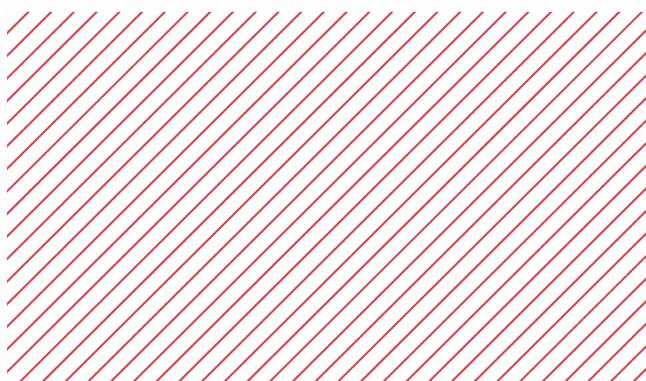


академия
больших
данных



HW02: MapReduce



Описание работы и критерии оценивания

В данной задаче мы будем подсчитывать среднее значение (аналог `pumpry.mean`) и дисперсию (аналог `pumpry.var`) для сета из N кусков данных с помощью map-reduce парадигмы. Маппер функция будет применяться нами к кортежам вида (ck, mk, vk) , где ck - размер `chunk_size`, mk -среднее данного `chunk` и vk -его дисперсия. Редюсер функция должна скомбинировать результаты среднего значения и дисперсии величины:

$$m_i = \frac{c_j m_j + c_k m_k}{c_j + c_k},$$
$$v_i = \frac{c_j v_j + c_k v_k}{c_j + c_k} + c_j c_k \left(\frac{m_j - m_k}{c_j + c_k} \right)^2$$

Максимально возможное количество баллов за работу: 130 баллов (30 дополнительных баллов начисляется за исполнение кода программы на языке Java):

За правильное выполнение map-reduce части для подсчета среднего значения начисляется 50 баллов и также 50 баллов можно получить за map-reduce подсчета дисперсии указанной величины.

Бонусы и штрафы:

- **100%** за плагиат в решениях (всем участникам процесса)
- **30%** за посылку решения в течение недели после deadline

Этапы задания

1. Загрузите датасет по ценам на жилье Airbnb, доступный на kaggle.com:
<https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>
2. Подсчитайте среднее значение и дисперсию по признаку "price" стандартными способами ("чистый код" или использование библиотек). Не учитывайте пропущенные значения при подсчете статистик.
3. Используя Python или Java, реализуйте скрипты mapper.py и reducer.py для расчета каждой из двух величин. В итоге у вас должно получиться 4 скрипта: 2 mapper и 2 reducer для каждой величины.
4. Проверьте правильность подсчета статистик методом map-reduce в сравнении со стандартным подходом
5. Результаты сравнения (то есть, подсчета двумя разными способами) для среднего значения и дисперсии запишите в файл .txt. В итоге, у вас должно получиться две пары значений (стандартного расчета и map-reduce)- одна пара для среднего, другая - для дисперсии.
6. Итоговый файл архива с выполненным заданием должен включать в себя сам код, а также результаты его работы. Архив присылать на адрес muzalevsky.ds@gmail.com