Université d'Orléans **2025/2026**
Master Physique Fondamentale et Applications (PFA)

Applied machine learning: Data preparation
Problem sheet 1

1. Images

1. Load the datasets:
   (a) Load the MNIST dataset into a 2D numpy array.
   (b) Load the Letter dataset into a 4D numpy array.
2. Normalize and regularize the datasets.
3. Process the labels:
   (a) Load the labels for both datasets into separate numpy arrays.
   (b) Encode the class labels as 1-hot encodings.

2. Text

1. Implement the byte pair encoding (BPE) tokenizer.
2. In addition:
   (a) Write a function to encode (text to list of token ids).
   (b) Write a function to decode (list of token ids to text).
3. Use your BPE tokenizer to process the tiny Shakespeare dataset:
   (a) Tokenize 10% of the dataset.
   (b) Use a vocabulary size of 1000.
   (c) Include all characters (a-z and A-Z) as well as all numbers 0-9 in the initial vocabulary.
4. (Optional) Repeat task 2 using the tiktoken library and compare the results with your implementation.

3. Bag of Words

1. Construct bag of words models for the following:
   (a) The second 10% of the tiny Shakespeare dataset.
   (b) The third 10% of the tiny Shakespeare dataset.
   (c) The headlines dataset.
2. Compare and contrast the three bag of words models.