Volume 10, Issue 1
April 2017
ISSN: 1946-1836

Journal of Information Systems Applied Research

In this issue:

The **Journal of Information Systems Applied Research** (JISAR) is a double-blind peer-reviewed academic journal published by **ISCAP,** Information Systems and Computing Academic Professionals. Publishing frequency is currently semi-annually. The first date of publication was December 1, 2008.

JISAR is published online (http://jisar.org) in connection with CONISAR, the Conference on Information Systems Applied Research, which is also double-blind peer reviewed. Our sister publication, the Proceedings of CONISAR, features all papers, panels, workshops, and presentations from the conference. (http://conisar.org)

The journal acceptance review process involves a minimum of three double-blind peer reviews, where both the reviewer is not aware of the identities of the authors and the authors are not aware of the identities of the reviewers. The initial reviews happen before the conference. At that point papers are divided into award papers (top 15%), other journal papers (top 30%), unsettled papers, and non-journal papers. The unsettled papers are subjected to a second round of blind peer review to establish whether they will be accepted to the journal or not. Those papers that are deemed of sufficient quality are accepted for publication in the JISAR journal. Currently the target acceptance rate for the journal is about 40%.

Questions should be addressed to the editor at editor@jisar.org or the publisher at publisher@jisar.org. Special thanks to members of AITP-EDSIG who perform the editorial and review processes for JISAR.

2017 AITP Education Special Interest Group (EDSIG) Board of Directors

JOURNAL OF
INFORMATION SYSTEMS APPLIED RESEARCH

An Interactive Toolbox For Twitter Content Analytics

Musa Jafar
musa.jafar@manhattan.edu

Marc Waldman
marc.waldman@manhattan.edu

Manhattan College
Riverside, NY

Abstract

In this paper we present a simple and easy to use toolbox that can be used for social media content analytics in the world of Twitter. The toolbox was developed primarily for researchers with minimal computing background who wish to visually analyze the content of tweets (text and the associated metadata such as screen-names, hashtags, mentions, etc.) across the twitter-defined timeline or a user-specified timeline. The toolbox is open source and built on top of the R programming platform, R-Shiny and R-word cloud. The toolbox uses a word cloud approach to visualize both the metadata and the N-gram text sequences that make up the tweets collection (the tweets corpus). Filter mechanisms of the toolbox allow the researcher to control for the type and amount of data displayed in the associated word clouds – allowing for a finer resolution of analysis.

Keywords: Text-Analytics, Visual-Analytics, twitter, R-shiny, word-cloud, N-grams.

1. INTRODUCTION

Social media sites are a rich source of data for researchers and practitioners to analyze. This is especially true of Twitter as it provides a "real-time window into the opinions, hopes, beliefs, complaints and dreams of its users, and the insights that it aggregates can provide marketers, product developers, sales, digital journalists, sociologists, educators -- really, the entire enterprise -- with deep, rich and spontaneous feedback on virtually any topic" (Li et al. 2013).

Each tweet, although only 140 characters in length, has an associated collection of interesting metadata that includes the author's username (screen-name), timestamp, geo-location (if enabled), hashtags, retweet-count, favorite-count, etc. A researcher's dataset will typically consist of a large number of related tweets – typically related by hashtags, keywords or authors. We will refer to this dataset of tweets as the tweets-corpus.

Typically acquiring the tweets to form the tweets-corpus is far from trivial. Twitter does not provide a non-programmatic mechanism to easily download the public tweets of any user (except your own tweets). A number of companies (Gnip, exporttweet.com, Twitonomy.com) do provide downloads of public tweets for a fee. However, a number of software tools are available that allow one to freely acquire tweets (within Twitter-specified limits). Tweets can also be captured by "screen-scraping" software or by utilizing the Twitter API – which provides access to tweets via a programming language such as Python or R. The Twitter Capture and Analysis Toolset (Borra et al. 2014) provides a freely available software distribution that can be used to capture tweets – however it requires users to have some familiarity with system administration concepts as the package is meant to be deployed in a Linux environment. It also provides summaries in the form of pie charts and line graphs.

Once acquired, a tweets-corpus presents an analysis challenge to those individuals who are not comfortable with programming. Many freely-available software libraries are available to assist in the analysis of tweets (and other social media data and free-form text). However, many require the user to have extensive previous programming skills. We created our tweets-corpus analysis toolbox to help address this issue – allowing individuals without the programming background to visually analyze the corpus and associated metadata.

2. RELATED WORK

Analysis of twitter data is certainly not an unexplored topic. Many powerful tools have been developed to analyze social media data (including Twitter). Several of these tools require no programming background and provide rich insight into the data. However, this analysis is almost exclusively focused on the social media metadata – how often was your tweet retweeted, how many people are following you, how many people mentioned a specific hashtag, etc. To date, there has been little focus on easy-to-use tools for the analysis of tweets text - more specifically tools that allow a researcher to visualize the central themes of a set of tweets and how these themes evolved over time.

The paper by Zimmer and Proferes (Zimmer et al. 2014) provides an overview of how Twitter data is being used by researchers. It found that content analysis was the dominant form of analysis performed on tweets. Content analysis, as defined in their paper, is one "where text within a tweet was used in part of the analysis in some way". We envision our toolbox will be used mainly for content analysis but it also supports analysis of some of the metadata associated with the tweets corpus. Word clouds (sometimes also referred to as tag clouds) are the central visualization elements of our analysis tool.

Word Clouds

We use a word cloud approach (Viégas et al. 2007, 2008, 2009) to visualize the N-grams and the metadata of a tweets-corpus. An N-gram is defined as a continuous sequence of N-words in some block of text. For example, in the phrase "Mary had a little lamb" – there are five 1-grams (or sequences of one word) in the text. The word "Mary" is the first 1-gram and "lamb" is the fifth one-gram. During N-gram analysis "insignificant" words such as "the", "a" and "is" are frequently dropped. "Mary had" and "little lamb" are among the more interesting 2-grams in the phrase.

There are numerous freely-available tools to generate word clouds. They vary from simple web-based tools such as wordle.net to sophisticated code libraries (e.g. the R-based word cloud package) that allow users to calibrate almost all aspects of cloud creation. We use the R-based word cloud in combination with RWeka and tm packages to facilitate the analysis beyond the 1-gram (Feinerer et al. 2008, 2015; Hornik, et al. 2009; Fellows 2014).

Word clouds, by their very nature, provide "big-picture" insight into a corpus of text. Words occurring with greater frequency are placed in a larger or more dominant font or color. Examples of word clouds used in our toolbox can be found in subsequent sections below. Almost all the word cloud tools we have examined operate solely at the 1-gram level – graphically illustrating the frequency of each word in a corpus. However, our toolbox supports word cloud visualization of larger N-grams. The toolbox provides user interface controls that allow the user to specify N-gram size (up-to 4-grams) with timeline and screen-name filter options. More information is provided in the implementation sections below.

Natural Language Processing

Although our toolbox is meant to assist the researcher to perform intelligent analysis on the tweets text we, as of yet, do not perform any form of sophisticated natural language processing on the text. This type of processing might take the form of sentiment analysis or other form of natural language understanding wherein the toolbox attempts to classify the text in some manner – for example, classify the kind of speech act intended by the creator of the tweet (Searle 1969) or labeling a tweet as either positive or negative toward a particular company, product, situation or an event.

Traditional theory, algorithms and tools that have been developed to analyze text from a corpus linguistics perspective (McEnry et al. 2012; Bird et al. 2009; Feinerer et al. 2008, 2015; Miller 1995) do not necessarily work well for social media based text. Social media text is typically grammatically incorrect, uses words that are not in the dictionary and embeds symbols, urls, hashtags, mentions and emoji(s). This makes it impractical to apply standard corpus linguistics algorithms and tools to analyze the content. (Maynard 2012) outlined many of the challenges that face social media text analytics.

Social Media Text Analytics

Social data is about the speech act itself, its background and its illocutionary effect. Until recently, except for the few that were documented and historically speaking, almost all speech acts went unnoticed. Those that have been archived are in forms and formats that are hard to access and subsequently hard to analyze. The internet and the underlying social media technologies however, provides us with platforms to express thoughts, say what is in our mind, make a comment, state a belief or express an opinion, an emotion or a desire. It also allowed for those thoughts to be captured and stored in digital formats that are retrievable, searchable, indexable, presentable and analyzable. At any given moment in time, those platforms allowed for the spawning of many social media networks and the forking of multiple communities.

Social media text bundles have a social active characteristic, are spur of the moment, do not easily conform to the traditional natural language processing rules, syntax and grammar. It uses language in a way not governed by traditional rules, it is free flowing, ambiguous and less rules bound. Social text is about the intention of the speaker, has a performative function and it is communication centric. In the case of Twitter, a tweet is created on the fly, it has a time component, a social aspect component, an intertextuality component and a para-textual component. When bundled together, a tweet collection becomes a corpus where each tweet is a rich document surrounded by a bundle of metadata (timestamp, hashtags, mentions, originality, status, geo-location, etc. within the bundle), the corpus is multi-dimensional and various aspects of it are analyzable (Ferragina et al. 2015; Metaxas et al. 2015).

3. THE CASE FOR VISUAL TEXT ANALYTICS

Well-designed visualizations are intuitive, insightful, hypothesis generating, help dispel myths, enable discoveries, emphasize a point of view or help discover patterns in almost every aspect of knowledge of our world. Terms like the "thinking eye" and the "seeing brain" date back to the Swiss-German artist Paul Klee and have been extensively used in the data visualization literature. The daunting challenge in social media text analytics is to make the content of a tweets-corpus visually available in a useful and presentable way for a researcher who is not a programmer, however expert in the domain content of the tweets-corpus. As John Tukey (Tukey 1977), the great statistician stated: "The greatest value of a picture is when it forces us to notice what we never expected to see".

Our toolbox allows an individual who is interested in performing ad-hoc analytics on a tweets-corpus to interactively and visually analyze it and its metadata, across multiple dimensions. We follow standard visualization principles. The user (browser-based) interface is built around the R-word cloud, R-Shiny package and its widgets (Chang et al. 2016), the computing engine (server back-end) is built around R, R text-analytics packages tm (Feinerer et al. 2008) and RWeka (Hornik et al. 2009).

As previously stated, we use a word cloud approach to visualize the N-grams and the metadata of a tweets-corpus. We apply timeline and metadata filters to allow the individual expert to get a better understanding of the context within which the content (tweets) was initially published. In our current implementation the word cloud itself is not yet interactive – you cannot directly manipulate and interact with the generated word cloud (drill down on an N-gram to view the corresponding tweets). This is a limitation of the R-word cloud package and the R-graphing system – which we may not use in future versions of the toolbox. However, we allow for the interactive control of the frequency ranges, N-gram count and font range. The toolbox allows the user to interactively and easily slice the content across a timeline and filter by the tweet author (screen-name).

Most forms of data analysis typically follow a three-phase process. First is the data collection phase wherein the necessary data is collected, possibly from multiple sources which can be a combination of API(s), screen-scrapping, purchase, etc. The second phase, which we will refer to as the data repurposing and cleaning phase, typically involves error checking and transforming the data into a format that can be cleanly loaded into the software being used for analysis. For example, the creation date and time of a tweet is usually in GMT(Greenwich Mean Time), converting it to the current locale may be necessary for meaningful analysis; hashtags and mentions are part of the body of the tweets, they may need to be extracted out for further analysis. The final phase is where the analysis actually takes place – data is typically filtered, summarized and visualized. Our toolbox is meant to assist in this third phase - to make the data analytics phase (of a tweets-corpus) generic, repeatable, interactive, intuitive and easy to use. The data collection and data repurposing phases are beyond the scope of this paper. The current implementation assumes that the tweets-corpus has been created using the twitteR package (Gentry 2015). However, if the data has been acquired using other methods (Python-twitter API, purchased through a twitter subsidiary or screen-scraped from the twitter homepage) we do assume that the data has been converted into an R-dataframe format form that is compatible with twitteR reference 'status' class list (Gentry 2015). This format is shown in Figure 1 - the underlined attributes are the ones that are currently required by the toolbox, the hashtags and the mentions are extracted from the body of the tweet's text.

4. THE INTERACTIVE TOOLBOX

The retrieval, in-depth manipulation, text analytics, and presentation of the results require dexterity in programming which is lacking in most of us. Our main goal is to abstract the analytics phase and make it generic, repeatable, interactive, intuitive and easy to use. This allows a researcher from a non-computing discipline to achieve their goal of gaining an understanding, insight and knowledge of the content through iterative interactive analytics without needing to learn how to write or repurpose code. Allowing the none-algorithmic, non-programmer to visually analyze tweets-corpus across multiple dimensions.

Table 1 briefly summarizes some of the datasets (tweets-corpora) that were used during the development and testing (piloting) of the toolbox. In the following sections we present the various components of the toolbox and their functionality. All screenshots used in this paper were taken while utilizing the datasets in the previously mentioned table.

The User Interface

Upon startup, the toolbox presents the user with three tabs ("Load & View a Tweets Dataset", "N-Gram Analysis", and "Metadata Analysis") located near the top of the screen (Figures 2, 3, 4 and 5). The "Load & View a Tweets dataset" tab allows a user to import a tweets-corpus. Once loaded, the tweets-corpus is displayed in tabular form and can be filtered (to allow selective viewing of tweets) based on word or regular-expression pattern matching. Figure 2 shows the result of the regular expression filter "^I.* pray" being applied to Pope Francis tweets-corpus. This regular expression is filtering for all tweets (in the corpus) that include the capital letter "I" at the beginning of a tweet followed by the word "pray" somewhere in the text. Note that this regular expression allows an arbitrary number of words to appear between letter "I" and the word "pray". Filtering can also be done via the tweets' author (screen-name) and the date of the tweets. This can be seen in Figure 3 which shows the "Dump Stoli, Dump Russian Vodka" tweets-corpus being filtered on the screen-name "fakedansavage".

The tool's "N-Gram Analysis" tab provides access to interface elements that allow the user to visualize the text of the tweets-corpus using word clouds. Figure 4 shows a 3-gram-based word cloud of the "Dump Stoli, Dump Russian Vodka" tweets-corpus. Note that the user interface elements below the word cloud allow the user to control the N-gram level (1, 2, 3 or 4-gram), as well as limits on (1) "Token Frequency Range" to allow for the filtering-out of very high and very low frequencies, (2) the Maximum Number of N-gram Tokens displayed to control for which tweets data will be included in the word cloud, and (3) "Font Size Scale" to penalize very high and reward very low frequency tokens in the display. Filtering can be done on screen-name and/or tweets corpus slice of the date range. Additional elements (font size and token frequency) govern how the word cloud is rendered – the font size of the cloud elements and the number of elements appearing in the word cloud, the most frequent tokens are displayed with the highest "Font Size" on the scale.

Varying the N-gram level can provide significant insight into the tweets-corpus. Two, three and four-gram analysis provides a more contextual usage of the words in the tweets and therefore more insight into a thread of interrelated tweets. We have found that 2, 3 and 4-gram analysis is particularly useful in the analysis of the tweets of competing groups – two or more groups with opposing messages. The toolbox facilitates this visual comparison of multiple word clouds via the "WordCloud in a new window" selection box that is found directly above the timeline filter in the user interface (see Figures 4 & 5). A use of this feature is discussed in greater detail in the application section below.

The tool's "Metadata Analysis" tab provides access to the interface elements that allow the user to visualize the metadata (screen-names, mentions, hashtags and applications used) associated with the tweets-corpus. Figure 5 shows a metadata analysis of the Narendra Modi tweets-corpus. In this figure we see a word cloud corresponding to the individual's (screen-names) that are referenced (mentioned) in the tweets. The prime minister mentioned himself (@narendramodi) most frequently (206 times). The screen-name @pmoinidia is mentioned 82 times while @un is mentioned only 37 times. The screen-name metadata analysis gives the user of the toolbox insight into the inner circles and the interest of the prime minister. The toolbox also allows for the analysis and the comparison of the hashtags, and underlying application used to emit the tweet (iPhone, Galaxy, Blackberry, the kind of twitter-app, etc.).

5. THE APPLICATIONS OF THE TOOLBOX

While developing and testing the toolbox we collaborated with a number of our colleagues (across multiple disciplines) to utilize the toolbox in academic research and in the classroom. In this section we briefly describe some of this work in the hope that it will inspire others to use the toolbox for similar or new areas of research and pedagogy.

Colleagues in the Management Department utilized the toolbox in the development of a crisis management study on how executives at Stoli Group USA handled a social media crisis that began in the summer of 2013 when "Dan Savage", an LGBT activist and sex advice columnist (https://twitter.com/fakedansavage), called for a boycott of Stolichnaya vodka because of its perceived Russian origins. The Russian Government was being sharply criticized for the passing of discriminatory anti-gay laws. Savage's Twitter messages were intended to show solidarity with, and to draw international attention to the plight of, the gay, lesbian, bi-sexual and transgender (LGBT) community in Russia. The tweets that comprised the case study were purchased directly from Twitter's subsidiary, Gnip (https://www.gnip.com/). The dataset provided access to all relevant historical tweets (those containing hashtags such as #dumpstoli and #dumprussianvodka) over a 40 day period between July 23 - September 1, 2013. Utilizing the toolbox our colleagues found several trends in the tweets - who the influential tweeters were; what was the sentiment of the related tweets (those utilizing the relevant hashtags) and the main geographic locations that the tweets were emanating from. Figure 6 shows a 3-gram word cloud comparison of the Stoli company tweets and Dan Savage tweets during the same study-related time period. Note the sharp difference in the message and tone in the 3-grams of the two competing voices. An analyst can also compare the content in Figure 6 to the overall content in Figure 4 for the same time period. It is clear that the Dan Savage's message is winning over the Stoli-team message.

During the Pope's recent (2015) visit to the United States a colleague in the Religion department utilized the 3-gram word clouds (of the Pope's tweets) generated by our toolbox to facilitate classroom discussions. After showing the word clouds (containing 3-gram phrases) to the students, the instructor found that "An individual phrase or, more often than not, a number of closely connected phrases led them [the students] to begin a conversation about a topic that was forming in their minds but they had not yet articulated up to that point."

Another colleague in the Religion department is utilizing the toolbox to visualize the tweets of India's Prime Minister Narendra Modi. This colleague is planning on having his students utilize the toolbox in a class on topics in contemporary religion and science.

An ad hoc Session with the ToolBox

While analyzing the Stoli-tweets, a colleague asked the question: What was Mr. Savage tweeting about the day before he published the "dumpstoli" "dumpRussianVodka" hashtags and Avatar? The 2013-07-23 tweets were not part of the Gnip dataset that we acquired, also when we acquired the dataset, our queries filtered on hashtags and not screen-name(s). Using the advanced search feature of twitter.com we captured these tweets. We then parsed the html and scrapped the tweet-id(s). Knowing the ids of the tweets and using the twitteR API we acquired the data and loaded it into the toolbox. The process provided us with an insight into Mr. Savage's interests at large and what triggered his interest in creating the "dumpStoli" and "dumpRussinaVodka" hashtags and avatar (Figure 3). Visualizing the 2013-07-23 tweets in the toolbox (Figure 7), allowed us to find answers to questions we did not previously propose to investigate. We were also able to compare the content with his tweets as they related to Stoli-case (Figure 4). We were able to track the announcement of the "dumpRussianVodka" and "dumpstoli" avatar, we were also able to find out where it all started, as it was part of Mr. Savage's timeline (Figure 3). On "2013-07-24 01-39-50" when he retweeted "it's time to call a BOYCOTT OF THE 2014 OLYMPICS IN RUSSIA…" from @BrianKentMusic, 12 hours later (2013-07-24 13:51:54) he announced his avatar. These are the types of capabilities we want to give to the investigator of the content to facilitate their analysis. To allow them to perform "Intelligence Analysis" through the "Search and Filter", "Schematize", "Build Case", "Tell Story", "Make a Presentation" (Pirolli et al. 2005).

Another ad hoc Session with the Tool

Another question that was raised: Has Pope Francis' message changed between Christmases? With our toolbox and using filters (Dec. 20-31st) of 2013, 2014, and 2015 (Figure 8); the data provided an abundance of insight into the overarching themes. In 2015 messages about prosecution and suffering predominated as compared to the 2013 and 2014 messages.

The Implementation of the Toolbox

In designing the toolbox we paid a close attention to the "Sense making Model for Intelligence Analysis" (Pirolli et al. 1999, 2005; Card et al. 1999). A Model-View-Controller (MVC) pattern approach was used to build the toolbox. The user interface (the View), is built on top of R-Shiny, R-data table and R-word cloud packages. The R-word cloud package is used to render the N-grams of the text and the metadata. The back-end server (the Controller), is built on top of basic R, RWeka (Hornik et al. 2009), R-tm text mining package (Feinerer et al. 2013) and R-helper packages. The R-Shiny server (Chang et al. 2016) manages the reactivity of the user interface and encapsulates the set of utility functions needed to interface with the data (Model) to perform the text-analytics functions on the tweets-corpus. The Model encapsulates the data and the set of functions used to retrieve, query, clean and manipulate a tweets corpus.

6. FUTURE WORK

Work continues on improving the usability and the feature set of our text analytics toolbox. Our future plans include accepting a tweets-corpus in both CSV and JSON formats, enabling live querying of twitter so that a twitterer and their circle of friends information can be displayed for a given screen-name, incorporating a higher resolution view of the timeline to the hour and possibly minute or even the second level is a high priority, including user interface elements to allow the exclusion of selected screen-names from a tweets-corpus analysis and geo-location analysis where geo-location is provided. The R generated word cloud is currently not interactive. As part of the long term enhancement of the toolbox, we want to implement an interactive Web-GL, SVG-based word cloud wherein a researcher can drill down using the tokens displayed in the word cloud itself to further investigate the underlying content and corresponding tweets.

7. CONCLUSIONS

Twitter data is playing an increasing role in research conducted across a variety of fields. In 2010, the Library of Congress and Twitter signed an agreement for the Library of Congress to retrieve and store all of the public tweets and on an ongoing basis. Eventually the tweets will become available "in a comprehensive useful way" (Library of Congress 2013). However, it is not clear as to when and how this data trove will become available. This also emphasizes the need for new approaches and different analytics tools to be developed.

We have implemented an extensible toolbox for the visual intelligent analysis of a tweets-corpus. This toolbox makes the analysis process repeatable, and the results are replicable. It was primarily designed for individuals who typically possess no programming background such as social scientists and digital journalists. The tool's contributions include visual N-gram (1, 2, 3 and 4 -gram) analysis of a tweets-corpus in combination with filter mechanisms that can be used to refine the granular level of analysis. The toolbox is open source and freely available.

Finally, one common misunderstanding of the framework is that it just produces pretty word clouds. While the framework does indeed display aesthetically appealing word clouds, its power however is in the analytics back-end that enables the visualization of content. Social events fade away as fast as they come to life. This framework allows for the instant analytics of social media content within the context of Twitter. An article, an event, a comment that spurred a storm of Twitter-based reaction can be instantly analyzed and its background investigated by a digital journalist, a commentator, or an opinion article writer. Our toolbox facilitates the contextualization of the parts.

8. REFERENCES

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. *O'REILLY Media*.

Borra, E., & Rieder, B. (2014). Programmed method developing a toolset for capturing and analyzing tweets. *International Journal of Information Management* 66(3), 262-278.

Card, S., Mackinlay, J., & Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Think. (1st Ed.). *Morgan Kaufmann*.

Chang, W., Cheng, J., et al. (2016, January 12). Shiny: Web Application Framework for R. *R package version 0.12.2. CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/shiny/index.html

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software* 25(5), 1-54.

Feinerer, I., & Hornik, K. (2015, July 03). tm: Text Mining Package. *R package version 0.6*. *CRAN R-Project:* Retrieved January 2016 from https://cran.r-project.org/web/packages/tm/index.html

Fellows, I. (2014, June 13). wordcloud: Word Clouds. *R package version 2.5*. *CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/wordcloud/index.html

Ferragina, P., & Santoro, F. (2015). On Analyzing Hashtags in Twitter. *9th International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, Oxford, UK, 110-119.

Gentry, J. (2015, July 29). twitteR: R Based Twitter Client. *R package version 1.1.9*. *CRAN R-Project:* Retrieved January 14, 2016 from https://cran.r-project.org/web/packages/twitteR/index.html

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics* 24(2), 225-232.

Li, C., Etlinger, S., Live, R., Jones, A., et al. (2013). Twitter's IPO: An Analysis of Opportunities and Threats. Retrieved January 23, 2016 from http://www.altimetergroup.com/2013/10/twitters-ipo-an-analysis-of-opportunity-and-threats/

Library of Congress. (2013), January 31. "Update on the Twitter Archive at the Library of Congress," Retrieved January14, 2016 from https://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Maynard, D., Bontcheva, K. & Rout, D. (2012). Challenges in developing opinion mining tools for socialmedia. Proceedings of @ NLP can u tag# usergenerat-edcontent ?! Workshop at LREC 2012, Turkey.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics. *Cambridge, UK: Cambridge University Press*.

Metaxas, P. T., Mustafaraj, E., Wong, K., et al. (2015). What do Retweets indicate? Results from User Survey and Meta-Review of Research. *International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, Oxford, UK, 658-661.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* 38(11), 39-41.

Pirolli, P., & Card, S. (1999). Information Foraging. *Psychological Review* (106), 643-657.

Pirolli, P., & Card, S. (2005). The Sensemaking Process and Leverage Points for Analyst Technology. *International Conference on Intelligence Analysis* MITR, McLean, VA, 1-6.

Searle, J. R. (1969). Speech Acts. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press.

Tukey, J. (1977). Exploratory Data Analysis. Readings: *Adison Wesley*.

Viégas, F. B., Wattenberg, M. H., et al. (2007). Many Eyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics* 13(6), 1121-1128.

Viégas, F. B., & Martin Wattenberg, M. (2008). Tag Clouds and the Case for Vernacular Visualization. *ACM Interactions* 15(4), 49-52.

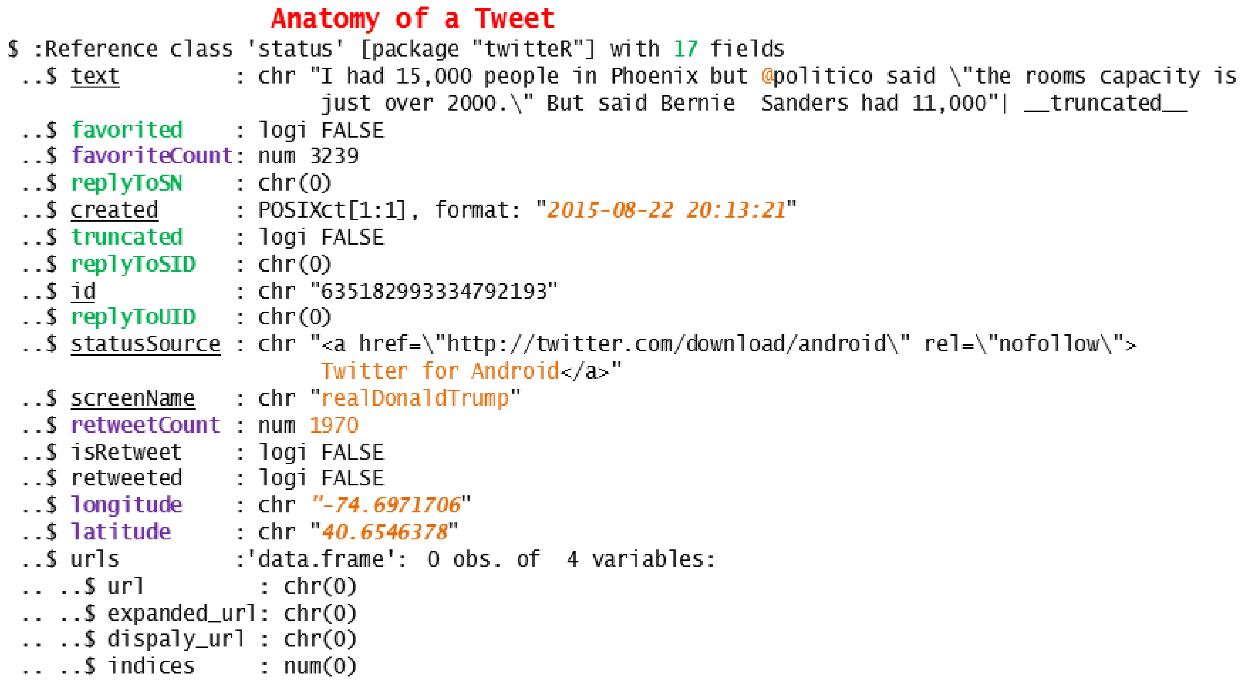
Viégas, F. B., Wattenberg, M., & Feinberg, J. (2009). Participatory Visualization with Wordle. *IEEE Transactions on Visualization and Computer Graphics* 15(6), 1137-1144.

Zimmer, M., & Proferes, J. N. (2014). A topology of Twitter research: disciplines, methods, and ethics. Journal of Information Management 66(3), 250-261.

Editor's Note:

This paper was selected for inclusion in the journal as a CONISAR 2016 Distinguished Paper. The acceptance rate is typically 7% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2016.

Appendices and Annexures

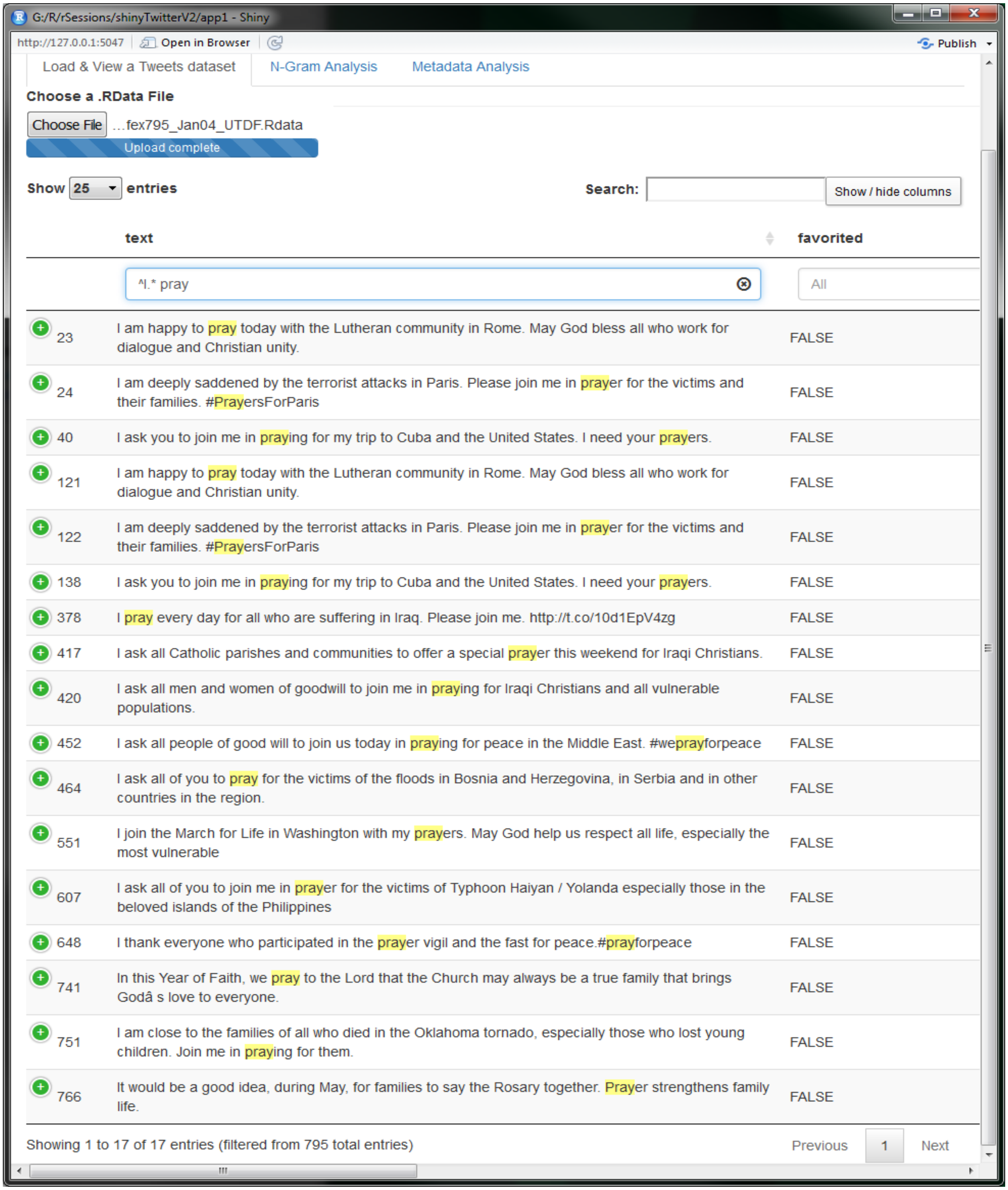


```
                    Anatomy of a Tweet
$ :Reference class 'status' [package "twitteR"] with 17 fields
 ..$ text          : chr "I had 15,000 people in Phoenix but @politico said \"the rooms capacity is
                           just over 2000.\" But said Bernie  Sanders had 11,000"| __truncated__
 ..$ favorited     : logi FALSE
 ..$ favoriteCount: num 3239
 ..$ replyToSN     : chr(0)
 ..$ created       : POSIXct[1:1], format: "2015-08-22 20:13:21"
 ..$ truncated     : logi FALSE
 ..$ replyToSID    : chr(0)
 ..$ id            : chr "635182993334792193"
 ..$ replyToUID    : chr(0)
 ..$ statusSource  : chr "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">
                           Twitter for Android</a>"
 ..$ screenName    : chr "realDonaldTrump"
 ..$ retweetCount  : num 1970
 ..$ isRetweet     : logi FALSE
 ..$ retweeted     : logi FALSE
 ..$ longitude     : chr "-74.6971706"
 ..$ latitude      : chr "40.6546378"
 ..$ urls          :'data.frame':  0 obs. of  4 variables:
 .. ..$ url         : chr(0)
 .. ..$ expanded_url: chr(0)
 .. ..$ dispaly_url : chr(0)
 .. ..$ indices     : num(0)
```

Figure 1: The structure of a twitteR tweet

Tweet Subject	Tweet Count	Timeline	Acquisition Method	Comments
Pope Francis	795	2013-03-17→ 2015-12-31	twitteR API	Pope Francis tweets
Narendra Modi	4,034	2014-12-10→ 2015-12-31	twitteR API	Prime Minister Modi tweets
John Stewart	3,010	2015-08-08→ 2015-08-09	twitteR API	Last Night of the Daily Show
Elton John	4,099	2015-03-16	twitteR API	Dolce & Gabbana Feud
Keith Olberman	1,238	2015-02-17→ 2015-02-24	twitteR API	Feud with Penn State & suspension by ESPN
Saudi Leaks	5,671	2015-6-20	twitteR API	Comments on Saudi leaks
Dump Stoli, Dump Russian Vodka	53,954	2013-07-23→ 2013-09-01	Purchased+ Python +R+ ScreenScrapping	Tweets were Purchased from GNIP for a case study
Sarah Palin	16,347	2009-11-19→ 2015-09-11	twitteR API	Mrs. Palin timeline & Palin related tweets.
Donald Trump	30,264	2015-06-28→ 2015-09-10	twitteR API	Mr. Trump's timeline & Trump related tweets
1600 Pennsylvania Ave Washington DC	16,061	2016-01-01→ 2016-01-13	Google Maps +Twitter API(s)	We used Google Maps API & Twitter geo-location tags

Table 1: Tweets-corpora used in testing the toolbox



Choose a .RData File

...fex795_Jan04_UTDF.Rdata

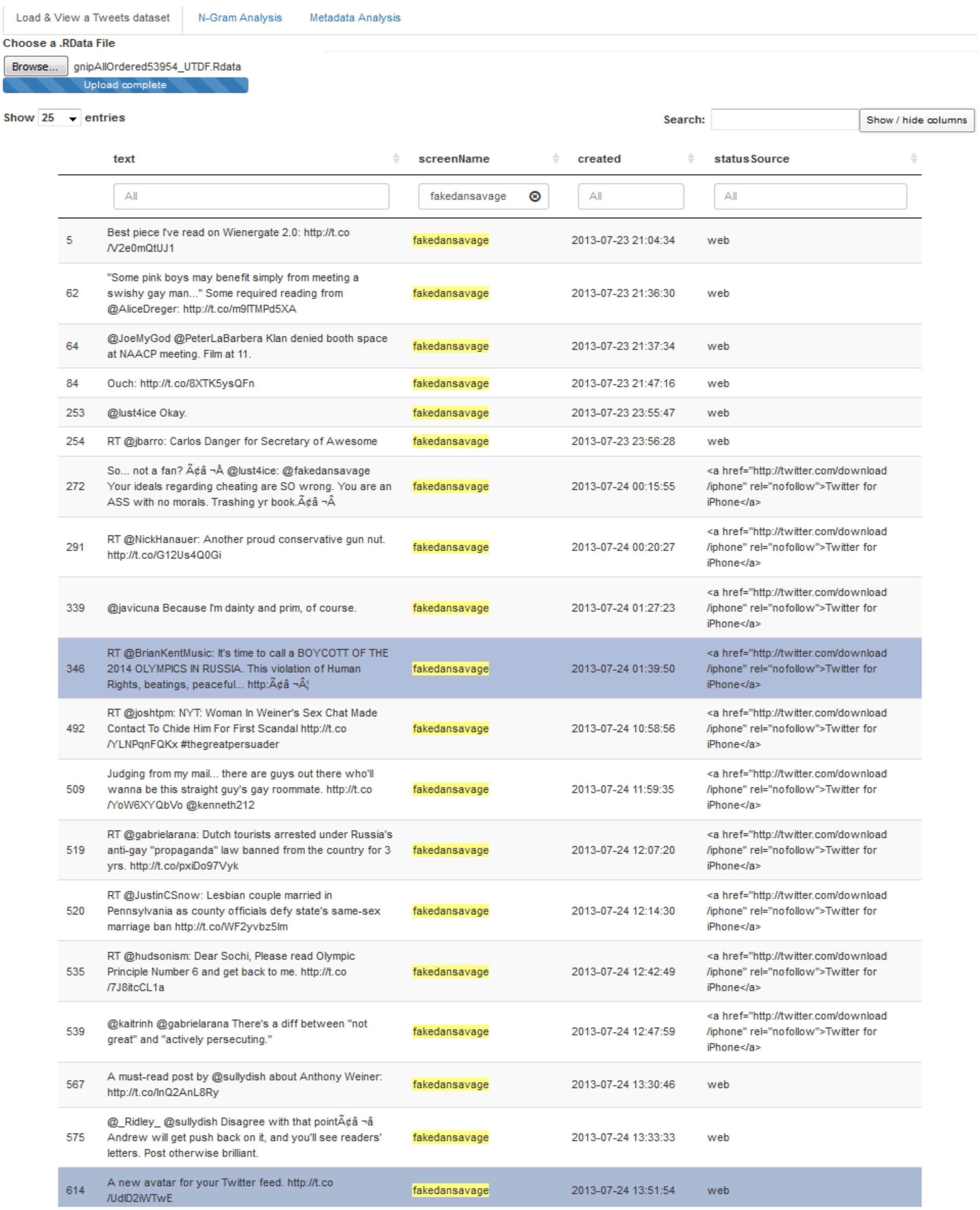
Show 25 entries

Search:

	text	favorited
	^I.* pray	All
23	I am happy to pray today with the Lutheran community in Rome. May God bless all who work for dialogue and Christian unity.	FALSE
24	I am deeply saddened by the terrorist attacks in Paris. Please join me in prayer for the victims and their families. #PrayersForParis	FALSE
40	I ask you to join me in praying for my trip to Cuba and the United States. I need your prayers.	FALSE
121	I am happy to pray today with the Lutheran community in Rome. May God bless all who work for dialogue and Christian unity.	FALSE
122	I am deeply saddened by the terrorist attacks in Paris. Please join me in prayer for the victims and their families. #PrayersForParis	FALSE
138	I ask you to join me in praying for my trip to Cuba and the United States. I need your prayers.	FALSE
378	I pray every day for all who are suffering in Iraq. Please join me. http://t.co/10d1EpV4zg	FALSE
417	I ask all Catholic parishes and communities to offer a special prayer this weekend for Iraqi Christians.	FALSE
420	I ask all men and women of goodwill to join me in praying for Iraqi Christians and all vulnerable populations.	FALSE
452	I ask all people of good will to join us today in praying for peace in the Middle East. #weprayforpeace	FALSE
464	I ask all of you to pray for the victims of the floods in Bosnia and Herzegovina, in Serbia and in other countries in the region.	FALSE
551	I join the March for Life in Washington with my prayers. May God help us respect all life, especially the most vulnerable	FALSE
607	I ask all of you to join me in prayer for the victims of Typhoon Haiyan / Yolanda especially those in the beloved islands of the Philippines	FALSE
648	I thank everyone who participated in the prayer vigil and the fast for peace.#prayforpeace	FALSE
741	In this Year of Faith, we pray to the Lord that the Church may always be a true family that brings Godâ s love to everyone.	FALSE
751	I am close to the families of all who died in the Oklahoma tornado, especially those who lost young children. Join me in praying for them.	FALSE
766	It would be a good idea, during May, for families to say the Rosary together. Prayer strengthens family life.	FALSE

Showing 1 to 17 of 17 entries (filtered from 795 total entries)

Figure 2: Load & View a Tweets dataset tab: Pope Francis tweets, with a regular expression filter



Choose a .RData File

gnipAllOrdered53954_UTDF.Rdata

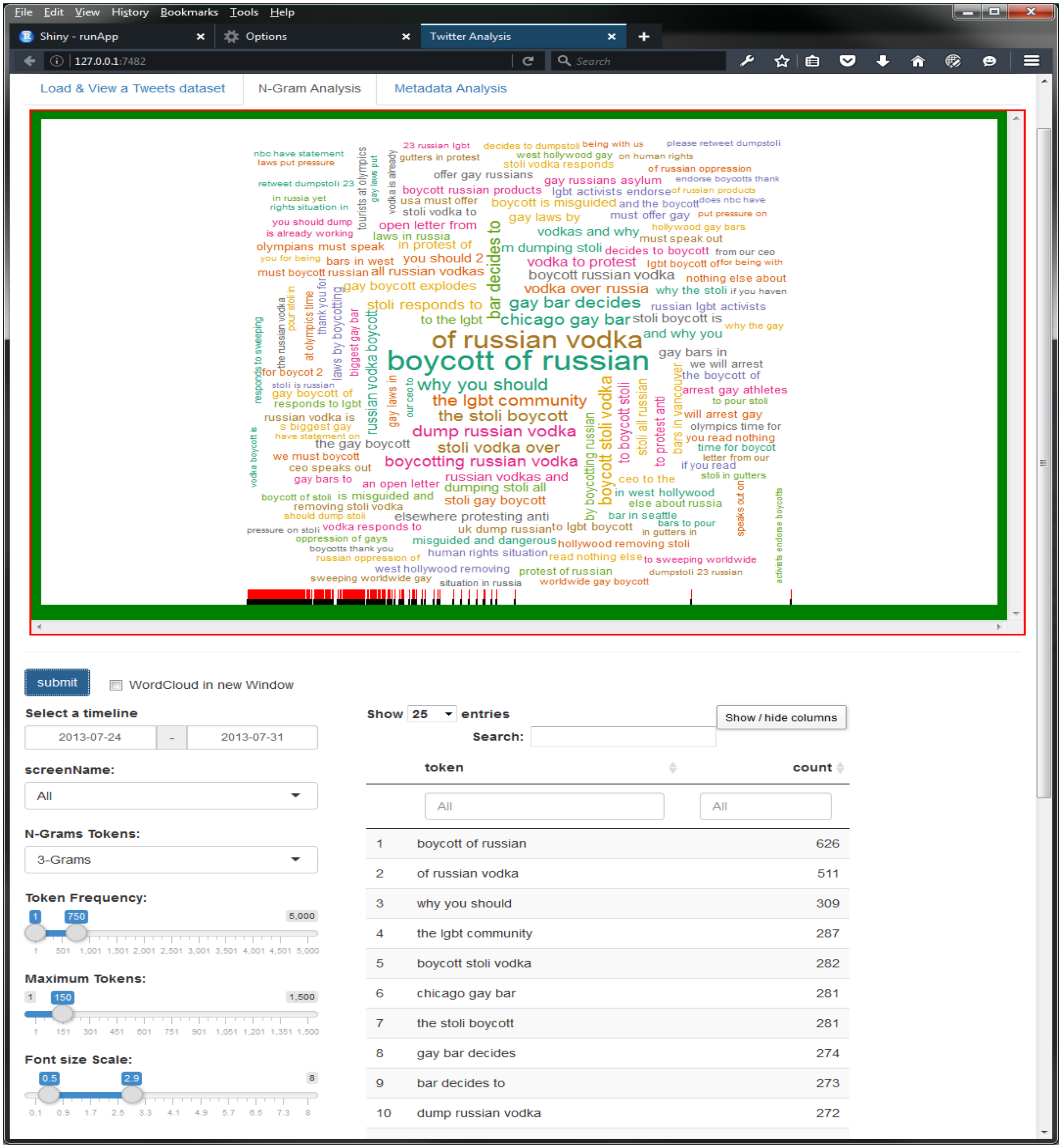
Upload complete

Show 25 entries

Search:

	text	screenName	created	statusSource
	All	fakedansavage	All	All
5	Best piece I've read on Wienergate 2.0: http://t.co/V2e0mQtUJ1	fakedansavage	2013-07-23 21:04:34	web
62	"Some pink boys may benefit simply from meeting a swishy gay man..." Some required reading from @AliceDreger: http://t.co/m9lTMPd5XA	fakedansavage	2013-07-23 21:36:30	web
64	@JoeMyGod @PeterLaBarbera Klan denied booth space at NAACP meeting. Film at 11.	fakedansavage	2013-07-23 21:37:34	web
84	Ouch: http://t.co/8XTK5ysQFn	fakedansavage	2013-07-23 21:47:16	web
253	@lust4ice Okay.	fakedansavage	2013-07-23 23:55:47	web
254	RT @jbarro: Carlos Danger for Secretary of Awesome	fakedansavage	2013-07-23 23:56:28	web
272	So... not a fan? Ã¢â ¬Å @lust4ice: @fakedansavage Your ideals regarding cheating are SO wrong. You are an ASS with no morals. Trashing yr book.Ã¢â ¬Â	fakedansavage	2013-07-24 00:15:55	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
291	RT @NickHanauer: Another proud conservative gun nut. http://t.co/G12Us4Q0Gi	fakedansavage	2013-07-24 00:20:27	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
339	@javicuna Because I'm dainty and prim, of course.	fakedansavage	2013-07-24 01:27:23	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
346	RT @BrianKentMusic: It's time to call a BOYCOTT OF THE 2014 OLYMPICS IN RUSSIA. This violation of Human Rights, beatings, peaceful... http:Ã¢â ¬Â¦	fakedansavage	2013-07-24 01:39:50	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
492	RT @joshtpm: NYT: Woman In Weiner's Sex Chat Made Contact To Chide Him For First Scandal http://t.co/YLNPqnFQKx #thegreatpersuader	fakedansavage	2013-07-24 10:58:56	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
509	Judging from my mail... there are guys out there who'll wanna be this straight guy's gay roommate. http://t.co/YoW6XYQbVo @kenneth212	fakedansavage	2013-07-24 11:59:35	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
519	RT @gabrielarana: Dutch tourists arrested under Russia's anti-gay "propaganda" law banned from the country for 3 yrs. http://t.co/pxiDo97Vyk	fakedansavage	2013-07-24 12:07:20	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
520	RT @JustinCSnow: Lesbian couple married in Pennsylvania as county officials defy state's same-sex marriage ban http://t.co/WF2yvbz5lm	fakedansavage	2013-07-24 12:14:30	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
535	RT @hudsonism: Dear Sochi, Please read Olympic Principle Number 6 and get back to me. http://t.co/7J8itcCL1a	fakedansavage	2013-07-24 12:42:49	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
539	@kaitrinh @gabrielarana There's a diff between "not great" and "actively persecuting."	fakedansavage	2013-07-24 12:47:59	<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
567	A must-read post by @sullydish about Anthony Weiner: http://t.co/lnQ2AnL8Ry	fakedansavage	2013-07-24 13:30:46	web
575	@_Ridley_ @sullydish Disagree with that pointÃ¢â ¬â Andrew will get push back on it, and you'll see readers' letters. Post otherwise brilliant.	fakedansavage	2013-07-24 13:33:33	web
614	A new avatar for your Twitter feed. http://t.co/UdID2iWTwE	fakedansavage	2013-07-24 13:51:54	web

Figure 3: Mr. Savage’s timeline when dumpstoli, dumpRussianVodka avatar was tweeted



□ WordCloud in new Window

Select a timeline
2013-07-24 - 2013-07-31

screenName: All

N-Grams Tokens: 3-Grams

Token Frequency: 1 – 750 (1 … 5,000)

Maximum Tokens: 1 – 150 (1 … 1,500)

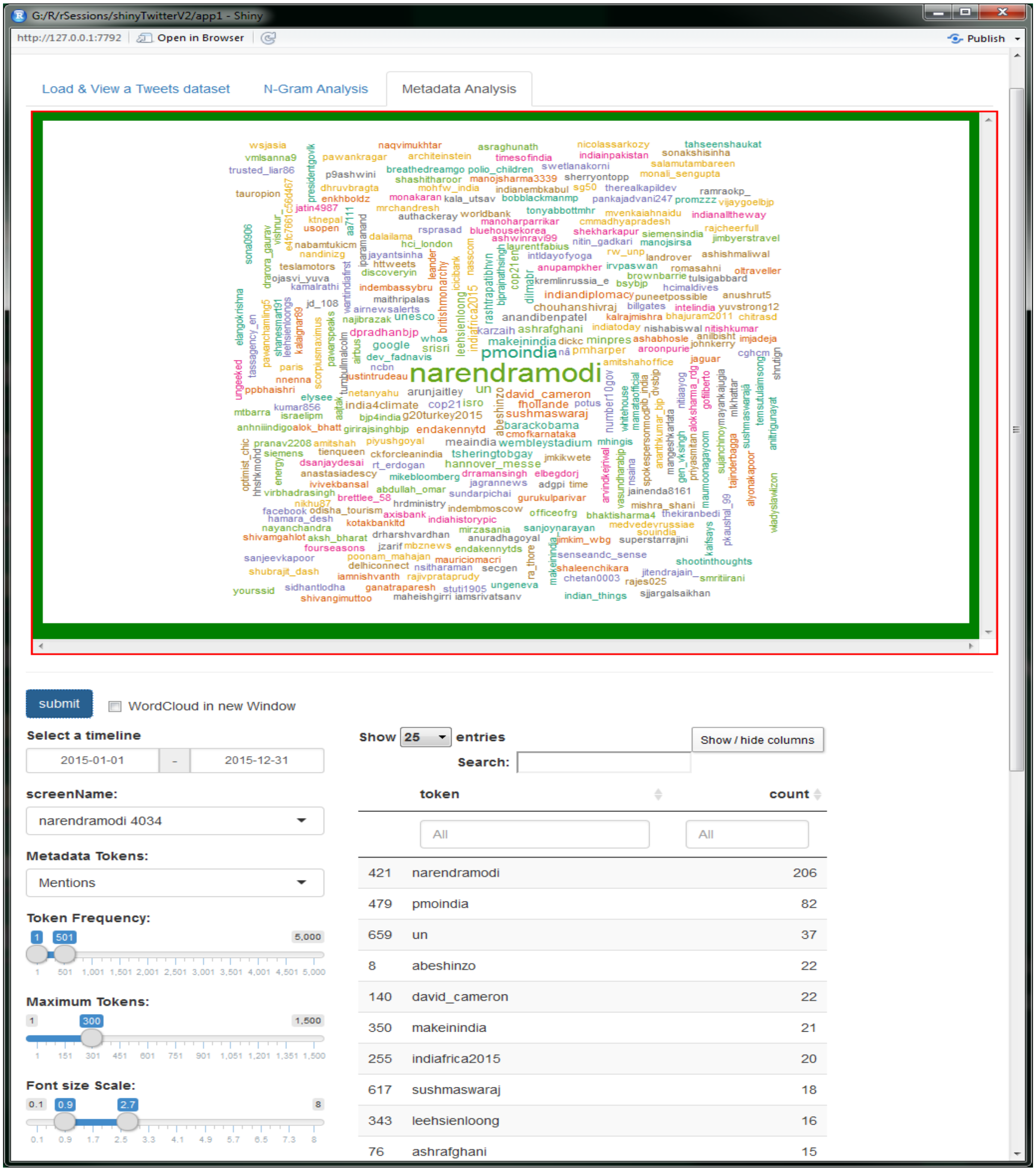
Font size Scale: 0.5 – 2.9 (0.1 … 8)

Show 25 entries

Search:

	token	count
	All	All
1	boycott of russian	626
2	of russian vodka	511
3	why you should	309
4	the lgbt community	287
5	boycott stoli vodka	282
6	chicago gay bar	281
7	the stoli boycott	281
8	gay bar decides	274
9	bar decides to	273
10	dump russian vodka	272

Figure 4: N-grams Analysis Tab: View all 3-grams of text in the dumpstoli tweets dataset



Select a timeline

2015-01-01 - 2015-12-31

screenName:

narendramodi 4034

Metadata Tokens:

Mentions

Token Frequency:

1 – 501 (range 1 to 5,000)

Maximum Tokens:

300 (range 1 to 1,500)

Font size Scale:

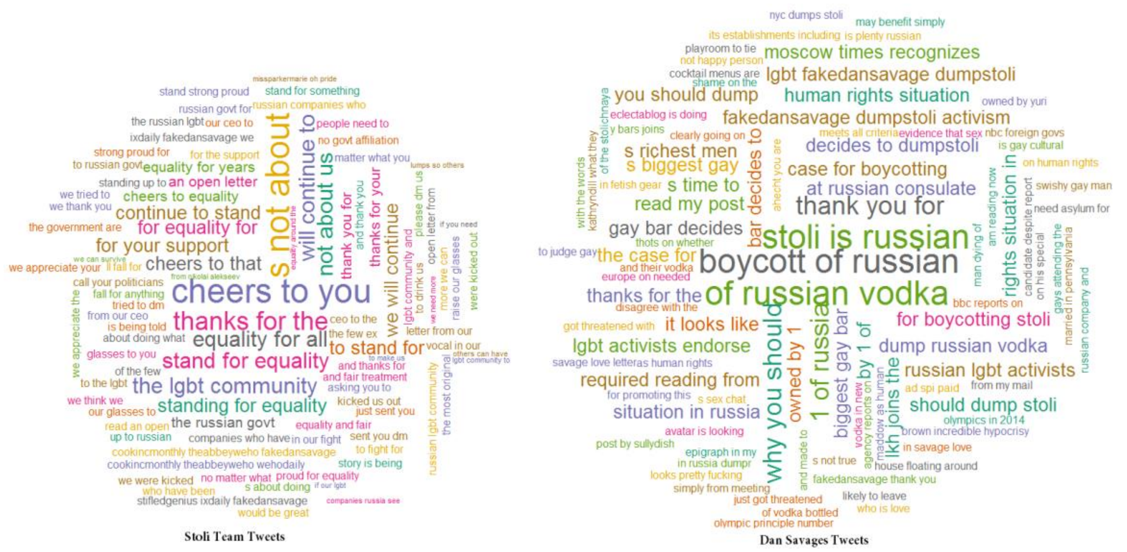
0.9 – 2.7 (range 0.1 to 8)

Show 25 entries

Search:

	token	count
	All	All
421	narendramodi	206
479	pmoindia	82
659	un	37
8	abeshinzo	22
140	david_cameron	22
350	makeinindia	21
255	indiafrica2015	20
617	sushmaswaraj	18
343	leehsienloong	16
76	ashrafghani	15

Figure 5: Metadata Analysis tab: P.M. Narendra Modi mentions



Stoli Team Tweets

Dan Savages Tweets

Figure 6: Comparing 3-grams of relevant tweets between Jul-24 & Jul-31 of 2013

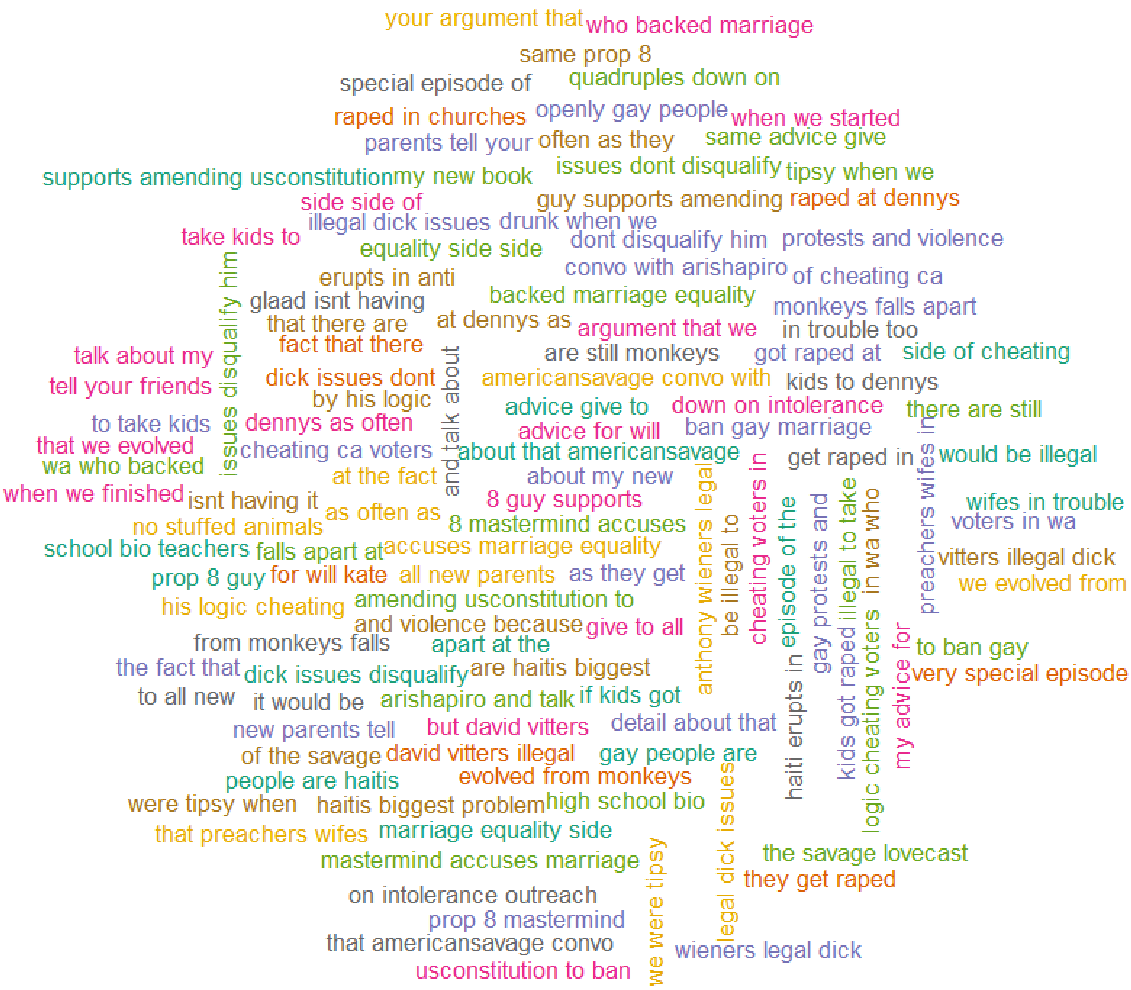


Figure 7: 3-gram tweets of Mr. Savage on 2013-07-23
To preserve space, we only displayed the word clouds

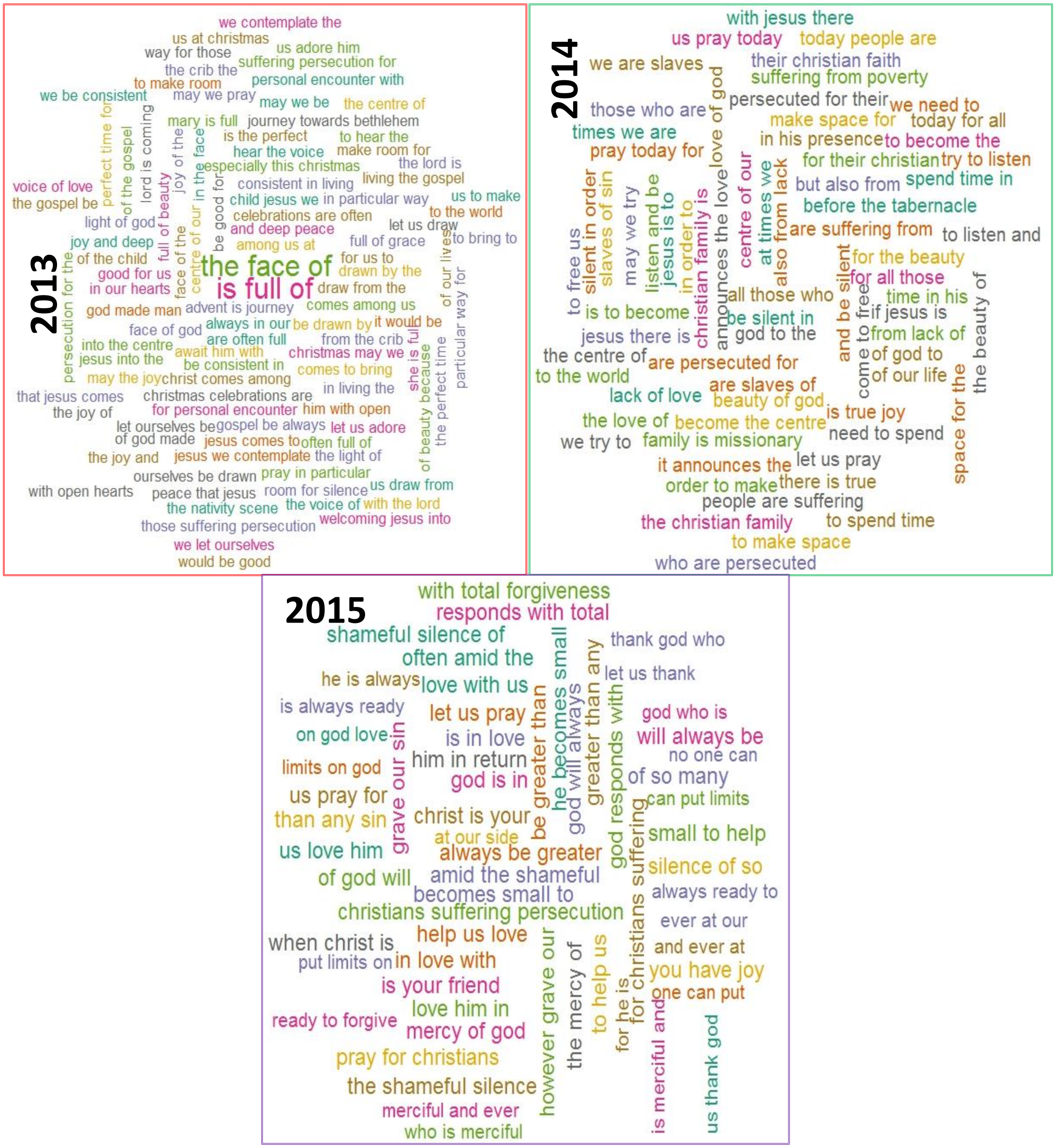


Figure 8: A 3-gram tweets of Pope Francis around Christmas 2013, 2014 & 2015