این کد دارای یک کلاس به اسم Maze می باشد که یک طول و عرض می گیرد و هزارتویی رندوم می سازد. نقطه شروع خانه 0 و 0 و خانه هدف خانه قطری آن می باشد

برای اینکه گارانتی بشود که هزارتو جواب دارد، یک DFS رندوم زده می شود از شروع تا هدف. اجازه داده نمی شود که در آن مسیر دیواری گذاشته بشود.

این کد همچنین 2 جایزه با امتیاز +3 به صورت رندوم در هزارتو قرار می دهد.

با هزار اپیزود train می شود (بستگی به هایپر پارامتر ها دارد. اگر اپسیلون زیاد باشد، برای اطمینان از جواب بهینه، باید مقدار بیشتری train بشود).

دو نمودار کشیده می شود. یکی از آن ها دارای چهار subplot است که در آن نمودار حرارتی هر یک از حرکات را برای هر خانه نشان می دهد.

در نمودار دوم بهترین action را برای هر خانه نشان می دهد.

هرچقدر رندوم بیشتر باشد، اپیزود هم باید بیشتر بشود. اگر هر دو زیاد بشوند، جواب بهینه تر می شود زیرا که گرفتن جایزه ها باید اثبات بشود برای ایجنت که چیز خوبی است (از انجایی که reward جایزه ها کم است، به زور می ارزد که ایجنت به آن سمت برود).

مقدار آلفا هرچه بیشتر باشد، مقادیر جدید دارای ارزش بیشتری هستند. از دید کلی، این خوب است. اما بعضی مواقع که حرکات رندوم بیشتر می خواهیم، نباید به این صورت باشند. (باید مساوی باشند ارزش ها. یعنی a = 0.5)

گاما هم در این مسئله باید باید مقدار زیادی داشته باشد. هیچ سودی در کم بودن آن نیست در این مسئله.