**Capstone Mini Project: Data Wrangling for Twitter Sentiment Analysis**

**About the data:**

Our data was extracted from Twitter by creating a Twitter app via <http://apps.twitter.com> to obtain URLs, tokens, and keys necessary for extraction. We then used the searchTwitteR function (from the twitteR package) to get tweets that contained the word “cosrx” in the English language. We saved the extracted tweets into a .csv file (tweetsDF.csv).

It includes 1,100 tweets that were created from February 14 to 24, 2018. Unfortunately, as most Twitter users don’t turn on their location, 1,099 out of the 1,100 observations for the variables longitude and latitude are NA.

The data contains 17 variables in total, including the tweet text, the user’s Twitter handle, date of tweet creation, whether or not a tweet was favorited or retweeted, and how many favorites or retweets it received.

**Data wrangling**

Since our analysis will focus tweet sentiments, we are only using the variables **X** (an int to be used as an index) and **text** in the data frame. Thus, the goal is is to clean up the text with regex, removing Twitter handles, punctuation, RT headers and links.

We also removed stop words and custom stop words, to get rid of gibberish and be left with words that are relevant.

Since this is a huge corpus of text, the tweets were cleaned up to the point that nonsensical words no longer showed up in exploratory data analysis (getting top words, creating wordclouds, and running a Latent Dirichlet allocation).

The cleaned up data set was then written into a separate .csv file (clean\_tweets.csv).

The last step was tokenizing the words to prepare for text mining.