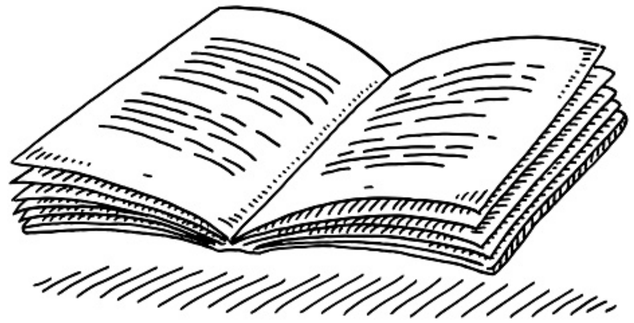
Работа с текстами

Елена Кантонистова

elena.kantonistova@yandex.ru
ekantonistova@hse.ru



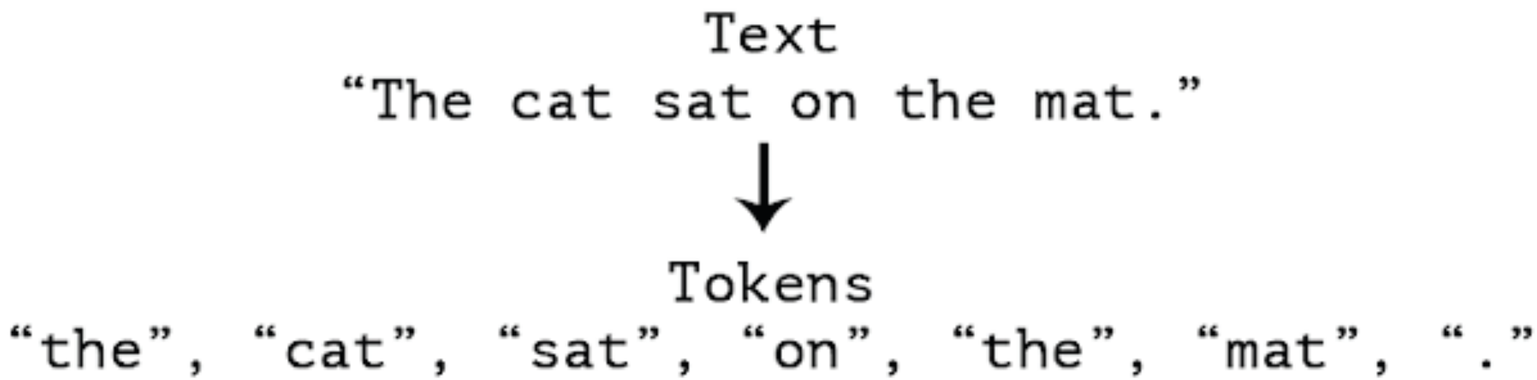
ВШЭ 2023

Терминология

- *Документ* - текст
- *Корпус* - набор документов
- *Токен* – формальное определение “слова”; токен может не иметь смыслового значения (например, “12fdh” или “авыдшл”), но обычно отделен от остальных токенов пробелами или знаками препинания

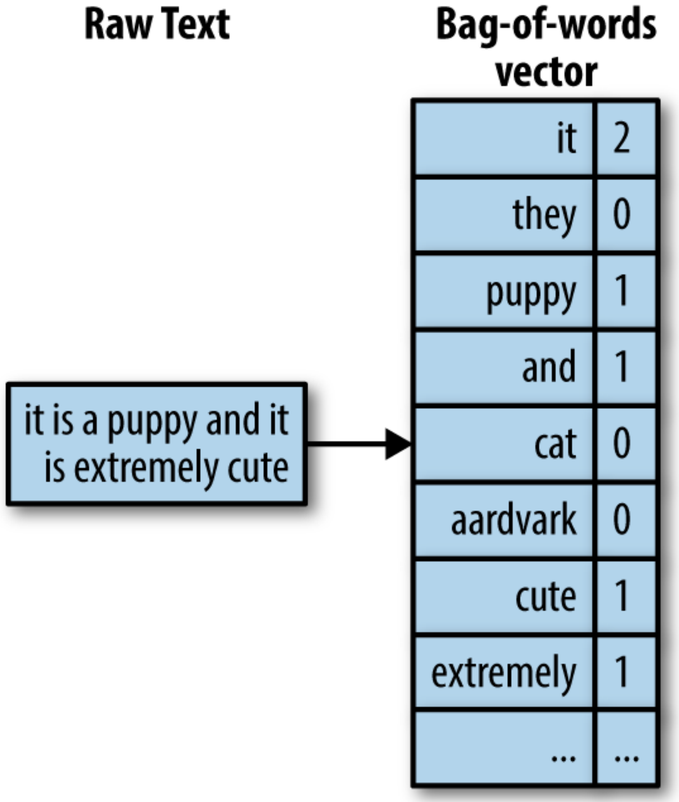
Токенизация текста

Чтобы работать с текстом, необходимо разбить его на токены. В простейшем случае токены – это слова (а также наборы букв, знаки препинания и т.д.).



Bag of words (мешок слов)

- По корпусу создадим словарь из всех встречающихся в нем слов (можно убрать общеупотребительные часто встречающиеся слова и очень редкие слова).
- Каждое слово закодируем вектором, в котором стоит единица на месте, соответствующем месту этого слова в словаре, все остальные компоненты вектора – 0.
- Для кодирования документа сложим коды всех его слов.



Bag of words (пример)

Пусть корпус состоит из следующих документов:

- D1 - “I am feeling very happy today”
- D2 - “I am not well today”
- D3 - “I wish I could go to play”

Кодировка этих документов будет такой:

	I	am	feeling	very	happy	today	not	well	wish	could	go	to	play
D1	1	1	1	1	1	1	0	0	0	0	0	0	0
D2	1	1	0	0	0	1	1	1	0	0	0	0	0
D3	2	0	0	0	0	0	0	0	1	1	1	1	1

Bag of words

Используя bag of words (BOW), мы теряем информацию о порядке слов в документе.

Пример: векторы документов “**I have no cats**” и “**No, I have cats**” будут идентичны.

Tf-idf

- Слова, которые редко встречаются в корпусе, но присутствуют в документе, могут оказаться важными для характеристики документа
- Слова, которые встречаются во всех документах, наоборот, не важны.

Tf-idf

Tf-idf слова $\boldsymbol{t}$ в документе $\boldsymbol{d}$ из корпуса $\boldsymbol{D}$:

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

- $tf(t, d)$ - частота вхождения слова t в документ d
- $idf(t, D)$ - величина, обратная частоте, с которой слово t встречается в корпусе D (обычно от нее еще берут логарифм)

Tf-idf

D1: He is a lazy boy. She is also lazy.

D2: Neeraj is a lazy person.

	He	She	lazy	boy	Neeraj	person
D1	0.06	0.06	0	0.06	0	0
D2	0	0	0	0	0.1	0.1