Project Proposal: DataFrame Toolbox for Working with Tabular Data

Author: Irena Torosyan

Supervisor: Norayr Chilingaryan

Affiliation: American University of Armenia, Data Science Department

Date: January 26, 2024

1. Introduction

Working with data often involves repetitive tasks on structured formats including CSVs, Excel spreadsheets, JSON, and more. This project aims to provide a user-friendly DataFrame tool, incorporating a DataFrame class that offers functionality similar to the popular pandas library. This combination strikes a balance between simplicity and flexibility.

2. Objectives

- Develop a user-friendly, efficient DataFrame class for reading, writing, and modifying data.
- Implement essential functionalities like find/replace, data filtering, and basic statistical analysis.
- Explore advanced features like regular expression capabilities and dynamic column typing for increased flexibility.
- Provide various file format support (e.g., CSV, ODS, XLSX, JSON) for enhanced usability. Provide an interface for loading, working with, and saving the data in various formats.

3. Target Audience

Data analysts, data scientists, system administrators, and anyone who regularly works with simple datasets in CSV-like formats.

4. Project Scope

This project will be divided into two main phases:

Phase 1: Core Functionality

- Develop the core DataFrame class and parsing engine.
- Implement functionalities for reading, writing, and basic data manipulation (e.g., filtering, sorting).
- Allow loading and saving data in different formats.
- Integrate find/replace with basic string operations.
- Conduct testing and finalize documentation for core functionalities. Compare the performance to tools with a similar functionality (e.g. pandas library)

Phase 2: Advanced Features

- Develop a simple regex library for more sophisticated data modification.
- Implement dynamic column typing based on header information.
- Continue to expand file format support.
- Conduct further testing and expand documentation.

5. Deliverables

- A functional DataFrame class.
- A commandline utility.
- Comprehensive documentation with usage instructions.
- Project report.
- Open-source release.

6. Timeline

- **Month 1**: Design, research, and framework development.
- **Month 2**: Core functionality implementation and testing. Documentation and initial testing and comparison with similar tools.
- **Month 3**: Advanced features development and testing.
- **Until Capstone Defense**: Code refinement and final testing. Finalizing the results in the report.

7. Conclusion

This project addresses the need for a lightweight tool that efficiently handles tabular data in various formats. The blend of simplicity and flexibility has the potential to improve data processing and user experience in various use cases. We are committed to developing a valuable tool that can evolve into a community-driven project, benefiting a wide range of data users.