Hackathon: Wall Street Journal Sentiment Analysis

Intro

In a sentiment analysis for the Wall Street Journal for the 2018 to early 2020 time period, we try four different methods for processing sentiment as positive, neutral, or negative over time as based on scores returned from these methods.

Synergy Report

Our team demonstrated great synergy for this assignment. The team discussed the plan of action during class, then the members coordinated virtually. Wen Si and Kiran Jaura generated the initial concept for the project. They subsequently shared their thoughts with Glen Cooper and Marcus Sianan, who both consented because they found the topic/concept to be interesting, important, and relevant to the course. Wen was most responsible for the design of the product. Kiran put the most effort into refining and improving upon the structure that Wen devised, and consulted with Glen and Marcus throughout the process (e.g., ensuring that the visualizations were as accessible and interpretable as possible for the audience). As team coordinator, Kiran handled much of the group communication, and was assisted by Glen and Marcus. Specifically, all members interacted via a series of emails in order to make sure that everyone was on the same page and that there was unanimous consent to all modifications of the code. Lastly, Marcus drafted the synergy report and it was reviewed by all members of the team. Their input was incorporated into the final draft. Overall, the group had a positive, collaborative experience.

Data Preparation

```r
rm(list = ls()) # clear the environment
setwd("C:/Users/19728/Desktop/EPPS 6356/Wen_Files")

# load packages
library(tm)
```

```
Loading required package: NLP
```

```r
library(NLP)
#library(austin)
library(ggplot2)
```

```

Attaching package: 'ggplot2'
The following object is masked from 'package:NLP':

    annotate
```

```r
library(readr)
library(tidytext)
library(dplyr)
```

```

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
library(tidyr)
library(lubridate)


Attaching package: 'lubridate'
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
library(quanteda)

Warning in .recacheSubclasses(def@className, def, env): undefined subclass
"unpackedMatrix" of class "mMatrix"; definition not updated
Warning in .recacheSubclasses(def@className, def, env): undefined subclass
"unpackedMatrix" of class "replValueSp"; definition not updated
Package version: 3.2.3
Unicode version: 13.0
ICU version: 69.1
Parallel computing: 8 of 8 threads used.
See https://quanteda.io for tutorials and examples.

Attaching package: 'quanteda'
The following object is masked from 'package:tm':

    stopwords
The following objects are masked from 'package:NLP':

    meta, meta<-
library(quanteda.textmodels)
library(quanteda.textplots)
library(stm)

stm v1.3.6 successfully loaded. See ?stm for help.
 Papers, resources, and other materials at structuraltopicmodel.com
library(syuzhet)

#load data

wsj_data <- read_csv("ISDSA_WSJ DATA 1.csv")

Rows: 5367 Columns: 6
── Column specification
────────────────────────────────────────────────────────
Delimiter: ","
chr (5): Title, Authors, documentType, Month, text
dbl (1): Year

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.
doc.corpus <- corpus(wsj_data)
```

```r
#summary(doc.corpus)

# remove stop words
my.dfm <- dfm(doc.corpus, remove_numbers = TRUE, remove_punct = TRUE,
remove_symbols = TRUE, remove = stopwords("english"))
```

```
Warning: 'dfm.corpus()' is deprecated. Use 'tokens()' first.
Warning: '...' should not be used for tokens() arguments; use 'tokens()'
first.
Warning: 'remove' is deprecated; use dfm_remove() instead
```

```r
#my.dfm


###### Sentiment Analysis by Time#####
library(readxl)
wsj_data1 <- read_excel("ISDSA_WSJ DATA.xlsx")
wsj_txt <- iconv(wsj_data1$Text)

wsj_sent <- get_nrc_sentiment(wsj_txt)
```

```
Warning: `spread_()` was deprecated in tidyr 1.2.0.
Please use `spread()` instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
generated.
```

```r
wsj_szvector = get_sentiment(wsj_data1$Text, method =  "syuzhet")

# get sentiments

wsj_bvector = get_sentiment(wsj_data1$Text, method =  "bing")
wsj_afvector = get_sentiment(wsj_data1$Text, method =  "afinn")
wsj_nrcvector = get_sentiment(wsj_data1$Text, method =  "nrc", lang
="english")


attach(wsj_data1)
library(lubridate)
wsj_data$date = mdy(wsj_data1$entryDate)
```

```
Warning: 3 failed to parse.
```

```r
new_b<-cbind(wsj_bvector, "Bing")
new_b<-cbind.data.frame(new_b,wsj_data$date)
colnames(new_b)<-c("Score","Method","Date")
new_afv<-cbind(wsj_afvector,"Affin")
new_afv<-cbind.data.frame(new_afv,wsj_data$date)
colnames(new_afv)<-c("Score","Method","Date")
new_nrc<-cbind(wsj_nrcvector,"NRC")
new_nrc<-cbind.data.frame(new_nrc,wsj_data$date)
colnames(new_nrc)<-c("Score","Method","Date")
new_sz<-cbind(wsj_szvector,"Syuzhet")
new_sz<-cbind.data.frame(new_sz,wsj_data$date)
colnames(new_sz)<-c("Score","Method","Date")
new<-rbind(new_b,new_afv,new_nrc,new_sz)
new<-as.data.frame(new)
new$Score<-as.integer(new$Score)

h_line <- mean(new$Score)
```

Visualization

```
library(ggplot2)

ggplot(new, aes(Date,Score)) + geom_line(aes(color=Method)) +
facet_wrap(~Method, ncol = 2) + theme(legend.position="none",axis.text.x =
element_text(angle = 90)) + xlab("Date")+ylab("Score") +ggtitle("WSJ
Sentiment Scores by Different Sentiment Methods") + scale_color_manual(values
= c("Affin" =
"goldenrod","Bing"="forestgreen","NRC"="steelblue","Syuzhet"="maroon")) +
geom_hline(aes(yintercept=0))
```

```
Warning: Removed 12 row(s) containing missing values (geom_path).
```

In a sentiment analysis for the Wall Street Journal for the 2018 to early 2020 time period, we try four different methods for processing sentiment as positive, neutral, or negative over time as based on scores returned from these methods.

We created the visual to aid in the following objectives:

1) show baseline of 0 for neutral sentiment to show how methods differ from each other in regards to how distant they are from 0

2) assist user in identifying which methods show more positive results, which ones show more negative results, and which ones have greater variance vs being packed more tightly together toward 0 (neutral)

We hope this visual will assist sentiment analysis researchers in choosing the right method for their project while understanding how the results can be affected by the method.

Details on sentiment analysis methods in R found here: https://arxiv.org/pdf/1901.08319

Note: mean of all Scores for all time periods was slightly positive: 2.25545.