Final Project, Spring 2018

Python for Data Management And Analytics

Guidelines

- The final project is team-based
- Each team can consist of 2-3 students
- All written reports should be submitted in PDF format
- All code must be submitted as Python Jupyter Notebook scripts

Milestones

- **(Due 10/22)** The abstract is just a short paragraph describing:
  - The project name and member(s)
  - The research question you have chosen
  - Why is the research question interesting to you
  - What is your overall plan for data analysis (this may change as you work on your project, but we want you to at least start planning ahead)
- **(Due 11/12)** Data story

  Note: by this point you must merge, clean, and transform your data
  - The title and member(s)
  - The research question
  - Why is the research question interesting to you
  - Feature selection: the list of variables that you are planning on using in your analysis and the justification explaining why you are using these particular variables.
  - Data story: tell a story of your data using descriptive statistics and visualizations.
- **(Due 12/10)** Final report
  - The final paper should be 6-8 pages in length following and contain the following sections:
    i. Abstract
    ii. Introduction
    iii. Methodology
    iv. Results
    v. Discussion
    vi. References

- You should describe your problem, approach, dataset, data analysis, evaluation, discussion, references, and so on, in sufficient details, and you need to show supporting evidence in tables and/or figures.
- You need to provide captions for all tables and figures.
- You should also briefly describe how each member contributes to the total work in the end of the report.

The final project will be graded largely based on the final report.

Late submissions will not be accepted.

The Data

You can use the data sources listed below, but you can also find any relevant datasets on your own if you think that they are useful to your analysis.

- FBI Crime Data:
  - https://crime-data-explorer.fr.cloud.gov/
  - https://github.com/fbi-cde/crime-data-frontend/blob/master/README.md
  - https://www.perspectiveapi.com/
- Centers for Disease Control Data: https://www.cdc.gov/datastatistics/index.html
- World Health Organization Data: https://www.who.int/gho/database/en/
- US Government Open Data: https://www.data.gov/
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.html
- FiveThirtyEight - all kinds of datasets from virtually every domain: https://data.fivethirtyeight.com/
- Amazon Web Services (AWS) datasets: https://registry.opendata.aws/
- Google public datasets: https://cloud.google.com/bigquery/public-data/
- Wikipedia datasets: https://en.wikipedia.org/wiki/Wikipedia:Database_download
- Kaggle: https://www.kaggle.com/

Evaluation Rubric

- Technical strength / implementation: 30%
- Experimental evaluation: 30%
- Explanation of results: 30%
- Overall project management: 10%

Technical strength and experimental evaluation

Technical strength: The pedagogical goal of this project is for you to practice how to design, implement and evaluate data mining techniques and how to apply them in practical scenarios. In order to show the technical strength in your project, you need to provide sufficient details in your written report. Specifically, what did you try to accomplish? What method (data mining techniques) did you adopt? What are the strengths of your method? Describe the systematic approach how you achieve your specific goal, and justify your method with supporting argument through evaluation.

Experimental evaluation: Describe your experimental results with systematic evaluation in your oral presentation and written report. Specifically, describe the evaluation criteria, including performance measures and baseline methods. Explicitly state the performance of your method, and provide whatever results and insight you have gained from the evaluation.

Honor code: You may consult any papers, books, online references, or publicly available implementations for ideas and code that you may want to incorporate into your project, as long as you clearly cite the sources in your code and your written report.

Implementation

You need to submit Python code for your project that is working and is able to produce the same experimental results as described in your final report. Hence, the experimental procedure and the testing data must be submitted together with necessary documentation describing how to run the experiment.

Overall project management

You need to submit required materials to meet each milestone as described in this document. These milestones are to help you seek resources you may need, make sure you develop a concrete direction, and better manage your project. Before each milestone, if you find it difficult to meet the requirement, you should make an appointment and discuss with me in advance.