

1 Introduction

1.1 Context

A mysterious software company (Acme) needs your help to automate the answering of security questions. This company provides you with a dataset of past questions/answers and wants you to develop a model capable of predicting the answer to any question from this data.

1.2 Dataset

You have two data files :

- A file `training.json` provided by Acme. It is a database extract containing (among other) a *question* field and an *answer* field
- A `test.csv` file including :
 - A column *question_id*, a unique identifier of the question
 - A column *question* the question text. You need to predict the answer to this question.

1.3 Assignment

Your submission must include at least :

- A `predictions.csv` file containing your results
- The code used to generate this file
- A report containing your answers and an explanation of your approach

If you code on a notebook, you can also use it for your report.

You can write your answers in French or English, as you prefer. The programming language used for the code must be Python.

You are free to use any library (including pre-trained models) as long as you explain your approach and the reasons for your choices. Special attention will be given to the clarity of explanations.

2 Instructions

You have complete freedom to define your methodology. Your report should clearly outline your thought process, analyses, choices, and conclusions. Examples of expected elements might include :

- Exploratory data analysis (e.g., distribution analysis, visualization).
- Pre-processing choices and justification.
- Choice of learning paradigm (e.g. supervised, unsupervised, reinforcement learning...).
- Model selection strategy and justification.
- Evaluation methodology and chosen metrics.
- Explanation on the resources used to tackle the problem (articles, code libraries, etc.)

However, these elements are examples only and should not limit your creativity or approach. The key objective is to showcase a coherent and justified approach tailored to the given problem.