Обработка естественного языка

Сергей В. Аксёнов,

к.т.н., доцент каф. Автоматизации обработки информации,

Томский университет систем управления и радиоэлектроники

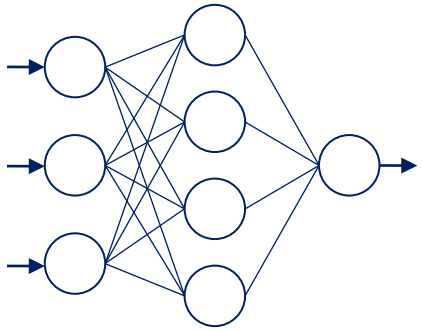
Томск-2024

Задачи обработки естественного языка

Задача	Описание
Токенизация	Деление текстового корпуса на неделимые единицы
Устранение неоднозначности слов	Определение правильного значения слова
Распознавание именованных сущностей	Выделение сущностей (имен, компаний, лекарств, городов, и т.д.) из текстового корпуса
Морфологическая разметка	Определение частей речи в предложении и их аннотирование
Классификатор предложений	Отнесение текстов к определенным классам
Генерация языка	Генерация новых текстов с помощью примеров
Системы вопросов и ответов	Чат-боты, поиск информации и представление знаний
Машинный перевод	Преобразование текстов с одного языка в другой

Примеры - 1

Image-to-Text



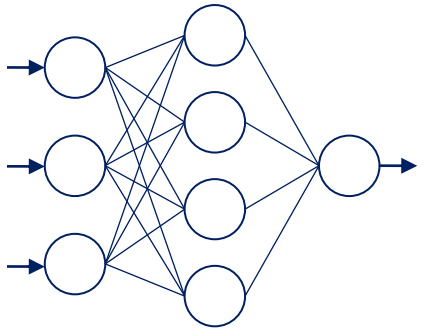
«A woodpecker sits on a fallen birch tree.»

Вход

Выход

Классификация текстов

"This was definitely the best bag I’ve bought."



«A positive review»

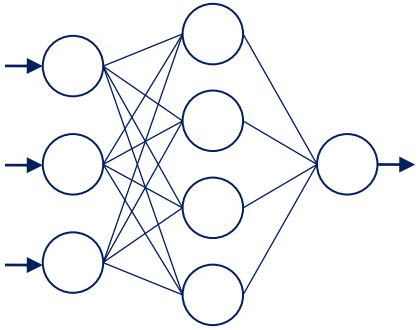
Вход

Выход

Примеры - 2

Text-to-Image

Cat barista, making coffee.

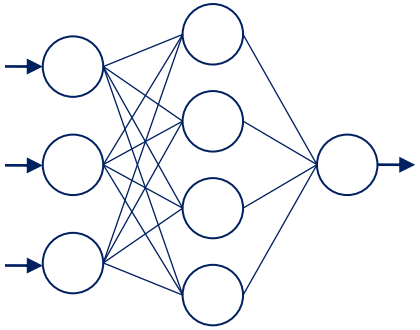


Вход

Выход

Text-to-3D

A blue poison-dart frog sitting on a water lily.



Вход

Выход

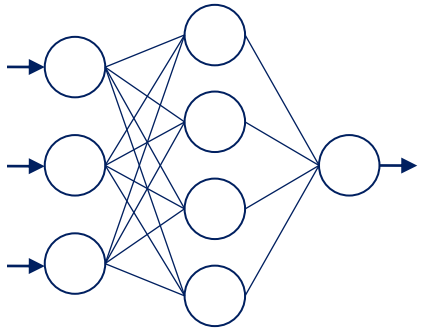
Source: https://www.midjourney.com/

Source: https://research.nvidia.com/labs/dir/magic3d/

Примеры - 3

Перевод текста

Cats have excellent night vision and can see at one sixth the light level required for human vision.



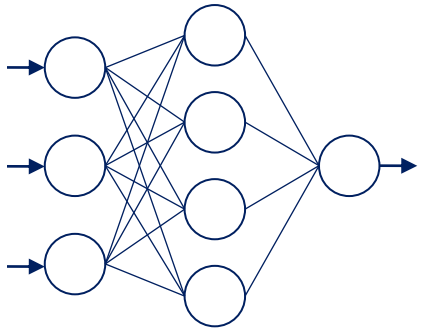
«Кошки обладают превосходным ночным зрением и могут видеть на уровне одной шестой уровня освещенности, необходимого для человеческого зрения.»

Вход

Выход

Генерация текста

Describe this fantastic game.



«One of the best moments in the game comes after a lengthy puzzle section that requires you to transport between different time periods to alter the present.»

Вход

Выход

Source: https://openai.com/

Source: https://translate.google.com/

Кодирование текстовых данных

Процесс преобразование текстов на естественном языке в числовые значения, которые компьютер может понимать.

Вход: Blood is composed of blood cells suspended in blood plasma.

Кодирование: 'blood' → 0, 'is' → 1, 'composed' → 2, 'of' → 3, 'cells' → 4, 'suspended' → 5, 'in' → 6, 'plasma' → 7

Выход: 0, 1, 2, 0, 3, 4, 5, 6, 0, 7

Токенизация

“From my earliest youth I realized that my nature was a mass of contradictions.”
(“And Then There Were None”, Agatha Christie)

1. Буквы: 'F', 'r', 'o', 'm', ' ', 'm', 'y', ' ', 'e', 'a', 'r', 'l', 'i', 'e', 's', 't’, …
2. Части слова: 'Fr', 'o', 'm', ' ', 'my', ' ', 'ea', 'r', 'l', 'ie', 'st’, …
3. Слова: 'From', 'my', 'earliest', 'youth', 'I', 'realized', 'that', 'my', 'nature’, …
4. N-граммы (Несколько соседних слов): 'From my', 'my earliest', 'earliest youth', 'youth I', 'I realized', 'realized that', 'that my', 'my nature', …

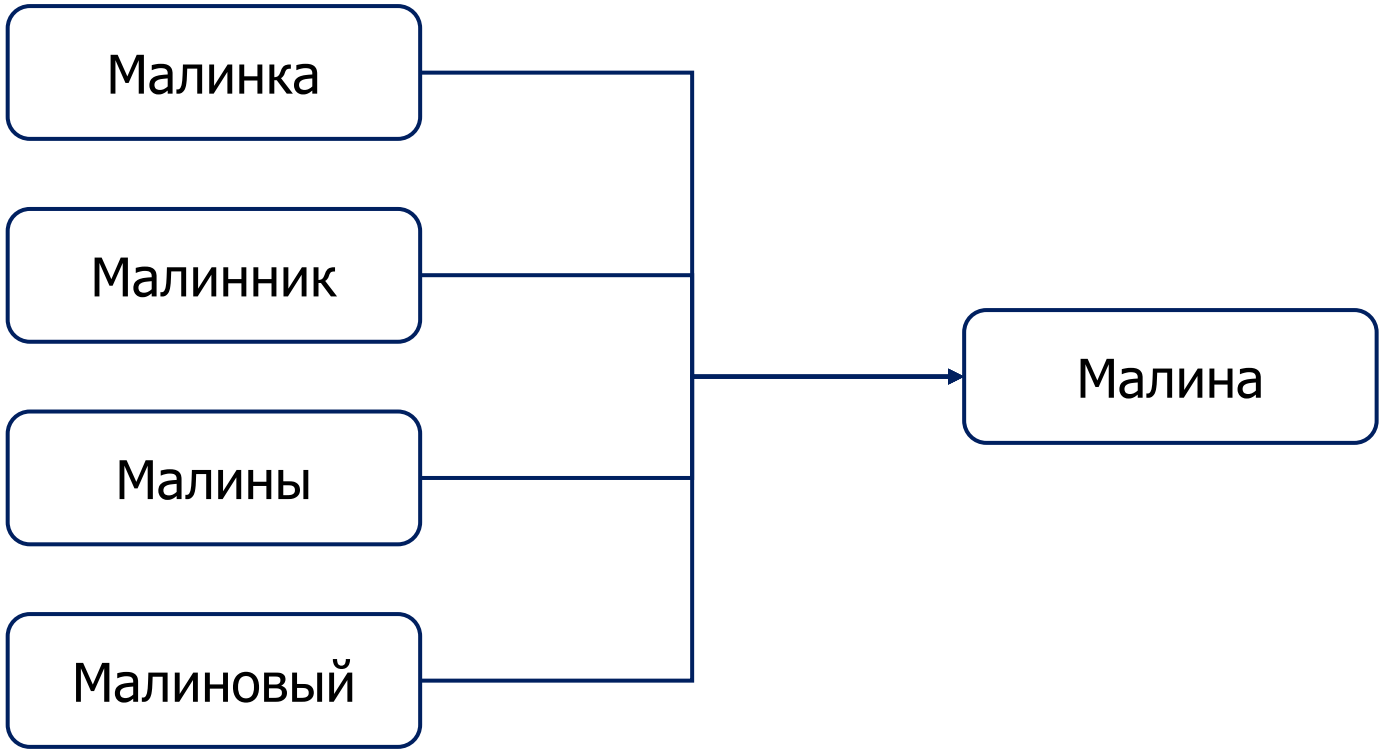
Некоторые вопросы (но не все)

1. Многозначность слов: "back" (Noun, Verb, Adjective), "date" (Noun, Verb), "fair" (Noun, Adjective), "present" (Noun, Verb, Adjective), …

2. Синонимы: "conversation" ↔ "talk", "amazing" ↔ "incredible", "begin" ↔ "start", "help" ↔ "aid"

3. Словоформы: "reason" → "reasonable", "happy" → "unhappy", "nature" → "natural" → "naturally" → "unnaturally"

Сокращение словоформ к одной базовой форме



Стемминг

- ✓ Алгоритмы стемминга находят окончания и суффиксы слов, сравнивая токены входных слов с заранее определенным списком общих окончаний и суффиксов.
- ✓ Затем стеммер удаляет из слова все найденные строки символов окончаний и суффиксов, если последнее не нарушает каких-либо правил или условий, связанных с ними.
- ✓ Некоторые стеммеры пропускают полученные фрагменты через дополнительный набор правил для исправления деформированных корней.

Стемминг: примеры

Анализируемые предложения:

1. “Life appears to me too short to be spent in nursing animosity or registering wrongs.” (“Jane Eyre”, Charlotte Brontë)

2. “I am the happiest creature in the world. Perhaps other people have said so before, but not one with such justice.” (“Pride and Prejudice”, Jane Austen)

Слова после стемминга:

1. 'life', 'appear', 'to', 'me', 'too', 'short', 'to', 'be', 'spent', 'in', 'nurs', 'animos', 'or', 'regist', 'wrong'

2. 'i', 'am', 'the', 'happ', 'creatur', 'in', 'the', 'world', 'perhap', 'other', 'peopl', 'have', 'said', 'so', 'befor', 'but', 'not', 'one', 'with', 'such', 'justic'

Лемматизация

- ✓ Лемматизация — это более существенная процедура анализа по сведению морфологических вариантов к одной словарной базовой форме.
- ✓ Практическое различие между стеммингом и лемматизацией заключается в том, что, когда стемминг просто удаляет общие окончания и суффиксы из конца токенов слова, лемматизация гарантирует, что выходное слово представляет собой существующую нормализованную форму слова (лемму), которую можно найти в словаре.
- ✓ Поскольку лемматизация направлена на выведение базовых форм словаря, она требует более надежного морфологического анализа, чем стемминг.

Лемматизация: Примеры

Анализируемые предложения:

1. “Life appears to me too short to be spent in nursing animosity or registering wrongs.” (“Jane Eyre”, Charlotte Brontë)

2. “I am the happiest creature in the world. Perhaps other people have said so before, but not one with such justice.” (“Pride and Prejudice”, Jane Austen)

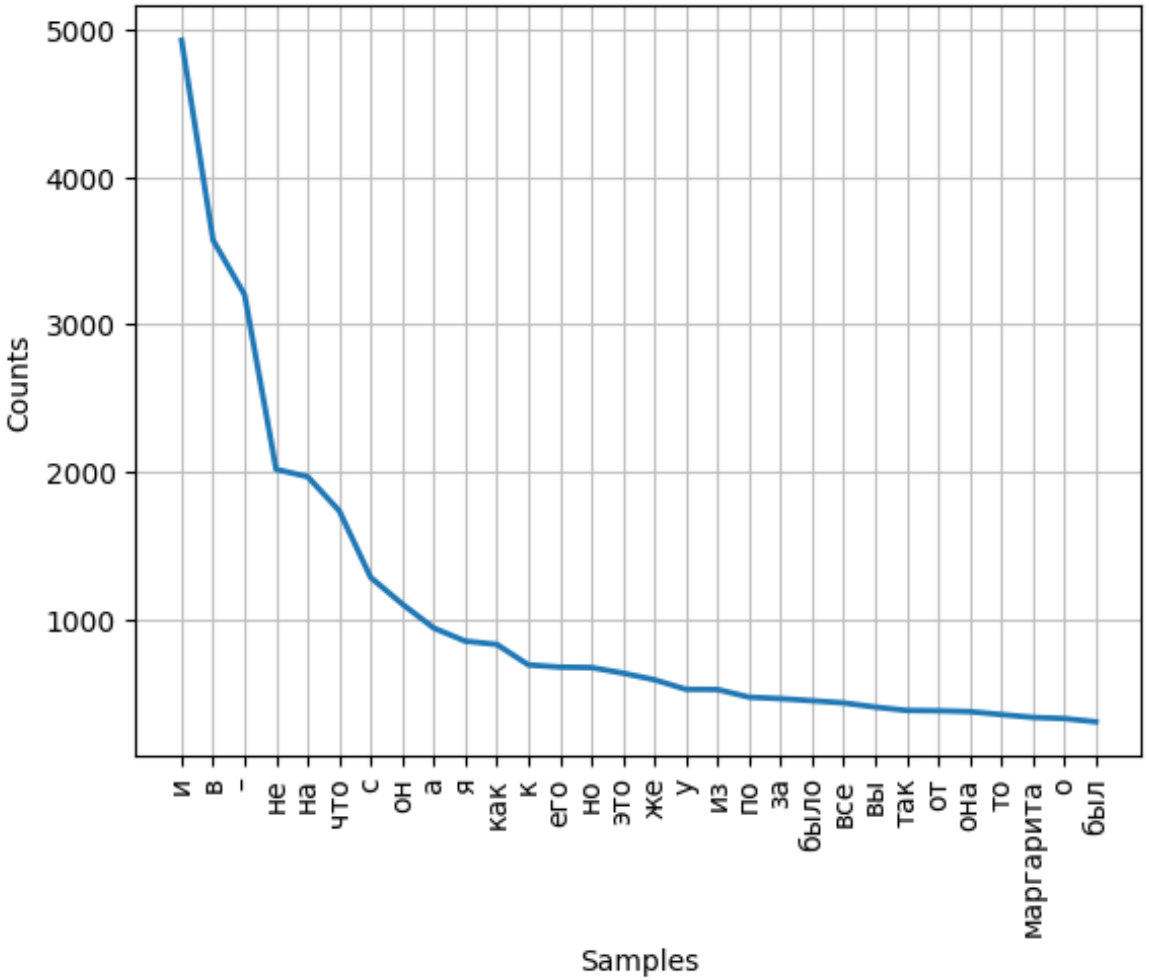
Лемматизированные слова:

1. 'life', 'appear', 'to', 'me', 'too', 'short', 'to', 'be', 'spend', 'in', 'nursing', 'animosity', 'or', 'registering', 'wrong'

2. 'i', 'am', 'the', 'happy', 'creature', 'in', 'the', 'world', 'perhaps', 'other', 'people', 'have', 'say', 'so', 'before', 'but', 'not', 'one', 'with', 'such', 'justice'

Частотный анализ слов

«Мастер и Маргарита», М.А.Булгаков

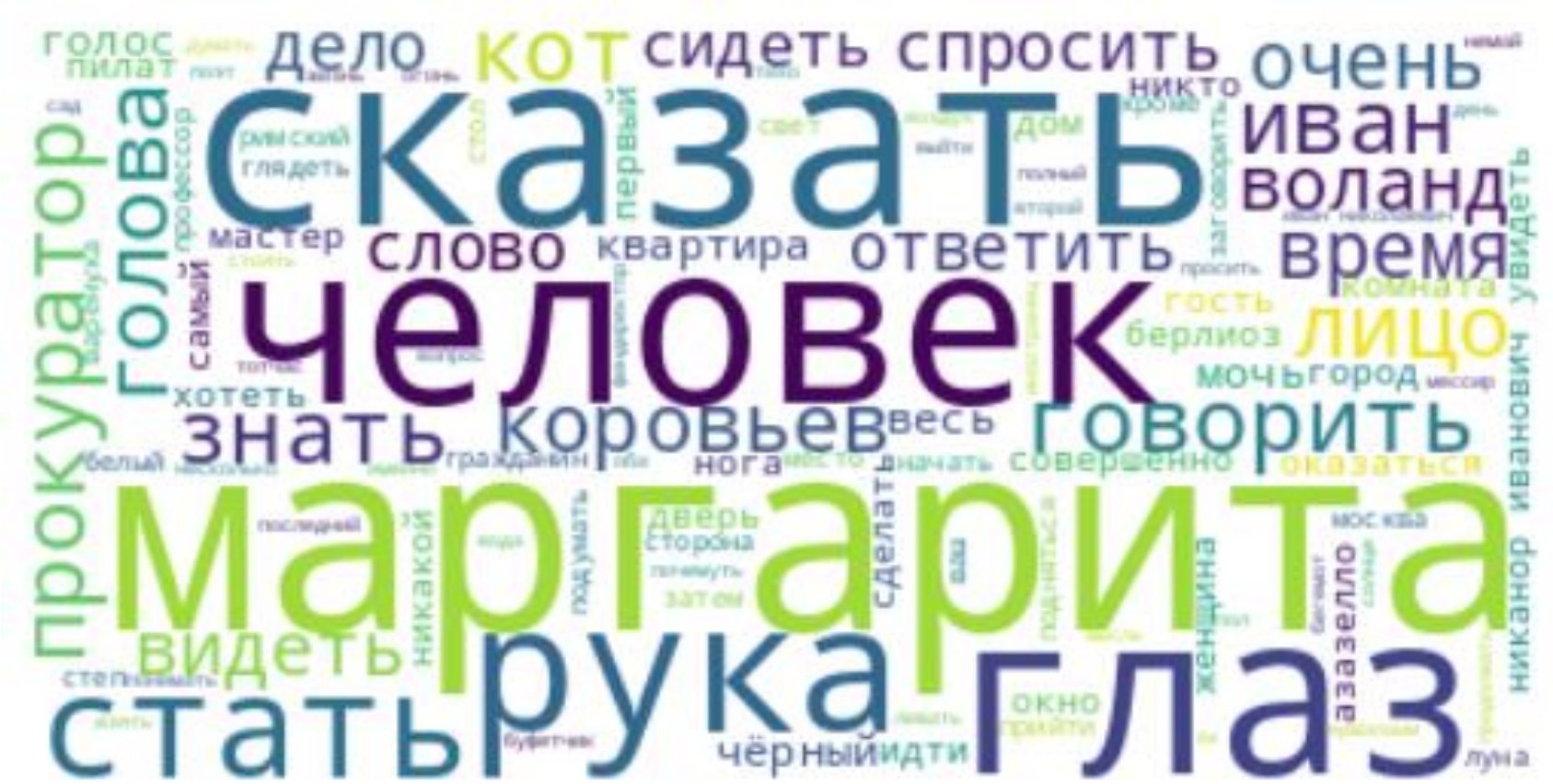


13 самых употребимых слов после удаления стоп-слов:

('маргарита', 519),
('сказать', 413),
('рука', 346),
('глаз', 325),
('человек', 322),
('иван', 283),
('ответить', 281),
('стать', 268),
('говорить', 249),
('знать', 242),
('прокуратор', 224),
('голова', 223),
('воланд', 217)

Облако слов: Пример

«Мастер и Маргарита», М.А. Булгаков



Унитарное кодирование

Вход: «Там некогда гулял и я: Но вреден север для меня» («Евгений Онегин», А.С. Пушкин)

Стоп-слова: там, и, я, но, для, меня.

Удаление стоп-слов + лемматизация: «некогда», «гулять», «вредный», «север»

Кодирование:

Токен				
некогда	1	0	0	0
гулять	0	1	0	0
вредный	0	0	1	0
север	0	0	0	1

Мешок слов (Bag-of-Words): Пример

Словарь: 'blood', 'cell', 'plasma', 'presence', 'computer', 'quantum', 'mechanical', 'compose', 'take'

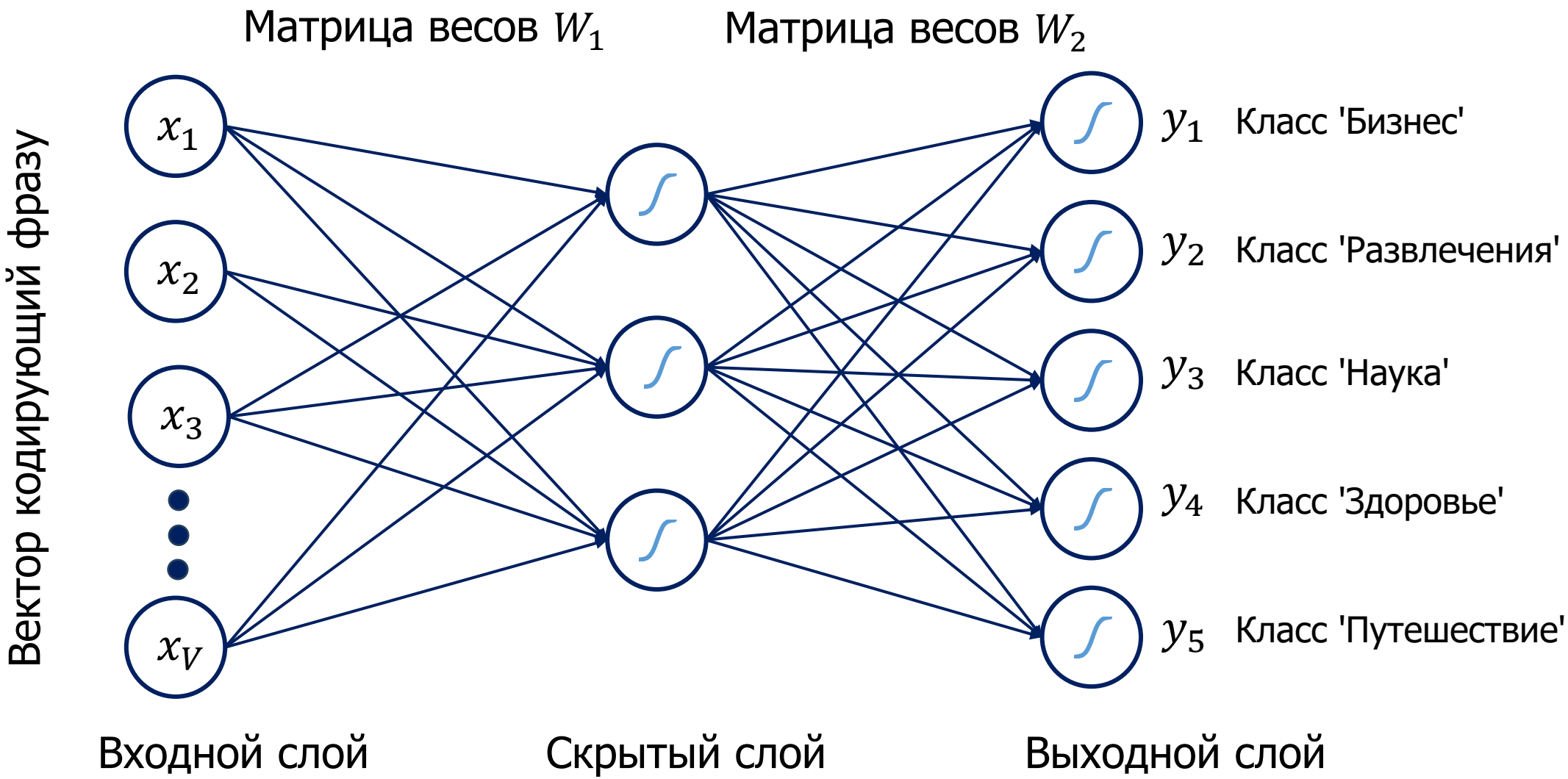
1. Blood is composed of blood cells suspended in blood plasma.
2. Red blood cells and white blood cells make about forty percent of the blood.
3. The presence of the red blood cells gives the blood a deep-red shade.
4. A quantum computer is a computer that takes advantage of quantum mechanical phenomena.
5. A quantum computer leverages quantum superposition and entanglement.

blood	cell	plasma	presence	computer	quantum	mechanical	compose	take
3	1	1	0	0	0	0	0	0
3	2	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
0	0	0	0	2	2	1	0	1
0	0	0	0	1	2	0	0	0

Примеры выборок

Вход	Метка
«Extremely rude staff. They don’t resolve any customer issues effectively and I had a bad experience with this company! I don’t highly recommend them to anyone looking for quality products and excellent service.»	Отрицательный отзыв
Представление «мешок слов»: (0, 1, 0, 0, 1, 0, … , 0)	0
«Five stars all the way! The service, the product, and the overall experience were outstanding.»	Положительный отзыв
Представление «мешок слов»: (0, 0, 1, 0, 0, 1, … , 0)	1

Простой классификатор



Пример кодирования

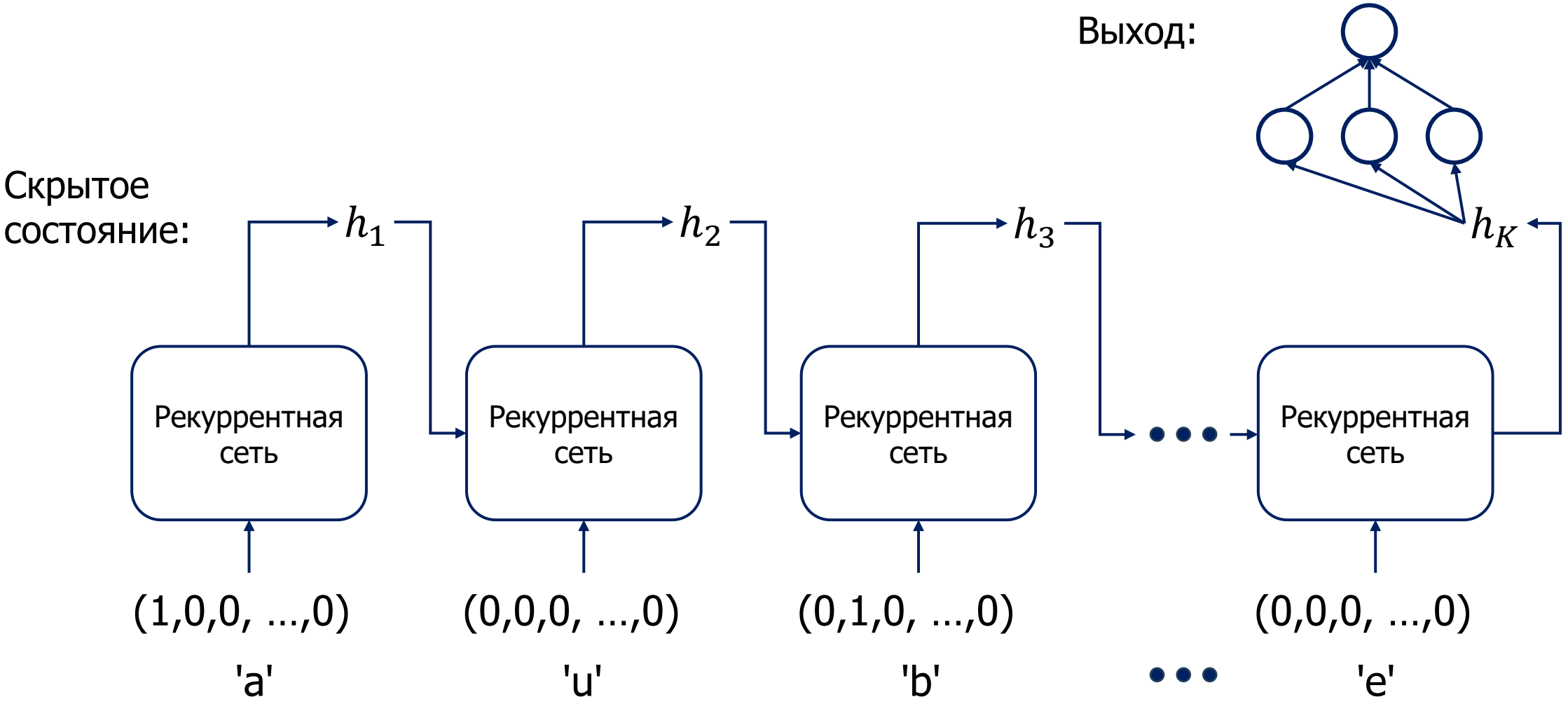
Имена: 'Aubree’ → Женское, 'Jane’ → Женское, 'Ann’ → Женское, 'Jack’ → Мужское

'Aubree’ (Токенизация) → 'a', 'u', 'b', 'r', 'e', 'e'

'a' → 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'u' → 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
'b' → 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'r' → 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
'e' → 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
'e' → 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Метки классов: Женское имя → 1, Мужское имя → 0

Активация нейросети

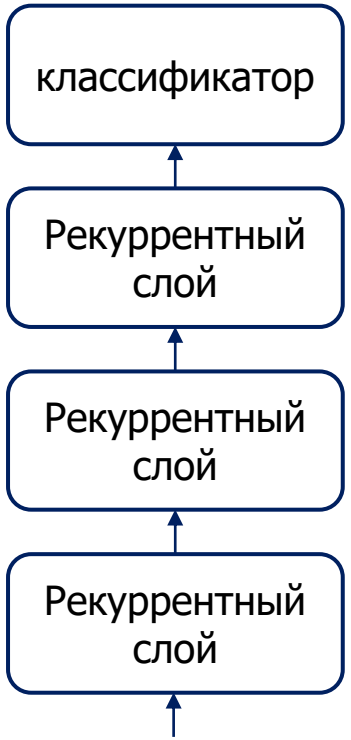


Более сложные зависимости в данных

Объединённое скрытое состояние=
скрытое состояние 1 & скрытое состояние 2

скрытое состояние 2←'a'←'u'←'b'←'r'←'e'←'e'

'a'→'u'→'b'→'r'→'e'→'e'→скрытое состояние 1



Двунаправленные представления

Стек рекуррентных слоёв

Частота термина-обратная частота документа (TF-IDF)

Частота термина (частота использования термина t в документе d):

$$TF(t,d) = \frac{n_t}{\sum_k n_k}$$

Обратная частота документа (мера, оценивающая, какую долю информации предоставляет слово):

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

Мера TF-IDF:

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

D - Корпус документов

n_t - сколько раз термин t встречается в документе

Пример TF-IDF

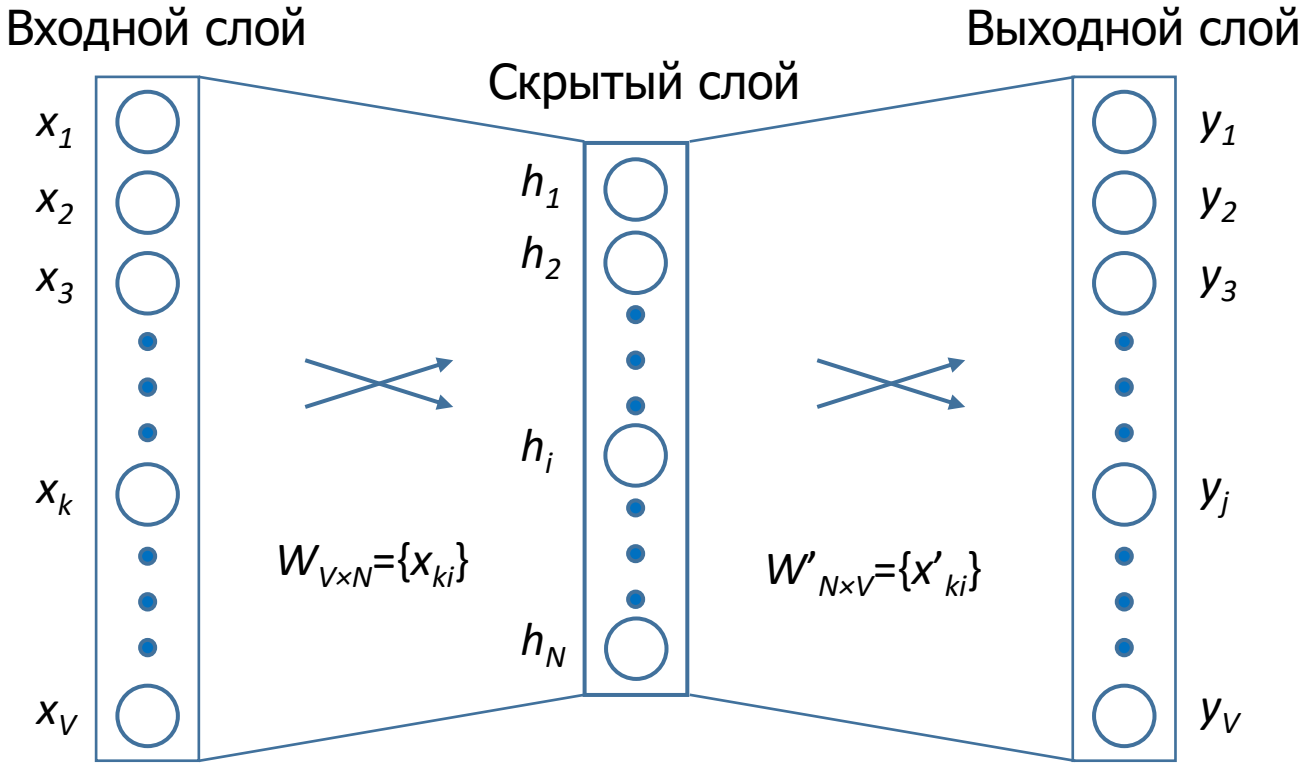
1. *Эскимосы живут на Севере*.
2. *Эскимосы хотят кушать свежую рыбу*.
3. *Рыба хочет кушать корм*.
4. *Свежая рыба очень вкусная*.

Слово «рыба» встречается в трёх документах.
TF(слово: *«рыба»*, документ: «*Рыба хочет кушать корм*.») = 1 /4 = 0.25.
IDF(слово: *«рыба»*, документы 1-4) = log10 (4/3) ≈ 0.125.
TF-IDF(слово: *«рыба»*, документ: «*Рыба хочет кушать корм*.», документы 1-4) ≈ 0.25 x 0.125 ≈ 0.031.

#	эскимос	кушать	вкусный	рыба	свежий	корм	на	жить	север	хотеть	очень
1	0.075	0	0	0	0	0	0.15	0.15	0.15	0	0
2	0.06	0.06	0	0.025	0.06	0	0	0	0	0.06	0
3	0	0.075	0	0.031	0	0.15	0	0	0	0.075	0
4	0	0	0.15	0.031	0.075	0	0	0	0	0	0.15

Предсказание следующего слова

- Задача предсказания слова по предыдущему слову
- Вход – one-hot вектор предыдущего слова
- Функция активации выходного слоя – softmax
- *W* или *W*– матрицы эмбеддингов



V - Размер словаря

N - Размер векторного вложения

Контекст (окружающие токены)

«В человеке должно быть всё прекрасно: и лицо, и одежда, и душа, и мысли.» («Дядя Ваня», А.П. Чехов)

должно быть всё **прекрасно** и лицо и $m = 3$

человеке должно быть всё **прекрасно** и лицо и одежда $m = 4$

В человеке должно быть всё **прекрасно** и лицо и одежда и $m = 5$

m - Количество слов слева и справа от целевого слова.

Примеры. Одинаковые контексты

В комнате котёнок наблюдал за рыбкой.
Из миски мой котёнок пил теплое молочко.
Там очень мило котёнок спал в корзинке.
Вика и Юля играли с котёнком в детской спальне.

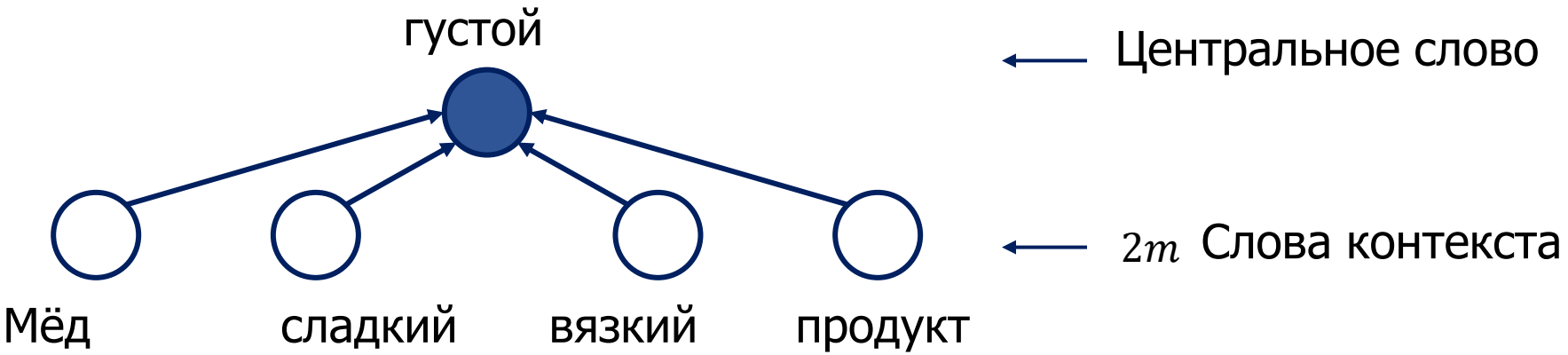
В комнате кот наблюдал за рыбкой.
Из миски мой кот пил теплое молочко.
Там очень мило кот спал в корзинке.
Вика и Юля играли с котом в детской спальне.

Марта увидела прекрасную сумку на витрине магазина.
Вера была самой прекрасной невестой на свете.

Марта увидела гламурную сумку на витрине магазина.
Вера была самой очаровательной невестой на свете.

Word2Vec: Непрерывный мешок слов (CBOW)

P(”густой” |”мёд”, ”сладкий”, ”вязкий”, ”продукт”)



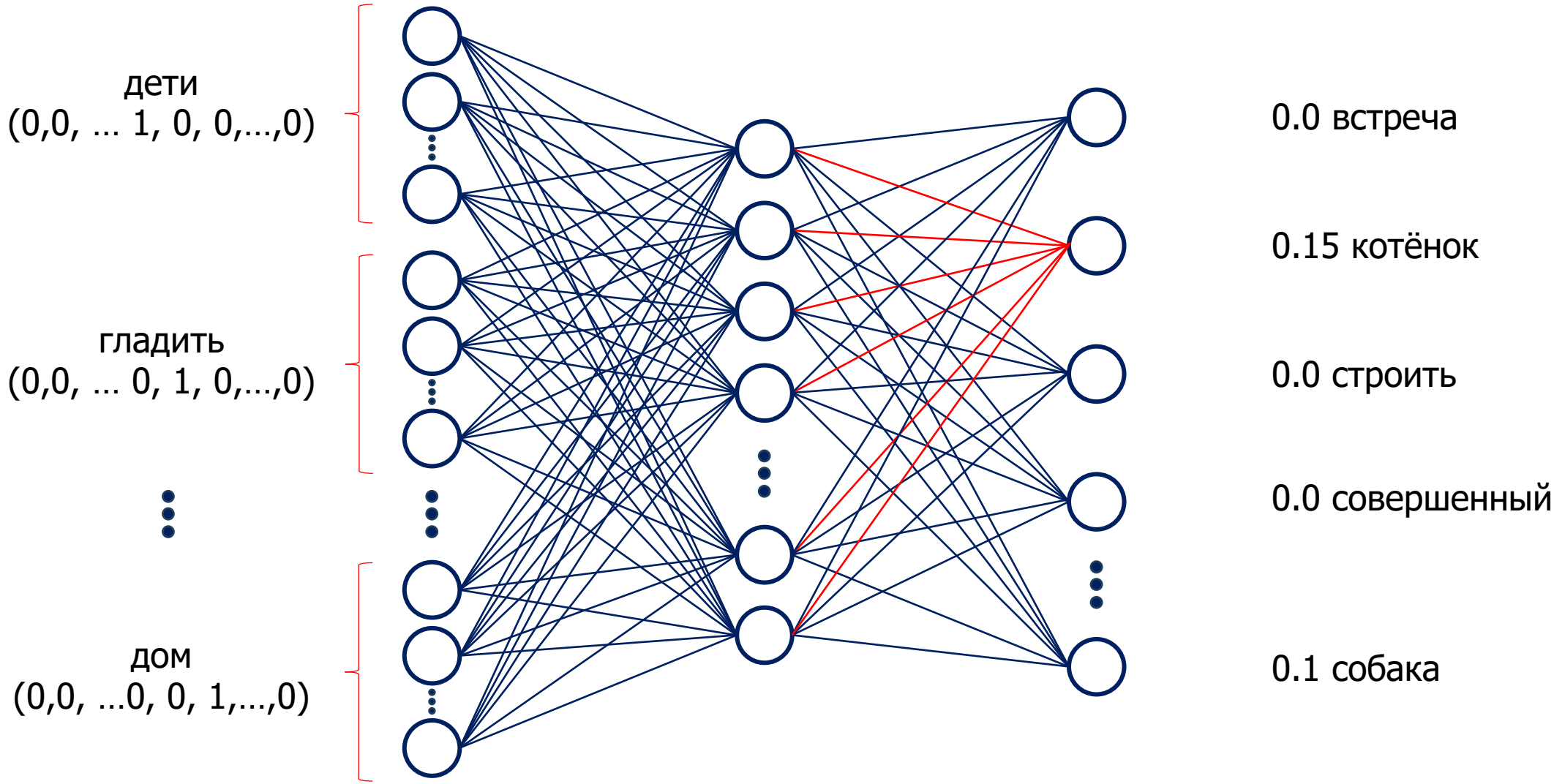
$$P(w_c|w_{o_1},\dots,w_{o_{2m}}) = \frac{\exp\left(\frac{1}{2m}u_c^\top(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}})\right)}{\sum_{i\in v}\exp\left(\frac{1}{2m}u_i^\top(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}})\right)} \qquad \mathcal{W}_o = \{w_{o_1},\dots,w_{o_{2m}}\}$$

$$\bar{\mathrm{v}}_o = \frac{(\mathrm{v}_{o_1}+\cdots+\mathrm{v}_{o_{2m}})}{2m} \qquad P(w_c|\mathcal{W}_o) = \frac{\exp(u_c^\top\bar{\mathrm{v}}_o)}{\sum_{i\in v}\exp(u_i^\top\bar{\mathrm{v}}_o)}$$

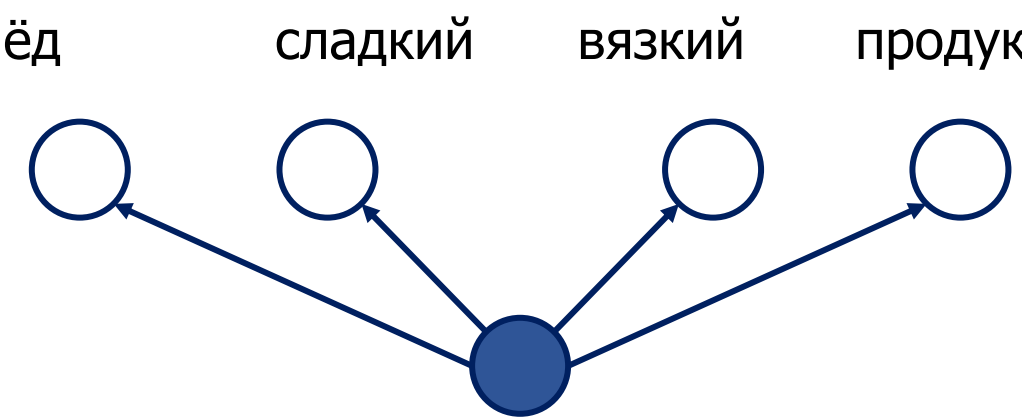
Функция потерь: $$J(\theta) = \prod_{t=1}^{T} P(w^{(t)}|w^{(t-m)},\dots,w^{(t-1)},w^{(t+1)},\dots,w^{(t+m)})$$

Весовые коэффициенты

Дети гладили маленького пушистого котёнка на лужайке перед домом.



Word2Vec: Skip-Gram



$2m$ Слова контекста

Центральное слово

P(”мёд”, ”сладкий”, ”вязкий”, ”продукт” |”густой”)

P(”мёд”|”густой”) · P(”сладкий”|”густой”) · P(”вязкий”|”густой”) · P(”продукт”|”густой”)

$$P(w_o|w_c) = \frac{\exp(u_o^\top \mathrm{v}_c)}{\sum_{i\in\nu}\exp(u_i^\top \mathrm{v}_c)}$$

v_i - представление центрального слова

u_i - представление контекстного слова $\nu = \{0, 1, \dots, |\nu| - 1\}$

Функция потерь: $J(\theta) = \prod_{t=1}^{T} \prod_{-m\le j\le m, j\neq 0} P(w^{(t+j)}|w^t)$ T - Длина текстовой последовательности

Обучение Skip-Gram

$$-\log\big(J(\theta)\big) = -\sum_{t=1}^{T} \prod_{-m\le j\le m, j\neq 0} \log\Big(P\big(w^{(t+j)}\big|w^{t}\big)\Big)$$

$$\log\big(P(w_o|w_c)\big) = u_o^{\top}\mathrm{v}_c - \log\Big(\sum\nolimits_{i\in v}\exp(u_i^{\top}\mathrm{v}_c)\Big)$$

$$\frac{\partial\log(P(w_o|w_c))}{\partial \mathrm{v}_c} = u_o - \frac{\sum_{i\in v}\exp\big(u_j^{\top}\mathrm{v}_c\big)u_j}{\sum_{i\in v}\exp\big(u_i^{\top}\mathrm{v}_c\big)} = u_o - \sum_{j\in v} P(w_j|w_c)u_j$$

FastText

fastText разбивает текст на n-граммные компоненты.

Примеры:

1. The bill has been sent back to comittee.
2. The decision was made by all the Northwest Wildlife Commitee members.
3. Jane chaired the Secondary Education Committe.

'comittee' → 'co', 'mi', 'tt', 'ee'
'committee' → 'co', 'mm', 'it', 'te', 'e'
'committe' → 'co', 'mm', 'it', 'te'
'committee' → 'committee', 'co', 'mm', 'it', 'te', 'e'

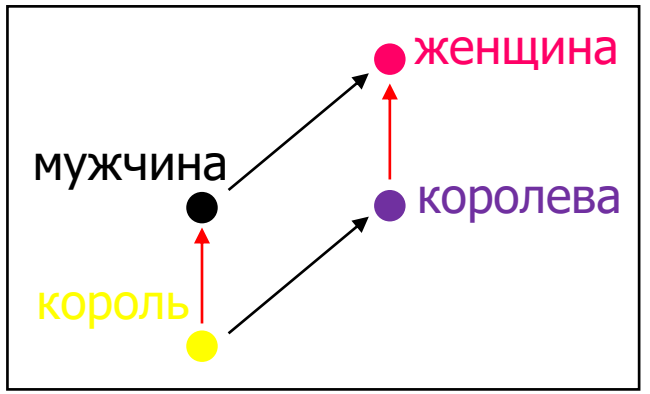
fastText запоминает веса каждой n-граммной компоненты вместе со целым токеном слова.
Каждый токен/слово будет выражаться как сумма и среднее значение его n-граммных компонентов.

Пример: «Королева-Женщина=Король-Мужчина»

	одушевлен.	кошачьи	человек	пол	монархия	глагол	множ. числ.
кошка	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
котёнок	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
собака	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
дома	-0.8	-0.4	-0.5	0.1	-0.9	-0.3	0.8
мужчина	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
женщина	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
король	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
королева	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

7D→2D

7D→2D



Слово

Векторное представление

Сокращение размерности

Двумерные представления

Языковые модели (Language Model)

- Языковая модель позволяет оценить вероятность следующего слова в последовательности и оценить вероятность всей последовательности слов.
- Пример: Какое слово в последовательности вероятнее:

  Кошки ценятся человеком за умение забавлять …

  1) детей, 2) грызунов, 3) природу
- Какая последовательность вероятнее:

  Кошки ценятся человеком за умение забавлять детей.

  Забавлять человеком за кошки умение детей ценятся.