Social Media Data Science Pipelines Project 1: Data Collection System

September 14, 2022

1 Introduction

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*. For this class, there will be (at least) three data sources you will collect. The first is Twitter, which you are required to collect via the "Sample stream" (`https://developer.twitter.com/en/docs/twitter-api/v1/tweets/sample-realtime/overview/get_statuses_sample` which provides an approximately 1% sample of Tweets in real time.

The second data source you will collect will be Reddit. The volume of Reddit comments is to high for you to realistically collect all of it, therefore, your crawler should be oriented around collecting data from a set of specific subreddits. The datasets your crawler collects from should not be hard coded, and new subreddits should be able to be added in a somewhat dynamic fashion.

You will decide what the third data source is. While for research purposes, "novel" data sets are often preferred, for our purposes, we just want to make sure that: 1) it is possible to collect the data you propose collecting, 2) there is enough data available to perform meaningful measurements and analysis, 3) that 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, Jeremy 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 Clasroom: `https://classroom.github.com/a/1xr7NKZ6`
   - Due 11:59PM Monday, Sept. 26th, 2022.
2. Project implementation.
   - GitHub Classroom: `https://classroom.github.com/a/OZfXH9g5`
   - Due 11:59PM Sunday, Oct. 23rd, 2022
3. Project report.
   - GitHub Classroom: `https://classroom.github.com/a/KJat2mPG`
   - Due 11:59PM Sunday, Oct. 23rd, 2022

3.1 Project Proposal

The purpose of your proposal is to ensure that: 1) you are not attempting to do something impossible, 2) you are not attempting to do something illegal, 3) you are not attempting to do something too easy. Your proposal should provide enough information that Jeremy can read it and have a rough idea of what it is you plan to do, and with enough detail that Jeremy can help you avoid pitfalls that he has experienced in the past.

To this end, I suggest your proposal have several sections:

- A description of the third data source you wish to collect from.
- A rough sketch of how you intend to collect the data. E.g., is there a Web API? If so, what methods does the API have that will help you collect data.
- 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.

3.2 Project Implementation

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`). **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 Jeremy 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, *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.

3.4 On Twitter Data

The 1%, even though it is a small fraction of overall Twitter activity, can still be very large. Although you do not need to store all of it (e.g., you can filter on certain keywords you are actually interested in using), you *should* keep track of how many tweets per hour you receive from the stream. Project two will ask you to plot the number of tweets received per hour from the 1% stream for a given date range, so make sure you are able to answer that question!

4 Grading

- Proposal is worth 25 points.
- Implementation is worth 50 points.
- Final report is worth 25 points.