Introduction

Chapter 1

Realization

In order to implement the proposed solution, we have proposed a proof of concept which is a web application built with streamlit that demonstrates the different features that can be offered.

The use case that we consider in this proof of concept has the following specifications:

- The user can select a dataset to be added and indexed.
- The user can have the previously indexed datasets.
- The user can select a dataset, that we call query dataset, and measure its similarities with the other ones.
- The computed similarities are presented in two sections:
    1. Dataset similarity representation: It orders the dataset by their similarity scores with the query datasets.
    2. Column similarity representation: It shows the details of the column similarities between the query dataset and a dataset that the user selects.

1.1 Technologies and tools used

In order to implement the proposed solution, we have opted for the following technologies, tools and frameworks:

1.1.1 Python

Python is a structured, open source, multi-paradigm, multi-platform programming language that runs on all major operating systems and computing platforms. By offering high-level tools and an easy-to-use syntax, it greatly optimizes the productivity of programmers and has become a leading language in exploratory data analysis and software development.

Python have been chosen for implementing this project for the following reasons:

- It provides a set of libraries, in the machine learning domain, which simplifies the development of our project.
- Python manages its resources (memory, file descriptors) without the intervention of the programmer, by a reference counting mechanism.

- The Lab team of Talend has already used Python in most of their projects, this makes it easier to understand, collaborate, scale and integrate our solution in the future.
-

We follow up by presenting some of the used libraries during this projet:

Numpy

A library used to manipulate matrices, multidimensional arrays, vectors and polynomials, it has a large number of mathematical functions that can be applied directly to the structures mentioned above. In this project we have used numpy to process matrices and arrays.

Jax

Following its definition in its official repository (Google, 2022), Google JAX is a machine learning framework for transforming numerical functions, the goal of using this framework is to take advantage from its version of Numpy that provides us with the possibility of doing the different calculations on GPU.

Pandas

Matplotlib

IGraph

Plotly

Bibliography

Google (2022). Jax. https://github.com/google/jax.