James Prillaman
CS 491
Dr. Che
Map Reduce Project

Map Reduce

Program Structure:

Class App:

This contains my main function that runs the program. It takes two arguments of input DNA file and output directory. It sets the configurations for the hadoop mapreduce job.

Class DNAGenerator:

Contains a LinkedList of Strings that contains every possible motif that can be made from given string size.

Method generate():
-generates every possible string combination of a motif of given string length and stores it in the LinkedList allDna.
-This uses an iterative approach by increment the index corresponding to the available letters for each character in the motif.

Class MotifMapper:

-Uses DNAGenerator to generate each possible motif combination.
-For each motif combination it compares each character of the motif with the sub string of equal length for every possible location of the sub string by incrementing the starting position of the given input dna sequence.
-Each motif is mapped to a key-value map of where key is the motif string and value is the distance between motif and the best sub string match found(determined by lowest distance)

-This will map each motif to the distance of its best match for each substring.

Class MotifReducer:

contains:
-linked list of strings for each motif
-linked list of int for each total distance

Method reduce():
-Goes through the map created my mapper and collects all distances for each single motif.
-Computes the total distance by adding each sequence's best distance for single motif.
-Stores the motif in a linked list and totaldistance in linked list at matching index.

Method cleanup()
-Is run once after all reducing is finished.
-has linked list of strings for motifs with lowest total distance
-Starts with lowest distance as first motif with corresponding first total distance in each respective linked list at index 0;
-increments and if total distance is found with small total distance than current best, list of best motifs is emptied and new best motif is added to empty list. Best distance set to current distance
-After best motifs are found, each motif in the linked list tempMotifs is printed to output file.
-These printed motifs are the equally best found motif for all given dna sequences.

Learned:

I learned the structure of how mapreduce works. I also applied it to a realistic problem and got hands on experience at implementing it using hadoop. I can use this structure on future programs that can be fit into the map reduce structure. This structure can drastically cut down on processing time and thus be very useful and beneficial to any problem set it can be assigned to.

I also gained incite and experience using clustering techniques with hadoop. This is the first time I have build software that takes advantage of this technology. I learned useful skills in debugging and configuring virtual cluster software.

I also discovered a iterative approach to creating every combination of a string for a given size by tracking the index in a library of each character in the string. Traditionally and in this lab, I originally tried to solve the problem using recursion. Given the large data set being created, I ran into a stack overflow. I can use this iterative approach on future problems dealing with combination when recursion is not an option.