

주식의견 분석

- 스파크 데이터프레임을 이용한 대용량 데이터 처리-

개요

- 동일한 조건과 환경임에도 불구하고, 성별 및 연령별로 수익률과 회전율에 차이가 발생.
- 이는 곧 투자자들의 투자 성향이 성별 및 연령별로 나뉜다는 것을 의미함.
- 현재 주가지수에 대한 투자자의 의견 데이터를 통해서 인사이트를 얻고자 함.

데이터 수집 및 전처리

데이터 수집 및 전처리

데이터 수집

- 주요 주식 커뮤니티에서 게시글 혹은 댓글 크롤링.

**재원 유**
2 분전

이번 상승은 테슬라로 시작해 테슬라로 끝난다 테슬라 올려서 물린놈들 또 몰타기 시켜서 지옥으로 끌고가면서 7k스닥

답글  0  0 신고하기

**재원 유**
방금 전









상승률은 테슬라보다 팅슬라쪽이 더 좋을것

답글  0  0 신고하기

**뽀빠맨**
10 분전

뽀빠맨이 우려하는 1월초 상단으로의 움직임발생시 뜬금포로 매수포 수익률이 달달하겠네 그걸 대비한 테슬라매수지만 연말까진저가구조는 지켜줘야 시나리오가맞다고다 매수는 연말마지막 거래일 로또매수도 좋아보인다 기한은1월12일까지 카이 아님말고~

답글  0  0 신고하기

4829927	매매공		미국 주식 포트폴리오 매매 일지 75일차
4829925	일반		코스피 개미를 패닉셀 보니까 미주겔 놈들이랑 차이...
4829924	일반		중 ㅋㅋ 줄이니까모 ㅋㅋ 르니까 [1]
4829923	일반		환율 1250 가능?? [1]
4829922	일반		이런 경기침체는 ㄹㅇ 처음 같음
4829921	일반		voo 배당금 입겔 ㅅㅅㅅㅅ [5]
4829920	일반		N챌린지는 해볼필요도없겠네 ㅋㅋ
4829919	일반		아니 너네 똑똑함? [4]

	Date	Title	View	Reference	Market	UserAge	UserGender
0	2022-10-31	알루알루 ~ 뽀 가신다 ~	0.0	investingcom	nasdaq	4050	M
1	2022-10-31	ㄷㄱㄷ~	0.0	investingcom	nasdaq	4050	M
2	2022-10-31	노뽀형 룡 어디다 걸었어? 평단 몇이야?	0.0	investingcom	nasdaq	4050	M
3	2022-10-31	노뽀형 기절 했어? 숨서! 숨! ㅎㅎㅎ	0.0	investingcom	nasdaq	4050	M
4	2022-10-31	답이 늦었네요 예쎈피 주봉 기준입니다	0.0	investingcom	nasdaq	4050	M

데이터 수집 및 전처리

데이터 수집

- 약 1,000만 개의 의견 수집 완료.

Date	Title
2022-10-31	알루알루 ~ 뽕 가신다 ~
2022-10-31	ㄷ ㅎ ㄸ ~
2022-10-31	노뽕형 롱 어디다 걸었어? 평단 몇이야?
2022-10-31	노뽕형 기절 했어? 숨서! 숨! ㅎ ㅎ ㅎ
2022-10-31	답이 늦었네요 예썬피 주봉 기준입니다
...	...
2020-04-11	ig.com wallstreet weekend 이요. 근데 거래량도 작고 비공식이라...
2020-04-11	담주 기대되네요.
2020-04-11	그거 적중도 그냥 쓰레기임
2020-04-11	좀더 올리자. 그래야 10%하락 할거 아니냐. 여기서 찔끔 5%빼고 다시 올리고 하...
2020-04-11	레버는 오일이 아직 기회가 있어보임.
9698141 rows × 1 columns	

데이터 수집 및 전처리

데이터 전처리

- 특수 문자 제거 후, string 타입의 데이터에 대해 명사만 추출

Date	Title
2022-10-31	알루알루 ~ 뽕 가신다 ~
2022-10-31	ㄷ ㅎ ㅏ ~
2022-10-31	노뽕형 룡 어디다 걸었어? 평단 멋이야?
2022-10-31	노뽕형 기절 했어? 숨서! 숨! ㅎㅎㅎ
2022-10-31	답이 늦었네요 예쎄피 주봉 기준입니다
...	...
2020-04-11	ig.com wallstreet weekend 이요. 근데 거래량도 작고 비공식이라...
2020-04-11	담주 기대되네요.
2020-04-11	그거 적중도 그냥 쓰레기임
2020-04-11	좀더 올리자. 그래야 10%하락 할거 아니냐. 여기서 찔끔 5%빼고 다시 올리고 하...
2020-04-11	레버는 오일이 아직 기회가 있어보임.

9698141 rows × 1 columns



Date	Title
2022-10-31	[아]
2022-10-31	[]
2022-10-31	[노, 뽕, 평단, 이]
2022-10-31	[노, 뽕, 기절, 숨서, 숨]
2022-10-31	[답, 예쎄, 피, 주봉, 기준]
...	...
2020-04-11	[거래량, 공식, 정도, 보]
2020-04-11	[담, 주, 기대]
2020-04-11	[적중, 쓰레기]
2020-04-11	[하락, 할거, 선물, 사람]
2020-04-11	[레, 버, 오일, 기회]

데이터 수집 및 전처리

데이터 전처리

- 각 일자별로 그룹화 진행

1) List extend / append : 약 30시간 소요

2) Pandas groupby : 약 40분 소요

3) Spark Dataframe groupby : 약 30초 소요

Date	collect_list(Title)
2022-10-31 00:00:00	[['야'], [], ['노', ...]
2022-10-30 00:00:00	[['조선', '총기', '합법...
2022-10-29 00:00:00	[['충격', '뚝', '영상' ...]
2022-10-28 00:00:00	[['오늘', '골드', '재미...
2022-10-27 00:00:00	[['달러', '원', '웨이' ...]
2022-10-26 00:00:00	[['상', '폐', '전', ...]
2022-10-25 00:00:00	[['저항', '것', '숫', ...]
2022-10-24 00:00:00	[['쌈'], [], [], [...]
2022-10-23 00:00:00	[['새'], ['포', '공파...
2022-10-22 00:00:00	[['폭락', '안심', '때' ...]

only showing top 10 rows

데이터 수집 및 전처리

데이터 전처리

- 각 키워드가 2음절 이상인 경우만 추출 후, 키워드 빈도 카운트

```
[2022-10-28 00:00:00| ('애플', 1499)| ('오늘', 1001)| ('아마존', 542)| ('실적', 478)| ('나스닥', 441)| ('인텔', 352)| ('메타', 346)| ('미국', 334)| ('주식', 307)| ('금리', 290)| ('하락', 273)| ('매수', 242)| ('상승', 241)| ('이유', 224)| ('사람', 214)| ('지수', 212)| ('달러', 209)| ('생각', 205)| ('지표', 201)| ('테슬라', 198)| ('시작', 188)| ('다우', 179)| ('발표', 176)| ('새끼', 167)| ('국장', 162)| ('이번', 159)| ('반등', 156)| ('지금', 151)| ('차트', 151)| ('간다', 149)| ('시발', 146)| ('시장', 145)| ('침체', 139)| ('다음', 136)| ('채권', 131)| ('수익', 129)| ('하네', 126)| ('프로', 125)| ('주가', 125)| ('양전', 124)| ('시간', 122)| ('바닥', 122)| ('매도', 120)| ('캘리', 118)| ('폭락', 117)| ('개미', 117)| ('경기', 117)| ('예상', 115)| ('충이', 112)| ('중국', 111)| ('매매', 108)| ('정도', 104)| ('병신', 102)| ('테크', 101)| ('인플레', 99)| ('선물', 98)| ('감사', 98)| ('거지', 98)| ('일본', 97)| ('선거', 92)| ('가자', 92)| ('연준', 92)| ('때문', 90)| ('기업', 87)| ('반도체', 87)| ('가이던스', 87)| ('물가', 86)| ('이다', 86)| ('분기', 82)| ('인상', 81)| ('투자', 81)| ('중간', 80)| ('푸틴', 80)| ('바이든', 78)| ('양봉', 78)| ('본장', 78)| ('요즘', 77)| ('가격', 77)| ('경제', 77)| ('내일', 77)| ('11월', 74)| ('이상', 74)| ('나락', 73)| ('포지션', 73)| ('자리', 73)| ('여자', 73)| ('기대', 72)| ('스위칭', 72)| ('아이폰', 72)| ('코스', 71)| ('한국', 70)| ('파월', 69)| ('반영', 69)| ('하방', 67)| ('느낌', 66)| ('이제', 65)| ('12월', 64)| ('마소', 64)| ('미주', 63)| ('하루', 62)]  
[2022-10-27 00:00:00| ('메타', 1434)| ('오늘', 677)| ('애플', 410)| ('나스닥', 259)| ('주식', 247)| ('실적', 234)| ('금리', 206)| ('미국', 194)| ('새끼', 187)| ('달러', 185)| ('이유', 169)| ('인텔', 169)| ('지표', 154)| ('테슬라', 150)| ('시발', 143)| ('상승', 141)| ('하락', 133)| ('사람', 133)| ('구글', 121)| ('생각', 117)| ('매수', 116)| ('발표', 114)| ('반도체', 113)| ('마소', 104)| ('시작', 102)| ('내일', 101)| ('다우', 96)| ('경제', 95)| ('하네', 85)| ('충이', 83)| ('양전', 82)| ('아마존', 82)| ('채권', 82)| ('시장', 79)| ('지수', 76)| ('지금', 74)| ('예상', 74)| ('요즘', 73)| ('차트', 73)| ('주가', 72)| ('병신', 68)| ('이번', 67)| ('미주', 67)| ('반등', 66)| ('국장', 65)| ('엔비디아', 63)| ('실업', 63)| ('간다', 62)| ('트위터', 62)| ('본장', 62)| ('침체', 62)| ('테크', 61)| ('환율', 61)| ('운지', 60)| ('프로', 59)| ('중국', 58)| ('인상', 58)| ('캘리', 57)| ('매도', 57)| ('커버', 57)| ('경기', 57)| ('개미', 55)| ('평단', 54)| ('속보', 54)| ('보니', 54)| ('민지', 54)| ('한국', 54)| ('기업', 52)| ('폭락', 51)| ('때문', 50)| ('10년', 50)| ('물가', 49)| ('호재', 48)| ('만원', 48)| ('메타버스', 48)| ('유럽', 47)| ('바이든', 46)| ('파월', 46)| ('가격', 45)| ('이랑', 45)| ('정도', 45)| ('인플레', 45)| ('선거', 44)| ('여자', 44)| ('이다', 43)| ('매매', 42)| ('이상', 42)| ('수익', 42)| ('거지', 42)| ('바닥', 41)| ('시간', 41)| ('노스', 40)| ('니들', 39)| ('스위칭', 38)| ('스타', 38)| ('게이', 36)| ('양봉', 36)| ('미래', 36)| ('연준', 36)| ('보잉', 35)]
```


EDA

EDA

모든 단어들에 대한 워드클라우드

: 모든 단어들을 일렬로 나열.

총 23,062,728개의 단어를
추출하였음.

Date	Title
2022-10-31	알루알루 ~ 뽕 가신다 ~
2022-10-31	ㄷ ㅎ ㅏ ~
2022-10-31	노뽕형 룡 어디다 걸었어? 평단 몇이야?
2022-10-31	노뽕형 기절 했어? 숨서! 숨! ㅎ ㅎ ㅎ
2022-10-31	답이 늦었네요 예쎄피 주봉 기준입니다
...	...
2020-04-11	ig.com wallstreet weekend 이요. 근데 거래량도 작고 비공식이라...
2020-04-11	담주 기대되네요.
2020-04-11	그거 적중도 그냥 쓰레기임
2020-04-11	좀더 올리자. 그래야 10%하락 할거 아니냐. 여기서 찔끔 5%빼고 다시 올리고 하...
2020-04-11	레버는 오일이 아직 기회가 있어보임.
9698141 rows × 1 columns	



0
0 평단
1 기절
2 숨서
3 예쎄
4 주봉
...
23062723 할거
23062724 선물
23062725 사람
23062726 오일
23062727 기회
23062728 rows

: 수집한 모든 단어들에 대한 워드클라우드 생성.

- 양전 코스시작 매도 계좌 속보매매단타
평단 생각미국 코로나 중국지수 주주주가한국 시장
국장매수 금리 오늘 나스닥 테슬라
시드 애플 하락 내일상승 지금이번 차트 주식
이상 선물 잡주반등 수익 조정 폭락프로 종목
호재 본장 요즘 얼마 다음 투자 추천
파월

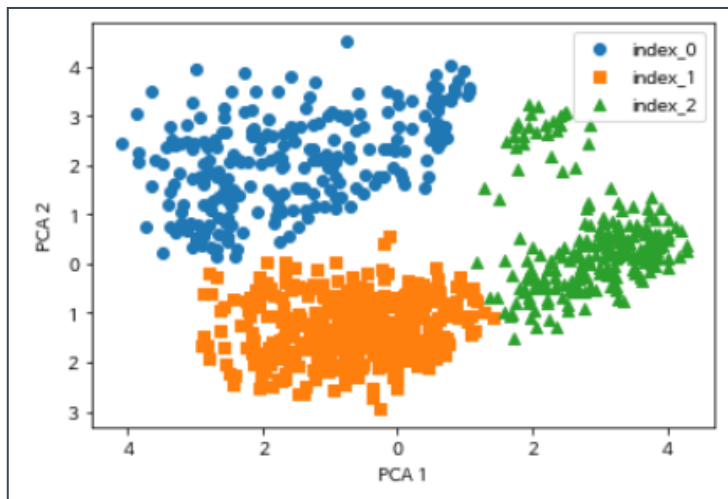
EDA

군집분석

- 각 일자별 게시글 집합을 하나의 문서로 생각하고 군집분석 진행. (K-Means)

1) Bag Of Words

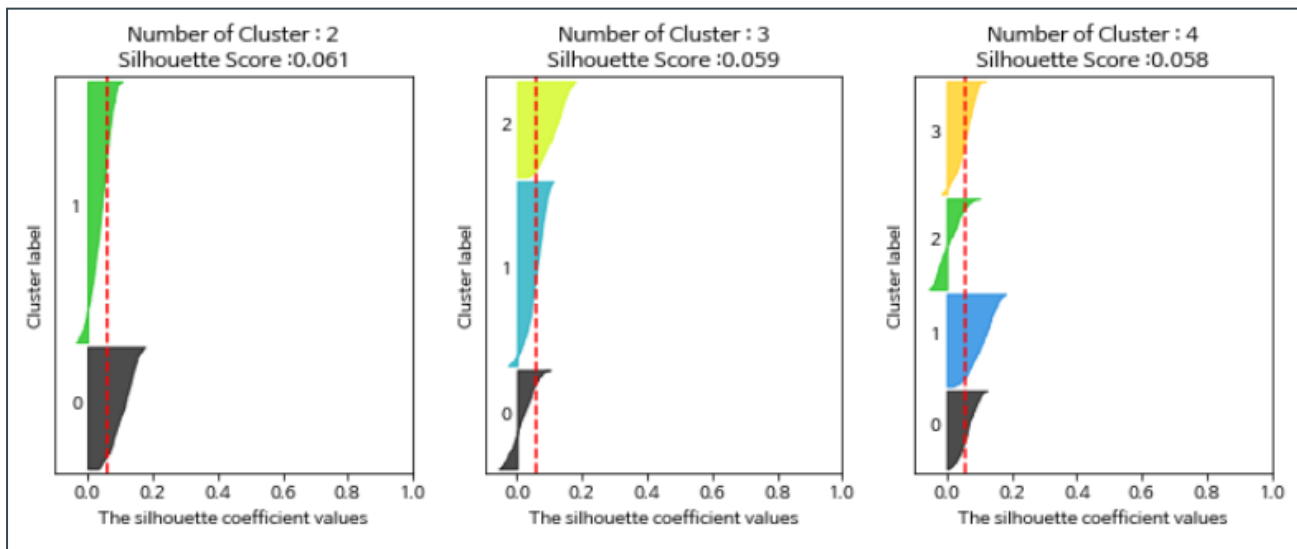
PCA 기법을 통해 2차원으로 축소 후, 군집분석 결과를 시각화



EDA

군집분석 평가

- 여러개의 클러스터링 갯수를 List로 입력 받아 각각의 실루엣 계수를 면적으로 시각화



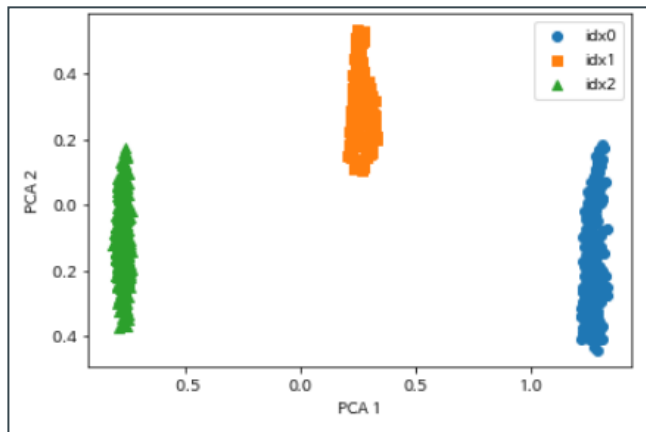
EDA

군집분석

- 각 일자별 게시글 집합을 하나의 문서로 생각하고 군집분석 진행. (K-Means)

2) TF-IDF

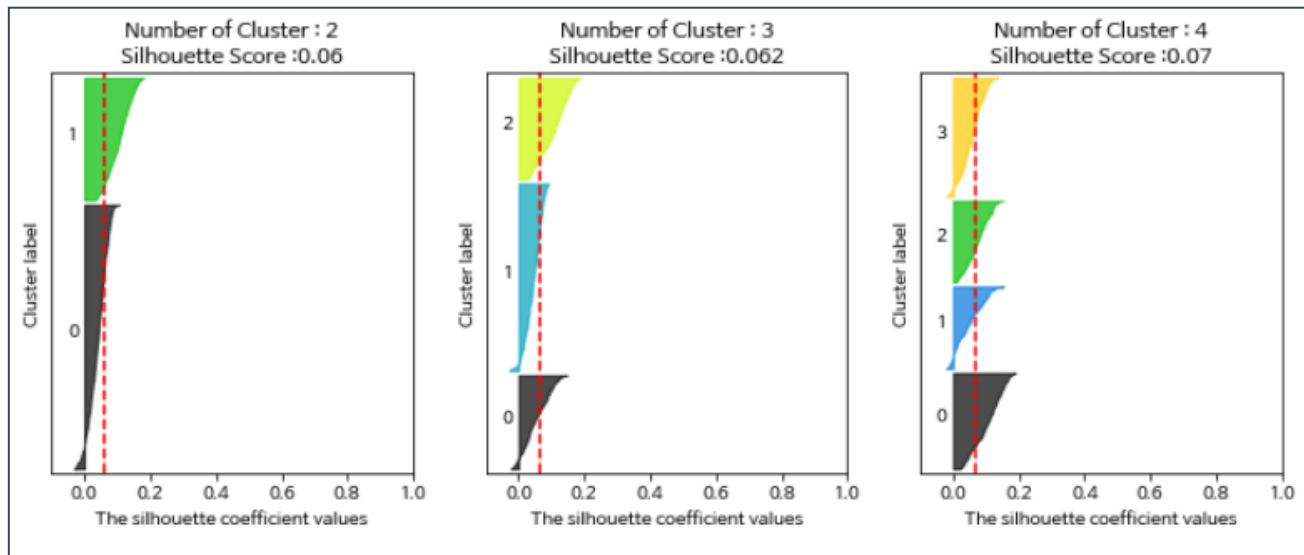
PCA 기법을 통해 2차원으로 축소 후, 군집분석 결과를 시각화



EDA

군집분석 평가

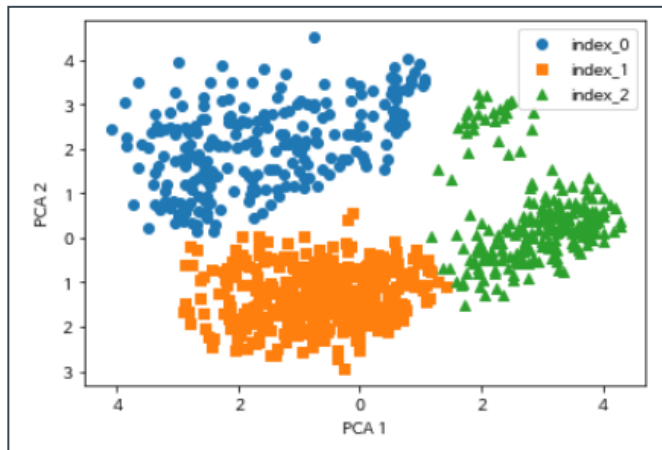
- 여러개의 클러스터링 갯수를 List로 입력 받아 각각의 실루엣 계수를 면적으로 시각화



EDA

군집 별 핵심단어 추출

- 각 군집 별 핵심단어 Top10 확인.



```
##### Cluster 0
Top features: ['하락', '이번', '이유', '오늘', '주식', '지금', '정도', '생각', '간다', '새끼']
=====
##### Cluster 1
Top features: ['시작', '미국', '새끼', '생각', '사람', '나스닥', '오늘', '간다', '지금', '이유']
=====
##### Cluster 2
Top features: ['나스닥', '주식', '사람', '매수', '새끼', '미국', '오늘', '생각', '이유', '시작']
=====
```


결론

결과

- EDA / 군집 분석을 통해 주식 의견 데이터로부터 인사이트를 얻을 수 있었음.

한계점

- 좀 더 세밀한 데이터 전처리(불용어 제거 등)가 필요.
- 추후 라벨링을 통해 지도학습 / 감성분석 진행해볼 필요가 있음.
- 시계열 데이터 분석 고려.

감사합니다.