What are Twitter Users Talking about Starbucks?

Sentiment And Network Analysis of Starbucks
Section C, Team C49
Team Members: Cara You, Juan Brando, Pritisha Punukollu, Julie Liu, Zhicheng Zhang
Date: 10/12/2019

Sentiment And Network Analysis of Starbucks

Part 1 Business Objective

Sentiment analysis is very useful in social media analysis, as it helps us develop a deeper understanding of the public opinion of a brand, product, or even hot topics in society. It is also widely recognized that consumers' opinions, experience, and feedback are vital for companies to attract new customers, retain current customers and build customer loyalty. In our project, we are interested in how Twitter users think about the world's largest coffee company: Starbucks. By using sentiment and network analysis, we will try to answer the questions below:

- What kind of topics related to Starbucks are twitter users talking about?
- What are consumers' attitudes towards Starbucks?
- Are there some complaints from customers? If so, what are they?
- How can Starbucks improve its customer services?
- Are there some marketing opportunities on twitter?

To answer these questions, we extract 1200 tweets using Twitter API through R and then perform the analysis, visualize results and draw conclusions.

Part 2 Sentiment Analysis

Data Cleaning and Preparation

As for sentiment analysis, we first clean our data following the steps below:

1. clean the text of special characters such as symbols and emoticons
2. convert data to lower case for analysis
3. remove punctuations
4. remove numbers
5. remove common words as they do not add any informational value
6. remove URLs (https://etc.)
7. remove white spaces
8. clean keyword "starbucks" from the text

Create Term Document Matrix

After cleaning the data, we create a term document matrix to record every document and their corresponding terms. The first 10 rows of 20 tweets are shown below:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
amazon	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
card	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
cash	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
enter	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
fall	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
gift	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
hello	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
paypal	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
want	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
win	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0

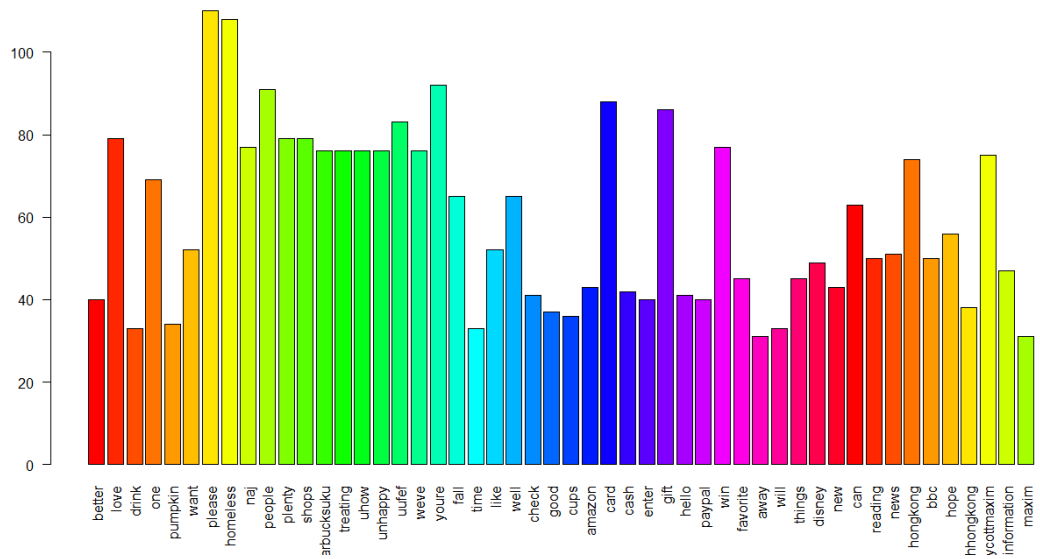
Term Document Matrix (10 rows, 20 documents)

Next, we will visualize this term document matrix to obtain more insights from it.

Data Visualization

1) Bar Plot

Since there is a large number of terms and many of them show up with a low frequency, we create a subset containing only terms showing up more than 30 times. And then we draw a map of the terms:



Distribution of Terms ("starbucks" removed)

We can see that "please" and "homeless" are one of the most frequent terms, which is related to the event that happened in the UK recently: a Starbucks barista tries to boot a homeless man. "Gift" and "card" also go together to be another group of hot terms, which is related to the recent Starbucks and Amazon gift card campaign. "Hong Kong" and "Boyycottmaxium" are related to the protest in Starbucks stores in Hong Kong recently.

2) Word Cloud

Word cloud is a more direct way to show the frequency of words. We only keep the terms that appears more than 5 times and we rotate some words for aesthetic purposes. The result is shown below:

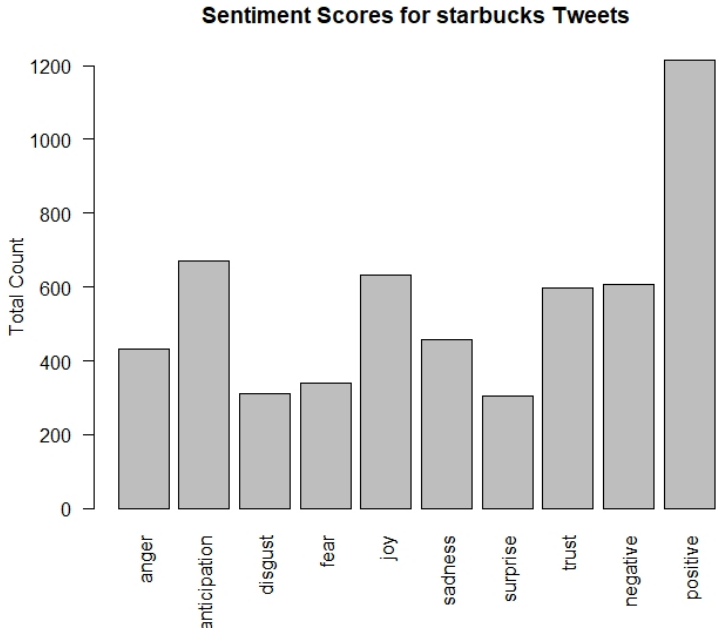
3) Summary

According to the barplot and word cloud, the most common topics related to Starbucks that people are talking about on Twitter recently are:

1. “homeless”: People talk a lot about this because there is a video showing that a Starbucks barista in England trying to boot a homeless man as he chowed down on a meal purchased by a kindhearted bystander outside the coffee chain. This event may generate some negative effects on the company’s brand image.
2. “hongkong”: People mention “Hong Kong” because many Starbucks stores were attacked in Hong Kong by some protesters.
3. “disney”: Starbucks and Disney cooperate together to sell Disney-themed merchandise.
4. “amazon”: Amazon is mentioned along with Starbucks because there are lots of people on twitter reposting the ads to win gift cards of Amazon and Starbucks. The tweet is:
   “Want to #win $50 in an #Amazon gift card, #Starbucks gift card, or #Paypal cash? Enter to win the Hello Fall #giveaway at @SouthernMomLove!”

Sentiment Analysis Results and Explanation

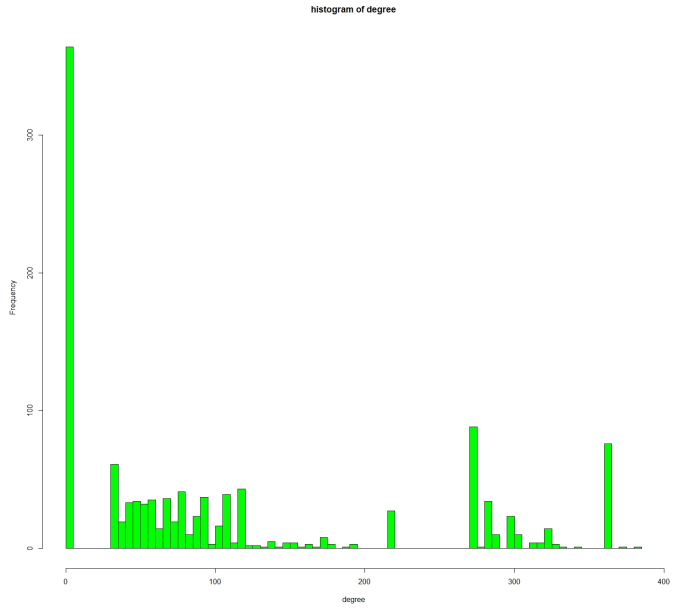
In this part, we use nrc_sentiment dictionary to obtain sentiment scores for each 1200 tweets. There are 8 emotions contained in the dictionary.



In this bar plot, we can tell that positive emotions (anticipation, trust, joy, etc.) are more than negative ones (anger, disgust, fear, sadness), which indicates a relatively good consumer perception.

Part 3 Social Network Analysis

Degree Histogram

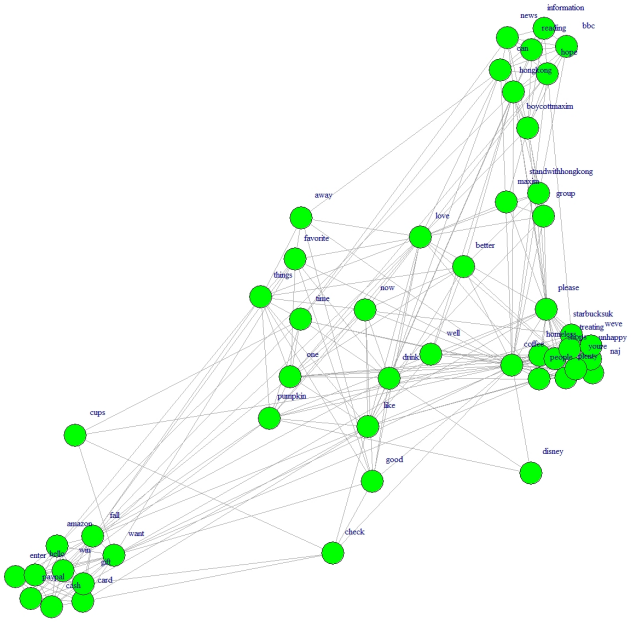


From the histogram above, we can tell that the degree's frequency is unevenly distributed. Specifically, the histogram is right skewed. This means that a majority of the tweets have a smaller degree. Also, many terms have 0 degrees, which means

they have no connection with other terms. Meanwhile, there are also many terms that have about 100 and 270 degrees.

Network Diagrams

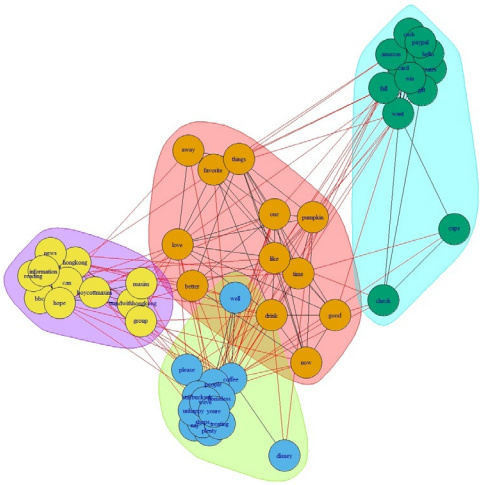
It is hard to interpret only by histogram, so we use more explicit diagrams to show the network.



Original Network Diagram

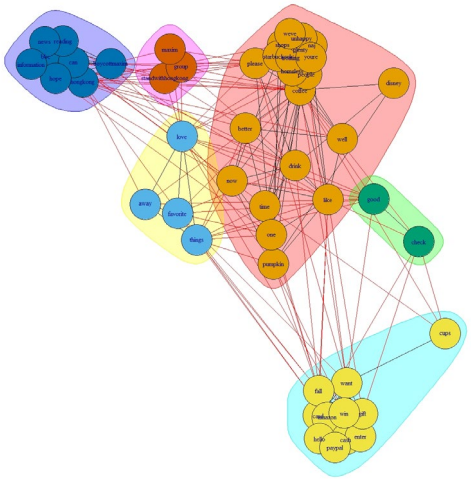
This is a network diagram about terms that people use in each document. Here, we have the same four clusters: homeless, Boycott, Disney, and gift cards, as we analyzed previously.

We also want to create communities to cluster terms and make their features more clear:



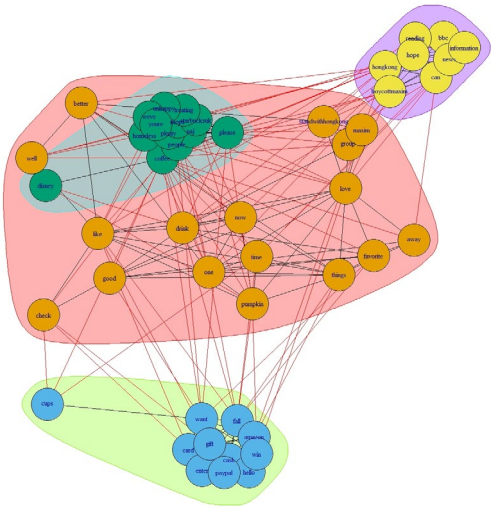
Network Diagram with Community Creation

The community creation plot is based on edge betweenness. For 4 different dense areas, we have 4 different clusters.



Network Diagram by using propagating lab

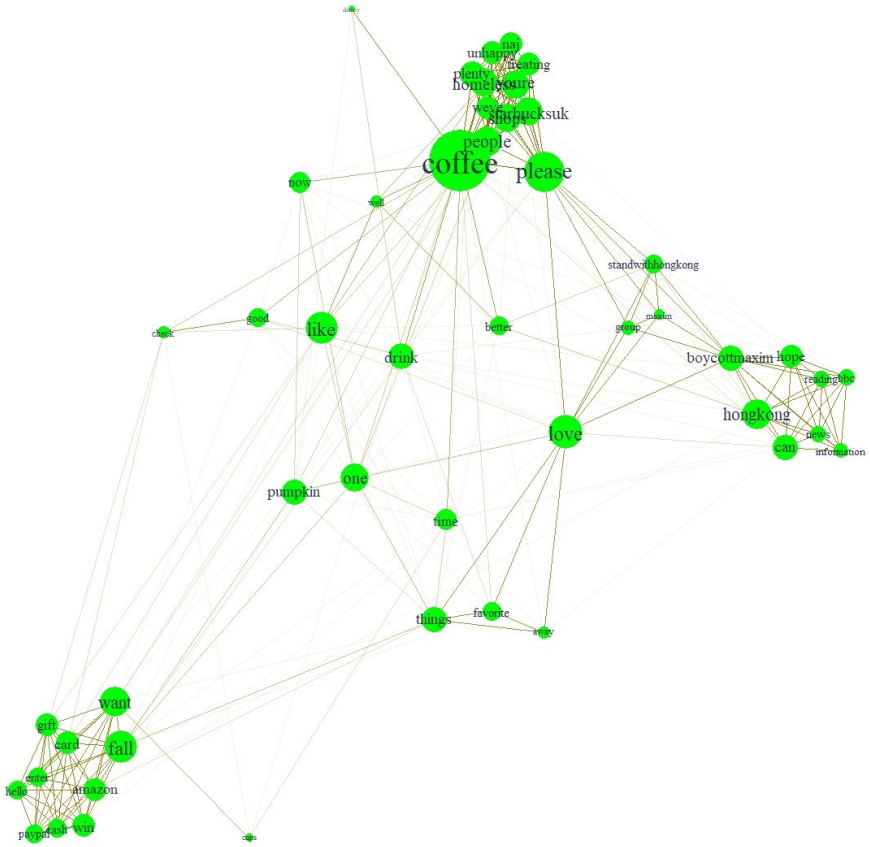
We then plot Label Propagation which re-assigns labels to nodes that each node takes the most frequent label of its neighbors, but with more clusters compared to the community creation plot.



Network Diagram by using greedy algorithm

Next, we plot greedy fashion graph, which merges clusters together as label propagation but with more frequent labels of its neighbors in each cluster compared to label propagation.

By using size to represent degrees, we can also draw a map like:



This takes into consideration the time that each term appears in the document.

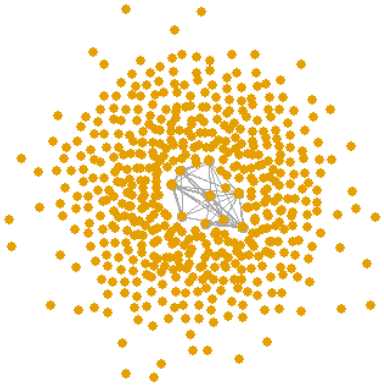
Summary

From the diagrams above, we can tell that:

1. Term “homeless” tends to appear with some negative words like “unhappy”, which means customers may be unsatisfied with the homeless news mentioned above.
2. Term “hongkong” is closely related to terms like “boycott”, “hope”, “bbc”, “news”, which are closely related to the protesting event in Hong Kong.
3. “amazon” always show up with “gift”, “card”, “win”, which verify the marketing activity mentioned above in the Sentiment Analysis part.
4. “disney” is discussed with “well”, “drink”, which may represent that Starbucks’ cooperation with Disney is successful.

Better Visualization

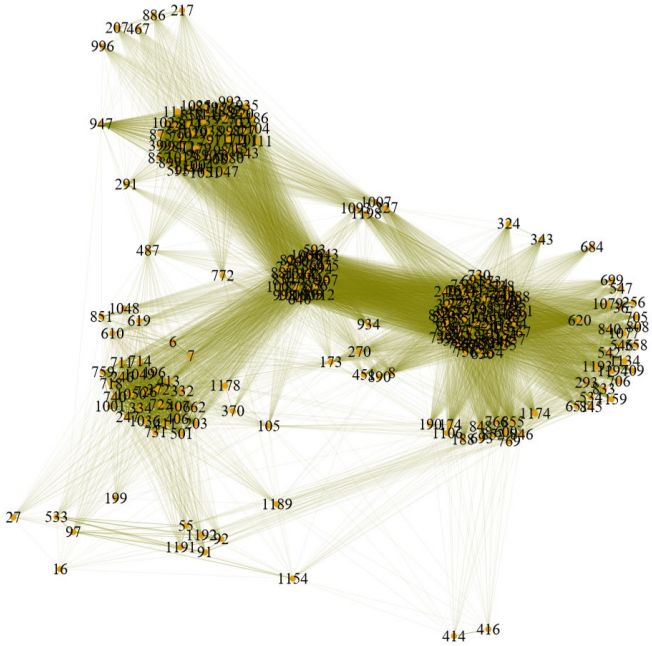
1. Remove Outliers



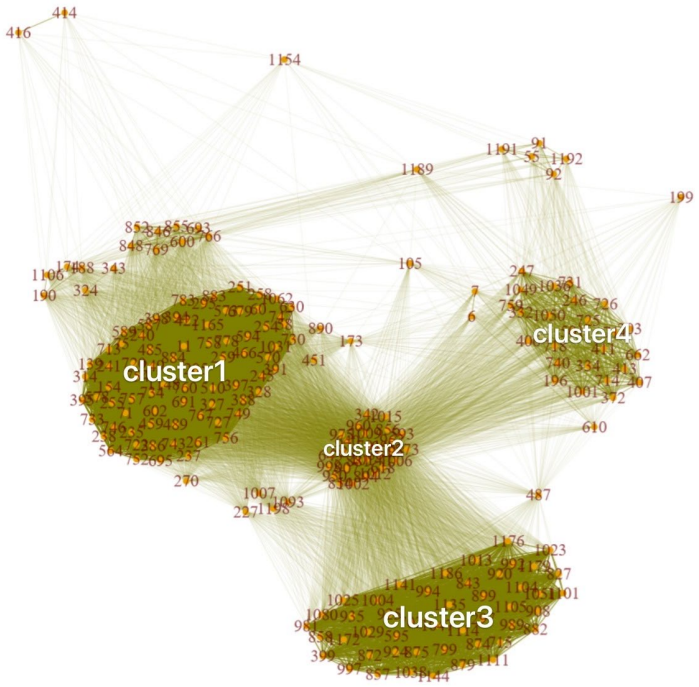
The graph shows that there are many outliers that have no connection with others (degree=0). So we remove these outliers and include only those that are more connected in the graph. The new graph is shown below:

2. Remove Less-connected terms

The graph on the right shows the terms that are connected to others. However, we want to improve the graph by only keeping those that have more than 2 edges. The new graph is shown below:



3. Final Network Graph



From the final network graph, we can tell that there are 4 closely-connected clusters. Here are some samples for each cluster:
Sample of Cluster1:
"I went to @starbucks this morning for my favorite #hotchocolate. I saw numerous people picking up their online & in https://t.co/PoCyhcHVou"
Sample of Cluster2:
"RT @_hilda1122_: Dear Starbucks, you deserve a better agent.\n#Boycott #HongKong #HK #Starbucks https://t.co/QfKhVNBmq0"
Sample of Cluster3:
"Best shopping combo #Disney #Starbucks https://t.co/HQjv3Uhkii"
Sample of Cluster4:
"@Starbucks Offers Free Digital News at US Stores for a Limited Time #starbucks https://t.co/PPPKXa6sMZ https://t.co/oow4Qlui8D"

Next, we also want to see what are these clusters are and how they contributed to the whole dataset's sentiment analysis outcome. To achieve this goal, we are going to do sentiment analysis for each cluster.

Cluster based Sentiment Analysis

We divided our 46 most frequent keywords into five main clusters as follows:
Cluster1:

```
"better"            "can"                "reading"
"news"              "hongkong"           "bbc"
"hope"              "standwithhongkong" "boycottmaxim"
"information"       "maxim"
```

Cluster2:

```
"love"     "drink"    "pumpkin"  "like"     "good"     "favorite"
"away"     "things"   "disney"   "new"
```

Cluster3:

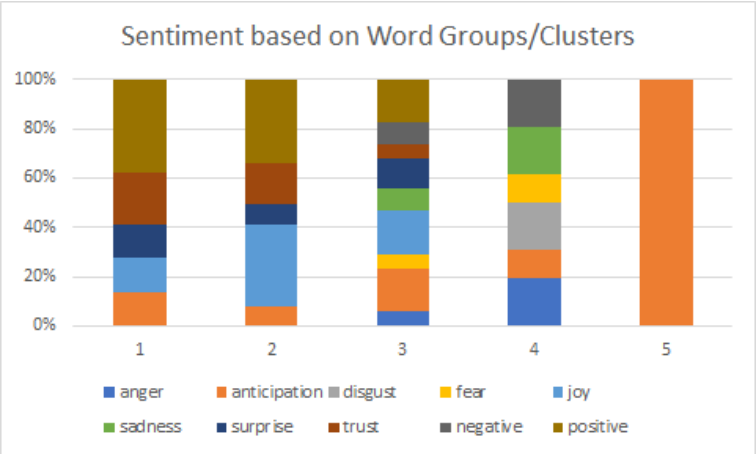
```
"want"   "fall"   "check"  "amazon" "card"   "cash"   "enter"  "gift"
"paypal" "win"
```

Cluster4:

```
"please"       "homeless"     "naj"          "people"
"plenty"       "shops"        "starbucksuku" "treating"
"uhow"         "unhappy"      "uufef"        "well"
```

Cluster5:

```
"time" "cups" "will"
```



We can see that not all the sentiments for all clusters are accurately represented by this method because it fails to assign enough sentiment to words like “boycottmaxim” and “standwithhongkong”. The reason may be that these words are not recognized by the sentiment analysis method. If they are captured in the right way, there should be more negative sentiments for cluster 1. The clusters with more descriptive words fare better. In particular we can see that with Cluster 4. This represents the buzz around the homeless man incident in the UK. There was an overwhelming outrage against the incident which is also reflected in the chart through anger”,”sadness”,”disgust”, and “negative”. Cluster 2 is a positive cluster which also is evident from its bar.

Overall, sentiment and network analysis seem to be fairly accurate in its account of all the buzz surrounding Starbuck on twitter recently.