**Название курса:**

**Искусственный интеллект: основы интеллектуального анализа данных и обработка текстов на естественном языке (Artificial Intelligence: Data Mining basics and Natural Language Processing.**

**Лекторы:**

**Котик К.В., ведущий инженер по обработке текстов на естественном языке в «Just AI»,**

**Гребенюк А.А., д.э.н., заместитель директора по научной работе ВШССН МГУ имени М.В. Ломоносова.**

**Краткая аннотация курса:**

Индустрия, наука, образование за несколько десятков лет накопили огромные массивы данных, которые ещё недавно было практически невозможно обрабатывать одновременно. В свою очередь текущий прогресс вычислительных мощностей позволяет за достаточно короткое время проанализировать гигантские массивы информации, причём не только проанализировать, но и построить по ним предиктивные модели:

* какая будет погода?
* выдать ли человеку кредит?
* что человеку ответить на его сообщение в банковском приложении?
* посоветовать покупку в маркетплейсе,
* отсортировать поисковую выдачу по релевантности,
* сделать автомобиль автопилотируемым.

Сегодня данные – это новая «нефть». С помощью современного анализа данных создаются новые продукты, выявляются те проблемы, о которых ранее не было известно.

Цель курса – сформировать представление о Data Mining и NLP, разобрать этапы решения задачи интеллектуального анализа данных, продемонстрировать соответствующий инструментарий для построения предсказательных моделей на реальных данных.

**Программа курса:**

1. Введение: что сейчас понимается под «искусственным интеллектом» (ИИ), мотивация применения таких технологий, перспективные практические направления развития ИИ, эффект для нашего общества.

*Количество часов: 2.*

1. История ИИ. Что такое интеллектуальный анализ данных (Data Mining). Кто такой Data Scientist?

*Количество часов: 2.*

1. Работаем с Pandas – Python-библиотека для анализа и работы с таблицами.

*Количество часов: 4.*

1. Основные этапы интеллектуального анализа данных: этапы решения задачи с помощью машинного обучения.

*Количество часов: 4.*

1. NLP. Подходы к обработке текстов. Bag of Words. Embeddings. Дистрибутивная Семантика.

*Количество часов*: *4 часа.*

1. NLP. Нейронная сеть (NN). Pipeline обработки текстов с помощью NN. Подходы к обработке текстов. Основные NN-блоки для обработки текста.

*Количество часов*: *4 часа.*

1. Практика. Классификация текстов. Задача сентимент-анализа. Классификация отзывов.

*Количество часов*: *2 часа.*

Всего часов: 24.

Время проведения занятий: 17:00 – 18:30.

**Проект для зачёта.**

* 1. Собрать отзывы об **объекте** (отелях/ресторанах/тур агентствах/компаниях) в таблицу со следующими полями:
1. id отзыва (уникальный идентификатор отзыва);
2. id объекта;
3. название объекта;
4. краткое описание объекта (может быть пропуск);
5. дата отзыва;
6. текст отзыва;
7. выделенные плюсы в отзыве (может быть пропуск);
8. выделенные минусы в отзыве (может быть пропуск);
9. url-ссылка на объект;
10. тэги объекта (может быть пропуск);
11. пользовательский рейтинг (если есть рейтинг по нескольким категориям, то тоже его включить);
12. дополнительная информация (может быть пропуск);
13. дополнительные рейтинги;
14. эмоциональный окрас отзыва (позитивный, негативный или нейтральный).
	1. Количество отзывов (оно же количество строчек в таблице): не меньше **500**. Важно, чтобы среди отзывов были как позитивные, так и негативные в приблизительно в равных пропорциях.
	2. Сбор отзывов делать по объектам (выбрали объект (например, отель на TripAdvisor, и собираете по нему все отзывы).
	3. Проанализировать полученные тексты и построить модель сентимент- анализа c помощью материалов, полученных на занятии. Формат итогового результата: Jupyter notebook и таблица c отзывами.