

## Автоматическая обработка текста

**26.12.2024.** Начало в 14-00. М203. Ссылка для подключения будет выслана перед началом экзамена.

Устный экзамен, 2 вопроса, 10-15 минут на подготовку.

### Темы:

1. Предобработка текста. Токенизация, лемматизация, стемминг, стоп-слова.
2. Векторное представление текста. Мешок слов, TF-IDF. Сходство текстов в векторном пространстве.
3. Эмбединги. Символьные эмбединги, эмбединги слов, предложений, документов. Контекстуализированные и неконтекстуализированные эмбединги.
4. Языковая модель: определение, оценка качества. Классические языковые модели: примеры.
5. Языковая модель: определение, оценка качества. Нейросетевые языковые модели: примеры.
6. Моделирование последовательностей: типы, примеры задач, примеры подходов.
7. Классификация текстов. Постановка задачи, примеры подходов.
8. Рекуррентные нейронные сети: RNN, GRU, LSTM. Основные принципы, примеры применения для анализа текстов.
9. Сверточные нейронные сети. Основные принципы, примеры применения для анализа текстов.
10. Архитектура «трансформер». Основные принципы, примеры применения для анализа текстов.
11. BERT и модели семейства BERT. Основные принципы. Обучение и дообучение с помощью BERT.
12. Question Answering. Типы задач, наборы данных, примеры подходов.
13. Суммаризация текстов. Экстрактивная и абстрактная суммаризация. Наборы данных, примеры подходов.
14. Симплификация текстов. Наборы данных, примеры подходов.
15. Лингвистические семантические ресурсы. Определение лингвистического ресурса, примеры ресурсов.
16. Синтаксический анализ текста. Грамматики зависимостей и составляющих. Трибанки. Основные подходы к синтаксическому парсингу. Оценка качества парсинга.
17. Генеративные языковые модели. Модели семейства GPT. Основные принципы, примеры применения для анализа текстов.
18. Промпт-инжиниринг. Zero-shot learning. Обучение на примерах. Определения, примеры применения для анализа текстов.