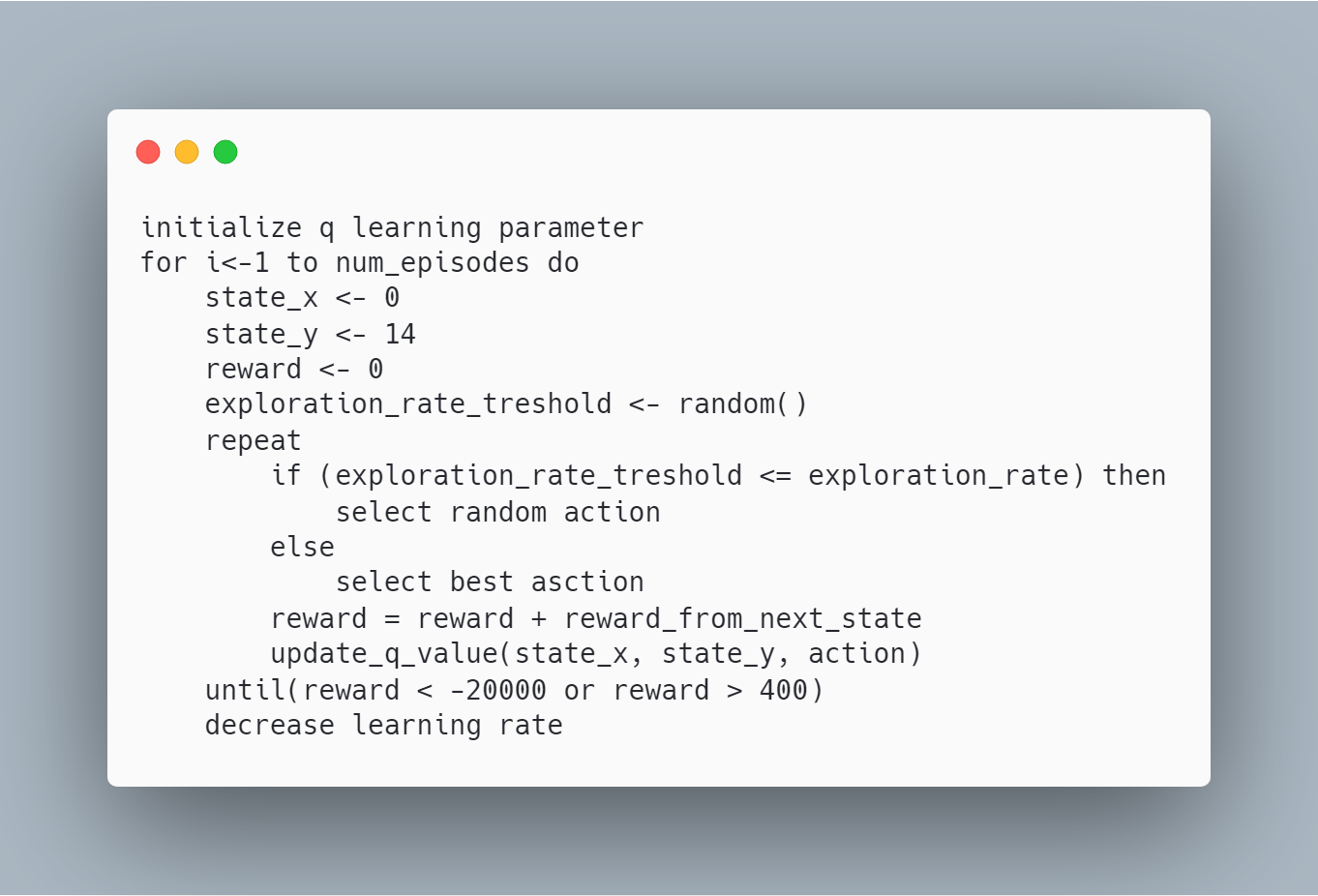
1. **Analisis**

Terdapat sebuah gridword berukuran 15 x 15 yang berisi point reward untuk setiap gridnya. Dari gridword tersebut, akan dicari *optimum policy* sehingga agent pada gridword yang dimulai dari grid (14,0) dapat mencapai grid (0,14) dengan menghasilkan reward yang paling optimum.

1. **Strategi Penyelesaian Masalah**

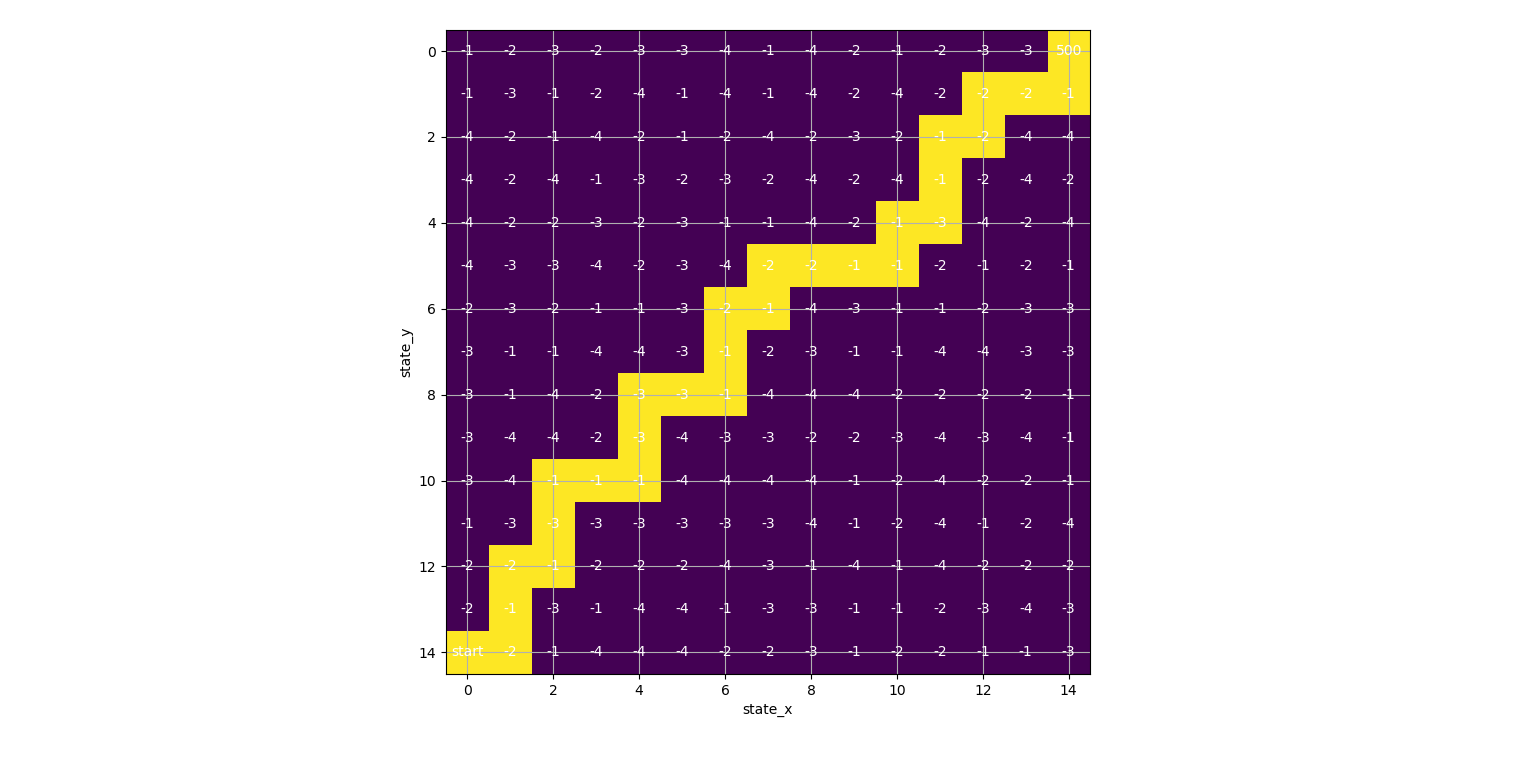
Sebelum membangun algoritma untuk mendapatkan *optimum policy*, **reward** dari *gridworld* yang disediakan disimpan ke dalam *file* csv. Kemudian algoritma untuk mendapatkan optimum policy adalah sebagai berikut:



Inisialisasi q learning parameter terdiri dari:

* Alpha (learning rate) = 0.4
* Gamma = 0.9
* Exploration\_rate = 1
* Num\_episodes = 1000

Parameter learning rate tersebut dipilih agar **q value** sebelumnya berpengaruh untuk perhitungan perubahan **q value** selanjutnya. Gamma tersebut dipilih untuk membuat state selanjutnya sangat berpengaruh terhdapat reward akhir. Exploradio rate tersebut dipilih agar saat *episode* atau iterasi awal program lebih banyak mengeksplorasi data dibanding exploitasi data. Semua data tersebut dipilih untuk menghasilkan *optimum policy*. Kemudian untuk setiap iterasi *episode* dipilih exploration rate treshold antaran 0 sampai 0,4. Artinya eksploitasi dapat dilakukan jika exploration\_rate sudah mencapai nilai 0,4. Iterasi tiap episode akan berhenti jika **total reward** yang di dapatkan kurang dari -**20000** atau lebih dari **400**. Berikut hasil dari progam yang dibuat:



Nilai reward yang paling optimum adalah **455**.