David P. Stonehouse

University of Maryland University College

DATA 650 – Big Data Analytics

Fall 2016, Section 9040, Professor Gortcheva

Assignment 2: Analyze Kombucha Twitter Feed

## Introduction/Purpose

The Southside Medical Clinic prides itself on taking a balanced approach concerning advice to patients who are considering new trends in nutrition. Many of their patients get excited about a new trend that has many followers, but for which there are limited number of clinical studies. In these cases, they seek to understand the patient’s point of view. One new topic is the consumption of kombucha, a fermented drink deemed to have benefits related to “healthy bacteria”. By studying twitter feeds that relate to kombucha, they hope to be able to more effectively engage and influence patients following the kombucha topic on social media.

## Data Exploration

The data for this project was retrieved from IBM’s Insights for Twitter service. This service allows the user to search for a term used in tweets (in our case, “#kombucha”) and pull back the results into database tables. The source data is a subset (random 10% sample) of the full set of tweets someone could download directly from Twitter. The benefit of using this subset is that IBM enriches the data with insights gained from natural language processing (NLP), accomplished with IBM Social Media Analytics. Insights for Twitter attempts to determine the sentiment of the message (positive, negative, etc.) and gender of the sender. The returned data also includes the term used to determine the sentiment. This data (over 8000 tweets) was loaded into IBM’s dashDb database service.

After loading the data, dashDB presents some basic statistics about the tweet set. From Figure 1, it is seen that there has been a steady stream of tweets with #kombucha over the past two years. The monthly tally has not fallen below 250 from the end of 2014 through last month.

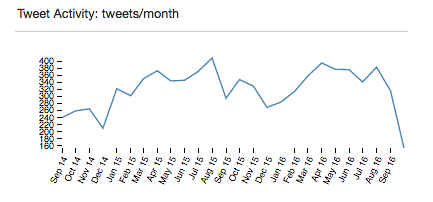


Figure . Kombucha tweets per month in Insights repository.

Looking at Figure 2, many tweets did not indicate the user’s country. For those that did, the majority were from the USA, followed by Canada.

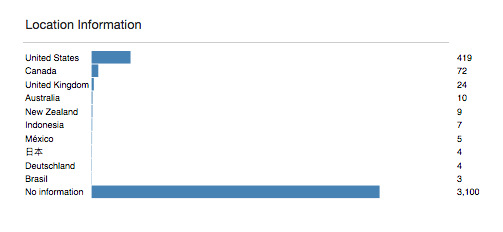


Figure . Locations for senders of kombucha tweets.

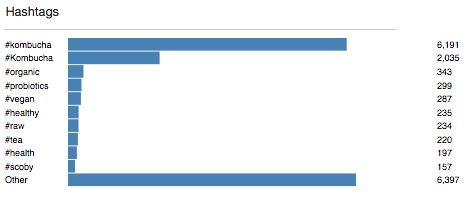
The hashtag information from Figure 3 reflect the relationship between kombucha and some other trends. For example, kombucha is part of a family of products that have probiotics (healthy bacteria). Also, some subset of kombucha tweets are also related to words like “vegan”, “organic”, “raw”, and “healthy”.

Figure . Hashtags in kombucha tweets

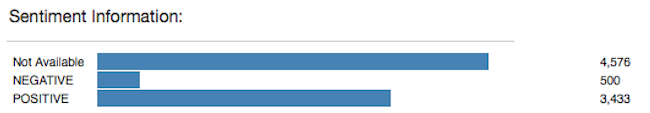
Figure 4 shows the last graph displayed after loading, concerning the sentiment polarities determined by Insights For Twitter. No sentiment was determined for over half the tweets. For those that had a sentiment polarity value, positive sentiments outweighed negative ones 7-to-1.

Figure . Sentiments in kombucha tweets.

The loading process produced a series of tables with the base name “KOMBUCHA” (see Table 1).

Table . Insights for Twitter tables

|  |  |
| --- | --- |
| **Table Name** | **Description** |
| KOMBUCHA\_TWEETS | Main table: text, URL, |
| KOMBUCHA\_HASHTAGS | The hashtags found in the messages, one row per hashtag. |
| KOMBUCHA\_LINKS | Expansion of tinyURL links in messages to the full URL. |
| KOMBUCHA\_LOCATIONS | Lat/Lon location – user profile, message transmission. |
| KOMBUCHA\_MEDIA | Expanded URLs media locations |
| KOMBUCHA\_SENTIMENTS | Polarity and text that indicate sentiment |
| KOMBUCHA\_USERS | User Id, name and screen name |

All of the tables have the key MESSAGE\_ID (e.g., “tag:search.twitter.com, 2005: 600637624315813888”), so that the data from different tables can be joined together as needed.

An example set of tweet texts from the KOMBUCHA\_TWEETS table is shown in Table 2.

Table . Example tweet texts.

|  |
| --- |
| **MESSAGE\_BODY** |
| RT @craftyleftdee: Slumped brown #Kombucha bottle for spoon rest #ecofriendly #etsymntt CraftyleftDee http://t.co/Qt75VOMuY8 via @Etsy |
| #Kombucha: Thrill Hill, Ginger Berry, Thai Temple (Lemon Grass, Lime Leaf) |
| I'm really enjoying this Ginger #kombucha from @livekombucha - and the cap says "we're betterâ€¦ http://t.co/MEfCKBrNXX |
| Homemade #kombucha first taste. So freaking delicious! This is my fermented beverage for the #GreyCup |
| Burnt cookies, #buchi\_kombucha, and #bongojava coffee = good morning! #wakeup #kombucha #paleo #grainfree http://t.co/TPis5SlC8a |
| RT @dbzweier: #kombucha for breakfast, lunch and dinner gets expensive. So I decided to brew my own: http://t.co/GkXbWtC5vy |

Table 3 shows example user information from KOMBACHA\_TWEETS. Note again that USER\_GENDER is a field generated by IBM Social Media Analytics.

Table . Example rows of user information

|  |  |  |  |
| --- | --- | --- | --- |
| **USER\_ GENDER** | **USER\_SCREEN\_NAME** | **USER\_SUMMARY** | **USER\_LOCATION\_DISPLAY\_NAME** |
| female | adams\_madeline | An artist in Alabama. |  |
| unknown | AileenMcGraw | A little heart and a whole lotta floral. | Evanston, IL / Elsewhere |
| unknown | amphore\_oz | Producer of Sydney's finest Ginger Brew Kombucha & the super-probiotic, dairy-free, sugar-free Coco-Kefir. | Sydney, Australia |
| female | annedooner | minneapolis. athleta. W hotel. travel. fitness. fashion. food. Instagram: lizannedooner | Edina, MN |
| unknown | AquaPamela | Tillbaka till grundlÃ¤ggande | estocolmo |
| Female | bAdLadyVet | #Veteran #Student #Millennial #MotivatingTeenSpirit #Volunteer #Advocate #Feminist does NOT = man haterâœŒï¸ #Videographer #SocialEntrepreneur | California, USA |

Some examples of negative and positive sentiment words that Insights for Twitter detected are shown in Table 4.

Table . Kombucha sentiment information

|  |  |
| --- | --- |
| **SENTIMENT\_POLARITY** | **SENTIMENT\_TERM** |
| NEGATIVE | nightmares |
| NEGATIVE | problem |
| NEGATIVE | ran out of |
| NEGATIVE | miserable |
| POSITIVE | best |
| POSITIVE | Love |
| POSITIVE | tasty |
| POSITIVE | beneficial |

## SQL Queries

Several SQL queries were run to gain insights into the data. The first one (Figure 5), is the first in a series of three queries to understand the negative sentiment messages.

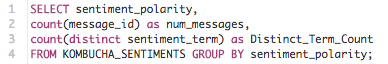
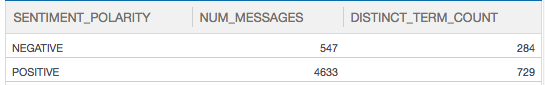


Figure . Query 1: Pattern of unique terms for each sentiment polarity.

The results (Table 5) indicate that over half (284 out of 547) of the negative tweets were analyzed with a term only used once. In contrast, almost 4000 tweets used repeated positive terms. That could make for a good word cloud.

Table . Results for unique sentiment terms.

The next query (Figure 6) found the top terms related to negative sentiment tweets.

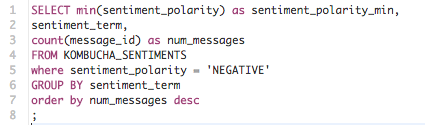
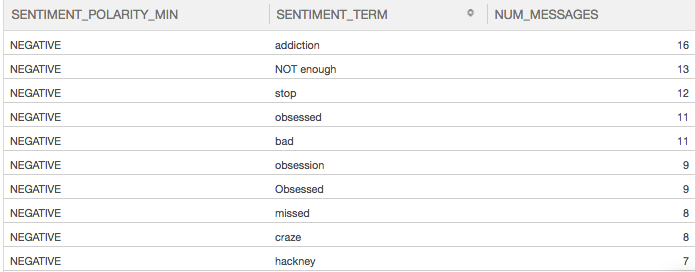


Figure . Query 2: top terms used to determine negative sentiment.

The results (Table 6) indicate that “obsessed”, “addiction”, “not enough” and “stop” are the top words. It is interesting to note that Insights for Twitter makes use of “not” as a contextual influence to determine sentiment.

Table . Results for message counts for the top negative terms.



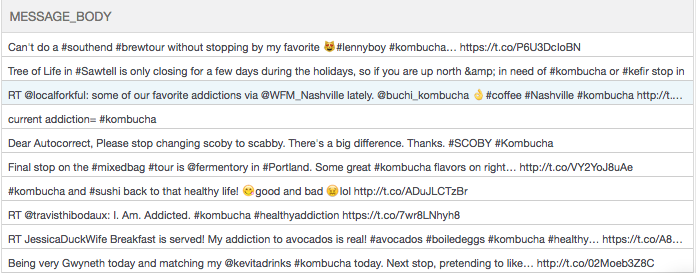
The third query (Figure 7) retrieves the messages that contain some of the top negative terms.



Figure . Query 3: Find matching messages for the top negative sentiment terms.

The results are enlightening. It appears that Social Media Analytics is either employing shallow NLP algorithms or still has a way to be trained for finding sentiment in tweets. As an example, “stop” was the negative sentiment, even though it is part of the phrase “next stop”, so it is a noun. The noun form of “stop” has no negative connotation. If this sample is any indication, there are likely very few negative sentiments in these tweets. The twitter crowd just loves their kombucha!

Table . Tweets containing the top negative sentiment terms.



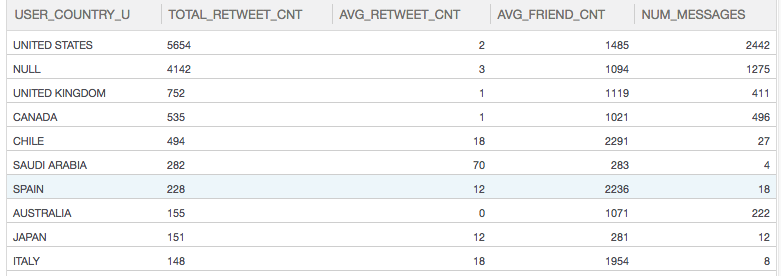
Switching topics, the next query (Figure 8) aims to find out which countries do the most retweeting. In support of this query, a view called SENTIMENT was created. The code to create that view is in the appendix.



Figure . Query 4: List the top countries for retweeting.

The results (Table 8) show the USA as the top country, which is not a surprise since Americans tweeted the most about kombucha. An interesting note is that the few messages from Saudi Arabia kombucha tweeters were retweeted 70 times. Similarly, influence is expanded in Chile and Spain.

Table . Top 10 countries for retweeting.



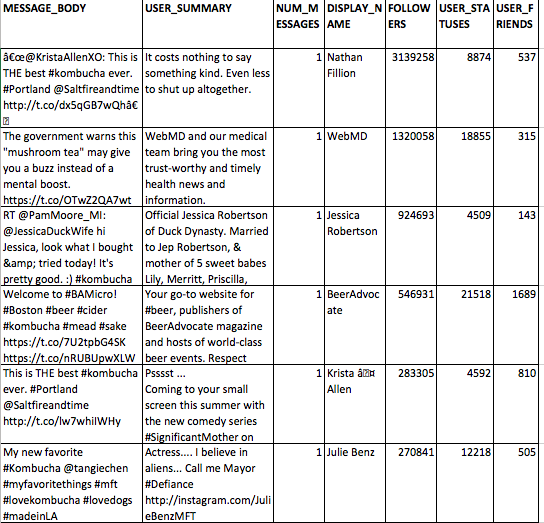
The last query (Figure 9) is an attempt to determine the set of influencers who are tweeting about kombucha. These are people who have potential for a very large influence based on their number of followers, friends or their number of status posts. The cutoff values were chosen based on basic statistical analysis in R. See the “Analysis in R” section for more details.

It is interesting to note in the results (Table 9) that the number two most followed user in this set of tweets is WebMD, which clearly has reservations about kombucha. Interesting in this case that these most influential people just tweeted once about kombucha. But with millions of followers, this is possible the introduction of the topic to a new set of people.



Figure . Query 5: Get the user summary and an example tweet for the biggest influencers. The biggest influencers are determined by their number of followers, statuses and friends.

Table . Example tweet and user summary text for the 6 users with the most followers.



## Loading and Validation in R

In dashDB, an R script was created using template code created using an interactive tool. After selecting a table and the desired columns to import into R, dashDB produces R code to load it into a data frame. In order to validate this load, users should run the R summary command. The R summary command An important step before validation was to use the as.factor function to specify that fields related to gender, country and sentiment polarity were factors. After that, the summary results could be compared with the load results from when dashDB loaded the data from Insights for Twitter. An example sanity check is that the tweets per month values as seen in R (Figure 10) match what was seen in the summary statistics graph earlier.

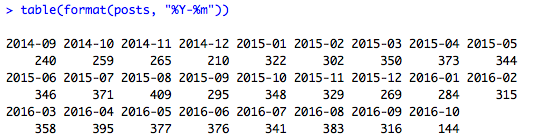


Figure . Table of number of kombucha posts by month.

## Analysis in R

After the data was loaded to R, there were a number of insights gained from doing basic statistical analysis of variables. For example, the summary information related to influencers in Figure 11 were used to set the cut off point for the SQL query in the previous section for retrieving influencers. As will be shown graphically later, there are a few outliers for each variable that greatly skew the data.

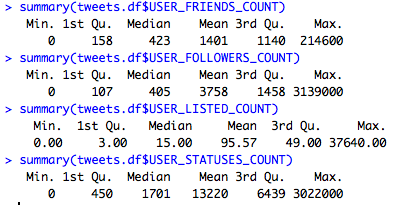


Figure . Summary information for variables that define influencers.

A table called influencers was created specifically to analyze the words used by influencers in their self-description (USER\_SUMMARY) and in an example tweet.



Figure . Word cloud from the user summaries of influencers.

The word cloud in Figure 12 gives a number of insights on who these influencers are. Many of them are passionate about food. They are also specifically into health things like *natural, yoga, organic, fitness, lifestyle, vegan.* They are influencers on Twitter at least in part because it is complementary to another relational activity, like *marketing, business, blogging* and *coaching*. But they are also passionate – note the words *live, love, lover.*  It is interesting to note that *beer* and *kombucha* have about the same level of presence.



Figure . Word cloud from the single sample post from each of the influencers.

The word cloud made by using one tweet from each influencer as input gives some more insights. It is interesting, that these people seem to be writing about the process: *ferment, make, brew, batch, scobi, home, homemade,* and *recipe.* This seems to confirm what Zelman (2016) talks about: many people are brewing kombucha at home. They mention a number of words associated with lifestyle things – *organic, organic, benefit, health, vegan,* From this it would seem that the word on Twitter is that kombucha has benefits and fits well with people pursuing some of these other lifestyle choices.

The clinicians were interested to understand whether there were any weekly or seasonal patterns to kombucha tweets. It does seem that Sunday is significantly quieter (Figure 12), while Tuesday to Friday seem to be pretty even.

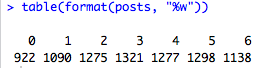


Figure . Number of kombucha tweets by day of the week

The day with the most kombucha tweets was on July 7th, 2015. More were sent on that day than in many whole months.

../../Desktop/Screen%20Shot%202016-10-16%20at%205.36.25%20PM.png

Figure . Busiest day for kombucha tweets.

Looking at Figure 13, there does not seem to be any interesting seasonal, quarterly or monthly patterns.

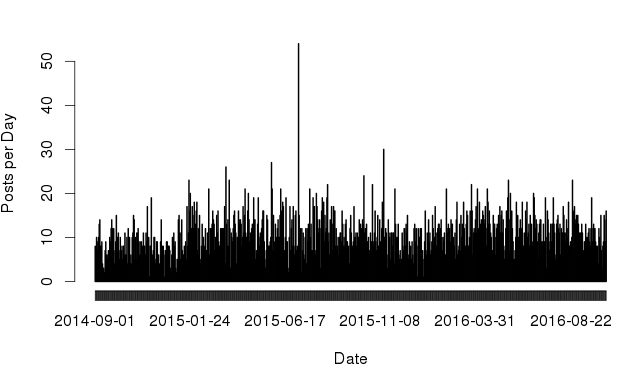


Figure . Plot of kombucha tweets for each day.

The clinicians were also interested in patterns related to gender. The analysts hoped to do analysis of variation (ANOVA) related to gender for the “influencer variables”: number of followers, number of friends, and number of statuses. Figure 15 and Figure 16 illustrate how a few outliers can dominate analysis of these variables. Just by removing the small number of outliers with more than 300,000 followers, the mean and standard deviations change from being drastically different between male, female and unknown gender classifications, to the numbers being quite similar. An attempt was made to do ANOVA on number of followers vs. gender, as well as number of friends, but there were no results that were shown to have significance.

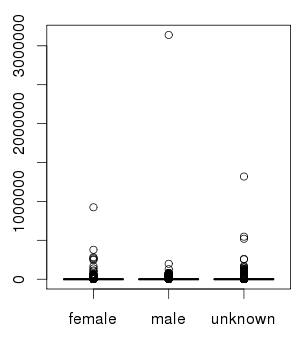
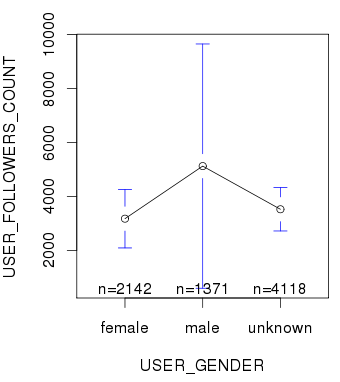


Figure . Plots of gender vs. number of followers: a) mean and standard deviation b) full distribution.

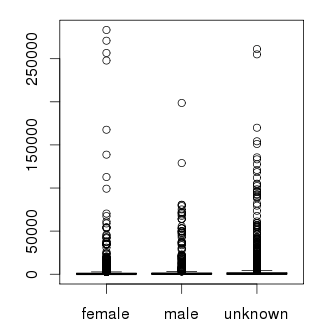
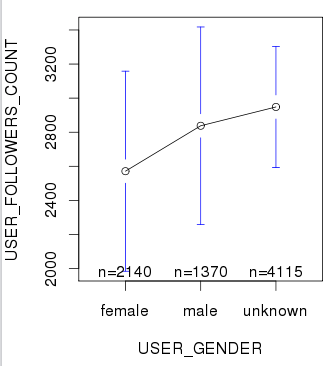


Figure . Plots of gender vs. number of followers with outliers removed: a) mean and standard deviation b) full distribution.

The clinicians were very interested to know how much these trends were taking globally, as that was likely to confirm to patients that viability of the claims. Figure 17 shows the distribution around the world by country and gender. It is interesting to see how the gender gap in interest seems to be significantly bigger in Canada.

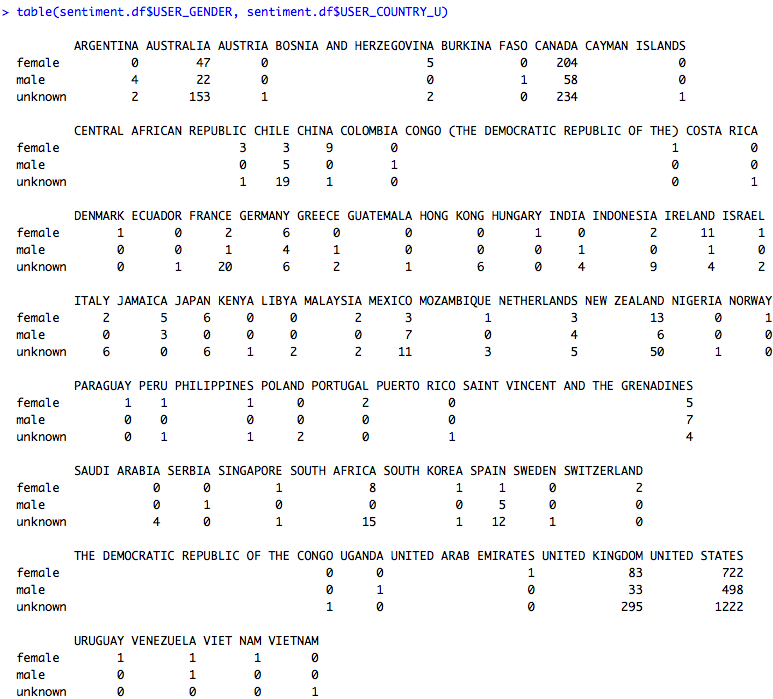


Figure . Distribution of messages by gender and country.

## Accomplishments

As a result of this article, the clinic is aware that kombucha is a global phenomenon. While there seem to be more women tweeting about kombucha, both men and women are in positions of great influence. The trend for kombucha seems solid. The number of monthly kombucha tweets jumped up in twitter feeds starting at the beginning of 2015 and has stayed steady.

A key result was with regard to people who had many followers, friends and/or posts (called “influencers” for this discussion). The word cloud studies related to those with a lot of influence on twitter would seem to indicate a couple things. One is that influencers express passion, they love food, and kombucha is an outworking of an interest in lifestyle and health. Another is that there does seem to be a lot of interest in home brewing of kombucha, which is a key concern among critics.

The ratio of positive to negative tweets about kombucha is seven to one according to IBM’s Social Media Analytics (SMA). However, the approach taken by SMA seems suspect for this particular topic. In the initial sampling of tweets classified as negative, none of them were negative, and many were clearly positive.

## Limitations and Future Research

One area of interest would be to see if there is a way to harness Watson to do sentiment analysis. This would be a nice improvement on the shallow NLP done by Social Media Analytics (SMA). As mentioned before, SMA does not appear to be even looking to determine the sentence structure and applying that semantic information to evaluating terms. A term like “stop” as a noun should not be a negative sentiment word. The Watson application that competed on Jeopardy does that well.

Another enhancement would be to evaluate the full set of tweets on this topic. Insight for Twitter only gives access to 10% of the tweets, while a direct download from Twitter would provide access to all of them. In our case, a ten-fold jump could potentially provide a view on more countries interested in kombucha.

Looking at the MESSAGE\_BODY field, many tweets have emoticons. This subset could be analyzed as another factor to determine sentiment.

The study above focused a lot on number of followers. It would be interesting to study “kombucha channels” – users who tweet a lot about kombucha, regardless of how influential they are. These might represent people like local sellers. That might give an idea of how kombucha is marketed at a local level.

A study of just the hashtags could give a different insight into related trends.

## Conclusions

The clinic should use the twitter information to look for ways that the users confirm that kombucha is a healthy addition to their diet. There should be a low expectation of finding true negative views about kombucha on Twitter.

## References

IBM, 2016, “About Insights for Twitter”. Retrieved from <https://console.ng.bluemix.net/docs/services/Twitter/twitter_overview.html#about_twitter>

UCLA, “How can I subset a data set?”, retrieved from http://www.ats.ucla.edu/stat/r/faq/subset\_R.htm

IBM, 2016, “Getting started with dashDB”. Retrieved from <https://console.ng.bluemix.net/docs/services/dashDB/dashDB.html>

Zelman, K., 2016. “The Truth about Kombucha”, retrieved from http://www.webmd.com/diet/features/truth-about-kombucha

## Appendix: Supporting Information

**SQL code:**

The following code was used to create the sentiment view:

create or replace view sentiment

as (

select s.message\_ID, s.sentiment\_polarity,

m.type,

s.sentiment\_term,

t.user\_gender,

upper(t.user\_country) as user\_country\_u,

t.user\_friends\_count,

t.user\_followers\_count,

t.user\_statuses\_count,

t.message\_retweet\_count,

t.message\_favorites\_count

from Kombucha\_tweets t

join Kombucha\_sentiments s on t.message\_id = s.message\_id

left join Kombucha\_media m on t.message\_id= m.message\_id

);

SELECT sentiment\_polarity,

count(message\_id) as num\_messages,

count(distinct sentiment\_term) as Distinct\_Term\_Count,

count(sentiment\_term) as Term\_Count

FROM KOMBUCHA\_SENTIMENTS GROUP BY sentiment\_polarity;

SELECT min(sentiment\_polarity) as sentiment\_polarity\_min,

sentiment\_term,

count(message\_id) as num\_messages

FROM KOMBUCHA\_SENTIMENTS

where sentiment\_polarity = 'NEGATIVE'

GROUP BY sentiment\_term

order by num\_messages desc

;

select message\_body

from kombucha\_tweets

where message\_body like '%addiction%' or

message\_body like '%stop%' or

message\_body like '%bad%' or

message\_body like '%obsessed%'

;

-- View for influencers

-- Note: can't have order by for

create or replace view influencers

as (

select

max(message\_body) as message\_body,

user\_summary,

count(message\_id) as num\_messages,

min(user\_display\_name) as display\_name,

min(user\_followers\_count) as followers,

min(user\_statuses\_count) as user\_statuses,

min(user\_friends\_count) as user\_friends

from kombucha\_tweets

where

(

message\_language = 'en' and

(user\_followers\_count > 3800 or

user\_statuses\_count > 13000 or

user\_friends\_count > 1400)

)

group by user\_summary

);

-- Just the query for influencers

select

max(message\_body) as message\_body,

user\_summary,

count(message\_id) as num\_messages,

min(user\_display\_name) as display\_name,

min(user\_followers\_count) as followers,

min(user\_statuses\_count) as user\_statuses,

min(user\_friends\_count) as user\_friends

from kombucha\_tweets

where

(

message\_language = 'en' and

(user\_followers\_count > 3800 or

user\_statuses\_count > 13000 or

user\_friends\_count > 1400)

)

group by user\_summary

order by followers desc

;

select s.user\_country\_u,

sum(s.message\_retweet\_count) as total\_retweet\_cnt,

integer(avg(s.message\_retweet\_count)) as avg\_retweet\_cnt,

integer(avg(s.user\_friends\_count)) as avg\_friend\_cnt,

count(message\_id) as num\_messages

from sentiment s

group by s.user\_country\_u

order by total\_retweet\_cnt desc

;

**R source code:**

This appendix has listings for several code modules:

* LoadKombucha.R – loading from dashDB
* Kombucha\_plots.R – message time analysis, also plots for country and gender
* Kombucha\_influencers.R – analysis of the key influencers
* Kombucha\_anova.R – plots of mean/standard deviation, full distribution and failed attempt at ANOVA.

# LoadKombucha.R

# Initial loading of dataframes from dashDB

library(ibmdbR)

mycon <- idaConnect("BLUDB", "", "")

idaInit(mycon)

sentiment.df <- as.data.frame(ida.data.frame('"DASH5590"."SENTIMENT"')[ ,c('MESSAGE\_FAVORITES\_COUNT', 'MESSAGE\_ID', 'MESSAGE\_RETWEET\_COUNT', 'SENTIMENT\_POLARITY', 'SENTIMENT\_TERM', 'TYPE', 'USER\_COUNTRY\_U', 'USER\_FOLLOWERS\_COUNT', 'USER\_FRIENDS\_COUNT', 'USER\_GENDER', 'USER\_STATUSES\_COUNT')])

influencers.df <- as.data.frame(ida.data.frame('"DASH5590"."INFLUENCERS"')[ ,c('DISPLAY\_NAME', 'FOLLOWERS', 'MESSAGE\_BODY', 'NUM\_MESSAGES', 'USER\_FRIENDS', 'USER\_STATUSES', 'USER\_SUMMARY')])

hashtags.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_HASHTAGS"')[ ,c('HASHTAG', 'MESSAGE\_ID')])

links.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_LINKS"')[ ,c('EXPANDED\_URL', 'MESSAGE\_ID', 'URL')])

locations.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_LOCATIONS"')[ ,c('MESSAGE\_ID', 'MESSAGE\_LOCATION', 'USER\_LOCATION')])

media.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_MEDIA"')[ ,c('IMAGE\_URL', 'MEDIA\_ID', 'MESSAGE\_ID', 'SOURCE\_MESSAGE\_ID', 'TYPE', 'URL', 'VIDEO\_URL')])

kombucha.sentiments.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_SENTIMENTS"')[ ,c('MESSAGE\_ID', 'SENTIMENT\_POLARITY', 'SENTIMENT\_TERM')])

tweets.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_TWEETS"')[ ,c('MESSAGE\_ACTION', 'MESSAGE\_BODY', 'MESSAGE\_COUNTRY', 'MESSAGE\_COUNTRY\_CODE', 'MESSAGE\_FAVORITES\_COUNT', 'MESSAGE\_GENERATOR\_DISPLAY\_NAME', 'MESSAGE\_ID', 'MESSAGE\_INREPLYTO\_URL', 'MESSAGE\_LANGUAGE', 'MESSAGE\_LOCATION', 'MESSAGE\_LOCATION\_DISPLAY\_NAME', 'MESSAGE\_POSTED\_TIME', 'MESSAGE\_RETWEET\_COUNT', 'MESSAGE\_URL', 'USER\_CITY', 'USER\_COUNTRY', 'USER\_COUNTRY\_CODE', 'USER\_DISPLAY\_NAME', 'USER\_FAVORITES\_COUNT', 'USER\_FOLLOWERS\_COUNT', 'USER\_FRIENDS\_COUNT', 'USER\_GENDER', 'USER\_ID', 'USER\_IMAGE\_URL', 'USER\_LISTED\_COUNT', 'USER\_LOCATION\_DISPLAY\_NAME', 'USER\_REGISTER\_TIME', 'USER\_SCREEN\_NAME', 'USER\_STATE', 'USER\_STATUSES\_COUNT', 'USER\_SUB\_REGION', 'USER\_SUMMARY', 'USER\_URL')])

users.df <- as.data.frame(ida.data.frame('"DASH5590"."KOMBUCHA\_USERS"')[ ,c('MESSAGE\_ID', 'USER\_ID', 'USER\_NAME', 'USER\_SCREEN\_NAME')])

# Validate the data for a couple of the tables

# These should match what is seen from the summary graphs when

# data was loaded into dashDB

tweets.df$USER\_COUNTRY <- as.factor(tweets.df$USER\_COUNTRY)

summary(tweets.df$USER\_COUNTRY)

tweets.df$MESSAGE\_COUNTRY <- as.factor(tweets.df$MESSAGE\_COUNTRY)

summary(tweets.df$USER\_COUNTRY)

tweets.df$USER\_GENDER <- as.factor(tweets.df$USER\_GENDER)

summary(tweets.df$USER\_GENDER)

summary(tweets.df)

str(tweets.df)

kombucha.sentiments.df$SENTIMENT\_POLARITY <- as.factor(kombucha.sentiments.df$SENTIMENT\_POLARITY)

kombucha.sentiments.df$SENTIMENT\_TERM <- as.factor(kombucha.sentiments.df$SENTIMENT\_TERM)

summary(kombucha.sentiments.df)

str(kombucha.sentiments.df)

# Kombuch\_plots.R

# For creating plots and tables related to message time,

# and relating user’s gender and country

# Have to run LoadKombucha.R first

library(ibmdbR)

con <- idaConnect('BLUDB','','')

idaInit(con)

# Analyze country vs. gender ----------------------------------------------

table(tweets.df$USER\_GENDER)

summary(tweets.df$MESSAGE\_COUNTRY)

table(tweets.df$USER\_GENDER, tweets.df$MESSAGE\_COUNTRY)

table(sentiment.df$USER\_GENDER, sentiment.df$USER\_COUNTRY\_U)

table(sentiment.df$USER\_COUNTRY\_U)

# Filter down to countries that had more than 10 tweets

x <- table(sentiment.df$USER\_COUNTRY\_U)

x[x>10]

x2 <- table(sentiment.df$USER\_GENDER, sentiment.df$USER\_COUNTRY\_U)

# Time analysis of Tweets -------------------------------------------------

posts <-strftime(tweets.df$MESSAGE\_POSTED\_TIME, '%Y-%m-%d')

posts <-as.Date(posts)

table(format(posts, "%Y-%m"))

table(format(posts, "%w"))

pie(table(kombucha.sentiments.df$SENTIMENT\_POLARITY))

table(format(posts))

which.max(table(posts))

plot(table(posts), xlab="Date", ylab="Posts per Day")

#plot(table(format(posts, "%Y-%m")), xlab="Date", ylab="Posts per Month")

### Kombucha\_influencers.R

### Analysis on subset of tweets from most influential tweeters

# Have to run LoadKombucha.R first

# install.packages("wordcloud")

# install.packages("tm")

library(wordcloud)

library(tm)

#docs <- Corpus(VectorSource(influencers.df$USER\_SUMMARY))

docs <- Corpus(VectorSource(influencers.df$MESSAGE\_BODY))

# Remove Punctuation and Special Characters -------------------------------

docs <- tm\_map(docs, removePunctuation)

# data.frame(text=unlist(sapply(docs, '[', "content")), stringsAsFactors = F)

docs <- tm\_map(docs, content\_transformer(tolower))

# Remove numeric characters -----------------------------------------------

docs<-tm\_map(docs, removeNumbers)

# Remove Stopwords --------------------------------------------------------

docs<-tm\_map(docs, removeWords, c(stopwords("english"), "kombucha", "amp"))

inspect(docs)

# Strip WhiteSpace --------------------------------------------------------

docs<-tm\_map(docs, stripWhitespace)

inspect(docs)

# Stemming ----------------------------------------------------------------

#install.packages("SnowballC")

library(SnowballC)

docs <- tm\_map(docs, stemDocument)

docs.tdm <- TermDocumentMatrix(docs)

docs.tdm.m <- as.matrix(docs.tdm)

docs.tf <- rowSums(docs.tdm.m)

docs.tf <- sort(docs.tf, decreasing=TRUE)

# View the top 20 words

# Example code from Data Camp: print(term\_frequency[1:10])

print(docs.tf[1:40])

plot.new()

library(wordcloud)

wordcloud(names(docs.tf), docs.tf, max.words=100, rot.per=0.5, random.order = FALSE)

text(x=0.5, y=0.2, "Sentiment Terms Word Cloud")

pal <- brewer.pal(8, "Dark2")

# pal <- brewer.pal(9, "BuGn")

pal <- pal[-(1:2)]

wordcloud(names(docs.tf), docs.tf,

# scale=c(8,.3),

min.freq=2,max.words=50,

random.order=T, rot.per=.15, colors=pal,

vfont=c("sans serif","plain"))

# Kombucha\_anova.R –

# (Failed) attempt to do analysis of variation for kombucha tweets.

# No combinations or filtered sets came close to a reasonable p-value.

# Have to run LoadKombucha.R first

# install.packages("multcomp")

library(multcomp)

# install.packages("gplots")

require(gplots)

summary(tweets.df$MESSAGE\_RETWEET\_COUNT)

summary(tweets.df$USER\_FRIENDS\_COUNT)

summary(tweets.df$USER\_FOLLOWERS\_COUNT)

summary(tweets.df$USER\_LISTED\_COUNT)

summary(tweets.df$USER\_STATUSES\_COUNT)

attach(tweets.df)

plotmeans(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER)

boxplot(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER)

detach(tweets.df)

tweets.subset.df <- subset(tweets.df,

(tweets.df$USER\_FOLLOWERS\_COUNT < 300000))

summary(tweets.subset.df$USER\_FOLLOWERS\_COUNT)

attach(tweets.subset.df)

plotmeans(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER)

boxplot(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER)

boxplot(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER, log = "y")

friend.fit <- aov(USER\_FOLLOWERS\_COUNT ~ USER\_GENDER)

summary(friend.fit)

TukeyHSD(friend.fit)

par(mar=c(5,4,6,2))

tuk <- glht(friend.fit, linfct=mcp(USER\_GENDER="Tukey"))

plot(cld(tuk, level=.05),col="lightgrey")

detach(tweets.subset.df)

attach(tweets.df)

plotmeans(USER\_FRIENDS\_COUNT ~ USER\_GENDER)

boxplot(USER\_FRIENDS\_COUNT ~ USER\_GENDER)

detach(tweets.df)

tweets.subset.df <- subset(tweets.df,

(tweets.df$USER\_FRIENDS\_COUNT < 50000))

summary(tweets.subset.df$USER\_FRIENDS\_COUNT)

attach(tweets.subset.df)

plotmeans(USER\_FRIENDS\_COUNT ~ USER\_GENDER)

boxplot(USER\_FRIENDS\_COUNT ~ USER\_GENDER)

detach(tweets.subset.df)

friend.fit <- aov(USER\_FRIENDS\_COUNT ~ USER\_GENDER)

summary(friend.fit)

TukeyHSD(friend.fit)

par(mar=c(5,4,6,2))

tuk <- glht(friend.fit, linfct=mcp(USER\_GENDER="Tukey"))

plot(cld(tuk, level=.05),col="lightgrey")

detach(tweets.subset.df)