# LOWES_Q42019_NLP

Isabel Villa Mendoza De Guzman

2024-03-09

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(textdata)
library(tidytext)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(RColorBrewer)
```

```
#Load the Text File
lowes_text <- read_file("LOWQ42019.txt")
```

```
#Check the number of characters in the file
nchar(lowes_text)
```

```
## [1] 60561
```

```r
summary(lowes_text)
```

```
##    Length     Class      Mode
##         1 character character
```

```r
# view
lowes_text %>% stringr::str_sub(1, 500)
```

```
## [1] "Lowe’s Companies Inc (NYSE:LOW)\r\nQ4 2019 Earnings Call\r\nFeb 26, 2020, 9:00 a.m. ET\r\nConte
```

```r
#Count Characters
lowes_text %>% stringr::str_count()
```

```
## [1] 60344
```

```r
#Look for Keywords
lowes_text %>% stringr ::str_detect(c('sales', 'growth', 'Pro', 'increase'))
```

```
## [1] TRUE TRUE TRUE TRUE
```

```r
#Count the number of matches of a substring
lowes_text %>% stringr :: str_count('growth')
```

```
## [1] 32
```

```r
# Where is this keyword mentioned?
lowes_text %>% stringr :: str_locate_all('growth')
```

```
## [[1]]
##       start   end
##  [1,]  1998  2003
##  [2,]  2412  2417
##  [3,]  3070  3075
##  [4,]  3345  3350
##  [5,]  3369  3374
##  [6,]  3744  3749
##  [7,]  4205  4210
##  [8,]  4381  4386
##  [9,]  4883  4888
## [10,]  4976  4981
## [11,]  5642  5647
## [12,]  5976  5981
## [13,]  7750  7755
## [14,]  9305  9310
## [15,] 10498 10503
## [16,] 16482 16487
## [17,] 25180 25185
## [18,] 27690 27695
## [19,] 27825 27830
```

```
## [20,] 28167 28172
## [21,] 28223 28228
## [22,] 29960 29965
## [23,] 32780 32785
## [24,] 35638 35643
## [25,] 39862 39867
## [26,] 49113 49118
## [27,] 51660 51665
## [28,] 55436 55441
## [29,] 55987 55992
## [30,] 56453 56458
## [31,] 56523 56528
## [32,] 56606 56611
```

```
#View surrounding text (requires regex)
lowes_text %>% stringr:: str_extract_all(".{20}(growth).{20}")
```

```
## [[1]]
##  [1] "ively balanced with growth of positive 2.8% in"
##  [2] "ured in-store sales growth more than we antici"
##  [3] "s.com lagged market growth, delivering comp gr"
##  [4] " slowed our dot com growth to resolve those is"
##  [5] "ivering a 2.6% comp growth almost exclusively "
##  [6] "o unlock additional growth when Lowe's.com sal"
##  [7] "rivers of our sales growth in the fourth quart"
##  [8] " for our U.S. sales growth with our Pro comps "
##  [9] "ce and drive future growth with this critical "
## [10] "we saw a consistent growth across the business"
## [11] "ed comparable sales growth in the fourth quart"
## [12] "nt comparable sales growth of 2.6% in the four"
## [13] "h double-digit comp growth supported by streng"
## [14] "only drove top-line growth, but also positione"
## [15] "support our healthy growth in 2020. A strong c"
## [16] "pport for continued growth in home improvement"
## [17] "g high-single-digit growth, which will provide"
## [18] "ntribution to sales growth. For 2020, we expec"
## [19] "ared comments, a 3% growth underperformed in a"
## [20] " a lot of inventory growth, much better in-sto"
## [21] "f confidence in the growth of the business bas"
## [22] "ers, average ticket growth, frequency of visit"
## [23] "re brick and mortar growth of the business? An"
## [24] "have more sustained growth in our productivity"
## [25] "otable step down in growth relative to where t"
## [26] "re seeing continued growth in luxury vital pla"
## [27] "e lines -- signs of growth now in our soft flo"
## [28] "years. We're seeing growth in new products. So"
```

```
#Change to a tibble (tidy dataframe)
#Put all the text from mda file to token1
lowes_token1 <- tibble(lowes_text)

# Tokenize
lowes_token1 <- lowes_token1 %>% tidytext::unnest_tokens(output=word, input=lowes_text, token='words',
```

```r
view(lowes_token1)

# add order of the words
lowes_token1 <- lowes_token1 %>% mutate(order = row_number())

#count tokens
lowes_token1 %>% nrow()
```

```
## [1] 10024
```

```r
#first few words
lowes_token1[1:15,]
```

```
## # A tibble: 15 x 2
##    word      order
##    <chr>     <int>
##  1 lowe's        1
##  2 companies     2
##  3 inc           3
##  4 nyse:low      4
##  5 q4            5
##  6 2019          6
##  7 earnings      7
##  8 call          8
##  9 feb           9
## 10 26           10
## 11 2020         11
## 12 9            12
## 13 00           13
## 14 a.m          14
## 15 et           15
```

```r
# count the number of matches of a substring
lowes_token1 %>% dplyr::filter(word == str_sub('growth')) %>% count()
```

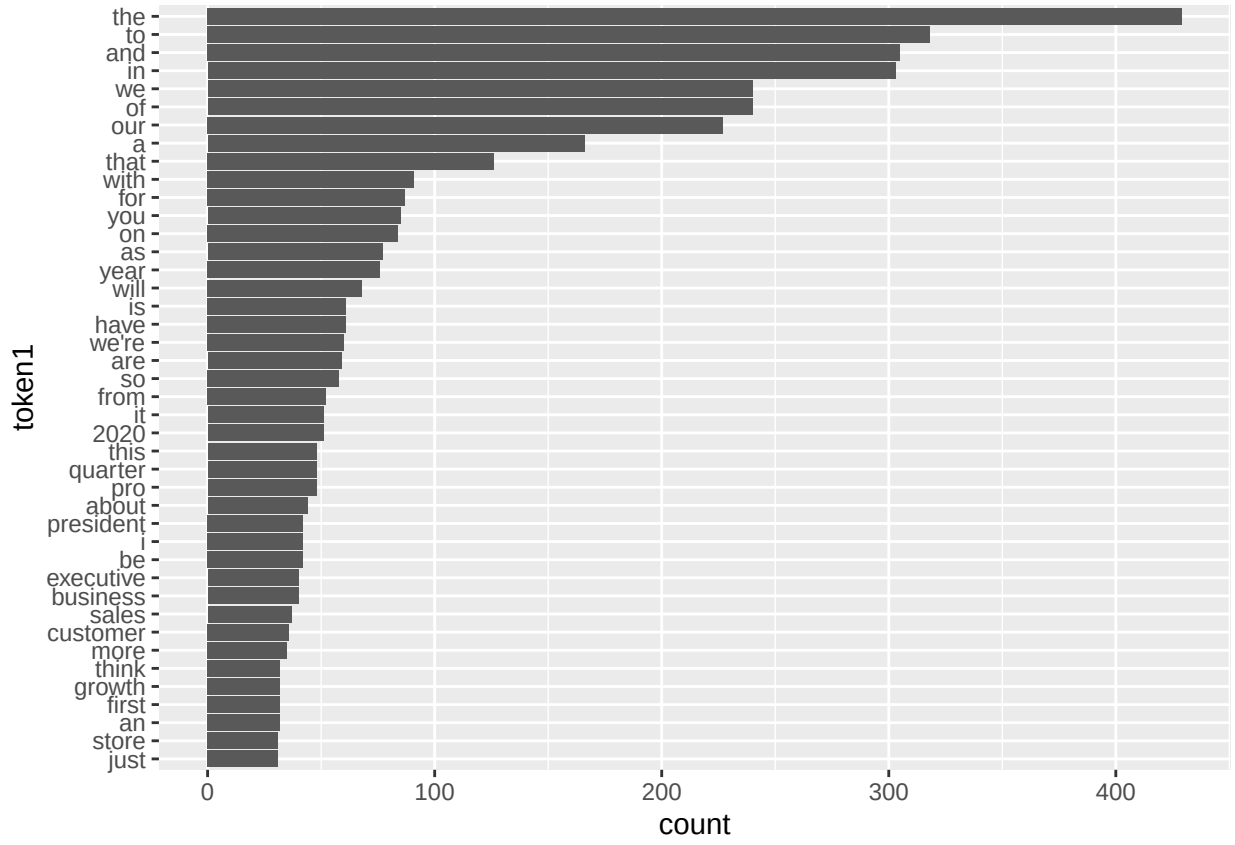
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    32
```

```r
#Where is this keyword mentioned?
lowes_token1 %>% dplyr::filter(word == str_sub('growth'))
```

```
## # A tibble: 32 x 2
##    word   order
##    <chr>  <int>
##  1 growth   294
##  2 growth   362
##  3 growth   467
##  4 growth   505
```

```
##  5 growth   508
##  6 growth   575
##  7 growth   655
##  8 growth   680
##  9 growth   764
## 10 growth   782
## # i 22 more rows
```

```r
#remove stop words
# Look at the most important frequent words
lowes_token1 %>%
  group_by(word) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  filter(count > 30) %>%
  mutate(token1 = reorder(word, count)) %>%
  ggplot(aes(x=count, y=token1)) +
  geom_col()
```




```r
#Load custom stopwords
custom_stop_words <- read_csv("stop_words_list.csv", col_names = FALSE)
```

```
## Rows: 1087 Columns: 1
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (1): X1
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
view(custom_stop_words)
```

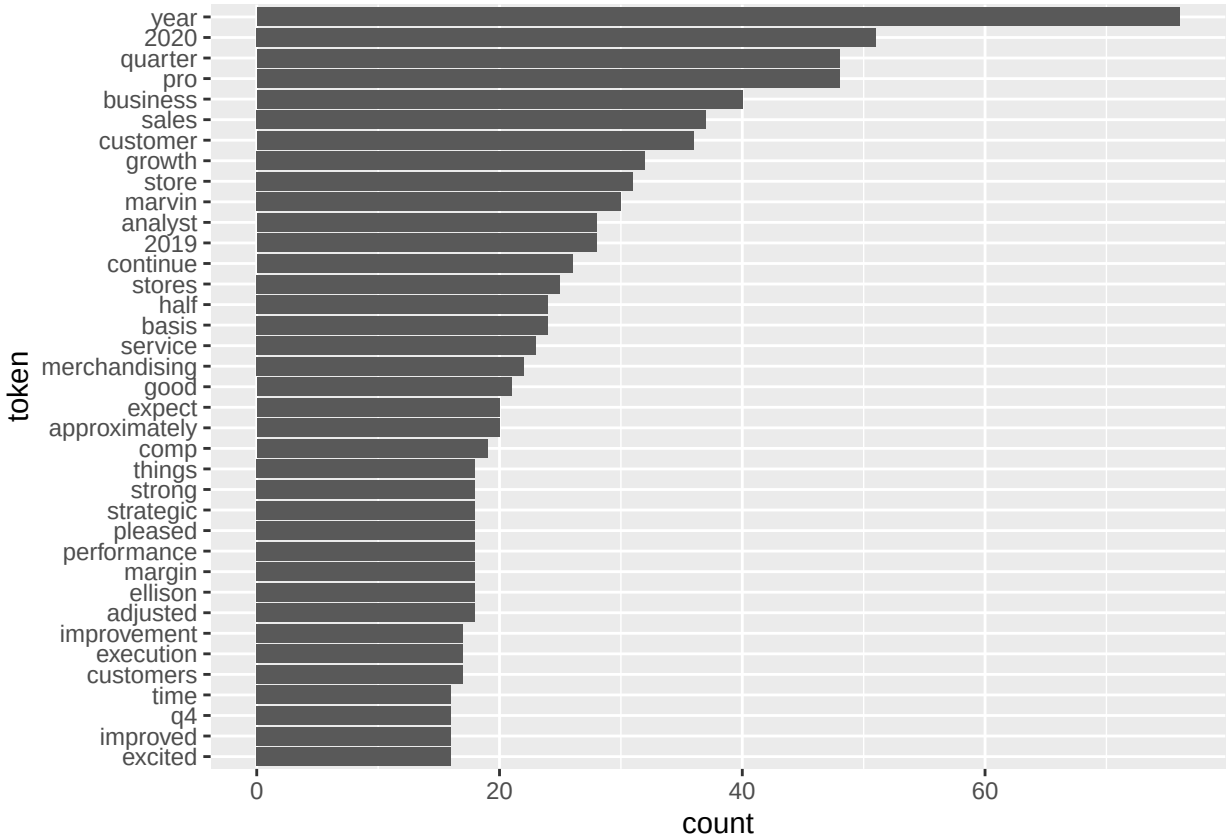
```
library(dplyr)

#remove stop words
lowes_token1 <- lowes_token1 %>%
  anti_join(custom_stop_words, by = c('word'='X1'))

lowes_token1 %>% nrow()
```

```
## [1] 4428
```

```
lowes_token1 %>%
  group_by(word) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  filter(count > 15) %>%
  mutate(token = reorder(word, count)) %>%
  ggplot(aes(x=count, y=token)) +
  geom_col()
```



```r
# look at similar words
arrange(lowes_token1, word)[1:274,1]
```

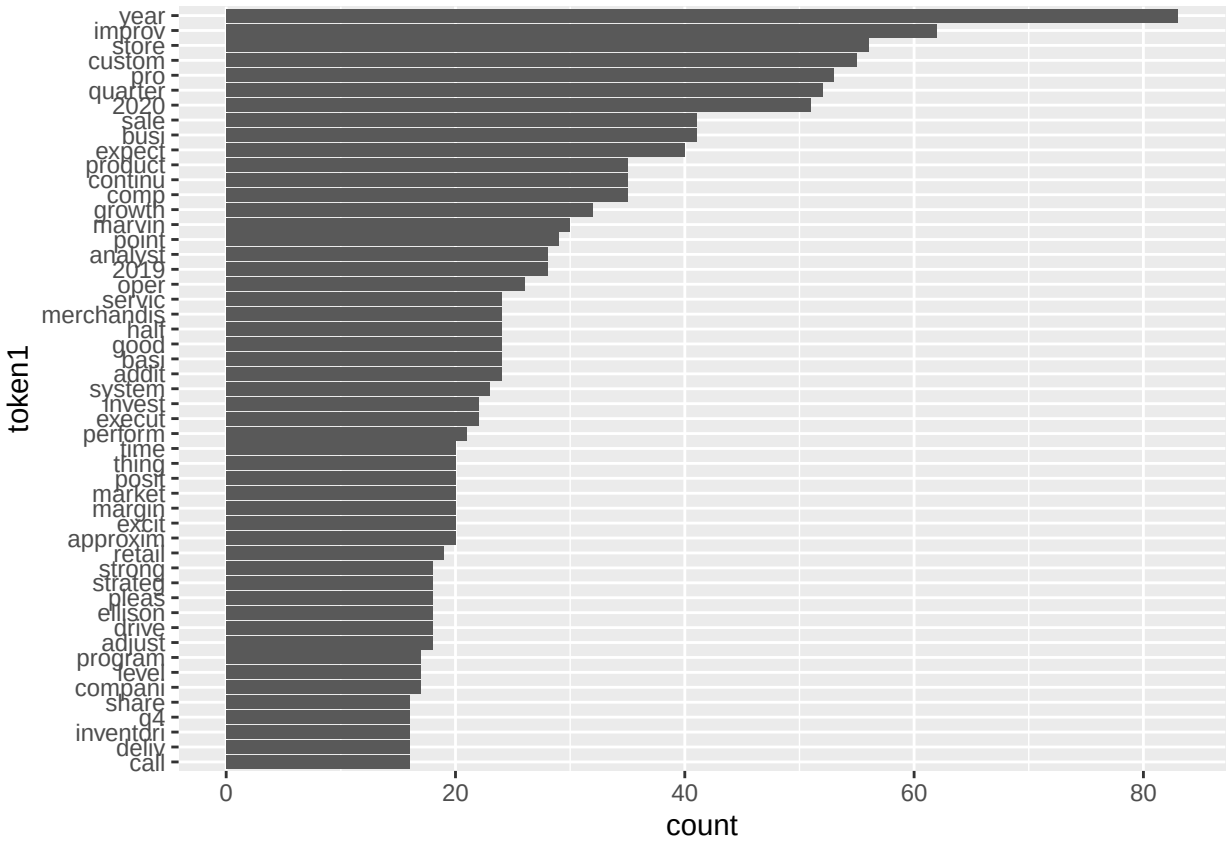
```
## # A tibble: 274 x 1
##    word
##    <chr>
##  1 0.5
##  2 0.66
##  3 0.7
##  4 0.94
##  5 0.94
##  6 00
##  7 1.5
##  8 1.6
##  9 1.7
## 10 1.8
## # i 264 more rows
```

```r
#Stem the tokens
stemmed <- lowes_token1 %>% mutate(stem = SnowballC::wordStem(word))
view(stemmed)

# look at similar words now
arrange(stemmed, word)[1:274,1]
```

```
## # A tibble: 274 x 1
##    word
##    <chr>
##  1 0.5
##  2 0.66
##  3 0.7
##  4 0.94
##  5 0.94
##  6 00
##  7 1.5
##  8 1.6
##  9 1.7
## 10 1.8
## # i 264 more rows
```

```r
stemmed %>%
  group_by(stem) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  filter(count > 15) %>%
  mutate(token1 = reorder(stem, count)) %>%
  ggplot(aes(x=count, y=token1)) +
  geom_col()
```




```
# Key words
set.seed(77)
 library(wordcloud)
stemmed %>%
  group_by(word) %>%
  summarize(count = n()) %>%
  with(wordcloud(words=word, freq=count, min.freq=1, max.words=70, random.order=F, rot.per=0.30, colors=
```


```
# load finance sentiment list and explore it
lm_dict <- tidytext::get_sentiments('loughran')

# view dictionary
lm_dict %>% group_by(sentiment) %>% summarize(count = n())
```

```
## # A tibble: 6 x 2
##   sentiment    count
##   <chr>        <int>
## 1 constraining   184
## 2 litigious      904
## 3 negative      2355
## 4 positive       354
## 5 superfluous     56
## 6 uncertainty    297
```

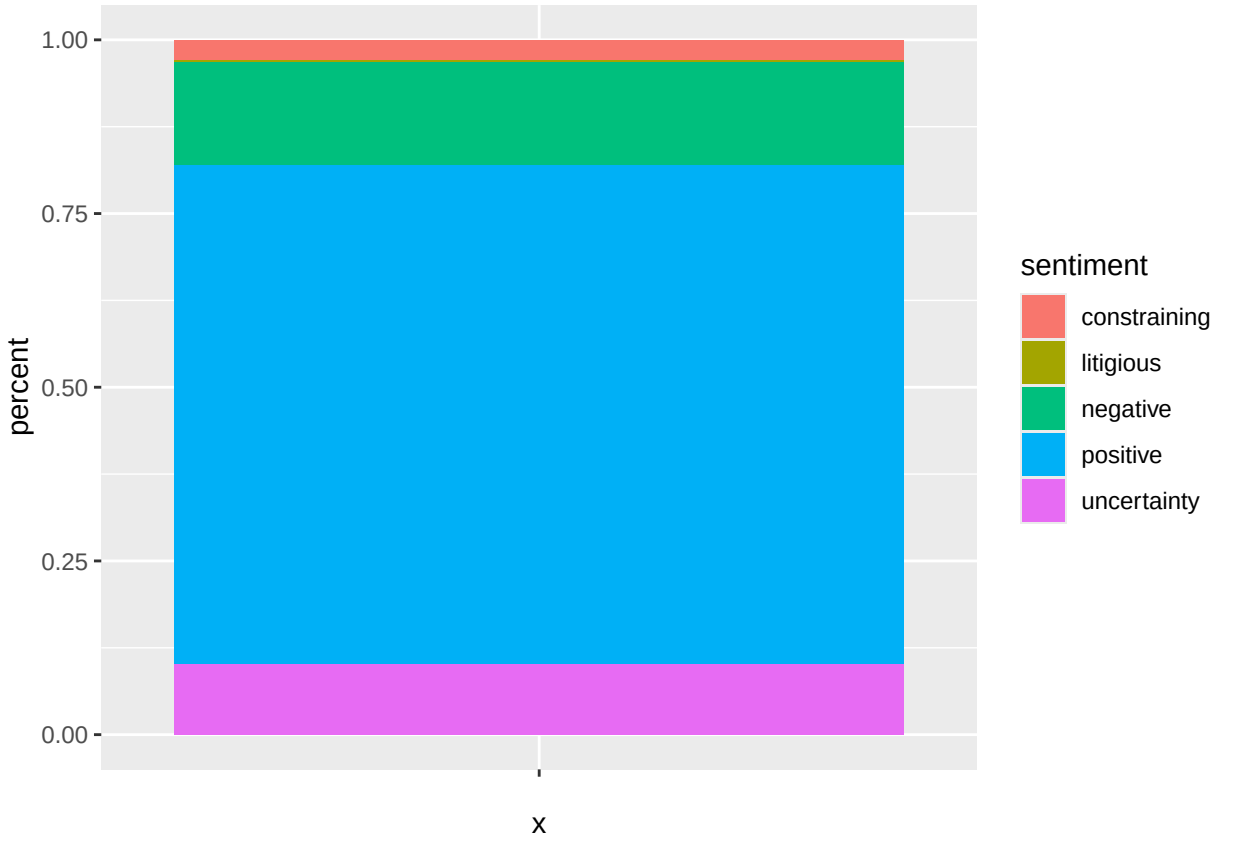
```
# Add sentiment
sentimented <- stemmed %>%
  inner_join(lm_dict, by = 'word')
```

```
## Warning in inner_join(., lm_dict, by = "word"): Detected an unexpected many-to-many relationship betw
## i Row 80 of `x` matches multiple rows in `y`.
## i Row 2906 of `y` matches multiple rows in `x`.
## i If a many-to-many relationship is expected, set `relationship =
##   "many-to-many"` to silence this warning.
```

```r
# Explore totals
sentimented %>%
  group_by(sentiment) %>%
  summarize(count = n(), percent = count/nrow(sentimented))
```

```
## # A tibble: 5 x 3
##   sentiment    count percent
##   <chr>        <int>   <dbl>
## 1 constraining    11 0.0296
## 2 litigious        1 0.00269
## 3 negative        55 0.148
## 4 positive       267 0.718
## 5 uncertainty     38 0.102
```

```r
sentimented %>%
  group_by(sentiment) %>%
  summarize(count = n(), percent = count/nrow(sentimented)) %>%
  ggplot(aes(x='', y=percent, fill=sentiment)) +
  geom_bar(width=1, stat='identity')
```




# Sentiment Analysis Uncertainty at 11%, Positive at 60%, Negative at 15%, Litigious at 10%, Constraining at 4%.