

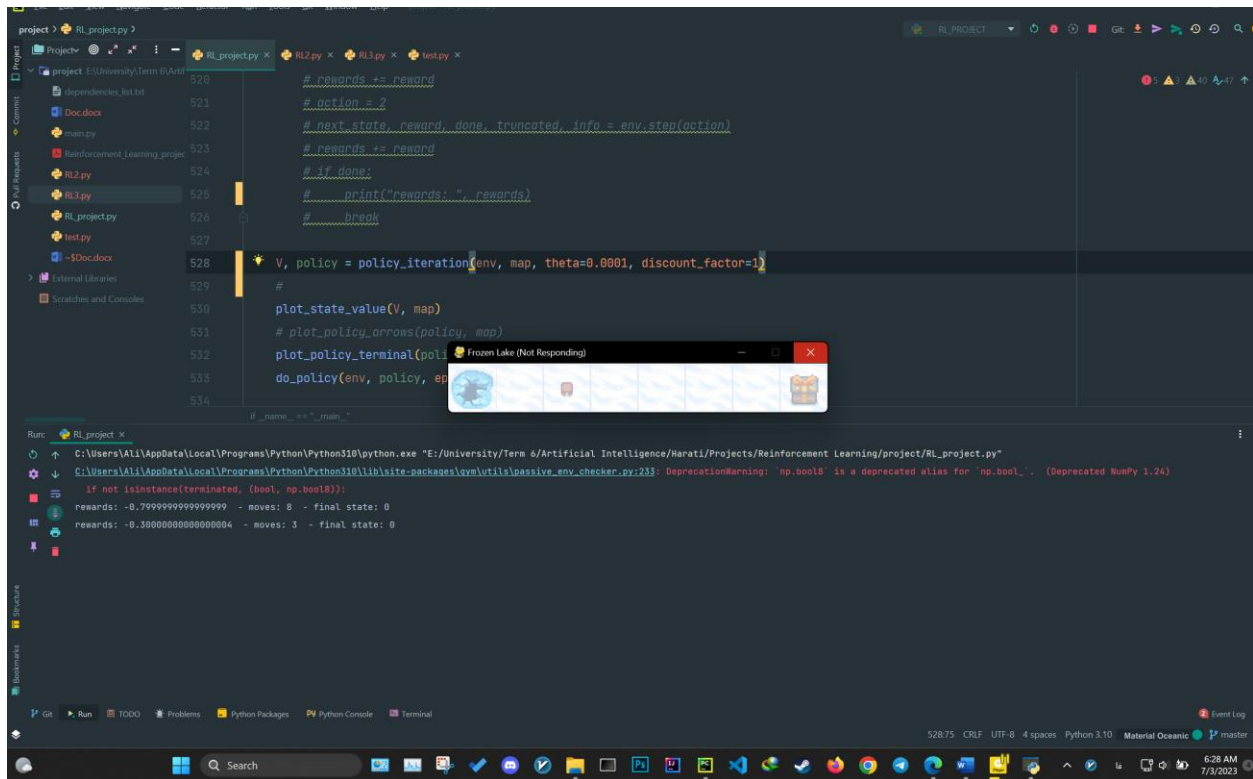
بخش دوم – ارزیابی و بررسی الگوریتم ها

پروژه RL

علی حمیدزاده ۹۹۱۲۷۶۲۵۲۹ – محمد افشاریان

سناریو ۱:

- با مقدار گاما = ۱: کرش میکنه و نمودار state-value نمیده!



```
520 # rewards += reward
521 # action = 2
522 # next_state, reward, done, truncated, info = env.step(action)
523 # rewards += reward
524 # if done:
525 #     print("rewards: ", rewards)
526 #     break
527
528 V, policy = policy_iteration(env, map, theta=0.0001, discount_factor=1)
529 #
530 plot_state_value(V, map)
531 # plot_policy_arrows(policy, map)
532 plot_policy_terminal(policy, map)
533 do_policy(env, policy, episode=1000)
534
535 if __name__ == "__main__":
```

Run: RL_project x

C:\Users\Ali\AppData\Local\Programs\Python\Python310\python.exe "E:/University/Term 6/Artificial Intelligence/Harati/Projects/Reinforcement Learning/project/RL_project.py"

C:\Users\Ali\AppData\Local\Programs\Python\Python310\lib\site-packages\jupyter_console\output.py:233: DeprecationWarning: 'np.bool' is a deprecated alias for 'np.bool_'. (Deprecated NumPy 1.24)

if not isinstance(terminated, (bool, np.bool_)):

rewards: -0.7999999999999999 - moves: 8 - final state: 0

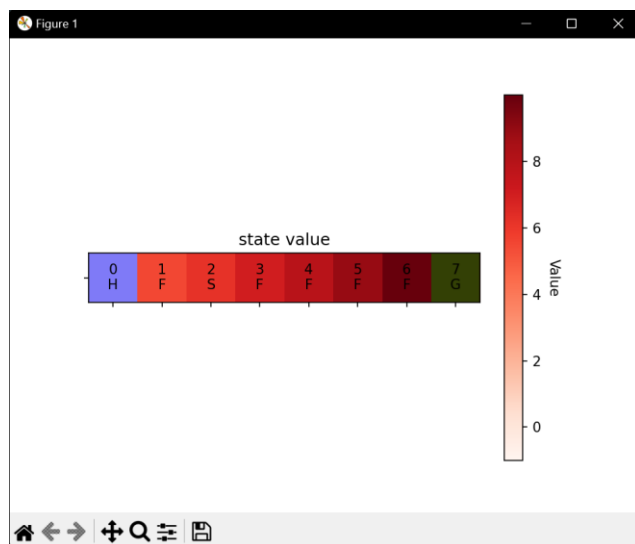
rewards: -0.30000000000000004 - moves: 3 - final state: 0

Event Log

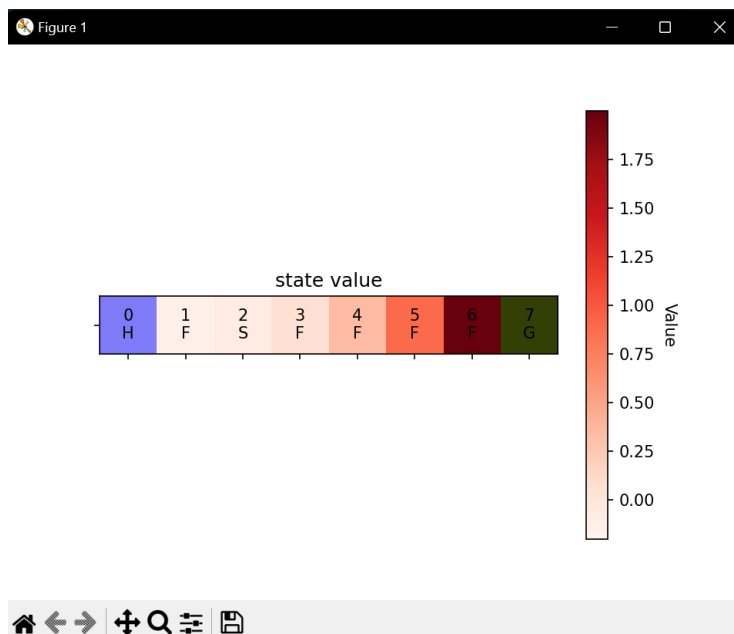
528.75 CRLF UTF-8 4 spaces Python 3.10 Material Oceanic master

6:28 AM 7/3/2023

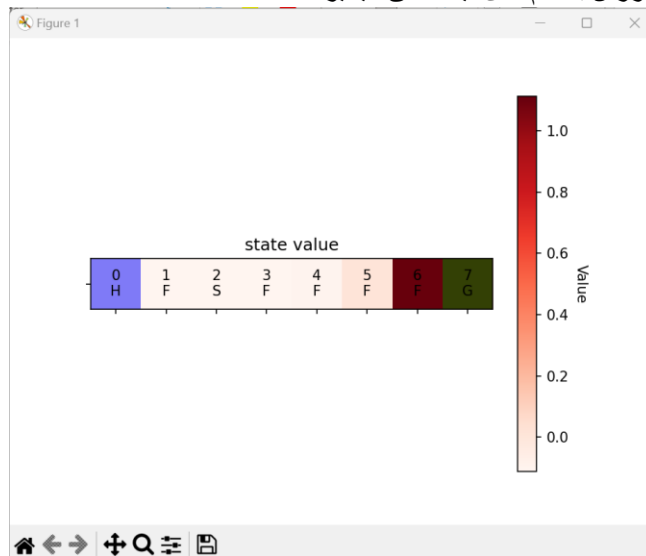
- با مقدار $\gamma = 0.9$: در هنگام محاسبه تابع ارزش در فرآیند ارزیابی سیاست، ضریب تخفیف به پاداش‌های آینده اعمال می‌شود. این ضریب باعث کاهش ارزش پاداش‌های به‌دست آمده در زمان‌های آینده می‌شود. ضریب تخفیف صفر نشان‌دهنده این است که عامل فقط به پاداش‌های فوری توجه می‌کند و از پاداش‌های آینده غافل می‌شود. از سوی دیگر، ضریب تخفیف برابر با یک به این معنی است که عامل پاداش‌های آینده را با پاداش‌های فوری به‌صورت یکسان ارزیابی می‌کند.



- با مقدار $\gamma = 0.5$: عامل پاداش‌های آینده را کمتر از ارزش پاداش‌های فوری (الان) ارزیابی می‌کند.



- مقدار گاما = $0/1$: کمترین ارزش به گام های آینده تعلق میگیره

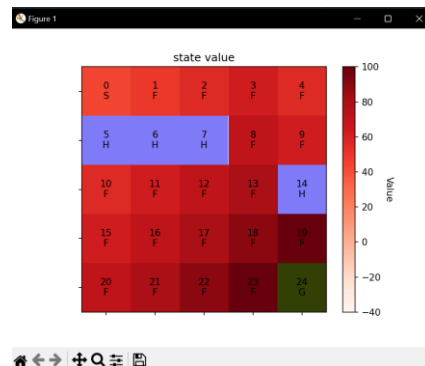


علت تفاوت بین نتایج، ارزشی هست که به گام های آینده در مقابل تجارب حال داده میشود.

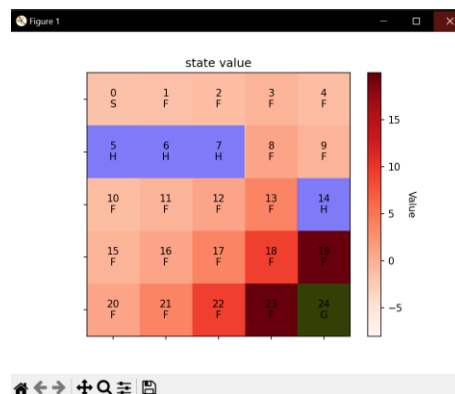
سناریو ۲:

- گاما = 1 : کرش میکنه

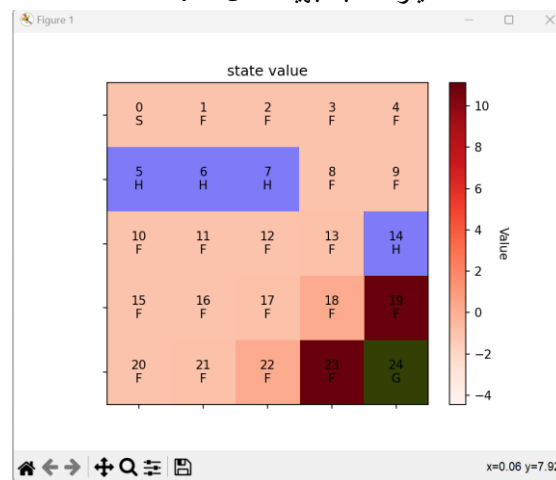
- گاما = $0/9$:



- گاما = $0/5$:

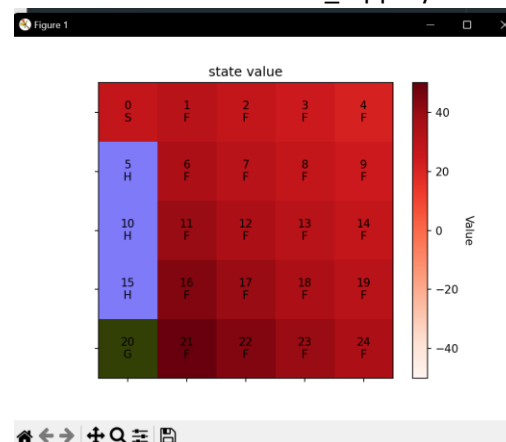


- گاما = 0.1 : علت تفاوت در مقادیر همانطور که گفته شد به دلیل تفاوت در میزان ارزش گذاری گام های آینده در مقابل گام های حال میباشد. به نظر میاد که گاما = 0.5 میتونه نسبتاً بهینه عمل کنه.

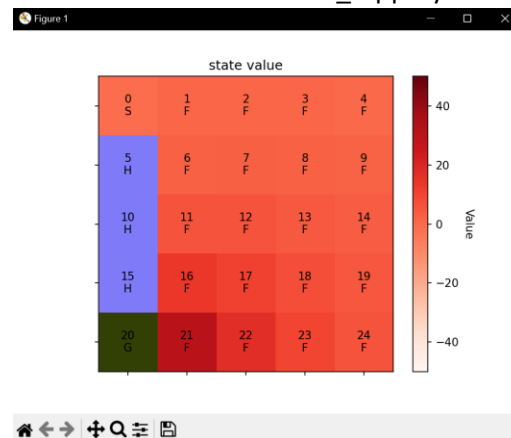


سناریو ۳:

:ls_slippery = False

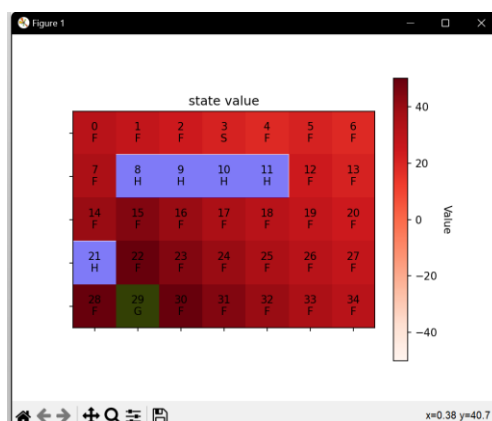


:ls_slippery = True

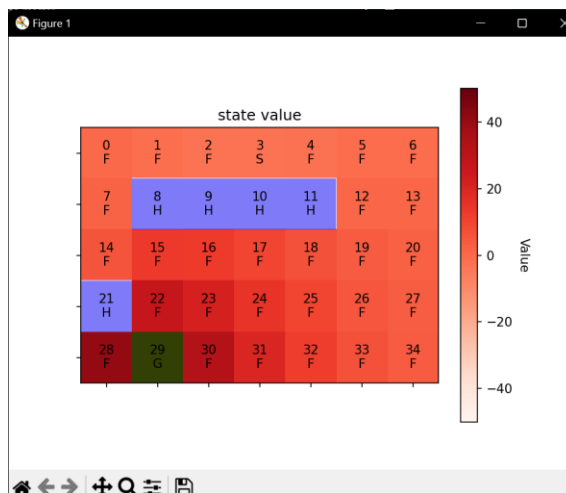


سناريو ۴:

:Is_slippery = False



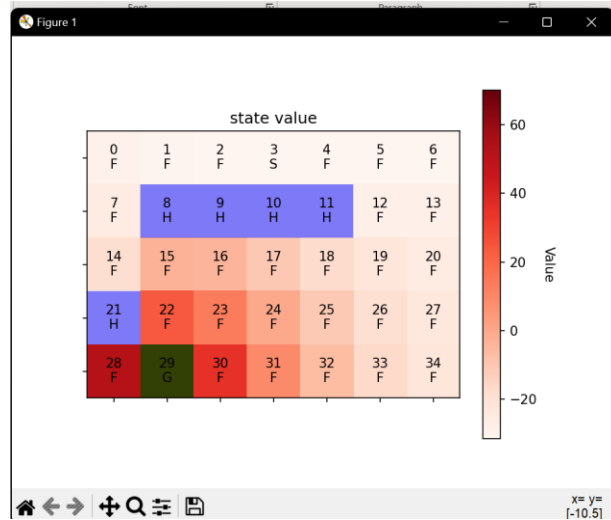
:Is_slippery = True



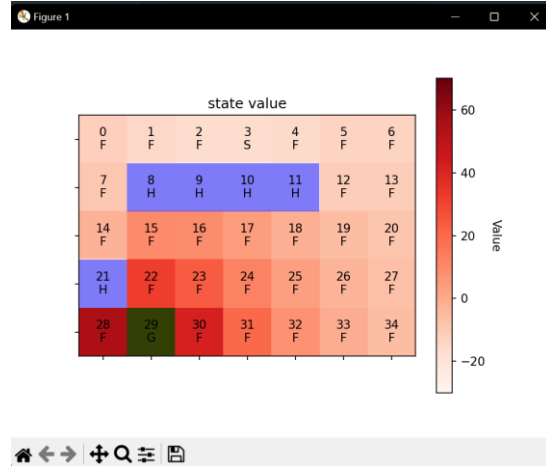
سناريو ٥:

Is_slippery=true

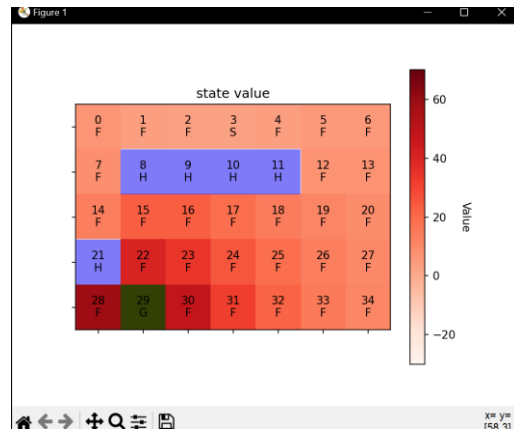
:Move_reward=-4



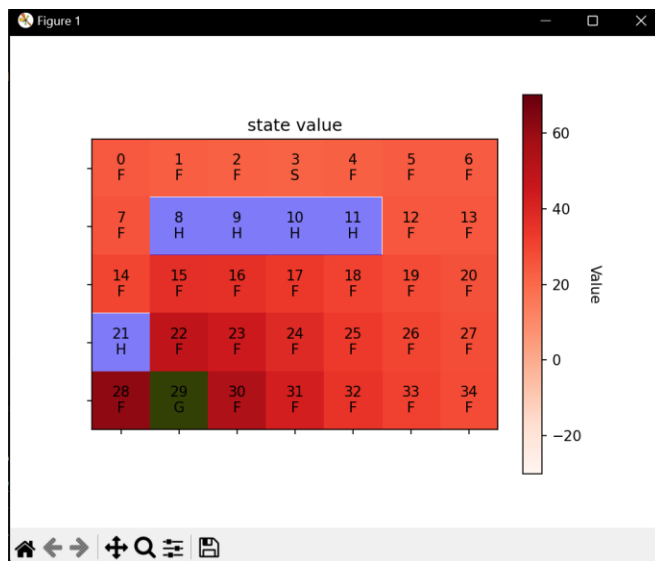
:Move_reward=-2



:Move_reward=0

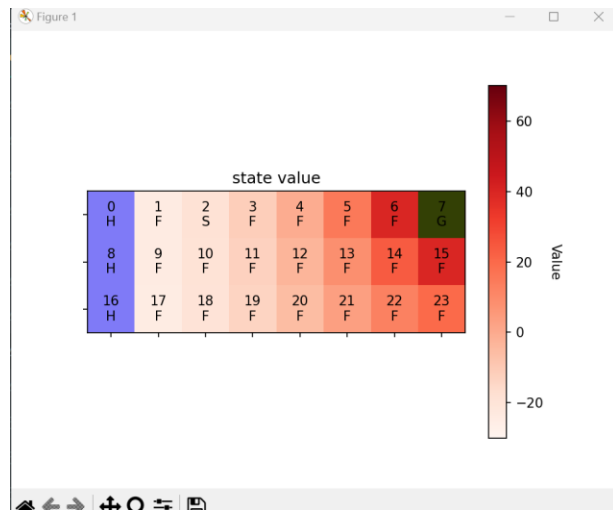


:Move_reward=2

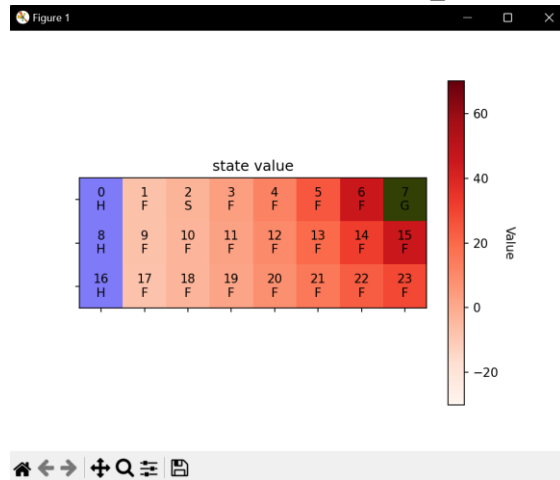


سوال ٦:

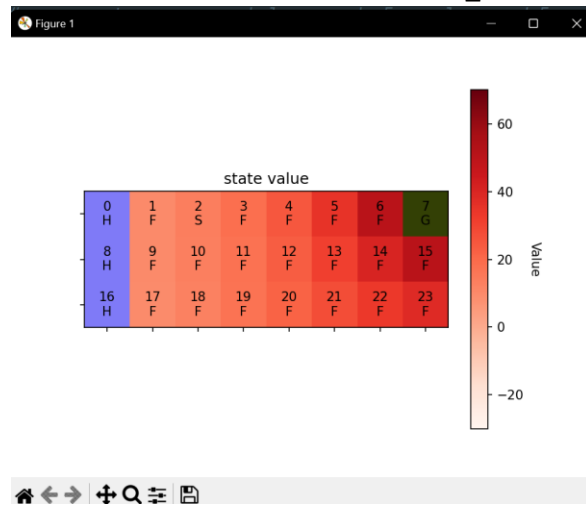
Is_slippery=true
:Move_reward=-4



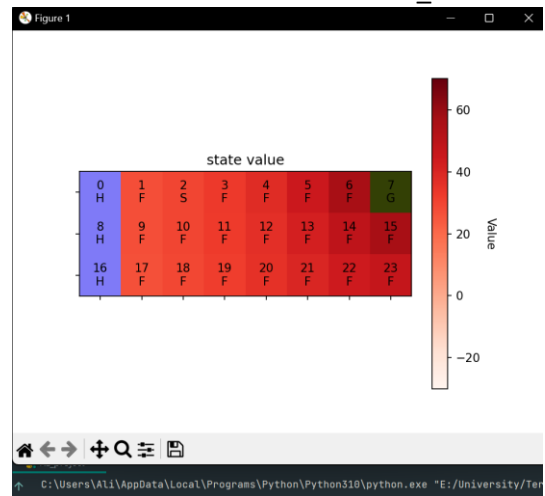
:Move_reward=-2



:Move_reward=0



:Move_reward=2



سناریو ۷:

الف: سیاست بدست آمده از policy_iteration

