EXTRA PROBLEM consider the following mini-grid (rewards shown on left, state names shown on right).

+4	A	B	+16

L	A	B	R

In this scenario, the discount is $\gamma = 1$. The failure probability is actually $f = 0$, but, now we do not actually know the details of the MDP, so we use reinforcement learning to compute various values. We observe the following transition sequence (recall that state X is the end-of-game absorbing state):

s	a	s'	r
A	*Right*	B	0
B	*Right*	R	0
R	*Exit*	X	16
B	*Right*	R	0
R	*Exit*	X	16
A	*Right*	B	0
B	*Left*	A	0
A	*Left*	L	0
L	*Exit*	X	4

(q) [2 pts] After this sequence of transitions, if we use a learning rate of $\alpha = 0.5$, what would Q-learning learn for the Q-value of all state-action pairs? Remember that $Q(s, a)$ is initialized with 0 for all (s, a).
Hint: Q-learning update rule is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

$$Q(L, Exit) = 2$$

$$Q(A, Left) = 0$$

$$Q(A, Right) = 2$$

$$Q(B, Left) = 1$$

$$Q(B, Right) = 4$$

$$Q(R, Exit) = 12$$

(r) [2 pts] If those transitions were generated by a SARSA algorithm with ϵ-greedy action selection, what is the Q-values of all state-action pairs learned by SARSA? Remember that $Q(s, a)$ is initialized with 0 for all (s, a).
Hint: SARSA update rule is

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

$$Q(L, Exit) = 2$$

$$Q(A, Left) = 0$$

$$Q(A, Right) = 0$$

$$Q(B, Left) = 0$$

$$Q(B, Right) = 4$$

$$Q(R, Exit) = 12$$

The critical point happens at the 6-th transition $(A, Right, B, 0)$. Q-learning will see the non-zero Q-value of $(B, Right)$ and hence update $Q(A, Right)$ to non-zero. In contrast, SARSA will use the transition tuple $SARS'A' = (A, Right, 0, B, Left)$ but $Q(B, left) = 0$ and hence $Q(A, Right)$ remains 0 after this step.