|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | |  | | |
|  | |  | | |
| 텍스트 마이닝  *비트코인* | | | | |
|  |  | | |  | |
|  |  | | | 이현호 | |

내용

[배경 및 목적 3](#_Toc520327780)

[수집한 데이터의 형태 3](#_Toc520327781)

[전체 데이터 셋 3](#_Toc520327782)

[일별 단어 빈도 4](#_Toc520327783)

[데이터의 연관성 9](#_Toc520327784)

[네트워크 키워드 분석 9](#_Toc520327785)

[감성분석 10](#_Toc520327786)

[결론 및 느낀점 11](#_Toc520327787)

텍스트 마이닝

비트코인

# 배경 및 목적

비트코인와 블록체인이 사회적으로 대두되면서 많은 양의 기사와 컬럼이 쏟아져 나오고 있습니다. 많은 양의 기사와 컬럼은 보편적으로 사회에 큰 영향을 미치는 것으로 알려져 있는데, 그것이 사실인지에 대해 객관적으로 알아보고자 텍스트 마이닝을 시도하게 되었습니다.

텍스트 마이닝의 대상은 네이버의 기사로 선정하였고, 7일간 약 2000개의 기사로 실행하였습니다. 각 일자별로 추출된 단어에 대해 빈도분석을 실행하여, 비트코인 및 다른 가상화폐와 기사의 연관성에 대해 알아보는 작업을 할 것입니다.

# 수집한 데이터의 형태

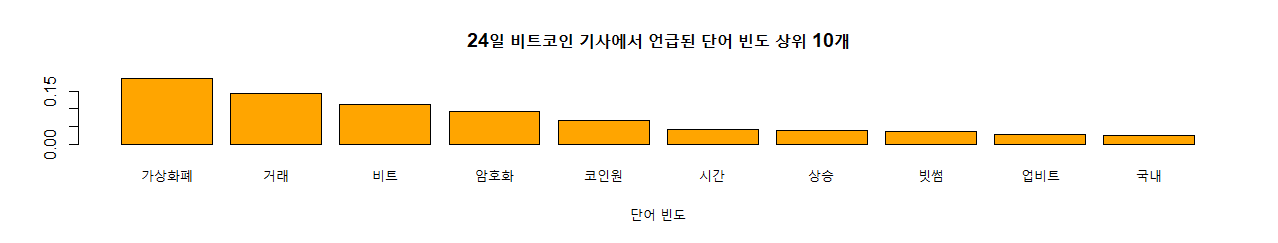
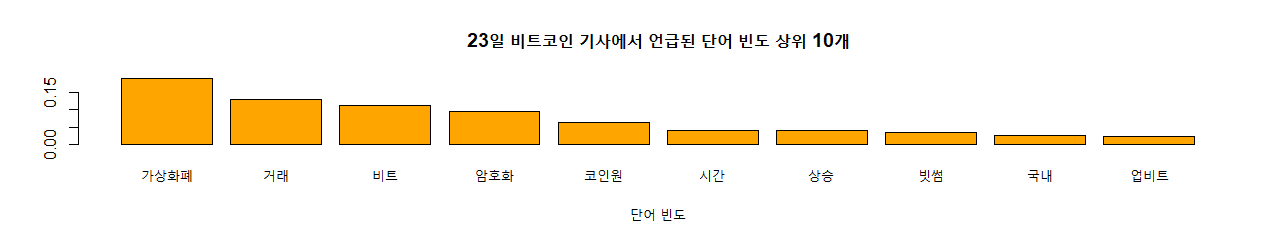
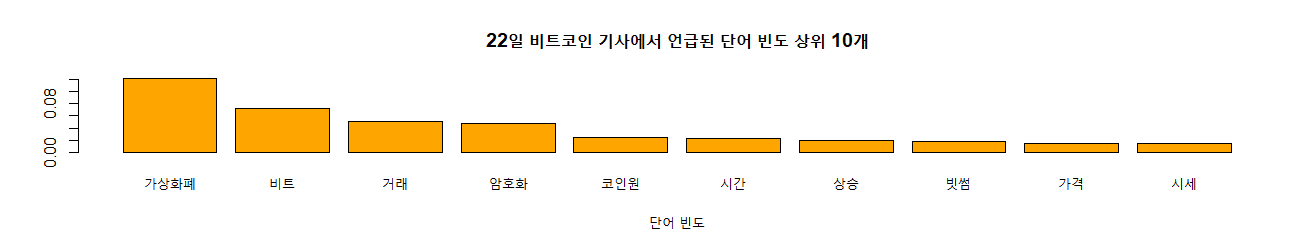
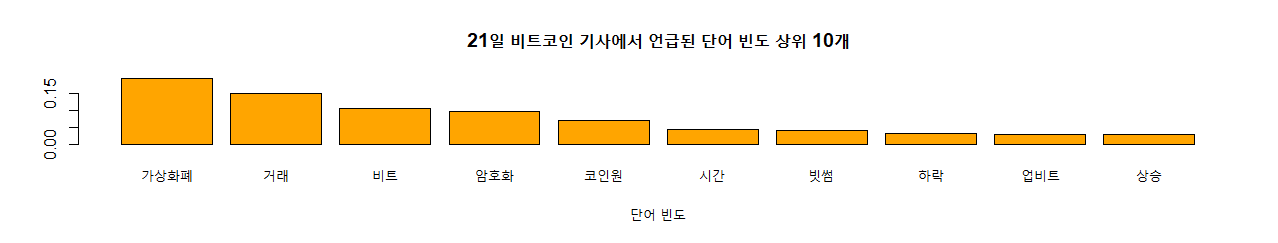
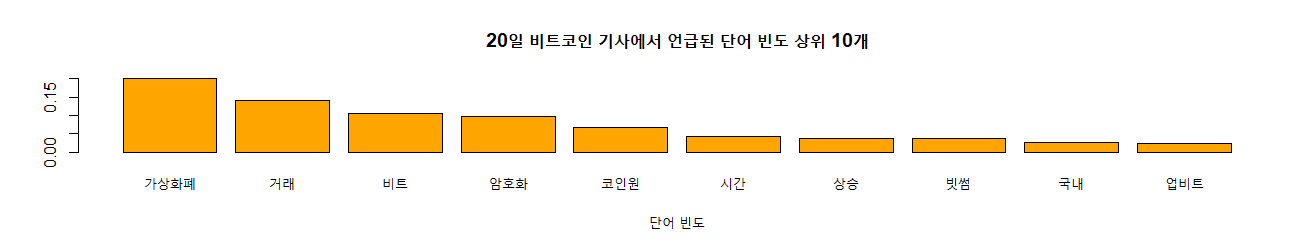
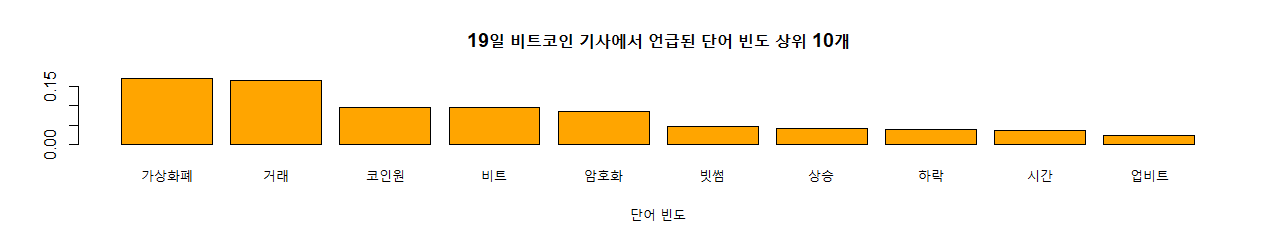
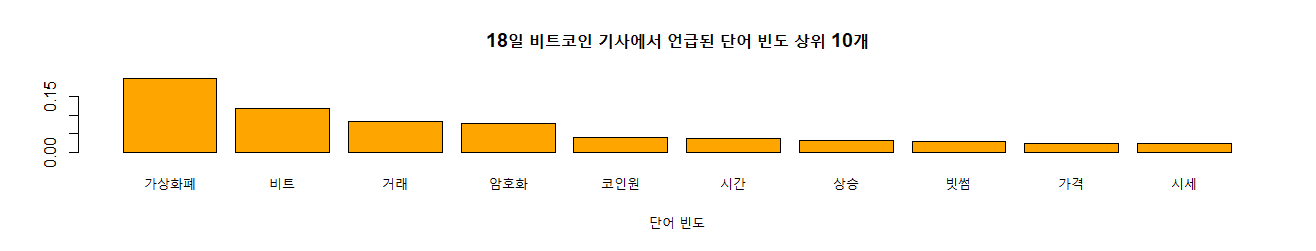
## 전체 데이터 셋

|  |  |
| --- | --- |
| C:\Users\renz\Downloads\Desktop\비트코인 시각화 워드클라우드2.png | 7일 간 2000개의 뉴스를 워드클라우드로 나타낸 그림입니다.  주요 단어로는 거래, 암호화, 코인원, 상승, 하락 등의 단어가 보이고 있습니다.  또한, 비트코인과 관련있는 단어인 it 혹은 코인거래소, 비트코인 외 다른 코인의 이름, 뉴욕 등이 확인 되고 있습니다.  거래와 코인원의 경우 비트코인의 특성상 기사에 잘 드러나는 것이 당연 |
| 하다고 느껴지지만, 암호화라는 단어가 빈도가 굉장히 크게 나타나는 것이 매우 흥미롭습니다.  즉 기사에서 긍정적, 부정적의미에 관계없이 암호화라는 말을 자주 차용한다는 점은 가상화폐에서 | |

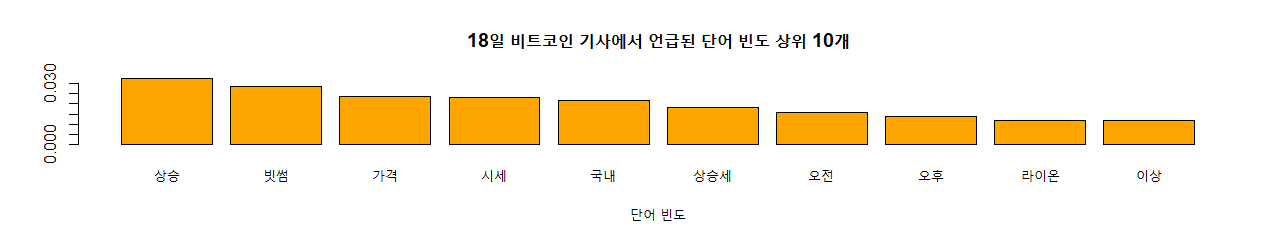
암호화를 굉장히 심도있게 다룬다는 점을 알 수 있습니다.

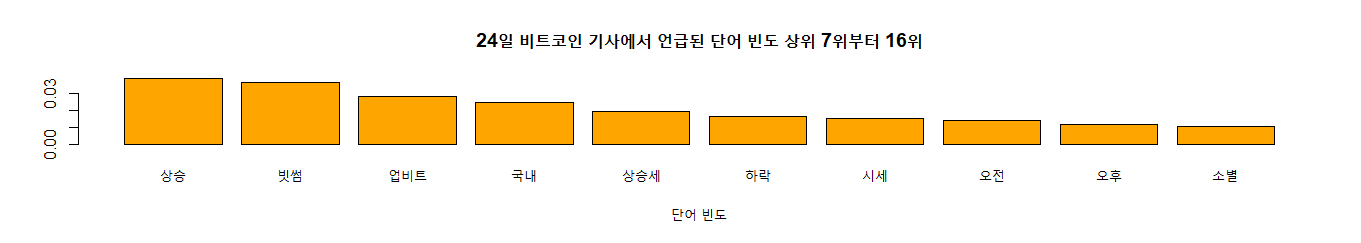
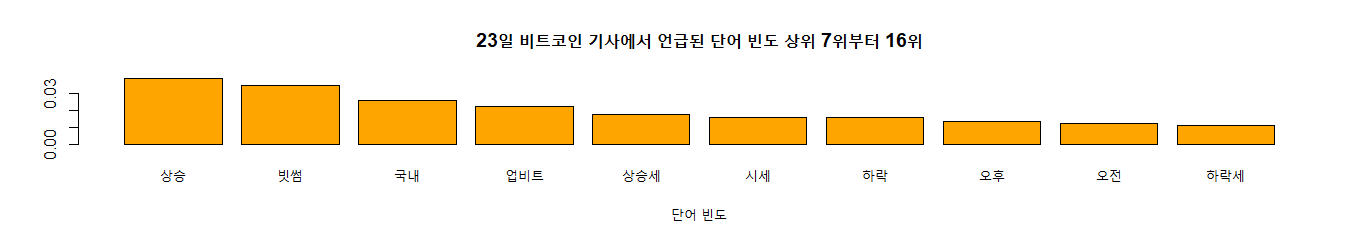
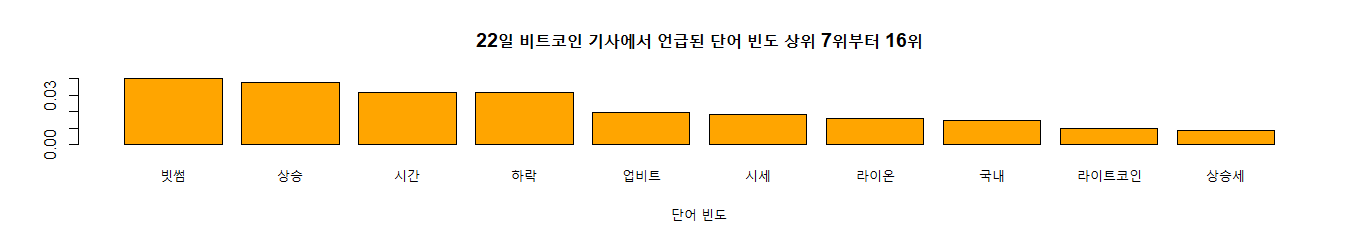
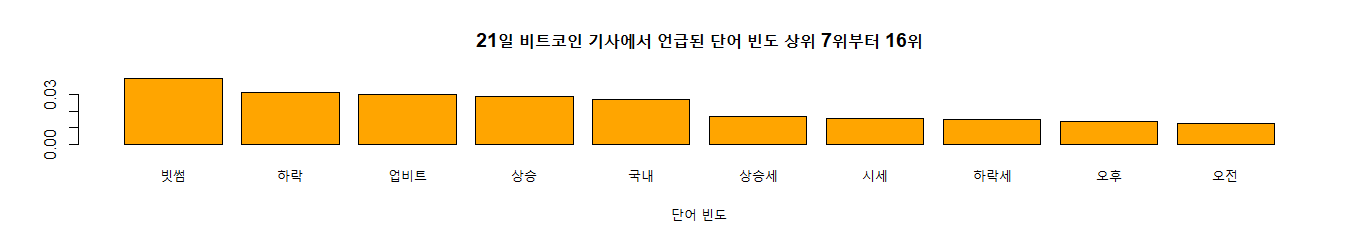
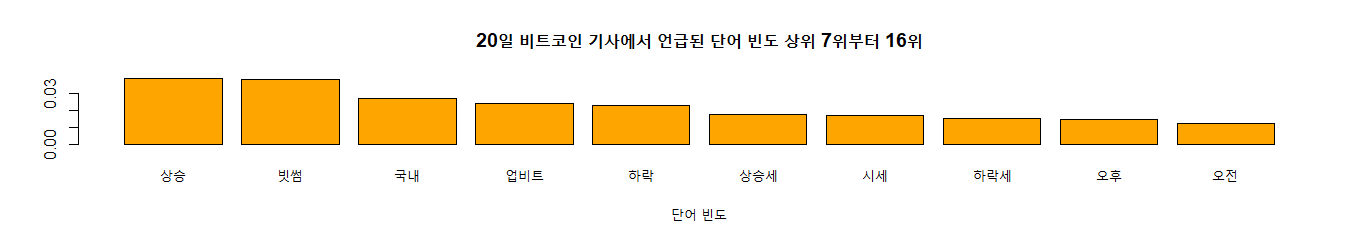
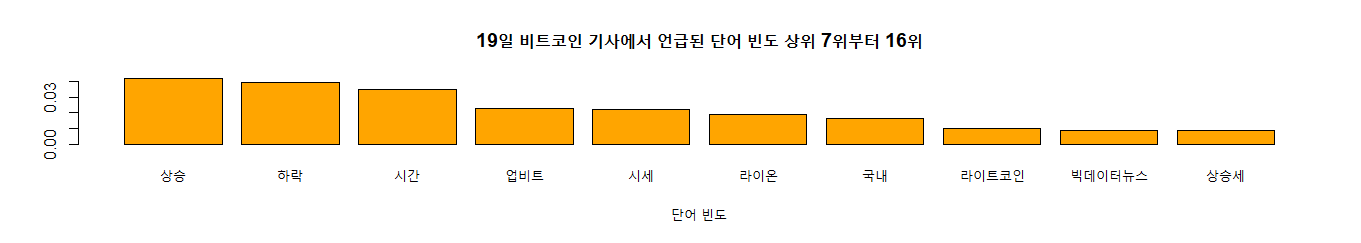
## 일별 단어 빈도

|  |  |
| --- | --- |
|  |  |
| 18일 | 19일 |
|  |  |
| 20일 | 21일 |
|  |  |
| 22일 | 23일 |
|  | 각 단어에 대한 일별 빈도수를 알 수가 있습니다. 재미있는 점은 가상화폐와 암호화, 거래는 거의 항상 높은 빈도수를 유지하고 있다는 점입니다. 또한 일별 빈도수를 통해 일마다 변화하는 키워드에 대해 알 수 있습니다.  그러나, 워드클라우드는 단점도 많습니다. 수치화된 정보를 알 수 없고, 상대적인 그림이기에, 정확히 어느정도로 변하는지에 대한 정보도 없습니다.  그래서 다음과 같은 그래프를 그려봤습니다. |
| 24일 |

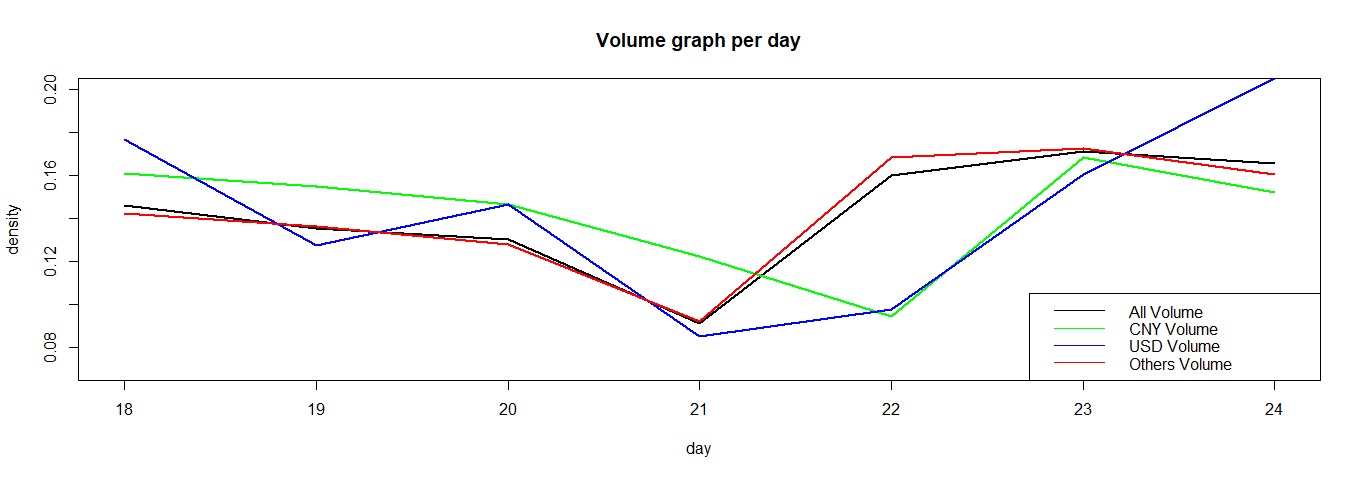


절대도수로 막대그래프를 그린 결과, 각 일별로의 빈도수를 알 수 있습니다. 상위 6개는 위에서 본 워드클라우드와 흡사하지만 상위 7위부터 10위까지는 다른 것을 알 수 있습니다. 상위 6개는 거의 흡사하므로 6개를 빼고 그래프를 그려보았습니다.

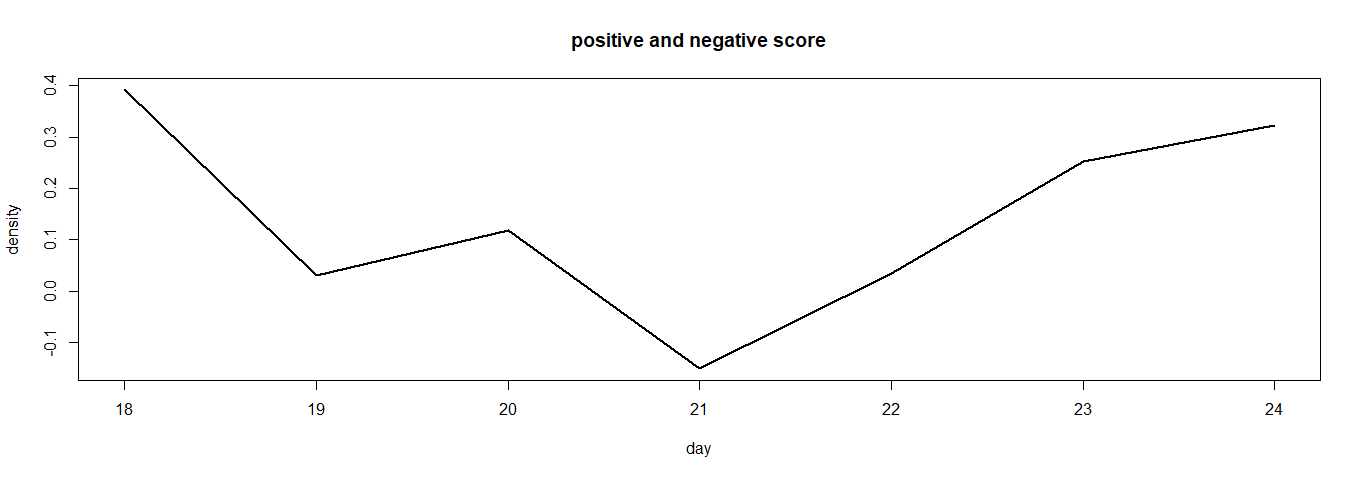




상승세와 하락세, 상승, 하락 같은 단어들도 등장하는 것을 볼 수 있습니다. 위와 같은 단어들이 실제로 가상화폐의 등락폭과 비교하기 위해 각 단어에 점수를 부여하고 수치화 하였습니다.



해당 그래프는 일별 코인 유동량을 도식화 한 그래프입니다.

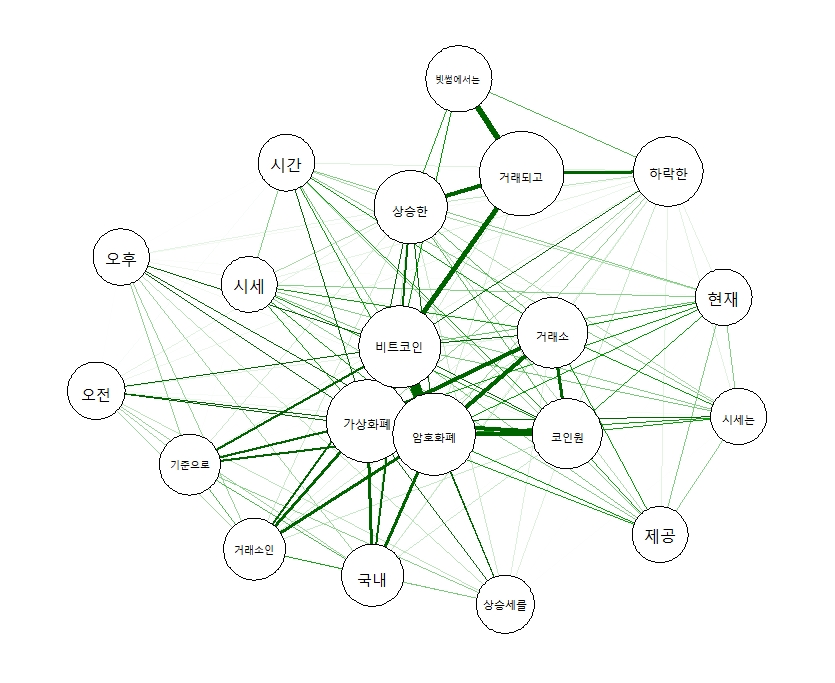


위의 그래프는 뉴스기사를 일별로 분류하여 감성분석한 스코어점수입니다. 놀랍게도 기사의 감성분석 스코어점수와 가상화폐의 움직임이 매우 유사함을 알 수 있습니다.

감성분석과 실제 주가 움직임은 뒤에서 더 자세히 살펴 보겠습니다.

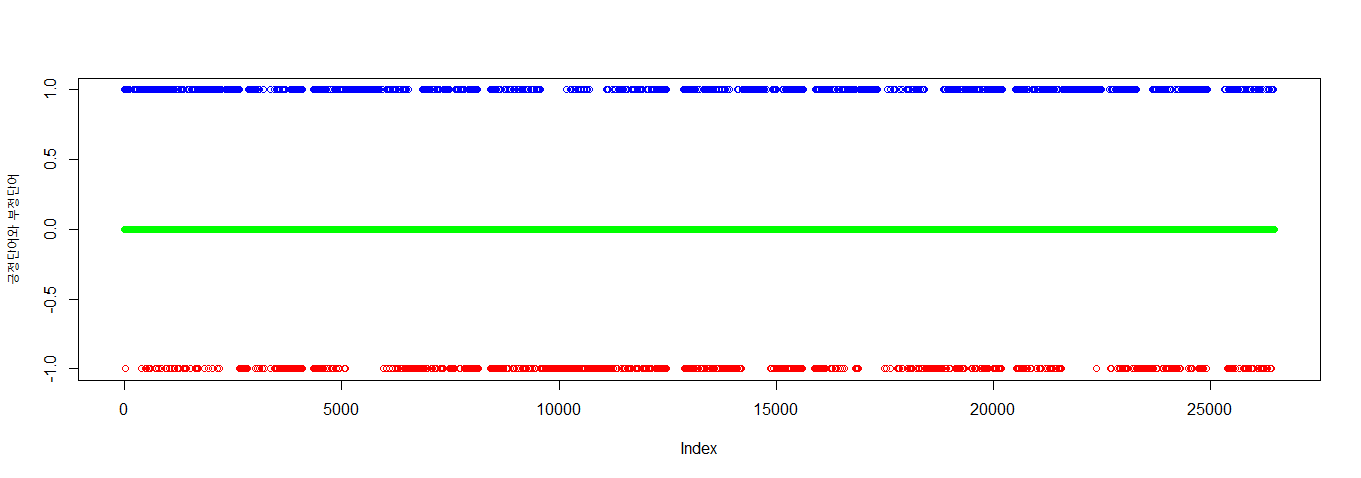
# 데이터의 연관성

## 네트워크 키워드 분석



네트워크 분석을 하여 연관 단어 즉 동시출현단어를 분석하였습니다. 국내의 암호화폐가 상승세인 만큼 상승에 대한 키워드가 많은 것을 확인 할 수 있고, 상승과 연결되는 단어들을 볼 수 있습니다.

# 감성분석

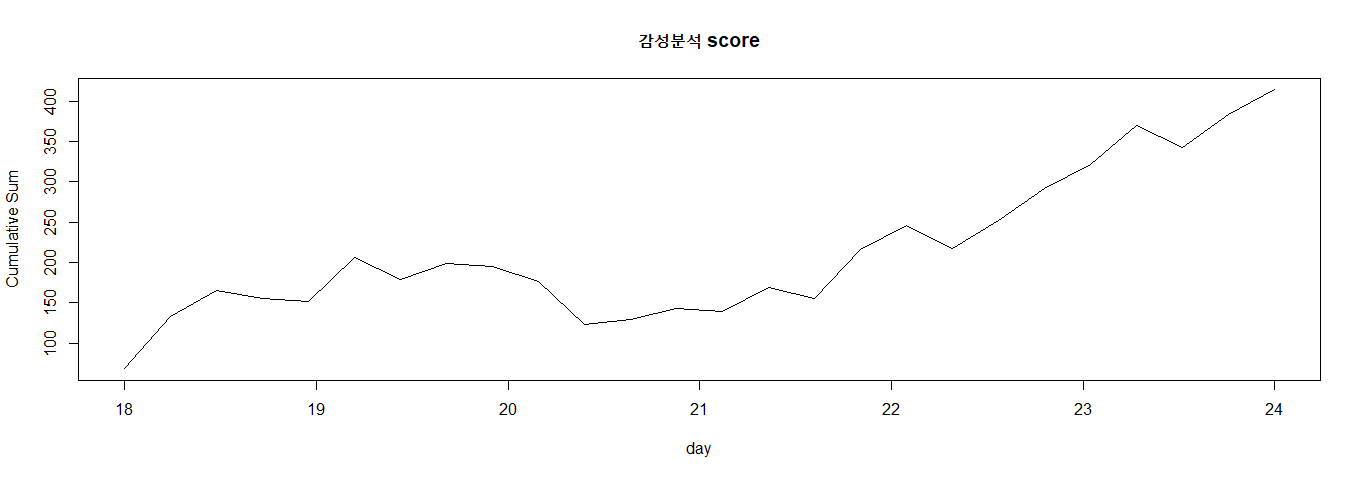




위의 그래프는 파란색은 긍정단어, 초록색은 중립단어 , 빨간색은 부정단어로 볼 수 있습니다.

밑의 그래프는 긍정단어와 부정단어 감정분석 결과화면을 보다 직관적으로 나타내어 좀 더 보기 쉽게 만들었습니다. Index 즉 요일마다 긍정단어, 부정단어 빈도를 쉽게 알 수 있습니다.

|  |  |
| --- | --- |
|  | 감성분석 결과를 빈도로 보는 원 그래프입니다. 중립단어가 매우 많은 것으로 보아 객관적으로 판단했을 때, 감성사전이 부족한 것으로 보입니다. 보완을 하여 최소한 1:1:2는 되어야 된다고 생각합니다. |



실제 비트코인의 시세와 감성분석 스코어점수를 누적합 한 결과를 비교해보았습니다. 매우 흡사한 패턴 즉 추세는 비슷하나, 중간중간 감소폭과 증가폭이 크게 다른 것을 볼 수 있습니다.

# 결론 및 느낀점

언론과 가상화폐는 매우 밀접한 관계에 있음을 알게 되었습니다. 가상화폐에 등락폭에 대해 언론은 매시각 긍정, 부정단어를 사용하여 기사를 쓰고 있었으며, 기사에 대한 영향은 매우 강하게 나타나는 것을 확인 할 수 있었습니다. 그 말은 사람들이 가상화폐에 대한 정보를 언론에 의해 빠르게 접해 빠른 대응을 할 수도 있지만, 잘못된 정보라면 큰 손실을 볼 수도 있다는 것을 보여줍니다. 즉 언론이 정확하고 신속하게 보도 하는 것은 맞지만, 오차가 아예 없는 부분은 아니니 맹신하면 안된다는 결론을 내렸습니다.

이번 보고서를 준비하며, 텍스트마이닝에 대해서 좀 더 알고, 배경을 알 게 되었습니다. 아쉬운 점은 연관성을 볼 때, 군집이라던지 다양한 시각화를 해보고 싶었지만, 여러가지 문제점에 봉착하여 못한 부분이 아쉽게 느껴졌습니다.

한학기 동안 열심히 가르쳐 주셔서 감사합니다.