Nome: Miller Biazus
Cartão: 00187984

Q-Learning

O algoritmo implementado deveria, a partir de um estado inicial S, percorrer o ambiente definido, evitando cair no penhasco e utilizando apenas movimentos CIMA, BAIXO, ESQUERDA, DIREITA.

Implementação:

Para rodar o algoritmo utilize, no prompt de comando:

```
python qlearning.py
```

Atenção: O trabalho foi feito utilizando-se Python 3. Versões mais antigas do Python podem não rodar o arquivo.

O aprendizado é feito com 50 sessões de treinamento. Este valor pode ser modificado pela variável global TRAINING_SESSIONS.

Em cada sessão de treinamento, o agente explora o ambiente definido e, para cada movimento, recebe uma recompensa. Caso se movimente rumo ao penhasco, recebe uma recompensa muito baixa (-100) e volta ao estado inicial. Caso tente sair do ambiente (fora da grade), continua no mesmo local e não recebe recompensa nem recalcula sua matriz de aprendizado Q. O movimento para o estado objetivo possui uma recompensa alta (+100).

O valor de Gamma foi definido em 0.9, ou seja, o agente irá considerar recompensas futuras com maior peso.

Quando terminam as sessões de treinamento, a matriz de conhecimento Q é percorrida a partir do estado inicial utilizando algoritmo guloso que consome o movimento com o maior valor de Q:

1. Estado atual <= estado inicial
2. A partir do estado atual, procura a ação que possui o maior valor de Q.
3. Estado atual <=próximo estado
4. Repetem-se 2 e 3 até que Estado atual == estado objetivo.

A sequência de estados do algoritmo acima é a melhor sequência de estados a ser percorrida com base na matriz Q aprendida.

A matriz Q é gerada baseando-se nas posições do grid: cada linha da matriz Q representa um estado possível, enquanto cada coluna representa uma ação (linha 1 x coluna 1 representa o valor para ir para cima no estado 1). Para isso, leva-se em consideração as posições da grade do ambiente representada abaixo:

37	38	39	40	41	42	43	44	45	46	47	48
25	26	27	28	29	30	31	32	33	34	35	36
13	14	15	16	17	18	19	20	21	22	23	24
1	2	3	4	5	6	7	8	9	10	11	12

A tabela de recompensas da grade é a seguinte:

-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-100	-100	-100	-100	-100	-100	-100	-100	-100	-100	100

Para 50 sessões de treinamento, a matriz Q gerada deve ser similar à seguinte (em destaque o caminho seguido pelo algoritmo guloso):

Posição	Cima	Baixo	Esquerda	Direita
1	15.52	0	0	-99.86
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	15.4	12.92	0	18.41
14	18.27	-99.86	15.45	21.62
15	21.4	-99.86	18.23	25.27
16	24.85	-99.86	21.34	29.46
17	28.78	-99.86	24.72	34.33
18	32.9	-99.86	28.93	40.11
19	37.97	-99.86	33.02	47.56
20	43.51	-99.86	39.78	55.45
21	48.46	-99.86	45.49	64.82
22	57.52	-99.86	53.34	74.68
23	63.53	-99.33	61.1	86.94
24	71.04	99.49	69.95	0
25	12.84	15.52	0	18.27
26	15.43	18.4	15.4	21.46
27	18.31	21.62	18.27	25.02
28	21.47	25.25	21.41	28.97

29	24.98	29.35	24.94	33.37
30	28.92	34.46	28.79	38.32
31	33.33	40.66	32.97	44.12
32	38.27	47.79	38.15	50.52
33	43.77	55.98	43.64	58.11
34	49.89	64.95	49.89	66.75
35	57.28	75.54	56.89	76.29
36	58.08	87.68	65.51	0
37	0	15.4	0	15.43
38	0	18.27	12.84	18.31
39	0	21.45	15.43	21.5
40	0	24.97	18.24	25.06
41	0	28.89	21.42	29.02
42	0	33.35	24.95	33.49
43	0	38.43	28.99	38.44
44	0	44.05	33.36	44.06
45	0	50.69	37.99	50.44
46	0	58.14	43.56	57.88
47	0	66.27	49.61	65.33
48	0	75.78	56.46	

O cálculo de tais valores é feito seguindo a fórmula:

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

,sendo $\max_{a'} Q[s',a']$ a ação do estado seguinte com valor máximo de recompensa.

A partir do caminhamento na tabela de conhecimento, segue o resultado na grade:

. = posição vazia
# = penhasco
0 = caminho deixado pelo agente

Posição Inicial

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 1: Cima

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

0	.	.	.	.	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 13: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	.	.	.	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 14: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	.	.	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 15: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	.	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 16: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	.	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 17: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	.	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 18: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	.	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 19: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	.	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 20: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	.	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 21: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	0	.	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 22: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	0	0	.
0	#	#	#	#	#	#	#	#	#	#	.

Posição 23: Direita

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	0	0	0
0	#	#	#	#	#	#	#	#	#	#	.

Posição 24: Baixo

.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
0	0	0	0	0	0	0	0	0	0	0	0
0	#	#	#	#	#	#	#	#	#	#	0