Data Management

Contents

Raw Data

```r
# 1. Political

#truthsocial
raw_truths <- read.csv(here("data/political_data", "truths_new.csv"))

#twitter
raw_tweets <- read.csv(here("data/political_data", "tweets.csv"))


# 2. Financial

#S&P500
data_loader(symbol="SPY")

#STOXX50
data_loader(symbol="VGK")

#CSI 300 (China)
data_loader(symbol="ASHR")
```

Cleanup

```r
#MAKE FUNCTIONS TO CLEANUP EASY

# 1. Tweets
tweets = raw_tweets

#only keep original Tweets
tweets <- tweets %>% filter(isRetweet != "t")
tokens <- tokens(tweets$text)
dfm <- dfm(tokens)

#cleanup
tweets = data.frame(tweets$date,tweets$text)
colnames(tweets) = c("timestamp","tweet_text")
tweets$timestamp = as.POSIXct(tweets$timestamp,format = "%Y-%m-%d %H:%M:%S")
second(tweets$timestamp) = 0

#nrc_scores <- get_nrc_sentiment(complete_data$posts)


# 2. Truths
truthsbackup <- truths_processer(raw_truths)
truths = truthsbackup

#cleanup
truths <- truths %>% filter(media != 1)
truths = data.frame(truths$date_time_parsed,truths$post)
```

```r
colnames(truths) = c("timestamp","truths_text")
truths$timestamp = as.POSIXct(truths$timestamp,format = "%Y-%m-%d %H:%M:%S")
second(truths$timestamp) = 0

# Merging social media data since it is not overlapping
names(truths)[names(truths) == 'truths_text'] <- 'tweet_text'
social = rbind(tweets,truths)
social <- social[order(social$timestamp, decreasing=F), ]


# 3. Financial

#remove index
SPY = raw_SPY[-1]
VGK = raw_VGK[-1]
ASHR = raw_ASHR[-1]

#rename financial columns to add symbol
colnames(SPY)[-1] <- paste0("SPY", colnames(SPY)[-1])
colnames(VGK)[-1] <- paste0("VGK", colnames(VGK)[-1])
colnames(ASHR)[-1] <- paste0("ASHR", colnames(ASHR)[-1])
```

Building Additional Variables

Volatility By Hour

```r
SPY = r.vol_hourly(SPY,merge=T)
```

Adding Dummies

```r
# 1. Add count for tweets

# 2. Sentiments

# 3. Dummy Tweet

# 4. Dummy Important Word

# 5. Dummy Emotional Word
```

Data Save

```r
#financial
write.csv(SPY, here("data/mothership/SPY.csv"), row.names=F)
write.csv(VGK, here("data/mothership/VGK.csv"), row.names=F)
write.csv(ASHR, here("data/mothership/ASHR.csv"), row.names=F)

#social media
write.csv(social, here("data/mothership/social.csv"), row.names=F)
```

Merging All Data

First Merge

```r
#run script to load and merge all data
rm(list=ls())
source(here("helperfunctions/fulldata_loader.R"))
```

Using Alpha Vantage API

```r
library(alphavantager)

av_api_key(Sys.getenv("ALPHAVANTAGE_API_KEY"))

#for past month
data=av_get("ASHR", av_fun = "TIME_SERIES_INTRADAY", interval = "1min",
              adjusted="false", extended_hours="false", outputsize = "full")

#for a particular month
data2=av_get("SPY", av_fun = "TIME_SERIES_INTRADAY", interval = "1min",
              adjusted="false", extended_hours="false",
              month="2025-04", outputsize = "full") #create loop for more



write.csv(data,"~/ASHR.csv", row.names = T) #saves to documents
write.csv(data2,"~/SPY-2025-04.csv", row.names = T)
```

```r
library(alphavantager)

av_api_key(Sys.getenv("ALPHAVANTAGE_API_KEY"))

year = "2022"
months = c("01","02","03","04","05","06","07","08","09","10","11","12")
market = "SPY"

for (t in 1:length(months)) {
    date = paste(year, months[t], sep="-")
dataloop = av_get(market, av_fun="TIME_SERIES_INTRADAY",interval="1min",
                  adjusted="false", extended_hours="false",
                  month=date, outputsize="full")
filename = paste(market,date,sep="-")
filename = paste("~/", filename, sep="")
filename = paste(filename,".csv",sep="")
write.csv(dataloop,filename)
}
```

Tutorials

Manual smp500: https://cafim.sssup.it/~giulio/other/alpha_vantage/index.html#orgaaf54ef

AlphaVantageR Tutorial: https://github.com/business-science/alphavantager/blob/master/man/av_get.Rd

Intra-Day Analysis: https://arxiv.org/html/2406.17198v1

Symbols Explanation

- ONEQ = NASDAQ Composite
- SPY = S&P500
- SMI = Swiss Market Index
- VTHR = Russell 3000 (US)
- VTI = CRSP US Total Market Index
- VGK = Euro Stoxx 50
- ASHR = basically china