Информационный поиск и обработка текстов на естественном языке

Чувакин С. Н.

8 сентября 2019 г.

10.09.2019

- Знакомство + о курсе + организационные моменты + учесть пожелания.
- Введение в NLP: что это такое, зачем business (and not only) application. Виды задач и их решения.

17.09.2019

- Простая векторизация текстов. BOW, n-grams, tf-idf. Лемматизация, стемминг.
- Практика на питоне. Рассмотрим gensim и sklearn.

24.09.2019

- Откуда брать данные? Готовые датасеты vs. самостоятельный сбор.
- Работа со строками в питоне. Regex, bs4, selenium.

01.10.2019

- Морфоанализ, POS tagging.
- Практика на питоне

8.10.2019

- Введение в классификация текстов. Topic modeling,sentiment analysis, простые рубрикаторы.
- Практика на питоне. LSI.
- ДЗ: LSI своими руками, без готовых решений на gensim, sklearn.

15.10.2019

- Классификация текстов - вероятностные подходы LDA, RP (Random Projections), HDP (Hierarchical Dirichlet Process). Метрики качества.
- Практика с gensim.

22.10.2019

- Sentiment analysis.
- практика на питоне.

27.10.2019 Сессия

- Задание на скачку постов с кинопоиска (или с роспотребнадзора, который реализован на AJAX). Их классификая и выделение тем.
-

05.11.2019

- Текстовая схожесть. word2vec, seq2vec, doc2vec
- Практика на питоне.

12.11.2019

- Извлечене информации из текста. NER, dependency matrix.
- Анализ готовых решений и их имплементация.

19.11.2019

- Simple-chatbots. Conversational vs. goal oriented.
- Пробуем написать своего чат бота.

26.11.2019

- Машинный перевод.
- Машинный перевод. Encoders-decoders.

03.12.2019

- Глубокое обучение.
- Обзор современных методов и решений.

10.12.2019

- Трансформеры.
- Обзор современных решений.

17.12.2019

- Подготовка к финальному проекту.
-