Rayan forsat

610398155

Assignment One

1. تعداد حالت ها وابسته به ابعاد بازی است. میتوان خانه های بلاک را حذف کرد

2. State: مختصات خانه ای که روی آن هستیم

حرکت به سمت بالا راست چپ پایین :Action

اگر به پرچم برسیم جایزه 50 میگیریم: Reward

اگر به تارگت برسیم جایزه 1- 1/len(flags) میگیریم

اگر به پرچم برسیم جایزه len(flags)/1 میگیریم

به ازای هر حرکت boardSize/1 مجازات میشویم

به از ای حرکت به خانه ای که تا به حال در آن بوده ایم boardSize/2 مجازات میشویم

3. گاما نرخ یادگیری را تعیین میکند. اگر گاما صفر باشد جایزه صرفا به انتخاب ما بستگی دارد و مهم نیست به چه state ای میرویم.. و اگر بینهایت باشد دیگر مهم نیست چه action ای انجام میدهیم.

آلفا نرخ یادگیریست. اگر صفر باشد q-table بدون تغییر باقی میماند و اگر 1 باشد بدون توجه به یادگیری های گذشته q-table را آپدیت میکند.

حالا نتیجه را برای مقادیر مختلف برسی می کنیم.

برای مقادیر alpha = 0.25, 0.5, 0.75 زمانی که آموزش را شروع می کنیم هرچه alpha را کمتر می کنیم تعداد iteration های بیشتری طول میکشد تا q-table به راه حل برسد. اشتباهات بیشتری انجام میدهد اما از اشتباهاتش درس میگیرد. برای alpha = 0.75 در اکثر مواقع با وجود آپدیت سریع مسیرها اما بعد از چند itter منحرف میشود و دوباره برمیگردد.

برای مقادیر gamma= 0.25,0.5,0.75 بیش

4. میتوانیم با تابع printQtable در main جدول q هارا چاپ کنیم.

تغییرات reward در زیر آمده است. رنگ سفید بیشترین reward یک state است

	-inf	Step: -22	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf
-inf	-inf	Iters: 75	-inf -0.03	-0.03	-0.03	-0.02	-0.02	-0.03	-0.03	-0.04 -inf
	0.00		-0.02	-0.)3	-0.03	-inf	0.01	-0.02	-0.03	-0.03
	0.02	-inf	-0.03	-0.02	-0.03	107	0.03	-0.03	-0.03	-0.03
-inf	-0.02	-0.02 -0.02	-0.03 _{-0.03}	-0.03	-0.03 -inf		-inf -0.02	-0.02 -0.02	-0.03	-0.03 -inf
	0.04	-0.03	-0.03	-0.12	-0.03		0 01	-0.02	-0.02	-0.03
	0.03	-0.02	-0.02	0.0	-0.03		- 1.03	-0.02	-0.03	-0.03
-inf	-0.02	-0.02	-0.03	0.03	-0.02 -inf		-inf -0.02	-0.03	-0.03	-0.03 -inf
	-inf	-inf	-0.03	-inf	-inf		-inf	-0.02	-0.02	-0.03
			-0.03			-inf		- 0.03	-0.02	-0.03
			-inf -inf			-inf -inf		-inf _0.02	-0.03	-U.Cinf
			-inf			-0.04		-inf	-inf	-0.02
	-inf	-inf		-inf		-0.04				- 1.03
-inf		-0.02 -inf		-inf		-inf				inf
				-int		-int				-int
	-0.02	-(.03 0 05		-0.03		-0.04				0.02
27.5	-0.03			-0.03		-0.03	+inf	-inf	-inf	0.03
-inf	-0.03	-0.02 -inf		-inf -inf		-inf _nos	-0.03	-0.03	-0.03	-0.02 -inf
	-0.02	-(.03		-0.03		2.02	-0.03	- 0.03	-0.03	0.02
	-0.03	-0.01	-inf	-0.02	-inf	-1.03	-0.02	0.02	-0.03	0.03
-inf	-0.03	-0.03	-0.02 -0.03	-0.02	-0.02 -0.03	-0.02	-0.03	-0.02	-0.02 _{-0.03}	-0.02 -inf
	0.02	-(,02	-0.02	-0.02	-0.02	-0.02	-inf	-inf	-inf	-inf
	0.03	-0.02	-0.02	-41.02	-0.02	-0.02				
-inf	-0.03	-0.03 -0.03	0.09	0.00	-0.02	-0.02 -inf				
	-02	-inf	-inf	-inf	-inf	-inf				
	- 1.03						-inf	-inf	-inf	-inf
-inf	-inf						-inf -0.01	O.Ot	-0.01 0.00	0.00 -inf
1	-0.02						-).03	-inf	0.4	0.14
- I	-0.03	-inf	-inf	-inf	-inf	-inf	0.02	3 (000)	4.00	0.00
-int	-000	-0.02	-0.02	-0.03 -0.02	-0.03 -0.02	-0.03 -0.02	-0.03 -inf		-inf 0.68	0.00
	-inf	-inf	-inf	-inf	-inf	-inf	-inf		-inf	-inf

5. ابتدا با تابع generate_random_env) در main یک محیط رندوم ایجاد می کنیم. سپس تابع train_random) را فراخوانی می کنیم.

مثال:

با انجام چند مثال از محیط های رندوم متوجه میشویم مدل در رسیدن به flag هایی که در مسیر رسیدنش به target نیست یا از بقیه flag ها دور افتاده ناتوان است. برای مثال board زیر را درنظر بگیریم

	-inf	Step:-inf	-inf	-inf	-inf	-inf	-inf		-inf	-inf
-inf	-0.01	Itoma 169 -0.01	0.02	-0.03	-0.02 0.01	0.02 0.11	-0.03 inf		-inf -0.03	-0.03 -inf
	-inf	-0.02	0.01	-inf	-0.02	-0.02	-0.07		-0.03	-0.03
	1000	-0.03	-0.03	1000	-0.02	-0.02	-0 03	-inf	-0.02	-0.03
		-inf -0.02	-0.0i -inf		-inf -0.02	-0.02	-0.02	-0.03 _0.01	-0.02	-0.02 -inf
		-inf	-0.01		-0.02	-inf	-0.02	-0.02	0.70	-0.03
			-0.03	-inf	-0.02		-0.03	-0.02	-0 02	-0.02
			-inf -0.02	-0.03	-0.02 _{-inf}		-inf -0.02	-0.02	-0.02	-0.02 inf
			0.01	-0.02	-0.02		-0.02	-0.02	-0.02	-4.01
Towns to	-inf	-inf	- 1.03	-0.02	-0.02	-inf	-0.02	-0.02	0.03	-0.03
-inf	-0.03	-0.02	-0.01	-0.02	-0.02	-0.02	-0.02 -0.01	-0.02 -001	-0.03 -0.02	-0.03 -inf
j	-0.01	-inf	-inf	-inf	-0.02	-0.02	-inf	-0.02	-0.02	-6 01
	-0.03				-0.02	-0.02		-0.01	-0.02	-0.03
-inf	-inf				-inf -0.02	-0.02 -inf		-inf -0.02	-0.02	-0.02 -lof
	-0.01				-0.02	-0.03		-0.03	-0.02	-0.0.
	-0.03	-inf	-inf	-inf	-0.02	-0.02	-inf	-0.02	-0.02	-0.02
-in	0.01	-0.03 -0.02	-0.03	-0.03	-0.03	-0.03 -0.02	-0.03	-0.02	-0.03 -0.02	-0.02 -in
	-0.02	-0.01	-inf	-0.02	-0.02	-0.02	-0.03	-0.02	-inf	-0.0
	-0.03	-0.03		-0.02	-0.02	-0.02	-0.05	-0.02		-0.08
-inf	-0.03	-0.03 -inf		-inf -0.03	-0.02	-0.02	-0.02	-0.03 -inf		-inf -nf
	-inf	-0.01		-0,02	-0 02	-0.03	-0.02	-inf		0.04
		- 0.03	-inf	-0.02	-0 01	-0.03	-0.02		-inf	-0.02
		-inf -0.02	-0.02	-0.02	-0.02	-0.02	-0.03 -inf		-inf 0.00	0.01 inf
		-0.01	-0.02	-0.02	-4.03	-0.02	-inf		0.12	0 17
		- 0.03	-0.02	-0.02	-41.02	-0.02			-0.01	0 01
		-inf -0.01	-0.02	-0.02	-0.02	-0.02 -inf			0.02	0.09 -inf
		-0.01	0.02	-0.02	-0.02	-0.02			014	(.58
	-inf	-0.02	- 1.02	-0.02	-0.02	-0.02	-inf	-inf	0.00	0.00
-inf	-0.03	-0.02	-0.02 _{0.12}	-0.02	-0.02	-0.03	-0.02	-0.01	0.00 U.54	000 -inf
	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf

مدل در مسیرهای مختلف سعی میکند flag هارا بگیرد اما موفق به گرفتن تمام آنها نمی شود.

مشکل کاملا اساسی در گرفتن flag هایی است که در بن بست قرار دارند چون بعد از گرفتن آنها باید برگردیم.. اما با بازگشت reward منفی میگیریم. دوباره به بن

بست وارد میشویم تا flag را بگیریم اما flag ای وجود ندارد. انقدر اینکار را انجام میدهیم تا دیگر هرگز وارد بن بست نشویم.