Exercise 1 - Mars-Rover control

In this example the Mars-Rover learns to find a good to optimal policy. Let's compare SARSA, first visit MC and Q-Learning on the following Markov decision process (MDP):

s_1	s_2	s_3	s_4	s_5	s_6	s_7
1						10

The MDP dynamics is stochastic. The rover can take action a_1, which should bring it one state state or action a_2, which should bring it one state left. The dynamics is such, that with a probability of 0.5 the rover ends in the intended state. With a probability of 1/3, the motor does not react and the state does not change and with probability 1/6 a mars storm transports the rover to the opposite state. The episodes start randomly in the non-terminal $s_2 - s_6$. Episodes terminate either on the extreme left or the extreme right (blue). When an episode terminates on the right, a reward of $+10$, when it terminates on the left a reward of 1 occurs; all other rewards are zero. For all experiments run the algorithms for 3000 episodes. Let the learning rate α decay inversely to the number of episodes to 0.01 in around 1000 episodes. Use the ϵ-greedy tactic, where ϵ decays in the same manner to 0.1 in around 2000 episodes.

- Show the values of each non-terminal state as a function of the episodes for all three methods. What can you observe (moving average)?

- Show a graph, where the learned policy during the training are evaluated. Do this by using the "success rate" meaning if the episode ended up in the right terminal state. Which method shows the fasted success increase (moving average)?

- Plot the mean reward per episode during the training (moving average).

★ Plot the moving average of the regret of these methods during the training. The regret is defined as the difference of the current Q to the optimal Q^*.

★ Include double Q-Learning in the experiments. What do you see?