


---

Работа с текстовыми  
данными.



---

# Содержание

1. Решаемые задачи
2. Проблемы работы с текстом
3. Counter
4. TFIDF
5. WORD2VEC
6. Меры близости
7. Наивный байесовский классификатор

# Основные задачи

- Машинный перевод
- Классификация текстов
  - Фильтрация спама
  - По тональности
  - По теме или жанру
- Кластеризация текстов
- Извлечение информации
  - Фактов и событий
  - Именованных сущностей
- Вопросно-ответные системы
- Суммаризация текстов
- Генерация текстов
- Распознавание речи
- Проверка правописания
- Оптическое распознавание символов
- Пользовательские эксперименты и оценка точности и качества методов

# Проблемы работы с текстом

Сколько слов в этом предложении?

- На дворе трава, на траве дрова, не руби дрова на траве двора.

# Проблемы работы с текстом

Сколько слов в этом предложении?

- На дворе трава, на траве дрова, не руби дрова на траве двора.

**\*\* 12 токенов\*\*** : На, дворе, трава, на, траве, дрова, не, руби, дрова, на, траве, двора

**\*\* 8 - 9 типов\*\*** : Н/на, дворе, трава, траве, дрова, не, руби, двора.

**\*\* 6 лексем\*\*** : на, не, двор, трава, дрова, рубить

Токен и тип

**\*\* Тип \*\*** – уникальное слово из текста

**\*\* Токен \*\*** – тип и его позиция в тексте

# Проблемы работы с текстом

Компьютеры работают с числами

Слово -> Число ?

	Слово 1	Слово 2	...	Слово N
Документ 1	1	0		2
Документ 2	0	0		0
...				
Докумен K	0	1		0

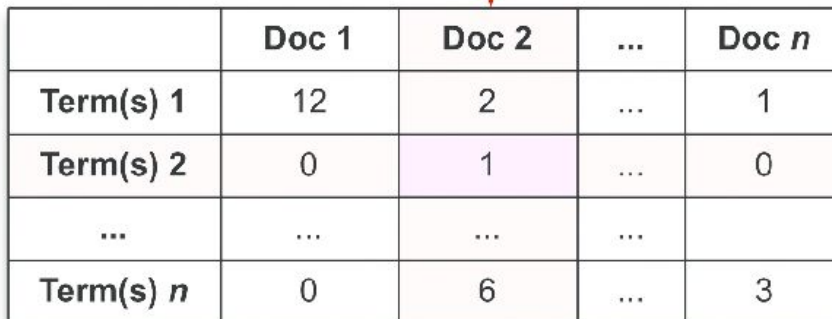
Посчитаем слова в документах.

# TF IDF

**TF** (*term frequency* — частота слова) — отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа.

**IDF** (*inverse document frequency* — обратная частота документа) — инверсия частоты, с которой некоторое слово встречается в документах коллекции. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение IDF.

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$



	Doc 1	Doc 2	...	Doc n
Term(s) 1	12	2	...	1
Term(s) 2	0	1	...	0
...	...	...	...	
Term(s) n	0	6	...	3

# TF IDF

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

$$\text{idf}(t, D) = \ln \left( \frac{|D|}{|\{d \in D : t \in d\}|} \right)$$

Выбор основания логарифма в формуле не имеет значения, поскольку изменение основания приводит к изменению веса каждого слова на постоянный множитель, что не влияет на соотношение весов.

Большой вес в TF-IDF получают слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Таким образом, мера TF-IDF является произведением двух сомножителей:

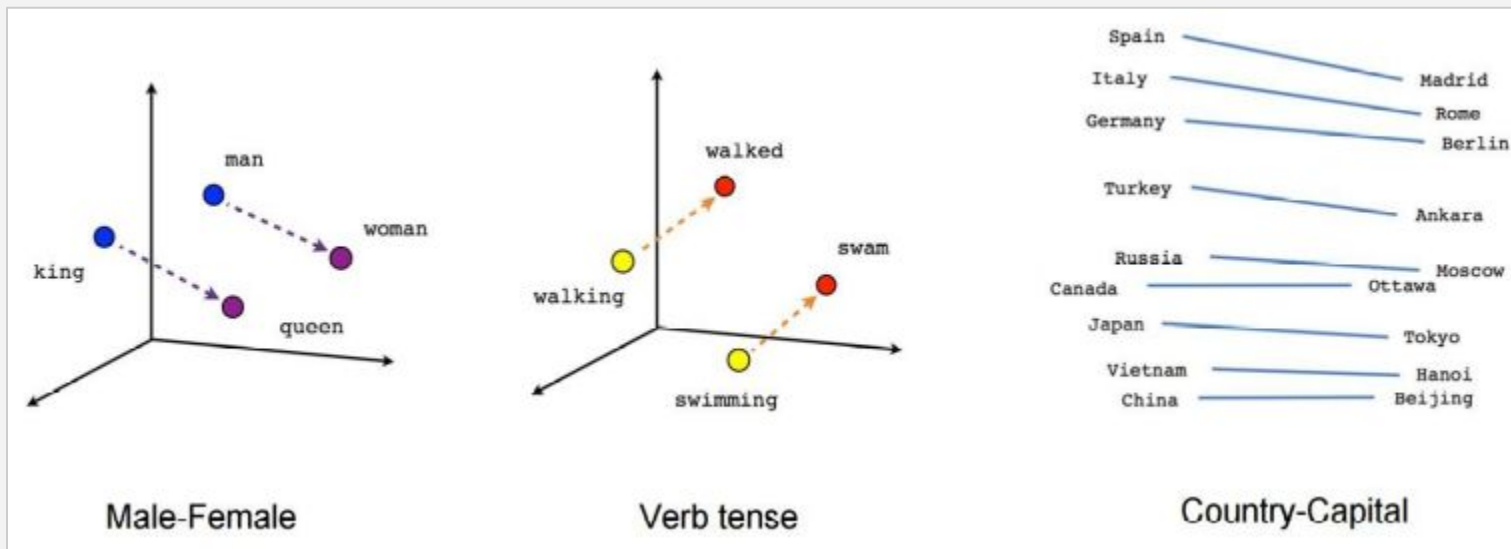
$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$



# Векторное представление слов

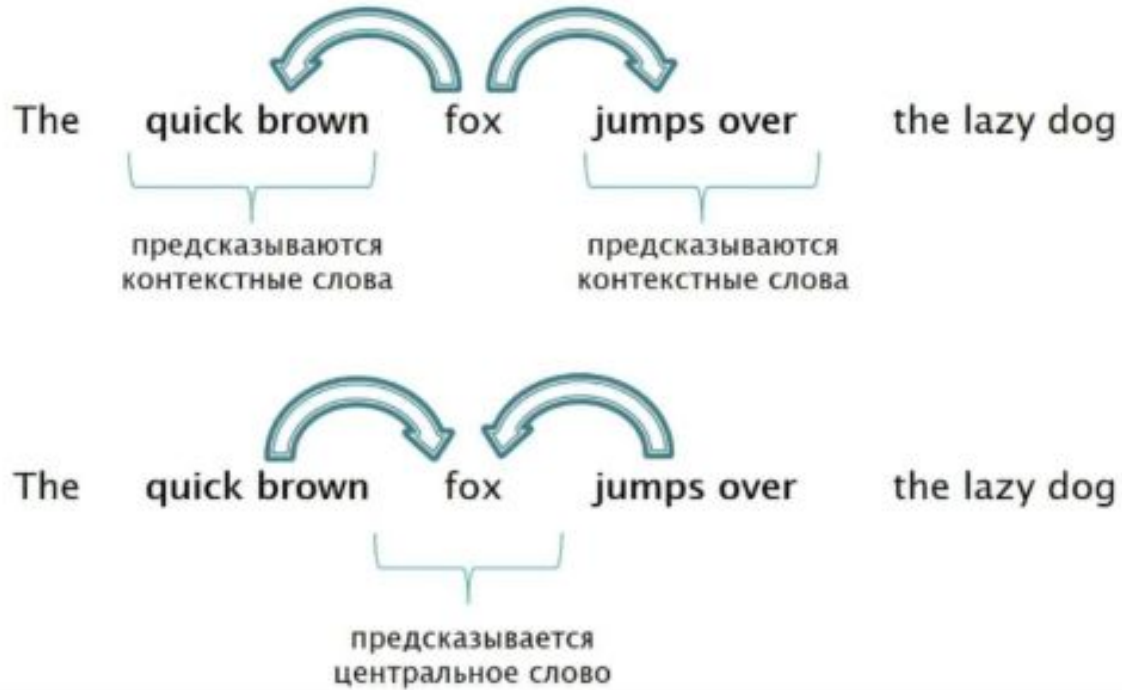
Традиционные методы - Bag of Words	Word Embeddings
<ul style="list-style-type: none"><li>• one hot encoding</li><li>• Каждое слово в словаре представляется одной единицей в большом словаре</li><li>• Информация в контексте не используется</li><li>• Например если в словаре 10000 слов и слово Hello – 4-е слово в словаре, то оно представляется вектором 000100....000</li></ul>	<ul style="list-style-type: none"><li>• Представляет каждое слово как точку в пространстве с фиксированной размерностью</li><li>• Unsupervised, строится на основе большого корпуса текста</li><li>• К примеру слово Hello может быть представлено: [0.4, -0.11, 0.55, 0.3 . . . 0.1, 0.02]</li></ul>

# Векторное представление слов



$$\text{vector}[\text{Queen}] = \text{vector}[\text{King}] - \text{vector}[\text{Man}] + \text{vector}[\text{Woman}]$$

# word2vec



# Мера близости

**Коэффициент сходства** - безразмерный показатель сходства сравниваемых объектов.

Эвклидова норма

$$\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$$

$$\rho_2(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Манхэттенская норма

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

$$\rho_1(x, y) = \|x - y\|_1 = \sum_{i=1}^n |x_i - y_i|$$

Косинусное расстояние

$$\langle x, y \rangle = \|x\| \|y\| \cos(\alpha) \implies \cos(\alpha) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$$

**Мера семантической близости** - это особая мера близости, предназначенная для количественной оценки семантической схожести лексем, например, существительных или многословных выражений.

- расстояние Левенштейна - минимальное количество односимвольных операций (вставки, удаления, замены)
- расстояние Хэмминга - число позиций, в которых соответствующие символы двух слов одинаковой длины различны
- коэффициент Жаккара - это мера, основанная на использовании информации о множестве общих символов. равна отношению пересечения двух множеств к их объединению

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

# Наивный байесовский классификатор

**Наивный байесовский классификатор** — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости.

**Формула Байеса:**

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)},$$

где

$P(A)$  — априорная вероятность гипотезы  $A$

$P(A | B)$  — вероятность гипотезы  $A$  при наступлении события  $B$  (апостериорная вероятность);

$P(B | A)$  — вероятность наступления события  $B$  при истинности гипотезы  $A$ ;

$P(B)$  — полная вероятность наступления события  $B$ .

Достоинством наивного байесовского классификатора является малое количество данных, необходимых для обучения, оценки параметров и классификации.

$$\arg \max [P(Q_k) \prod_{i=1}^n P(x_i | Q_k)]$$

$$P(Q_k) = \frac{\text{число документов класса } Q_k}{\text{общее количество документов}}$$

$$P(x_i | Q_k) = \frac{\alpha + N_{ik}}{\alpha M + N_k} \text{ — вхождение слова } x_i \text{ в документ класса } Q_k$$

$N_k$  — количество слов входящих в документ класса  $Q_k$

$M$  — количество слов из обучающей выборки

$N_{ik}$  — количество вхождений слова  $x_i$  в документ класса  $Q_k$

$\alpha$  — параметр для сглаживания

# Наивный байесовский классификатор

Спам:

- «Путевки по низкой цене»
- «Акция! Купи шоколадку и получи телефон в подарок»

Не спам:

- «Завтра состоится собрание»
- «Купи килограмм яблок и шоколадку»

Задание: определить, к какой категории отнести следующее письмо:

- «В магазине гора яблок. Купи семь килограмм и шоколадку»

Оценка для категории «Спам»:

$$\frac{2}{4} \cdot \frac{2}{23} \cdot \frac{2}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} \cdot \frac{1}{23} = 5,87E-10$$

Оценка для категории «Не спам»:

$$\frac{2}{4} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{2}{21} \cdot \frac{1}{21} \cdot \frac{1}{21} \cdot \frac{1}{21} = 4,44E-9$$

	Слово	Кол-во вхождений в «Спам»	Кол-во вхождений в «Не спам»	$P(x_i \text{Спам})$	$P(x_i \text{Не спам})$
Слова из обучающей выборки	Путевки	1	0		
	Низкой	1	0		
	Цене	1	0		
	Акция	1	0		
	Купи	1	1	$\frac{1+1}{14+9}$	$\frac{1+1}{14+7}$
	Шоколадку	1	1	$\frac{1+1}{14+9}$	$\frac{1+1}{14+7}$
	Получи	1	0		
	Телефон	1	0		
	Подарок	1	0		
	Завтра	0	1		
	Состоится	0	1		
	Собрание	0	1		
	Килограмм	0	1	$\frac{1+0}{14+9}$	$\frac{1+1}{14+7}$
	Яблок	0	1	$\frac{1+0}{14+9}$	$\frac{1+1}{14+7}$
	Магазине	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$
	Гора	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$
	Семь	0	0	$\frac{1+0}{14+9}$	$\frac{1+0}{14+7}$