The objective of this assignment is to implement, compare, and discuss the performance of three algorithms: SARSA, Q-learning, and Double Q-learning in navigating a maze environment, for the single agent case and for a multi-agent case.

Consider the following 2 grids worlds in which an agent can have 4 actions: up, down, left and right:

Gridworld A

					▨				
					▨				
S					▨		G		
					▨				

Gridworld B

S							G		
0	0	0	1	1	1	2	2	1	0

In grid A the grey cells are obstacles, while in grid B there is a crosswind upward through the middle of the grid. In the regions affected by the wind the resulting next states are shifted upward by the wind. The strength of the wind is given below each column, in number of cells shifted upward. For example, if the agent is one cell to the right of the goal, then the action left takes it to the cell just above the goal.

We treat this example as an undiscounted episodic task, with constant rewards of -1 until the goal state is reached. For the goal state the reward is +1. Initial Q(s,a) values are 0.

Task 1. Implement SARSA, Q-learning, and Double Q-learning in Gridworld A and Gridworld B, with different values for ε in the ε-greedy policies and different values for the learning rate. You have to consider at least 3 different values for each parameter.

Compare the performance of the 3 algorithms in each case using the following metrics:

- Convergence Time - the number of episodes required for each agent to consistently find an optimal or near-optimal path to the goal.
- Path Efficiency - the average number of steps taken to reach the goal over a set of trials.
- Robustness - consistency in performance across test scenarios

Plot the learning curves for the 3 algorithms in each case.

Discuss and interpret the results: discuss the observed behaviors, highlighting the strengths and weaknesses of each algorithm in the context of the maze navigation task.

Task 2. Consider Gridworld B in which there are 3 agents, each agent using the same algorithm – you choose the one you prefer (SARSA, Q-learning or Double Q-Learning). The setting is the same as in Task 1 except that in this case the agents receive a reward of 10 if they reach the goal at the same time step and a reward of -0.5 if they reach the goal at different time steps. This approach tries to model (in an extremely simplistic way) the cooperative behavior of the agents.

Implement the algorithm (you chose) for the 3 agents, plot the learning curves, discuss and interpret the results.