Problem 5 [30 points]. Carry out policy iteration over the MDP example covered in class with R given in Table 2 and $\gamma = 0.9$. For a state s, if $R(s) = \pm 1$, s is a terminal state. For the transition model, assume that the agent has 0.9 probability of going to the intended direction and 0.1 probability of moving to the left. For example, if the agent is at the lower left corner (coordinates $(1, 1)$) and intends to go right, then it will reach $(2, 1)$ with 0.9 probability and $(1, 2)$ with 0.1 probability. If a target cell is not reachable, then the corresponding probability goes back to the current cell. For example, if the agent is at $(3, 3)$ and is trying to go up, then with 0.1 probability it goes to $(2, 3)$ and with 0.9 probability it is stuck at $(3, 3)$. For your answer you should provide:

a) [15 points]. The first two iterations of your computation.

b) [15 points]. The converged rewards and the extracted policy. For this problem, you need to provide last two iterations showing that the value changes are within 0.001 for all cells.

Table 2: Reward R for a 4×3 grid world

-0.05	-0.05	-0.05	+1
-0.05	OBS	-0.05	-1
-0.05	-0.05	-0.05	-0.05

As a suggestion, you should complete the first question manually to make sure you will be able to do so, for obvious reasons :). For solving the second, it is perhaps better to do it using a program, perhaps using Python or excel.