Name: 2025-03-14

NLP 2 LLM – Exam

1. You have to do a complete pre-processing on your documents to apply traditional NLP methods.

   Document 1: *"Cats running in the garden."*
   Document 2: *"The dog barks loudly when the letter carrier starts to run!"*
   Document 3: *"My friend had a very cute dog and cat."*

   - Normalize the text by removing punctuation, converting it to lowercase, removing stopwords [*"a", "the", "and", "to"*], and applying lemmatization.
   - Tokenize the documents using whitespace tokenization.
   - Write the Bag of Words representation for all documents.

2. Calculate the TF-IDF for *"dog"* and *"nice"* for each document.

   Document 1: *"My dog is a nice dog"*
   Document 2: *"I like this dog"*
   Document 3: *"The weather is nice"*

3. Your company wants to build a supervised Machine Learning (ML) model for spam detection in emails.

    - What are the key pre-processing steps you would apply to the text data before training a model?
    - Name and briefly describe two methods to convert text into numerical features.
    - Which ML model would you choose for this task and why?
    - Name two evaluation metrics suitable for this classification task and explain whether there is a benefit to using one over the other.

4. In a transformer model trained for the text classification task, the hidden state of the decoder (the tensor which is the output of the decoder and the input of the classification head) is of shape $(16, 1000, 512)$.

   - What do the numbers 16, 1000 and 512 correspond to?
   - What is the shape of the tensor outputted by the classification head?

5. The following formula is an important formula in the transformer architecture:

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Answer the following questions assuming the model only uses one attention head:

- What is the complete name of the above formula ?
- What are the names of the matrices Q, K and V ?
- Assuming we are using this formula inside a **self-attention** module, give an example of possible shapes of the matrices Q, K, V.
- Assuming we are using this formula inside a **cross-attention** module, give an example of possible shapes of the matrices Q, K, V where not all of the three matrices have the same shape.

6. We are given the following batch composed of two tokenized sequences:

The	dog	is	hungry	<EOS>	<PAD>	<PAD>
Do	not	cross	<EOS>	<PAD>	<PAD>	<PAD>

- Write down the padding masks M_1^{pad}, M_2^{pad} (as matrices) of the two sentences in the batch.
- Write down the causal masks $M_1^{causal}, M_2^{causal}$ (as matrices) of the two sentences in the batch.
- Write down the attention masks M_1, M_2 obtained by combining the padding masks with the causal masks, for the two sentences in the batch.
- How would you encode the information in the padding masks in a more efficient way?
- How would you encode the information in the causal masks in a more efficient way?