Homework 3

Answer 3 of the 4 questions below.

1. a. Implement the code of the function self_attention, below

```
def self_attention(queries, keys, values, mask=None):
    """Implement vanilla self-attention"""
    pass
```

b. Implement the code of the function multi_head_attention

```
def multi_head_attention(x, num_heads=8):
    """Implement multi-head attention with head splitting"""
    pass
```

c. Apply both functions to the polysemy example sentences you created as an answer to question 2 in HW2.

2. In HW2 question 1 – you applied Word2Vec to a set of sentences.

a. Apply self attention to the same set of sentences and visualize the attention patterns of the words: bank, rose, lead, book and file.

b. Compare with Word2Vec results from HW2, while quantifying how attention captures contextual meaning vs. static embeddings.

3. Consider a transformer with 8 attention heads. Show mathematically why splitting attention into multiple heads can be more effective than using a single attention mechanism with the same total computational cost.

4. a. Implement sinusoidal positional encoding.

b. Analyze the impact of sequence length on its performance.