중간고사 텍스트 데이터 분석

주식 뉴스 기사 분석

빅데이터경영통계 20192780 유광열

A+

INDEX

TEXT DATA ANALYSIS

Introduction

TEXT DATA ANALYSIS

기업의 뉴스 제목을 분석하여 주식의 상승과 하락 예측, 투자 타이밍 이해

[현황]

- 코로나 19의 장기화
- 연일 집값 상승
- 빠른 재산 증식

⇒ 폭발적인 주식 거래 증가

[문제점]

⇒ But, 손해 보는 사람 증가

✓ 전문성 없는 투자

✓ 투기성 투자

✓ 정보 없는 투자

세부주제

- 텍스트 데이터 분석을 통해 상승과 하락의 많이 나오는 단어를 파악
- 감성분석을 사용하여 상승과 하락을 예측
- 주제분석을 통하여 상승과 하락의 관련 주제를 파악 타이밍을 파악

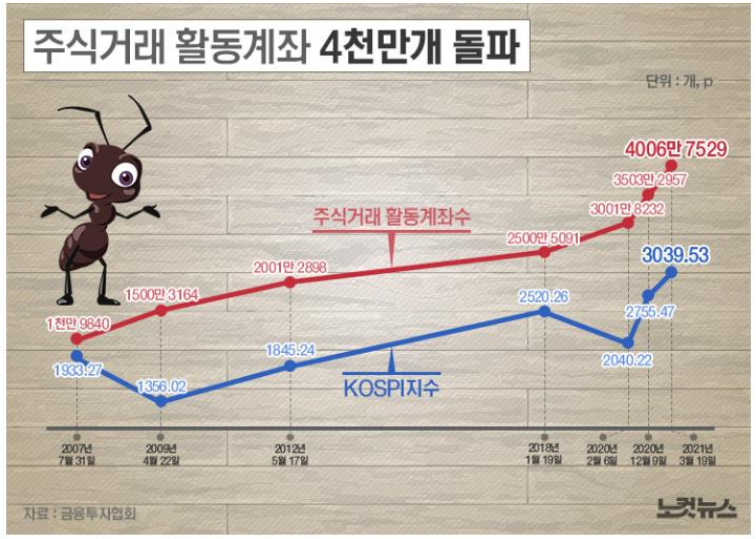
전 국민 주식투자 시대..계좌수 4천만개 돌파

올해만 500만개 증가

'성인 1인당 1계좌'...주식계좌수 첫 4000만개 돌파

특징주

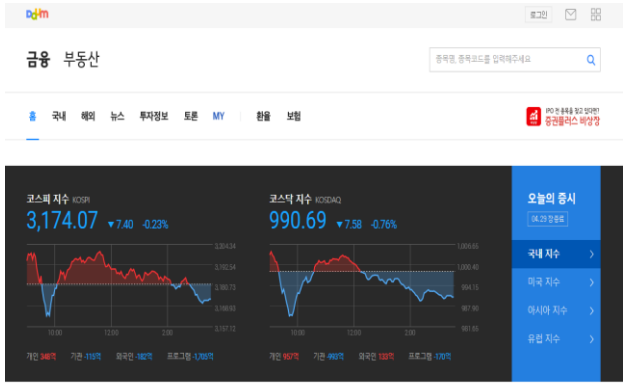
'주식 열풍' 증시 관련 대금만 5경, 전년대비 10.5% 증가



2.Web Crawling

TEXT DATA ANALYSIS

데이터 수집 사이트: 다음증권



Crawling Module: Selenium

1차 수집 데이터: 10000개

2차 수집 데이터: 3200개

데이터 저장: EXCEL

1,2차 excel파일

Labeling: 주식 차트보고 기입

(감성분석에서 자세히 설명)

추가적인 부분

1. Try, except를 사용 예외 처리
2. 검색어 입력 자동 페이지 이동
3. For문을 사용한 페이지 넘김 자동화

사이트 접속 → 키워드 입력 (삼성전자) → 제목, 시간, 신문사 데이터 수집 → 데이터 저장

4/12, 4/28일 총 2번에 걸려 시행

3.Data Processing

TEXT DATA ANALYSIS

▶데이터 결합:

- 4월 12일 데이터 4월 28일 데이터 결합
- 중복 값 제거

▶데이터 전처리:

- Re.sub을 사용하여 한글과 영어 소문자 외에 제외 (데이콘: NLP전처리 참고)

▶데이터 Labeling:

- 상승:1 하락:0 유지 제거

- 유지는 주말 또는 공휴일

(예외는 하루 존재 -> 삭제)

- New_labeling 생성 (감성분석에서 설명)

불용어 처리

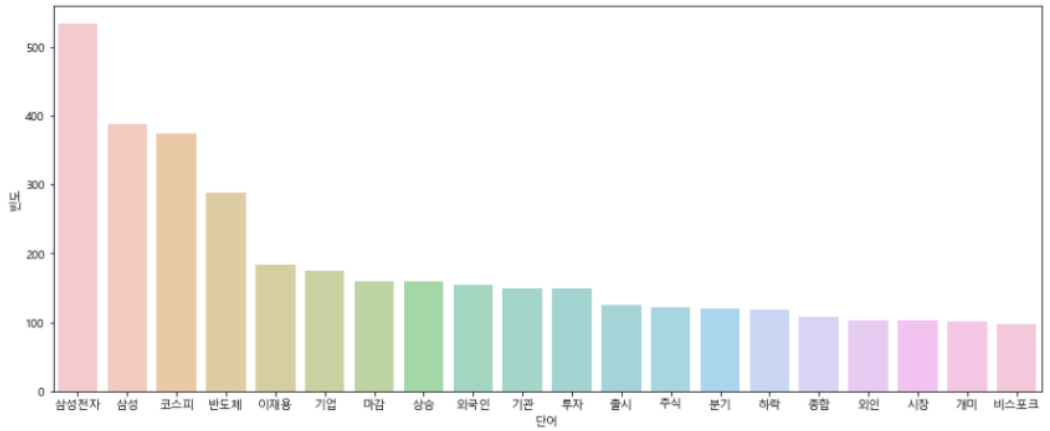
:총20개 부문을 살펴보고 처리

- 전체, label:0 , label:1,new_label:0, new_label:1
- Countervectorizr,TF_IDF, kiwi,stanza 사용

->총 20개 부문에서 많이 겹치는 단어 불용어 처리

불용어 :[삼성전자, 삼성, 코스피, 반도체]

Ex-stanza)



4.TDM & Word Cloud

TEXT DATA ANALYSIS

Label을 나누지 않은 전체 데이터

- CounterVectorizer
- TfidVectorizer
- Kiwi
- Stanza (사용예정)

-> 4개의 모델을 사용하여 감성분석에서 비교 후 사용

	단어	빈도
649	이재용	185.917710
133	기업	176.859576
247	마감	164.029774
414	상승	163.941886
600	외국인	158.719156
...	...	...
861	코트라	5.977592
339	보호청	5.926586
138	기초	5.858054
95	구내	5.714662
986	회계	5.688605

시각화

- 단어 구름 시각화
- 앞서 사용한 단어 막대 그래프 사용

분석 내용:

label에 구분 없이 단어 구름을 봤을 때 이재용, 기관, 상승,마감, 외국인 등이 눈에 띄게 보인다. 그러나 저 단어구름만 보고는 어떤 단어가 어디서 어떻게 많이 나왔는지는 알아보기 어렵다. 따라서 label과 new_label로 나누어 상승일 때와 하락일 때 어떤 단어가 많이 등장했는지 알아보자.

4.TDM & Word Cloud

TEXT DATA ANALYSIS

Original_Label | New_Label

상승

하락

Original

상승:

- 외국인과 기관 투자자들의 투자가 상승함을 알 수 있다.
- 강세, 돌파, 출발, 글로벌, 출시와 같은 긍정적인 단어들이 포함되어 있다.

하락:

- 이건희 회장이 별세함에 따라 주주총회가 열리고 상속세와 같은 문제로 주식이 하락함을 알 수 있다.
- 이재용 부회장이 재판을 받아 부정적인 영향을 끼치는 것을 알 수 있다.

New_label

상승: 위와 비슷하다.
하락: 외국인이라는 단어가 눈에 띄는 것을 볼 수 있다.

5.Sentiment Analysis

TEXT DATA ANALYSIS

New_label:

- 기존의 Label은 뉴스의 영향력을 잘 반영하지 못한다

-> 기사가 나오고 영향력이 미칠 때가지 3일 정도 소요

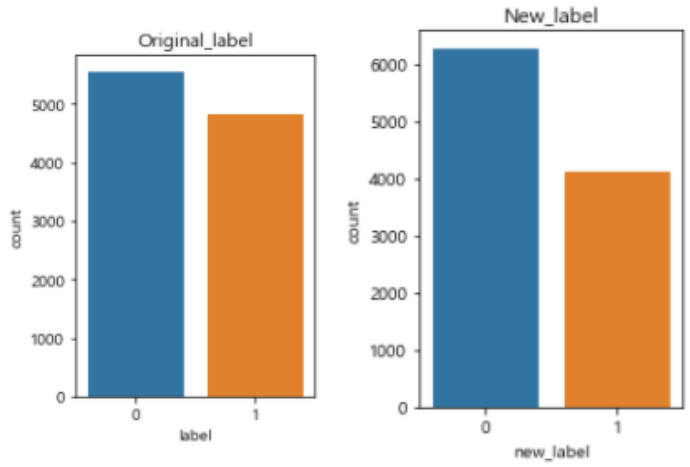
- 또한 상승세에서 잠시 주춤하는 하락 일 수 있다.

⇒ 따라서 이를 반영하는 새로운 Label 필요

⇒ 앞뒤 이틀 중에 최빈값으로 Label값을 대체함

	년도	월	일	label	new_label
0	2021	2	15	1	1
1	2021	2	16	1	1
2	2021	2	17	0	1
3	2021	2	18	0	0
4	2021	2	19	1	0
5	2021	2	22	0	0

Original_Label VS New_Label



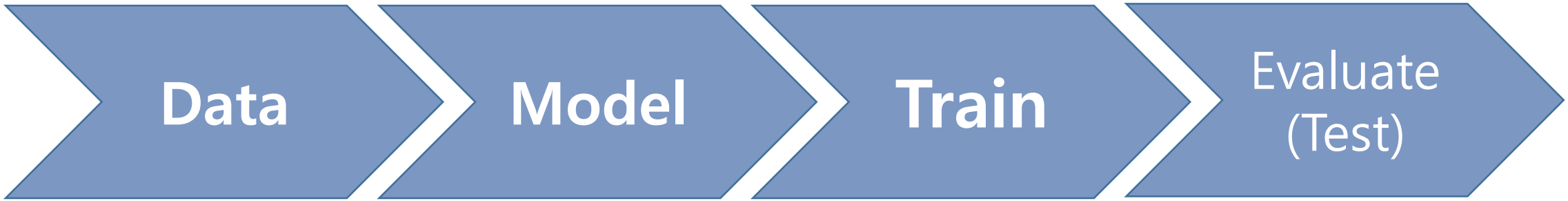
Original_Label VS New_Label

- Label값의 변동은 컸으나 비율은 많이 변화하지 않을 알 수 있다.
- 두 Label값 모두 상승 하락의 비율이 10%로 이내이므로 그래도 둔다.

	상승	하락
Org	47%	53%
New	40%	60%

5.Sentiment Analysis

모델 학습 과정



- Org_Label vs New_Label
- TDM
  - Countervectorizer
  - TfidfVectorizer
  - Kiwi
  - Stanza

- One_Layer_perceptron
- ActivationFuction:sigmoid
- Optimizer: Nadam
- Epochs = 100s

- Early_stopping:

->1%로 증가 X -> 10회 반복 후 종료

- ModelCheckpoint:

->Model에 accuracy가 최대일 때 저장

5.Sentiment Analysis

Best Choice

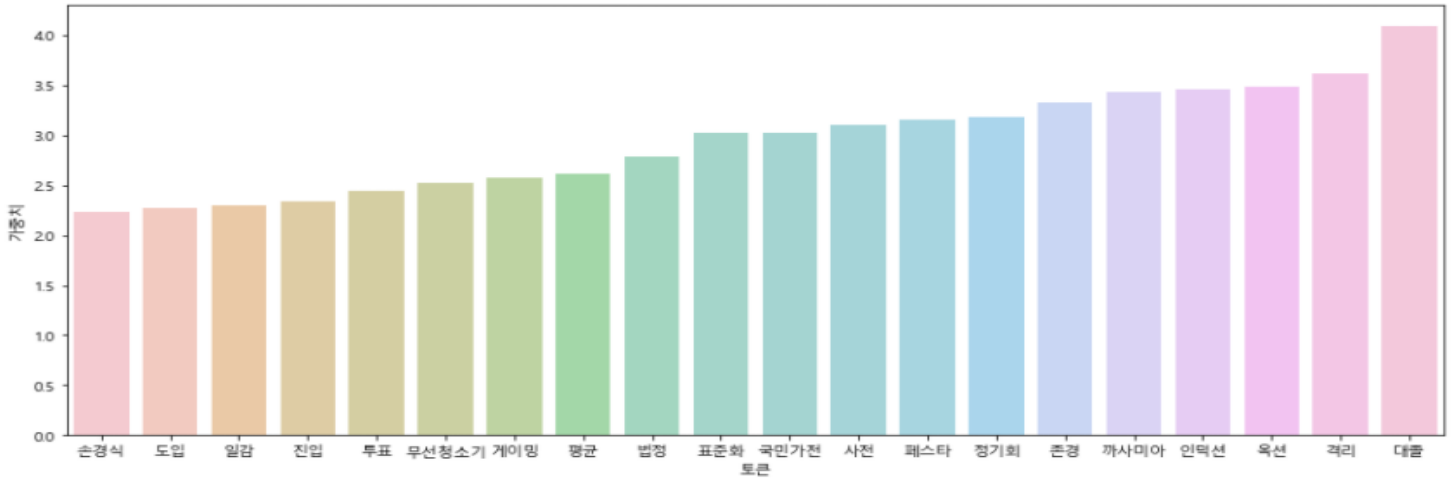
	Original Label	New Label
성능	0.6731	0.6854
해석	난해함	더 쉬움

	Counter Vectorizer	Tfidf vectorizer	Kiwi	Stanza
성능	0.6773	0.6603	0.6327	0.6854
해석	보통	보통	제일 난해함	제일 쉬움

Best Choice: New_Label & Stanza
Evaluate(Test) Accuracy: 0.6854

5.Sentiment Analysis *(긍정단어 & 해석)*

TEXT DATA ANALYSIS



- 대졸: 대졸자 등 연봉 11% 파격 인상
- 격리: 인도법인 항공비,자가격리 비용 지원
- 인덕션: 비스포크 홈 및 비스포크 제품 인덕션 출시
- 까사미아: 까사미아와 협업 매장 오픈
- 국민가전 & 페스타 : 국민가전 페스타 진행
- 무선청소기: 제트봇AI 매출 3배 성장
- 투표: 전자투표 시행
- 손경식: 이재용 사면 요구

→

- 처음 단어만 보고 어떠한 영향을 미치는지 알 수 없음
- 긍정이라고 나온 단어+ 삼성전자로 검색

↓

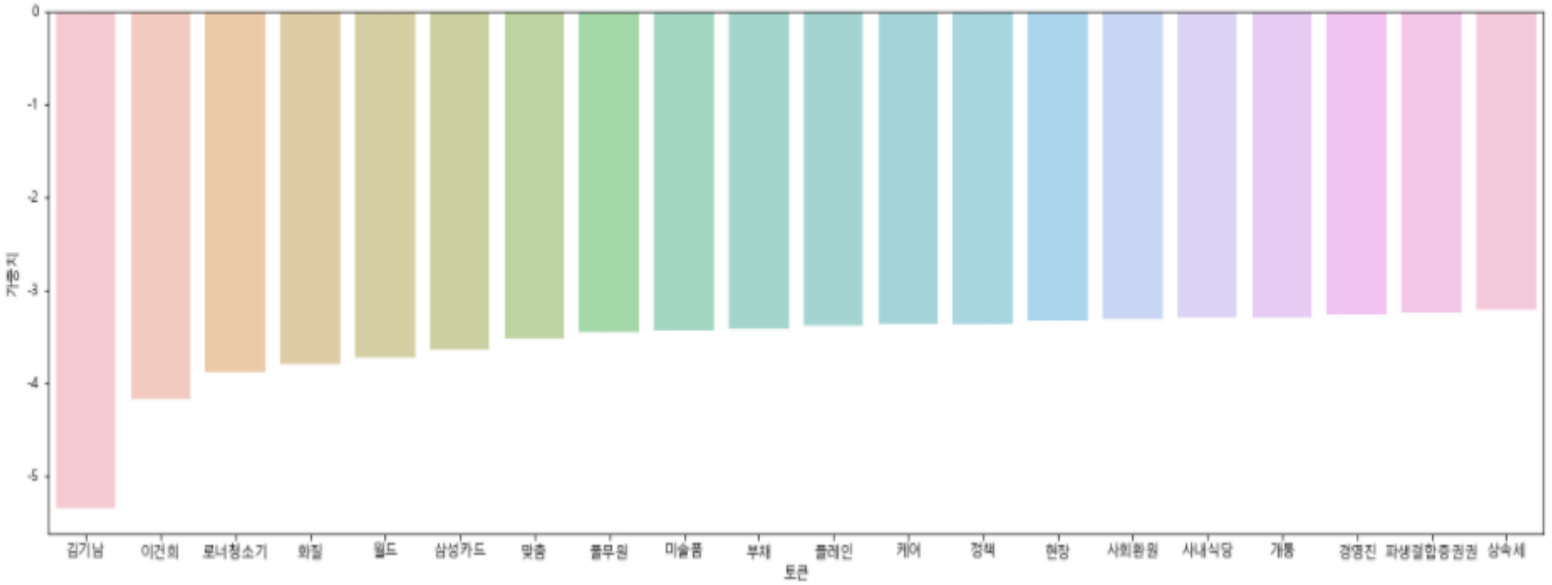
[해석]: 주식 상승에 영향을 미치는 요인

1. 파격적인 연봉 인상, 자가격리 비용 지원 사회적 이슈 증가
1. 신제품 출시와 그로 인한 매출 증가
2. 다른 기업과 협업
3. 국민가전 페스타와 같은 이벤트와 그로인한 매출 증가

5. 사회적으로 지위가 높은 사람의 발언 (그로인한 사회적 이슈)

5.Sentiment Analysis *(부정단어 & 해석)*

TEXT DATA ANALYSIS



[해석]: 주식 하락에 영향을 미치는 요인

1. 대기업의 오너의 별세
2. 유산 상속의 문제
3. 경영진 교체
4. 법적 다툼
5. 대표의 건강

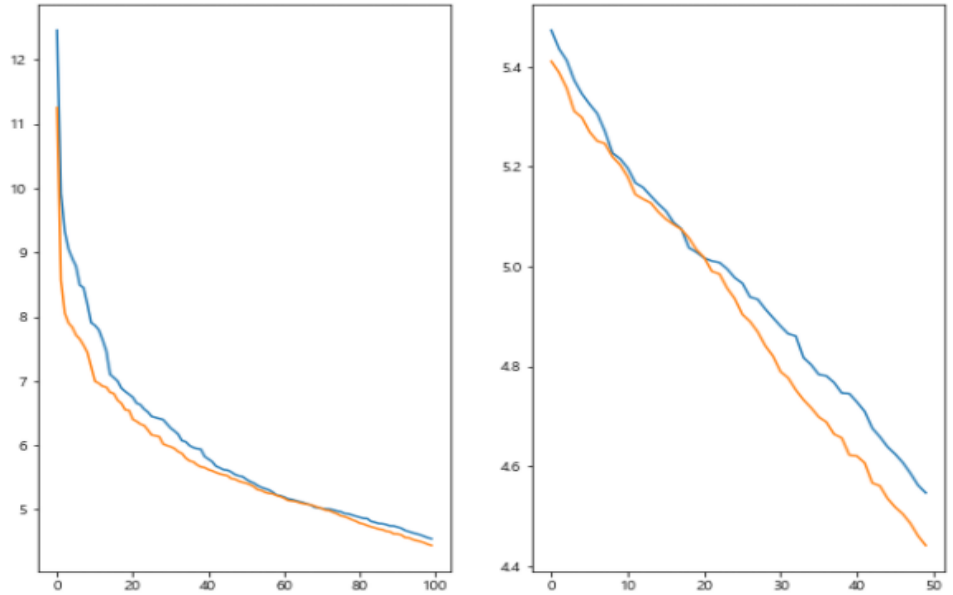
- 김기남: 삼성전자 부회장 김기남 이건희 회장 재산 상속 계획 발표
- 이건희: 삼성전자 회장 별세
- 미술품 & 사회환원 & 경영진 & 상속세: 이건희 회장 별세
- 경영진: 이재용 부회장 재판과 수술

→

이건희 회장의 별세와 이재용 부회장의 재판 및 건강 악화로 삼성전자 하락세임을 알 수 있다.

6.Topic Analysis *(LSA)*

TEXT DATA ANALYSIS



기존 단어분포 Vs Random으로 발생한 데이터

⇒ 60에서 70사이에 파랑색과 주황색이 만남

⇒ 이후로 설명력의 변화가 크게 없음

최종 차원의 수: 68번

주제분석:

- LSA & 병렬차원
- LSA_Rotator
- NMF
- LDA

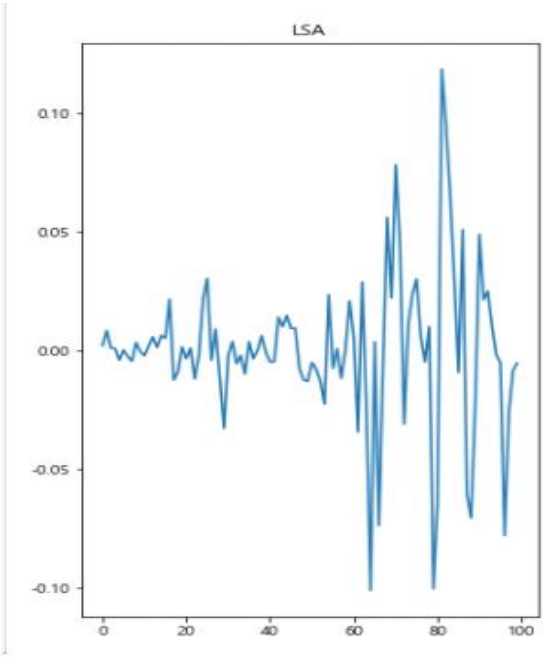
4가지를 사용하여 주제분석 실행

※ LDA를 제외한 나머지는 함수화를 시켜 topic입력 시 결과값 도출

6.Topic Analysis *(LSA & NMF & LSA_Rotator)*

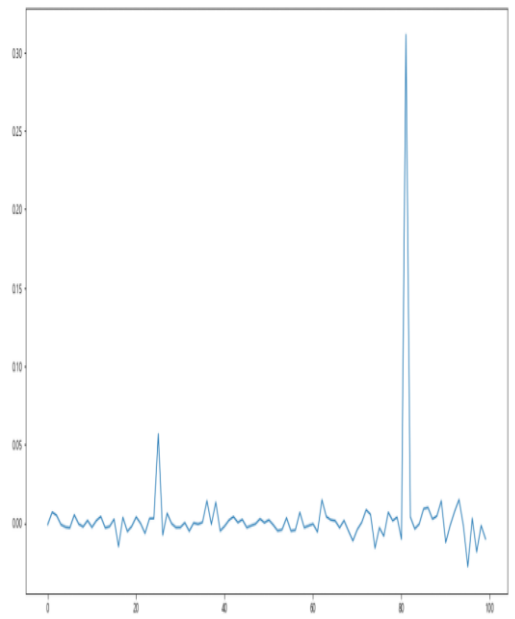
TEXT DATA ANALYSIS

ex) 감성분석에서 제일 긍정 단어 대졸



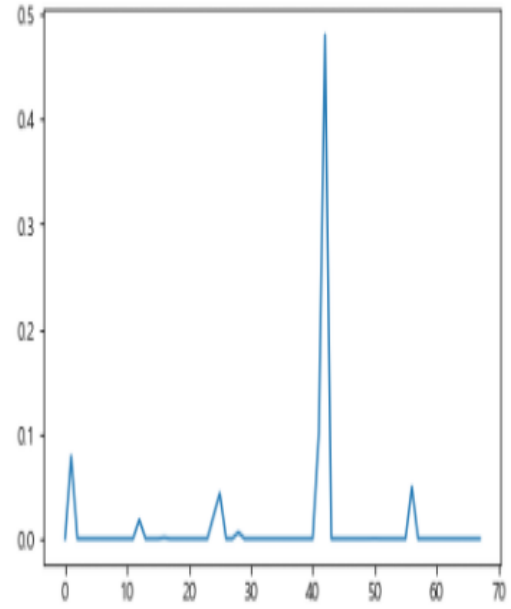
	word	loading
112	규모	0.046514
414	상승	0.047607
116	글로벌	0.048989
194	단체	0.056923
790	직원	0.062903
60	계약	0.071436
490	순매수	0.072716
632	이건희	0.077803
689	장비	0.078644
73	공급	0.084831
453	성장	0.091467
57	경제	0.110516
226	돌파	0.111065
572	연봉	0.125543
905	평균	0.140227
554	억원	0.143991
675	임금	0.178373
506	시장	0.181108
658	인상	0.209220
598	올해	0.459112

주제 차원:25



	word	loading
247	마감	0.052389
88	과제	0.052694
225	돌입	0.053329
979	확대	0.054653
55	경쟁	0.054775
356	분기	0.055812
839	출발	0.069645
501	시대	0.081632
576	열리	0.089885
620	유일	0.090691
244	롯데	0.107301
445	선정	0.134148
389	사원	0.144310
808	채용	0.153238
115	그룹	0.153522
518	신입	0.270636
208	대졸	0.311398
410	상반기	0.376621
505	시작	0.399879
85	공채	0.402732

주제 차원:81



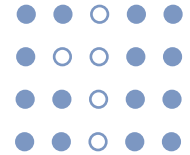
	word	loading
67	고문	0.330827
288	무보수	0.336800
405	삼성전기	0.350845
233	등기	0.356482
642	이상	0.383399
678	임직원	0.409552
483	수준	0.415197
597	올리	0.465537
208	대졸	0.480302
586	오르	0.531524
997	혹자	0.547017
677	임원	0.652061
110	권오현	0.675589
54	경영진	0.992168
780	지난해	1.707037
790	직원	2.014232
675	임금	2.497153
905	평균	2.608052
658	인상	3.674006
572	연봉	4.892972

주제 차원:42

※ 주제에 대한 차원과 해당 토픽이 다름을 알 수 있다.

6.Topic Analysis *(LDA분석)*

TEXT DATA ANALYSIS



```
[('갤럭시', 0.12106081), ('세계', 0.10496523), ('판매', 0.09796596), ('국가', 0.058723938), ('절반', 0.045509923), ('할인', 0.040880
07), ('최대', 0.038629513), ('진행', 0.038569797), ('위하', 0.03733282), ('최태원', 0.029482836)]

[('제출', 0.16951731), ('이익', 0.084058166), ('영업', 0.07992986), ('최고치', 0.05180974), ('신규', 0.04301932), ('이끌', 0.0368849
67), ('만전자', 0.034099597), ('고민', 0.033189885), ('공략', 0.026929699), ('마켓', 0.026149506)]

[('사회', 0.14205928), ('세금', 0.06717212), ('전쟁', 0.04790096), ('공식', 0.04664193), ('정부', 0.035678256), ('마련', 0.03040874
7), ('억원', 0.029791178), ('백신', 0.027199972), ('개월', 0.027061425), ('기업공시', 0.026123138)]

[('순매수', 0.08217767), ('코스닥', 0.065990366), ('공모', 0.06421989), ('관련', 0.056749485), ('재산', 0.04768064), ('등락', 0.0468
01444), ('팔자', 0.040086087), ('하회', 0.03910949), ('종목', 0.035025246), ('기관', 0.03180047)]

[('로너청소기', 0.08670887), ('소송', 0.05068413), ('거래소', 0.046933364), ('유일', 0.046258174), ('순매', 0.044613484), ('향방',
0.044460427), ('내주', 0.035892744), ('이벤트', 0.030871473), ('가격', 0.027877089), ('늘리', 0.027416969)]

[('핵심', 0.048936892), ('맞춤', 0.047188208), ('국민연금', 0.04548797), ('이사', 0.044990577), ('기회', 0.042116966), ('관심', 0.04
2027693), ('성장', 0.040423322), ('수혜', 0.03438809), ('회의', 0.03121529), ('기대감', 0.02982843)]

[('기관', 0.08649751), ('외국인', 0.07627271), ('투자', 0.06399883), ('매도', 0.044558126), ('외인', 0.04190858), ('동반', 0.0408559
03), ('매수', 0.038787093), ('안착', 0.034469835), ('코로나', 0.03395013), ('오후', 0.032428645)]

[('이건희', 0.3480799), ('기부', 0.09287699), ('조원', 0.0882044), ('역대', 0.07441295), ('최대', 0.044247862), ('오늘', 0.03406846
5), ('최고', 0.031790853), ('사상', 0.03122146), ('지키', 0.03023134), ('경신', 0.021265324)]

[('이재용', 0.14467101), ('부회장', 0.076279745), ('카카오', 0.053846158), ('시장', 0.034674175), ('오전', 0.029060753), ('단계', 0.
028947525), ('액면분', 0.02715094), ('빅데이터', 0.0270206), ('복귀', 0.026631221), ('시즌', 0.02533232)]

[('의료', 0.10204994), ('공개', 0.081855334), ('규모', 0.069271855), ('바이든', 0.04349426), ('인텔', 0.041198067), ('가전', 0.03891
1834), ('공급', 0.037165496), ('장비', 0.034419302), ('투자', 0.030984337), ('배터리', 0.024947949)]
```

- LDA 분석:
- Num_topics = 25
- Compile: token_pattern
- Filter_extream no_below:10, no_above:0.9

[긍정적 해석]

- 삼성전자 영업이익 최고치
- 외인,개인 주식 매수 주가 회복
- 인텔 부사장 방, 배터리 가전 투자
- 비스포크 가전 제품 출시

[부정적 해석]

- 이건희 회장 역대 최대 기부
- 삼성전자 주주총회

6.Topic Analysis *(긍정단어)*

TEXT DATA ANALYSIS

긍정 주제분석 키워드: 대졸, 상승, 까사미아, 목표

대졸
상승,공채,계약,성장,
연봉,임금,인상,채용,
전망,신입,흑자

상승
출발,회복,외국인,
기관,금리,마감,
대형주,매수세,전환

까사미아
수익,목표,주가,억원,
최고,성장,역대,주목,
매출,오픈,흑자,이상

목표
지난해,수익,배터리,
매출,주가,성장,속도,
확대,배당,성장

[해석]:

주가상승에 대한 긍정적 단어와 연관성 (감정분석 & 단어구름)

1. 상반기 파격적인 임금 상승으로 인한 공채 지원 상승 등 사회적 이슈 증가
2. 외국인, 기관 투자자들의 투자 증가로 주식 회복으로 전환 & 1분기 마감
3. 까시미아와 협동으로 오픈 목표 수익 역대 최고 성장 최고 매출 주목 받음
4. 지난해 수익 대비 삼성전자 매출 증가 빠른 속도로 성장세

6.Topic Analysis *(부정단어)*

TEXT DATA ANALYSIS

부정 주제분석 키워드: 이건희,이재용,상속세,정책

이건희
지분,납부,사회한원,
미술품,역대,기증,상
속,유산,발표,상속세

이재용
사면,회계,수사,출석,
구치소,투약,불법,
프로포폴,수술,재판

상속세
지분,납부,사회환원,
역대,미술품,상속,
유산,기증

정책
규제,대책,사태,생산,
대란,부족,지원,공급

[해석]:

주가하락에 대한 부정적 단어와 연관성 (감정분석 & 단어구름)

1. 이건희 회장이 죽으며 재산을 사회에 환원, 유산,상속세 등 문제 발생
2. 이재용 회장의 여러 죄로 수사와 재판 중, 건강 악화 수술 예정
3. 이건희 회장이 남긴 많은 재산의 상속세와 재산의 사회환원
4. 코로나19로 인한 반도체 생산 재료 공급 물량부족, 정책 규제

7.Conclusion

*주가 상승과 하락 예측 & 투자 타이밍

[주가 상승: 긍정적인 영향]

- 신제품 출시와 그에 따른 매출이 증가
- 기관, 외국인, 외인 투자자들이 급증
- 연금 인상, 비용 지원 등 긍정적인 사회적 이슈
- 다른 기업과 협업과 그에 따른 매출
- 기업의 이벤트로 인한 매출 상승

[주가 하락: 부정적인 영향]

- 그룹 회장의 별세
- 상속세와 지분 배분으로 인한 혼란스러움
- 경영진 변경으로 인한 주주총회
- 대표 경영진의 법적 공방과 건강악화

*아쉬운점

- 해당 기업의 부정적인 영향이 회장의 별세로 한정적이었다.
- 주식 거래 자체가 종목이 제한적이라 다른 종목에 대한 일반화는 불가능하다고 판단된다.
- 여러 기업과 종목을 분석해 보지 못했다.

*결론: 주가 상승과 하락에 미치는 영향을 뉴스기사를 통한 텍스트 데이터 분석으로 알아보았다. 주식 거래 시 해당 기업의 정보를 뉴스 기사를 통해 알고, 긍정적인지 부정적인지 알아야 할 필요가 있다. 텍스트 데이터를 통해 분석한 내용은 실제로 주가 형성에 많은 영향을 끼친다. 뉴스 기사가 주가 형성에 많은 영향을 끼치는 만큼 해당 기업의 정보에 귀를 기울여야 할 필요가 있다.

8. Self_Fedback

TEXT DATA ANALYSIS

항목	점수	평가 근거
서론	2/2	그래프와 뉴스자료 사용으로 이해도와 전문성을 높임
데이터 수집	3/3	Try/except문을 사용하여 예외처리를 함 For문을 사용하여 자동화 시킴 Driver에 접근 하지 않아도 Ipynb파일에서 자동으로 실행
전처리	3/3	Re.sub을 사용하여 전처리를 문자 외것을 전처리함 기존 Label문제점을 발견하고 새로운 Label를 생성 불용어 처리서 다방면을 보고 불용어 처리를 함
단어 빈도	3/3	단어구름을 한가지 방법이 아닌 여러 방법을 사용하여 만듬 Label값에 따라 Word Cloud를 분석함 막대 그래프도 사용하여 더 정확한 수치화한 시각화를 함
감성분석	3/3	새로운 여러 Label과 TDM으로 감성부석 실시 Early_stopping과 Checkpoint사용

8. Self_Fedback

TEXT DATA ANALYSIS

항목	점수	평가 근거
주제분석	3/3	다양한 방법을 사용한 주제분석을 진행함 LDA를 제외한 여러 함수를 자동화 하여 빠른 결과물을 도출함
결론	2/2	해당 주제와 목적에 맞는 결과물을 도출함 서론 및 본론에 충실하게 작성함 아쉬운 점을 통해 부족한 점을 찾음
합계	19/19	

출처	
데이터 전처리	https://dacon.io/codeshare/1808
Early_stopping & Model_Checkpoint	https://3months.tistory.com/424
Re.compile	https://notebook.community/zzsza/Datascience_School/19.%20%EB%AC%B8%EC%84%9C%20%EC%A0%84%EC%B2%98%EB%A6%AC/04.%20%EB%AC%B8%EC%84%9C%20%EC%A0%84%EC%B2%98%EB%A6%AC https://datascienceschool.net/03%20machine%20learning/03.01.03%20Scikit-Learn%EC%9D%98%20%EB%AC%B8%EC%84%9C%20%EC%A0%84%EC%B2%98%EB%A6%AC%20%EA%B8%B0%EB%8A%A5.html

Thankyou

A+