Predictive Analytics World / Deep Learning World
Extra Exercise – Q-Learning

1. The equation below uses a 1-step lookahead, because it explicitly considers only one reward, then uses the Q estimate at time t+1. Modify the equation to perform 2-step lookahead instead. This will explicitly consider two rewards, then use the Q estimate at time t+2. (The trick is to handle the gammas properly.) Implement 2-step lookahead, and compare its results.

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \Big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \Big]$$