CS 415/515 Social Media Data Science Pipelines Project 1: Data Collection System

Last updated: September 10, 2025

1 Introduction

Carefully read the *entirety* of this document. Deviations from the instructions in this document *will* be penalized.

If you have questions or need clarification on something in this project description, you *must* include the relevant text from this project description you are asking about in your email, or we will assume that you haven't fully read the project description and *will not* respond. Lack of response to emails that do not follow this requirement will not be considered an excuse for deviating from the directions outlined in this document.

The first step to building a data science pipeline is collecting data. In this project, you will be building a continuously operating data collection system. The data that this system collects will be used in the remaining projects, thus effort on this project is crucial.

2 Project Description

Data collection is arguably the most important part of data science. Simply put, without data, there is no data science to perform. Data collection systems often seem simple on the surface, but the reality of different data sources on the Web means that there are often numerous road blocks that must be dealt with along the way. As mentioned previously, this project will serve as the foundation for the remaining projects in the class: if this project is not completed, it is extremely unlikely you will be able to complete the remaining projects.

Before you set about building your data collection system, you need to decide *what data you are going to collect*. Depending on your status (undergraduate or 4+1), you have different requirements for the data you collect:

- Undergraduate: You must collect data from at least **two** different data sources.

- 4+1: You must collect data from at least **three** different data sources.

Since the project is a group project, the requirements are the same for all members of the group. A group is considered 4+1 if at least one member of the group is a 4+1 student.

You are encouraged to focus on the three platforms we have covered in class: BlueSky, 4chan, and YouTube. You also have the freedom to choose other platforms, but you have to justify your choice in your proposal.

The data collection system must be able to collect data continuously for the duration of the class and support projects 2 and 3. While for research purposes, "novel" data sets are often preferred, for our purposes, we want to make sure that:

- It is possible to collect the data you propose collecting
- There is enough data available to perform meaningful measurements and analysis
- Data is continuously being generated (i.e., not a single snapshot in time)

Once your data source is decided on, you can build the collection system. Unfortunately, there is not any single right way to build a data collection system. However, there are many potentially bad decisions that experience can help you avoid. Since the majority of you do not have that experience, we will be providing face-to-face feedback to each group throughout the class.

3 Project Deliverables

There are three deliverables for this project.

1. Project proposal
   - GitHub Classroom Section: `https://classroom.github.com/a/NzHZBi4t`
   - Due 11:59PM, Sep. 18th, 2025. **NB: This is a hard cut off and you will not be able to commit to this repository when the due date passes without a late penalty.**
2. Project implementation.
   - GitHub Classroom Section: `https://classroom.github.com/a/wzZcDhBB`
   - Due 11:59PM, Oct. 14th, 2025
3. Project report.
   - GitHub Classroom: `https://classroom.github.com/a/MeNol0uK`
   - Due 11:59PM, Oct. 14st, 2025. **NB: This is a hard cut off and you will not be able to commit to this repository when the due date passes without a late penalty.**

3.1 Project Proposal

The purpose of your proposal is to ensure that:

- You are not attempting to do something impossible
- You are not attempting to do something illegal
- You are not attempting to do something too easy

Your proposal should provide enough information that we can read it and have a rough idea of what it is you plan to do, and with enough detail that we can help you avoid pitfalls we have experienced in the past.

To this end, we your proposal must include at least the following:

- A description of your data sources you wish to collect from, including the specific API calls you will use, etc.
- A rough sketch of how you intend to collect the data. You must also include a diagram that shows your system architecture. Make sure you include any libraries you intend to use.
- Some ideas with respect to measurements and analysis you have in mind.
- Some "napkin math" estimates of how much data you think will be collected each week.

Your proposal should be one to two pages. **Your report must conform to the two column ACM 'sigconf' format** available here: `https://www.acm.org/publications/proceedings-template` and *must be submitted as PDF*. If your proposal does not conform to this format, or you submit something besides a PDF then you will receive a zero.

You are encouraged to use the LaTeX template. For easy compilation and collaboration with your teammates, you can use Overleaf (`https://www.overleaf.com`). You can also use the Word template, but make sure the submission is a PDF, otherwise you will receive a zero.

3.2 Project Implementation

Your data collection system *must* run continuously for the duration of the class.

You will be required to submit the code that you wrote to collect data. Please note that while you are generally free to use whatever language and libraries you want, there is a restriction with respect to crawling frameworks. I.e., you may not use any crawling frameworks. You are allowed to use an HTTP client, but nothing like Scrapy (`https://scrapy.org`).

You will receive a zero on your implementation if you use any unauthorized library.

If there is any confusion as to whether or not you are allowed to use a particular library, the onus is on you to ask about it.

3.3 Project Report

Once your data collection system has been implemented, you will submit a two to four page report describing your implementation in full, *as well as a preliminary exploration of the data.* Your report should indicate any changes that happened since your proposal, any challenges you faced, etc. Your report *must* also contain at least one plot that indicates how much data is being collected over time. You should also include updated projections on how much data you are likely to collect. This is *very* important because some of you might choose a data source that requires getting the storage limits of your VM raised.

4 Grading

- Proposal is worth 25 points with the following rubric
  - **-1,000 points:** Not a PDF in ACM format.
  - **-1,000 points:** Missing `HONESTY.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing `CREDITS.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing AI usage statement if any generative AI tools were used (please refer to the course syllabus for instructions).
  - **10 points:** Data source description (*specific* APIs used, etc.)
  - **10 points:** Data collection description (including a diagram).
  - **5 points:** Napkin math of how much data is expected.
- Implementation is worth 50 points.
  - **-1,000 points:** Unauthorized libraries used.
  - **-1,000 points:** Missing `HONESTY.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing `CREDITS.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing AI usage statement if any generative AI tools were used (please refer to the course syllabus for instructions).
  - **30 points:** Continuously collected data
  - **10 points:** Infrastructure for data source 1
  - **10 points:** Infrastructure for data source 2
- Final report is worth 25 points.
  - **-1,000 points:** Not a PDF in ACM format.
  - **-1,000 points:** Missing `HONESTY.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing `CREDITS.md` (please refer to the course syllabus for instructions).
  - **-1,000 points:** Missing AI usage statement if any generative AI tools were used (please refer to the course syllabus for instructions).
  - **10 points:** Final system description.

  - **10 points:** Plot showing data collection over time.
  - **5 points:** Updated projections of data collection volume.

NB: You *must* commit to the correct repository or that component of your project will not count as submitted! E.g., if you put your implementation in the report repository it will be treated as if you did not submit your implementation at all.