Data Wrangling & Visualization: Project Proposal

Interactive Visualization of `UMAP` Algorithm on User Data

Nikita Zagainov, Dmitry Tetkin, Nikita Tsukanov

February 2025

1 Project Vision and Goal

This project aims to create an interactive website for explaining one of the most powerful dimensionality reduction algorithms, `UMAP` (Uniform Manifold Approximation and Projection).

We aim to provide a comprehensive explanation of the algorithm, and demonstrate its application to table data, which can be chosen by the user

2 Data

The data for our project is arbitrarily chosen by the user, a few restriction apply to the data:

- The data should be in a table format
- The data should be in a `.csv` format
- The data should contain at least 2 columns (for `2D` visualization) or 3 columns (for `3D` visualization)
- All columns should be numerical or categorical (arbitrary data type). Irrelevant columns will be ignored
- Optionally, the user can provide a label column in form of additional `.csv` file, which will be used for coloring the data points

For example purposes, we will provide `MNIST` (can be found here) dataset in `.csv` format, and visualization of this data.

3 App Architecture & Pipeline

We omit data scrapping step, as it is not required for this project. All other obligatory steps are provided:

1. **Data Cleaning and Preprocessing:** to make sure that our engine works with the data, first step of our pipeline cleans and preprocesses the input data, and performs necessary checks
2. **UMAP Algorithm:** this step applies `UMAP` algorithm to the input data and streams the resulting embeddings on each iteration to the visualization app. For algorithm implementation, we will use original `UMAP` implementation as backbone with some modifications (control over NN search algorithm, embedding streaming) in `NumPy`, `SciPy`
3. **Data Delivery:** for streaming purposes, we will use `REST API` to deliver the embeddings to the frontend. For implementation, we will use `FastAPI` library
4. **Visualization App:** the embeddings are streamed to the visualization app, which will display the process in real time in 2 or 3 dimensions, allowing user to interact with the visualization (zoom, pan, rotate, etc.). Fronted app will use `D3.js` library for visualization

4 Visualization Features

Proposed features of the app:

- We aim to provide as much interactivity as possible, allowing the user to insert their own data in `.csv` format
- Optionally, the user can provide a label column in form of additional `.csv` file, which will be used for coloring the data points
- For better understanding of the algorithm, we allow user to set desired parameters of the `UMAP` algorithm which will be applied to the data (2 or 3 dimensions, number of neighbors, etc.)
- Real-time process of the algorithm will be displayed on the page, with ability to pause, rewind, fast-forward, and navigate through the embedding space

5 Project Timeline and Milestones

We cannot be certain about the exact timeline of the project, but we provide a rough estimation of the project milestones:

- **Week 1-2:** Project Setup, gathering with a team, finding hosting service, creating basic frontend page
- **Week 3:** `UMAP` algorithm implementation, possibly some optimizations and testing
- **Week 4-5:** Creating pipeline for effective data delivery
- **Week 6-7:** Creating frontend visualization app
- **Week 8:** Final testing and deployment, presentation preparation