



# Методы машинного обучения

ИУ-5, магистратура, 2 семестр,  
весна 2021 года



# Методы предобработки и векторизации текстов



# Предобработка текста

- Для построения моделей машинного обучения необходимо предварительно предобработать текст. На этапе предобработки, как правило, используется только лингвистическая информация. Предобработка может включать следующие шаги:
  1. Уровень символов:
    - Токенизация – разбиение текста на токены (как правило, предложения или слова).
  2. Уровень слов:
    - Нормализация текста – лемматизация или стемминг.
    - Если текст получен путем генерации из аудио, то возможно к словам придется применять фонетические алгоритмы и искать наиболее похожие в словаре слова с использованием расстояния Левенштейна.
    - Частеречная разметка (Part-Of-Speech tagging, POS-tagging).
    - Выделение (распознавание) именованных сущностей, named-entity recognition (NER).
  3. Уровень предложений:
    - Разбор предложения, выделение семантических ролей.
- Пример предобработки текста.

# Модель мешка слов – простейший вариант векторизации

- Модель мешка слов.
- Использование CountVectorizer для построения мешка слов.
- Метрика TF-IDF.
- Использование TfidfVectorizer для построения мешка слов.
- Пример решения задачи анализа тональности текста с помощью модели мешка слов.



# Векторные представления слов и документов

- Для рассматриваемых подходов используется термин «неглубокие векторные представления», так как в настоящее время существуют более сложные модели векторизации (например, BERT), основанные на глубоком обучении.
- Обзор моделей
- Модель word2vec
  - статья на русском языке, оригинальная статья Т.Миколова, пример визуализации
- Модель Glove
  - оригинальная статья
- Модель fasttext
  - Официальный сайт , оригинальная статья Т. Миколова
- По аналогии с векторными представлениями для слов, можно строить векторные представления для документов
  - Модель doc2vec, оригинальная статья.
- Примеры использования моделей.