[첨부 1]

[캡스톤디자인 신청서]

■ 연구과제

과 제 명	주가 분석 및 종목 감성분석을 통한 종목 추천	참여학기	2020 년 1 학기

■ 신청자

팀 명	STOCK				팀구성 총인원	총 3 명
구 분	성명	학번	소속학과	학년	연락처	이메일
대표학생	박승혜	2017103081	응용수학과	4	010-9141-3439	tmdgp1204@naver.com
참여학생	송재원	2013103432	수학과	4	010-5898-1596	rearsilre@gmail.com
	이상윤	2017110264	소프트웨어융합학과	4	010-6541-5667	syl11121@khu.ac.kr

■ 지도교수

지도교수	성 명	이대호	직 급	전임교수
	소속대학	소프트웨어융합대학	소속학과	소프트웨어융합학과

■ 붙임

[양식1] 수행계획서
[양식2] 팀구성원 명단 (2인 이상 팀인 경우 해당)
[양식3] 결과보고서
[양식4] 과제 요약보고서

본인(또는 팀)은 상기와 같이 캡스톤디자인을 신청하며, 이를 성실히 수행하겠습니다.
학습에 불성실하였거나, 중도포기 시에는 낙제 성적 부여함에 이의가 없음을 서약합니다.

일자 : 2020 년 4 월 10 일

신청자(또는 팀 대표) 박 승 혜 (서명)

[양식 1] 수행계획서

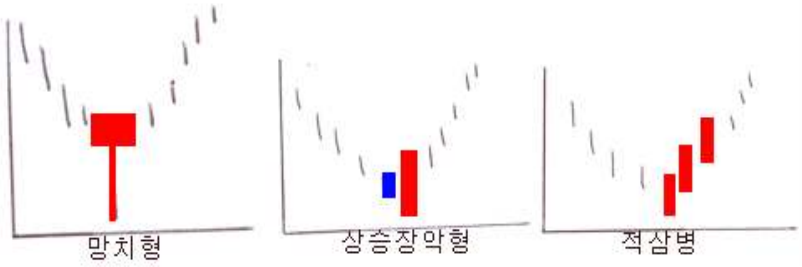
[캡스톤디자인 수행계획서]

과 제 명	주가 분석 및 종목 감성분석을 통한 종목 추천

1. 과제 개요

가. 과제 선정 배경 및 필요성

주식은 과거의 종가/시가/거래량 등으로 주가의 흐름을 읽는 것이 가능합니다. 예를 들어 망치형 패턴이 출현하는 경우는 전일의 하락세로 인해 떨어지던 주가가 어느 지점에 다다른 후 주가가 충분히 하락했다고 생각하는 사람이 등장하여 최저가를 찍고 다시 올라가는데, 그 주가가 첫 거래보다 오른 상태에서 장이 마치는 경우입니다. 이러한 경우는 당일의 상승세를 반영하여 다음날의 주가 또한 오를 가능성이 높습니다. 이외에도 상승장악형, 적삼병 등 주가 정보를 분석하면 주가의 단기 예측이 가능합니다.



[그림 1] 주가가 상승할 것으로 예측되는 봉차트 패턴 (출처:[경제신문 읽는법])

[그림 3] 재무상태표와 손익계산서

또한 기업 자체의 실적 정보를 참고하면 주식의 가격을 예측할 수 있습니다. 매출이나 실적이 우수한 기업의 경우, 부도가 날 가능성이 낮으므로 안정적인 종목이 됩니다. 이러한 기업은 주식을 사고자 하는 사람이 많아지고, 주가는 상승하는 모습을 보이게 됩니다. 기업 자체의 실적 정보는 재무상태표나 손익계산서 등으로부터 얻을 수 있습니다.

최근 코로나의 영향으로 바이오 분야의 주식이 상승하고 있는 것과 같이 주가는 기업 또는 산업마다 발생하는 사회적 사건에도 영향을 받습니다. 이로 인해 특정 사건 전후에는 같은 산업군에 해당하는 주식이 대체적으로 유사한 차트 구조를 보일 것이라고 예측할 수 있습니다. 각 기업의 경영 활동에 대한 정보 또한 주가에 큰 영향을 주기 때문에 이러한 정보를 얻을 수 있는 뉴스로부터 기업 또는 산업 간 미치는 영향과, 각 기업별 주가의 변화를 분석할 수 있습니다.

나. 과제 주요내용

코스피 200의 각 산업군별 대표 종목을 4가지씩 선정하고, 각 종목별 과거의 주식 정보를 입력값으로 하여 주가 분석을 진행합니다. 여러 가지 수치적인 분석 방법을 시도하고 각각의 정확도를 MAPE를 사용하여 측정합니다. 추가적으로, 기업의 실적 정보를 포함하여 분석할 경우 기존의 방법에 비해 예측값이 어떻게 변화하는지 확인합니다.

각 종목에 대한 뉴스로부터 가장 많이 등장하는 단어들을 추출하고, 단어사전을 구축합니다. 단어사전 내의 단어들이 긍정적인 의미를 나타내는지, 부정적인 의미를 나타내는지 확인하고, 긍정값을 부여합니다. 뉴스 본문을 문장 단위로 나누고, 문장 내 사전에 포함된 단어들의 긍정값 곱하여 해당 문장의 긍정값을 부여합니다. 이를 통해 해당 기사의 긍정값을 계산합니다. 위에서 진행하였던 주가 분석에 계산된 기사의 긍정값들을 입력값으로 추가하여 예측의 정확도를 측정하고, 추가하기 전과 비교하여 뉴스기사의 오피니언마이닝이 주가에 영향을 미치는지 판단합니다.

최종적으로 결정된 주가 분석 방법으로 이후의 주가를 예측합니다. 가장 상승률이 높을 것으로 예측되는 산업군과 해당 종목을 찾고 추천합니다. 뉴스기사의 오피니언 마이닝에서 기업의 발생 사건이외에 사회적 사건이 발생하는 경우에도 분석이 가능한지 확인하기 위해 '코로나'라는 변수가 없는 경우와, 이를 포함하여 학습을 진행한 경우의 예측 결과를 비교합니다.

2. 과제의 목표

가. 최종결과물의 목표 (정량적/정성적 목표를 정하되, 가능한한 정량적 목표로 설정)

코스피 200에 해당하는 주식 종목들을 11개 산업군 마다 시가총액 순으로 4개의 대표 기업들을 선정하여 총 44개 기업을 대상으로 분석을 진행합니다. 주가의 수치적인 분석은 LSTM, GRU, CONV 1D 등의 분석방법을 사용하여 가장 적절한 방법을 찾고, 오피니언 마이닝이 유의미한 결과를 가져오는지 분석합니다. 도출된 결과들을 통해 대표 기업들의 주가를 예측하고 좋은 결과를 내는 산업군에 대한 종목을 추천하고, 실제 추천결과를 검증합니다. 검증의 정확도는 MAPE를 사용하고, 0.3을 정확도값의 목표로 설정합니다.

나. 최종결과물의 세부내용 및 구성

특정 기간을 설정하여 해당 기업에 대한 기사 정보를 오피니언마이닝을 통해 당일의 기사가 긍정/중립/부정인지 수치적으로 값을 도출합니다. 이후 해당 값을 당일의 주가 정보와 함께 LSTM, GRU, CONV 1D를 사용하여 각각의 방법마다 주가를 예측하고, 정확도를 계산합니다. 도출된 주가들에 대해 PER, ROE 등의 회사정보를 추가하여 수정된 주가를 예측하고, 정확도를 계산합니다. 회사정보를 추가하지 않은 예측값과 추가한 예측값을 비교하고, 가장 좋은 결과를 내는 분석방법을 찾아 예측을 진행하여 상승률이 가장 높을 것으로 예측되는 산업군에 대한 종목을 추천합니다.

3. 기대효과 및 활용방안

사회적으로 기업 및 산업의 발전 흐름이 어떻게 흘러가고 있는지를 파악하고, 주가를 수치적으로 분석함으로써 주식 시장의 흐름을 이해하고, 리스크를 최소화할 수 있는 종목을 찾을 수 있습니다.

4. 수행방법

가. 과제수행을 위한 도구적 방법 (활용 장비, 조사 방법론 등)

i) 주가의 수치분석

가) ARIMA (Auto Regressive Integrated Moving Average)

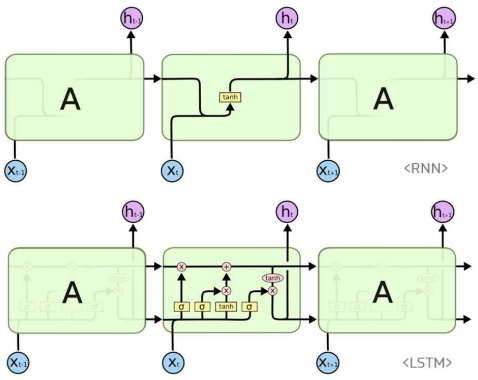
자기회귀와 이동평균 및 시계열을 비정상성을 모두 고려하는 모델입니다. 과거의 관측 값과 오차를 사용하여 현재의 시계열 값을 설명하는 ARMA모델을 일반화 한 모델으로 분기/반기/연간 단위로 다음 지표를 예측하거나 주간/월간 단위로 지표를 리뷰하여 모니터링 하는데 사용되는 기법입니다. 분석 대상이 다소 비안정적인 시계열의 특징을 보여도 적용이 가능합니다.

일반적인 형태는 $X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_t - p = Z_t + \theta_1 Z_t - 1 + \cdots + \theta_q Z_t - q, t = 0, \pm 1, \pm 2, \cdots$로 표현됩니다. 이 때, X_t는 ARIMA를 통해 예측하고자 하는 데이터, Z_t는 백색잡음으로 모두 독립적이고 동일하게 분산된 확률변수를 사용합니다.

나) Prophet

Facebook에서 개발한 시계열 예측 모델으로 확률적이고 이론적인 모형이 아닌 경험적 규칙을 사용하는 단순회귀 모형입니다. 주기적인 특성을 갖는 시계열 데이터에서 가장 잘 동작하며 non-daily data를 다룰 수 있습니다. Accurate and fast, fully automatic, Tunable forecasts, available in R or Python 이라는 장점들을 가지고 있습니다. Python에서 import Prophet을 통해 사용할 수 있습니다.

다) LSTM, CONV1D



RNN은 히든노드가 방향을 가진 엣지로 연결되어 순환구조를 이루는 인공신경망의 한 종류로 순차적으로 등장하는 데이터 처리에 적합한 모델입니다. RNN은 관련 정보와 그 정보를 사용하는 지점 사이의 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하되는 vanishing gradient problem이 발생하게 됩니다. LSTM은 RNN의 일종으로 이러한 문제를 극복하기 위해 고안된 것으로 RNN의 히든 state에 cell-state를 추가한 구조입니다. cell-state가 일종의 컨베이어 역할을 하여 오래 경과

하는 state에도 그래디언트의 전파가 비교적 잘 됩니다.

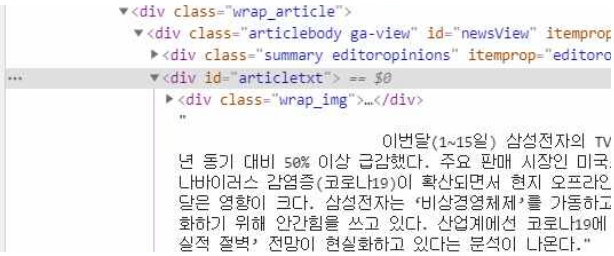
Conv1D는 필터를 이용하여 지역적인 특징을 추출하는 컨볼루션 레이어입니다. LSTM 구현시 해당 레이어를 사용하여 분석을 진행합니다.

ii) 오피니언마이닝을 통한 기사 분석 및 주가 예측

가) BeautifulSoup(기사 크롤링/본문 가져오기)

크롤링이란 해당 페이지 리소스를 이용한 웹 데이터 마이닝 기법으로 웹 상의 다양한 정보를 자동으로 검색하고 색인하기 위해 사용됩니다. HTML 페이지를 가져와 HTML/CSS 등을 파싱하고, 필요한 데이터만 추출합니다. HTML의 태그를 파싱하여 필요한 데이터만 추출할 수 있게 해주는 라이브러리로서 BeautifulSoup가 있습니다.

BeautifulSoup는 pip install을 사용하여 설치가 가능한 파이선 모듈로, import를 통하여 사용합니다. urllib의 .request를 사용하여 원하는 URL의 HTML 소스를 가지고 옵니다. 웹에서 원하는 데이터가 있는 위치의 class나 태그, ID 등을 확인한 후, BeautifulSoup의 find_all() 등의 함수 인자로 넣어 정보를 받습니다.



예를 들어 위의 web으로부터 기사 본문을 가지고 오기 위해 찾은 id는 "articletxt"이므로 이 id를 사용하면 아래의 본문 텍스트를 얻을 수 있습니다.

나) 오피니언마이닝

오피니언마이닝이란, 특정 제품 및 서비스를 좋아하거나 싫어하는 이유를 분석하며 특정 사안에 대해 대중이나 여론이 어떻게 변하는지 확인하는 기법입니다. 긍정, 부정, 중립을 표현하는 단어 정보를 추출하고, 세부 평가 요소와 그것이 가리키는 오피니언의 연결관계를 포함한 문장을 인식하여 긍정, 부정, 중립 표현의 수 등으로 의견을 분류합니다. 이를 통해 해당 문장이 어떤 평가를 나타내고 있는지 사용자의 선호도를 제시하는 방법입니다.

주로 별점 등을 활용하여 별점이 매우 높은 리뷰의 고빈도 단어를 긍정으로, 별점이 매우 낮은 리뷰의 고빈도 단어를 부정 표현으로 추출합니다. 그러나 이러한 활용 데이터나 기존에 구축된 사전 등의 리소스가 없는 경우 수작업을 통해 단어의 정보를 추출합니다.

위의 단계에서 구축된 어휘 정보를 사용하여 규칙기반방법과 통계기반방법을 동시에 활용하여 대량의 레이블이 부착된 학습데이터를 생성하고 다양한 기계학습 알고리즘을 적용하여 오피니언으로 구성된 문장을 인식합니다. 유의미한 문장들을 긍정/부정 표현의 차를 통해 평가합니다.

뉴스와 주가의 관계를 분석할 때, 뉴스의 긍정/부정을 주가를 사용하여 평가하는 것은 전후관계가 맞지 않다고 판단하여 뉴스 자체가 해당 기업에 대해 긍정적인 내용을 담고있는지, 부정적인 내용을 담고있는지로 판단하여 단어를 분류하도록 합니다.

iii) 주식의 종목 선정 – GICS 산업분류

GICS 산업분류란 글로벌 지수산출기관인 S&P와 MSCI가 1999년 공동으로 개발한 증시전용 산업분류체계로서 투자분석, 포트폴리오 및 자산관리에 있어 세계적으로 가장 널리 활용되는 산업분류체계입니다. GICS 산업분류체계는 전세계 산업을 포괄하며 경제섹터(11), 산업군(24), 산업(69), 하위산업(158)의 4단계의 계층구조로 되어있습니다.

본 연구에서는 경제섹터(11)에 해당하는 에너지, 소재, 산업재, 자유소비재, 필수소비재, 건강관리, 금융, 정보기술, 커뮤니케이션서비스, 유틸리티, 부동산의 산업군 분류를 사용합니다.

나. 과제수행 계획

i) 주가의 수치분석

가) 데이터의 수집

학습 set은 2018년 1월 1일부터 2019년 9월 30일까지의 데이터를, test set은 2019년 10월 1일 이후의 데이터를 사용합니다. 해당 시기의 주가정보는 investing.com에서 제공하는 과거 데이터 csv파일을 사용합니다. 각 기업별 종가/시가/고가/저가/거래량/변동%의 정보를 얻을 수 있습니다.

각 기업별 네이버 금융페이지의 기업실적정보를 웹크롤링을 사용하여 가지고옵니다.

나) 전일 주가 정보를 사용한 주가의 분석

우선 주가데이터만을 가지고 ARIMA, Prophet 등의 시계열 분석 모델을 통해 일정기간 동안의 주가 흐름을 예측합니다. 이후 일별 주가데이터와 오피니언마이닝을 통해 나온 일별 기사의 긍부정 결과를 합하여 LSTM, Conv1D 모델들을 통해 2차적으로 주가 흐름을 예측합니다. 이와 해당 회사의 실적정보를 결합하여 거시적인 주가흐름과 예측결과가 더 부합할 수 있도록 조정하는 단계를 추가합니다.

다) 산업군 별 주가의 분석

주가가 내부적인 요소에만 영향을 받는게 아니므로 동일 섹터에 존재하는 회사들의 주가변동 혹은 다른 섹터들의 주가변동에 영향을 받는지, COSPI지수와 같이 전체적인 흐름과 각 회사의 주가흐름이 일치하는 정도 등 주가에 영향을 줄 수 있는 여러가지 외부적인 요소들을 parameter화 시켜 수치분석 모델링 시 입력 parameter로 사용합니다

ii) 오피니언마이닝을 통한 기사 분석

가) 감성 사전에 들어갈 단어 분류하기

선정된 각 종목별로 종목 명이 제목/본문에 포함된 2018년 1월 ~ 2019년 9월의 뉴스 기사를 크롤링으로 가져옵니다. 한국경제, 매일경제, 머니투데이의 세 언론사로부터 기사를 가져오고, 가져온 데이터를 제목/날짜/본문 순으로 나열하여 csv파일로 저장합니다. 본문의 텍스트를 형태소 단위로 분리하여 불필요한 어휘와 의미를 갖지 않은 단음절 체언 및 용언을 제거합니다.

나) 감성 사전 컨셉 구축

본문의 긍/부정 여부를 기반으로 다양한 글들에서 나온 단어 출현 빈도를 기반으로 한 긍/부정 가중치 부여 방식은 본 주제에 선행관계가 맞지 않으므로 다른 방법을 강구합니다.

기본적인 초기 긍/부정적 단어집은 KNU 한국어 감성사전을 사용합니다 . 기본적으로 해당 사전에 등재되어 있는 가중치는 전부 제거하고 부호만 남겨둡니다. 등재된 단어들을 인풋값으로 네이버 국어사전에 검색하여 유의어와 반의어를 크롤링하고 검색 단어를 기준으로 동의어는 기준과 부호를 동일하게, 반의어는 기준과 부호를 반대로 부여하여 단어정보를 추가하는 형식으로 진행합니다.

다) 감정 분석 단계

기사의 분석은 문장 단위로 진행이 되며, 본문 내의 문장별 분석 값의 평균(총 합으로 사용시 기사의 길이에 따라 높은 값이 도출될 수 있음)으로 해당 기사의 긍정수치로 설정합니다. 문장별 분석은 문장 내에 등장하는 단어들의 긍정값 부호를 모두 곱하여 문장의 분석값으로 설정합니다. 예를 들어, "삼성전자는 17일 뉴스룸을 통해 태블릿PC 신제품인 갤럭시 탭 S6 라이트를 공개했다."라는 문장에서 '삼성전자', '신제품', '공개'의 세 단어는 모두 긍정값을 가져 해당 문장은 +1의 값을 갖게 됩니다.

"이번달(1~15일) 삼성전자의 TV·생활가전 매출이 전년 동기 대비 50% 이상 급감했다."

(+1) × (+1) × (-1) = -1

위의 문장은 '삼성전자', '매출', '급감'의 단어 중 '삼성전자'와 '매출'은 긍정, '급감'은 부정의 값을 갖기 때문에 해당 문장은 -1의 값을 갖게 됩니다. 이외 아무 정보도 얻을 수 없는 문장은 0의 값을 갖게 됩니다. 0의 값을 갖는 문장을 제외하고 긍정/부정의 값을 갖는 문장들의 긍정값을 평균내어 기사의 최종적인 긍정수치로 사용합니다.

이후 특정 날짜별 기사들의 긍정수치들을 입력값으로 포함하여 수치 분석을 진행합니다.

5. 추진일정

순번	추진내용		3월	4월	5월	6월	비고
1	데이터 수집	주가 정보 수집		o			
		기업 정보 수집 크롤링					
2		뉴스 기사 크롤링		o			공동
3	수치 분석						
4							
5	오피니언 마이닝 (기사 분석)	데이터 전처리 (형태소 분석)		o			박승혜 이상윤
6		감성사전 구축			o		박승혜 이상윤
		문장별 긍정수치 측정 코드 작성			o		박승혜
		기사별 긍정수치 측정 코드 작성			o		이상윤

※ 문서 작성 시 순서번호는 1. 가. 1) 가) 순으로 기재

[양식 2] 팀구성원 명단

[캡스톤디자인 팀구성표]

■ 과제명

과 제 명	주가 분석 및 종목 감성분석을 통한 종목 추천	참여학기	2020 년 1 학기
팀 명	STOCK	팀 총인원	3 명
대표학생	성명 : 박 승 혜 연락처 : 010-9141-3439	지도교수	이대호 교수님

■ 팀 구성원

구분	구성원 명단					
신청자 1 (대표자)	성 명	박승혜	학 번	2017103081	학년	4
	소속대학	응용과학대학	소속학과	응용수학과		
	휴대전화	010-9141-3439	이 메 일	tmdgp1204@naver.com		
신청자 2	성 명	송재원	학 번	2013103432	학년	4
	소속대학	이과대학	소속학과	수학과		
	휴대전화	010-5898-1596	이 메 일	rearsilre@gmail.com		
신청자 3	성 명	이상윤	학 번	2017110264	학년	4
	소속대학	소프트웨어융합대학	소속학과	소프트웨어융합학과		
	휴대전화	010-6541-5667	이 메 일	syl11121@khu.ac.kr		
신청자 4	성 명		학 번		학년	
	소속대학		소속학과			
	휴대전화		이 메 일			
신청자 5	성 명		학 번		학년	
	소속대학		소속학과			
	휴대전화		이 메 일			