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. Their patients get excited about a new trend that has many followers, but limited 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 “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 deep 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. 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.

After loading, 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 500 starting the end of 2014. From Figure 2, many tweets did not indicate the user’s country. For those that did, the majority were from the USA, followed by Canada. 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”. Finally, Figure 4 shows that 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.

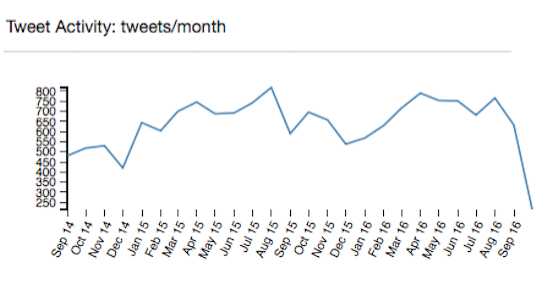


Figure . Kombucha tweets per month in Insights repository.

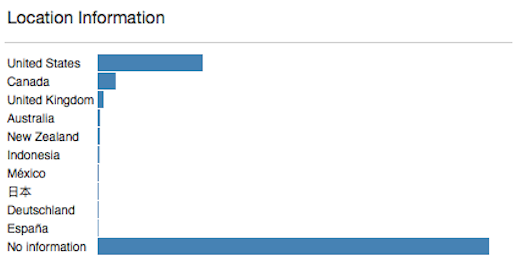


Figure . Locations for senders of kombucha tweets.

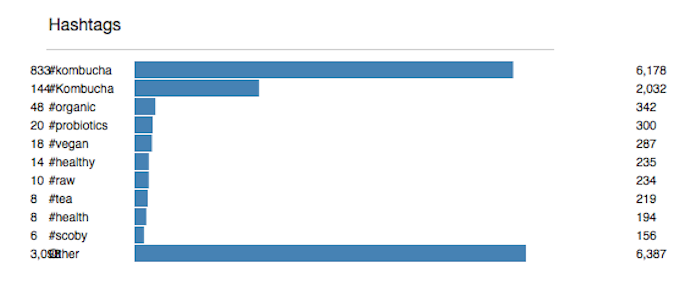


Figure . Hashtags in kombucha tweets

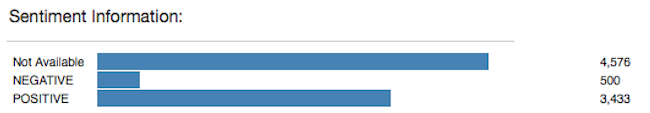


Figure . Sentiments in kombucha tweets.

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 sentiment words that Insights for Twitter detected are shown in Table 2.

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 |

## Results

appendix has a listing of words that correlate with “trailwood” and “birch”, two of the top agenda item words.

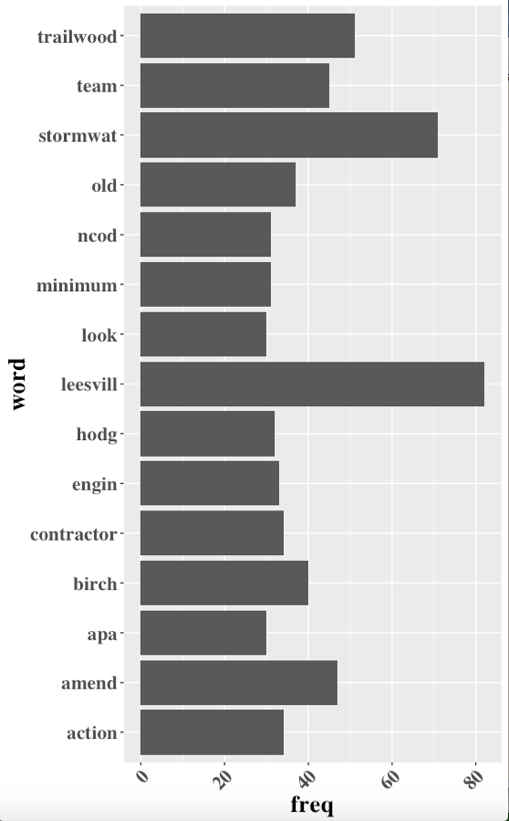
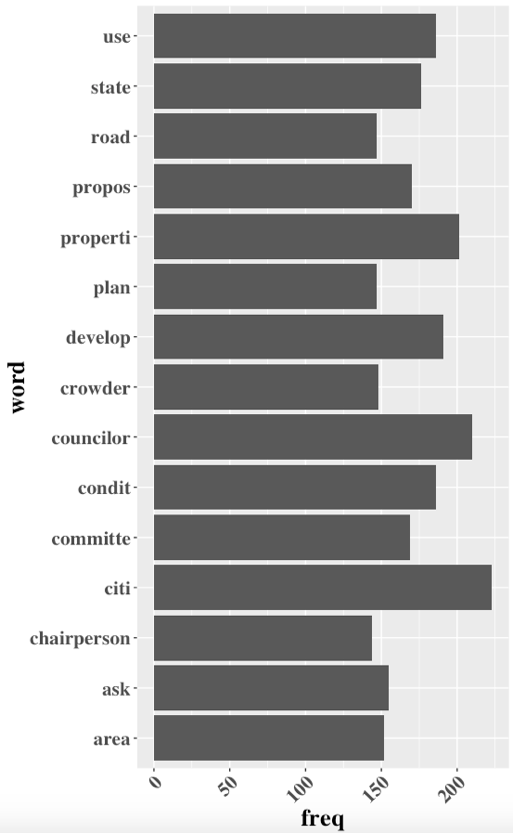


Figure . Term frequency plots: a) all terms b) after removing terms used in every document

## Limitations and Future Research

much easier to understand. E.g., “properti” is restored back to “properties”. However, changes related to version 3.3.1 of R made stem completion difficult to use.

## Conclusions

come up with topics even more targeted to mid-town communities of most interest to the readers.

## References

IBM, 2016, “About Insights for Twitter”. Retrieved from https://console.ng.bluemix.net/docs/services/Twitter/twitter\_overview.html#about\_twitter

## Appendix: Supporting Information

**Selected logs from R code:**

**R source code:**

Now try with tf-idf

kfit <- DoKMeans(m.tf.idf.transpose2,8)