EC 40, INTRODUCTION TO REINFORCEMENT LEARNING
CODING EXERCISE 1

In this assignment, you will implement policy evaluation and policy iteration. We consider a gridworld MDP, where we have a 4×4 grid, and at each step you can move up, down, left, or right. If the move leads you out of the grid, you stay where you are. Each transition has -1 reward. The upper-left and bottom-right corner are terminal states. The discount factor equals one.

The attached code below will compute the value function of a policy which chooses an action uniformly at random, provided you fill in the line that begins with

###

correctly. Your first task is to do so. Your second task are to modify this code to implement value iteration.

If you find the code below confusing, please feel free to ignore it completely and write your own code which, on this MDP (i) evaluates the "choose an action at random" policy (ii) finds the optimal policy and value function through value iteration.