НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#1

Нижний Новгород, 2019

- **Рассадин Александр ~~Георгиевич~~**
- **mail:** arassadin@hse.ru
- **github:** HSE_AMI15_NLP-2019

$$О_{итог} = 0.8 \cdot О_{накопленная} + 0.2 \cdot О_{экзамен}$$

- Задачи компьютерной лингвистики. Парсинг сайта с помощью lxml. Pandas. XML.
- Предобработка текстов: токенизация, разбиение на предложения, нормализация, стемминг, лемматизация, удаление стоп-слов.
- Поиск подстроки в строке и дистанция редактирования. Модель мешка слов. Обратный индекс, TF-IDF. Разреженные матрицы. Расстояние в метрическом пространстве.
- Классификация текстов с помощью sklearn. Анализ точности классификации: ошибки первого и второго рода, accuracy, precision, recall, F-score, ROC, AUC, Confusion Matrix.
- Марковский процесс и N-граммная языковая модель. OpenCorpora. Перплексия.
- Задача тегирования и скрытая марковская модель. Применение TreeTagger и Spacy для POS tagging.
- Тематическое моделирование: LSA, LDA, Gensim. Применение PCA (sklearn) к матрице TF-IDF для получение факторных матриц.
- Векторное представление слов. Введение в нейронные сети. word2vec.
- Neural Machine Translation.

Лекция #1.

- Задачи компьютерной лингвистики. Парсинг сайта с помощью lxml. Pandas. XML.

Natural Language Processing - “обработка естественного языка”. Это работа с текстами. Работа с устной речью не является предметом NLP.

Задачи, решаемые NLP:

- проверка правописания;
- “разбор” предложений;
- определения принадлежности текста определенному автору, тематике;
- сравнение текстов;
- суммаризация текста;
- определение тональности текста;
- перевод между языками;
- генерация текста и др.

С развитием методов и алгоритмов задачи NLP становились все более изощренными. Кроме разбора по членам предложения системы научились извлекать из текста именованные сущности. Вместо конструирования summary из готовых предложений исходного текста - синтезировать собственные предложения, компактно представляющие текст. Тексты стало можно автоматически классифицировать по тематике, определять эмоциональную направленность текста. Наконец, появились алгоритмы синтеза текста в ответ на поступающий внешний текст так, чтобы поддерживать некоторый диалог. Чат-боты, которые ранее строились на основе правил, также стали использовать NLP на основе машинного обучения.

- Цели NLP - извлечение из текста полезной, целевой информации для записи или визуализации, а также построение нового текста, например ответа на вопрос.
- Основной проблемой в NLP является NLU (Natural Language Understanding) - понимание текста на естественном языке. Мы будем говорить об NLU, как о комплексе методов и алгоритмов для трансформирования текстов в активные действия, соответствующие семантике текстов в соответствии с задачей реакции на поток предложений, сформулированной заранее. Активные действия могут быть сами процессом генерирования некоторого ответного текста или процессами управления некоторыми ресурсами или просто командами создания некоторой новой структуры данных, соответствующей тексту.

Formal vs Natural

```
SELECT name, address
FROM businesses
WHERE business_type = 'pub'
AND postcode = '50121'
```

vs

Where is the nearest pub?

Домашняя работа #1.

Программа автоматического сбора новостного корпуса

Задача.

1. Собрать не менее 200 текстов в каждой из 4х или более новостных категорий
2. Разбить тексты на предложения, сохранить в `XML` в виде индекса слов

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ