CS444 Markov Decision Processes

1. Consider the MDP represented by the following simple maze:

A .75			S						B 1.0

The robot in this maze starts at location S. At each time step he has the option of taking one of two actions LEFT, or RIGHT. These actions deterministically move the robot one space to the left or right. When the robot exits state A, he receives a reward of 0.75. When he exits state B he receives a reward of 1.0. Any action taken in states A or B end the trial: the robot exits the maze and no more steps are taken. Assume a discount factor of $\gamma = .9$. As a reminder, here is the equation for the utility of a state under the optimal policy in an MDP:

$$V(s) = \max_a \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma V(s'))$$

- In each square of the maze, write the utility (expected discounted reward under the optimal policy) of being in that state. States A and B are already completed. (1pt)
- In each square of the maze (other than A and B) indicate the optimal action for the robot to take.

2. Fill out the following tables with the appropriate Q-values for each action in each state. Assume the same reward function and transition model as the previous question. As a reminder, the constraint equation for Q-values is:

$$Q(s,a) = \sum_{s'} P(s'|s,a)(R(s,a,s') + \max_{a'} \gamma Q(s',a'))$$

<table>
<tr><td>A
.75|.75</td><td>|</td><td>|</td><td>S
|</td><td>|</td><td>|</td><td>|</td><td>|</td><td>|</td><td>B
1.0|1.0</td></tr>
</table>

3. In the previous two questions, the robot's actions were deterministic, which is not really in the spirit of Markov Decision Processes. For this question, we will consider the same environment, but we will assume that the robot's actions have probabilistic outcomes. In particular, we will assume that the robot's actions only succeed with a probability of .9, otherwise the robot stays in the same location. (Actions taken in states A and B are still deterministic).

   As a reminder, the update rule for value iteration can be written as follows:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

   Perform three steps of value iteration given the following initial values:

V_0

A			S						B
0	0	0	0	0	0	0	0	0	0

V_1

A			S						B

V_2

A			S						B

V_3

A			S						B