INNOVIUS

Technical Case Study

Objective: Develop a solution to compute and visualize similarity scores between companies based on their descriptions.

Background: We often need to analyze and compare various companies in our database to make informed investment decisions. Being able to quantify how similar companies are based on their business descriptions can provide valuable insights into market trends, competition, and potential investment opportunities.

Data: Attached in the email is a CSV containing a dataset of ~55k company descriptions. The dataset includes:

- Company ID
- Name & Website
- Company Description(s) – we often get similar data attributes from various data-sources, use your discretion on how to use them
- Industry Tags: Top & Secondary Category
- Employee Count

Note: Missing values are a feature, not a bug of the case study dataset. We often deal with incomplete data from vendors, either due to vendor coverage, budget constraints, or incorrect entity mapping & resolution.

Tasks

1. **Data Preprocessing**
   - Clean and preprocess the text data.
   - Explain the preprocessing steps and their importance.

2. **Feature Extraction**
   - Convert text data into a suitable numerical format using techniques like TF-IDF, word embeddings, or others.
   - Discuss the chosen method and its impact on the similarity analysis.

3. **Similarity Score Calculation**
   - Implement a method to calculate similarity scores between company descriptions.
   - Provide a rationale for the chosen similarity measure.

4. **Data Engineering & Scalability**
   - Design a data pipeline that can efficiently process and update similarity scores as new companies are added to the database. *(Note, there is no need to create anything here, just looking for systems & data engineering thinking)*
   - Address scalability and performance considerations.

5. **Visualization & Output**

Note: Two options below are not exhaustive, feel free to choose an output that best showcases your ability to think through & execute on this problem.

Option 1: R/Python notebook that demonstrates the workflow and outputs similarity scores in a clear and interpretable format + design document outlining the architecture of the solution, data flow, and user interface design for a web application.

Option 2: Simple web application (e.g., using Streamlit, Flask, Vercel) that allows users to select companies and view their similarity scores with other companies. If choosing this approach, be sure to document & share the underlying code behind the webapp.

In your final submission, also include a brief discussion of next steps & improvements.

Evaluation Criteria

This case is designed to assess a wide range of skills relevant to a data engineering role in a VC setting, from technical proficiency in data processing and analysis to the ability to conceptualize and design scalable systems.

- ✓ **Code Quality:** Readability, use of best practices, and documentation.
- ✓ **Data Handling:** Effectiveness of data preprocessing & feature extraction.
- ✓ **Algorithm Design**: Correctness and efficiency of the similarity score calculation.
- ✓ **Scalability**: Consideration of scalability and performance in the data pipeline design.
- ✓ **Output Quality:** Clarity and usability of the final output (notebook, web application, and/or conceptual write-up).