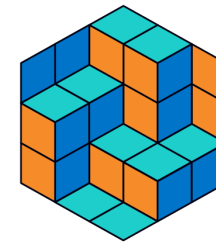


Курс: генерация рассказов

Часть 1: Токенизация и word2vec

Давайте знакомиться!

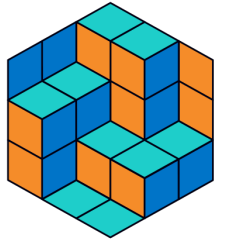


Миша
студент магистратуры МКН СПбГУ



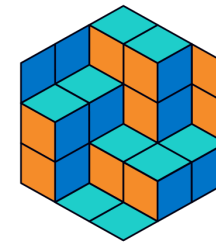
Влад
студент магистратуры МКН СПбГУ

Что такое NLP?

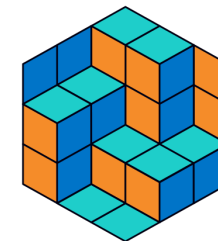


Что такое NLP?

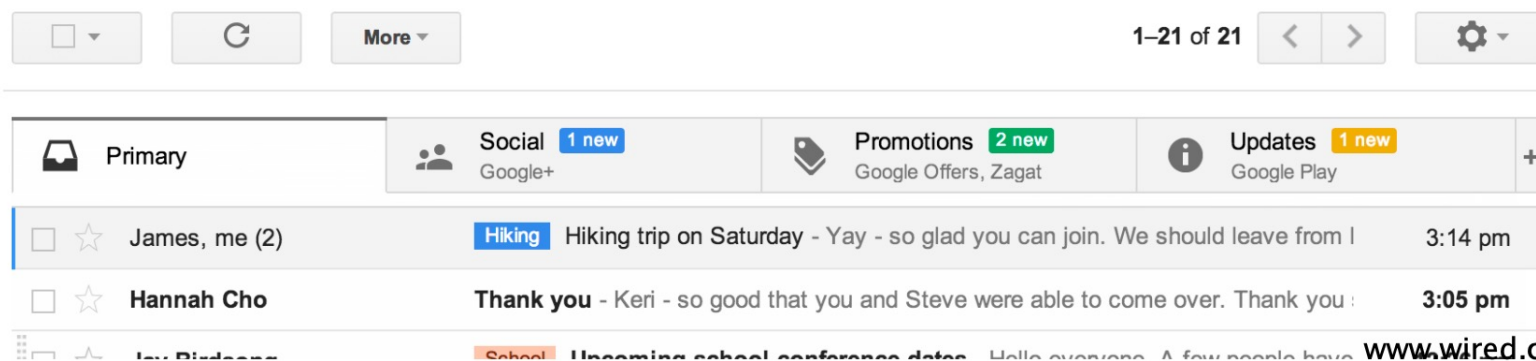
- NLP - **Natural Language Processing**



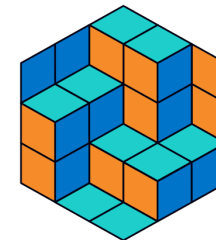
Какие задачи решает NLP?



- Классификация



Какие задачи решает NLP?



- Предсказание релевантной рекламы (и не только рекламы)

Meatloaf That Conquers the Mundane

FEB. 13, 2015

City Kitchen
By DAVID TANIS

Email
Share
Tweet
Pin
Save
More

I was raised on Midwestern meatloaf. My mother's dependable recipe did not vary: Ground beef, grated onion and carrot and a little oatmeal were the main ingredients, along with a dash of "seasoned salt." A ribbon of bottled chili sauce ran down a gully in the center.


Served hot, accompanied by Tater Tots, it was dinner. Served cold for lunch, it was always a sandwich on white bread, with potato chips on the side. It was usually moist and tasty but never remarkable, and there was no way you could call it anything but meatloaf.

Do I harbor a kind of nostalgia for it? Yes. But would I use that recipe now? I think not.

I have a friend from Brussels who loves to entertain. Of his dinner party repertoire, one dish is most requested and admired. It is pain de veau, served with a vermouth-splashed mushroom sauce. In French, it sounds elegant. Translated into English — veal loaf — it sounds dull.

The Italian word for meatloaf is polpettone. (Polpette are Italian meatballs; polpettine are meatballs, too, but more diminutive.) This substantial family-size meatball, whether ovoid or elongated, plain or fancy, served with tomato sauce or not, is beloved both in Italy and in Italian communities throughout the world. Aside from its melodic, polysyllabic name, polpettone is always well seasoned, prepared with care and served with gusto.

It is usually a combination of different kinds of ground meat, typically beef, pork and veal



Evan Sung for The New York Times

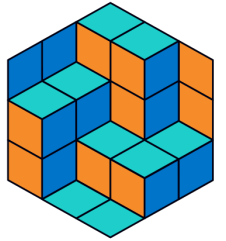
RELATED COVERAGE

City Kitchen: How to Make Polpettone, Step by Step FEB. 13, 2015

RECIPES FROM COOKING

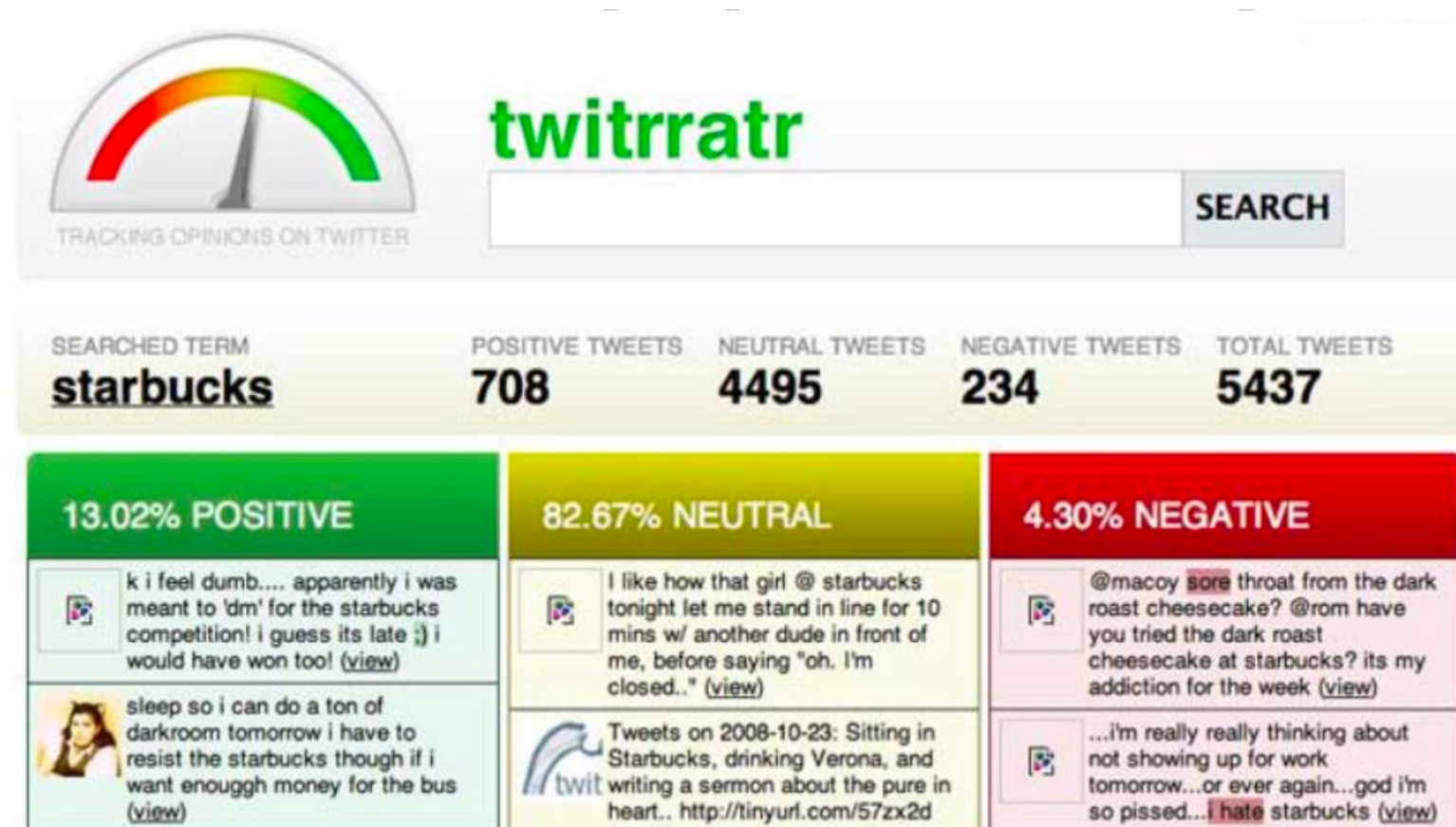
Polpettone with Spinach and Provolone

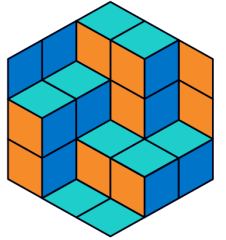
related
articles



Какие задачи решает NLP?

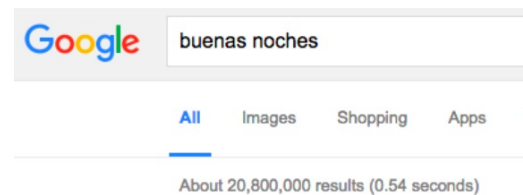
- Определение токсичных и хороших текстов





Какие задачи решает NLP?

- Машинный перевод

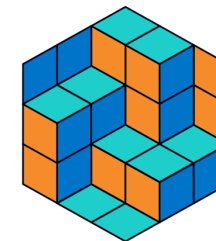


Open in Google Translate



Facebook translation, image credit: Meedan.org

Какие задачи решает NLP?



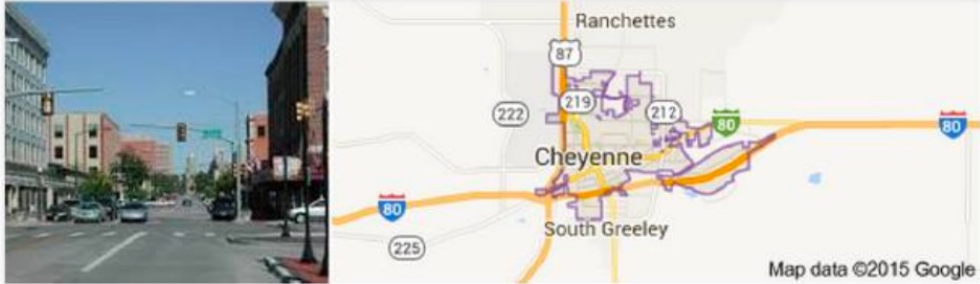
- Ответы на вопросы

What's the capital of Wyoming?

Web Maps Shopping Images News More Search tools

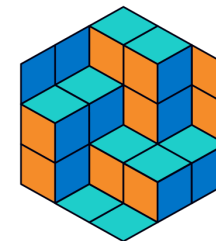
About 984,000 results (0.54 seconds)

Wyoming / Capital



Cheyenne

Какие задачи решает NLP?



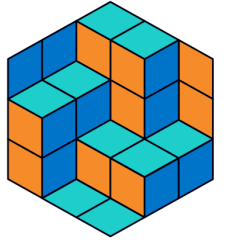
- Определение ролей в предложении

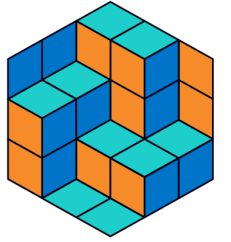
NER DEFINITION

Luke Rawlence PERSON joined Aiimi ORG as a data scientist in Milton Keynes PLACE, after finishing his computer science degree at the University of Lincoln. ORG

Что мы будем изучать?

- word2vec
- Bag of words
- seq2seq
- Encoder/Decoder
- RNN
- BERT

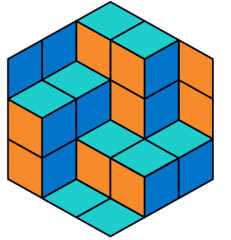




Глава 1

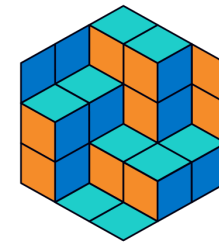
Как компьютер видит текст?

Определений

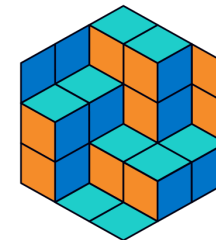


Токен – так мы будем называть любое слово или цельный знак пунктуации
‘.’ и ‘...’ - разные токены!

Как можно делить текст?



Как можно делить текст?

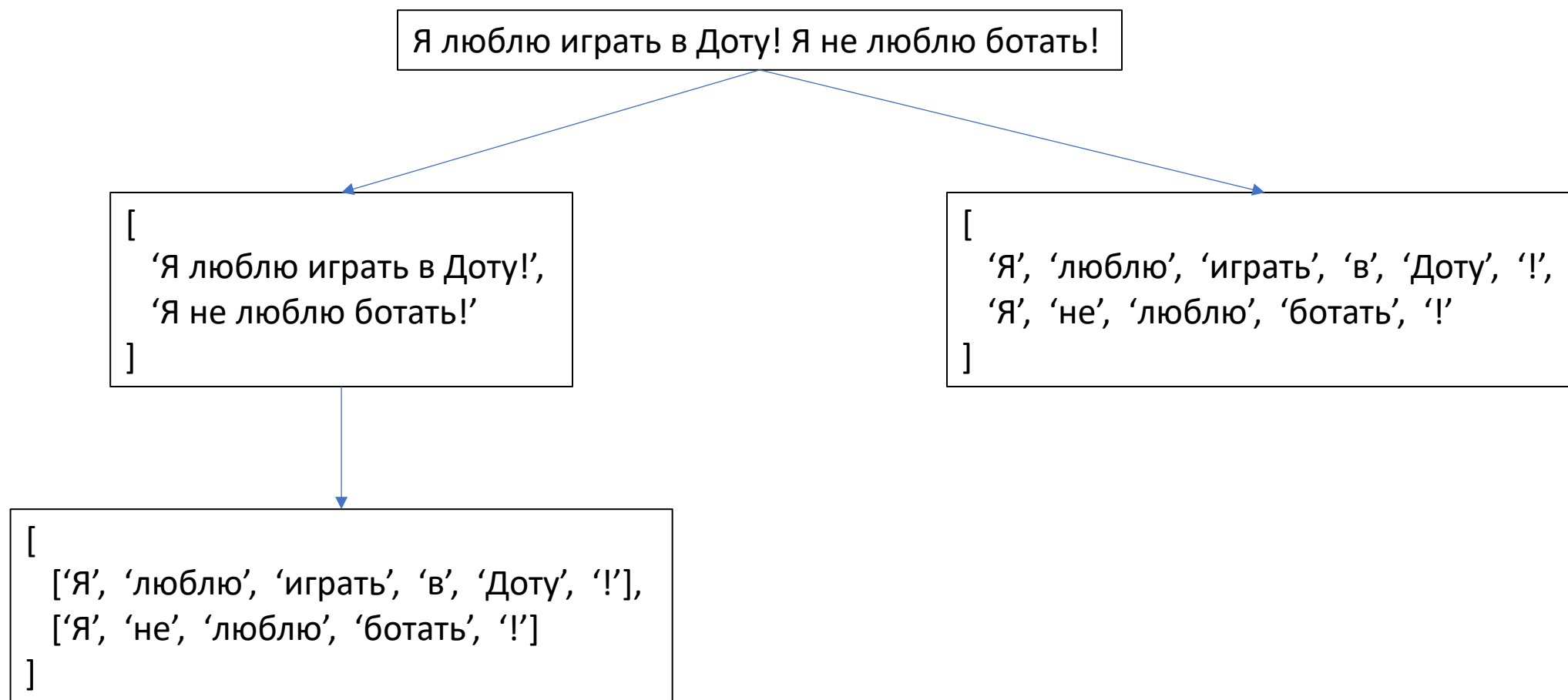
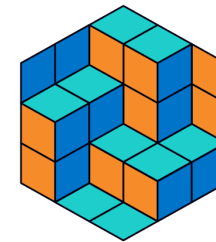


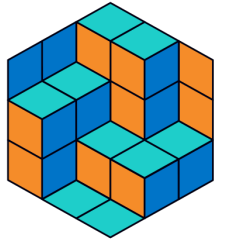
Я люблю играть в Доту! Я не люблю ботать!

[
 'Я люблю играть в Доту!',
 'Я не люблю ботать!'
]

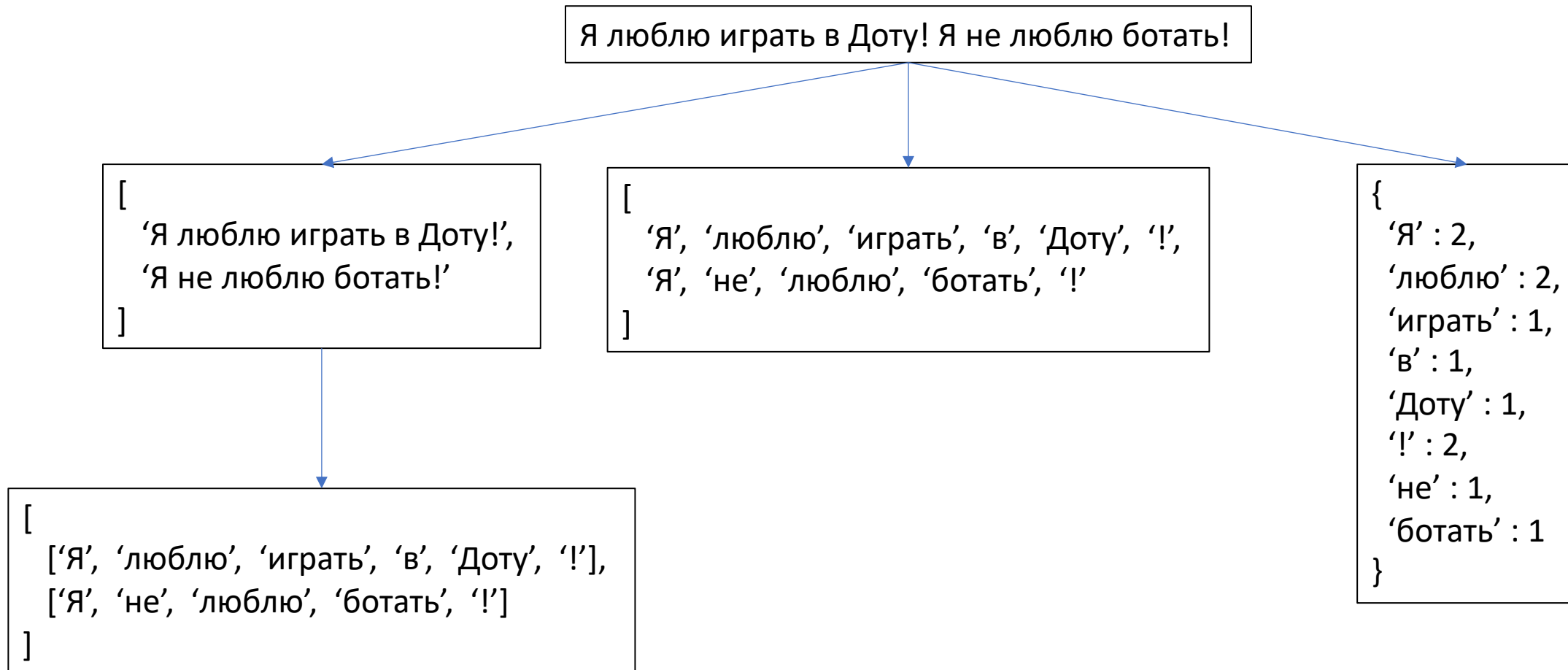
[
 ['Я', 'люблю', 'играть', 'в', 'Доту', '!'],
 ['Я', 'не', 'люблю', 'ботать', '!']
]

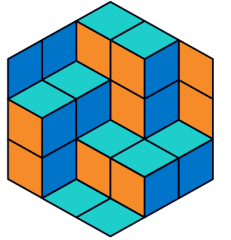
Как можно делить текст?





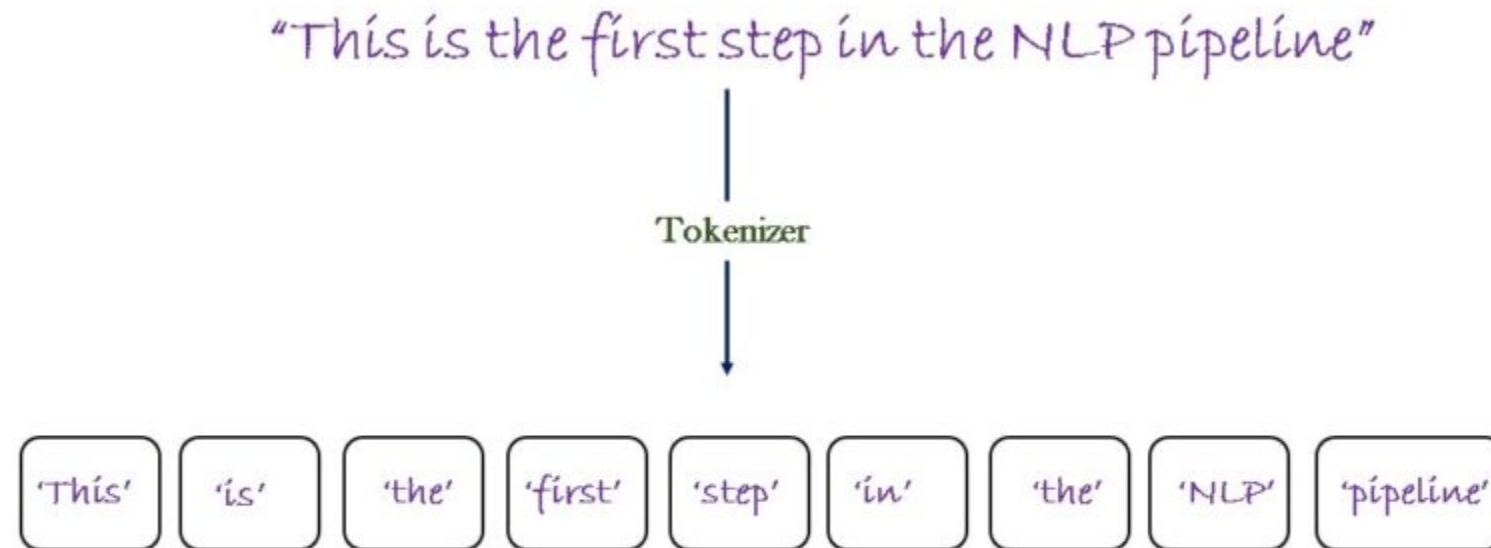
Как можно делить текст?

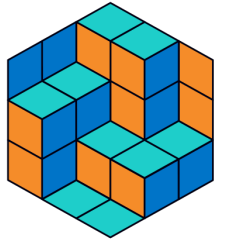




Как можно делить текст?

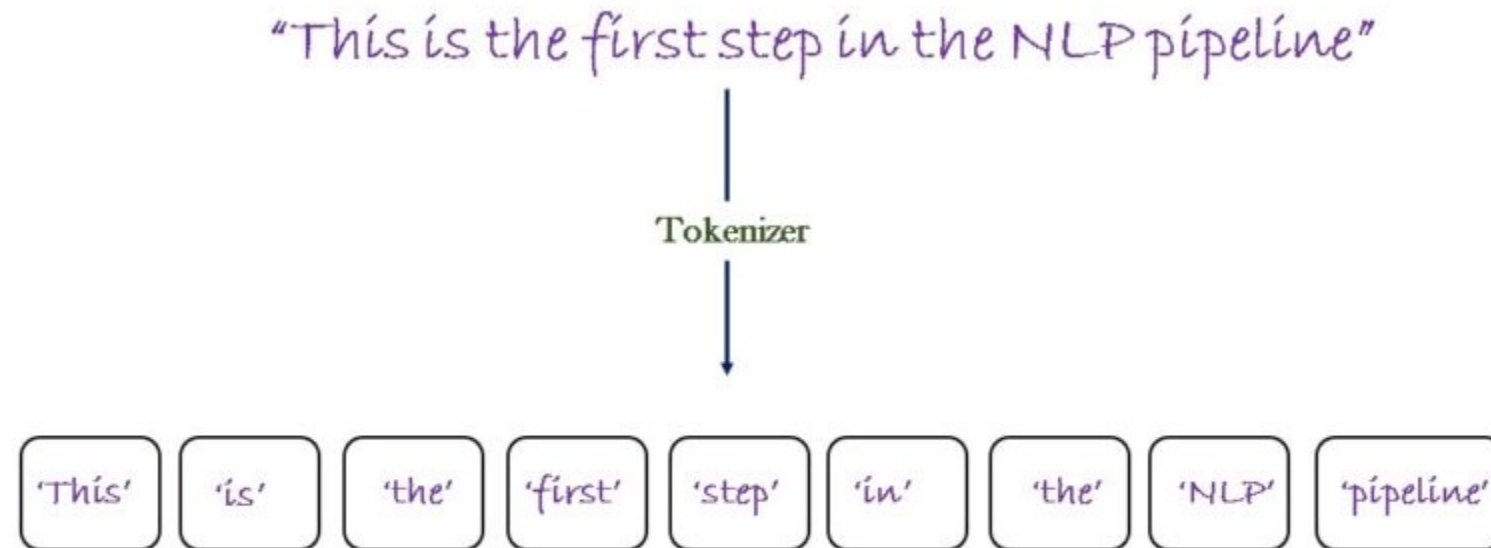
- Процесс деления предложения на токены - Токенизация



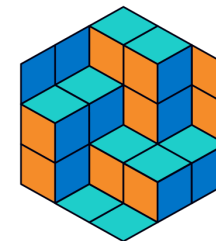


Как можно делить текст?

- Процесс деления предложения на токены – Токенизация
- Фактически – это несколько усложненный *split*

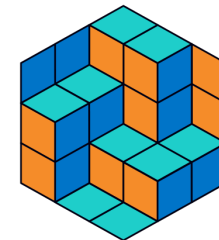


Как можно хранить текст?



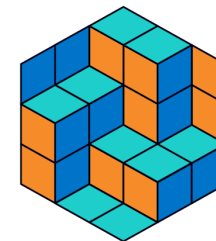
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.

Как можно хранить текст?



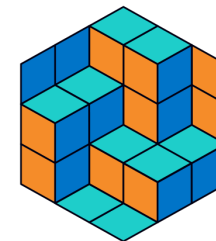
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов?

Как можно хранить текст?



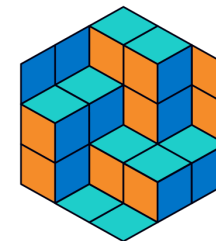
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов? **Плохо**
Это очень долго, затратно, и совершенно не понятно, как с этим работать

Как можно хранить текст?



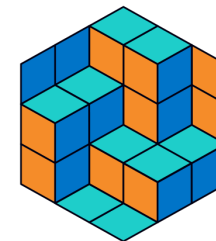
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов? **Плохо**
- Вектором предложений?

Как можно хранить текст?

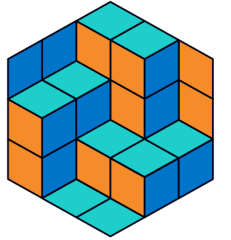


- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов? **Плохо**
- Вектором предложений? **Еще хуже**
Причины все те же

Как можно хранить текст?



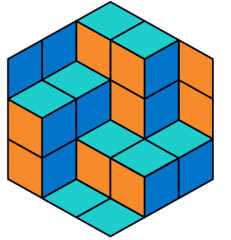
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов? **Плохо**
- Вектором предложений? **Еще хуже**
- Вектор чисел?



Как можно хранить текст?

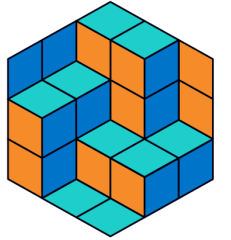
- Следующий вопрос: мы поделили предложение на токены, и теперь хотим это грамотно хранить и работать с этим.
- Вектором слов? **Плохо**
- Вектором предложений? **Еще хуже**
- Вектор чисел? **В точку**

Как можно хранить текст?

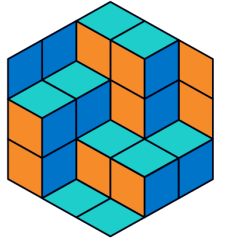


- А как человек понимает, что значит слово?
Благодаря чему мы в детстве учились говорить?

Как можно хранить текст?



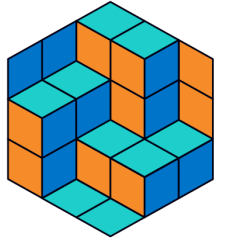
- А как человек понимает, что значит слово?
Благодаря чему мы в детстве учились говорить?
- **Контекст**
- Может ли «машина» понять контекст?



Глава 2

Word Embedding

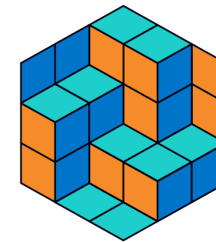
Определение



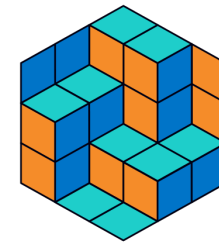
- *Word embedding* – вещественнозначный вектор, сопоставляемый слову.

Мотивация

- Задача: ищем похожих людей



Мотивация

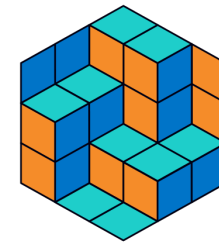


- Задача: ищем похожих людей



Проведем опрос

Мотивация



- Задача: ищем похожих людей

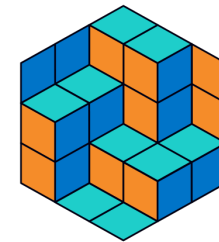


Проведем опрос



Жестко отвечаем

Мотивация



- Задача: ищем похожих людей



Проведем опрос

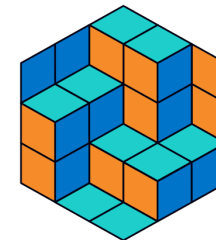


Жестко отвечаем

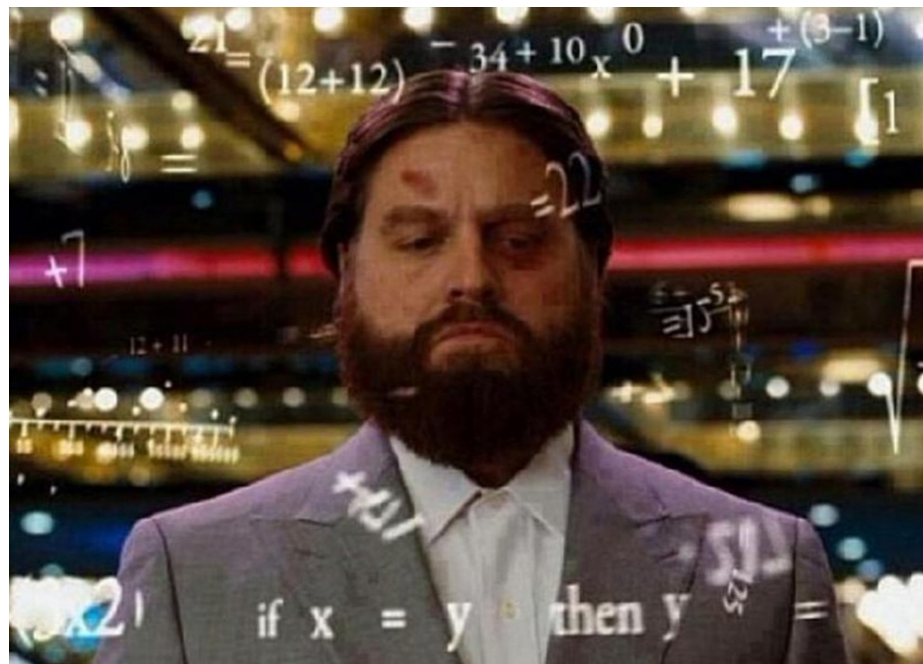


Устрашающе анализируем

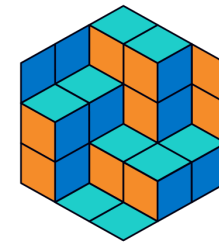
Мотивация



- Задача: ищем похожих людей
- Как найти похожих людей?



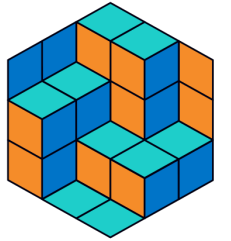
Мотивация



- Задача: ищем похожих людей
- Как найти похожих людей?
Верно! Расстояние между векторами.

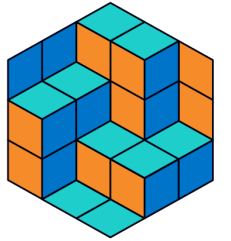


word2vec



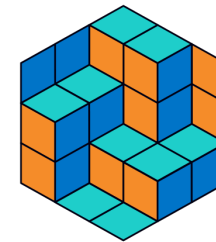
- Похожая логика лежит в основе word2vec
НО! Нужно определить, что будет «вопросами в опросе»

word2vec



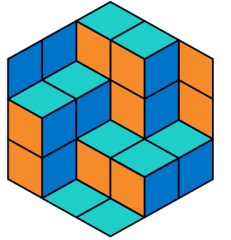
- Похожая логика лежит в основе word2vec
НО! Нужно определить, что будет «вопросами в опросе»
- Вопросы – контекст.
- Как правило, word2vec работает через **cbow** или **skip gram**

word2vec. Skip gram



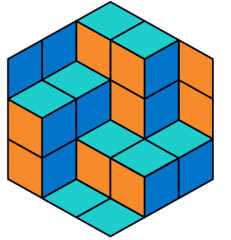
- Разберемся на примере пословицы
«Русский медленно запрягает да быстро едет»

word2vec. Skip gram



- Шаг 1: Произведем one hot кодировку:
 1. Составим список уникальных токенов - **словарь**
 2. Каждому токenu из словаря сопоставим вектор длины словаря так, что на i -м месте стоит 1 только если мы сейчас рассматриваем i -й токен из словаря. Остальные нули
- Для нашей пословицы:
 - Словарь: ['Русский', 'медленно', 'запрягает', 'да', 'быстро', 'едет']
 - Сопоставление:
 - 'медленно' -> [0,1,0,0,0,0]
 - 'едет' -> [0,0,0,0,0,1]

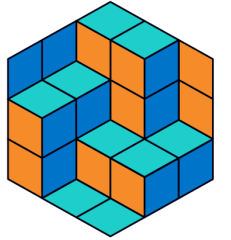
word2vec. Skip gram



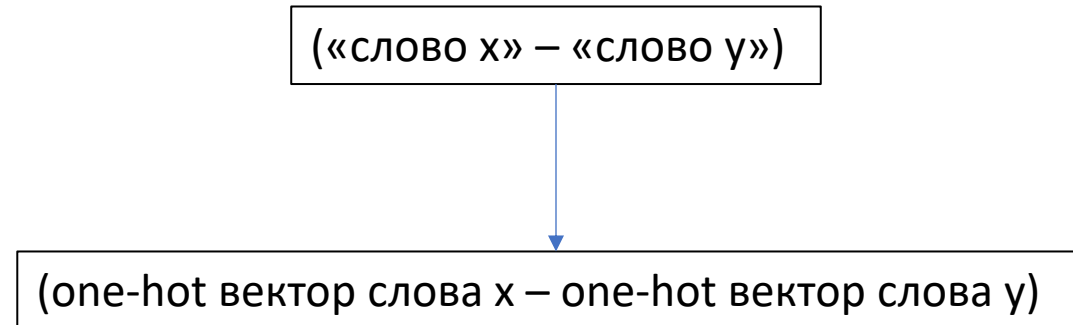
- Шаг 2: пройдем окном фиксированной длины по предложению и составим «контекстные пары» вида «рассматриваемое слово» – «слово в окне»
- Пример:

Исходный текст	Тренировочные сэмплы
Русский медленно запрягает да быстро едет	(русский, медленно) (русский, запрягает)
Русский медленно запрягает да быстро едет	(медленно, русский) (медленно, запрягает) (медленно, да)
Русский медленно запрягает да быстро едет	(запрягает, медленно) (запрягает, да) (запрягает, быстро)
Русский медленно запрягает да быстро едет	(да, запрягает) (да, быстро) (да, едет)

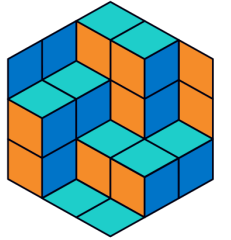
word2vec. Skip gram



- Шаг 3:



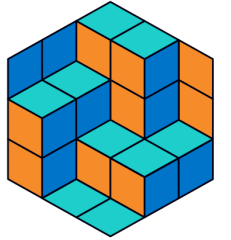
word2vec. Skip gram



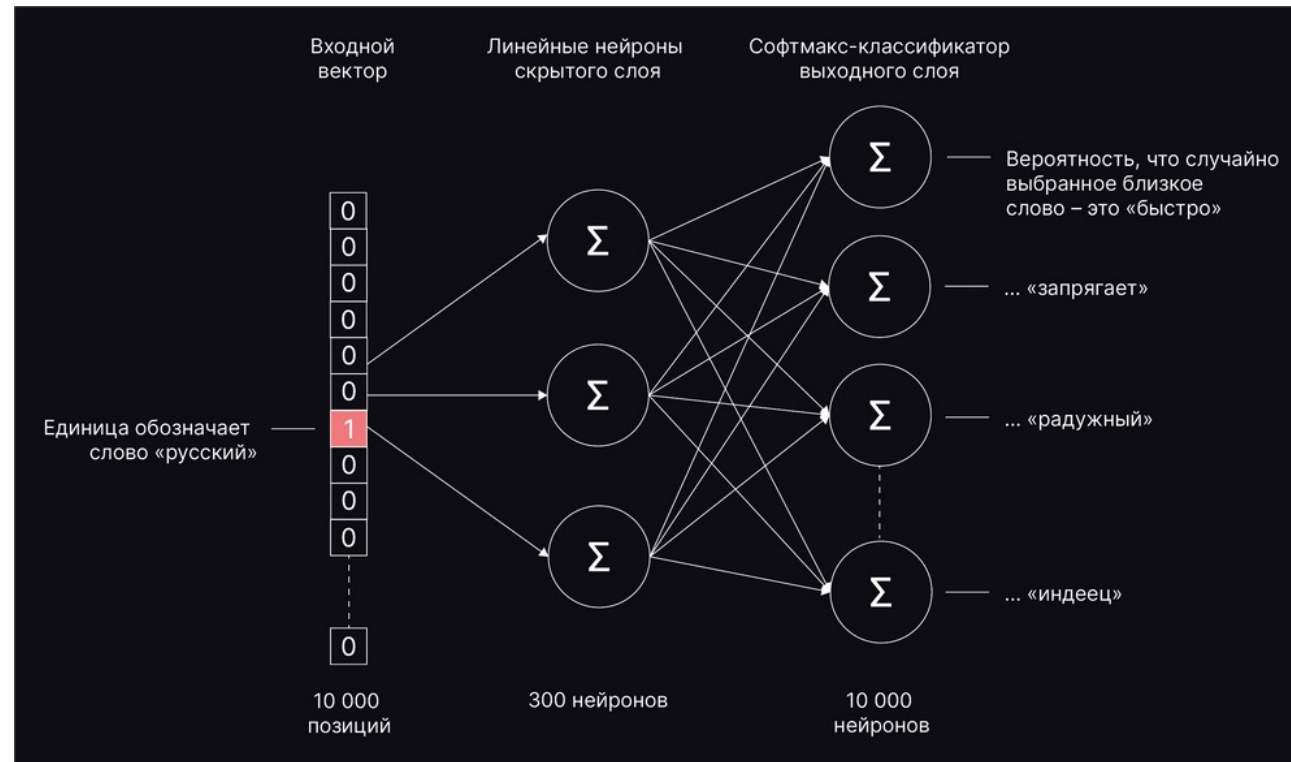
- Шаг 4: обучим модель (на самом деле не обязательно нейронную сеть – можно регрессию) на данных вида:

(one-hot вектор слова x – one-hot вектор слова y)

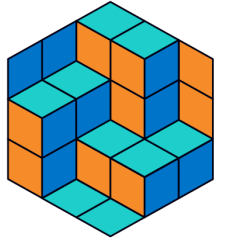
word2vec. Skip gram



- Шаг 4: Предположим, что наш словарь размера 10000. Тогда получится что-то вроде:



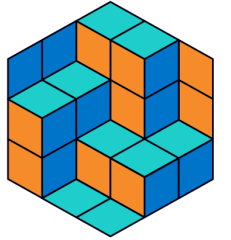
word2vec. Skip gram



- Вопрос: что делать с размерностью? Если словарь будет большого размера – то и вектор получится огромный...



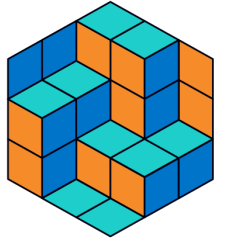
word2vec. Skip gram



- Как итог, слово «король» выглядит так:

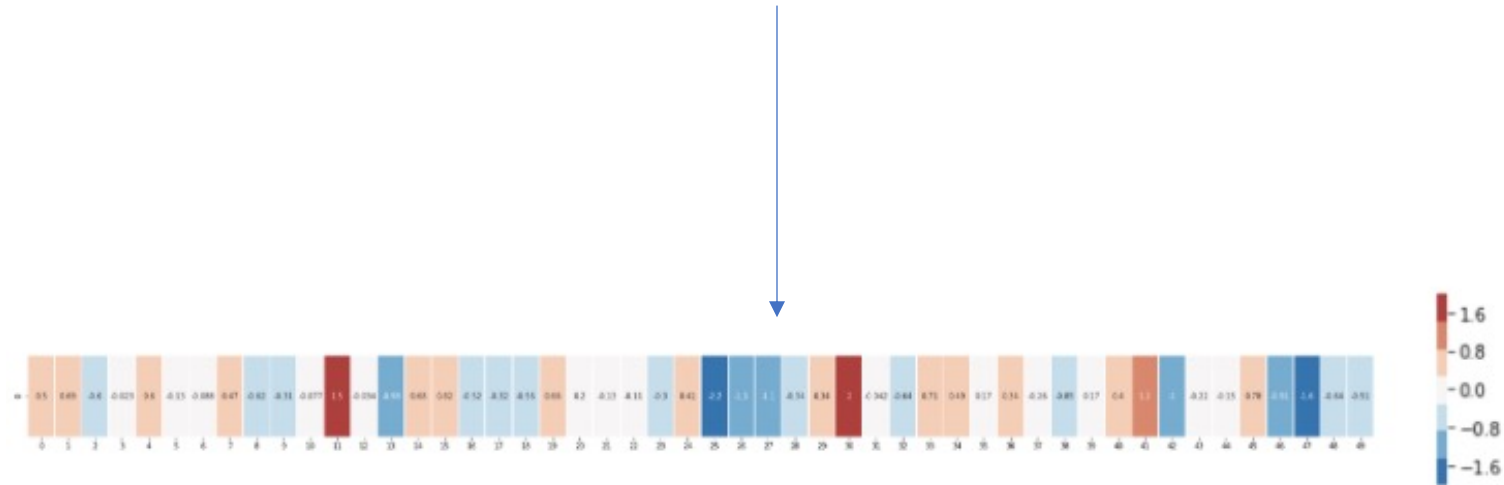
```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 ,  
0.47377 , -0.61798 , -0.31012 , -0.076666, 1.493 , -0.034189, -0.98173 ,  
0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961 ,  
-0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 ,  
-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 ]
```

word2vec. Skip gram

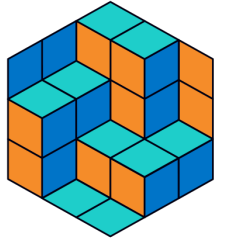


- Как итог, слово «король» выглядит так:

```
[ 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 ]
```



word2vec. Skip gram



“king”



“Man”



“Woman”



To be continued...

