# АНАЛИЗ ТОНАЛЬНОСТИ ТЕКСТА

# ПЛАН

МФТИ

- Примеры, применения и трудности
- Простая постановка задачи и простое решение
- Наблюдения из практики
- Возможные постановки задачи
- Данные

# ПРИМЕРЫ, ПРИМЕНЕНИЯ И ТРУДНОСТИ

# ПРИМЕР НА АНАЛИЗ ТОНАЛЬНОСТИ

МФТИ

- Я купил этот телефон две недели назад.
  Всё изначально было хорошо.
  Отличный звук, батарея жила долго.
  Но вчера он перестал работать.
- Объективные и субъективные предложения
- Характеристика текста в целом и отдельных предложений
- Характеристики: общее впечатление, звук, батарея

# ПРИМЕНЕНИЯ SENTIMENT ANALYSIS

МФТИ

- Для потребителя: анализ отзывов на товары
- Для организаций: замена опросов и фокус-групп
- Политика: результаты выборов и мнение избирателей
- Биржевые торги: анализ оценок экспертов и предсказание курсов

# SA: ОСНОВНЫЕ ТРУДНОСТИ

МФТИ

- Тексты от пользователей отличаются от текстов, прошедших редактуру
- Люди используют различные наборы слов в зависимости от пола, возраста, страны проживания...
- Слова меняют эмоциональную окраску в зависимости от предмета описания
- Сарказм
- Каждый сайт с отзывами навязывает некоторую модель написания текста

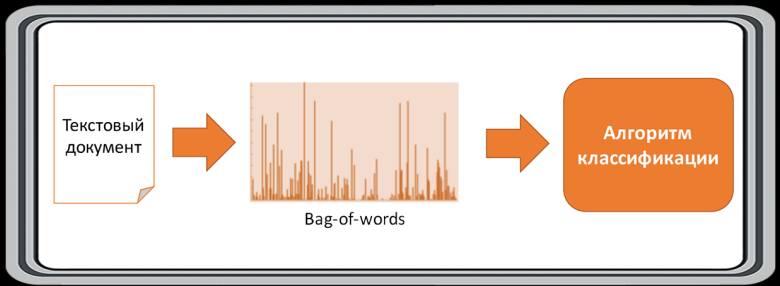
# ПРОСТАЯ ПОСТАНОВКА ЗАДАЧИ И ПРОСТОЕ РЕШЕНИЕ

# ПРОСТАЯ ПОСТАНОВКА ЗАДАЧИ

МФТИ

- $2$ класса — позитивный и негативный
- Признаки — обычный мешок слов

# ПРОСТОЕ РЕШЕНИЕ ЗАДАЧИ



# НАБЛЮДЕНИЯ ИЗ ПРАКТИКИ

# ЛЕММАТИЗАЦИЯ И СТЕММИРОВАНИЕ

МФТИ

- Лемматизация — приведение в нормальную форму:
  - летел → лететь
  - самолетами → самолет
  - идешь → идти
  - шел → идти
- Стеммирование — выделение основы слова:
  - летел → лет
  - самолетами → самолет
  - идешь → ид
  - шел → шел
- Часто не улучшает качество в сентимент-анализе

# ОТРИЦАНИЯ

МФТИ

- Отзывы разной тональности с одинаковым представлением в модели мешка слов:
  - Это лучшая модель, экран не отвратительный, как в прошлой
  - Это не лучшая модель, экран отвратительный, как в прошлой

- Простейший способ – объединять с частицей «не» в один токен:
  - не отвратительный $\rightarrow$ не_отвратительный
  - не лучшая $\rightarrow$ не_лучшая

# РАЗГОВОРНЫЙ ЯЗЫК, ОШИБКИ

МФТИ

- Частоты буквенных n-грамм вместо частот слов позволят похожим образом учитывать в тексте разные варианты написания одного слова

- Пример с триграммами:
  - ужасно $\rightarrow$ (ужа, жас, асн, сно)
  - ужаааасно $\rightarrow$ (ужа, жаа, ааа, аас, асн, сно)

- В текстах с этими словами будет хотя бы три общих токена

# ПУНКТУАЦИЯ

МФТИ

- Сижу в кино на «Вспомнить все» :)
- Вчера купил новый айфон, сложно описать эмоции словами :(
- Очень рекомендую эту модель!

# ВОЗМОЖНЫЕ ПОСТАНОВКИ ЗАДАЧИ

# МУЛЬТИКЛАССОВАЯ КЛАССИФИКАЦИЯ

МФТИ

- Классы отзывов:
  - Положительные
  - Негативные
  - Нейтральные

- Проблема: отнести негативный к нейтральному — не так плохо, как в позитивному

# ПРОГНОЗИРОВАНИЕ ОЦЕНКИ

МФТИ

- Обучать алгоритм предсказывать не класс, а оценку
- В этом случае, конечно, надо решать задачу регрессии
- Плюс: алгоритм начинает чувствовать разную цену ошибок
- Минус: плохо интерпретируемый функционал качества

- Два класса — позитивные и негативные отзывы
- Когда не уверены — говорим, что отзыв «без яркой эмоциональной окраски»
- При доработке:
  - Повышаем качество вне серой зоны
  - Уменьшаем её размер

# УРОВНИ SA

МФТИ

- Документ
  - Положительное или отрицательное мнение или отношение выражает данный документ?
- Предложение
  - Предположение: «маленький документ», содержащий только одно мнение
  - Фактически — промежуточный этап

# УРОВНИ SA

МФТИ

- Предложение
  - Предположение: «маленький документ», содержащий только одно мнение
  - Фактически — промежуточный этап

- Аспект

  Примеры:
  - отличный звук
  - батарея живет долго
  - дисплей яркий

МФТИ

# ДАННЫЕ

# ВАРИАНТ 1: ВЗЯТЬ ГОТОВЫЙ ДАТАСЕТ

МФТИ

- https://www.cs.cornell.edu/people/pabo/movie-review-data/
- http://www.sananalytics.com/lab/twitter-sentiment/
- http://inclass.kaggle.com/c/si650winter11/data
- http://nlp.stanford.edu/sentiment/treebank.html

# ВАРИАНТ 2: ПАРСИТЬ САЙТ С ОТЗЫВАМИ

МФТИ

- Примеры:
  - сайты с отзывами на фильмы
  - сайты интернет-магазинов
  - сайты с отзывами на работу организаций и компаний

# ИТОГ

МФТИ

- Примеры, применения и трудности
- Простая постановка задачи и простое решение
- Наблюдения из практики
- Возможные постановки задачи
- Данные