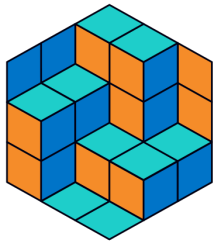
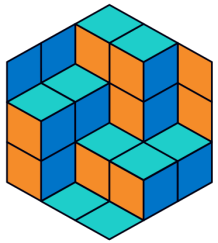
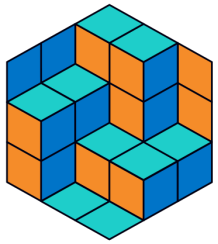


Курс: генерация рассказов
Часть 2: bag of words & N-gramm



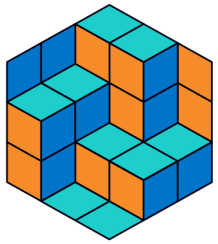
Глава 1
bag of words

Обратите внимание



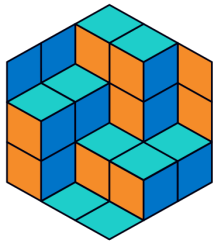
Bag of words – буквально «мешок слов», а значит работать мы будем **со словами, не с токенами**

Какие задачи решает?



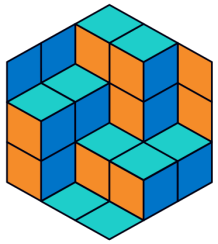
- Классификация
- Определение тональности

Определение



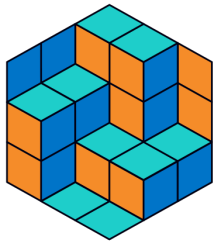
Корпус – так называется набор всех текстов из датасета.

Bag of words. Описание



Суть метод – создать неупорядоченный список слов, после чего для каждого текста из корпуса посчитать количество вхождения каждого слова в текст.

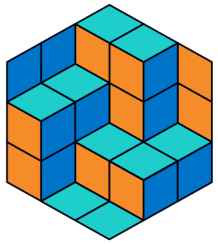
Bag of words. Описание



- Шаг 1: Сформировать корпус текстов

```
[   Текст,              Текст Поменьше,      большой Текст,
    Еще.Текст,          И еще текст,         Еще больше Текста ]
```

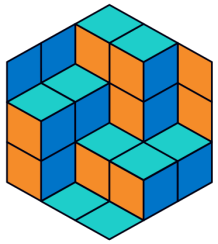
Bag of words. Описание



- Шаг 1: Сформировать корпус текстов
- Шаг 2: Все приводим в нижний регистр

```
[  текст,          текст поменьше,     большой текст,
   еще.текст,      и еще текст,        еще больше текста  ]
```

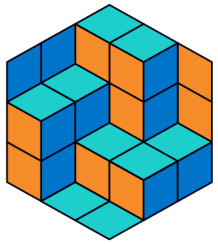
Bag of words. Описание



- Шаг 1: Сформировать корпус текстов
- Шаг 2: Все приводим в нижний регистр
- Шаг 3: Удаляются знаки пунктуация и «стоп-слова»

```
[  текст,          текст поменьше,     большой текст,
   еще текст,      еще текст,          еще больше текста  ]
```

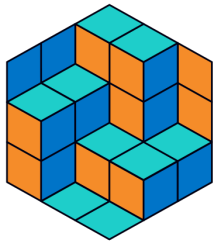
Bag of words. Описание



- Шаг 1: Сформировать корпус текстов
- Шаг 2: Все приводим в нижний регистр
- Шаг 3: Удаляются знаки пунктуация и «стоп-слова»
- Шаг 4: Производится стеминг – все слова к начальной форме

```
[  текст,          текст маленький,    большой текст,
   еще текст,      еще текст,          еще больший текст  ]
```

Bag of words. Описание

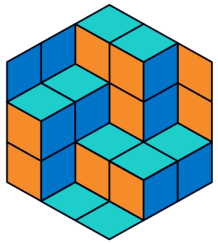


- Шаг 1: Сформировать корпус текстов
- Шаг 2: Все приводим в нижний регистр
- Шаг 3: Удаляются знаки пунктуация и «стоп-слова»
- Шаг 4: Производится стеминг – все слова к начальной форме
- Шаг 5: Определяем словарь

```
[ текст,          текст маленький,    большой текст,
  еще текст,      еще текст,          еще больший текст ]
```

```
[текст, маленький, большой, еще, больший]
```

Bag of words. Описание

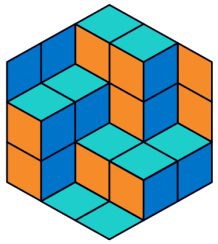


- Шаг 1: Сформировать корпус текстов
- Шаг 2: Все приводим в нижний регистр
- Шаг 3: Удаляются знаки пунктуация и «стоп-слова»
- Шаг 4: Производится стеминг – все слова к начальной форме
- Шаг 5: Определяем словарь
- Шаг 6: Для каждого текста считаем вектор вхождений слов

```
[ текст,          текст маленький,    большой текст,
  еще текст,      еще текст,          еще больший текст ]
```

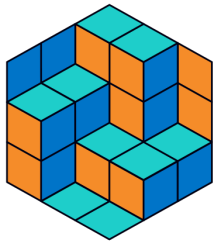
```
[текст, маленький, большой, еще, больший]
```

Bag of words. Описание



текст	[1,0,0,0,0]
текст маленький	[1,1,0,0,0]
большой текст	[1,0,1,0,0]
еще текст	[1,0,0,1,0]
еще текст	[1,0,0,1,0]
еще больший текст	[1,0,0,1,1]

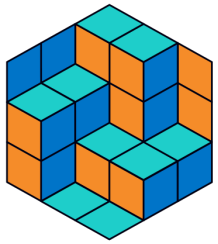
Bag of words. Описание



- Шаг 7 (опционально): уменьшаем размерность, переставая учитывать слова, которые очень редко встречаются

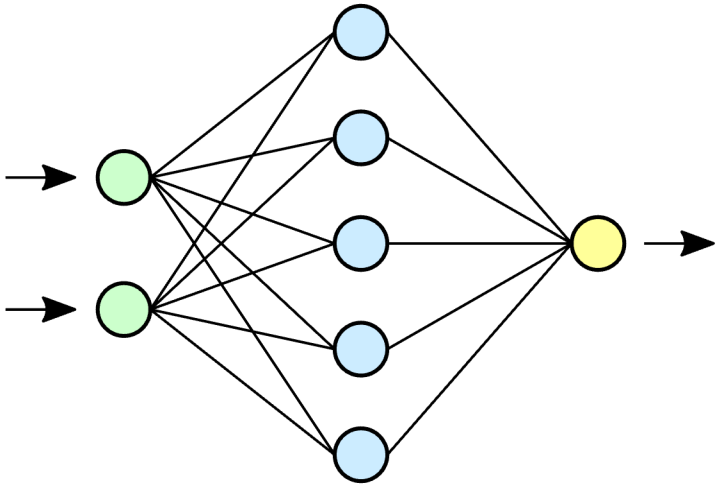
текст	→	[1,0,0,0,0]	→	[1,0]
текст маленький		[1,1,0,0,0]		[1,0]
большой текст		[1,0,1,0,0]		[1,0]
еще текст		[1,0,0,1,0]		[1,1]
еще текст		[1,0,0,1,0]		[1,1]
еще больший текст		[1,0,0,1,1]		[1,1]

Bag of words. Описание

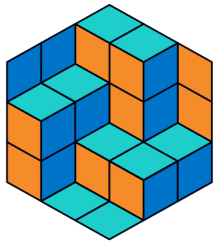


- Шаг 8 (последний): Запускаем обучаться модель

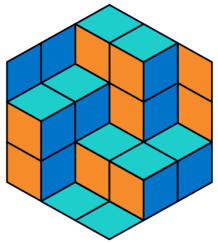
```
[1,0]
[1,0]
[1,0]
[1,1]
[1,1]
[1,1]
```



Какие задачи НЕ решает?

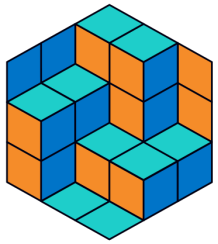


- Перевод
- Подстановка слова
- Любую другую задачу, где надо что-то генерировать



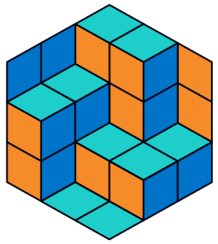
Глава 2
N-gramm. Language model.

Какие задачи решает?



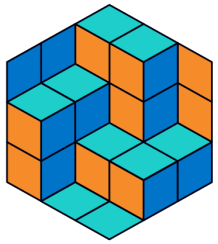
- Подстановка слова
- Генерация предложения

N-gramm



- Это вероятностная модель
- Поэтому может получиться что-то интересное
- …а может получиться что-то очень странное
- …а может вообще ничего не получиться

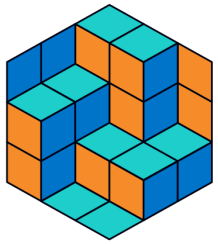
N-gramm



- Это вероятностная модель
- Поэтому может получиться что-то интересное
- …а может получиться что-то очень странное
- …а может вообще ничего не получиться

- Но обо всем по порядку!

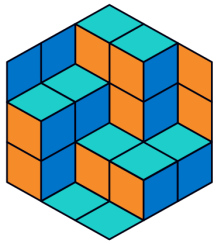
N-gramm



- Что такое вероятность получения предложения?
- То есть:

$$P(x_1, x_2, x_3, x_4, x_5)$$

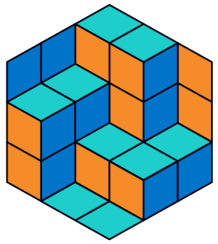
N-gramm



- Что такое вероятность получения предложения?
- То есть:

$$\underbrace{P(x_1, x_2, x_3, x_4, x_5)}_{\text{Слова}}$$

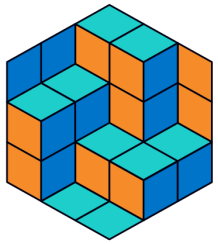
N-gramm



- Что такое вероятность получения предложения?
- То есть:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1) * P(x_2, x_3, x_4, x_5|x_1)$$
$$= P(x_1) * P(x_2|x_1) * P(x_3, x_4, x_5|x_1, x_2)$$
$$= \cdots$$
$$= P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

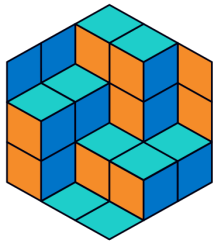


- Что такое вероятность получения предложения?
- То есть:

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1) * P(x_2, x_3, x_4, x_5|x_1)$$
$$= P(x_1) * P(x_2|x_1) * P(x_3, x_4, x_5|x_1, x_2)$$
$$= \cdots$$
$$= P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

- Получается, что вероятность предложения зависит от вероятности каждого слова, при условии фиксированных предыдущих

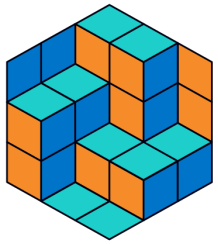
N-gramm



Пусть первое слово мы знаем – его вероятность 1

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

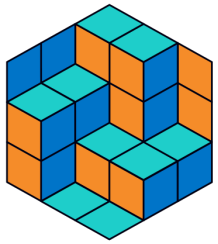


Пусть первое слово мы знаем – его вероятность 1

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

Тогда откуда взять это?

N-gramm



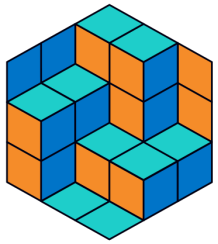
Пусть первое слово мы знаем – его вероятность 1

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

Тогда откуда взять это?

- Верно, из датасета!

N-gramm

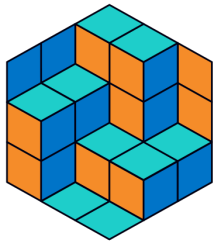


Датасет

```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



Датасет

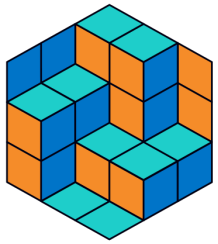
```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

→ Я

Пусть первое слово «Я», какие могут быть слова дальше, и какая у них вероятность?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



Датасет

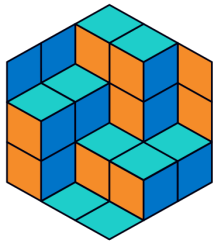
```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

→ Я люблю

Пусть первое слово «Я», какие могут быть слова дальше, и какая у них вероятность?
А дальше?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



Датасет

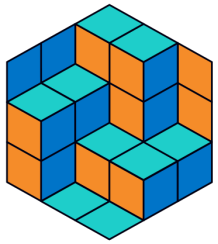
```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

```
Я люблю  решать (0.5)
         играть (0.5)
```

Пусть первое слово «Я», какие могут быть слова дальше, и какая у них вероятность?
А дальше?
Дальше?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



Датасет

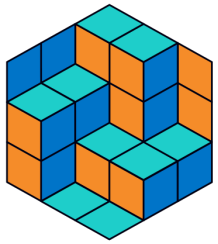
```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

→ `Я люблю решать` `математику (0.5)` `задачи (0.5)`

Пусть первое слово «Я», какие могут быть слова дальше, и какая у них вероятность?
А дальше?
Дальше?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



Датасет

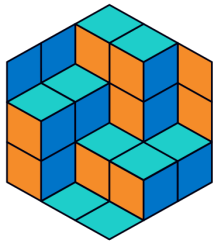
```
Я люблю решать математику
Я люблю играть в доту
Я люблю играть в кс
Я люблю решать задачи по математике
```

→ Я люблю решать математику

Пусть первое слово «Я», какие могут быть слова дальше, и какая у них вероятность?
А дальше?
Дальше?
Все!

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

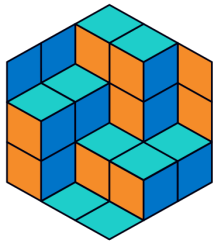
N-gramm



- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

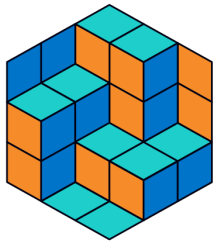


- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

Я

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



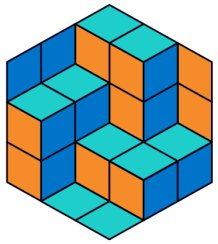
- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

Я

Я люблю

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

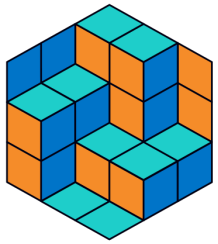


- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

```
Я
Я люблю
Я люблю решать
```

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

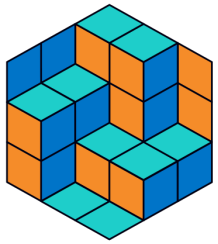


- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

```
Я
Я люблю
Я люблю решать
Я люблю решать задачи
```

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

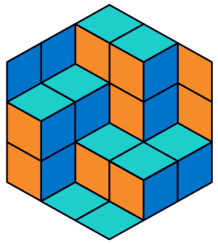


- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

```
Я
Я люблю
Я люблю решать
Я люблю решать задачи
Я люблю решать задачи по
```

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm

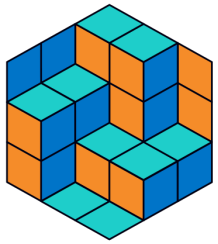


- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

```
Я
Я люблю
Я люблю решать
Я люблю решать задачи
Я люблю решать задачи по
Я люблю решать задачи по математике
```

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



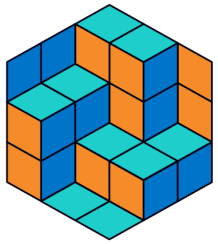
- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

Я
Я люблю
Я люблю решать
Я люблю решать задачи
Я люблю решать задачи по
Я люблю решать задачи по математике

В случае, если бы в датасете было
«Я никогда не любил решать задачи по биологии»
То в каких местах вероятности бы изменились?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

N-gramm



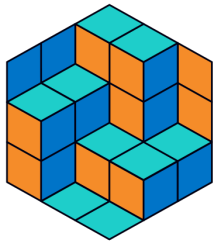
- Принцип N-gramm ровно такой же –только мы не смотрим назад дальше, чем на N слов
- Для N = 3:

Я
Я люблю
Я люблю решать
Я люблю решать задачи
Я люблю решать задачи по
Я люблю решать задачи по математике

В случае, если бы в датасете было
«Я никогда не любил решать задачи по биологии»
То в каких местах вероятности бы изменились?

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_1, x_2, x_3, x_4)$$

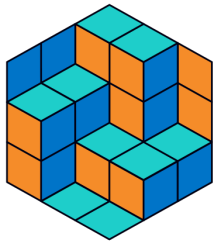
N-gramm



- Итого, формула примет вид:

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_2, x_3, x_4) * P(x_6|x_3, x_4, x_5) * \cdots$$

N-gramm

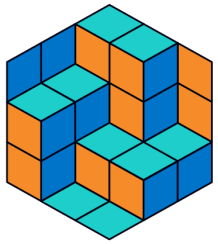


- Итого, формула примет вид:

$$P(x_1) * P(x_2|x_1) * P(x_3|x_1, x_2) * P(x_4|x_1, x_2, x_3) * P(x_5|x_2, x_3, x_4) * P(x_6|x_3, x_4, x_5) * \cdots$$

- И для реализации такой модели достаточно хранить
  - Все последовательные тройки слов
  - Для каждой тройки хранить возможные за ней слова и их вероятности

Вопросы на подумать



- Можно ли вместо слов использовать Токены?
- Что делать, если получится тройка, после которой мы не видели ни одного слова?
- Как выбрать N?

Еще вопросы?