프로젝트 보고서

목차

1. 프로젝트 주제
2. 프로젝트 진행 상황
3. 이상치 탐지
4. 뉴스 데이터

1. 프로젝트 주제

- 코스피 200에 속하는 종목들의 일일 거래데이터 (종가, 거래량 , 거래대금, 상장주식수 등)를 활용하여 이상치 탐지모델을 구현합니다.

- 각 종목별 이상치를 지닌 날짜들을 이상치 모델을 이용하여 추출을 합니다.

- 추출된 날짜들을 활용해 과거에 기업의 주식 거래에 영향을 끼쳤던 사건들을 뉴스 데이터를 활용해 보여줍니다.

예시)

날짜	시가	고가	저가	종가	거래량	거래대금	등락률	시가총액	상장주식수	diff	pct_change
날짜											
2021-02-04	2021-02-04	19350	24200	18850	22600	18348535	406894578750	17.1	2788278367400	123375149	0.236726	0.170984

(포스코 인터내셔널의 2001년 이후 최대 일일수익률을 기록한 날짜의 거래 데이터)

포스코인터내셔널, 전기차 부품 사업 기대감에 17% 급등..기관 130억 샀다

특징주

포스코인터내셔널이 전기차 부품 사업 기대감에 급등했다. 4일 **포스코인터내셔널**은 전일 대비 3300원(17.10%) 오른 2만2600원에 마감했다. 장중 주가는 2만4200원까지 올라 52주 최고가를 경신했다. 이날 기관은 홀로 **포스코인터내셔널** 130억원을 사들이며 주…

머니투데이 | 2021.02.04 | 다음뉴스

(해당 날짜 포스코인터내셔널의 뉴스)

2. 프로젝트 진행상황

뉴스 데이터

- Scrapy를 이용하여 Daum에서 언론사별 뉴스 크롤링 구현 완료

- 회사 사명 변경 데이터 추가

주식 데이터

- Yahoo Finance를 활용해 2000년 01월 01일 이후 코스피 종목별 종가, 저가, 고가, 시가, 거래량을 가진 데이터 베이스 구축

- 4/1 : PYKRX를 활용하여 Yahoo Finance에서 제공하지 않는 종목별 일별 거래대금 데이터, 코스피 일일 거래대금 데이터 추가

- 4/8 : PYKRX를 활용하여 각 종목별 거래량 회전율, 거래대금 회전율을 계산하기위해 일일 상장주식수, 일일 시가총액 데이터 추가

이상치 탐지

- 이상치탐지 모형을 위한 변수 선택
  - ■ 초기에 거래대금, 거래량, 일일 수익률을 이용해 이상치 탐지를 하려고 하였습니다.

  - ■ 하지만 전반적인 주식시장 상황 (코로나로 인한 주식시장 붕괴, 서브프라임 모기지 사건 등)에 개별주식들이 영향을 크게 받습니다.
    그래서 코스피 시장과 개별 종목사이 상대적으로 비교할수 있는 지표를 생각 해보는 것이 중요합니다.

  - ■ 현재 까지 선택 변수 : 거래량 회전율, 일별 수익률

3. 이상치 탐지

- 거래대금 회전율과 거래량 회전율

코스피 200 종목들의 개별종목 거래대금 회전율과 거래량 회전율사이 상관계수 평균 : 0.999

개별 종목 거래대금 회전율과 거래량 회전율의 상관계수가 높은 값이 나오기 때문에 두 변수 모두 사용할 이유가 없다고 생각되어 거래량 회전율을 이상치 탐지 모델의 변수로 사용하기로 하였습니다.

- Behavioral finance : the psycology of investing, T Hens, A Meier

"This article claims that if the media spotlights a particular stock, it is more likely to attract investor attention."

투자자의 관심도를 판단하는 지표가 저희 프로젝트를 위한 이상치에서 중요한 요인입니다. 그래서 저희는 투자자의 관심도를 판단할수 있는 지표를 거래량회전율 및 일일 수익률로 정하였습니다.

- 논문 "Trading Volume : Definitions, Data Analysis, and Implications of portfolio theory , Andrew W. Lo and Jiang Wang" 속에서 거래량 회전율에 대한 정의

- 개별주식 회전율 계산 공식

- Number of trades per period
- Share volume, X_{jt}
- Share turnover,

$$\tau_{jt} \equiv \frac{X_{jt}}{N_{jt}}$$

- 포트폴리오의 회전율 계산방식

Share-Weighted Turnover Ratio

$$\tau_t^p \equiv \sum_{j=1}^{J} \omega_{jt}^p \tau_{jt} .$$

상장 주식수의 크기를 기반으로 가중치 계산

Value-Weighted Turnover Ratio

$$\tau_t^{VW} \equiv \sum_{j=1}^{J} \omega_{jt}^{VW} \tau_{jt}$$

시가총액의 크기를 기반으로 가중치 계산

Equal-Weighted Turnover Ratio

$$\tau_t^{EW} \equiv \frac{1}{J} \sum_{j=1}^{J} \tau_{jt}$$

모든 구성종목의 회전율에 동일한 가중치 부여

- “A Trading Volume Benchmark: Theory and Evidence, Paula A. Tkac” 논문

해당 논문에서는 시장 포트폴리오의 Turnover Ratio를 Value-Weighted Turnover Ratio로 계산하였습니다.

$$\mathrm{TO}_t^i = \alpha^i + \beta^i \mathrm{TO}_t^m + \epsilon_t^i.$$

TO_t^M : *시장 포트폴리오의 t시점 거래량 회전율*
TO_t^i : *i 종목의 t시점 거래량 회전율*

귀무가설 : a = 0 , b = 1

위의 귀무가설을 기각을 하게된다면 “Information Trading이 존재한다고 볼 수 있다” 라고 설명합니다.

$$\epsilon_t^i = \mathrm{TO}_t^i - \alpha^i - \beta^i \mathrm{TO}_t^m,$$

그래서 Abnormal Trading Turnover을 다음과 같이 정의하였고 Abnormal Trading Turnover는 event-specific trading을 고립시킬 수 있는 지표라고 설명합니다.

추가로 읽을 논문

- “How investor attention affects stock returns? Some international evidence, Sergen Akarsu and Omur (2021)”

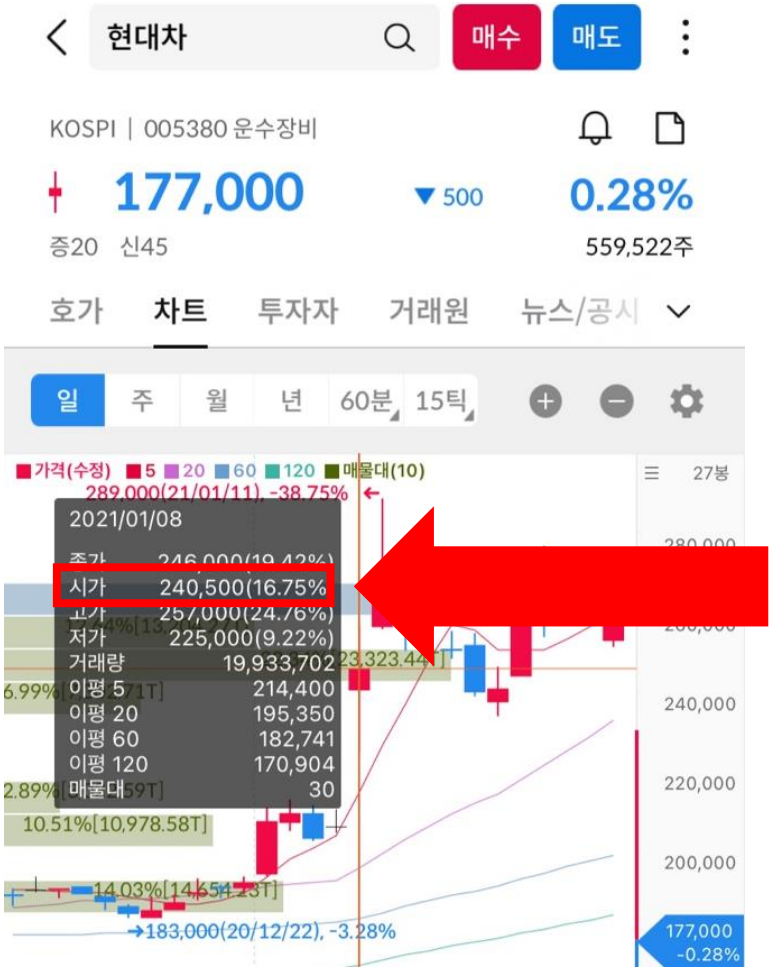
- “Does Investor Attention Affect Trading Volume In The Brazilian Stock Market?, Heloisa Elias de Souza (2018)”

- “Financial Attention , Nachum Sicherman (2016)”

4. 뉴스 데이터

거래 데이터에 있어서 이상치라고 판단된 날의 뉴스들을 뽑아서 요약하여 일목요연하게 보여줌. (코스피 200에 해당하는 종목만)

고려할 점 : “**시가**”에 따라 전 날 or 당일 장이 열리기 전(9:00 이전)에 이슈가 있는지도 확인해 봐야한다.



해결책 : 1일 기준을 15:30분부터 다음날 15:30으로 정하였다.
Ex) 1월 7일 오후 6시 30분에 나온 뉴스는 1월 8일로 간주한다.

■ **첫 번째 task**

- 뉴스를 전부 크롤링한 다음에 DB 구축 후에 날짜에 맞게 뽑는다. ← 이 방향성으로, 왜냐하면 이상치는 모델을 수정함에 따라 정해지는 날짜가 바뀔 수도 있으니, 전체 종목 DB를 구축한 다음, 상황에 맞게 바로 뽑아서 쓰는 것이 좋다.

또한, 위에서 말했듯이 어떤 날엔 당일의 뉴스만, 어떤 날엔 그 전날 및 다른 날의 뉴스도 뽑아야 하는 경우도 있으니, 전체 DB를 구축하는 것이 좋다.

1. 다음 뉴스에서 검색하고 시간별로 나열한 다음, 쭉 가져오는 방법.

2. 대안) 빅카인즈에서 키워드 검색 후, 키워드에 종목 이름이 포함되는 뉴스 가져오는 방법. -> 존재하지 않는 링크의 가능성이 있음

- 뉴스 DB 구축

종목별 뉴스를 정하는 기준

➔ 제목에 종목을 포함하는 것이 아니라, 내용에 종목의 이름이 들어가는 것으로 일단 정해보았다.

Ex) 다른 경쟁사의 영향으로 인해서, 다른 대안 종목의 가격이 내려가면, 그 뉴스도 해당 기업의 사건이라 볼 수 있다.

주의할 점 : 종목의 이름은 중간에 바뀐 경우가 있어서(사명 변경), **이름이 바뀌기 전**의 뉴스도 포함해야 한다.

- 어느 시기부터 크롤링을 해야 할까?

➔ 회사의 상장일 기준으로부터 시작.

만약 상장일이 2000년 1월 1일 전이라면, 2000년부터 시작해도 될 것이라고 판단했습니다.
Why? : 한국의 언론사 인터넷 서비스 첫 시작은 1995년 3월 중앙일보이며,

인터넷 뉴스의 활발한 배포는 2000년 이후에 더욱 활발하게 되었다고 생각하여,
우선적으로 하는 건 내부 회의를 통해 2000년 이후 뉴스로 정했습니다.

Test로 다음 뉴스에 있는 1년 간의 모든 뉴스를 크롤링해보았으며, 개수는 약 470만 개였으며, 시간은 4일 정도 소요되었습니다.

■ 두 번째 task.

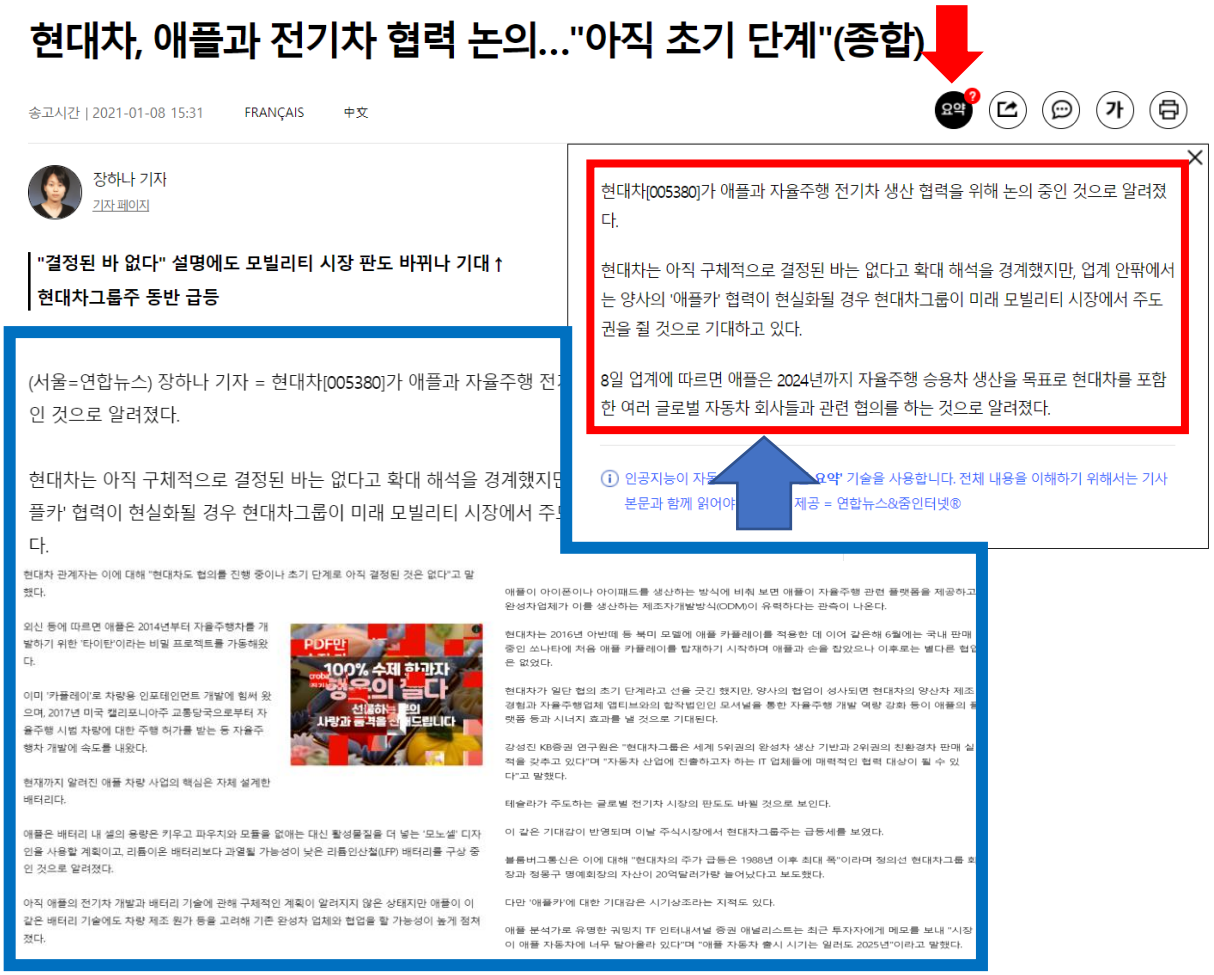
> 애플이 자율주행 전기차 개발을 위해 현대자동차그룹에 협력을 제안하고 논의중인 것으로 알려졌다.
>
> 현대차는 8일 "다수의 기업으로부터 자율주행 전기차 관련 공동 개발 협력을 요청받고 있으나, 초기단계로 결정된 바 없다"고 공시했다. 업계에 따르면 애플은 현대차 뿐 여러 글로벌 자동차 회사들과 관련 협의를 진행하고 있다.
>
> 이날 국내 한 매체는 애플과 현대차가 애플카 출시를 위해 협상을 진행중에 있으며 검토가 마무리 된 상태라고 보도했다.
>
> 애플과 현대차의 협력 소식이 전해지면서 8일 오전 현대차와 현대모비스 등 현대차 관련 주가가 급격히 뛰었다.

해당 뉴스를 어떻게 요약할 것인가?

- **기존 서비스 확인**

연합 뉴스

➔ 세줄 요약문, 구글 딥러닝 기술 “**bert**” 응용



- **Bert**란?

구글에서 개발한 NLP(자연어처리) 사전 훈련 기술이며, 특정 분야에 국한된 기술이 아니라 **모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model**입니다.

추후에, 이 모델을 이용하여 뉴스 요약을 할 예정입니다.

한국어 버전의 BERT인 KoBERT 모델이 있는데, 이 모델을 활용하면 더 효과적으로 task를 완료할 수 있을 것이라고 생각합니다.