HW11_Markdown

2024-04-25

Manay Divatia

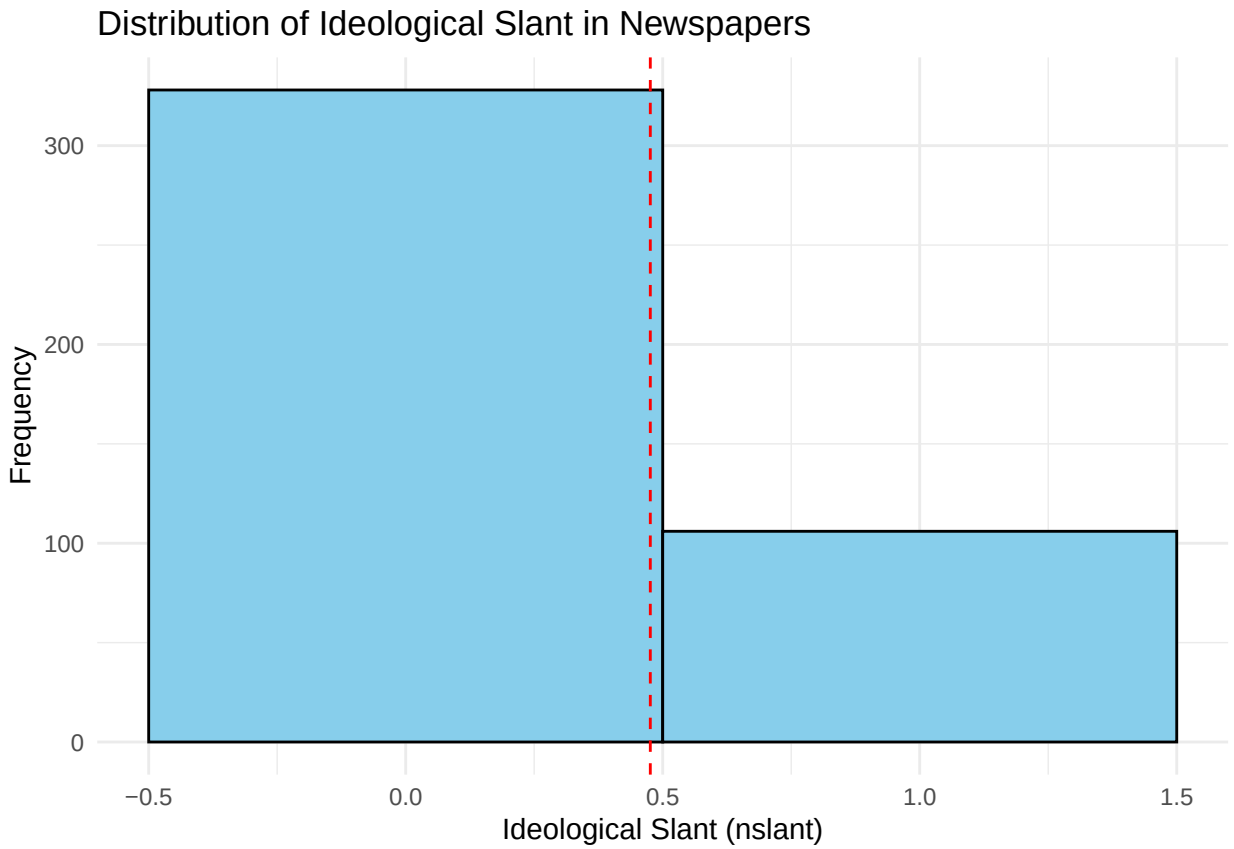
md46245

a

```
load("~/Downloads/newspapers.RData")
library(ggplot2)
str(papers)
```

```
## 'data.frame':    434 obs. of  6 variables:
##  $ newsid  : int  10 11 12 19 22 28 30 33 34 36 ...
##  $ paper   : chr  "Anchorage Daily News" "Fairbanks Daily News-Miner" "Juneau Empire" "The Anniston S
##  $ city    : chr  "Anchorage" "Fairbanks" "Juneau" "Anniston" ...
##  $ state   : chr  "AK" "AK" "AK" "AL" ...
##  $ nslant  : num  0.518 0.546 0.522 0.448 0.465 ...
##  $ district: num  NA NA NA NA NA NA NA NA NA NA ...
```

```
ggplot(papers, aes(x = nslant)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  geom_vline(aes(xintercept = median(nslant)), color = "red", linetype = "dashed") +
  labs(title = "Distribution of Ideological Slant in Newspapers",
       x = "Ideological Slant (nslant)",
       y = "Frequency") +
  theme_minimal()
```




```
left_most_slant <- papers[which.min(papers$nslant), ]
right_most_slant <- papers[which.max(papers$nslant), ]

cat("Newspaper with the largest left-wing slant:", left_most_slant$paper, "\n")
```

```
## Newspaper with the largest left-wing slant: Chicago Defender
```

```
cat("Newspaper with the largest right-wing slant:", right_most_slant$paper, "\n")
```

```
## Newspaper with the largest right-wing slant: The Daily Sentinel
```

The newspaper with the largest left-wing slant was Chicago Defender. The newspaper with the largest right-wing slant was The Daily Sentinel.

b

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
terms <- dtm$dimnames$Terms
#freq <- colSums(as.matrix())
#dtm_df <- data.frame(terms = terms, freq = dtm$ncol)
#dtm_df <- dtm_df[order(-dtm_df$freq), ]

#wordcloud(words = dtm_df$terms, freq = dtm_df$freq, max.words = 20)
#left_most_newspapers <- papers[order(papers$nslant)][1:round(nrow(papers)/10), ]
#right_most_newspapers <- papers[order(papers$nslant, decreasing = TRUE)][1:round(nrow(papers)/10), ]
#dtm_df <- as.data.frame(dtm)
```

```
#left_most_dtm <- dtm_df[rownames(dtm_df) %in% left_most_newspapers$newsid, ]
#right_most_dtm <- dtm_df[rownames(dtm_df) %in% right_most_newspapers$newsid, ]

#wordcloud(words = rownames(left_most_dtm), freq = colSums(left_most_dtm), max.words = 20)
#wordcloud(words = rownames(right_most_dtm), freq = colSums(right_most_dtm), max.words = 20)
```

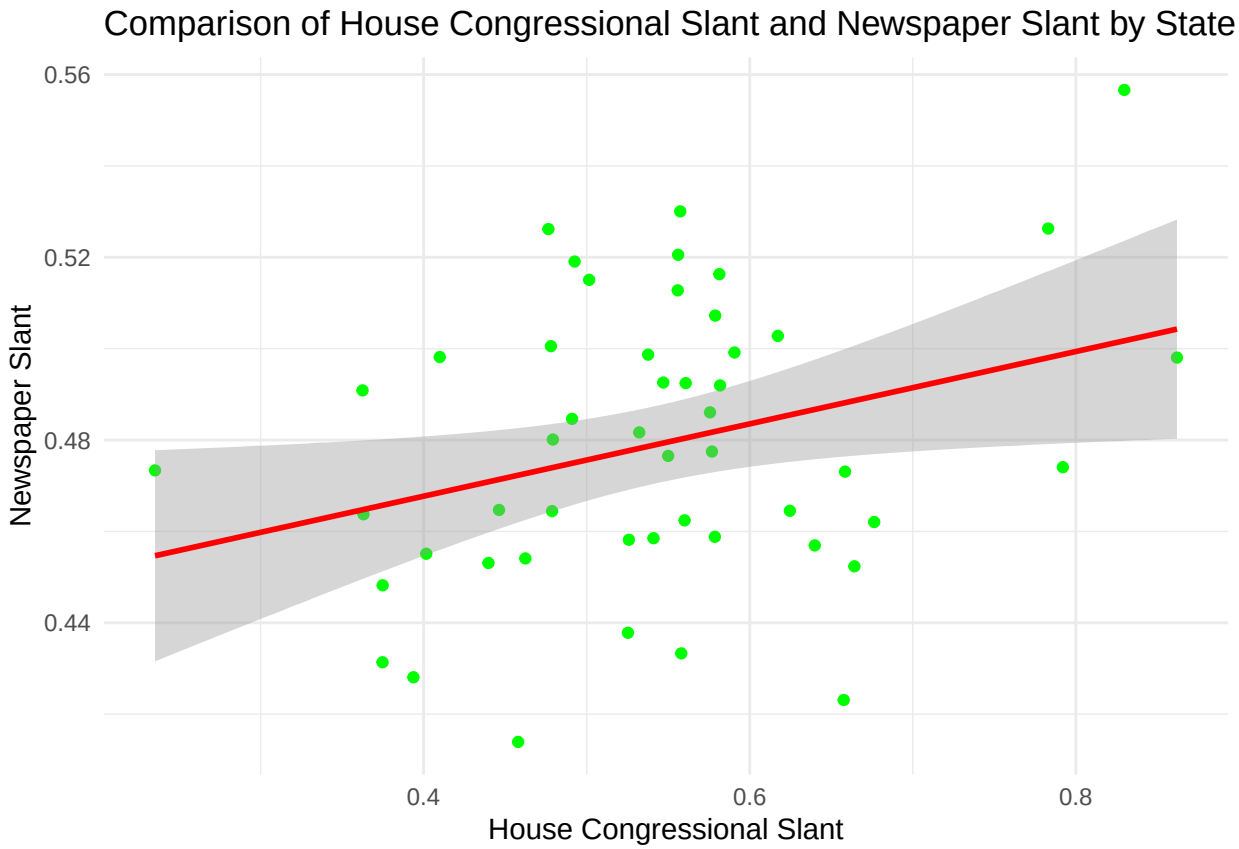
A lot of the words differ because they talk about different things. However, the general tone of the topics are more or less the same, it is just the specific words that are completely different.

c

```
house_avg_slant <- aggregate(cslant ~ state, data = subset(cong, chamber == "H"), FUN = mean)
senate_avg_slant <- aggregate(cslant ~ state, data = subset(cong, chamber == "S"), FUN = mean)
newspaper_avg_slant <- aggregate(nslant ~ state, data = papers, FUN = mean)

ggplot() +
  geom_point(data = house_avg_slant, aes(x = cslant, y = newspaper_avg_slant$nslant[2:51]), color = "gr
  geom_smooth(data = house_avg_slant, aes(x = cslant, y = newspaper_avg_slant$nslant[2:51]), method = "
  labs(title = "Comparison of House Congressional Slant and Newspaper Slant by State",
       x = "House Congressional Slant",
       y = "Newspaper Slant") +
  theme_minimal()
```

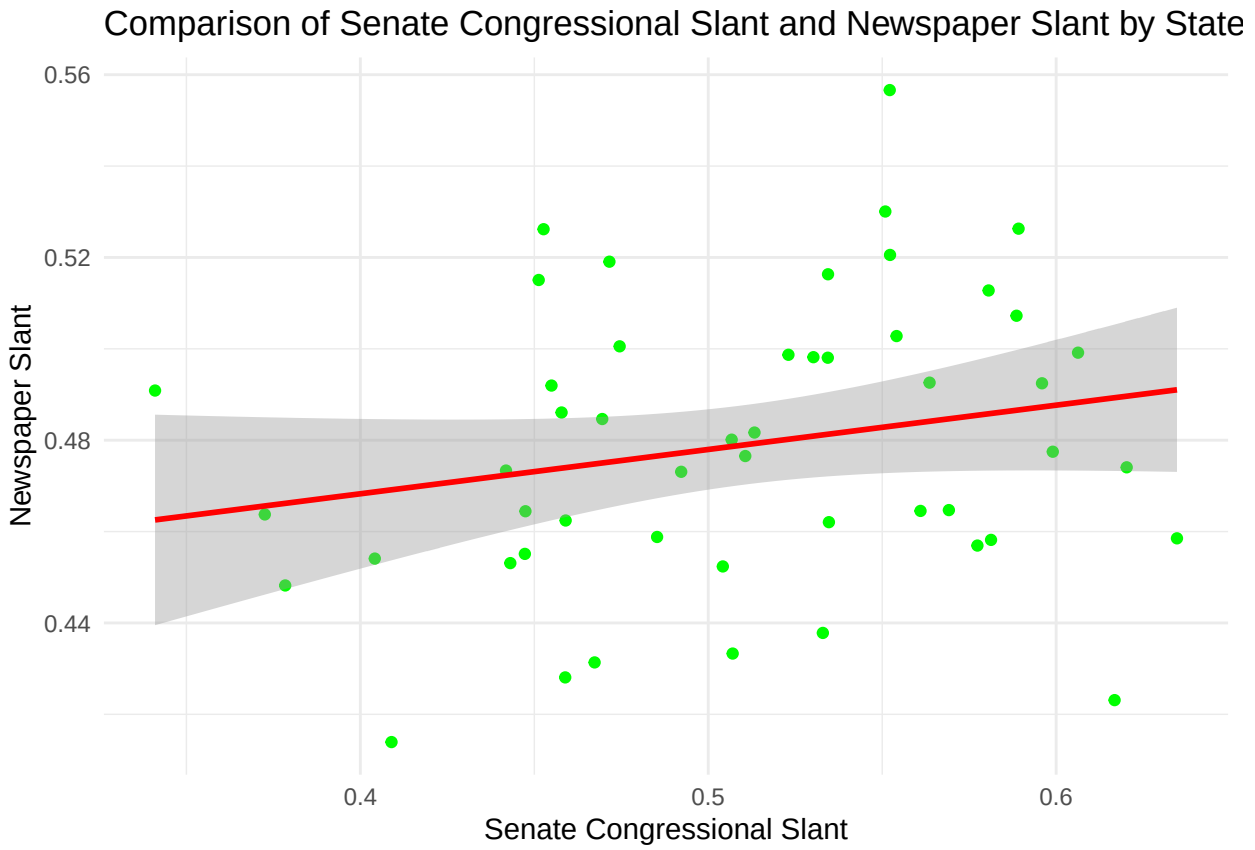
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot() +
  geom_point(data = senate_avg_slant, aes(x = cslant, y = newspaper_avg_slant$nslant[2:51]), color = "g
```

```
  geom_smooth(data = senate_avg_slant, aes(x = cslant, y = newspaper_avg_slant$nslant[2:51]), method =
  labs(title = "Comparison of Senate Congressional Slant and Newspaper Slant by State",
       x = "Senate Congressional Slant",
       y = "Newspaper Slant") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



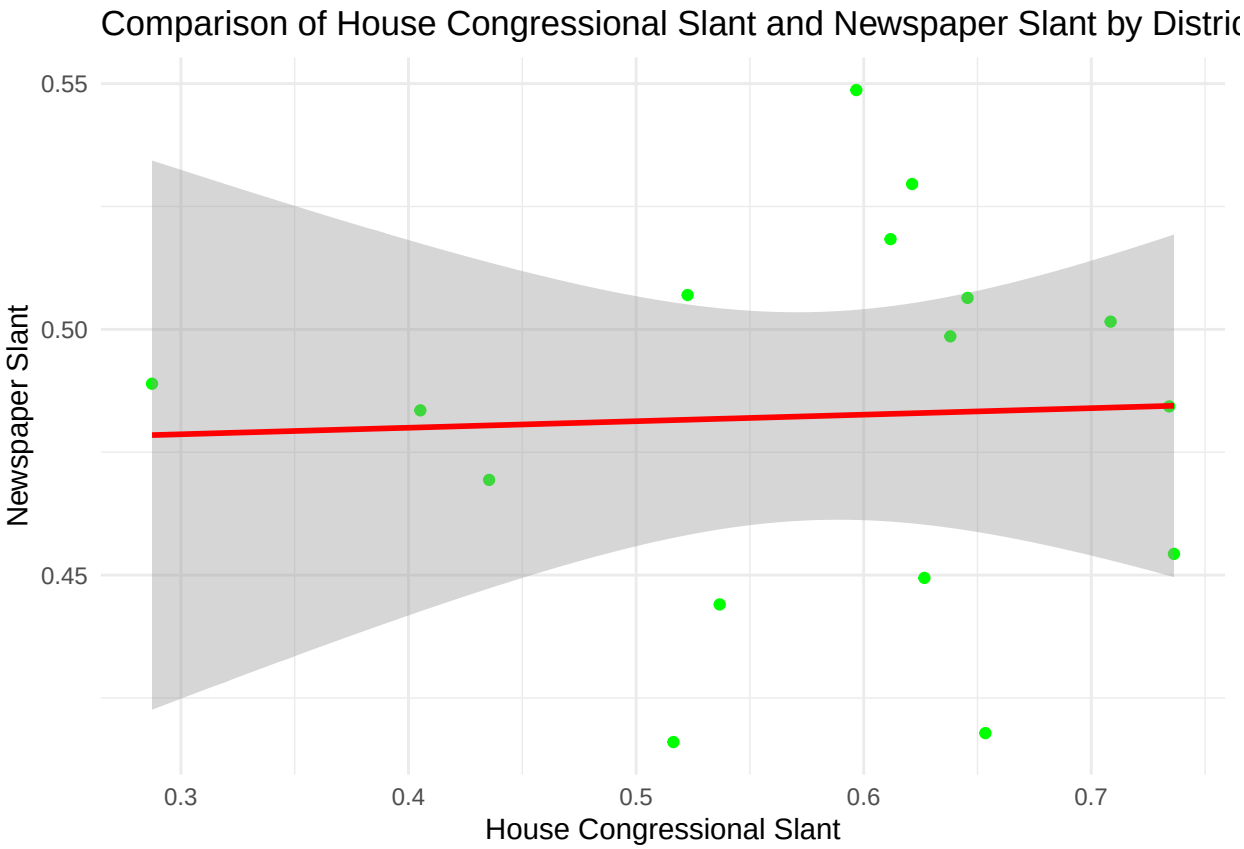
From the plots, I don't think we can make any conclusive claims about whether or not the newspapers are directly influenced by the political language of elected officials. While, the regression shows a positive correlation we can see a large variance band which tells us how unsure we are about the regression line.

d

```
cong_tx <- subset(cong, state == "TX")
papers_tx <- subset(papers, state == "TX")
merged_tx <- merge(cong_tx, papers_tx, by = c("district", "state"))
house_avg_slant_tx <- aggregate(cslant ~ district, data = subset(merged_tx, chamber == "H"), FUN = mean)
newspaper_avg_slant_tx <- aggregate(nslant ~ district, data = merged_tx, FUN = mean)

ggplot() +
  geom_point(data = house_avg_slant_tx, aes(x = cslant, y = newspaper_avg_slant_tx$nslant), color = "gr
  geom_smooth(data = house_avg_slant_tx, aes(x = cslant, y = newspaper_avg_slant_tx$nslant), method = "
  labs(title = "Comparison of House Congressional Slant and Newspaper Slant by District in Texas",
       x = "House Congressional Slant",
       y = "Newspaper Slant") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



These results tell a different story but it isn't a conclusive story. The variance is very high so I think we can't make any definitive conclusions based on the results of this graph.

e

```
#library(tm)
#dtm_tdm <- TermDocumentMatrix(dtm)

#tfidf <- weightTfIdf(dtm_tdm)
#tfidf_df <- as.data.frame(as.matrix(tfidf))

#home_news_tribune <- tfidf_df[rownames(tfidf_df) == "Home News Tribune", ]

#top_terms <- sort(home_news_tribune, decreasing = TRUE)[1:10]
#print(top_terms)
```