Milestone I (SIADS 591 & 592) Project Guidelines

Version 2020.11.20.1.CT

Introduction

The project component of Milestone I (SIADS 591 & 592) is designed to give you the opportunity to apply the skills, tools, techniques, and knowledge that you've learned in the courses that you've completed to date in the MADS program to a new problem. The project comprises five components:

- team formation and pre-proposal (via Google Sheets)
- draft proposal (via Google Docs)
- peer review of proposals (commenting on Google Docs)
- instructor-reviewed proposals (via Coursera)
- final report (via Coursera)

We recommend strongly that you work in pairs. Although this recommendation can be relaxed in exceptional cases, it has been our experience that working in pairs balances high-quality work with the amount of work each person needs to do to complete the project.

The formation of teams and selection of topics and datasets will be done via Slack and coordinated via this Google Sheet.

Overall, the goal of this project is to give you the opportunity to bring your unique interests, creativity and ingenuity to our data world, demonstrating your ability to apply the skills and concepts that you learn in the course on real datasets of your choice. In particular, your project should show off your ability to take a minimum two datasets that have different formats and/or access methods and manipulate them in order to combine them and extract a useful byproduct: something more than you could have gotten from either dataset by itself. The manipulation could involve filtering, format conversion, handling missing or noisy data, matching records from one data source with corresponding records in the other, and so on, and must make significant use of programming tools we've covered in prior MADS courses.

Before starting work on the actual project, you'll write a short project proposal that is meant to be a high-level summary and does not need to contain technical details or code. You will then review your peers' proposals, and they will review yours. You can adjust your proposal based on input you receive and you'll then submit it to the course instructors for grading and approval. Requiring the proposal to be submitted early is intended to get you thinking about questions and datasets you're interested in, and how those might be answered with the tools you've been exposed to in previous classes.

Technology Choices

This course differs from your other MADS courses in many ways including technology. We have created a Jupyter environment for you that is functionally equivalent to SIADS 516, which is a superset of the base MADS environment, and you can access that environment via the "ungraded lab assignment" in Coursera. You can use that environment or choose to use any of the environments from courses you have already completed to build and test data manipulations and visualizations for your project. Alternatively, you can use your own locally installed environment. Another possibility is to use Google Colaboratory, which may facilitate collaboration.

Draft Proposal Guidelines

You should use the following outline for your proposal, which shouldn't need more than about one page. The proposal draft (which will be reviewed by your peers; see below) is worth up to 10 points; the revised proposal (which will be reviewed by the course instructors; also see below) is also worth up to 10 points.

1. Summarize your proposed project in a few sentences.

- What is your proposed project and why are you proposing it?
- What are the question(s) you want to answer, or goal to achieve?

2. Describe two different data sources you plan to access, manipulate and bring together. The data sources must require different access mechanisms and/or use different data formats. (For example, you might pick one data source that uses a Web API that returns JSON, and the second might use SQL to query a database, or fetch and parse an HTML page.)

For each data set, you should summarize these properties:

- Name
- Short description (i.e., 1-3 sentences)
- Size (in records and/or bytes)
- Location (give the URL or other access method)
- Format
- Access method

3. Describe with 1-3 sentences for each point below what data manipulation is likely to be needed:

- What initial processing will have to be done on each?
- How will you combine the datasets, and what will be produced as output?
- What new information will result from combining them?

4. Describe in 1-3 sentences one interesting visualization you could include in a final presentation and report that would show the value/answer a particular question in your final output dataset (that would not be possible with either of the original datasets alone).

5. Indicate the contribution that each team member will make to the project.

Your proposals will be reviewed by two peers from the class and you will take those into consideration when you revise your proposal for review by the instructional team.

Your draft proposal should be a Google Doc that you share via the coordination (see link above). Please note that you must enable commenting on your Google Doc to receive peer reviews.

You can propose an alternative project structure with prior approval from the teaching team.

Peer Review Guidelines

You will receive credit for completing reviews of your peers' proposals. Each review is worth up to 5 points toward your final grade. Note that your proposal grade will not be based on the content of your peers' reviews. The purpose of the peer reviews is threefold: (1) to gain experience reviewing proposals, (2) to learn about other work going on in class, and (3) to get feedback on how to make your project better. Your reviews should be several sentences long and should take into consideration the following points:

- **professional:** what would a co-worker think about your review?
- **pleasant:** courtesy goes a long way
- **helpful:** what sort of advice would you want?
- **scientific:** focus on facts, not opinions
- **realistic**: keep scope in mind
- **empathetic**: how would you feel if you received the review you wrote?
- **organized:** make it easy for the recipient to follow your train of thought

A useful approach when writing peer reviews is "two stars and a dog" approach. In other words, highlight two things that the authors did well and identify one area where they might spend some time improving their work (and make constructive suggestions about how to do so).

Peer reviews will take the form of comments on the Google Doc containing the proposal and will be coordinated via the Google sheet.

Final Proposal Guidelines

Taking into consideration the feedback from your peers' reviews of your draft proposal, you will revise your proposal and submit it to the course instructors for review. The course instructors may make suggestions to your plans and will give the official "go-ahead" to proceed with your project.

You will submit the proposal as a PDF document via Coursera.

Project Report Guidelines

First and foremost, following the individual original work policy clearly stated at the start of the course, the topic and questions you ask in your project must be of your own invention. **If you used ideas from a particular web site or previous project, or did your project as part of an existing research collaboration,**

you must identify your sources and/or collaborators and provide links and citation(s) where appropriate.

As a guide, the report should be around 8-10 pages depending on space used for any visualizations, tables, etc.

The format of the report is semi-flexible - you can include additional information, but at a minimum it should have the following sections:

1. Motivation (5 points): (a) Briefly state the nature of your project and why you chose it. (b) What specific question or goal did you try to address?

2. Data Sources (10 points): Describe the properties of the two dataset(s) or API services you used. Be specific. Your information at a minimum should include but not be limited to:

- where the datasets or API resources are located,
- what formats they returned/used,
- what were the important variables contained in them,
- how many records you used or retrieved (if using an API), and
- what time periods they covered (if there is a time element)

For example, if you downloaded data or used API services, you should state the specific URLs to those files or resources. It should require zero effort on my part to find and access the exact resources you used if I need to do so.

3. Data Manipulation Methods (30 points): For each of your two sources, describe how you manipulated the data. For example:

- How specifically did you need to manipulate the data?
- How did you handle missing, incomplete, or incorrect data?
- How did you perform conversion or processing steps?
- What variables and steps did you use to join the two data resources to perform your data analysis?
- Briefly describe the workflow of your source code and what the main parts do.
- What challenges did you encounter and how did you solve them?

4. Analysis and Visualization (25 points):

- A key goal of this project was bringing together two different data resources to answer an interesting question or find a new insight that could not have been answered with either data resource alone (which you summarized in part 1). Now describe the analysis steps you performed on your combined dataset to address that goal/question. Be specific, and include references to key functions or parts of your code.
- What interesting relationships or insights did you get from your analysis?

- What didn't work, and why?
- To summarize your findings, include at least one visualization (chart, plot, tag cloud, map or other graphic) that summarizes your analysis.

5. Statement of Work (0 points)

- You must include a statement that describes the contribution that each team member made to the project.

Project Report Bonus Points

At the discretion of the instructor up to 10 bonus points will be awarded for especially high-quality, creative or insightful projects.

Project Report Submission

Please submit a zip file containing following files:

- Your project report, as a single PDF document. Remember, the project report should be no more than 10 pages. Keep this in mind if you are planning to generate the PDF from a Jupyter notebook. We strongly recommend that you use Google Docs, Word, or some other word processing package to generate your final PDF.
- All source code files/scripts (Python, or any other code) used for your project in a `source/` folder in your zip file.
- Working URLs that point to either (a) the actual data/API resources you used or (b) if the datafile is over 10 Mb or not available in file form, create a sample file containing the first 100 records.

As part of the grading the teaching team may attempt to reproduce your results using your code and data, and you are expected to assist with this if we request it.