**목 차**

**1. 프로젝트 개요**

1.1 연구주제 및 배경

1.2 종속 변수 독립변수

1.3 프로젝트 방향

**2. 자료 수집**

2.1 파이썬을 이용한 텍스트 크롤링 및 라벨링

2.2 Yahoo Finanace Kakao Corp stock historical data

**3. 자료 분석**

3.1 Machine Learning에 활용할Bert 소개

3.2 Google colab을 활용한 Machine Learning

3.3 R-studio를 활용한 분석 결과

3.4 Outlier 제거 후R-studio를 활용한 재분석 결과

**4. 한계점 및 소감**

4.1 한계점

4.2 개선사항

**5. 자료 출처 및 참고자료**

**1. 프로젝트 개요**

1.1 연구 주제 및 배경: 네이버 종목토론 게시판유저들이 올리는 정보는 카카오 주가를 예측할까?

1.2 종속 변수: 카카오 주가, 독립변수: 카카오종목토론게시판 유저들의 게시글 긍정 부정 비율

1.3 프로젝트 방향



빨간 색 박스 안에 있는 글 들은 카카오 주가에 대한 종목게시판 유저들의 의견이다. 해당 종목에 대한 의견은 누구나 상관없이 네이버에 로그인 만 하면 쓸 수 있다. 나의 연구 목표는 게시판 유저들이 쓰는 종목에 대한 정보와 의견들이 해당 종목을 예측 할 수 있냐는 것이다. 연구는 다음과 같이 진행 하였다. 해당 글들을 파이썬으로 크롤링 하여 google colab을 통해 머신러닝을 시키면 해당 날짜 카카오에 대한 종목토론게시판 유저들의 의견이 긍정적인지 혹은 부정적인지 알 수 있게 된다. 예를 들어 9월 18일 장이 끝난 15시 30분 이후부터 19일 장 개장 전 8시 59분까지의 게시판 이용자들의 의견이 긍정 60프로 부정 40프로로 나뉘면 실제로 9월 18일 카카오 주가가 다수의 의견에 따라 상승하여 움직이는지 조사하였다.

**2. 자료 수집**

2.1 파이썬을 이용한 텍스트 크롤링 및 라벨링



위에 보이는 것은 네이버 종목토론 게시판의 유저들이 쓴 게시판 글을 엑셀로 저장하는 코드이다. 한 페이지 당 20개의 글을 1000페이지까지 가져 왔다. 즉 2만개의 데이터를 네이버로부터 가져왔다. 가져오는데 총 8시간 정도 소요 되었으며 카카오 주가뿐만 아니라 다른 종목에 대해 조사를 할 수 있게 종목번호를 다른 것으로 입력하면 다른 회사에 대한 게시판 유저들의 의견도 갖고 올 수 있게 코딩 하였다.



위에 보이는 것이 네이버 종목토론 게시판으로부터 가져 온 텍스트 데이터다. 해당 날짜에 유저들이 어떤 의견을 내는지 보기 위해 글을 쓴 날짜와 게시글의 제목을 갖고 왔다. 이 데이터를 바로 기계학습을 시킬 수는 없다. 사람 손으로 하나씩 해당 글이 카카오 주가에 연관된 정보 인지, 또한 카카오에 대해 부정적인 의견을 갖고 있는 의견인지 혹은 긍정적인 의견을 갖고 있는 의견인지 연구자가 판단하여 라벨링을 해야 한다.



위에 보이는 것이 라벨링이 된 엑셀 파일 이다. 12월 1일에 한 유저가 단 글인 “아침장에 부지런히 올려라” 라는 글은 카카오가 오를 것이라고 아침 장에 생각을 하고 보유 해야 한다는 의견으로 판단하여 긍정을 의미하는 1을 붙여줬다. 반대로 카카오 버리고 셀트리온 갔다는 뜻은 카카오가 더 이상 상승여력을 보이지 않는다는 부정적인 의견으로 판단하여 부정을 의미하는 0을 붙여줬다. 그 외 라벨링이 되어 있지 않은 데이터 들은 해당 게시 글이 카카오와 연관되어 있는지 판단하기 힘들어 라벨링을 하지 않았다. 예를 들어 “모더나 긴급 승인”이라는 글은 카카오와 연관점을 찾지 못하여 라벨링을 하지 않았다. 허나 “모더나 긴급승인으로 인해 언택트 주 카카오는 하향세를 탈 것” 이라는 글은 카카오에 대해 부정적인 데이터 이므로 0을 붙여줬다. 이런 식으로 2만 개의 데이터 중에 5000개를 라벨링 하였다. 본래 정확도를 올리기 위해서는 다다익선이지만 종목 게시판 데이터를 읽으면 읽을수록 중복되는 글이 너무 많고 실제로 기계학습을 시켜봤을 때 정확도를 의미하는 epochs accuracy 가 78%로 크게 낮지 않아 5000개에서 중지하였다.

2.2 Yahoo Finanace Kakao Corp stock historical data

종속변수 이자 Y 값인 카카오 주가를 8월 31일부터 Yahoo Finance 에서 갖고 왔다.

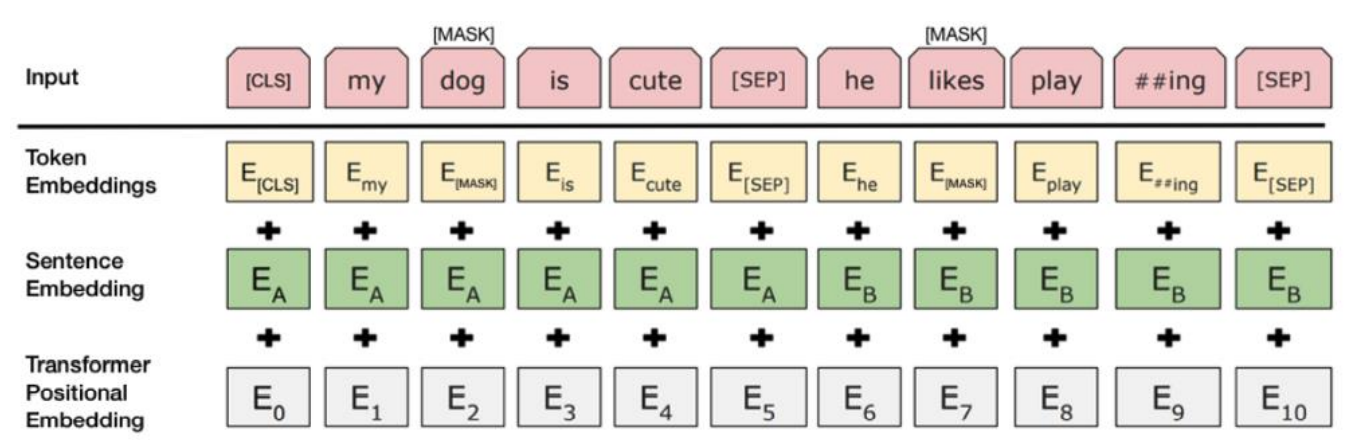
**3. 자료 분석**

3.1 Machine Learning에 활용할Bert 소개

Bert 모형을 완벽하게 100퍼센트 이해하기에는 관련 학과 비전공 학부생인 나로서는 불가능하다. 엄밀히 말하면 Bert 모형은 Machine Learning에 속하는 Deep Learning 분야이다. 내가 프로젝트를 진행하면서 이해한 만큼 소개를 하겠다. Bert 모형은 최근에 개발된 NLP 모델 즉 자연어 처리 모델이다. 범주형 비정형 변수인 텍스트를 벡터화 시켜 정형 데이터로 전환시켜 기계가 학습해 새로운 데이터가 들어왔을 때 예측하는 것이다. 우리가 수업시간에 배웠던 회귀분석의 매우 고차원 적이고 첨단적인 형식이라고 이해하면 편하다. 즉, 아주 쉽게 설명하자면 글을 컴퓨터가 이해할 수 있는 숫자형 데이터로 만들어 컴퓨터에 학습시키고 이 학습된 내용을 바탕으로 새로운 텍스트가 들어왔을 때 새로운 텍스트가 어떤 내용을 담고 어떤 감정을 담고 있는지 예측하는 모델이다. 자연어 처리 모델은 시중에 많이 존재하지만 Bert를 선택한 이유는 데이터 Embedding 작업이 모델에 내재되어 있어 비전공자들도 쉽게 활용할 수 있기 때문이다. 파이썬을 조금이라도 다룰 수 있는 사람은 코드가 비록 길긴 하지만 누구나 활용할 수 있는 Machine Learning 분석 방법이다.



예시를 들어보면 “이번 주 얼마까지 떨어질까요?” 라는 문장을 위에 모델에 내장되어 있는 범주형 데이터를 벡터 형으로 전환시키는 작업을 통해 기계가 학습하는 것이다.



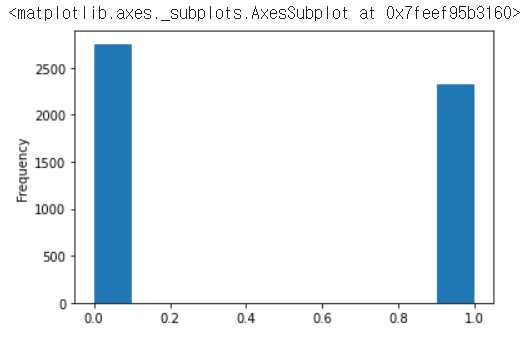
Bert는 위에 보이는 그림과 같은 방법으로 학습을 한다. 형태소 별로 문장에 포함된 글자들을 찢은 다음에 이를 벡터화 시켜 각각의 값들을 모두 더해 학습하는 방식이다. 이해하기 너무 어려우면 Machine Learning을 통해 문장의 성향을 컴퓨터가 학습된 데이터를 바탕으로 판단하고 결과값을 알려준다 라고 생각하면 된다.

3.2 Google colab을 활용한 Machine Learning

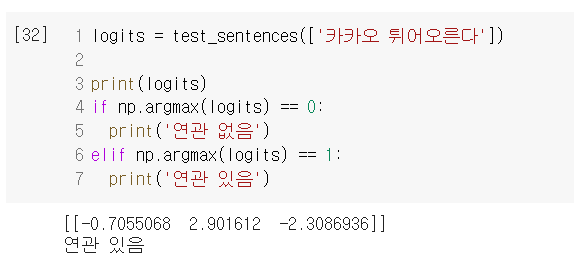
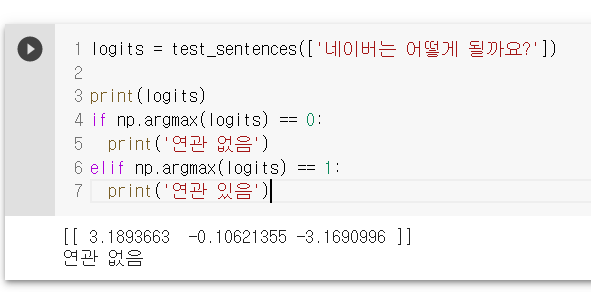
프로젝트를 진행하면서 기계학습을 총 두 번 시켰다. 첫 번째 기계학습은 카카오에 연관된 텍스트 데이터 인지 아닌지에 대해 먼저 기계학습을 시켜 2만개의 데이터 중에 카카오와 실제로 연관되어 있지 않은 데이터들을 걸렀다. 코드가 너무 길어 이미지 파일로는 첨부하지 못하고 코드를 첨부하겠다.



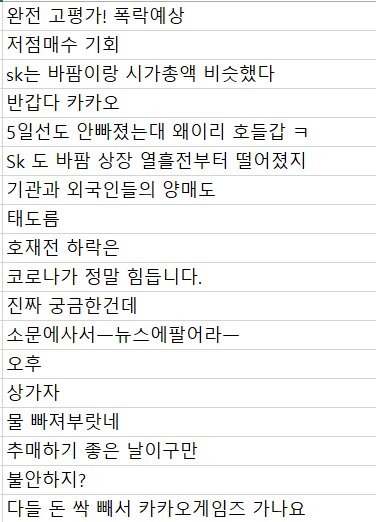
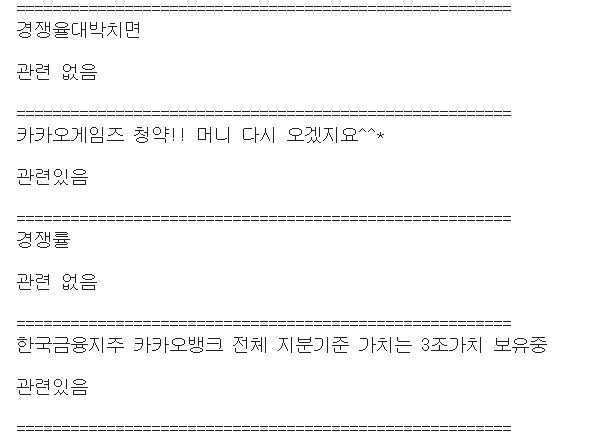
위에 보이는 엑셀 파일은 카카오에 관련 있는 정보라고 생각되는 것을 1, 관련 없다고 생각 되는 것을 0으로 라벨링 하여 기계에게 학습시킬 데이터 이다.



실제로 5000여개의 라벨링한 데이터 중에 카카오에 관련된 정보를 포함하고 있는 게시물은 전체 게시물의 50%도 되지 않는 것을 확인 하였다. 위에 히스토그램이 카카오에 관련된 정보를 먼저 판별하는 것이 우선적이라는 거의 증명이 된다.



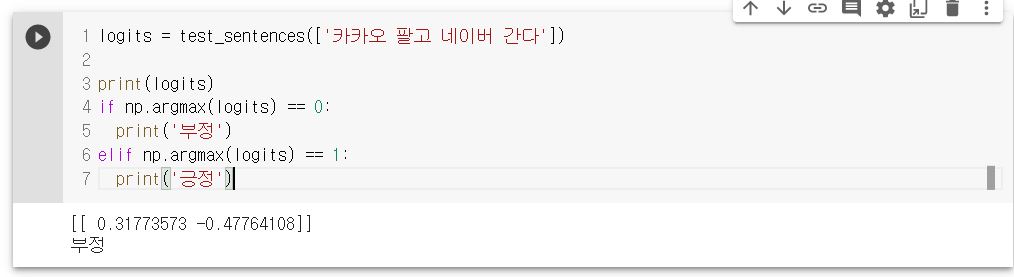
위의 엑셀 파일을 학습시키면 위와 같은 결과가 나오게 된다. 네이버에 관련된 정보는 기계가 학습한 데이터를 바탕으로 판단하여 카카오와 관련이 없다는 값인 0으로 판단을 하고 카카오에 관련된 데이터는 1로 판단을 하게 된다.

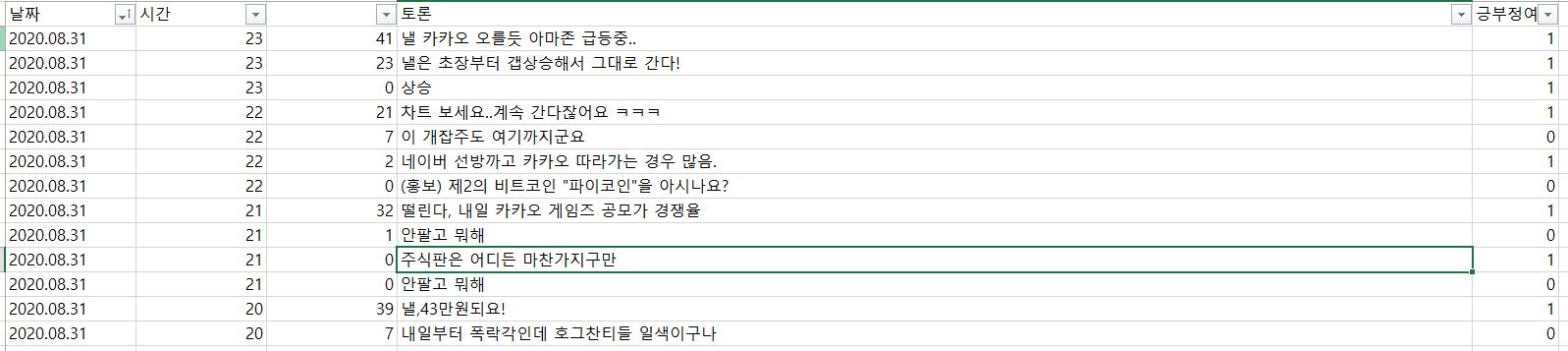
20000여개의 뽑은 데이터 중 기계학습을 통해 카카오에 관련된 데이터는 6000여개에 불과하다는 것을 기계학습을 통해 확인하였다. 이제 이 6000여개의 데이터를 긍정 인지 부정인지 판별하겠다. 기계 학습을 시키는 과정과 방법은 카카오에 관련된 정보를 판별하는 방법과 동일하다.



위에 보이는 엑셀은 카카오 주가에 관한 긍정 부정 의견을 학습시키기 위한 학습데이터다. 위에있는 데이터를 앞에서의 기계학습 방식과 동일한 방식으로 학습을 시키면 6000여개의 데이터의 긍정 부정 여부를 Bert 모델이 알려준다. 아래는 학습이 잘 되었는지 실험한 결과 값이다..

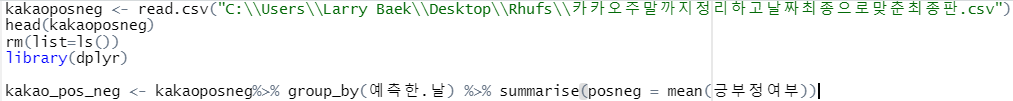
  

예시로 선택된 문장들의 긍정 부정 의도를 잘 예측한 것으로 파악할 수 있다.

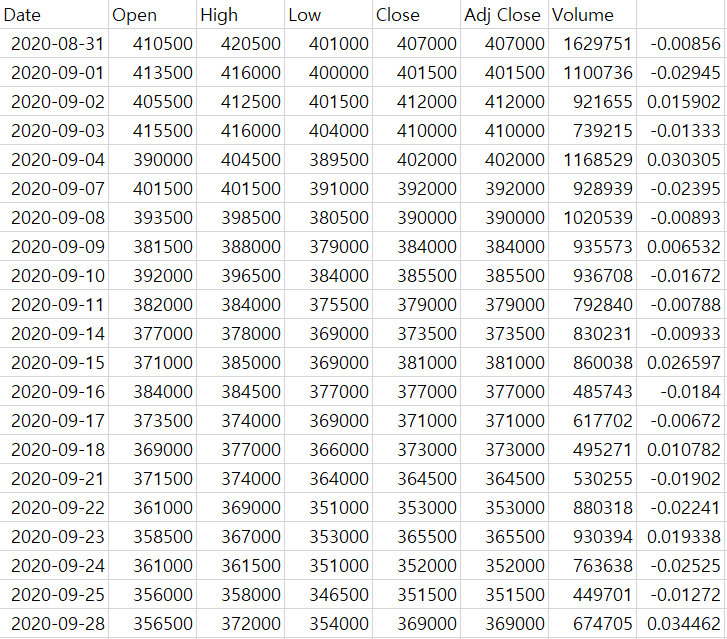
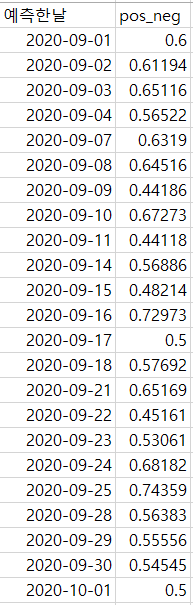


위에 보이는 엑셀 파일이 카카오와 관련이 있다고 기계가 파악한 6000여개의 데이터의 긍정 부정 여부를 최종적으로 파악한 데이터다. 이제 6000여개의 데이터를 날짜 별로 긍정이 몇 퍼센트 나왔고 부정이 몇 퍼센트 나왔는지 파악해야 한다.

3.3 R-studio를 활용한 분석 결과

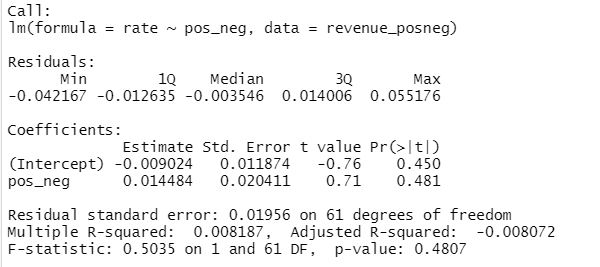


위에 보이는 것은 R-studio코드 인데 dplyr라는 package를 깔아서 날짜 기준으로 긍정, 부정의 평균을 낸 것이다.

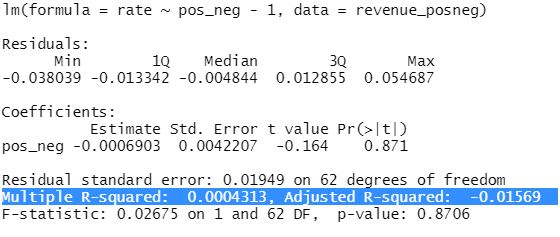


왼편에 보이는 데이터는 날짜 별로 긍정, 부정 둘 중에 어떤 의견이 더 우세한지 보여준다. 데이터를 해석해보면 2020년 8월 31일에는 종목토론게시판의 긍정의견이 60프로 부정의견이 40프로 차지한다는 것을 알 수 있다. 따라서 9월 1일에 주가가 오를 것이라는 긍정적인 의견 전체 의견 중 60프로다. 반면에 9월 10일에는 긍정 44프로 부정 56프로로 부정의견이 더 우세하다는 것을 알 수 있다. 그러므로 9월 11일에 주가 하락을 예측하는 의견이 대다수 라는 것을 알 수 있다. 예측한 날은 추석연휴, 주말, 공휴일 모두 반영한 것이다. 이제 왼편의 데이터를 Yahoo Finance에서 카카오 주식가격 정보를 갖고 와서 대조해 보면 종목토론 게시판 유저들의 의견이 실제로 주가를 예측하는지 알 수 있다. 오른쪽 데이터는 Yahoo Finance에서 갖고 온 카카오 주가 데이터다 오른편 끝 빨간색 상자 안에는 log(장 마감 가격/ 개장 시 가격)한 것 이다. 따라서 해석해보면 8월 31일 카카오 주가는 실제로 0.85% 하락한 것을 알 수 있다.

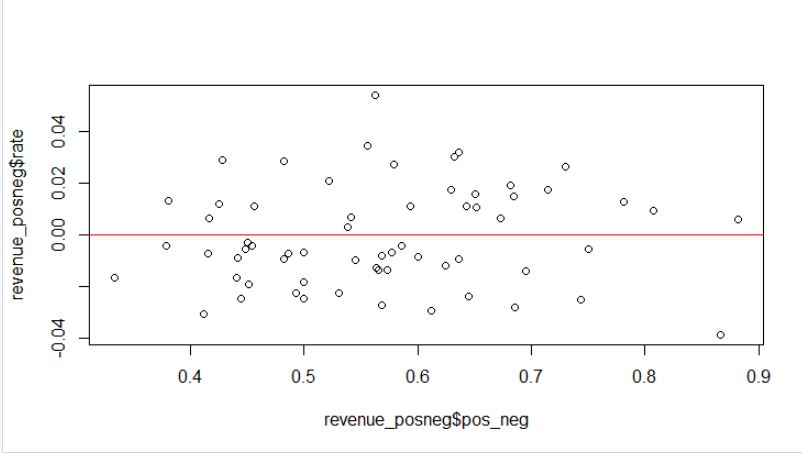


상단에 위치한 엑셀 데이터는 실제 주가 등락률과 종목토론게시판 유저들의 카카오 종목에 대한 긍정 부정 의견을 합친 데이터다. 이제 이 데이터를 R을 이용하여 회귀분석을 해보면 종목토론 게시판 이용자들의 예측력을 시험해 볼 수 있다. 

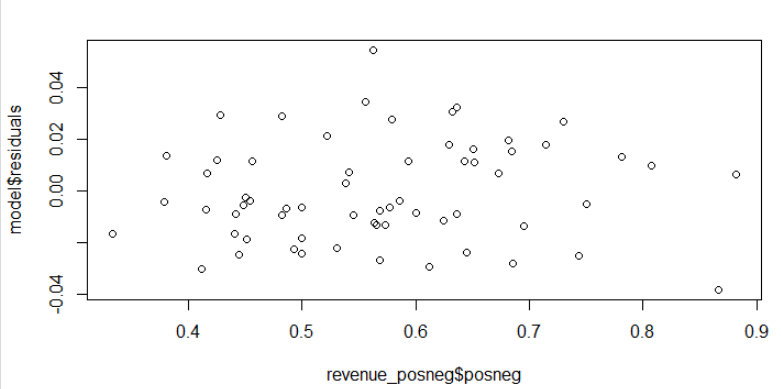
위의 검증 값을 확인해보면 종목토론게시판 이용자들의 의견은 예측력이 전혀 없다고 해석 할 수 있다. P-value 와 R-squared 값은 전혀 유의미한 값을 보여주지 못하고 절편 값과 긍정 부정 비율을 의미하는 pos\_neg 비율 모두 유의미 하지 않다고 R에서 보여주고 있다. 혹시 몰라서 유의미 하지 않은 절편 값을 빼보았다.



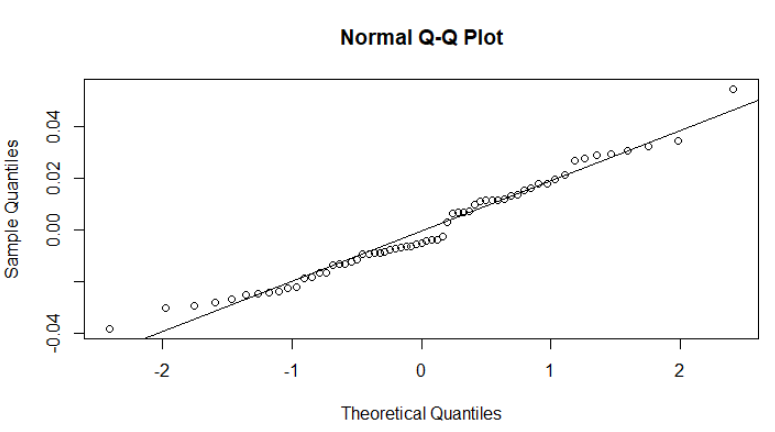
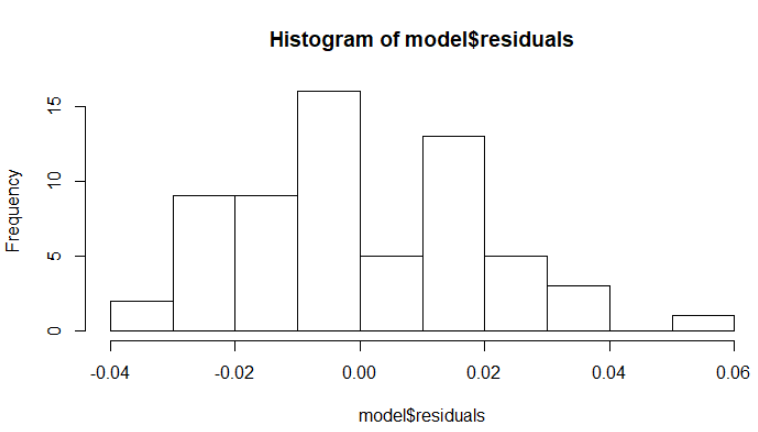
더 유의미하지 않아진 것을 확인 했다. 선형관계가 전혀 존재하지 않는 다는 것을 보여주고 있다. Scatter plot 을 그려봐서 선형관계가 전혀 없는지 확인해 봤다.



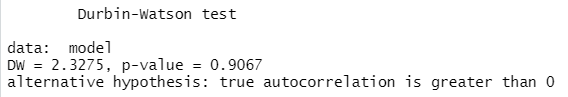
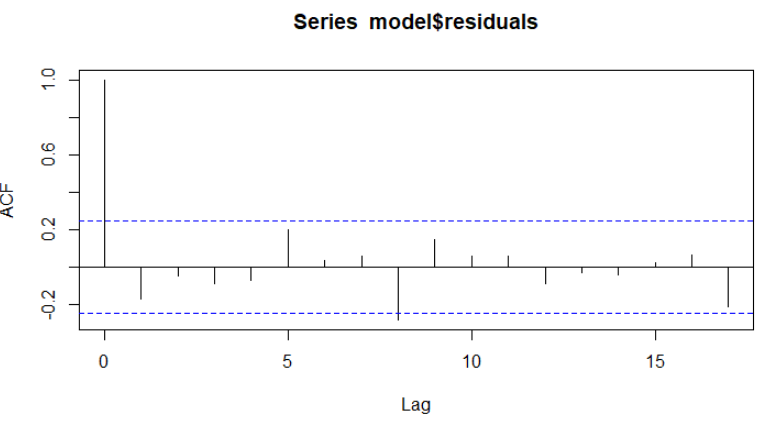
눈으로 봤을 때도 선형관계가 없는 것을 확인 할 수 있다. 혹시 다른 가정이 문제 인지 잔차 검정을 시행해봤다.



잔차 plot을 찍어봤는데 scatter plot 과 유사한 것을 확인했다. 즉 선형관계가 존재하지 않기 때문에 찍혀 있는 모든 분포들이 잔차로 인식되는 것 같다. 일단 잔차들이 점차 커지거나 작아지는 현상은 육안으로 보이지 않고 일정한 패턴을 띄는 것 같지도 않다. 실제로 검증을 해봤다.



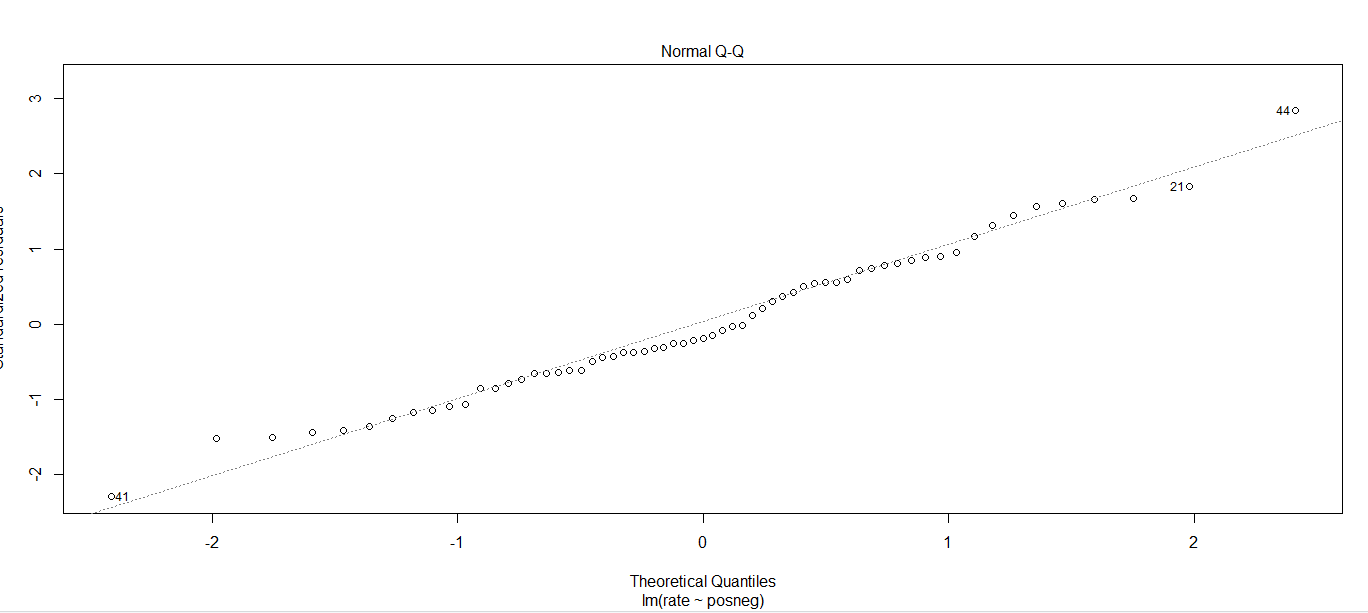
표본이 더 많으면 정규분포의 형식이 될 가능성이 있는 히스토그램을 보여주고 있다. QQ-plot을 찍어봤을 때도 outlier 몇 개를 제외 하면 선 위에 대체로 존재하고 있다고 봐줄 수도 있다.

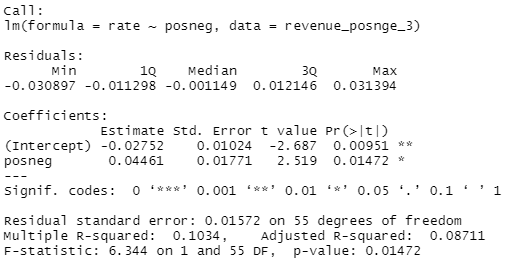


자기상관 문제를 확인해 봤을 때도 8번째 값을 제외 하고는 다 파란선 안 쪽으로 들어오는 것을 확인했다. Durbin-Watson 검정도 2에 가까운 값을 보여주고 있으며 p-value 또한 0.05보다 훨씬 큰 값을 보여주고 있기 때문에 자기상관문제는 없다고 판단해도 무방해 보인다. 즉 다른 가정의 문제 때문에 선형관계가 존재하지 않는 것이 아니고 처음부터 주가와 종목토론 게시판 의견 사이에는 상관관계가 없다. 즉, 종목토론 게시판 유저들은 주가의 등락을 맞추지 대체적으로 맞추지 못하는 것이다. 이를 보고 우리가 판단 할 수 있는 것은 인터넷 여론을 보고 무작정 해당 종목에 투자 하면 안 된다는 것이다.

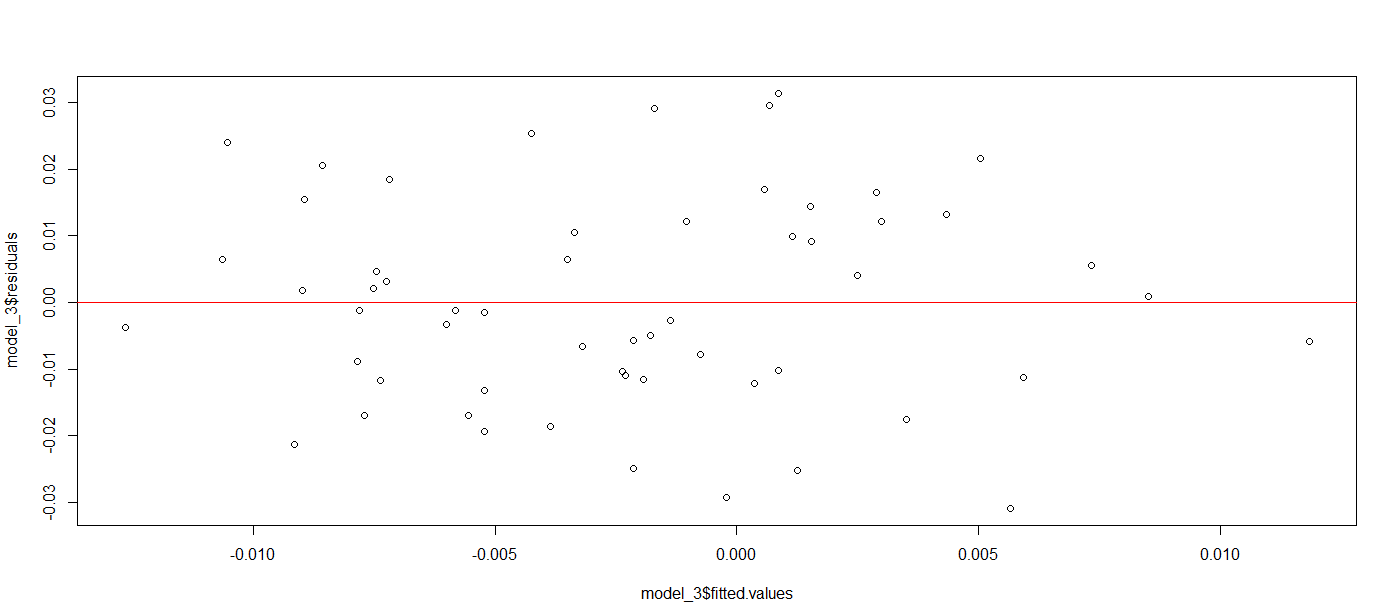
3.3 Outlier 제거 후 R-studio를 활용한 재분석 결과

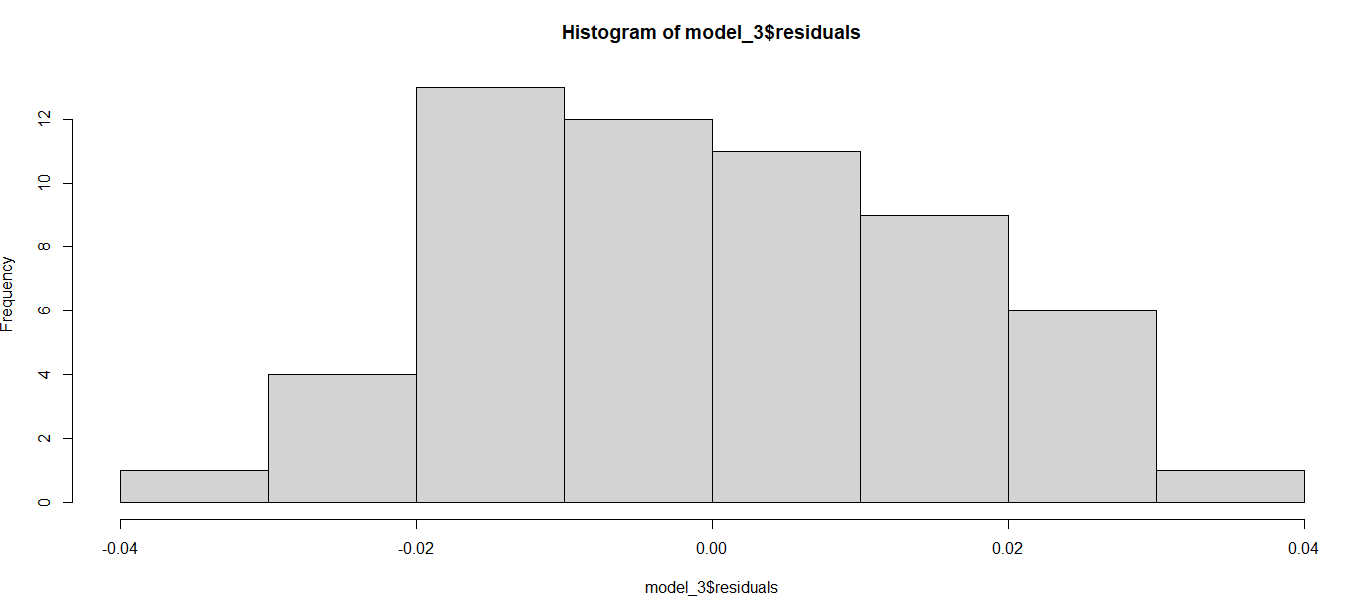
아웃 라이어 제거 전 R-studio를 통해 plot 을 찍어봤을 때 44, 21, 41,51,38,59 번째 행이 outlier라고 알려주고 있어서 제거해 본 후 단순회귀분석을 돌려봤다.

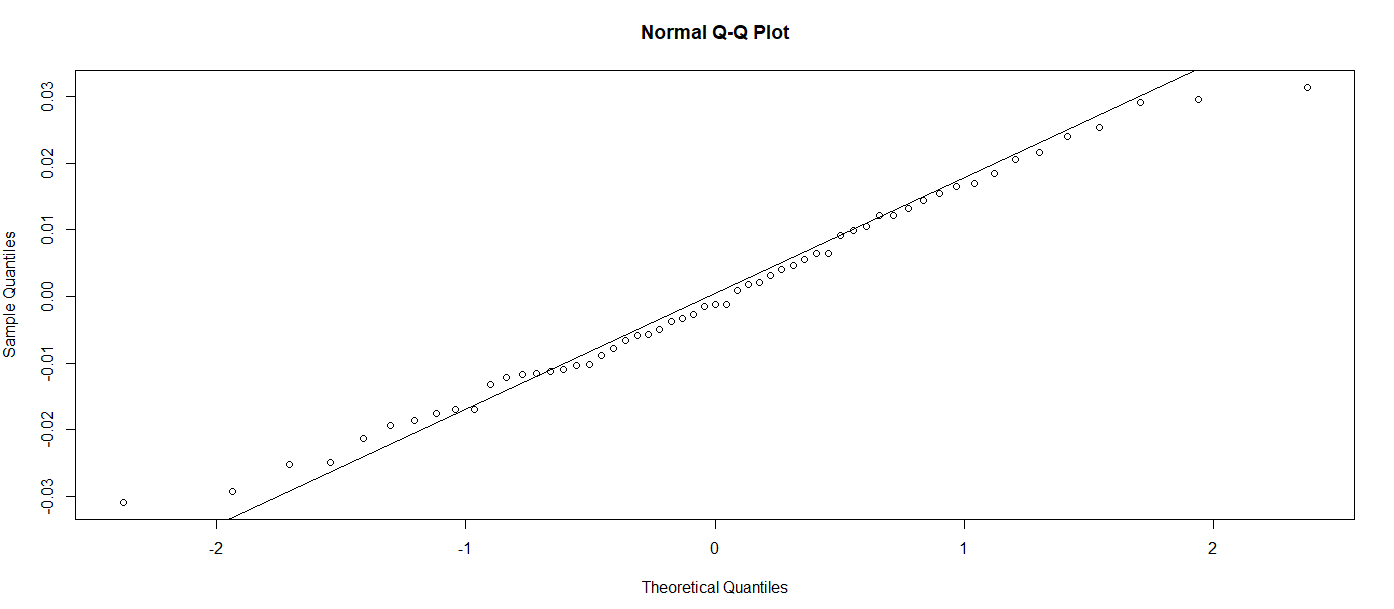


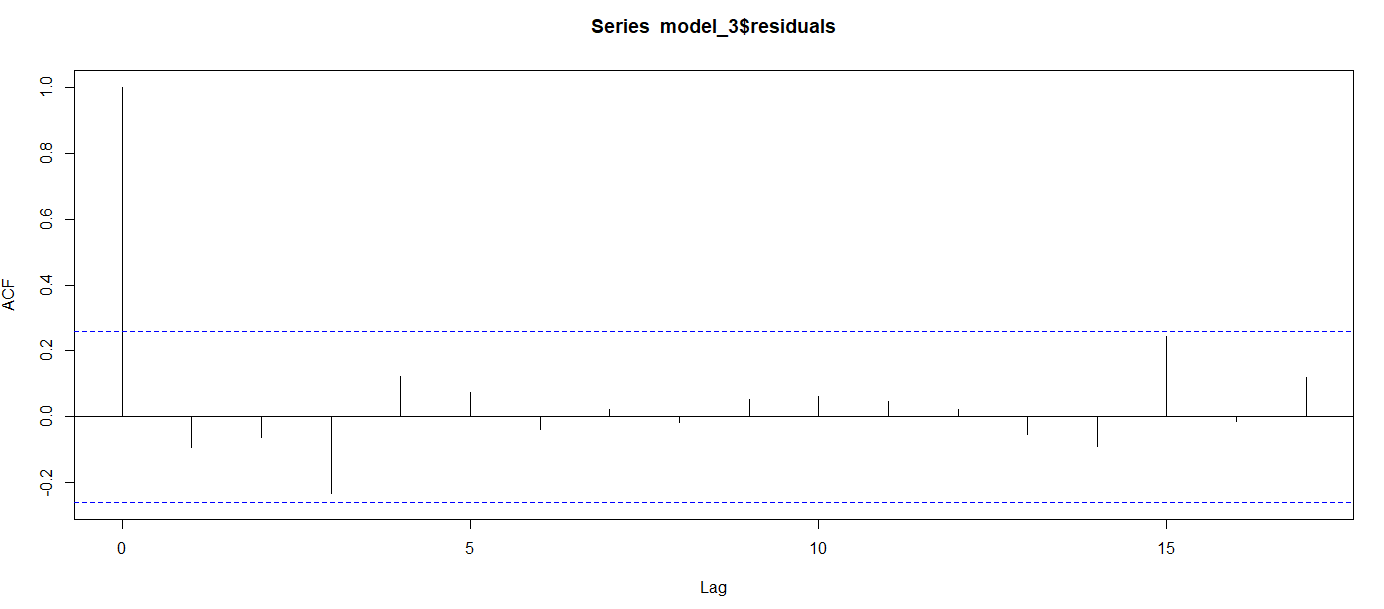


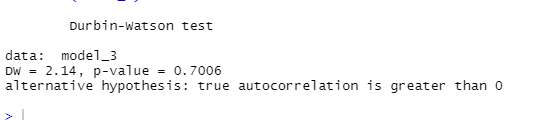
Outlier 제거 후 y 절편과 긍부정 비율 모두 카카오 주가와 유의미한 상관관계가 있다고 통계적으로 보여주고 있다. 따라서 이상치 제거 후 최종 모형으로는 y = -0.02752 + 0.04461X1 + e 이라는 공식을 도출할 수 있다. 타겟 변수와 독립변수 모두 자연 log가 씌워져 있기 때문에 해석은 다음과 같이 할 수 있다. 카카오 종목토론 게시판의 긍정, 부정 비율이 1% point 증가할 때마다 카카오 주가는 평균적으로 0.04%point 증가한다고 해석할 수 있다. 다음으로는 outlier 제거 후에도 단순선형회귀모형의 성립을 위한 가정이 만족하는지 확인해 보겠다.









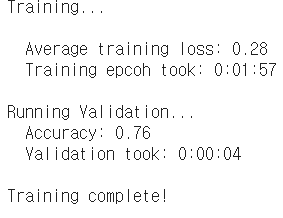


**위에 제시한 도표들을 통해 미루어 볼 때 단순선형회귀모형 분석을 위한 가정에 아무 문제가 없다는 것을 확인 할 수 있다.**

**4. 프로젝트 개요**

4.1 한계점

첫 번째로 한국어는 알파벳 기반의 언어처첨 띄어쓰기 단위로 명확하게 형태소가 구분되지 않아 자연어 처리 후 결과가 100퍼센트 정확하지 않다.



Bert 모형을 통해 예측된 값이 모형이 자체적으로 분석한 결과 76%프로의 정확도를 보인다고 이야기 해주고 있다. 데이터 라벨링을 업계에서 많은 사람들을 동원해서 수십만에서 수백만 건을 진행하면 해소될 문제이기는 하지만 혼자서 프로젝트를 진행하기 때문에 많은 수의 라벨링을 진행하기에는 많은 제약사항이 있었다.

두 번째로 종목토론 게시판의 특성이 주식에 대한 정보만 올리는 것이 아니라 자신의 아무 의견을 자유롭게 개진할 수 있기 때문에 관련 없는 정보들이 많이 올라온다. 프로젝트를 처음 진행했을 때는 대한민국을 대표하는 회사인 삼성전자로 진행하려고 했지만 삼성전자 게시판을 라벨링하면서 제일 많이 본 단어는 윤석열과 추미애다. 따라서 프로젝트를 진행하는데 무의미 하다고 판단하여 카카오로 진행하였다. 하지만 카카오 역시 인기가 많은 주식이라 다양한 사람들이 다양한 의견을 개진하였고 윤석열, 추미애, 문재인 등 관련 없는 주가와 상대적으로 관련도가 떨어지는 문장들이 상당수 확인 되었다.



약 2만개의 데이터 중 위에 보이는 거처럼 연관 없는 데이터가 1만4000여개 확인 되었다. 따라서 게시판은 처음부터 예측력이 있다고 기대하기에는 무리가 있다.

세 번째로는 게시판 글을 몇 명이 작성하는지 확인하는데 불가능하다. 네이버는 현재 아이디를 \*\*\*\*표시로 같은 아이디 인지 같지 않은 아이디인지 확인하기 어렵게 해놨기 때문에 한 명이 중복해서 다른 글을 계속 올리는지 아닌지 파악하는데 어려움이 존재한다. 따라서 1명이 100개의 부정적인 의견을 개진하는지 100명이서 100개의 긍정적인 의견을 개진하는지 확인하기 어렵다. 이는 연구진행에 있어서 큰 걸림돌이었다.

마지막으로는 게시판 이용자들은 당일 주식시장 혹은 당일 기준으로 마지막 주식 시장의 등락여부에 민감하게 반응하여 종목토론 게시판에 긍정적인 의견 혹은 부정적인 의견을 개진하는 패턴을 확인하였다. 예를 들어 9월 28일에 카카오 주가가 많이 올랐다면 다음 개장일인 9월 29일 카카오 주식에 대한 긍정적인 의견이 9월 28일 15시 30분부터 9월29일부터 8시 59분까지 많이 올라온다. 즉 미래에 대한 예측이 아니라 과거 주가 등락여부에 여론이 반응을 하는 것이다.

4.2 개선사항

시간이 더 많고 금전적인 여유가 된다면 코스피 200 지수에 속해 있는 모든 종목토론 게시판의 정보를 가져와서 수 백만 건의 라벨링을 진행하여 다음날 주가가 몇 퍼센트 변동하는지가 아니라 상승과 하락만 파악하고 싶다. 즉 주가 상승 혹은 하락이 대세 긍정, 부정 의견과 대조해서 상관관계가 있는지 조사해보고 싶다. 즉 종속변수를 범주형으로 해서 프로젝트를 더 심도있게 진행해보고 싶다. 또한 네이버 종목 토론 게시판뿐만 아니라 다음에도 비슷한 형태의 게시판이 존재하고 인터넷 상에 존재하는 다양한 게시판의 게시글을 크롤링하여 분석해 보고싶다.

**5. 자료 출처 및 참고자료**

**Yahoo Finance Kakao Cop stock data (August 8.31 ~ December .12.1)**

네이버 카카오 종목토론 게시판 (8월 31일부터 12월 1일 게시판 제목)