Assignment: Reinforcement Learning

In this assignment you will apply reinforcement learning to the 4x3 grid shown below. You may use any programming language to implement the required algorithms (although Python is preferred and recommended).

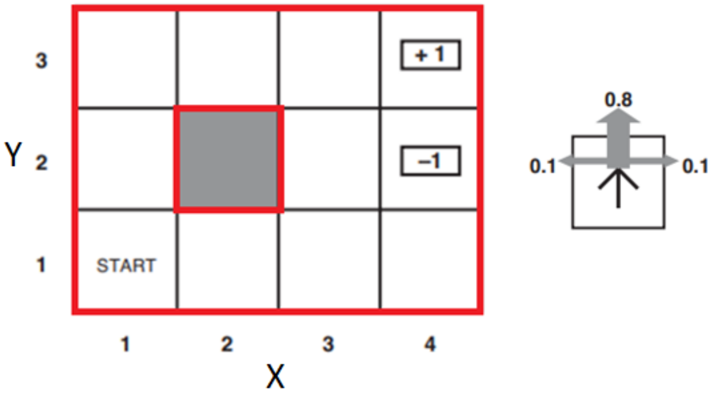

Figure 1: 4x3 world. Red lines show walls.

The rules of the environment are as follows.

1. An agent is supposed to start in cell (1,1).
2. Available actions, $A(s) = \{up,\ down,\ left,\ right\}$, where state s is some cell (x, y) with x being the column number and y being the row number.
    a. An action results in the intended movement with probability 0.8 but can cause movement in one of the perpendicular directions with probability 0.1 each.
    b. Bumping into a wall results in staying in the same cell.
3. $R(s)$ denotes the reward received in state s. $R\big((4{,}3)\big) = +1, R((4{,}2)) = -1, and\ for\ any\ other\ (x, y),\ \ R(x, y) = -0.04.$
4. The agent will exit the environment if any of the two terminal states, i.e., $(4{,}3)\ or\ (4{,}2)$, is reached.
5. The objective of the agent is to maximize the total sum of rewards received. Assume that rewards are additive. Use a discount factor $\gamma = 0.9$, wherever applicable.
6. The environment is fully observable – the agent knows exactly which sell it is in

Deliverables:

1. Answer to Q1 below in the form of a report in PDF format. Please copy and paste the tables from this document as needed.
2. Source code in Zipped format (this should be separate from the PDF mentioned above).

Q1

Find the true utilities $U(s)$ and the optimal policy π^* by implementing the following algorithms

i. policy iteration
ii. value iteration.

Use discount factor $\gamma = 0.9$. Fill Tables 1 and 2 below with the values you obtain for each algorithm.

Hint: Use the rules of the environment given above (especially Rule 2) to compute the transitions probabilities $P(s' \mid s, a)$ that specify the probability of reaching state s' if action a was executed in state s.

Table 1: Utility values. In each cell enter the value of $U(s = (x, y))$

	3				
Y	2				
	1				
		1	2	3	4
			X		

Table 2: Optimal policy. In each cell, enter the $action \in \{up, down, left, right\}$ as specified by the optimal policy π^*.

	3				
Y	2				
	1				
		1	2	3	4
			X		