NLP HW3

1.

1.

a. can be interpreted as a categorical probability distribution because if we view it as a categorical probability distribution with n categories such that each category is I and the probability is then we can see it has all the needed properties:   
 , .

b. The categorical distribution α puts almost all of its weight on some αj when the dot product between the query q and a specific key kj is significantly larger than the dot products between the query and all other keys, indicating a strong similarity or match between q and kj compared to the other keys.

c. The output C will be very close to for the j from the last question.

d. In intuitive terms, this means that when the query q closely matches or aligns with a specific key kj compared to the other keys, the resulting output c will be strongly influenced by the vector vj associated with that key. It implies that the model focuses its attention and assigns a higher weight to the key that exhibits a stronger similarity or relevance to the query, resulting in a more pronounced impact on the final output.

2.

a. We will observe M = .

b. We will observe .