금융 빅데이터 전문가 1기

연예뉴스와 주가와의 관계분석

- YG 엔터테인먼트의 뉴스 감성분석

조원 : 홍창수(조장)
이종석
김진수
최용준
진정화

목 차

1. 개요 및 연구 동기

개요 및 연구 동기

- 뉴스 감성분석을 통한 개별주가와의 상관관계 분석

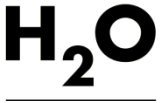
  : 연예뉴스와 엔터테인먼트 주가와의 관계 분석

- 엔터테인먼트 주가: YG엔터테인먼트에 대해 연구
- 주요분석 절차 : 뉴스 스크랩핑, 형태소 분석, 감성 사전 구축, 감성분석, YG 주가와의 상관관계 분석 및 추가분석 진행

프로젝트에 사용된 도구

- 뉴스 데이터 크롤링

- 텍스트 분석 및 통계 분석

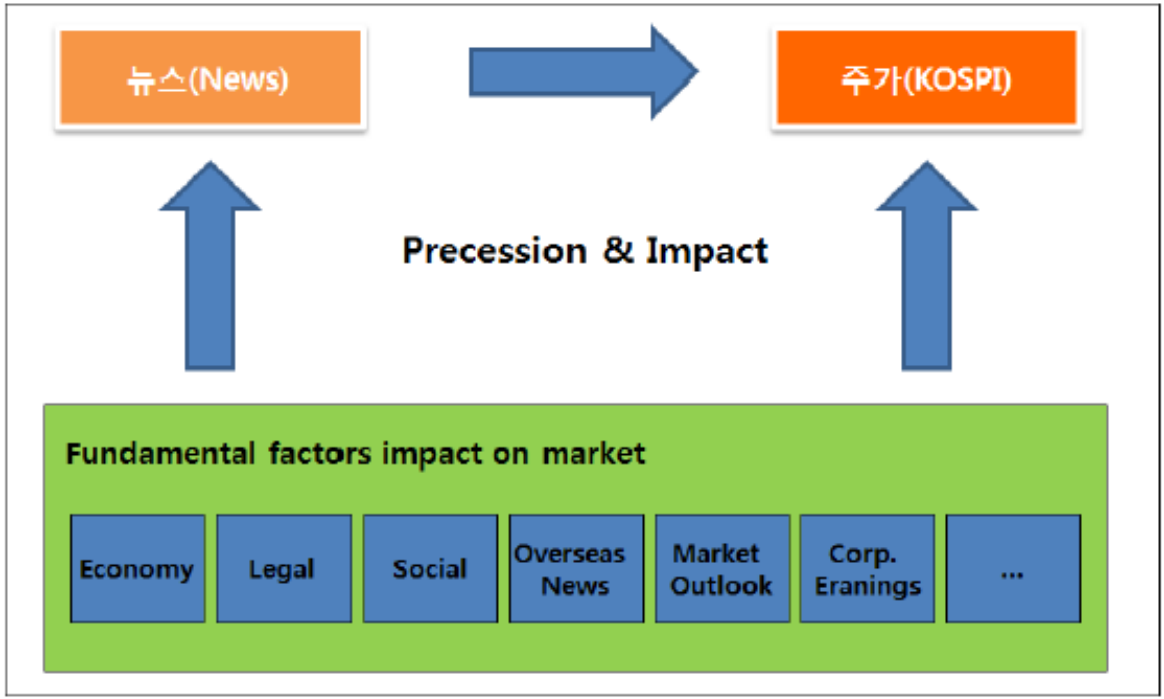


- 예비분석(EDA) 및 추가분석

- 보고 및 시각화

2. 관련 연구

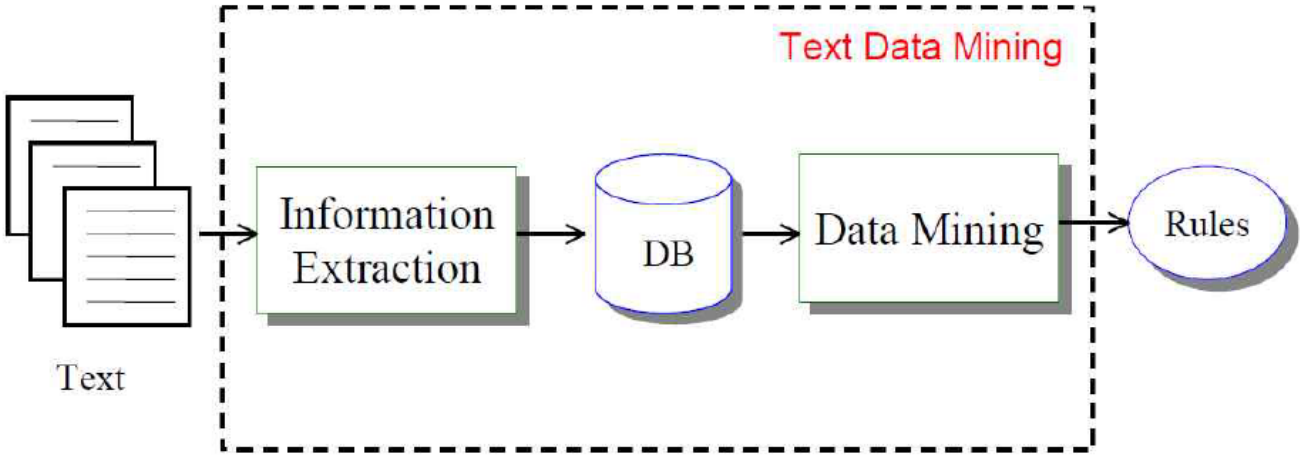
1.1 뉴스와 주가와의 관계(영향요인)



출처: 김유신 외(2011)

- 뉴스 데이터가 주식의 가격 및 거래량에 영향 미침

1.2 텍스트 마이닝과 정보추출

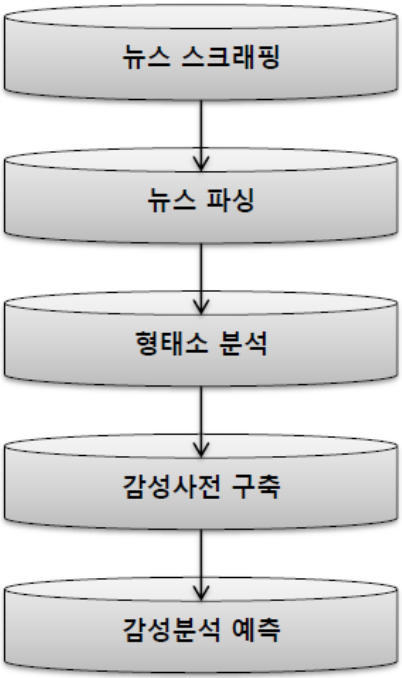


출처: 김유신 외 (2011)

- 인터넷과 데이터베이스 발달로 오피니언 마이닝 연구 활발
- 오피니언 마이닝: 문서의 의견정보 추출 분석(뉴스, 영화평 등)

3. 데이터 수집 및 연구방법

3.1 데이터수집 및 분석절차



- 네이버 금융뉴스: 스크랩핑
- 스포츠경제: 파싱(연예뉴스)
- 행태소 분석 : 명사 추출
- 감성사전구축 : 긍정/부정어
- 감성분석: 긍/부정 횟수와 주가

3.2 뉴스데이터 크롤링

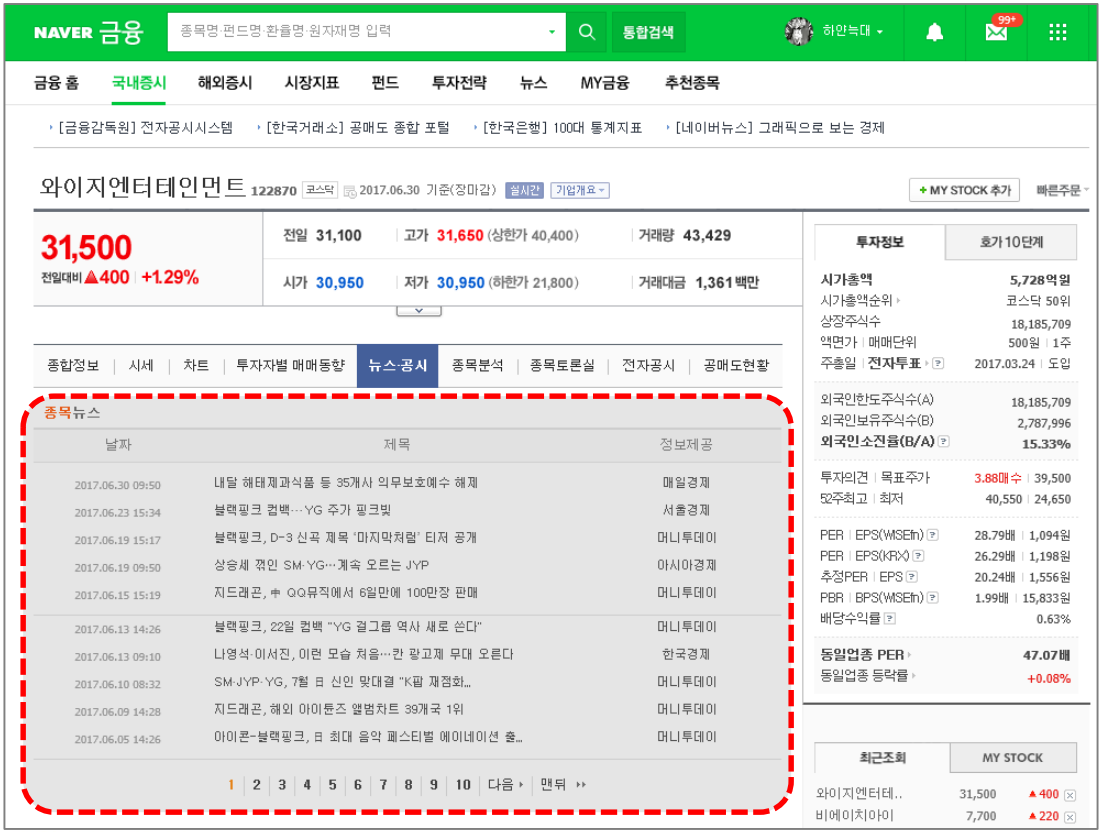
- 분석을 위한 엔터테인먼트 뉴스정보 수집

수집1. NAVER 금융의 종목뉴스 크롤링
cf. item_code = '122870' #

수집2. 연예뉴스 사이트의 주요키워드 선별후,
뉴스검색 크롤링

3.2.1 NAVER 금융의 종목뉴스 크롤링

- 와이지엔터테인먼트 종목뉴스 화면



3.2.1 NAVER 금융의 종목뉴스 크롤링

- 와이지엔터테인먼트 종목뉴스 크롤링 소스코드

```
# 종목별 코드
item_code = '041510'    # 에스엠
item_code = '035900'    # JYP Ent.
item_code = '122870'    # 와이지엔터테인먼트


def getScopePage(item_code):
    """종목별 뉴스범위(시작/끝페이지)"""

    nmenu_finance = 'http://finance.naver.com'
    sub_tap_05 = nmenu_finance + '/item/news_news.nhn'

    url = sub_tap_05 + '?code=' + item_code
    html = urlopen(url)
    soup = BeautifulSoup(html, "lxml")

    tmp = soup.find_all('td', 'pgRR')
    tmp2 = tmp[0].find_all('a')
    tmp3 = tmp2[0]['href']
    tmp4 = tmp3[tmp3.find("page="):]
    e_page = int(re.findall('[0-9]+', tmp4)[0])

    return s_page, e_page
```

```
def getArticleList(item_code, page=1):
    """종목별 해당페이지 기사리스트"""
    url = sub_tap_05 + '?code=' + item_code + '&page=' + str(page)
    html = urlopen(url)
    soup = BeautifulSoup(html, "lxml")
    # print('url:', url)

    article_list = soup.find_all('tr')
    article_su = len(article_list)

    s_line = 2  # empty, column명 제외
    e_line = article_su

    articles = range(s_line, e_line)
    page_df = pd.DataFrame()

    for num in articles:
        tmp = article_list[num].find_all('td')

        if len(tmp) > 0:
            tmp1 = tmp[0].find_all('span')
        else:
            continue

        if len(tmp1) > 0:
            a_dttm = tmp1[0].get_text()
            a_date = re.findall('[0-9]+.[0-9]+.[0-9]+', a_dttm)[0]
            a_time = re.findall('[0-9]+:[0-9]+', a_dttm)[0]

            tmp1 = tmp[1].find_all('a')
            a_title = tmp1[0].get_text()

            a_link = tmp1[0]['href']
            if a_link.startswith('http'):
                pass
            else:
                a_link = nmenu_finance + a_link

            a_media = tmp[2].get_text()

    column_nm = ['Date', 'Time', 'Title', 'Media', 'Link']
    page_df = pd.DataFrame(page_dict, columns=column_nm)

    return page_df
```

3.2.1 NAVER 금융의 종목뉴스 크롤링

- 와이지엔터테인먼트 종목뉴스 크롤링 소스코드 2

```
def getTermList(item_code, total_df, date1, date2):
    """기간별 기사리스트"""

    s_news_df = total_df.sort_values(by=['Date', 'Time'], ascending=[False, False])

    save_file = 'data/s_news_' + item_code + '.csv'
    s_news_df.to_csv(save_file, sep=',', encoding='UTF-8')

    rnews_df = pd.read_csv(save_file)
    del rnews_df['Unnamed: 0']

    today = datetime.today()
    init_date = '2000.01.01'
    last_date = today.strftime('%Y.%m.%d')

    s_date = init_date if date1 == None  else date1
    e_date = last_date if date2 == None  else date2

    snews_df = rnews_df[(s_date <= rnews_df['Date']) & (rne

    snews_df.to_csv(save_file, sep=',', encoding='UTF-8')

    return snews_df
```

```
def getNewsList(item_code, start_dt=None, end_dt=None):
    """종목별 전체뉴스 기사리스트"""

    news_df = pd.DataFrame()
    snews_df = pd.DataFrame()
    total_df = pd.DataFrame()
    start_page, end_page = getScopePage(item_code)

    for page in range(start_page, end_page + 1):
        page_df = getArticleList(item_code, page)

        if page == 1:
            total_df = page_df
        else:
            total_df = pd.concat([total_df, page_df], axis=0)

    snews_df = getTermList(item_code, total_df, start_dt, end_dt)
    news_df = getFinalList(item_code, snews_df)
    print('Naver 증권뉴스 수집이 완료되었습니다!')

    return snews_df
```

3.2.1 NAVER 금융의 종목뉴스 크롤링

- 와이지엔터테인먼트 종목뉴스 크롤링 데이터 (=10,000건)

```
In [75]: snews_df = getNewsList(item_code)
```

```
1706 http://finance.naver.com/item/news_read.nhn?article_id=0000550999&office_id=008&code=035900&page=171
1707 http://finance.naver.com/item/news_read.nhn?article_id=0000275709&office_id=018&code=035900&page=171
1708 http://finance.naver.com/item/news_read.nhn?article_id=0000536922&office_id=008&code=035900&page=171
1709 http://finance.naver.com/item/news_read.nhn?article_id=0000533700&office_id=008&code=035900&page=171
1710 http://finance.naver.com/item/news_read.nhn?article_id=0000528045&office_id=008&code=035900&page=172
1711 http://finance.naver.com/item/news_read.nhn?article_id=0000268605&office_id=018&code=035900&page=172
1712 http://finance.naver.com/item/news_read.nhn?article_id=0000268036&office_id=018&code=035900&page=172
1713 http://finance.naver.com/item/news_read.nhn?article_id=0000523517&office_id=008&code=035900&page=172
1714 http://finance.naver.com/item/news_read.nhn?article_id=0000265323&office_id=018&code=035900&page=172
1715 http://finance.naver.com/item/news_read.nhn?article_id=0000523398&office_id=008&code=035900&page=172
1716 http://finance.naver.com/item/news_read.nhn?article_id=0000522840&office_id=008&code=035900&page=172
1717 http://finance.naver.com/item/news_read.nhn?article_id=0000262786&office_id=018&code=035900&page=172
1718 http://finance.naver.com/item/news_read.nhn?article_id=0000258158&office_id=018&code=035900&page=172
1719 http://finance.naver.com/item/news_read.nhn?article_id=0000514638&office_id=008&code=035900&page=172
1720 http://finance.naver.com/item/news_read.nhn?article_id=0000513713&office_id=008&code=035900&page=173
1721 http://finance.naver.com/item/news_read.nhn?article_id=0000249189&office_id=018&code=035900&page=173
1722 http://finance.naver.com/item/news_read.nhn?article_id=0000503393&office_id=008&code=035900&page=173
1723 http://finance.naver.com/item/news_read.nhn?article_id=0000498213&office_id=008&code=035900&page=173
1724 http://finance.naver.com/item/news_read.nhn?article_id=0000244890&office_id=018&code=035900&page=173
Naver 증권뉴스 수집이 완료되었습니다!
```

```
In [77]: snews_df.head()
```

Out[77]:

	Date	Time	Media	Link	NewsTitle	NewsArticle
0	2017.06.10	08:32	머니투데이	http://finance.naver.com/item/news_read.nhn?ar...	SM·JYP·YG, 7월 日 신인 맞대결 "K팝 재점화 나선다"	[머니투데이 김건우 기자] [걸그룹 트와이스를 시작으로 블랙핑크 등 일본 데...
1	2017.06.02	08:42	머니투데이	http://finance.naver.com/item/news_read.nhn?ar...	장 마감 후 주요 공시-1일	[머니투데이 박계현 기자] -능심은 신준호 회장이 보유하던 회사 주식 10만...
2	2017.06.02	04:12	한국경제	http://finance.naver.com/item/news_read.nhn?ar...	[코스닥 기업공시] (1일) 기가레인 등	▲기가레인=최대주주가 김정곤 외 12명에서 케플러밸류파트너스 외 12명으로 ...
3	2017.06.01	17:33	이데일리	http://finance.naver.com/item/news_read.nhn?ar...	JYP Ent.,서울 강동구 토지 및 건물 취득	[이데일리 최정희 기자] JYP Ent.(035900)는 서울시 강동구 성내...
4	2017.06.01	16:57	머니투데이	http://finance.naver.com/item/news_read.nhn?ar...	JYP엔터, 202억원에 서울 강동구 토지·건물 매입…사옥 확보 목적	[머니투데이 김지훈 기자] 제이와이피엔터테인먼트는 서울시 강동구 성내동 44...

3.2.2 연예뉴스 주요키워드별 뉴스검색

- 연예전문미디어 엔터테인먼트 기사 화면

리스트 페이지

상세 페이지

3.2.2 연예뉴스 주요키워드별 뉴스검색

- 연예전문미디어 엔터테인먼트 기사 크롤링 소스코드

```
def getPageList(page=1):
    page_df = pd.DataFrame()

    site = 'http://www.sporbiz.co.kr/news/'
    sect_code = 'S1N1'   # 연예
    view_type = 'sm'
    order_by = 'I'
    sc_word = 'PSY'   # 싸이, PSY

    url = site +
'articleList.html?page={page}&total={total}&sc_word={sc_word}&sc_section_co
ode}&sc_order_by={order_by}&view_type={view_type}'.format(
        page=str(page),
        total=total,
        sc_word=sc_word,
        sect_code=sect_code,
        view_type=view_type,
        order_by=order_by
    )

    html = urlopen(url)
    soup = BeautifulSoup(html, "lxml")

    page_list = soup.find_all('td', 'ArtList_Title')
    list_cnt = len(page_list)

    for num in range(list_cnt):
        news = page_list[num]
        seq = (page - 1) * page_cnt + num + 1
        # print(seq)

        tmp2 = news.getText()
        tmp3 = site + news.find('a')['href']

        view_html = urlopen(tmp3)
        view_soup = BeautifulSoup(view_html, "lxml")

        tmp4 = view_soup.find('span', 'sec1')
        tmp4 = '' if tmp4 == None else tmp4.getText()

        tmp5 = view_soup.find('span', 'sec2')
        tmp5 = '' if tmp5 == None else tmp5.getText()

        tmp6 = view_soup.find('div', 'headline')
        tmp6 = tmp6.find('h3').getText()

        tmp7 = view_soup.find('span', 'info')
        tmp7_1 = tmp7.find_all('em')[0].getText()
        tmp7_2 = tmp7.find_all('span')[1].getText()
```

```
        tmp8 = view_soup.find('div', 'art_body')
        tmp8 = tmp8.getText()
        tmp8 = str.replace(tmp8, 'Wr', '')
        tmp8 = str.replace(tmp8, 'Wn', '')
        comment1, comment2 = tmp8.find('<!--'), tmp8.find('-->')
        if comment1 > 1:
            tmp8 = tmp8[:comment1]

        news_seq.append(seq)
        news_title.append(tmp2)
        news_link.append(tmp3)
        news_sec1.append(tmp4)
        news_sec2.append(tmp5)
        news_headline.append(tmp6)
        news_reporter.append(tmp7_1)
        news_date.append(tmp7_2)
        news_text.append(tmp8)

    page_dict = dict()
    page_dict['Seq'] = news_seq
    page_dict['Title'] = news_title
    page_dict['Link'] = news_link
    page_dict['Sec1'] = news_sec1
    page_dict['Sec2'] = news_sec2
    page_dict['Headline'] = news_headline
    page_dict['Reporter'] = news_reporter
    page_dict['Date'] = news_date
    page_dict['Text'] = news_text

    column_name = ['Seq', 'Title', 'Link', 'Sec1', 'Sec2', 'Headline', 'Reporter',
'Date', 'Text']
    page_df = pd.DataFrame(page_dict, columns=column_name)

    return page_df
```

3.2.2 연예뉴스 주요키워드별 뉴스검색

- 연예전문미디어 엔터테인먼트 기사 크롤링 소스코드 2

```
def getScopePageList(start_page, end_page):
    news_df = pd.DataFrame()

    for page in range(start_page, end_page + 1):
        page_df = getPageList(page)

        if page == start_page:
            news_df = page_df
        else:
            news_df = pd.concat([news_df, pag

        if page % 10 == 0:
            time.sleep(1)

    return news_df
```

```
def getNewsList(news_max=200):
    news_df = pd.DataFrame()

    page_cnt = 20
    start_page = 1
    end_page = int((news_max - 1) / page_cnt) + 1

    news_df = getScopePageList(start_page, end_page)
    news_df.set_index('Seq', inplace=True)
    del news_df['Title']

    today = datetime.today()
    today = today.strftime('%Y%m%d%H%M%S')
    save_file = 'data/news_' + str(end_page * page_cnt) + '_' + today + '.csv'
    news_df.to_csv(save_file, sep=',')
    # news_df.to_csv(save_file, sep=',', encoding='UTF-8')
    print('Success News Crawling !!')

    return news_df
```

3.2.2 연예뉴스 주요키워드별 뉴스검색

- 연예전문미디어 엔터테인먼트 기사 크롤링 데이터

```
In [12]: news_df = getNewsList(12501, 12600)
         news_df.tail()

626 http://www.sporbiz.co.kr/news/articleList.html?page=626&total=&sc_section_code=S1N1&sc_order_by=I&view_type=sm
627 http://www.sporbiz.co.kr/news/articleList.html?page=627&total=&sc_section_code=S1N1&sc_order_by=I&view_type=sm
628 http://www.sporbiz.co.kr/news/articleList.html?page=628&total=&sc_section_code=S1N1&sc_order_by=I&view_type=sm
629 http://www.sporbiz.co.kr/news/articleList.html?page=629&total=&sc_section_code=S1N1&sc_order_by=I&view_type=sm
630 http://www.sporbiz.co.kr/news/articleList.html?page=630&total=&sc_section_code=S1N1&sc_order_by=I&view_type=sm
Success News Crawling !!
```

Out[12]:

Seq	Link	Sec1	Sec2	Headline	Reporter	Date	Text
12596	http://www.sporbiz.co.kr/news/articleView.html...	연예	가요/음반	인피니트 월드투어, 예매 시작 동시 20만명 몰려	심재걸 기자	2016.01.20	그룹 인피니트가 월드투어 서울 앵콜 공연을 12분 만에 매진시켰다.19일 오후 8시...
12597	http://www.sporbiz.co.kr/news/articleView.html...	연예	가요/음반	'강남스타일'뮤비, 25억뷰 돌파 대기록	심재걸 기자	2016.01.20	싸이의 '강남스타일' 뮤직비디오가 신기록 행진을 멈추지 않고 있다.20일 유튜브에서...
12598	http://www.sporbiz.co.kr/news/articleView.html...	연예	엔터테인먼트	[스타분석]서강준, 물오른 4년차 '치인트'로 폭발	황지영 기자	2016.01.20	▲ 서강준 인스타그램배우 서강준을 향한 연기 호평이 이어지고 있다.분명한 것은 하루...
12599	http://www.sporbiz.co.kr/news/articleView.html...	연예	트렌드 기획	[트렌드 기획] 설현 독주? '좌시하지 않겠다'	심재걸 기자	2016.01.20	설현의 독주를 마냥 지켜만 보지 않을 세력이 있다. '국민 여동생' 타이틀로 여풍을...
12600	http://www.sporbiz.co.kr/news/articleView.html...	연예	트렌드 기획	[트렌드 기획] 지금은 '설현'시대	심재걸 기자	2016.01.20	'설현'시대가 도래했다. 대세는 설현이다. 길거리에는 설현의 전신 사진이 블록마다 ...

YG엔터테인먼트 소속 연예인 관련 키워드 크롤링 후 데이터 저장

4. 실증 분석

4.1 단어 추출 및 긍/부정 사전

YG엔터테인먼트 연예기사 크롤링 한 기사에서 명사 추출

YG엔터테인먼트 연예기사 크롤링 1,800 여건

Seq	Date	Link	Sec1	Sec2	Headline	Reporter	Text
1	2017.06.20	http://www	연예	엔터테인먼	'10살 연하	김은혜 기자	▲ 에이미/사진=연합뉴스[한국스포츠경제 김은혜]?방송인 에이미가 자살 시도를 한 것
2	2017.06.20	http://www	연예	가요/음반	크러쉬X백	정진영 기자	[한국스포츠경제 정진영]?'세이팝 알쌈 콘서트'의 세트리스트 일부가 공개됐다.공연 주최
3	2017.06.20	http://www	연예	엔터테인먼	에비뉴535	정진영 기자	[한국스포츠경제 정진영]?클럽 앤써가 클럽, 힙합, 디제잉 등 다양한 문화가 공존하는 에
4	2017.06.20	http://www	연예	가요/음반	'팝의 황제	정진영 기자	[한국스포츠경제 정진영]?'월간 팝스인멜론' 에 고(故) 마이클 잭슨이 이름을 올렸다.뮤직
5	2017.06.20	http://www	연예	엔터테인먼	설현, "박보	이성봉 기자	▲ 설현이 박보검과 함께 찍은 사진을 소셜미디어에 게재했다./사진=설현 인스타그램[한
6	2017.06.20	http://www	연예	TV	심은하 시(	김지영 기자	▲ 심은하 / 사진=연합뉴스[한국스포츠경제 김지영]?심은하와 시어머니의 관계에 대한
7	2017.06.20	http://www	연예	TV	"데뷔의 주	김미유 기자	[한국스포츠경제 김미유]?지난 주 16일 케이블 채널 엠넷의 <프로듀스 101 시즌2>이 성
8	2017.06.20	http://www	연예	엔터테인먼	'워너원' 라	이성봉 기자	▲ '워너원' 강다니엘(아래), 라이관린/사진=Mnet '프로듀스101 시즌2' 공식 페이스북[한
9	2017.06.20	http://www	연예	엔터테인먼	하민호 "유	김지영 기자	▲ 방송에서 유선호 연습생을 비주얼 멤버로 뽑는 하민호 / 사진=Mnet '프로듀스101 시
10	2017.06.20	http://www	연예	이런씨네	[이런씨네]	양지원 기자	'박열' 리뷰[한국스포츠경제 양지원]?영화 '박열'(28일 개봉)은 우리가 익히 알고 있는 독
11	2017.06.20	http://www	연예	엔터테인먼	정혜성, 남	이현아 기자	[한국스포츠경제 이현아]?배우 정혜성이 우유 광고 모델로 발탁됐다.정혜성은 남양유업
12	2017.06.20	http://www	연예	엔터테인먼	오상진, 국	이현아 기자	[한국스포츠경제 이현아]?방송인 오상진이 전세계 대학생으로 대상으로 영어 토론대회를
13	2017.06.20	http://www	연예	TV	OCN '구해	이현아 기자	[한국스포츠경제 이현아]?OCN 오리지널 드라마 '구해줘'가 첫 티저 영상을 공개했다.'구
14	2017.06.20	http://www	연예	TV	박하나, '이	이현아 기자	[한국스포츠경제 이현아]?박하나가 MBN 미식프로그램 '이맛이야' MC로 전격 합류했다.
15	2017.06.20	http://www	연예	엔터테인먼	'리얼' 김수	이성봉 기자	▲ 영화 '리얼' 배우 김수현/사진=키이스트 공식 인스타그램[한국스포츠경제 이성봉] 배
16	2017.06.20	http://www	연예	영화 뒷담화	[영화 뒷담	이성봉 기자	▲ 사진 = 영화 '꿈의 제인'[한국스포츠경제 이성봉]?"이상과 현실 차이를 느낀 사례를 스
17	2017.06.20	http://www	연예	TV	윤두준-용	이현아 기자	[한국스포츠경제 이현아]?그룹 하이라이트의 윤두준과 용준형이 데뷔 이래 처음으로 상
18	2017.06.20	http://www	연예	엔터테인먼	'프듀101'	최지윤 기자	▲ '프듀101' 시즌2 사무엘[한국스포츠경제 최지윤] '프로듀스101' 시즌2 김사무엘이 사
19	2017.06.20	http://www	연예	TV	'라디오스타	이현아 기자	[한국스포츠경제 이현아]?모델 이소라가 과거 수영복 차림으로 관광객들 앞에서 그야말
20	2017.06.20	http://www	연예	영화	[E-핫스팟]	양지원 기자	[한국스포츠경제 양지원] 올해로 첫 회를 맞이한 한중국제영화제 측이 남다른 각오를 전
21	2017.06.20	http://www	연예	엔터테인먼	설리의 알	김의기 기자	▲ 죽어가는 장어 촬영한 가수 겸 배우 설리/사진=설리 인스타그램[한국스포츠경제 김

명사 추출

```
library(KoNLP)
# 1. 텍스트 파일 가져오기#marketing =
file("D:\\bigdata\\test\\test2.txt", encoding="UTF-8")
....

# -- 2. 폴 단위 단어 추출
lword = Map(extractNoun, marketing2)
length(lword)
lword = unique(lword)
length(lword

lword = sapply(lword, unique
length(lword
str(lword) lword

## list 데이터가공
install.packages("data.table")
require(data.table)
head(rbindlist(data))

l <- list()
list <- list()
for(i in 1:105){
  for (j in 1:length(lword[[i]])){
      list[[j]] <-  data.frame(기사 =i,단어순=j,단어=lword[[i]][j])
    }
  ...
```

4.1 단어 추출 및 긍/부정 사전

추출 된 명사를 기반으로 긍정, 부정 사전을 구축

긍정	부정
1 사랑	1 횡포
2 기약	2 성폭행
3 친분	3 종영
4 신화	4 열애설
5 뮤직비디오	5 추문
6 프로젝트	6 대작
7 축제	7 실괘
8 본상	8 논란
9 음원차트	9 일베
10 선정	10 하락
11 평정	11 입대
12 팬미팅	12 취소
13 단독 콘서트	13 몰락
14 광고왕	14 분계
15 1위	15 한탄
16 음원	16 퇴출
17 스포트라이트	17 강제집행
18 쇼케이스	18 파장
19 화두	19 하차
20 한솥밥	20 교통사고
21 최고	21 공포감
22 출연	22 집행
23 신기록	23 성추행
24 행운	24 혼선
25 라인업	25 곤욕
26 괴물	26 신음
27 빌보드	27 욕설

1,200여건 600여건

4.2 일자별 감성점수 집계 및 코스닥 정보 가져오기

```
# ------------------------------
# -- 4) 감성 분석을 위한 긍/부정 카운트 집계
# ------------------------------
result=sentimental(data[,2], posDic.final, negDic.final)
result = cbind(result, date_tx)   # 날짜추가

#-------------날짜별 점수 집합계-----------------
result$score <- as.numeric(result$score)

result <- aggregate(result$score, by=list(date_tx=result$date_tx), FUN=sum)
#-----------------------------------------------------

sentimental = function(sentences, posDic, negDic){

   scores = laply(sentences, function(sentence, posDic, negDic) {

      sentence = gsub('[[:punct:]]', '', sentence) # 문장부호 제거
      sentence = gsub('[[:cntrl:]]', '', sentence) # 특수문자 제거
      sentence = gsub('\\d+', '', sentence)        # 숫자 제거
      sentence = tolower(sentence
      word.list = str_split(sentence, '\\s+')
      words = unlist(word.list

      pos.matches = match(words, posDic)
      neg.matches = match(words, negDic)
      pos.matches = !is.na(pos.matches
      neg.matches = !is.na(neg.matches)
      score = sum(pos.matches) - sum(neg.matches)  # 긍정 - 부정
      return(score)
   }, posDic, negDic)

   scores.df = data.frame(score=scores, text=sentences)
    return(scores.df)
}
```

```
# YG엔터 주식 가져오기
output <- g.hist2("KOSDAQ:122870", "2015-03-23", "2017-06-22")

#------------데이타 조인(감정점수집계와 테이블조인--------------
join_result <- join(result, output, type="left")

write.xlsx(join_result, "D:/bigdata/프로젝트/YG_join.xlsx")

g.hist2 <- function(ticker, start_date, end_date){

   ticker_m <- gsub(":","%3A",ticker)

   lct<-Sys.getlocale("LC_TIME") #set locale
   Sys.setlocale("LC_TIME", "C")

   start_date = as.Date(start_date) #set date
   end_date = as.Date(end_date)

   fmtstr <- "%b+%d%%2C+%Y"  #ex) Jan+01,+2017
   start_date_fmt <- as.character(format(start_date, fmtstr))
   end_date_fmt <- as.character(format(end_date, fmtstr))

   url = 'http://www.google.com/finance/historical?q='
   url =
paste(url,ticker_m,'&startdate=',start_date_fmt,'&enddate=',end_date_fmt,'&num=200',sep='')

   output <- as.data.frame(readHTMLTable(url,stringsAsFactors=FALSE)[4])
   output[, 1] <- as.Date(output[, 1], format="%b %d, %Y")
   names(output) <- c("date", "Open","High","Low","Close","Volume")
   for (k in 2:6){
      output[, k] <- as.numeric(gsub(",", "", output[, k]))
   }
   end <- as.Date(output[length(output[, 1]), 1])
   ll <- 200 - length(output[, 1])
   if ((end-ll)<start_date) end <- start_date
   if (end > start_date) {
      output <- rbind(output, g.hist2(ticker,start_date,as.character(end-1)))
   }
    return(output)
}
```

기초 테이블 생성 (xlsx)

date	score	Open	High	Low	Close	Volume
2015-03-23	9	45550	46800	45550	46700	156434
2015-03-24	9	46800	47000	46250	46400	66564
2015-03-25	39	47100	47150	46100	46900	129813

4.3 실증분석: 1차 및 2차 분석(2015.3~2017.6)

분석을 위한 테이블 가공(xlsx)

date	score	Open	High	Low	Close	Volume	종가+5일 평균	거래량+5일 평균	감정점수+5일 평균
2015-03-23	9	45550	46800	45550	46700	156434	**46660**	**136212.8**	**16.4**
2015-03-24	9	46800	47000	46250	46400	66564	**46670**	**139192.8**	**18.4**
2015-03-25	39	47100	47150	46100	46900	129813	**46770**	**181696.6**	**20.2**

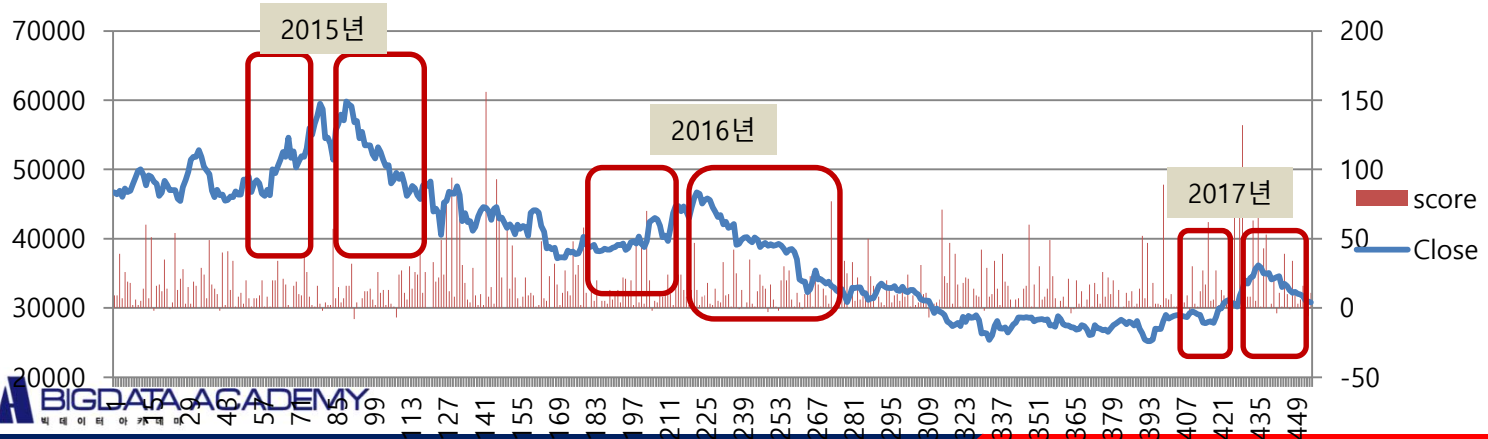
1차 분석

전체기간을 대상으로 감정점수와 종가 및 거래량 관계 분석

분석결과 거래기간 범위가 넓고 환경변수가 많아 상관 관계에 큰 의미가 없음

2차 분석

년도별 지속적으로 1~3개월 이상 상승 또는 하락한 기간을 산정하여 분석



□ 상관관계 분석(Correlation matrics)

상승: 2015.7.6~8.17(1.5개월)

	Score	Close	Volume	clo_5avg	vol_5avg	sco_5avg
score	1	-0.02754	-0.076	0.25829	0.20163	0.40564
Close	-0.02754	1	0.45208	-0.55479	-0.06814	-0.65091
Volume	-0.076	0.45208	1	-0.20033	0.32058	-0.35839
clo_5avg	0.25829	-0.55479	-0.20033	1	0.46527	0.56613
vol_5avg	0.20163	-0.06814	0.32058	0.46527	1	-0.00535
sco_5avg	0.40564	-0.65091	-0.35839	**0.56613**	-0.00535	1

하락: 2015.9.1~11.12(2.5개월)

	score	Close	Volume	clo_5avg	vol_5avg	sco_5avg
score	1	-0.12545	-0.09916	-0.18917	-0.11039	0.36282
Close	-0.12545	1	0.25356	0.71991	0.61602	-0.2988
Volume	-0.09916	0.25356	1	0.24912	0.43605	0.04143
close_avg	-0.18917	0.71991	0.24912	1	0.87249	0.0678
vol_5avg	-0.11039	0.61602	0.43605	0.87249	1	0.29207
sco_5avg	0.36282	-0.2988	0.04143	0.0678	**0.29207**	1

상승: 2016.2.23~5.25(3개월)

	score	Close	Volume	clos_5avg	vol_5avg	sco_5avg
score	1	0.30873	0.04487	-0.1552	0.22727	0.4939
Close	0.30873	1	0.39192	-0.2424	0.22966	0.38533
Volume	0.04487	0.39192	1	-0.09114	0.42385	0.00943
clos_5avg	-0.1552	-0.2424	-0.09114	1	0.36912	0.2074
vol_5avg	0.22727	0.22966	0.42385	0.36912	1	0.38554
sco_5avg	0.4939	0.38533	0.00943	0.2074	**0.38554**	1

하락: 2016.5.26~8.30(3개월)

	score	Close	Volume	clo_5avg	vol_5avg	sco_5avg
score	1	-0.21385	0.13295	-0.27986	0.02331	0.29716
Close	-0.21385	1	-0.35931	0.81017	-0.32806	-0.37635
Volume	0.13295	-0.35931	1	-0.28098	0.71978	0.12366
clo_5avg	-0.27986	0.81017	-0.28098	1	-0.14617	-0.16103
vol_5avg	0.02331	-0.32806	0.71978	-0.14617	1	0.20373
sco_5avg	0.29716	-0.37635	0.12366	-0.16103	**0.20373**	1

상승: 2017.3.3~5.18(2.5개월)

	score	Close	Volume	clos_5avg	volu_5avg	scor_5avg
score	1	0.30873	0.04487	0.3452	0.31662	0.54111
Close	0.30873	1	0.39192	0.96411	0.4331	0.53223
Volume	0.04487	0.39192	1	0.36909	0.49835	0.07059
clos_5avg	0.3452	0.96411	0.36909	1	0.54156	0.59766
vol_5avg	0.31662	0.4331	0.49835	0.54156	1	0.42326
sco_5avg	0.54111	**0.53223**	0.07059	**0.59766**	**0.42326**	1

하락: 2017.5.22~6.22(1개월)

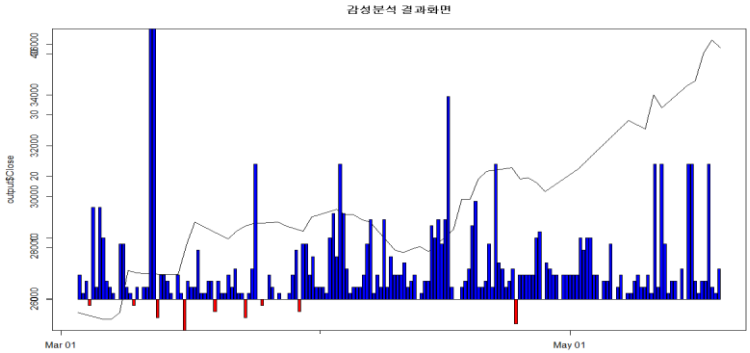
	score	Close	Volume	clo_5avg	vol_5avg	sco_5avg
score	1	0.14865	0.28132	0.07868	0.064	0.51883
Close	0.14865	1	0.37828	0.76094	0.84105	0.21921
Volume	0.28132	0.37828	1	0.43177	0.64251	0.45693
clo_5avg	0.07868	0.76094	0.43177	1	0.8677	0.52595
vol_5avg	0.064	0.84105	0.64251	0.8677	1	0.44916
sco_5avg	0.51883	0.21921	**0.45693**	**0.52595**	**0.44916**	1

연속적으로 1~3개월 이상 상승 또는 하락한 기간을 산정하여 분석한 결과 거래량 및 종가와 상관관계가 있는 것으로 분석됨

□ 상관관계 분석(Correlation matrics)

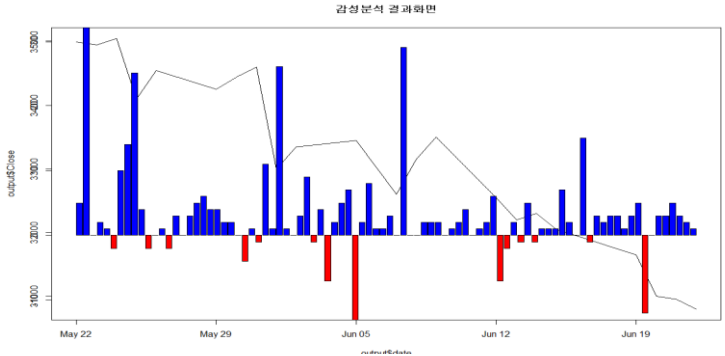
상승: 2017.3.3~5.18(2.5개월) 25,450 ~ 36,150원

	score	Close	Volume	clos_5avg	volu_5avg	scor_5avg
score	1	0.30873	0.04487	0.3452	0.31662	0.54111
Close	0.30873	1	0.39192	0.96411	0.4331	0.53223
Volume	0.04487	0.39192	1	0.36909	0.49835	0.07059
clos_5avg	0.3452	0.96411	0.36909	1	0.54156	0.59766
vol_5avg	0.31662	0.4331	0.49835	0.54156	1	0.42326
sco_5avg	0.54111	**0.53223**	0.07059	**0.59766**	**0.42326**	1



하락: 2017.5.22~6.22(1개월) 35,000 ~ 30,800원

	score	Close	Volume	clo_5avg	vol_5avg	sco_5avg
score	1	0.14865	0.28132	0.07868	0.064	0.51883
Close	0.14865	1	0.37828	0.76094	0.84105	0.21921
Volume	0.28132	0.37828	1	0.43177	0.64251	0.45693
clo_5avg	0.07868	0.76094	0.43177	1	0.8677	0.52595
vol_5avg	0.064	0.84105	0.64251	0.8677	1	0.44916
sco_5avg	0.51883	0.21921	**0.45693**	**0.52595**	**0.44916**	1



긍정	161	60% 감소	긍정	64
부정	9	65% 증가	부정	14
중립	18		중립	12

*감정점수: 일별 뉴스에서 긍정/부정어 count 하여 집계한 점수

□ 분석 제약요건 및 설명

- 3개의 엔터테인먼트 기업과 그 기업에 소속된 연예인을 대상으로 연애기사, 트윗, 페이스북의 Raw Data를 수집을 시작하였으나 시간과 가공, 처리에 많은 시간이 소요되어 1개 기업을 대상으로 수집

- 수집기간(2015.3~2017.6) 까지 대상으로 분석을 시작하였으나 연관성이 없는 것으로 분석되어 분석기간을 특정 이벤트가 있는(연속적으로 1~3개월 주가가 상승 또는 하락하는 기간)을 산정하여 분석하였음

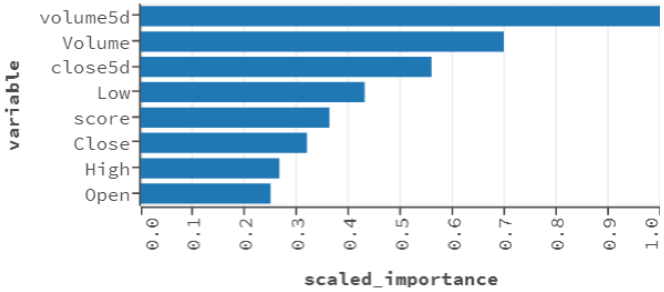
- 상관관계 분석을 위해 주요 변수(감정점수, 종가, 거래량)의 5일 평균값으로 분석

4.4 추가 분석 실행

□ 추가분석 실행(H20 활용 : Random Forest분석)

• **H2O DRF(Distributed Random Forest) 알고리즘 분석**

▾ VARIABLE IMPORTANCES



▾ OUTPUT - VARIABLE IMPORTANCES

variable	relative_importance	scaled_importance	percentage
volume5d	50814.7695	1.0	0.2575
Volume	35441.9805	0.6975	0.1796
close5d	28400.2168	0.5589	0.1439
Low	21867.6426	0.4303	0.1108
score	18418.3223	0.3625	0.0933
Close	16220.9805	0.3192	0.0822
High	13522.7441	0.2661	0.0685
Open	12664.7881	0.2492	0.0642

기간(2017.3.3~6.22)

종속변수: sco_5day

5. 결과 및 시사점

결과 및 시사점

- 연예뉴스 감성분석을 통한 YG엔터테인먼트 주가와의 상관관계 분석 진행.
- 전체 기간(2015.3~2017.6)을 대상으로 한 분석에서 상관관계 큰 의미가 없었음.
- 2015년, 2016년, 2017년 상승 또는 하락한 기간을 산정하여 분석한 결과 거래량 및 종가와의 상관관계 있는 것으로 분석됨(데이터수집, 분석기간에 민감)
- 연구의 한계: YG엔터테인먼트 1개 종목 분석. 감성사전 구축의 어려움. 연예기사 특성 고려 부분.

참고문헌

- 빅데이터 활용서 1 (R을 이용한 중·고급 데이터 분석)
  김경태, 안정국 외 1명, 시대인, 2017.1

- 뉴스와 주가: 오피니언 마이닝을 통한 지능형 투자의사결정모형, 김유신, 김남규, 정승렬, 지능정보시스템학회, 2011.12

- 뉴스와 공시정보에 기반한 주가 변동 예측시스템, 왕인내, 신현아, 권혜진, 최예송, 한국정보과학회, 2016.12