project-1

September 14, 2024

1 DSCI 511: Data Acquisition and Pre-Processing Term Project Phase 1: Scoping a data set

1.1 The big picture

Welcome to your term project! This is the first portion of a two-part, open-ended team assignment that will culminate in a presentation during the last week of class or the regularly scheduled final exam period. Overall, this term project is intended to provide some open-ended experience with building a complex dataset and making it available. Specifically, all projects for this course will entail the following:

- The construction, acquistion, integration, enrichment, and distribution of a project-motivated and computationally significant dataset.

The first report on your team's project will constitue a discussion of what the dataset is that you want to build/access/create, why you believe it will be possilbe to conduct, and how long you believe it will take to build, in addition to discussion of the sorts of tasks that will be involved. Additionally, this inital project planning report should speculate and provide examples potential dataset uses, whether academic or commercial.

Note: All reports should inclue a high level abstract/discussion in a tone that is set for a completely diverse audience.

Later on, a final report will recap progress at the task you're group has come up with, specifically revisiting what you *though* the dataset development would take, as compared to the actual work involved and obstacles encountered.

Important: because your project reports will have discussion intermingled with data and code as output, I not only request the submission of your work in Jupyter Notebooks format, but additionly recommend conducting your work as a group collaboratively in Jupyter notebooks.

1.2 This is only a guideline

While I will provide some idea of structure and expectation for your project it is important to note that this is an intentionally open-ended project. Hence, no specific rubric is provided. The courses of different projects will require overcoming different obstacles, and success in a data science project is ultimately a (partial) function of a team's abaility to adapt to project needs. However, all work should be well documented, articulately presented, and justified. If at any point it is unclear what to do or how to represent your project's work, please do not hesitate to ask your instructor for direction.

1.3 Your team

The first thing you'll have to do in this phase is organize into a project team. Data science is often conducted in teams, with different team members covering the diversity of knowledge and skills relevant to the different areas that a project must support to succeed. Even though our course is only focused on early-phase data science tasks (data set development), be sure to consider the strengths of your teamates and interests for gaining experience in dataset construction—if you want extensive experience with web scraping, pitch a project about this with a few other interested points. It will help to discuss interests. Be sure to write out the names of the project team's members in your first report and answer the two questions:

1. What areas/skills/domains does the team member presently identify with?
2. Into which areas/skills/domains would the team member like to grow?

1.4 Your topic

The course of your project will be determined by two things:

1. the motivations present in your project's team and
2. the data your project is able to pull together.

Thus, choosing your topic is closely tied to both your team and the data you are able to identify. To start, discuss the domain interests present on your project team. Te get you on your way, let's start with two questions:

1. Is there an aspect of the IoT, natural world, society, literature, or art, etc. that you would like to investigate computationally through what might be considered 'data'?
2. What sort of data-medium are you interested to work with?—For example: transaction records, stock prices, memes and online conversations, open-domain poems, congressional records, News Articles, songs and popularity, Associated Press Images, transit records, call logs, CCTV footage, etcetera.

Whatever the direction you set for your project please make sure you document it well, keeping track of how its objectives and strategies change as you encounter available materials and other existing work.

1.5 What you're responsible for in this phase

Ok, so here's the goal again for phase 1. You must:

- scope a computationally tangible artifact—heretofore known as the data set—whose study is expected to satisfy goals pertaining to the project's topic of interest.

This phase of the project will set expectations and a work plan for your project's open-ended work. Not only should you scope the collection of your dataset, but determine what mode's of distribution will be possible once its produces. Will you have to distribute access code, or will you be able to directly provide links to stored data.

Ultimately, the completion of your poject will produce raw materials for other folks (possibly you) interested in trying out analysis applications in future coursework (DSCI 521). So, as you identify a potential data set be sure to be realistic about what is possible to collect and how you can preprocess it for use! Ultimately, please make sure that some portion of your target data are guarenteed to be collectable. However, it's okay to try for some data that are a reach, just document any un- or partially successful efforts in your report and discuss what obstacles prevented those data from being collected.

1.5.1 Things I'll be looking for in a Phase 1 report

- a background report on the team's members, their self-identified skills, and individual contributions
- a discussion of what you would like to your data to do/hope it is good for
- an exhibition of a sample of your data—show me it exists and what it looks like, even if very raw
- a discussion of who might be interested in your data set
- a discussion of how your data is limited and could be improved
- a discussion of how your data were created, e.g., people texting, The Earth's molten core spinning, etc.
- a discussion of what sort of access rights presently exist on your data and how/if you will make them available

As a heads up, by the end of the term and in your final report I'll be looking for things like - a data dictionary or README.md that describes what is present in the data set and where or how to access - code that documents the construction of your data—I should be able to re-construct/re-access it! - code that allows me or someone else to interact with your data set - tables and figures indicating the size and variety present in your data

Note: These are not exhaustive lists of topics or tasks worth covering in your project. In general, if there's something interesting about your dataset, whether relating to its construction, existence, representative population or *anything else*, then be sure to document it!

[ ]: