НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#2

Нижний Новгород, 2019

Лабораторная работа #1.

Программа автоматического сбора новостного корпуса

Задача.

1. Собрать не менее 200 текстов в каждой из 4х или более новостных категорий с одного из новостных сайтов:
   - ria.ru
   - newsru.com
   - nn.ru
   - life.ru
   - bbc.com
   - lenta.ru
   - tass.ru
   - …
2. Сохранить корпус в формате `XML`

Лекция #2.

- Предобработка текстов: токенизация, разбиение на предложения, нормализация, стемминг, лемматизация, удаление стоп-слов.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ