Управление разно-структурированными большими данными

к.т.н. Брюхов Д.О. (dbriukhov@ipiran.ru)

Hadoop Map Reduce (Домашнее задание 1)

Общие требования

- Задание выполняется в виде MapReduce приложения на Java или Python (допускается использование и других языков программирования).
- Срок выполнения задания – **22 ноября**.
- Выполненное задание должно быть прислано на почту в виде архива (например, содержащего проект Eclipse).
- Имя должно быть названо: <Family_Name>_MapReduce_Var<#>.rar (zip, gz, и.т.д.)
    - <Family_Name> - Ваша фамилия
    - <#> - номер варианта
    - Например: Briukhov_MapReduce_Var4.rar

Платформа Hadoop

Выполнение задания осуществляется на любой платформе Hadoop (например, IBM BigInsights, HortonWorks).

Если домашний компьютер позволяет запускать виртуальные машины дома, можно скачать виртуальную машину и работать с ней.

Если такой возможности нет, то можно запускать программы локально с помощью библиотеки MRJob

Использование виртуальной машины

Скачать виртуальную машину

Например, https://www.cloudera.com/downloads/hortonworks-sandbox/hdp.html

Установить Python

```
yum install python-pip
```

Установить MRJob

```
pip install mrjob
```

Варианты заданий

1. Реализовать алгоритм TF*IDF (на четверку)

Входными данными являются: директория, содержащая несколько текстовых файлов, и поисковая строка, состоящая из нескольких слов. Результатом является упорядоченный (по среднему tf*idf) список файлов.

Формат представления входных и выходных данных выбирается Вами.

2. Реализовать алгоритм поиска кратчайшего пути в графе

Формат представления входных и выходных данных выбирается Вами.

Требования:

- Ребра графа должны содержать положительные веса.
- (Дополнительно) Рассмотреть вариант с отрицательными весами

3. Реализовать алгоритм PageRank

Формат представления входных и выходных данных выбирается Вами.

4. Реализовать метод кластеризации K-means

Формат представления входных и выходных данных выбирается Вами.

Требования:

- Не использовать Евклидову метрику в качестве метрики расстояния между точками.
- Использовать не меньше 3 измерений
- (Дополнительно) Использовать другие данные (не координаты точек), например, изображения

5. Реализовать произвольный алгоритм

Формат представления входных и выходных данных выбирается Вами.

Требования:

- Алгоритм должен содержать по крайней мере 2 шага MapReduce
- Алгоритм сперва нужно согласовать со мной