coreservlets.com – Hadoop Course

First MapReduce Job

In this exercise, you will have a chance to implement MapReduce jobs. These jobs will get you acquainted with the framework.

Approx. Time: 60 minutes

Perform

1. Develop a MapReduce job that will count up each unique token:
    a. Persist a file with tab-separated results, one token and corresponding occurrence count per line:

    ```
    Airline   20
    Airport   7
    ...
    ```

    b. Perform tokenization using Java's `StringTokenizer` (just like lecture's examples)
    c. Use `/training/data/war_and_peace.txt` as input to your job; the file already exists in HDFS
2. Develop a MapReduce job that given a text file input will produce the two counts: (1) Number of tokens whose character length is greater than or equal to five characters (2) Number of tokens whose character length is less than five characters
    a. Persist results to a file that should look something like this:

    ```
    greaterOrEqualsToFiveChars    236865
    lessThanFiveChars   329372
    ```

    b. Perform tokenization using Java's `StringTokenizer` (just like lecture examples)
    c. Use `/training/data/war_and_peace.txt` as input to your job; the file already exists in HDFS

Solution

1. The code can be found in the Solutions project:

   ```
   mapRed.firstJob.WordCountMapper.java
   mapRed.firstJob.WordCountReducer.java
   mapRed.firstJob.WordCountTool.java
   ```

   To run the code

   ```
   $ yarn jar $PLAY_AREA/Solutions.jar \
       mapRed.firstJob.WordCountTool \
       /training/data/war_and_peace.txt \
       /training/exercises/mapRed/firstJob/ex1
   ```

2. The code can be found in the Solutions project:

   ```
   mapRed.firstJob.LengthDividerCountMapper.java
   mapRed.firstJob.LengthDividerCountTool.java
   mapRed.firstJob.WordCountReducer.java
   ```

   To run the code

   ```
   $ yarn jar $PLAY_AREA/Solutions.jar \
       mapRed.firstJob.LengthDividerCountTool \
       /training/data/war_and_peace.txt \
       /training/exercises/mapRed/firstJob/ex2
   ```