FINAL PROJECT
COMPUTATIONAL METHODS FOR DATA SCIENCE
FALL SEMESTER 2022

The main goal of this final project is to learn new algorithms on top of those learnt from the lecture, and to gain the ability to judge and compare these new algorithms to the existing ones. Here is a list of components that will be required in the final project.

1. **A Revision on Traveler Salesman Problem.** *(10 points)* The first part of this project needs you to review the famous Traveler Salesman Problem (TSP). It is a classic optimization problem that many metaheuristic algorithms try to implement. In this section, please write down the following:

   (a) What is TSP is?

   (b) What are the difficulties in TSP?

   (c) What are the applications of TSP in real life?

2. **Data Collection.** *(20 points)* It is well-known that 7-Eleven is the most abundant convenient store in Taiwan. The logistic among all stores is always very complicated. Now we only focus on 47 7-Eleven stores with organic vegetables in 5 districts of Taipei City, and look for the optimal path through these 31 stores, so that the delivery van can ship all vegetables in the shortest path. The 5 districts are: Xinyi (6 stores), Da'an (7 stores), Zhongshan (11 stores), Zhongzheng (4 stores), and Wanhua (3 stores). We assume Node 00 is both the starting and ending nodes of the path. Find the distance matrix by the following steps below.

   (a) Go to Google Map and measure the pairwise store distances, and form a distance matrix.

   (b) Consider a complete graph with each node as a convenient store, and the edge between two nodes contains the distance information. Then the question becomes how to draw a line to pass through all nodes on this graph, so that the distance is the shortest.

   In this part, you need to return the 31×31 distance matrix with zero diagonals and all values in the off-diagonal entries are the pairwise distances between two stores.

3. **A Start Initialization by Rejection Sampling.** *(20 points)* In addition to the efficiency of an optimization method, a smart starting location is always a key to the success. In this part, we use the rejection sampling to help generating some good-quality initial particles for the optimization method.

   (a) First, we need to know the size of our domain. Assume a computer can enumerate at most 0.5 billion sequences of random order. How many distinct nodes it can actually handle (not including the starting and ending nodes) in the path?

   (b) Write a simple program to generate a sequence of random order from 01 to 30.

   (c) Generate 1000 sequences and calculate their distances (NOTE: start from Node 00, through the path, and back to Node 00). Report their average distances and set it as the threshold.

   (d) Initialization Step. Generate 1000 initial sequences with "good" distances by the spirit of rejection sampling, i.e. if the generated sequence has shorter distance than the threshold, accept it, otherwise, do something to decide whether to accept it.

4. **An Implementation of Known Metaheuristic Methods.** *(40 points)* We learn many metaheuristic algorithms that can be implemented for TSP. In this section, please implement the following algorithms in TSP.

   (a) Genetic Algorithm.

   (b) Simulated Annealing.

   NOTE that the initialization step has been done in Question 3(d). You need to state the following results in this part:

   (a) State all necessary information, including the particle definition, objective function, goal, constraints, etc. The number of iteration is set at 1000.

   (b) Write a pseudo-code on two algorithms in text.

   (c) Submit two workable functions that input the distance matrix and all necessary parameters, and output the optimal sequence with its distance.

   (d) Write down in text the suggested paths (optimal sequences) by two algorithms, together with their distances and the computing time.

5. **Summary and Discussion.** *(10 points)* In the last part of this final project, please provide the following information:

   (a) Summarize your comparison results on two methods.

   (b) State the advantages and disadvantages of two methods shown in this application to TSP.

   (c) State at least one potential improvement on the best method to make the algorithm even better on this TSP application.

   (d) State the potential improvement of the initialization method, either an improvement from the current rejection sampling, or another method that replaces the rejection sampling.

6. **Low-Rank Approximation on Distance Matrix.** *(Bonus 20 points)* In the first part of the course, we learn many matrix manipulation techniques that may help to reduce the matrix computations. Although there exists low-rank sparse decomposition methods for adjacency matrix (i.e distance matrix in our case) of a graph, comment on why these methods cannot help in our problem? Or if you think it can help reducing the rank (i.e. low-rank approximation), demonstrate how it works by using the 7×7 sub-matrix (i.e. stores in Zhongzheng and Wanhua districts).