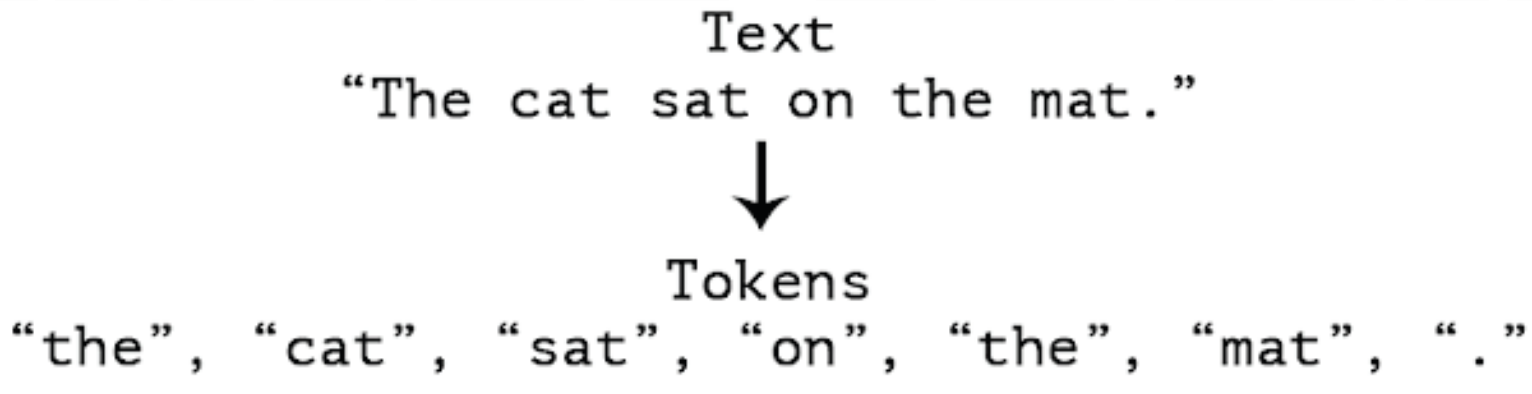
РАБОТА С ТЕКСТОВЫМИ ДАННЫМИ

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

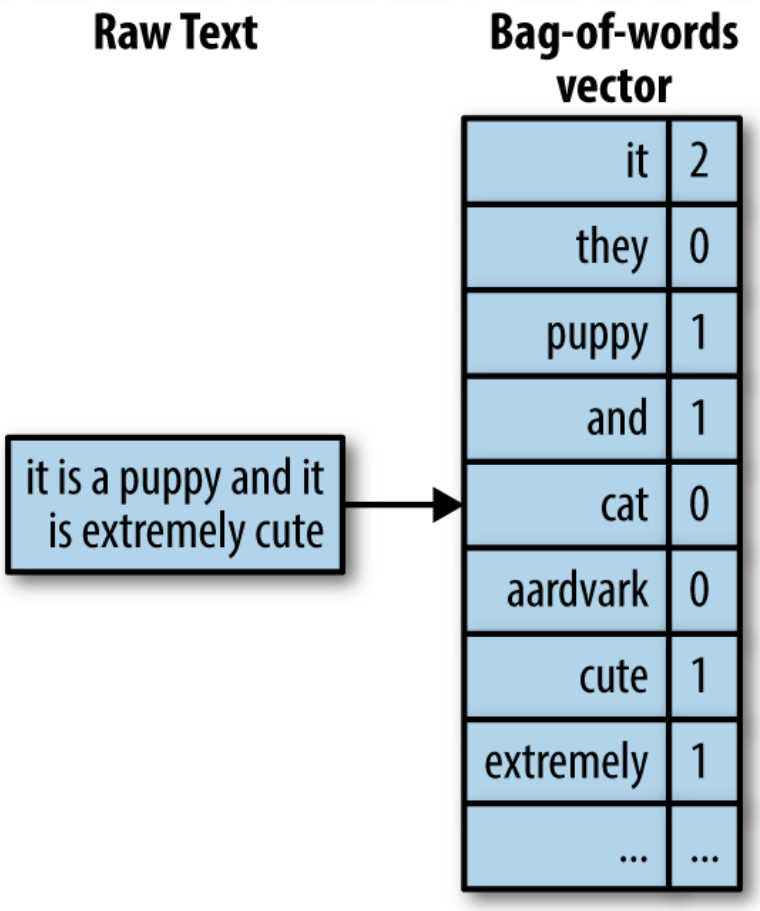
ТОКЕНИЗАЦИЯ ТЕКСТА

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.

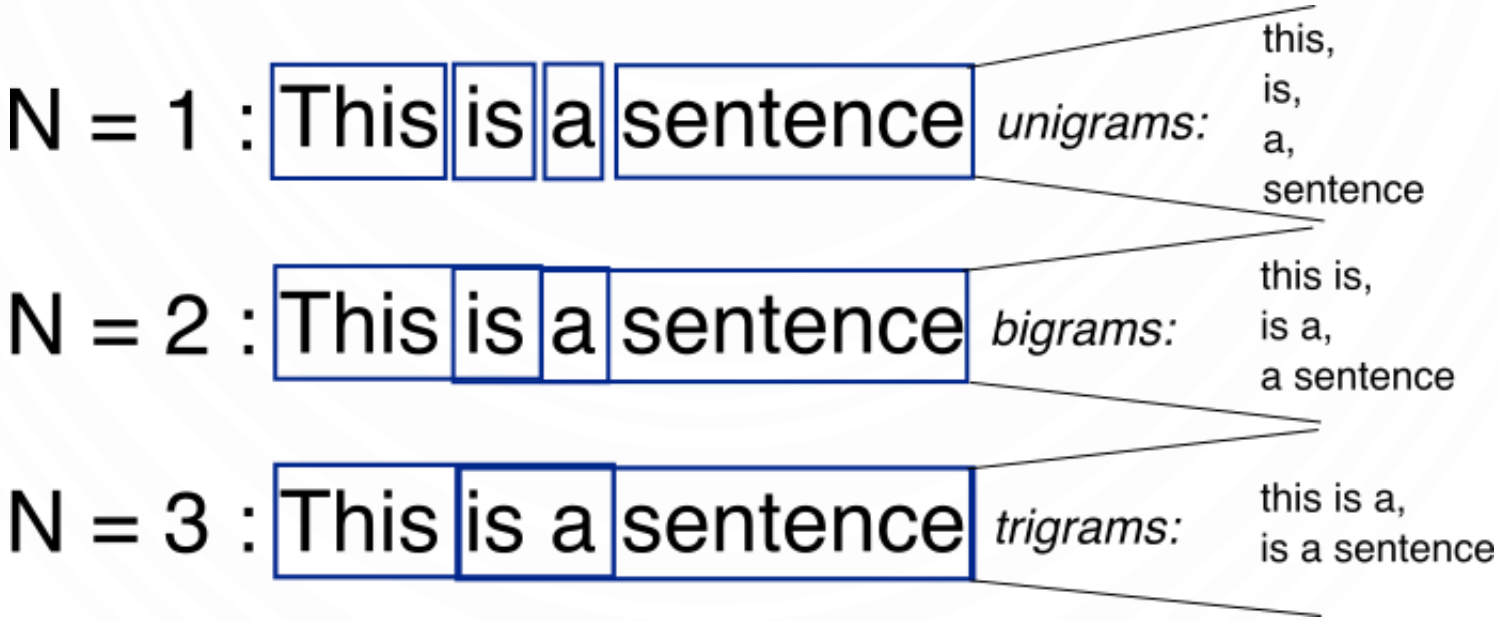
Пример: векторы документов “I have no cats” и “No, I have cats” будут идентичны.

N-GRAM BAG OF WORDS

В качестве слов в словаре можно использовать:

- N-граммы из букв (наборы букв длины N в слове)
- N-граммы из слов (наборы фраз длины N в документе)

Такой подход поможет учесть сходственные слова и опечатки.



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 (term frequency – inverse document frequency):

- *$tf(t,d)$ - частота вхождения слова t в документ d:*

$$\boldsymbol{tf(t,d)} = \frac{\boldsymbol{n_t}}{\sum_{\boldsymbol{k}} \boldsymbol{n_k}} = \frac{\text{число вхождений слова } \boldsymbol{t} \text{ в документ}}{\text{общее число слов в документе}}$$

$tf(t,d)$ показывает важность слова t в документе d.

TF-IDF

- $tf(t, d)$ - *частота вхождения слова* t *в документ* d:

$$\boldsymbol{tf(t, d)} = \frac{\boldsymbol{n_t}}{\sum_k \boldsymbol{n_k}} = \frac{\text{число вхождений слова } \boldsymbol{t} \text{ в документ}}{\text{общее число слов в документе}}$$

$tf(t, d)$ *показывает важность слова* t *в документе* d.

- $idf(t, D)$ - *величина, обратная частоте, с которой слово* t *встречается в документах корпуса* D.

$$\boldsymbol{idf(t, D)} = \log \frac{|\boldsymbol{D}|}{|\{\boldsymbol{d_i} \in \boldsymbol{D} \mid \boldsymbol{t} \in \boldsymbol{d_i}\}|},$$

$|D|$ — число документов *в корпусе,*

$|\{d_i \in D \mid t \in d_i\}|$ - *число документов, в которых встречается слово* t

Учёт idf *уменьшает вес часто используемых в корпусе слов.*

TF-IDF

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

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

Пример:

Дана коллекция D из $10000000 = 10^7$ документов, в 1000 из них встречается слово “заяц”. В данном документе d из коллекции 100 слов, и слово “заяц” встречается 3 раза.

$$tf(\text{заяц}, d) = \frac{3}{100} = 0{,}03$$

$$idf(\text{заяц}, D) = \log\left(\frac{10^7}{10^3}\right) = 4$$

Поэтому $tfidf(\text{заяц}, d, D) = 0{,}03 \cdot 4 = 0{,}12$.

ИНТЕРПРЕТАЦИЯ ЛИНЕЙНОЙ МОДЕЛИ

text	label
отвратительное обслуживание был у меня вклад в...	0
мнение о банке изменилось в худшую сторону это...	0
банк поступил красиво у меня дебетовая карта б...	1
прошу принять меры по исправлению ситуации бан...	0
спокойно и качественно пользуюсь услугами альф...	1

ИНТЕРПРЕТАЦИЯ ЛИНЕЙНОЙ МОДЕЛИ

- 0.99 accuracy на обучении
- 0.93 accuracy на валидации

Примеры весов

```
спасибо 15.3812631501
приятно 10.195153067
благодарность 8.75099611487
оперативность 7.9119980712
быстро 7.20768729913
всегда 6.49503091778
оперативно 6.36190679808
большое 6.02762583473
доволен 5.86536526776
отзыв 5.64047141286
помощь 5.43980835894
поблагодарить 5.19673514028
```

```
претензию -3.84736026948
не работает -3.89934654597
два -3.9180675684
звонков -3.99518600488
готовности -4.00435284458
говорят -4.10305804728
дозвониться -4.10647379932
пусть -4.20500663563
видимо -4.32809243057
не -4.59523464931
звонки -4.63261991797
отказ -4.90228031373
```