Data Structures and Algorithms Assignment

Social Network Simulator

Documentation

Reece Jones

Overview

An overview of the project structure is as follows:

* Root directory: main entry point, user interface, and some miscellaneous files.
* dsa directory: generic data structures and algorithms required throughout the project.
* network directory: social network structure and associated operations.
* tests directory: files containing unit tests.

Hopefully the project’s code and structure are straightforward and self-explanatory. Apart from the data structure implementations, of which some are quite involved (particularly the hash table), a logical approach to the social network modelling and operations has led to most of the code being fairly simple.  
The first notable comment on the code overall is the use of some Python-specific “tricks”; for example, dynamic object attribute manipulation and iterator manipulation. Some of these usages are perhaps a bit “fancy” or “clever”, but I have tried to use them with good reason, as will be explained in this documentation.  
Another, similar point is the heavy use of type hints. As I developed the code with an IDE supporting static type analysis, the type hints made life a lot easier in spotting errors before runtime. An additional benefit is making it easier for a reader to understand what a function or variable does/is.

Main entry point (SocialSim.py)

The application’s main entry point is unremarkable. It simply checks the command line arguments and invokes the correct program mode, either simulation\_mode.py or interactive\_mode.py.

Simulation mode (simulation\_mode.py)

This file contains the high-level logic for the simulation mode of the application. The module assumes that the command line arguments (in sys.argv) are correct, from which is parses the program parameters, loads the social network, and begins simulation.

Perhaps the most complex part of this module is the checking when the simulation has completed. As the application has no user interaction in this mode, nothing prevents it from running infinitely and it needs to determine for itself when the simulation should finish. Due to the evolution model specified for this assignment, and since the social network gains no new posts over time, eventually it will have evolved to a stable state. In this state, every person has liked every post they possibly can, and follows every person they possibly can, and further simulation is a no-op. I consider this the “completed” state of the simulation, and thus the simulation should stop here.  
In order to detect this state, we must consider the nature of the network evolution model. Note that a person only ever evolves to be directly connected to people and posts they are transitively connected to, i.e. at the beginning of the simulation, there exists a path from every person to every other person and post to which they can evolve to be directly connected with. Knowing this, a graph traversal can be performed, following edges that represent “X follows Y”, “X likes Y” and “X posted Y”, to enumerate all future follows and post likes. (This effectively solves this simulation in one step, but since step-by-step logging is required, the full simulation must still be run.) These connections can be counted and saved, so that later the network state may be compared to detect if the simulation is complete.

This is the job of the annotate\_solution() function, which runs once at the start of the simulation and “annotates” each person with the number of people they follow and posts they like in the fully-evolved state. A recursive depth-first search is used for simplicity; however, I believe a breadth-first search would also work. As noted in the code comments, I believe recursion is acceptable here, as the maximum possible search depth is equal to the longest simple path through the network, which for our purposes is likely to be low. I believe the default maximum recursion limit in Python is around 1000, for which it should be noted that a simulation of a network that large likely takes on the order of hours to complete anyway (despite my best efforts, re-implementations of data structures in Python is inevitably rather slow).  
This function contains the dynamic object attribute manipulation mentioned in the “Overview” section. The \_visited attribute is dynamically added to vertices as part of the traversal to mark them as previously traversed, and then is removed at the end of the traversal. The annotations done by annotate\_solution() are also dynamically created attributes. The reason I have chosen to dynamically add/remove them as opposed to having them as persistent attributes is simple: they are not required anywhere else in the codebase. Rather than pollute the graph vertex classes, I have utilised the gifts Python has granted and confined their existence to this one module.

One final note on simulation\_mode.py is the extra statistics logging which can be enabled with the STATS\_ENABLED constant. This is solely for the collection of data used for the investigation and may be ignored for typical use.

Interactive mode (interactive\_mode.py)

This file contains the high-level logic and user interfacing for the application’s interactive mode. There is nothing particularly remarkable about this module, except that it contains many if and try/catch statements.

SocialNetwork class (and related) (network/network.py)

The social network is represented by the SocialNetwork class, which stores instances of the Person and Post classes, which represent people and posts in the network. Obviously, a social network is a graph, but rather than have a separate, dedicated, generic graph class, I have opted to implement SocialNetwork, Person, and Post such that together they form an implicit graph structure. I chose this route for simplicity and maximum conciseness – every piece of extra code is more time, more testing, and more potential for bugs. Ultimately this simple design is not only concise, but also quite elegant, logical, and easy to use.

Since natively interpreted Python is not exactly fast, and most of bulk CPU-time-wise of the application would be operations on the network, the implementation of these classes was based almost entirely on desired performance. When choosing data structures and algorithms, one must have *some* desired characteristics in mind, which I have chosen to be speed and only speed. I argue that memory usage in this case is not typically a problem until one runs out of memory, so if the application does not use exorbitant amounts without reason, memory “inefficiency” is not be that much of a concern.  
In any case, the networks loaded and simulated for the purposes of this assignment are not likely to be huge, and as mentioned previously, ultimately Python’s sub-optimal speed is likely to be the limiting factor in running a simulation before modern-day computers’ memory is.

The SocialNetwork class stores a hash table of people keyed by name (names are assumed/required to be unique), as quick lookup of people by their name is often required – particularly the code that loads a network file, which relies on fast person lookup in order to not be terribly inefficient. It also stores a count of the number of posts in the network so that can be accessed in constant time, since there is no one data structure that stores all the posts (they are stored with the person that posts them, as will be explained shortly).  
Finally there are two externally-supplied attributes \_expected\_people and \_expected\_posts, which are solely for performance optimisation. They indicate the expected total number of people and posts in the network, respectively. As the people and posts are frequently accessed via hash tables, these parameters give a starting capacity for those hash tables such that resizes (which have time complexity) may be avoided.  
That is all that the SocialNetwork class explicitly stores. The application and simulation are heavily based on operations on people, as such the Person class supplies the majority of operations supported by the network.