Graphical user interface, text, application

Description automatically generated

Photo by Tracy Le Blanc from Pexels

## Steps to Complete the Assignment:

1. Load two of the four provided CSV files.
2. Identify the relevant columns, data dimensionality and other attributes which may be useful for NLP based analysis.
3. Process the corpora strings.
4. Apply NLP class, book, online research or package documentation methods to fully understand each individual corpus. Examples may include but are not limited to
   1. Word clouds
   2. Pyramid plots
   3. Sentiment
   4. Topic modeling
   5. Word network
   6. Dendrograms
   7. Associations
   8. Ngram tokenization
5. Retain all aspects of the analytical efforts in your R script.
6. Create a poiwerpoint presentation of 8-12 slides of the most interesting and relevant EDA outputs. Items can be descriptions of topics, data tables, and visuals. The slides need to be professional and presented as if to a non-technical coworker therefore explanations of the data and steps taken are expected.

## Assignment Deliverables include

1. R scripts for data processing, modeling & exploration “lastName\_TM\_SeriesA\_case.R”
   1. Your script(s) must account for all aspects of the material in your presentation to ensure the presentation is data driven (no cheating with Excel or other tools!) and other analytical aspects which you have decide to omit from the presentation to demonstrate topic fluency but may not have been insightful.
2. PowerPoint of any visualizations, findings and descriptions of non-technical results as if presented to your non technical coworkers. “lastName\_TM\_SeriesA\_case\_.pptx”
   1. NO narration is needed for the powerpoint.

## Data Information

There are a total of 4 CSV files. You are expected to choose two of most interesting for exploration, applying NLP methods including comparison and contrasts among and between the two. The data was collected using Twitter’s API and has 90 columns. The data is a snapshot in time and not meant to be a full corporal understanding of the subject. For your information the following code was employed in collection.

library(rtweet)

topicA <- search\_tweets(q = 'Democrats', n = 1000)

topicB <- search\_tweets(q = 'Republicans', n = 1000)

topicC <- search\_tweets(q = 'Dunkin', n = 1000)

topicD <- search\_tweets(q = 'Starbucks', n = 1000)

write\_as\_csv(topicA, 'dems.csv')

write\_as\_csv(topicB, 'reps.csv')

write\_as\_csv(topicC, 'dunks.csv')

write\_as\_csv(topicD, 'sbux.csv')

## Criteria for Success

The case material will be evaluated according to the following criteria.

## **Organization of content**–25pts Logical ordering of ideas, analytical artifacts, applicable visualizations in slides

## **Scripting**- 25pts R Code is well organized, concise, and free from error, with consistent code style and appropriate comments

## **Completeness** – 50pts Understood, and mined the data for relevant insights/recommendations employing multiple methods from lecture, class, research or package documentation