MINI PROJECT 5: WORKING WITH TEXT

A great part of the information about the world comes to us as text. To be able to process, analyse and generate text automatically, often we need to convert it into numeric data sets (vectors).

The process of converting or transforming data into a set of vectors is called vectorization. **Text vectorisation** is an essential prerequisite of the modern Natural Language Processing (NLP) and Understanding (NLU), maintained by the Generative AI.

The **objectives** of this project are:

- understanding the basic concepts and use of text vectorisation and vector similarity
- gaining experience in implementation of methods, algorithms, and libraries for working with text in BI and Python programming.

Your **tasks** are the following:

1. Collect and load text documents from various sources of one domain – e.g. some of txt, doc, csv, json, pdf files, web pages, or data frame attributes.
2. Extract, clean, and transform the text from the sources, to prepare it for vectorisation.
3. Vectorise and store the clean text in a software structure.
4. Create a simple interactive prototype of application, which can input a text from a user and output the top three related pieces of texts, stored earlier, applying **vector similarity** approach.
5. Optionally, integrate your application with LLM (large language model) for improving the quality of the language operations.
6. Suggest various implementations of such an application.

Enjoy you first personal language assistant!