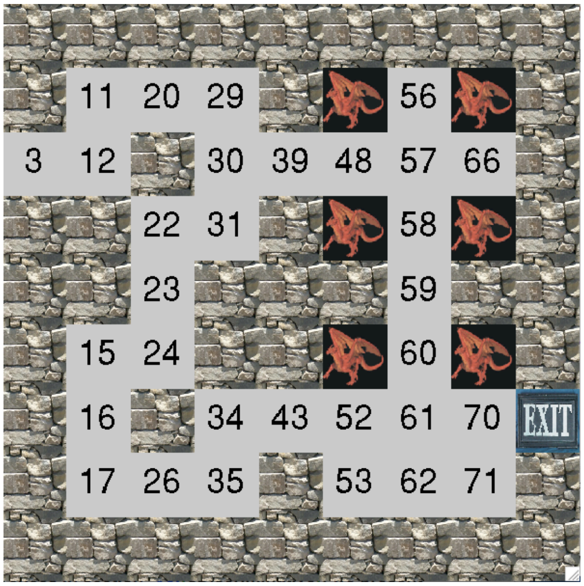
9.4 Policy and value iteration

In this problem, you will use policy and value iteration to find the optimal policy of the MDP demonstrated in class. This MDP has $|\mathcal{S}| = 81$ states, $|\mathcal{A}| = 4$ actions, and discount factor $\gamma = 0.9925$. Download the ASCII files on the course web site that store the transition matrices and reward function for this MDP. The transition matrices are stored in a sparse format, listing only the row and column indices with non-zero values; if loaded correctly, the rows of these matrices should sum to one.

(a) Compute the optimal policy $\pi^*(s)$ and optimal value function $V^*(s)$ of the MDP using the method of *policy iteration*. (i) Examine the non-zero values of $V^*(s)$, and compare your answer to the numbered maze shown below. The correct solution will have positive values at all the numbered squares and negative values at all the squares with dragons. Fill in the correspondingly numbered squares of the maze with your answers for the optimal value function. Turn in a copy of your solution for $V^*(s)$ as visualized in this way. (ii) Interpret the four actions in this MDP as (attempted) moves to the WEST, NORTH, EAST, and SOUTH. Fill in the correspondingly numbered squares of the maze (on a separate print-out) with arrows that point in the directions prescribed by the optimal policy. Turn in a copy of your solution for $\pi^*(s)$ as visualized in this way.

(b) Compute the optimal state value function $V^*(s)$ using the method of *value iteration*. For the numbered squares in the maze, does it agree with your result from part (a)? (It should.) Use this check to make sure that your answers from value iteration are correct to at least two decimal places.

(c) **Turn in your source code for the above questions.** As usual, you may program in the language of your choice.


11 20 29 56
3 12 30 39 48 57 66
22 31 58
23 59
15 24 60
16 34 43 52 61 70 EXIT
17 26 35 53 62 71