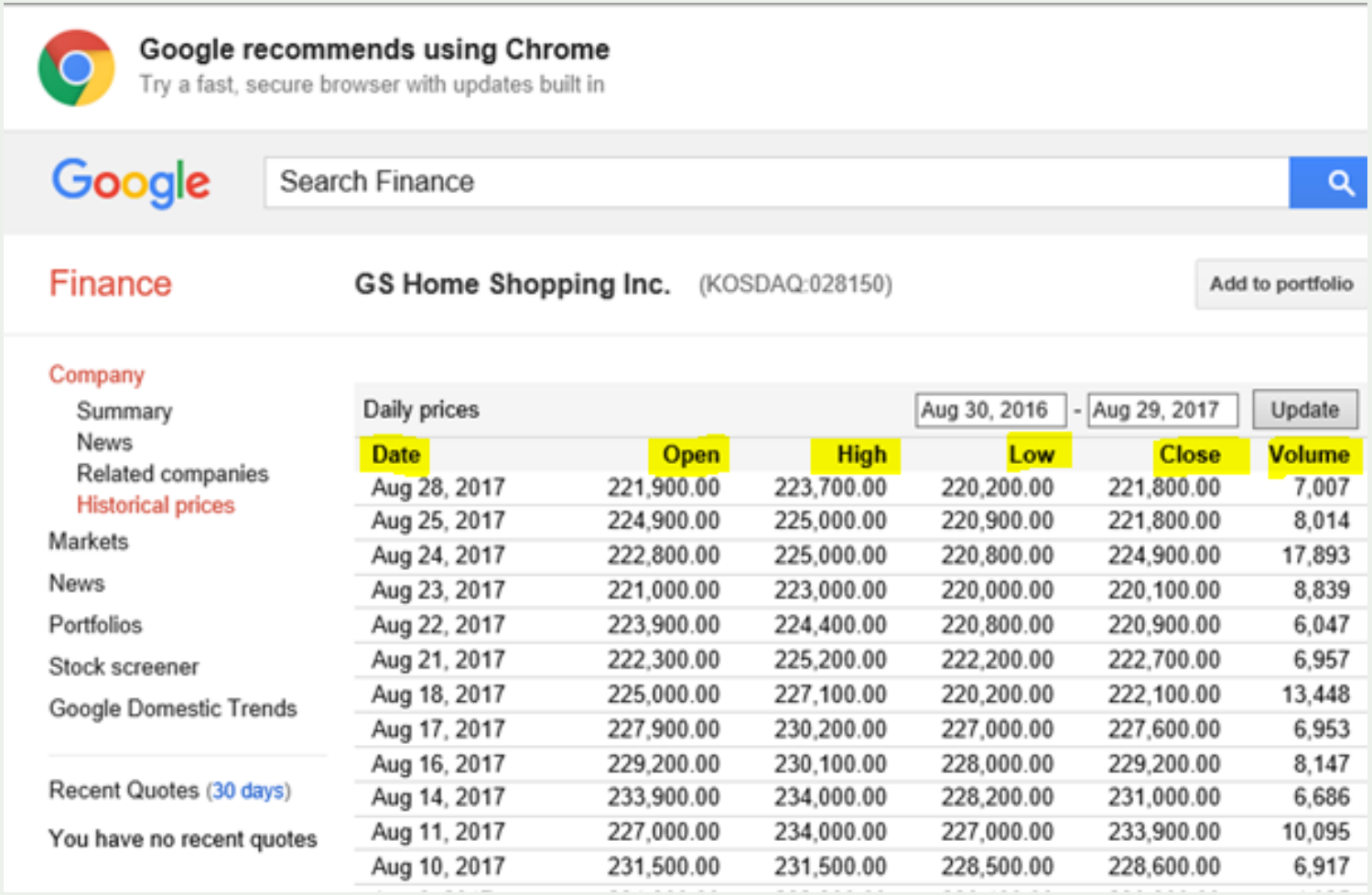
뉴스 텍스트 마이닝

Summary

1. Google finance gs 홈쇼핑 주식데이터 크롤링
2. Date별 주식이 오르면 up 내리면 down표시
3. Daum뉴스에서 gs 홈쇼핑 뉴스데이터 크롤링
4. 주식, 뉴스데이터 table merge
5. 테이블 결측치 처리
6. 웹에서 긍/부정 사전 다운로드
7. 크롤링한 뉴스데이터와 비교하여 긍정 사전에 해당하는 단어가 많으면 긍정기사 부정단어에 해당하는 단어가 많으면 부정적 기사로 분류
8. 주식 up down과 뉴스의 긍/부정 비교

구글 finance 자료 크롤링

-2015/08 ~ 20017/08 3년치 주식데이터 크롤링



Google recommends using Chrome
Try a fast, secure browser with updates built in

Google Search Finance

Finance GS Home Shopping Inc. (KOSDAQ:028150) Add to portfolio

Company
- Summary
- News
- Related companies
- Historical prices

Markets
News
Portfolios
Stock screener
Google Domestic Trends

Recent Quotes (30 days)
You have no recent quotes

Daily prices Aug 30, 2016 - Aug 29, 2017 Update

Date	Open	High	Low	Close	Volume
Aug 28, 2017	221,900.00	223,700.00	220,200.00	221,800.00	7,007
Aug 25, 2017	224,900.00	225,000.00	220,900.00	221,800.00	8,014
Aug 24, 2017	222,800.00	225,000.00	220,800.00	224,900.00	17,893
Aug 23, 2017	221,000.00	223,000.00	220,000.00	220,100.00	8,839
Aug 22, 2017	223,900.00	224,400.00	220,800.00	220,900.00	6,047
Aug 21, 2017	222,300.00	225,200.00	222,200.00	222,700.00	6,957
Aug 18, 2017	225,000.00	227,100.00	220,200.00	222,100.00	13,448
Aug 17, 2017	227,900.00	230,200.00	227,000.00	227,600.00	6,953
Aug 16, 2017	229,200.00	230,100.00	228,000.00	229,200.00	8,147
Aug 14, 2017	233,900.00	234,000.00	228,200.00	231,000.00	6,686
Aug 11, 2017	227,000.00	234,000.00	227,000.00	233,900.00	10,095
Aug 10, 2017	231,500.00	231,500.00	228,500.00	228,600.00	6,917

R코드- 구글finance data 크롤링

```
#google finance data 크롤링
#https://www.google.com/finance/historical?cid=8212221665534 43&startdate
= Apr+7 %2C+ 2015 &enddate= May+3 %2C+ 2017 &num= 200
#https://www.google.com/finance/historical?cid=8212221665534 43&startdate
= Jan+1 %2C+ 2017 &enddate= May+17 %2C+ 2017 &num= 200
startdate<-"2015/08/10"
enddate<-"2017/08/12"
month<-
c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec")
startyear<-as.numeric(str_sub(startdate,1,4))
startmonth<-month[as.numeric(str_sub(startdate,6,7))]
startday<-as.numeric(str_sub(startdate,9,10))
endyear<-as.numeric(str_sub(enddate,1,4))
endmonth<-month[as.numeric(str_sub(enddate,6,7))]
endday<-as.numeric(str_sub(enddate,9,10))
num<-200
start<-0
```

```
url<-"https://www.google.com/finance/historical?cid=8212221665534 43&startdate="
continue <- TRUE
table<-data.frame()

while(continue)
{
  url1<-
url%>%paste(.,startmonth,"+",startday,"%2C+",startyear,"&enddate=",endmonth,"+",e
ndday,"%2C+",endyear,"&num=",num,"&start=",start,sep="")
  table_url<-read_html(url1)%>%html_nodes("div#prices")%>%html_nodes(".gf-
table.historical_price")
  plus<-html_table(table_url)[[1]]
  for(i in 2:6) plus[,i]<-as.numeric(gsub(",","",plus[,i]))
  table<-rbind(table,plus)
  if(nrow(plus)<200) continue <- FALSE
  start<-start+num
}

table_month<-factor(str_sub(table$Date,1,3),levels = month,labels=1:12)
table_day<-str_sub(table$Date,5,6)
table_day<-ifelse(str_sub(table_day,2)==",",str_sub(table_day,1,1),table_day)
table_year<-str_trim(str_sub(table$Date,8,12))
```

```
date<-as.Date(paste(table_year,table_month,table_day),format="%Y%m%d")
table_result<-cbind(date,table[,2:6])
date20170811<-c(table_result$Close,0)
date20170810<-c(0,table_result$Close)


#updown 컬럼 만들기 -
updown<-ifelse(date20170810-date20170811>0,"up","down")
updown[1]<-NA
updown<-updown[1:496]
updown<-factor(updown,levels=c("down","up"))
table_result<-cbind(table_result,updown)
table_result
```

구글 finance 주식 자료 크롤링

```
> table_result
          date   Open   High    Low  Close Volume updown
1   2017-08-11 227000 234000 227000 233900  10095   <NA>
2   2017-08-10 231500 231500 228500 228600   6917     up
3   2017-08-09 231800 232900 230100 230800   4025   down
4   2017-08-08 228200 234500 228200 233900   6076   down
5   2017-08-07 232000 232000 228600 230500   6723     up
6   2017-08-04 229400 237000 228100 231600  15969   down
7   2017-08-03 229600 233900 227100 228000  10975     up
8   2017-08-02 233800 233800 228700 228800  10308   down
9   2017-08-01 236100 236100 231500 232700   8863   down
10  2017-07-31 232000 236600 225500 235500  13809   down
11  2017-07-28 237800 240000 234800 234800  12691     up
12  2017-07-27 238300 240000 235600 239500  11392   down
13  2017-07-26 234900 237800 231300 237700  12830     up
14  2017-07-25 234400 234500 229500 233900   8007     up
15  2017-07-24 229000 232600 227300 232100  12644     up
16  2017-07-21 228200 231900 226900 227300   9691     up
17  2017-07-20 229900 230100 226800 226800  10929     up
18  2017-07-19 225000 229800 223800 229800   9571   down
19  2017-07-18 225200 225400 223200 225200   9135     up
20  2017-07-17 228500 228500 223900 225100  12605     up
21  2017-07-14 223800 230000 223700 228500  13138   down
22  2017-07-13 220000 227000 220000 225900  14706     up
23  2017-07-12 222700 222700 218100 220000  14797     up
24  2017-07-11 223500 223900 222500 222700  12494   down
25  2017-07-10 225000 226000 222200 223000   7209   down
```

뉴스 크롤링

-2015/08 ~ 2017/08 3년간 GS홈쇼핑의 뉴스자료 크롤링

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
f1	url	date	title	article													
0	http://new	#######	GS홈쇼핑	GS홈쇼핑은 코렐 비전 파이렉스 등의 식기 브랜드를 갖고 있는 기업 월드키친의 지분 9.1%를 인수했다고 12일 발표했다. GS홈쇼핑과 미국 사모펀													
1	http://new	#######	물걸레 로	물걸레 로봇청소기 에브리봇에는 바퀴가 없다. 대신 바닥에 달린 동그란 물걸레 두 개가 어떻게 회전하느냐에 따라 에브리봇이 종횡무진 움직이며													
2	http://new	#######	청와대 참	청와대 참모진이 보유 주식을 매각한 것으로 나타났다. 1급 이상 고위공직자는 업무 관련 기업의 주식을 보유할 수 없도록 공직자윤리법이 규정하고													
3	http://new	#######	탈모닷컴	탈모방지에 도움을 주는 스테디셀러 TS샴푸를 선보이는 탈모닷컴(대표 장기영)은 TS샴푸 1000억원 판매와 100만 고객을 돌파했다고 7일 밝혔다.개													
4	http://stoc	#######	나노스 어	가 5거래일째 급등하며 시가총액 8위로 올라섰다.19일 오전 9시7분 현재 는 전날보다 3650원(25.26%) 오른 1만8100원에 거래중이다. 지난 13일 거													
5	http://new	#######	허창수 "불	"지난 일을 잊지 않고 잘 살펴서 일의 지침으로 삼는다."허창수 GS그룹 회장(사진)이 19일 서울 역삼동 GS타워에서 열린 3분기 GS 임원 모임에서 단													
6	http://new	#######	GS '열린	GS그룹은 구성원들이 원활하게 소통할 수 있는 '열린 조직문화' 정착에 힘쓰고 있다. 일과 삶의 조화를 통해 조직의 활력과 생산성은 물론 개인의 삶													
7	http://new	#######	집밥같은	40~50대 주부들의 가정간편식(HMR) 구매가 빠르게 늘고 있다. 제품 선택의 폭이 넓어진 데다 맛도 좋아지면서 1인가구나 맞벌이 부부뿐 아니라 직													
8	http://stoc	#######	코스피 外	코스피지수가 사상 처음으로 7개월 연속 상승했다.30일 코스피지수는 전날 대비 0.16% 하락한 2391.79를 기록했다. 이날 하락했지만 이달 코스피지													
9	http://stoc	#######	코스피 23	코스피지수가 2380선으로 밀려나 하락 흐름을 이어가고 있다.3일 오후 1시22분 현재 코스피지수는 전날보다 4.36포인트(0.18%) 내린 2387.43에 거													
10	http://stoc	#######	코스피 24	코스피가 개인의 매수세에 힘입어 이틀째 2410선에서 상승세를 지속하고 있다. 와 하이닉스는 나란히 오르고 있다.14일 오전 9시13분 현재 코스피지													
11	http://new	#######	GS 허창수	허창수 GS 회장은 "고객 니즈의 변화를 빠르게 파악하고 유연하게 대응해 지속적으로 고객에게 새로운 가치를 제공해야 하며 내부적으로는 원가 경													
12	http://stoc	#######	브이티 코	지엠피의 자회사 브이티 코스메틱은 지난 1일 GS홈쇼핑에서 방송한 'VT 블루 콜라겐 팩트'가 연속 완판 행진을 이어갔다고 3일 밝혔다VT 블루 콜라													
13	http://stoc	#######	코스피 9일	코스피지수가 9거래일 만에 하락했다.이로써 코스피의 '최장 연속' 사상 최고치 기록(8거래일) 경신도 다음으로 미뤄졌다. 코스피의 기존 연속 최고치													
14	http://stoc	#######	코스피 장	코스피지수가 개장 직후 사상 최고치 기록을 갈아치운 뒤 숨을 고르고 있다. 오후 들어서 지수는 외국인과 기관의 공방 탓에 보합권을 맴돌고 있는													
15	http://new	#######	화제의 '자	요즘처럼 더운 여름 발뒤꿈치는 여름 샌들을 신기 전 필수적으로 챙겨야 할 부분이다. 허영게 일어난 각질은 샌들이나 슬리퍼를 신으면 더욱 도드라													
16	http://new	#######	GS홈쇼핑	GS홈쇼핑이 코렐·비전·파이렉스 등으로 유명한 키친웨어 제조사 월드키친에 투자를 진행했다.GS홈쇼핑은 12일 미국계 사모펀드인 코넬캐피털과 월													

뉴스 크롤링

-2015/08 ~ 2017/08 3년간 주식데이터와 뉴스데이터를 date 기준으로 table merge

	date	Open	High	Low	Close	Volume	tom	f1	url	title
306	2016-05-10	190800	198900	190500	194000	23175	down	326	http://news.hankyung.com/article/2016051096251	70번의 실패서 배웠다
307	2016-05-11	194100	196900	190300	192900	15056	up	318	http://news.hankyung.com/article/201605112665i	편의점 실적 악화에도
310	2016-05-12	192000	197000	192000	197000	17336	down	320	http://news.hankyung.com/article/2016051263601	해외서 파는 상품에도
309	2016-05-12	192000	197000	192000	197000	17336	down	317	http://news.hankyung.com/article/2016051263681	"중기 사장이 비행기
317	2016-05-18	189800	191900	188300	191000	7443	down	344	http://news.hankyung.com/article/201605181284i	웨매지 않은 옷 인기..
320	2016-05-19	190500	190900	187100	190000	7625	up	346	http://news.hankyung.com/article/201605194619g	세원ITC "사이먼스캇
324	2016-05-25	189100	191000	187000	187100	10538	down	343	http://news.hankyung.com/article/201605250857a	무더위/미세먼지로 야
327	2016-05-30	NA	NA	NA	NA	NA	NA	345	http://news.hankyung.com/article/201605290285i	'저가' 이미지 벗은 홈
331	2016-06-01	185300	186500	181400	185500	6728	up	341	http://news.hankyung.com/article/201606018268g	'정용진표 집밥' PB 피
330	2016-06-01	185300	186500	181400	185500	6728	up	340	http://news.hankyung.com/article/201606019567i	이마트 롯데에서 PB
334	2016-06-07	NA	NA	NA	NA	NA	NA	328	http://news.hankyung.com/article/2016060603191	편의점에 IoT 기술 접
336	2016-06-07	NA	NA	NA	NA	NA	NA	357	http://news.hankyung.com/article/2016060608851	일본 수출길 막힌 중기

테이블 결측 처리

-주가데이터 없는 경우 전날 주가로 날짜를 바꿔 해당날짜 이후로 주가를 예측할 수 있도록 함.

	date	Open	High	Low	Close	Volume	tom	f1	url	title
1	2015-08-10	199100	199100	193100	194500	8579	down	NA	NA	NA
2	2015-08-11	194000	197000	191000	191000	9226	down	NA	NA	NA
3	2015-08-12	191100	194000	188600	189800	18388	down	NA	NA	NA
4	2015-08-13	191000	191000	187200	188000	17768	up	NA	NA	NA
5	2015-08-14	NA	NA	NA	NA	NA	NA	555	http://news.hankyung.com/article/201508122585g	홈쇼핑 이젠 홈쇼핑 아닌 '모바
6	2015-08-17	186200	190800	186000	190000	14667	down	554	http://news.hankyung.com/article/2015081609061	사이판 크루즈 탈까...'아시아의

	date	Open	High	Low	Close	Volume	tom	f1	url	title
1	2015-08-10	199100	199100	193100	194500	8579	down	NA	NA	NA
2	2015-08-11	194000	197000	191000	191000	9226	down	NA	NA	NA
3	2015-08-12	191100	194000	188600	189800	18388	down	NA	NA	NA
4	2015-08-13	191000	191000	187200	188000	17768	up	NA	NA	NA
5	2015-08-13	NA	NA	NA	NA	NA	NA	555	http://news.hankyung.com/article/201508122585g	홈쇼핑 이젠 홈쇼핑 C
6	2015-08-17	186200	190800	186000	190000	14667	down	554	http://news.hankyung.com/article/2015081609061	사이판 크루즈 탈까...

뉴스 크롤링

-isnews변수생성 : 해당날짜에 뉴스가 있으면 1 없으면 0 으로 표시

	date	Open	High	Low	Close	Volume	tom	isnews
1	2015-08-10	199100	199100	193100	194500	8579	down	0
2	2015-08-11	194000	197000	191000	191000	9226	down	0
3	2015-08-12	191100	194000	188600	189800	18388	down	0
4	2015-08-13	191000	191000	187200	188000	17768	up	1
5	2015-08-13	191000	191000	187200	188000	17768	up	0
6	2015-08-17	186200	190800	186000	190000	14667	down	1
7	2015-08-18	190100	190700	187600	187600	7938	up	1
8	2015-08-19	187100	206300	186000	206200	36964	down	1
9	2015-08-19	187100	206300	186000	206200	36964	down	1
10	2015-08-19	187100	206300	186000	206200	36964	down	1
11	2015-08-19	187100	206300	186000	206200	36964	down	1
12	2015-08-19	187100	206300	186000	206200	36964	down	1

뉴스기사 긍부정 판단

<한글 긍부정 사전 출처>

-https://github.com/The-ECG/BigData1_1.3.3_Text-Mining/blob/master/dictionary.zip

-Github에서 만들어진 Text-Mining에 필요한 긍 부정사전을 google translation한 자료

뉴스기사 긍부정 판단

-isnew==1 인것만 가져와서 긍부정 사전과 비교
-Github에서 가져온 긍부정 사전을 이용하여 뉴스 article에 나타난 긍정단어와 부정단어의 개수 sum
-긍정단어가 많으면 positive article, 부정단어가 많으면 negative article 로 판단
-주식데이터의 up, down 과 뉴스데이터의 positive, negative 비교

```
> pos_neg_news_num
          date   Open   High    Low  Close Volume updown isnews pos_num neg_num pos_ar
1   2015-08-13 191000 191000 187200 188000  17768     up       1       14        8
2   2015-08-17 186200 190800 186000 190000  14667   down       1       22       14
3   2015-08-18 190100 190700 187600 187600   7938     up       1       17       14
4   2015-08-19 187100 206300 186000 206200  36964   down       1        9       11
5   2015-08-19 187100 206300 186000 206200  36964   down       1       13       14
6   2015-08-19 187100 206300 186000 206200  36964   down       1        5        4
7   2015-08-19 187100 206300 186000 206200  36964   down       1        9       16
8   2015-08-19 187100 206300 186000 206200  36964   down       1       12       14
9   2015-08-20 206100 206100 197000 197300  13321   down       1        7        8
10  2015-08-21 188000 198200 187700 197200  12127   down       1       12       20
11  2015-08-21 188000 198200 187700 197200  12127   down       1       10       14
12  2015-08-21 188000 198200 187700 197200  12127   down       1        8       11
13  2015-08-24 192000 195800 189200 191400  14519   down       1        8       16
14  2015-08-24 192000 195800 189200 191400  14519   down       1       16        7
15  2015-08-24 192000 195800 189200 191400  14519   down       1       14       11
```