〔별지 제5호서식〕

SW캡스톤디자인 과제 결과보고서

과제분야	☐ 기업연계형 과제 (업체명 : 티쓰리큐(주))						
과제유형	☐ 시제품 개발 및 제작 ☑ 분석, 연구, 실험, 논문 ☐ 디자인 개발 및 제작 ☐ 기타( )						
교과목명	빅데이터컴퓨팅실무(캡스톤디자인)						
과제명	특정 기업 뉴스기사의 평판과 주식 시세의 상관성 및 예측						
팀명	SUNOE						
팀장	노00						
참 여 학 생 명 단							
연번	**소속학과(전공)**	**학번**	**학년**	**성별**	**성명**	**연락처**	**E-mail**
1	인공지능·빅데이터공학과	17155651	3	남성	노00		
2	인공지능·빅데이터공학과	19127029	3	남성	이순주	010-8872-5715	soonju0304@cu.ac.kr
3	인공지능·빅데이터공학과	17155890	3	여성	이00		
4	도서관학과	19132348	3	남성	서00		
5							
6							
7							
8							
9							
10							

집행금액(원)	800,000			
과제수행기간	2021년 09월 ~ 2021년 12월 (4개월)			
참여기업 멘토	**소속**	티쓰리큐(주)	**성명**	서민관
	연락처	02-6344-7660		
교과목 담당교수	**소속**	인공지능·빅데이터공학과	**성명**	이종혁
	연락처	053-850-2882		

상기의 내용과 같이 SW캡스톤디자인 과제 결과보고서를 제출합니다.

별첨 : 1. SW캡스톤디자인 과제 수행 결과보고서 1부

2. SW캡스톤디자인 과제 수행 결과물 1부. 끝.

2021년 12월 17일

과제수행팀 팀장 : (인 또는 서명)

교과목 담당교수 : 이 종 혁 (인 또는 서명)

대구가톨릭대학교 SW중심대학사업단장 귀하

〔별지 제5호서식〕 - 〔별첨 1〕

DCU
DAEGU CATHOLIC UNIVERSITY

SW캡스톤디자인 과제 수행 결과보고서

SW
대구가톨릭대학교
소프트웨어중심대학사업단

1. 과제 수행 결과보고

과 제 명	특정 기업 뉴스기사의 평판과 주식 시세의 상관성 및 예측

□ 과제 개요 및 필요성

- 과제 개요
  - 현 세대에 들어서서 인터넷과 정보통신 기술, 미디어가 발달함에 따라 주식은 그에 대한 지식을 가진 사람들만의 종목이 아닌 남녀노소가 투자할 수 있는 개념으로 확장되었음.
  - 주식투자에 있어서 비전문가들에게 어떤 정보가 좀 더 효율적인 주식투자를 하는데 도움이 되는지 생각해 보았을 때, 실시간으로 방대한 양의 정보가 업데이트 되는 뉴스기사를 분석하여 기업에 대한 사람들의 인식, 평판을 조사한 후 점수화하여 실제 그 점수가 기업의 주가와 얼마나 상관성이 있는지, 나아가 상관성이 있다면 미디어의 평판을 통해 주가예측 또한 가능할 것이라고 생각하여 주가에 대한 더 나은 데이터를 제공하기 위하여 과제를 수행하고자 함.
- 필요성
  - 온라인상의 실제 기업에 대한 사람들의 인식과 주가의 상관관계를 찾을 수 있게 되면 그 정보는 다양한 가능성이 있음.
  - 주가 예측을 통하여 주식투자를 하는데 있어서 사람들에게 유용한 정보를 전달 할 수 있음.
  - 온라인상에서 기업에 대한 사람들의 평판을 점수로 나타내서 한눈에 알아 볼 수 있음.

□ 과제의 개발 방법 및 과제 수행 과정

- 분석을 실시하기 위한 4개의 파일이 존재함.

1. 뉴스기사 본문 데이터 및 주가 데이터 추출을 위한 파일
   - 분석을 위한 특정 기업의 뉴스기사 본문을 수집하고 해당 기업의 10일 전까지의 주식 시세 (종가) 데이터를 불러와 본문 데이터는 각 날짜 별 폴더에 txt 파일로, 주식 시세 데이터는 csv 파일로 저장하는 파일.
   - 웹크롤러를 이용하여 기업명 키워드를 입력하면 오늘부터 10일 전까지의 해당 기업 네이버 뉴스 첫 페이지 10개의 기사 본문 내용을 가져온다.

<기업명을 입력하면 주식 종목명 코드를 반환한다>

```
기업명을 입력하세요 : 삼성전자
삼성전자의 종목명 코드는 005930 입니다.
```

   - 오늘 날짜의 뉴스기사 10개, 어제 뉴스기사 10개 ••• 10일 전 날짜의 뉴스기사 10개의 웹 데이터를 수집하여 총 뉴스기사 100개의 웹 데이터를 수집한다.

<해당 기업의 오늘날짜에 해당하는 뉴스기사 10개 주소>

```
<20211203 의 삼성전자에 대한 뉴스 기사 입니다>
https://www.chosun.com/politics/politics_general/2021/12/03/BAXAIYN6YJCENH3IJO2IXSME6A/?utm_source=naver&
http://news.mt.co.kr/mtview.php?no=2021120310003464068
https://view.asiae.co.kr/article/2021120318191801186
http://news.heraldcorp.com/view.php?ud=20211203000635
http://www.newsis.com/view/?id=NISX20211203_0001674923&cID=10803&pID=14000
https://www.news1.kr/articles/?4512057
https://www.hankyung.com/finance/article/2021120330276
https://biz.chosun.com/it-science/ict/2021/12/03/TN3PTT3M2FEMNDKSR5QZFETQLQ/?utm_source=naver&utm_medium=
http://moneys.mt.co.kr/news/mwView.php?no=2021120217368060053
https://www.etoday.co.kr/news/view/2083872
```

1. 과제 수행 결과보고

과 제 명	특정 기업 뉴스기사의 평판과 주식 시세의 상관성 및 예측

□ 과제의 개발 방법 및 과제 수행 과정

- 수집된 뉴스기사 웹 데이터는 정규표현식을 이용하여 특수문자와 영어를 제외한 한글만을 남기도록 정제한다.
- 정제된 본문 데이터를 저장하기 위하여 뉴스기사가 추출된 날짜를 이름으로 하는 디렉터리를 만든다,
- 각 날짜별 디렉터리에 뉴스기사 본문 데이터를 txt 파일로 저장한다. 그렇게 되면 10일간의 날짜 폴더별로 10개의 txt 파일이 만들어지며, 각 텍스트 파일은 각 뉴스기사의 본문 데이터를 담고 있다.
- 뉴스기사가 작성된 날짜의 주식 시세를 확인하여야 하기 때문에 뉴스기사가 작성된 날짜 리스트를 이용하여 키워드로 입력한 특정 기업의 종목코드를 알아낸 후, 종목코드와 뉴스기사가 작성된 날짜를 주식 시세 정보를 반환하는 함수에 변수로 넣어 날짜별 종가 정보만 인덱싱하여 데이터프레임으로 저장한다.
- 분석 단계에서 변수로 이용하기 위하여 날짜별 종가 테이블을 final_prices.csv 파일로 저장한다.

2. 뉴스기사 본문 데이터 내의 단어 추출 및 리스트 저장

- 뉴스기사가 긍정적인 기사인지 부정적인 기사인지를 판단 및 분석하기 위하여 우선 본문 내의 단어를 추출하여 리스트로 저장해야 함.
- 각 폴더 별 저장된 뉴스기사 본문 데이터를 불러온 후, konlpy 패키지의 twitter 클래스를 이용하여 본문 데이터에서 명사들을 추출한 뒤, 단어 리스트로 저장한다.
- 단어 리스트가 만들어지면 해당 뉴스기사가 작성된 날짜명 디렉터리에 csv 파일로 저장한다.
- 반복문을 이용하여 날짜별 10번을 반복하고 10일간의 뉴스 데이터를 저장하기 위하여 총 100개의 단어 리스트를 생성하게 된다.

3. 각 뉴스기사가 긍정적 내용인지 부정적 내용인지에 대한 점수 부여

- 우선 긍정, 부정 판단을 위하여 웹상에서 한국어 긍/부정어 단어 리스트 자료를 내려 받아 데이터프레임 형식으로 불러온다.
- 긍/부정어 단어 리스트를 이용할 수 있도록 토큰화를 시켜준다.
- 긍/부정어 단어 리스트를 분석에 사용하기 위하여 긍정 단어 리스트는 1, 부정 단어 리스트는 0으로 라벨링을 한 뒤, 긍정어 테이블와 부정어 테이블을 합친다.
- 두 번째 과정(뉴스기사 본문 단어 추출 프로그램)에서 저장된 뉴스기사의 단어 리스트를 불러온다.

1. 과제 수행 결과보고

과 제 명	특정 기업 뉴스기사의 평판과 주식 시세의 상관성 및 예측

□ 과제의 개발 방법 및 과제 수행 과정

- 뉴스기사 점수를 저장할 ArticleScore 딕셔너리에 10개의 원소를 0으로 초기화 한다.
- 반복문을 사용하여 날짜별 10개의 뉴스기사 중 뉴스기사 내 단어 리스트를 긍/부정어 테이블과 비교하여 긍정어 중 일치하는 단어가 있으면 1점을 부여, 부정어 중 일치하는 단어가 있으면 -1점을 부여하여 총 합산된 결과를 ArticleScore 딕셔너리에 점수를 저장한다.
- 날짜별로 뉴스기사 개수에 맞게 10번 반복하면 ArticleScore 딕셔너리 내의 모든 원소에 점수가 매핑되고 해당 딕셔너리를 날짜폴더에 csv파일로 저장한다.
- 위와 같은 과정을 뉴스기사를 추출한 날짜별로 반복하면 모든 폴더 내에 ArticleScore.csv 파일이 만들어지고, 회귀분석을 위한 모든 데이터가 준비된다.

4. 회귀분석

- 회귀분석은 단순선형회귀분석기법을 이용하였다.
- 분석에 필요한 데이터를 불러오기 위하여 뉴스기사 점수는 'articleScore.csv' 파일, 주식 종가는 'final_prices.csv'와 날짜 데이터는 'articleDateList.csv'를 읽어들인다.
- 귀무가설 h0 - "특정 기업에 대한 뉴스기사 평가와 주식시세는 상관이 없다."
  대립가설 h1 - "특정 기업에 대한 뉴스기사 평가에 따라 주식시세가 변한다."
  를 설정한다.
- 뉴스기사 점수에 따른 주식 시세를 분석 및 예측하기 때문에 독립변수를 뉴스기사 점수로 설정, 종속변수를 주식 시세 종가로 둔다.
- 분석을 시작하기 전에 반복문을 사용하여 각 날짜별 뉴스기사 점수의 평균을 독립변수 리스트에 저장한다.
- 주식시세의 종가는 읽기 쉽도록 10,000 대의 자리에서 1000을 나누어준다.
- 뉴스기사 평균점수 리스트와 주식시세 리스트를 합친 테이블을 만들어 회귀분석표를 출력한다.
- 회귀분석표에 나타난 기울기와 절편을 이용하여 회귀모델을 만들어낸다.
- 최종적으로 상관계수와 결정계수, 회귀식을 출력한다.
- 최초 분석했을 당시의 뉴스기사 점수와 주식시세 종가의 상관계수는 0.132로, 양의 상관성을 보였다. 즉, 뉴스기사 점수가 증가함에 따라 주식 시세도 증가하는 추세라는 것을 알 수 있다.
- 회귀식은 Y = 0.011*X + 7.7 으로 뉴스기사 점수가 1점 증가할 때 마다 110원이 증가할 것이라고 예측하였다.
- 두 변수의 상관계수가 양의 값을 나타내므로 뉴스기사와 점수 사이의 상관성은 양의 상관을 보인다 즉, 뉴스기사 점수가 증가함에 따라 주식시세 또한 증가한다고 볼 수 있으며
  대립가설 h1 - "특정 기업에 대한 뉴스기사 평가에 따라 주식시세가 변한다." 를 채택하였다.
  하지만, 결정계수가 충분히 크지 못하므로 모델의 설명력이 높다고 할 수 없으며 데이터의 수 또한 부족하여 정확도가 떨어지는 것으로 판단된다. 그러므로 좀 더 정확한 분석을 하기 위해서는 충분히 많은 데이터를 사용하는 것과 다양한 분석기법을 이용하여 비교하는 것이 좋을 것 이라는 결론을 내렸다.

1. 과제 수행 결과보고

과 제 명	특정 기업 뉴스기사의 평판과 주식 시세의 상관성 및 예측

□ 과제의 개발 방법 및 과제 수행 과정

```
전체 날짜의 평균점수 : [0.0, 0.1, 1.9, 0.4, 0.9, 0.1, 0.4, -1.3, -0.6, -0.7]
   price  score
0   7.56    0.0
1   7.63    0.1
2   7.74    1.9
3   7.74    0.4
4   7.82    0.9
5   7.69    0.1
6   7.68    0.4
7   7.70   -1.3
8   7.76   -0.6
9   7.78   -0.7
```

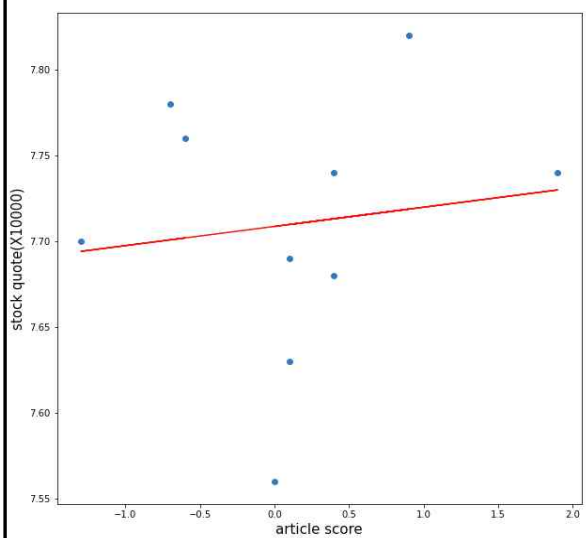
〈변수를 포함한 테이블, 1열은 종가, 2열은 뉴스기사 평균점수〉

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                 -0.106
Method:                 Least Squares   F-statistic:                    0.1411
Date:                Thu, 16 Dec 2021   Prob (F-statistic):              0.717
Time:                        18:06:56   Log-Likelihood:                 12.232
No. Observations:                  10   AIC:                            -20.46
Df Residuals:                       8   BIC:                            -19.86
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.7087      0.025    303.158      0.000       7.650       7.767
score          0.0112      0.030      0.376      0.717      -0.057       0.080
==============================================================================
Omnibus:                        1.280   Durbin-Watson:                   0.753
Prob(Omnibus):                  0.527   Jarque-Bera (JB):                0.601
Skew:                          -0.578   Prob(JB):                        0.740
Kurtosis:                       2.671   Cond. No.                         1.24
==============================================================================
```

〈점수에 따른 종가 변화율의 회귀분석 테이블〉

```
 두 변수의 pearson 상관계수 :  0.132

기울기 : 0.011

절편 : 7.709

추정된 회귀식 : Y = 0.011 * X + 7.709

결정계수 :  0.0

뉴스기사의 점수가 10 점이라면 예측되는 주식 시세는 78204.528 원 입니다.
```

〈분석을 통하여 도출된 값〉



〈추정된 회귀 직선〉

2. 활용방안 및 기대효과

□ 활용방안

- 뉴스기사 데이터를 정리하여 단어별로 나열해 놓은 것이기 때문에 해당 기사가 특정 기업에 대하여 어떤 반응을 보이고 있는지 파악할 수 있다.
- 해당 프로그램을 이용하여 특정 기업의 현재 평판을 쉽게 알아내고 주식 매매를 할 시에 도움을 줄 수 있다.
- 키워드를 바꿔서 한 대상이 아닌 여러 대상의 기업에 대하여 분석할 수 있기 때문에 기업들 사이의 주가를 비교할 수 있고 여러 기업의 평판을 분석할 수 있다.

□ 기대효과

- 우선 특정 기업의 뉴스기사를 데이터로 사용하기 때문에 기업의 평판을 보기 쉽게 정리하여 알아낼 수 있으므로 뉴스를 읽지 않고도 해당 기업의 평판이 어떠한지 유추해 볼 수 있음.
- 뉴스기사와 주가사이의 상관성이 존재하기 때문에 뉴스기사 데이터를 이용하여 주식투자를 할 때 도움이 될 수 있음.
- 특정 기업의 과거 평판과 점수의 증감률을 보고 해당 기업에 대하여 사람들이 어떻게 평가하고 있는지, 앞으로 어떻게 평가될 것인지를 예상할 수 있음.

□ 기타

- 프로젝트를 진행하고 완성하며 놓쳤던 부분과 개선사항에 대한 내용
- 뉴스기사 점수에 따른 주식의 변화를 분석하기 위해서는 주식의 종가가 아닌 주식 상승률을 종속변수로 하였으면 좀 더 적합한 모델이 나왔을 것이다.
- 많은 양의 뉴스기사를 데이터로 추출하게 된다면, 분명 중복되는 뉴스기사가 있을 것이며 그러한 중복데이터를 처리하는 방안을 생각해 봐야 한다.
- 최근의 데이터만으로 주식의 변동률을 예측하는 것은 정확하지 않으므로 더욱 많은 데이터와 데이터의 날짜 스펙트럼을 넓게 잡아야 조금 더 정확한 예측이 가능할 것 같다.
- 뉴스기사에 긍정점수와 부정점수를 부여할 때, 단어의 빈도수와 더불어 긍정점수와 부정점수의 강도를 조절하여 부여하였다면 조금 더 넓은 범위의 점수가 도출 될 수 있었을 것이다.
- 긍정적인 데이터가 연속으로 나온다면 해당 기업에 대한 평가가 대부분 좋은 것이므로 추가적인 점수를 부여하는 방법이 있었다면 뉴스기사 점수를 이용한 예측의 정확도가 오를 수 있을 것 같다.

3. 예산 집행내역

구분	사용항목	구입물품	사용목적(상세기재)	금액(원)
기술지도비	멘토링	멘토링 기술지도	과제수행을 위한 멘토 및 전문가 기술지도	400,000
과제운영비	재료구입	AWS(Amazon Web Services)	과제수행을 위한 분석용 클라우드 서버 구입	400,000
합 계				800,000

※ 중간보고 시 제출한 집행내역을 포함한 전체 예산 집행내역 작성

〔별지 제6호서식〕

DCU DAEGU CATHOLIC UNIVERSITY

SW캡스톤디자인 만족도 조사서

SW 대구가톨릭대학교 소프트웨어중심대학사업단

SW캡스톤다자인 교과목 교육과정 및 지원에 대한 학생 여러분의 만족도를 조사하고자 합니다.
평가의 내용은 이후 실시되는 교육과정에 반영되어, 더 내실 있는 교육과정이 진행되는데 참고자료가 될 것입니다.

교과목명	빅데이터컴퓨팅실무	학 과	인공지능빅데이터공학과
성 별	남	학 년	3

1. SW캡스톤디자인 교과목 수행평가 (체크해 주세요.)

평가내용	매우 그렇다	그렇다	보통이다	아니다	매우 아니다
	5	4	3	2	1
교과목에 대한 취지 및 의도를 이해하고 있는가?	●				
과제의 목표가 과제 내용과 적합한가?		●			
팀원의 임무가 적절히 분배되고, 수행되었는가?		●			
시간 내에 목표하던 일들이 완료되었는가?	●				
과제의 내용이 다양한 관점에서 분석되었는가?		●			
과제 수행 시 문제점은 정확히 파악되었는가?	●				
과제에 대한 문제점 해결의 대안은 적절히 제시되었는가?	●				
과제 결과물에 경제적, 기술적 측면의 결론이 포함되어 있는가?	●				
합 계	5	3	0	0	0

2. SW캡스톤디자인 교과목 지원 만족도 평가 (체크해 주세요.)

평가내용	매우 그렇다	그렇다	보통이다	아니다	매우 아니다
	5	4	3	2	1
교과목 참여로 인해 문제 해결 능력향상 및 자기계발에 도움이 되었다.	●				
교과목 과제 지원비가 적당하였다.	●				
과제 수행 기간은 적당하였다.			●		
SW캡스톤디자인 교과목 수강을 다른 사람에게 권유하겠다.		●			
기회가 된다면 SW캡스톤디자인 교과목을 다시 수강할 의사가 있다.		●			
합 계	2	2	1	0	0

3. 기타 건의사항을 자유롭게 기재해 주세요.