WSTĘP DO SZTUCZNEJ INTELIGENCJI

Ćwiczenie 5 – Algorytm Q-learning

JAKUB KWAŚNIAK 331396

Wstęp

1. Zaimplementować algorytm Q-learning, a następnie użyć go do wytrenowania agenta rozwiązującego problem Cliff Walking https://gymnasium.farama.org/environments/toy_text/cliff_walking/
2. Stworzyć wizualizację wyuczonej polityki i umieścić ją w sprawozdaniu. Wzór wizualizacji https://gymnasium.farama.org/tutorials/training_agents/FrozenLake_tuto/#visualization

1. Zaimplementowany został zachłanny algorytm Q-learning (Epsilon-Greedy), który wyznacza najlepszą dla siebie akcję w danym stanie według formuły Bellmana:

   Bellman's Equation:

   Update Q(s,a):= Q(s,a) + lr [R(s,a) + gamma * max Q(s',a') - Q(s,a)]

   delta = [R(s,a) + gamma * max Q(s',a') - Q(s,a)]

   Q(s,a):= Q(s,a) + lr * delta

   ****

   w prostych przypadkach takich jak cliff walking często w pierwszych epizodach (kiedy świat nie jest jeszcze oceniony) max Q(s',a') == Q(s,a) więc gamma nie powinna być równa 1

   W zaimplementowanym środkowsiku z pakietu gymnasium nagrody są nas†epujące:

   Każdy krok -> -1

   Dojście do celu -> 0

   Spadnięcie z klifu -> -100

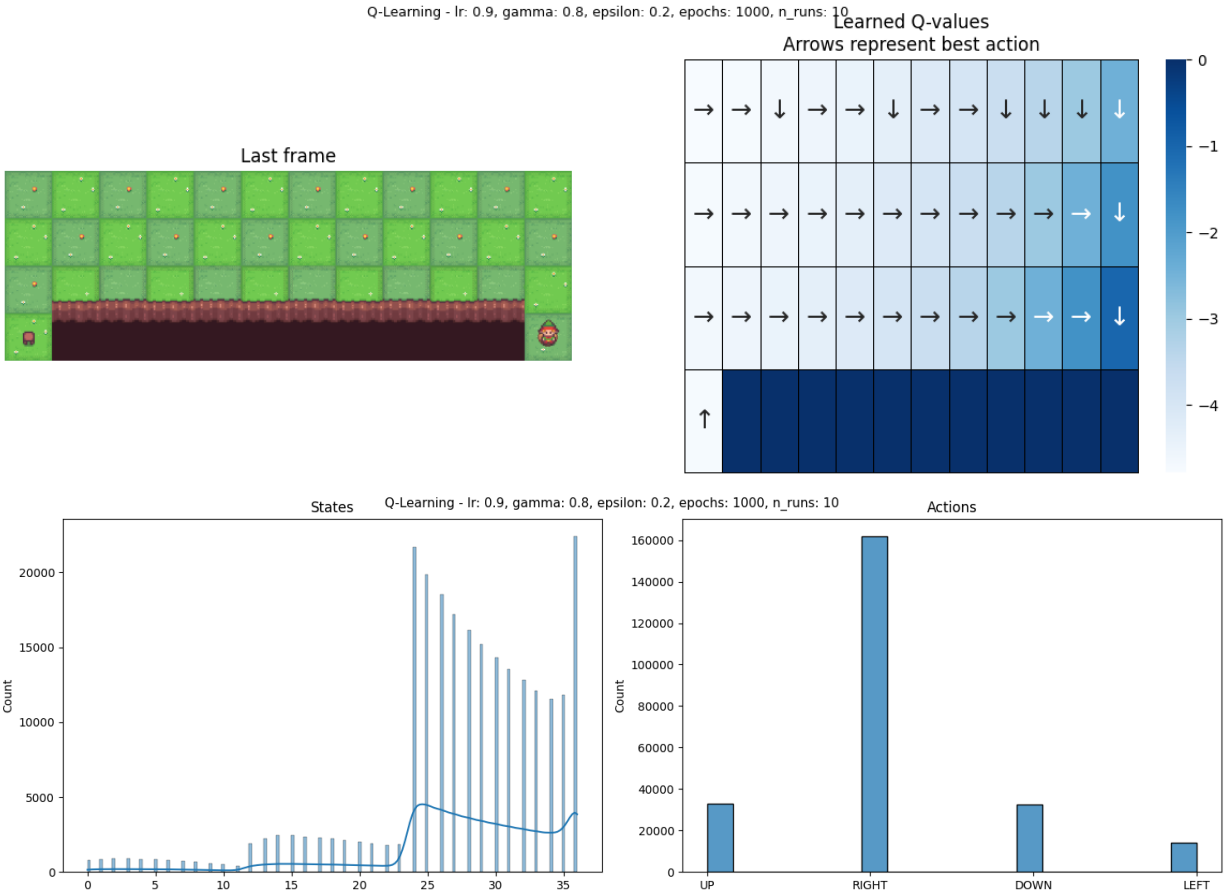
   Dlatego ustawienie gamma=1 uniemożliwiłoby algorytmowi ocene środowiska I naukę dobrej taktyki

   ****

   Użycie Epsilon – Greedy pozwala w pewnym stopniu rozwiązać problem „Exploration-exploitation dilemma" – jako parametr epsilon przyjmuje wartość będącą prawdopodobieństwem eksploracji, zwyczajowo jest to mała wartość by uczeń mógł eksploatować najlepszą znalezioną część środowiska a jednocześnie czasami eksplorował resztę środowiska i nie zamykał się na inne możliwości – może znajdzie równie dobrą lub lepszą część środowiska.

Wyniki

Eksperyment 1:



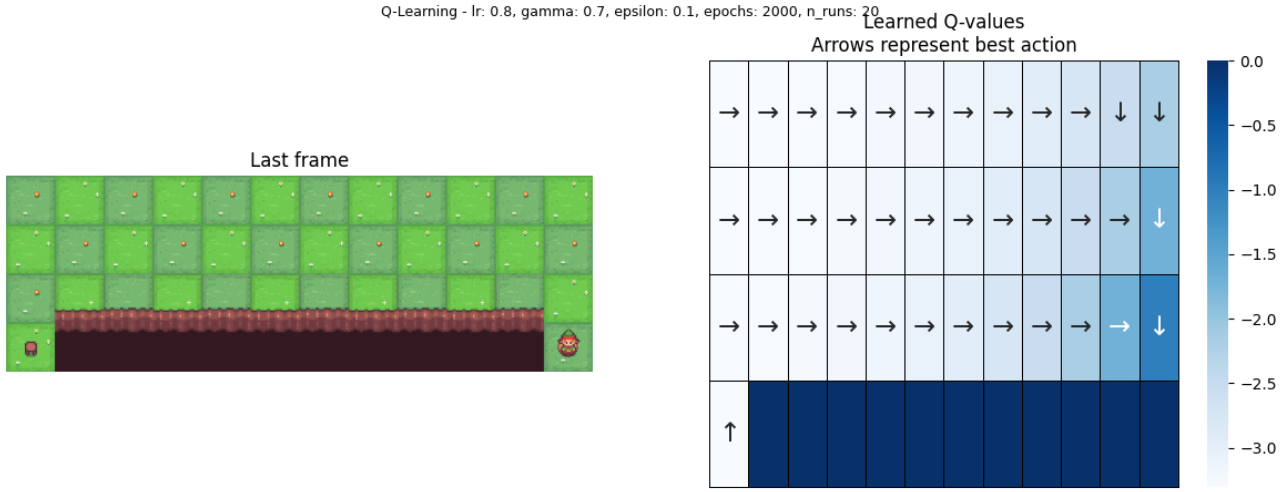
Qtable

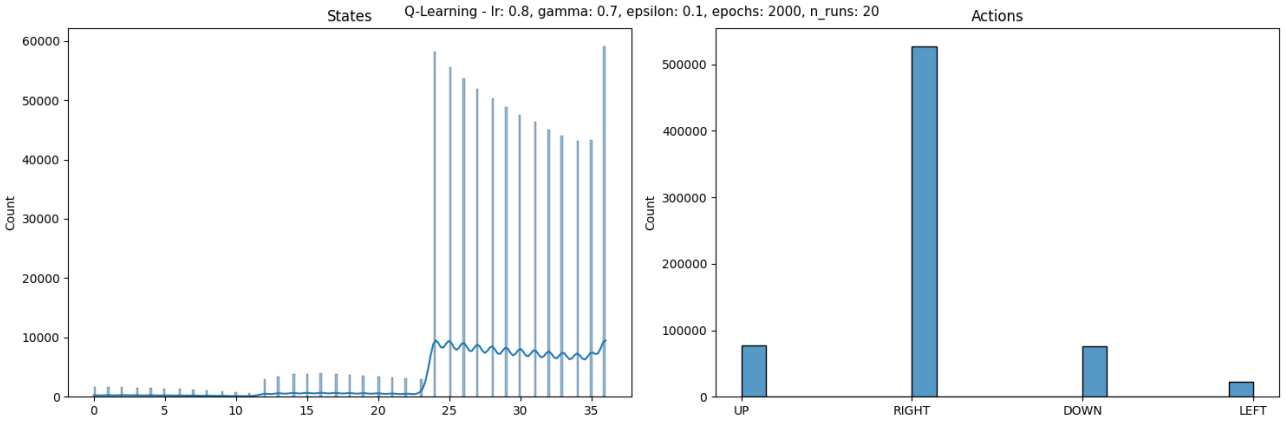
```
[[  -4.79856173   -4.77741172   -4.7787826    -4.80137707]
 [  -4.76627962   -4.72509933   -4.72510254   -4.77943245]
 [  -4.70653806   -4.65640051   -4.65639986   -4.75084546]
 [  -4.62565676   -4.57050263   -4.57050307   -4.69251911]
 [  -4.54975341   -4.46312892   -4.46312899   -4.60967224]
 [  -4.43082614   -4.32891134   -4.32891134   -4.53187019]
 [  -4.29539208   -4.1611392    -4.1611392    -4.35467725]
 [  -4.10643579   -3.951424     -3.951424     -4.24984891]
 [  -3.8815899    -3.68928      -3.68928      -4.10021206]
 [  -3.6151166    -3.3616       -3.3616       -3.85862969]
 [  -3.27214525   -2.952        -2.952        -3.57479846]
 [  -2.88472775   -2.90325998   -2.44         -3.21767825]
 [  -4.81449272   -4.72512209   -4.72512209   -4.78008201]
 [  -4.77849103   -4.65640262   -4.65640262   -4.78001824]
 [  -4.72508177   -4.57050327   -4.57050327   -4.72512183]
 [  -4.6563953    -4.46312909   -4.46312909   -4.65639827]
 [  -4.57044583   -4.32891136   -4.32891136   -4.57039029]
 [  -4.46312374   -4.1611392    -4.1611392    -4.4631286 ]
 [  -4.32890434   -3.951424     -3.951424     -4.32891127]
 [  -4.16113827   -3.68928      -3.68928      -4.16113869]
 [  -3.94585857   -3.3616       -3.3616       -3.95141985]
 [  -3.68923787   -2.952        -2.952        -3.68927964]
 [  -3.36159998   -2.44         -2.44         -3.36159946]
 [  -2.95199384   -2.43999992   -1.8          -2.95199997]
 [  -4.78009767   -4.65640262   -4.78009767   -4.72512209]
 [  -4.72512209   -4.57050327 -103.78009767   -4.72512209]
 [  -4.65640262   -4.46312909 -103.78009767   -4.65640262]
 [  -4.57050327   -4.32891136 -103.78009767   -4.57050327]
 [  -4.46312909   -4.1611392  -103.78009767   -4.46312909]
 [  -4.32891136   -3.951424   -103.78009767   -4.32891136]
 [  -4.1611392    -3.68928    -103.78009767   -4.1611392 ]
 [  -3.951424     -3.3616     -103.78009767   -3.951424  ]
 [  -3.68928      -2.952      -103.78009767   -3.68928   ]
 [  -3.3616       -2.44       -103.78009767   -3.3616    ]
 [  -2.952        -1.8        -103.78009767   -2.952     ]
 [  -2.44         -1.8          -1.           -2.44      ]
 [  -4.72512209 -103.78009767   -4.78009767   -4.78009767]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]]
```

Cumulative Rewards

	Episodes	Rewards	Steps	cum_rewards	map_size
0	0	-866.0	173.0	-866.0	4x12
1	1	-1437.0	447.0	-2303.0	4x12
2	2	-653.0	158.0	-2956.0	4x12
3	3	-275.0	77.0	-3231.0	4x12
4	4	-44.0	44.0	-3275.0	4x12
...	...	...	...	...	...
9995	995	-16.0	16.0	-108151.0	4x12
9996	996	-19.0	19.0	-108170.0	4x12
9997	997	-13.0	13.0	-108183.0	4x12
9998	998	-13.0	13.0	-108196.0	4x12
9999	999	-238.0	40.0	-108434.0	4x12

Eksperyment 2:





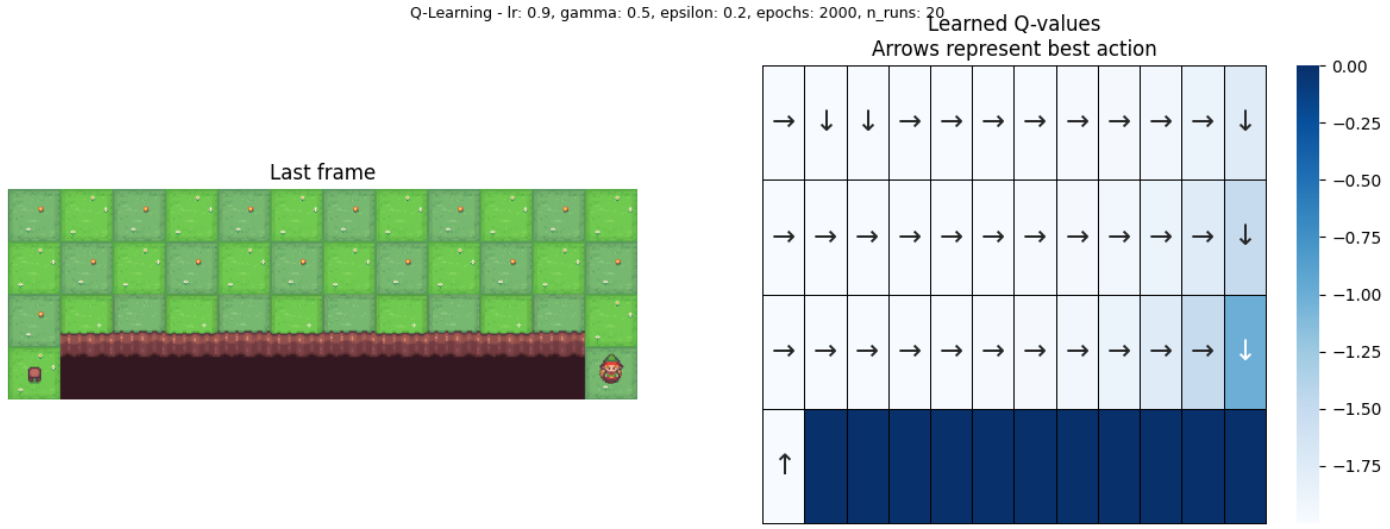
Qtable

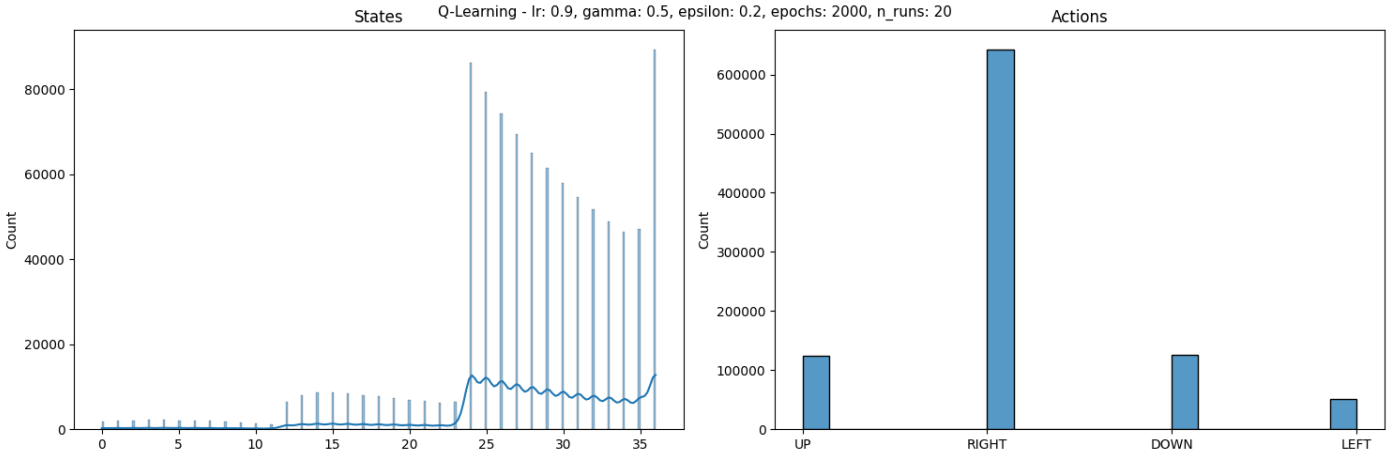
```
[[  -3.30404283   -3.30285886   -3.30724161   -3.30481546]
 [  -3.29949508   -3.29579959   -3.29755676   -3.3028931 ]
 [  -3.28993866   -3.28380758   -3.28382005   -3.29386627]
 [  -3.27349714   -3.26535106   -3.26623946   -3.27818769]
 [  -3.24770803   -3.23805532   -3.23820355   -3.26916839]
 [  -3.21845516   -3.19815318   -3.19848691   -3.22871627]
 [  -3.16761751   -3.14083524   -3.14091436   -3.20263571]
 [  -3.10978442   -3.05863626   -3.05873024   -3.1365486 ]
 [  -3.00139488   -2.94108126   -2.94109053   -3.07208961]
 [  -2.88122739   -2.77306742   -2.77307537   -2.9368804 ]
 [  -2.69128056   -2.53299655   -2.53299262   -2.77875136]
 [  -2.29597689   -2.28138357   -2.18999999   -2.43991229]
 [  -3.30831945   -3.301037     -3.301037     -3.30848193]
 [  -3.30326102   -3.28719571   -3.28719571   -3.30870983]
 [  -3.29250852   -3.26742244   -3.26742244   -3.29964667]
 [  -3.27847186   -3.23917492   -3.23917492   -3.28485522]
 [  -3.25908415   -3.19882131   -3.19882131   -3.26642959]
 [  -3.2328314    -3.1411733    -3.1411733    -3.23328171]
 [  -3.19525134   -3.058819     -3.058819     -3.19228358]
 [  -3.13435293   -2.94117      -2.94117      -3.12953143]
 [  -3.05127697   -2.7731       -2.7731       -3.0545467 ]
 [  -2.91942065   -2.533        -2.533        -2.92703244]
 [  -2.75461334   -2.19         -2.19         -2.74774221]
 [  -2.50463882   -2.18438753   -1.7          -2.47628405]
 [  -3.3107259    -3.28719571   -3.3107259    -3.301037  ]
 [  -3.301037     -3.26742244 -102.3107259    -3.301037  ]
 [  -3.28719571   -3.23917492 -102.3107259    -3.28719571]
 [  -3.26742244   -3.19882131 -102.3107259    -3.26742244]
 [  -3.23917492   -3.1411733  -102.3107259    -3.23917492]
 [  -3.19882131   -3.058819   -102.3107259    -3.19882131]
 [  -3.1411733    -2.94117    -102.3107259    -3.1411733 ]
 [  -3.058819     -2.7731     -102.3107259    -3.058819  ]
 [  -2.94117      -2.533      -102.3107259    -2.94117   ]
 [  -2.7731       -2.19       -102.3107259    -2.7731    ]
 [  -2.533        -1.7        -102.3107259    -2.533     ]
 [  -2.19         -1.7          -1.           -2.19      ]
 [  -3.301037   -102.3107259    -3.3107259    -3.3107259 ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]]
```

Cumulative Rewards

	Episodes	Rewards	Steps	cum_rewards	map_size
0	0	-1498.0	310.0	-1498.0	4x12
1	1	-893.0	299.0	-2391.0	4x12
2	2	-52.0	52.0	-2443.0	4x12
3	3	-48.0	48.0	-2491.0	4x12
4	4	-45.0	45.0	-2536.0	4x12
...	...	...	...	...	...
39995	1995	-13.0	13.0	-102033.0	4x12
39996	1996	-13.0	13.0	-102046.0	4x12
39997	1997	-15.0	15.0	-102061.0	4x12
39998	1998	-13.0	13.0	-102074.0	4x12
39999	1999	-15.0	15.0	-102089.0	4x12

Eksperyment 3:





Qtable

```
[[  -1.99992895   -1.99987792   -1.99987793   -1.99991949]
 [  -1.99985888   -1.99975586   -1.99975586   -1.99990275]
 [  -1.99972761   -1.99951172   -1.99951172   -1.99984119]
 [  -1.99945005   -1.99902344   -1.99902344   -1.99963212]
 [  -1.99896332   -1.99804687   -1.99804688   -1.99932922]
 [  -1.99798912   -1.99609375   -1.99609375   -1.99890351]
 [  -1.99571417   -1.9921875    -1.9921875    -1.99799439]
 [  -1.99145251   -1.984375     -1.984375     -1.99560304]
 [  -1.98279157   -1.96875      -1.96875      -1.99200894]
 [  -1.9666849    -1.9375       -1.9375       -1.98067256]
 [  -1.93282019   -1.875        -1.875        -1.96640919]
 [  -1.86616634   -1.86101123   -1.75         -1.92456297]
 [  -1.99993879   -1.99975586   -1.99975586   -1.99987793]
 [  -1.99987793   -1.99951172   -1.99951172   -1.99987793]
 [  -1.99975586   -1.99902344   -1.99902344   -1.99975586]
 [  -1.99951172   -1.99804688   -1.99804688   -1.99951172]
 [  -1.99902344   -1.99609375   -1.99609375   -1.99902344]
 [  -1.99804687   -1.9921875    -1.9921875    -1.99804687]
 [  -1.99609375   -1.984375     -1.984375     -1.99609375]
 [  -1.9921875    -1.96875      -1.96875      -1.9921875 ]
 [  -1.984375     -1.9375       -1.9375       -1.984375  ]
 [  -1.96875      -1.875        -1.875        -1.96875   ]
 [  -1.9375       -1.75         -1.75         -1.93749996]
 [  -1.87499896   -1.75         -1.5          -1.875     ]
 [  -1.99987793   -1.99951172   -1.99987793   -1.99975586]
 [  -1.99975586   -1.99902344 -100.99987793   -1.99975586]
 [  -1.99951172   -1.99804688 -100.99987793   -1.99951172]
 [  -1.99902344   -1.99609375 -100.99987793   -1.99902344]
 [  -1.99804688   -1.9921875  -100.99987793   -1.99804688]
 [  -1.99609375   -1.984375   -100.99987793   -1.99609375]
 [  -1.9921875    -1.96875    -100.99987793   -1.9921875 ]
 [  -1.984375     -1.9375     -100.99987793   -1.984375  ]
 [  -1.96875      -1.875      -100.99987793   -1.96875   ]
 [  -1.9375       -1.75       -100.99987793   -1.9375    ]
 [  -1.875        -1.5        -100.99987793   -1.875     ]
 [  -1.75         -1.5          -1.           -1.75      ]
 [  -1.99975586 -100.99987793   -1.99987793   -1.99987793]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]
 [   0.            0.            0.            0.        ]]
```

Cumulative Rewards

	Episodes	Rewards	Steps	cum_rewards	map_size
0	0	-1807.0	520.0	-1807.0	4x12
1	1	-27.0	27.0	-1834.0	4x12
2	2	-215.0	116.0	-2049.0	4x12
3	3	-179.0	80.0	-2228.0	4x12
4	4	-48.0	48.0	-2276.0	4x12
...	...	...	...	...	...
39995	1995	-130.0	31.0	-221996.0	4x12
39996	1996	-229.0	31.0	-222225.0	4x12
39997	1997	-13.0	13.0	-222238.0	4x12
39998	1998	-17.0	17.0	-222255.0	4x12
39999	1999	-15.0	15.0	-222270.0	4x12

Obserwacje i wnioski

1. Stan 35 – w którym odnotować można najwięcej wykonanych kroków/akcji podczas treningu to stan początkowy, w którym gracz/uczeń pojawia się zaczynając epizod, natomiast stan 24, który ma równie wysoką liczbę odwiedzin to jedyny stan do którego można przejść ze stanu początkowego (nie wpadając z klifu), są to dwa kroki które uczeń musi zrobić by kontynuować naukę. Najczęściej odwiedzane stany są w przedziale 24-36, czyli stany bliskie klifu – reprezentujące najlepszą ścieżkę. Pokazuje to że uczeń w trakcie treningu uczył się najlepszej ścieżki i eksploatował jej możliwości z niewielką eksploracją pozostałych miejsc.
2. Również najczęściej podejmowane przez ucznia akcje to pójście w prawo, co jest zgodne z wizualizacją na heatmap (średnio w większości stanów z oceny środowiska wychodzi, że największą wygraną (tu: jako najmniejszą stratę) zapewni pójście w lewo)
3. Współczynnik gamma (w kodzie: discount_factor) wpływa na to jak istotny w ocenie wybranej akcji/przejścia ze stanu s0 -> s1 jest nagroda jaką można dostać robiąc krok s1->s2. W skrócie powinna być gamma != 1 – jeśli gamma byłaby równa 1 to algorytm nie brałby w ogóle pod uwagę przy wyborze przejścia przyszłych możliwych nagród, z kolei nawet przy dosyć niskim współczynniku gamma (jak gamma=0.5) a odpowiednio dobranych pozostałych hiperparametrach algorytm w niewielkim stopniu modyfikuje q-table (Eksperyment 3) a jest w stanie nauczyć się najelpszej strategii.
4. Algorytm Q-learning jest dobrym wyborem w przypadku rozwiązywania niezbyt skomplikowanych problemów w prostych środowiskach – np. w takim jak Cliff Walking, dodatkowo w małych środowiskach nauka jest bardzo szybka. Uczeń ‘zapamiętuje’ najlepszą ścieżkę (jako wartości qtable) i podczas każdego epizodu/epoki modyfikuje ją na podstawie nowych ‘obserwacji’ w ten sposób wyznaczając optymalne rozwiązanie. Natomiast algorytm nie dałby dobrych rezultatów w bardziej złożonych problemach – gdzie np. środowisko modyfikuje się w każdej epoce.