615 Twitter Analysis Project about Nike and Adidas

Laura Lin

12/18/2017

Introduction

Nike and Adidas are two popular brands producing sport goods. Therefore, it may be interesting to see how people talk about these two brands in different ways. In this project, I will be analyzing the tweets using "nike"/"adidas" as hashtags. I will first compare their content. And I will see if the content of tweets related to a certain brand will change as time goes.

1 Set Up Server

```
api_key <-  "HnY07OC8D50vmeDMUuI740Ska"
api_secret <- "vSn8V6HmReeEsmpM79aBOYXad1F5HkitdyUuWVfmrpfVH8B7rx"
access_token <- "927639479191564293-d5ks0pqdx0TLVS2ra6CEUz7xjbCpPiu"
access_token_secret <- "KJrXp4HLFUeWRIphDbroZcyQ9qgULe20gw0uXXncwwVsJ"


setup_twitter_oauth(api_key, api_secret, access_token, access_token_secret)
```

```
## [1] "Using direct authentication"
```

```
## Warning in strptime(x, fmt, tz = "GMT"): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/America/New_York'
```

2 Acquire Data from Twitter

2.1 Get tweets with hashtag #nike and #adidas.

```
tweets1 <- searchTwitter("#nike", n=5000, lang="en")
```

```
## [1] "Rate limited .... blocking for a minute and retrying up to 119 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 118 times ..."
```

```
tweets2 <- searchTwitter("#adidas", n=5000, lang="en")
```

2.2 Transform the list of tweets into dataframe

```
tweets1.df <- twListToDF(tweets1)
tweets2.df <- twListToDF(tweets2)
```

3 Clean and Organize Tweets

3.1 Remove punctuations and numbers

```
df.text1 <- str_replace_all(tweets1.df$text, "[^[:alpha:]]", " ")
df.text2 <- str_replace_all(tweets2.df$text, "[^[:alpha:]]", " ")
```

3.2 Transform text into corpus

```
df.text1 <- Corpus(VectorSource(df.text1))
df.text2 <- Corpus(VectorSource(df.text2))
```

3.3 Convert the text to lower case

```
df.text1 <- tm_map(df.text1, content_transformer(tolower))
df.text2 <- tm_map(df.text2, content_transformer(tolower))
```

3.4 Remove english common stopwords

```
df.text1 <- tm_map(df.text1, removeWords, stopwords("english"))
df.text2 <- tm_map(df.text2, removeWords, stopwords("english"))
```

3.5 Remove own-defined stop word

These are the stop words that I defined by myself after manually going through the word list.

```
df.text1 <- tm_map(df.text1, removeWords, c("https","pic","took","jayteep","fzgsmh","select","check","v
df.text2 <- tm_map(df.text2, removeWords, c("https","giving","one","get","check","amp","official","want
```

3.6 Eliminate extra white spaces

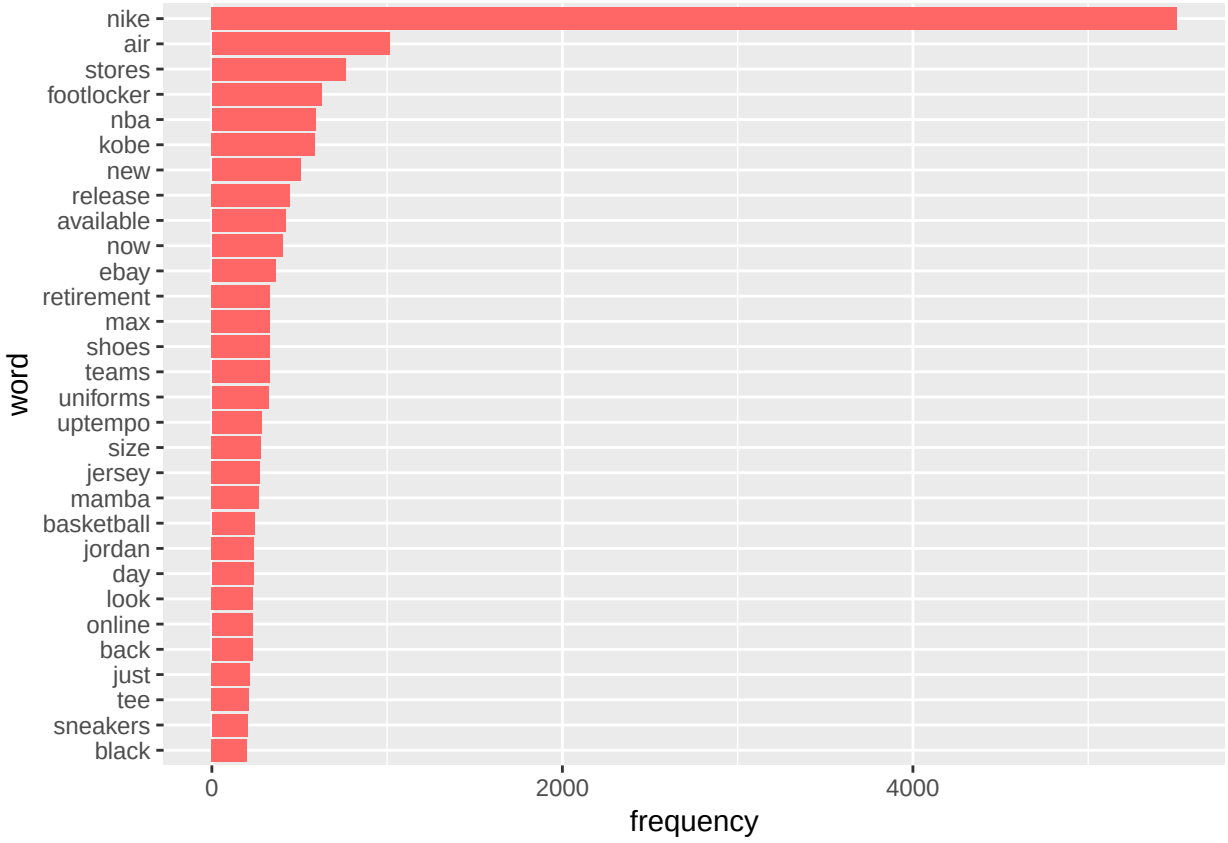
```
df.text1 <- tm_map(df.text1, stripWhitespace)
df.text2 <- tm_map(df.text2, stripWhitespace)
```

4 EDA

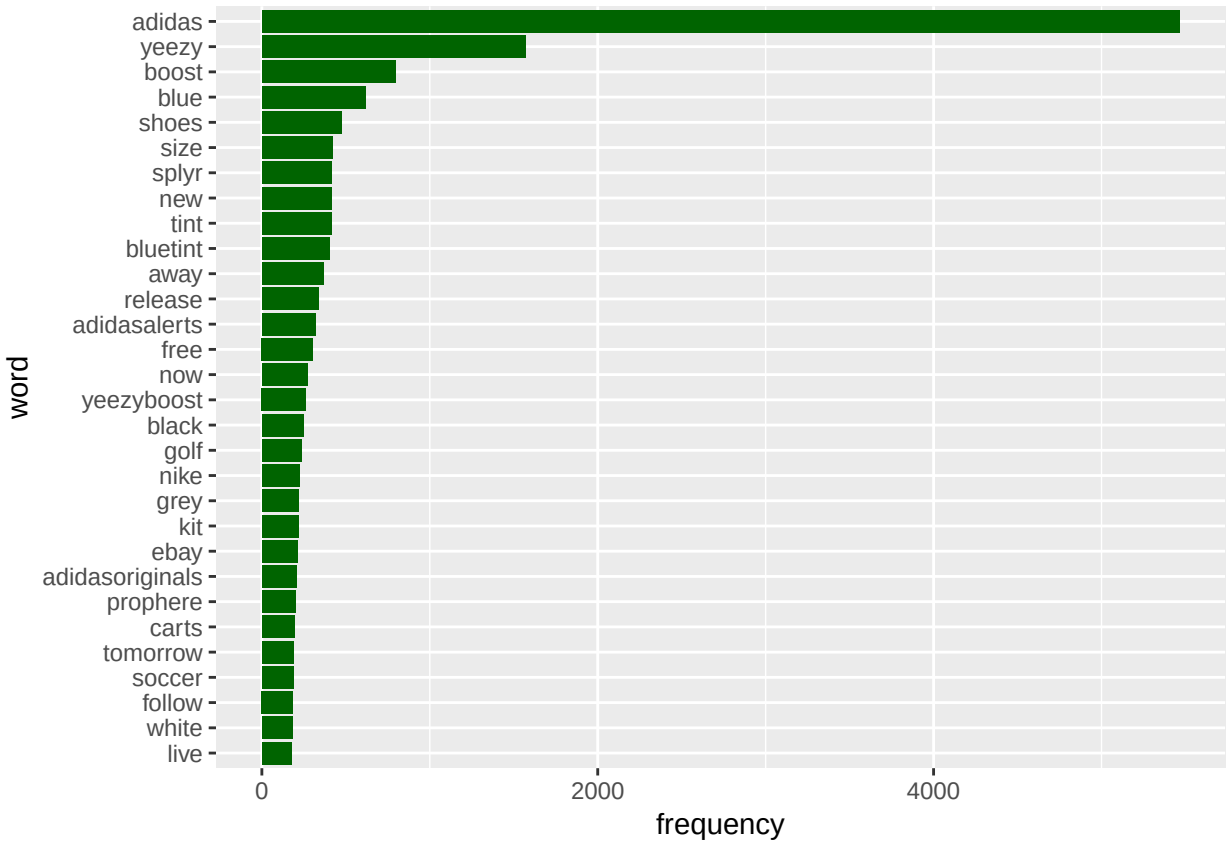
4.1 Word Frequency Visualization of Top 30 Most Frequent Words

```
dtm1 <- TermDocumentMatrix(df.text1)
m1 <- as.matrix(dtm1)
v1 <- sort(rowSums(m1),decreasing=TRUE)
d1 <- data.frame(word = names(v1),frequency=v1)
```

```
top1 <- head(d1,30)
top1$proportion <- top1$frequency/sum(top1$frequency)

dtm2 <- TermDocumentMatrix(df.text2)
m2 <- as.matrix(dtm2)
v2 <- sort(rowSums(m2),decreasing=TRUE)
d2 <- data.frame(word = names(v2),frequency=v2)
top2 <- head(d2, 30)
top2$proportion <- top2$frequency/sum(top2$frequency)

top1$word <- factor(top1$word, levels = top1$word[order(top1$frequency)])
ggplot(data=top1, aes(x=word, y=frequency)) + geom_bar(stat="identity",fill = "#FF6666") + coord_flip()
```




```
top2$word <- factor(top2$word, levels = top2$word[order(top2$frequency)])
ggplot(data=top2, aes(x=word, y=frequency)) + geom_bar(stat="identity",fill = "dark green") + coord_fli
```



4.2 Word Cloud

```
wordcloud(words = d1$word, freq = d1$frequency, min.freq = 100, scale = c(5, 1),
          max.words=500, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```


```
wordcloud(words = d2$word, freq = d2$frequency, min.freq = 100, scale = c(5, 1),
          max.words=500, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(12, "Paired"))
```


```
d1[d1$word == "adidas",]
```

```
##          word frequency
## adidas adidas       128
```

```
d2[d2$word == "nike",]
```

```
##      word frequency
```

```
## nike nike       226
```

Here we can see most of the words are the names of products.They are mostly about the shoes and sport shirts of these two brands, which match the types of their popular products.

An unexpected dicovery is that "nike" is the 20th most frequent word appearing in adidas-related tweets. However, "adidas" is only the 57th most frequent word appearing in nike-related tweets. This is interesting because it shows that people like mentioning nike when they tweet about adidas but not exactly the opposite. It means that they like to compare adidas with nike when they talk about adidas but won't really think about adidas while talking about nike. This could be an implication that nike is more of a representative brand of sport goods than adidas, which should be a useful discovery for both two companies.

5 Comparison of Nike and Adidas

5.1 Analysis of Common Words

```
new1 <- top1[top1$word == "new",]
now1 <- top1[top1$word == "now", ]
release1 <- top1[top1$word == "release",]
ebay1 <- top1[top1$word == "ebay", ]
common1 <- rbind(new1,now1,release1,ebay1)

new2 <- top2[top2$word == "new",]
now2 <- top2[top2$word == "now", ]
release2 <- top2[top2$word == "release",]
ebay2 <- top2[top2$word == "ebay", ]
common2 <- rbind(new2,now2,release2,ebay2)

data.frame(word=common1$word,pro_in_nike=common1$proportion,pro_in_adidas=common2$proportion)
```

```
##      word pro_in_nike pro_in_adidas
## 1     new  0.03118452    0.02633253
## 2     now  0.02483668    0.01717339
## 3 release  0.02748675    0.02143493
## 4    ebay  0.02224824    0.01342068
```

We can see that the words "new","now","release" and "ebay" are all in the list of top 30 most frequent words of the tweets using "nike" and "adidas" hashtags. From the first three words, we can get the information that people like tweeting on product release of nike and adidas. Generally speaking, the proportion of these three words are higher for nike than adidas.

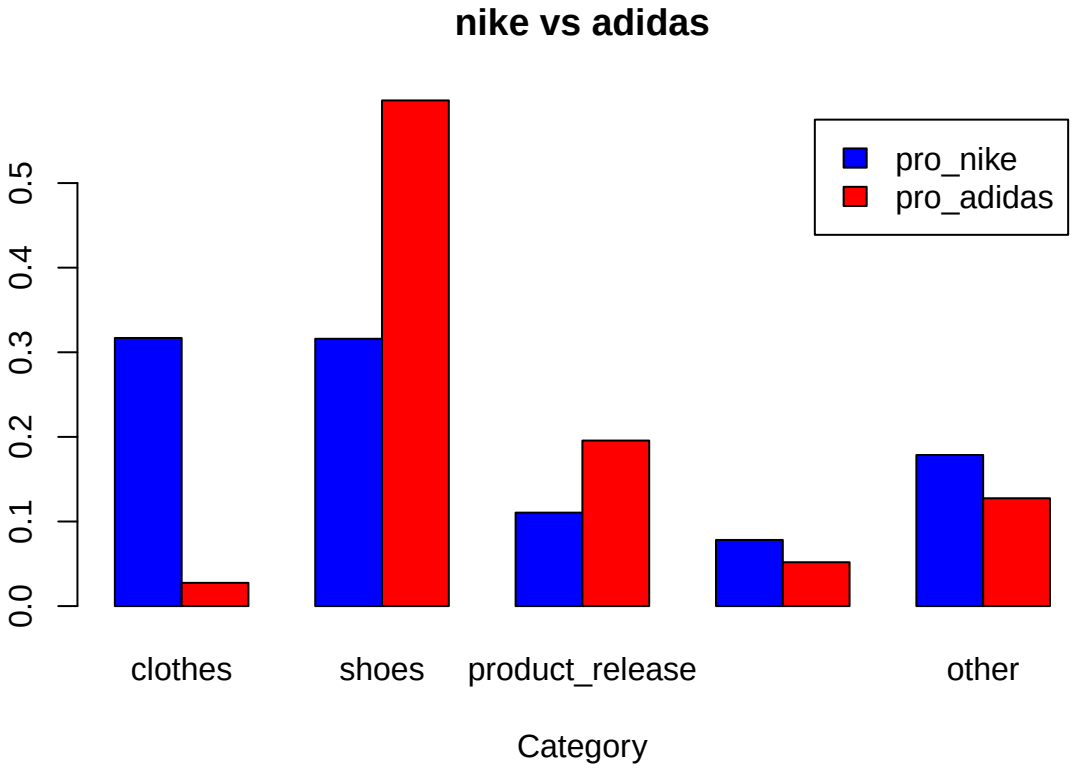
In addition to this,"ebay" is the 13th most frequent word appearing in nike-related tweets and is the 23th most frequent word appearing in adidas-related tweets. These are valuable information. From this we may guess that ebay is a commonly used platform for selling and buying nike and adidas products. The usage of it should be popular enough so people tweet a lot about it.

5.2 Difference of Focus of Tweets

After looking for the meaning of each of the top 30 most frequent words,I manually categorize the words into five categories: clothes,shoe,product release,third-party platform and other.

The shirt category contains words about the styles of clothes,for example "uniforms"" and "jersey". The shoe category contains words of the names of shoe such as "air","boost" and "tint". The product release category contains words like "new" and "release". The third-party platform category contains names of the third-party like "footlocker"" and "ebay". The last category-other contains more general words such as "online", "size" and "stores".

```
shirt1 <- sum(top1[c(2,4,5,20,22,26,27,28),c(1,2)]$frequency)
shoe1 <- sum(top1[c(3,8,14,15,16,18,21,23,25,29),c(1,2)]$frequency)
event1 <- sum(top1[c(9,10,11),c(1,2)]$frequency)
third1 <- sum(top1[c(7,13),c(1,2)]$frequency)
other1 <- sum(top1[c(6,12,17,19,24,30),c(1,2)]$frequency)
nike_diff <- c(shirt1,shoe1,event1,third1,other1)
nike_diff <- nike_diff/sum(nike_diff)

shirt2 <- sum(top2[15,c(1,2)]$frequency)
shoe2 <- sum(top2[c(2:7,11,18,19,22,27,29,30),c(1,2)]$frequency)
event2 <- sum(top2[c(8,10,12,13,16,26),c(1,2)]$frequency)
third2 <- sum(top2[c(14,23),c(1,2)]$frequency)
other2 <- sum(top2[c(9,17,21,25,28),c(1,2)]$frequency)
adidas_diff <- c(shirt2,shoe2,event2,third2,other2)
adidas_diff <- adidas_diff/sum(adidas_diff)

diff <- data.frame(word=c("clothes","shoes","product_release","third_party","other"),pro_nike=nike_diff

counts <- t(as.matrix(select(diff, pro_nike, pro_adidas)))
barplot(counts, names.arg = diff$word,main="nike vs adidas",
xlab="Category", col=c("blue","red"),
legend = rownames(counts), beside=TRUE)
```



As we can see, the distributions look very different for the two brands. For tweets on products, adidas-related tweets focus much more on shoes than clothes while the distributions of shoe and shirt look pretty even for Nike. Therefore, we may say that people who tweet about adidas are muchn more interetsed in discussing its shoes than its clothes. And the interest on clothes and shoes of people who tweet about nike are pretty much the same. From this we can find out which kind of products of these two brands is more popular in people's discussion.

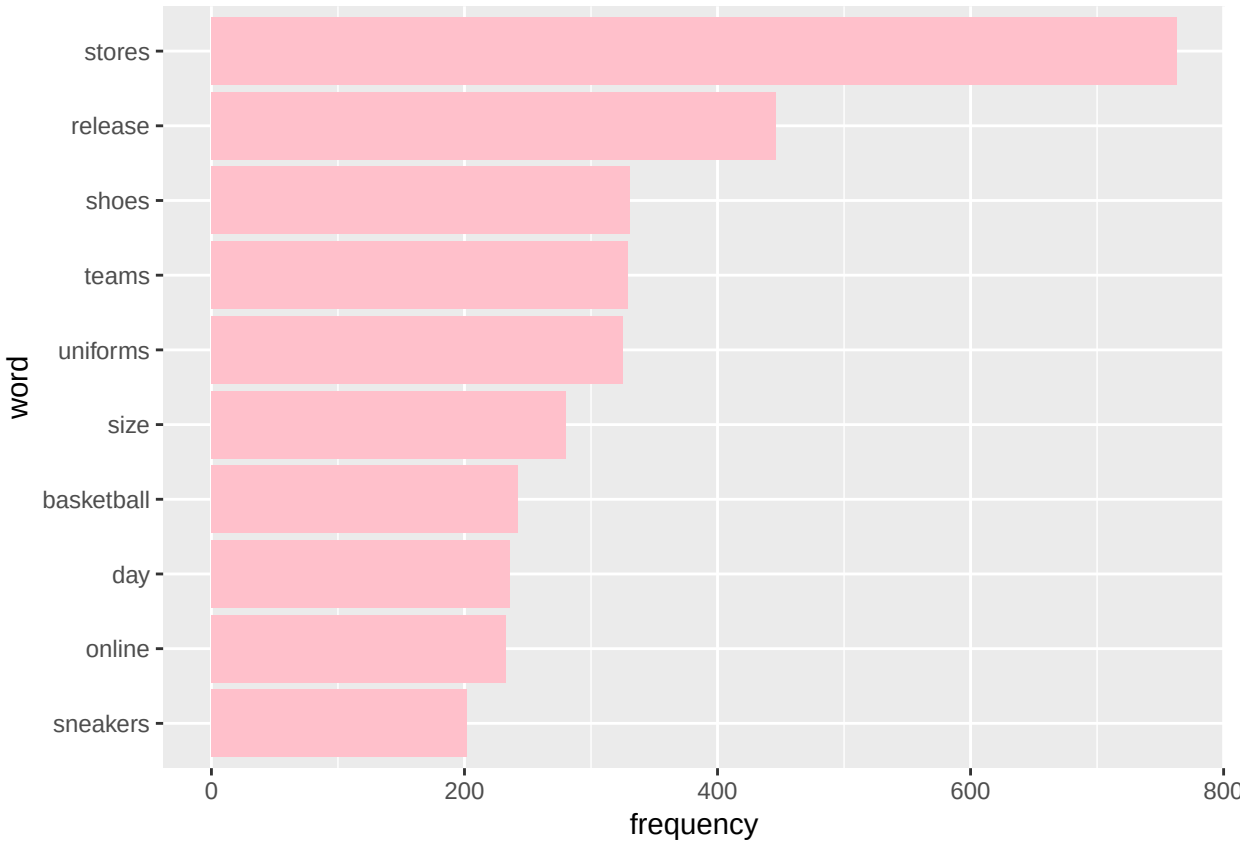
Also we can see that people who tweet about adidas care more about its product release than people who tweet about nike. From this we may infer that the product release of adidas attract more attention than nike's.

The proportion of tweets of nike about third-party platform is slightly higher than adidas'. This could be because the transactions of nike products on third-party platform are more common than adidas's.

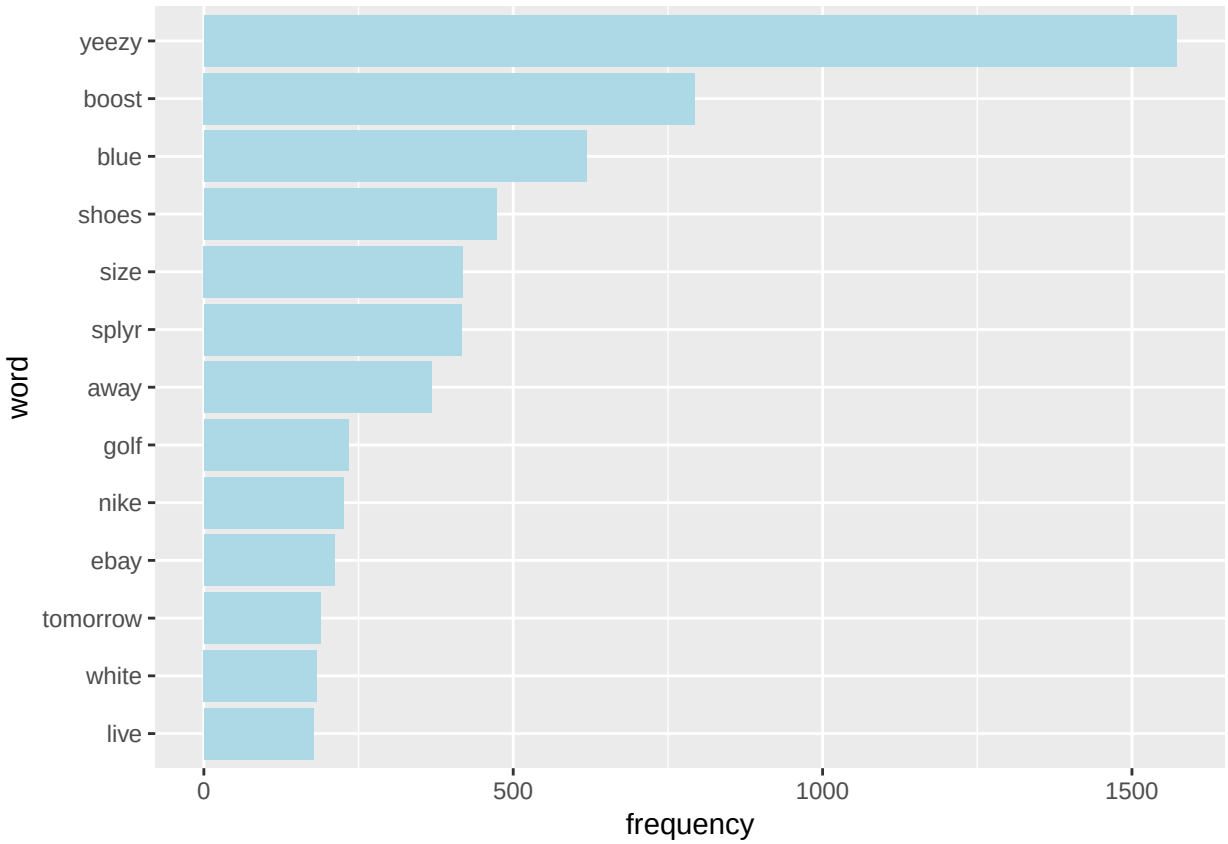
5.3 Diffrent Frequent Words about Shoes

Since we can see that words related to shoes appear most frequently in tweets under both nike and adidas, we will take out the words about shoes in the top 30 most frequent words to see which kind of things people like talking about on their shoes.

```
shoe1 <- top1[c(3,8,14,15,16,18,21,23,25,29),c(1,2)]

shoe2 <- top2[c(2:7,11,18,19,22,27,29,30),c(1,2)]

shoe1$word <- factor(shoe1$word, levels = shoe1$word[order(shoe1$frequency)])
ggplot(data=shoe1, aes(x=word, y=frequency)) + geom_bar(stat="identity",fill = "pink") + coord_flip()
```




```
shoe2$word <- factor(shoe2$word, levels = shoe2$word[order(shoe2$frequency)])
ggplot(data=shoe2, aes(x=word, y=frequency)) + geom_bar(stat="identity",fill = "light blue") + coord_fli
```



There are both similarity and difference between this two graphs. Similarity are that both graph contain names of the shoes like "air","jordan" for nike and "yeezy","splyr" for adidas. Difference is that there are many words about color in the adidas graphs that the nike graph doesn't contain. Its shows that when people tweet about adidas' shoes, they not only use the names of the shoes to refer to them but also use a certain color to describe a set of shoes.

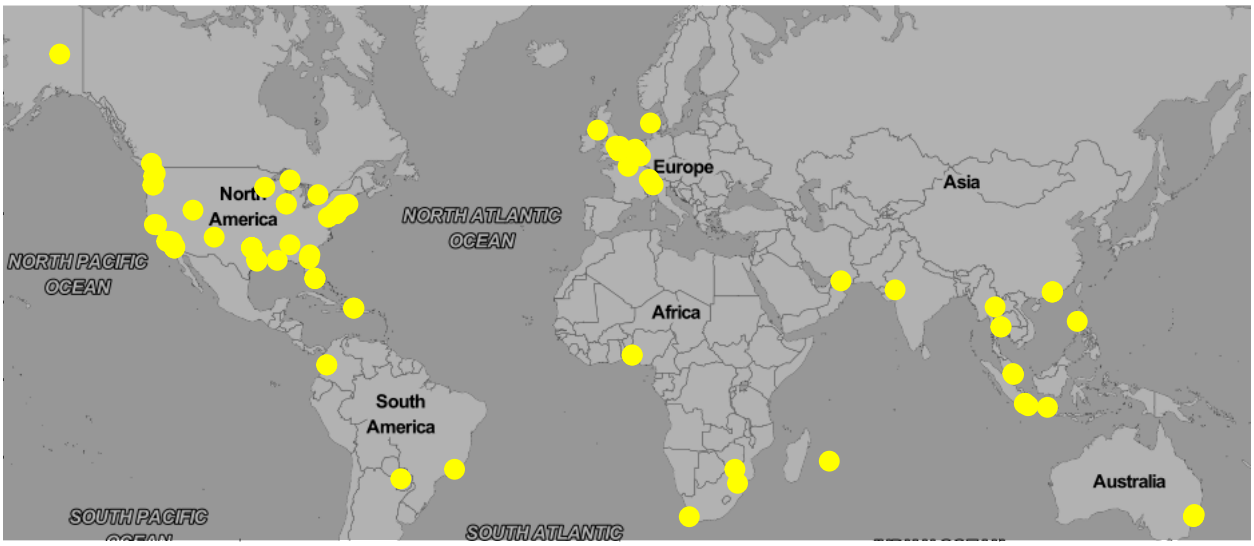
The way that people mention the product may not seem important. But it actually reflects the expressions of products in their minds. For example, when people directly use the name of shoes to point to their interested shoes, they want to talk about that specific style, which corresponds to a specific function, apperance and even price. On the other hand, if people use color names to refer to shoes, it means that they just want to talk about shoes in that color without anything specific on their functions or other perspectives.

6 Map Distribution

```
map1.df <- tweets1.df[!is.na(tweets1.df$longitude),]
map1.df$longitude <- as.numeric(map1.df$longitude)
map1.df$latitude <- as.numeric(map1.df$latitude)
qmplot(longitude, latitude, data = map1.df,
       colour = I('yellow'), size = I(3), darken = .3,zoom=2)
```

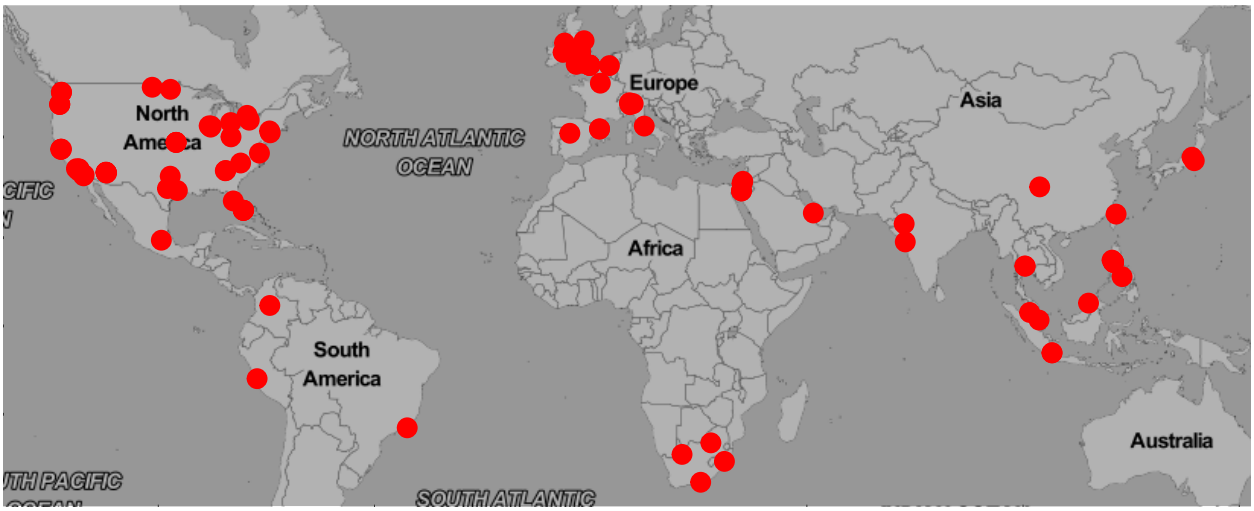
```
## Map from URL : http://tile.stamen.com/toner-lite/2/0/0.png
## Map from URL : http://tile.stamen.com/toner-lite/2/1/0.png
## Map from URL : http://tile.stamen.com/toner-lite/2/2/0.png
```

```
## Map from URL : http://tile.stamen.com/toner-lite/2/3/0.png
## Map from URL : http://tile.stamen.com/toner-lite/2/0/1.png
## Map from URL : http://tile.stamen.com/toner-lite/2/1/1.png
## Map from URL : http://tile.stamen.com/toner-lite/2/2/1.png
## Map from URL : http://tile.stamen.com/toner-lite/2/3/1.png
## Map from URL : http://tile.stamen.com/toner-lite/2/0/2.png
## Map from URL : http://tile.stamen.com/toner-lite/2/1/2.png
## Map from URL : http://tile.stamen.com/toner-lite/2/2/2.png
## Map from URL : http://tile.stamen.com/toner-lite/2/3/2.png
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```




```
map2.df <- tweets2.df[!is.na(tweets2.df$longitude),]
map2.df <- subset(map2.df,screenName!="Sennandy")
map2.df$longitude <- as.numeric(map2.df$longitude)
map2.df$latitude <- as.numeric(map2.df$latitude)
qmplot(longitude, latitude, data = map2.df,
       colour = I('red'), size = I(3), darken = .3,zoom = 2)
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```



The map distribution of tweets related to nike/adidas are similar. Most of the tweets are posted by people in North America, which is reasonable because both nike and adidas are American brands. The second largest group locates on Western Europe, which also makes sense because nike and adidas are popular in Europe too. The rest are posted by people in Asia, Africa and South America, which are located pretty disperse without too much pattern.

7 Analysis on Nike-How Word Frequency Changes as Time Changes

Theoretically speaking, the attention on new product release will become higher as the time approches the release date. Here I would like to see if we can infer on any future release based on the the change of peoples' interest in December in 2017 by visualizing the change of word frequency of product release related words.

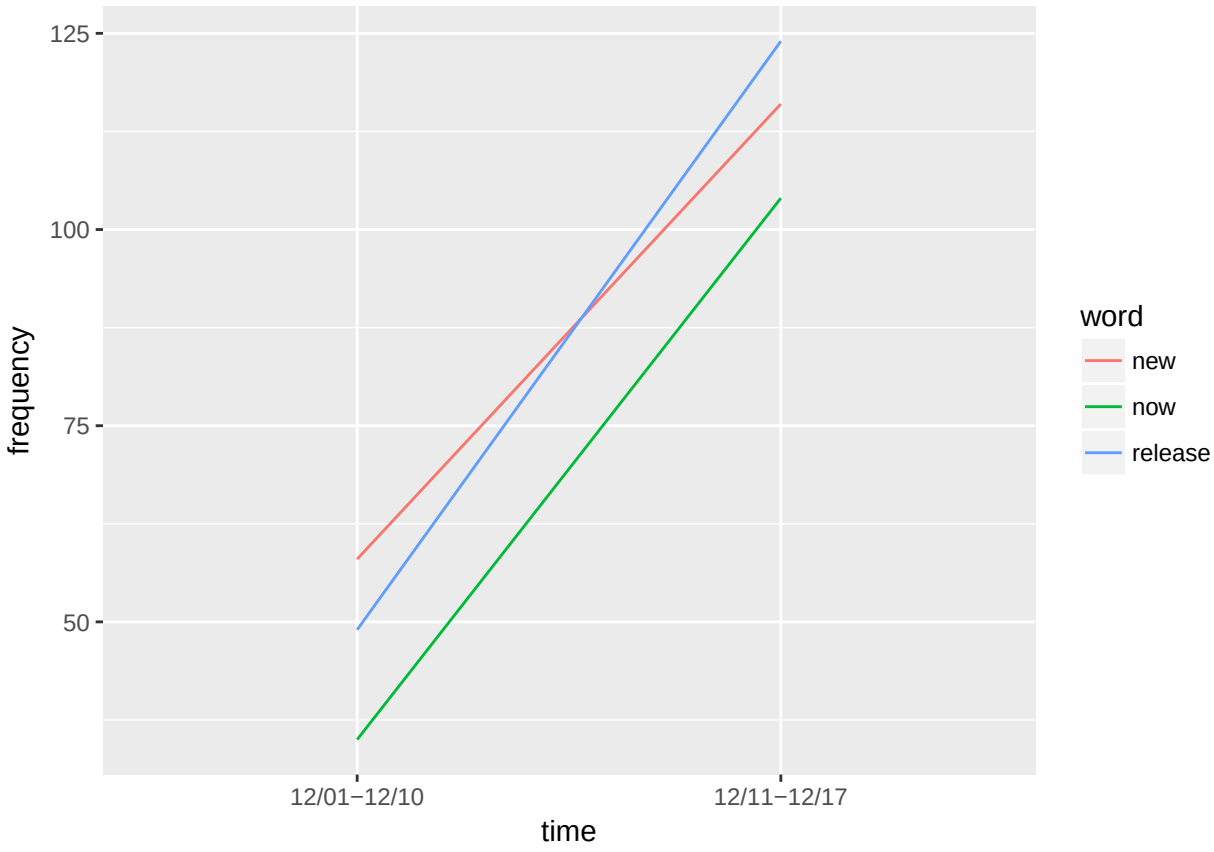
```
#tweets acquiring
time1_1 <- searchTwitter("#nike", n=2000, lang="en", since='2017-12-01', until='2017-12-10')
```

```
## Warning in doRppAPICall("search/tweets", n, params = params,
## retryOnRateLimit = retryOnRateLimit, : 2000 tweets were requested but the
## API can only return 928
```

```
time1_1.df <- twListToDF(time1_1)

time2_1 <- searchTwitter("#nike", n=2000, lang="en", since='2017-12-11', until='2017-12-17')
```

```
## [1] "Rate limited .... blocking for a minute and retrying up to 119 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 118 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 117 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 116 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 115 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 114 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 113 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 112 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 111 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 110 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 109 times ..."
## [1] "Rate limited .... blocking for a minute and retrying up to 108 times ..."
```

```
## [1] "Rate limited .... blocking for a minute and retrying up to 107 times ..."
time2_1.df <- twListToDF(time2_1)

#tweets cleaning
df.text1.t1 <- str_replace_all(time1_1.df$text, "[^[:alpha:]]", " ")
df.text2.t2 <- str_replace_all(time2_1.df$text, "[^[:alpha:]]", " ")

df.text1.t1 <- Corpus(VectorSource(df.text1.t1))
df.text2.t2 <- Corpus(VectorSource(df.text2.t2))

df.text1.t1 <- tm_map(df.text1.t1, content_transformer(tolower))
df.text2.t2 <- tm_map(df.text2.t2, content_transformer(tolower))

df.text1.t1 <- tm_map(df.text1.t1, removeWords, stopwords("english"))
df.text2.t2 <- tm_map(df.text2.t2, removeWords, stopwords("english"))

df.text1.t1 <- tm_map(df.text1.t1, removeWords, c("https","pic","took","jayteep","fzgsmh","select","che
df.text2.t2 <- tm_map(df.text2.t2, removeWords, c("https","giving","one","get","check","amp","official"

df.text1.t1 <- tm_map(df.text1.t1, stripWhitespace)
df.text2.t2 <- tm_map(df.text2.t2, stripWhitespace)


dtm.t1 <- TermDocumentMatrix(df.text1.t1)
m1.t1 <- as.matrix(dtm.t1)
v1.t1 <- sort(rowSums(m1.t1),decreasing=TRUE)
d1.t1 <- data.frame(word = names(v1.t1),frequency=v1.t1)

dtm.t2 <- TermDocumentMatrix(df.text2.t2)
m2.t2  <- as.matrix(dtm.t2)
v2.t2 <- sort(rowSums(m2.t2),decreasing=TRUE)
d2.t2 <- data.frame(word = names(v2.t2),frequency=v2.t2)

#pick the words about product release
new.t1 <- d1.t1 [d1.t1 $word == "new",]
now.t1 <- d1.t1 [d1.t1 $word == "now", ]
release.t1 <- d1.t1 [d1.t1 $word == "release",]
all.t1 <- rbind(new.t1,now.t1,release.t1)
all.t1$time <- rep("12/01-12/10",3)

new.t2 <- d2.t2[d2.t2$word == "new",]
now.t2 <- d2.t2[d2.t2$word == "now", ]
release.t2 <- d2.t2[d2.t2$word == "release",]
all.t2 <- rbind(new.t2,now.t2,release.t2)
all.t2$time <- rep("12/11-12/17",3)

all <- rbind(all.t1,all.t2)

ggplot(all,aes(x = time,y = frequency,group = word,color = word)) + geom_line()
```



As we expect, the frequency of words–"new", "now" and "release" become much higher in the late December, from which we can assume that the release will be in the beginning of 2018. Nike did a successful advertising because people are talking about the release. This method could be used by companies to detect the change of people's interest on their new product. If there is not much change as the release date approches, the reason could be either the company doesn't do enough advertising so people don't know about it or the advertising is done so early compared with the release date that people's interest already faded.