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

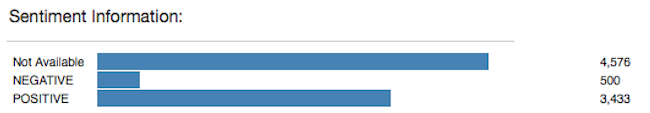
## Introduction/Purpose

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 and pull back the results into database tables. The source data is a subset of the full set someone could download directly from Twitter. The benefit of this source is that IBM enriches the data with insights gained from deep natural language processing (NLP), accomplished with IBM Social Media Analytics. The analysis attempts to determine the sentiment of the message and gender of the sender. The data returned also includes the term used to determine the sentiment.

Some initial statistics of the data are shown in



Some examples of sentiment words that Insights for Twitter detected are shown in Table 1.

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.

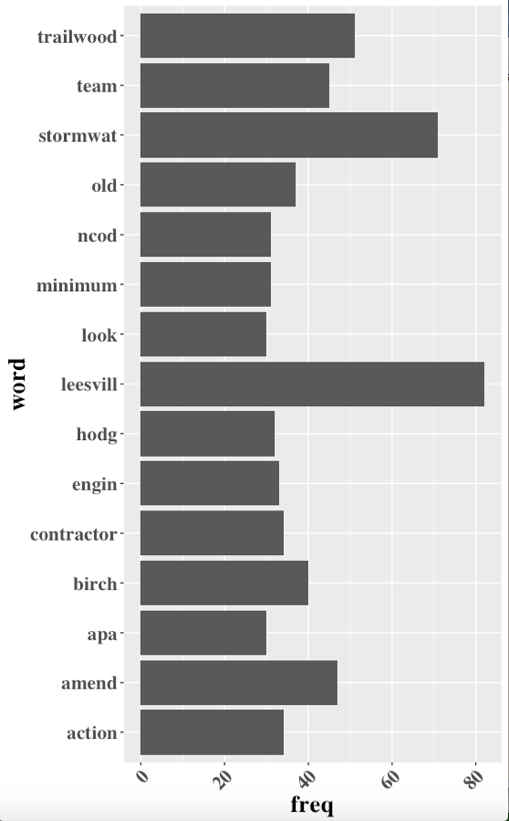
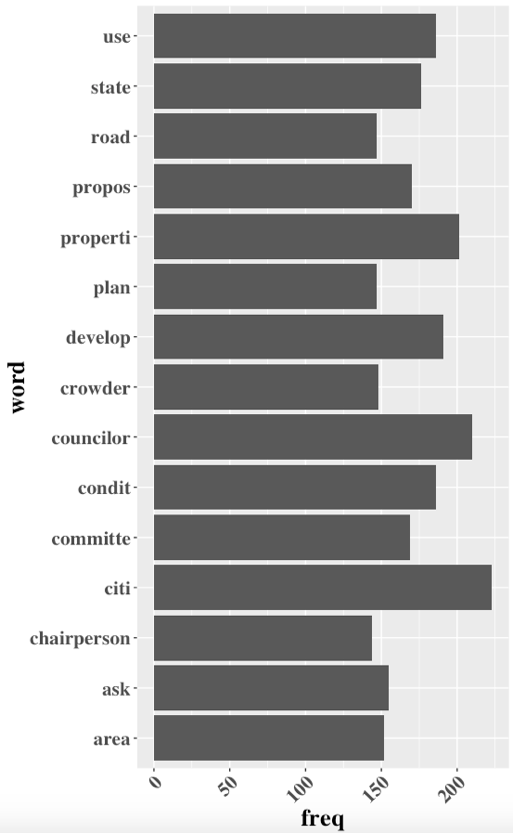


Figure . Term frequency plots: a) all terms b) after removing terms used in every document

## 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)