Homework 3

To complete the problems on this assignment, use the following definition of this markov decision process:

-2	+3	
(c)	(d)	-2
(a)	(b)	+1

MDP Definition

- Living reward, R(s), always returns -0.1
- Transition probabilities, P(s'|a, s): 0.6 chance of going the direction you choose, 0.2 chance each of going to the left or right instead of the chosen direction. If you run into a wall, you don't move (s' is the same as s).
- Discount factor, γ = 0.9
- Equation for utility update in value iteration:

$$U_{i+1}(s) = R(s) + \gamma * max_{a \in A(s)} \sum_{s'} P(s'|a, s) U_i(s')$$

- Equation for utility update in policy iteration:

$$U_{i+1}(s) = R(s) + \gamma * \sum_{s'} P(s'|\pi_i(s), s) U_i(s')$$

Problems

1. Do 2 rounds of value iteration. Start with utilities at 0 for each state. For each round, find the max action, and then use that action to update the utility.

2. Do 2 rounds of policy iteration. Start with utilities at 0. For each round, use a max to find the policy that is being followed (break ties in the order: up, right, down, left). If the policy doesn't differ from the previous round, stop. Otherwise, update the utilities two times using that policy.