Aluno: Evandro Matheus Schmitz

Respostas da Atividade 1

Representação do cenário:

	0	1	2	3	4	5	6
0			Base	Base	Base		
1							
2				Objeto			
3							
4							
5	Agente						

1. Modelagem do MDP:

(a) Apresente a modelagem de estados considerada, bem como a quantidade de estados presentes no MDP. Inclua na contagem os estados não-válidos;

R: O MDP pensado para este exercício consistem em armazenar 3 informações para representar o estado do agente: a posição do **Agente**, a posição do **Objeto** quando o **Agente** tiver agarrado ele (se mover a direita ou esquerda do **Objeto**) e se o estado é final (se a posição do **Objeto** for igual a uma das células de **Base**).

O **Agente** não pode sair do espaço do mundo (6 x 7) nem atravessar paredes (células pretas) ou passar por cima do **Objeto** (se mover para a célula que ele está ocupando). Quando o **Agente** pega o **Objeto**, mas mesmas restrições de movimentação em relação ao mundo e as paredes são aplicadas ao **Objeto**, que passa a se mover junto com o **Agente**, ou seja, o **Agente** só se mover se ele puder se mover e mover o **Objeto** junto. Considerando só o **Agente** e a movimentação dentro do mundo existem 9 estados inválidos (8 paredes e a célula do Objeto) e 33 estados válidos. Quando o **Agente** pega o **Objeto** só existem 27 estados válidos (linhas 0 até 3 tirando 1 parede), visto que a linha 5 se torna inacessível, dados 15 estados inválidos.

(b) Apresente a modelagem das ações que o agente pode executar;

R: As restrições da atividade dizem que o agente pode executar 5 ações:

- Ir para cima (up), diminuindo em 1 (-1) a sua posição em uma determinada linha;

- Ir para baixo (down), aumentando em 1 (+1) a sua posição em uma determinada linha;
- Ir para a direita (right), aumentando em 1 (+1) a sua posição em uma determinada coluna;
- Ir para a esquerda (left), diminuindo em 1 (-1) a sua posição em uma determinada coluna;
- Ficar parado (nothing), onde o agente não muda de posição;

Agarrar o **Objeto** é algo feito automaticamente pelo **Agente** se o mesmo estiver a direita ou a esquerda do **Objeto**, por isso esta ação não precisou ser modelada.

(c) Apresente a modelagem da função de recompensa, com as situações em que o agente é recompensado bem como a magnitude da recompensa. Justifique as suas escolhas.

R: A função de recompensa determina que o **Agente** receberá uma pontuação de +1 se o **Objeto** chegar a base, ou seja, caso o **Agente** agarre o **Objeto** e o transporte até base. O **Agente** receberá uma pontuação de -1 para cada outro movimento que não seja entregar o **Objeto** na base. As pontuações foram escolhidas com base nos exemplos passados em aula.

2. Configuração dos Experimentos

(a) Apresente os valores de taxa de aprendizagem (alfa) e fator de desconto (gamma) do algoritmo de aprendizagem Q-Learning;

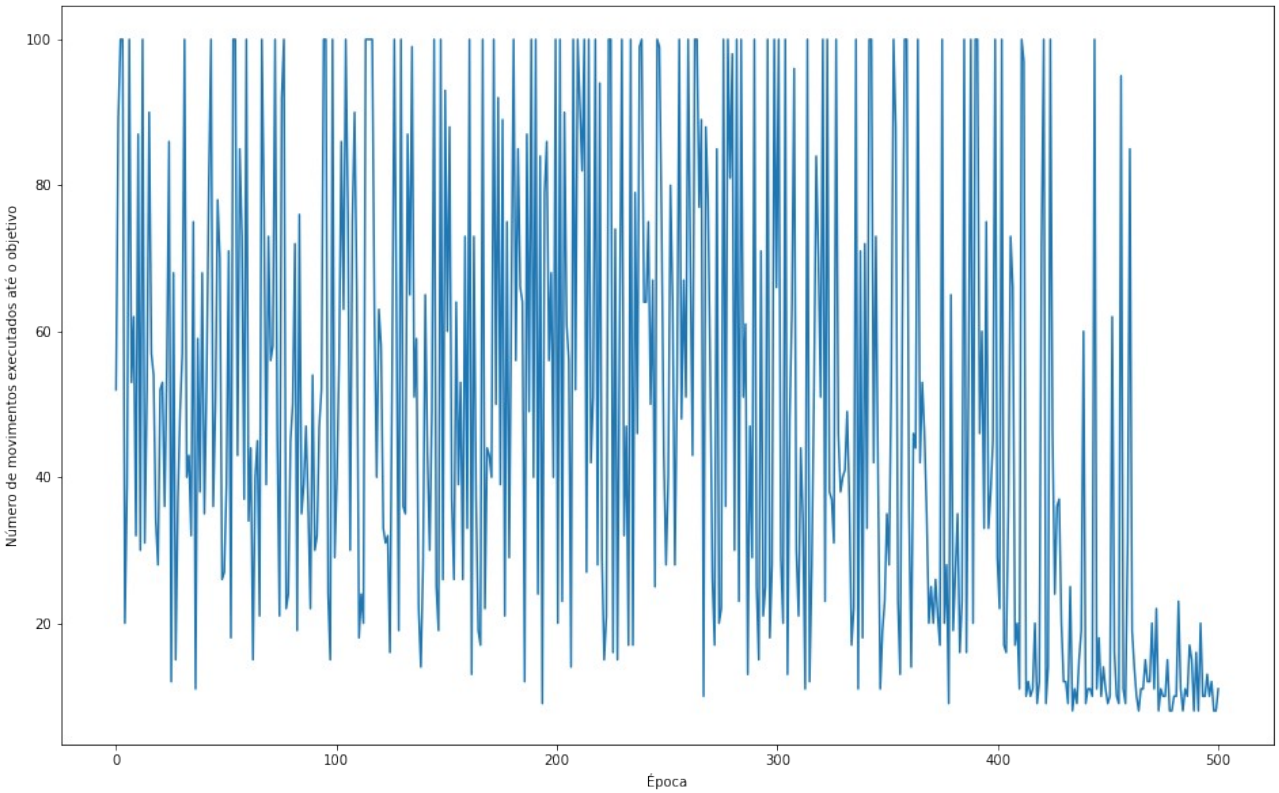
R: Os valores usados no experimento foram um alfa de 0,2 e um gamma de 0,9. A política de exploração (epsilon) foi de 0,3, ou seja, a cada passo o **Agente** tem 30% de chance de escolher uma ação aleatória e em 70% para escolher a ação que mais lhe beneficia, com base no seu conhecimento.

(b) Apresente as configurações do horizonte de aprendizagem, que é representado pela quantidade máxima de passos de tempo por episódios, quantidade máxima de episódios, e política de exploração ao longo do tempo;

R: A quantidade máxima de passos de tempo foi 100, o total de máximo de episódios foi 500.

3. Resultados Experimentais

(a) Apresente a curva de convergência, representada pela quantidade de passos (timesteps) necessários para resolver a tarefa ao longo do tempo (episódios).



R: Este gráfico pode variar conforme a execução, pois o agente é ensinado a chegar na base com o objeto. Se for aumentado o número de épocas será possível ver até a formação de vales, épocas onde o agente demora muito seguido de um período onde ele demora pouco, voltando a demorar muito.

(b) Apresente o tempo de processamento necessário para resolver o problema:

R: O tempo necessário varia conforme a execução do experimento, mas fica na casa de alguns segundos, para as configurações apresentadas. Em duas instâncias ele ficou entre 6 à 7 segundos.