2023.09

유사도 기반 양방향 주식 종목 추천

- 국내 주식 기반 미국 주식 추천 서비스
- 미국 주식 기반 국내 주식 추천 서비스

CHAEGPT

윤지영 채소연 채희지

CONTENTS

01 아이디어 개요

1. 아이디어 배경 3
2. 아이디어 소개 4

02 데이터 소개

1. 데이터 목록 5
2. 데이터 설명 6

03 적용 모델/코드 상세 설명

1. Flow Chart 15
2. 모델 소개 16
3. 소스코드 22
4. 알고리즘 흐름도 24

04 서비스 설명

1. 서비스 사용 예시 25

01 아이디어 개요

아이디어 배경

아이디어 소개

국내 주식 시장의 어려움

대다수의 주식 투자자들은 국내 주식과 미국 주식 중 본인의 투자 목표 및 성향에 따라 하나의 주식 시장만을 이용합니다. 신한금융투자에 따르면, 2030 주식 투자자 중 단 20%만이 해외 주식을 매매한 경험이 있다고 합니다(<<서울경제>>, <20대 10명 중 2명, 해외 주식 투자한다>). 이렇듯 국내 주식 투자자에 비해 해외 주식 투자자의 비율은 현저히 낮습니다. 그렇지만 국내 주식 투자자들에게도 미국 주식 시장을 눈여겨볼 필요성이 존재합니다. 이는 국내 시장이 미국 시장의 변화에 민감하게 반응하기 때문입니다. 국내 시장의 이러한 특성은 주식에 입문하는 투자자들에게 어려움으로 다가옵니다. 주식 용어를 익히고, 매일 변화하는 내수시장을 파악하는 것도 벅찬데 미국 시장까지 고려해야 하기 때문입니다. 저희는 이 점에 주목하여 서비스를 기획하였습니다.

국내와 미국 주식 시장의 차이

국내 주식은 하루동안 주가가 움직일 수 있는 폭이 정해져 있습니다. 가격 변동의 상한과 하한이 있어 한 종목의 주가가 전날 종가를 기준으로 30% 넘게 상승 또는 하락할 수 없습니다. 한편 미국 증시시장에는 가격 변동 폭의 제한이 없어, 하루에도 500%, 1000%를 상회하는 상승 종목이 나올 수도 있고, 정반대의 경우도 자주 발생합니다. 따라서 고위험 고수익을 추구하는 투자자들은 미국 시장의 거래에, 그렇지 않은 사람들은 비교적 안정적이고 쉽게 파악이 가능한 국내 시장의 거래에만 참여하는 경향이 있습니다. 국내 주식 시장은 비교적 작은 규모이며 주로 한국 기업들의 주식으로 구성되어 있습니다. 반면, 미국 주식 시장은 세계에서 가장 큰 규모를 자랑하고 있고 상장되어 있는 기업들의 국적도 다양합니다.

해결 방안

미래에셋증권을 포함한 여러 증권사에서 해외 주식 시장으로의 접근을 돕기 위한 다양한 방안이 마련되고 있습니다. 예컨대 올해 4월 미래에셋증권과 네이버 클라우드에서 AI 요약 서비스가 개발되었습니다. 그럼에도 불구하고 미국 주식 시장으로의 접근은 아직 어렵습니다. 위에서도 언급했듯이, 아직 해외 주식 시장에 투자하는 사람은 그리 많지 않습니다.

해외 주식 시장으로의 유입을 늘리려면, 언어의 장벽을 허물고 정보에 대한 접근성을 높여야 합니다. 이를 위해선 각 시장이 갖는 특성은 물론이고, 세부 종목에 대한 구체적인 정보를 제공해야 합니다. 또한 투자자들로 하여금 다양한 종목들에 노출되게 하여 폭 넓은 투자가 가능하도록 해야 합니다. 더불어 투자자들이 각각의 종목을 비교할 수 있도록 한다면, 더 나은 투자의 경험을 제공할 수 있을 것입니다. 해외 주식에 접근할 기회가 늘어나고 있는 현 시점에서, 이러한 서비스를 제공한다면 더욱 많은 투자자들이 유입될 것입니다.

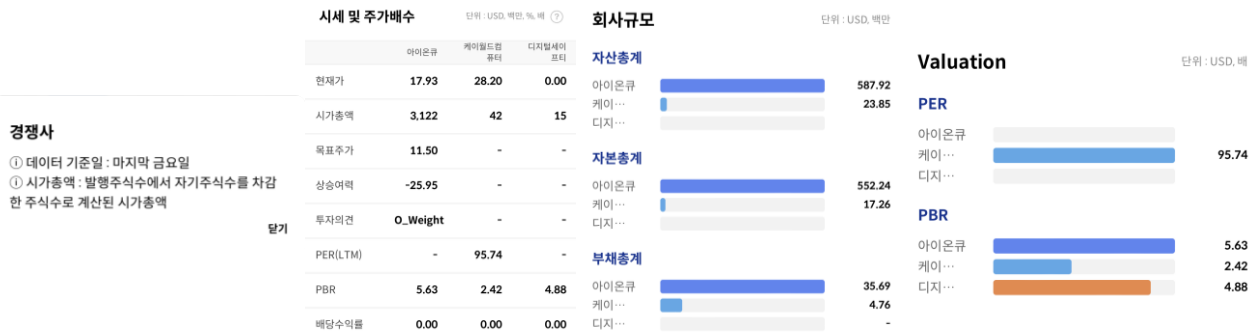
01 아이디어 개요

아이디어 배경 | 아이디어 소개

저희 서비스는 '미국 주식 기반 국내 주식 추천 서비스'와 '국내 주식 기반 미국 주식 추천 서비스'로 구분하여 제공됩니다. 국내와 미국 주식 시장은 위에서 언급한 투자리스크, 시장 규모, 다양성과 같은 여러 측면에서 차이가 있습니다. 두 시장의 차이를 착안하여 각 시장으로의 투자가 가지는 장점은 살리고 단점은 상호 보완하여, 양방향 유사도 기반 주식 종목 추천 서비스를 구축하고자 합니다.

미국 주식 기반 국내 주식 추천 서비스

미국 주식 기반 국내 주식 추천 서비스의 경우, M-STOCK 앱의 경쟁사 분석에서 유사한 형태로 제공되고 있습니다. 그러나 분석 결과가 어떤 기준에 따라 제공되는지 사용자가 파악하기 어렵다는 문제점이 있었습니다. 아래와 같이 경쟁사 기준을 확인했을 때 시가총액에 대한 내용만을 확인할 수 있었습니다. 또한 뉴스, 투자 심리 등과 같은 정보가 아예 반영되지 않아 주식에 영향을 미치는 정보들이 종합적으로 고려되고 있다고 보기 어려웠습니다.



▲ M-STOCK 앱 > 메뉴 > 주식 > 투자 정보 > 미국리서치 > 종목분석 > 경쟁사분석

국내 주식 기반 미국 주식 추천 서비스

국내 주식 기반 미국 주식 추천 서비스는 미국 주식 시장을 잘 알지 못하는 투자자들을 대상으로 하는 서비스입니다. 국내 시장의 경우, 상장된 기업에 대한 정보를 다양한 경로를 통해 쉽게 얻을 수 있습니다. 반면 미국 주식 시장의 경우, 이를 입수하기가 상대적으로 어렵고, 입수한 정보들은 주로 영어로 되어 있어 이해하기 어렵습니다. 이로 인해 해외 주식에 입문하는 투자자들에게는 국내 주식과는 다른 분명한 진입장벽이 존재합니다.

추천 서비스를 위해 수집한 국내와 미국의 기업들에 대한 정보와 유사도를 활용하여, 미국 주식 기반 국내 주식 추천 서비스를 종합적으로 개선하고자 합니다. 또한 미국 시장에 대한 진입장벽을 낮추기 위해, 비교적 친숙한 국내 기업들에 대한 정보를 기반으로 유사한 미국 기업에 대한 정보를 제공해주는 국내 주식 기반 미국 주식 추천 서비스를 구축하고자 합니다.

02 데이터 소개

데이터 목록

분류	데이터 이름	출처
산업 분류	미국 산업 분류	Finviz 크롤링
	국내 산업 분류	인포스탁 크롤링 & FinanceDataReader라이브러리(Python)
재무 정보	미국 상장 기업 재무 정보	Yahooquery 라이브러리(Python)
	국내 상장 기업 재무 정보	CompanyGuide 크롤링
뉴스 데이터	미국 상장 기업 뉴스 데이터	로이터(Reuters) 크롤링
	국내 상장 기업 뉴스 데이터	다음(Daum) 뉴스 크롤링
주식 토론방	미국 주식 토론방	Stocktwits 크롤링
	국내 주식 토론방	네이버 금융 - 종목 토론방 크롤링
주가 정보	미국 상장 기업 일별 주가 정보	Yfinance 라이브러리(Python)
	국내 상장 기업 일별 주가 정보	Yfinance, Pykrx 라이브러리(Python)

02 데이터 소개

데이터 설명

산업 분류 미국과 국내에 공통된 산업 분류 기준을 새로 구성하여 유사도 계산 지표 중 하나로 활용.

크롤링

미국

미국의 경우, 'Finviz' 사이트를 크롤링하여 데이터셋을 구축하였습니다. Finviz는 미국 주식 시장의 섹터별 전체적인 흐름을 사용자들이 한눈에 제공할 수 있도록 제공해준다는 장점을 지닌 사이트입니다. 따라서 미국의 주식 시장인 'AMEX', 'NASDAQ', 'NYSE'에 속한 기업들에 대한 정보를 크롤링하여 하나의 파일로 구성하였습니다.

국내

국내의 경우, 'MSTOCK'의 [테마/업종/시세정보]의 섹터별 분류 정보가 인포스탁(infostock)에서 분류한 정보를 활용하기 때문에 동일하게 진행하기 위해 인포스탁 사이트를 크롤링하여 하나의 파일로 구성하였습니다.

[미국 - 크롤링 데이터셋 - 7,294 X 10]

Ticker	Company	Sector	···	Volume
AA	Alcoa Corporation	Basic Materials	···	3,649,055
⋮				
YORW	The York Water Company	Utilities - Regulated Water	···	38,398

[국내 - 크롤링 데이터셋 - 5,685 X 2]

소분류	기업
2차전지	LG에너지솔루션
⋮	
U-Healthcare(원격진료)	라이프시맨틱스

전처리

미국

먼저, 불필요한 열들을 제거해주었습니다. 이후 수집된 종목 중 ETF 종목들의 경우, 국내 데이터에는 ETF 종목들이 수집되지 않았다는 점, ETF는 특정 지수를 추종하는 구성종목들로 구성된 상장지수 펀드로 그 성격이 일반적인 종목들과는 다르다는 점에 기인하여 삭제하였습니다.

이후 Market 정보를 추가하기 위해 'FinanceDataReader' 라이브러리를 활용하여 'NASDAQ', 'AMEX', 'NYSE' 시장에 상장된 종목에 대한 정보를 불러왔습니다. 크롤링하여 구축한 데이터셋과 불러온 시장 종목 정보를 비교해보니 크롤링 데이터셋에만 있는 종목들이 약 120개 가량 존재하였습니다. 따라서 직접 확인해보니 크롤링한 데이터셋에서의 종목명과 불러온 시장 종목에 나타난 종목명의 형태가 달랐던 것도 있었고, OTC Market에 상장된 종목이 크롤링된 경우도 있어 수작업으로 수정하였습니다.

02 데이터 소개

데이터 설명

산업 분류 미국과 국내에 공통된 산업 분류 기준을 새로 구성하여 유사도 계산 지표 중 하나로 활용.

전처리

국내

인포스탁의 분류 기준에는 국내의 주요 산업에 특화된 기준(ex. DMZ 평화공원, 남북경협, 정치/외교_윤석열 등)이 존재하였으며, 국내와 미국의 연경점을 찾는 저희의 서비스에는 이러한 기준이 불필요하다고 생각하여 제거하였습니다. 제거된 분류 기준에만 속해 있던 기업들은 직접 기업의 사업분야 정보를 찾아 적절히 산업 분류를 진행하였습니다.

또한 SPAC 종목들이 다수 있었는데, SPAC 종목은 기업합병을 진행하는 과정에서 생기는 종목으로 기업합병 이후에는 해당 종목이 거래가 이루어지지 않는다는 점에서 그 성격이 일반적인 종목들과는 다르다는 점에 기인하여 삭제하였습니다.

이후 다른 데이터들과의 작업을 위해, FinanceDataReader 라이브러리의 KRX 데이터셋을 활용하여 종목코드를 매칭하는 과정에서 누락된 종목이 약 800개가 있다는 사실을 발견하였습니다. 누락된 종목들에는 우선주, KONEX에 상장된 종목 등이 있었습니다. 우선주의 경우 유사도를 계산할 때 고려되는 특징들이 일반 종목과 동일하기 때문에 삭제하였고, 나머지 종목들에 대해서는 추가적으로 앞서 진행했던 산업 분류 과정을 동일하게 적용하여 데이터셋에 추가하였습니다.

최종 데이터셋

미국_산업분류(최종).csv [4882 X 4]

Ticker	Company	Industrials	Market
AA	Alcoa Corporation	Aluminum	NYSE
⋮			

> 산업분류 파일을 기준으로 주가 데이터를 수집하는 과정에서 'BREZR'이라는 티커의 주가 데이터가 하루치만 수집 되어 분석에 사용할 수 없기에 해당 행을 제거해주었습니다

> 산업분류 파일을 기준으로 Stocktwits에서 토로방 데이터를 수집하는 과정에서 수집되지 않는 종목 한 개 또한 분석에 사용할 수 없어 해당 행을 제거해주었습니다.

그 결과 4882 X 4 형태로 최종적으로 구성되었습니다.

국내_산업분류(최종).csv [3976 X 4]

종목코드	기업	소분류	시장
'360070'	탑머티리얼	Specialty Chemicals	KOSDAQ
⋮			

> 산업분류 파일을 기준으로 주가 데이터를 수집하는 과정에서 거래 정지된 종목들이 다수 발견되어서 해당 종목의 행은 삭제하였습니다.

그 결과 3976 X 4 형태로 최종적으로 구성되었습니다.

02 데이터 소개

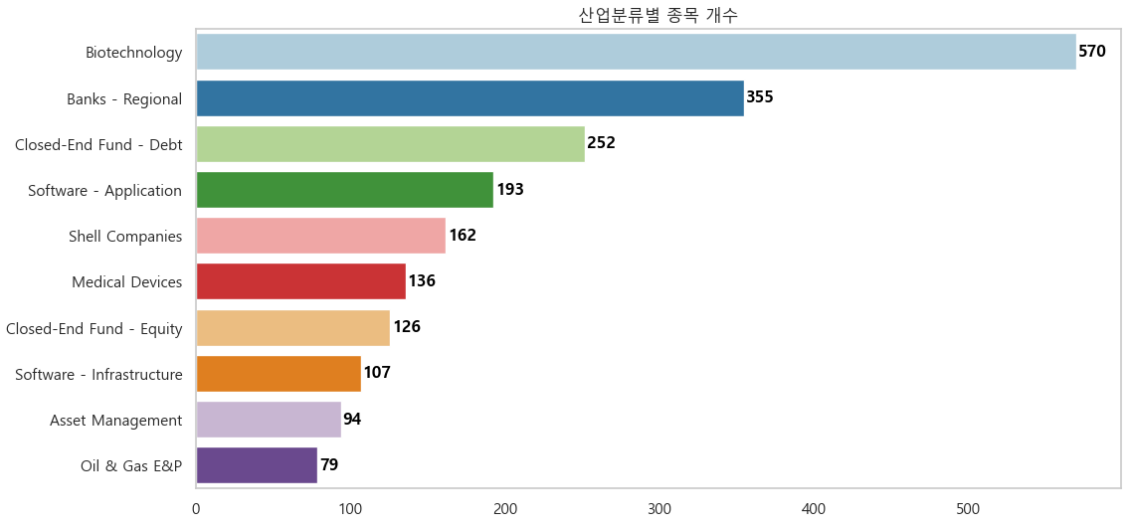
데이터 설명

산업 분류 미국과 국내에 공통된 산업 분류 기준을 새로 구성하여 유사도 계산 지표 중 하나로 활용.

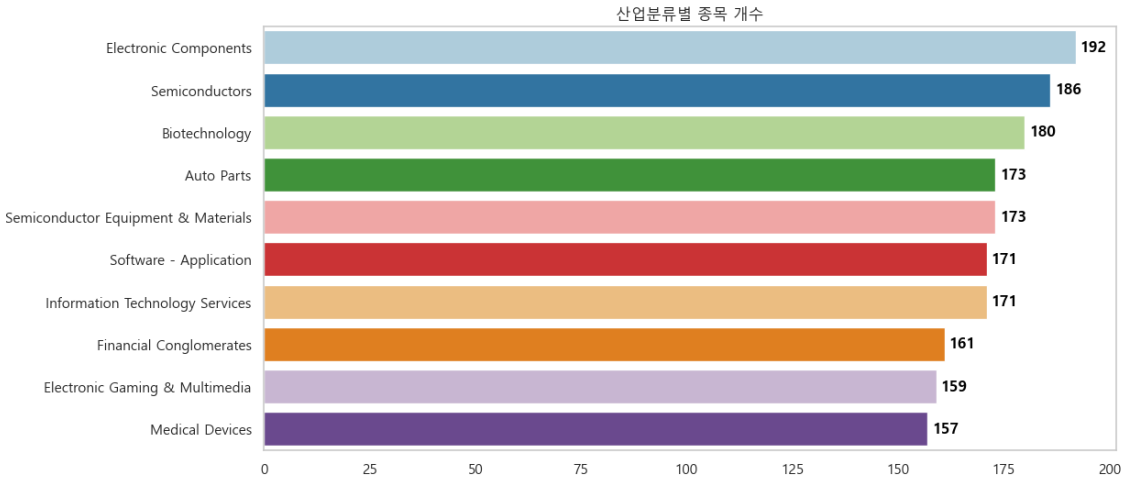
EDA

최종 산업분류의 개수는 총 126개로 각 산업분류별 포함된 종목의 개수를 볼 수 있습니다. 이 페이지에서 126개의 산업분류를 모두 나타내는 데 한계가 있어 상위 10개만 가져왔습니다. 자세한 정보는 '소스코드/대시보드 과정 노트북/시각화.ipnyb' 파일에서 확인하실 수 있습니다.

미국



국내



02 데이터 소개

데이터 설명

주가 데이터 각 기업별로 2년치의 데이터를 수집하여 차트 분석 및 정보 제공으로 활용

크롤링

미국

미국의 경우, Python의 'yfinance' 라이브러리를 활용하여 약 2년치의 주가 데이터를 수집하였습니다. Yfinance 라이브러리는 증권 데이터 수집용 라이브러리 중 하나로 야후 파이낸스(Yahoo Finance)에서 크롤링한 데이터를 제공합니다. Yfinance 라이브러리가 가장 다양한 정보를 보유하고 있기 때문에 사용하였습니다.

코드를 구성할 때 datetime 라이브러리를 활용하여 특정일의 전일부터 2년 전까지의 주가 데이터를 수집할 수 있도록 설정해주었습니다.

또한 yfinance에서 제공되는 주가 데이터에는 등락률이 포함되어 있지 않아 등락률을 계산하는 코드를 작성하여 'Change'라는 열을 추가해주었습니다.

네이버 클라우드의 Object Storage에 각 기업별 2년치의 주가데이터를 포함한 파일이 저장되도록 하였습니다.

국내

국내의 경우, 'KOSPI', 'KOSDAQ', 'KONEX'에 상장된 기업들을 수집하였는데 yfinance는 'KONEX'에 상장된 종목에 대한 정보는 보유하고 있지 않아 추가로 pykrx 라이브러리도 사용하였습니다.

Yfinance에서 국내 종목의 주가데이터를 가져오기 위해서는 종목코드 뒤에 해당 종목이 속한 시장의 약자를 붙여줘야 합니다(ex. 005930.KS). 따라서 코스닥에 상장된 종목은 '.KQ'가, 코스피에 상장된 종목은 '.KS'가 추가되도록 코드를 작성하였습니다. Pykrx는 종목코드만 요구되기 때문에 KONEX에 상장된 종목은 위 과정은 따로 진행하지 않았습니다.

미국 데이터 처리와 동일하게 yifnance를 사용하는 경우, 등락률 계산 후 'Change'라는 열을 추가해 주었고 pykrx의 경우 등락률에 대한 정보가 포함되지만 열의 이름이 '등락률'로 되어 있었기 때문에 'Change'로 열이름을 변경해주었습니다.

국내 기업 역시 각 기업별로 2년치의 주가데이터를 포함한 파일이 클라우드에 저장되도록 진행하였습니다.

최종 데이터셋

NAVER_주가데이터.csv [09.01 기준]

Date	Open	High	Low	Close	Volume	Change	Ticker
2021-08-31	430000	439500	427000	439000	768823	2.570096	035420
⋮							
2023-08-31	217500	219000	213000	214500	808546	-1.37931	035420

02 데이터 소개

데이터 설명

재무정보 국내, 미국 재무정보를 기간별로 각각 수집하여 차트 분석 및 정보 제공으로 활용

크롤링

미국

미국의 경우, Python의 'Yahooquery' 라이브러리를 활용하여 2022년 6월부터 2023년 3월까지(2023.08 기준) 4분기 재무정보 데이터를 수집하였습니다. 'Yahooquery' 라이브러리는 Yahoo Finance의 재무 데이터에 접근할 수 있는 라이브러리입니다. Yfinance를 포함한 이외의 라이브러리에서는 최근 데이터를 제공하지 않고 있지만, Yahooquery는 최근까지의 다양한 재무정보 항목들을 보유하고 있어 해당 라이브러리를 사용하였습니다.

재무정보는 Income Statement(손익계산서), Balance Sheet(재무상태표), Cash Flow(현금흐름표), Valuation Measures(투자지표)까지 4가지 종류로 나뉩니다. For문으로 미국 주식 종목들에 대해 all_financial_data 옵션을 사용하였고, 총 342개 재무정보 변수에 대한 데이터를 수집하였습니다. 전체 변수들 중 국내의 변수와 겹치는 변수 15개를 데이터프레임에 반영하였습니다. 국내와 미국 주식 유사도 계산 시 반영할 재무정보 변수는 총 17개입니다. 반영된 15개 변수 이외에 2개의 변수는 파생변수로 따로 계산해주었기 때문에, 다음 페이지의 파생변수 및 전처리 파트에서 설명하겠습니다.

all_financial_data 옵션을 사용하면 Valuation Measures(투자지표)에 해당하는 항목들은 수집되지 않아 추가적으로 valuation_measures() 옵션을 사용해 필요한 항목들을 추가로 데이터프레임에 저장하였습니다. 투자지표를 제외한 항목들은 2022년 6월부터 2023년 3월까지 분기별 자료가 전부 수집이 되었으나, 투자지표의 경우 국내 종목들과 달리 2022년 12월 데이터만 국내와 겹치는 것으로 확인되어 해당 시기만 반영하였습니다.

국내

국내의 경우, Python의 'Requests' 라이브러리를 통해 Fnguide에서 제공하는 Company Guide 사이트의 재무제표 카테고리에 있는 포괄손익계산서, 재무상태표, 현금흐름표에 대한 정보들을 가져왔습니다. KRX에 등재되어 있는 국내 주식 전종목에 대해 HTML 내용을 문자열 형태로 추출하여 데이터를 가져왔습니다. 기간은 CompGuide에서 22년 6월부터 23년 3월까지 분기별로 제공하고 있어 해당 기간의 재무정보 데이터를 가져왔습니다. 세 개의 재무제표 종류에 따라 각각 하나의 dataframe으로 저장하였습니다.

M-STOCK과 FnGuide에서는 국내 종목들의 투자지표 항목에 대해서 18년 12월부터 22년 12월까지 최근 5년간의 12월달 및 2023년 3월의 투자지표 데이터를 제공하고 있었습니다(2023.08 기준). 따라서 유사도 계산을 위해서 기간을 통일하고자 미국 재무정보 데이터를 연도별로 설정하고 확인해보니, 투자지표에 반영할 항목들 중 겹치는 시기는 2022년 12월밖에 없어 해당 시기의 데이터만 반영하였습니다. Company Guide의 투자지표 카테고리 중 제공하는 정보 중 기업가치 지표에 대한 내용을 가져왔으며, 데이터를 가져오는 방식은 재무제표 카테고리에서 불러오는 방식과 동일합니다.

02 데이터 소개

데이터 설명

재무정보 국내, 미국 재무정보를 기간별로 각각 수집하여 차트 분석 및 정보 제공으로 활용

[미국 – 투자지표 제외 원본 – 19,080 X 342]

Ticker	asOfDate	매출액	···	투자활동	재무활동
AA	2022/06	3.644000e+09	···	-9.300000e+07	-349000000.0
⋮					
YORW	2023/03	1.540100e+07	···	-1.058100e+07	4522000.0

[국내 – 손익계산서 원본 – 10,220 X 45]

종목코드	date	매출액	매출원가	···	한글종목약명	시장구분
A098120	2022/06	154.0	129.0	···	마이크로컨텍솔	KOSDAQ
⋮						
A238490	2023/03	24.0	18.0	···	힘스	KOSDAQ

파생변수 생성 및 전처리

미국

all_financial_data를 통해 손익계산서, 재무상태표, 현금흐름표에 해당하는 변수는 총 342개였습니다. 손익계산서에 해당하는 변수들 중에서는 매출액, 매출원가, 매출총이익, 판매비와관리비, 영업이익, 당기순이익까지 6개의 항목을 담았습니다. 재무상태표에서는 자산과 부채 두 항목을 담고, 자본과 일치하는 변수는 찾을 수 없었기 때문에 "자산 – 부채 = 자본 " 식을 통해 파생변수로 반영하여 세 변수를 고려하였습니다. 현금흐름표에서는 영업활동, 투자활동, 재무활동 변수를 반영하였습니다. 변수명이 영어로 설정되어 있어 우리말로 바꿔주고, 국내 재무제표와 날짜를 통일하고 TTM(Trailing Twelve Months: 지난 12개월 동안의 데이터)에 해당하는 데이터는 국내 재무제표에서는 확인할 수 없어 반영하지 않았습니다.

투자지표의 경우 PER, PSR, PBR, EV/EBITDA 항목을 반영하고 변수명과 날짜를 국내와 통일하였습니다. 미국 투자지표 변수들은 앞에서 언급했듯이 2022/12 데이터만 존재하기 때문에 추가로 국내 투자지표의 기간과 동일하게 2019/12, 2020/12, 2021/12에 해당하는 열들을 생성하였습니다. 또한 Gower Distance를 위한 dataframe에 활용하기 위해, 날짜(asOfDate)에 해당하는 열을 인덱스 및 컬럼 조정을 통해 "날짜+변수명 " 의 형태로 만들어주었습니다.

또한 추가로 ROE라는 파생변수를 생성하였습니다. ROE(Return on Equity)란 우리말로 자기자본이익률으로, "순이익 / 자기자본 * 100"의 식으로 계산되는 기업의 중요한 수익성 지표 중 하나입니다. 각 날짜별로 당기순이익과 자본 변수를 이용해 ROE 파생변수를 생성해주었습니다. 또한 Yahooquery 라이브러리의 경우 Yahoo Finance에서 데이터를 가져오기 때문에 사이트에 방문하여 확인한 결과, 비율 데이터를 제외하고는 All numbers in thousands로 단위가 1000씩 생략되어 있어 필요한 변수들에 대해서 곱해주었습니다. 중간에 merge하는 과정들을 거쳐 최종적으로, 총 변수는 17(변수 개수) X 4(4개 시점) = 68개로 통일하였습니다.

02 데이터 소개

데이터 설명

재무정보 국내, 미국 재무정보를 기간별로 각각 수집하여 차트 분석 및 정보 제공으로 활용

국내

국내의 경우, 미국 재무제표와 동일하게 손익계산서에서 6개, 재무상태표에서 3개, 현금흐름표에서 3개 변수를 반영하였습니다. 미국과 달리 국내의 경우 자본 변수까지 고려가 되어 있어 따로 계산을 진행하지 않았습니다. 또한 url 페이지를 통해 가져오는 과정에서 재무제표 종류가 달라 개별적으로 데이터를 불러 활용했습니다. 미국 재무제표와 같은 방식으로 투자지표를 제외한 재무제표 종류들부터 "날짜+변수명"으로 변수명을 맞춰주었습니다. 투자지표도 동일한 방식이나, 날짜가 달라 변수명은 차이가 있습니다. 재무 정보 네 종류 모두 한글종목약명과 시장구분이라는 변수는 날짜 구분이 필요 없어 해당 이름으로 열 하나씩만 남겨주었습니다. '기업' 열에서 중복되는 행들을 제거해줘야 총 기업 개수가 동일하게 맞춰줄 수 있어 해당 과정 진행 후 merge하였습니다.

미국 재무제표와 동일한 방식으로 ROE 변수를 총 4개의 날짜에 맞춰 생성해주었습니다. 추가적으로 국내와 미국의 화폐 가치를 통일해줘야 했기 때문에, 데이터를 크롤링해온 사이트를 확인해보니 비율 데이터를 제외하고는 수치에 억(원) 단위가 빠져 있었습니다. 따라서 각 데이터에 억 단위를 곱하고 환율(09/09 기준)을 곱해주는 함수를 통해 최종 데이터프레임을 완성하였습니다. 미국과 동일하게 날짜나 기업 열을 제외하고 최종 재무정보로 고려될 변수는 68개입니다.

최종 데이터셋

미국 재무제표.csv [08.31 기준]

Ticker	Company	Industrials	Market	2022/06 매출액	...	2022/12 ROE	2023/03 ROE
AA	Alcoa Corporation	Aluminum	NYSE	3644000000000	...	-5.69081	-3.669
			⋮				
CWEN.A	Clearway Energy, Inc.	Utilities - Renewable	NYSE				

국내 재무제표.csv [08.31 기준]

소분류	기업	종목코드	시장	2022/06 매출액	...	2022/12 ROE	2023/03 ROE
Specialty Chemicals	탑머티리얼	360070.KQ	KOSDAQ		...	2.173913	-1.282051
			⋮				
Biotechnology	바이오솔루션	086820.Kq	KOSDAQ		...		

02 데이터 소개

데이터 설명

주식 토론방 유사도 계산 지표 활용 및 긍/부정도와 워드클라우드를 통한 투자심리 정보 제공

크롤링

미국

미국의 종목 토론방은 대표적으로 야후 파이낸스(Yahoo Finance), 인베스팅닷컴(investing.com), Stocktwits가 있는데, 이 중 가장 활성화된 곳인 Stocktwits 크롤링을 통해 데이터를 수집하였습니다. Stocktwits의 종목 토론방 분위기가 SNS처럼 타사용자의 댓글에 답글을 달아 의견을 공유하는 것이 활성화되어 있는 것을 확인하였고, 댓글 뿐만 아니라 대댓글(답글)도 모두 크롤링하기로 결정하였습니다.

Stocktwits는 스크롤을 내리면 이전 댓글 목록이 생성되는 식으로 페이지가 구성되어 있는데, 기업에 대한 최근 반응(최대 7일)에 대한 데이터를 가져와서 이를 반영하는 것이 중요하다고 판단하였습니다. 따라서 최근 Hot Trending 기업 순위를 보며 가장 활성화되어 있는 기업 토론방들의 댓글 개수를 파악하였고, 크롤링 시간, 최근 여론 반영 등을 종합적으로 고려해보아, 1,800개의 댓글을 가져왔을 때 투자자들의 심리 파악을 위한 데이터를 수집할 수 있다고 판단하여 종목별로 크롤링을 진행하였습니다.

국내

국내의 경우, '네이버 금융' 사이트에서 제공되는 종목 토론방의 글들을 크롤링하였습니다. 네이버 종목 토론방의 형식은 게시글을 게시하는 형식으로 되어 있어 날짜, 제목, 내용, 조회수, 공감, 비공감 데이터를 수집하여 하나의 데이터셋으로 구성하였습니다. 기간은 특정일의 전일부터 7일동안의 데이터를 수집하도록 설정하였습니다. 또한 종목별 토론방 활성화 정도에 따라 수집된 데이터 양이 천차만별하다는 특징을 지니고 있었고, 따라서 7일 동안의 데이터를 수집하는 중에 페이지 수가 100페이지(약 2,000개 데이터)를 넘어가면 수집을 중단하도록 설정하였습니다. 7일이란 기간은 특정일을 기준으로 '투자자들의 심리 파악'을 하기 위해 정한 숫자로, 투자자들의 심리 파악을 위한 데이터가 충분히 갖춰졌다면 꼭 그 기간이 지켜지지 않아도 된다고 판단하였기 때문입니다.

최종 데이터셋

AA_20230831.pickle [08.31 기준]

날짜	내용	구분
2.01 AM	$AA this is getting interesting now over $30 B…	댓글
⋮		
Aug 01, 2023 1:26 PM	$AA Bullish Bullish	대댓글

NAVER_주식토론방.xlsx [08.31 기준]

날짜	제목	조회	공감	비공감	내용	종목코드	기업
2023.08.29	하 떨어질 때 더..	740	17	8	낼 폭등하겠네…	035420	NAVER
⋮							
2023.08.25	예상가	97	4	0	개 떡락이다…	035420	NAVER

02 데이터 소개

데이터 설명

뉴스 데이터 유사도 계산 지표 활용 및 긍/부정도와 중요도를 통한 기업의 여론/평판 정보 제공

크롤링

미국

미국의 종목별 최근 이슈, 중요한 정보 등을 제공받을 수 있는 언론 사이트로 로이터(Reuters)를 선정하였습니다. 크롤링을 진행할 때 시간적으로 한계가 있어 중복된 정보를 가져오기보다는 언론사에서 중요 뉴스라고 판별한 기사를 기재한다는 점, 그리고 M-STOCK에서 AI뉴스요약 서비스를 제공하는 언론사에 로이터가 있다는 점을 모두 종합적으로 고려하였습니다. 최근 6개월 종목별 최근 이슈를 가져오는 것을 목표로, 종목 토론방과 유사하게 검색량이 많은 기업들의 기사 개수를 확인해보았고, 140~160개 정도의 뉴스기사를 가져왔을 때 원하는 기간에 해당되는 기사를 가져올 수 있었습니다.

기업리스트를 크롤링할 때 가져온 공식 기업명 표기 중 'Co.', 'Inc'. 'Corp.' 등과 같이 약어를 기호로 추가한 형태의 기업들이 많았는데, 기사 보도 시에는 해당 기호를 사용하지 않고 기업명만을 사용한 것을 확인하였습니다. 기호를 포함하여 검색하였을 때 일부 기사가 제외되는 것을 막기 위해 이러한 기호를 제거하는 전처리를 진행하였습니다.

국내

국내의 경우, 다음(Daum) 뉴스에서 기사를 크롤링하도록 하였습니다. 그러나 다음 뉴스는 한 검색어에 대해 최대 80페이지(기사 800개)까지만 제공하였고, 기사 양이 많은 기업의 경우에는 '전체'가 아닌 '다음뉴스' 카테고리를 보더라도 80페이지 안에 원하는 기간인 한 달 동안의 뉴스를 얻을 수 없다는 한계가 있었습니다. 국내 주식 토론방과 같은 이유로 기업의 최근 이슈들을 통해 평판을 반영할 수 있는 데이터가 충분히 갖춰졌다면 꼭 그 기간을 지키지 않더라도 이후에 이 데이터를 활용할 때 그 정보가 반영될 것이라 생각하였고, 최대 80페이지까지의 기사를 크롤링하도록 하였습니다.

최종 데이터셋

Airbnb_20230831.pickle [08.31 기준]

언론사	검색어	기사종류	제목	날짜	기사내용	링크
REUTERS	Airbnb	World	Apartment fire that killed 7 in Mon…	August 29, 2023	Aug 28(Reuters) - A quick-moving ea…	https://...

⋮

DSC인베스트먼트_20230831.pickle [08.31 기준]

검색어	신문사명	제목	날짜	기사내용	링크
DSC인베스트먼트	전자신문	[AI TECH+ 2023]크라우드웍스, 데이터…	2023. 8. 31. 14:52	₩n₩n₩n₩n₩n 크라우드웍스 부스…	http://v.daum.net/v/2023...

⋮

02 데이터 소개

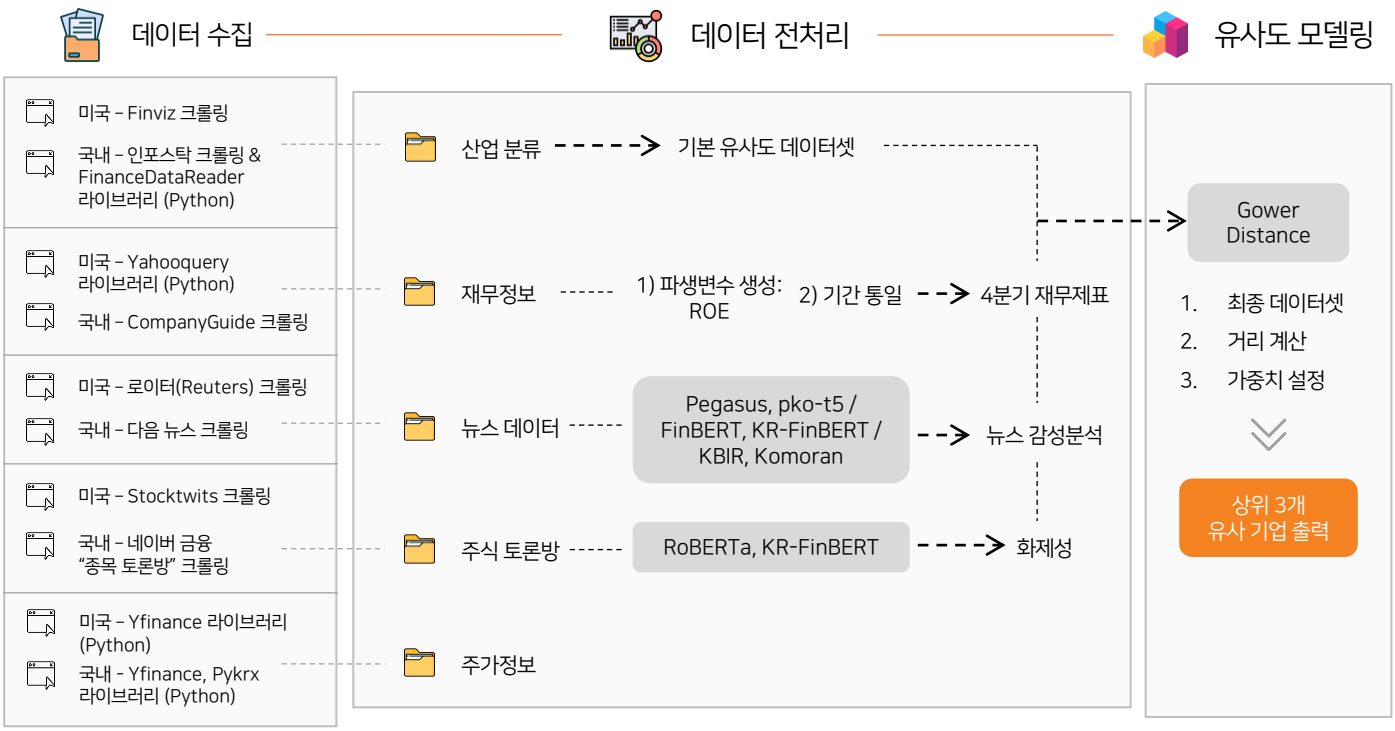
Flow Chart

데이터 수집과 전처리, 모델 적용 및 분석 후 결과 도출까지의 흐름을 Flow Chart를 통해 한 눈에 확인하실 수 있습니다.

데이터 수집의 경우, 각 데이터별로 앞에서 소개해드린 수집 방법을 통해 '네이버 클라우드'의 서버에 저장하였습니다. 산업 분류 데이터를 제외한 모든 데이터에 대하여 자동화를 진행하여 업데이트 되어 수집된 정보들이 자동으로 클라우드 서버에 저장되도록 구성하였습니다.

데이터 전처리의 경우, 클라우드에 저장된 데이터들을 불러오는 과정을 함수화하여 각 데이터를 이후 유사도 계산에 활용할 수 있도록 적절하게 처리한 후 추가적으로 기존 데이터를 활용하여 파생변수를 생성하였습니다. 해당 과정 또한 유기적으로 연결되어 자동적으로 진행되도록 모두 구성하였습니다.

마지막으로 유사도 모델링의 경우, 전처리 과정을 마친 후 최종적으로 미국과 국내에 대해 각각의 데이터셋을 구성하여 데이터의 특징을 반영하여 거리가 계산되도록 하였습니다.



다음 장에서는 위 Flow Chart에서는 보여지지 않은 모델 구성과 처리 과정에 대하여 자세히 설명드리겠습니다.

03 적용 모델 상세 설명

모델 소개 – 국내 뉴스 데이터

T5 v1.1 OOV(Out-of-Vocabulary)를 해결한 BBPE(Byte-level Byte Pair Encoding)을 사용한 모델

요약

사용 이유

뉴스 데이터를 사용하게 되는 감성분석 태스크의 경우, 뉴스 데이터 전문을 모두 사용할 시 불필요한 내용들이 섞여 정확한 감성분석/키워드 추출이 어려울 것이라 판단하였습니다. 또한 해당 기업명이 들어간 문장만 추출한다면 해당 변수가 뉴스의 전체적인 감정/내용을 내포하지 못할 수도 있다고 판단하여, 기사를 요약하여 사용하기로 하였습니다. 모델 선정 조건으로 결과의 완성도가 물론 중요하지만, 로드할 모델 수가 많은 만큼 한정된 서버의 용량 내에서 모델을 작동시켜야 했기 때문에 모델의 사이즈(파라미터 개수)와 모델이 추론을 할 때 소요되는 시간 또한 고려해야 했습니다. 이를 종합적으로 고려하였을 때, 한국어 전용 데이터로 't5 v1.1' 모델을 학습시킨 'pko-t5' 모델을 사용하기로 하였습니다. 추가적으로 AIHUB에서 제공한 '요약문 및 레포트 생성 데이터 ' 에는 약 50,000 건의 뉴스기사, 보도자료의 요약 데이터가 존재하는데 현재 진행하는 뉴스 요약 태스크에 도움이 될 것이라고 판단하여 해당 데이터로 파인튜닝(Fine-tuning)을 진행한 모델의 학습 가중치를 사용했습니다.

전처리 / 결과

전처리를 위해서 뉴스 기사 텍스트에서는 페이지 구조 형식상 문단 넘어가기가 과도하게 설정되어있는 경우가 많아 이를 찾아 제거하였고 글자 외의 특수기호들을 제거하였습니다. 모델에 기사 내용을 입력으로 넣으면 'generated_text'라는 key와 요약된 텍스트(value)가 JSON 데이터 형식으로 출력됩니다. 여기에서 텍스트 부분만 추출하여 '요약'열로 저장하여 이후에 사용하였습니다.

Example.

검색어: 금호건설

제목: 회사채 발행 속속 성공하는 건설사들…"미래 리스크 대비" [투자360]

기사내용(전처리): 연합 헤럴드경제신동윤 기자 원자재 가격 상승과 부동상
프로젝트파이낸싱PF 부실 우려 속에서 향후 발생할 수 있는 리스크 대비에 나섰던
건설사들이 속속 회사채 발행에 성공하고 있다30일 금융투자업계에 따르면 최근 …

summary [{'generated_text': '건설 경기가 바닥을 치고 반등할 것이란 기대 심리 덕분에
건설채에 투자하려는 투자자들이 늘면서 회사채 시장도 활기를 띨 것이란 기대감이
높아지는 상황이다.'}]

'건설 경기가 바닥을 치고 반등할 것이란 기대 심리 덕분에 건설채에 투자하려는 투자자들이 늘면서 회사채 시장도 활기를 띨 것이란 기대감이 높아지는 상황이다.'

요약 column

03 적용 모델 상세 설명

모델 소개 – 국내 뉴스 데이터, 국내 주식토론방 데이터

KR-FinBert BERT 모듈 중 한국 금융 분야에 특화된 모듈

감성분석

사용 이유

요약된 뉴스와 네이버 종목 토론방 댓글들을 사용하여 감성분석을 진행하였습니다. 이때, 사용한 모델은 KR-FinBERT 모델로, 이는 BERT 모델 중 한국 금융 분야에 특화된 모듈입니다. 기업명을 키워드로 한 뉴스 데이터의 특징상, 경제 카테고리로 분류된 경우가 많아 FinBERT가 적합할 것이라고 판단하였습니다. 또한 여러 뉴스 기사와 기업들의 분석 레포트 등을 이용하여 전이학습(Transfer learning)시킨 모델의 가중치를 사용하여 감성분석 성능 향상에 도움이 되리라고 판단하였습니다. 종목 토론방의 경우, [네이버 금융] 사이트에 있는 토론실 형식으로, 기재된 댓글 유형을 보았을 때 해당 기업의 실적, 최근 이슈들에 대한 내용이 많아 경제 관련 단어들을 보고 추론이 가능한 모델이 필요하다고 판단하여 금융 분야에 특화된 KR-FinBERT 모델을 동일하게 사용하기로 하였습니다.

전처리 / 결과

전처리를 위해, 뉴스 기사 텍스트에서 페이지 구조형식 상 문단 넘어가기가 과도하게 되어있는 경우가 많아 이를 찾아 제거하였고, 종목 토론방 데이터 같은 경우에는 이모티콘과 같이 특수기호도 많아 이러한 불용어도 제거하였습니다. 전처리된 텍스트를 KR-FinBERT 모델에 입력하게 되면, '긍정', '중립', '부정' 세 가지 감정라벨 중 어떤 라벨에 속하는지에 대한 'label' 값과, 그 정도의 값이 'score' 값이 딕셔너리로 반환됩니다. 이를 각 데이터에 '감성분석'이라는 열로 추가하여 저장한 후 이후에 사용할 때는 이 두 가지의 값을 각각 'label' 열과 'score' 열로 분리하여 사용하게 됩니다.

daum_news_pipe.pickle

검색어	신문사명	날짜	제목	기사내용	링크	요약	감성분석
금호 건설	헤럴드 경제	2023-08-30 16:33:00	회사채 발행 속속 성공하는…	연합 헤럴드경제신동윤 기자 원자재 가격…	http://daum.net/v/2023083016...	건설 경기가 바닥을 치고 반등할 것이란 기…	[{'label': 'positive', 'score': 0.9872146844…}]

⋮

결과 활용

뉴스 데이터의 감성분석 결과의 경우, 이후에 유사도를 계산할 때 사용되기 때문에 추가적인 처리가 필요했습니다. 기업별로 세 가지의 감정 라벨 값의 개수, 세 가지의 감정 지수 값의 평균을 모두 저장합니다. 이때, 기업별로 뉴스 개수의 차이가 크기 때문에 비율과 값이 같다고 하더라도, 뉴스 기사의 개수에 따라 다르게 가중치를 주는 것이 필요하다고 판단하였습니다. 이를 위해 전체 기업에서 각 기업의 뉴스 기사 개수 비율을 계산하여 감성분석 결과값에 모두 곱해주었습니다. 이후 서비스 구현에서 시각화에 사용하기 위해 감정 비율의 합이 100이 되도록 각 기업별로 비율을 새롭게 계산한 열을 따로 저장해 두었습니다.

국내_유사도_최종_데이터셋.csv 중 [금호건설] 뉴스 관련 열

score_mean_negative	score_mean_neutral	score_mean_positive	ratio_negative	ratio_neutral	ratio_positive	new_ratio_negative	new_ratio_positive	new_ratio_neutral
1.961648	0.76539	0.548075	0.639872	0.438101	0.130976	0.52928	0.108339	0.362382

⋮

03 적용 모델 상세 설명

모델 소개 – 국내 뉴스 데이터, 국내 주식토론방 데이터

Komoran

Java 기반의 대표적인 오픈소스 형태소 분석기

키워드 추출

사용 이유

처음에는 키워드 추출 모델로 한국어를 학습시킨 생성형 AI 및 대형 언어 모델을 사용하여 '키워드 추출' 태스크를 쿼리로 넣는 방법을 고려하였으나, 대부분 모델의 크기가 서버 용량에 비해 컸으며 무엇보다 추론 시간이 길다는 점에서 크롤링한 모든 뉴스 본문 데이터를 시간 안에 처리하기 어려울 것이라고 판단하였습니다. 따라서 형태소 분석기를 사용하여 명사를 추출하는 것으로 키워드 추출을 대체하여 진행하기로 결정하였습니다.

지도학습 기반 한글 형태소 분석기에는 대표적으로 KoNLPy 패키지의 여러 모듈이 있습니다. 모듈에는 Kkma(꼬꼬마), Okt, hannanum, Mecab, Komoran으로 총 5가지가 존재합니다.

형태소 분석기를 사용하는 뉴스 데이터와 종목 토론방 데이터의 특징을 살펴보면, 하나의 데이터프레임에 약 12,000개가 넘는 행이 존재하며 한 행당 포함되어 있는 글자 수가 다양하게 분포되어 있었습니다. 따라서 모든 행의 데이터에 대해 형태소를 분석하기 위해서는 속도가 상대적으로 짧아야 했습니다. 또한 대부분의 글들이 '주식'과 관련된 단어로 구성되어 있고 그 양이 많아 사용자 사전을 활용해 단어를 저장하는데 한계가 있어 단어를 인식하는데 있어 좀 더 유동적인 모듈을 필요로 했습니다.

이러한 특징을 모두 반영하여 비교적 나은 결과를 도출할 수 있는 모듈이 바로 'Komoran'이었습니다. Komoran은 여러 어절을 하나의 품사로 분석 가능함으로써 형태소 분석기의 적용 분야에 따라 공백이 포함된 고유명사(영화 제목, 음식점명, 노래 제목, 전문 용어 등)를 더 정확하게 분석할 수 있다는 장점이 있습니다. 뿐만 아니라 분석 속도가 빠르다는 장점 또한 존재합니다.

전처리 과정

KoNLPy 패키지는 JAVA 기반의 형태소 분석기로 사용자의 로컬 컴퓨터에 알맞은 JAVA 환경이 구축되어 있어야 활용할 수 있습니다. 따라서 해당 컴퓨터에 JAVA와 JPye1-py3를 설치한 뒤 저장된 JPye1 버전을 JAVA_HOME으로 설정한 후 사용해야 합니다. 이 부분은 컴퓨터에 환경에 따라 상황이 다르니 코드를 실행 시 이 점을 꼭 참고해 주시기 바랍니다.

뉴스 데이터와 종목 토론방 데이터 모두 감성분석을 진행한 후 감정라벨이 추가된 결과 데이터를 활용하였습니다. 종목 토론방 데이터는 먼저 제목과 본문의 내용을 합친 후, 이후 과정은 뉴스 데이터와 동일하게 진행하였습니다.

불용어를 제거한 후 'Komoran'을 활용하여 명사를 추출하여 기존 데이터프레임의 새로운 열에 결과를 추가해주었습니다. 이후 감정 Label 별로 데이터를 분리시킨 뒤 글자 수가 한 글자인 경우는 제외해 주고 python의 Counter 함수를 활용하여 각 단어의 빈도 수를 계산하였습니다.

종목 토론방 데이터의 경우, 빈도 수 상위 50개를 추출하여 python의 wordcloud 함수를 활용하여 결과 화면의 워드 클라우드에 활용하였습니다.

03 적용 모델 상세 설명

모델 소개 – 미국 뉴스 데이터

PEGASUS GSG(Gap sentence generation)을 사용하여 중요 문장 단위로 학습을 수행하는 요약 모델

요약

사용 이유

미국 뉴스 또한 국내와 같이, 기사의 본문 요약 태스크를 진행하였습니다. 영어를 다루는 모델은 한국어를 학습시킨 모델보다 용량 부분에서의 한계가 적었으며, 추론 시간이 효율적인지와 Output 형태가 규칙적으로 추출됨에 따라 쉽게 대용량 데이터를 처리 가능한지가 주요 고려 요소였습니다. 따라서 요약(Summarization) 분야에 특화된 모델을 사용하기로 하였고, 문장 단위로 마스킹(Masking)을 진행하여 학습하는 PEGASUS(Pre-training with Extracted Gap-sentences for Abstractive Summarization) 모델을 선정하였습니다. 이 모델 또한 구글에서 제공한 요약 데이터(Pegasus-xsum)로 파인튜닝 (Fine-tuning)을 마친 모델을 사용하였는데 이 데이터에는 대량의 CNN 기사, 경제 메일 등이 포함되어 Financial 분야에 특화되도록 학습을 시킨 모델이며, 종목별 뉴스를 요약하는 태스크에 부합하다고 판단하였습니다.

전처리 / 결과

크롤링한 기사내용의 형식을 보면, 기사 머리글로 기사의 언론사, 나라 등과 같은 기사 작성 정보가 들어있었고, 이러한 정보가 기사 요약 성능을 악화시키는 것을 확인하여 머리글 부분을 모두 제거하였습니다. 또한 특수기호를 추가적으로 제거하여 최종적으로 전처리된 기사내용을 입력으로 넣으면 'summary_text'라는 key와 요약된 텍스트(value)가 JSON 데이터 형식으로 출력됩니다. 여기서 텍스트 부분만 추출하여 '요약'열로 저장하여 이후에 사용하였습니다.

FinBERT Financial 분야에 특화된 BERT 모델

감성분석

사용 이유

국내 뉴스의 감성분석에서 'KR-FinBERT'를 선정했던 동일한 이유로 미국 뉴스의 감성분석 태스크에도 FinBERT 모델을 사용하기로 하였습니다. 이 모델은 Financial PhraseBank 데이터를 사용하여 감성분석 태스크 학습을 진행하였으며 기존 BERT 모델보다 financial 도메인에서 우수한 성능을 보였다고 합니다.

전처리 / 결과

요약을 진행한 '요약 ' 열들의 텍스트를 모델에 입력으로 넣으면, '긍정', '중립', '부정' 세 가지 감정라벨 중 어떤 라벨에 속하는지에 대한 'label' 값과, 그 정도의 값이 'score'값이 딕셔너리로 반환됩니다. 이를 각 데이터에 '감성분석'이라는 열로 추가하여 저장한 후 이후에 사용할 때는 이 두 가지의 값을 각각 'label' 열과 'score' 열로 분리하여 사용하게 됩니다.

Example. **검색어**: Altimmune

기사내용: Altimmune Inc₩'s (ALT.O) experimental obesity drug helped reduce weight by over 10% on average in a mid-stage trial, the company said on Tuesday, but safety concerns sent its shares tumbling more than 50%. Most patients experienced nausea and vomiting …

summary Shares fall more than 50% after mid-stage trial fails to meet goals. Altimmune is targeting a global obesity treatment market expected to hit $50 billion

Sentiment Analysis [{'label': 'negative', 'score': 0.90323…}]

03 적용 모델 상세 설명

모델 소개 – 미국 뉴스 데이터, 미국 종목토론방 데이터

KBIR RoBERTa 구조를 차용하여 키워드 추출 태스크를 학습하는 모델

키워드 추출

사용 이유

미국 뉴스의 본문 내용을 사용하여 키워드 추출을 위한 모델로 KBIR 모델을 사용하였습니다. 이는 RoBERTa 모델의 구조에 infilling head와 replacement classification head를 추가하여 키워드 추출과 같은 'token-classification', 즉 문장의 전체적인 맥락을 나타낼 수 있도록 하는 키워드 리스트를 추출하도록 학습시킨 모델입니다. KBIR 모델을 뉴스 기사 데이터로 파인튜닝(Fine-tuning)시킨 모델의 가중치를 사용하여 뉴스 키워드 추출이라는 태스크의 성능 향상에 도움이 되도록 하였습니다.

전처리 / 결과

입력 데이터는 뉴스 기사 본문내용으로, 요약을 진행할 때 적용한 기사 머리글/특수기호 제외 전처리를 동일하게 적용하였습니다. 전처리된 텍스트를 모델에 입력하면 결과값으로 BIO 태깅(B-KEY, I-KEY, O), 키워드 등이 JSON 형식으로 추출되며 여기에서 키워드 리스트인 'word' 값만 추출하여 '키워드' 열로 저장하였습니다. 일부 키워드 앞에 'Ġ' 라는 특수기호가 함께 추출되는 것을 확인하여 해당 기호를 제외하고 정제된 순수 키워드를 '키워드_리스트 ' 라는 열에 저장하여 최종 키워드 리스트를 저장하였습니다.

usa_news_pipe.pickle 중 Altimmune 일부 결과

검색어	기사종류	날짜	링크	…	키워드_리스트
Altimmune	Business	2023-03-22	https://www.reuters.com/business/healthca...		['safety', 'concerns', 'nausea', 'vomitting', 'side', 'effects', 'Securities', 'reduced' …]

기사내용, Ticker, 요약, 감성분석

Twitter - RoBERTa 대량의 Twitter 데이터로 학습시킨 RoBERTa-base 감성분석 모델

#감성분석

사용 이유

이전에 사용한 모델들은 활용 도메인이 financial, 혹은 매체가 뉴스라는 점에 집중하였지만, 미국 주식토론방인 Stocktwits는 그 성격이 조금 다르다고 판단하였습니다. 이는 국내의 네이버 종목토론방은 [네이버 금융] 페이지 내에 있는 토론방으로 각 종목의 주가 정보, 최근 이슈들을 관련 용어들을 사용하며 글을 작성하는 경우가 많았기 때문입니다. 반면 Stocktwits의 경우에는 하나의 고유한 사이트로 사용자들이 자유롭게 사진과 영상과 함께 다양한 기업들을 태그하며 글을 작성하고, 답글 기능을 통해 의견을 공유하는 SNS 형태와 더욱 유사했습니다. 사용하는 용어들 또한 신조어와 같은 인터넷 용어들의 비중이 컸으며, 이를 파악하여 실제 인터넷 사용자들의 심리, 감성을 분석하는 것이 중요하다고 판단하였습니다. 따라서 RoBERTa 모델을 대량의 Twitter 데이터로 파인튜닝(Fine-tuning)시킨 모델을 사용하여 감정분석을 진행하였습니다. 이때 사용자들이 함께 태그한 기업들이 '$'와 함께 글의 앞부분에 첨부가 되기 때문에 이 리스트를 '관련기업' 열로 저장하였습니다. 그리고 전처리한 텍스트를 모델에 입력하여 감정 'label', 'score' 값을 얻을 수 있었습니다.

03 적용 모델 상세 설명

모델 소개

Gower-distance

거리 계산을 통한 유사도 기반 추천 종목 선정하기

#유사도

최종 데이터셋 merge

국내/해외의 모든 섹터분류, 재무제표, 뉴스, 종목토론방 데이터를 합쳐 추천 종목을 선정하기 위한 최종 데이터셋을 만들었습니다.

먼저, 국내/해외 각각 분류된 섹터에 따라 기업별로 어떤 섹터에 속하는 지에 대한 열을 추가하여 기본 데이터셋을 만들었습니다. 국내의 경우, 한 기업이 여러 섹터에 분류되어 있는 경우가 많기 때문에 이를 반영하기 위해 총 147개의 섹터를 원-핫 인코딩(One-hot Encoding)으로 표현하여 147개의 열로 만들었습니다. 해당되는 섹터의 열의 값은 1, 아니면 0으로 저장됩니다.

그 다음, 최근 4분기의 재무제표 데이터에 대한 열을 병합합니다. 재무제표의 목록으로는 매출액, 매출원가,… 등 총 17개가 존재하며 각 항목의 4분기, 즉 68(17 X 4) 개의 열을 추가하게 됩니다.

뉴스 데이터의 열은 감정 분석을 한 결과인 감정 라벨의 비율, 스코어 값, 비율을 백분율로 다시 계산한 비율 총 3가지로 이루어져 있으며 각각의 항목마다 'positive', 'neutral', 'negative' 세 감정라벨의 값이 존재하기 때문에 9(3 X 3) 열이 저장되게 됩니다.

마지막으로, 종목토론방의 '화제성' 열입니다. 사용자가 서비스를 이용하는 날짜를 기준으로 수집된 데이터 중 최근 날짜에 더 많은 가중치를 부여하여 화제성을 계산하도록 하였습니다. 먼저, 사용자가 서비스를 이용하는 날짜를 받아 각 행의 '날짜' 열 값의 차이값을 추출합니다. 그 후 전체 값의 max 값을 도출한 뒤 차이값들을 max값에서 빼 줍니다. 이때 최소값은 무조건 1값으로 선정하기 위해 추가로 1을 더해줍니다. 이렇게 되면 최근 날짜에는 높은 값이, 예전 날짜에는 낮은 값이 계산됩니다. 이러한 방식으로 가중치 열을 추가적으로 저장하였습니다.

따라서 최종 데이터셋은 225개의 열(147 + 68 + 9 + 1)로 이루어져 있으며 국내/해외 각각 기업의 개수만큼 (2544, 225), (4882,225) 크기의 데이터로 완성됩니다.

추천 종목 선정 흐름

데이터의 거리를 계산하는 방법으로 가워 거리(Gower-distance)를 사용하였는데, 이는 범주형/연속형 변수들을 모두 다루기 위함입니다. 이때, 사용자가 직접 어떤 항목에 가중치를 줄지 설정을 할 수 있도록 하기 위해 225개의 항목이 있는 가중치 딕셔너리를 정의하여 사용자가 자유롭게 설정하는 선택 필터에 따라 (ex. 전체종합, 재무제표_종합, 감정지수, 화제성…) 해당 key의 가중치 값을 바꾸도록 하였습니다.

기준이 되는 하나의 Target 기업이 국내 기업이라고 가정하면 그 기업의 벡터와 해외 데이터셋 내 모든 기업들 간의 가워 거리를 계산하게 됩니다. 계산 과정에서 이전에 구한 가중치를 곱하여 원하는 필터링 항목의 영향력을 키우게 되며, 최종적으로 4882개의 거리 계산값에서 작은 순서대로 3개의 기업을 선정하여 해당 기업들을 추천하게 됩니다. 기업 추천 결과 예시는 '소스코드/유사도계산 과정 노트북/유사도계산 예시.ipynb' 파일에서 보실 수 있습니다.

03 적용 모델 상세 설명

소스코드 설명

전체 과정 클라우드 자동 업로드

알고리즘 흐름

데이터셋을 구축하기 위해 필요한 raw 데이터 대부분이 크롤링 데이터이기 때문에 데이터 수집부터, 이를 저장하고, 로드하는 데에 많은 시간이 소요됩니다. 따라서 네이버 클라우드 플랫폼의 Object Storage를 활용하여 언제든지 필요한 데이터를 불러올 수 있도록 크롤링 데이터부터 전처리 결과, 모델링 결과에 대한 데이터들을 날짜별로 중간 과정마다 클라우드에 저장하도록 하였습니다.

저희의 서비스는 매일 새롭게 업데이트 되는 뉴스 기사, 주가 정보 등을 반영하여 데이터셋을 업데이트해야 하는데, 이를 위해 [main.py] 파일의 'gower_sim_main(…)' 함수를 매일 자정에 실행시킬 수 있도록 합니다. 이때, 미국과의 시차로 인해 미국 데이터의 경우에는 그 전날의 자정부터 해당되는 데이터를 가져오도록 하였습니다. 그러면 재무제표를 제외한(재무제표는 분기마다 업데이트 된 후 financial_main() 함수로 크롤링하기) 주가정보, 뉴스, 주식토론방 등의 하루치 데이터를 해당 날짜의 폴더를 자동으로 생성하여 모두 클라우드에 저장하게 됩니다.

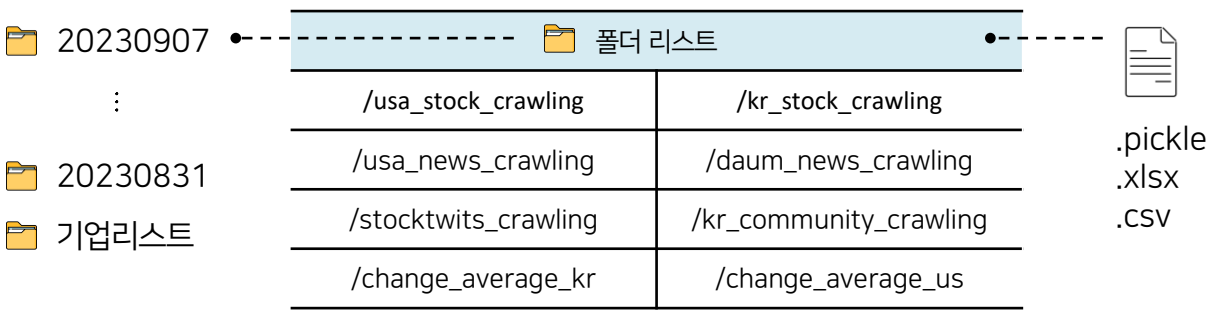
크롤링이 끝나면 뉴스와 종목토론방을 이용한 감성분석 등의 태스크가 진행이 되는데. 이때 기업별로 각각 하나의 파일로 크롤링된 피클(pickle), 엑셀(xlsx) 파일들을 클라우드에서 불러와 일자별로 취합을 한 후 다시 국내/미국, 뉴스/종목토론방 별로 하나의 결과 데이터셋으로 저장하여 같은 폴더에 저장되게 됩니다.

이 모든 작업이 끝나면 추천 종목 선정을 위한 데이터셋 취합이 시작되는데, 이때 뉴스/종목토론방 관련 데이터를 불러올 때 오늘 새롭게 업데이트된 데이터들을 어제의 데이터프레임과 병합한 후, 데이터프레임의 '날짜' 열에서 가장 오래된 날짜에 해당되는 행들을 삭제하여 최근 기간의 정보가 반영되는 것을 유지하도록 합니다. 이렇게 완성된 최종 데이터셋은 클라우드에 업로드한 후, 오늘의 종목 추천 서비스에 사용이 됩니다.

[main.py]파일의 eda_main() 함수는 종목 추천 서비스의 결과로 나온 기업에 대한 시각화 대시보드 구성을 위해 데이터를 처리하는 함수입니다. 급등/급락 기간을 추출하고 해당 기간의 이슈를 키워드로 나타내거나, 평균 등락률을 계산하는 함수가 포함되어 있으며, 추천 종목으로 선정된 기업의 상세 정보를 제공하기 위해 필요한 함수입니다.

Object Storage

Example) 2023-09-07에 main.py 실행



03 적용 모델 상세 설명

소스코드 설명

등락률 계산 주가 데이터의 '등락률' 열 활용

차트 분석

주가 데이터와 뉴스 데이터를 활용하여 급등&급락 기간의 이슈에 대해 분석하여 사용자들이 과거 이슈와 그 반영 정도에 대해 한 눈에 파악할 수 있도록 제공하는 서비스입니다. 해당 서비스 제공을 위해서 먼저 주가데이터를 통해 급등&급락 기간을 추출하고 뉴스 데이터의 키워드 추출을 통해 해당 기간의 이슈를 파악합니다.

먼저, 차트 분석에는 6개월 치의 주가데이터만 활용하기 때문에 사용자가 서비스를 이용하는 날짜를 받아 해당 날짜를 기준으로 6개월 동안의 주가 데이터를 불러옵니다.

불러온 주가 데이터의 '등락률(Change)'열의 특징을 살펴보면, 예를 들어 '2023-08-27'의 등락률이 1.47이라고 한다면 이건 전날인 '2023-08-26'일의 종가를 기준으로 27일의 종가가 1.47% 올랐다는 의미입니다.

이 점을 고려하여 급등 기간 추출의 경우, 해당일의 등락률이 전일과 비교하여 음수에서 양수로 변했다면 해당일의 전날을 'Start Date'로 설정하고 연속적으로 양수이다가 다시 음수로 변하는 시점의 전일을 'End Date'로 설정합니다. 급락 기간 추출의 경우, 급등과 반대로 해당일의 등락률이 전일과 비교하여 양수에서 음수로 변했으면 해당일의 전날을 'Start Date'로 설정하고 연속적으로 음수이다가 다시 양수로 변하는 시점의 전일을 'End Date'로 설정합니다. 이렇게 하여 기간(Duration)을 계산하여 그 기간이 1일 이상인 데이터만 수집한 뒤 해당 기간의 평균 등락률을 계산합니다. 최종적으로 평균 등락률의 절댓값을 비교하여 상위 4개의 기간을 추출합니다.

그리고, 기사내용에서 키워드 추출을 마친 'daum_news_pipe.pickle', 'usa_news.pickle' 파일을 불러옵니다. 타겟이 되는 기업의 행만을 추출한 후 4개의 기간별로 키워드를 모두 합친 합산 키워드 리스트를 만듭니다. 그중, 상위 3개의 키워드를 추출하여 첫 번째 키워드에 대해서는 기간(Duration) 내에서 처음으로 등장한 시점, 그 시점의 뉴스 요약 기사, 해당 기사의 링크를 함께 저장합니다.

뉴스 데이터에서 추출된 키워드와 해당 뉴스 및 링크 값이 주가 차트에서 급등&급락 기간에 표시되도록 시각화를 진행하였습니다. 해당 시각화는 '시각화.ipynb' 파일에서 확인하실 수 있습니다.

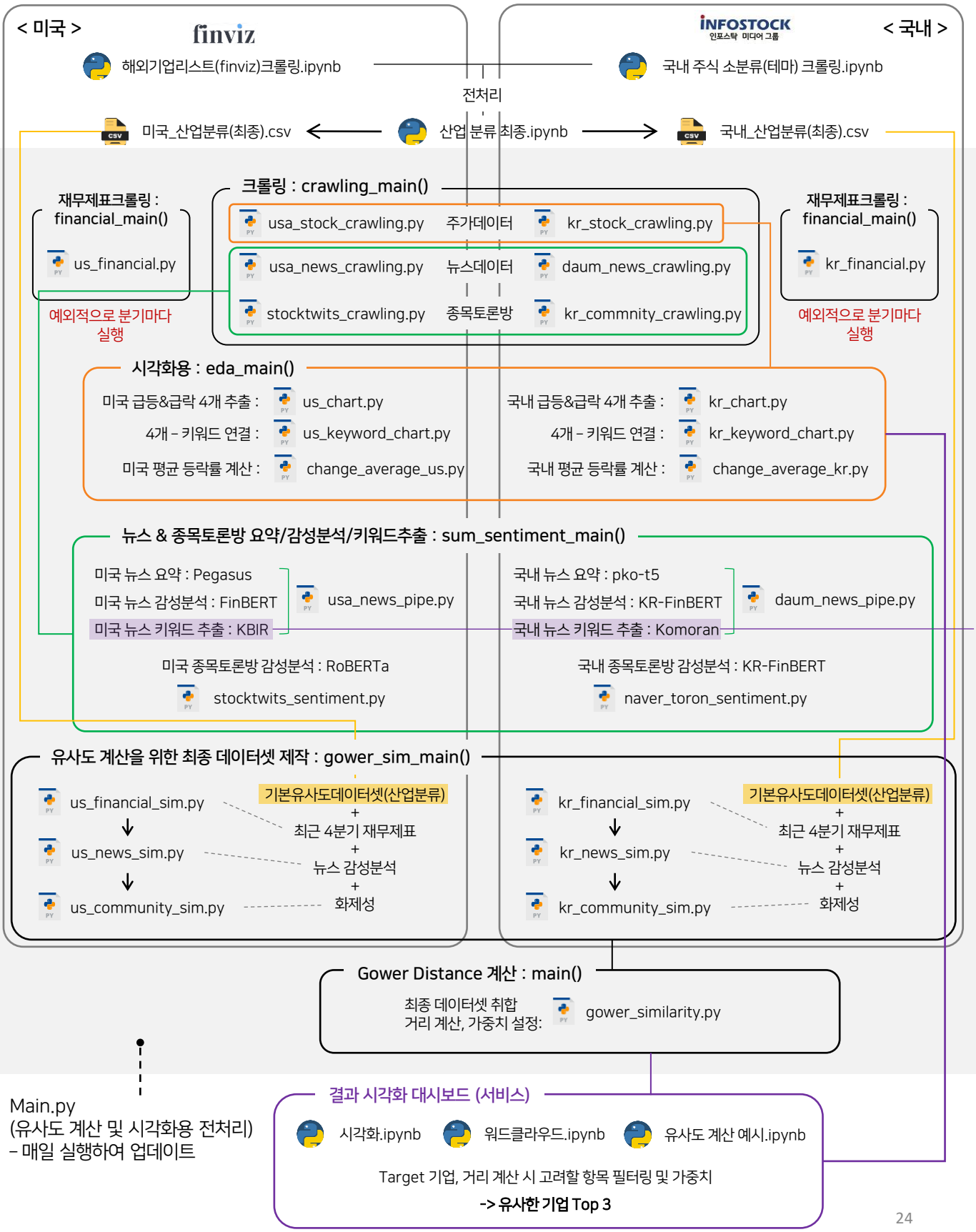
평균 수익률

평균 수익률의 경우, 1개월 간, 2개월 간, 3개월 간으로 분리하여 총 3개의 데이터를 사용자들에게 제공합니다. 사용자가 서비스를 이용하는 날짜를 받아 해당일을 기준으로 약 1개월, 2개월, 3개월 간 각각의 평균 등락률을 계산합니다.

해당 서비스를 통해 사용자들이 투자를 결정하는데 있어서 도움을 주고자 합니다.

03 적용 모델 상세 설명

서비스 알고리즘 흐름도



04 서비스 사용 제안에 따른 사용 예시

서비스 사용 예시

저희 서비스의 기본 원리는 고객이 기준이 되는 종목과 세부 지표들을 선택하면 국내 주식 종목과 해외 주식 종목의 유사도를 계산하여 유사도가 높은 상위 3개의 종목을 추천 종목으로 제공해주는 것입니다.
'[빅데이터 부문/채지피티] 프로토타입.pdf' 파일을 바탕으로 사용자가 서비스를 이용하는 흐름 순서대로 자세히 설명드리겠습니다. (그래프 시각화 부분은 '소스코드/대시보드 과정 노트북/시각화.ipynb' 참고)

1. 사용자 선택

먼저, 사용자가 국내 주식을 기반으로 해외 주식 종목에 대해 추천 받을 것인지, 해외 주식을 기반으로 국내 주식 종목에 대해 추천 받을 것인지 선택합니다. 그 후 종목 검색 란에 추천의 기준이 되는 종목을 검색하여 선택합니다.

세부사항으로는 유사도 계산 시 사용자가 중점적으로 고려하고자 하는 지표들이 있습니다. 대분류는 [종합, 산업분류, 재무제표, 뉴스, 투자심리]로 구성되어 있습니다. '종합'을 선택하면 모든 지표를 추가적인 가중치 부여없이 동등하게 고려하게 되는 것이고, 그 외 분류를 선택하면 선택된 지표들에 적절한 가중치가 부여됩니다. 또한 대분류에 속하는 소분류로 세분화된 선택을 할 수도 있는데, '재무제표'에는 17개의 세부 지표로 구성되어 있습니다. > MockUp 2p

사용자가 'NAVER' 기업에 대해서 '뉴스 > 감정비율'에 대해 중점적으로 유사한 기업들을 추천받고 싶다고 선택하였다는 가정 하에 이후 결과들에 대해 설명 드리겠습니다.

2. 추천 결과

저희의 유사도 계산 결과에 따르면 유사도 상위 3개의 기업으로 'CRGE', 'CR', 'ASST' 기업이 선정되었습니다. 첫 번째 화면에서는 각 기업에 대해 'NAVER'와 비교해서 '산업 분류', '뉴스', '투자 심리' 기준에 대해 유사 정도를 사용자가 한 눈에 확인할 수 있도록 제공합니다. > MockUp 3p

또한 '한 눈에 보기' 파트에서는 'NAVER'와 나머지 세 기업에 대한 각 세부사항별 자세한 정보를 확인할 수 있도록 제공합니다. 여기서 '재무제표'의 경우, 사용자가 재무제표의 세부 지표들을 유사도 계산 시 고려했으면 좋겠다고 선택하였을 때, 해당 지표들에 대해 색깔을 달리 하여 사용자가 직관적으로 결과를 이해할 수 있도록 제공합니다. '뉴스' 정보에서는 유사도 계산 과정에서 사용한 정보인 '감정 분포'와 '감정 점수'에 대하여 4개 기업의 정보를 한 눈에 비교하여 파악할 수 있도록 제공합니다. 또한 '투자심리'의 경우, 사용자의 서비스 이용일을 기준으로 과거 3일간의 종목토론방의 화제성을 그래프로 표현하여 한 눈에 그 정도를 확인할 수 있도록 제공합니다. > MockUp 7-8p

3. 상세 보기

저희 서비스는 추가적으로 4개의 종목에 대한 각각의 '상세보기'를 제공합니다. 해당 서비스에 대해서는'CRGE' 기업의 상세보기를 기준으로 설명 드리겠습니다.

사용자가 '상세 보기' 항목을 누르게 되면, 해당 기업에 대한 재무제표 정보 및 시각화, 뉴스의 감정분석 및 차트 분석, 종목 토론방을 활용하여 제공되는 워드 클라우드 결과, 주가 데이터를 활용한 기간별 수익률 등에 대한 정보가 제공됩니다. '상세 보기'에서 제공되는 항목은 4개 기업 모두 동일하게 적용됩니다. >MockUp 4p

04 서비스 사용 제안에 따른 사용 예시

서비스 사용 예시

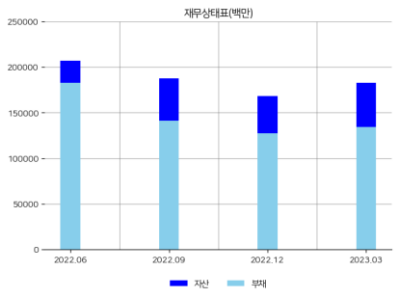
3. 상세보기

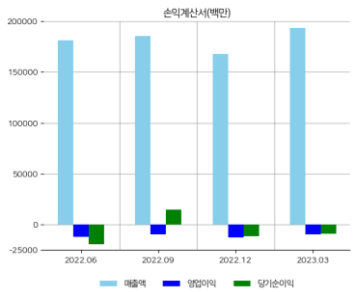
먼저, 재무제표 정보 및 시각화 부분부터 자세히 살펴보겠습니다.

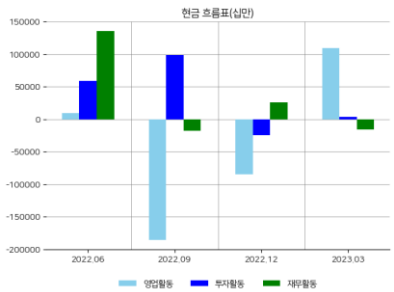
재무제표의 경우, 유사도 계산에 활용한 데이터를 동일하게 제공합니다. 대신 요약 보기에서는 확인할 수 없었던 분기별 변화 흐름에 대해서 사용자가 쉽게 파악할 수 있도록 각 재무제표 항목별로 시각화를 하여 제공합니다.

재무상태표의 경우, 자산총액과 부채총액에 대하여 2022년 6월부터 각 분기별 정보를 막대그래프로 확인할 수 있도록 제공합니다. 저희가 분석을 진행할 당시, 2023년 3월에 제공된 재무제표가 가장 최근에 업데이트된 정보여서 해당 정보를 활용하여 아래 첨부된 자료와 같이 시각화하였습니다. >MockUp 4p

손익계산서의 경우 매출액, 영업이익, 당기순이익에 대하여, 현금흐름표의 경우, 영업활동, 투자활동, 재무활동에 대하여 분기별 정보에 대해 막대그래프로 한 눈에 변화를 이해할 수 있도록 제공합니다.

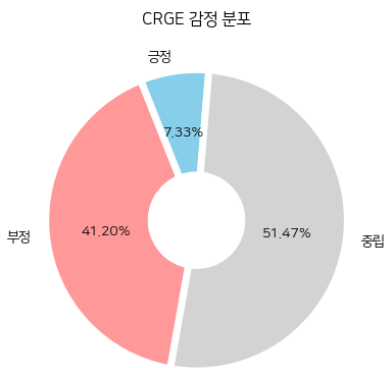


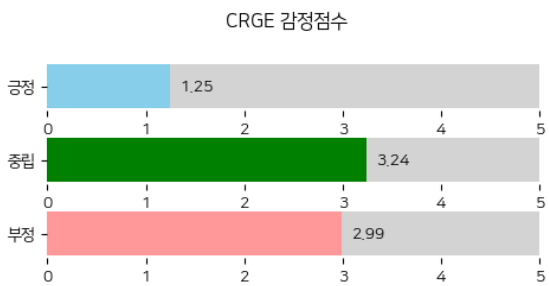




다음, 뉴스 데이터의 경우, 감성분석을 통해 도출된 '긍정, 부정, 중립'의 전체적인 퍼센테이지를 '감정 분포'라는 제목으로 제공합니다. 이를 통해 사용자는 수집된 뉴스들에 대하여 전체 뉴스에서 각 감정의 분포 정도를 확인하여 투자를 판단하는데 활용할 수 있습니다.

또한 감성분석을 통해 도출된 '긍정, 부정, 중립'에 대한 각 감정의 score 값의 평균을 도출하여 '감정 점수'라는 제목으로 제공합니다. 'CRGE'라는 기업을 확인해보면 '긍정' 감정의 평균 score 값은 1.25로 전체 뉴스의 7.33% 정도 되는 긍정적 뉴스들의 실질적인 긍정 점수는 5점 만점에 1.25점으로 상대적으로 낮은 점수를 부여 받았다고 해석할 수 있습니다. 이러한 정보를 통해 사용자들은 해당 기업의 호재와 악재의 정도를 파악하여 투자 판단에 활용할 수 있습니다. >MockUp 5p





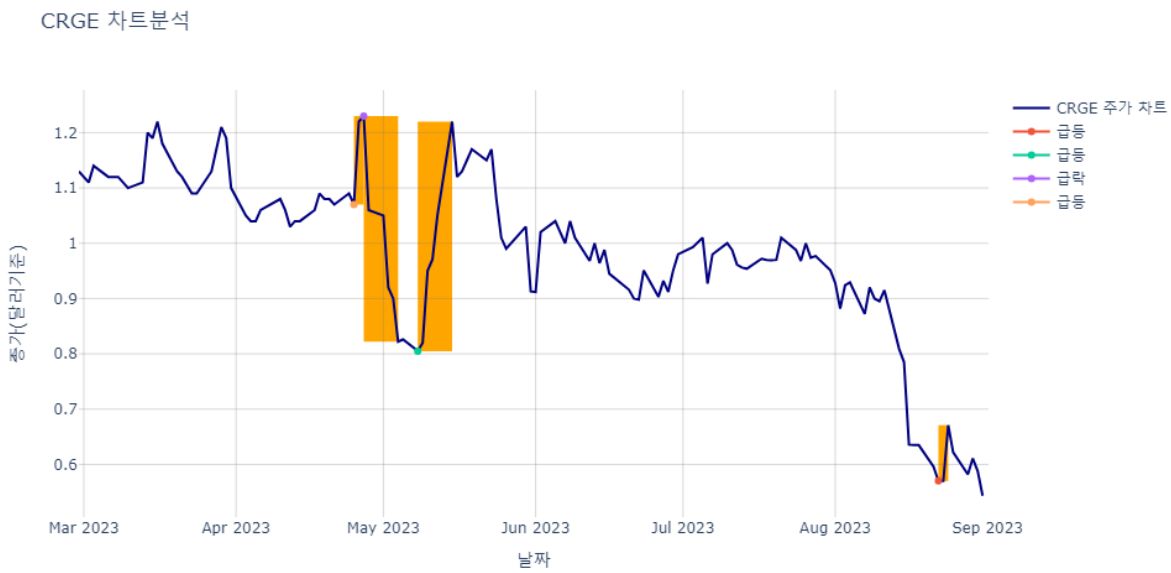
04 서비스 사용 제안에 따른 사용 예시

서비스 사용 예시

3. 상세보기

차트 분석의 경우 수집된 주가 데이터에서 사용자가 서비스를 이용하는 시점으로부터 6개월전까지의 주가데이터를 활용하여 추출된 '급등&급락' 상위 4개의 기간과 관련 있는 뉴스 키워드를 제공합니다. 이를 통해 사용자들은 해당 기업이 과거에 어떤 이유로 얼마 정도의 급등&급락이 이루어졌는지 한 눈에 확인하고 투자 판단에 활용할 수 있습니다. 아래 시각화 자료에서는 주황색 박스에 속한 부분이 상위 4개의 급등&급락 기간입니다. Python의 'plotly' 패키지를 활용하여 해당 부분에 마우스를 가져가면 해당 기간에 대한 정보를 확인할 수 있도록 구성하였으니 자세한 정보는 '소스코드/대시보드 과정 노트북/시각화.ipynb' 파일을 확인해주시면 됩니다.

>MockUp 6p



종목 토론방의 경우, 감성 분석을 통해 도출된 '긍정, 부정, 중립' 총 3가지의 감정별로 키워드들을 추출하여 워드 클라우드를 진행한 결과 화면을 제공합니다. 이를 통해 사용자들은 전체 게시글에서 각 감정들의 분포와 해당 감정에서 나타나는 키워드들을 한 눈에 확인할 수 있습니다. >MockUp 6p

< 긍정 >

< 중립 >

< 부정 >

마지막으로 기간별 수익률 정보의 경우, 1개월, 2개월, 3개월 간의 평균적인 등락률을 계산한 결과를 제공합니다. 해당 정보를 통해 사용자들에게 투자를 판단하는데 있어 도움을 주고자 하였습니다. >MockUp 6p

실제로 구현되었을 때의 서비스 화면은 '서비스 프로토타입.pdf'와 제출한 링크를 참고해주시면 됩니다.

링크 : https://chaegpt.netlify.app

05 부록

출처

[6p]
심우일, “20대 10명 중 2명, 해외 주식 투자한다”, 서울경제, 2021.07.29, https://www.sedaily.com/NewsView/22P3MKI0NG
[16p]
이유진 외, “특허 문서를 위한 형태소 분석기 비교 평가“, (국립군산대학교, 한국정보기술학회, Proceedings of KIIT Conference , 2019), 264-265.
[17p]
장주현, 김재윤, “KR-FinBERT 뉴스 감성분석을 활용한 KOSPI 주가지수 예측”, (순천향대학교, 2023년도 한국통신학회 동계종합학술발표회논문집, 2023), 1,142-1,143.
안재준 외, “BERT 감성분석과 기술분석을 결합한 주식시장 예측에 관한 연구”, (연세대학교, 대한산업공학회 추계 학술대회 눈문집, 2021), 1,630-1,637.
[18p]
Jiang Tingsong and Zeng Andy, “Finanial sentiment analysis using FinBERT with application in predicting stock movement”, (University of Pennsylvania, University of Rochester, CS224U project, 2023), 1-5.
Zhang, Jingqing et al., “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, (Computer Science - Computation and Language, arXiv:1912.98777)
[19p]
Kaeley Harsimrat, Qiao Ye and Bagherzadeh Nader, “Support for Stock Trend Prediction Using Transformers and Sentiment Analysis”, (University of California, Irvine, IISES 18th Economics & Finance Conference, London, n.d.), 1-7p.
You, Lan a, b et al., “ASK-RoBERTa: A pretraining model for aspect -based sentiment classification via sentiment knowledge mining”, (a School of Computer Science and Information Engineering, Hubei University, Hubei Wuhan, 430062, China, In Knowledge-Based Systems, 2022), 109511.
Ao Xiong et al., “News keywords extraction algorithm based on TextRank and classified TF-IDF”, (Beijing University of Posts and Telecommunications, China, Beijing, China, international Wireless Communications, 2020), 1,364 – 1,369.
[20p]
Ranalli Monia, Rocci Roberto, “A comparison between methods to cluster mixed-type data: Gaussian mixtures versus Gower distance.”, (Dipartimento di Scienze Statistiche, Università di Roma `La Sapienza'”, Statistical learning and modeling in data analysis---methods and applications, 2021), 163-172.