WSI

Ćwiczenie 6 – uczenie się ze wzmocnieniem

Prowadzący: mgr inż. Mikołaj Markiewicz

Wykonał: Jan Kaniuka

Numer indeksu: 303762

Treść zadania – zaimplementować algorytm *Q-learning*

Zebrać i przedstawić na wykresie liczbę wykonanych kroków i naliczoną karę/nagrodę w kolejnych epokach.

Problem do rozwiązania **to znalezienie drogi z punktu 'S' do punktu 'F' w "labiryncie" / świecie z przeszkodami.** Rezultatem działania algorytmu powinna być ścieżka w postaci: **(1,1)->(0,1)->...->(2,3)** oraz ww. wykres.

Przykładowe mapy powinny być czytane na starcie programu z jakiegoś formatu np. *ASCI*.

Założenia:

- do wyboru akcji zastosowano **strategię ε-zachłanną**
- nagroda dla celu wynosi +100, dla ściany (przeszkody) -100, a dla elementów ścieżki labiryntu -1 (jest ujemna, aby agent szukał najkrótszej ścieżki *maksymalizując* sumaryczną nagrodę)

Metoda testowania rozwiązania:

Rozwiązanie testowano na różnych labiryntach przygotowanych przez generator online (*https://www.dcode.fr/maze-generator*). Program działa dla labiryntów, gdzie '**#**' oznacza ścianę, a '**.**' wolny obszar. Po wygenerowaniu labiryntu należy kliknąć przycisk *copy* (pierwsza ikona na prawo od *Results*) i wkleić labirynt do pustego pliku w formacie *.txt* oraz zamienić dwa dowolne znaki na **S** oraz **F**. Program sam usunie ostatnią (pustą) linię oraz znaki końca linii i utworzy macierz nagród.

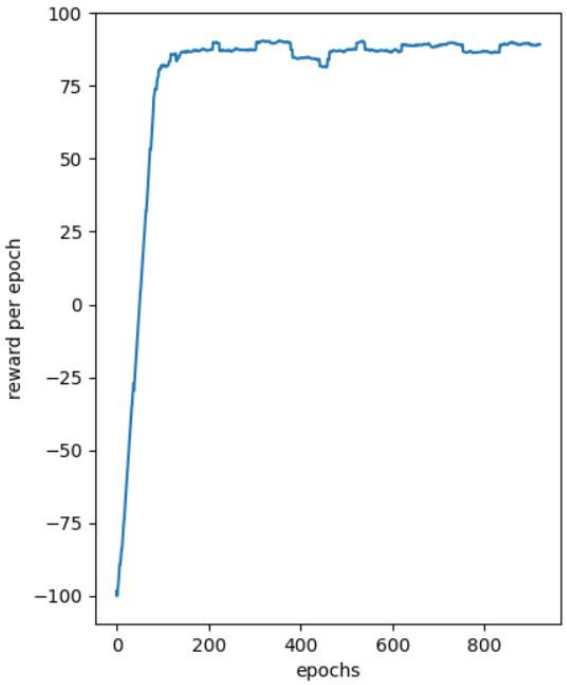
Raport z przeprowadzonych eksperymentów:

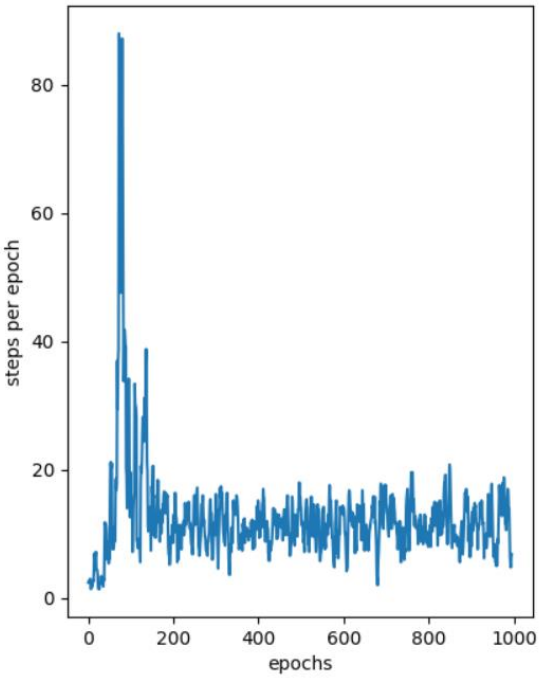
Program znajduje ścieżkę od punktu **S** do **F**. Agent uczy się metodą „prób i błędów" jak znaleźć najkrótszą drogę do celu i nie uderzać w ściany.

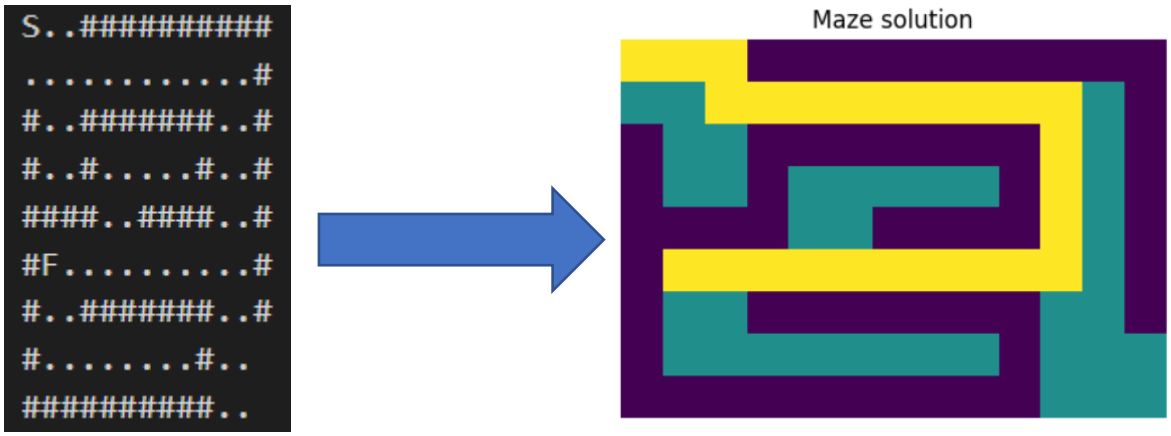
- Na początku widoczny jest wzrost wartości sumy nagród w danej epoce. Liczba kroków również jest największa na początku działania algorytmu – agent dokonuje **eksploracji** i zdobywa wiedzę o konsekwencjach różnych akcji.
- Potem wartość sumy nagród oraz liczby wykonanych kroków stabilizuje się (dalsze, niewielkie zmiany wartości obu wskaźników są spowodowane niezerową wartością

parametru *epsilon*, przez co czasami wybierana jest losowa, a nie najlepsza akcja). Na tym etapie następuje **eksploatacja** – agent dociera do stanów, w których łatwo o wysokie nagrody.

Parametry eksperymentu: $\gamma = 0.9,\ ß = 0.8\ , \varepsilon = 0.001\ , epochs = 1000$







Wnioski, obserwacje, spostrzeżenia:

- Algorytm *Q-learning* dobrze sprawdził się w problemie znalezienia ścieżki w labiryncie. Jest intuicyjny i prosty w zrozumieniu.
- Wartość parametru γ (dyskonto) nie może być zbyt niska, aby algorytm nie patrzył jedynie na najbliższą nagrodę – powinien cechować się „*długowzrocznością*”.
- Wzrost wartości parametru ε powoduje zwiększenie zdolności eksploracyjnych, co potencjalnie wydłuża proces uczenia się.