Обработка естественного языка

Фитц С.Ю.

4 ноября 2020 г.

4.11.2020

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

11.11.2020

- Простая векторизация текстов: BOW, n-grams, tf-idf, tokenization, lemmatization, stemming.
- Практика на python: gensim, nltk и spacy.

18.11.2020

- Откуда брать данные? Готовые датасеты vs. самостоятельный сбор.
- Работа со строками в python: regex, bs4, selenium, scrapy.
- Задание на скачку постов с сайта, который реализован на AJAX. Их классификая и выделение тем.

25.11.2020

- Введение в классификацию текстов.
- Topic modeling.
- Sentiment analysis.
- Практика на python.
- ДЗ: LSI своими руками, без готовых решений на gensim, sklearn.

2.12.2020

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

9.12.2020

- Текстовая схожесть. word2vec, seq2vec, doc2vec
- Практика на python.
- Simple-chatbots: conversational vs. goal oriented.
- Проект: Пробуем написать своего чат бота.

16.12.2020

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

23.12.2020

- Машинный перевод: encoders-decoders.
- Глубокое обучение: LSTM, BERT и GPT
- Обзор современных методов и решений.

30.12.2020

- Если что-то не успеем