Visualization Dashboard for single-cell RNA transcriptomics data

LDATA2010 Project

Data is becoming increasingly important in our society, while datasets are becoming larger and larger every day. Visualizing those datasets is an important part of data science, and various algorithms have been designed to help the visualization of big data.

In this project, you will have the opportunity to create a *specialized* visualization dashboard. This includes getting familiar with different chart types, exploring different visualization algorithms, depicting multiple views of the data, etc. Obviously the aim of the project is *not* to develop a generic visualization software. Instead you are asked to implement only the features necessary for the dataset provided, with a careful look at user experience and user interaction.

1 Dataset: single-cell RNA transcriptomics in the brain of a rat

ℹ Dataset Context

RNA transcriptomics is a technology allowing biologists to read the RNA transcripts within individual cells. This allows the biologists to analyze the expression of specific genes in cells at a given time, which can reveal intra-tissular heterogeneities that would remain hidden with other methods.
Datasets produced with transcriptomics are often large, with hundred of thousands of cells (observations) and hundreds or thousands of genes (features). Biologists require data analytics tools to visualize their data and build their intuition on the studied cells.

In this dataset, biologists read the expression levels of 300 genes in 23,822 different cells of a rat brain. The cells belong to three general families: non-neurons, excitatory neurons, and inhibitory neurons. Further distinctions between cells appear within each of these families for a total of 133 cell types.

The first 300 columns in transcriptomics_data.csv contain the expression level of genes as floating point values. The second to last column contains the identifier of the cell type, taking integer values in [0, 132]. The last column contains a hexadecimal code for the colour given by biologists to each cell type.

1.1 Exploratory analysis

Your first task is to explore and understand the purpose and contents of the file. Without this knowledge, it will be difficult to create a dashboard appropriate to the dataset. However, there's no need to become an expert !

You should also be able to answer basic questions like : Is the dataset complete ? What should you do with missing data? How should you deal with categorical values ?

Before starting to design your visualization tool, you should ask yourself what questions a user might ask himself. What does the user want to discover from the data ? What are the important relationships to know ?

2 Basic Features (Deadline 1)

You will develop a software with the following features:

1. A user will be able to display multiple views of the data, to explore different facets of its structure using different chart types.
2. You will enable the user to compute some basic properties and metrics of the dataset. The user will have the possibility to highlight these properties on one or multiple plots.
3. The user will be able to filter the data according to some values or attributes. This should update all the visible plots, metrics and properties.
4. The user should be able to select some data in one plot to update the other plots accordingly.
5. The user should be able to visually see which genes are correlated.

User Interface

While the user interface (the placement of views, charts, buttons, ...) is left up to you, be careful that it is an integral aspect in the design of the application.

3 Clustering and Dimensionality Reduction (Deadline 2)

Once you have implemented the global plotting capabilities of your interface, you will now have to dive deeper in advanced aspect of information visualization. For this, we ask that you implement different views of the data using clustering and dimensionality reduction algorithms:

- Clustering : What can you use clusters for ? How will you show them in your interface ? Can you use cross-filtering with a hierachical clustering ? How does changing the parameters of the algorithms change the clustering ?
  - Implement K-Means
  - Implement a hierarchical clustering algorithm of your choice
  - Implement a density-based clustering algorithm of your choice
- Dimensionality reduction : Can a clustering algorithm help you in choosing an effective DR algorithm ? Do you detect visual clusters when you visualize a DR algorithm ? What parameter should you choose and which should you enable the user to change ?
  - Implement a linear DR algorithm
  - Implement a non-linear DR algorithm of your choice

4 Report

In addition to your software, you are asked to provide a small report (maximum 7 pages) detailing the features that you have implemented. In particular,

- You can write your report as a user guide for your software.
- Explain and justify your design choices.
- Cite the toolboxes, sources and papers that you employed. There is no restriction on the sources you use nor on the papers that you read, but you have to cite them.
- Reasonably detail the algorithms that you have employed (e.g. by providing an overview of each one of them without the practical implementation details) and justify why you chose them.
- Provide examples on how to use your software, illustrating its capabilities in terms of scaling, interaction and visualization.
- Give some ideas on how to improve your software. Which features might be worth implementing in future versions? How could you make your software more scalable?

5 Project Structure

The project will be structured in three parts, with a deadline and a deliverable for each part.

5.1 Global analysis and first draft of final UI

- Deadline : 08/11/2022
- Deliverable : An analysis of the data using standard plots, as shown in the exercise session.
- One-on-one sessions : In order to start on good foundations, each group will have a (15 minute) one-on-one with the teaching assistant on the 08/11/2022 or 11/11/2022, during the exercise session. It will happen either face-to-face or on Teams, depending on choice. During this meeting, you should already have :
  - An understanding of the dataset;
  - A User Interface helpful for exploratory analysis
  - A sketch/draft of the final user interface. You might for example use pen-and-paper or `https://excalidraw.com/` to draw the layout of your application.

5.2 Clustering and Dimensionality Reduction

- Deadline : 25/11/2022
- Deliverable : Complete UI with both global dataset analysis, dimensionality reduction applied to the dataset and clustering visualization and report detailed in Section 4
- Peer review : After this deadline, you will receive two projects to review individually. You will then receive those reviews (2-4) to help you finalize the project.

5.3 Final Deliverable

- Deadline : 16/12/2022
- Deliverable : Complete project and report, corrected according to peer review
- Project presentation : During the exam, you will be asked to present your visualization tool.

6 Practical information

- **Groups:** You can complete the project alone or by groups of 2 students. Please join a group on Moodle (even if you're alone).
- **Programming language:** you can use the one you prefer (Python, Matlab, R, etc.), but Python is recommended. You can use all the toolboxes, packages, modules, etc., that you find relevant. You can use toolboxes to help you designing the interactive user interface, but you cannot just rely on an already existing interface.
- **Final Deadline for the project submission on Moodle:** Friday December 16, 15pm. Submit one .zip file per group, containing report and code.
- Do not hesitate to ask questions to the teaching assistants before or after your planned one-on-one, for instance to define what you plan to implement, the network metrics that you could evaluate, etc.

ℹ installation of the software

Don't forget to add a requirements.txt file containing the necessary libraries, one way to build requirements.txt is by using *pipreqs*. The software should be easy to install at the moment of evaluation.
Also, all file paths should be relative, not absolute!