CS425/591 Individual Project 3: Motif Search

Problem: Implement the median string Motif search algorithm (on slide #34, posted at mycourses.siu.edu) MapReduce on the Hadoop platform you already built for Project 2 on amazon AWS; and through test your program using the input sequence data file, named "promoters_data_clean.txt", provides.

Hints: Supposedly, each running instance of your *Map* function works on one sequence (one line) of the input dataset. Your *Map* function loops over all the enumerated (candidate) median string of length L (as an input parameter), steps through the sequence and finds the best match (i.e., with minimum matching distance) with each candidate median string; the output key/value pairs of the map functions are pairs of a median string and the best match (corresponding to **minimum matching distance**) in the sequence. Your *Reduce* function shall loops over all the median strings and sum up the total matching distances with each median word. The median string with the minimum total matching distance is the desired median string, and the corresponding matches in all input sequences are the motifs sought.

Output: The output of your program must include the following items (each in a separate column): the found median string (of length 8), the motif (i.e., the found best matching subsequence in each input sequence), the motif's position index in the input sequence, the local matching distance (between the median string and the motif found in one sequence), the total matching distance, and the sequence's id (i.e., its sequential order in the input file) if your output does not follow the same order of input sequences in the input dataset.

Submission Requirement:

(1) A brief report with a description of your program structure and discussion/comment (short or long) on any insight you have gained from doing this project.

(2) The required output described above (as a separate file).

(3) The source code of you program (only the code that you wrote) with necessary comments.

(4) Two screenshots taken before and after you run your program.

Please submit all required files to mycourses.siu.edu.