Tidy Tuesday

PMAP 8551, Spring 2025

ARYA DEVKOTA

March 29, 2025

Table of contents

#Task 1: Loading Data

```
library(tidyverse)
library(ggplot2)
library(tidytext)
library(ggtext)
library(syuzhet)
library(wordcloud2)
library(ggrepel)
library(webshot2)
library(htmlwidgets)
AMZannualreport <- read_csv("report_words_clean.csv")
```

#Task 2: Cleaning Data

```
AMZannual_filtered <- AMZannualreport |>
  filter(year >= 2005 & year <= 2015) |>
  mutate(word = as.character(word)) |>
  group_by(year, word) |>
  summarize(frequency = n(), .groups = "drop")
#Here, I'm simply finding the total frequencies of each word.


word_flags <- AMZannual_filtered |>
  group_by(word) |>
  summarise(entries_acrossyears = n_distinct(year) > 1, .groups = "drop")
```

```
AMZannual_filtered <- AMZannual_filtered |>
  left_join(word_flags, by = "word")
#I am now creating a seperate tibble that displays a TRUE if the word itself
appears in multiple years,
#and FALSE if it doesn't. For example, if the word "amazon" appears in both
2014 and 2015,
#both the entry for 2014 and the entry for 2015 would display a "TRUE", as it
is found in multiple years.



#Finally, I want to now create two columns with the syuzhet package; one
calculating the sentiment score,
#and one categorizing the sentiment score by one of four categories with
criteria spanning
#from -1 to 1.
AMZreport_sentiment <- AMZannual_filtered |>
  mutate(
    sentiment_score = get_sentiment(word, method = "syuzhet"),
    sentiment_category = case_when(
      sentiment_score > 0 & sentiment_score < 0.5  ~ "Pleasant",
      sentiment_score >= 0.5 & sentiment_score <= 1  ~ "Positive",
      sentiment_score > -0.5 & sentiment_score < 0  ~ "Negative",
      sentiment_score >= -1 & sentiment_score <= -0.5 ~ "Concerning",
      TRUE ~ "Neutral"
    )
  )
```

#Task 3: Visualization #1

```
word_freq <- AMZreport_sentiment |>
  select(word, frequency) |>
  distinct()


wc_amzannual <- wordcloud2(data=word_freq, size=1.6)
```

#Task 4: Visualization #2

```
#Here, I am manually selecting the top 10 frequently-appearing neutral words
and top 10 words of
#all of my other sentiment categories, in order to visualize more 'diverse'
data from my tibble.
#If I were to simply slice the first 25 words without any specifications as
I've done below,
#the majority of the data I'd be visualizing would be 'Neutral'.
amz_graph <- AMZreport_sentiment |>
```

```
   filter(year %in% c(2005, 2015))

top_neutral_words <- amz_graph |>
  filter(sentiment_category == "Neutral") |>
  group_by(year) |>
  arrange(desc(frequency)) |>
  slice_head(n = 10) |>
  ungroup()

top_additional_words <- amz_graph |>
  filter(sentiment_category %in% c("Positive", "Negative", "Concerning",
"Pleasant")) |>
  group_by(year, sentiment_category) |>
  arrange(desc(frequency)) |>
  slice_head(n = 10) |>
  ungroup()

top_words <- bind_rows(top_neutral_words, top_additional_words) |>
    spread(key = year, value = frequency) |>
  mutate(
    frequency_diff = `2015` - `2005`,
    label_color = case_when(
      !is.na(frequency_diff) & frequency_diff > 0 ~ "blue",
      !is.na(frequency_diff) & frequency_diff < 0 ~ "#fb5500",
      is.na(`2005`) | is.na(`2015`) ~ "black",
      TRUE ~ "black"
    )) |>
  gather(key = "year", value = "frequency", `2005`, `2015`)

WordFreqSentPlot <-
ggplot(top_words, aes(x = sentiment_score, y = frequency, color =
sentiment_category)) +
  geom_point(alpha = 0.8, size = 2) +  # Adjusted point size
  geom_text_repel(aes(label = word), size = 3, max.overlaps = 30,
                  box.padding = 0.3, point.padding = 0.5, segment.color =
"black", color = top_words$label_color, segment.linetype="dotted") +
  scale_x_continuous(
    limits = c(-1, 1),
    breaks = seq(-1, 1, by = 0.2)
  ) +
   scale_y_continuous(
    limits = c(0, 450),
    breaks = seq(0, 450, by = 75),
    expand = expansion(mult = c(0.05, 0.1))) +
  labs(
    title = "Comparing Word Frequency by Sentiment Score (2005 vs. 2015)",
    subtitle = "Based on textual data pulling from Amazon Annual Report(s).
    <br>Of those that appear across annual reports, which words have
    <span style='color:#fb5500;'>decreased</span> in frequency overtime?<br>
    Which have <span style='color:blue;'>increased</span>?",
```

```
    x = "Sentiment Score",
    y = "Word Frequency",
    size = "Frequency",
    color = "Sentiment Category"
  ) +
  facet_grid(~year) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 25, face = "bold"),
    plot.subtitle = element_markdown(size = 18),
    axis.title.x = element_text(size = 18, face = "bold"),
    axis.title.y = element_text(size = 18, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1, size=11),
    axis.text.y = element_text(angle = 0, hjust = 1, size=11),
    strip.text = element_text(size = 16, face = "bold"),
    strip.background = element_rect(color = "black", size = 0.5),
    panel.border = element_rect(color = "black", size = 0.5)
  )
```

#Save all Visualizations

```
ggsave("words_freq-sent_amz.svg", plot = WordFreqSentPlot, device = "svg",
width = 17, height = 7)
ggsave("words_freq-sent_amz.pdf", plot = WordFreqSentPlot, width = 17, height
= 7)

saveWidget(wc_amzannual,"wordcloud.html")
#I have to save it as an HTML first in order to snapshot my
#wordcloud as a raster file.
webshot("wordcloud.html", file = "wc_amzannual.png")

webshot("wordcloud.html", file = "wc_amzannual.pdf")
```