Data Mining UROP Report

Ding Ma and Duy Ha Khuong

1/14/2015

During the first week of IAP, we got started on the Data Mining Exploration Task for the MIT Mobile Experience Lab Millennial Observatory. Our overarching goal is to determine what millennials (young adults ages 18-30) are saying about aspects of urban life, in particular sentiments about Transportation. In order to do this, we decided to split up the data collection from two main social media sources: Twitter and Facebook. We will focus on Megacities such as San Francisco, New York City, and Boston, as well as various foreign cities. Below is a summary of our findings so far, as well as steps for further research.

**Twitter**

The Problem

For the first week, the problem we were trying to solve is primarily data collection through web scraping. We felt that once we figure out how to get the data, analyzing the data would be much easier once we were given more direction on what kind of stuff to pay attention to. Thus, we focused mainly on getting the raw data, and we will try to make conclusions out of the data later. For Twitter, we tried to extract tweets that fit a certain search query that we were looking for, in an efficient way such that similar processes could be used to replicate the results easily.

Tools/Techniques Used

The main programming language used was Python, and within Python, there were many packages that were installed for data collection and analysis. The tweepy package contained many functions for getting data directly from Twitter, using the Twitter API. I signed up for a Developer’s account using my Twitter account, and was able to get access to the API. The pattern package allowed you to enter a sentence, and it would return the sentiment of whether the sentence was positive or negative.

We programmed a Python file, transportationTweets.py, for performing the web crawl and storing the data in a .csv file, which can later be edited in Excel. We have sections of code that can be commented out, which represent the different cities and the search queries. Eventually, we will streamline the code so that it will be easier to pass in new entries. One example of a search query is:

bostonTweets = tweepy.Cursor(api.search,

q = "Transportation OR MBTA OR subway OR amtrak OR bus OR commute OR rail OR parking OR walking OR biking OR port OR boat OR airport OR uber OR cab OR cab",

geocode = "42.3581,-71.0636,20mi",

since = "2015-01-04",

until = "2015-01-09",

).items()

In the parameter q, we can pass in what the search query should be. “OR” indicates that we are searching for either one of these words, or a combination of them. For each city, we should vary some of the city-specific words such as MBTA, but the general words relating to transportation can be kept the same. The geocode indicates the location of the city, in latitude and longitude coordinates. The 20 miles indicates the radius of search. The since and until allow you to specify from what dates you want to search from. Note that Twitter search API only allows you to access back a week’s worth of data. If you want to go back further, you need to find an alternative way of doing so, such as going into each user’s profile individually and extracting all the tweets from certain users.

-Tweepy: <http://www.tweepy.org/>

-Pattern: <http://www.clips.ua.ac.be/pattern>

-Twitter API: <https://dev.twitter.com/rest/public>

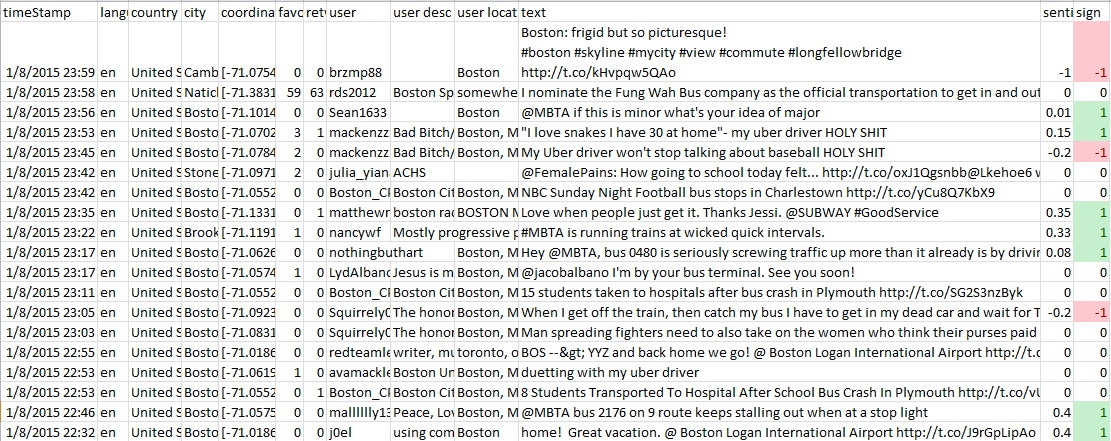
-Useful tutorial on web crawling on Twitter: <http://knightlab.northwestern.edu/2014/03/15/a-beginners-guide-to-collecting-twitter-data-and-a-bit-of-web-scraping/>

Pros and Cons

The pros of using Twitter were that there were lots of data readily available. Because a wide majority of Tweets (90%) are public, it is one of the most accessible social media sites, when compared to other social networks such as Facebook that have lots of private users. Also, the Python program runs pretty quickly and gets a fair amount of data. There are many data fields that can be collected, which include timestamp, language, country, city, coordinates, favorites, retweets, user, userDesc, userLoc, text, and sent, all of which provide opportunities for further analysis. Also, now that we already have a general code skeleton for scraping from Twitter, making modifications for new cities or search queries is fairly simple.

One of the major cons of how we performed our search query that it was dependent on either the tweet’s or the user’s location being specified. However, only about 1% of all tweets are geotagged with the location. Thus, it is missing the other 99% of the tweets, and we must assume that our sample size will be large enough to make meaningful conclusions. Otherwise, there isn’t really a good way to determine what location a tweet is from, for general transportation queries such as “subway,” because a tweet might not necessarily contain the location of interest. However, we could also expand our search by including the name of the city in the search query, rather than searching by geotagged locations, and seeing if this provides more results. Additionally, there isn’t an easy way to specify the user’s age in the search query. Thus, the only way to determine the age is to go into each user’s description and seeing if they provide that information.

Results

We collected data from Boston, New York, San Francisco, Glasgow, and Paris, and stored them each in a separate Excel file. I calculated the average sentiment for all the tweets about transportation each city, and they were as follows: 0.0078, 0.0299, 0.1090, and 0.0241 (Paris tweets have not been sentiment analyzed yet). I also added a new column that converts each sentiment number to a sign, either +1 (positive), 0 (neutral), or -1 (negative), and made it into easy-to-see colors. Below is a screenshot of the Excel File:

Conclusion

Because we were focusing on the first part of data collection, we could not draw too many conclusions yet. However, we could see that just from the raw data, the overall average sentiment for each of the cities was positive. This might be because this includes the tweets from the transportation organizations themselves, which might contain overly positive words that make the data more biased. Also, it was interesting to note that Boston had the least positive sentiments, followed by New York, followed by San Francisco. This makes sense because people generally have the opinion that public transportation in Boston is not very good, while transportation in San Francisco is pretty good. The various differences between cities will be looked into more in-depth later.

Further Steps

This week, we are planning to investigate the Twitter data more in-depth to try to draw more meaningful conclusions. First of all, instead of simply looking at whether a tweet’s sentiment is positive or negative, we should try to extract more out of the contents of the tweet, to find out the major opinions that people have on certain aspects of urban life. One way of doing this is to perform a word count analysis of the most frequently used words in the tweets. For example, we could narrow the search down to just the query “MBTA” in order to find out what are the most common phrases said about the transportation system. We will try to ensure that all the tweets we are analyzing are from millennials by trying to figure out the user’s age from the user’s profile description, and, if necessary, using Python packages to try to predict the age based on his or her previous posts.

If we have more time, we can start to investigate other factors besides Transportation, such as wellbeing (exercise, health, happiness), food (restaurants, markets, cooking), public parks (green spaces, recreational spaces), art and culture, and work/employment (job search, unemployment, etc. In general, we will try to be able to extract meaningful conclusions out of the data, rather than just having the raw data without any explanation. We can try comparing how aspects differ between different cities, and whether certain cities have more positive sentiments than others on certain topics. Additionally, we will look into other possible sources of data, such as newspapers. We can extract contents or summaries of newspaper articles and try to determine what are the most common things people talk about, especially opinions that people are quoted as saying.

**Week 2 Progress**

As mentioned in the “further steps” above, one of the key factors is discriminating between tweets from people who fall in the millennial age range (18-30), and other tweets. This is because our study focuses specifically on what millennials think about various aspects of city life. On Twitter, there is not a easy way to obtain an user’s age, as that information is not publicly available. Some users put information in their description that would indicate their age, such as being a college student. However, less than 5% of all profiles have such information. Thus, we needed to figure out a way to figure out the user’s age just from the tweet itself.

We decided to use a python package that contains a machine learning classifier, **Ageanalyzer,** to analyze each tweet’s text, and based on the words and phrasing in the text, determine what age group it likely came from. This package, called **uclassify**, automatically returns the percentage likelihood of a certain text being written by a certain demographic. We simply take the age range with the greatest probability that the text came from, as the age of that Twitter user. We choose to include the age ranges 13-17, 18-25, and 26-35 as valid age ranges to be considered part of our sample set. (We include 13-17 because it is difficult to distinguish between teenage and young adult speech from just a single tweet.) We ran into the minor issue of needing to create an account on UClassify to access the API read and write keys, and we were limited to below a certain number of classifications per day allowed, but this was easily overcome by creating multiple usernames and using those new API keys.

Python package: <https://pypi.python.org/pypi/uclassify/0.1.0>

UClassify website: <http://www.uclassify.com/browse/uClassify/Ageanalyzer>

When we ran this classifier on multiple tweets, about 50% of all the tweets we decided to keep as being from a millennial age group. This is fairly accurate, as we know from published data that 40% of all Twitter users are in the age 18-35, and in our classifier we over-sampled a bit. We considered going through each user’s tweets and analyzing all the text from their first 200 tweets, but this proved to be too time-consuming (2 seconds per person), as well as getting blocked by Twitter API for accessing too much user information too quickly.

We basically add this “age” into another column of the output .csv file. Once we create the file, we need to make sure the users are actually real people, not the transportation agency or a news account dedicated to track transportation. Thus, we can sort by the user in Excel, and then manually erase all the tweets that are from non-people organizations (this should be pretty easy and quick to tell manually, from both the user name and description.) Another issue we ran into was sometimes, our search query yielded results in different languages, that do not use the word in the context we want to (for example, “MTA” is an actual word in another language). Thus, we decided to specify lang = “en” in our search.

The entire code of transportationTweets.py was made more streamlined and easy-to-understand, with everything broken down into different methods. First, we set up the keys for both the Twitter and UClassify API. We have a dictionary of city to its geocode , in order to search by location, as well as a dictionary of user to age to be used later to store the ages of all the unique users. The main method we are calling is initialize(city, query), which we call by typing in our desired city and query from IDLE (or other Python runner) once the program is running. This basically creates the .csv file to be stored in, does the search query, then calls scrape(tweet, fileWriter) for each tweet in searchResults. Scrape is the method that retrieves data from the tweet, and writes it into the .csv file. Within this method, we call checkEncode(text) to make sure the text can be encoded into a format storable in Excel. WE also call sentiment(text) to analyze the Tweet text’s sentiment. We also call checkTweetAge(text, user) to use UClassify to determine the age of this user.

Below are some of the preliminary results from comparing various different search queries for the three different cities in the U.S., Boston, New York City, and San Francisco. Keep in mind that Twitter only allows us to get the past 5 days or so’s worth of data using the API search. The population of Boston is 6, New York City is 82, and San Francisco is 7.77. % positive means the % positive out of the tweets that actually had an opinion.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Search Query | Number tweets | # normalized for population | Average sentiment | % positive |  |
| MBTA | 562 | 93.7 | -0.145 | 50.3% |  |
| MTA | 867 (timed out) | 10.6 | 0.023 | 59.4% |  |
| Caltrain | 594 | 76.4 | 0.059 | 62.5% |  |

“subway OR train OR rain OR metro”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 112 | 18.7 | 0.025 | 47.0% |  |
| New York City | 858 | 10.5 | 0.041 | 62.7% |  |
| San Francisco | 108 | 13.9 | 0.058 | 71.7% |  |

“bus”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 110 | 18.3 | -0.044 | 39.0% |  |
| New York City | 348 | 4.24 | 0.021 | 56.1% |  |
| San Francisco | 87 | 11.2 | 0.046 | 55.2% |  |

“uber or lyft”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 72 | 12 | 0.045 | 64.9% |  |
| New York City | 131 | 1.6 | 0.069 | 63.2% |  |
| San Francisco | 39 | 5.0 | 0.100 | 77.3% |  |

“taxi OR cab”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 15 | 2.5 | 0.078 | 62.5% |  |
| New York City | 150 | 1.83 | 0.007 | 52.6% |  |
| San Francisco |  |  |  |  |  |

“biking OR bikes OR bike OR hubway/citibike”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 22 | 3.67 | 0.077 | 66.6% |  |
| New York City | 52 | 0.63 | 0.132 | 80.6% |  |
| San Francisco | 31 | 3.99 | 0.085 | 63.2% |  |

“ferry OR boat OR ship”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 32 | 5.33 | 0.040 | 57.9% |  |
| New York City | 122 | 1.49 | 0.121 | 70.4% |  |
| San Francisco | 25 | 3.22 | 0.184 | 68.8% |  |

“airport OR plane”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| City | Number tweets | # normalized for population | Average sentiment | % positive |  |
| Boston | 165 | 27.5 | 0.037 | 60% |  |
| New York City | 316 | 3.85 | 0.051 | 65.1% |  |
| San Francisco | 86 | 11.1 | 0.095 | 70.5% |  |

**Further Analysis**

We will put everything we have so far in nice-looking graphs and charts for data visualization. We want to not just be able to compare the numbers, but also get more qualitative data on what exactly people are saying about each aspect of transportation. To do this, we will do a word count study of the most commonly used words. We will do this for each Excel file. Also, we can analyze factors beyond transportation, to include other facets of urban life such as wellbeing (exercise, health, happiness), food (restaurants, markets, cooking), public parks (green spaces, recreational spaces), art and culture, and work/employment (job search, unemployment, etc. We will also look at alternative sources besides just Facebook and Twitter, for example, newspaper articles and possible blog posts.

**Week 3 Progress**

One of the things that was mentioned was, what do we do about image-based data from Twitter? This is something that cannot be easily analyzed by a computer, so we decided to find a way to store the images, and then view them manually later. I found this useful code for image crawling using tweepy. Basically, for each tweet, I analyze whether it contains a picture, and if it does, store it in a file, as well as the original text of the tweet as a text file so they can easily be compared, and see whether the text is sarcastic. Later, we’ll figure out a way to compile all these pictures and texts into an Excel file. The python package, **urllib**, was installed in order to extract images from their URL.

<https://gist.github.com/zmwangx/6191892dc6a96baebdce>

However, we could not get it working properly. The images would only store properly if we extracted a tweet from a user, not a search query, for some reason. Also, we noticed that the percentage of tweets that actually contain an image was very small, less than 2% of all tweets. After we collect the images, we would have had to manually go through them anyways to analyze them, as it is very difficult for a computer program to tell what a picture is about. Instead, we can just run a simple topsy search, and select photos only, to check out the relevant photos. Thus, we decided to halt this endeavor.

Another thing we wanted to look at was whether we could use a third-party platform such as **topsy**, a search engine for Twitter, to extract more tweets than just from the past week. Unfortunately, the topsy API requires payment to be able to use (such as $0.50/1000 tweets, based on how many tweets we scrape per second or minute. We will discuss the merits of getting an account for this. In order to get 10,000 tweets (which is the minimum we should aim for in order to have a large enough sample size), we would need to pay $5.00. And this would need to be paid for every single search query we put in. Thus, this might not be a cost-effective solution. Although the current method does not return a large enough sample size, the data should be good enough for most purposes, as we just want to analyze what majority of people think about aspects of city life, not necessarily every single person’s views.

<http://api.topsy.com/doc/>

Next, we looked at what if instead of searching by geotagged location, we include the city in the search query? For example, instead of searching for “boston”, “subway”, instead search for “boston AND subway” so that it searches for the word boston in the search query. A comparison of these two methods for the past week’s worth of tweets resulted in 37 tweets for “boston AND subway,” compared to only 16 tweets for “boston”, “subway.” Let’s take a look at this for New York, which shows similar results, and there were actually too many results that the rate limit for scraping tweets was exceeded!

Additionally, we want to be able to find out what exactly people are saying about various aspects of city transportation, not just the sentiment of whether it was positive or negative. We use the package **nltk,** which contains tools for word processing and other natural language functions. This way, we will be able to determine the greatest word counts for a given search parameter.

<http://www.nltk.org/>

**Overall Conclusions**

A lot of the scraped data is affected by recent activity in the past weeks’ news. For example, the blizzard generated much buzz on Twitter, with many forms of transportation shutting down. This might lead to skewed results, as it might not be indicative of the state of the actual transportation at the time. In order to mitigate this, I decided to limit the date range to the week just before the snowstorm, scraping data from “2015-01-20” to “2015-01-26”.

Unfortunately, it seems that there is no way to get a bigger sample size needed for robust sentiment analysis. The maximum number of tweets allowed to be scraped before the API rate limit is exceeded is around 700, and so it is very difficult to obtain more tweets, especially also considering the fact that you can only search for the previous week’s data to begin with. However, there are still some general trends that came up. See **comparisons.xlsx** for all the data.

Below is a chart of the percentage of tweets that are positive, color-coded by city. As you can see, Boston tends to have lower sentiment than San Francisco, followed by New York. In fact, taking the averages of the sentiments across all categories, Boston is only 74.2% positive, as compared to 85.7% for New York and 78.5% for San Francisco. Thus, this supports the view that millennials in Boston do not approve of their transportation as strongly as those in New York, which is widely regarded as having one of the best public transportation systems in the country.

Here are the specific data for each search category. Note that in these tables below, “% positive” indicates what percentage of tweets with a sentiment are positive; neutral tweets are ignored.

**Subway**

Looking at the number of tweets (normalized for population), we can see trends in which types of transportation are more popular in which cities: Many more people take the subway in Boston (155.8) and New York (203.0) as compared to San Francisco (16.7). In terms of sentiment from the extracted tweets, it seems residents of New York have a much more positive opinion of the subway system, with 90.6% of the tweets being positive, as compared to only 72.2% of the tweets being positive for Boston. This makes sense, as the NYC subway system is usually seen as much more efficient than Boston’s. For San Francisco, there weren’t enough tweets to have any meaningful conclusions.

It seems that a lot of the scraped data is affected by recent activity in the past week. In New York, there was recently a raise from $2.50 to $2.75 for a subway ride, which everyone was talking about. This leads to a skew in the word frequencies analysis (and thus is not included here). Unfortunately, there is no way around this, as the Twitter API only allows you to search for tweets from the past week. Thus, this data should be taken with a grain of salt.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive | KeyWords |
| Boston | 935 | 155.8 | 49 | 0.0227 | 72.2% | city (3) |
| New York | 16643 | 203.0 | 39 | 0.126 | 90.6% |  |
| San Fran | 130 | 16.7 | 50 | -0.0655 | 0.0% | fatal(3), hobbled(3), incident(3), remains(3), system(3) |

**Uber**

Uber seems to be very popular in Boston, having an overwhelming number (1125.2), compared to New York (55.9) and San Francisco (123.0). This makes sense, as taxis are abundant in New York City and there is less need for Uber, relative to the number of people who live there. What was surprising is that even the raw number of tweets was higher in Boston than New York, demonstrating the overwhelming choice of Uber in Boston. People tended to be more satisfied with Uber in Boston and New York, as compared to San Francisco (only 66.1% of tweets positive).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive | KeyWords |
| Boston | 6751 | 1125.2 | 35 | 0.157 | 88.6% | data(60), drivers(46), city(26), better(25), Lyft(20), hours(20), plan(20), transportation(20), urban(20), paid(19), work(19) |
| New York | 4587 | 55.9 | 31 | 0.124 | 88.4% | drivers(20), more(17), than(14), around(13), flights(13) minute(13), promises(12), cabbies(12) |
| San Francisco | 956 | 123.0 | 40 | 0.126 | 66.1% | Reveals(79), Mind-Boggling(77), Statistic(77), Hate(67), Skeptics(67), taxi(64), bigger(42) |

**Taxi**

For taxis, a surprisingly large number of tweets were from Boston, even though New York actually has more taxi usage. This might be because more people in Boston tended to tweet in general, a trend that was true in a majority of the search queries. However, Boston’s sentiment was less positive compared to that of New York and San Francisco, which makes sense, given that taxi transportation isn’t one of Boston’s strengths. As you can see from the keywords, people were talking about alternative services such as Uber and Lyft, indicating that many are making the switch from hailing a cab to simply requesting a ride from their smartphone.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive | KeyWords |
| Boston | 713 | 118.8 | 53 | 0.0218 | 72.2% | Uber(30), Lyft(16), Just(12), Why(12), data(11), drivers(8), sue(7), City(6), How(6), Live(6), help(6), over(6), owners(6), understand(6) |
| New York | 2588 | 31.6 | 46 | 0.119 | 86.8% | Driver(20), Life(16), Visualized(16), citylab(16), driver(11), |
| San Fran | 341 | 43.9 | 50 | 0.122 | 85.2% | Uber(88), taxi(66), than(51), bigger(43), market(29), $(25), industry(22) |

**Airport**

For airports, more people talk about it in Boston and San Francisco than New York. This might be due to the fact that New York’s airports are not very close to Manhattan, which is where a lot of the activity in the city is. Also, people in New York may prefer to use alternative methods, such as bus or train, to travel to nearby cities, as traffic and the distance to the airport might be deterrence. It seemed New York’s airports, however, had higher sentiment than both Boston and San Francisco.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive | KeyWords |
| Boston | 3554 | 592.3 | 59 | 0.0482 | 73.1% | Logan(187), flight(9), home(8), out(8), back(7), security(6), come(6) |
| New York | 7336 | 89.5 | 41 | 0.123 | 85.1% | JFK(25), LaGuardia(15), JetBlue(11), like(11), from(9), snow(8), plan(8), Terminal(7) |
| San Fran | 2909 | 374.4 | 53 | 0.0613 | 70.6% | SFO(44), for (18), Hilton(9), Bayfront(8), from(6), travelling(4) |

**Ferry**

Ferries are much more prevalent in San Francisco than either Boston or New York. This is directly as a result of San Francisco being in a bay, and having multiple interesting places to go via ferry. There wasn’t enough data to draw any meaningful conclusions about the content of the tweets, so ferries are not as significant as other forms of transportation in mega-cities.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive | KeyWords |
| Boston | 102 | 17.0 | 49 |  |  | whale(3), terminal(3), exam(2), dead(2), found(2), study(2) |
| New York | 684 | 8.3 | 26 | 0.175 |  | terminal(3), Space(3) |
| San Francisco | 645 | 83.0 | 55 | 0.0292 |  | Building(4), Japanese(2), unknown(2), warships(2), docked(2), date(2) |

**Parking**

There were an overwhelming amount of tweets about parking for Boston compared to NYC or SF, which makes sense, given that those two cities are more crowded an people don’t drive into the city as much, preferring to take public transportation instead. As for the sentiment, there wasn’t enough data to draw conclusions.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive |  |
| Boston | 1119 | 186.5 | 48 | 0.0458 | 71.7% | …(22), via(20), permits(16), system(13), Event(11), Guarantee(11), Unlimited(11), strain(11), report(10), free(6) |
| New York | 2703 | 33.0 | 24 | 0.104 | 81.4% | lot(13), Knicks(10), PASS(7), Ticket(7), car(6), fucking(6), GEICO(5), Off-duty(5), cop(5) |
| San Fran | 420 | 54.1 | 45 | 0.0298 |  | 49ers(4), carsharing(3), hacked(3), how(3), regs(3), Airport(2) |

**Transportation**

I tried searching the query “transportation”, but it did not achieve any meaningful results.

**Walking**

It seems that Boston is the most walkable city, followed by New York, and lastly, San Francisco. For Boston, the sentiment was lower, but that data might have been skewed due to the negative recent news about someone getting hit while walking or similar type of incident.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | # of Tweets | # Normalized | Topsy Sentiment | Avg Sentiment | % Positive |  |
| Boston | 138 | 23.0 | 41 | 0.0948 | 67.3% | walking(64), dog(17), school(15), Man(14), bus(14), Police(13), hit(12), ATLAS(9) |
| New York | 970 | 11.8 | 23 | 0.122 | 81.9% | around(29), streets(19), new(17), celebrity(16), sightings(16), through(12) |
| San Fran | 107 | 1.3 | 43 | 0.182 | 80.0% | around(8), with(6), love(5), via(4), hills(3) |

**Further Analysis**

Although Twitter was not as powerful of a source as we had hoped, due to the API crawl rate limitations, we were able to get a general idea of how young people ages 18-35 feel about the transportation aspect of urban life. Because we could not obtain the amount of data required for the sample size to be considered comprehensive, Twitter should not be the primary or sole source of analysis; rather, it should be used to supplement and support existing findings. A similar type of study could be done for other aspects of urban life besides transportation. For example, wellbeing can be investigated by using keywords relating to exercise, health, and happiness.