ПРЕДОБРАБОТКА ТЕКСТА

ПРЕДОБРАБОТКА ТЕКСТА

МФТИ

- Разбиение текста на отдельные «слова» (токены)
- Приведение слов к начальной форме (нормализация)

ТОКЕНИЗАЦИЯ

МФТИ

- Разбиение текста на отдельные «слова»
- Пример:

  Текст (от лат. textus — «ткань; сплетение, связь, паутина, сочетание») — зафиксированная на каком-либо материальном носителе человеческая мысль; в общем плане связная и полная последовательность символов.

ТОКЕНИЗАЦИЯ

МФТИ

- Разбиение текста на отдельные «слова»
- Пример:

  Текст (от лат. textus — «ткань; сплетение, связь, паутина, сочетание») — зафиксированная на каком-либо материальном носителе человеческая мысль; в общем плане связная и полная последовательность символов.

ТОКЕНИЗАЦИЯ

МФТИ

- Разбиение текста на отдельные «слова»
- Пример:

  Текст (от лат. textus — «ткань; сплетение, связь, паутина, сочетание») — зафиксированная на каком-либо материальном носителе человеческая мысль; в общем плане связная и полная последовательность символов.

ТОКЕНИЗАЦИЯ

МФТИ

- Приведение к нижнему регистру
  - Но регистр может нести информацию: «ООО» и «ооо»
- Замена всех знаков препинания и прочих символов на пробелы
  - Правильно ли это для сложных составных слов? («красно-чёрный»)
  - Смайлы могут нести информацию
- Каждое слово объявляется отдельным токеном
  - Некоторые наборы слов должны рассматриваться как одно: «Нижний Новгород», «к.т.н.»

ТОКЕНИЗАЦИЯ

МФТИ

- В некоторых языках слова пишутся без пробелов
- Китайский:
- Необходима сегментация текста на слова

НОРМАЛИЗАЦИЯ

МФТИ

- Приведение слов к начальной форме
- «машинное» → «машинный»
- «шёл» → «идти»
- Форма слова не всегда несёт в себе полезную информацию
- Может быть важно сократить количество различных слов
- Два подхода: стэмминг и лемматизация

СТЕММИНГ

МФТИ

- «Стрижка» окончаний слов по набору правил
- Не всегда имеет смысл: «был», «есть», «будет»

ЛЕММАТИЗАЦИЯ

МФТИ

- Приведение слов к начальной форме
- На основе словаря
- Если слова нет в словаре, то строится гипотеза о способе изменения слова
- Работает медленнее, чем стэмминг

РЕЗЮМЕ

МФТИ

- Предобработка текста состоит из токенизации и нормализации
- При токенизации следует учитывать особенности задачи
- Два подхода к нормализации: стэмминг и лемматизация