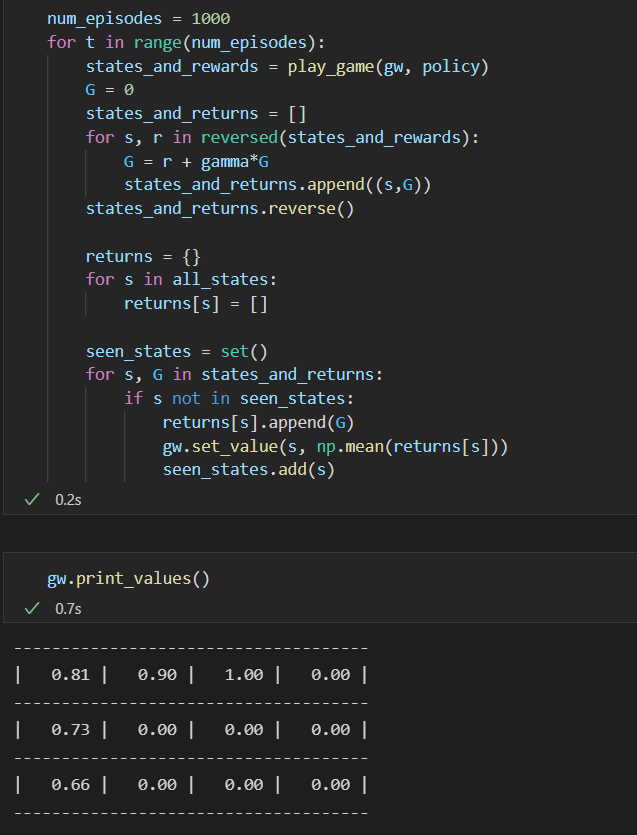
In policy evaluation and value iterations, we assumed that the agent knows everything about the environment. Thus, the agent has access to the model of the environment. That means the agent knows the consequences of actions on states. In real world applications, the agent may not know anything about the environment. Therefore, we need an approach where the agent does not need to know about the environment and gain knowledge after interacting with it. First-visit Monte Carlo method is one of these approaches that does not need the knowledge of the immediate rewards. For week 4 lab, we are going to use First-visit Monte Carlo prediction to calculate the values of states. So, basically, First-visit MC function returns states that’ve been visited and their rewards, and how it works is by starting at default starting node, (0,0) in this case, then while the node is not converged it will get an action from the policy, get the reward then move to the next state, and add it to a list called “states and rewards”, then it’ll calculate the return of states, it will insure that only calculate first seen or first visit states, and ignore the states that have been visited before. The lab also has a function to play the game 10 times, the results are the same as one episode because it follows a policy and calculate first visit states which will ignore in next episodes and it does not matter how many times we run the function, same results come up. I have tried to play the game 1000 times and same values come up, because the grid and the policy are deterministic.



However, not all states values are calculated because we followed a policy where we started at initial state that determined before, which is (0,0). The last part of the lab is to start from random state in the grid that is not a barrier, or a terminal then run the MC function and get the reward and (G)s values.

I went further to try to change the discount factor (gamma) to 0.7, to see if it is going to change procedure, it turned out as expected to change the values to the 0.7 discount, and the rest is the same. To sum up, is first visit MC method, all states are not discovered if we follow a given policy, and it will discover all states starting from a random state. Returns is only calculated to first visit state, then ignore the state in next iteration. Below is the values of the grid with 0.7 penalty.

