뉴스 키워드 노출 빈도수에 따른 기업 주가 영향 분석

2017104003 양지수

2017104039 한서흔

**요약**

최근 주식시장을 향한 사람들의 관심이 커지면서, 주식의 가치를 판단할 때 어떤 것을 참고해야 하는지에 대한 고민 또한 증가하고 있다. 주가를 움직이는 원인에는 여러 요소가 있지만, 그중 뉴스가 큰 영향을 준다고 알려져 있다. 따라서 우리는 기업과 종합지수에 관한 뉴스의 빈도수와 주가 간의 상관관계가 존재한다는 가설을 세워, 이를 뒷받침하기 위한 빅데이터 분석 연구를 진행한다.

# 1. 서론

## 1.1. 연구 배경

최근 주식시장을 향한 사람들의 관심이 커지고 있다. 금융투자협회에 따르면 현재 주식거래로 활동하고 있는 계좌는 2021년 3월 19일 기준 40,067,529개로 집계됐다.[[1]](#endnote-1) 주식거래 활동 계좌는 코로나19 사태 직후인 지난해 3월 6일 3천만 개를 넘어선 뒤, 약 1년 만에 1천만 개가 늘었다. 올해 들어서도 주식 열풍이 이어지면서 지난해 말 3,548만 개보다 500만 개 가까이 늘었다. 이러한 관심의 증가는 사람들이 주식의 가치를 판단할 때 어떤 것을 참고해야 할지 고찰하게 한다. 주가는 여러 요인에 의해 움직이지만, 그중 뉴스가 큰 영향을 주는 경우를 볼 수 있다. 뉴스는 사회의 여러 사건, 각국의 경제, 정치 상황, 기업의 행보 등 여러 정보를 전달하는 매체이다. 현재 스마트폰이 대중화되면서 사람들은 자신이 원하는 주제의 뉴스를 언제 어디서나 실시간으로 접할 수 있게 되었고, 이는 사람들이 뉴스의 흐름에 더욱 민감하게 반응하며 빠르게 변화하는 경제를 손쉽게 따라잡을 수 있게 만들어 주었다. 예를 들어, 2020년 코로나로 인해 진단 키트의 수요가 늘어나자 코로나 진단 키트를 전문적으로 만드는 기업인 씨젠의 노출이 뉴스에서 증가하기 시작했다.[[2]](#endnote-2) 그에 따라 해당 기업의 주가가 같이 증가했음을 확인할 수 있다. 또 다른 예시로, 2021년 초 현대자동차와 애플과의 자동차 협업 이슈가 제기됨에 따라 현대자동차의 주가가 상승[[3]](#endnote-3)하는 모습을 보였다. 하지만 기업에서 직접 이 사실을 부정하고 뒤이어 협상이 결렬되었다는 뉴스가 여럿 등장하자 주가가 곧바로 하락하는 추세[[4]](#endnote-4)를 보였다. 따라서 본 연구에서는 주가와 관련된 뉴스의 노출 빈도수와 주가 간의 상관관계에 대해 분석하고자 한다.

## 1.2. 연구목표

기업과 관련된 뉴스의 데이터를 직접 수집하고 내용을 분석해, 뉴스의 핵심 키워드와 주가 동향의 연관성을 찾는 연구를 진행하는 것을 목표로 한다.

코스피 지수와 업종이 서로 다른 코스피 지수에 소속된 기업 4개를 선택하여 2020년 10월부터 2021년 3월까지 작성된 뉴스를 수집한다. 수집한 뉴스의 제목에서 키워드를 추출해 주가와 직접적으로 연관이 있는 핵심 키워드를 선별한다. 수집한 키워드를 각각 해당 지수와 기업의 주가의 등락과 연관 지어 키워드의 극성을 사전적 의미가 아닌 기업과 지수별로 특화된 극성으로 설정하여, 해당 키워드가 노출된 빈도와 주가가 유의미한 상관관계를 가지는지 비교분석을 진행한다. 마지막으로 코스피 지수와 기업간의 상관관계를 비교하여 코스피 지수가 각 기업에 미치는 영향을 알아본다.

# 2. 관련 연구

## 2.1. KoNLPy

KoNLPy(Korean NLP in Python)[[5]](#endnote-5)은 한국어 정보처리를 위한 파이썬 패키지이다. KoNLPy 패키지 안에는 카이스트 semantic web 연구실에서 개발한 Hannanum, 서울대 IDS 연구실의 Kkma, Shineware에서 개발한 Komoran, Mecab-ko등이 포함되어 있다. 이를 이용하여 뉴스에 포함된 한국어 정보들을 원하는 품사에 따라 처리할 수 있다. 샘플데이터를 각 분석기를 이용해 분석해 본 후 소요 시간과 분석 정확도를 고려하여 적절한 분석기를 선택한다.

## 2.2. KNU 한국어 감성 사전

감성 사전은 웹사이트나 소셜미디어에서 특정 주제에 대한 여론이나 정보를 수집, 분석해 평판을 도출하는 빅데이터 처리 기술이다. KNU 한국어 감성 사전[[6]](#endnote-6)은 군산대학교에서 개발한 한국어 감성 사전으로 표준국어대사전 뜻풀이의 감성을 Bi-LSTM을 사용해 분류하였다. 해당 사전을 통해 추출한 뉴스 키워드의 감성을 분류한다.

## 2.3. Gephi

소셜 네트워크 분석 소프트웨어(SNA 소프트웨어)는 수치적 또는 시각적 표현을 통해 네트워크의 특징을 설명함으로써 소셜 네트워크의 정량적 또는 정성적 분석을 용이하게 하는 소프트웨어이다.[[7]](#endnote-7) 소셜 네트워크의 시각적 표현은 네트워크 데이터를 이해하고 분석 결과를 전달하는 데 중요하다. SNA 소프트웨어의 종류에는 AllegroGraph, Gephi, Graph-tool 등이 있는데, 그중 NetBeans 플랫폼에서 Java로 작성된 오픈 소스 네트워크 분석 및 시각화 소프트웨어 패키지인 Gephi를 통해 조사한 키워드에 대한 분석을 시각적으로 나타낸다.

## 2.4. 기존 연구의 문제점 및 해결 방안

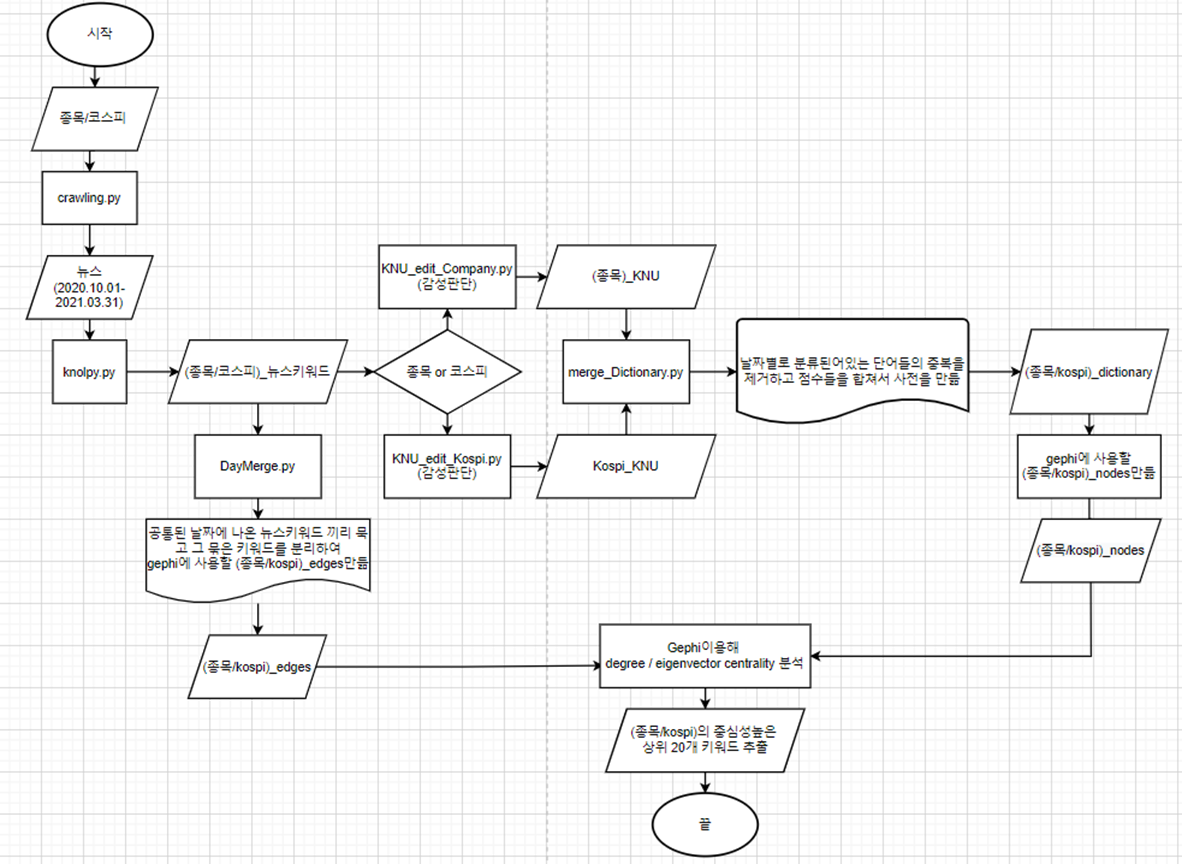
기존 연구에서는 빅데이터 분석 기법 중 오피니언 마이닝을 이용하여 비정형 뉴스 컨텐츠의 감성을 분석하고 수치화하여 뉴스가 주식 시장에 미칠 영향을 분석하고, 이를 기반으로 투자정보를 생성하는 지능형 투자의사결정 모형을 제시하였다. 다만 이 연구에는 한계점이 존재했는데, 먼저 뉴스와 주가를 비교하기에 충분한 데이터 양을 사용하지 않았다. 또한 개별종목의 주가가 아닌 종합주가지수를 비교의 대상으로 하고 있어, 기업에 대한 투자의사결정 정보를 제공하지 못했다.

위의 문제에 대한 해결방안을 다음과 같이 마련했다. 먼저 해당 연구에서 진행한 데이터보다 많은 양의 데이터를 분석한다. 기존연구에서는 3개월간의 뉴스를 조사하였지만 본 연구에서는 2020년 10월부터 2021년 3월까지 총 6개월에 해당하는 뉴스를 조사할 예정이다. 분석할 데이터의 폭이 넓어짐에 따라 결과에 대한 정확도는 자연스럽게 높아질 것으로 예상된다. 또한 KoNLPY를 이용하여 명사단위로 정제한 데이터들을 뉴스 주가와 비교하여 단어에 주가 맞춤형 성격을 부여하고 이를 종합지수가 아닌 개별 주가와 연관 지어, 뉴스 키워드와 기업간의 연결점을 직접적으로 찾고 이를 투자의사결정에 참고가 되는 정보로 사용할 수 있는지 탐구한다.

# 3. 프로젝트 내용

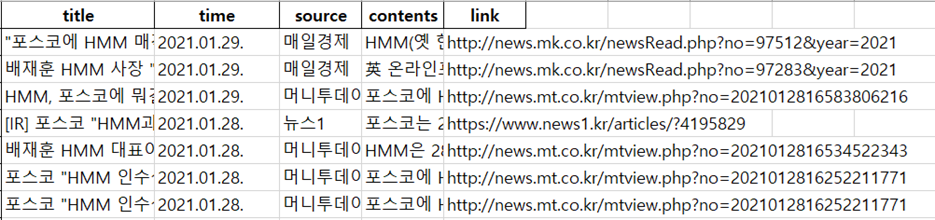
## 3.1. 시스템 설계

**3.1.1. flowchart**



<Flowchart>

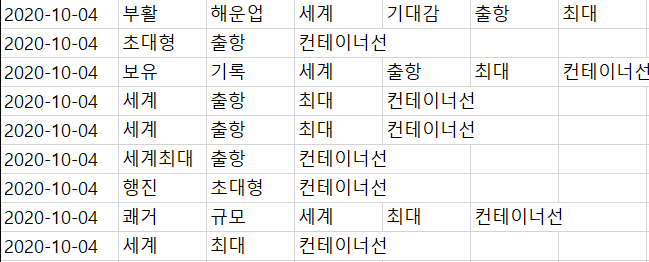
### 3.1.2. 데이터 수집



<HMM 기업 뉴스 크롤링 결과의 일부>

코스피, 그리고 업종이 서로 다른 종목을 4개 선택하여 총 6개월 동안 네이버 포털 사이트에 코스피 혹은 해당 종목의 기업을 검색했을 시 나오는 뉴스의 제목 데이터를 크롤링하여 가져온다. 개발 언어는 Python을 사용, Excel 형태로 자료를 추출한다. 총 15142개의 뉴스를 수집하여 사용했다. (코스피: 9199개, HMM: 1016개, LG디스플레이: 450개, 현대차: 3431개, SK하이닉스: 1046개)

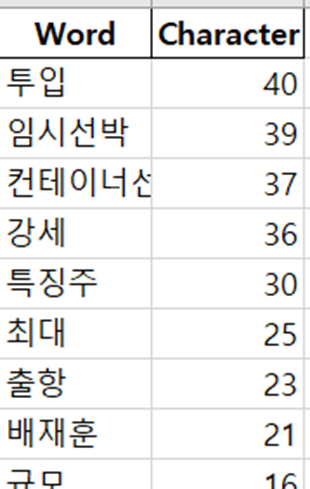
### 3.1.3. 데이터 정제



<KoNLPy 결과>

수집한 데이터를 KoNLPy패키지 중 hannanum을 사용해 자연어 처리하여 형태소를 분석한다. 숫자나 접속사, 조사와 같은 분석에 불필요한 단어들은 제거한 뒤 분석에 사용 가능한 단어들을 정리하여 저장한다. 최종적으로 각각의 기업뉴스의 발행일과, 뉴스제목 속 키워드로 이루어진 데이터 파일을 완성한다.

### 3.1.4. 데이터와 주가 비교



<특화된 감성사전>

KHU 한국어 감성사전을 이용해 수집한 키워드들의 감성을 판별한다. 양수는 긍정, 음수는 부정을 뜻한다. 코스피의 키워드와 기업의 키워드의 분리 기준에는 차별성을 둔다.

기업 키워드의 경우, 각 개별업종 기업의 뉴스 키워드와 주가의 수치를 비교한다. 기본적으로 뉴스는 다음날 주가에 영향을 준다고 정해두지만, 만일 주가가 당일 시가 대비 고가와 저가가 ±2%인 경우, 그 날의 뉴스가 당일주가에 영향을 주었다고 판단해 당일 뉴스 키워드의 극성에 ±1을 부여한다. 그렇지 않을 경우, 다음 날의 주가 등락률을 확인하여 양수일 경우 +1, 음수일 경우 -1을 극성에 부여한다. 금, 토, 일의 뉴스일 경우 월요일의 주가에 영향을 미친다고 판단한다.

코스피 키워드의 경우, 당일에 발행된 뉴스에 해당 날의 코스피 지수에 대한 내용이 담겨 있어 같은 날의 코스피의 뉴스 키워드와 코스피 지수의 등락률을 비교해 코스피의 등락률이 양수일 경우 +1을, 음수일 경우 -1의 값을 더한다.

위 과정을 통해 주식 시장에 특화된 감성 사전을 구축한다.

### 3.1.5. 중심성 분석

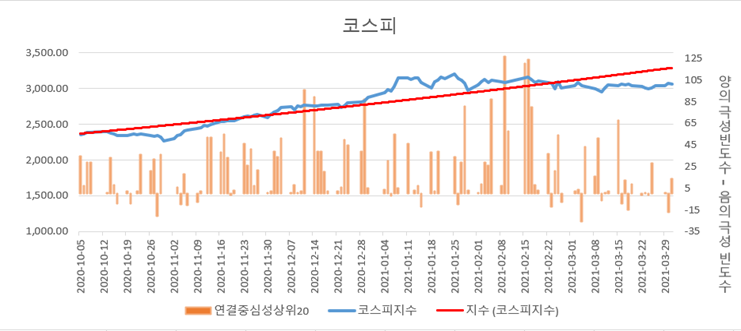
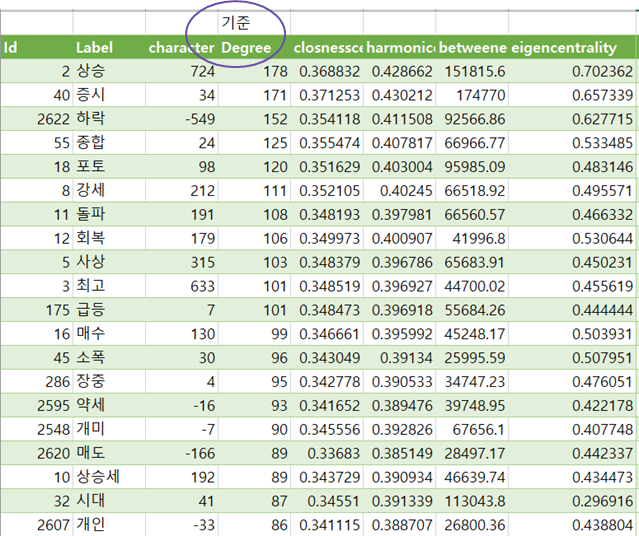
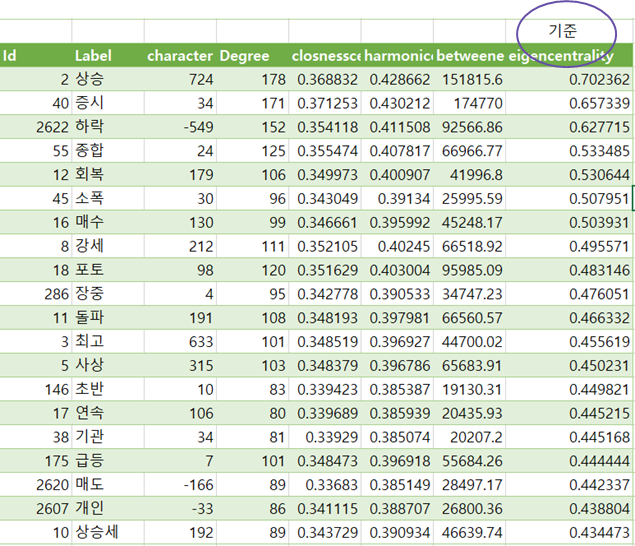
뉴스 키워드 자료에서 중심성을 분석할 노드와 엣지 파일을 Python을 통해 분류한다. 노드 파일에 앞선 과정에서 만든 특화사전의 극성을 Character라고 설정해 Attribute를 추가해서 Gephi의 partition기능을 이용해 붉은 계열은 단어사전에서의 극성이 양수(긍정)인 것, 푸른 계열은 음수(부정)인 것을 표현한다.

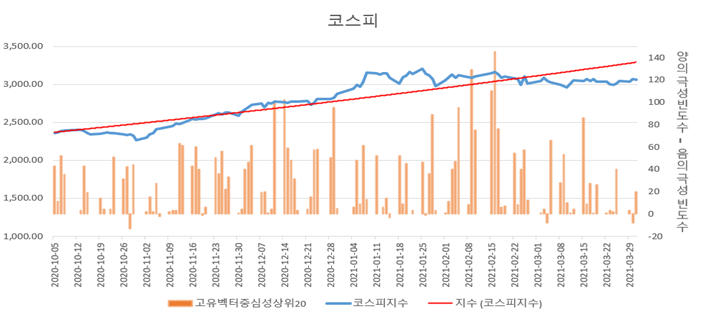
Gephi를 이용하여 4가지의 중심성 수치를 얻는다. 얻은 결과 값 중에서 연결 중심성과 고유벡터 중심성의 결과가 프로젝트에 적합하다고 판단하여, 그 결과 중 상위 가중치 값이 큰 상위 20개의 단어를 뽑는다. 각 중심성의 결과로 나온 20개 단어의 극성을 기업의 주가의 등락률과 비교한다. 3일 연속으로 극성이 양수값일 때, 다음날 주가의 상승 여부를 확인해 정확도를 판별한다.

## 3.2. 구현

### 3.2.1. 코스피 (코스피)

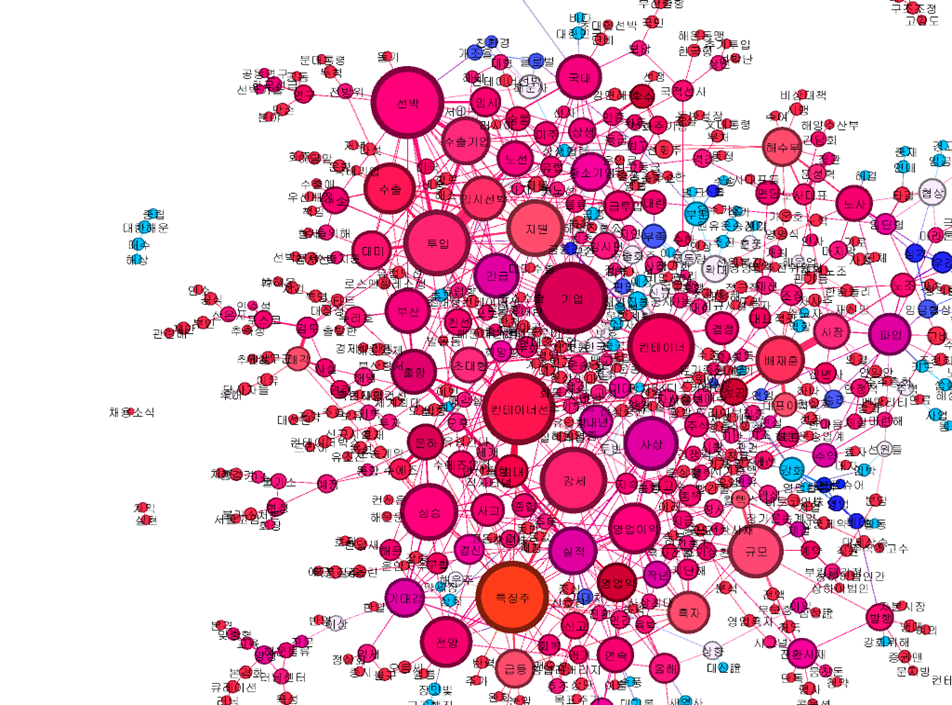
<코스피 데이터를 Gephi에 넣어 시각화한 결과>

<연결 중심성과 고유벡터 중심성 수치 결과 중 가중치 값이 큰 상위 20개의 단어표>

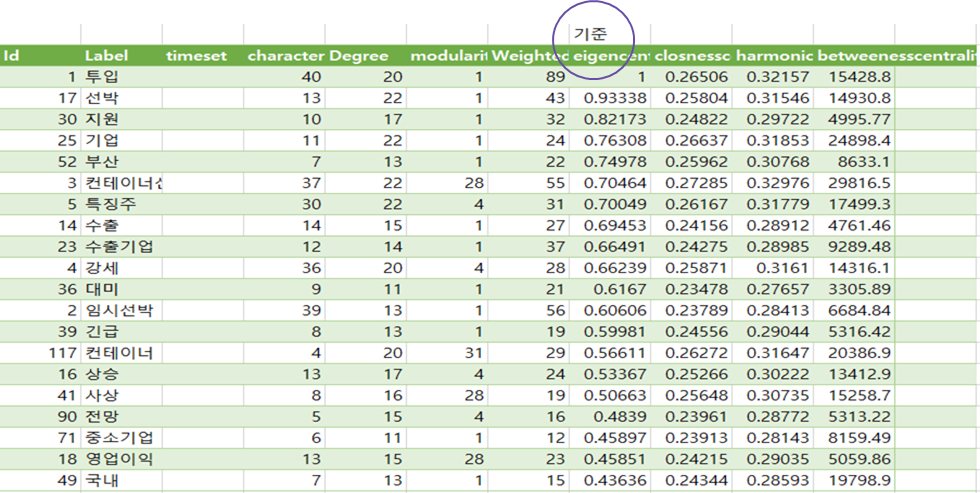
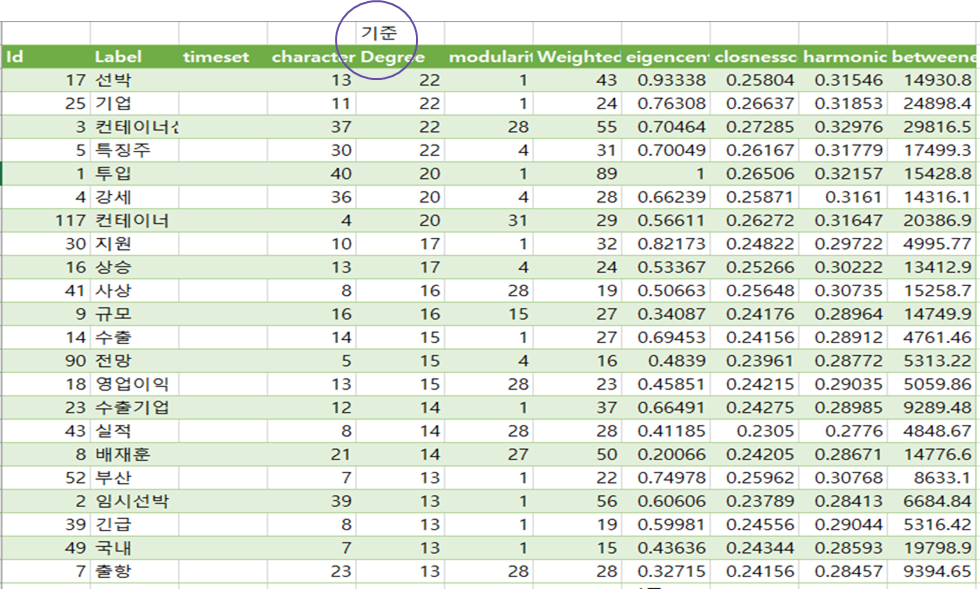
<코스피의 키워드 빈도와 주가 비교 그래프>

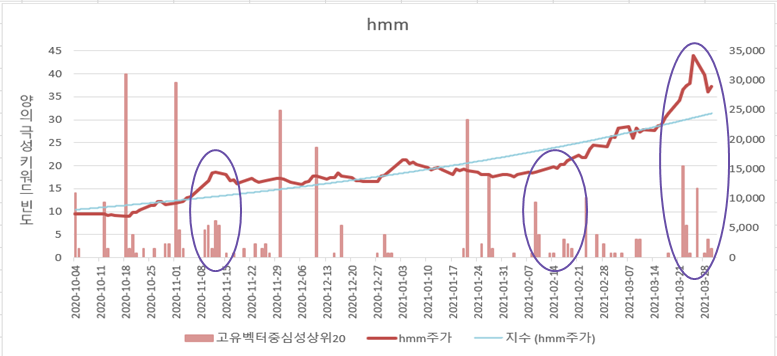
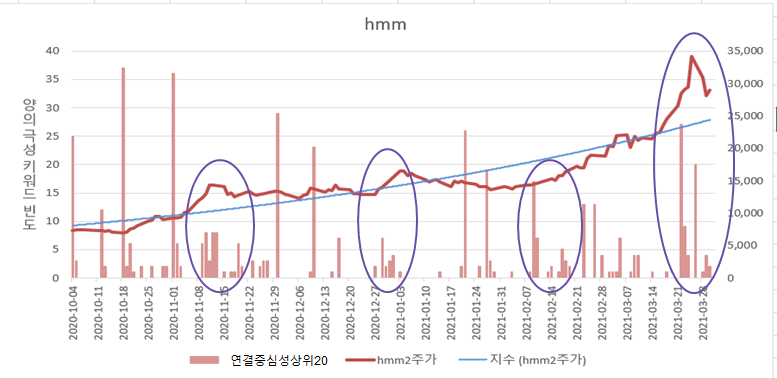
연결 중심성을 이용하여 중심을 찾았을 경우, 20개 중 15개는 양의 극성, 5개는 음의 극성을 띠었다. 고유벡터 중심성을 이용하여 중심을 찾았을 경우, 상위 20개의 키워드 중 17개는 양의 극성, 3개는 음의 극성을 띠었다. 긍정의 성격을 띤 단어와 부정의 성격을 띤 단어가 모두 있기 때문에, 긍정적인 단어와 부정적인 단어를 따로 분리하여 빈도수를 구하고 그 두 결과를 빼서 결과적으로 우세한 성향을 띠도록 하였다. 키워드의 빈도와 성격, 그리고 코스피의 성격을 비교해보면 6개월 동안 양의 극성이 우세하였고 코스피 지수 또한 우상향하는 추세를 보였다는 것을 알 수가 있다.

### 3.2.2. HMM (기업)



<HMM 데이터를 Gephi에 넣어 시각화한 결과>

<연결 중심성과 고유벡터 중심성 수치 결과 중 가중치 값이 큰 상위 20개의 단어표>



<HMM의 키워드 빈도와 주가 비교 그래프>



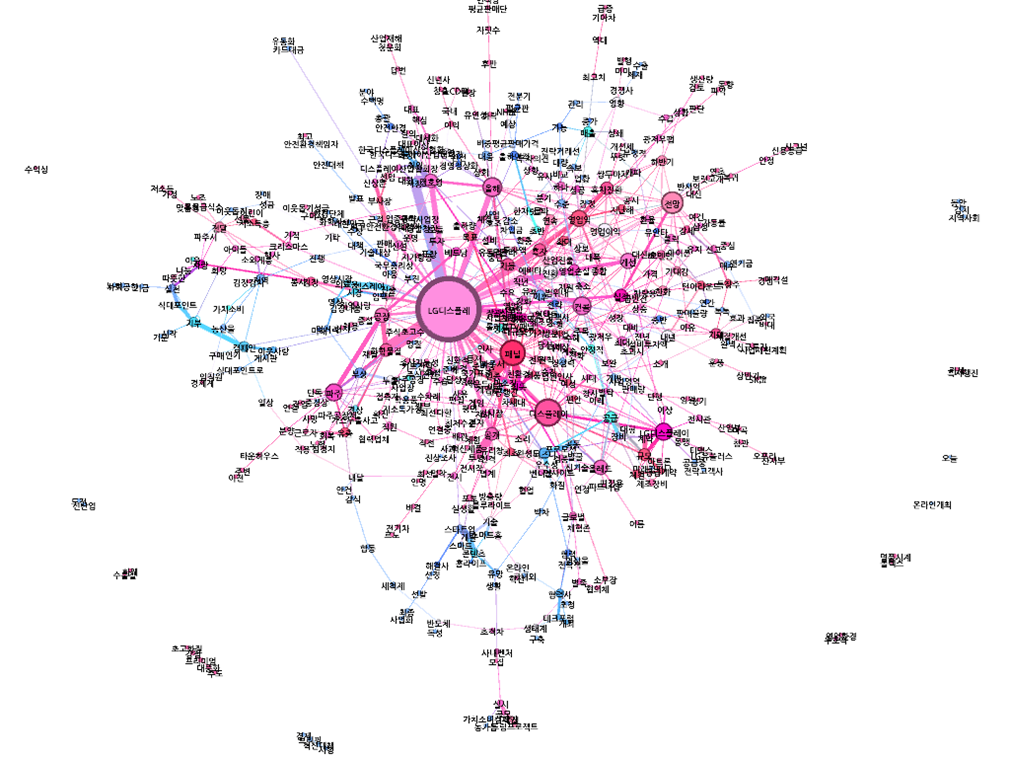
<주가의 상승 여부를 통한 정확도 판별>

빈도를 수치화하여 나타내어 보면 3일이상 긍정의 의미를 가진 뉴스가 나왔을 때 다음날 주가가 상승할 확률은 고유벡터 중심성을 이용했을 때는 총 17회 중 16회 (94.1%) , 연결중심성을 이용했을 때는 총 16회 중 15회 (93.75%)였다.

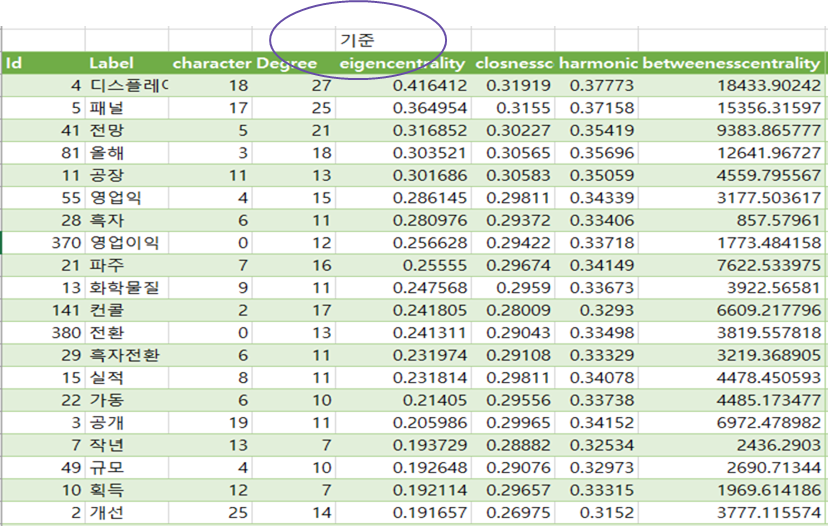
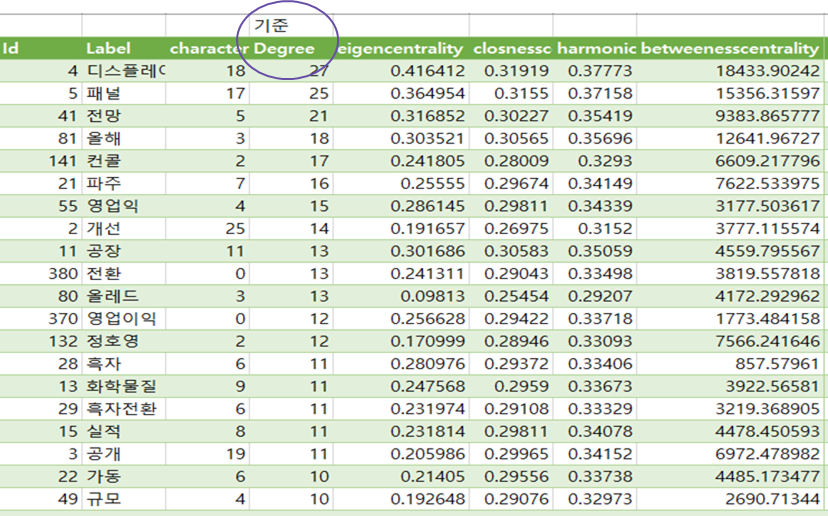
코스피와 기업의 주가를 비교해보면 주식 시장이 열린 121일 중 69일이 일치하였으며 이 비율을 코스피의 해당 기업에 대한 영향력(K)이라고 가정하였다. 이 가정을 바탕으로 코스피의 키워드 값에 K를 곱하고 그 값을 종목의 키워드 값과 합하여 종합적인 중심성을 이용하였다.

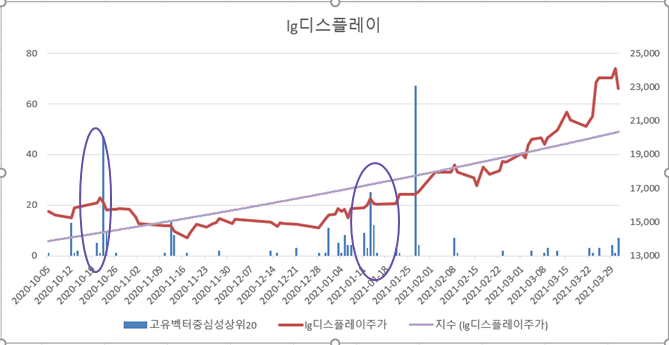
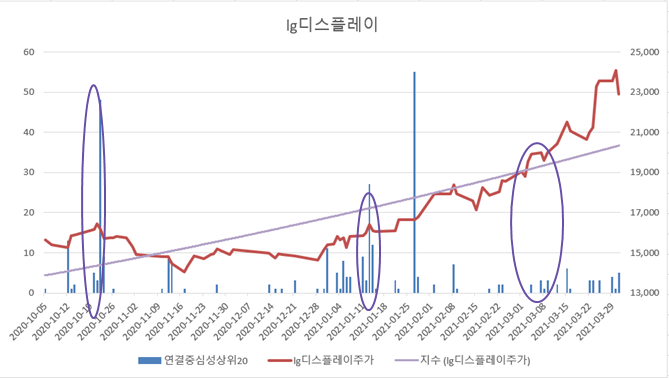
그 결과 확인 할 수 있는 키워드 빈도의 횟수는 증가하였지만 기업뉴스 키워드의 중심성만 사용했을 때보다는 94.1% → 61%, 93.75% → 57.5%로 정확도가 떨어지는 것을 확인할 수 있다.

### 3.2.3. LG디스플레이 (기업)



<LG디스플레이 데이터를 Gephi에 넣어 시각화한 결과>

<연결 중심성과 고유벡터 중심성 수치 결과 중 가중치 값이 큰 상위 20개의 단어표>

<LG디스플레이의 키워드 빈도와 주가 비교 그래프>



<주가의 상승 여부를 통한 정확도 판별>

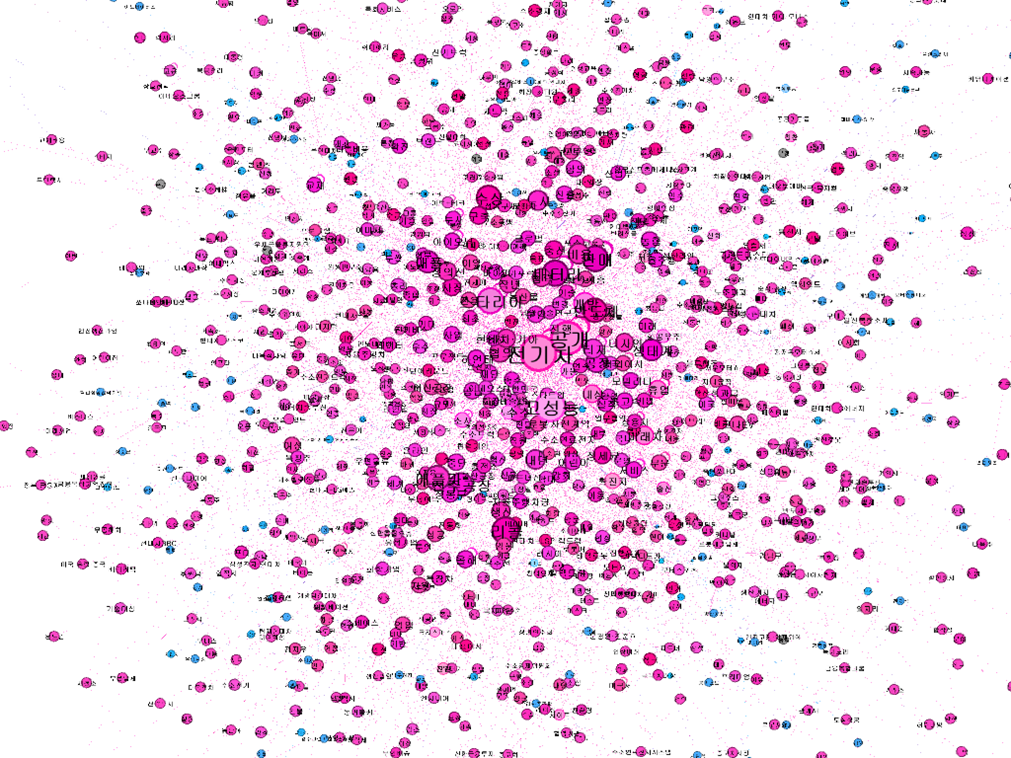
키워드를 연결 중심성, 고유벡터 중심성을 이용하여 중심을 찾았을 경우, 상위 20위의 키워드 중 18개는 양의 극성 2개는 0(중립)을 가졌다. 이를 통해 6개월 동안 긍정적인 키워드들이 높은 영향력을 가졌다는 것을 알 수 있고, 실제 주가를 보면 6개월동안 계속 증가하는 추세였다는 것을 알 수 있다. 또한 키워드의 빈도와 주가를 보면 하루의 높은 빈도 수 보다는 여러날 꾸준히 노출 빈도의 밀집도가 높을 때 주가 상승과 관련이 있다는 것을 알 수가 있다.

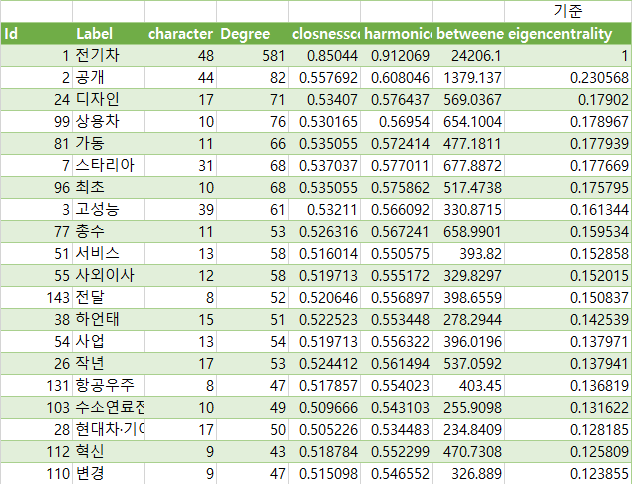
빈도를 수치화하여 나타내어 보면 3일이상 긍정의 의미를 가진 뉴스가 나왔을 때 다음날 주가가 상승할 확률은 고유벡터 중심성을 이용했을 때는 총 5회 중 3회 (60%), 연결중심성을 이용했을 때는 총 9회 중 5회 (55.5%)였다.

코스피와 기업의 주가를 비교해보면 주식 시장이 열린 121일 중 57일이 일치하였으며 이 비율을 코스피의 해당 기업에 대한 영향력(K)이라고 가정하였다. 이 가정을 바탕으로 코스피의 키워드 값에 K를 곱하고 그 값을 종목의 키워드 값과 합하여 종합적인 중심성을 이용하였다.

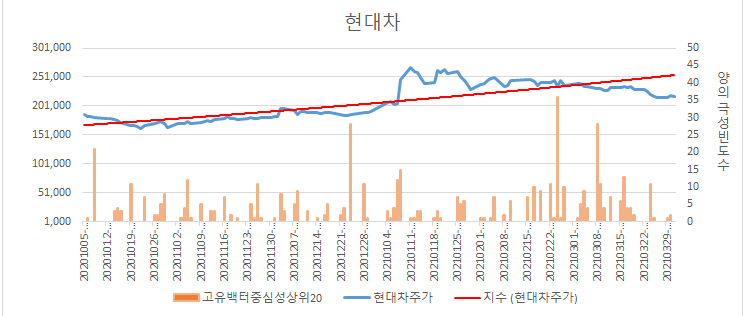
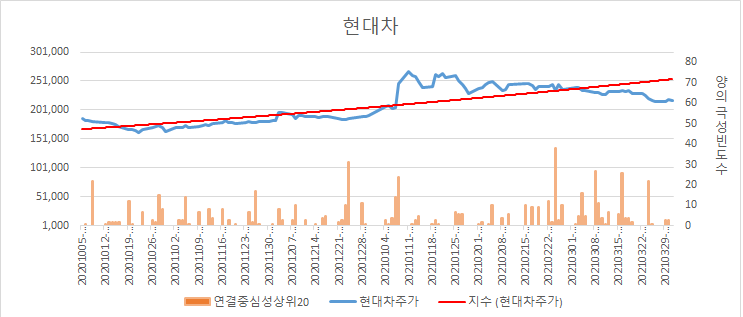
그 결과 확인 할 수 있는 키워드 빈도의 횟수는 증가하였지만 기업뉴스 키워드의 중심성만 사용했을 때보다는 정확도가 60% → 48%, 55.5% → 53%로 떨어지는 것을 확인할 수 있다.

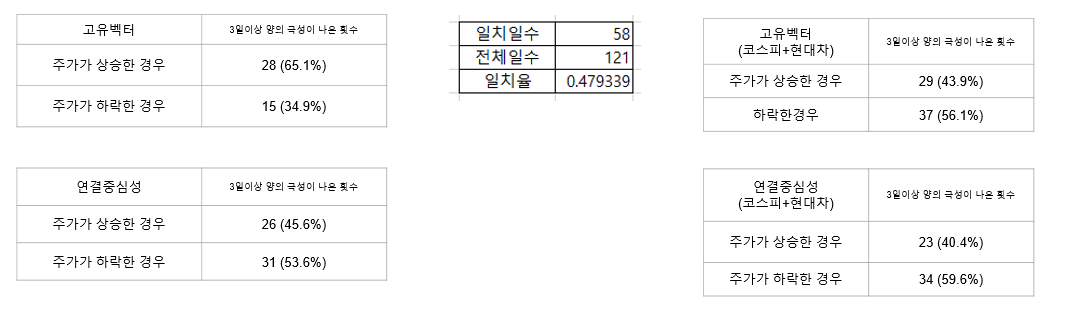
### 3.2.3. 현대차 (기업)



<현대차 데이터를 Gephi에 넣어 시각화한 결과>

<연결 중심성과 고유벡터 중심성 수치 결과 중 가중치 값이 큰 상위 20개의 단어표>

<현대차의 키워드 빈도와 주가 비교 그래프>

<주가의 상승 여부를 통한 정확도 판별>

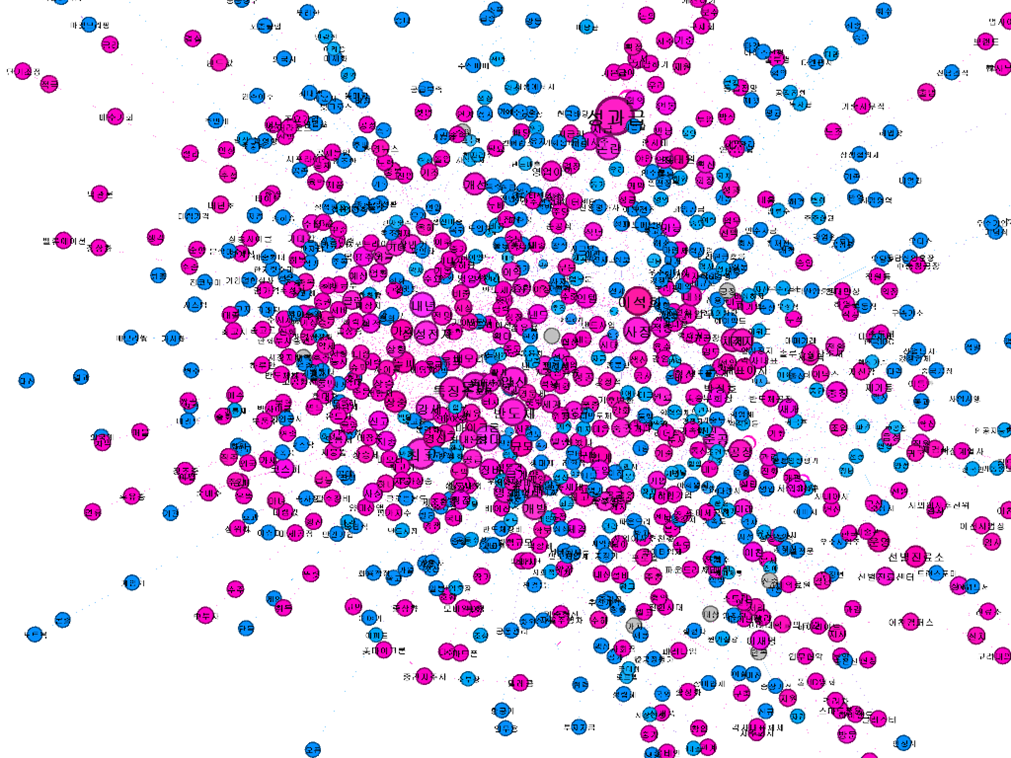
키워드를 정도 중심성, 고유벡터 중심성을 이용하여 중심을 찾았을 경우, 상위 20위의 키워드 모두 양의 극성을 가졌다. 이를 통해 6개월 동안 긍정적인 키워드들이 높은 영향력을 가졌다는 것을 알 수 있다. 키워드의 빈도와 주가를 보면 여러날 꾸준히 노출 빈도의 밀집도가 높을 때 주가 상승과 관련이 있으며, 양의 극성빈도수가 높은 특정 날들 이후로 주가가 소폭 혹은 대폭 상승하는 경우가 많음을 알 수 있다.

빈도를 수치화하여 나타내어 보면 3일이상 긍정의 의미를 가진 뉴스가 나왔을 때 다음날 주가가 상승할 확률은 고유벡터 중심성을 이용했을 때는 총 43회 중 28회 (65.1%), 연결중심성을 이용했을 때는 총 57회 중 26회 (45.6%)였다.

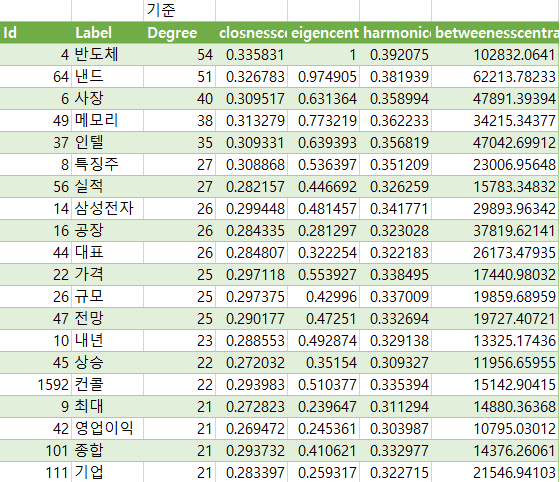
코스피와 기업의 주가를 비교해보면 주식 시장이 열린 121일 중 58일이 일치하였으며 이 비율을 코스피의 해당 기업에 대한 영향력(K)이라고 가정하였다. 이 가정을 바탕으로 코스피의 키워드 값에 K를 곱하고 그 값을 종목의 키워드 값과 합하여 종합적인 중심성을 이용하였다.

그 결과 확인 할 수 있는 키워드 빈도의 횟수는 증가하였지만 기업뉴스 키워드의 중심성만 사용했을 때보다는 65.1% → 43.9%, 45.6% → 40.4%로 정확도가 떨어지는 것을 확인할 수 있다.

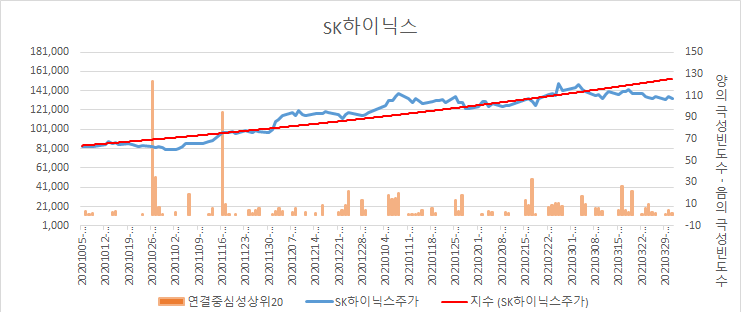
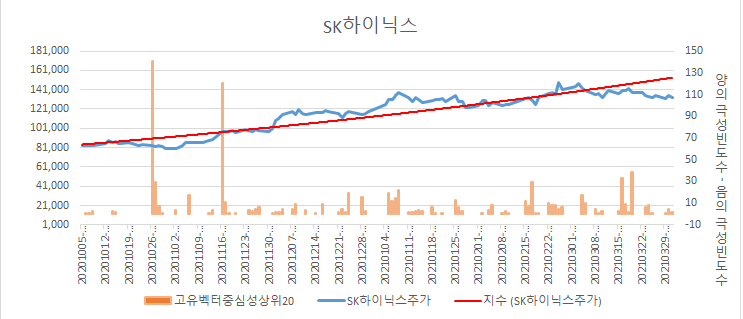
### 3.2.4. SK하이닉스 (기업)



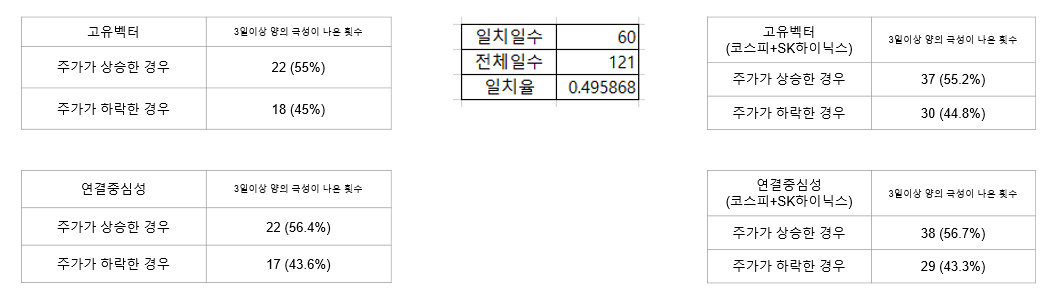
<SK하이닉스 데이터를 Gephi에 넣어 시각화한 결과>



<연결 중심성과 고유벡터 중심성 수치 결과 중 가중치 값이 큰 상위 20개의 단어표>



<SK하이닉스의 키워드 빈도와 주가 비교 그래프>

<주가의 상승 여부를 통한 정확도 판별>

키워드를 정도 중심성, 고유벡터 중심성을 이용하여 중심을 찾았을 경우, 상위 20위의 키워드 중 19개가 양의 극성을 가지고 1개가 음의 극성을 가졌다. 이를 통해 6개월 동안 부정적인 키워드보다는 긍정적인 키워드들이 높은 영향력을 가졌다는 것을 알 수 있다. 6개월간의 주가를 살펴보면 상승하는 추세를 보였고, 키워드의 빈도와 주가를 비교했을 시 노출 빈도의 밀집도가 높을 때 주가가 상승하는 모습을 확인할 수 있었다.

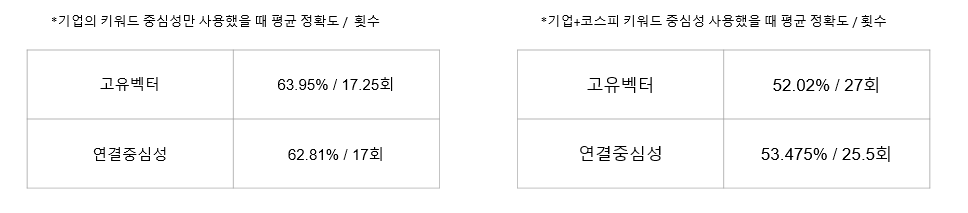
빈도를 수치화하여 나타내어 보면 3일이상 긍정의 의미를 가진 뉴스가 나왔을 때 다음날 주가가 상승할 확률은 고유벡터 중심성을 이용했을 때는 총 40회 중 22회 (55%), 연결중심성을 이용했을 때는 총 39회 중 22회 (56.4%)였다.

코스피와 기업의 주가를 비교해보면 주식 시장이 열린 121일 중 60일이 일치하였으며 이 비율을 코스피의 해당 기업에 대한 영향력(K)이라고 가정하였다. 이 가정을 바탕으로 코스피의 키워드 값에 K를 곱하고 그 값을 종목의 키워드 값과 합하여 종합적인 중심성을 이용하였다.

그 결과 확인 할 수 있는 키워드 빈도의 횟수는 증가하였고, 기업뉴스 키워드의 중심성만 사용했을 때보다 55% → 55.2%, 56.4% → 56.7%로 정확도가 비슷하게 측정되는 것을 확인 할 수 있다.

# 4. 프로젝트 결과

## 4.1. 연구 결과



<평균 정확도와 횟수>

조사한 6개월의 기간 동안 4개의 기업 모두 양의 극성을 가진 키워드들이 우세했으며, 실제로 주가 그래프 또한 우상향하는 모습을 보였다. 이때 키워드의 빈도수와 주가를 비교해보았을 때 한 번 언급되었을 때보다 연속적으로 언급되었을 경우 주가가 상승하는 경향을 보인다고 추측했다. 이를 증명하기 위해 키워드의 빈도수를 수치화하여 주가와 비교해 보았고, 실제로 여러 날에 걸쳐 긍정적인 키워드가 노출 되었을 때 주가가 상승하는 경향을 확인할 수 있었다.

코스피와 각 기업들의 키워드 극성을 비교할 시, 코스피의 영향력을 추가할 경우 기업이 가지고 있는 키워드의 다양한 조합 표본을 보여주지만 기업의 키워드만을 다뤘을 때에 비해 정확도가 낮아짐을 확인할 수 있었다.

따라서 결론적으로 기업의 주가는 뉴스 키워드의 빈도수에 영향을 받아 변동되며, 그 키워드가 어떤 성격을 가지느냐에 따라 상승과 하락의 방향이 정해짐을 알 수 있다.

## 4.2. 추후 연구 방향

본 연구에서는 4개의 기업의 6개월간 뉴스를 이용해 데이터를 수집하였고, 그것을 바탕으로 단어사전을 만들어 해당 단어사전에 포함된 단어의 노출빈도와 단어의 성격, 그리고 주가간의 연관성을 분석하였다. 추후 연구에서는 수집하는 뉴스 데이터의 기간을 늘려 단어사전의 신뢰도와 정확도를 더욱 높일 것이다. 또한 기업의 수를 4개의 기업에서 코스피 200에 포함된 기업들로 확대하여 다양한 기업에 각각 맞는 단어사전을 만들고, 다앙한 종목에 대한 판단을 할 수 있게 되는 것이 목표이다.

# 5. 참고문헌

1. 김영배, “주식계좌 수 4천만개 돌파···1년만에 1천만개 늘어”, 한겨레, 2021년 3월 23일자. [↑](#endnote-ref-1)
2. 이영란, “"코로나 진단키트 주목"...씨젠 · 바디텍메드, '장중 주가 껑충', 초이스경제, 2020년 10월 8일자. [↑](#endnote-ref-2)
3. 이영란, “美 애플과 협력?...현대차 · 기아차, '장중 주가 급등'”, 초이스경제, 2021년 2월 3일자. [↑](#endnote-ref-3)
4. 권유정, “현대차그룹, 애플카 쇼크에 시총 13조원 증발”, 조선비스, 2021년 2월 8일자. [↑](#endnote-ref-4)
5. 박은정, 조성준. “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지.” 제26회 한글 및 한국어 정보처리 학술대회 논문집, p. 133-134, 2014. [↑](#endnote-ref-5)
6. 박상민, 나철원, 최민성, 이다희, 온병원. ” Bi-LSTM 기반의 한국어 감성사전 구축 방안.” 지능정보연구, 24(4), p.219, 2018. [↑](#endnote-ref-6)
7. Wikipedia, https://ko.wikipedia.org [↑](#endnote-ref-7)