НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#1, #2

Нижний Новгород, 2018

- **Рассадин Александр ~~Георгиевич~~**
- **mail:** arassadin@hse.ru
- **github:** HSE_AMI14_NLP-2018

Составляющие зачета:

1. Минимум 3 посещения
2. Все КР
3. Все ДР

EZ зачет:

1. ~~почти~~ Всегда ходить
2. Все сдавать вовремя

- КР можно закончить дома
- 1 неделя на ДР

- GIt. Python. Pandas
- XML. JSON. Парсинг новостного сайта
- Предобработка корпуса. Лемматизация. Дистанция редактирования
- Обратный индекс
- TF-IDF. Разреженные матрицы
- Косинусное расстояние. BIRM
- Анализ моделей классификации. Ошибки 1го, 2го рода, точность, полнота, F-score… (sklearn)
- Марковский процесс. N-граммы. Правдоподобие и перплексия. OpenCorpora
- Тегирование и POS. HMM, TreeTagger, Spacy
- Тематическое моделирование (x2)
- Векторное представление слов, skip-gram, w2vec. Введение в нейронные сети (x2)

Практика #1.

Git. Python. Pandas.

Практика #2.

XML. JSON. Парсинг новостного сайта.

Контрольная работа #1.

Программа автоматического сбора новостного корпуса

Задача.

1. Используя `lxml` распарсить один из новостных сайтов:
    - RT
    - BBC
    - РИА
    - Рамблер
    - ТАСС
    - …
2. Собрать не менее 50 текстов в каждой из 4х или более новостных категорий

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