Twitter Data Analysis Using Twarc

Analyzing Twitter JSON data

In order to answer research questions using the Twitter data collection you created in the fourth tutorial in this series, you will need to be able to analyze Twitter JSON data. This tutorial explains how to get an idea of what your dataset looks like by performing a simple analysis using Twarc. This snapshot of your dataset can help further define your research questions, and might even lead you to some new ones. For the purposes of this introductory tutorial, we will analyze an existing Twitter dataset from UNLV, the 1 October Twitter Data Collection.

Difficulty level: Beginner

Optimized for: Mac users

Prerequisite(s)

- Tweet JSON
- Command Line
- Collection Design
- Collection with Twarc
- Collection Documentation
- Collection Ethics
- Cleaning Your Data

Tutorial Key

- Command Line arguments will be displayed in this format
- 🎉The party popper emoji signals the end of each set of instructions 🎉

Lesson objectives

- Use Twarc to analyze UNLV's 1 October Twitter Data Collection

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
- 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.
- 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

Introduction

Twarc was created by the team behind Documenting the Now, a project aimed at promoting ethical collection and archival processes for social media data. This tutorial will walk you through how to use Twarc to perform a simple analysis of a collection of Twitter data. The collection we will explore in the tutorial was created by Thomas Padilla at UNLV Libraries and contains over 14 million tweets collected around the mass shooting that occurred in Las Vegas on October 1, 2017.

Download the 1 October Twitter Data Collection

UNLV Access: If you are a UNLV faculty, student, or staff member, you can access the private, full collection.

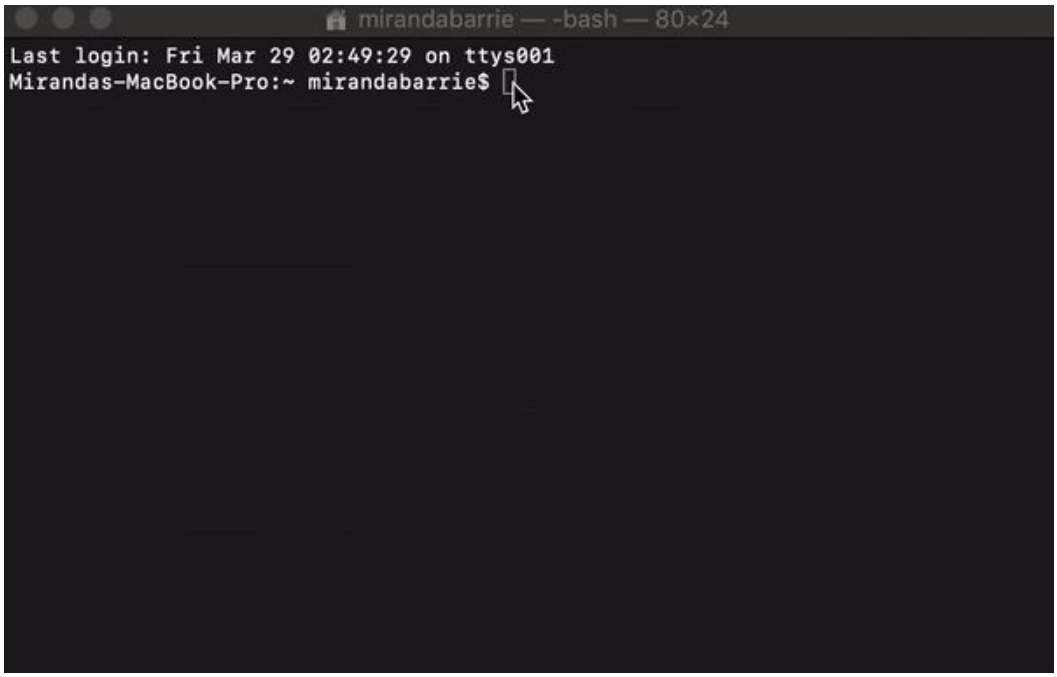
Step 1: Download the collection

1. Start by visiting the UNLV library site to access the private, full dataset. Then download the zipped folder, 'oct1_postdedupe.zip'.

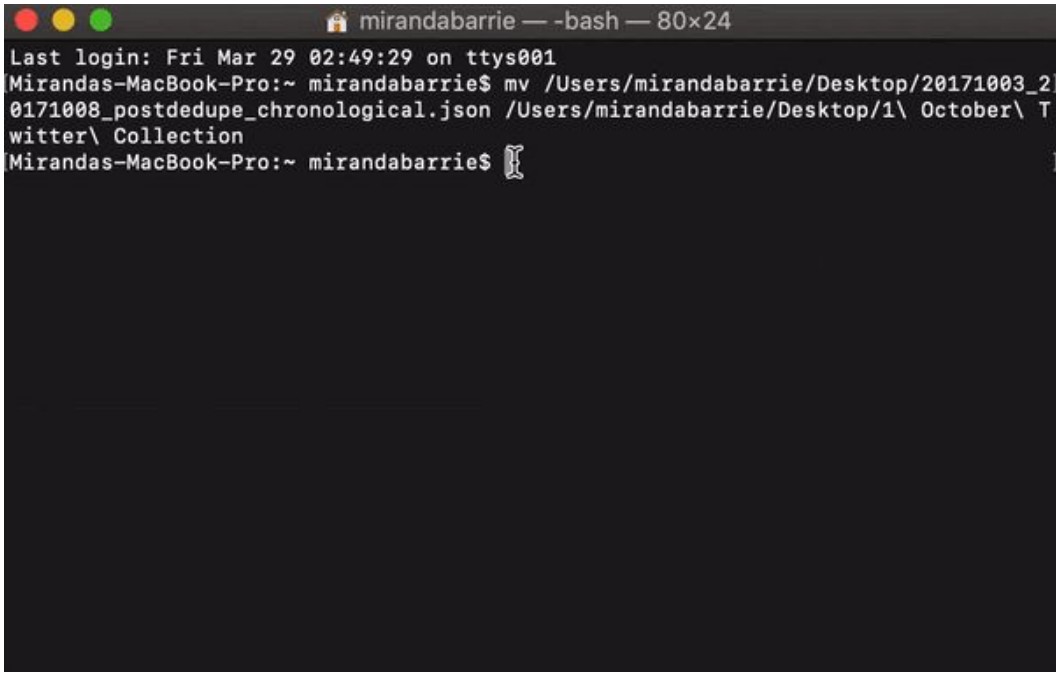
2. Unzip the 'oct1_postdedupe' folder. The file we will be working with today is '**20171003_20171008_postdedupe_chronological.json**'.

3. Create a folder titled '1 October Twitter Collection'.

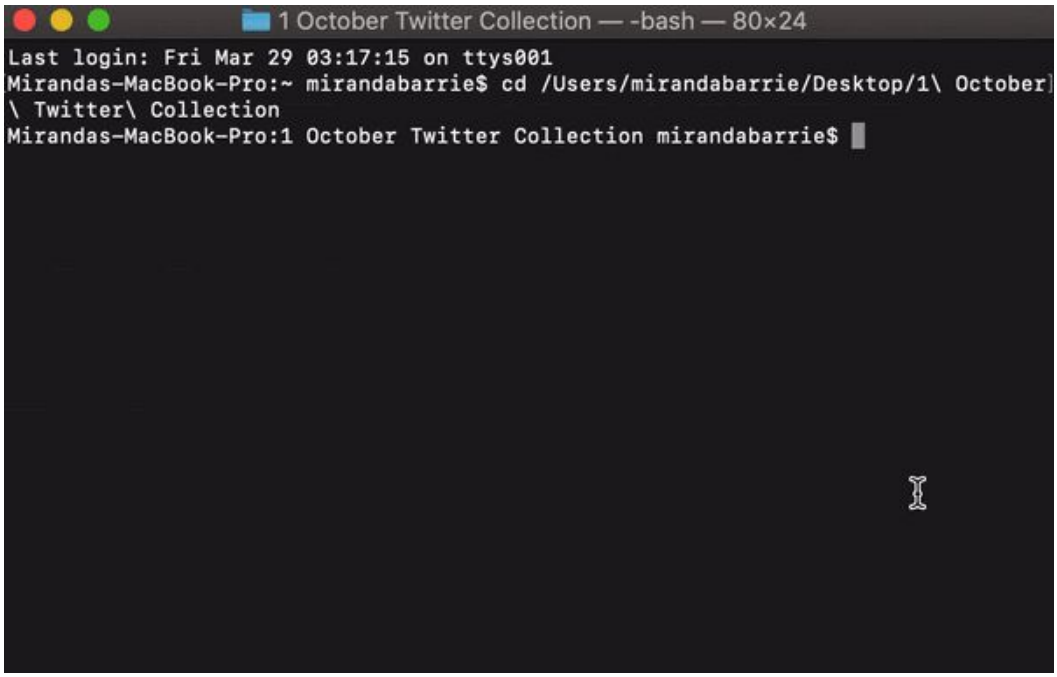
4. Open Terminal and move the '**20171003_20171008_postdedupe_chronological.json**' file to the '1 October Twitter Collection' folder you just created using the move command.



5. Navigate to your '1 October Twitter Collection' directory using the change directory command.



6. Rename the '20171003_20171008_postdedupe_chronological.json' file to '1_october_tweets.json' using the move command.



🎉Great job! You successfully downloaded the 1 October Twitter Collection!🎉

Public Access: All other users may access the public set of tweet identifiers by clicking here. Follow the instructions below to download the tweet IDs and then 'rehydrate' them using Twarc. You can learn how to rehydrate Twitter data in the fourth tutorial in this series.

It is important to note that Twitter limits users to 900 API status/lookup requests every 15 minutes. Each request can hydrate up to 100 Tweet IDs using the statuses/lookup REST API call. This means that every 15 minutes you will only be able to hydrate 90,000 tweets.

900 requests x 100 tweets = 90,000 tweets/15 min
= 360,000 tweets/hour

The full dataset of 14,108,104 tweets will take approximately 39 hours to hydrate.

Important: Rehydrating the tweet IDs will most likely not return the full dataset. Tweets and accounts that have been deleted since the collection period and tweets from now-private accounts will not be returned.

Step 1: Download the collection

1. Create a folder on your Desktop titled '1 October Twitter Collection'.

2. Visit the UNLV Library site to access the public tweet IDs. Save the IDs as a text file to your '1 October Twitter Collection' folder you created on your Desktop.

3. Change the name of the text file to 'tweet_ids_1_october'.

   🎉Great job! You successfully downloaded the 1 October Twitter Collection tweet IDs!🎉

Step 2: Use Twarc to 'hydrate' the collection

1. Open the Terminal application (Located in the applications folder)

2. Change directories by starting the command `cd` and then dragging your '1 October Twitter Collection' folder from your Desktop into Terminal. Hit return to complete the command once you have dragged the folder into terminal successfully.
   Tip: Make sure to leave a space between the command **cd and the filepath.*

3. Make sure you are in the right directory by entering the command `pwd`. You should be in your '1 October Twitter Collection' directory.

4. Hydrate your dataset by entering the following command:

   ```
   twarc hydrate tweet_ids_1_october.txt > 1_october_tweets.json
   ```

   🎉Nice work! You now have your tweets in JSON format ready to go in your '1 October Twitter Collection' folder. You can always check your folder to confirm your .json file is there.🎉

Clone Twarc GitHub Repository

In order to use Twarc utilities, you will need to clone a copy of the Twarc GitHub repository into your local directory.

It's as simple as entering the following command into the command line.

```
git clone https://github.com/DocNow/twarc.git
```

🎉Look at you go! You now know how to clone a git repository.🎉

Twitter Data Analysis Using Twarc

Now that you've downloaded either the private (UNLV access) or public dataset and read the 'readme.txt' file, you are ready to begin analyzing the 1 October Twitter Data Collection with Twarc. Twarc has a number of helpful utilities that you can use to do a basic analysis of Twitter data. Remember to take it slow and double check that you have entered the correct command.

Getting started

1. Open the Terminal application

2. Change directories by starting the command `cd` and then dragging your '1 October Twitter Collection' folder into Terminal. Hit return to complete the command once you have dragged the folder into terminal successfully.
   Tip: Make sure to leave a space between the command **cd and the filepath.*

Number of Tweets

To count the number of tweets in the dataset, enter the command below. The command `wc -l` will count the number of lines in the collection. Since each Tweet is a single line of JSON, performing this command will return the total number of tweets in the collection.

Remember: It will take a few minutes for the process to finish. You can tell when a process is complete when it returns to the shell ($) prompt.

```
wc -l 1_october_tweets.json
```

```
14108104 1_october_tweets.json
```

Check: If you downloaded the private dataset, you should have 14,108,104 Tweets. If you downloaded the public dataset, this number will likely be smaller.

Number of Users

To count the number of unique users in the dataset, enter the command below.

```
python ~/git/twarc/utils/users.py 1_october_tweets.json > 1_october_users.txt
```

To sort the unique users in the dataset by the number of tweets sent by users during the collection period, enter the following command.

```
cat 1_october_users.txt | sort | uniq -c | sort -n > 1_october_users_sorted.txt
```

To find the total number of unique users counted, enter the following command.

```
cat 1_october_users_sorted.txt | wc -l
```

To find the users with the most tweets in the collection, enter the following command.

```
tail 1_october_users_sorted.txt
```

Check: If you downloaded the private dataset, you should see 4,771,228 unique users counted. If you downloaded the public dataset, this number will most likely be smaller.

Number of Hashtags

To find the total number of unique hashtags used in the dataset, enter the following commands.

```
python ~/git/twarc/utils/tags.py 1_october_tweets.json > 1_october_tweets_hashtags.txt
```

```
cat 1_october_tweets_hashtags.txt | wc -l
```

To find the top ten hashtags used in the dataset, enter the following command.

```
head 1_october_tweets_hashtags.txt
```

Check: If you downloaded the private dataset, you should see 70,531 unique hashtags counted. If you downloaded the public dataset, this number will most likely be smaller.

🎉Nice work! You have finished the tutorial.🎉