Tarika Gadh, Lauren Mendez, Regine Fae Serafico

Professor Li

MIS3640-01

November 4, 2017

Analyzing tweets of the Houston Astros and Los Angeles Dodgers

Overview:

Given the recent victory of the Houston Astros over the Los Angeles Dodgers in the world series, we decided to analyze the tweets containing each team's name to each to get a sense of the audience's mood towards the outcome. We predicted in general the Astros would have more positive tweets since this is their first world series victory since the beginning of their franchise. We found the most common words used in tweets referring to each team, and conducted natural language processing on these words to better gage the mood of the outcome.

Implementation:

Since we are trying to gage the twitter audience's mood towards the outcome of the world series, we wanted to observe how the mood compared when referring to the Astros against tweets referring to the Dodgers. After requesting 50 tweets, for each team, we ran a function to find the most common words within the tweets. Upon running this function, we realized the most common words were stop words, names, and RT, which appears when a tweet is retweeted.

```
def stop_words(hist):
    stop_words = set(stopwords.words('english'))

    #filtered_tweets = [w for w in word_tokenize if not w in stop_words]
    hist_1 = []
    for w in hist.keys():
        if w not in stop_words:
            hist_1.append(w)
    return hist_1
```

While we could've used the package nltk to not only run sentiment analysis, but also remove stop words, we opted to keep the stop words. The code below shows the function that would've removed the words. However, had we ran this code, we would've lost the ability to count most common words since this code appends words into hist_1 that have not been seen before and are not stop words. Thus, the count for each word becomes one.

Since we still wanted a count to figure out the most common words, we decided to keep the stop words by not running the function above and instead manually omitted these before conducting the sentiment analysis. Since stop words are so common, along with the names of the teams observed, the cities of the respective teams, emojis, and words relating to twitter, such as RT, twitter, and specific links, we increased the number of tweets retrieved and changed the frequency number to fifty. Below is an illustration generated through tableau depicting the top fifty most words within the 200 tweets collected for each team. Finally, we chose the top ten most common significant words to conduct a sentiment analysis using the nltk package. Choosing these top words to analyze gave us a better sense of the overall emotions behind each team's outcome.

Results:

Upon completing our sentiment analysis for tweets about each team, we obtained the following output: Astors, {'neg': 0.0, 'neu': 0.835, 'pos': 0.165, 'compound': 0.9905}; Dodgers {'neg': 0.197, 'neu': 0.803, 'pos': 0.0, 'compound': -0.9831}. Given the nature of baseball a sport, we found the common mention of "trump" amongst the "dodgers" tweets strange. However, upon further research, scrolling through the tweets we retrieved with our keys, we realized that in fact people were tweeting about the LA Dodgers and President Trump (see below). Similarly, we also found tweets pertaining to "justice." As we had predicted, the "astros" tweets were more positive than the "dodgers" tweets. However, the results weren't as extreme as anticipated.

It's import to consider why both sets of tweets were highly neutral. Because the users of twitter and baseball are usually of younger age, they express themselves highly through emojis and punctuation. A sentence of "The Astros won the world series," illustrates less emotion than "The Astros won the world series! :):)" Yet, we aren't accounting for emojis or punctuation. We take them out in order to do the sentiment analysis which may be part of the reason why neutrality was so high for both sets of tweets.

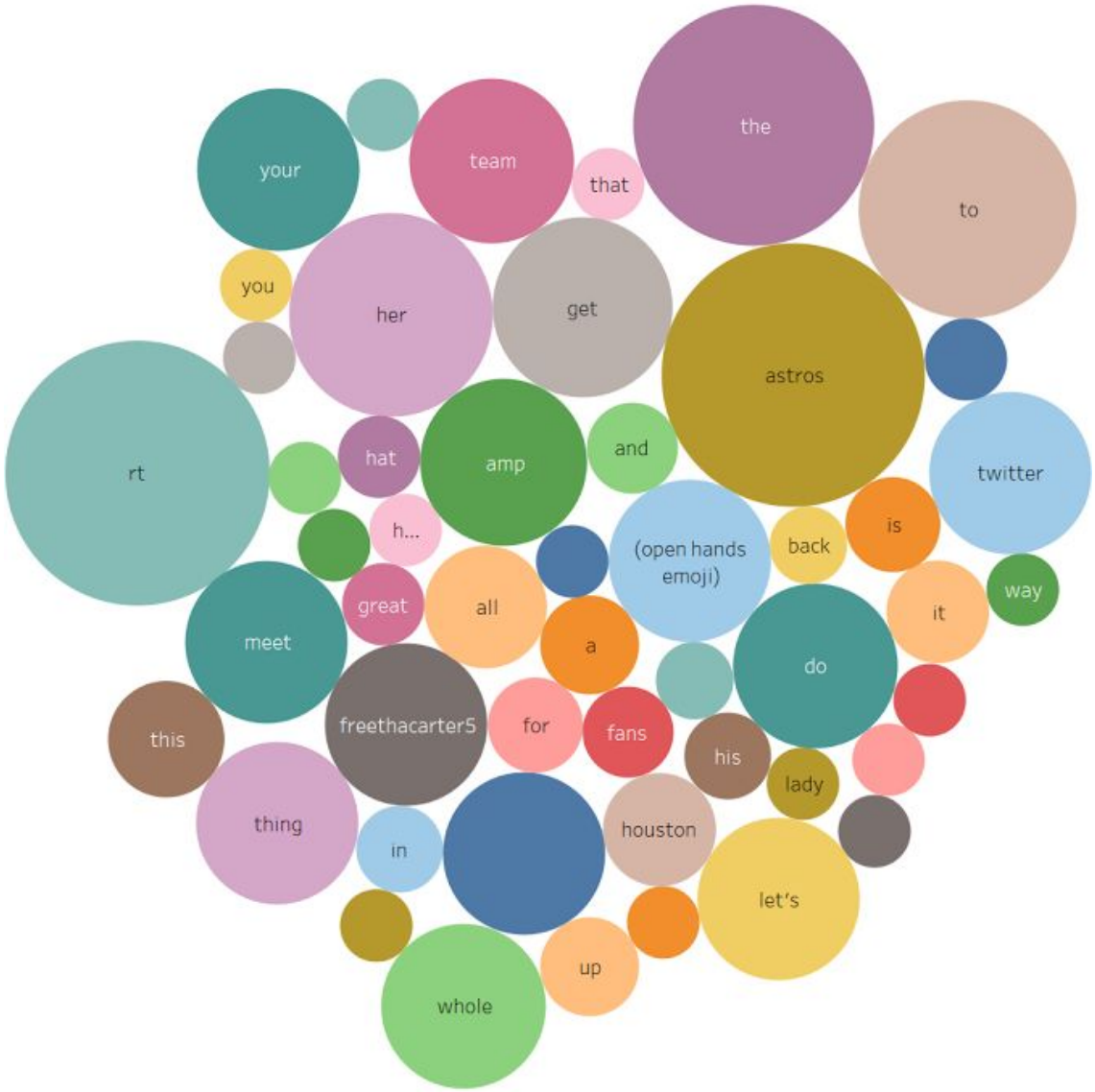
Reflection:

In order to get a lower neutrality score and higher emotional read, perhaps we could have replaced the emojis with their names. Each emoji has a name that can be found here. For example, this emoji 😀 could've been replaced with its name, "grinning face." Furthermore, perhaps partitioning and training the tweets could've lead to us possibly predicting whether a tweet was positive or negative given the set of words. We wished we would've known more about removing certain items from the tweets collected. For example, there were links associated

with retweeting, which can be seen in the top 50 list, which we wish we could have removed through coding. In terms of teamwork, we all worked well and divided the work equally. We found it easier to meet and work on one laptop, as we got our tokens from twitter and set up our base code, and later work individually on pieces of the assignment. We communicated on groupme about what we were doing and hope to continue this for future assignments.

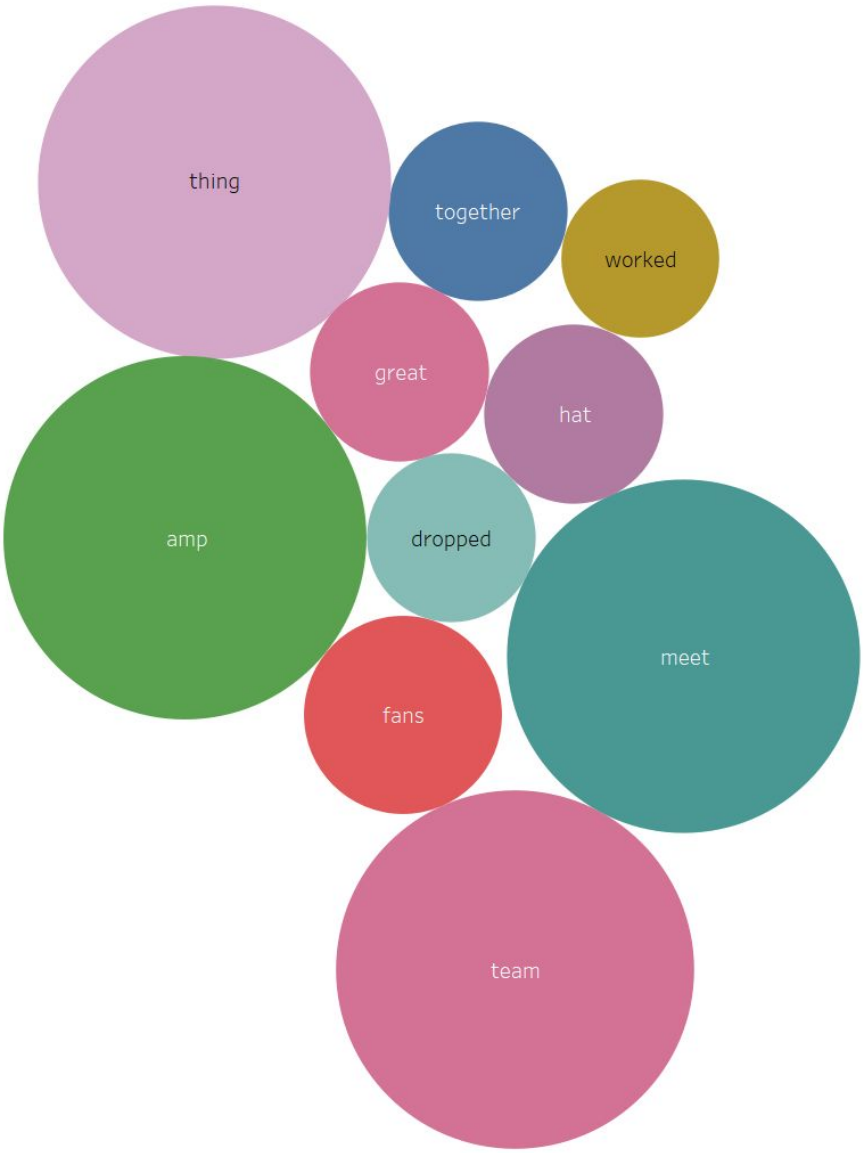
Appendix:

Top 50 most common words amongst 200 tweets containing “astros”:



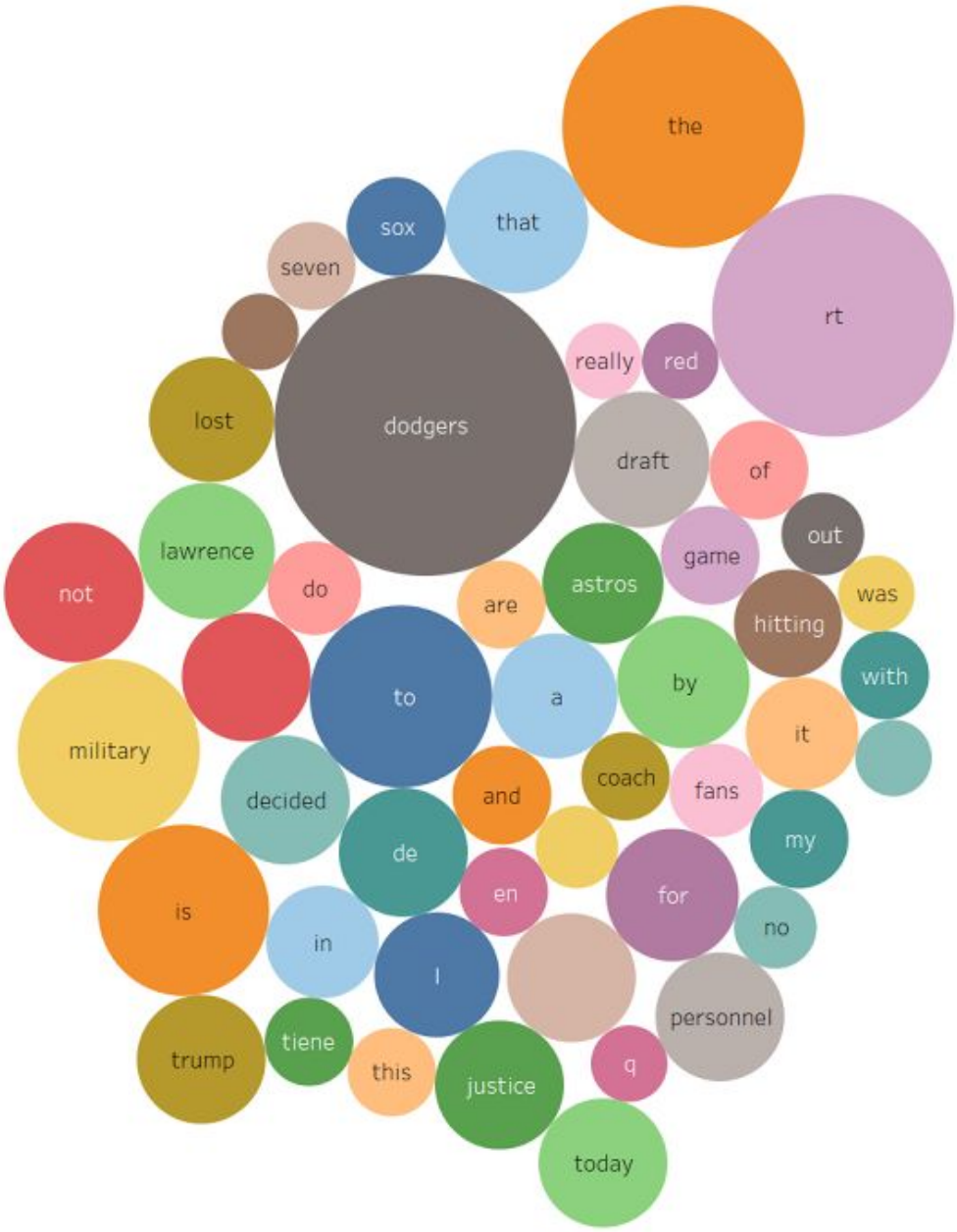
Word. Color shows details about Word. Size shows sum of Count. The marks are labeled by Word.

Top 10 significant words amongst 50 most common words amongst 200 tweets containing “astros”:



Word. Color shows details about Word. Size shows sum of Count. The marks are labeled by Word. Details are shown for Word. The view is filtered on Word, which excludes 40 members.

Top 50 most common words amongst 200 tweets containing “dodgers”:



Word. Color shows details about Word. Size shows sum of Count. The marks are labeled by Word.

Top 10 significant words amongst 50 most common words amongst 200 tweets containing “dodgers”:

Word. Color shows details about Word. Size shows sum of Count. The marks are labeled by Word. The view is filtered on Word, which excludes 40 members.

Tweets containing “dodgers” and “trump”

Tweet containing “dodgers” and “justice”

Raw Data

Top 50 most common words from 200 collected tweets for the Houston Astros:
rt 93
astros 92
the 77
to 63
her 55
get 43
amp 37
whole 36
team 36
do 36
🍊 35
your 35
twitter 35
thing 35
meet 35
let’s 35
https://t.co/0na8nrogxl 35
freethacarter5 35
all 20
this 18
houston 17
it 14
up 13
a 13
is 12
for 12
fans 11
and 11
in 10
his 10
together 9

hat 9
great 9
dropped 8
back 8
😂😂😂 7
şapkasının 7
şapkası 7
şampiyonluk 7
yürüyüşünü 7
you 7
worked 7
way 7
ulaştırılması 7
that 7
taylanozmutlu 7
lady 7
kentmurphy 7
izlerken 7
h… 7
The highlighted words are non-stopwords, used for sentiment analysis

Top 50 most common words from 200 collected tweets for the Los Angeles Dodgers:
dodgers 93
the 60
rt 60
to 34
military 34
is 30
that 21
not 20
lawrence 19
draft 19
for 18
by 18
trump 17
today 17
personnel 17
justice 17
https://t.co/qgdmorrnyl 17
discovered 17
decided 17
de 17
lost 16

i 16
a 16
astros 15

it 13
in 13
hitting 12
sox 10
of 10
my 10
game 10
and 10
fans 9
do 9
with 8
tiene 8
this 8
en 8
coach 8
are 8
7 8
out 7
no 7
darvish 7
worldseries 6
was 6
schoolboy 6
red 6
really 6
q 6

The highlighted words are non-stopwords, used for sentiment analysis