تمرین سری سوم:
Markov Decision Process

لطفاً به نکات زیر توجه کنید**:**

- در صورتی که به اطلاعات بیشتری نیاز دارید می‌توانید به صفحه‌ی تمرین در وب‌سایت درس مراجعه کنید.
- این تمرین شامل یک سوال برنامه‌نویسی می‌باشد، بنابراین توجه کنید که حتماً موارد خواسته‌شده در سوال را رعایت کنید.
- ما همواره هم‌فکری و هم‌کاری را برای حلِ تمرین‌ها به دانشجویان توصیه می‌کنیم. اما هر فرد باید تمامی سوالات را به تنهایی به اتمام برساند و پاسخ ارسالی باید توسط خود دانش‌جو نوشته شده باشد. اگر با کسی هم‌فکری کردید نام او را ذکر کنید.
- برای ارسال پاسخ‌های خود از راهنمای موجود در صفحه‌ی تمرین استفاده کنید.
- هر سؤالی درباره‌ی این تمرین را در گروه درس مطرح کنید و یا از دستیاران حلِ تمرین بپرسید.
- نمره این سری تمرین از 100 محاسبه میشود.

- آدرس صفحه‌ی تمرین:
- آدرس گروه درس:

برای اجرای main.py از پایتون 3 استفاده کنید .

سوال یک) *ربات با زندگی ای حوصله سر بر* (40 نمره)

جهانی را در نظر بگیرید که از n x m خانه تشکیل شده است (ماتریس به ارتفاع n و عرض m):

- رباتی در این دنیا زندگی میکند که میتواند با اعمال action (شمال، جنوب، شرق و غرب) از خانه ای به خانه ی دیگر حرکت کند.
- نتیجه اعمال اکشن ها قطعی (deterministic) نیست. (تابع T در بدنه کلاس MDPProblem پیاده سازی شده است.)
- حرکت از خانه ای به خانه ی دیگر هزینه (Living reward) دارد.
- خانه هایی وجود دارند که در آنها فقط اکشن خروج اعمال میشود و با ورود به این خانه ها ربات reward نهایی (میتواند خوب و یا بد باشد) را گرفته و بازی خاتمه میابد.

مثالی از جهان 3x3

		+1EXIT
		-1EXIT

پیاده سازی کنید:

- 1) تابع compute_policy را در فایل mdp_problems به گونه ای کامل کنید که مقادیر π_k و V_k را در هر iteration محاسبه کند. (Value Iteration)
- تمامی پارامتر ها و توابع لازم از قبل پیاده سازی شده و در فایل پروژه توضیح داده شده اند.
- برای بررسی درستی الگوریتم خود کامند های زیر را اجرا کنید:

```
$python main.py type=smallWorld
$python main.py type=largeWorld
```

2) در هر دو حالت smallWorld و largeWorld، بعد از چند لوپ policy ها converge میشوند؟ (در قسمت تئوری پاسخ دهید)
3) در هر دو حالت smallWorld و largeWorld، بعد از چند لوپ V^* ها converge میشوند؟
4) از مقایسه 2 و 3 چه نتیجه ای میتوان گرفت؟

سوال دو) *دوران قرنطینه* (هر کدام از سوال ها 15 نمره)

ربات ما قرنطینه شده و در جدال با خود به سر میبرد و نمی داند کی فیلم تماشا کند و کی درس بخواند! با استفاده از MDP او را کمک کنید! جدول Transition به گونه زیر میباشد:

From State	Action	Probability	Result
Bored-Useless	Study	1.0	Bored,Productive
	Watch Movie	0.8	Bored, Useless
		0.2	Entertained, Useless
Bored-Productive	Study	1.0	Bored,Productive
	Watch Movie	0.8	Entertained, Productive
		0.2	Bored, Useless
Entertained-Useless	Study	1.0	Bored,Productive
	Watch Movie	0.8	Bored, Useless
		0.2	Entertained, Useless
Entertained-Productive	Study	1.0	Bored,Productive
	Watch Movie	0.8	Entertained, Useless
		0.2	Bored, Useless

جدول Rewards به گونه زیر میباشد:

R(*,*,State)	Has Reward Value
Bored-Useless	0
Bored-Productive	1
Entertained-Useless	1
Entertained-Productive	2

نکته: گاما را عددی بین 0 تا 1 فرض کنید.

سوال دو-یک) Policy ای را در نظر بگیرید که همیشه Study را انتخاب میکند.
$V^{\pi}(Bored-Useless)$ را بر حسب گاما محاسبه کنید.

سوال دو- دو) فرض کنید:

V*(Entertained-Productive) = A
Q(Bored-Useless, Study) = B (B>A)
Q(Bored-Useless, Watch Movie) = C

$Q\,(Bored-Productive,\ Watch\ Movie)$ را حساب کنید.

سوال دو-سه) مقادیر $V_0^*(state)$ ، $V_1^*(state)$ و $V_2^*(state)$ را محاسبه کرده و Policy در iteration دوم را بنویسید.

سوال دو-چهار) تفاوت های MDP و Expectimax را بیان کنید.