ARTIFICIAL INTELLIGENCE SEARCH ASSIGNMENT

This is the assignment for the sub-module *Artificial Intelligence Search* of the module *Software Methodologies*. It is to be completed and handed in **via DUO** by **2 p.m.** on **Friday 15th December 2017**.

Basic instructions

You are to implement ***two different algorithms***, using techniques studied during the lectures, but possibly also additional algorithms you have devised or discovered for yourself, to solve the ***Travelling Salesman Problem (TSP)***. Your implementations should seek to obtain the best Travelling Salesman tour that you can, given a collection of cities and their distances.

The intrinsic difficulty of the underlying algorithms that you choose will impact upon the marks awarded. The obvious contenders for algorithms to implement from the course are:

- a brute-force search
  as you are doubtless aware, a brute-force search for the TSP will only work for very, very small sets of cities
- a basic greedy algorithm
  the most obvious one is 'build the tour by moving to the nearest city from wherever you happen to be', i.e., 'nearest neighbour'
- a best-first search without heuristic data
  depending on how you choose your evaluation function, you can implement a breadth-first search, a depth-first search, and various other algorithms
- a **greedy best-first search with heuristic information**
  you'll have to build your own heuristic functions, though, as none are supplied
- **A* search**
  again, you'll have to build your own heuristic functions; also, note that an A* search with a 'good' heuristic is optimal and also that the TSP is **NP**-hard
- a hill-climbing search
  this is very easy to implement once you decide upon the appropriate coding of the TSP as a search problem

- a **simulated annealing** algorithm
  there is much scope for experimentation

- a **genetic algorithm**
  there are many ways in which you can solve the TSP using a genetic algorithms and there is much scope for experimentation.

The algorithms written in **bold** are algorithms that I regard as 'more involved' and you should aim to implement at least one of these 'more involved' algorithms. ***More credit will be given for implementations of 'more involved' algorithms*** (see the mark scheme below); besides, 'more involved' algorithms will probably give better results and so you will pick up more marks for the quality of the tours found too (again, see the mark scheme below). Apart from the algorithms above, there are many other algorithms that you might implement but which are not covered in this course. If you wish to implement one of these algorithms then this is absolutely fine and I encourage you to explore. If you are unsure as to whether an algorithm that you wish to implement is 'more involved' or not then please contact me. Other algorithms implemented in the past have included *Ant Colony Optimization* and *Beam Search*.

You are strongly advised to develop your own implementations using the pseudo-code given in lectures rather than search the web for related implementations.

You should ensure that you have implemented your algorithms *well before* the end of Michaelmas Term so as to give yourself time to experiment on the Travelling Salesman instances upon which your implementations are to be evaluated. The time allocated for experimental work will give you ample time to fine-tune your implementations, in the light of experimental experience, so as to obtain the best results you can. The tours provided by your implementations to these instances should be submitted, along with a report of your implementations and your experimentation (what is expected is described in more detail below).

The data-files

The actual instances of the Travelling Salesman problem (more precisely, the symmetric Travelling Salesman problem, where the distance from city x to city y, denoted (x, y), is always the same as the distance from city y to city x, that is, (y, x)) will be given in the following form (the cities are always named $1, 2, \ldots, n$).

`NAME = <string` = name-of-the-data-file>,
`SIZE = <integer` n = the number of cities in the instance>,

<list-of-integers $d_1, d_2, d_3, \ldots, d_m$> where the list consists of
the distances between cities $(1, 2), (1, 3), \ldots, (1, n)$,
then the distances between cities $(2, 3), (2, 4), \ldots, (2, n)$,
$\ldots$
and finally the distance between the cities $(n - 1, n)$.

Commas ‘,’ are used as delimitters in data-files; carriage returns, end-of-line markers, spaces, etc., should be ignored.

So, for example, the instance with 5 cities where: the city 1 is at the origin; the city 2 is 3 miles north; the city 3 is 4 miles east; the city 4 is 3 miles south; and the city 5 is 4 miles west, and all distances are the Euclidean distances between cities, is encoded as the city-file `AISearchsample.txt`:

```
NAME = AISearchsample,
SIZE = 5,
3, 4, 3, 4,
5, 6, 5,
5, 8,
5
```

With reference to the remark above, re: delimitters, the above city-file could well be presented with no carriage returns, etc., as simply

```
NAME = AISearchsample,SIZE = 5,3,4,3,4,5,6,5,5,8,5
```

Note that in general a given instance need not be based on the Euclidean distances of a collection of cities on the plane. You should assume that a given distance between two cities is a non-negative integer (and so it could well be the case that the distance between two distinct cities is 0).

First task: Your first task is to be able to read in the data from a city-file and store it internally for use by your implementation. I strongly recommend that you write some code to read the data from a city-file into a two-dimensional list whose (i, j)th entry stores the distance from city i to city j. So, with the city-file `AISearchsample.txt`, above, if the list is called *city* then $city[2][3]$ has the value 5 as does $city[3][2]$. Also, there are carriage returns and other control characters in the city-files. It is your job to ignore these characters. The reason these characters are there is because this is how data is presented in the real world: as ‘dirty data’. I would advise reading the city-files character by character and stripping out characters that are not as you would expect from a well-formatted city-file. I have supplied a sample city-file, `AISearchtestcase.txt` (that is dirty), for you to experiment on.

The data-files that you will have to execute your code on will be called

- `AISearchfile012.txt`
- `AISearchfile017.txt`
- `AISearchfile021.txt`
- `AISearchfile026.txt`
- `AISearchfile042.txt`
- `AISearchfile048.txt`
- `AISearchfile058.txt`
- `AISearchfile175.txt`
- `AISearchfile180.txt`
- `AISearchfile535.txt`.

The numeric digits denote the number of cities in the particular instance.

Your tour-files

You are expected to derive the best tours that you can with your implementations and to provide the actual tour in each case. ***What follows next is extremely important!*** Your tour-file should be named by prefixing the name of the city-file with the prefix '`tour`' and should have the following form:

> `NAME = <string =` name-of-TSP-instance-file>,
> `TOURSIZE = <integer` n `=` integer-giving-the-number-of
> -cities-in-the-tour>,
> `LENGTH = <integer` len `=` integer-giving-length-of-tour>,
> `<list-of-integers` $x_1, x_2, x_3, \ldots, x_n$`>` where $x_1, x_2, x_3, \ldots, x_n$
> is a list of the cities in the order of the tour (with the final hop
> being from city x_n back to city x_1).

So, a tour-file for the city-file `AISearchfile012.txt` will be the file named `tourAISearchfile012.txt` and possibly of the form:

```
NAME = AISearchfile012,
TOURSIZE = 12,
LENGTH = 93,
1,3,4,6,8,5,2,9,10,11,7,12
```

I have supplied a sample tour-file, `tourAISearchtestcase.txt` (for the sample city-file `AISearchtestcase.txt`).

I also supply a small Python program for you to verify that your tour-files are both correctly formatted and such that the tour you present does indeed have the length that you claim.

USE THIS PROGRAM TO VERIFY YOUR TOUR-FILES!

Last year, 13 students did not bother to do this and their incorrect tour-files resulted in them scoring zero marks for their tours.

Material to be submitted

You are expected to hand in a folder named with your username, e.g., `dcs0ias`, and within which:

- there are 2 folders named

  `TourfileA` and `TourfileB`.

  Each folder should contain 10 tour-files, named

  `tourAISearchfile012.txt`, `tourAISearchfile017.txt`, etc.,

  detailing the best tours that particular implementation has found in the respective collection of cities (Tourfile A contains the tours from your first algorithm and Tourfile B the tours from your second).

- there is a pdf file, entitled `dcs0iasreport.pdf`, in the folder named `dcs0ias` including a detailed description of your implementations (each marked as A or B, as appropriate), tabulations of the results produced, an explanation of your experiments, and critical comments (the report, about which I say more when I discuss the mark scheme, should be no more than 4 pages of A4, in a single column format, and no smaller than 11 pt. font; there is no need for cover and contents pages, etc.)

- there is another folder, called `dcs0iasrest`, in the folder `dcs0ias` containing your implementations and anything else of relevance.

Ensure that your material is handed in as above with files and folder so named (you can zip everything together into one file if you like).

Why are only 4 pages allowed? Paraphrasing your work is an important skill, especially in industry. More often than not, you have a short period of time or a short document in order to get your ideas across as best you can. What is important is that you use the 4 pages to include salient and important data. The whole point of restricting the length is to give you practice at focussing your ideas and presenting yourself succinctly and cogently.

Mark scheme

Credit will be given for good working implementations and good experimental results as demonstrated in your detailed account, structured as follows. There are 20 marks available in total, apportioned as follows (though your final mark will be returned as a percentage).

- Full and clear descriptions of ***your implementations***, focussing on the different implementation issues arising (do not include your code in your report). Focus on specific implementation details such as choice of data structures or data representation, use and implementation of probabilistic methods, technical aspects of specific algorithms such as the implementation of crossover and mutation in a genetic algorithm, etc. (do not include your code in your report). There are [3 marks] available for a 'more involved' algorithm and [1.5 marks] available for an easier algorithm. So, there are potentially [6 marks] available based on the descriptions of aspects of your implementations. ***Do not just copy the pseudo-code of your algorithms from the slides. I am interested in implementation details, not algorithmic details.***

- A thorough (tabulated) description of your results so that you specify the lengths of the best tours obtained (of course, these lengths are witnessed by the tour-files that you have submitted). The better the tours you find, the better the marks. There are [6 marks] available as regards the quality of your tours. Note that although I want to see the lengths of the tours found for both your implementations, I'll award the marks on the basis of the overall best tours found (so, if you had one algorithm that gave really good tours and one that didn't, but might be interesting in other ways, then you'll score on the basis of your good tours).

- Details of your experiences with your implementations and the fine-tuning and experimentation that you undertook in order to try and improve performance. There are [8 marks] available as regards experimentation. The fact that there are more marks available for experimentation than anything else should tell you that I expect you

> not only to implement two basic algorithm but also to experiment and investigate varying different parameters within your implementations. You should undergo experimentation systematically and give a clear account of the things that worked and the things that didn't. Be adventurous and try different things.

You will be given marks only for the content of the report. Do not expect me to look in other files or folders for missing details. These other folders are supplied just for me to clarify your results if I choose to do so.

It is absolutely crucial that you conform to the above format. As I explained above, I automatically check that your tours are indeed of the lengths you state and if I can't do this because you have not formatted files properly then you will score no marks for tour quality.

In general, I am looking for a *good understanding* of the algorithms you have implemented and *some ingenuity* in manipulating these implementations so as to try and get better results. I will reward ingenuity even if the method you have chosen to implement does not in general give good results. However, I will also reward students who get good results. If you choose to implement a basic greedy algorithm, for example, then you may not score so well, because this algorithm is easy to implement. However, you might gain extra credit if you were to try and improve your basic greedy algorithm by some ingenuity. The choice is yours as to the algorithms you implement. As regards your implementation descriptions, I am looking for well written, concise, and informative presentations, and as regards running your implementations, I am looking for 'innovative tinkering' to try and get improvements.

As an example scenario, suppose that you are a student who does not enjoy the content of this course or who is short of time (for whatever reason!) and who has decided to implement two algorithms that are not 'more involved'. If the implementations are reasonably well written up then you might score 2/3 marks (the most you could have scored would have been 3 as you didn't implement a 'more involved' algorithm). Perhaps you fiddled around a bit with your implementations and did manage to improve them with a little ingenuity; so, maybe you picked up 3.5/6 marks for experimentation. Finally, because you originally chose basic algorithms, perhaps your original tours were not that good and would have only scored 3/6 marks. However, your experimentation did improve things so that you scored 3.5/6 marks for tour quality. This would give you a total of 9/20 = 45%. This is an absolute bare minimum if you wish to pass this coursework. I am expecting more!

Some remarks and hints

As mentioned earlier, the following methods are available to you (as studied in the course):

- brute-force search
- basic greedy algorithm ('nearest-neighbour')
- best-first search without heuristic data
- greedy best-first search
- A* search
- hill-climbing search
- simulated annealing
- genetic algorithm.

There is a little bit of work to do as regards the implementation of each method and here is a little bit of help.

Brute-force search

Here is a hint as to how all tours in an n-city Travelling Salesman instance can be generated. Each tour is stored in a 1-dimensional list T of size n, where the elements are all distinct and come from $\{1, 2, \ldots, n\}$. The tour stored in T is $T[1], T[2], \ldots, T[n], T[1]$. In fact, we can similarly store a tour of the $m \leq n$ cities $\{1, 2, \ldots, m\}$ in $T[1], T[2], \ldots, T[m]$, with $T[m+1] = T[m+2] = \ldots = T[n] = 0$.

Our procedure `gen(`T`,`m`)` takes as input a tour of m cities, $x_1, x_2, \ldots, x_m, x_1$, say, held in the list T (as above), and proceeds as follows.

- If $m = n$ then we compare the length of the tour T (of n cities) with the length of the shortest tour found so far and if T is shorter then we remember T and its length.
- If $m < n$ then `gen(`T`,`m`)` generates all tours of the cities $\{1, 2, \ldots, m, m+1\}$ by inserting the value $m+1$ in location 1, then location 2, …, then location m, then location $m+1$ of T (so that the cities coming after $m+1$ are 'shifted' along the array). Interleaved with generating each tour, which we refer to as T', we recursively call the procedure `gen(`T'`,`$m+1$`)`.

In more detail, `gen(`T`,`m`)` is as follows.

```
if $m$ == $n$ then
   calculate the length of the tour $T$ and if it is shorter
   than the best tour found so far, remember $T$ and its
   length as the best tour found so far
else
   for $i$ = 1 to $m+1$ do
      $T'$ = $T$ with the city $m+1$ inserted into location $i$
      call gen($T'$,$m+1$)
      $T'$ = $T$ with the city $m+1$ removed from location $i$
fi
```

Thus, the following pseudo-code generates and tests all possible tours, given the distance-file of n cities.

```
$T$ = $[1, 0, 0, \ldots, 0]$      %initialize tour $T$ as [1]
$m$ = 1                   %initialize number of cities of tour $T$ as 1
call gen($T$,$m$)
output the shortest tour found
```

Although I don't recommend that you implement a brute-force search, brute-force searches are good for checking optimal values in small cases; also generating all combinatorial possibilities comes up regularly and it is useful to know how to do this. If you do implement a brute-force search then you can always choose to kill an execution and take the best tour you have found up until that point as a guide.

Best-first, greedy best-first and A* search

The Travelling Salesman Problem needs to be realised as a search problem. One way of doing this is to have the set of all lists of distinct cities as the states together with the lists of n distinct cities augmented with the start city (and so a state is a list of between 0 and n cities, or a list of $n+1$ cities where the first n are distinct and the last city equals the first). There is one action with a state t' being a successor of a state t if the list t' is the partial tour t extended with one new city (not appearing in t), or if t has length n and t' is t augmented with the first city of t. The step-cost associated with any transition from state t to state t' is the cost of moving from the final city in the list t to the final city of the list t'. The initial state is the list t_0 consisting of just the start city and a goal state is a list of $n+1$ cities. An optimal solution is thus a path from the initial state to a goal state of minimal cost.

There are a number of heuristic functions available for the Travelling Salesman Problem. One of these is the heuristic h where, given a state $t = x_1, x_2, \ldots, x_r$, $h(t)$ is defined to be the minimal step-cost of moving to a state of the form $t' = x_1, x_2, \ldots, x_r, x_{r+1}$ (that is, always move to the nearest legal city from where you are; if there is no city to move to then $h(t) = 0$).

Another heuristic is as follows. Given some state t, your heuristic function $h(t)$ is the sum of the distance of the closest city c that has not been visited to the last city of the partial tour t plus the distance of any other unvisited city (different from c) to the start city (if there aren't enough unvisited cities to apply this rule then the heuristic value is 0). This heuristic reflects that you want your next city to be visited to be close to the current city but that you don't want to be left with a city that is a long way from the start city.

Yet another heuristic h is as follows. Given a state $t = x_1, x_2, \ldots, x_m$, $h(t)$ is defined to be the minimal step-cost of moving to a state t', where t' is t with some new city inserted somewhere within t, e.g., if $t = 1, 5, 4, 7$ then t' might be $1, 5, 6, 4, 7$. If this heuristic is to be used then the transition function (above) needs to be amended to allow such transitions.

Bear in mind that A* search gives optimal solutions (under mild circumstances) and so unless you have a brilliant heuristic function (which is unlikely) then this method will only work on small instances. I have no idea how the above heuristics will pan out on the instances given!

Hill-climbing search and simulated annealing

Here, the states might be the set of all possible tours of n cities, and one state t' might be a successor of another state t if a swap of the positions of two (or more!) of the cities in the tour t results in the tour t'. The heuristic cost function of a state might be the length of the tour. There are numerous other definitions of a successor function.

Genetic algorithms

In order to formulate the Travelling Salesman Problem for solution by a genetic algorithm, we need to define our population. One way of doing this is to define the population as a set of tours of n cities, represented as strings of length n. The fitness of a member of the population might be just the length of the tour. We now need to come up with a notion of mutation and crossover. One way of defining a mutation is just to randomly swap the positions of two cities within a tour (though there are many others). Defining a notion of crossover is more difficult. However, given two tours $t = x_1, x_2, \ldots, x_n$ and $t' = x'_1, x'_2, \ldots, x'_n$, we could define a new tour as follows.

- Randomly choose some $i \in \{1, 2, \ldots, n-1\}$ and form the strings

$$s = x_1, x_2, \ldots, x_i, x'_{i+1}, x'_{i+2}, \ldots, x'_n$$

and

$$s' = x'_1, x'_2, \ldots, x'_i, x_{i+1}, x_{i+2}, \ldots, x_n$$

(note that these might not be tours as some cities might be missing and some repeated).

- Scan through s and make a list of the cities not appearing in s and a list of the locations containing repeated cities (these lists have the same length). Replace every repetition with a missing city (according to some user-defined strategy). Do the same for s'.

- Hence, we obtain two tours s and s', and we take the crossover as the shortest one.

For those really interested, there is a paper:

- P. Larranaga, C.M.H. Kuijpers, R.H. Murga, I. Inza and S. Dizdarevic, Genetic Algorithms for the Travelling Salesman Problem: A Review of Representations and Operators, *Artificial Intelligence Review* **13** (1999) 129–170

that discusses genetic algorithms for the TSP.

Please note: the above hints are just suggestions and you might care to come up with your own ideas. Also, it will be up to you to (experimentally) vary parameters (e.g., the different probabilities in a genetic algorithm or a simulated annealing algorithm) to improve your solutions.