**Report**

**Name: Chen Xingyi UID: 3036198102**

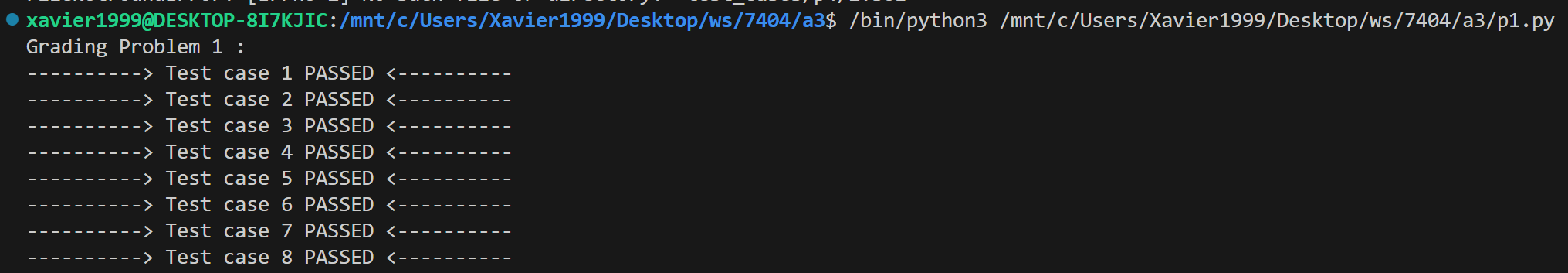
**Parser:**

Completed (Runtime < 1s).

**P1:**

Completed. (Runtime < 1s).

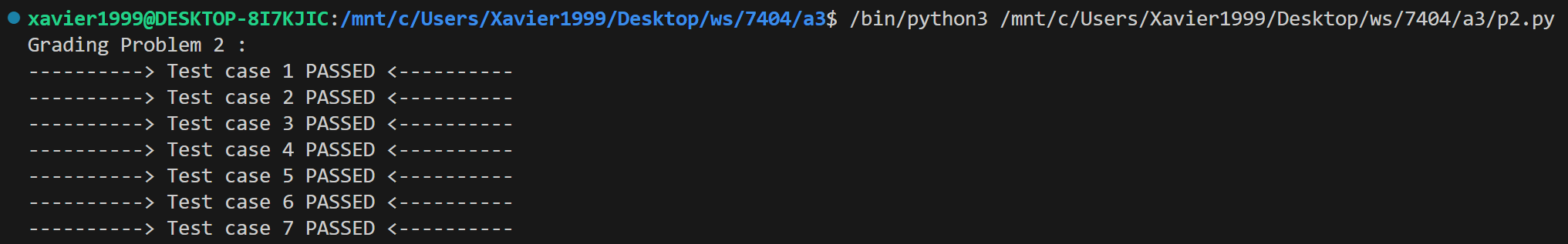
Challenge: Hardly any challenge.



**P2:**

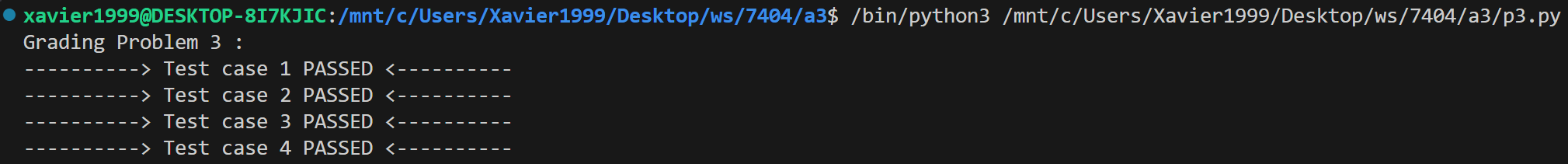
Completed. (Runtime < 1s).

Challenge: Hardly any challenge.



**P3:**

Completed. (Runtime < 1s).

Challenge: Hardly any challenge. 

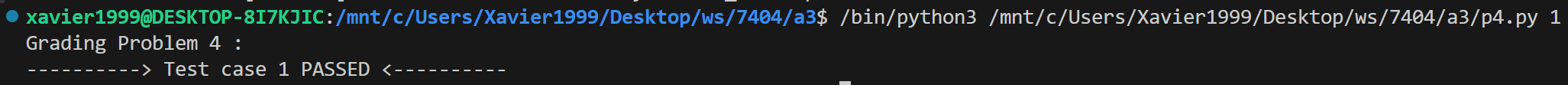
**P4:**

Completed. (Runtime < 1s).

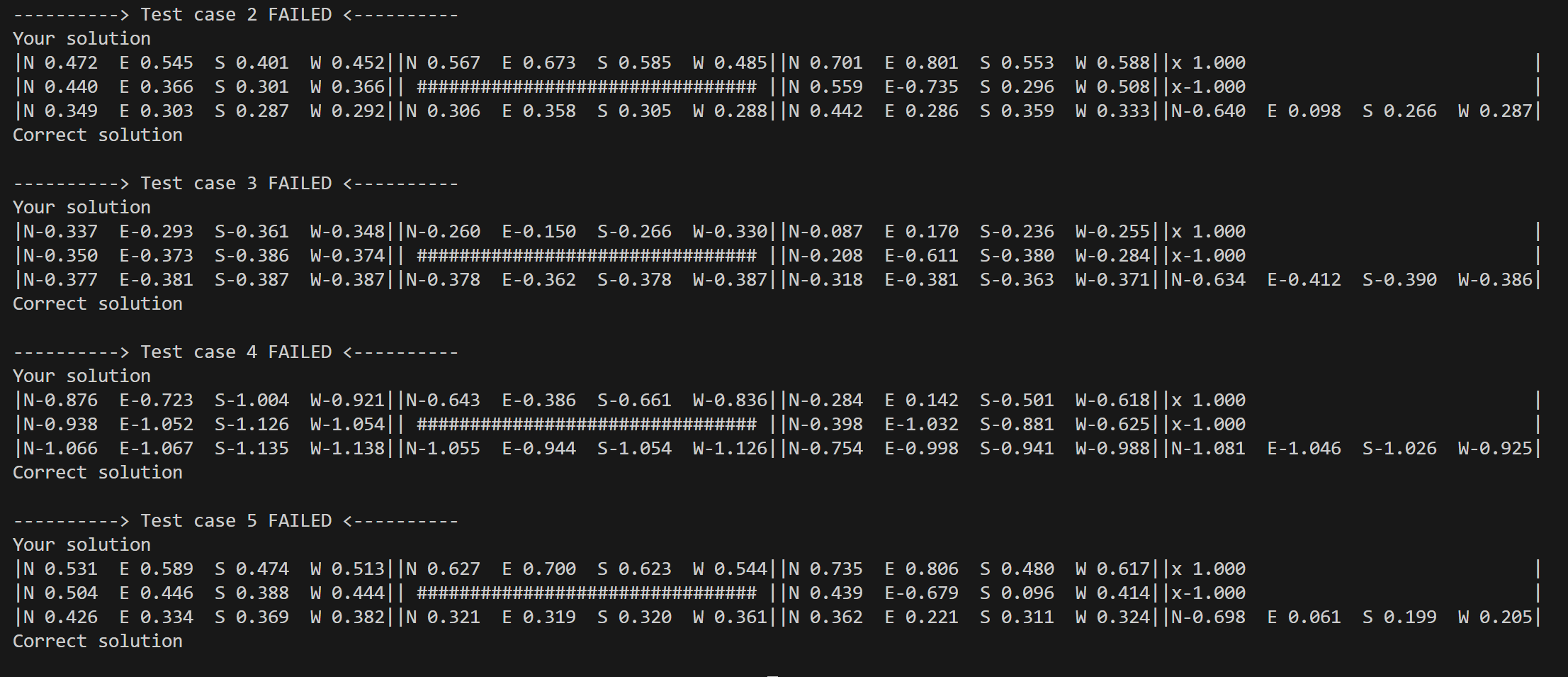
Challenge:

My algorithm successfully identifies the optimal policy approximately 65% of the time. However, upon closer examination, I noticed that deviations from the optimal policy tend to occur when the 3rd column of the 3rd row contains the action N instead of W.

**Analysis:** The random directions selected during the learning process can significantly influence the resulting optimal policy. Considering that each action incurs a negative living reward, if a particular direction is chosen more frequently, its associated Q value tends to decrease. This effect is particularly pronounced at the outset when the learning rate is high. The cumulative impact of living rewards may lead to a substantial reduction in the Q value for that specific direction. Consequently, the algorithm might favor an alternative direction as the optimal policy for a given state if its Q value surpasses that of the originally preferred direction. Q-Value Impact: The living reward’s influence can cause a substantial reduction in the Q-value for a specific direction. Consequently, if the Q-value for the optimal policy’s direction is smaller than that for other actions, the algorithm may choose an alternative direction as the best policy for a given state.



The converged policy and the converged Q values for the left test cases in p4:



**approximate number of hours: 9 hours**