НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра прикладной математики и информатики

Компьютерная лингвистика. Практика.

#3

Нижний Новгород, 2019

Лабораторная работа #2.

Предобработка собранного корпуса

Задача.

1. Предобработать собранный новостной корпус: выполнить очистку от несловарных токенов, выполнить стемминг и лематизацию.
2. Собрать словарь.
3. Сохранить тексты, заменив слова на соответствующие индексы в словаре.

Лекция #3.

- Поиск подстроки в строке и дистанция редактирования
- Модель мешка слов
- Обратный индекс, TF-IDF
- Разреженные матрицы
- Расстояние в метрическом пространстве

Лабораторная работа #3.

Подсчет TF-IDF для своего корпуса

Задача.

1. “Руками” вычислить TF-IDF для своего корпуса

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ