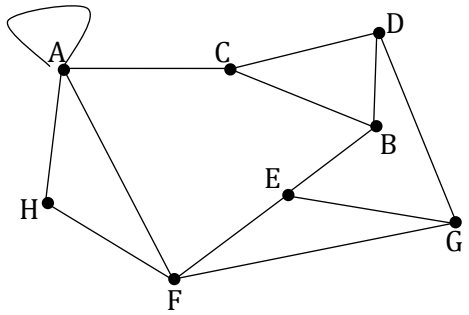
According to the given graph, a robot try to find the exit which is represented by A. When this robot use Q learning, show the first four updates in trainig procedure by starting H node. Note: we can use the learning parameter (γ) as 0.8, and the four random action as A, H, F,A, and F.



At first, we should prepare R matrix according to reward node. Reward node and its neighbours can be 100 points.

$$R = \begin{bmatrix} 100 & - & 0 & - & - & 0 & - & 0 \\ - & - & 0 & 0 & 0 & - & - & - \\ 100 & 0 & - & 0 & - & - & - & - \\ - & 0 & 0 & - & - & - & 0 & - \\ - & 0 & - & - & - & 0 & 0 & - \\ 100 & - & - & - & 0 & - & 0 & 0 \\ - & - & - & 0 & 0 & 0 & - & - \\ 100 & - & - & - & - & 0 & - & - \end{bmatrix}$$

Then start Q matrix with only zeros.

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1. When the robot goes to node A;
   Q(H, A) = R(H, A)+γ max(Q(A,A), Q(A,C), Q(A,F), Q(A,H))=100+0.8 max(0,0,0,0)=100

2. When the robot goes to node H;
   Q(A, H) = R(A, H)+γ max(Q(H,A), Q(H,F))=0+0.8 max(100,0)=80

3. When the robot goes to node F;
   Q(H, F) = R(H, F)+γ max(Q(F,A) , Q(F,E) , Q(F,G) , Q(F,H))=0+0.8 max(0,0,0,0)=0

4. When the robot goes to node A;
   Q(F, A) = R(F, A)+γ max(Q(A,A), Q(A,C), Q(A,F), Q(A,H))=0+0.8 max(0,0,0,80)=64

5. When the robot goes to node F;
   Q(A, F) = R(A, F)+γ max(Q(F,A) , Q(F,E) , Q(F,G) , Q(F,H))=0+0.8 max(64,0,0,0)=51.2

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 51 & 0 & 80 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 64 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 100 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$