Problem

4) a)

$$K = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \\ 0 & 0 & 10 \end{bmatrix} \qquad V = \begin{bmatrix} 1 & 0 \\ 10 & 0 \\ 100 & 5 \\ 100 & 6 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 10 & 0 \end{bmatrix}$$

$$QK^T = \begin{bmatrix} 0 & 0 & 100 & 100 \\ 0 & 100 & 0 & 0 \\ 100 & 100 & 0 & 0 \end{bmatrix}$$

$$QK^TV = \begin{bmatrix} 0 & 0 & 100 & 100 \\ 0 & 100 & 0 & 0 \\ 100 & 100 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 10 & 0 \\ 100 & 5 \\ 1000 & 6 \end{bmatrix} = \begin{bmatrix} 110000 & 1100 \\ 1000 & 0 \\ 1100 & 0 \end{bmatrix}$$

$$\text{Softmax}(QK^T)V = \begin{bmatrix} \sim 1 & \sim 0 \\ \sim 1 & \sim 0 \\ \sim 1 & \sim 0 \end{bmatrix}$$

4b)

The attention function is used to solve the bottleneck problem in recurrent model of neural networks and the vanishing gradient problem

The Soft max function is used to convert the scores to probabilities

So that they sum up to 1

$$\text{Softmax}(x) = \frac{e^{x_i}}{\sum_{x} e^{x_i}}$$