

## Задание к лабораторной работе №8

### Алгоритмы обработки текстовых данных (15-16 недели)

Задание 1: Построить облако тегов по тексту.

Требования:

- Для построения тегов использовать только имена существительные в именительном падеже (если не получится – любые слова, но в своей начальной форме)
- Длина слов, подлежащих анализу – не менее 3 символов(можно и 4)
- Указывать статистику вхождения слова в текст (и всех его форм)

Задание 2: Реализовать алгоритм вычисления разницы между двумя текстами (антиплагиат)

Требования:

- Для сравнения используется 10 исходных текстов размером от 500 до 1000 слов.
- Показать разницу между текстами. Показать общее число совпадений, процентное соотношение совпадений и все оформить в виде инфографики. (Опциональное пожелание – вывести наиболее подходящий текст и все совпадения подсветить зеленым цветом)
- Время для анализа по всем текстам – не более 10 минут. (Можем опять соревнование на самую быструю реализацию провести)

Общие требования:

- Приложение должно иметь графический интерфейс
- Файл для анализа обязательно должен свободно выбираться.
- Используемые языки программирования: C#, C++, Delphi.

Немного теории:

Полезные ссылки:

<http://habrahabr.ru/post/114997/>

[http://ru.wikipedia.org/wiki/Алгоритм\\_Рабина\\_—\\_Карпа](http://ru.wikipedia.org/wiki/Алгоритм_Рабина_—_Карпа)

[http://wiki.asoiu.com/index.php/Поиск\\_плагиата\\_методом\\_шинглов](http://wiki.asoiu.com/index.php/Поиск_плагиата_методом_шинглов)

[http://e-maxx.ru/algo/duval\\_algorithm](http://e-maxx.ru/algo/duval_algorithm)