4/19/18

Dr. Gopathy Purushothaman

Data 650 9040

Assignment #3 - Claritin Sentiment Analysis Among Genders

Wesley Clark

iamwesleyclark@gmail.com

(301) 254-9395

Introduction and Background

An analysis was conducted to review sentiment data and tweets surrounding a popular allergy medication, Claritin. The dataset used in this analysis was public dataset available through data.world. This data has all tweets that contain a reference to Claritin for October 2012(Claritin Twitter, 2016). A sentiment analysis has been conducted that tags the sentiment into 1 of 5 categories depending on how the author felt about Claritin. This dataset was comprised of 4900 instances. Each instance represented a separate tweet. The dataset was further divided into 17 columns. In order, the columns are INTERACTION_ID, ARTICLE_URL, CONTENT, TIME, RELEVANT, SENTIMENT, GENDER, DIZZINESS, CONVULSIONS, HEART PALPITATIONS, SHORTNESS OF BREATHE, HEADACHES, DRUG EFFECT DECREASED, ALLERGIES WORSE AFTER TAKING DRUG, BAD INTERACTION BETWEEN CLARITIN AND ANOTHER DRUG., NAUSEA (MADE THE PERSON NAUSEOUS), T ABLE TO SLEEP). INTERACTION_ID is a unique identification tag that is different for each entry and serves to be able to reference each instance independently. ARTICLE_URL is the URL in which the tweet was derived from. CONTENT is the text contained within the tweet. TIME is comprised of the time and date of each instance. RELEVANT describes whether or not the tweet was in english or not. If so the category is marked as ‘yes’, and if not it is labelled as ‘non_english’. SENTIMENT is a numeric input that is used to display the results of the sentiment analysis. The sentiment has 5 categories, each of which is described in a later section. This category is largely the study of this analysis. GENDER was the following column and was divided into ‘male’ or ‘female’. The next 10 columns are complications or side effects that are possible as a result of taking Claritin. They are

all divided into ‘yes’ or ‘no’ categories. These columns include DIZZINESS(whether or not the subject experienced a sensation of spinning around losing ones balance), CONVULSIONS(sudden violent involuntary and irregular movement of the body), HEART PALPITATIONS(an irregular oscillation of heartbeat), SHORTNESS OF BREATHE(the experience of shallow or rapid breath), HEADACHES(experience of pain or aching sensation in the cranial area), DRUG EFFECT DECREASED(a determination of whether or not Claritin maintained or decreased in terms of efficacy over time), ALLERGIES WORSE AFTER TAKING DRUG(the experience of allergies worsening even after treatment), BAD INTERACTION BETWEEN CLARITIN AND ANOTHER DRUG.(the unwelcome experience of a synergistic effect caused by taking Claritin in addition to another drug), NAUSEA (MADE THE PERSON NAUSEOUS)(stomach and head malady characterized as experience before vomit. The last column T ABLE TO SLEEP). This column was meant to indicate insomnia(loss of sleep). As the title was not considered fitting, the data was deleted and reloaded to reflect the proper column name, INSOMNIA.

Programmatic Approaches

To analyze this data IBM Db2, RStudio and SQL were used. IBM Db2 integrated SQL and RStudio and was the main platform for analysis. This proved to be an effective way of analyzing data as multiple methods of analysis were present within a single application. This platform is an efficient way to handle many data needs and is marketed under the IBM and Watson umbrella. Many benefits to this service are realized. In this analysis alone, the ability to analyze data using RStudio as well as SQL directly integrated into the platform made the experience smooth and

fluid. Although the interchangeability of RStudio and SQL was a nice feature of IBM Db2, the most impressive feature was the ability to upload large quantities of data. This is a unique challenge that does not have any single silver bullet solution. After downloading the dataset used in this analysis, it was uploaded onto IBM Db2 Warehouse on Cloud. This upload took about 2 minutes and occurred on April 19th. All 4900 rows were read and loaded. 0 Rows were rejected and 0 errors were reported(). The method of investigation included an initial inspection into the dataset through SQL. By querying the dataset a better understanding of how the dataset was comprised as well as what values to expect within each column was done. Additionally, subqueries were performed to look even further into correlations within the data. To accomplish this 35 queries were conducted. The first query served to verify that the data was loaded correctly by counting the instances of INTERACTION_ID. An output of 4900 achieved this. The next query was performed independently on each of the 8 side effects including DIZINESS, HEART PALPITATIONS, SHORTNESS OF BREATHE, DRUG EFFECT DECREASED, ALLERGIES WORSE AFTER TAKING DRUG, BAD INTERACTION BETWEEN CLARITIN AND ANOTHER DRUG, NAUSEA, INSOMNIA and HEADACHES. This mostly served to get a better understanding of what was present within each column and obtain a better overall view of how the dataset was constructed. One example of how this proved to be worthwhile occurred during a count of HEADACHES(headaches SS). It was noticed on this query that the possible outputs were ‘no’ and ‘ye’. It was assumed that ‘ye’ was short for yes, and another query was run to verify that this was correct(HeadachesYESS). Next, GENDER and RELEVANT were counted and grouped allowing a better view of the data to be complete. Once the data was surveyed, the next step was to group the sentiments. This was done via query and

subquery. The data was next loaded into the RStudio environment. To accomplish this the ibmdbR(https://cran.r-project.org/web/packages/ibmdbR/ibmdbR.pdf) package was loaded. This package is required to use R in combination with IBM Db2. Next proper credentials were input at the onset of the data load(data credentials SS). This included creating the variables dsn_driver, dsn_database, dsn_hostname, dsn_port, dsn_protocol, dsn_uid, and dsn_pwd. These credentials were then input into a new variable, conn_path. Next conn_path was called to idaConnect and saved as mycon. Eventually the database was initialized using the idaInit function on mycon. This allows for existing RODBC connection to initialize the iDA in-database analytics functions(https://www.rdocumentation.org/packages/ibmdbR/versions/1.49.0/topics/idaInit). Next the Claritin variable was created and the data was saved as a data.frame so that any SQL query could be run. This was done using the idaQuery function. The query used to accomplish this was a simple SELECT ALL FROM CLARITIN_SIDEFFECTS. This completed the data loading process in R. Several commands were now run to check that the database was loaded properly into RStudio.

- nrow(Claritin) (nrowClaritingSS) displayed 4900 rows.
- idadf(mycon, "SELECT count(*) FROM CLARITIN_SIDEFFECTS”). The idadf function allowed the above query to be done resulting in a count displaying 4900 entries
- dim(Claritin) output the dimensions of the database. We observed 4900 rows with 17 columns.
- str(Claritin) revealed that the data is a saved in data.frame. It also enabled 4900 instances of 17 variables and examples of each to be seen.
- summary(Claritin) displayed a summary of the dataset, including a breakdown of each column.

- Because SQL is directly integrated into the IBMdB2 environment, a query was run in the SQL editor to confirm the results received in R. Each instance of INTERACTION ID was counted, revealing 4900 instances.
- Additional commands of idaShowTables(), and idaExistTable('CLARITIN_SIDEFFECTS') was completed to ensure the existence of the correct tables.

Data Analysis

The data was confirmed to be properly loaded at this point. The next step was to continue with the analysis by conducting a SQL query with idadf(). The data was then loaded into a table display a clean output of sentiment quantities. Now that the data has been correctly analyzed, it was time to continue with graphic vizualization. A pie chart was conducted first, and a bar graph displayed next. These two forms of visualization give a telling overview of the data and the sentiments within. It is clear that the sentiments were dominated by more neutral responses, many of which fall into the sentiment classification of 4. A second bar graph was done which separated each sentiment by color, as well as included the NA values. Not all tweets contained sentiments, and it was worthwhile to include a second graph with this present. A continued analysis of the tweets was done to inspect each sentiment. Through both SQL and R it was clear that sentiments 4 and 5 were positinve(5 being the most positive), 3 a neutral stance, 1 and 2 negative sentiment, with one being the worst. The positive sentiments were grouped using another idadf query including a where statement setting sentiment equal to 4 or 5. Next neutral following the same logic with sentiment 3. Finally negative with sentiment 1 and 2.

The next step was to prepare a corpus for text mining. This will enable different kinds of language visualizations. It was chosen to obtain a word cloud with corresponding terms analysis, as well as a network of terms with corresponding adjacency matrix. A corpus was created, white space, punctuation, numbers, stop words were all removed. Next a matrix was created as a document term matrix. This matrix describes the frequency of words in a collection of documents. The word cloud was next conducted. As expected, the single biggest word present is claritin, followed by the stem of allergi. Sparse terms were removed as well as word frequencies discovered to plot this. Once the word Cloud was conducted a new term matrix was done. This time the sparse terms was set to 0.95, and the as.matrix function was applied to the corpus. A term matrix was displayed showing the frequencies of the most common words as they relate to eachother in an adjacency matrix. Finally, this information was plotted in a network of terms analysis. Further analysis was conducted to discover how gender played a part in the sentiment of tweets. This was done with both SQL and R. Each gender contained a subquery among each sentiment to reveal the breakdown of male and female sentiments for all tweets. The goal of this portion of the analysis was to notice if there was a difference between the male and female populations. At this point it was clear that additional analysis was needed.

Sentiment	Frequency	Female/Male Count	Percent Female
1	87	59/20	74.7%
2	641	412/166	71.3%
3	1421	712/456	61.0%
4	2328	1000/857	54.9%
5	191	118/59	66.7%

It was shown that females tweeted significantly more than males. Also striking was that the negative sentiments were more often tweeted by females than their male counterparts.

Limitations and Conclusions

This study proved to be rewarding as it challenged me to use both old and new approaches into one analysis. One accomplishment that I found rewarding was to be able to use SQL and R together in the IBMdB2 platform. Additionally, being able to correlate how people tweet and how different sexes tweet was accomplished. Not only did the women tweet more than the men, the women expressed negative sentiments at a higher frequency then the men. A continued study would be to research potential reasons this is true. It would be interesting to see if this is true in other fields, outside of medicine. One possible explanation is that more women were taking Claritin then men, or perhaps Claritin was targeting women with specific marketing. Until more analysis is done on how different genders tweet, it is difficult to draw concrete conclusions in this regard. Another limitation of this study was the data itself. This data was all collected on Twitter, and may be inherently biased and not reflective of actual numbers. It has become increasingly clear that not all Twitter accounts are even real people(Wang H, 2012). For this reason, it was deemed that correlating FDA numbers to twitter numbers may not be fair. The average person tweeting may be predisposed to certain inclinations. The fact that a person is tweeting at all could be indicative of access to technology. Another limitation to this study was not knowing enough about the individuals that were tweeting. It would have been very useful to know things such as the age, nationality and race of each instance. This would have enabled much more in depth analysis, and likely yielded significantly stronger correlation as well as more

confidence in the results. Another limitation occurred in the NLP aspect of this study. It was noticed upon manual inspection of the tweets that some tweets were sarcastic, or being used as a joke. These tweets were misclassified as positive or negative, when in fact the tweeter may have never even used Claritin themselves. For continued analysis it is recommended that different kinds of data be collected outside of twitter for analysis. A survey of 1000 Claritin users would be sufficient to compare to this data and reveal how accurate it is. Additionally, it is recommended that more detailed information be collected moving forward.

Appendix A

A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. (n.d.). Retrieved from https://dl.acm.org/citation.cfm?id=2390490

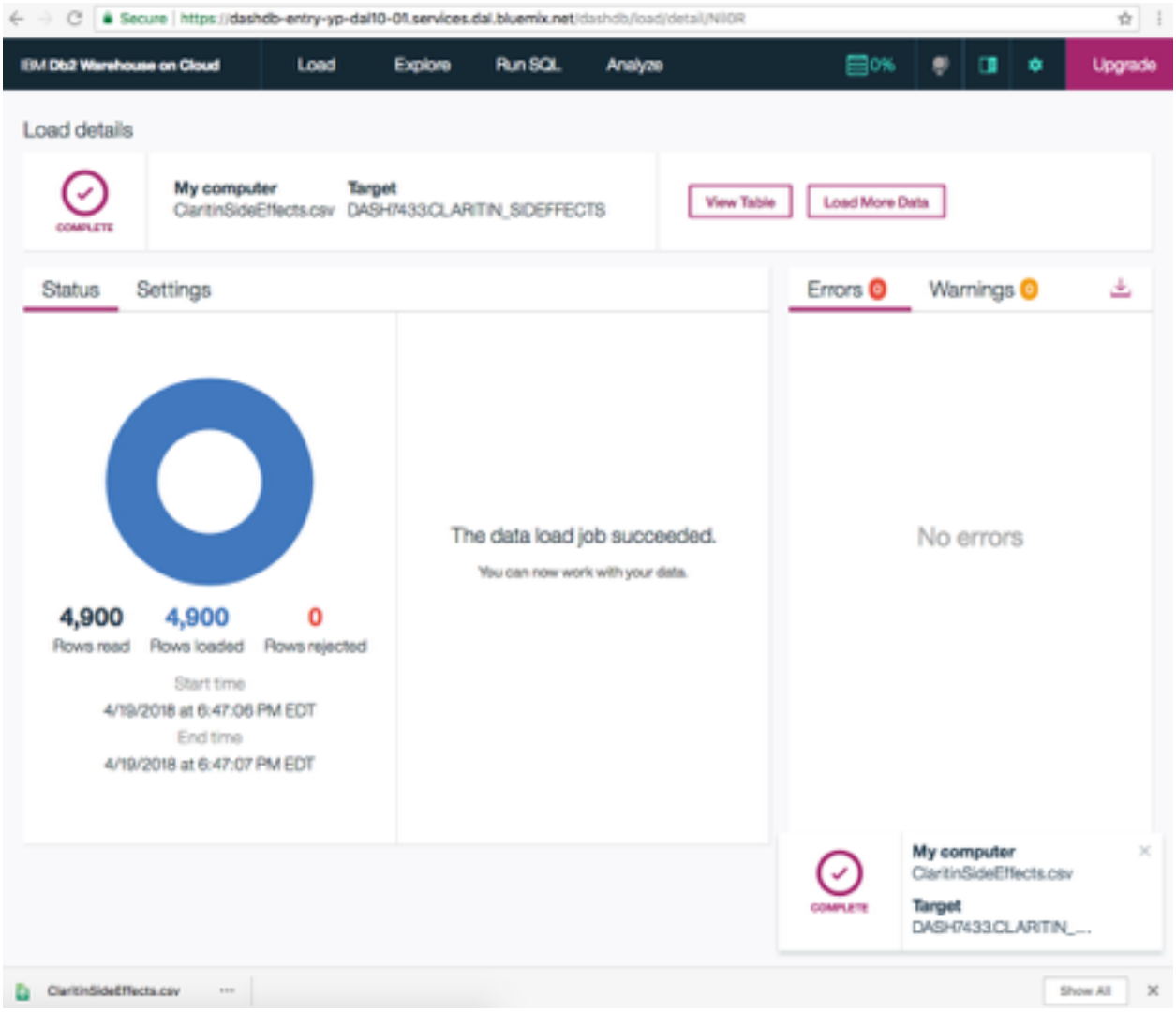
Claritin Twitter - dataset by crowdflower. (2016, November 21). Retrieved from https://data.world/crowdflower/claritin-twitter

IBM DB2 for i. (2012, January 13). Retrieved from https://www-03.ibm.com/systems/power/software/i/db2/benefits.html

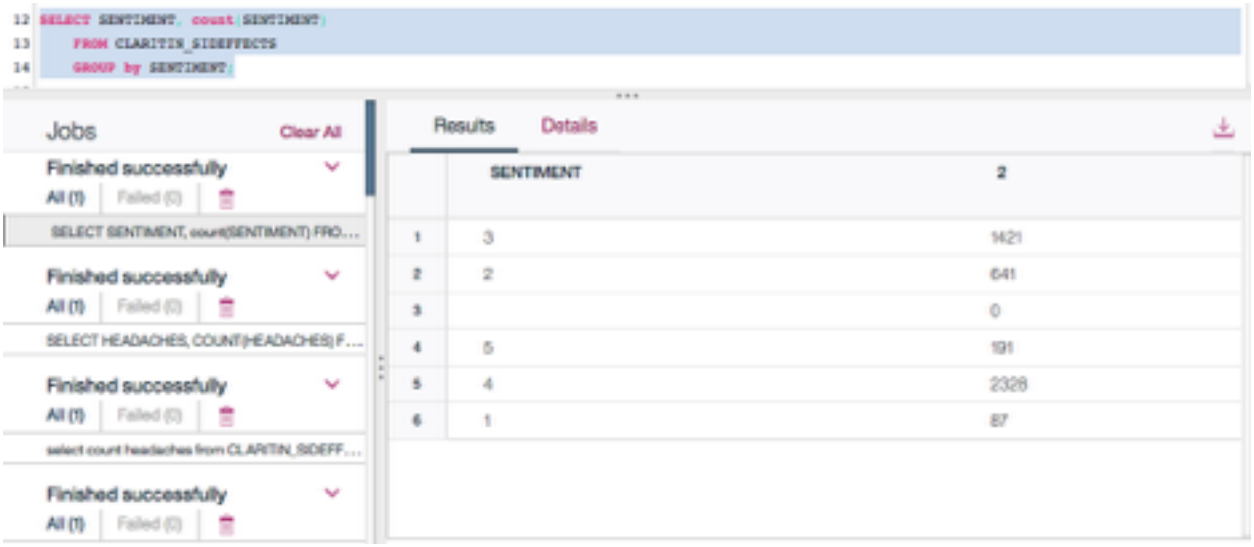
Sentiment analysis of Twitter data. (n.d.). Retrieved from https://dl.acm.org/citation.cfm?id=2021114

Appendix B

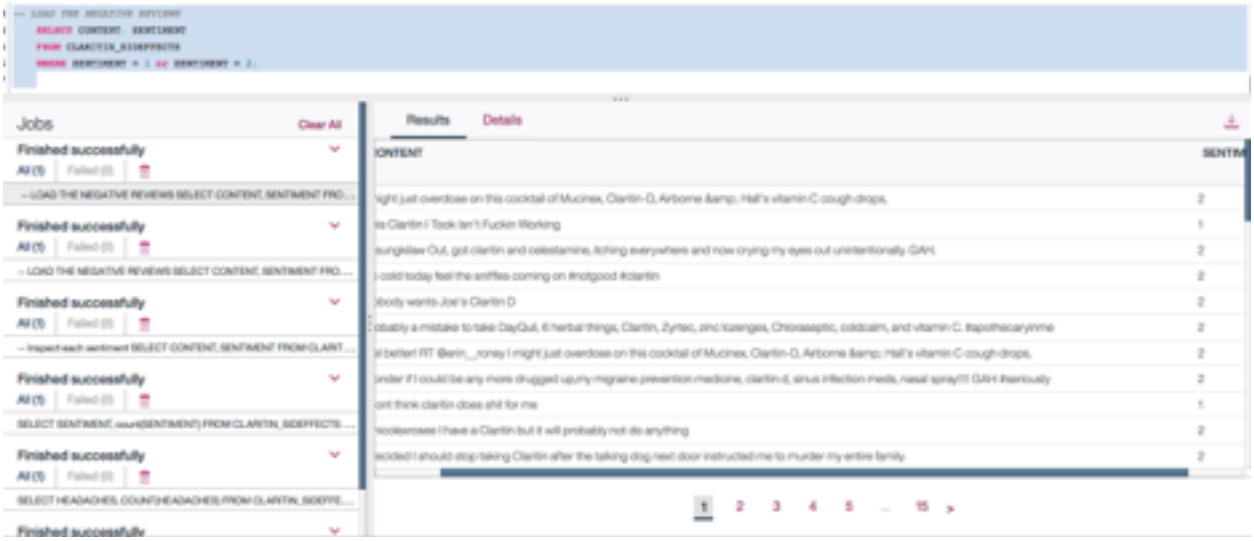
Data Load



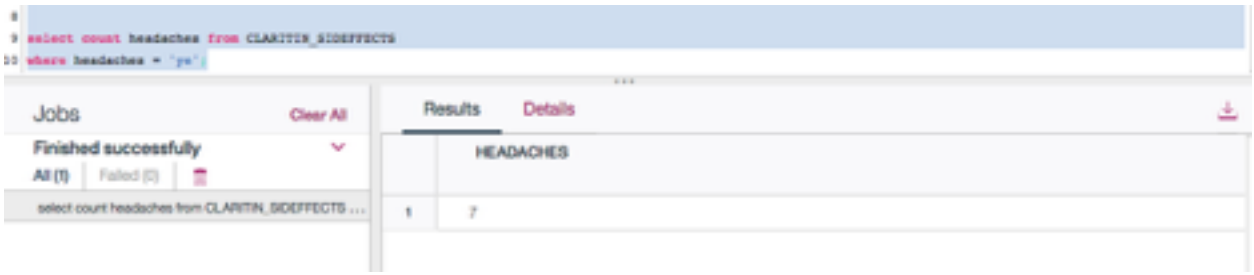
Sentiment Query



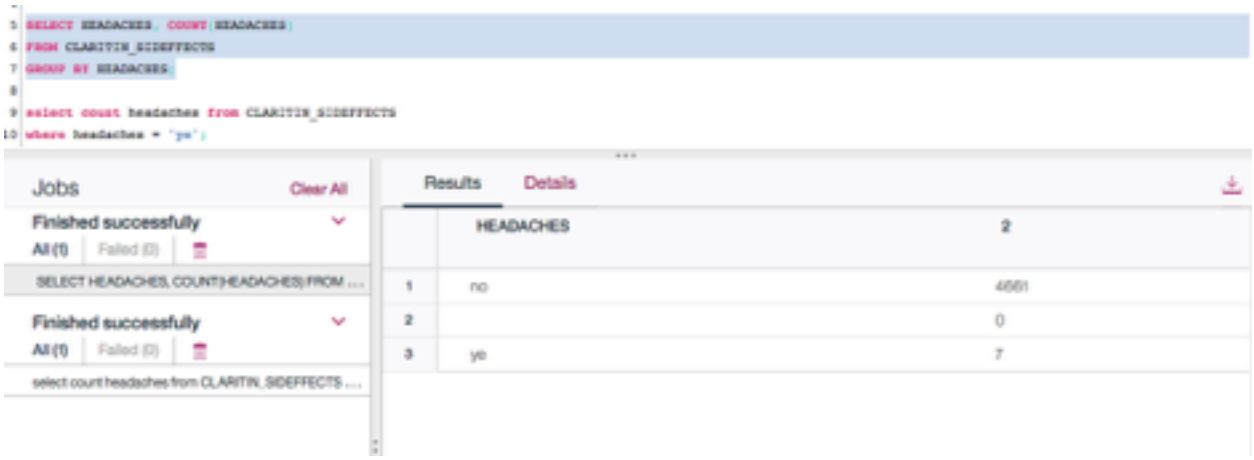
Query Example and Output



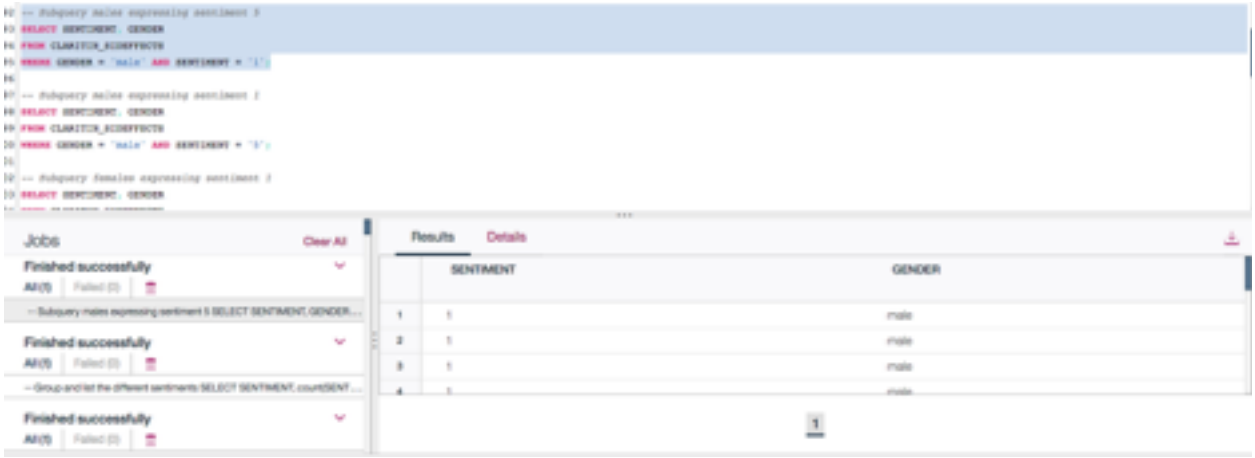
Headaches Query



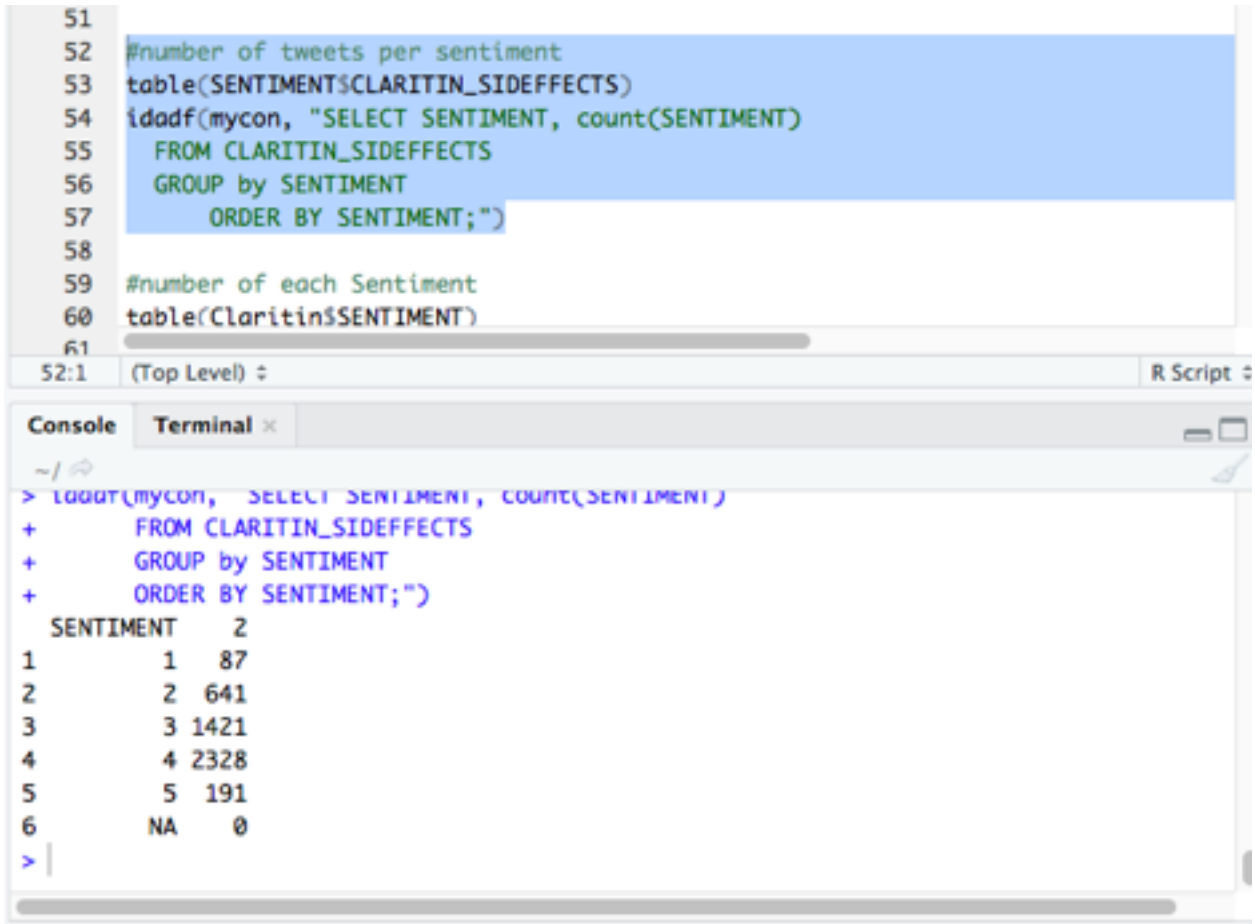
Headaches ‘ye’ Query



Male Sentiment 1 SubQuery



Sentiment Group Query in RStudio



Claritin Document Term Matrix and Preview of Word Cloud Matrix

	claritin	drug	nieanadya	overthecount	seafood	stay	store	time	act
1	1	1	1	1	1	1	1	1	
2	1	0	0	0	0	0	0	0	
3			0	0	0	0	0	0	
4	1	0	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	
6	1	0	0	0	0	0	0	0	
7	1	0	0	0	0	0	0	0	
8	1	0	0	0	0	0	0	0	
9	1	0	0	0	0	0	0	0	
10	1	0	0	0	0	0	0	0	

column 0: rownames

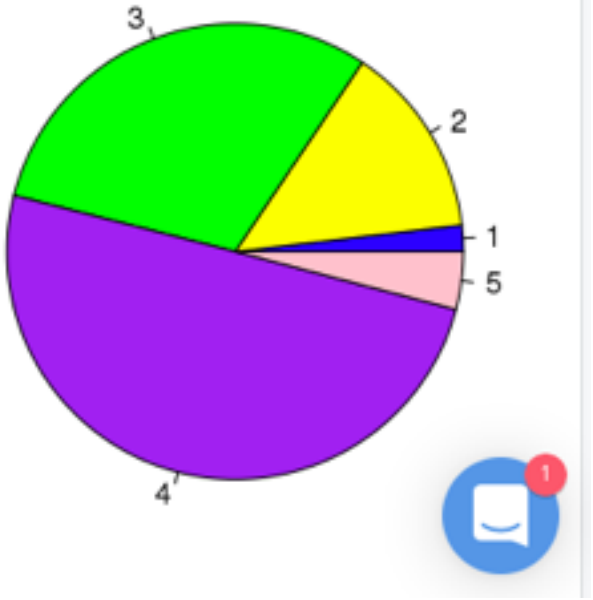
Showing 1 to 11 of 4,900 entries

termDocMatrix | ClaritingSE.R* | m | positive | termMatrix

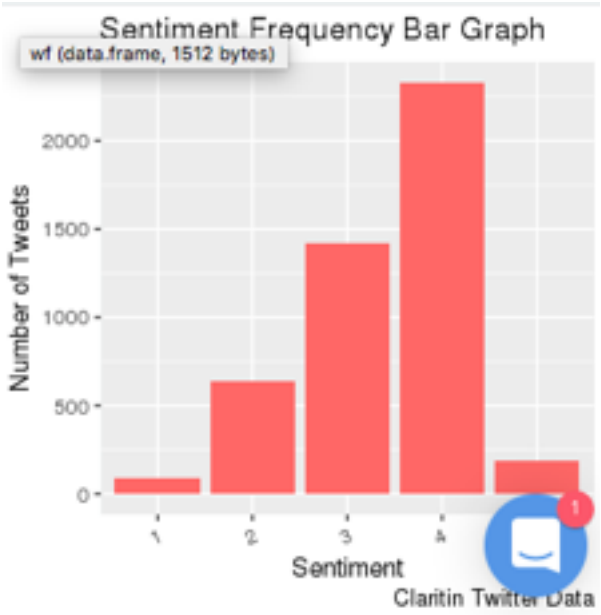
Filter

	claritin	drug	nieanadya	overthecount	seafood	stay	store	time	act

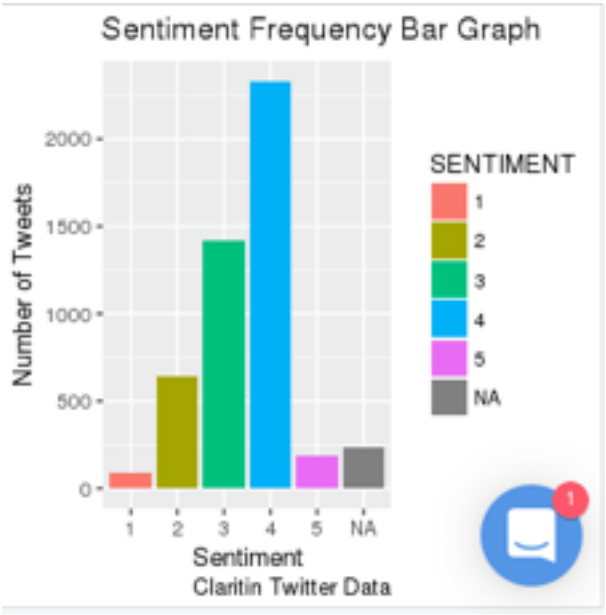
Sentiment Pie Chart



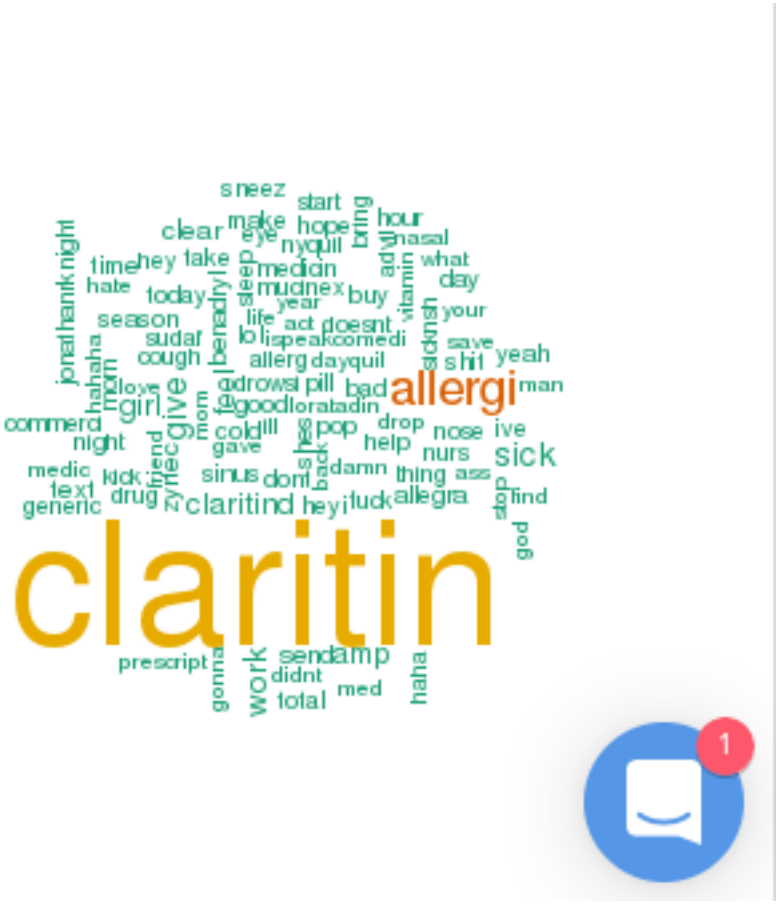
Sentiment Bar Graph 1



Sentiment Bar Graph 2



Claritin Word Cloud and DTM

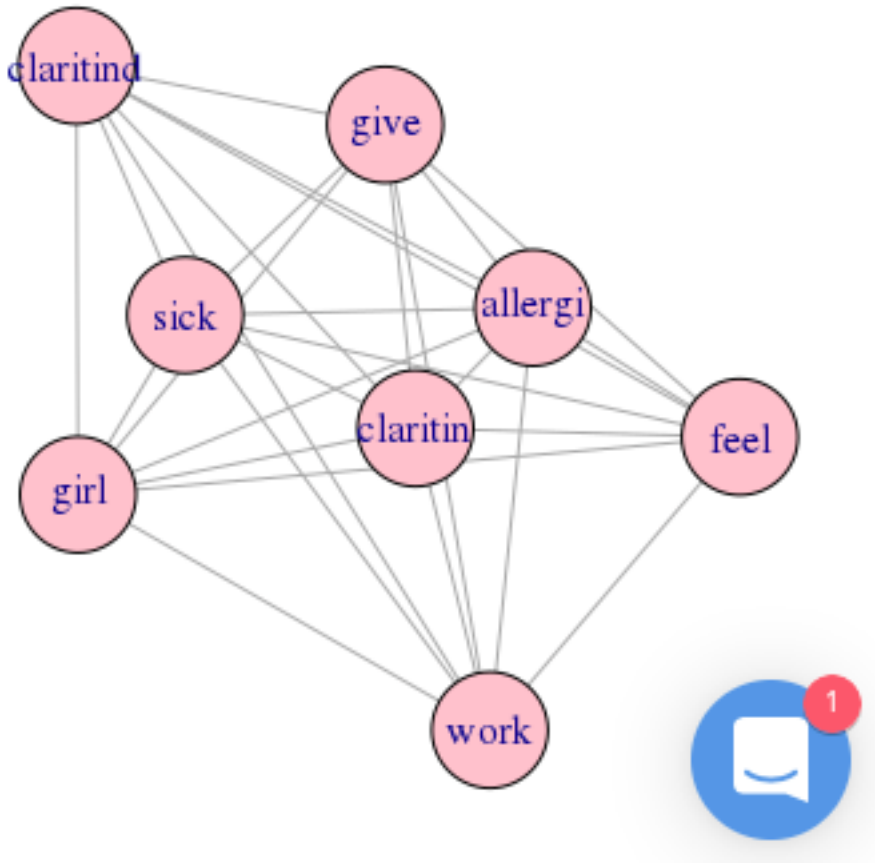


	claritin	drug	nieanadya	overthecount	seafood	stay	store	time	act
1	1	1	1	1	1	1	1	1	
2	1	0	0	0	0	0	0	0	
3			0	0	0	0	0	0	
4	1	0	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	
6	1	0	0	0	0	0	0	0	
7	1	0	0	0	0	0	0	0	
8	1	0	0	0	0	0	0	0	
9	1	0	0	0	0	0	0	0	
10	1	0	0	0	0	0	0	0	

column 0: rownames

Showing 1 to 11 of 4,900 entries

Claritin Vertex and Adjacency Matrix



	claritin	allergi	work	claritind	give	feel	girl	sick
claritin	4458	744	296	7	310	207	180	264
allergi	744	800	50	37	130	29	9	12
work	296	50	319	13	3	16	1	3
claritind	7	37	13	285	65	12	66	102
give	310	130	3	65	381	17	32	104
feel	207	29	16	12	17	226	7	18
girl	180	9	1	66	32	7	274	169
sick	264	12	3	102	104	18	169	404