

```
{ "created_at": "Thu Apr 06 15:24:15 +0000 2017", "id_str": "850006245121695744", "text": "1\\ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps://t.co/XweGngmx1P", "user": { "id": 2244994945, "name": "Twitter Dev", "screen_name": "TwitterDev", "location": "Internet", "url": "https://dev.twitter.com/", "description": "Your official source for Twitter Platform news, updates & events. Need help? @TwitterDev", "entities": { "url": { "urls": [ { "url": "https://t.co/XweGngmx1P", "expanded_url": "https://t.co/XweGngmx1P", "unwound": { "url": "https://t.co/XweGngmx1P", "title": "Building the Future of the Twitter API Platform" } } ] }, "user_mentions": [ ] } } { "created_at": "Thu Apr 06 15:24:15 +0000 2017", "id_str": "850006245121695744", "text": "1\\ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps://t.co/XweGngmx1P", "user": { "id": 2244994945, "name": "Twitter Dev", "screen_name": "TwitterDev", "location": "Internet", "url": "https://dev.twitter.com/", "description": "Your official source for Twitter Platform news, updates & events. Need help? @TwitterDev", "entities": { "url": { "urls": [ { "url": "https://t.co/XweGngmx1P", "expanded_url": "https://t.co/XweGngmx1P", "unwound": { "url": "https://t.co/XweGngmx1P", "title": "Building the Future of the Twitter API Platform" } } ] }, "user_mentions": [ ] } } { "created_at": "Thu Apr 06 15:24:15 +0000 2017", "id_str": "850006245121695744", "text": "1\\ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps://t.co/XweGngmx1P", "user": { "id": 2244994945, "name": "Twitter Dev", "screen_name": "TwitterDev", "location": "Internet", "url": "https://dev.twitter.com/",
```

# {Cleaning Your Twitter Data}

## Cleaning Your Data

### *Organizing a collection of Twitter Data*

In the [fourth tutorial](#) in the 1 October Twitter Data tutorial series, you learned how to create a collection of Twitter data. Before working with Twitter JSON data, you may want to organize your collection. Depending on the purpose of your collection and desired research outcomes, sorting your tweets chronologically and removing duplicate tweets might make your dataset more amenable to computational methods. How you organize your collection is completely up to you - this tutorial simply provides you with the tools necessary to get organized. If you haven't already, check out the first six tutorials in this series to get familiar with [Twarc](#), a command-line tool for archiving Twitter data.

**Difficulty level:** Intermediate

### Materials

- [Collection Documentation Checklist+](#)
- [Download Sample Collection Tweet ids](#)

### Prerequisite(s)

- [Tweet JSON](#)
- [Command Line](#)
- [Collection Design](#)
- [Collection with Twarc](#)
- [Collection Documentation](#)
- [Collection Ethics](#)

### Lesson objectives

- Sort a collection of Twitter data chronologically
- Remove retweets from a collection of Twitter data
- Remove tweets before a set date for a collection of Twitter data

## Key Terms

- Terminal - OS X Command Line
  - A text interface for your computer. Terminal receives commands, and then passes those commands on to the computer's operating system to run.
- Twarc
  - A command line tool and python library
- Python
  - The programming language that Twarc is developed in
- JSON - JavaScript Object Notation
  - A minimal, human-readable format for structuring data. Twitter data is in JSON format.
- **Hydrate** - Twarc Command
  - Reads a file of tweet identifiers and write out the tweet JSON for them using Twitter's [status/lookup API](#).

## Table of Contents

---

Lesson objectives	1
Key Terms	2
Table of Contents	2
<b>Introduction</b>	<b>2</b>
<b>Sort Your Collection Chronologically</b>	<b>3</b>
<b>Remove Retweets From Your Collection</b>	<b>5</b>
<b>Remove Tweets Before Certain Date</b>	<b>6</b>
<b>Update Your Collection Documentation</b>	<b>7</b>

# Introduction

---

In order to complete this tutorial, you are going to need a collection of Twitter data. If you completed the [fourth tutorial](#) in this series, you should have your own collection ready to go. If not, you can hydrate a collection of tweets created for the purposes of this tutorial. To learn how to hydrate Twitter data, check out [this tutorial](#).

## [Download Twitter IDs - 'Vegas' 3.20.2019](#)

*Tip: Save your hydrated file of tweets as **vegas\_tweets.json**; the filename 'vegas\_tweets.json' is referenced in the tutorial.*

### Collection Information

The search term 'Vegas' was used to create the sample collection for this tutorial. The original dataset contained 32,636 tweets. Please note that once hydrated, the dataset may contain significantly fewer tweets than the original. The first Tweet in the collection was sent March 20, 2019 at 01:10:06 UTC. The last Tweet in the collection was sent March 21, 2019 at 00:30:28 UTC.

# Sort Your Collection Chronologically

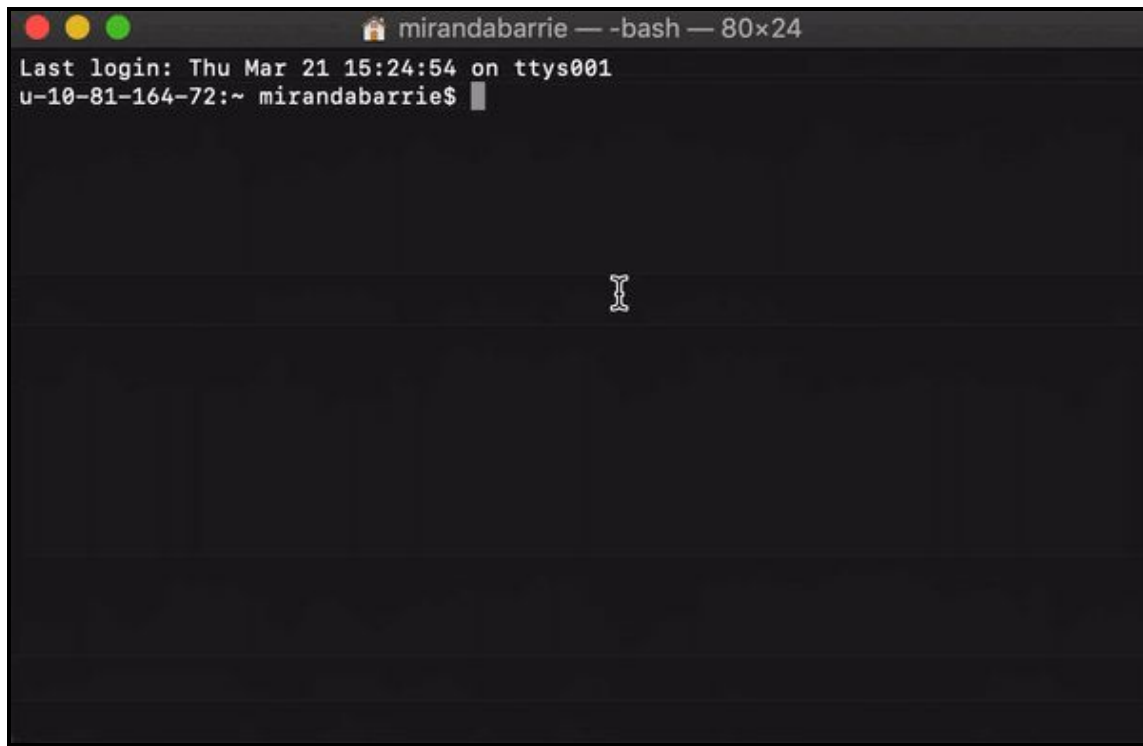
---

You might want to know when the first tweet and last tweets in your collection were sent out. To figure that out, you are going to need to sort your collection by id, which is the same thing as sorting your tweets by their individual timestamps.

Go ahead and navigate to the directory that contains your Twitter data in Terminal. I'm going to use the sample collection in the following examples. Remember, you can drag the folder into Terminal if you don't know the filepath.

**Important: your commands may look different from the commands below if you are using your own data.**

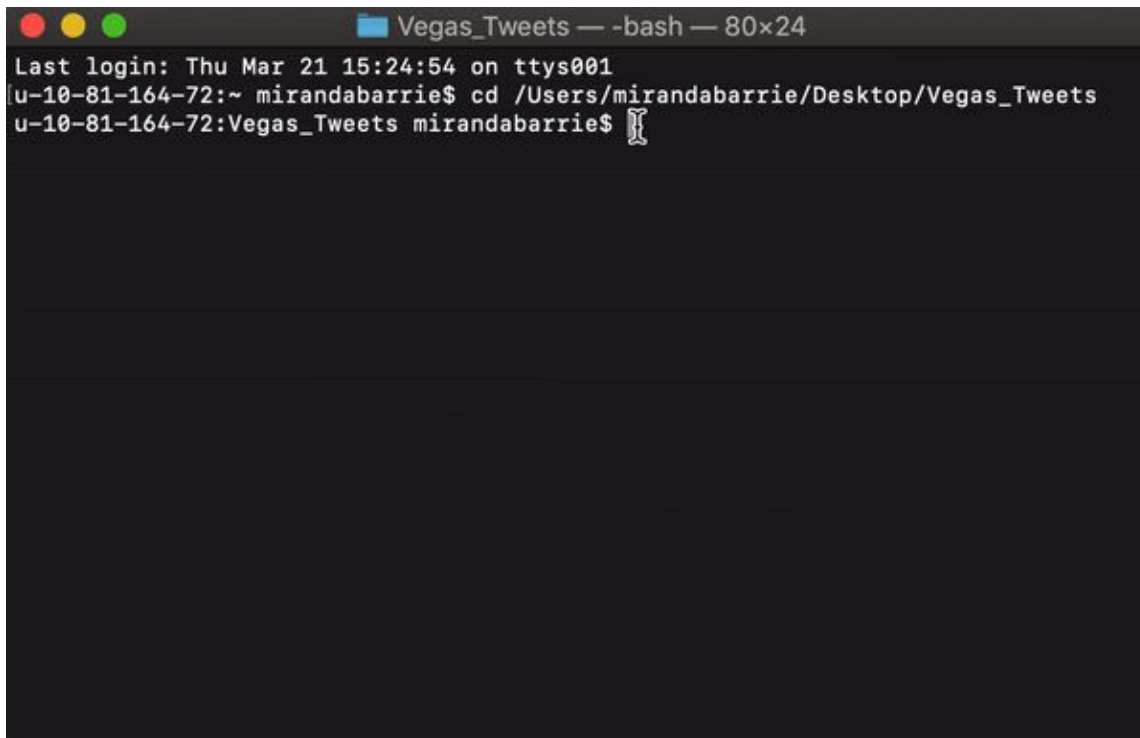
```
cd Vegas_Tweets
```



```
mirandabarrie — -bash — 80x24
Last login: Thu Mar 21 15:24:54 on ttys001
u-10-81-164-72:~ mirandabarrie$
```

Now we're going to sort the collection by id, which is the same thing as sorting the collection by time.

```
python ~/git/twarc/utils/sort_by_id.py vegas_tweets.jsonl > vegas_tweets_chronological.jsonl
```



```
Vegas_Tweets — -bash — 80x24
Last login: Thu Mar 21 15:24:54 on ttys001
u-10-81-164-72:~ mirandabarrie$ cd /Users/mirandabarrie/Desktop/Vegas_Tweets
u-10-81-164-72:Vegas_Tweets mirandabarrie$
```

*Tip: You can view the first and last tweets in the dataset by using the `head` and `tail` commands. To learn more, check out the [second tutorial](#) in this series.*

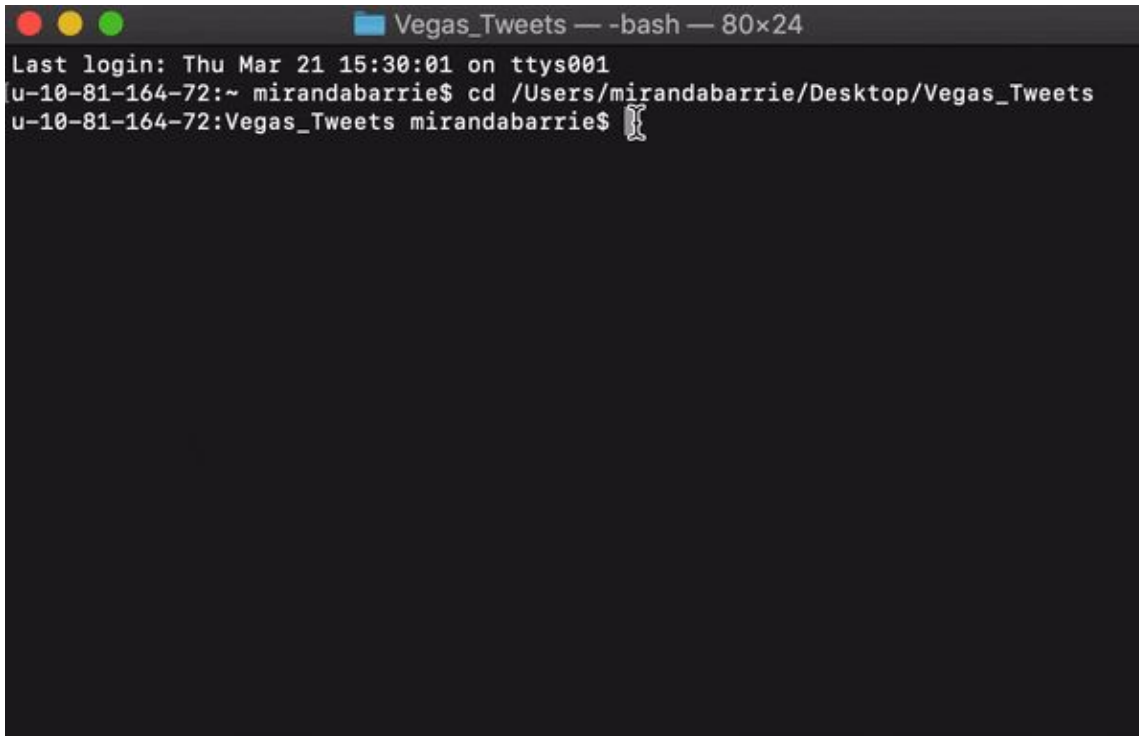
🎉 Nice work! Your tweets are now sorted chronologically. 🎉

## Remove Retweets From Your Collection

Some researchers may want to include the entire conversation in their dataset, which would include retweets. Others may want to focus on original content. If you are in the latter group, you may want to remove retweets from your collection of Twitter data.

Assuming you are still in the directory that contains your collection, enter the following command to remove retweets from your dataset.

```
python ~/git/twarc/utils/noretweets.py vegas_tweets.jsonl > vegas_tweets_norts.jsonl
```

A terminal window titled "Vegas\_Tweets" with a window size of 80x24. The terminal shows the following text: "Last login: Thu Mar 21 15:30:01 on ttys001", "u-10-81-164-72:~ mirandabarrie\$ cd /Users/mirandabarrie/Desktop/Vegas\_Tweets", and "u-10-81-164-72:Vegas\_Tweets mirandabarrie\$". The prompt is followed by a cursor icon.

```
Vegas_Tweets — -bash — 80x24
Last login: Thu Mar 21 15:30:01 on ttys001
u-10-81-164-72:~ mirandabarrie$ cd /Users/mirandabarrie/Desktop/Vegas_Tweets
u-10-81-164-72:Vegas_Tweets mirandabarrie$
```

🎉 Nice work! You have removed retweets from your collection. 🎉

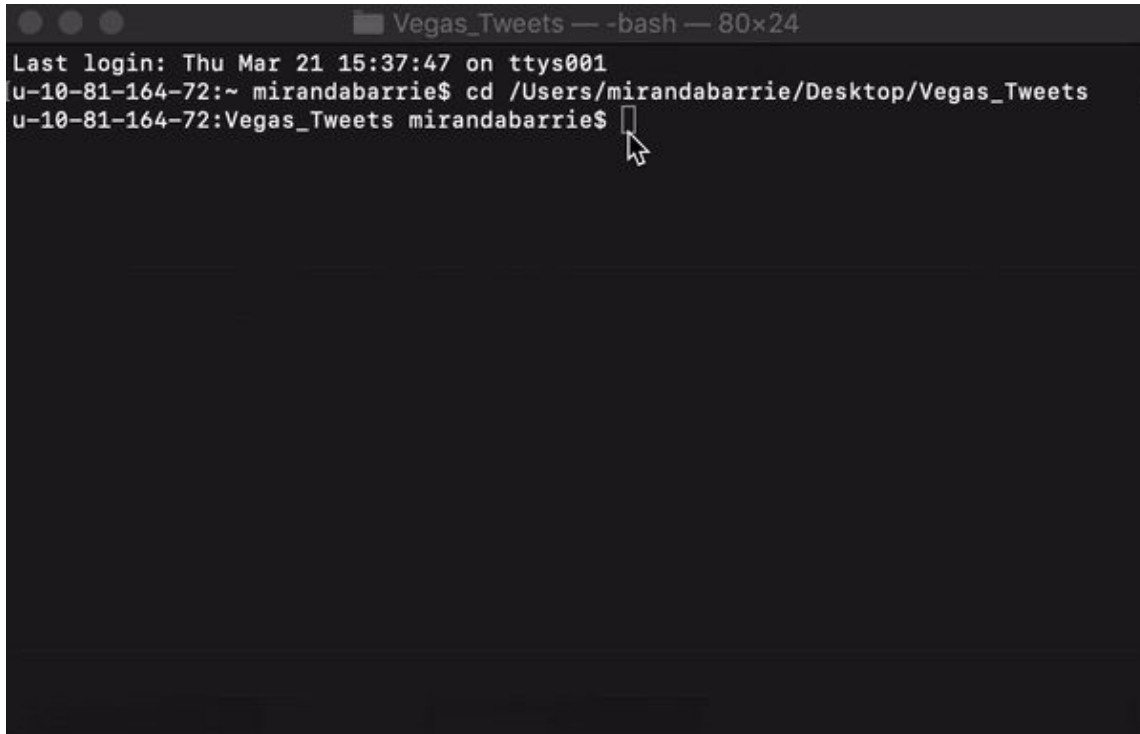
## Remove Tweets Before Certain Date

In the [fourth tutorial](#) in this series, we recommended collecting as much Twitter data as possible. You can always get rid of unwanted data, but it is more difficult (and expensive) to collect historical data. A good example of this is the 1 October Twitter Data Collection. The mass shooting that occurred in Las Vegas took place on October 1, 2017. The UNLV Libraries Twitter data collection contains tweets related to the search term 'vegas' beginning on September 29th and ending October 7th, 2017. Some researchers may find the entire dataset useful, but others may want to only study tweets that were sent during the days following the shooting. Controlling for the date allows you to reduce the size of your collection which, in some cases, can facilitate the process of analyzing the data.

**Important:** Even if you are tweeting from Las Vegas, tweets are recorded in Coordinated Universal Time (UTC). Whenever you consider the time/date of a tweet, remember that all tweets are recorded in UTC.

The first tweet in the original sample collection was sent March 20, 2019 at 01:10:06 UTC. The last tweet in the collection was sent March 21, 2019 at 00:30:28 UTC. We are going to remove any tweets that occurred before March 21, 2019. Enter the following command to do so.

```
python ~/git/twarc/utils/filter_date.py --mindate 21-march-2019 vegas_tweets.jsonl > vegas_tweets_mindate_03212019.jsonl
```



```
Vegas_Tweets — -bash — 80x24
Last login: Thu Mar 21 15:37:47 on ttys001
u-10-81-164-72:~ mirandabarrie$ cd /Users/mirandabarrie/Desktop/Vegas_Tweets
u-10-81-164-72:Vegas_Tweets mirandabarrie$
```

🎉 Nice work! Now you know how to narrow your collection by date. 🎉

## Update Your Collection Documentation

If you end up deciding to make any modifications to an original dataset, it is important to update your collection documentation. In this [updated Collection Documentation Checklist](#), you can document any changes you made to your collection in this tutorial. If you make changes beyond those already featured, add them to the spreadsheet. Your future self will thank you when you prepare to share your research and/or the collection itself.

Collection Documentation Checklist						
*Information below is not related to UNLV's 1 October Twitter Data Collection						
Collection Title	Collection Description	Twarc Command #	API	Twarc Command (Complete Query)	Initiated (Date)	Initiate (Time)
Vegas Collection	Tweets relevant to the search	Command 1	search	twarc search vegas > vegas_tweets.json	03/20/2019	5:30

[Download the Collection Documentation+ Checklist](#)

🎉 Congrats! You've finished the tutorial. Go you! 🎉