Collection Design

Curating a Collection of Twitter JSON Data

Before you start collecting Twitter data, it is important to clearly define your research question(s) to ensure that the tweets returned from Twitter's Application Programming Interface (API) meet your needs as a researcher. This tutorial will walk you through the collection design process and provides a checklist to help organize your collection. For the purposes of this tutorial, the process will assume that the collection will be created using Twarc, an open-source tool from Documenting the Now which allows users with Twitter developer access to collect Twitter data.

Difficulty level: Beginner

Prerequisite(s)

- None

Material(s)

- Open the Collection Design Checklist in a new tab

Lesson objectives

- Become familiar with Twitter's Terms of Service and Developer Agreement
- Understand the difference between Twitter's standard search, filter, and sample APIs
- Complete the Collection Design Checklist to design collections that meet the needs of the researcher

Key Terms

- Application Programming Interface (API)
    - The interface that allows software applications to communicate with one another
- JSON - JavaScript Object Notation
    - A minimal, human-readable format for structuring data. Twitter data is in JSON format.

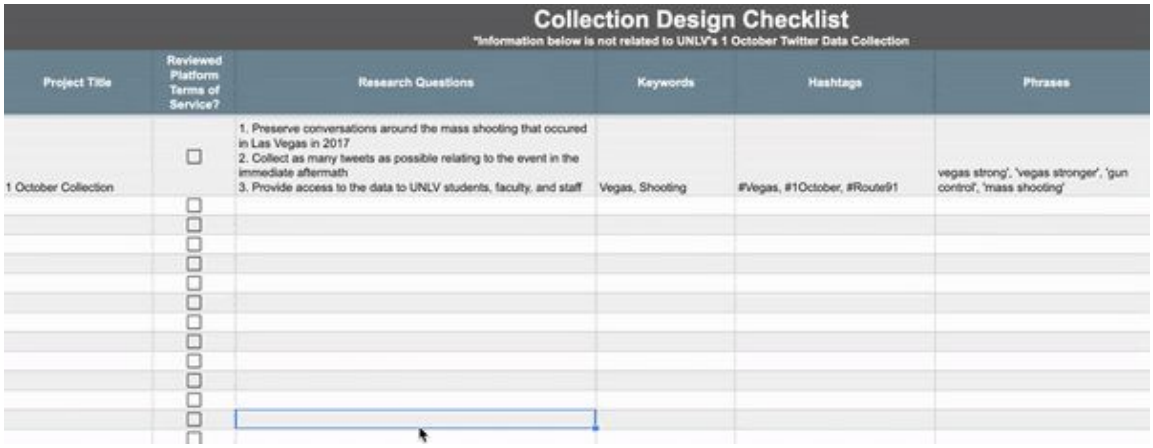
- Twarc
    - A command line tool and python library for collecting Twitter JSON data

Table of Contents

Introduction

Well-curated collections can help academic researchers more effectively answer their research questions. While there exist multiple approaches to curating and analyzing data, there are several items concerning collection design that anyone working with Twitter data should consider. We have combined these items into a checklist that will help you as you prepare to create your first collection. For the purposes of this tutorial, the collection design process of UNLV's 1 October Twitter Collection will be referenced to illustrate each checklist item. Follow the steps below to prepare yourself to create a Twitter data collection.

Important: If you haven't already, open the Collection Design Checklist in a new tab. Once you have finished the tutorial, make a copy of the spreadsheet and use it to prepare for future collections.



Collection Design Checklist

"Information below is not related to UNLV's 1 October Twitter Data Collection

Project Title	Reviewed Platform Terms of Service?	Research Questions	Keywords	Hashtags	Phrases
1 October Collection	☐	1. Preserve conversations around the mass shooting that occured in Las Vegas in 2017 2. Collect as many tweets as possible relating to the event in the immediate aftermath 3. Provide access to the data to UNLV students, faculty, and staff	Vegas, Shooting	#Vegas, #1October, #Route91	vegas strong', 'vegas stronger', 'gun control', 'mass shooting'
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				
	☐				

Collection Design Checklist

☑Step 1: Review Platform Terms of Service

Before starting the collection process, it is important to review Twitter's Terms of Service. Twitter data is, by nature, incredibly sensitive. Users may not realize that you are collecting and analyzing their content, which presents a number of ethical challenges. Take a moment to review the pages below to familiarize yourself with Twitter's expectations of both users and developers:

- Link to Terms of Service
- Link to Developer Agreement and Policy

As a researcher, you may want to share your data collection in the future. The pages above outline the legal guidelines for sharing Twitter data. Understanding the legal boundaries before collecting Twitter data will help to clarify the ways you can publish your findings while keeping with Twitter's Terms of Service.

1 October Collection: Adhering to Twitter's Terms of Service.
It was important to the creators of the collection that anyone accessing the data did so legally. The 1 October collection's 'readme' text file contains a use disclaimer that references Twitter's Terms of Service. This is to ensure that anyone accessing the dataset is aware of the legal implications of downloading and inappropriately sharing data contained in the collection.

☑Step 2: Define Research Question(s)

Clearly defining your research questions before collecting Twitter data will help you to curate a collection that meets your unique needs as a researcher. The purpose of your collection can be anything from a desire to archive and preserve conversations around a specific event, to finding additional data to support your work. Consider the questions below when defining your research question(s):

Why are you collecting?
Thinking about the purpose of your collection will help you plan ahead. If your goal is, for example, to understand the shared followers between two political accounts, you may not need a collection that contains tweets from each account, but rather a list of follower ids. Writing out your purpose may help you discover which tools you will need to build your collection, and which ones will provide you with data that is irrelevant to your work.

What are you collecting?
After defining the purpose of your collection, consider which of the following data points may be relevant to your research needs:

Keywords
Collect tweets based on a keyword or group of keywords

Hashtags
While keywords will pull in every mention of a specific word on Twitter, hashtags will only pull intentionally-tagged tweets

Phrases
Collect tweets based on a specific phrase

Location
Collect tweets sent from a specific location

1 October Collection: Purpose of Collection
The purpose of the 1 October Collection was to collect as many tweets as possible in the immediate aftermath of the mass shooting that occurred in Las Vegas in 2017. A large-scale collection would provide researchers on campus with a macro-level look at the event as seen from local, national, and international perspectives.

☑Step 3: Determine Collection Period

Now that you have defined your research questions, you can begin to think about the details that will provide you with a collection curated to your project goals. Setting a desired collection period before collecting will help prevent gaps in the dataset. It will also allow you to plan for storage and security-related issues discussed in the next step. Consider the following when determining your collection period:

Collection dates
Having an idea of which dates you would like to begin and end the collection can be useful. If you are collecting around political debates, for example, knowing that you want to cover the period leading up to the event as well as the discussions that follow can help you determine when you will need to start your collection.

Historical data
Purchasing historical Twitter data is expensive. While we will discuss this further in the steps that follow, it is important to know that the standard (free) access to existing tweets limits you to a 7-day search window. Determining the collection period beforehand can help to prevent missing that window and spending large portions of your research budget on historical data.

Real-time data
Collecting Twitter data in real-time, like historical data, comes with its challenges. For researchers that do not wish to pay for access to stream data, a limit is placed on the amount of data they can collect in real-time. This limit presents a challenge for researchers looking to collect data around massive, short-term events.

1 October Twitter Dataset Collection Period
Since the purpose of the 1 October collection was to collect as many tweets as possible in the immediate aftermath of the shooting, a collection including tweets from a 7-day period before and after the event was conducted.

☑Step 4: Determine Collection Storage

After defining the purpose and determining the length of your collection, you can start to plan for the storage of your data. The sensitive nature of tweets calls for the secure storage of your dataset. When planning on where and how to store your data, consider the following:

Collection Size

While determining the exact size of your collection is not feasible, planning ahead for a smaller or larger collection is possible. After reviewing your research aims and desired collection period, you can get a better idea of whether you will be working with a smaller or larger set of data. For example, if you are aiming to collect tweets around a conference with 1,000 attendees, you can assume that your collection will be smaller than a collection built around a crisis event. The goal is not to determine the exact size of the collection, but rather to use your stated research purpose and collection dates to infer whether or not you will be dealing with large or small amounts of Twitter data.

Collection Storage

If you know that you will be collecting a large amount of Twitter data, it is important to make sure you have a sufficient amount of storage space. A failure to do so could result in gaps in the data if a collection is cut short due to insufficient storage space.

Collection Security

Twitter data is sensitive, and should be stored in a secure place. If you are based out of an institution, asking to store it on a secure computer may be your best option. If you are planning to store it on a personal laptop, create a plan to store your computer in a safe space when it is not in use. Remember, depending on the collection you may or may not have data that could put a user's life in danger. Proper and safe storage of your data will help you adhere to both the Twitter developer agreement and to ethical research practices.

1 October Collection: Collection Security and Storage

The 1 October collection is stored in the LIED Library. All private-access downloads are limited to UNLV faculty, students, and staff. All physical copies are stored in secure areas of the building. The public version of the dataset follows Twitter's guidelines for the sharing of data outside of the institution that facilitated the download.

☑Step 5: Plan Ahead

You are almost ready to select your tools and start collecting! Before you move on to step six, take some time to think about the future use of the data you are collecting and plan accordingly. In steps one through four, you have considered the terms of service of the platform, defined your research questions, determined the collection period, and identified storage solutions for your collection. Thinking ahead will prepare you to create a collection that is well-documented and can be shared with other researchers. Consider the following:

Future use

While you may have your research questions ready to go, it is important to consider future research needs. Earlier in this tutorial, it was mentioned that the cost of purchasing historical

Twitter data is high. When in doubt, collect more. Remember: you can always create smaller subsets of your collection.

Sharing your data
Decide whether or not your plan on sharing your data outside of your institution. Preparing beforehand to create and share a collection with other researchers will help determine how you will document your collection.

1 October Collection: Planning Ahead
The creator of the 1 October Collection intended it to serve as a part of the University's archival efforts around the mass shooting that occurred in Las Vegas. They knew that researchers from multiple disciplines could find this data useful, so it was important to them to design a collection that captured as many relevant tweets as possible. Their search returned over 14,000,000 tweets in a period of seven days.

☑Step 6: Select Tools

Depending on the purpose you outlined for your collection, different tools can be used to create it. There are web-based analytics services that will help you collect and analyze the data, as well as tools that will allow you to directly mine the Twitter API. For this series of tutorials we will be using an open-source tool from Documenting the Now: Twarc.

Twarc allows you to, with Twitter developer access, collect JSON data from several of Twitter's APIs. This tutorial will focus on the standard search, filter, and sample APIs. Keeping your collection purpose in mind, consider using one or more of these APIs to build your collection.

Search
Twitter's standard search API has a 7-day search endpoint. This means that, for any searches you conduct, it will only return tweets within 7 days of the query date. For research purposes, it is important to note that the use of a standard search API does not provide full-fidelity data. In addition to a strict window for collecting tweets, the data returned may not include every relevant tweet from the collection period.

Filter
The statuses/filter API returns public tweets as they occur. The standard access to the streaming API will limit the amount of tweets collected. If you are planning on collecting data around, for example, an internationally trending topic, the resulting dataset may not return all tweets.

Sample

The statuses/sample API returns a small 'random' sample of all public statuses. The randomness of the sample is determined by Twitter.

1 October Collection: Tools Used
The 1 October Collection was built using Twarc, a command line tool. The collection was created using Twarc's 'search' command, which uses Twitter's search/tweets API to download pre-existing tweets matching a given query.

☑Step 7: Begin Collecting

Congratulations! You have made it to the seventh and final step. The design process is unique for every collection, and you may find that tweaking the steps above may be necessary for the purpose of your collection. In the next tutorial, you will learn how to install Twarc and use it to create the collection you designed in this tutorial.

Check out the next tutorial in the series: Twitter Data Collection Using Twarc.

1 October Collection
The 1 October Collection contains 14,108,104 tweets sent from September 29th through October 7th, 2017. The collection was created by Thomas Padilla and prepared by Thomas Padilla and Miranda Barrie at the University of Nevada, Las Vegas Libraries.

🎉*Happy collecting!* 🎉