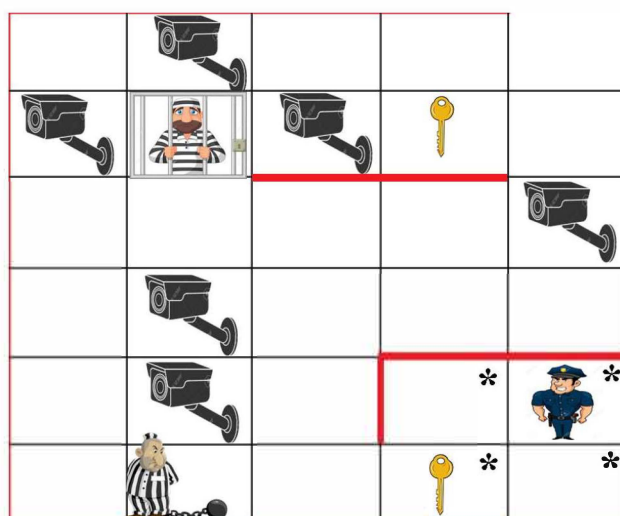




رهایی از انفرادی

چنگیز در زندان به رهایی‌دهنده از انفرادی معروف است! او به ازای هر زندانی که از انفرادی آزاد می‌کند مبلغ کلانی دریافت می‌کند. در این مسئله می‌خواهیم شریک جرم او شویم و مسیر بهینه‌ای که می‌تواند چنگیز را به زندانی محبوس در انفرادی برساند را یاد بگیریم!

نقشه‌ی زندان به صورت زیر است:



چنگیز در هر مرحله تنها می‌تواند یک خانه به بالا، پایین، چپ و یا راست حرکت کند. البته در صورتی‌که به دیوار نخورد! (خطوط قرمز دیوار هستند و عبور از آنها غیر ممکن است). وزنه‌ی سنگینی که به پای چنگیز زنجیر شده نه تنها امکان بالا رفتن از دیوار را از او گرفته، که حتی عبور میان خانه‌های نقشه را هم برایش دردآور کرده‌است.

چنگیز ابتدا بایستی خود را به کلید برساند. همان‌طور که در نقشه می‌بینید در زندان دو کلید وجود دارد که هر دو درب سلول انفرادی را باز می‌کنند. چنگیز کافی است یکی از این کلیدها را بردارد و سپس سراغ زندانی محبوس در انفرادی برود.

در برخی از خانه‌های مسیر دوربین نصب شده است، در صورتی که چنگیز در این خانه‌ها قرار بگیرد، تصویرش ضبط شده و به جرم پرسه زدن در زندان شکنجه می شود.

همانگونه که در نقشه می بینید، زندانبانی در حال نگهبانی دادن است. در صورتی که چنگیز وارد خانه‌ای از جدول که زندانبان در آن قرار دارد شود، دستگیر می شود و تا آخر عمرش به زندانی دورافتاده تبعید می شود و دیگر فرصت این شغل شریف را نخواهد داشت! زندانبان در هر لحظه با احتمال مساوی در یکی از چهار خانه ایست که با علامت * نشان داده شده اند.

فضای حالت و تابع پاداش را تعریف کنید و سپس نحوه یادگیری مسیر بهینه برای چنگیز را با روش $n\text{-step TD}(\lambda)$ برای لامبدای صفر و همینطور یک لامبدای غیرصفر که جواب مناسب می دهد پیاده سازی کنید. Performance و سرعت یادگیری چنگیز به ازای دو حالت مختلف را بررسی کنید.

نقطه شروع حرکت همواره همان خانه ایست که چنگیز در نقشه نشان داده شده است. مسیر بهینه مسیری است که کمترین جابجایی و همچنین کمترین احتمال و میزان جریمه و شکنجه شدن را برای چنگیز به همراه داشته باشد.