

Задачи на mapreduce.

<https://github.com/michaelfairley/mincemeatpy> -- python mapreduce framework

Для разделения текста на слова рекомендуется использовать пакет `nltk.tokenize`

Задача 1.

Посчитать богатство языка (количество слов, используемых каждым персонажем). Слова очистить от знаков пунктуации, привести в нижний регистр.

Данные: southpark

Формат результата: csv таблица: персонаж, богатство языка.

Задача 2.

Построить обратный индекс с частотностью слов. Слова очистить от знаков пунктуации, привести в нижний регистр.

Данные: sherlock

Формат результата:

	doc1	doc2	...
term1	n1	n2	..
term2	n3	n4	..
...

Задача 3.

Произведение двух квадратных матриц, все операции делать по модулю 97.

Данные: файл с матрицей в csv формате (генерируете сами).

Формат результата: файл в csv формате

Формат csv файлов:

матрица	строка	столбец	значение
a	0	0	2

a	0	1	1
b	0	2	3
...