Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Рубежный контроль №2

по дисциплине

«Методы машинного обучения»

Выполнил:

Студент группы ИУ5-21М

Королёва Ю.П.

Проверил:

Гапанюк Ю.Е.

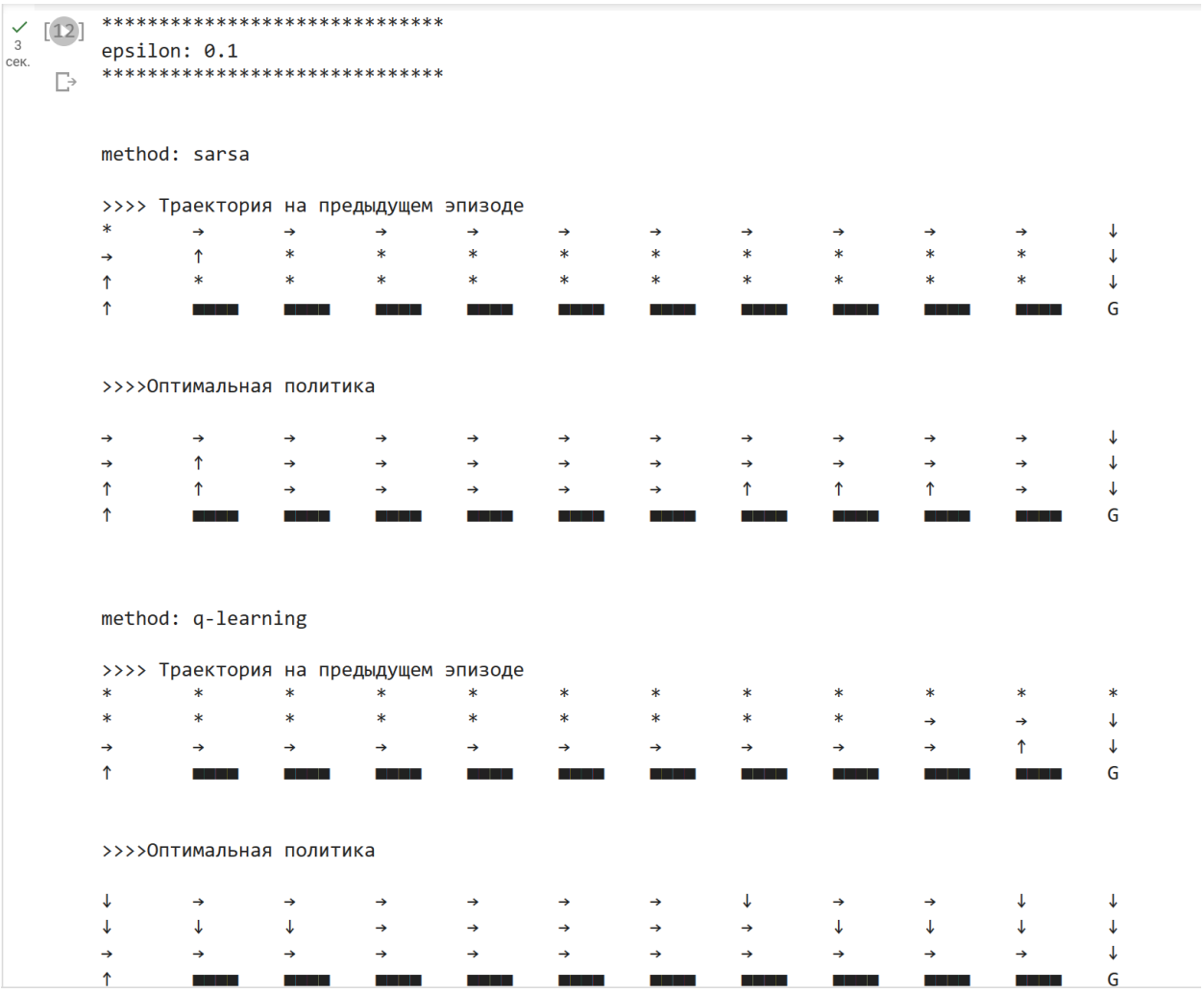
Москва, 2023

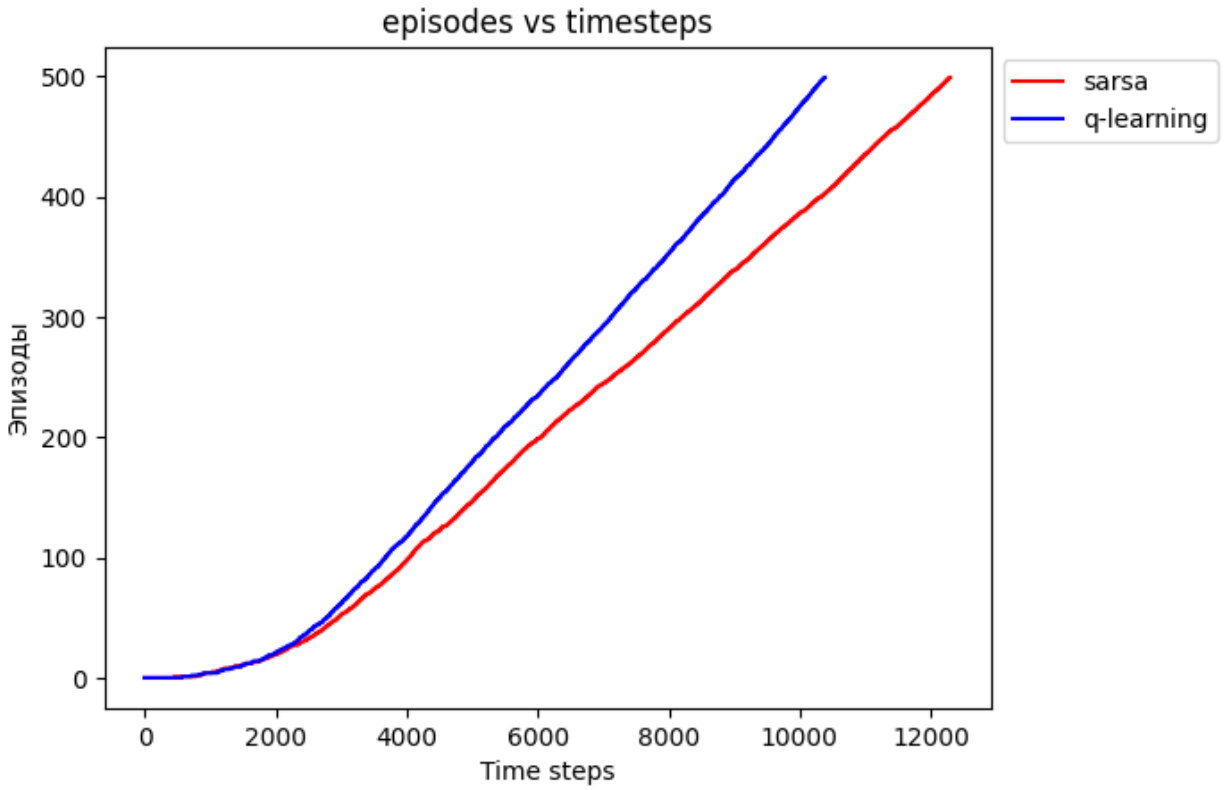
Для одного из алгоритмов временных различий, реализованных Вами в соответствующей лабораторная работе:

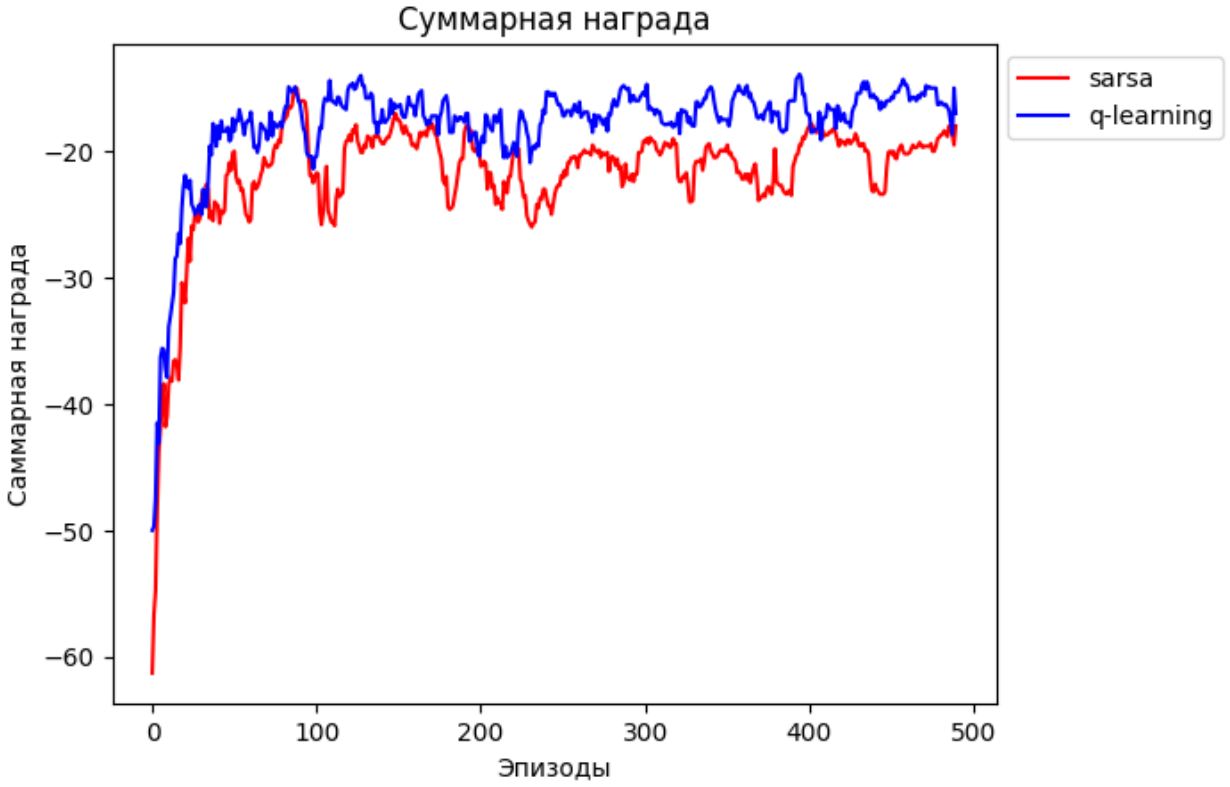
- SARSA
- Q-обучение
- Двойное Q-обучение

осуществите подбор гиперпараметров. Критерием оптимизации должна являться суммарная награда.

Меняя EPSILON=.05, посмотрим на графики.

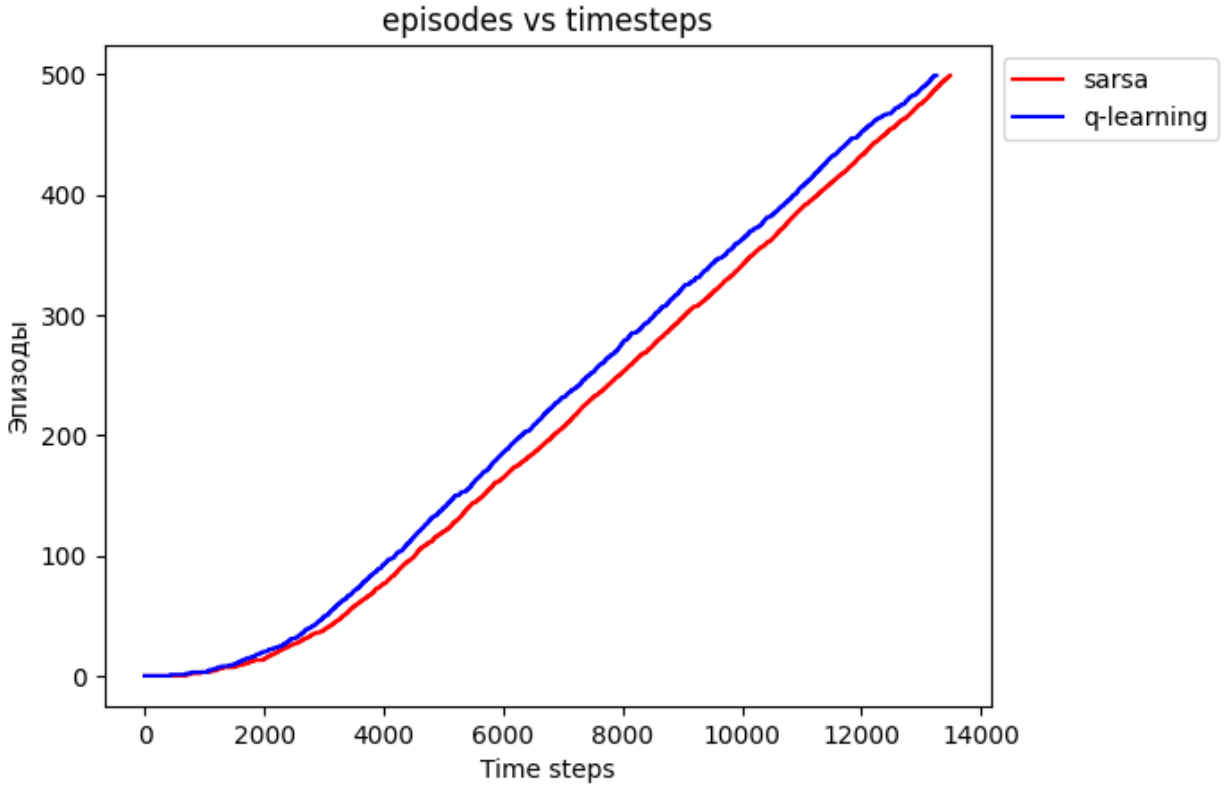



episodes vs timesteps
sarsa
q-learning
Эпизоды
Time steps
0
100
200
300
400
500
0
2000
4000
6000
8000
10000
12000

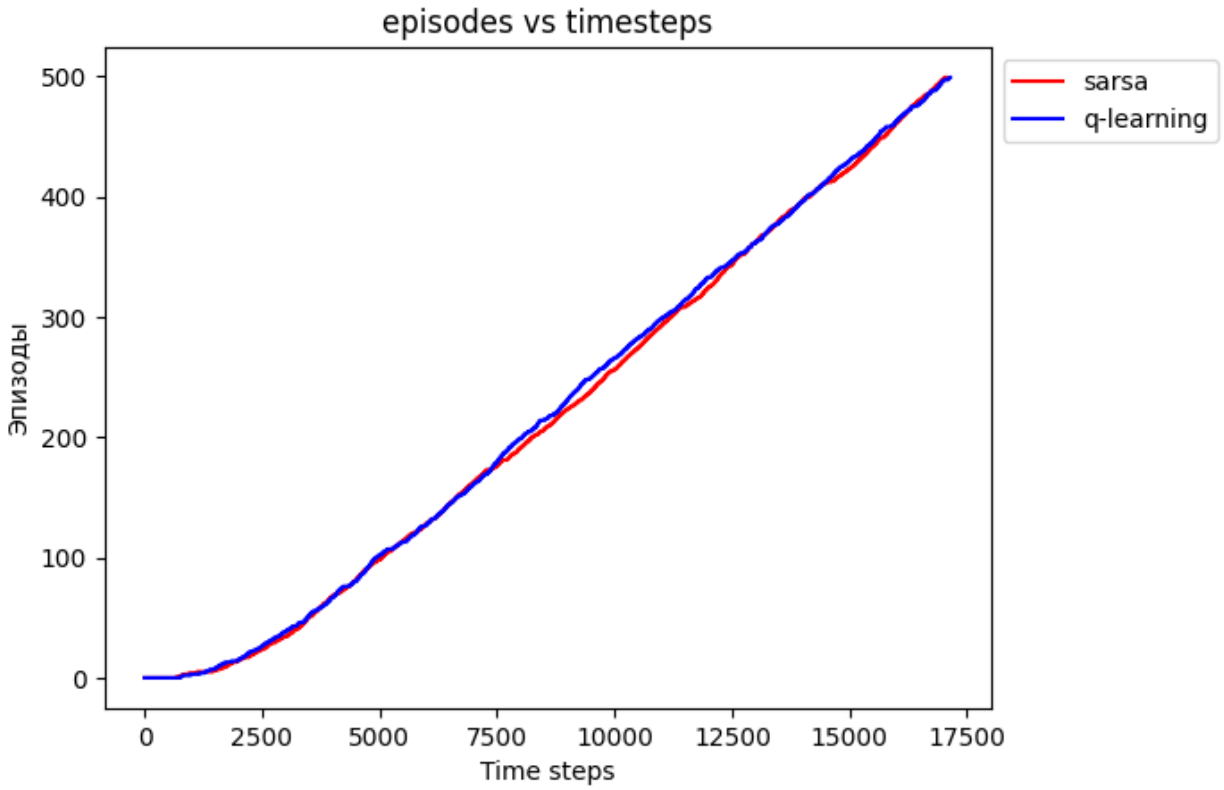

Суммарная награда
sarsa
q-learning
Саммарная награда
Эпизоды
−20
−30
−40
−50
−60
0
100
200
300
400
500

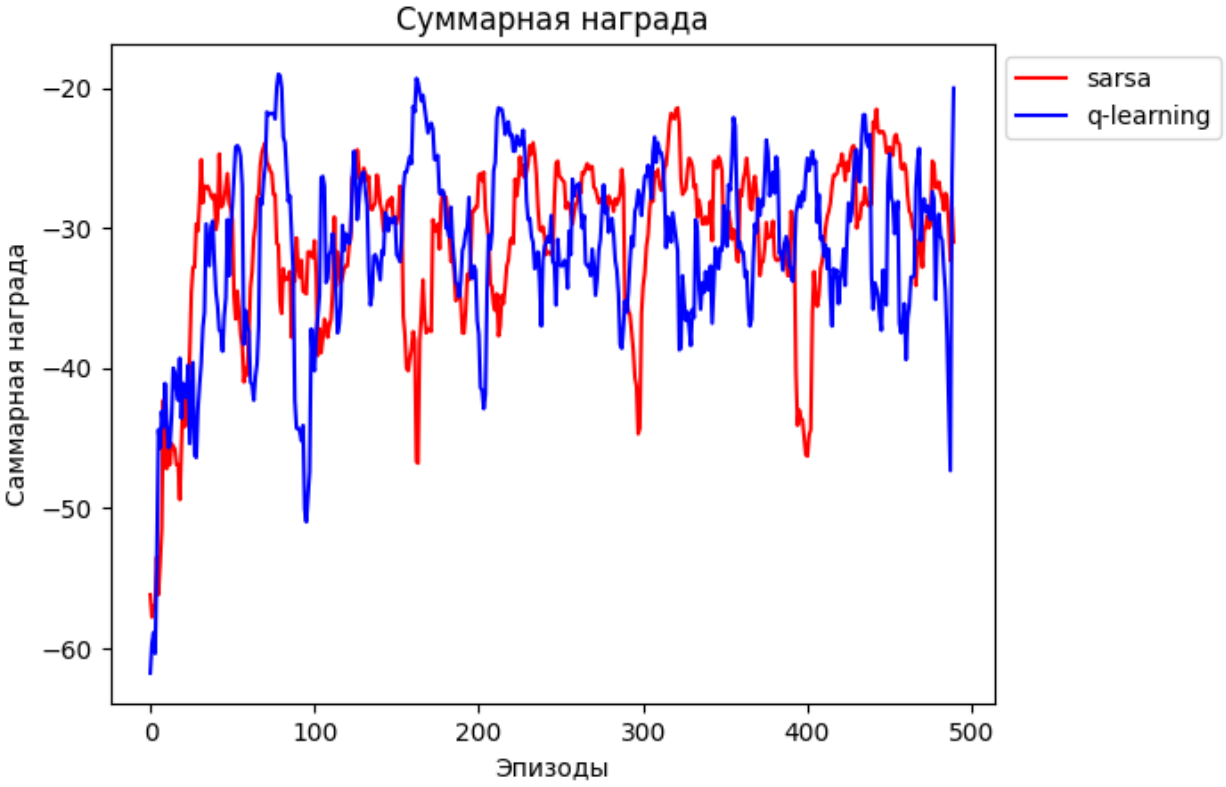
[13] ✓ 3 сек.

```
*****************************
epsilon: 0.2
*****************************


method: sarsa

>>>> Траектория на предыдущем эпизоде
*       *       *       *       *       *       *       *       *       *       *       *
*       →       →       →       →       →       →       →       →       →       ↓       *
→       ↑       *       *       *       *       *       *       *       *       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

→       →       →       →       ↓       ↓       ↓       →       →       ↓       →       ↓
↑       →       →       →       →       →       →       →       →       →       ↓       ↓
→       ↑       →       ↑       ↑       →       →       →       →       ↑       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


method: q-learning

>>>> Траектория на предыдущем эпизоде
*       *       *       *       *       *       *       *       *       *       *       *
*       *       ↓       *       *       *       *       *       *       *       *       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

→       →       ↓       →       →       →       →       ↓       →       ↓       ↓       ↓
↓       →       →       →       →       →       →       →       ↓       →       →       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G
```



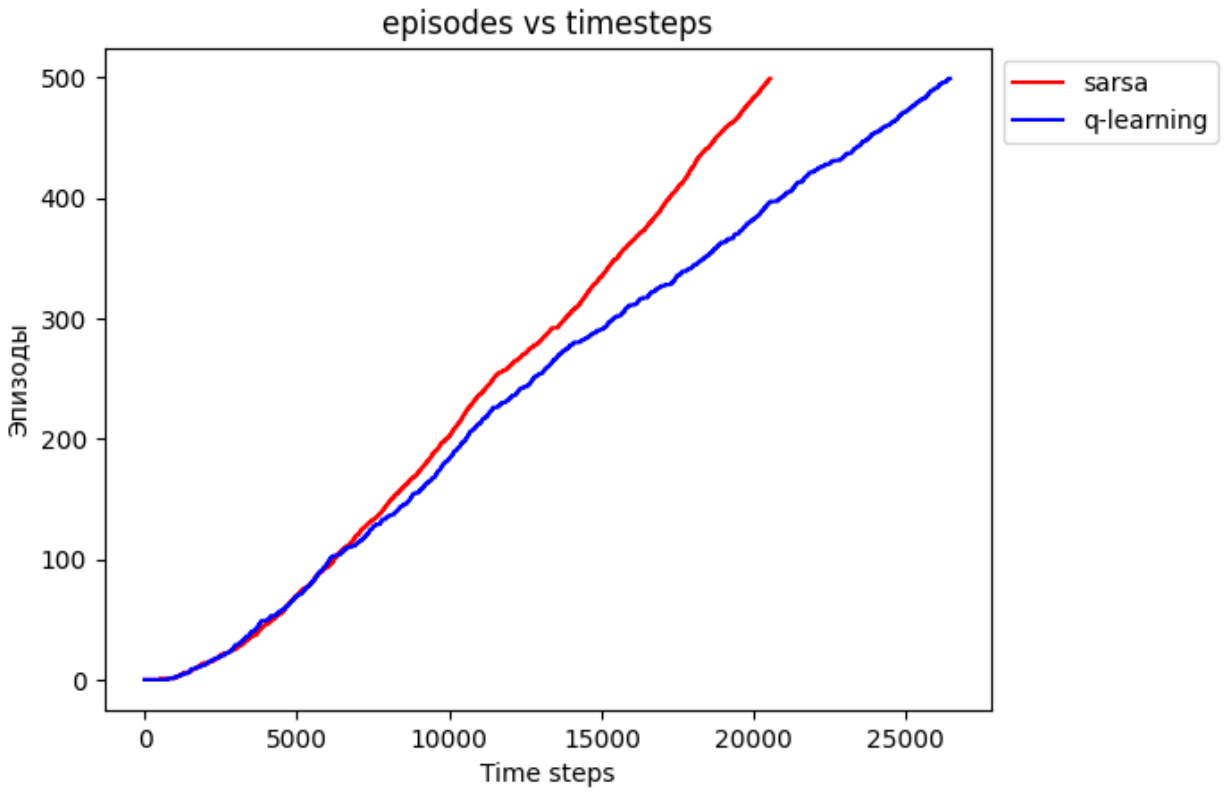

```
**********
[14] epsilon: 0.3
**********


method: sarsa

>>>> Траектория на предыдущем эпизоде
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       *       *       *       *       *       *       *       ↑       *       ↑       ↓
↑       *       *       *       *       *       *       *       *       *       *       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G


>>>>Оптимальная политика

→       →       →       →       →       →       →       →       →       →       ↓       ↓
↑       ↑       →       →       →       →       →       →       ↑       ↑       →       ↓
↑       ↑       ↑       ↑       →       →       ↑       →       ↑       →       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G


method: q-learning

>>>> Траектория на предыдущем эпизоде
*       *       *       *       *       *       *       *       *       *       *       *
*       *       *       ↓       *       *       *       *       *       *       *       *
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G


>>>>Оптимальная политика

→       ↓       →       →       →       →       →       →       →       →       →       ↓
↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G
```


episodes vs timesteps
sarsa
q-learning
Эпизоды
Time steps
0
100
200
300
400
500
0
2500
5000
7500
10000
12500
15000
17500


Суммарная награда
sarsa
q-learning
Саммарная награда
Эпизоды
−20
−30
−40
−50
−60
0
100
200
300
400
500

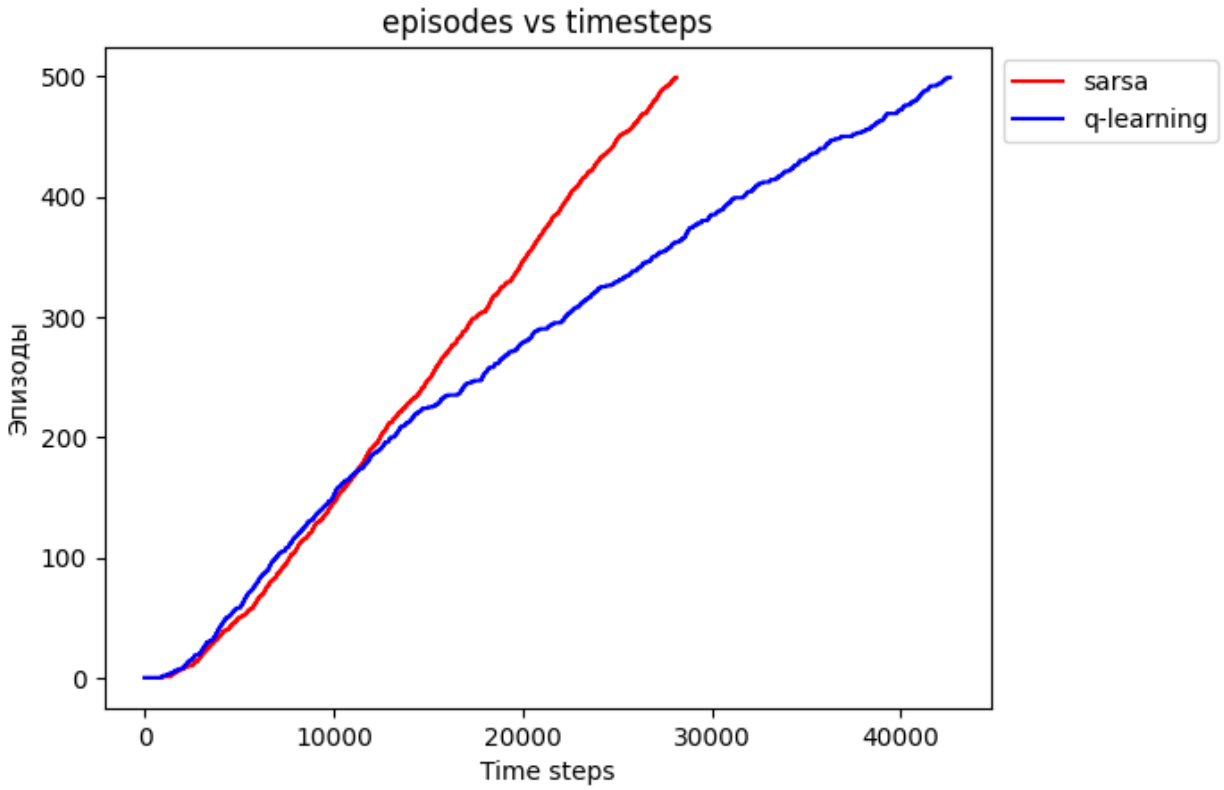
✓ 2 сек. [15]

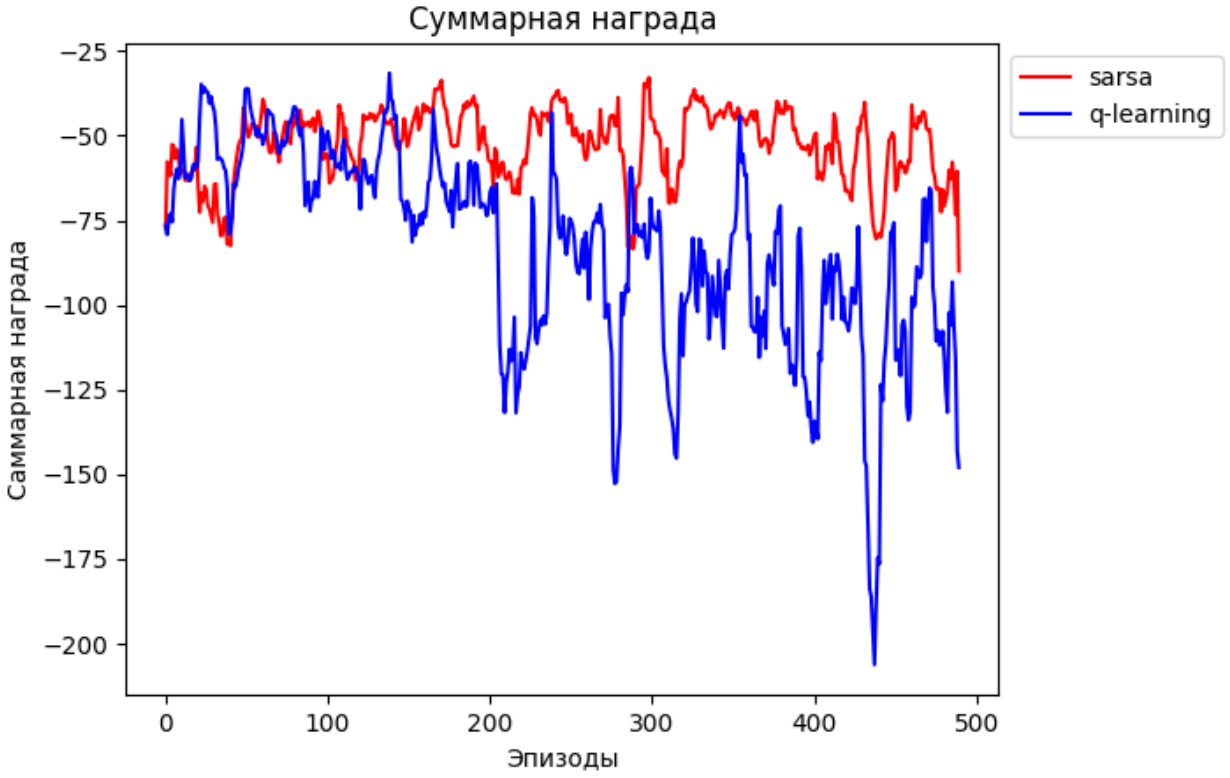
```
epsilon: 0.4
******************************


method: sarsa

>>>> Траектория на предыдущем эпизоде
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       *       *       *       *       *       *       *       *       *       *       ↓
↑       *       *       *       *       *       *       *       *       *       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

→       →       →       →       →       →       →       →       →       →       ↓       ↓
→       →       →       ↑       →       →       →       ↑       →       →       →       ↓
↑       →       ↑       ↑       →       ↑       ↑       ↑       ↑       ↑       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


method: q-learning

>>>> Траектория на предыдущем эпизоде
→       ↓       *       →       →       →       →       →       →       →       →       ↓
↑       ↓       *       *       ↑       *       *       *       *       *       *       ↓
↑       →       →       →       ↑       ↓       *       *       *       *       *       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

↓       →       →       ↓       →       →       ↓       ↓       →       ↓       ↓       ↓
↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G
```



✓ [16] 5 сек.

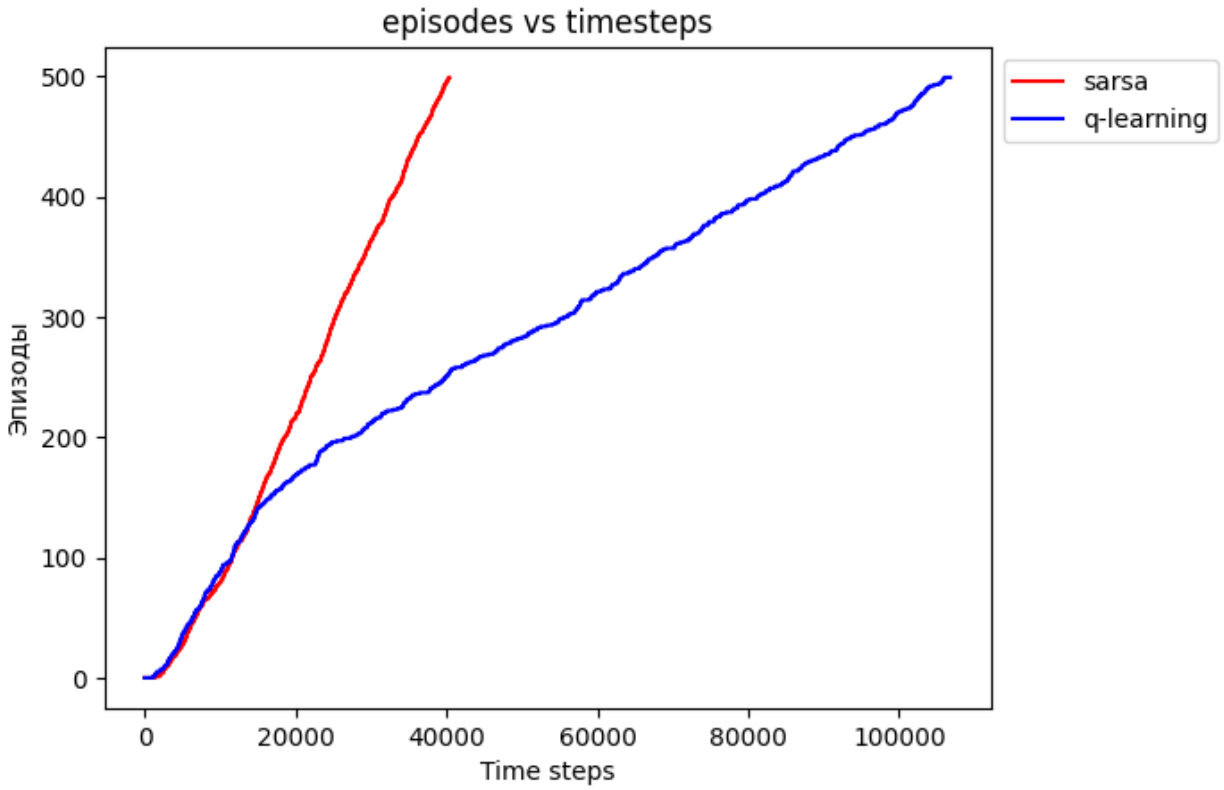
```
******************************
epsilon: 0.5
******************************


method: sarsa

>>>> Траектория на предыдущем эпизоде
→       →       ↓       ←       →       →       →       →       →       ↓       *       ↓
→       →       →       →       ↑       ←       *       ↑       ←       →       →       ↓
↑       *       *       *       ↓       *       *       *       *       *       *       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

↑       →       →       →       →       →       →       →       →       →       →       ↑
↑       →       →       →       ↑       ↑       →       ↑       →       →       →       ↓
↑       →       →       →       ↑       →       ↑       ↑       ↑       ↑       ↑       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


method: q-learning

>>>> Траектория на предыдущем эпизоде
↓       →       →       ↓       ↓       *       *       ↓       *       ↓       ←       ←
↓       ↑       ↓       ↓       ↓       *       ↓       ←       ←       →       →       ↓
→       →       →       →       →       →       →       →       →       →       ↑       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>>Оптимальная политика

↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G
```

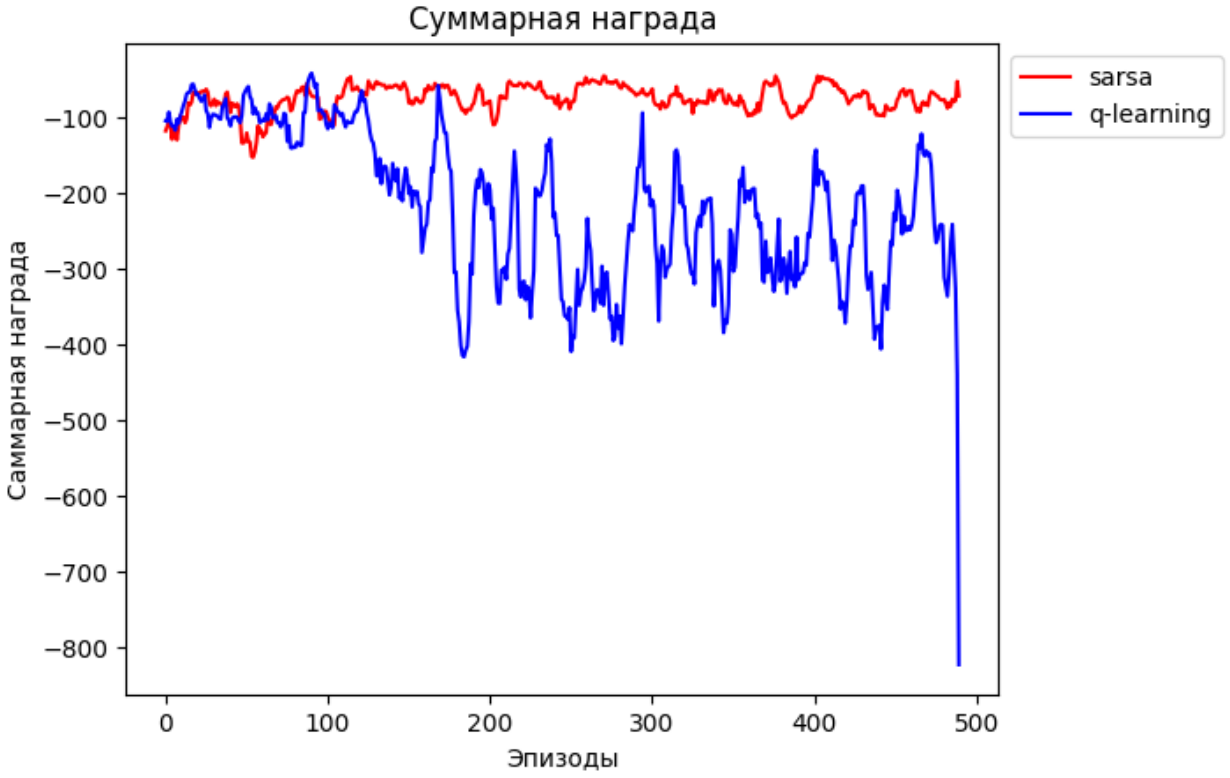

episodes vs timesteps
sarsa
q-learning
Эпизоды
Time steps
0
100
200
300
400
500
0
10000
20000
30000
40000


Суммарная награда
sarsa
q-learning
Саммарная награда
Эпизоды
−25
−50
−75
−100
−125
−150
−175
−200
0
100
200
300
400
500

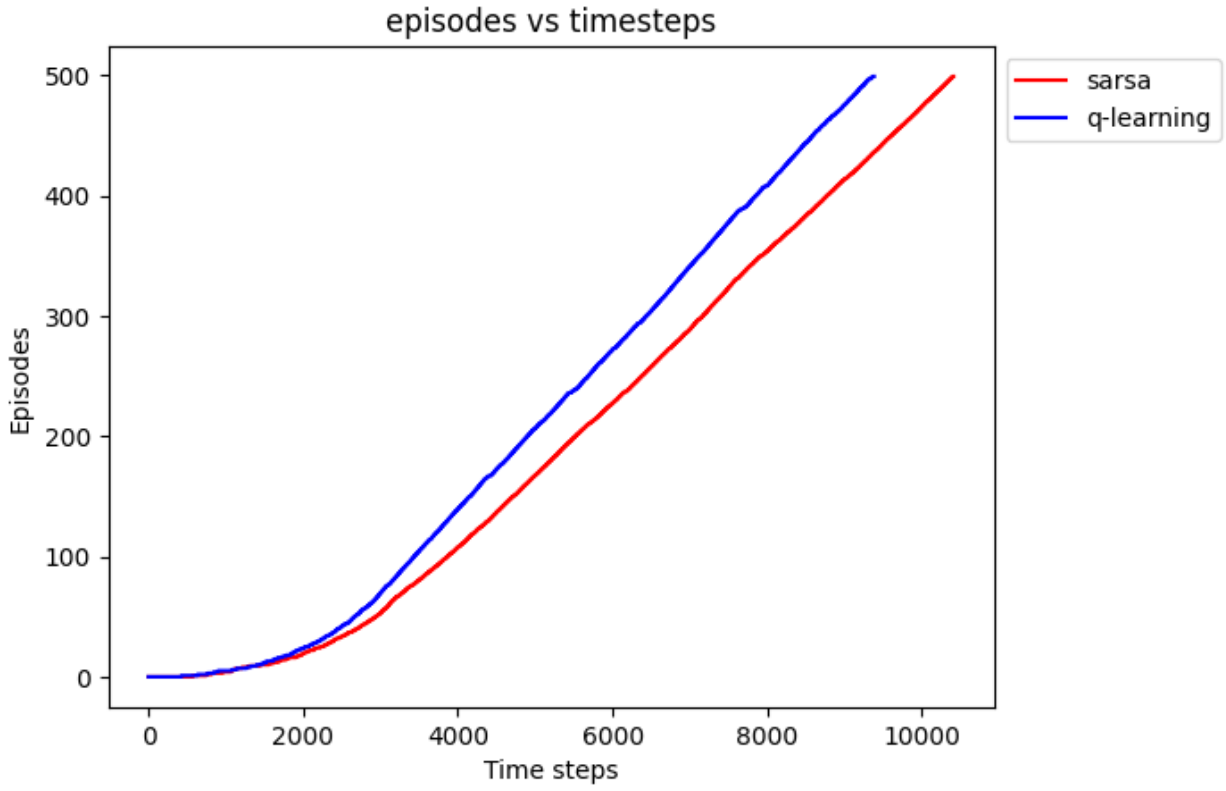
[17] ✓ 4 сек.

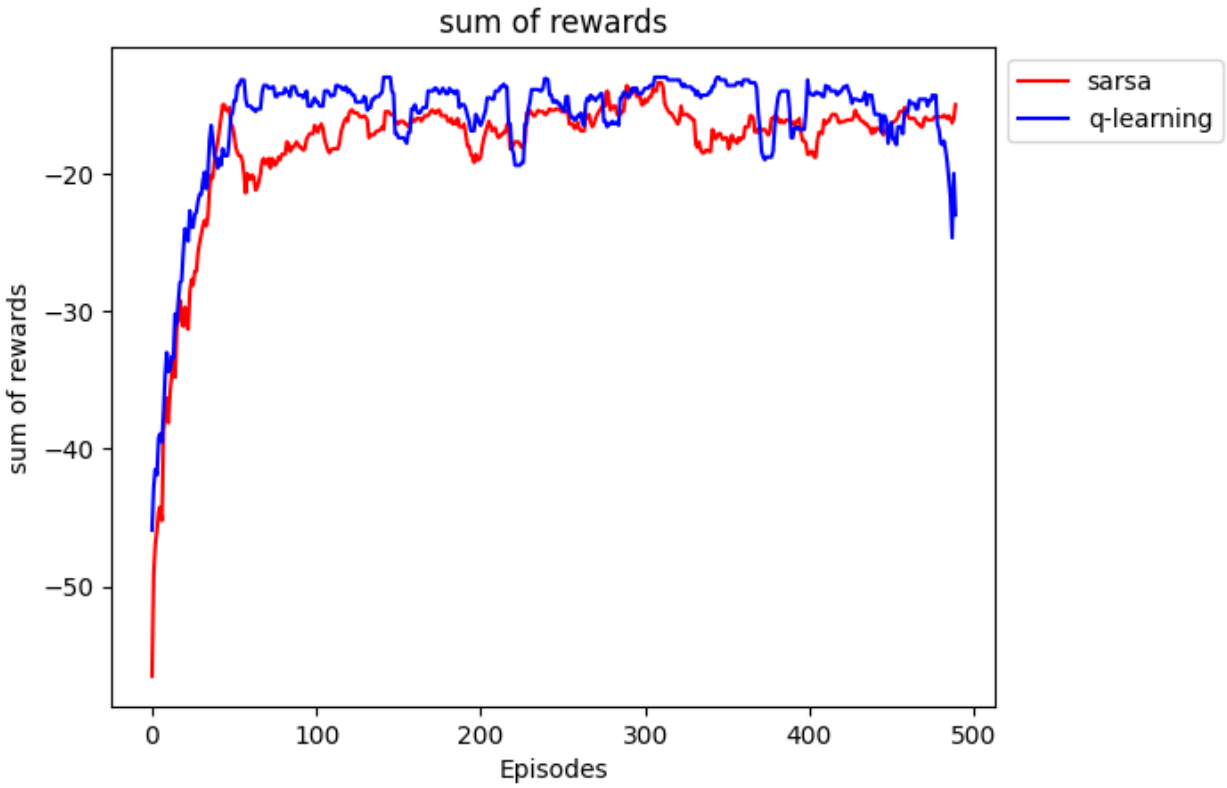
```
******************************
epsilon: 0.6
******************************


method: sarsa

>>>> Траектория на предыдущем эпизоде
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ←       *       ↑       *       *       →       ↑       *       *       →       ↓
↑       ←       →       ↓       *       *       *       *       *       *       *       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G


>>>>Оптимальная политика

→       →       ↑       →       →       →       ↑       →       →       →       →       ↑
→       →       ↑       →       →       →       ↑       ↑       →       →       →       ↓
↑       ↑       ↑       ↑       →       ↑       →       ↑       ↑       ↑       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G



method: q-learning

>>>> Траектория на предыдущем эпизоде
→       →       ↓       ↓       ↓       ↓       →       ↓       *       *       *       *
↑       ↓       →       →       ↓       ↓       ↓       ↓       ↓       *       *       *
→       →       ↑       ↑       →       →       →       →       →       →       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G


>>>>Оптимальная политика

↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   ■■■■■   G
```


episodes vs timesteps
Эпизоды
Time steps
sarsa
q-learning
0
100
200
300
400
500
0
20000
40000
60000
80000
100000


Суммарная награда
Саммарная награда
Эпизоды
sarsa
q-learning
−100
−200
−300
−400
−500
−600
−700
−800
0
100
200
300
400
500

```
epsilon: 0.05
******************************


method: sarsa

>>>> trajectory of the latest episode
*       *       *       *       *       *       *       *       *       *       *       *
→       →       →       →       →       →       →       →       →       →       ↓       *
↑       *       *       *       *       *       *       *       *       *       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>> optimal policy

→       →       →       →       →       ↓       →       →       →       ↓       →       ↓
→       →       →       →       →       →       →       →       →       →       ↓       ↓
↑       ↑       →       →       ↑       ←       ↑       ↑       →       ↑       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


method: q-learning

>>>> trajectory of the latest episode
*       *       *       *       *       *       *       *       *       *       *       *
*       *       *       *       *       *       *       *       ↓       *       *       *
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G


>>>> optimal policy

↑       →       →       ↑       →       →       ↓       →       ↑       →       ↓       ↓
→       →       →       →       ↑       →       →       →       →       ↓       ↓       ↓
→       →       →       →       →       →       →       →       →       →       →       ↓
↑       ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    ▬▬▬▬    G
```





Вывод:

Наилучший гиперпараметр EPSILON=.05. Дает максимальную суммарную награду. При данном параметре получается оптимальная стратегия.