РАБОТА
С ТЕКСТОВЫМИ ДАННЫМИ

ЗАДАЧИ АНАЛИЗА ТЕКСТОВ

МФТИ

- Предсказать рейтинг записи в блоге
- Определить эмоциональный окрас комментария
- Определить тематику научной статьи
- Сгруппировать новости по сюжетам
- Найти слова, похожие по смыслу на данное

ЗАДАЧИ АНАЛИЗА ТЕКСТОВ

МФТИ

- Выделить все упоминания имён в тексте
- Построить краткую аннотацию текста
- Построить модель, отвечающую на вопросы
- Сгенерировать новый текст, похожий на заданный набор текстов

ОСНОВНАЯ ЗАДАЧА

МФТИ

- Дан текстовый признак
- Преобразовать текст в матрицу «объекты-признаки»
- Описать текст произвольной длины фиксированным числом признаков

МЕШОК СЛОВ

МФТИ

- Пример: определение тематики текста (датасет 20newsgroups)

МЕШОК СЛОВ

МФТИ

- Пример: определение тематики текста (датасет 20newsgroups)

МЕШОК СЛОВ

МФТИ

- Пример: определение тематики текста (датасет 20newsgroups)
- Порядок слов не так важен для решения задачи
- Мешок слов — представление текста в виде счётчиков вхождений

РЕЗЮМЕ

МФТИ

- Основная цель при работе с текстами — признаковое описание

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

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

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

ОБУЧЕНИЕ МОДЕЛЕЙ НА ТЕКСТАХ

ПОДГОТОВКА ВЫБОРКИ

МФТИ

- Удаление редких и популярных слов

ПОДГОТОВКА ВЫБОРКИ

МФТИ

- Удаление редких и популярных слов
- Признаки: n-граммы + счётчики/TF-IDF

ПОДГОТОВКА ВЫБОРКИ

МФТИ

- Число признаков — $10^3 - 10^4$ и больше

ПОДГОТОВКА ВЫБОРКИ

МФТИ

- Число признаков — $10^3 - 10^4$ и больше
- Можно пробовать отбор признаков и понижение размерности

ОБУЧЕНИЕ МОДЕЛИ

МФТИ

- Случайный лес — низкая скорость обучения

ОБУЧЕНИЕ МОДЕЛИ

МФТИ

- Случайный лес — низкая скорость обучения
- Градиентный бустинг — проблемы из-за маленькой глубины деревьев

ОБУЧЕНИЕ МОДЕЛИ

МФТИ

- Случайный лес — низкая скорость обучения
- Градиентный бустинг — проблемы из-за маленькой глубины деревьев
- Наивный байесовский классификатор

ОБУЧЕНИЕ МОДЕЛИ

МФТИ

- Случайный лес — низкая скорость обучения
- Градиентный бустинг — проблемы из-за маленькой глубины деревьев
- Наивный байесовский классификатор
- Линейные модели используются чаще всего

ЛИНЕЙНЫЕ МОДЕЛИ И SGD

МФТИ

- Стохастический градиентный спуск позволяет читать с диска по одному объекту

ЛИНЕЙНЫЕ МОДЕЛИ И SGD

МФТИ

- Стохастический градиентный спуск позволяет читать с диска по одному объекту
- пока не выполнен критерий останова:
  $t =$ следующий текст
  для всех слов x в t :

$$w_{h(x)} = w_{h(x)} - \alpha \nabla_{w_{h(x)}} Q(w)$$

РЕЗЮМЕ

МФТИ

- n-граммы и мешок слов

РЕЗЮМЕ

МФТИ

- n-граммы и мешок слов
- Линейные модели и хэширование