ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ ИЗ ТЕКСТА

МЕШОК СЛОВ

МФТИ

- Текст можно анализировать без учёта порядка слов
- Достаточно знать, какие слова и сколько раз встретились

СЧЁТЧИКИ СЛОВ

МФТИ

- Пусть всего в выборке N различных слов: $\omega_1, \ldots, \omega_N$
- Кодируем тексты с помощью N признаков
- j-й признак — доля вхождений слова ω_j среди всех вхождений слов в документе

СЧЁТЧИКИ СЛОВ

МФТИ

- Пример: “текст состоит из слов”, “вхождения данного слова среди всех слов”

текст	состоит	слово	вхождение	данный	все
0.33	0.33	0.33	0	0	0
0	0	0.4	0.2	0.2	0.2

СЧЁТЧИКИ СЛОВ

МФТИ

- Стоп-слова — слова, которые встречаются очень часто и не несут в себе информацию
- Редкие слова имеет смысл удалять

TF-IDF

МФТИ

- Если слово часто встречается в документе, то оно важно для документа
- Если слово редко встречается в других документах, то оно важно для документа

TF-IDF

МФТИ

$$TDF - IDF(x, \omega) = n_{dw} \log \frac{\ell}{n_\omega}$$

- n_{dw} — доля вхождений слова ω в документ d
- n_w — количество документов, в которых есть слово ω

РЕЗЮМЕ

МФТИ

- Для извлечения признаков из текстов хорошо работает подход “мешок слов”
- Имеет смысл удалять редкие слова и стоп-слова
- TF-IDF учитывает все документы в выборке при вычислении важности слова