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