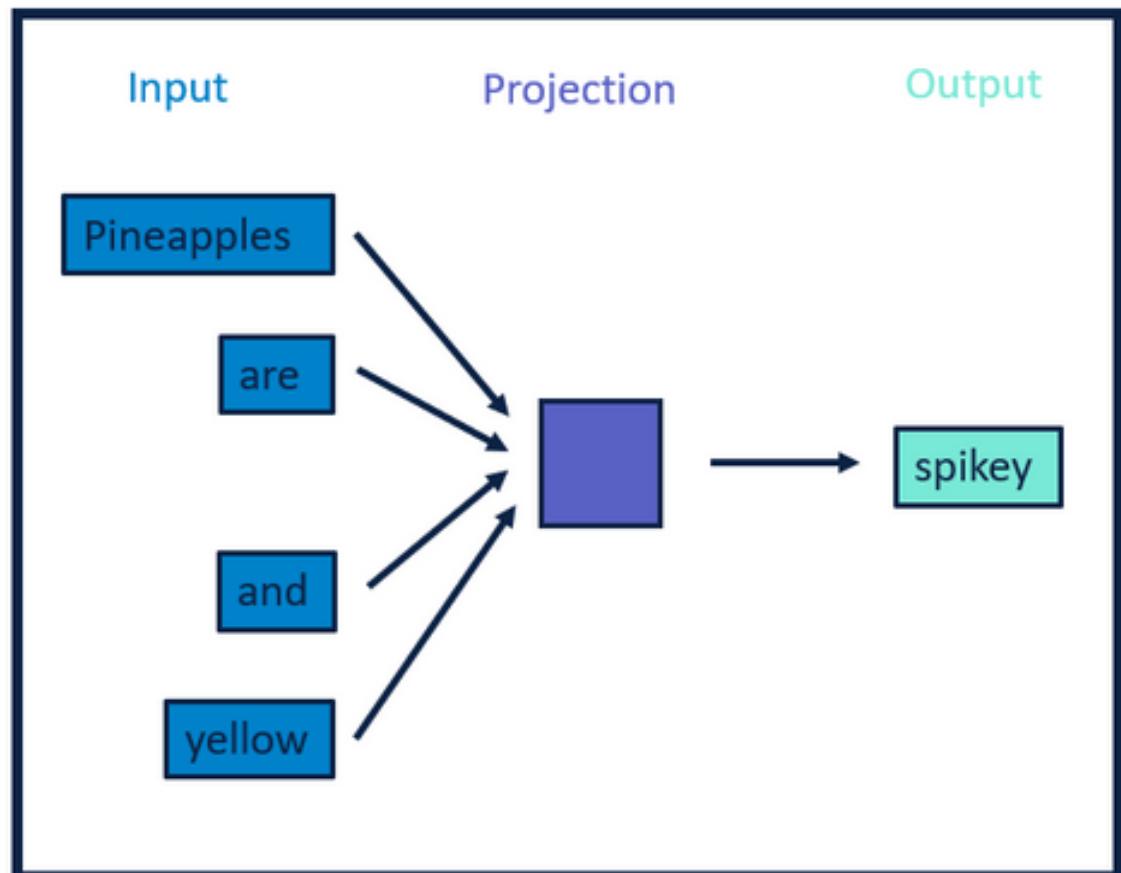
# Модель CBOW

- ▶ Тогда функция потерь word2vec модели запишем как:

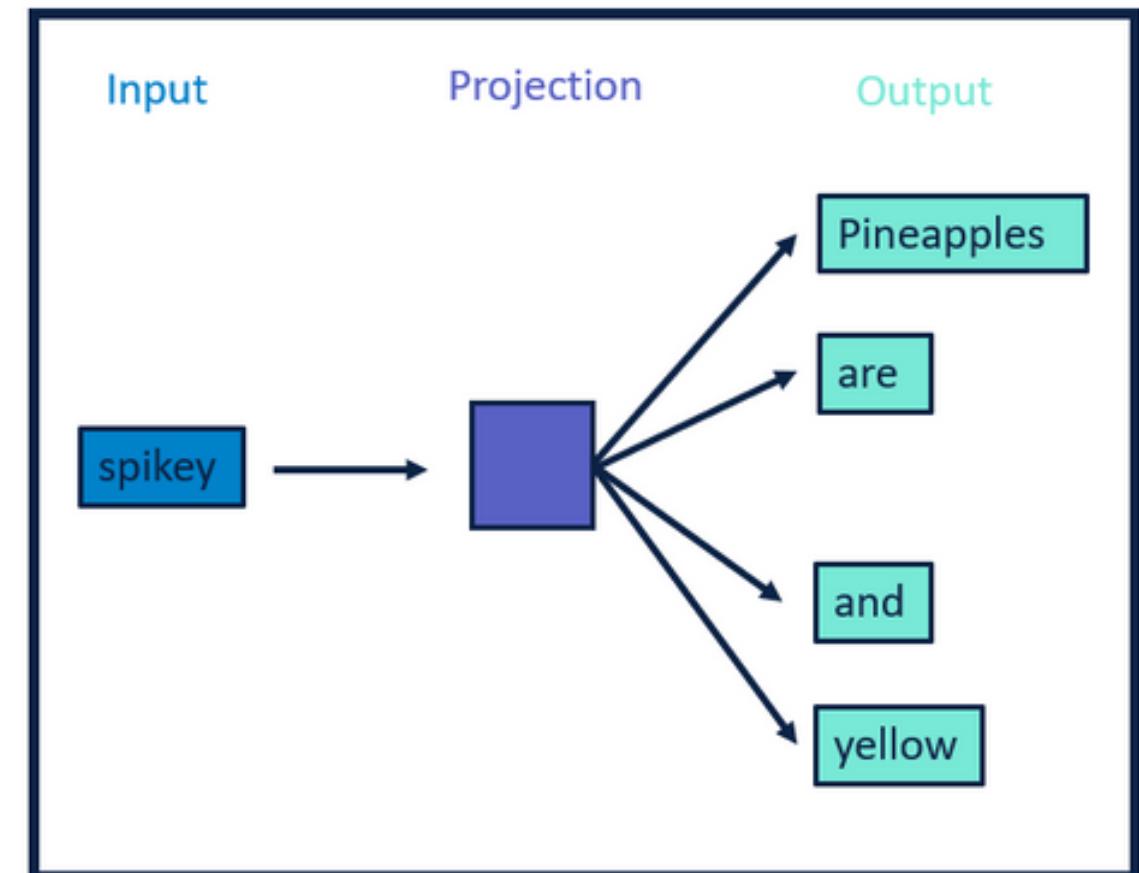
$$L = - \sum_{t=1}^T \log P(w_c | w_{o1}, w_{o2}, \dots, w_{o2m}) \rightarrow \min$$

- ▶ Т.е. сумма логарифмов вероятностей для всех последовательностей слов длиной  $2m$

# Word2vec



CBOW



Skip-gram

# Проблемы word2vec

- ▶ Не учитывает структуру слов (окончания, суффиксы, приставки)
- ▶ Не использует никакой априорной информации о разных формах одного слова



FastText

---

# Модель FastText

- ▶ Заменим каждое слово на  $n$  токенов
  - «руслан» -> (<руслан>, <ру, рус, усл, сла, лан, ан>)
- ▶ Обучаем векторы токенов:  $u_1, u_2, \dots, u_n$
- ▶ Вектор первоначального слова  $z_w = \sum_{i=1}^n u_i$
- ▶ Архитектура и обучение как в word2vec

# Пример



# Demo

## Word2Vec Demo

Training word embeddings in the browser using [TensorFlow.js](#)

**Place your text below:**

Hey there, Delilah. What's it like in New York city? I'm a thousand miles away. But, girl, tonight you look so pretty. Yes, you do. Times Square can't shine as bright as you. I swear it's true.

**Specify model and parameters to generate dataset:**

Model Skip-gram

The Word2Vec (Skip-gram) model trains words to predict their context / surrounding words.

Window size 3

Negative sampling?

**Generate dataset**

<https://remykarem.github.io/word2vec-demo/>