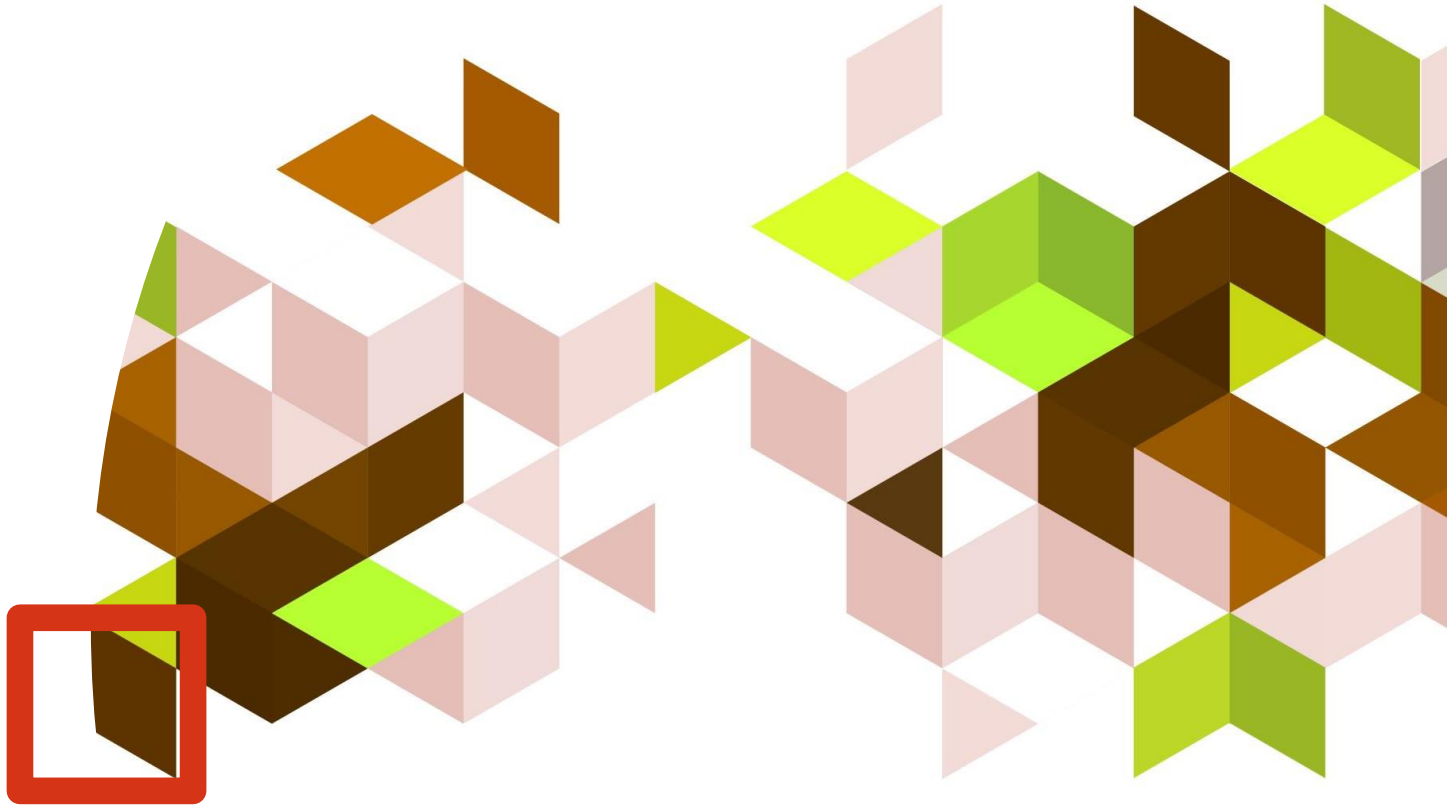
텍스트데이터분석 중간과제

빅데이터경영통계전공

20182791 권유진



목차

Ⅰ. 서론

최근 주식 투자 인구가 800만명을 넘어서고 있다. 예탁결제원의 12월 결산 상장법인 주식 소유자 기준에 따르면 주식 투자 인구는 2019년 말 619만 명에서 2021년 3월에는 835만명으로 추정돼 216만명의 사람들이 증가했다고 한다. 주식시장에서는 미래에 유망하다고 생각하는 분야에 돈을 투자하여 나중에 이 분야가 성공하여 주가가 상승하였을 때 이를 매도해 이득을 취하려고 한다. 이러한 이유로 최근 많이 언급되는 산업들이 있다. 바로 반도체, 자동차, 바이오, 배터리 등이다. 실제로 이러한 산업과 관련된 주식들은 최근 주가가 상승하는 경향을 보이고 있다.

위에서 말한 것과 같이 최근 주식시장에서는 자동차분야가 유망주로 떠오르고 있다. 그 중 나는 내가 요즘 관심을 갖고 있는 자동차 기업에 대해 분석해보려고 한다. 바로 전기차로 떠오르고 있는 기업 '기아(Kia)' 이다. 기아는 현재 국내증시 시가총액순위 11위이고(2021.04.23기준), 자동차 산업 시가총액은 2위이다(2021.04.23기준). 나는 기아의 주가가 어떠한 반응변수들에 영향을 받아 변동하는지 분석해보고자 한다.

주가가 오르고 내리는 것에는 여러가지 반응변수들이 있을 것이다. 특히 기아는 대부분 미래에 유망할 것이라는 기대를 갖고 투자를 하기 때문에 '정보' 가 중요하게 작용할 것이라고 생각을 했다. 그렇기 때문에 '소식' 과 '정보'를 얻을 수 있는 인터넷 기사가 어떠한 내용을 갖고 올라오는 지에 주가가 영향을 받을 것이라고 생각을 했다. 그래서 나는 주제를 '인터넷 기사를 통한 기아의 소식' 으로 선정했고 이를 분석해보고자 한다. 나는 기사에 어떠한 단어들이 많이 등장하는지, 과연 사람들은 어떠한 기사에 긍정적, 부정적 감정을 느끼는지, 어떠한 시기에 어떠한 주제가 많이 언급되었는지 등을 분석할 것이다. 그리고 감성분석, 주제분석을 통해 기사가 주가변동에 어떠한 영향을 미치는지 분석해보겠다.

Ⅱ. 본론 - ⅰ. 데이터 수집

‘네이버증권’ 사이트 웹스크래핑을 이용하여 데이터 수집을 하였다. ‘네이버 증권’ 에는 각 종목별로 정보가 모여있어 주가의 변동, 회사의 실적, 동일업종비교 등 여러가지 정보를 얻을 수 있어 많은 사람들이 이용한다. 그 중 종목과 관련된 인터넷 기사들을 웹스크래핑을 하였다.

‘https://finance.naver.com/item/news.nhn?code=000270’ 에 기아와 관련된 기사들이 모여있다. 주소의 query부분에 페이지와 관련된 부분이 없어 **selenium**의 ‘**click()**’ 메소드를 이용해 페이지를 바꾸려고 했지만, 해당 사이트에선 #news_frame(인터넷 기사 전체 부분)까지만 css선택자가 선택이 되고, 이것의 하위 class는 선택이 되지 않았다. 방법을 찾던 중에 #news_frame이 ‘/item/news_news.nhn?code=000270&page=&sm=title_entity_id.basic&clusterId=’ 로 이어진다는 것을 발견하였다. 이 주소에는 #news_frame의 내용만 있었고, 여기서는 하위css선택자가 선택이 되었다. 그래서 이 주소를 이용해 웹스크래핑을 하였다. 이 주소에서는 query부분에 page와 관련된 부분이 있어 이를 **포맷팅**과 **반복문**을 사용해 여러 페이지에 있는 기사들을 선택할 수 있었다.(**requests**, **lxml.html.fromstring**, **cssselect** 이용)

기사와 연결되는 class의 ‘href’ 특성들은 모두 상대주소를 갖고 있었다. 이를 상대주소에서 절대주소로 변환하기 위해 **urljoin** 메소드를 이용하였다. 그리고 **requests**패키지를 이용해 서버에 요청을 하고 **lxml**패키지를 이용해 정보를 추출하고 **cssselect**를 이용해 기사를 작성한 날짜와 기사내용을 스크래핑했다. 많은 공백들이 같이 스크래핑되었는데, **re**패키지로 정규표현식을 사용해서 이 공백들을 띄어쓰기를 제외하고는 모두 제거하였다. 또한 관련뉴스, 광고 등도 같이 스크래핑되었다. 다행히 이들이 기사의 하위class로 묶여 있어 css선택자 중 ‘**:not()**’ 을 이용해 제외하고 스크래핑을 시도했다. 하지만 **lxml**패키지에서는 **:not()** 선택자가 실행되지 않았다. 그래서 **BeautifulSoup**, **HTML-Translator**을 이용해서 시도해보았지만 이 또한 **:not()** 선택자가 실행되지 않았다. 그래서 나는 그냥 **lxml**패키지로 해당 css선택자를 불러와 **pandas**에 있는 **replace**메소드와 **반복문** 이용해 해당 css선택자에 있는 문자들을 ‘’ 으로 대체하는 방법을 사용하였다.

Ⅱ. 본론 - ii. 데이터 전처리

기사내용을 스크랩하여 데이터프레임을 작성했다. 데이터프레임에 주소, 날짜, 제목, 내용, label(감성)을 열로 정의했다. 이 중 날짜 열을 먼저 전처리하였다. 날짜 열의 데이터형식이 날짜가 아닌 문자형식으로 작성되어 있어, arrow패키지를 이용해 문자열을 받아와 날짜형으로 변환시켜주었다.

```
news.content[866]
```

```
' 이창환 기자 goldfish@asiae.co.kr<ⓒ경제를 보는 눈, 세계를 보는 창 아시아경제 무단전재 배포금지> '
```

그리고 스크래핑한 내용들을 훑어보던 중, 위와 같이 내용이 기자이름과 이메일 밖에 없는 기사들이 있었다. 그래서 주소를 통해 기사를 읽어보았더니 오른쪽 사진과 같이 속보의 경우 제목으로 정보를 전하고 내용이 위와 같이 의미가 없었다. 이러한 기사들을 찾기 위해 기사내용이 100글자 이하인 기사들을 찾아보았는데, 이 지점이 속보와 일반기사를 완벽하게 나누는 지점이었다. 그래서 내용이 100글자 이하인 기사의 내용에 기존 내용들을 없애고 제목을 입력하였다. 또한 같은 내용이 중복되는 기사들이 있어 pandas의 drop_duplicates메소드를 이용해 제거하였고, 행을 제거하다보니 index번호가 비어 reset_index메소드로 index번호를 다시 지정해주었다.

그 다음 형태소분석을 위한 함수를 만들어 전처리하였다. Kiwi()메소드를 이용해 한글로 이루어진 내용을 형태소분석을 실시한 후, 명사만을 선택하게 하였다. 시험삼아 한번 함수를 적용시켜본 결과, 이메일과 특수문자가 들어간 단어들이 발견되어, 이메일 제거를 위해 @이 포함된 단어의 경우 제외하였고, 나머지 특수문자는 re패키지를 통해 정규표현식을 이용해 특수문자만 제거하였다. 또한 기사 마지막에 위 사진과 같이 '무단전재 및 재배포 금지' 라는 문구들이 있었다. 이는 단어빈도분석에 지장을 줄 것 같았다. 무작정 기사에서 해당 단어들을 제외하기엔 기사 내용에 타격이 있을 것 같아 기사의 마지막 20글자 이내에 있는 경우 기사 내용에서 제외시켰다.

Ⅱ. 본론 - iii. 단어빈도 분석

단어빈도분석을 위해 단어문서행렬을 먼저 만들었다. CountVectorizer을 이용해서 만들었는데, 이때 전처리 과정에서 형태소분석과 분석에 불필요한 단어들을 걸러주는 함수 extract_n을 tokenizer로 지정해서 만들었다. 그리고 상대적으로 적은 문서에 나오면서 특정 문서에 자주 나온 단어들을 보기 위해 TfidfTransformer을 이용해 TF-IDF를 적용한 단어문서행렬도 만들었다.

CountVectorizer

TfidfVectorizer

옆 그림은 기사 전체의 단어빈도를 나타낸 단어구름이다. 왼쪽 그림은 CountVectorizer을 이용한 단어구름, 오른쪽 그림은 TfidfTransformer을 이용해 변환한 단어구름이다. 이 두 그림의 차이를 통해, '대', '만', '판매' 등의 단어가 아래 그림에서 크기가 커진 것을 보아, 특정 문서에 자주 나온 것을 알 수 있다. 그리고 '기아차', '현대차'는 아래 그림에서 크기가 작아졌으므로 많은 문서에 등장하여 단어빈도가 높았음을 알 수 있다.

단어구름을 보면, 회사의 상호인 '기아차', '현대차'가 많이 등장하였다. 그리고 차량판매 실적과 관련된 '판매', '차량', '고객', '대', '분기'등이 많이 등장하였다. 또한 회사의 상황을 알려주는 '코로나', '공장', 또 미래를 알려주는 '애플', '전기차', '모델 ' 이 많이 등장했음을 알 수 있다.

단어빈도순위로 분석해보면, 판매와 관련된 단어 '대', '판매', '원', '고객', '억'이 빈도가 가장 높게 나왔다. 이를 통해 회사 실적을 알려주는 기사가 가장 많음을 유추할 수 있다. 그 다음으로는 '전기', '차'가 많이 등장한 것으로 보아, 전기차관련기사가 그 다음을 이음을 알 수 있고, 그 다음으로 '코로나 ' 의 빈도가 높은 것을 보아 코로나의 영향과 관련된 기사의 수가 많음을 유추할 수 있다. 하지만 이는 유추일 뿐, 위 단어들이 같은 기사에 등장했음을 확신할 수 없다. 이와 관련된 분석은 뒤의 '주제분석'에서 자세히 진행하겠다.

Ⅱ. 본론 - iii. 단어빈도 분석

앞에서는 전체기사들의 단어빈도를 분석했다. 이번에는 왼쪽그림을 보면 주별로 주가가 올랐는지 내려갔는지 보여주는 그래프(주봉)가 있다. 이 그래프를 참고해서 주가가 올라간 주와 내려간 주의 기사를 나눴다. date열은 arrow패키지로 날짜형식으로 이뤄져 있다. 이를 .year과 .week을 이용해 '연도-주'로 나누고 이를 이용해 주가가 상승한 주의 기사인지 내려간 주의 기사인지 구분했다. 그리고 상승한 주에는 어떤 단어가 많이 나왔는지, 하락한 주에는 어떤 단어가 많이 나왔는지 분석해보았다. 이때 문서간의 차이를 보려는 것이 아니므로 CountVectorizer을 이용했다.

상승

하락

왼쪽 그림은 상승한 주와 하락한 주의 단어구름그림이다. 상승한 주에는 '디자인', '판매량', '브랜드', '증가', '기록', '위'라는 단어가 훨씬 많이 등장하였다. 또한 전기, 서비스 같은 단어도 더 많이 등장하였다. 이 그림에서는 상승한 주와 하락한 주의 빈도 차이를 명확히 보기 어렵다. 그래서 상승한 주의 단어빈도와 하락한 주의 단어빈도의 차이 값으로도 단어구름을 그려 차이를 더 잘 볼 수 있도록 했다.

왼쪽 그림은 상승한 주와 하락한 주의 단어빈도 차이를 나타낸다. '대', '만' 의 빈도 차이가 가장 크다. 그 외 판매와 관련된 '판매', '판매량' 그리고 수상과 관련된 '위', '선정', '부문' 전기차와 관련된 '배터리', '전기'가 상대적으로 많이 나왔다. 오른쪽 그림은 하락한 주와 상승한 주의 차이를 나타내는데 '분기', '지원', '이익', '가격', '억'이 가장 차이가 크다. 또한 '부담', '중단', '소송', '판결'도 차이가 크다.

상승-하락

하락-상승

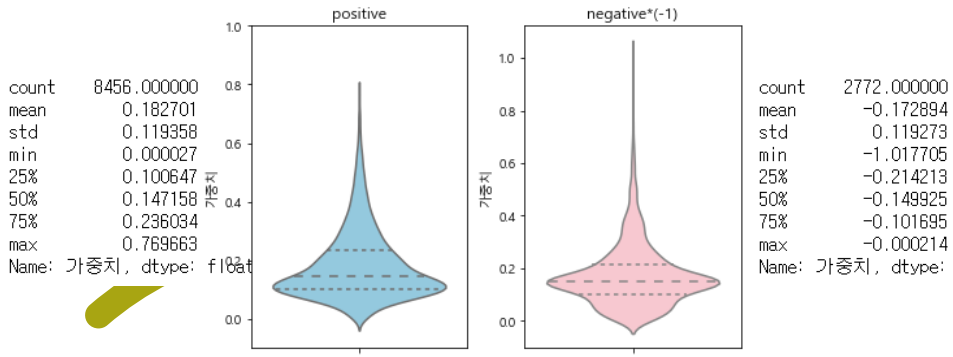
Ⅱ. 본론 - iv. 감성분석

감정분석을 하기에 앞서, 기사들마다 label열에 감성을 0과 1로 작성하였다. 기아에 긍정적인 영향을 끼치는 기사의 경우 1을 부정적인 영향을 끼치는 기사의 경우에는 0을 라벨링하였다. 그리고 value_counts로 각 값들의 개수를 세보았더니, 오른쪽과 같이 매우 불균형적으로 분포하고 있음을 알 수 있었다. 그렇기때문에 평가 척도를 accuracy로 하는 것은 적합하지 않다고 생각했고, 평가척도를 f1-score로 하였다.(https://aakashgoel12.medium.com/how-to-add-user-defined-function-get-f1-score-in-keras-metrics-3013f979ce0d 참조)

```
1    2214
0     581
Name: label, dtype: int64
```

감성분석에 사용할 모델로 로지스틱 모형을 사용하였다. 그리고 데이터의 수가 많다고 생각되지 않아 hold-out방법(train_test_split)을 사용하는 것보다 K-fold cross validation방법을 사용해 과적합을 줄이고 모든 데이터가 학습데이터와 평가데이터가 되도록 하였다. 이 때 모형은 tensorflow.keras를 통해 생성했고, cross validation은 scikit-learn의 Kfold를 이용해 교차검증하였다. 그 결과, holdout방법을 이용할때는 약 88%를 맴돌던 성능이 교차검증을 이용하니 90%가 넘는 성능을 갖게 되었다.

그 다음 가중치 분석을 위해 만들어 놓은 모델에 .weights를 통해 가중치를 추출한 후, 데이터프레임에 저장하였다. 그리고 분석을 위해 가중치가 양수인 단어들과 가중치가 음수인 단어들을 나누었다.




```
count    8456.000000
mean        0.182701
std         0.119358
min         0.000027
25%         0.100647
50%         0.147158
75%         0.236034
max         0.769663
Name: 가중치, dtype: float
```

```
count    2772.000000
mean       -0.172894
std         0.119273
min        -1.017705
25%        -0.214213
50%        -0.149925
75%        -0.101695
max        -0.000214
Name: 가중치, dtype:
```

본격적인 분석에 들어가기 앞서 가중치의 분포를 간단히 분석해보겠다. 왼쪽 그림은 가중치의 분포를 나타낸 violinplot이다. 왼쪽은 가중치가 양수인 분포이고, 오른쪽은 가중치가 음수인 분포로서, 가독성을 위해 해당 그래프에서는 양수로 조정했다. 가중치가 양수인 단어의 수가 8456개로 2772개인 음수인 단어보다 훨씬 많았다. 그에 비해 음수의 가중치는 -1까지 분포한 반면 양수인 가중치는 최대값이 0.77이었다. 그리고 IQR은 가중치가 음수인 쪽이 약 0.2만큼 더 작았다.

Ⅱ. 본론 - iv. 감성분석

긍정단어구름

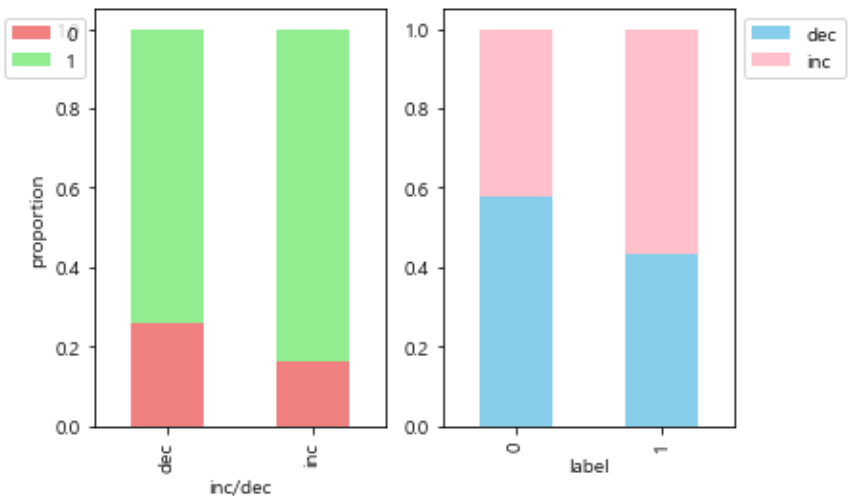
부정단어구름

왼쪽 그림은 가중치를 기준으로 그린 단어구름그림이다. 위 그림은 가중치가 양수인 단어, 아래 그림은 가중치가 음수인 단어들로 이루어졌다. 이 그림을 통해 단어들의 가중치의 크기 비교를 용이하게 할 수 있다.

먼저 가중치가 양수인 단어구름을 보면, 가중치가 가장 높은 단어는 '편의', '출시', '고객', '제공', '신규', '향상'이다. 이를 통해 유추해보면 신차가 출시했다는 기사나 제품의 편의성, 성능을 강조한 기사에 긍정적인 감정을 느낄 확률이 높다. 그리고 '미래', '기대', '상승세' 등의 단어를 사용한 회사의 미래를 긍정적으로 평가하는 기사 역시 감정에 좋은 영향을 미친다. 또한 '카니발'이라는 단어가 있는 것으로 보아 이 기사가 쓰여진 기간내에 기아는 카니발로 흥행했음을 유추할 수 있다. '글로벌', '캠페인', '대상' 단어의 가중치도 높게 나왔는데 판매활동을 제외한 다른 캠페인, 시상식도 긍정적인 영향을 미쳤다.

그 다음 가중치가 음수인 단어구름을 살펴보겠다. 언뜻 살펴보아도 매우 부정적인 단어들이 많다. 가중치가 가장 높게 나온 단어는 '중단', '급감', '감소', '파업'이다. 어떠한 것이 중단되거나 직원들이 파업을 할 때 그리고 어떠한 성적이나 무엇인가가 급감, 감소할 때 우리는 기사를 읽고 부정적인 감정을 느낄확률이 매우 높음을 알 수 있다. 그리고 요즘 시국에 맞게 '확진 ' 이라는 단어 역시 감정에 부정적인 영향을 미쳤다. 또 '블룸버그', '소하리'의 가중치도 컸는데, 블룸버그통신에서 기아의 좋지 않는 기사를 전했고, 소하리에 있는 공장에 좋지 않은 일이 있었음을 알 수 있다. 이 이외에도 '화재', '한파' 등 자연재해 역시 좋지 못한 영향을 끼치고 '임금', '성과급', '상여금' 등 좋은 영향을 미칠 것 같은 단어들 역시 음수 가중치를 갖고 있었다.

Ⅱ. 본론 - iv. 감성분석



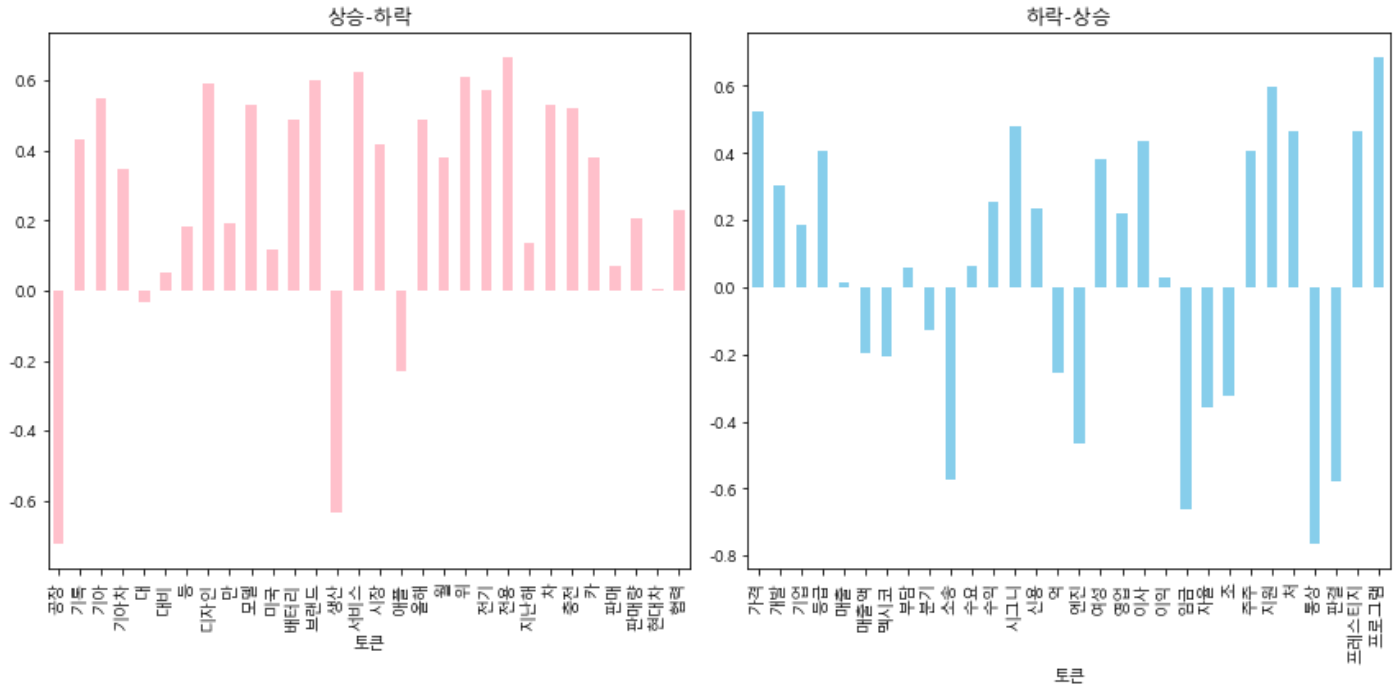

```
chi2_contingency(pd.crosstab(news.label,news['inc/dec']))
```

```
(40.33654382925258,
 2.137727280708424e-10,
 1,
 array([[ 268.56958855,  312.43041145],
        [1023.43041145, 1190.56958855]]))
```

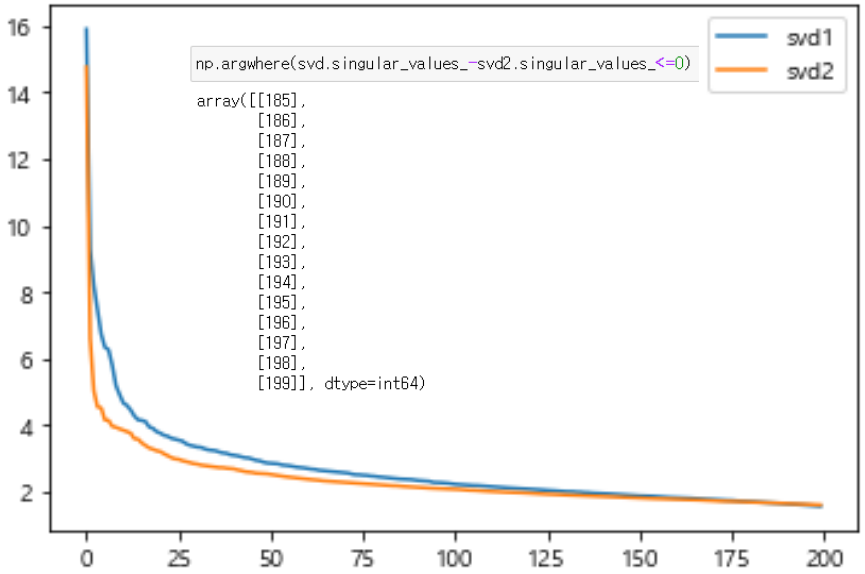
기사의 긍정 여부가 주가 상승에 영향을 미치는지 여부가 우리가 궁금한 점이다. 그래서 pivot_table와 barchart의 stacked parameter을 이용해 왼쪽과 같은 그림을 그렸다. 실제로 주가가 상승했을 때에 긍정적인 기사의 비율이 더 높게 나타났다. 또한 긍정적인 기사가 작성되었을 때 주가가 상승한 비율도 더 높게 나타났다. 이를 통해 서로 상관관계가 있다고 말할 수 있다.

이를 증명하기 위해 scipy.stats의 chi2_contingency메소드를 통해 카이제곱검정을 하였다. 그 결과, p-value가 매우 작게 나와 유의수준에 미치지 못해 두 변수가 서로 독립이 아님을 밝혀냈다.

오른쪽 그림은 상승한 주의 단어빈도와 하락한 주의 단어빈도의 차이가 가장 큰 상위 30개 단어들이다. 상승한 주의 기사에 많이 나온 단어들이 거의 양수의 가중치를 갖고 있으며, 하락한 주에 나온 기사에 많이 나온 단어들은 긍,부정을 띄는 단어들이 섞여 있지만, 상대적으로 부정을 띄는 단어들이 더 많다.



Ⅱ. 본론 - v. 주제분석



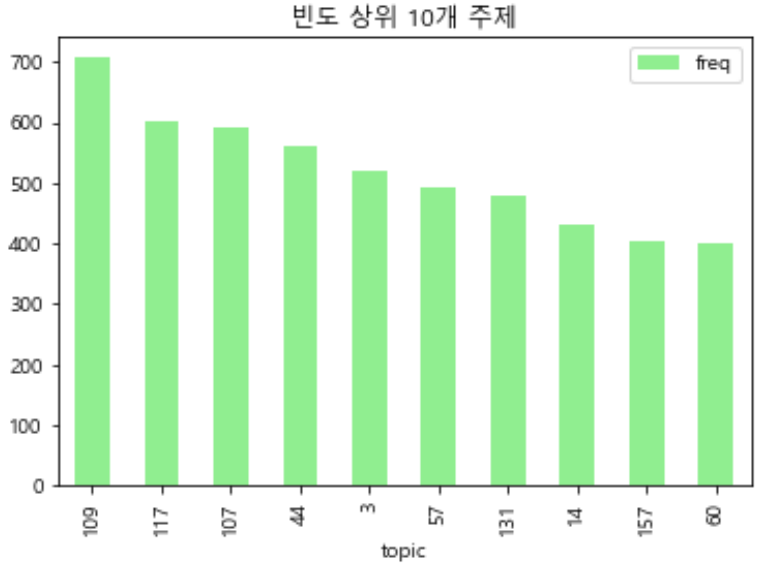
주제분석을 위해 문서의 주제를 몇 개로 나눌 지를 정해야 한다. 그렇기 위해 sklearn패키지의 TruncatedSVD를 이용해 LSA방법으로 단어문서행렬을 적합했다. 그리고 numpy의 random을 사용해 더미모델을 만든 후, 실제 모델의 그래프가 더미 모델 그래프보다 밑으로 내려가는 지점을 찾았다. 바로 185였다. 그래서 나는 차원의 수를 185개로 정하였다.

LDA를 사용해 주제분석을 실시했다. 모델은 gensim으로 선정해서 실시했는데, 이를 위해서는 단어문서 행렬을 corpus 형식으로 변환시켜줘야 한다. 기존의 만들어놨던 tdm과 단어 목록을 Sparse2Corpus, Dictionary를 이용해 gensim모델에 알맞게 변환했다.

그리고 모델링을 실시하였다. scikit-learn의 train_test_split메소드를 이용해 홀드아웃방법으로 corpus를 나누고, 학습데이터를 모델에 적합시켰다. 그리고 log_perplexity를 이용해 혼잡도를 기준으로 loss를 지정했는데, 반복문을 사용하여 검증데이터의 혼잡도를 측정해 성능의 증가세가 작아질 때까지 반복하였다.

그리고 전반적으로 모델을 평가하기위해 주제 응집도와 다양도를 계산해보았다. 주제 응집도는 Coherence Model을 사용해서 계산하였는데 0.80431이 나왔다. 주제 다양도는 주제별 상위 25개 단어 수를 기준으로 계산을 하였는데, 0.6744가 나왔다. 이 수치를 보았을 때, 충분히 주제안의 단어들끼리 유사하고 주제들 간에 관련이 크지 않아 주제분석에 문제가 없다고 생각했다.

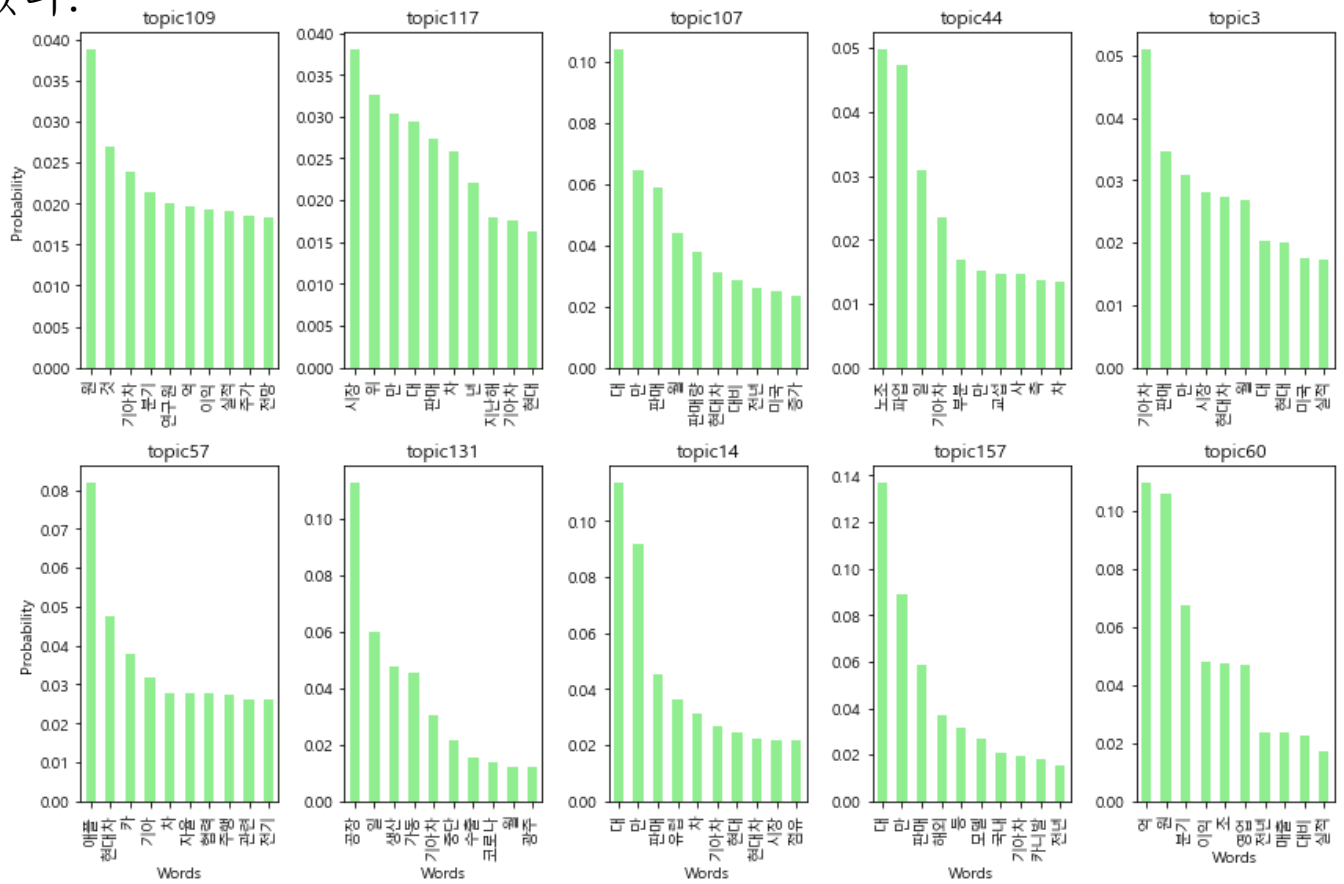
Ⅱ. 본론 - v. 주제분석



get_document_topics로 문서별 주제의 비율을 구한 후, 반복문과 pandas의 concat을 이용해 문서 전체에서의 주제의 빈도를 구했다. 좌측 그림은 빈도가 가장 높게 나타난 10개 주제의 빈도 시각화한 barchart이다.

109번 주제의 빈도가 700으로 가장 높게 나타났으며, 두번째로 높은 117번 주제의 빈도와 약 100의 차이가 난다.

이 주제들을 더욱 자세히 보기위해, show_topic을 이용해 주제별로 단어의 확률을 나타냈다.



빈도가 가장 높게 나온 109번 주제 먼저 살펴보면, 분기, 억, 이익, 실적, 주가, 전망이라는 단어의 확률이 높다. 기업의 상황을 보고 주가를 평가하는 주제로 예상된다.

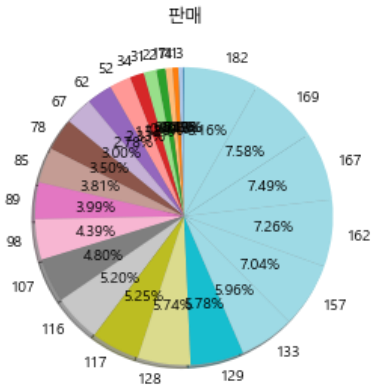
그리고 이어지는 주제들을 살펴보면, 판매, 노사갈등, 미래방향(전기차,자율주행), 코로나피해, 제품, 지난 기간 실적에 대한 주제의 빈도가 높게 나타났음을 알 수 있다.

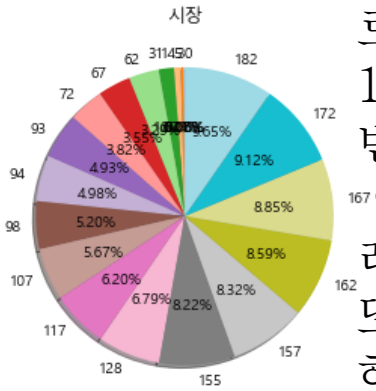
117번, 107번, 3번, 14번은 판매와 관련된 주제들이다. 주제 간에 약간의 차이가 있어 분리되었지만, 차원의 수가 더 작았더라면 판매에 관한 주제가 압도적 1위였음을 유추할 수 있다.

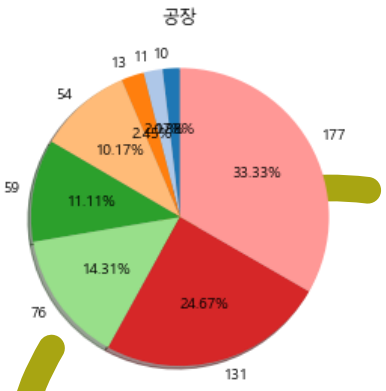
Ⅱ. 본론 - v. 주제분석

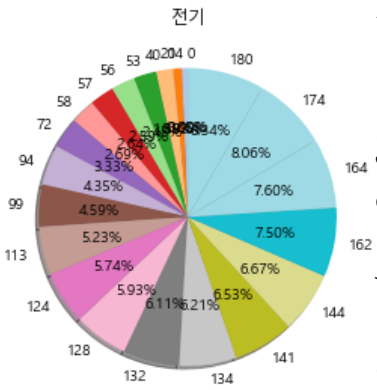
이 그림은 앞 단어빈도분석에서 살펴본 TfidfV-ectorizer로 만든 단어구름그림이다. 단어의 빈도가 높은 단어들의 주제를 살펴보겠다. 판매, 시장, 공장, 전기가 많이 등장하였으므로, 이 단어들의 주제를 분석해보겠다.

get_term_topics, token2id 메소드를 이용해 각 단어들이 어떠한 주제들에 포함되는지를 구해 pie chart를 그렸다. piechart를 분석해보면 단어 '판매' 와 '시장' 의 주제가 많이 겹치고 비슷한 분포를 갖고 있는 것이 보인다. 그리고 이 두 단어는 매우 많은 주제에 포함된다. 또한 한 주제에 많이 분포하는 것이 아닌, 많은 주제에 고르게 분포한다. 이와 다르게 '공장'은 적은 주제에 분포하며 고르게 분포하지 않는다. 177번주제와 131번 주제에만 약 58%가 분포한다. '전기' 역시 다양한 주제에 분포한다. 180번 주제에 약 9.3%, 174번 주제에 약 8.0%가 분포한다.







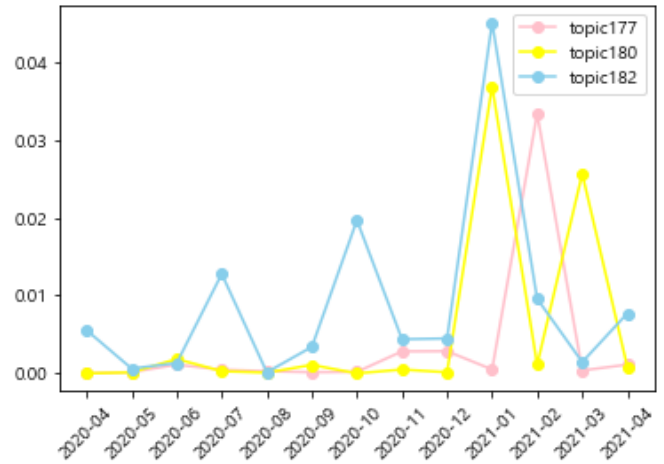


이 단어들이 가장 많이 등장한 주제의 월별로 등장한 정도를 살펴보기 위해 그래프로 나타냈다. 182번 주제의 경우 각 분기마다 2번째 달에 많이 등장하였다. 또한 2021년 1월에 꽤 많이 등장하였다. 분기별, 연도별로 판매 실적 등을 발표하기 때문으로 보인다.

177번 주제는 모두 낮게 나오다 21년 2월에 많이 등장하였다.

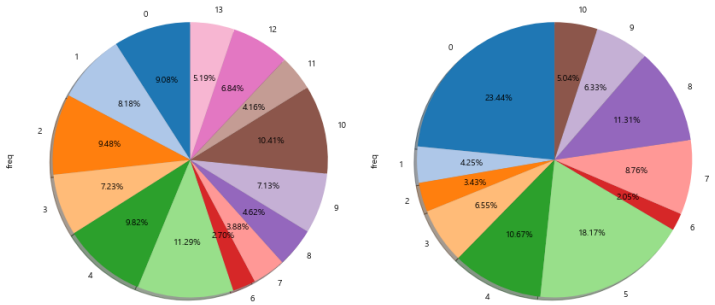
180번 주제는 모두 낮게 나오다가 21년 1월, 3월에 많이 등장하였다. 전기자동차 출시 관련 이슈들이 해당 달에 많이 등장하기 때문으로 보인다.

해당 주제들은 주로 21년에 많이 등장한 것을 알 수 있다.

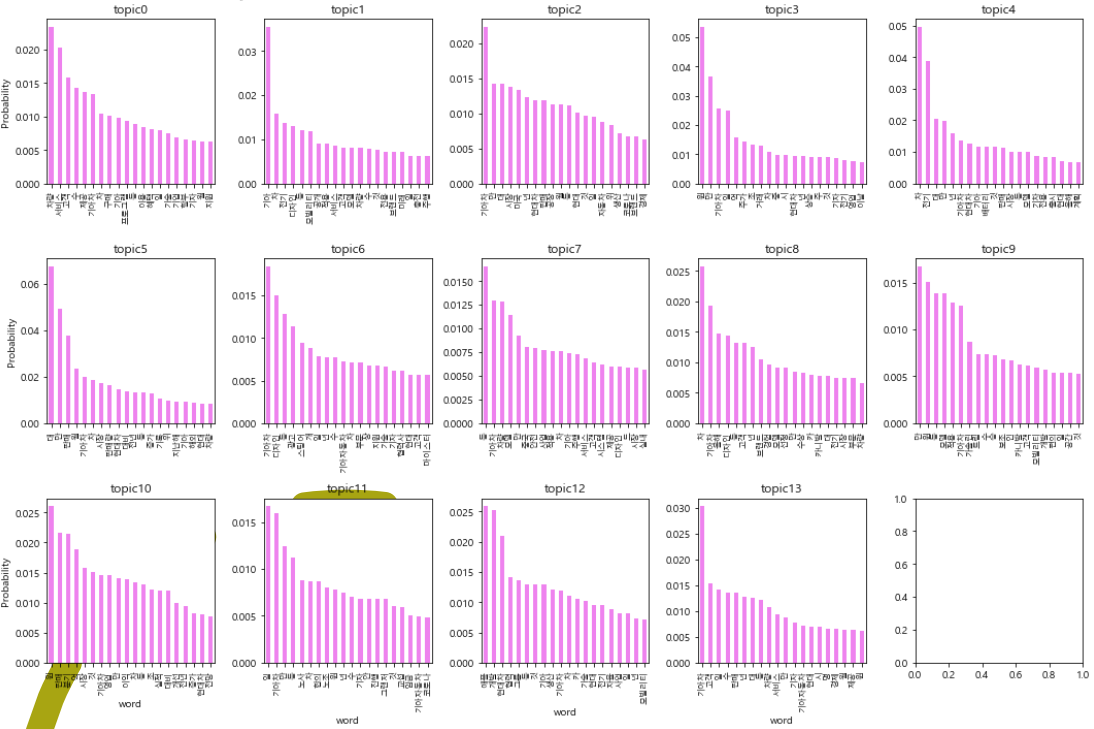


Ⅱ. 본론 - v. 주제분석

앞의 감성분석을 보면, 기사를 읽고 느끼는 감정(긍정/부정)이 주가 등락과 상관관계가 있음을 알 수 있었다. 지금부터는 감성에 따른 기사의 주제를 분석해보고자 한다. 이 역시 gensim모델을 이용해서 분석하였다. 전체 기사에서는 더미모델과 비교해보면 주제를 185개로 나누는 것이 가장 효율적이라고 나왔었는데, 긍정과 부정기사를 나눠서 적절한 주제 수를 찾아보았더니 대폭 줄어들은 14개, 11개였다.

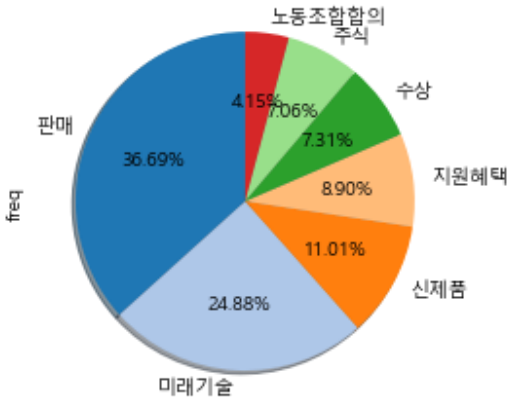


왼쪽 그림은 긍정, 부정 기사의 주제들의 분포이다. 긍정적인 기사는 고르게 잘 분포한 편이다. 대부분의 주제는 10%안팎을 차지한다. 반면 부정적인 기사는 0, 5번 주제가 23.4%, 18,2%로 높은 비율을 차지하고 있다. 그 다음으로 4, 7, 8번 주제가 10% 안팎으로 높은 비율을 차지하고 있으며, 나머지 주제들은 3~6.6%를 차지하고 있다.



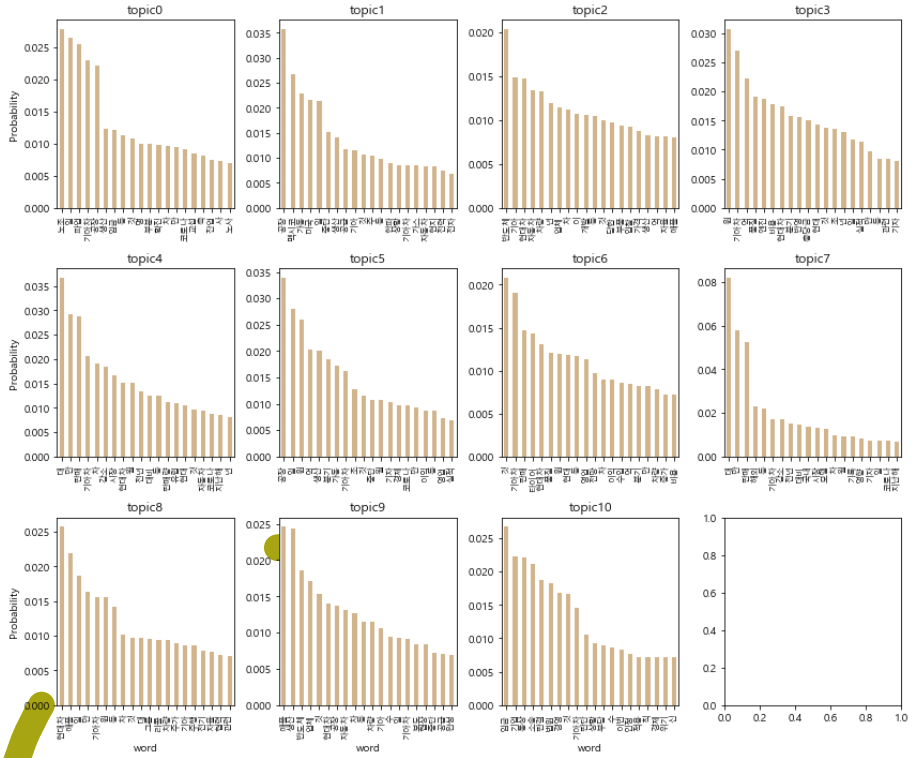
이제 긍정적인 기사는 어떠한 주제를 갖고 있는지 살펴보겠다. 0번 주제는 구매를 유도하기 위한 지원혜택에 관한 주제이다. 그리고 1, 4, 12번 주제는 미래기술(전기자동차, 자율주행)과 관련된 주제이다. 2, 5, 10, 13번 주제는 판매와 관련된 주제이고, 3번주제는 주식과 관련된 주제이다. 6, 8번 주제는 수상과 관련된 주제이다. 7, 9번 주제는 신제품과 관련된 주제이고 11번 주제는 노동조합과 좋은 관계를 유지한다는 소식에 관한 주제이다.

비슷한 주제별로 합쳐 piechart를 그렸다. 판매와 관련된 주제가 38,89%로 가장 높고 그 다음 미래기술, 신제품과 관련된 주제가 뒤를 잇는다.



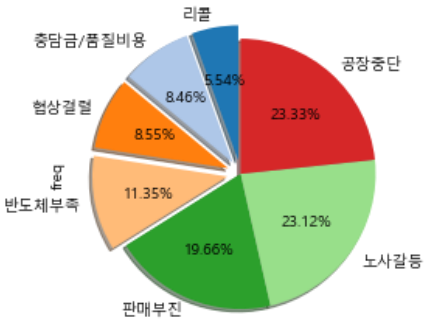
Ⅱ. 본론 - v. 주제분석

앞에서 부정적인 기사들의 주제 분포를 보았을 때 0번, 5번, 7번 주제가 가장 높은 비율을 차지하고 있었다. 오른쪽 그림은 이 3개의 주제가 월별로 언제 많이 등장했는지를 보여준다. 밑에서 설명하겠지만, 이 3개의 주제는 각각 노사갈등, 공장중단, 판매부진과 관련된 주제이다. 0번 주제는 평소에는 잠잠하다가 20년 9월, 11월에 많이 등장했다. 9, 11월에 노사갈등이 심했음을 알 수 있다. 5번 주제는 20년 4월에 가장 많이 등장했으며 나머지 기간에는 꾸준히 등장하고 있다. 코로나 바이러스 등의 원인으로 꾸준히 공장이 중단 되고 있음을 알 수 있다. 7번 주제는 20년 5, 7월에 많이 등장했고 21년 1월에 비교할 수 없게 많이 등장했다. 이를 통해 20년 1, 2분기에 판매가 부진했다가 3분기는 상황이 좋아졌음을 알 수 있다. 그리고 작년 한 해 판매가 매우 부진하여 21년 1월에 빈도가 매우 높음을 알 수 있다.



왼쪽 그림은 부정적인 기사들의 주제를 그래프로 나타낸 것이다. 0, 10번 주제는 노사갈등과 관련된 주제이다. 1, 5번 주제는 공장 가동 중단과 관련된 주제이다. 4, 6, 7번 주제는 판매 부진과 관련된 주제이다. 남은 2, 3, 8, 9번 주제는 모두 위기와 관련된 주제이다. 2번은 반도체 부족 상황 색이 더욱 강하고, 3번은 충당금, 품질비용 관련, 8번은 리콜, 9번은 협상결렬 관련한 성질이 강하다.

개별 주제별로 보았을 때에는 0, 5, 7번 주제가 가장 많았다. 하지만 비슷한 주제끼리 묶어서 보니, 위기관련 주제가 가장 높았고 그 다음이 공장중단, 노사갈등 주제가 약 23%로 비슷하게 많이 나왔고 그 다음은 판매 부진관련 주제가 많았다.



Ⅲ. 결론

웹스크래핑을 통해 데이터 수집 후, 전처리를 하고 단어빈도분석, 감성분석, 주제분석을 통해 기사의 내용이 주가의 변동과 어떠한 상관관계가 있는지 찾아보았다.

단어빈도분석의 결과를 보면, 주가가 상승한 주에서는 전기차, 판매, 수상과 관련된 단어들이 하락한 주보다 더 많이 등장하였다. 반대로 하락한 주에는 '분기', '지원', '이익', '가격', '억'이 가장 많이 등장하였다. 또한 '부담', '중단', '소송', '판결'도 많이 등장하였다.

감성분석의 결과를 보면, 카이제곱검정을 통해 주가 상승과 기사의 감성이 독립적인 관계가 아님을 밝혀냈다. 또한 주가가 상승한 주에 긍정적인 기사의 비율이 더 높게 나타났고, 주가가 하락한 주에는 부정적인 기사의 비율이 증가했음을 볼 수 있었다. 또한 긍정적인 기사가 작성되었을 때의 주가 상승한 경우의 비율이 부정적 기사의 비율보다 더 높게 나타났다. 그리고 주가가 상승한 주와 하락한 주의 단어 빈도를 보아도 긍정적 기사에 가중치가 음수인 단어가 적으며, 부정적 기사에 가중치가 음수인 단어들이 더 많이 나타났다. 이를 통해 주가 상승과 기사의 감성이 양의 방향의 상관관계를 가진다고 볼 수 있다.

주제분석의 결과, 긍정적인 기사는 대부분 판매와 미래기술, 신제품 관련 기사(약 72.5%)임을 알 수 있었다. 그리고 부정적인 기사는 크게 4가지 주제로 나눌 수 있는데, 위기상황, 공장가동중단, 노사갈등, 판매부진 순서로 비율이 높다는 것을 알 수 있었다. 이를 통해 주가 변동과 상관관계가 있는 인터넷 기사가 감성에 따라 어떤 주제들을 갖고 있는지 분석해 볼 수 있었다.

위 내용들을 종합해 보면 긍정적인 기사와 특정 주제들이 등장했을 때 주가가 상승하고 부정적인 기사와 특정 주제들이 등장했을 때 주가가 하락하는 관계가 있다고 볼 수 있다. 이는 인터넷 기사가 충분히 주가변동의 설명변수가 될 수 있음을 의미한다. 이를 활용해 다른 설명변수과 연관 지어 모델링을 한다면 주가의 변동을 예측할 수 있는 알고리즘을 구현하여 주식투자에 도움이 되는 프로그램을 만들 수 있을 것이다.

Ⅳ. 자기평가표

목차		점수	설명
Ⅰ. 서론		2/2	서론을 통해 주제, 현황, 문제점을 충분히 파악할 수 있음.
Ⅱ. 본론	ⅰ. 데이터 수집	3/3	수업시간에 배운 내용을 웹스크래핑에 충분히 반영하였고, 이에 더해 기사에 불필요한 내용을 제거하기 위해 추가적인 css선택자를 적용시키려 하였으나 그 과정에서 문제점을 발견해 해결을 위해 다른 여러가지 패키지들을 적용해 보았으며, 최종적으로 문제를 해결함.
	ⅱ. 데이터 전처리	3/3	Kiwipiepy패키지를 통해 형태소분석을 한 후, 기사 내용분석을 위해 명사만 걸러냄. 날짜데이터 이지만 문자형식인 데이터를 올바르게 변환시켰으며, 기사 분석에 발생하는 문제들을 해결함. 속보들의 내용이 없는 문제를 해결하였으며, 중복기사들을 제거함. 또한 분석에 차질을 일으키는 저작권표시 , 배포금지 문자 등을 기사 내용에 영향을 미치지 않고 제거함.
	ⅲ. 단어빈도 분석	3/3	단어의 빈도를 단어구름그림을 통해 잘 제시하고 해석했음. 이에 더해 주제에 알맞게 주봉데이터와 연계해서 주가변동에 따른 단어빈도를 제시했으며, 더욱 효과적으로 살펴보기 위한 방법을 개발하여 살펴봄.
	ⅳ. 감성분석	3/3	수업시간에 배운 내용을 충분히 반영하였으며, 데이터의 상황에 알맞게 데이터 분할, 평가척도를 사용함. 그리고 violin plot, bar chart, wordcloud를 사용해 가중치 분석을 잘 수행했음. 이에 더해 주제와 관련해 주가의 변동과 가중치의 관계를 보았으며, chi2_contingency를 이용해 타당성을 입증했음.
	ⅴ. 주제분석	3/3	gensim모델을 이용하여 주제분석을 함. 주제와 관련 지어 분석을 하기 위해 gensim패키지의 여러 메소드를 조사하여 응용해 데이터프레임을 만들고 시각화하여 분석하는 방법을 개발함.
Ⅲ. 결론		2/2	분석결과를 요약 후 주제와 관련되어 실행 방향 및 함의를 제시하였다.
총		19/19	