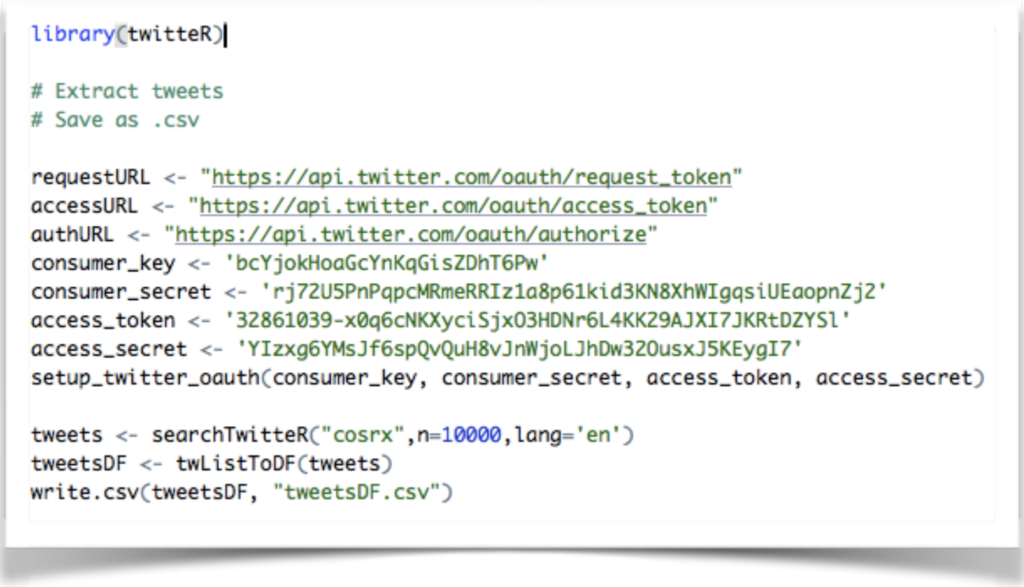
Data Story
COSRX Twitter Sentiment Analysis
By Cate de Leon

Social media is a fertile breeding ground when it comes to customer feedback about brands, and the beauty industry is not exempted from this. Many customers who don't bother to comment and rate on official websites freely share and exchange their opinions on social media. It is thus very useful for any business to be able to mine this data.

This project uses Twitter data. Using the twitteR function, we extracted tweets that mention COSRX — a cult-fave Korean beauty brand—in the English language. The searchTwitteR function was able to extract 1,100 tweets created from February 14 to 24, 2018.



```
library(twitteR)

# Extract tweets
# Save as .csv

requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
consumer_key <- 'bcYjokHoaGcYnKqGisZDhT6Pw'
consumer_secret <- 'rj72U5PnPqpcMRmeRRIz1a8p61kid3KN8XhWIgqsiUEaopnZj2'
access_token <- '32861039-x0q6cNKXyciSjxO3HDNr6L4KK29AJXI7JKRtDZYSl'
access_secret <- 'YIzxg6YMsJf6spQvQuH8vJnWjoLJhDw32OusxJ5KEygI7'
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)

tweets <- searchTwitteR("cosrx",n=10000,lang='en')
tweetsDF <- twListToDF(tweets)
write.csv(tweetsDF, "tweetsDF.csv")
```

Scope and limitation

Unfortunately, as most Twitter users don't turn on their location, 1099 out of the 1,100 tweets have NA values for the longitude and latitude values. This would have been useful for locating where most clients and fans of the brand are coming from.

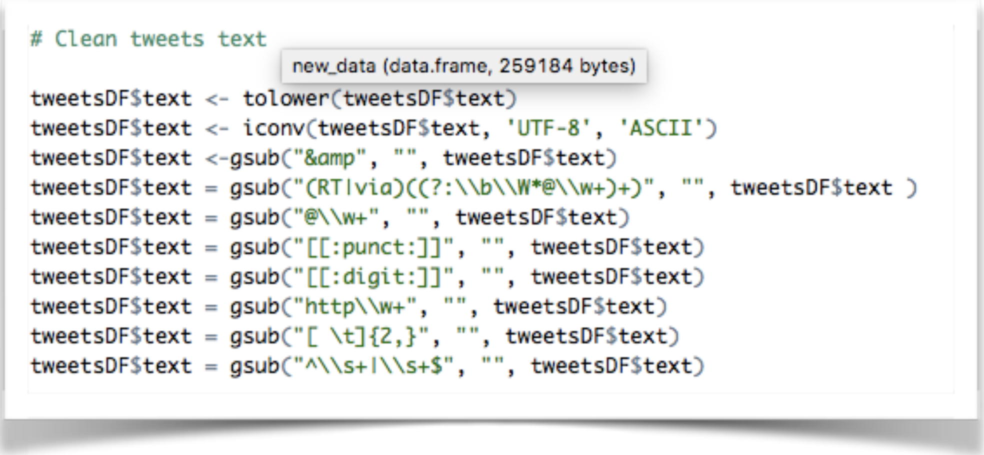
While the data frame contains 17 variables, including favorites, retweets, Twitter handles, and date created, we will be using only two:

1. **X** - Int values to be used as an index.
2. **text** - The text content of the tweets.

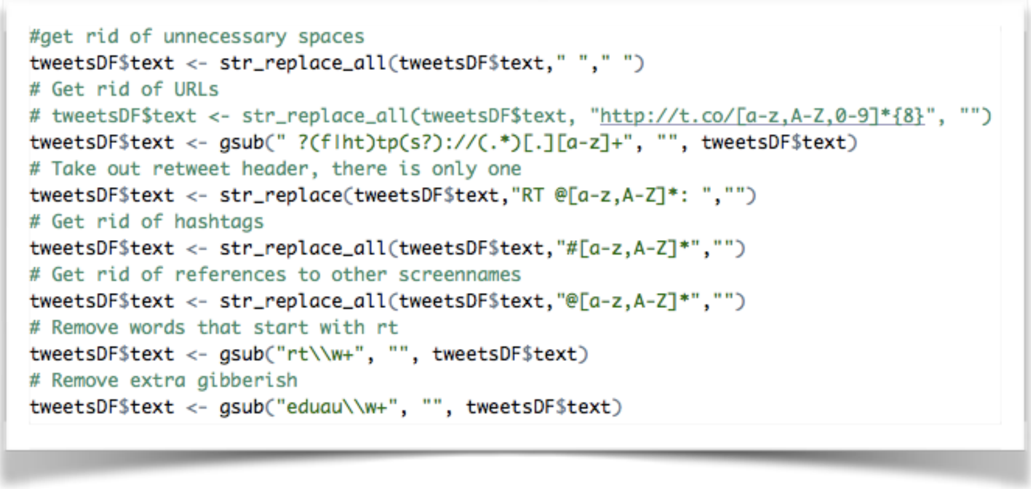
Note
Extracting such a data frame from Twitter can be done at any time. To keep the findings of this particular project consistent, the extraction code will be saved separately so as not to accidentally create duplicate data sets with different contents.

Data wrangling

Since we're dealing with a corpus of text, data wrangling will focus on regex to remove Twitter handles, links, hashtags, stop words, and gibberish.



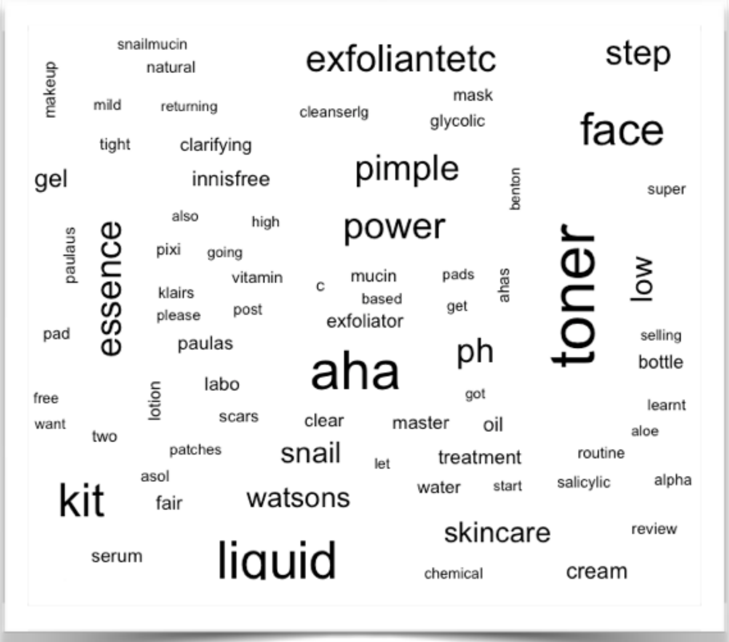
```
# Clean tweets text

tweetsDF$text <- tolower(tweetsDF$text)
tweetsDF$text <- iconv(tweetsDF$text, 'UTF-8', 'ASCII')
tweetsDF$text <-gsub("&amp", "", tweetsDF$text)
tweetsDF$text = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", tweetsDF$text )
tweetsDF$text = gsub("@\\w+", "", tweetsDF$text)
tweetsDF$text = gsub("[[:punct:]]", "", tweetsDF$text)
tweetsDF$text = gsub("[[:digit:]]", "", tweetsDF$text)
tweetsDF$text = gsub("http\\w+", "", tweetsDF$text)
tweetsDF$text = gsub("[ \t]{2,}", "", tweetsDF$text)
tweetsDF$text = gsub("^\\s+|\\s+$", "", tweetsDF$text)
```



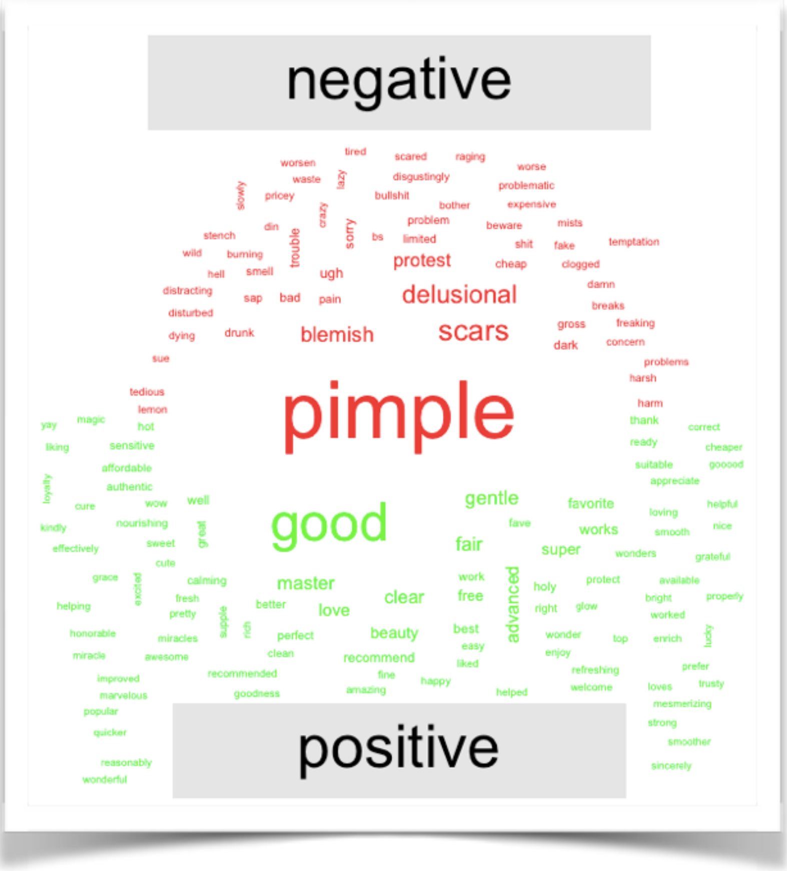
```
#get rid of unnecessary spaces
tweetsDF$text <- str_replace_all(tweetsDF$text," "," ")
# Get rid of URLs
# tweetsDF$text <- str_replace_all(tweetsDF$text, "http://t.co/[a-z,A-Z,0-9]*{8}", "")
tweetsDF$text <- gsub(" ?(f|ht)tp(s?)://(.*)[.][a-z]+", "", tweetsDF$text)
# Take out retweet header, there is only one
tweetsDF$text <- str_replace(tweetsDF$text,"RT @[a-z,A-Z]*: ","")
# Get rid of hashtags
tweetsDF$text <- str_replace_all(tweetsDF$text,"#[a-z,A-Z]*","")
# Get rid of references to other screennames
tweetsDF$text <- str_replace_all(tweetsDF$text,"@[a-z,A-Z]*","")
# Remove words that start with rt
tweetsDF$text <- gsub("rt\\w+", "", tweetsDF$text)
# Remove extra gibberish
tweetsDF$text <- gsub("eduau\\w+", "", tweetsDF$text)
```

First insights

A good place to start when it comes to figuring out what people are saying about a brand on social media is through the construction of wordclouds. These won't give you exact figures, but will give you key words and discussion topics at a glance:



We can also create a negative-positive wordcloud to pit the opposing sentiments against each other:

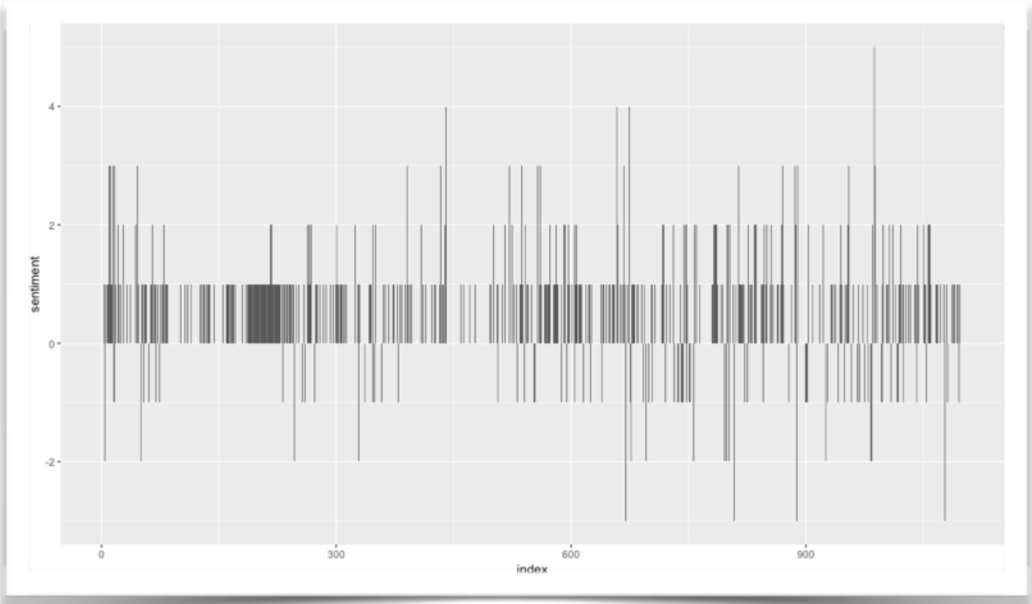


From these two wordclouds, we can already gather that the tweet texts will mainly be providing us with the following:

1. Most popular product names and substances
2. Competing brands
3. Brand/product feedback, although this comes at a much lower frequency than the first two.

Sentiment analysis

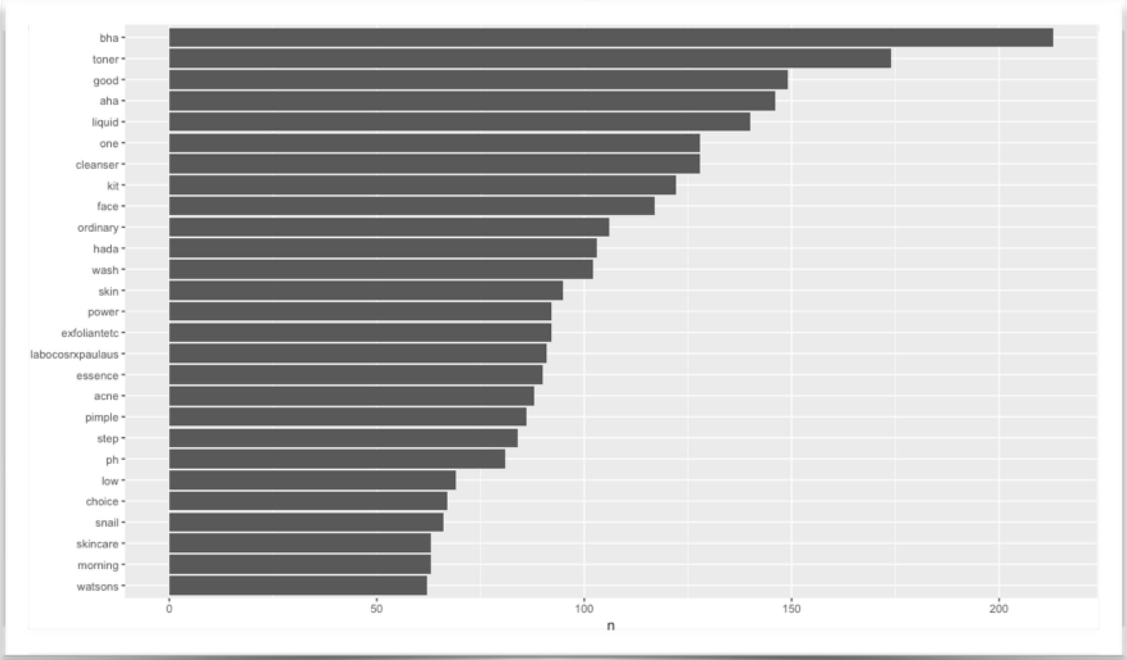
We performed a sentiment analysis on the text. However, as the top words tend to be parts of product names or blemishes themselves (e.g. **Good** Morning Facial Wash, **Advanced** Snail Mucin, **Acne Pimple** Master Patch), these findings should be taken with a grain of salt. For the most part, these don't reflect customers' opinions.



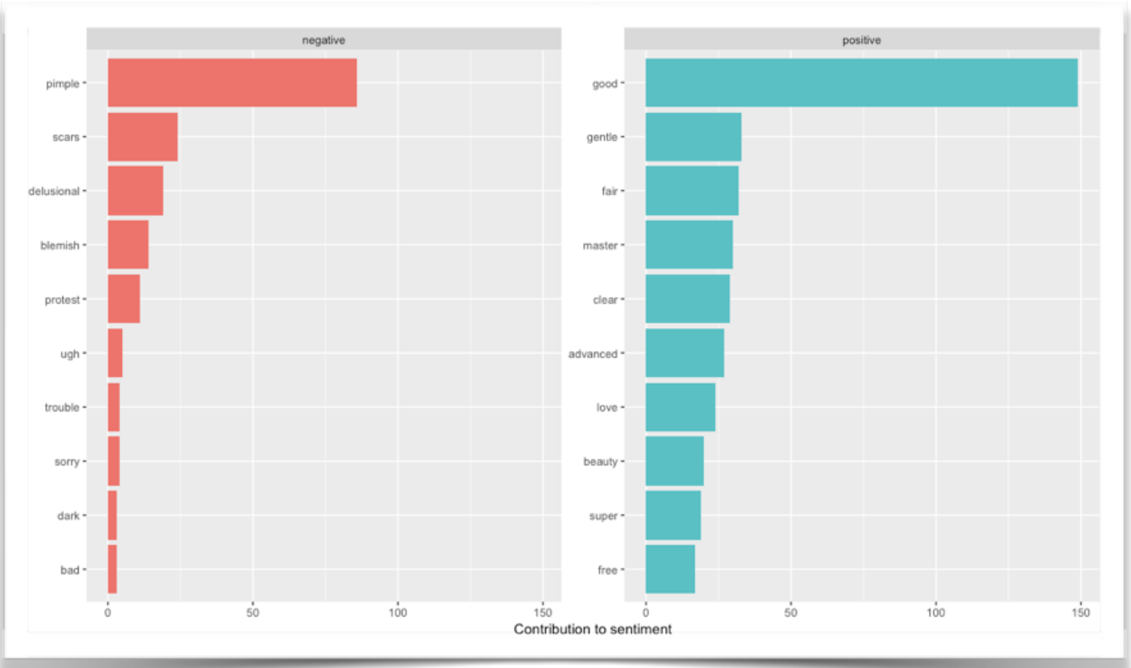
That being said, sentiments seem to be mostly neutral and skewed towards positive.

Quantifying findings

We can quantify these initial findings by getting the top words. After stripping the text of stop words and gibberish, these are the most frequent words that emerged:



We also gathered the top negative and positive words:



As previously mentioned in the Statistical Analysis, the two word counts support each other, the main finding being:

Customers are mostly concerned with pimples and skin blemishes

Products and chemicals that address pimples and acne, as well as the word pimple itself, easily emerge to the top.

This is perhaps the most significant finding of the study so far, and can be directly applied by the client to product development and social media marketing.

Secondary finding:

Other words that emerge to the top are competitor brands, mainly The Ordinary, Paula's Choice, and Hada Labo, which makes it easier for the COSRX to figure out who to keep an eye on, as these are the brands that enter most into the discussion and are pitted against their products.

Further step: Topic modelling

From here, a good step would be to run a Latent Dirichlet allocation. This will create bags of related and co-occurring words, to help us get a sense of the different topics being talked about. This will provide us with a more nuanced picture of the discussion, on top of what we have so far.