Лекция 13. Текстовый анализ и обработка естественного языка (NLP)
Тема: Bag-of-Words, TF-IDF, Word2Vec, Трансформеры

---

1. Введение

В современном мире огромная часть информации представлена в **текстовой форме** — новости, статьи, отзывы, электронные письма, посты в соцсетях. Для анализа такой информации применяется направление **обработки естественного языка (Natural Language Processing, NLP)** — совокупность методов и алгоритмов, позволяющих компьютерам понимать, интерпретировать и генерировать человеческий язык.

Цель NLP — извлечь смысл и структуру из текста для последующего анализа, классификации, перевода, резюмирования, диалога или генерации новых текстов.

---

2. Основные задачи NLP

- **Классификация текста** (определение темы, тональности, категории документа);
- **Анализ тональности** (sentiment analysis — определение эмоциональной окраски текста);
- **Распознавание именованных сущностей (NER)** — выделение людей, мест, организаций;
- **Машинный перевод** (Google Translate, DeepL и др.);
- **Резюмирование текстов**;
- **Диалоговые системы и чат-боты** (ChatGPT, Siri, Alexa).

---

3. Представление текста в виде чисел

Для машинного анализа текст нужно преобразовать в числовой формат — **вектор признаков**.
Существует несколько способов векторизации текста, от простых статистических до нейронных моделей.

---

4. Модель Bag-of-Words (Мешок слов)

Bag-of-Words (BoW) — один из первых и простейших способов представления текста.
Он рассматривает текст как **набор слов без учёта порядка**.

Принцип работы:

1. Создаётся словарь всех слов в корпусе.
2. Каждый документ представляется вектором, где каждый элемент — количество появлений слова.

Пример:
Тексты:

1. «Кошка сидит на окне»
2. «Собака лежит на полу»

Словарь: {кошка, сидит, окно, собака, лежит, пол, на}

Вектор для текста 1: [1, 1, 1, 0, 0, 0, 1]
Вектор для текста 2: [0, 0, 0, 1, 1, 1, 1]

Плюсы: простота, легкость реализации.
Минусы: не учитывает порядок слов, значения, синонимы; большие разреженные матрицы.

---

5. TF-IDF (Term Frequency — Inverse Document Frequency)

Модель **TF-IDF** улучшает BoW, учитывая **важность** слова для конкретного документа.
Она снижает вес часто встречающихся слов («и», «в», «это») и усиливает редкие, но значимые слова.

TF-IDF(t,d)=TF(t,d)×log⁡NDF(t)\text{TF-IDF}(t, d) = TF(t, d) \times \log \frac{N}{DF(t)}TF-IDF(t,d)=TF(t,d)×logDF(t)N

где:

- TF(t,d)TF(t, d)TF(t,d) — частота слова t в документе d;
- DF(t)DF(t)DF(t) — число документов, содержащих t;
- NNN — общее число документов.

Применение: классификация текстов, поиск, фильтрация спама.
Преимущества: простота, эффективность для традиционных моделей (SVM, Logistic Regression).
Недостатки: не учитывает контекст и порядок слов.

---

6. Word2Vec: нейронные векторные представления слов

Word2Vec — революционная модель, предложенная Google в 2013 году. Она преобразует слова в **векторы фиксированной длины**, где похожие по смыслу слова имеют близкие координаты.

Например:

- вектор(«король») – вектор(«мужчина») + вектор(«женщина») ≈ вектор(«королева»)

Два варианта архитектуры Word2Vec:

1. **CBOW (Continuous Bag-of-Words)** — предсказывает слово по контексту;
2. **Skip-Gram** — предсказывает контекст по слову.

Преимущества:

- отражает семантические и синтаксические связи слов;
- компактные и плотные векторы.

Недостатки:

- одно значение для слова (нет учёта контекста);
- требует большого корпуса данных.

---

7. Современные модели: Трансформеры

С появлением **архитектуры трансформеров (Transformers)** в 2017 году (Vaswani et al., *“Attention Is All You Need”*), NLP пережил революцию. Трансформеры позволяют моделировать контекст на уровне **всего предложения** с помощью механизма **внимания (Attention)**, который оценивает важность каждого слова по отношению к другим.

Основные идеи:

- Не используют рекуррентные связи, что ускоряет обучение.
- Работают параллельно с большими объемами текста.
- Учитывают контекст слова во всей фразе.

Популярные модели на базе трансформеров:

- **BERT** — для анализа контекста и классификации текста.
- **GPT (Generative Pre-trained Transformer)** — для генерации текста и диалогов.
- **T5, RoBERTa, XLNet** — специализированные модели для перевода, резюмирования, ответов на вопросы.

Преимущества:

- высокая точность и контекстная осведомлённость;
- универсальность — подходит для многих задач NLP.

Недостатки:

- требует больших вычислительных ресурсов;
- сложность интерпретации.

---

8. Сравнение моделей

Метод	Учитывает контекст	Плотность векторов	Применимость
BoW	❌	Разреженные	Простые задачи
TF-IDF	❌	Разреженные	Классификация, поиск
Word2Vec	Частично	Плотные	Семантический анализ
Трансформеры (BERT, GPT)	☑	Плотные	Современные NLP-задачи

---

9. Применение NLP в Data Mining

- **Анализ отзывов клиентов и социальных сетей**;
- **Фильтрация спама и фишинга**;
- **Поиск и кластеризация документов**;
- **Генерация текстов, заголовков, описаний**;
- **Автоматическая обработка обращений и чатов**;
- **Анализ новостных потоков и финансовых текстов.**

---

10. Заключение

Текстовые данные — один из самых богатых источников информации. Эволюция методов NLP — от простых статистических моделей (Bag-of-Words, TF-IDF) до нейронных сетей и трансформеров — позволила компьютерам понимать и создавать язык на уровне, близком к человеческому.

Сегодня трансформерные модели (BERT, GPT и др.) стали стандартом для анализа текста, обеспечивая высокую точность и универсальность решений.

Список литературы

1. Хэн, Дж., Камбер, М., Пей, Дж. *Интеллектуальный анализ данных: концепции и методы.* — М.: Вильямс, 2019.
2. Jurafsky, D., Martin, J. H. *Speech and Language Processing.* — Pearson, 2023.
3. Mikolov, T. et al. *Efficient Estimation of Word Representations in Vector Space.* — arXiv, 2013.
4. Vaswani, A. et al. *Attention Is All You Need.* — NeurIPS, 2017.
5. Devlin, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* — NAACL, 2019.
6. Brown, T. et al. *Language Models are Few-Shot Learners (GPT-3).* — NeurIPS, 2020.