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:

1. **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.

2. **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. It shows up that this choice is accurate if we take in consideration the number of matrix multiplication that we have to do in Cross-Polytope.

3. **Pandas**: Pandas is a python library that allows to easily manipulate data in form of data tables that we call DataFrame with column and row labels. It provides us with some interesting easily used features to read, analyze and complete some preprocessing operations.

4. **IGraph**: It is a library allowing to define and manipulate graphs, it is written in C programming language and can be used with Python or R. We use it in our solution to define the graphs representing the results of similarities between columns and between datasets

5. **Plotly**: Plotly is a complete library for creating static, animated and interactive visualizations in Python. It is used in our solution to represent the graphs defined with IGraph with an interactive figures in the streamlit application.

6. **Transformers**: We use the Transformers library which is a state of the art pre-trained model library for natural language processing (NLP). The library currently contains pre-trained model weights, user scripts and conversion utilities: Bert, Roberta...

1.1.2 Streamlit

It is an open source framework in Python language dedicated for data scientist and machine learning engineers to create complete web applications in a short time as it provides a set of components that we need to have an interactive layout. It also provides a compatibility with different Python libraries which make it easier for building demonstrations that are not exclusive to data scientists.

Stramlit has also a cloud platform that gives its users the possiblity to deploy, manage and share their apps.

1.1.3 TensorFlow Hub

It is a repository of trained machine learning models ready for downloading, fine-tuning or reusing in different projects, with the possiblity to be deployed everywhere.

We used this library to test different embedding models and choose the most suitable one in order to get a representation for the textual columns.

Bibliography

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