Your Map-Reduce

Воспользуемся моделью распределённых вычислений Map-Reduce для задачи подсчета количества слов в тексте(Word Count). Входные данные – текстовый файл greeneggsandham.txt, выходной файл – out.txt, который состоит из строк “слово – количество вхождений в файл”. Предположения – не будем учитывать регистр, слова состоят из английских букв, возможен символ ‘ - ‘ в словах.

Составим 4 функции:

1. map – шаг Map, составление записей вида: (слово, 1);
2. partition – промежуточный шаг, группировка по словам, итог: словарь вида (слово, [1,…,1]);
3. reduce – шаг Reduce, сортировка по ключам, суммирование количества появлений слов;
4. write_result – запись в файл

Использованный материал – https://habrahabr.ru/post/103467/

Рассмотрим решение по шагам:

1) **map:** считываем построчно файл, выделяем слова с помощью библиотеки re, создаем список из tuple’ов, возвращаем этот список;
2) **partition:** получаем список, заводим словарь(ключ – слово, значение – список из вхождений), заполняем словарь, возвращаем этот словарь;
3) **reduce:** получаем словарь частот, создаем выходной список, итерируемся по отсортированным ключам и формируем список, подсчитывая количество вхождений, возвращаем список из слов и их вхождений в исходный файл;
4) **write_result:** открываем файл для записи, сохраняем строки из слов и их вхождений в исходный файл