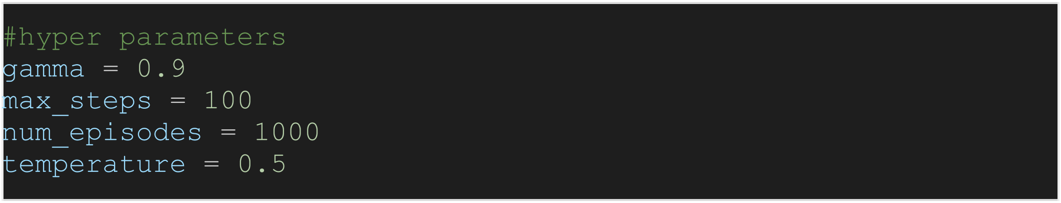
Training a Reinforcement Learning Agent for Reward Maximization in a Grid World Environment with Obstacles

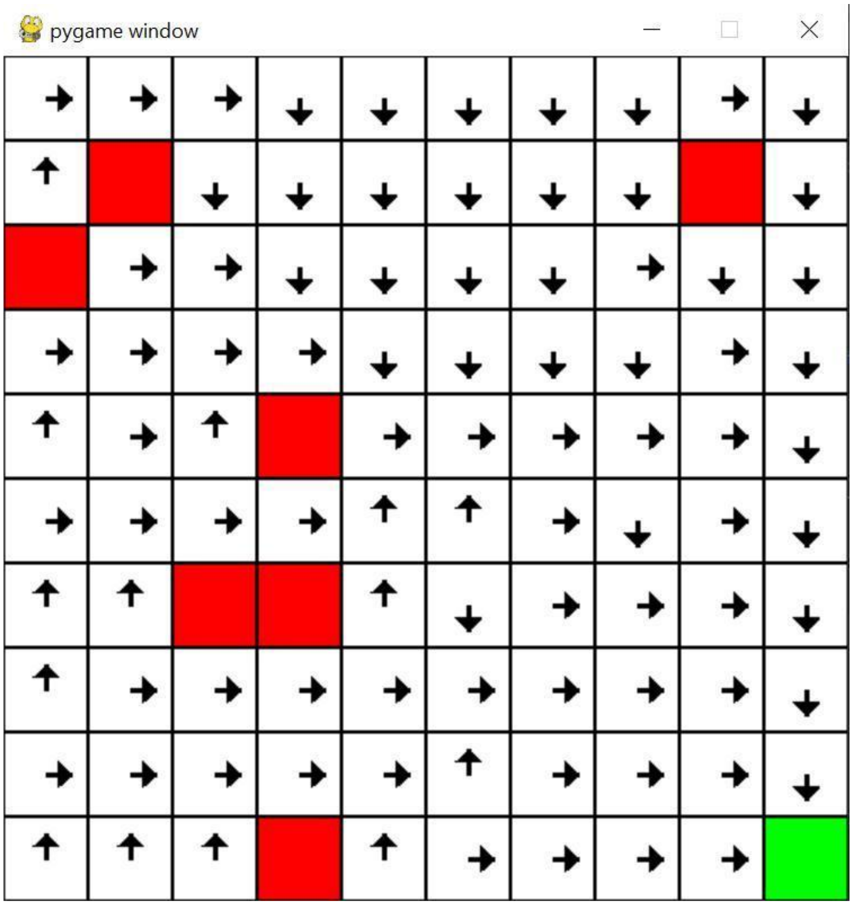
The code initializes a value function that is used to determine the value of each state and action in the grid. Boltzmann distribution is used to select an action for each state, and the agent updates the value function based on the reward received for each action and the maximum expected reward of the next state. The agent runs through the environment for each episode and receives a cumulative reward for each episode.

Parameters used for training the agent:



```
#hyper parameters
gamma = 0.9
max_steps = 100
num_episodes = 1000
temperature = 0.5
```

The visualization uses a grid size of 10. This particular agent has a preference for exploration. Obstacles are placed randomly throughout the grid. The arrows on the grid indicate the recommended actions to take at each cell.



→	→	→	↓	↓	↓	↓	↓	→	↓
↑	🟥	↓	↓	↓	↓	↓	↓	🟥	↓
🟥	→	→	↓	↓	↓	↓	→	↓	↓
→	→	→	→	↓	↓	↓	↓	→	↓
↑	→	↑	🟥	→	→	→	→	→	↓
→	→	→	→	↑	↑	→	↓	→	↓
↑	↑	🟥	🟥	↑	↓	→	→	→	↓
↑	→	→	→	→	→	→	→	→	↓
→	→	→	→	→	↑	→	→	→	↓
↑	↑	↑	🟥	↑	→	→	→	→	🟩

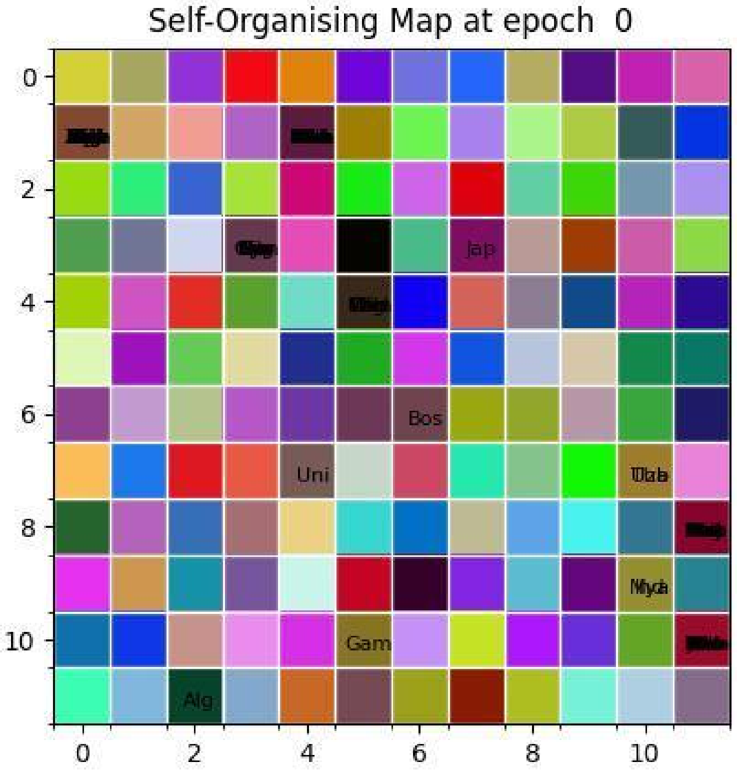
Clustering World Data using SOM:

The dataset that we have chosen is the world Happiness Report data set for the year 2019.

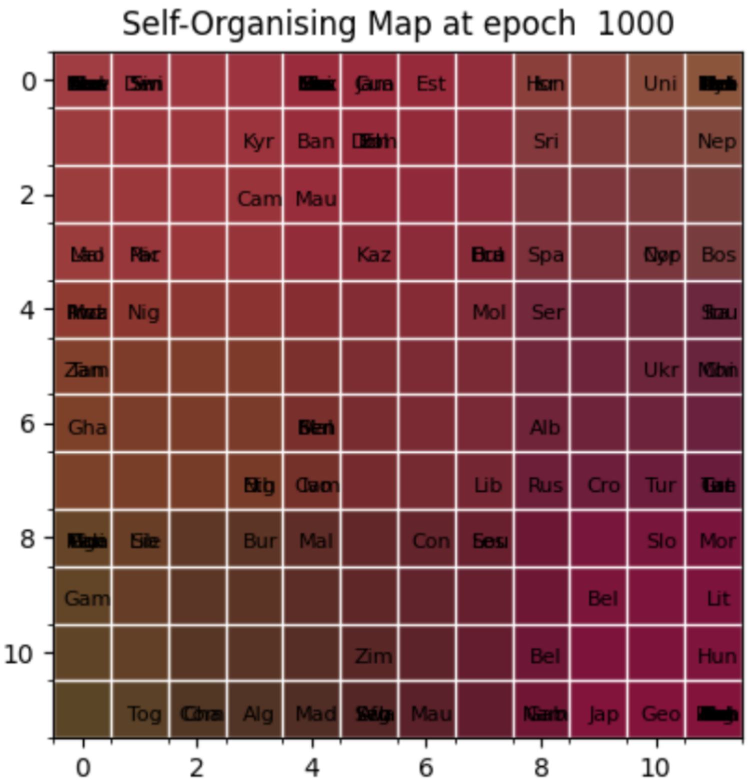
Our dataset contains the following six attributes:

- Log GDP per capita
- Social support
- Healthy life expectancy at birth
- Freedom to make life choices
- Generosity
- Perceptions of Corruption

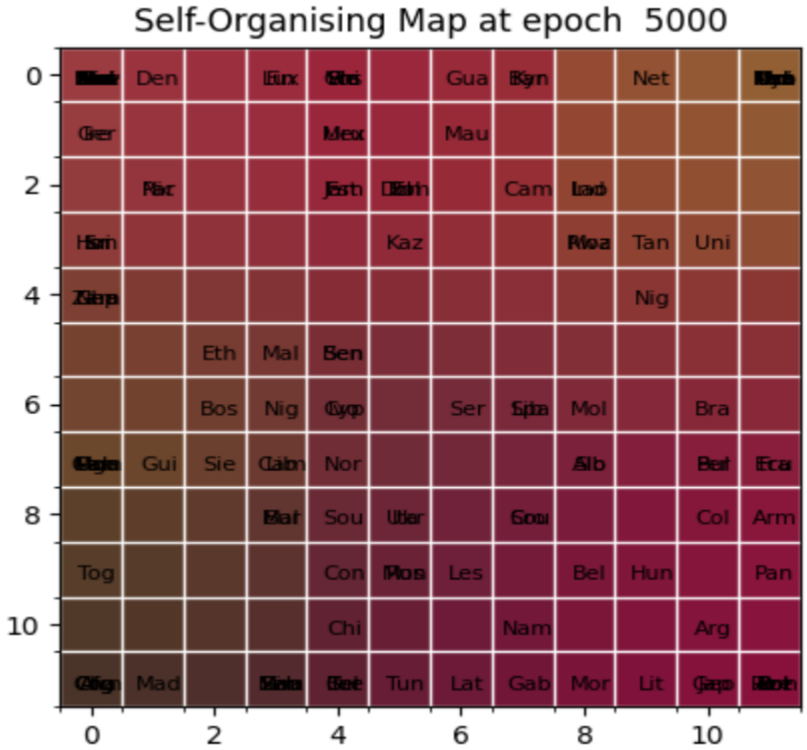
The values for each of the attributes are normalized between 0-1 where a higher score for each relates to a higher ranking of the country on the Happiness scale. These six values are mapped to RGB values using a color map of random colors and the average of the six values is set as the RGB weight value of the data point/country. The learning rate is set as 0.08 and the radius is 6 as per the 12 x 12 dimension of the SOM grid. The SOM is trained for 10000 epochs and the results obtained are as follows:



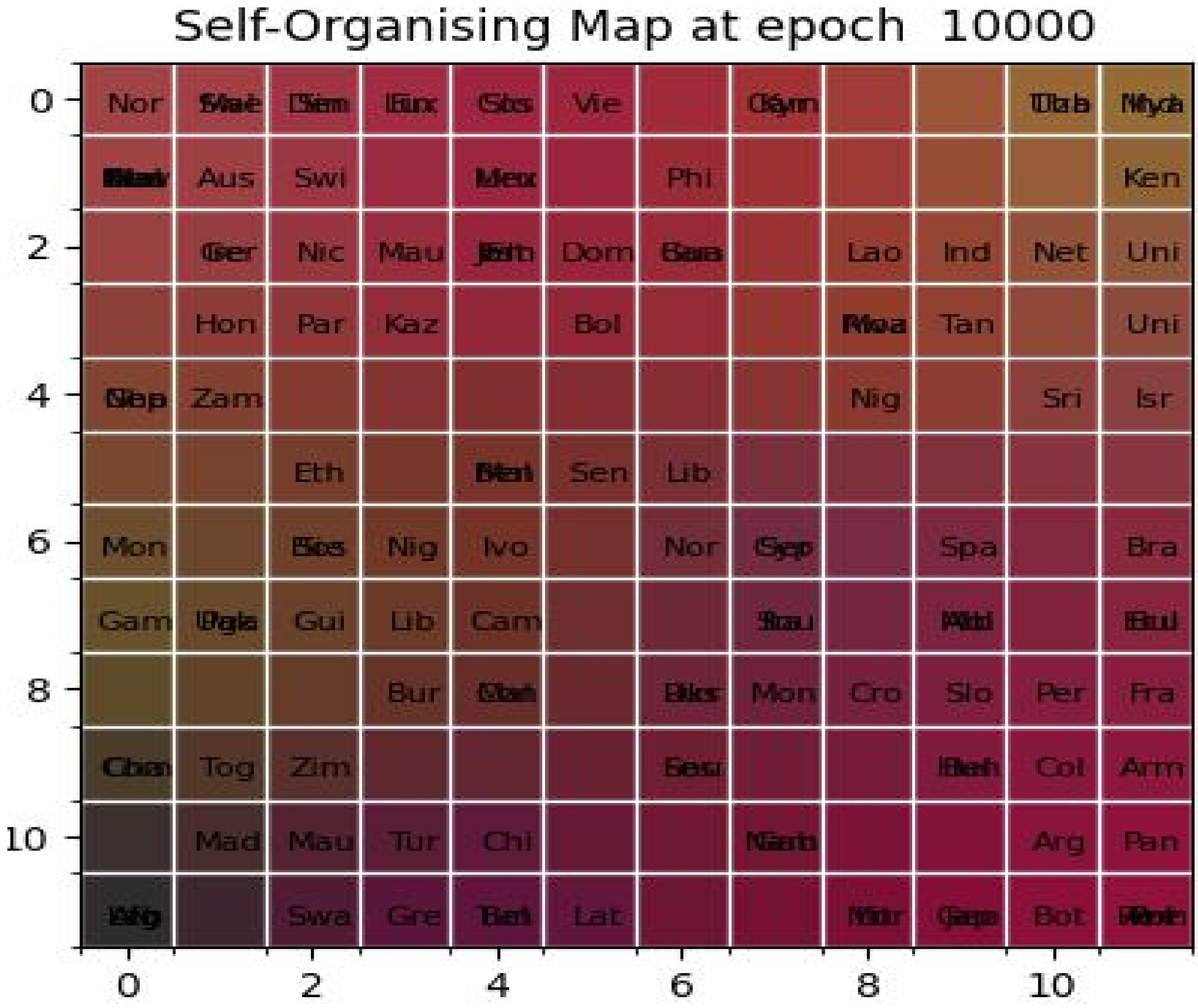
The SOM is initialized as above



Self-Organising Map at epoch 1000



Self-Organising Map at epoch 5000



It is observed that the data visibly clusters after the 10000 iteration but the clustering begins at 1000 iteration. It was also observed that a similar range of colors cover the SOM grid which is due to similarity in the attribute values of the data which maps them to similar weight values and BMUs. This is why a similar pattern of colors covers most of the grid even after random initialization.