Assignment 4, Problem 1: Attention (40 points)

CSC-496/696: Natural Language Processing and Text as Data

For this problem, you may turn in a hard copy of your solutions, submit a typeset solution, or submit a scan of your handwritten solutions. You can slip your solution under the door of DMTI 108A if I am not in my office.

Let's say you were interested in the sentence, "I like turtles." We will apply a simplified version of a single step of single-headed self-attention. By simplified, I mean the version we discussed on slide 35 of Lecture 15. Define the word embeddings as follows:

$$\overrightarrow{\text{I}} = x^{(1)} = \begin{bmatrix} 0.5 & 0.3 & 0.9 \end{bmatrix}$$
$$\overrightarrow{\text{like}} = x^{(2)} = \begin{bmatrix} 0.2 & 0.8 & 1.2 \end{bmatrix}$$
$$\overrightarrow{\text{turtles}} = x^{(3)} = \begin{bmatrix} 0.3 & 1.9 & 1.0 \end{bmatrix}$$

Task 1.1 (5 points)

Let's say we were trying to calculate $z^{(1)}$, the transformed feature associated with $x^{(1)}$. Which vectors are the query, keys, and values?

Task 1.2 (5 points)

Calculate the unnormalized dot-product attention scores between $x^{(1)}$ and all other vectors. In other words, calculate

$$\text{score}(x^{(1)}, x^{(1)})$$
$$\text{score}(x^{(1)}, x^{(2)})$$
$$\text{score}(x^{(1)}, x^{(3)})$$

Task 1.3 (15 points)

We'll now calculate the attention weights. In other words, calculate

$$\alpha_{11} = \text{softmax}\left(\text{score}(x^{(1)}, x^{(1)})\right)$$
$$\alpha_{12} = \text{softmax}\left(\text{score}(x^{(1)}, x^{(2)})\right)$$
$$\alpha_{13} = \text{softmax}\left(\text{score}(x^{(1)}, x^{(3)})\right)$$

Verify this is a proper probability distribution (i.e., all terms are between 0 and 1, and all terms sum up to 1).

Task 1.4 (5 points)

Using the previous parts, calculate $z^{(1)}$. In other words, calculate:

$$z^{(1)} = \sum_{t=1}^{3} \alpha_{1t} x^{(t)}$$

Remember, $z^{(1)}$ should be a vector.

Task 1.5 (10 points)

Using the same steps above, calculate $z^{(2)}$ and $z^{(3)}$. You don't need to verify that the attention weights add up to 1, although it could be a good consistency check.

Task 1.6 (Optional, 5 extra credit points)

Let's say we were to apply another step of this simplified single-headed self-attention. What would your three transformed features be?