Homework

- Use the CliffWalking domain from OpenAI gym
  - See Example 6.6, pg 132 in Sutton and Barto [2018]

- Modify the TD(λ) algorithm presented to implement SARSA(λ)
  - The only difference here is that there is an eligibility trace for each state-action pair!
  - Use ε-greedy policies with $\varepsilon = 0.1$ and a learning rate of $\alpha = 0.5$
  - Run SARSA(λ) on the domain for $\lambda = \{0, 0.3, 0.5\}$ for 200 episodes
    - Record the return for each episode
    - Average your returns over 100 runs

By next week's lecture, submit on Moodle:

1. Perform a single run of the algorithm. After each episode plot the value function (take $\max_a Q(s,a)$) learned so far as a heatmap for each λ side by side. This should result in 200 separate plots/images. Turn these images into an animation/video and submit it.
2. A combined plot of average return over time for the different values of λ. Include error bars/shading indicating variance in your results
3. Your code