CAPSTONE PROJECT FINAL REPORT
COSRX Twitter Sentiment Analysis
By Cate de Leon

INTRODUCTION

COSRX is a cult-fave Korean beauty brand that focuses on skincare, and has enthusiastic customers across Asia. Social media is rich with people's opinions on various products and services, and the beauty industry is definitely not exempted from this.

This project uses Twitter data and sentiment analysis to glean what customers and skincare enthusiasts are saying about the brand and how they can improve. This includes common skincare concerns, favorite products and trends—particularly when it comes to skincare chemicals—and competing products and brands that are often compared with and mentioned alongside COSRX's.

ABOUT THE DATA

Tweets about COSRX were gathered using the twitteR package. The data includes 1,100 tweets that were created from February 14 to 24, 2018 in the English language. As most Twitter users don't turn on their location, we don't have access to where most of these tweets came from. But beauty cult brands are patronized internationally, and as will become evident further in this project, many of COSRX's competitors are manufactured in other countries, such as Japan, Canada, the US, and the UK.

DATA EXTRACTION

We created a Twitter app via http://apps.twitter.com to obtain URLs, tokens, and keys necessary for extraction. We then used the searchTwitteR function to get tweets that mentioned "cosrx" and converted this into a data frame.

```
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 <- 'YIzxg6YMsJf6spQvQuH8vJnWjoLJhDw3ZOusxJ5KEygI7'
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")
```

The data contains 17 variables, including the user's Twitter handle, date of tweet creation, whether or not a tweet was favorited or retweeted, how many favorites or retweets it received, and the longitude and latitude of the tweet which, as previously mentioned, mostly had NA values (1099 out of 1100 observations).

***Note:** So as not to accidentally create and work on duplicate and differing data frames, the extraction code is saved in a separate .R file.

For this project, we only used two variables:

1. **X** - an index that organizes the text per tweet.
2. **text** - the actual tweets/text content.

DATA WRANGLING

The tweet text needed to be cleaned up to get rid of Twitter handles that were mentioned, links, RT headers, hashtags and gibberish.

```
# Clean tweets text
                    new_data (data.frame, 259184 bytes)
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)
```

We also removed stop words and custom stop words to further clean the text. Since we're dealing with a corpus of words, the data was cleaned up to the point that gibberish and unimportant words no longer showed up in the exploratory data analysis.

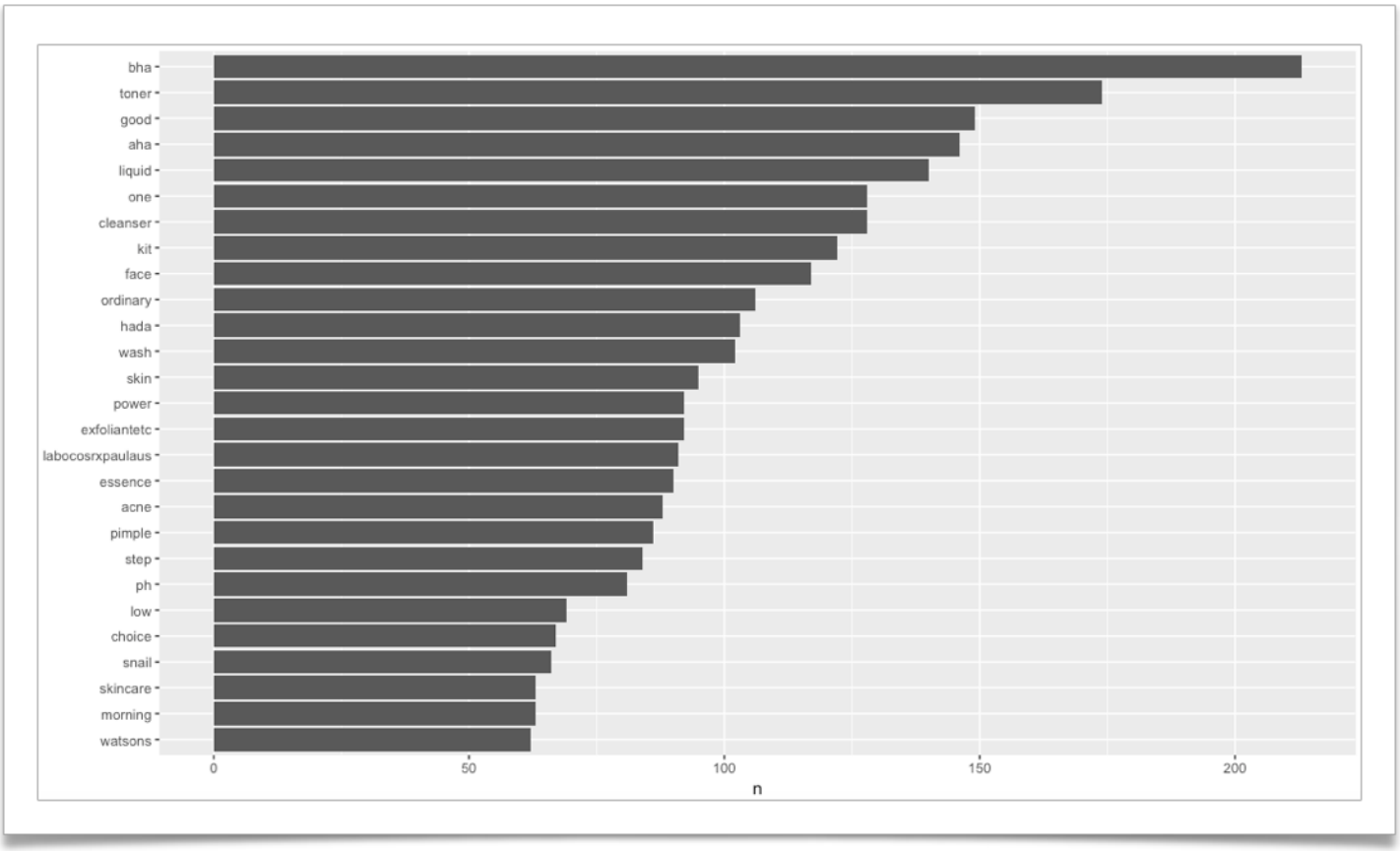
The last step to prepare the data was to create a subset to include only observations under the variables X and text. We then tokenized the words to prepare for text mining using the tidytext library.

```
# Create text and index subset
# Tokenize words

tidy_tweets <- select(tweetsDF, X, text) %>%
  unnest_tokens(word,text) %>%
  anti_join(stop_words)
```

EXPLORATORY DATA ANALYSIS

A good place to start gauging what people are saying about the brand is to see which are the most common words used.



Immediately we see that **BHA**'s frequency is the highest among the top words, with more than 200 appearances. BHA (beta hydroxy acid) is an exfoliating skincare chemical that is typically used to treat acne, as it is oil soluble and able to penetrate pores and clear them out.

This is followed by the words “toner” and in fourth place AHA—alpha hydroxy acid, which exfoliates the surface of the skin, leading to less dead skin cell build up, which can lead to irritation. These three top words form the name of one of COSRX’s bestselling products, the **AHA/BHA Clarifying Toner**, which helps to clear skin blemishes, both on the surface and beneath.

Next is the word “liquid” which is also attached to the words BHA and AHA in other COSRX product names (i.e. **BHA Blackhead Power Liquid** and **AHA 7 Whitehead Power Liquid**), which also clarify the skin.

Going back a bit, the #3 word is “good” which is part of COSRX’s bestselling facial wash, the **Low pH Good Morning Gel Cleanser**.

At number 6 and 8, we have the words “one” and “kit”, which are part of the product names **One Step Clear Pads** and **One Step Pimple Clear Kit**.

Going downwards from #10, we start to see the competition brands that are most often pitted against COSRX, such as **The Ordinary** (10), **Hada Labo** (11) which is known for its face washes (9 and 12), and **Paula’s Choice** (16).

Moving down further on the top words, we continue to see words that are part of previously mentioned products and brands: **power** (14), **step** (20), **ph** (21), **low** (22), **choice** (23), and **morning** (26).

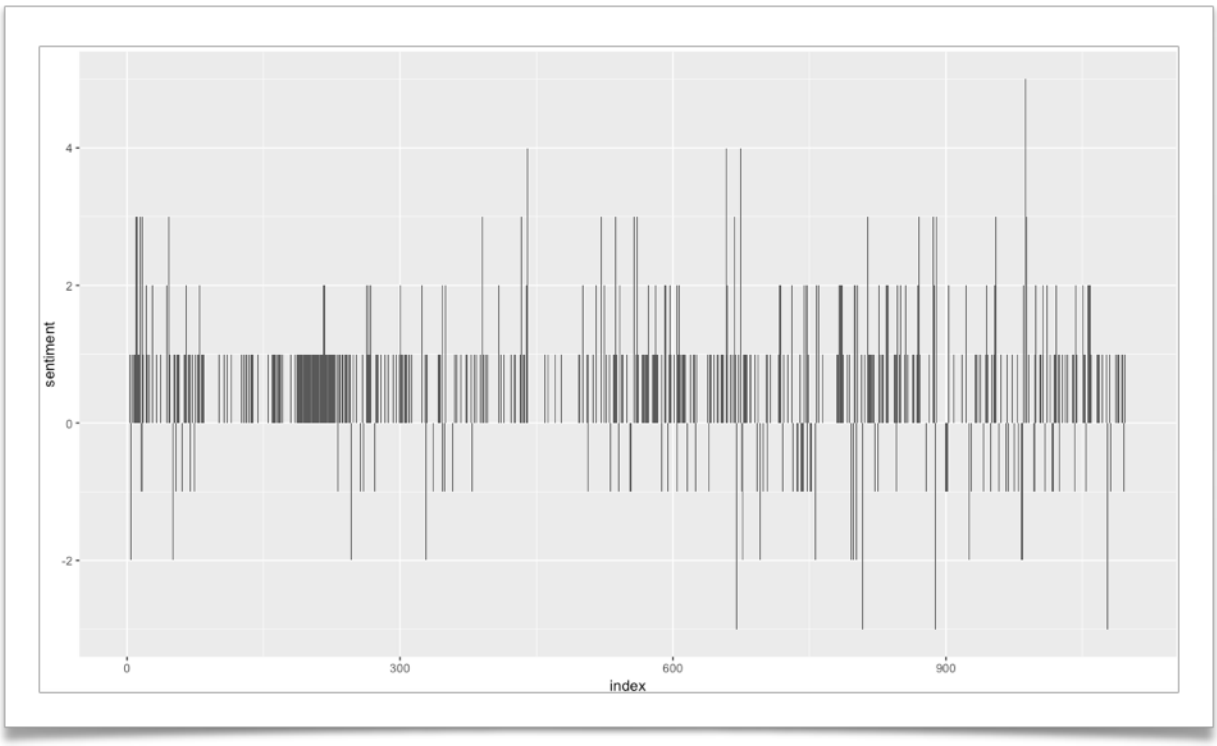
In addition, we have **exfoliant** (15), which describes what BHA and AHA essentially are. We have **essence** (17) and **snail** (24), which are part of the COSRX product name **Advanced Snail 96 Mucin Power Essence**. Snail Mucin is known for its ability to hydrate and heal the skin, which complements pimple treatments and helps heal acne scars.

We also see another pimple fighting product at numbers 18 and 19—**acne** and **pimple**—which are part of the product **Acne Pimple Master Patch**. These are stickers that draw out the gunk from blemishes and help heal pimples faster overnight.

At last place, with a little over 50 words, we have **watsons**, which is a drugstore/pharmacy where a lot of beauty products can be bought.

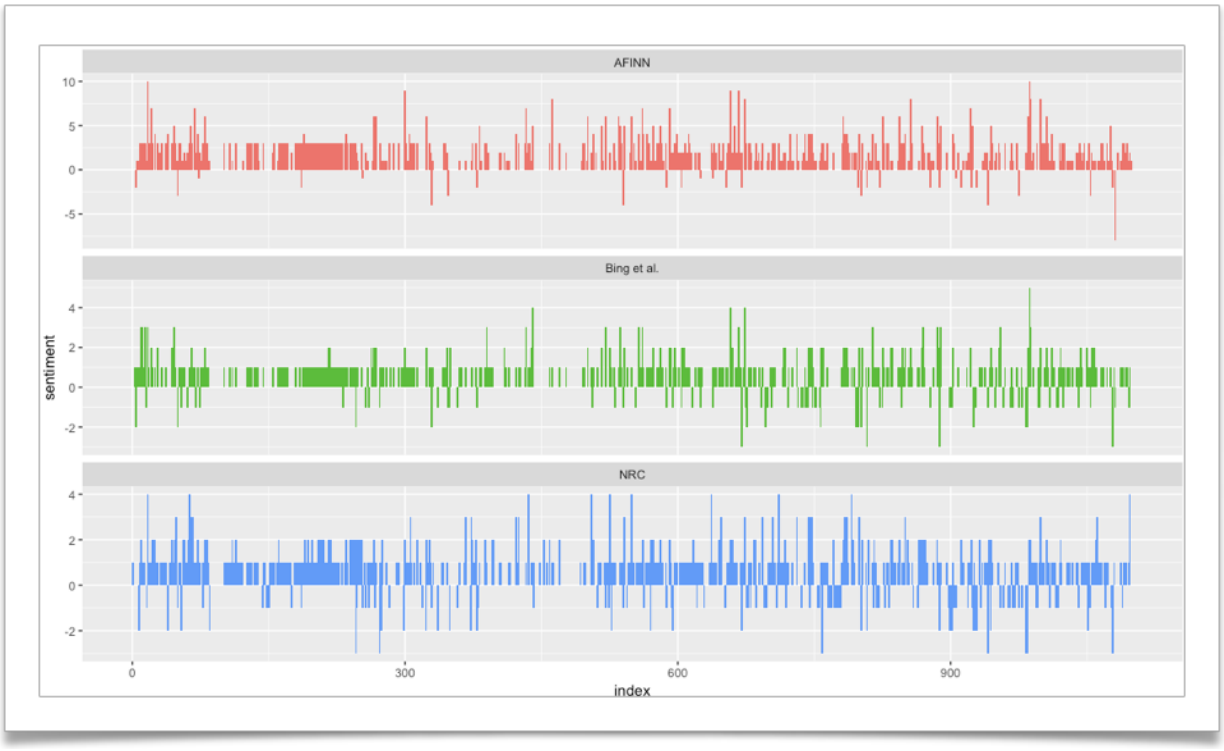
SENTIMENT ANALYSIS

Analyzing sentiments is a good way to see how tweeting customers feel about the brand in general. We start with Bing sentiment lexicon, as this simply rates the words as either positive and negative.

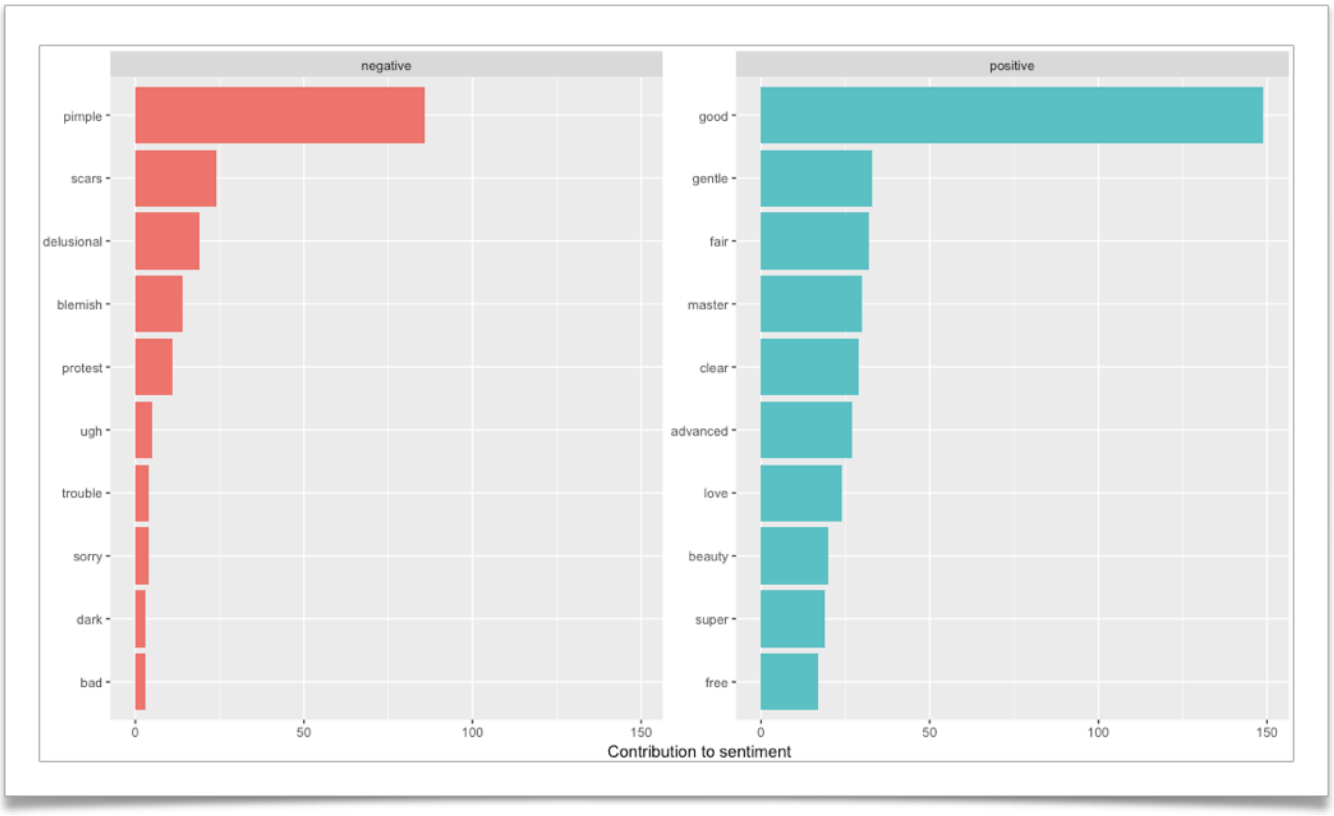


Sentiment seems to be mostly neutral and skewed towards positive. This is a cult fave brand after all.

We then compare how the different sentiment lexicons (Bing, Afinn, and NRC) analyze the text.



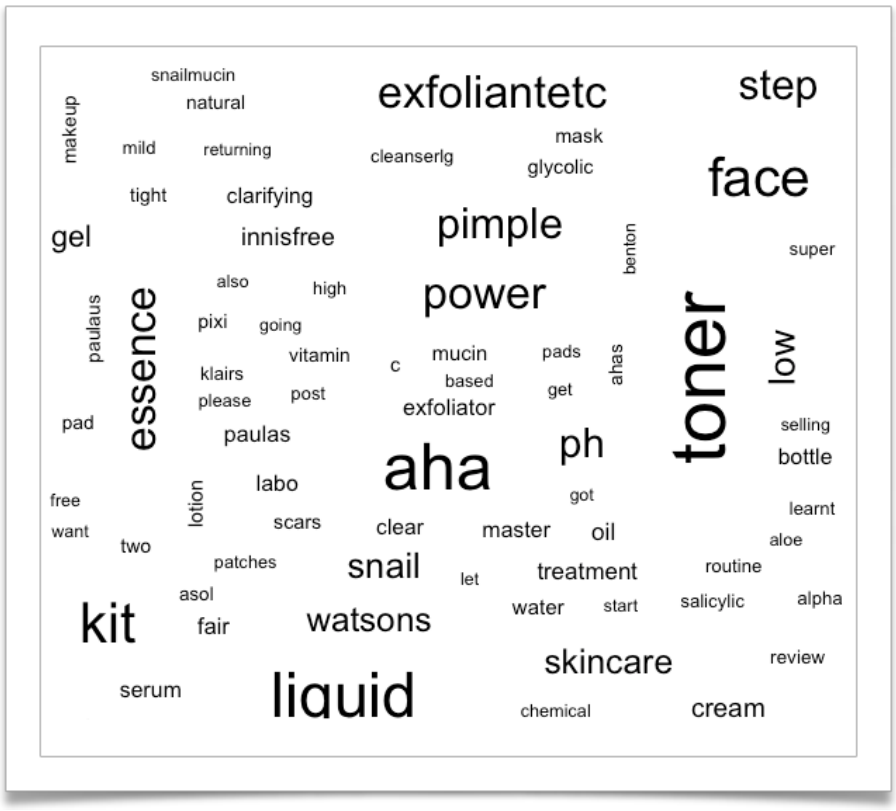
We also analyzed the most common positive and negative words.

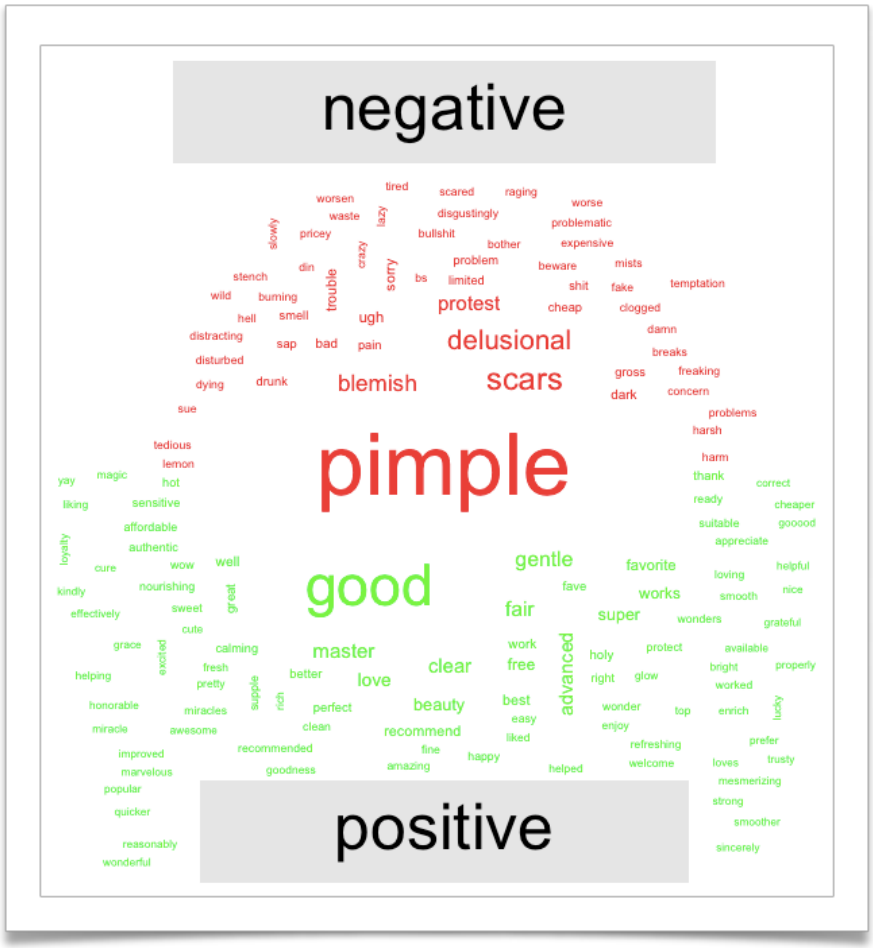


From here, we realize we must take the tidytext sentiment analysis with a grain of salt. The most common negative words seem to be skin problems, rather than feedback about the brand. Meanwhile, the most common positive words are part of COSRX product names instead of reviews (e.g. Low-pH **Good** Morning Gel Cleanser, Acne Pimple **Master** Patch, **Advanced** Snail 96 Mucin Power Essence, One Step Pimple **Clear** Pads, **Clear** Fit Master Patch). The word **pimple** could also be the top negative word due to this reason.

Many beauty brands use such positive/negative words as well in their products to better market the results customers can expect from usage, so sentiment analysis provided by tidytext lexicons may not be as effective when analyzing some beauty brands.

Another effective way to visualize most common words is through wordclouds. This gives us additional important words that may not appear in the top 10 or 20.





The positive vs. negative wordcloud obviously has more positive than negative words in it, even after product names and skin problems are accounted for.

From here, it's also obvious that pimples are customer's main skin concern.

TOPIC MODELLING

We run a Latent Dirichlet allocation (text2vec) on the text to get a sense of the different topics people talked about. This is done by grouping the words that often appear with one another and are relevant to each other. This gives us a more nuanced view of the discussion.

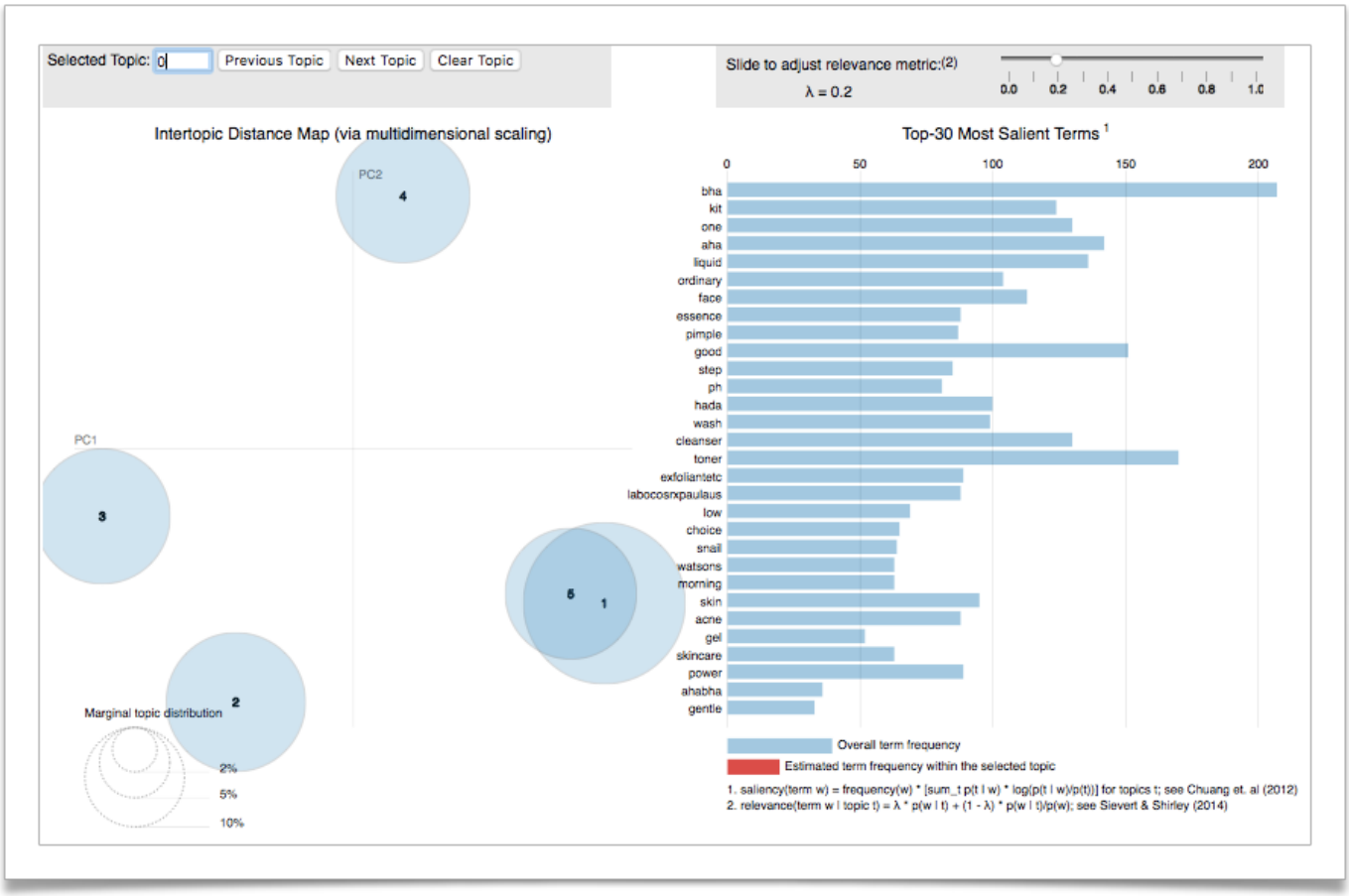
The first step is to tokenize the words and create a document term matrix:

```
# Topic modelling
# Run Latent Dirichlet allocation

tokens = tweetsDF$text %>%
  tolower %>%
  word_tokenizer
it = itoken(tokens, progressbar = FALSE)
v = create_vocabulary(it) %>%
  prune_vocabulary(term_count_min = 10, doc_proportion_max = 0.2)
vectorizer = vocab_vectorizer(v)
dtm = create_dtm(it, vectorizer, type = "dgTMatrix")
```

For this project, we set the number of topics to five.

```
set.seed(123)
lda_model = LDA$new(n_topics = 5, doc_topic_prior = 0.3, topic_word_prior = 0.0001)
doc_topic_distr =
  lda_model$fit_transform(x = dtm, n_iter = 1000,
                          convergence_tol = 0.0001, n_check_convergence = 25,
                          progressbar = FALSE)
barplot(doc_topic_distr[1, ], xlab = "topic",
        ylab = "proportion", ylim = c(0, 1),
        names.arg = 1:ncol(doc_topic_distr))
```



We then get the top 10 words per topic:

```
> lda_model$get_top_words(n = 10, topic_number = c(1L, 2L, 3L, 4L, 5L), lambda = 0.2)
      [,1]       [,2]          [,3]       [,4]         [,5]
 [1,] "ordinary" "ahabha"      "kit"      "ph"         "bha"
 [2,] "essence"  "gentle"      "one"      "low"        "aha"
 [3,] "choice"   "exfoliator"  "pimple"   "morning"    "liquid"
 [4,] "snail"    "mask"        "step"     "gel"        "face"
 [5,] "paulas"   "two"         "watsons"  "oil"        "hada"
 [6,] "mucin"    "tight"       "fair"     "cleansing"  "wash"
 [7,] "c"        "salicylic"   "master"   "labo"       "exfoliantetc"
 [8,] "scars"    "normal"      "advanced" "clear"      "labocosrxpaulaus"
 [9,] "love"     "cleanserlg"  "patch"    "water"      "toner"
[10,] "glycolic" "mild"        "pixi"     "lotion"     "bottle"
> lda_model$plot()
```

From this, we can draw the following topics:

Topic 1 is about brightening:

This is what is commonly achieved by the substances Snail Mucin (aside from hydration), exfoliants such as Glycolic Acid, and Vitamin C. Brightening is commonly desired by customers who are trying to fade scars, such as dark post-acne marks, or have dull skin.

The brands The Ordinary and Paula's Choice were also mentioned in this topic.

Topic 2 is about gently carifying the skin:

This is what is achieved by the substances AHA (alpha hydroxy acid), BHA (beta hydroxy acid), and salicylic acid (which is a form of BHA)—all of which are exfoliators. Their Low pH Good Morning Gel Cleanser contains a form of BHA and is known for being both effective and gentle/mild—which is COSRX's claim to fame with its product line.

Other words in this topic are mask, two, tight, and normal.

Topic 3 is about quick fixes for pimples:

COSRX's One Step Pimple Clear Pads, One Step Pimple Clear Kit, Acne Pimple Master Patch, and Clear Fit Master Patch (stickers that shrink pimples overnight) dominated this topic.

Other words mentioned were watsons (a drugstore/pharmacy), fair, advanced (most probably from the Advanced Snail 96 Mucin Power Essence—which helps heal damaged skin), and the brand Pixi.

Topic 4 is about cleansers.

Users pitted (or perhaps used in conjunction, as double cleansing with oil/micellar water, followed by a foaming cleanser is a popular regimen) COSRX's Low pH Good Morning Gel Cleanser and Hada Labo's Cleansing Oil (against/with) each other.

Other words in this topic are clear, water, and lotion—which can probably be alluded to COSRX's Oil-Free Ultra-Moisturizing Lotion. (Among beauty enthusiasts, moisturization is an important step to follow up cleansing).

Topic 5:

Topic 5 is the least clear cut of the five, but the top words comprise the product Hada Labo Face Wash.

It also mentions BHA, AHA, liquid, exfoliant, and toner, which can be alluded to COSRX's AHA/BHA Clarifying Toner and other AHA and BHA liquid products.

The brand Paula's Choice is also mentioned in this topic.

SUMMARY OF FINDINGS:

Top competing brands mentioned were:

The Ordinary (Canada)
Hada Labo (Japan)
Paula's Choice (US)
Innisfree (Korea)
Pixi (UK)
Benton (Korea)
Drunk Elephant (US)
Dear, Klairs (Korea)

Top skin problems:

Pimples
Scars
Blemishes
Blackheads
Dark spots
Clogged pores

Favorite skincare substances:

BHA (clarifying)
AHA (clarifying)
Snail Mucin (hydration and brightening)
Glycolic Acid (clarifying and brightening)
Vitamin C (brightening)
Aloe (hydration)
Birch sap (hydration)

COSRX products with the most organic social media mileage:

AHA/BHA Clarifying Toner (clarifying)
Low-pH Good Morning Gel Cleanser (clarifying)
BHA Blackhead Power Liquid (clarifying)
AHA 7 Whitehead Liquid (clarifying)
Acne Pimple Master Patch (clarifying)
Clear Fit Master Patch (clarifying)
Advanced Snail 96 Mucin Power Essence (hydration and brightening)
One Step Pimple Clear Pads (clarifying)
BHA Skin Returning A-sol (clarifying)
Galactomyces 95 Whitening Power (brightening)

Additional skincare regimen and results keywords:

Negative:	Postive:
Tedious / trouble / lazy / bother / tired	Gentle
Harsh	Spotless
Distracting	Miracles / magic / wonder
Disturbed	Glow

Negative:	Postive:
Delusional	Easy
Ugh	Reasonably / affordable / cheaper
Stench	Rich
Pain	Love
Bs	Works
Fake	Supple
Burning	Refreshing
Pricey	Bright
Problematic	Calming
	Nourishing
	Protect
	Smooth
	Free

Recommendations for client:

1. Keep an eye on developments by top competing brands, particularly Hada Labo, The Ordinary, and Paula's Choice. Hada Labo's cleansers were particularly popular in the discussion. The Ordinary and Paula's Choice were mentioned when it came to brightening.

2. Pimples and acne are the most popular skincare concerns (this finding is supplemented by the COSRX products that emerged on top—8 out of 10 cater to clearing up acne and blemishes). This is followed by scarring and dark spots. Hydration and moisturization matter to customers, but not as much as the first two. Keep this in mind when prioritizing developments and improvements, as well as social media marketing.

3. When it comes to skincare regimens, convenience is key. Customers don't like products that are tedious or too much trouble. The feel of products when applied to skin also matters (eg. gentleness vs. harshness, refreshing, calming).