



Министерство науки и высшего образования
Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего образования
"Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)"
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)_____

ОТЧЕТ

Лабораторная работа №5

«Обучение на основе временных различий»

по курсу «Методы машинного обучения»

ИСПОЛНИТЕЛЬ:

группа ИУ5-21М

Гришин И.А.

ФИО

подпись

"__" _____ 2023 г.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

ФИО

подпись

"__" _____ 2023 г.

Цель работы

Ознакомление с базовыми методами обучения с подкреплением на основе временных различий.

Задание

На основе рассмотренного на лекции примера реализуйте следующие алгоритмы:

- SARSA
- Q-обучение
- Двойное Q-обучение

для любой среды обучения с подкреплением (кроме рассмотренной на лекции среды Toy Text / Frozen Lake) из библиотеки Gym (или аналогичной библиотеки).

Выполнение

Для реализации алгоритмов была выбрана среда Taxi из библиотеки Gym. Агент может находиться в 25 позициях, пассажир может находиться в 5 позициях, и 4 позиции для места назначения = $25 \cdot 5 \cdot 4 = 500$ состояний системы.

Текст программы:

```
import numpy as np
import matplotlib.pyplot as plt
import gym
from tqdm import tqdm

# ***** БАЗОВЫЙ АГЕНТ *****

class BasicAgent:
    """
    Базовый агент, от которого наследуются стратегии обучения
    """

    # Наименование алгоритма
    ALGO_NAME = '----'

    def __init__(self, env, eps=0.1):
        # Среда
        self.env = env
        # Размерности Q-матрицы
```

```

self.nA = env.action_space.n
self.nS = env.observation_space.n
#и сама матрица
self.Q = np.zeros((self.nS, self.nA))
# Значения коэффициентов
# Порог выбора случайного действия
self.eps=eps
# Награды по эпизодам
self.episodes_reward = []

def print_q(self):
    print('Вывод Q-матрицы для алгоритма ', self.ALGO_NAME)
    print(self.Q)

def get_state(self, state):
    """
    Возвращает правильное начальное состояние
    """
    if type(state) is tuple:
        # Если состояние вернулось с виде кортежа, то вернуть только номер
состояния
        return state[0]
    else:
        return state

def greedy(self, state):
    """
    <<Жадное>> текущее действие
    Возвращает действие, соответствующее максимальному Q-значению
    для состояния state
    """
    return np.argmax(self.Q[state])

def make_action(self, state):
    """
    Выбор действия агентом
    """
    if np.random.uniform(0,1) < self.eps:
        # Если вероятность меньше eps
        # то выбирается случайное действие
        return self.env.action_space.sample()
    else:
        # иначе действие, соответствующее максимальному Q-значению
        return self.greedy(state)

```

```

def draw_episodes_reward(self):
    # Построение графика наград по эпизодам
    fig, ax = plt.subplots(figsize = (15,10))
    y = self.episodes_reward
    x = list(range(1, len(y)+1))
    plt.plot(x, y, '-', linewidth=1, color='green')
    plt.title('Награды по эпизодам')
    plt.xlabel('Номер эпизода')
    plt.ylabel('Награда')
    plt.show()

def learn():
    """
    Реализация алгоритма обучения
    """
    pass

# ***** SARSA
# *****

class SARSA_Agent(BasicAgent):
    """
    Реализация алгоритма SARSA
    """
    # Наименование алгоритма
    ALGO_NAME = 'SARSA'

def __init__(self, env, eps=0.4, lr=0.1, gamma=0.98, num_episodes=20000):
    # Вызов конструктора верхнего уровня
    super().__init__(env, eps)
    # Learning rate
    self.lr=lr
    # Коэффициент дисконтирования
    self.gamma = gamma
    # Количество эпизодов
    self.num_episodes=num_episodes
    # Постепенное уменьшение eps
    self.eps_decay=0.00005
    self.eps_threshold=0.01

def learn(self):
    """
    Обучение на основе алгоритма SARSA
    """
    self.episodes_reward = []
    # Цикл по эпизодам
    for ep in tqdm(list(range(self.num_episodes))):
        # Начальное состояние среды

```

```

state = self.get_state(self.env.reset())
# Флаг штатного завершения эпизода
done = False
# Флаг нештатного завершения эпизода
truncated = False
# Суммарная награда по эпизоду
tot_rew = 0

# По мере заполнения Q-матрицы уменьшаем вероятность случайного
выбора действия
if self.eps > self.eps_threshold:
    self.eps -= self.eps_decay

# Выбор действия
action = self.make_action(state)

# Проигрывание одного эпизода до финального состояния
while not (done or truncated):

    # Выполняем шаг в среде
    next_state, rew, done, truncated, _ = self.env.step(action)

    # Выполняем следующее действие
    next_action = self.make_action(next_state)

    # Правило обновления Q для SARSA
    self.Q[state][action] = self.Q[state][action] + self.lr * \
        (rew + self.gamma * self.Q[next_state][next_action] -
self.Q[state][action])

    # Следующее состояние считаем текущим
    state = next_state
    action = next_action
    # Суммарная награда за эпизод
    tot_rew += rew
    if (done or truncated):
        self.episodes_reward.append(tot_rew)

# ***** Q-обучение
*****

class QLearning_Agent(BasicAgent):
    ...

    Реализация алгоритма Q-Learning
    ...

    # Наименование алгоритма
    ALGO_NAME = 'Q-обучение'

def __init__(self, env, eps=0.4, lr=0.1, gamma=0.98, num_episodes=20000):
    # Вызов конструктора верхнего уровня

```

```

super().__init__(env, eps)
# Learning rate
self.lr=lr
# Коэффициент дисконтирования
self.gamma = gamma
# Количество эпизодов
self.num_episodes=num_episodes
# Постепенное уменьшение eps
self.eps_decay=0.00005
self.eps_threshold=0.01

def learn(self):
    """
    Обучение на основе алгоритма Q-Learning
    """
    self.episodes_reward = []
    # Цикл по эпизодам
    for ep in tqdm(list(range(self.num_episodes))):
        # Начальное состояние среды
        state = self.get_state(self.env.reset())
        # Флаг штатного завершения эпизода
        done = False
        # Флаг нештатного завершения эпизода
        truncated = False
        # Суммарная награда по эпизоду
        tot_rew = 0

        # По мере заполнения Q-матрицы уменьшаем вероятность случайного
        # выбора действия
        if self.eps > self.eps_threshold:
            self.eps -= self.eps_decay

        # Проигрывание одного эпизода до финального состояния
        while not (done or truncated):

            # Выбор действия
            # В SARSA следующее действие выбиралось после шага в среде
            action = self.make_action(state)

            # Выполняем шаг в среде
            next_state, rew, done, truncated, _ = self.env.step(action)

            # Правило обновления Q для SARSA (для сравнения)
            # self.Q[state][action] = self.Q[state][action] + self.lr * \
            #     (rew + self.gamma * self.Q[next_state][next_action] -
            self.Q[state][action])

            # Правило обновления для Q-обучения
            self.Q[state][action] = self.Q[state][action] + self.lr * \

```

```

        (rew + self.gamma * np.max(self.Q[next_state]) -
self.Q[state][action])

        # Следующее состояние считаем текущим
        state = next_state
        # Суммарная награда за эпизод
        tot_rew += rew
        if (done or truncated):
            self.episodes_reward.append(tot_rew)

# ***** Двойное Q-обучение
*****

class DoubleQLearning_Agent(BasicAgent):
    """
    Реализация алгоритма Double Q-Learning
    """
    # Наименование алгоритма
    ALGO_NAME = 'Двойное Q-обучение'

    def __init__(self, env, eps=0.4, lr=0.1, gamma=0.98, num_episodes=20000):
        # Вызов конструктора верхнего уровня
        super().__init__(env, eps)
        # Вторая матрица
        self.Q2 = np.zeros((self.nS, self.nA))
        # Learning rate
        self.lr=lr
        # Коэффициент дисконтирования
        self.gamma = gamma
        # Количество эпизодов
        self.num_episodes=num_episodes
        # Постепенное уменьшение eps
        self.eps_decay=0.00005
        self.eps_threshold=0.01

    def greedy(self, state):
        """
        <<Жадное>> текущее действие
        Возвращает действие, соответствующее максимальному Q-значению
        для состояния state
        """
        temp_q = self.Q[state] + self.Q2[state]
        return np.argmax(temp_q)

    def print_q(self):
        print('Вывод Q-матриц для алгоритма ', self.ALGO_NAME)
        print('Q1')
        print(self.Q)

```

```

print('Q2')
print(self.Q2)

def learn(self):
    """
    Обучение на основе алгоритма Double Q-Learning
    """
    self.episodes_reward = []
    # Цикл по эпизодам
    for ep in tqdm(list(range(self.num_episodes))):
        # Начальное состояние среды
        state = self.get_state(self.env.reset())
        # Флаг штатного завершения эпизода
        done = False
        # Флаг нештатного завершения эпизода
        truncated = False
        # Суммарная награда по эпизоду
        tot_rew = 0

        # По мере заполнения Q-матрицы уменьшаем вероятность случайного
        # выбора действия
        if self.eps > self.eps_threshold:
            self.eps -= self.eps_decay

        # Проигрывание одного эпизода до финального состояния
        while not (done or truncated):

            # Выбор действия
            # В SARSA следующее действие выбиралось после шага в среде
            action = self.make_action(state)

            # Выполняем шаг в среде
            next_state, rew, done, truncated, _ = self.env.step(action)

            if np.random.rand() < 0.5:
                # Обновление первой таблицы
                self.Q[state][action] = self.Q[state][action] + self.lr * \
                    (rew + self.gamma *
self.Q2[next_state][np.argmax(self.Q[next_state])] - self.Q[state][action])
            else:
                # Обновление второй таблицы
                self.Q2[state][action] = self.Q2[state][action] + self.lr * \
                    (rew + self.gamma *
self.Q[next_state][np.argmax(self.Q2[next_state])] - self.Q2[state][action])

            # Следующее состояние считаем текущим
            state = next_state
            # Суммарная награда за эпизод
            tot_rew += rew
            if (done or truncated):

```



```
self.episodes_reward.append(tot_rew)
```

```
def play_agent(agent):  
    ...  
    Проигрывание сессии для обученного агента  
    ...  
    env2 = gym.make('Taxi-v3', render_mode='human')  
    state = env2.reset()[0]  
    done = False  
    while not done:  
        action = agent.greedy(state)  
        next_state, reward, terminated, truncated, _ = env2.step(action)  
        env2.render()  
        state = next_state  
        if terminated or truncated:  
            done = True
```

```
def run_sarsa():  
    env = gym.make('Taxi-v3')  
    agent = SARSA_Agent(env)  
    agent.learn()  
    agent.print_q()  
    agent.draw_episodes_reward()  
    play_agent(agent)
```

```
def run_q_learning():  
    env = gym.make('Taxi-v3')  
    agent = QLearning_Agent(env)  
    agent.learn()  
    agent.print_q()  
    agent.draw_episodes_reward()  
    play_agent(agent)
```

```
def run_double_q_learning():  
    env = gym.make('Taxi-v3')  
    agent = DoubleQLearning_Agent(env)  
    agent.learn()  
    agent.print_q()  
    agent.draw_episodes_reward()  
    play_agent(agent)
```

```
def main():  
    run_sarsa()  
    #run_q_learning()  
    #run_double_q_learning()
```

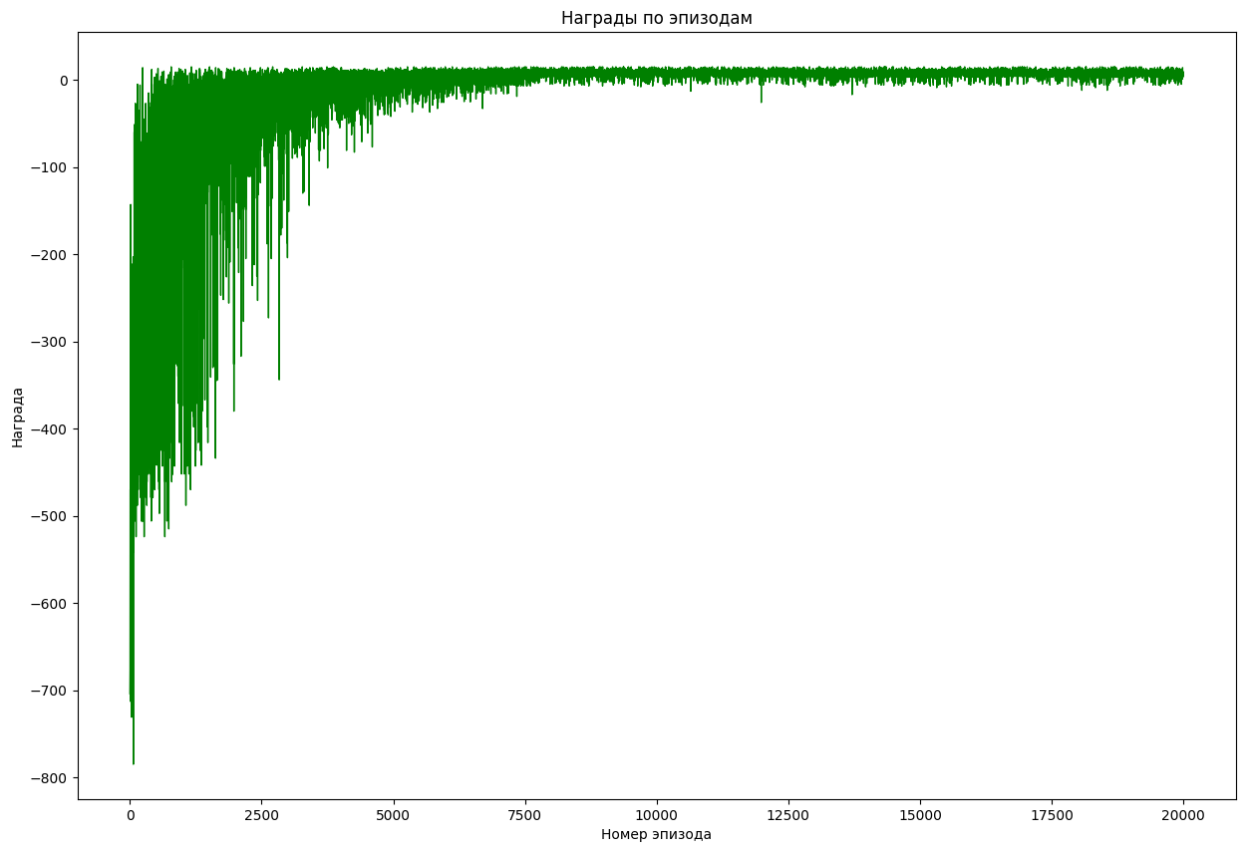
```

if __name__ == '__main__':
    main()

```

Результат выполнения

Алгоритм SARSA



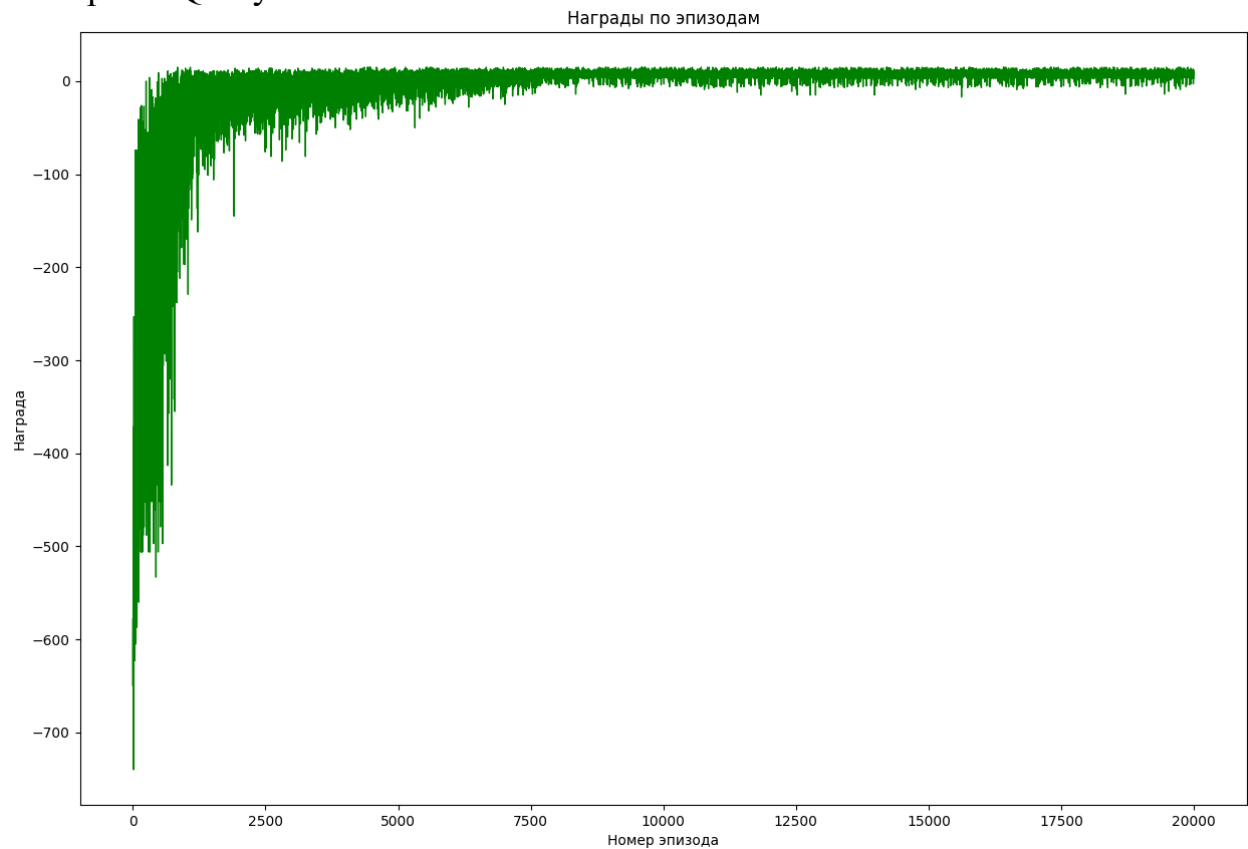
Вывод Q-матрицы для алгоритма SARSA

```

[[ 0.      0.      0.      0.      0.
  0.      ]
 [ -7.8898246 -6.24405781 -9.03233281 -5.11763441  7.92937842
 -12.84820021]
 [  3.47463823  2.01447102  2.14629857  6.01424493 12.99601202
 -1.52498709]
 ...
 [ -1.49036153  9.94564874 -1.97350244 -2.58223539 -7.82846602
 -7.99826697]
 [ -8.32978289 -4.06843975 -8.80247459 -8.46661662 -11.05405831
 -13.72079767]
 [  3.21273673  5.16601841  3.62966677 18.40293968 -0.98907943
 -2.58631448]]

```

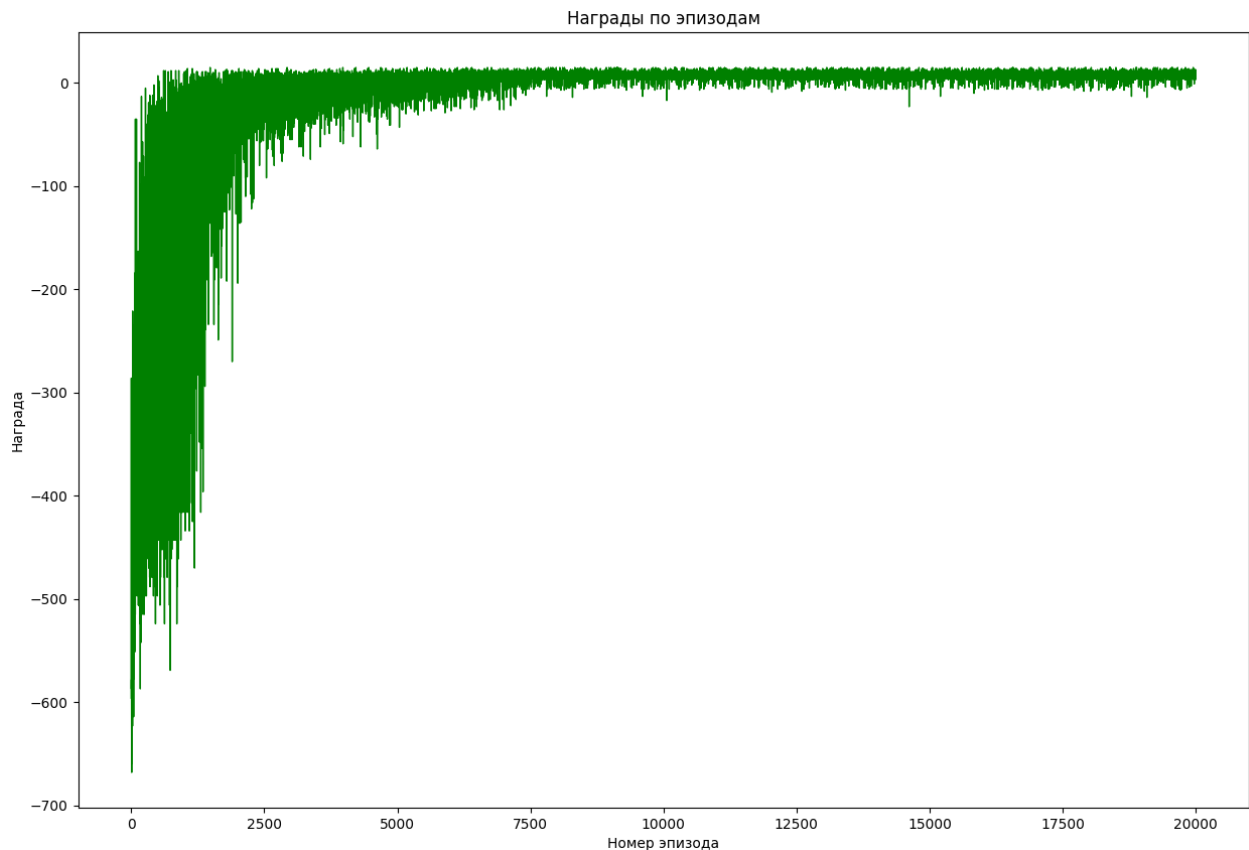
Алгоритм Q-обучение



Вывод Q-матрицы для алгоритма Q-обучение

```
[[ 0.          0.          0.          0.          0.          0.        ]
 [ 5.37044847  5.14514048  4.09492199  6.18245122  8.36234335 -2.63513937]
 [ 9.50426768 10.92607943 10.15614952 11.55273662 13.27445578  2.23687016]
 ...
 [-0.92334665 -0.4050141  -1.27313511 10.19935743 -3.97012192 -4.24053095]
 [ 0.43797046 -2.1744967  -1.51085877  9.33466713 -6.69681892 -7.41177538]
 [ 6.672264    6.41626915  6.56620228 18.59890138  1.40850918 -0.31545857]]
```

Алгоритм двойное Q-обучение



Вывод Q-матриц для алгоритма Двойное Q-обучение

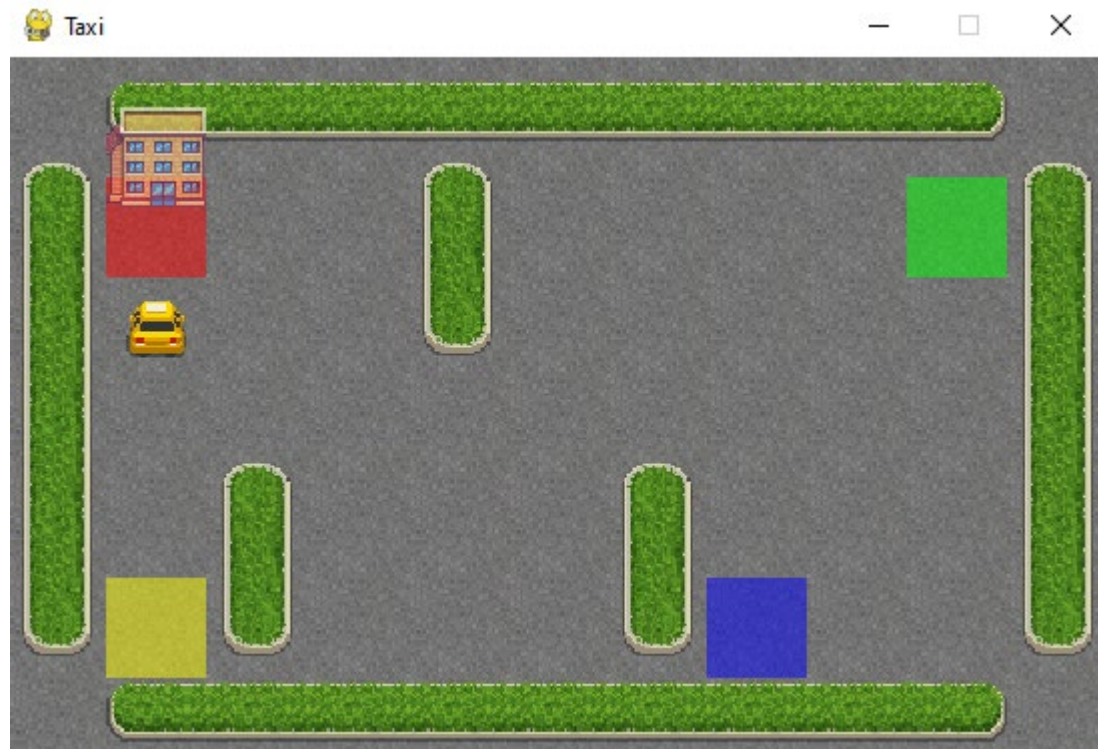
Q1

```
[[ 0.          0.          0.          0.          0.          0.         ]
 [-0.42674465  0.83654428 -1.44580562  1.49689447  8.36234335 -5.98640953]
 [ 2.53545344  3.72058179  2.66382996  8.13718952 13.27445578 -1.51145122]
 ...
 [-1.85128628  7.88112715 -1.41493144 -0.82158236 -1.89008713 -3.06191132]
 [-3.57644351 -3.30271779 -4.76477537  6.13869782 -4.87006677 -1.0098      ]
 [ 1.85095065  4.23843903  4.87884128 18.47786944 -0.15762315 -2.12408469]]
```

Q2

```
[[ 0.          0.          0.          0.          0.          0.         ]
 [ 2.8875773   1.26720629 -0.36970522  3.16735372  8.36234335 -6.45640768]
 [ 7.92224642  6.93994222  2.4851841   7.55510701 13.27445578  0.11459421]
 ...
 [-1.34521224  7.72763929 -1.67984498 -1.47400353 -5.05250003 -3.15322683]
 [-1.2971344  -3.57668889 -2.8353877   6.19967041 -7.67198904 -6.73399531]
 [ 4.32346577  1.9971451   1.90020708 18.31693341  0.75820926 -0.6166202  ]]
```

Пример работы агента:



Вывод

В ходе выполнения лабораторной работы мы ознакомились с базовыми методами обучения с подкреплением на основе временных различий, а именно алгоритмами SARSA, Q-Learning, Double Q-Learning.