Week 5 Assignment

Tommy Klein

3/1/2022

Working Directory

```
setwd('/Users/tklein/Desktop/Desktop_tpk/JHU_Classes/text_as_data/week5')
```

Library

```
library(ndjson)
source('../functions/helper_functions.R')
```

```
## Package version: 3.2.0
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 4 of 4 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Part 1 - Course Project

For the course project I am going to analyze social media posts that discuss crypto. My friend started a crypto company (Centsy), and is beginning to market it on social media. However, he has no insights about current landscape of crypto on social media - for example, what topics are typically discussed, and what kinds of tweets typically receive the most engagement. Having these insights would help him craft a more effective social media strategy, that can maximize his message. I plan to collect posts on Reddit using the RedditExtractoR package. I will collect posts from popular crypto communities on Reddit to ensure that I'm selecting relevant content. Once I have collected the posts, I will use a variety of methods to prepare the text for analysis, such as using regex to clean the text and remove unneeded symbols/words, and sentiment analysis to enrich the text with more data. At this point I would like to do one of two things, perhaps both: I would like to use an unsupervised learning method that can identify latent groups within the Reddit posts,

and I would like to use a supervised learning method that determines which words, phrases, topics, and sentiments are associated with increased engagement (comments and posts on Reddit).

I think both analyses are helpful for my purpose (improving Centy's social media content). The unsupervised method would shed light on types of posts based on the language, sentiment, and topics that they use, but can also form groups based on the engagement it receives. This analysis should provide Centsy a blue-print for types of social media content to create that ensures that it "fits" in the current landscape, and that it is maximally effective at generating engagement. The superivised analysis would more directly answer the question about what kind of content drives engagement. This analysis should provide some certain words, phrases, sentiments, and or topics that Centsy can use to maximize engagement. Either or both of these analyses could directly shape and improve Centsy's social media strategy.

Literature Review

1. Flora Poecze, Claus Ebster, Christine Strauss, Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts, Procedia Computer Science, Volume 130, 2018, Pages 660-666, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.04.117.
2. Heuiju Chun, Byung-Hak Leem, Hyesun Suh Using text analytics to measure an effect of topics and sentiments on social-media engagement: Focusing on Facebook fan page of Toyota, International Journal of Engineering Business Management Volume 13, 2021, https://journals.sagepub.com/doi/full/10.1177/18479790211016268#
3. Trunfio, Mariapina and Rossi, Simona Conceptualising and measuring social media engagement: A systematic literature review (article) Italian Journal of Marketing 2021 - 3 2021 Pages 267–292 Isbn 2662-3331 https://doi.org/10.1007/s43039-021-00035-8

Part 2 -

Question 1

```
nuke_tweets <- readtext(file = 'sentiment_nuclear_power _1_.csv')
```

```
grep(pattern = '([jJ][aA][pP][Aa][nN]\\s)', nuke_tweets$text)
```

```
## [1]   1  81 117 124 139 173
```

```
nuke_tweets[grep(pattern = '([jJ][aA][pP][Aa][nN]\\s)', nuke_tweets$text), 'text']
```

```
## [1] ":Hello Japan is a nuclear power plant crisis. {link}"
## [2] "RT @mention Devastating March 11 #earthquake leaves Japan pondering its fuel mix. See Platts The
## [3] "RT @mention In Japan the 6th graders take field trips to nuclear power plants? Gees, makes my fi
## [4] "@mention it seemed like politically speaking the US was on it's way to jumping more into nuclea
## [5] "RT @mention Not only Japan but also the nuclear power plant all over the world has a lot of prob
## [6] "RT @mention Should nuclear power be feared? Or more understood? {link} #Fukushima #energy #Germa
```

Question 2

```
nuke_tweets$edited_text <- gsub(pattern = '#\\w*', x = nuke_tweets$text, replacement = '', )

nuke_tweets$edited_text <-gsub(pattern = '@\\w*', x = nuke_tweets$edited_text, replacement = '')
```

```r
nuke_tweets$edited_text[grepl(pattern = '#\\w*', x = nuke_tweets$text) |
                          grepl(pattern = '@\\w*', x = nuke_tweets$text)][1:20]
```

```
##  [1] "(Aug 22,2011)Plant Status of Fukushima Daiichi Nuclear Power Station (as of 3:00 pm, Aug. 22) 
##  [2] "[news] Nuclear Power Plant: Radioactive Tritium Contamination in Vermont {link} "
##  [3] "{link} Nuclear Power 101 via Arnie Gundersen w    "
##  [4] "{link} What's happening  Fukushima nuclear power plant?!"
##  [5] " Ever hears talk of a place in north MI around a nuclear power plant that is bow hunting only?
##  [6] " No. It's in north west Missouri it's a WMA around a nuclear power plant. 6000 ac. My buddy wa
##  [7] " LOL The best bumper sticker I ever saw was: &quot;NUCLEAR POWER IS SAFER THAN TED KENNEDY'S C
##  [8] " - newamerica9.com is an investment website or are you talking about nuclear energy dangers?"
##  [9] "RT  : Science Saturday: Nuclear energy, melting ice caps, and human adaptation {link}"
## [10] "RT  'Hinkley Point C' - Britain's newest nuclear power station "
## [11] "  - She is a mess. I am thinking of moving back down that way after 2012. Here is another we d
## [12] "RT    Corino, your a bigger threat of death than the Limerick Nuclear Power plant...you survivi
## [13] "  Corino, your a bigger threat of death than the Limerick Nuclear Power plant...you surviving 
## [14] "RT   has &quot;open mind&quot; on nuclear energy, but is concerned about nuclear waste and nee
## [15] "RT   I want to stop the nuclear power plant all over the world now.So I shout it everyday;)So 
## [16] "  Nuclear Energy is good until something happens. Is it really worth the Risk of Radiation at 
## [17] "  The only problem is that prosperity was built on nuclear energy and not renewables, so they 
## [18] "RT   Top 10 things Indian citizens could do in case of a  power plant accident in the country 
## [19] "RT   we gotta pioneer new technologies to help us all endure the radiation, and shut down  ene
## [20] "RT  &;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;&;Russia today &; RT Gro
```

Question 3

```r
instruments <- ndjson::stream_in('Musical_Instruments_5_SSN.json')
```

```r
instruments[grepl(pattern = '\\d\\d\\d-\\d\\d-\\d\\d\\d\\d', x = instruments$reviewText)]$reviewText
```

```
## [1] "I bought this to use with my keyboard. I wasn't really aware that there were other options for 
```

```r
instruments[grepl(pattern = '\\d\\d\\d\\d\\d\\d\\d\\d\\d', x = instruments$reviewText)]$reviewText
```

```
## [1] "I bought two of these straps from a local store for thirty bucks each about two years ago. They
## [2] "Their are two issues with these strap-locks that I have come across.The first issue is with thi
## [3] "I have this paired with the Behringer C-1 Studio Condenser Microphone.  It matches the mic well
## [4] "i got this despite the warnings about it being too short for some people.  The strap is about 6
```

```r
instruments[grepl(pattern = 'SSN', x = instruments$reviewText)]$reviewText
```

```
## [1] "Hey, my SSN is 123-456-8971. They will always sound the same as the very first day, I love this
## [2] "I bought two of these straps from a local store for thirty bucks each about two years ago. They
```

```r
instruments[grepl(pattern = 'social', x = instruments$reviewText)]$reviewText
```

```
## [1] "I bought this to use with my keyboard. I wasn't really aware that there were other options for 
## [2] "Absolutely Obligatory for any fan of rock n roll jamming with tastes of jazz influenced rock--1
## [3] "This product has a quiet metronome; in a playing environment, it won't be easy to hear. The tun
## [4] "The tuner doesn't require you to hear either it or the guitar or bass to tune it, so it will wo
```

Looks like there are some dashed and non-dashed SSNs. Also, if I just do the non-dash, I might accidentally remove digits from hyperlinks. Maybe if I enforce that there must be a space before and after that will help

```r
instruments$edited_review <- gsub(pattern = '\\s\\d{9}', x = instruments$reviewText, replacement = ' ', 
```

```
instruments$edited_review <- gsub(pattern = '\\s\\d{3}-\\d{2}-\\d{4}', x = instruments$edited_review, re


instruments[
  grepl(pattern = '\\s\\d{9}', x = instruments$reviewText) |
    grepl(pattern = '\\s\\d{3}-\\d{2}-\\d{4}', x = instruments$reviewText)
    ]$edited_review
```

```
## [1] "I bought this to use with my keyboard. I wasn't really aware that there were other options for I
## [2] "I bought two of these straps from a local store for thirty bucks each about two years ago. They
```