Data 650 – Fall 2019

Assignment 3 – Sentiment Analysis of Coachella 2015

Kenneth Lulie

Professor Purushothman

University of Maryland University College

November 16, 2019

**Assignment 3: Part One**

1. **Introduction – Coachella 2015 Sentiment Analysis**

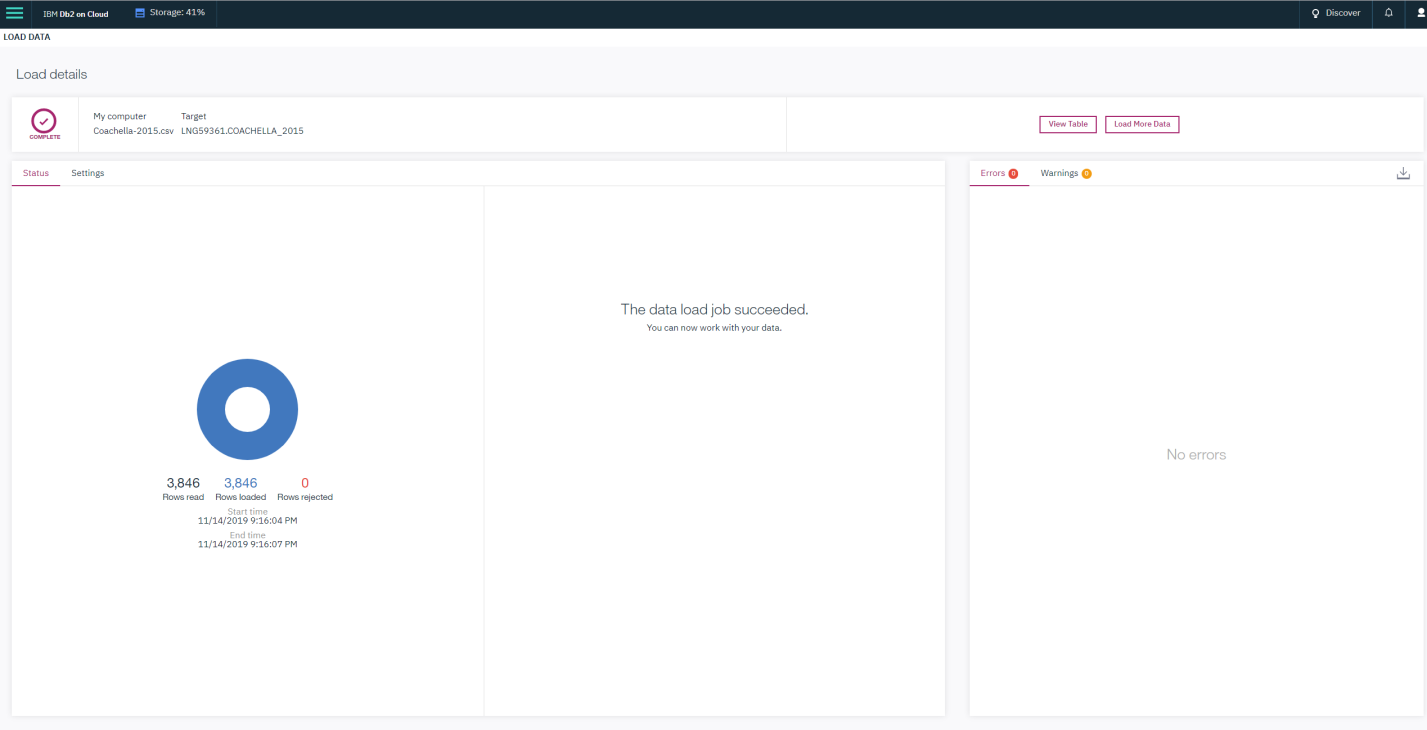
The Coachella Valley Music and Arts Festival is a famous music festival held annually (O’Leary, 2019). It is attended by quarter million people every year, and is hosted at the Empire Polo Club in the Coachella Valley in Colorado. The gross income from the event is typically over a hundred million dollars. The fees allow the event to hire the most popular musicians and bands.

On January 6, 2015 the lineup for the 2015 Coachella music festival was revealed, showing that the headliners for the event that year would be AC/DC, Jack White, and Drake (Coachella, 2015). Tweets including the hastag ‘#Coachella2015’ were harvested for the time period of January 6, 2015 through January, 2015. The purpose of this analysis will be to analyze these tweets for useful information to provide to the Coachella 2015 social media team on what the reception to the lineup announcement was on twitter for those two days.

The data used in this analysis is the text of 3846 tweets along with metadata including the sentiment of the tweet, the username of the poster, the number of retweets, the coordinates of the tweet, the time the tweet was created, the id of the tweet, the location of the tweet, and the timezone of the user who posted the tweet.

1. **Data Load of Tweets on IBM DB2 Cloud Web Console**

The data was loaded into the IBM DB2 Cloud web service as shown in the screenshot below. The status page showed that the data loaded successfully with all 3,846 rows imported and with no warnings or errors.

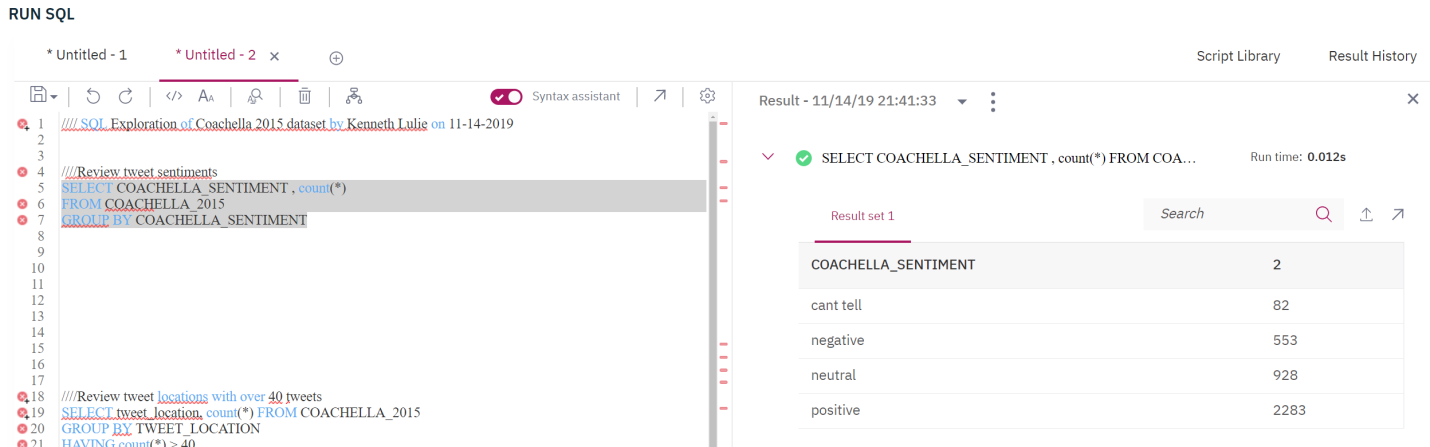


1. **Verification of Proper Data Load**

To verify the data was loaded correctly the number of rows was checked against the csv file which was uploaded, the date types of the columns were reviewed to ensure that they matched the expected data, and the data was previewed using the view feature to ensure the data appeared to be loaded correctly.

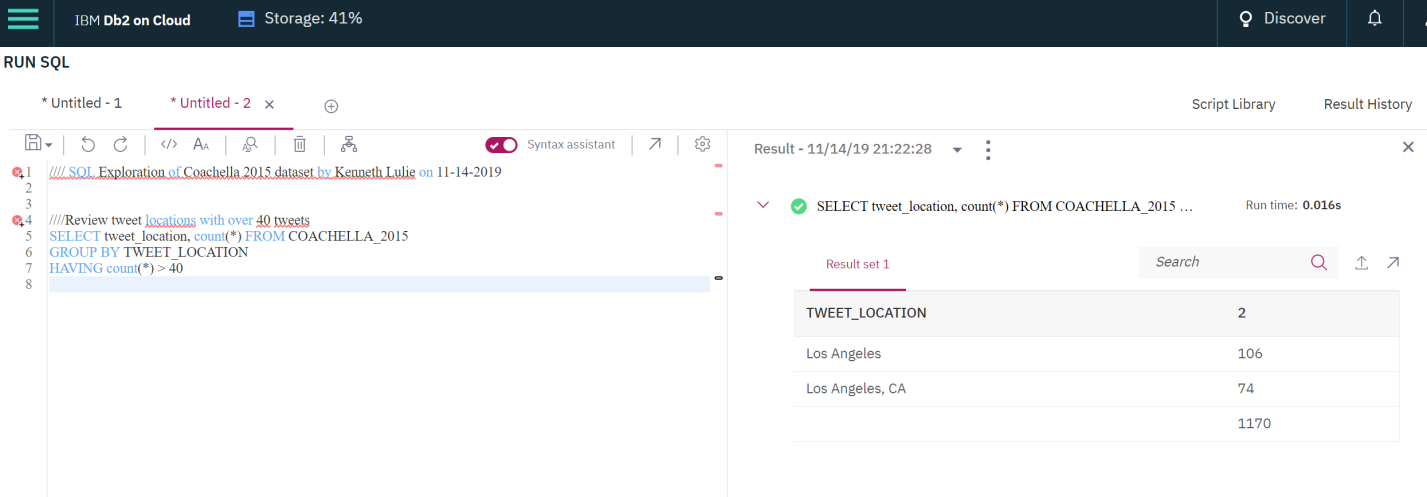
1. **Analysis of Data using SQL Editor in IBM Web Console**
   1. **Review of Number of Tweets per Sentiment**

The number of tweets of each sentiment were reviewed using an SQL query as shown below. From this review it is shown that the reception appeared to be overwhelming positive with about 60% of the tweets labeled as positive and only 14% labeled as negative.



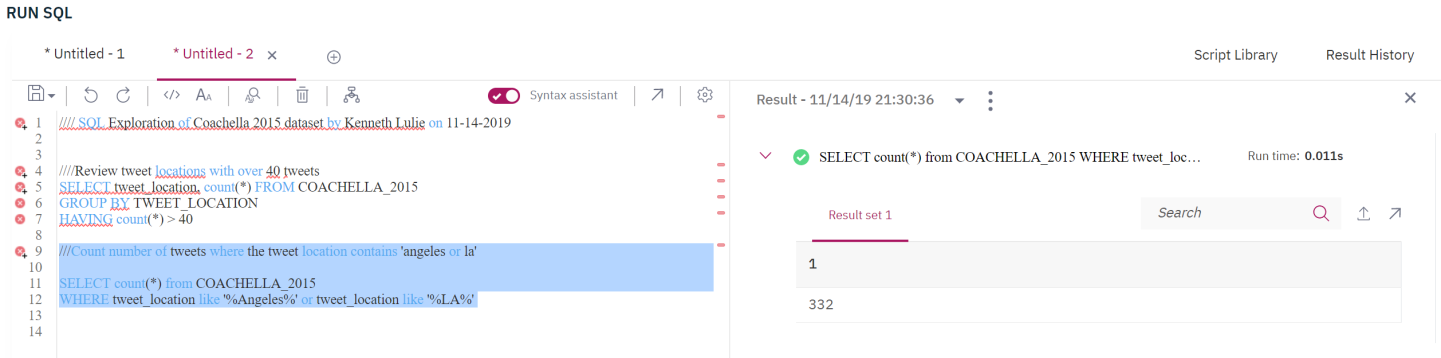
* 1. **Review of Number of Posts Per User Location with More than 40 Tweets**

The number of tweets per user location with more than 40 tweets was reviewed using an SQL query as shown below. From this review it is shown that the only user locations with more than 40 tweets were Los Angles and Los Angeles, CA. This raises concerns as to the data quality of the user location as it would be expected that both of these locations should be counted together. Based on this result it does not appear that the query is actually capturing all locations with more than 40 tweets. One explanation could be that the user location field is user entered which would explain inconsistencies in formatting. Additionally, 1170 tweets or about a third, have no user location at all.



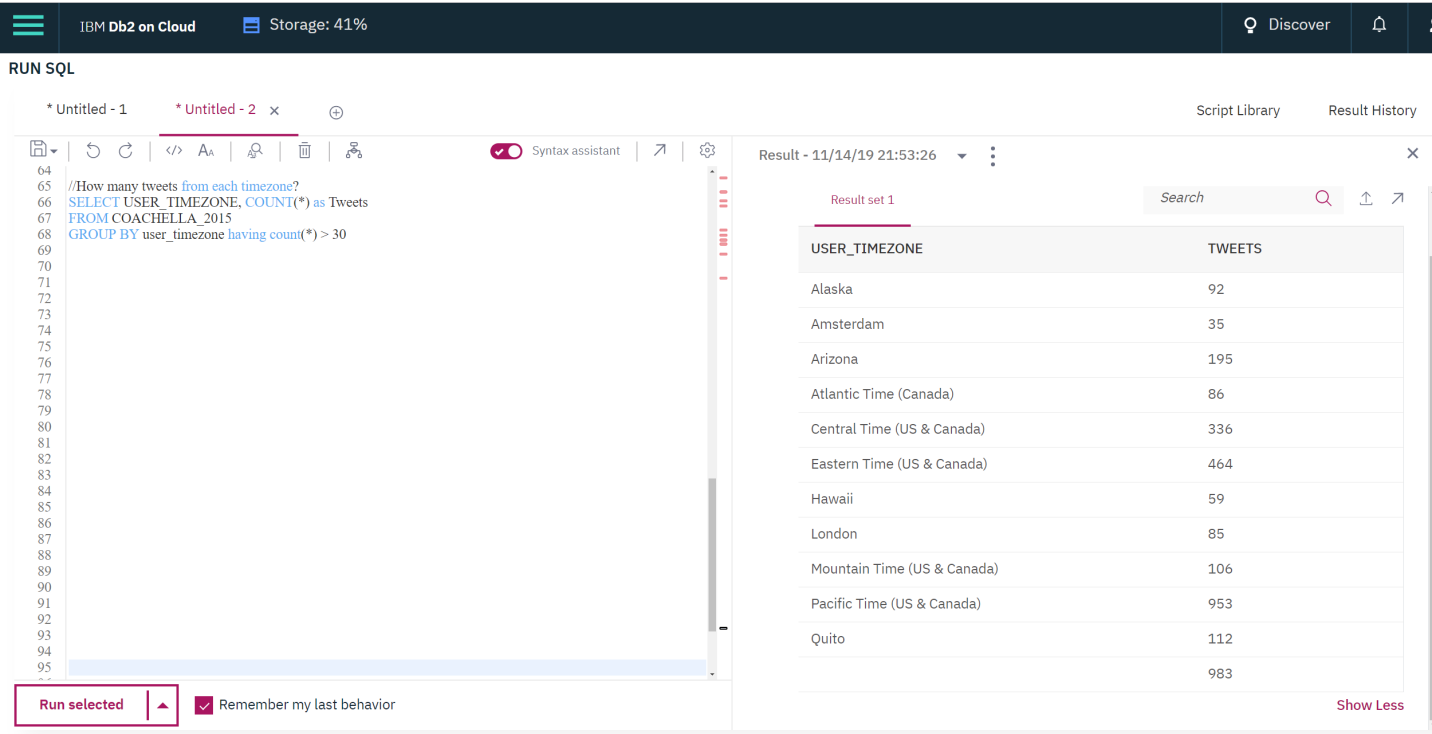
* 1. **Expanded Review of Tweets from Los Angeles**

The query below captures the number of all tweets that contain either ‘Angeles’ or ‘LA’ in the user location field. From the output it is shown that there are 332 tweets with this criteria, almost twice as many as detected by the above output. This implies that the formatting for user location is not consistent and that further exploration is required to understand specifics if asked.



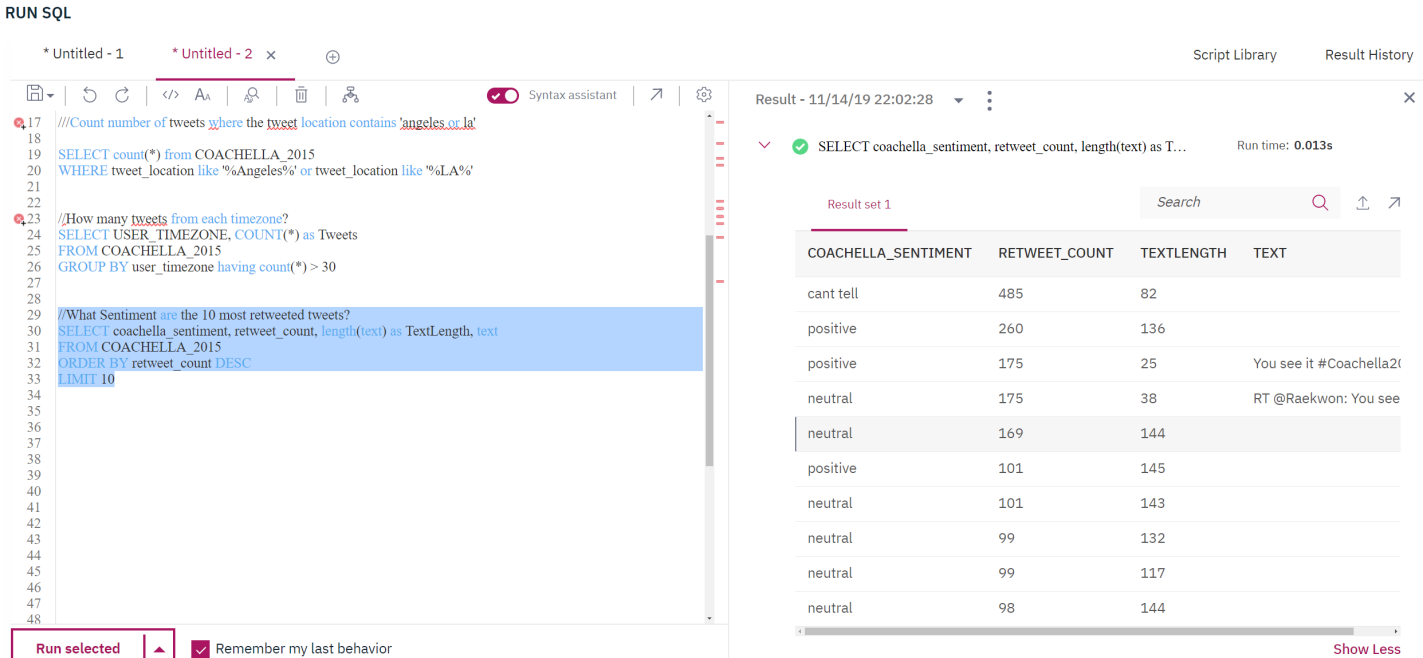
* 1. **Review of User Timezones**

The below query reviews for the number of tweets from each user Timezone with a minimum of 31 tweets. The output shows some results that are expected such as the bulk of the tweets coming from continental US time zones. However, it is interesting to see that Hawaii, London, Quito and Amsterdam also appeared in the list. This indicates that a small but significant number of people are interested in Coachella 2015 despite not being currently in the United States. These users are likely wealthy and anticipate having enough money to travel for the festival. These users may provide insight into the customer segment of the more wealthy attendees who will travel internationally.



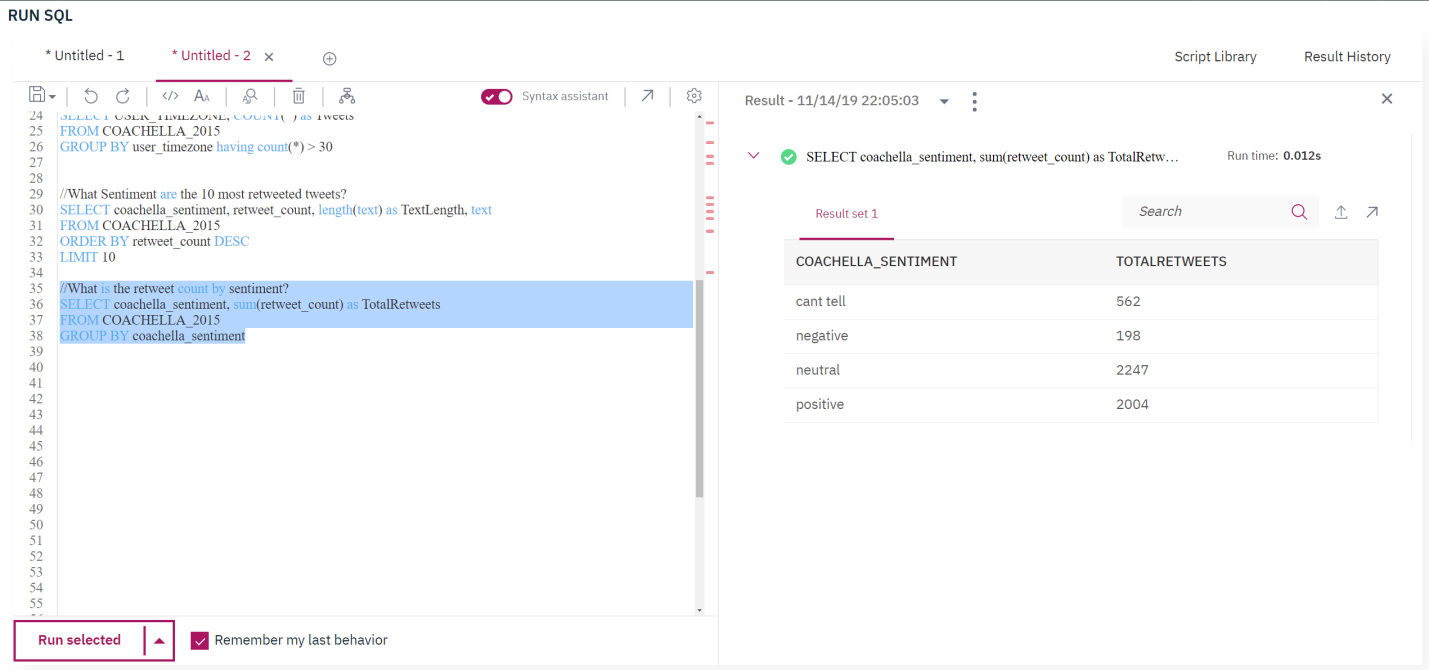
* 1. **Review of Most Retweeted Tweets**

The below query reviews for the text of the most retweeted tweets to allow for review of the content. Tweets with high retweet counts will be much more visible and influential than other tweets and provides valuable insight into what people are reading regarding Coachella. Unfortunately the IBM Web Console appeared to have difficulty displaying the text. The query was rerun with the length of the text appeared as ‘Text Length’ which verified that the tweets did have characters loaded and that the SQL editor was not properly showing them. However, the top ten retweeted tweets are all neutral or positive sentiment.



* 1. **Review of Total Retweets by Sentiment**

As retweets are more important than raw number of tweets in visibility to twitter, the number of retweets by each sentiment were reviewed. The output shows that most retweets were of neutral or positive tweets with only about 4% being for negative tweets. This appears to be very encouraging showing a positive reaction to the lineup announcement.



1. **Analysis of Tweet Data in R Studio**

**Verification of Proper Load in R Studio**

The data was loaded into IBM hosted R Studio and then verified using a header of the data, the number of rows, load status, and number of columns. The data was also manually reviewed using the R Studio viewer. See the R script for more details.

* 1. **Frequency Counts and Plots to Support Social Media Engagement Analysis**

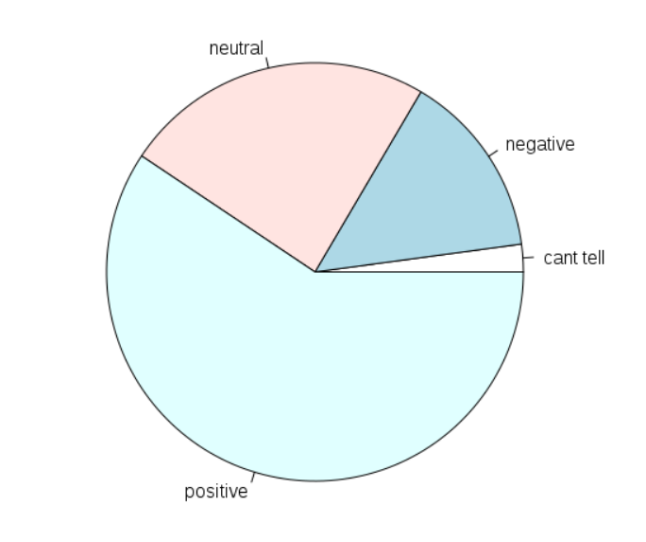
Analysis was conducted to create frequency counts, table and charts for the purpose of creating materials to provide to the Coachella 2015 social media engagement team to inform on the reaction to the headliner announcement for Coachella 2015 on January 6 and January 7, 2015.

**(a1) Review of Tweets by Sentiment**

To prepare statistics to share with the social media team, Table 1 and Figure 1 below were prepared to show the number of Tweets by each sentiment. The output shows the total number of Tweets by each sentiment as well as showing visually the proportion of each type. Visual representations of information can be very important to quickly allow comprehension by people who are not trained in analytics. From this analysis it shows that the reaction to the headliner announcement for Coachella 2015 was extremely positive.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 1. Number of Tweets by Sentiment | | | |
| Cant Tell | Negative | Neutral | Positive |
| 82 | 553 | 928 | 2283 |

Figure 1. Pie Chart of Tweet Sentiments

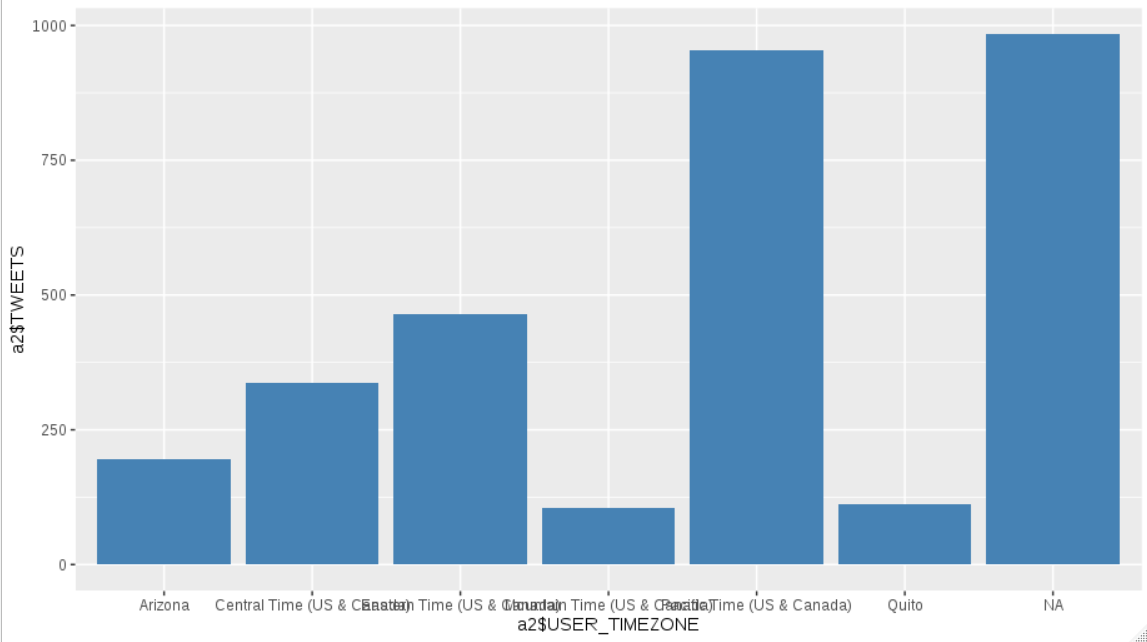
****

**(a2) Review of Tweets by Timezone**

The location of twitter users who are engaged with Coachella in social media can be valuable information to social media teams. The Table 2 and Figure 2 below show the number of tweets by each Timezone with other 100 tweets in reaction to the Coachella 2015 lineup announcement. The timezone with the most engagement was Pacific US which is expected as that is the most populous region close to the festival, although as expected Central, Mountain, Arizona and Eastern time also have significant numbers of Tweets related to this event. Interestingly Quito, which is a time-zone in Ecuador also has 112 tweets. The social media engagement team might be interested in following up on these users with further research.



Figure 2. Bar Chart of Tweets by Timezone

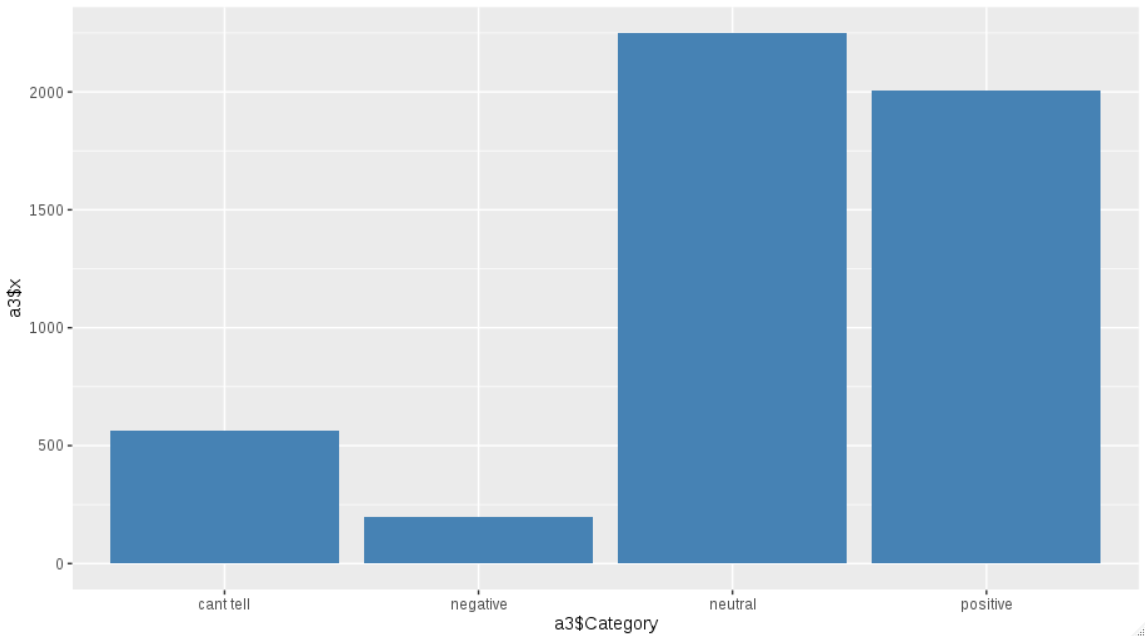
****

**(a3) Review of Retweets by Sentiment**

Tweets with a high number of retweets are usually from influential posters or with a popular or interesting message. It is therefore important for Social Media teams to review which twitter posts are getting the most retweets as these are the twitter posts with the most visibility to customers. The number of retweets by sentiment were pulled, and provided in Table 3 and Figure 3 below. They show that by far the most retweets are on neutral or positive sentiment tweets, and very few retweets for negative sentiment. This appears to be another indicator that the reception to the headliners is positive. Again, the numbers are shown in both numeric and visual form to aid in comprehension.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3. Number of Re-Tweets by Sentiment | | | |
| Cant Tell | Negative | Neutral | Positive |
| 562 | 198 | 2247 | 2004 |

Figure 3. Number of Retweets by Sentiment

****

**(a4) Review of Sentiment by Headliner.**

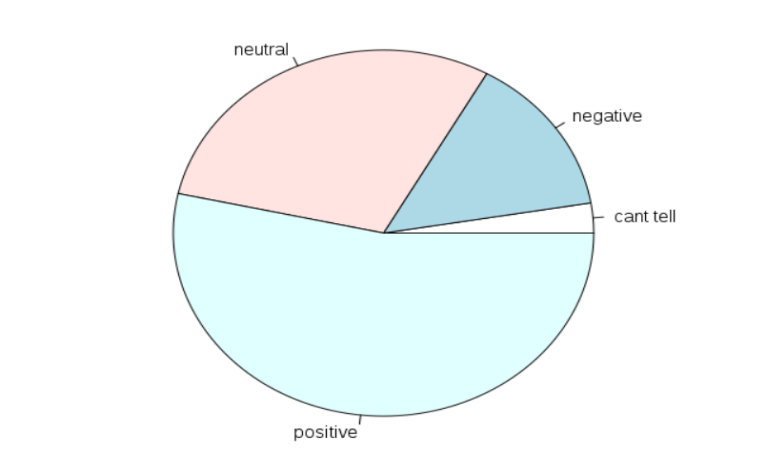
The headliners that were announced were the band AC/DC and the musicians Jack White and Drake. Understanding the reaction to each headliner is of extreme importance as they are the main draw to the event. Understanding if there is an unexpected reaction to each headliner can help avoid or respond to controversy or even help with negotiation.

**(a4 Part 1) – Review of Sentiment to Drake**

The number of twitter posts containing ‘Drake’ or ‘drake’ was pulled, and then analyzed. The total number of twitter posts was 303, with 14% of posts with a negative sentiment and 53% of posts with a positive sentiment. This appears to show a high engagement for Drake and a positive reaction. This is as expected as Drake is generally known to be a popular musician. The breakdown of each twitter post by sentiment is in Table 4 and Figure 4 below.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 4. Number of Tweets containing 'Drake' by Sentiment | | | |
| Cant Tell | Negative | Neutral | Positive |
| 8 | 43 | 90 | 162 |

Figure 4. Pie Chart of Sentiment of Drake Twitter Posts

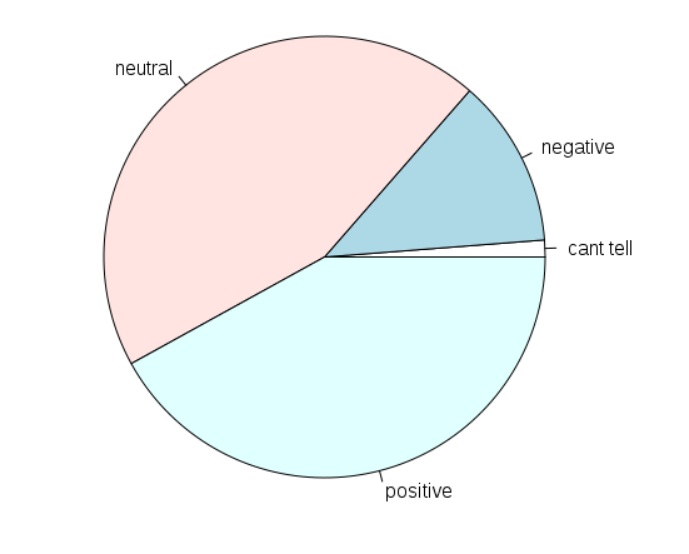
****

**(a4 Part 2) – Review of Sentiment to Jack White**

The number of twitter posts containing ‘White’ or ‘white’ was pulled, and then analyzed under the assumption that the post related to Jack White. The total number of twitter posts was 81, with 12.3% of posts with a negative sentiment and 42% of posts with a positive sentiment. This shows a much smaller reaction to Jack White than there was to Drake, assuming that similar proportions of posts relating to them were captured. This would be expected as he is generally not considered as popular as Drake. The breakdown of each twitter post by sentiment is provided in Table 5 and Figure 5 below.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 5. Number of Tweets containing 'Jack White' by Sentiment | | | |
| Cant Tell | Negative | Neutral | Positive |
| 1 | 10 | 36 | 34 |

Figure 5. Pie Chart of Sentiment of Jack White Twitter Posts

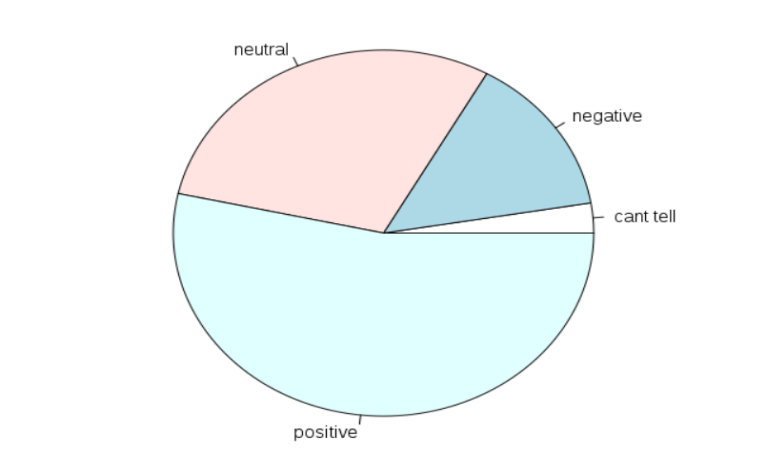
****

**(a4 Part 3) – Review of Sentiment to AC/DC**

The number of twitter posts containing ‘AC/DC’ or ‘AC/DC’ or ‘acdc’ or ‘ACDC’ was pulled, and then analyzed under the assumption that the post was related to AC/DC. The total number of twitter posts was 220, with 17.7% of posts with a negative sentiment and 50.5% of posts with a positive sentiment. This shows that the reaction to AC/DC appears to be bigger than Jack White, but less than Drake assuming similar proportion of twitter posts were captured. The breakdown of each twitter post by sentiment is provided in Table 6 and Figure 6 below.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 6. Number of Tweets containing AC/DC' by Sentiment | | | |
| Cant Tell | Negative | Neutral | Positive |
| 3 | 39 | 67 | 111 |

Figure 6. Pie Chart of Sentiment to ACDC

****

After reviewing each of the headliners, it appears that the reaction to Drake was the biggest, followed by AC/DC, followed by Jack White and that the reaction was generally positive for each of them.

* 1. **Text Mining Analysis of Tweets**

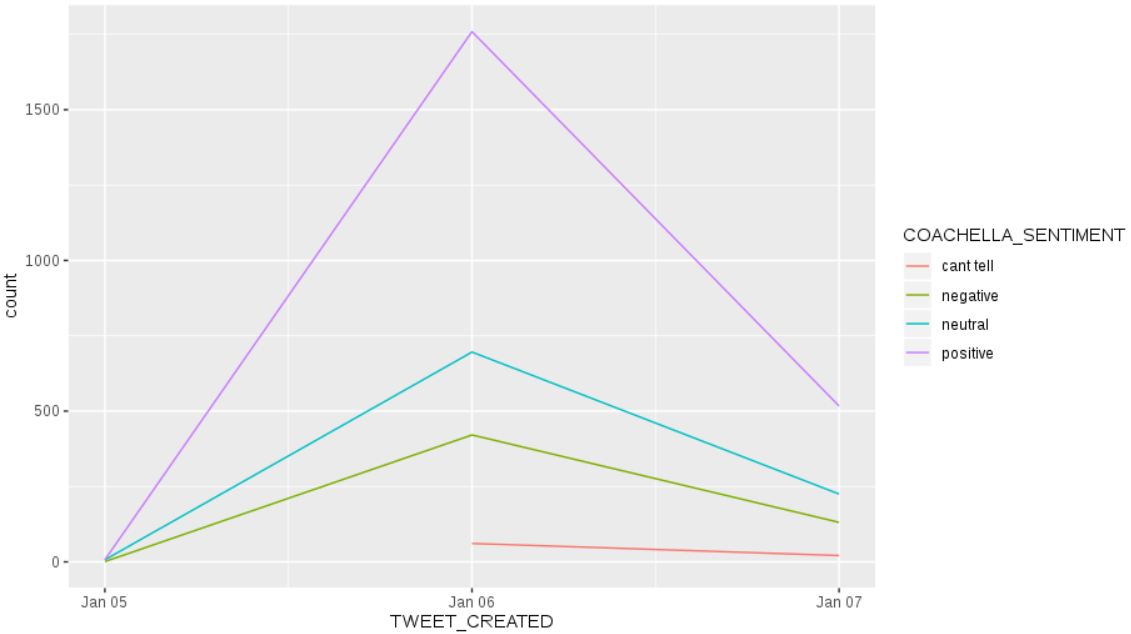
Various text mining analysis was done for the purpose of creating materials to provide additional analysis and insight to the Coachella 2015 social media team on the reaction to the announcement of the headliners for Coachella 2015.

Positive tweets had their text preprocessed for text mining. The process included removing invalid characters, removing URLs, removing punctuations, removing numbers, making the text all lower cases, removing stopwords of the ‘english’ and ‘smart’ lists, and then stemming the document which attempt to convert words with different endings but the same stem into just the stem. Then a document term matrix was created which is a matrix of each remaining word and the number of times it appeared per tweet.

**(b1) Review of Trend Analysis**

Trend Analysis, or reviewing how the sentiment changes overtime, is a classic tool for social media analysis. The sentiment for each day was plotted and is shown below. The chart shows that the most number of tweets were on January 6, with a smaller amount of January 7th. The proportion of sentiment also appears to remain fairly constant. This indicates that the reception was similar and there appears to be no unexpected backlash on the second day as sometimes can occur in the fast moving world of social media. Trend Analysis can be a valuable tool to present visually this information to Social Media teams.

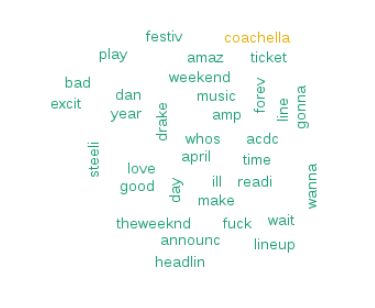
Figure 7. Trend Analysis of Sentiment for Tweets

****

**(b2) Word Cloud**

Word clouds are visual representations of the most frequent words in a corpus. Word clouds can be a quick visualization of what words are repeated most frequently, but have the problem that there is no information given as to what context the words are appearing in. They are best suited for initial exploration to show where to look next, or when the corpus is already well understood and the context can be inferred by the subject matter expert reviewing it. The word cloud below was prepared by removing sparse terms with a setting of 60%, and a minimum word frequency of 35. It is important to remember that this corpus only contains words from tweets identified as having a positive sentiment. The most frequent terms appear to be as expected including Coachella, ‘amaz’ the stem of ‘amazing, ticket relating to discussion about ticket sales, and line up. Interestingly while as expected ACDC and Drake appear, the other headliner Jack White does not appear in any form. However, ‘theweekend’ and ‘weekend’ both show up twice indicating a high level of excitement for that band.

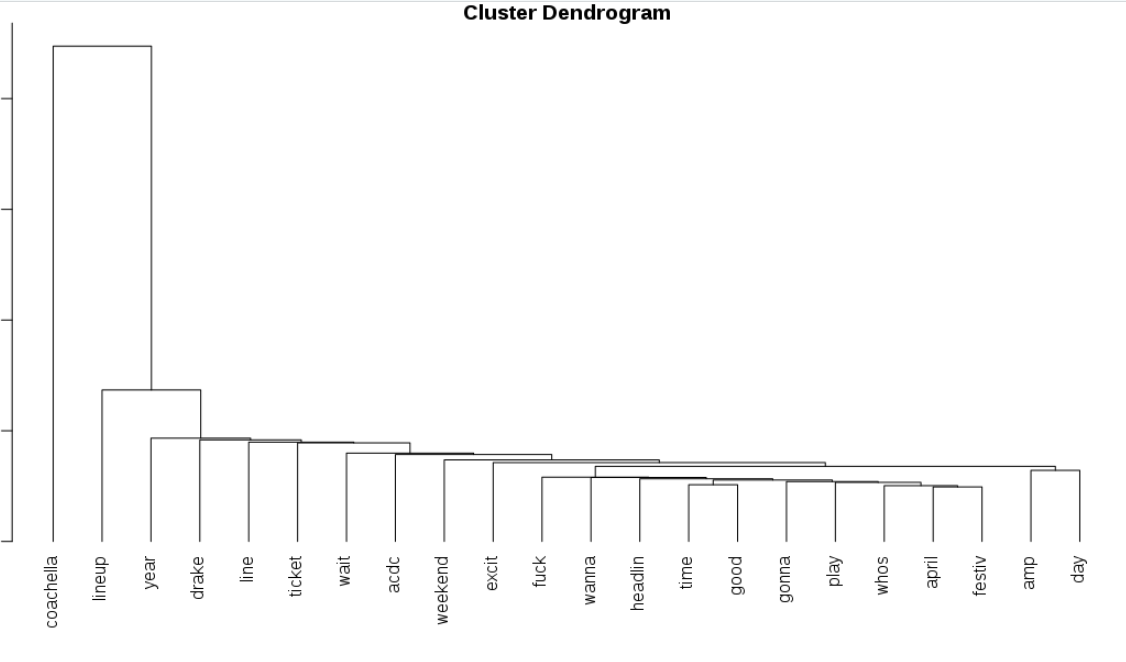
Figure8. Word Cloud of Most Frequent Terms in Positive Tweets

****

**(b3) Dendogram**

A dendogram provides a ‘tree’ like structure to show which words appear with other words. It can provide a useful visual to show context for frequently appearing words, although can be more difficult to interpret visually than other methods especially for people not trained in analytics or familiar with the method. The below dendogram was generated by removing sparse terms with a 98% setting. The dendogram shows that as expected the words which appear together in positive tweets are generally terms showing positivity and the bands the tweets are excited about. April shows up in this dendogram as Coachella 2015 was held in April. Presumably there is a common theme about the difficulty of waiting until April.

Figure 9. Dendogram of Positive Tweets

****

* 1. **Additional Analysis**

Additional analysis was conducted to research avenues that may provide value to a social media team attempting to react to the head liner announcement. These queries are designed to be more detailed than the information provided above, ideally provided as a supplement for any analysts on the social media team who would like to dive in deeper into the data. As such, the goal was to provide more information even when it may be more difficult to interpret.

**(c1) Provide Frequent Terms to Social Media Team**

All terms which appeared more than 50 times in tweets identified as positive were provided in the below table. As expected, these terms are related to the head liners and general excitement.

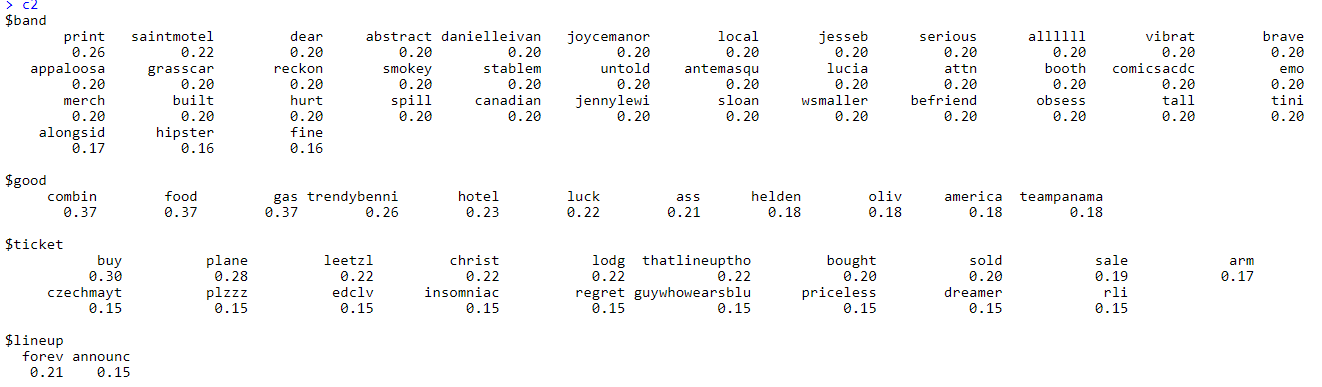
Table 6. Most Frequent Words in Positive Tweets

****

**(c2) Provide Words Commonly Associated with Key Ideas**

As the purpose of this analysis is to provide value about the reaction to the headliner announcement, the below table shows all words correlated to band, good, ticket, and lineup respectively with the correlation limit set to .15. From this analysis it is shown some combinations that may warrant follow up such as band and local indicating possible interest in local bands, and ticket and Christ which is an unexpected combination. There are several other combinations as well with unexpected words which are likely related to hashtags.

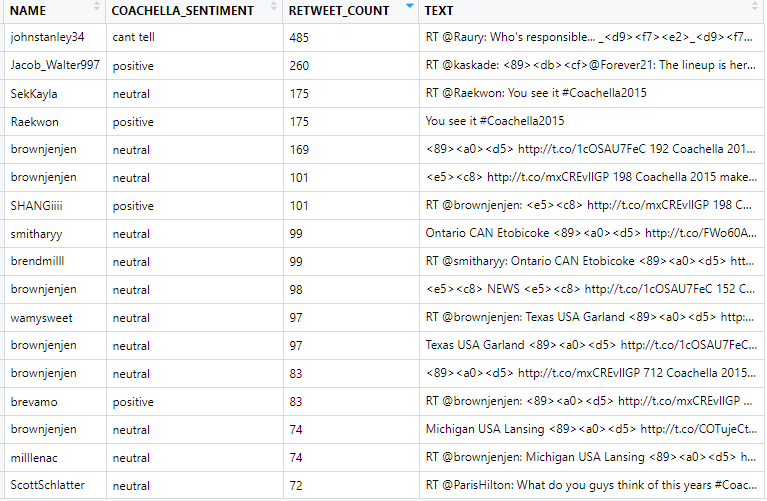
Table 7. Words Associated With Key Ideas

****

**(c3) Provide the Tweets Most Retweeted**

To efficiently invest the time of social media team-members, the Tweets with the highest engagement as measured by retweets is provided below. The social media team can review them and respond as appropriately to address questions or address concerns that may have been raised before irreparable reputational harm is caused. Ideally, this information would be provided hourly so the social media team can respond in a timely fashion with links directly to the post on twitter to assist the team with their use of time.

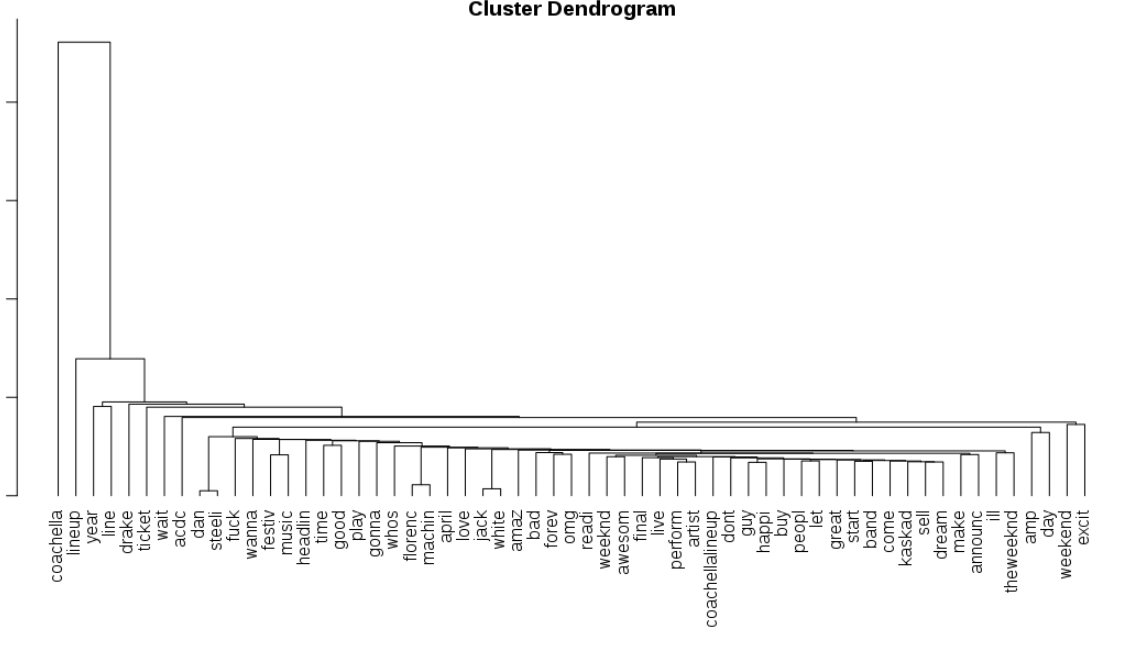
Table 8. Most Retweeted Coachella Announcement Tweets

****

**(c4) Expanded Dendogram**

An expanded Dendogram was produced with the sparse terms argument set to 99%, which results in a denser tree. The tree is more visually complex, but provides more context than the simpler tree originally produced does not. Some new word combinations that appear include ‘weekend’ and ‘excit’ short for excitement or exciting, ‘dan’ and ‘steeli’, ‘florenc’ and ‘machine’, and ‘ill’ and ‘theweekend’. These combinations can be explored and followed up on by the social media team if deemed interesting.

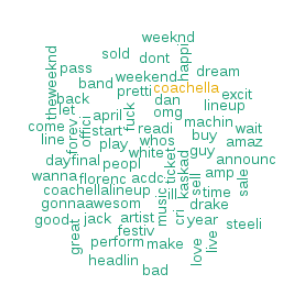
Figure 10. Expanded Dendogram of Positive Tweets

****

**(c5) Expanded Word Cloud**

An expanded word cloud was produced with the sparse terms set to max 80% empty space, and the minimum word frequency lowered to 20. The corpus was again of only from positive tweets. This creates a word cloud with more terms than the one originally produced. This word cloud can provide inspiration into new hashtags to follow up on.

Figure 11. Expanded Word Cloud of Positive Tweets

****

**Conclusion - Study Accomplishments and Limitations**

The study accomplished providing analysis of the line up announcement, as well as the reaction to each specific headliner which was announced. Additionally, intelligence was provided to the social media of specific items to follow up on such as unexpected word combinations, hashtags, and the most retweeted tweets. Value was provided both through high level visual aids as well as more detailed tables and more detailed figures.

However, the study had limitations which may have impacted the effectiveness of the study. Based on manual review, tweets may be misassigned in sentiment. The time period reviewed is limited at only two days. The twitter data provided is not standardized in terms of the user location field. Some analysis regarding terms may be incomplete as user entered text is inconsistent and searches for terms may not have captured different versions of the terms.

Further research could be done to improve the reliability and scope of the study. A look at tweets over a longer time period, more detailed analysis and term searches for each headliner, expanding the corpus to include tweets including just ‘#Coachella’ instead of only including tweets with the hashtag ‘#Coachella2015 could all benefit the study. However, the most urgent alteration would be a deeper review to adjust the sentiment calculation for the purpose of reviewing words used in relationship to reacting to popular culture. For example, when talking about popular media the word ‘ill’ or expletives are much more likely to be a positive sentiment than when talking about other subjects.

**References**

Coachella's 2015 lineup revealed. (2015, January 6). Retrieved from https://consequenceofsound.net/2015/01/coachellas-2015-lineup-revealed/#:~:targetText=Coachella has revealed its 2015,-J, and The Weeknd.

O'Leary, A. (2019, April 12). The annual Coachella festival has a pretty unexpected back story. Retrieved from https://www.mirror.co.uk/3am/celebrity-news/what-coachella-full-run-down-7785765.