Data set source:

This data set comes from scraping Twitter over a seven-day period (11-18 Oct 2020). I collected this data to support a local Montana business. From previous conversations with Paul Gladen (UM Blackstone Launchpad <https://www.umt.edu/launchpad/>), I knew of a local business specializing in a nonlethal deterrent (pepper spray) that would be interested in any insight into the sentiment around products such as theirs. I’m vague here because I don’t know if Paul wants me to put the company’s name out over Moodle, but I am happy to connect you if this is a project that you find interesting for practice/resume padding. To collect tweets about pepper spray, I selected five search terms that I thought would tie a tweet to sentiment concerning pepper spray: ‘teargas,’ ‘pepper spray,’ ‘mace,’ ‘nonlethal,’ and ‘self-defense.’

Data set description:

A seven-day scrape of Twitter’s API, using the search words listed above, captured 17,927 tweets (5.6MB). If you are using the code from my repo to perform subsequent pulls of similar data, you should note that I have set the iteration break per search term at 10,000 iterations. None of the search terms reached that limit over the seven days that I searched, but a different seven-day period may result in a larger number of returns. I was aiming for around 100,000 words, with the average tweet being 33 characters and the average word using 4.9-5.1 characters. The estimated 6.5 words-per-tweet meant I needed a little over 15,000 tweets to capture 100,000 words.

The schema for this data set is below:

|  |  |
| --- | --- |
| COLUMN NAME | DATA REPRESENTED |
| ‘HASHTAG’ | SEARCH WORD PRESENT IN TWEET |
| ‘CREATED\_AT | DATETIME OF TWEET |
| SCREEN\_NAME | SCREEN NAME OF USER |
| DESCRIPTION | USER-CREATED SELF-DESCRIPTION |
| FOLLOWER\_COUNT | NUMBER OF USER FOLLOWERS |
| FULL\_TEXT | CONTENT OF TWEET |

There are several limitations to consider when cleaning/analyzing this data set. First, columns are not always tab-delimited. Many users insert tabs into the full\_text field or description field, which will cause errors when relying solely on tabs to designate columns. It may be useful to establish columns using quotations as the individual tweet data are captured as a list with quotes encasing the text-centric fields such as full\_text. Second, Mace is a famous pepper spray brand and a medieval weapon, rapper, and company name. Using the search word ‘mace’ requires an evaluation of tweet content to ensure the tweet is aligned with pepper spray. Lastly, I used the column name ‘hashtag’ to designate search words, but it is worth noting that these are not #hastags in the sense of Twitter usage.

Descriptive statistics:

These data are presented in a tab-delimited text document. After case-folding and splitting by tab, there were 479,762 tokens. Removing stop words, the dataset contains 250,927 tokens, of which 23,161 are unique. The clean (case-folded/alphanumeric/excluding stop words) data demonstrate a lexical diversity of .092 with an average token length of 5.67. I have included a notebook entitled Pepper\_analysis if you would like to build upon this cleaning. The order of cleaning in the Pepper\_analysis notebook is what I found to work best to avoid splitting words into letters or returning errors.

Interesting question:

The most interesting question that I thought about with this data set is a sentiment analysis concerning the use of pepper spray. The opinions on Twitter appear to range from the use of pepper spray as a form of police brutality to pepper spray representing a life-saving nonlethal alternative in a country with too much gun violence. Some statistics quantifying those thoughts would be informative:

Next steps:

The tweets are grouped by the search word used to collect them.

My next steps would be:

1 Use nltk, textblob, etc. to qualify text content, eliminating unrelated usage of search words and classifying associated words with perceived sentiment.

2 Disaggregate search terms and visualize a time series of sentiment. Several more seven-day collection periods may be required.

3 Mark the time series with significant events such as crescendo in riots, etc. This is where location data for the tweets would be useful. Unfortunately, Twitter’s location and coordinate fields are user-dependent and often utilized for additional text rather than a true location (hence they were not included in this data set).