Q-Learning

Justification

For my implementation of the Q-Learning algorithm I added to the sweepers a hashmap with a key based on a tuple representing a state action and the value being a float representing the Q value for that particular key. This hashmap represents the sweepers Q-Table. I also added a Set which maintains which state actions performed already in a given iteration. This set is reset at the end of each iteration. This stops the sweeper from making the same state action twice in the same iteration.

Method parameters:

- Max number of ticks: 500
- Reward Values: Supermine=-100; Nothing=--20; Mine=100
- Discount factor: 0.5

These values were picked for no particular reason, only that the order of their values have to be a in that particular order Supermine<Nothing<Mine.

I made the number of ticks for each iteration to 500 because at 2000 ticks per generation if a single sweeper manages to survives and avoids the supermines for long only it learns all other mines won't. 500 ticks per generation mitigates this problem and speed up reiteration times.

I also end the iteration if all the mines have been collected, which is the behaviour when all sweepers are dead.

The given discount factor of 0.5 is convenient because it was half of the maximum it could be.

Method

1. For each sweep
    a. Pick the max next possible action that has not been perform from current state in this iteration. If all actions have been performed in this iteration, pick the max.
2. Perform the chosen max choice for all mines
3. For each mine on same tick
    a. Calculate immediate reward for the sweeper's new state
    b. Update the sweeper's previous state's Q value with the function $Q = R(x_p, y_p, a_p) + \gamma NPA(x, y)$ where $R(x_p, y_p, a_p)$ is the reward for the just performed state action, γ is the discount factor and $NPA(x, y)$ is the Q value of the next possible action of the current state.
4. Repeat from step 1 for each tick until end of iteration, or all mine are collected or until all sweepers are dead.

Results

Environment	Average Mines Collected	Average Deaths
Environment 1	1.33	0.2
Environment 2	0.796667	0.65
Environment 3	0.333333	2

Results for the 220 iterations for each environment can be found in the root solution folder in the following files:

- results QL – env1
- results QL – env2
- results QL – env3

They are essentially csv files the order is:

- Average mines gathered for that iteration
- Number of deaths for that iteration

The last 20 iteration of each environment were used for the results.