Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
  Initialize s
  Repeat (for each step of episode):
    Choose a from s using policy derived from Q ($e.g. \quad \varepsilon - greed$)
    Take action a, observe r, s'
    $Q(s, a) = Q(s, a) + \alpha[r + \gamma * max_{a'} Q(s', a') - Q(s, a)]$
    $s = s'$
    Until s is terminal