ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ ИЗ ТЕКСТА-2

УЧЁТ ПОРЯДКА ТОКЕНОВ

МФТИ

- «Мешок слов» никак не учитывает порядок слов
- Порядок слов важен: «нравится» и «не нравится»
- Учёт словосочетаний расширяет признаковое пространство
- Можно находить сложные закономерности простыми моделями

N-ГРАММЫ

МФТИ

- Наборы из n подряд идущих токенов
- Пример: «Наборы подряд идущих токенов»
  - Униграммы: наборы, подряд, идущих, токенов
  - Биграммы: наборы подряд, подряд идущих, идущих токенов
  - Триграммы: наборы подряд идущих, подряд идущих токенов

N-ГРАММЫ

МФТИ

- Наборы из n подряд идущих токенов
- К признакам добавляются счётчики или TF-IDF по всем n-граммам
- n — гиперпараметр, увеличение может привести к переобучению

БУКВЕННЫЕ N-ГРАММЫ

МФТИ

- В качестве токенов можно рассматривать буквы
- Признаки — счётчики/TF-IDF для буквенных n-грамм
- Позволяет учитывать смайлы, незнакомые формы слов и т.д.

SKIP-ГРАММЫ

МФТИ

- k-skip-n-граммы —наборы из n токенов, между соседними должно быть не более k токенов
- Пример: «Наборы подряд идущих токенов»
  - Биграммы: наборы подряд, подряд идущих, идущих токенов
  - 1-skip-2-граммы: наборы подряд, подряд идущих, идущих токенов, наборы идущих, подряд токенов

ХЭШИРОВАНИЕ

МФТИ

- $h(x)$ — хэш-функция с 2^n возможными значениями
- Используем 2^n признаков-счётчиков
- Каждое слово x заменяем на его хэш $h(x)$

ХЭШИРОВАНИЕ

МФТИ

- Позволяет сократить количество признаков
- Упрощает вычисление признаков
- Не требует хранения соответствия между словами и признаками

РЕЗЮМЕ

МФТИ

- n-граммы и k-skip-n-граммы
- Хэширование при подсчёте признаков