HW2\_201821479\_황혜린

201821479\_황혜린

2018년 11월 21일

1. 텍스트 마이닝을 통해 해결할 수 있는 문제를 스스로 선택하시오. 그리고 그런 문제를 선택한 이유를 상세히 서술하시오.

비트코인을 개발한 나카모토 사토시는 ‘비트코인의 핵심은 정부나 중앙은행 등 어떤 중앙집중적 권력의 개입 없이 작동하는 새로운 화폐를 창출하는 데 있다’ 고 말했다. 출처 :[네이버 지식백과] 비트코인 [bitcoin] (한경 경제용어사전, 한국경제신문/한경닷컴 )

급격하게 환율이 변동되는 등 세계경제가 불안하고 시장이 불확실한 상황에서는 상대적으로 안전한 자산에 투자가 몰린다고 한다.

2009년은 금융위기가 한창이던 시기에 달러화 가치 하락 우려가 겹치면서 비트코인이 대안 화폐로 주목받기 시작했다. 비트코인의 기반인 블록체인 기술은 분산형 장부 기록 데이터베이스 기술로서 금융거래에서 장부 책임자가 없는 거래 시스템이기 때문이다.

위 사실에 따라, 세계경제가 불안하고 국제 금융시장이 흔들린다면 비트코인의 가치는 상승하지 않을까? R을 통해 분석하여 궁금증을 해결해보고자 한다.

1. 분석에 필요한 기본 지식을 서술하시오. 선행적으로 분석할 작업이 있다면 수행하시오.

‘브렉시트(Brexit)’란 영국(Britain)과 탈퇴(Exit)의 합성어이다. 2016년 3월29일부터 시작되었으며 2016년 6월 23일 국민투표(탈퇴 51.9%, 잔류 48.1%)를 통하여 영국은 EU 탈퇴를 결정하였다.

아래 내용은 한국은행이 발행한 글 일부를 발췌한 것이다.

“Brexit가 실현되면 영국-EU간 무역관계 변화뿐 아니라 전세계 많은 국가들에게 다방면의 영향을 미칠 것으로 보입니다. 중략 Brexit 결정 이후 영국의 통화인 파운드화 가치가 크게 떨어지고, 세계 각국 금융시장이 한동안 출렁인 것도 이러한 우려가 반영된 것입니다. 앞으로 영국과 EU간 Brexit 협상의 전개 양상에 따라 국제금융시장에서 환율, 주가 등 가격변수와 자본 유출입의 변동성이 확대될 수 있는 만큼 이에 유의할 필요가 있습니다.”

빅데이터 분석도구 R을 이용하여 브렉시트가 일어났던 시기에 비트코인의 가치가 실제로 상승하였는지 분석해보고자 한다.

‘블록체인 시세’에 관한 기사들을 분석하여 워드클라우드로 나타내 보았다.

검색 시작: 2017.03.29 (브렉시트 시작일)

검색 완료: 2017.12.31

#install.packages("rvest")  
#install.packages("httr")  
#install.packages("KoNLP")  
#install.packages("wordcloud")  
#install.packages("tm")  
#install.packages("slam")  
  
headline <- NULL  
content <- NULL #변수들에 값 할당하기 전에 NULL지정   
  
library(rvest) #get 방식의http 주소를 사용하는경우

## Loading required package: xml2

library(httr) #post방식으로 데이터를 요청해서 가져오는 경우에 추가로 사용  
  
for (i in 1:250) {  
 url <- ("https://search.daum.net/search?nil\_suggest=btn&w=news&DA=PGD&cluster=y&q=%EB%B9%84%ED%8A%B8%EC%BD%94%EC%9D%B8+%EC%8B%9C%EC%84%B8&spacing=2&orgq=%EB%B9%84%ED%8A%B8%EC%BD%94%EC%9D%B8+tltp&sd=20170329000000&ed=20171231235959&period=u&p=")  
 urls <- paste(url, i, sep = "") #for문, paste함수를 사용하여 url조합  
 html\_source <- read\_html(urls) #read\_html함수를 사용하여 html페이지를 html\_sourse라는 변수에 저장  
   
 h <- html\_nodes(html\_source, 'a.f\_link\_b')%>%html\_text() #페이지에서 헤드라인부분 셀렉터의 텍스트부분을 h에 저장   
 headline <- append(headline, h) #append함수를 이용하여 headline에 h추가   
   
 c <- html\_nodes(html\_source, 'p.f\_eb.desc')%>%html\_text() #페이지에서 기사부분 셀렉터의 텍스트부분을 c에 저장  
 content <- append(content, c) #append이용하여 content에 c추가   
   
 bit <- data.frame(headline, content) #headline과 content를 컬럼으로 가지는 데이터프레임 bit 생성   
   
 bit[,1] <- gsub("[\n\r\t]", "", bit[,1])  
 bit[,2] <- gsub("[\n\r\t]", "", bit[,2]) #데이터 전처리   
   
   
}  
  
  
  
#headline 컬럼을 bit\_에 할당   
bit\_ <- bit$headline   
library(KoNLP)

## Checking user defined dictionary!

data <- sapply(bit\_, extractNoun, USE.NAMES = F) #KoNLP패키지로 명사 추출   
  
bit\_1 <- unlist(data) #데이터를 필터링하기 쉽게 리스트형태를 벡터로 반환   
bit\_chr <- as.character(bit\_1) #데이터를 필터링하기 위해 character형태로 만든다   
bit\_chr <- Filter(function(x){nchar(x)>1}, bit\_chr) #글자수가 1보다 큰 글자만 추출   
  
  
#데이터 전처리   
bit\_chr <- gsub("경찰", "", bit\_chr)  
bit\_chr <- gsub("[0-9]", "", bit\_chr)  
bit\_chr <- gsub("고교생", "", bit\_chr)  
bit\_chr <- gsub("사기극", "", bit\_chr)  
bit\_chr <- gsub("시황", "", bit\_chr)  
bit\_chr <- gsub("신고", "", bit\_chr)  
bit\_chr <- gsub("누구", "", bit\_chr)  
bit\_chr <- gsub("넌1]", "", bit\_chr)  
bit\_chr <- gsub("숨고르기", "", bit\_chr)  
bit\_chr <- gsub("시작됐", "", bit\_chr)  
bit\_chr <- gsub("김치", "", bit\_chr)  
  
#워드클라우드로 표현해주기 위해 데이터의 형태를 변환한다   
bit\_df <- data.frame(bit\_chr)  
bit\_tb <- table(bit\_df)  
  
#워드클라우드를 만들기 위해 사용할 패키지들   
library(wordcloud)

## Loading required package: RColorBrewer

library(tm)

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:httr':  
##   
## content

library(slam)  
  
#워드클라우드에 사용할 색 개수와 색깔   
pal <- brewer.pal(11, "Paired")  
  
  
wordcloud(names(bit\_tb), bit\_tb, min.freq = 50, max.words = 30, random.order = F, scale = c(10, 0.8), colors = pal)

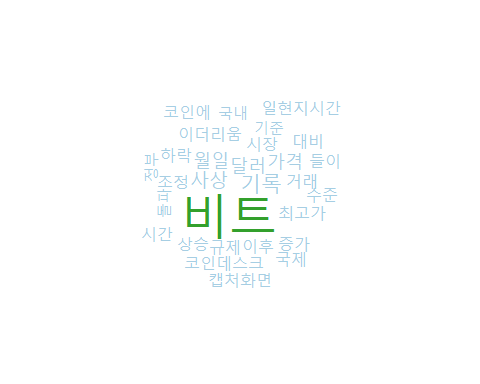


#bit\_tb라는 데이터 프레임을 활용해 50빈도수 이상, 최대 단어 30개를 추출. 높은 빈도수를 가진 단어는 가운데에 두고 스케일을 크기에 맞게 조절한 후 위에서 고른 색을 대입.  
  
  
head(sort(bit\_tb, decreasing = T), 15) #높은 빈도수부터 15개의 데이터 출력

## bit\_df  
## 비트 최고 화폐 가상 가격 달러 논란   
## 2504 2124 360 306 302 224 221 195   
## 분석 이더리움 암호화 상승세 신변 시간 사상   
## 191 188 186 185 184 183 182

위는 기사 제목을 분석한것이다. 워드클라우드와 head,sort함수를 통해 최고, 상승세와 같은 키워드가 많이 등장하는 것을 확인할 수 있다.

#content 컬럼을 cont\_에 할당   
cont\_ <- bit$content  
library(KoNLP)  
cont <- sapply(cont\_, extractNoun, USE.NAMES = F) #KoNLP패키지로 명사 추출   
  
cont\_1 <- unlist(cont) #데이터를 필터링하기 쉽게 리스트형태를 벡터로 반환  
cont\_chr <- as.character(cont\_1) #데이터를 필터링하기 위해 character형태로 만든다  
cont\_chr <- Filter(function(x){nchar(x)>1}, cont\_chr) #글자수가 1보다 큰 글자만 추출   
  
  
#데이터 전처리   
cont\_chr <- gsub("캡쳐화면", "", cont\_chr)  
cont\_chr <- gsub("[~!@#$%^&\*()-\_=+?:]", "", cont\_chr)  
cont\_chr <- gsub("[0-9]", "", cont\_chr)  
cont\_chr <- gsub("가상", "", cont\_chr)  
cont\_chr <- gsub("화폐", "", cont\_chr)  
cont\_chr <- gsub("오후", "", cont\_chr)  
cont\_chr <- gsub("오전", "", cont\_chr)  
cont\_chr <- gsub("우려", "", cont\_chr)  
cont\_chr <- gsub("지난달", "", cont\_chr)  
cont\_chr <- gsub("코인은", "", cont\_chr)  
cont\_chr <- gsub("코인데스크에", "", cont\_chr)  
cont\_chr <- gsub("신문", "", cont\_chr)  
cont\_chr <- gsub("CNBC", "", cont\_chr)  
cont\_chr <- gsub("보도", "", cont\_chr)  
cont\_chr <- gsub("코인과", "", cont\_chr)  
cont\_chr <- gsub("통화", "", cont\_chr)  
cont\_chr <- gsub("시세", "", cont\_chr)  
cont\_chr <- gsub("은", "", cont\_chr)  
  
  
#워드클라우드로 표현해주기 위해 데이터의 형태를 변환한다  
cont\_df <- data.frame(cont\_chr)  
cont\_tb <- table(cont\_df)  
  
  
#워드클라우드를 만들기 위해 사용할 패키지들  
library(wordcloud)  
library(tm)  
library(slam)  
  
  
#워드클라우드에 사용할 색 개수와 색깔  
pal <- brewer.pal(11, "Paired")  
  
  
wordcloud(names(cont\_tb), cont\_tb, min.freq = 50, max.words = 30, random.order = F, scale = c(10, 1), colors = pal)



#bit\_tb라는 데이터 프레임을 활용해 50빈도수 이상, 최대 단어 30개를 추출. 높은 빈도수를 가진 단어는 가운데에 두고 스케일을 크기에 맞게 조절한 후 위에서 고른 색을 대입.  
  
head(sort(cont\_tb, decreasing = T), 20) #높은 빈도수부터 20개의 데이터 출력

## cont\_df  
## 비트 기록 달러 사상 월일   
## 10257 2938 726 646 543 527   
## 가격 시장 규제 이후 이더리움 코인데스크   
## 510 386 384 381 377 364   
## 캡처화면 최고가 거래 조정 증가 기준   
## 361 359 355 347 343 267   
## 국내 돌파   
## 257 248

위는 기사 내용을 분석한것이다. 워드클라우드와 head,sort함수를 통해 최고가, 증가란 키워드가 자주 등장하는 것을 확인할 수 있었다.

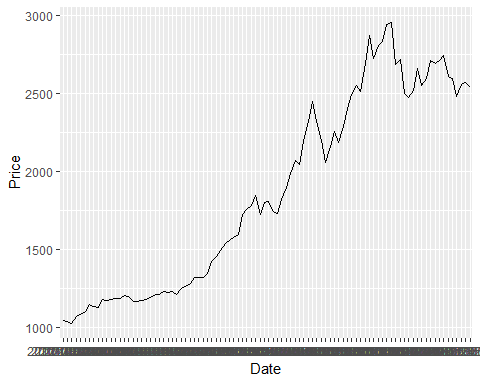
1. 분석을 수행하고 자신의 의견을 서술하시오. 다음은 2017년 3월 29일부터 12월까지 비트코인의 시가 변화이다. 실제로 브렉시트 이후 3개월동안 2배 이상 상승했다.

inBrexit <- read.csv("bit\_price.csv", header = T, stringsAsFactors = F) #csv파일을 읽어온다   
Price <- inBrexit$price #Price컬럼   
Date <- inBrexit$Date #Date컬럼   
  
  
library(ggplot2) #ggplot으로 그래프를 그린다

##   
## Attaching package: 'ggplot2'

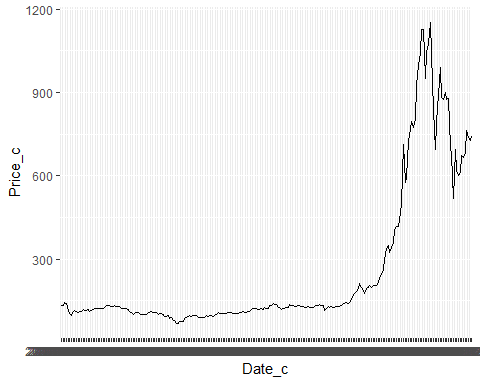
## The following object is masked from 'package:NLP':  
##   
## annotate

ggplot(inBrexit, aes(Date, Price, group = 1)) + geom\_line() #x축은 Date, y축은 Price, 이산으로 흩어진 숫자를 연속적으로 표현하기 위해 group=1을 사용한다. 선 그래프로 표현하였다.



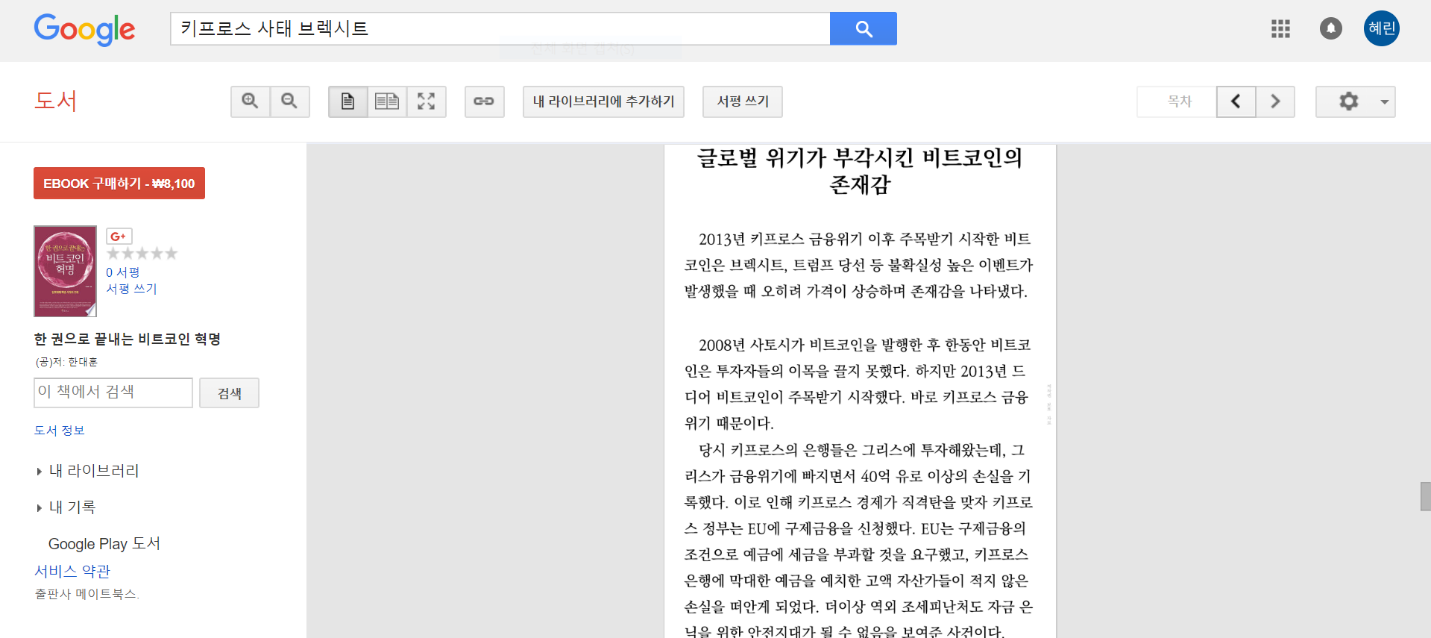
더불어 비슷한 사례인 2013년 3월 키프로스 사태 이후 2013년 비트코인의 시가 변화를 나타내보았다. 마찬가지로 사건이 발생한 후 연말, 시가가 6배에서 최대9배 상승했음을 알 수 있다. 점진적으로 증가한 이유는 초반에 비트코인 인지도가 낮아 그러한 것으로 예측된다.

cyprus <- read.csv("Cyprus.csv", header = T, stringsAsFactors = F) #csv파일을 읽어온다  
Date\_c <- cyprus$Date #Date컬럼   
Price\_c <- cyprus$Price #Price컬럼  
  
  
library(ggplot2) #ggplot으로 그래프를 그린다   
ggplot(cyprus, aes(Date\_c, Price\_c, group = 1)) + geom\_line() #x축은 Date, y축은 Price, 이산으로 흩어진 숫자를 연속적으로 표현하기 위해 group=1을 사용한다. 선 그래프로 표현하였다.



1. 문제3에서 제시한 의견의 근거를 서술하시오.

구글 검색을 통해 “한 권으로 끝내는 비트코인 혁명”이라는 책을 찾을 수 있었다. “글로벌 위기가 부각시킨 비트코인의 존재감”에서 브렉시트와 키프로스 사태를 예로 들어 설명한 것을 볼 수 있었다.



1. 문제해결과정에서 겪었던 어려움을 서술하시오.

브렉시트는 영국에서 일어난 일이기 때문에 google.uk나 YAHOO.uk의 기사를 크롤링하고 싶었으나 POST방식을 사용해서 크롤링하기 쉽지 않았다. 상대적으로 영향력이 작았던 우리나라에서 데이터를 구해 아쉬웠다. 그리고 워드클라우드로 표현하였을 때 빈도수가 부각되지 않아 어려움을 겪어 head,sort함수를 이용하여 추가로 빈도수를 나타내었다.

키프로스 사태를 크롤링하여 워드클라우드로 나타내고 싶었지만 2013년에 비트코인과 연관된 뉴스기사 데이터가 많이 부족해 차선책으로 ggplot 그래프로 나타내었다.

데이터 전처리 과정에서 분석과 관계없는 단어들이 많이 나타났고, KoNLP패키지의 형태소분류가 아쉬웠다.