Dialectica - Data Science Assignment

Introduction

Welcome to our Data Science challenge! This assignment is designed to test your skills in handling real-world data, performing exploratory data analysis (EDA), utilizing SQL databases, applying machine learning techniques using natural language processing (NLP), writing reports and writing clean and readable code according to standards.

Objective

Your task is to develop an application that processes provided datasets (in a SQLite db), integrates and cleans the data, performs an analysis, and builds a predictive model.

Data Description

You will be provided with two datasets in tables format inside a SQLite database. The datasets are as follows:

1. **Company Dataset 1**: Contains basic company information.
2. **Company Dataset 2**: Includes detailed attributes about companies.

https://drive.google.com/file/d/1bMYzu0pAFZz_nG7fsLogNsTlvg3ykix_/view?usp=share_link

Tasks

1. SQL Queries

- Craft SQL queries to answer the following questions:
  - Find the top 10 industries with the highest average number of employees, only considering companies founded after 2000 that have more than 10 employees. (CompanyDataset)
  - Identify companies in the 'Technology'-like industry that do not have effective 'homepage_text' and have fewer than 100 employees based on data merged from both datasets.
  - Rank companies within each country by their total employee estimate in descending order, showing only companies that rank in the top 5 within their country. (CompanyDataset)
- Ensure that your queries and db are optimized for performance (Use Indexes).

2. Data Integration and Database Insertion

- Merge the two company datasets into one using high-quality merging techniques.
- Ensure that the data is clean and well-prepared for analysis.
- Load the merged dataset into a new table in the SQLite database.
- Develop SQL queries to efficiently store and retrieve data.

3. Exploratory Data Analysis (EDA)

- Perform a detailed EDA on the merged dataset.
- Visualize key aspects of the data and generate insights that could aid in model building.
- Document your findings and hypotheses from the EDA.

4. Model Development

- Use the company data to train a model (or more) to predict the category of each company.
- Preprocess the text data, set up the model(s), and explain your choice of architecture and parameters.
- At least one of the models should be an LLM (e.g. Bert)

5. Model Evaluation

- Evaluate the performance of your model using appropriate metrics.
- Provide a detailed analysis of the model's performance and discuss any potential improvements.

6. Reporting

- Compile a report summarizing the methodologies used, insights gained, model performance, and any challenges faced during the tasks.
- Your report should be clear, well-organized, and easy to read.

Submission Guidelines

- Your final submission should include the source code, the SQLite database file, the final model file, and the report.
- Ensure all code is well-commented and organized. Your code should follow software design principles, being clear, easy to read and efficient. Make sure it is clear how to run your code (provide a Readme). We should be able to quickly test your code.
- You should upload the whole project in a Google Drive folder, grant read access and share the link with us in an email (include all persons listed in cc).
- The deadline for the submission is 5 days from the day you receive the assignment.

Evaluation Criteria

- Accuracy and efficiency of SQL queries.
- Thoroughness and insightfulness of the EDA.
- Performance and robustness of the predictive model.
- Quality and clarity of the final report.