CSCD25: Course Project Information

Schedule 1
Datasets 1
Milestone 2 details (due at 12:00pm (noon) on Monday, November 20) 4
Final project details (due at 11:59pm on Monday, December 4) 6

The course project gives you an opportunity to put your new skills to work! You will practice applying the entire data science process you have been learning about, from data preparation to exploration to modeling to interpretation and communication. You may work in groups of 1 or 2, with 2-person groups expected to be more ambitious in scope than solo projects. The remaining deliverables for the project are a short Milestone 2 report and a final report presented as a "data story".

Schedule

Milestone 2 report: due at 12:00pm (noon) on Monday, November 20
Final data story report: due at 11:59pm (midnight) on Monday, December 4

Datasets

The datasets available for this project are derived from Reddit posts and comments. Reddit is a large online social media platform oriented around discussion, which takes place in communities called *subreddits*. There are subreddits for almost anything you can imagine. Social media platforms like Reddit have played an increasingly central role in society, which has inspired a rich body of research that uses social media to both answer existing social questions and to answer new questions this new ecosystem poses. As such, there is a plethora of interesting and important questions waiting to be answered in these Reddit datasets.

Note that because this is real data of people interacting online, there are instances of negative, toxic, or alarming content—the kind of thing you may run into in the course of

browsing Reddit. We have taken steps to minimize this as much as possible. First, the main dataset we are providing includes the metadata of who comments where and when, but *not* what they say. Second, we have restricted the set of subreddits represented in the data (see below for details). Third, in the main dataset the usernames have been hashed. Examining the raw data, you may still come across the odd unsavoury content in the subreddit names.

There are three datasets available for your projects: a main dataset, a full dataset, and a text dataset.

- The **main** dataset is intended to satisfy most metadata-based research questions and be manageable in size,
- The **full** dataset contains much more metadata but is larger and thus more difficult to process,
- The **text** dataset contains the actual text of what people say and is manageable in size.

Main dataset (**Metadata only, curated set of subreddits, downsampled to ~1GB**)

The main dataset includes metadata for both comments and submissions on Reddit from January 2019–June 2021. It only includes data from a group of about 100 selected subreddits that contain minimally offensive content. Additionally, no text is included in this dataset, and usernames are hashed. It is a random 2% down-sample, so that it is only about 1GB in size.
main_comments (1.1 GB)
main_submissions (56 MB)

Note: The username hashing for the `main` project dataset was done incorrectly and the usernames "[deleted]" and "AutoModerator" were hashed when they should have been left in plain-text. If you wish to filter those usernames out (or otherwise analyze them), you can use the hashed values instead:
>>> hash_username("[deleted]")

'Io99IHkg-4QzX6xbKwbte0cuzp4='

>>> hash_username("AutoModerator")

'EA1r-K5p_lVBLesLhCFRrKOPN-I='

Full dataset (**Metadata only, "all" subreddits, very large and split by day**)

The full dataset contains metadata for both comments and submissions on Reddit from January 2019–June 2021. This dataset contains data for the top 5,000 SFW subreddits by activity, which have not been vetted (although 18+ communities have been removed), and full usernames are included. It is split up into daily CSV files, each of which is about 75MB.

Data available at: http://csslab.cs.toronto.edu/cscd25/

Text dataset (**Metadata + text, "all" subreddits, downsampled to ~1GB**)

The text dataset includes both the metadata and the text for both comments and submissions on Reddit from January 2019–June 2021. This dataset contains data for the top 5,000 subreddits by activity, which have not been vetted (although 18+ communities have been removed.) It also contains the textual content of comments and the titles/self-text of submissions. It is a random 1% down-sample, so that it is only about 1GB in size.

text_comments (3.5 GB)

text_submissions (364 MB)

Schema:

For both comments and submissions:

- *id:* a unique id for the item
- *score:* score of the item (upvotes minus downvotes, with some algorithmic 'fuzzing' applied)
- *author*: username of the user who posted the item, can be '[deleted]' if an item has been deleted from its authors' profile, or 'AutoModerator' if posted by the AutoModerator bot

- *subreddit*: name of the subreddit the item was posted in
- *created_utc*: time the item was posted, in Unix time

For comments only:

- *link_id*: id of the link to which this comment belongs
- *body*: textual content of the comment, in the 'text' dataset only

For submissions only:

- *is_self*: True if a submission is a text-only 'self-post', False if the submission is a link
- *domain*: domain of the link
- *title*: title of the submission, in the 'text' dataset only
- *selftext*: content of the self-post, in the 'text' dataset only

Milestone 2 details (due at 12:00pm (noon) on Monday, November 20)

For the Milestone, you will identify the project you will pursue, verify that it is feasible given the dataset you have chosen, preprocess the data, and conduct initial descriptive analyses. The goal is for you to decide on a clear research question, start working with the data, and set up the pipeline you will need to execute your final analyses.

We will support you in working with the Reddit datasets we have provided. You can incorporate other datasets to augment your project with additional data, if you want, but it is not necessary and we may not be able to support issues that arise with extraneous data sources.

As you begin your data science process, you should especially think about, and demonstrate your progress on, the following considerations:

- First and foremost, what is the question you intend to answer?

- What is the analysis strategy you will use in your pursuit of this question? What analyses, modeling steps, and visualization plans do you have? What alternatives did you consider but eventually drop in favour of your proposed plan?
- What dataset(s) will you use?
- How will you manage the size of the dataset(s) you have chosen?
- How will you filter, transform, and/or enrich the raw data to get it in a form appropriate for your project?

Deliverables:

1. A README.md file containing the detailed project proposal (up to 1000 words). Your README.md should contain:
    - Title: A memorable, accurate headline for your project.
    - Abstract: A 150-word description of your project idea and goals. What's the motivation behind your project? What story would you like to tell, and why?
    - Research Questions: A short list of research questions you would like to address during the project (with one being the primary research question).
    - Proposed additional datasets (if any): List the additional dataset(s) you want to use (if any), and some ideas on how you expect to get, manage, process, and enrich them. Show us that you've read the docs and some examples, and that you have a clear idea on what to expect. Discuss data size and format if relevant. It is your responsibility to check that what you propose is feasible.
    - Methods: What exploratory analyses, statistical summaries, modeling approaches, and visualization techniques do you plan to use to answer your research question(s)? You should address the considerations listed above here.
    - Organization within the team: If you're working in a team of 2, who will do what?

2. Jupyter notebook containing initial analyses and data handling pipelines. We will grade the relevance and quality of code, and quality of textual descriptions.

Grading rubric:
README.md (60%)
  Title/abstract (15%)
  RQs (15%)
  Methods (30%)

Jupyter notebook (40%)
  Relevance and quality of code (30%)
  Quality of textual descriptions (10%)

Final project details (**due at 11:59pm on Monday, December 4**)

Once you have completed the Milestone, you will execute the project you proposed and describe your project in a "data story". Data stories are a blog post or short article, with an important visual component, using data to tell a story and illustrate it effectively. You can be less formal here (although methods and math should then appear in the notebook), but more visual. You can pick your preferred platform option, but we encourage you to use Jekyll. You submit the story by providing a URL to it in your README file.

A single supporting notebook extending the one delivered for the Milestone is also expected and will be graded. The README in the Milestone should be updated and finalized. It should also detail the contributions of group members for 2-person groups.

Deliverables:

1. The final project repository containing your final notebook and README.md file. We will grade the correctness, quality of code, and quality of textual descriptions.
2. Data story. We will grade the quality of motivation and framing of the research questions, the quality of the analytical choices, and the quality of your presentation and communication.

Grading rubric:

Repository (50%)

Quality of data science process (35%)

Textual descriptions (15%)

Data story (50%)

Motivation/research question/setup: (7.5%)

Visualizations (12.5%)

Description of methods and analyses (20%)

Interpretations and discussion (10%)