Deepak Shine(dshine2)

MP1 – Part 2

1a. The end total reward is always 0. Because there is no randomness to how the agent picks its next move, it will continuously go to where it is programmed to. This means that it will go to the same locations. And it happens to be that these string of locations result in the agent never achieving any reward.

1b. Because the agent is now allowed to make random decisions it doesn't continuously go to the same spots. And because it now explores random paths, it can get out of the path that resulted in 0 total end reward each time that happened in part A

1c. The reward value is very low. This is because the chance of making a random choice is so high, the agent will make the greedy choice less frequently. And because it doesn't make the greedy choice as often, its reward isn't as high as it could be.

2a. Because the agent does not know where the thief is, it will continuously get intercepted by the thief, and will continuously receive penalties, which results in so many episodes resulting in negative reward values.

2b. Because the agent now knows where the thief is, it can avoid the thief which results in much higher end reward values for the episodes.

2c. I found that the best parameters were epsilon = .01, rate = .15. I first tried to find the best epsilon to pick and found that .1 resulted in the best balance between not repeatedly going the same way and also making sure that the agent pursued the greedy path often enough to maximize reward. Then I moved onto learning rate and found that it was harder to pinpoint. But .1 seem good because it made sure that we trusted our new Q values but not so much that in the case where the new Q values have a lot of error, that we don't completely lose all that we have learned so far.

3. Plotting the data w/ the y-axis being the total reward and the x-axis being the episode, I found that the reward increases logarithmically w/ the number of episodes.