初始化$Q(s,a)$

循环(对每个episode)

初始化s

循环(对每个episode的每个step)

按照ε贪婪策略基于Q选择s状态下要执行的动作a

执行动作a，观测得到r, s'

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{\alpha'} Q(s',a') - Q(s,a) \right]$$

$$s \leftarrow s'$$

直到s是结束状态