General AI exercises

1st problem. We are given a dataset that contains product reviews and a rating for each product. A product can be rated from one to five. Clearly, the review is highly correlated with the final score that the users give to the product. The dataset provided contains 10K samples. We would like you to analyze the dataset and build a model that is able to predict from a given review its score. For this purpose, the use of open-source libraries is encouraged.

You can download the dataset from https://app.goldenspear.com/ratings.csv

1. **Data analysis.** Plot the balance of classes and show the five most predominant words for each class.
2. **Data cleaning.** Have you performed data cleaning? If so, what kind of data cleaning and which tools have you used?
3. **Learning process.** Answer briefly these questions:
    a. What kind of features have you used?
    b. What model or models have you chosen? Why?
    c. What libraries have you used?
4. **Models validation.** Evaluate the performance of your estimator using some validation methods and answer these questions:
    a. What validation method have you chosen?
    b. What evaluation metric have you chosen?
    c. Write down your training and testing accuracies.
5. **Final summary**. Write down what would you have done if we had given you more time and data.

2nd problem. We are given a dataset that contains an unlabelled set of images and a smaller dataset that contains labeled images. Those images represent shirts from e-commerce, so usually, the background is plain and the product is the focus of the image. The unlabelled dataset provided contains 5K samples and the labeled one contains just 10 images, all **blue** shirts.

We would like you to analyze the datasets and build a model that is able to retrieve all the blue shirts from the 5K dataset. For this purpose, the use of open-source libraries is encouraged.

You can download the datasets from https://app.goldenspear.com/shirts.tar.gz

1. **Problem analysis.** Write down your solution proposal.
2. **Models validation.** As we don't have validation samples on the dataset, find a way to visually demonstrate the capacity of the model.
3. **Final summary**. Write down what would you have done if we had given you more time and data.