CS444 Q-Learning

1. The robot in the maze below 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 $\gamma = .9$.

   For this question, assume that we are using Q-learning to find a policy for this environment. Assume a learning rate of $\alpha = .1$. All Q-values are initialized to 0:

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

   As a reminder, the update equation for Q-learning is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

   - In the first trial, the agent randomly follows the following action sequence (starting from S):

     ⟨LEFT, RIGHT, LEFT, LEFT, LEFT, LEFT⟩

     What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|

   - In the second trial, the agent randomly follows the following action sequence (starting from S):

     ⟨LEFT, LEFT, LEFT, LEFT⟩

     What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|

   - In the third trial, the agent randomly follows the following action sequence (starting from S):

     ⟨LEFT, LEFT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT, RIGHT⟩

     What are the estimated Q-values after this trial?

A			S						B
|	|	|	|	|	|	|	|	|	|