Hmw III: SVD methods and Elections Data

2025-05-29

> **Important**
>
> - Due : May 29 2025
> - Work in pairs
> - Deliver your work through a github repository
> - Present your work (15 minutes) on *2025-06-03*

This homework is about

1. Using Matrix Factorization methods in Data analysis
2. Investigating voting patterns in Paris

I. **Voting Data**

Voting data per polling station can be obtained from a variety of websites.

- An example
- Another example
- Yet another one
- Opendatasoft API
- https://opendata.paris.fr
- https://data.smartidf.services/pages/data/
- https://data.opendatasoft.com

> **Note**
>
> Many datasets are available in several formats. When possible, use `parquet`. `parquet` files can be uploaded using package `arrow`.

Data concerning polling stations can also be gathered from various sources.

- https://github.com/datagouv/bureau-vote
- https://opendata.paris.fr

You first task will be to design an *extraction pipeline* to obtain the voting data you will analyse. You will gather data corresponding to different types of elections (Municipales, Régionales, Législatives, Européennes, Présidentielles) that took place since Year 2000.

title	rounds	year
European P	1	2004
Parliament	2	2007
European P	1	2019
Régionales	2	2021
European P	1	2024

II. **Data cleaning**

Some data cleaning may be necessary, for example

- Some parties changed their names during the last 25 years. Defining a mapping can facilitate the comparison of results from different elections
- Check that the names of `bulletins nuls`, `bulletins blancs`, … are consistent across the different datasets.

> **Note**
>
> Design a cleaning pipeline. Save the cleaned data.

III. **Applying Matrix Factorization Methods (SVD)**

For one election round, the outcome is summarized by a `tibble` where rows (individuals) are polling stations and variables/columns are the number of votes obtained by the different condidates/parties.

Perform PCA on different elections. Visualize and describe the result (attention, this is data analysis, not political science).

You may also perform CCA to compare different elections.

Feel free to combine different methods.

IV. **Clustering**

Perform clustering on the outcomes of the Principal Component Analyses.

References

- Advanced R Programming
  - Functional programming
  - S3
  - Meta programming
- Packages
- Programming with/for `ggplot2`
- Programming with `dplyr`
- `tidyeval` helpers
- Cheatsheets

Table 2: 🎓 Grading criteria

Criterion	Points	Details
Documentation/Report	45%	English/French 🖌
Presentation	40%	
Data gathering/cleaning pipelines	15%	