DS-GA 1007: Programming for Data Science

Final Project Requirements

The final project will be an opportunity to explore applications in your areas of interest. Groups can work on either

- Data Science Project
- Software Engineering Project

Throughout the project, groups will have to work collaboratively to meet goals – managing expectations along the way. Like we discussed in class, the version control system Git/GitHub will be useful for the coding component and the typesetting platform Overleaf will be useful for the reporting component.

Groups should contain 1, 2, or 3 members. If you would like to be assigned at random to a group, then please contact the instructors. If you would like to determine your group, then please post to Forums under the *Project* thread to contact classmates.

Timelines

October 31st: Project Proposal
November 28th: Project Milestone
December 15th: Project Report

October 31st: Project Proposal

By October 31th groups must upload a one page pdf file on Gradescope containing:

- Title
- Summary of Plans
  - Description of Problem / Application
  - General Approach to Solving Problem / Implementing Application
  - Suggested Experiments / Tests
- Group
  - Name and NetID of each member.
  - Member responsible for uploading submissions.

Groups could check the resources below for possible datasets.

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

November 28th: Project Milestone

By November 28th groups must upload on Gradescope a two page pdf file containing:

- Title
- Group Members
  - Name and NetID of each member.
  - Member responsible for uploading submissions.
- Background
  - Description of Problem / Application
  - Motivation for Problem / Application
  - References / Existing Software
- Plans
  - Description of Methodology / UML diagram
  - Proposed Experiments / Tests
  - Some Relevant Datasets / Packages

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

December 15th: Project Report

Groups will not be responsible for a presentation. By December 14th groups must upload a notebook (.ipynb file format) and pdf on Gradescope.

The notebook should describe the problem, the methodology and experiments used to understand the problem, evaluation of results, and possible next steps. More specifically, the notebook should be structured as follows:

1. Title
2. Group Members
   a. Name and NetID of each member.
   b. Member responsible for uploading submissions
3. Abstract
4. Background
   a. Description of Problem / Application
   b. Motivation for Problem / Application
   c. References / Existing Software
5. Results

    a. Description of Methodology / UML Diagram
    b. Experiments Conducted / Tests Performed
    c. Description of Datasets / Packages
6. Discussion
    a. Evaluation of Findings
    b. Possible Next Steps

See the template below for more information.

The pdf should summarize the notebook. The summary should motivate the problem, explain some aspects of the approach and implementation, and describe the outcomes of the experiments. The pdf should be limited to 3 pages in bulletin format (https://www.overleaf.com/latex/templates/tagged/newsletter). Groups can share their summaries with the class by electing to upload pdf's to https://wp.nyu.edu/pdsf19/

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

Final Project Evaluation:

The final project will be graded based on three main aspects:

1. adherence to guidelines
2. quality of the report
3. implementation of the code

While projects will not be assessed on the technicality of the problem, we will recognize efforts regarding

1. size and "cleanliness" of the datasets
2. sophistication of the code
3. relevance of the problem / application

A final report for a ***data science project*** should:

1. clearly state the problem, pointing which are the hurdles and issues to solve it;
2. clearly present the methodology employed to solve the problem, pointing out:
    i. the data sets used
    ii. the methods employed to (if necessary) handle missing data, transform data, combine data, etc.
    iii. the algorithms involved in the solution, as for example, SVM for classification, DBScan for clustering, etc.
3. present and discuss the results, highlighting the strengths and weaknesses of the proposed methodology

4. make some conclusion, emphasizing whether the chosen approach was success and, if not, why.

A final report for a ***software engineering project*** should:

1. clearly state the functionality of the application, pointing which are the hurdles and issues to solve it;
2. clearly present the methodology employed to implement the application, pointing out:
    a. the packages used
    b. the architecture of the application
    c. the tests made on the application
3. highlight the key parts of the code along with the use cases that informed the tests
4. make some conclusion, emphasizing what changes could make the application more usable or reliable

Template Data Science Project

Title: Project Title Here

Authors:

Name1, NetID1
Name2, NetID2
Name3, NetID3
Name4, NetID4

Abstract: This project focus on … we approached it using … and the results shows that our approach is a good alternative.

1. Introduction and Motivation

Presentation of the problem, its importance, and which are the difficulties involved on it.

2. Methodology

- The data sets involved and how they were “cleaned”.
- Mathematical and computational methods employed to solve the problem.
- Particular design decisions that you deem important when handling the problem.

3. Results

Description of the results, pointing how well the problem has been solved. Figures showing the results are, in general, better than tables.

4. Discussion

Which are the strengths and weaknesses of the proposed methodology. Are the results good enough? Can they be improved? Which are the limitations?

5. Conclusion

Summary of the problem, findings and limitations. Future work directions.

References:

[Joia et al. 2011] Joia, P., Coimbra, D., Cuminato, J. A., Paulovich, F. V., and Nonato, L. G Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12):2563-2571, 2011.

Resources

- Datasets on Amazon's AWS cloud
- Yelp Dataset Challenge
- NYC Open Data
- Data.gov
- UN Data
- Kaggle
- Quandl financial, economic, social datasets
- Face recognition, collaborative filtering, web ranking (see bottom, under "Projects")
    - See here for more collaborative filtering data
- 20 Newsgroups
- Blogs (with spam labels)
- Enron e-mail data set (see also here)
- Congress voting records
- Quota's meta list of datasets
- NYTimes news articles