

The background image is a wide-angle aerial photograph of a city at night. The city is densely packed with buildings of various heights, their windows glowing with warm light. A major river cuts through the city, with several bridges spanning it, their structures illuminated. The sky is dark, suggesting it's nighttime. The overall atmosphere is one of a bustling urban center.

2019 Big Contest [Innovation]

Tracking Topics: A new business framework combining LDA Topic Modeling & Card Transactions

Deep-Diving
Dongkyu Cho

Contents

01 Introduction

02 EDA

03 Data Analysis

04 Result

Introduction

미세먼지 분야 국내외 시장 동향 및 분야별 사례분석과 기술개발 현황

머리말

국내 미세먼지 문제가 심각한 수준으로 이어지면서 국민건강에 치명적인 위협을 가져오고 있어 이를 해결할 수 있는 근본적인 대책 마련이 시급한 상황으로, 세계보건기구 산하 국제암연구소는 미세먼지 중 블랙카본을 1급 발암물질로 규정하고 있다.

국내 미세먼지 농도와 공기 질은 세계 최하위권 수준이며, 서울의 미세먼지 농도는 선진국 주요 도시 중 가장 높은 것으로 나타났다. 세계 각국에서도 대기오염물질 배출 증가에 따라 대기질 개선을 위한 배출기준을 설정하여, 다양한 기술개발과 이를 활용한 미세먼지 문제 해결에 접근하고 있다.

정부는 2018년 126억 원의 범부처 미세먼지 국가전략프로젝트 사업 시행계획을 세워 본격적으로 추진하며, 고농도 미세먼지 예보의 정확도 향상을 위해 지속적인 투자를 진행할 계획이다.

이에 본 보고서는 국가적으로 이슈가 되고 있는 미세먼지 국내외 시장동향과 정책현황/ 분야별 저감사례 분석 및 기술개발 동향 등을 수록하였다.

각종 연구들에 따르면 더 이상 미세먼지는 우리 사회에서 피할 수 있는 존재가 아니며, 미세먼지로 인한 수 많은 기업과 정부 매출, 정책에 지대한 영향을 주고 있다.

실제로 이미 공기청정기, 마스크 등은 눈에 띌 정도로 폭발적인 성장세를 보이고 있으며, 기업 및 정부는 빠르게 대응하고 있다.

배경 1. 미세먼지 지각변동

-[공기정화, 필터산업의 시장동향],

경북테크노파크

-[미세먼지 국내외 관련 산업 정책 동향과 주요 핵심 사업 기술 및 시장 전망],

미래산업 리서치



- [매일경제] 주범 파악도 못하고… '과학' 빠진 미세먼지 대책
- [YTN] 미세먼지 고농도 시기 대비해 패러다임 싹 바꿔야

Introduction

배경 2-1 부족한 비정형 문자 데이터 연구

기업은 성공적으로 대중의 욕구를 읽었으며, 주요한 성과를 거둔 것으로 판단된다.

- [중앙일보] 뜨거워진 의류관리기 시장…LG독주에 지각변동 올까
- [신아일보] 미세먼지가 바꾼 주방풍경…‘안티 더스트’ 주방용품 인기

하지만 대부분의 연구들은 대중의 관심을 요인으로 한 데이터 분석을 시도하지 못하고 있으며,
피상적인 설문조사에만 그쳤다.

미세먼지가 바꾼 소비 행태 변화

2019년 4월



- 2019년 4월 하나금융경영연구소의 보고서 :
오프라인 매출은 미세먼지 농도가 아닌 뉴스 건 수에 긴밀한 관계가 존재.
- 온라인 데이터의 가능성을 보여줌과 동시에, 단순한 뉴스 건 수에만 집중.

Introduction

배경 2-2 부족한 비정형 문자 데이터 연구와 그 이유

1. 자연어 처리 과정은 비지도학습으로 처리되기 너무 어려움.
2. 학습을 위한 좋은 한국어 Dataset이 존재하지 않음.
 - BERT와 XLNET의 등장으로 자연어 처리 기술이 대폭 증가하였으나, Task oriented Dataset이 부족한 한국어는 fine - tuning 을 적용할 수 없다.
3. 방대함
 - 매우 광범위한 Data를 포괄하고 있는 관계로 정제하기 어려움.
4. 부족함
 - 가공된 NLP를 기반으로 딥러닝 모델을 수행하기에 Data가 너무 적음.



Is hard..

Especially for unsupervised data

[witanworld.com]

Introduction

연구목적

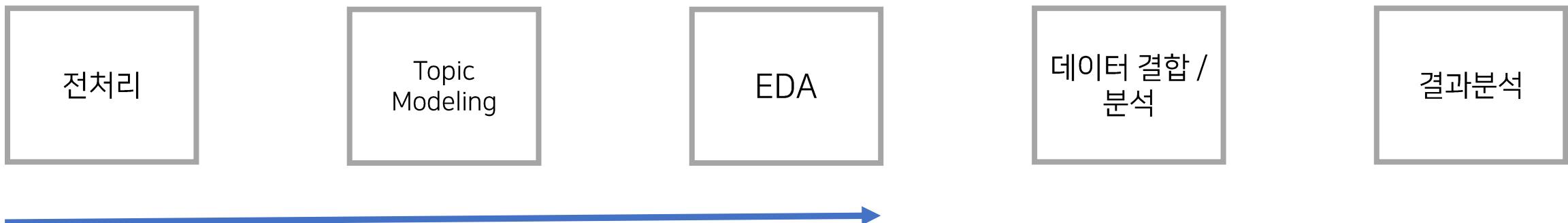
미세먼지 농도가 아닌, 자연어 데이터를 기반으로 오프라인 매출 요인분석.

가설

추출된 Topic 과 특정 인구분포의 업종 별 매출에 강한 상관관계가 존재.

온라인 상의 자연어 데이터와 보다 강한 상관관계가 존재.

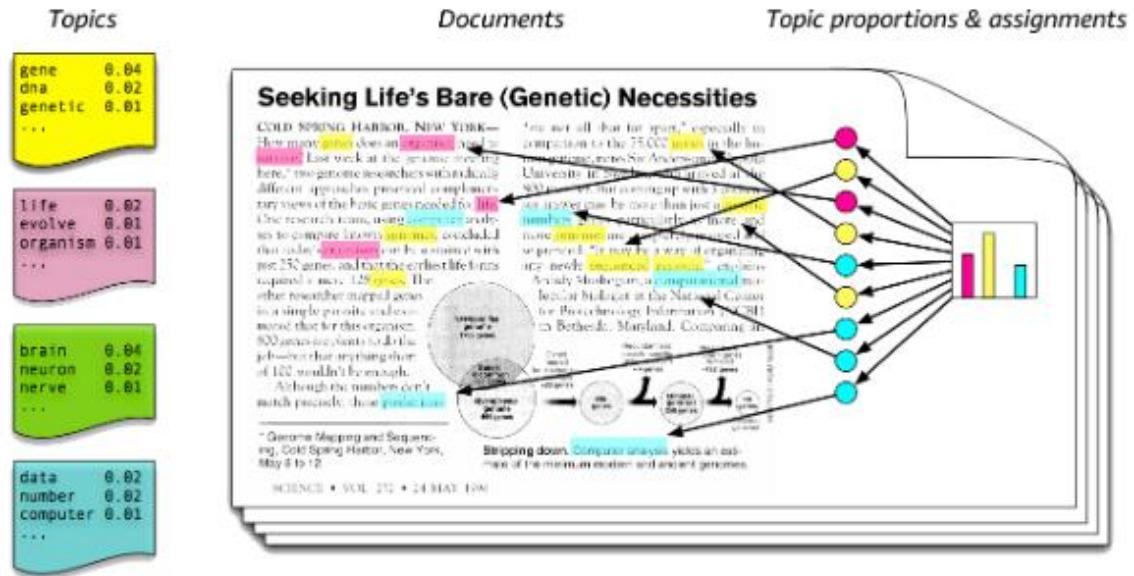
연구 진행 방향



Preprocessing & Feature-Extraction

LDA Topic Modeling

Latent Dirichlet Allocation (LDA)



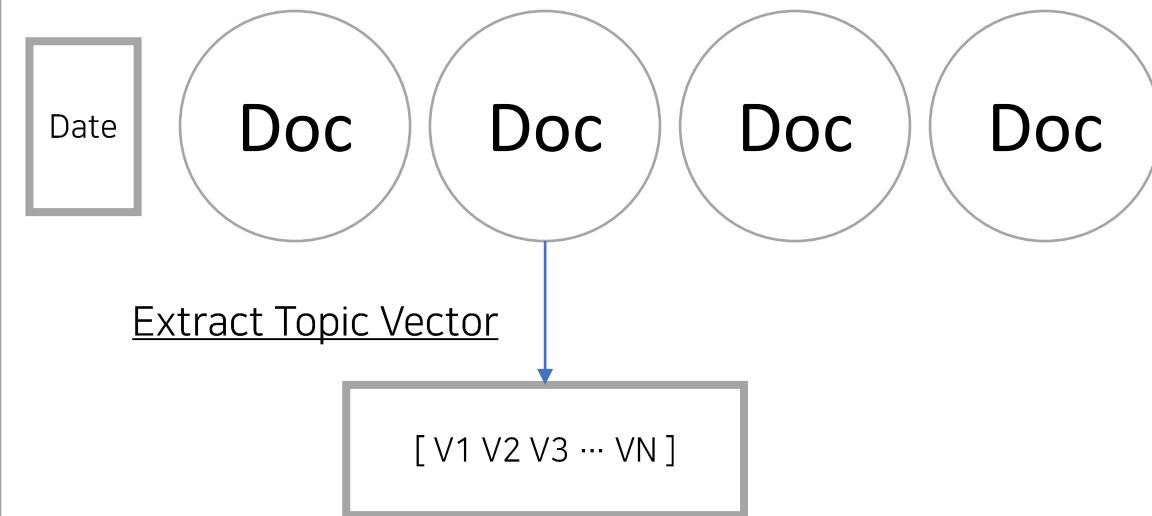
Document 별 Topic의 분포와 Topic의 말뭉치 분포를 추정하는 기술.

Text Summarization, Topic Tracking 등 유용하게 사용되고 있다.

사용 이유

1. 비지도 학습으로 써 Labeled Dataset 불필요
2. 수 많은 연구 및 활용으로 효용성 검증
3. 방대한 자료에 적합

본 연구 활용 방안

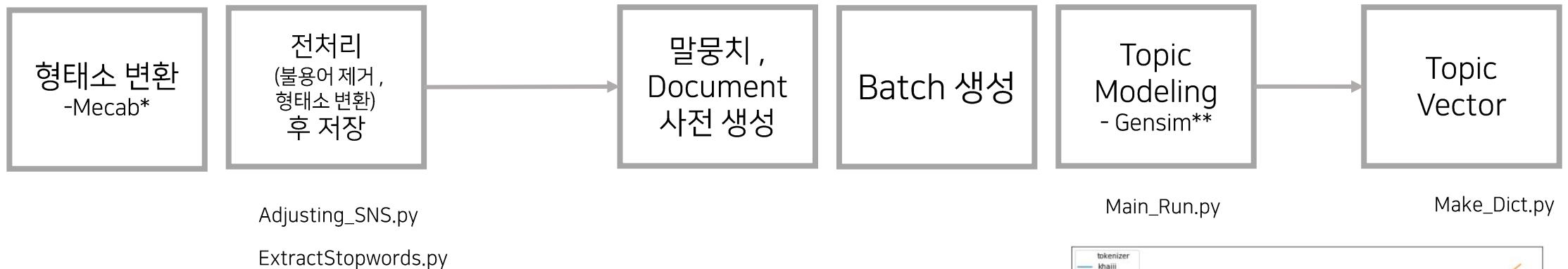


*N : Num of Topics

** Sum of all i , $V_i = 1$

LDA Topic Modeling

Process

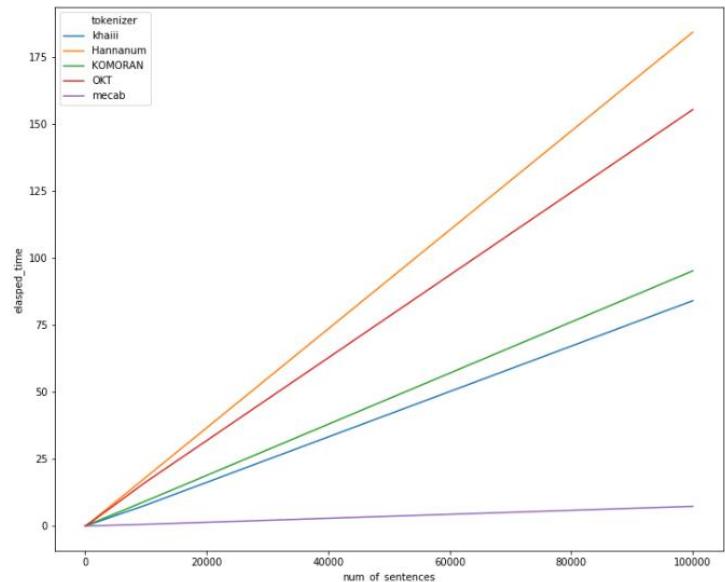


* Why Mecab?

Linux에서 지원하는 고성능의 한국어 형태소 분석기. Komoran과 함께 훌륭한 문장 분석 능력 보유했으나, Komoran의 버그와 선형적으로 증가하는 연산 속도로 Mecab 채택.

** Why Gensim?

Multicore LDA 기능 사용시 속도 향상 및 다양한 Parameter 지원.
다년 간의 field 검증으로 안정성 입증.



LDA Topic Modeling

Topic List

Topic 1 : 일상적인 내용 (대표 : 먹는데, 맛있, ㅋㅋ)

Topic 2 : 정부 미세먼지 정책 (대표 : 시민, 저감, 환경, 산업)

Topic 3 : 미세먼지 가전 및 용품 (대표 : 공기청정기, 마스크, 필터)

Topic 4 : 야외 체험 (대표 : 길, 체험, 사진, 놀이 공원)

Topic 5 : 정치관련 토픽 (대표 : 대통령, 중국, 원전)

Topic 6 : 피부미용 (대표 : 피부, 클렌징, 크림)

Topic 7 : 집, 살림 (대표 : 청소, 아파트, 가구)

Topic 8 : 미세먼지 관련 질환 (대표 : 건강, 치료, 질환, 원인, 호흡기, 비염)

Topic 9 : 기상예보 (대표 : 날씨, 오늘, 전국, 예보)

Topic 10 : 주가 및 주식 정보 (대표 : 투자, 시장, 기업, 매출)

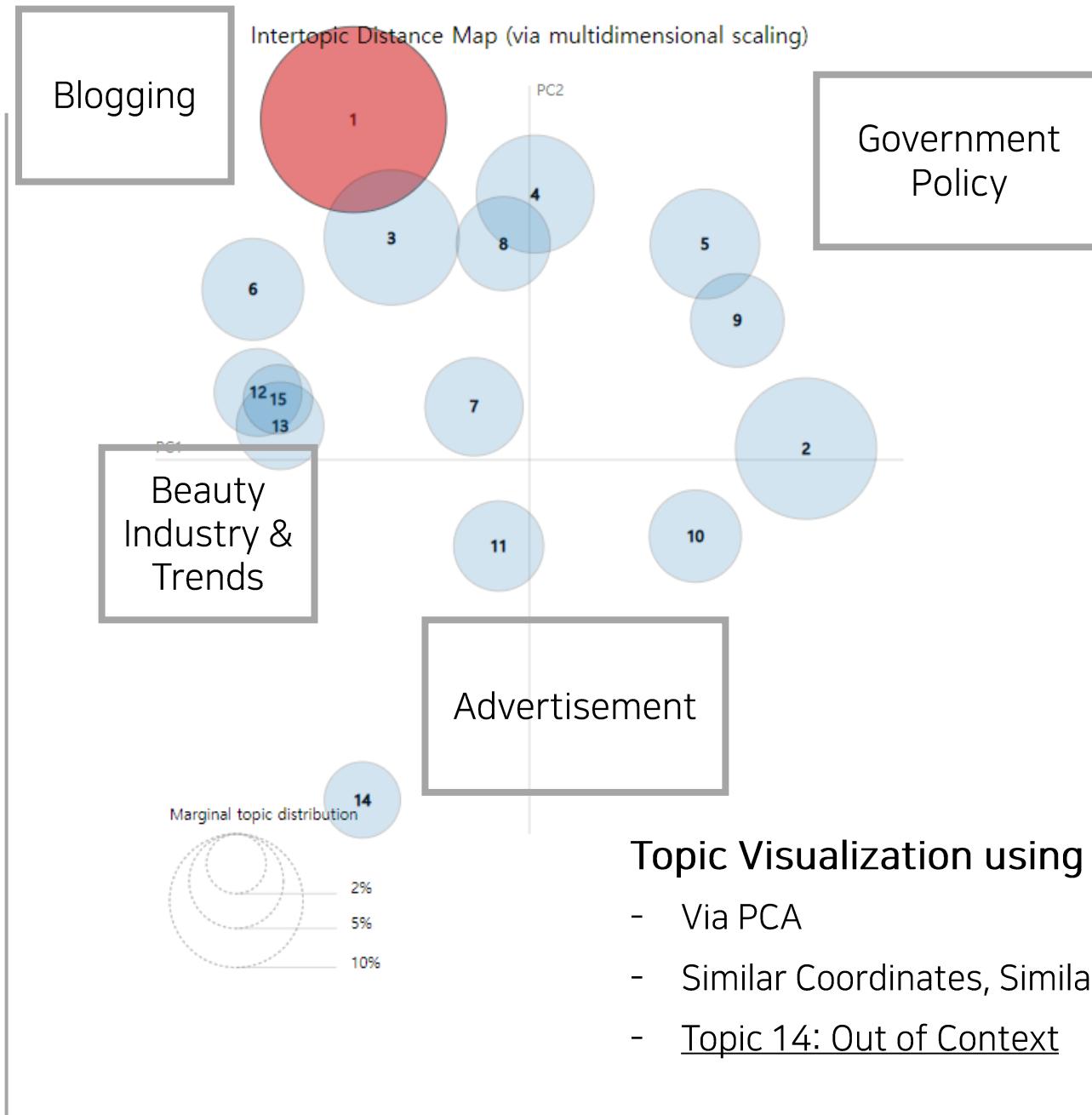
Topic 11 : 자동차 정보 및 거래 (대표 : 자동차, 차량, 등록, 세차)

Topic 12: 성형 및 시술 (대표 : 턱, 라인, 축소, 피지, 개선)

Topic 13: 광고 및 홍보 (대표 : 선물, 이벤트, 추가, 할인)

Topic 14: 제품 판매 안내 (대표 : 수량, 직거래, 택배, 별도)

Topic 15: 다이어트 및 운동 (대표 : 살, 다이어트, 체중)

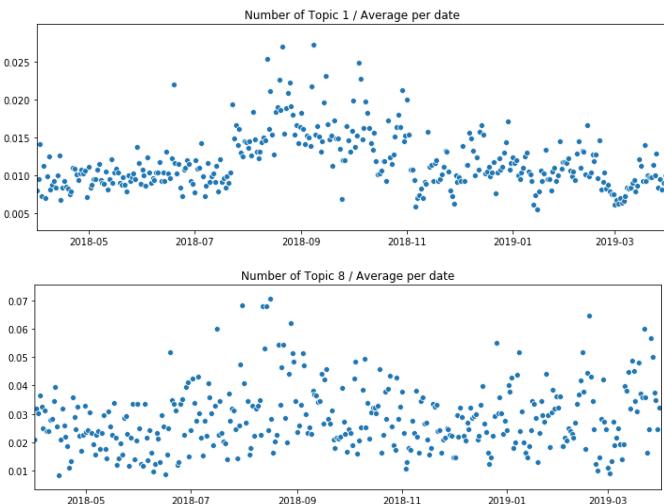
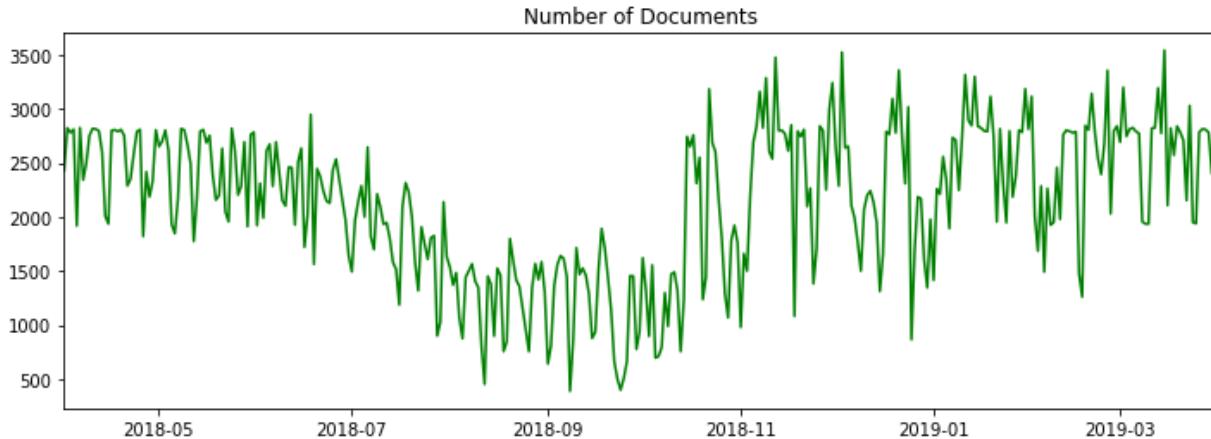


Topic Visualization using pyLDAvis

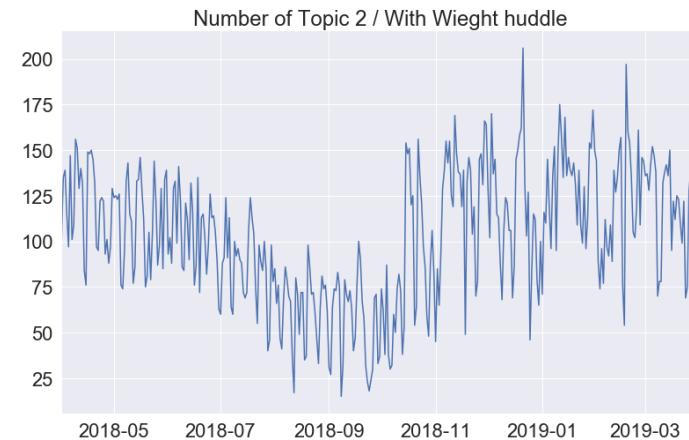
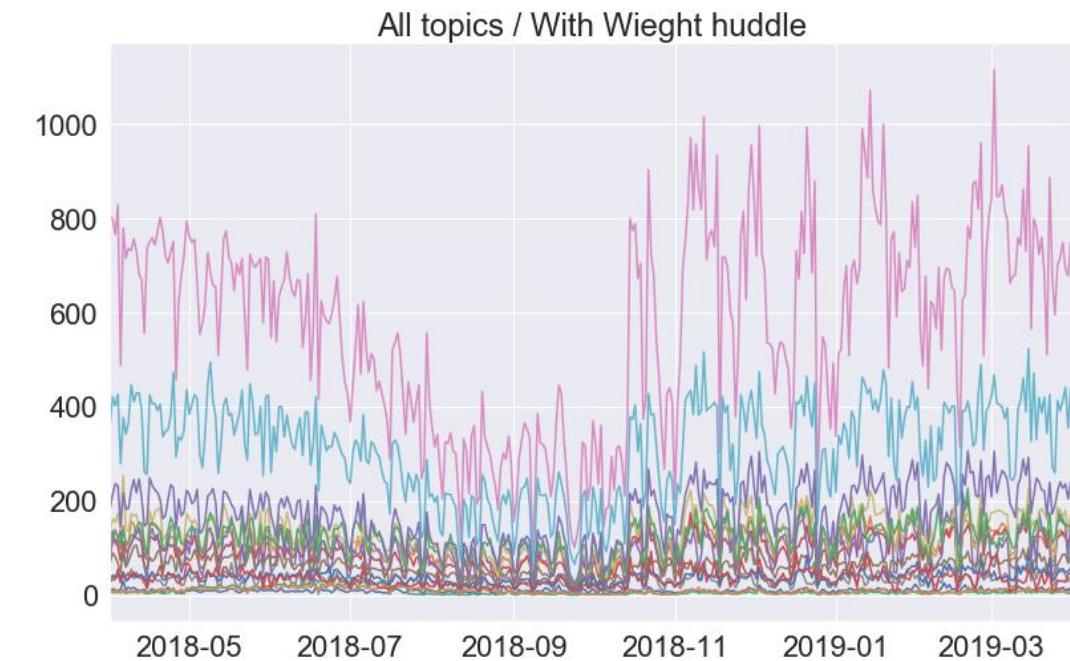
- Via PCA
- Similar Coordinates, Similar Contexts
- Topic 14: Out of Context

Exploratory Data Analysis

1. Document-Topic Vector Analysis



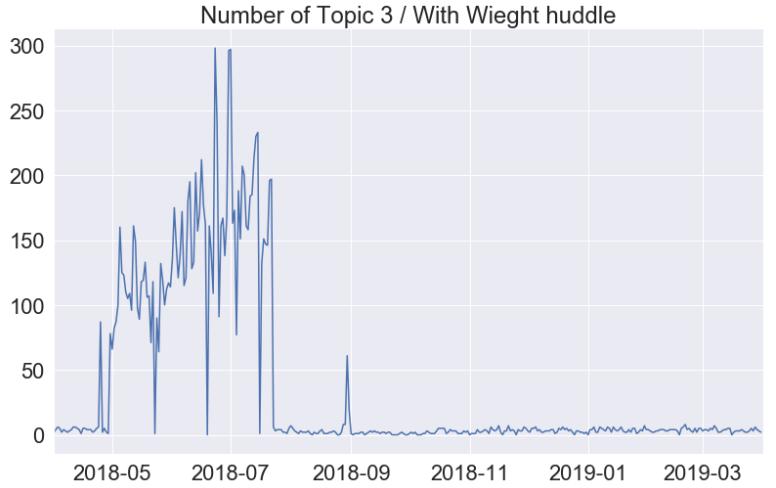
- 대부분의 Document는, 유의미한 Topic 몇 개와 나머지 Topic 보유.
- 따라서 Topic Vector 평균 시 Document 수가 가장 적은 여름~가을에 Topic 가장 높다.



- 평균은 적합한 기준이 아니라고 판단.
- 문서 내에서 최소의 기준 ($1/15$)를 넘어야 해당 Topic이 존재하는 것으로 판단하여 합산.

Exploratory Data Analysis

1. Document-Topic Vector Analysis



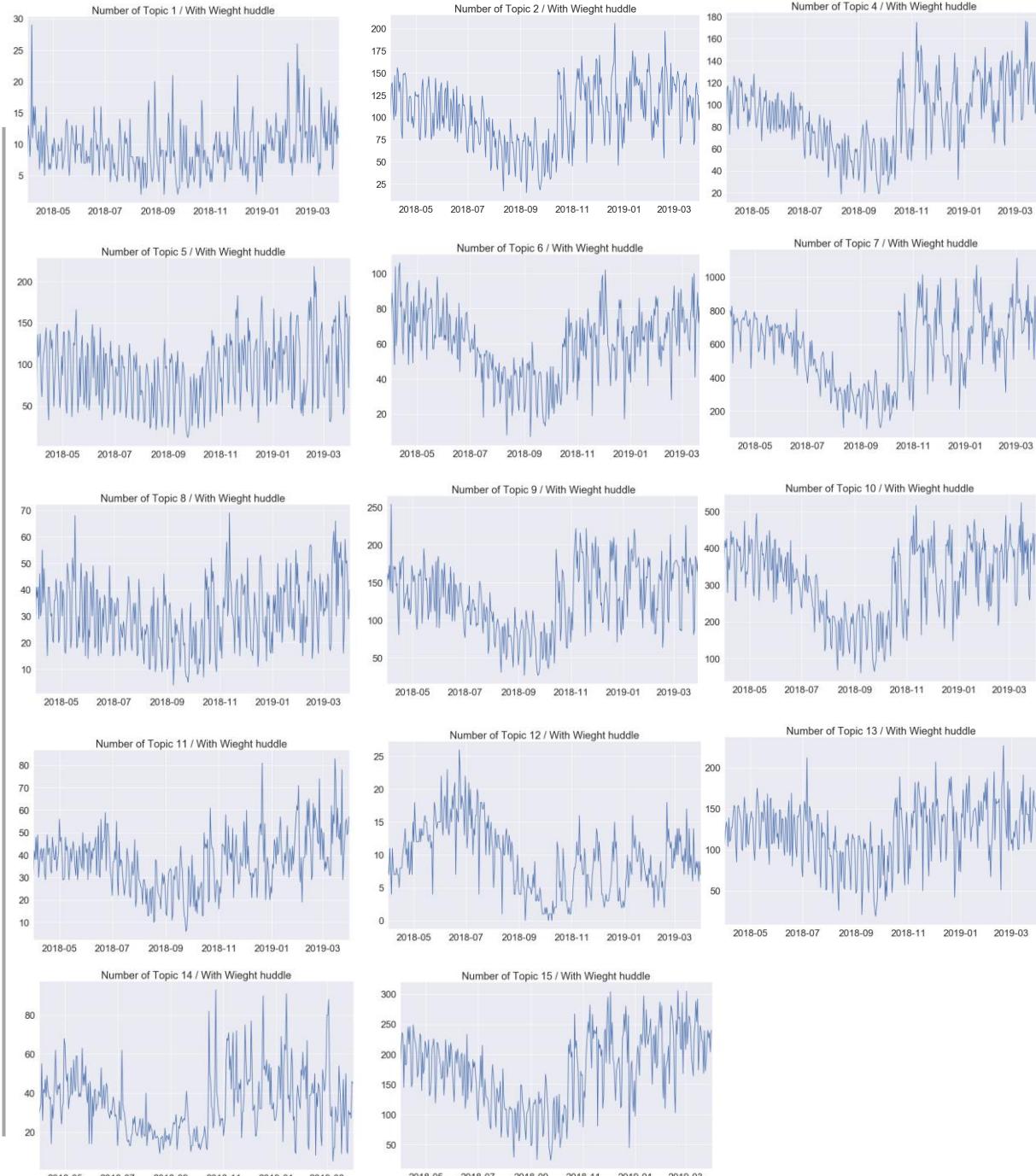
이제 에어콘없이 못사는 계절이 돌아왔어요!
에어콘을 딱 틀자마다 혁? 이 냄새 뭐야..?
동생도 딱 틀자마다 누나 에어콘 고고가자..
차안에서 알수없는 냄새들..
디퓨저도 넣어봤지만 이유를 알수 없었는데 에어필터가 문제일거라는
자동차박사친구의 말을 듣고 바로 바꾸기로했다!
비용조사서 바꿔야하나 했는데 셀프로 손쉽게 교환이 가능하다해서 셀프도전!

“
미세먼지 NO!!
”

엔진오일처럼 구동계에 영향을 주는 소모품은 아니지만
우리 건강을 위해 꼭! 필요한 소모품 중에 하나죠 ㅎㅎ
누구나 쉽게 알 수 있는 소모품 중에 하나이므로 짧게 가겠습니다
출발~!!

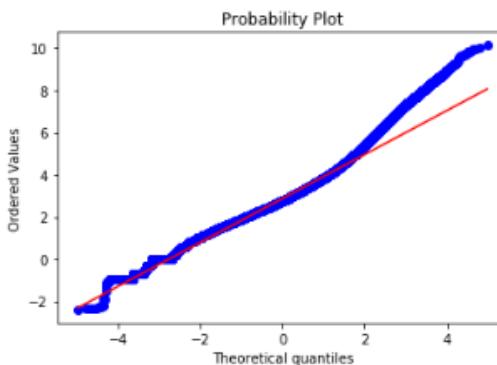
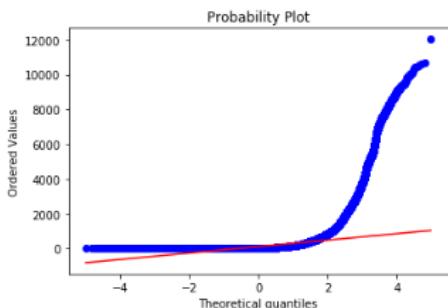
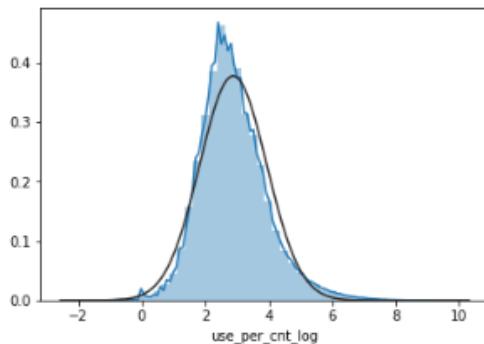
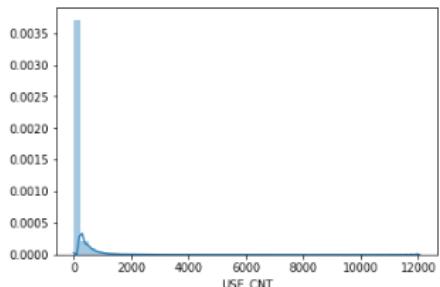
- 대부분의 Topic은 미세먼지 이슈와 비슷하게 겨울~봄에 상승하고 여름~가을에 낮아지나, Topic3만은 6월 ~ 7월에 급증하고 이후에는 폭발적으로 감소한다.

- 이는 에어컨을 처음 가동하는 시기와 맞물린다. Topic 3가 미세먼지 관련 가전 임을 고려해보면 에어컨 및 환풍기를 정비하면서 발생하는 것으로 추정되며, 대중들의 미세먼지 관심은 여름까지도 이어짐을 알 수 있다.



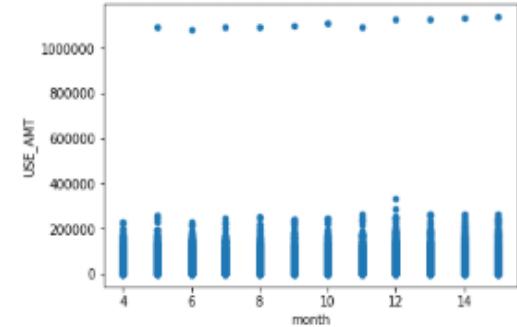
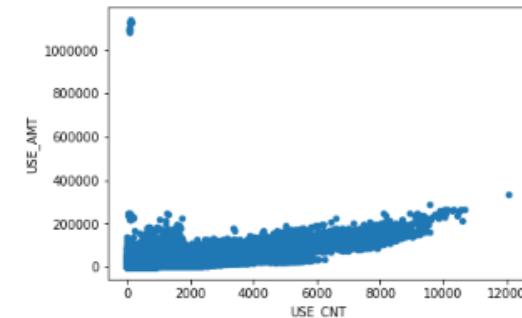
Exploratory Data Analysis

2. Card Data Analysis



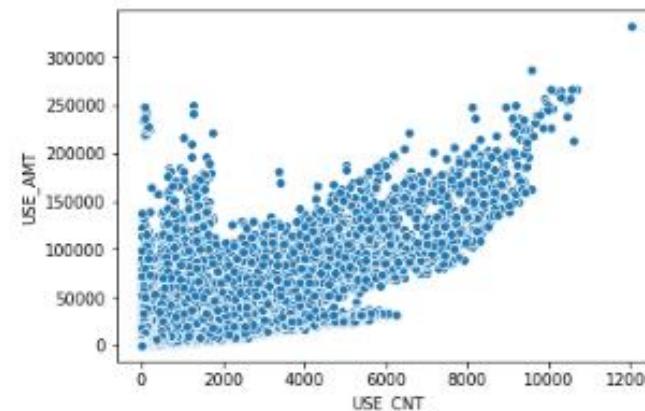
극단적인 데이터 분포의 경우 로그변환을 이용해 분포 분석

결과 정규성을 발견함으로써 데이터의 관계를 보다 능동적으로 분석 가능.



위와 같이 극단적인 경향성을 띠는 이상치를 발견.

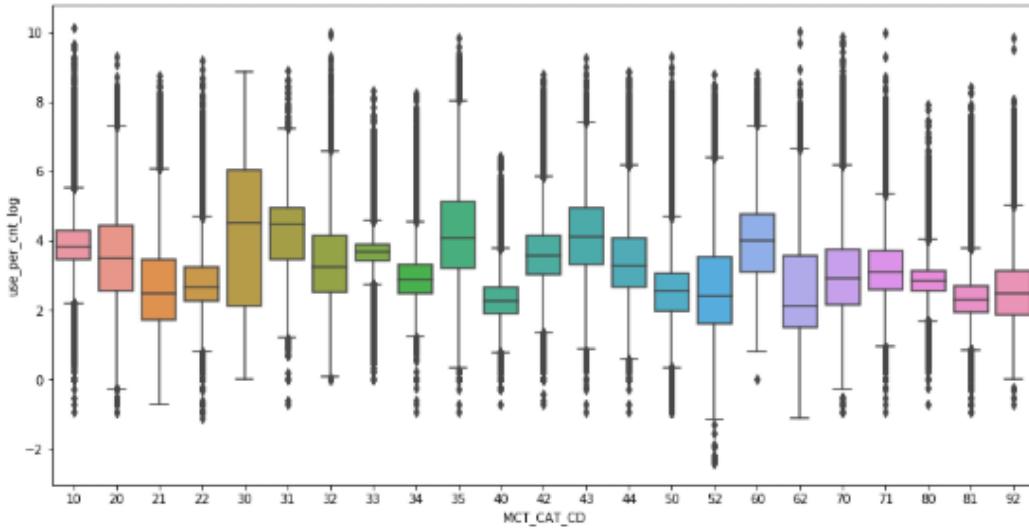
데이터 분석에 저해가 될 뿐만 아니라 상관관계 분석 능력도 저해.



제거 결과 양의 상관관계 발견

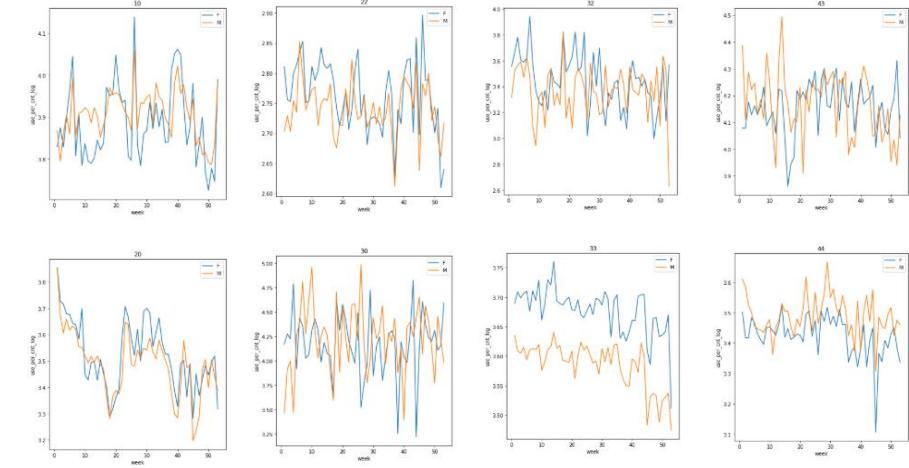
Exploratory Data Analysis

2. Card Data Analysis

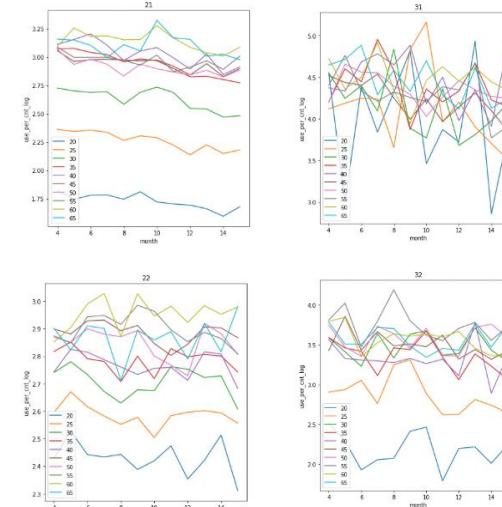


- 업종에 따른 분포가 상이하며, 특히 중앙값이 크게 다르다는 사실 확인.
- 모든 범주가 정규성을 따르지는 않는다고 보여지며, 이에 따라 순서 통계량을 활용하는 것이 적절하다고 여겨진다.
- 카드 매출 데이터는 시간의 흐름에 따라 기록되어 있으므로 시간의 경향성 요인을 제거하기 힘듬.

성별 카드 소비액 추이



연령대별 소비액 추이

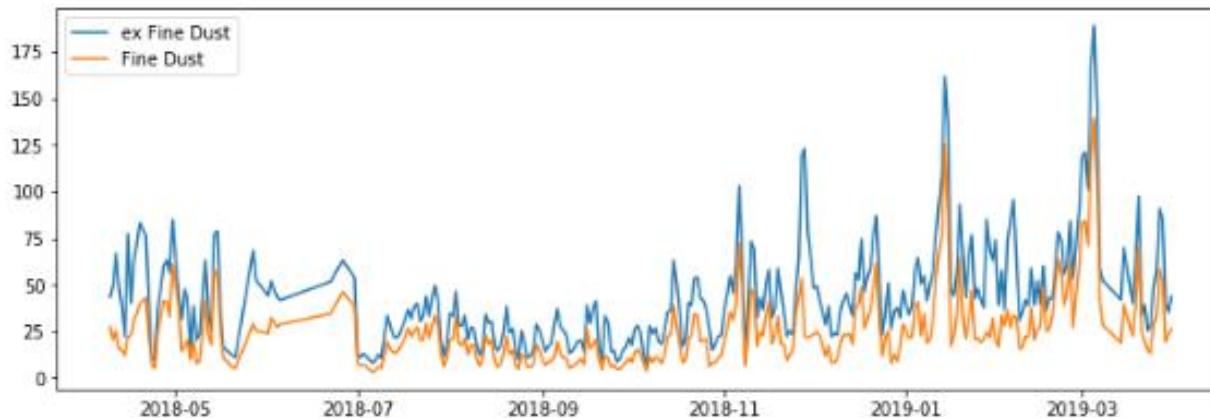


- 모두 연령대별, 성별 카드 매출액의 차이 확인 가능
- 그러나 위의 자료들은 시간의 경향성 문제로 타더 이상의 심고 있는 분석 불가능.

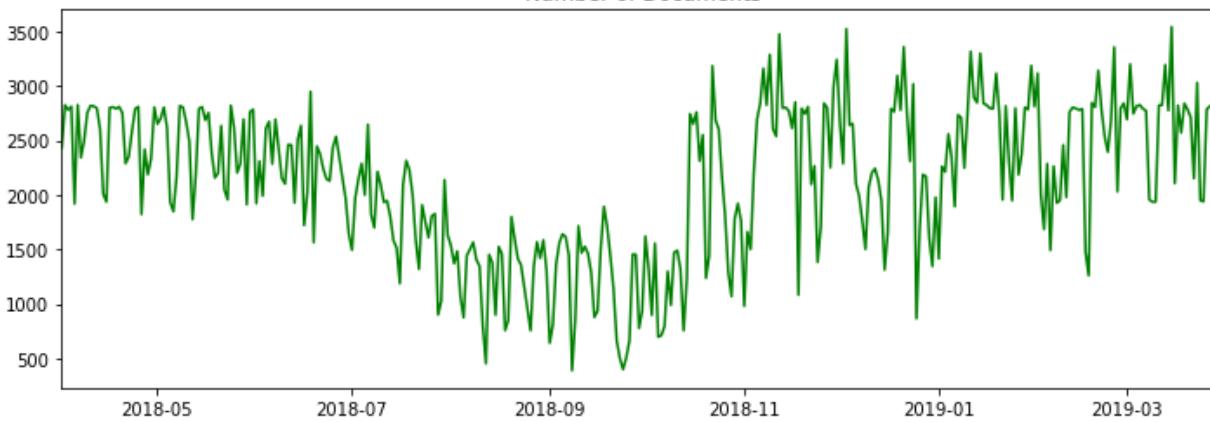
Exploratory Data Analysis

3. Dust Analysis

미세먼지 농도



Number of Documents



대체로 미세먼지 농도와 Topic, Documents 수는 어느 정도의 경향성이 있으나, 미세먼지 농도와의 상관관계는 없다.

특히 기업의 매출은 미세먼지 농도와의 상관관계를 판단하기 힘들다.

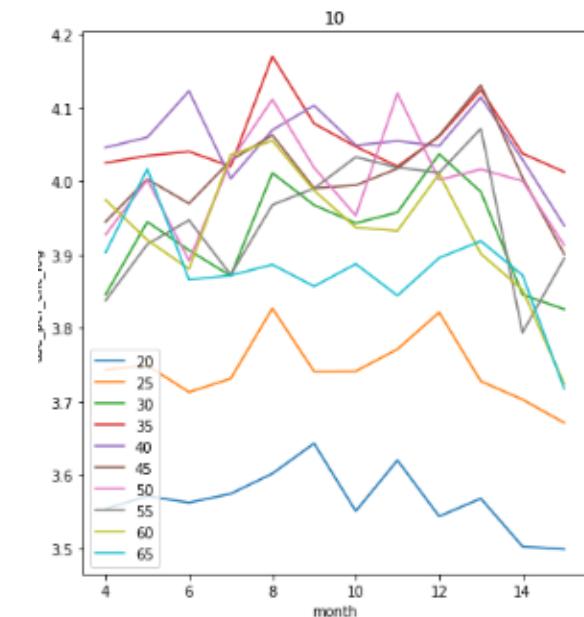
Number of Topic 13 / With Wieght huddle



Number of Topic 11 / With Wieght huddle



연령대 별 기간 별 숙박 업소 매출



Data Analysis

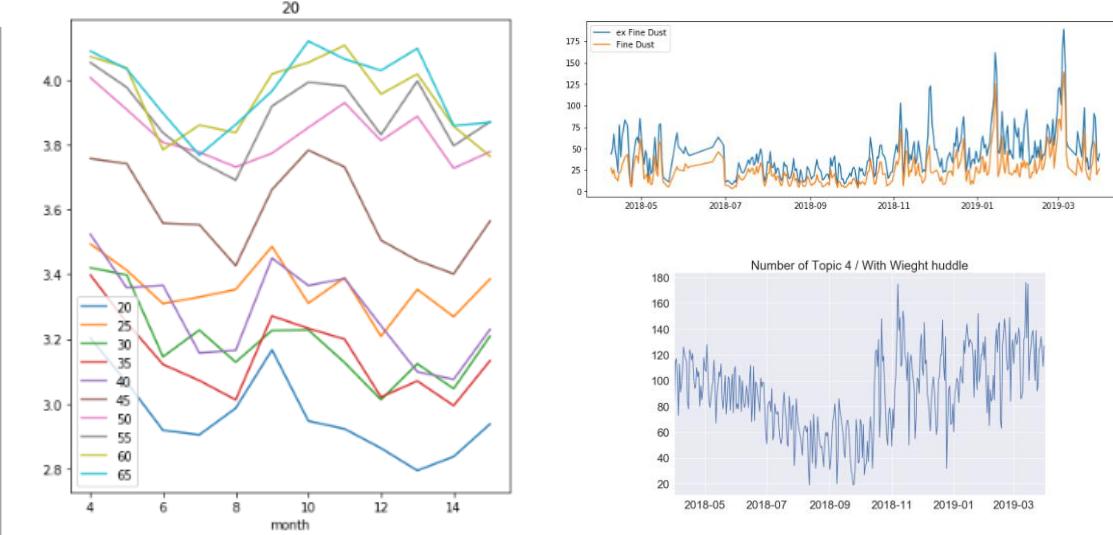
분석방법

기본적인 EDA를 통해서는 계절 / 일자 / 정치 / 경제의 영향력을 배제할 수 없다.
영향력 및 요인 분석 불가능.

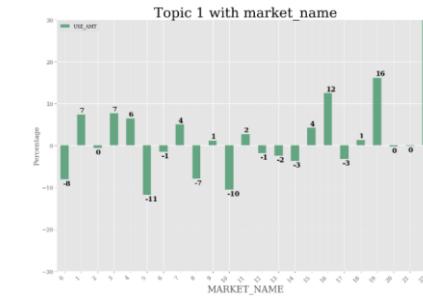
변인을 기준 삼아 기준 별 Dataset 재배열
예) Topic 1 상위 10% Dataset, 상위 20% Dataset…

상위 50% / 하위 50% Leverage 분석

- * Sampling 원리를 따른 것으로, 만일 변인과의 상관관계가 없다면 두 상위 N% Dataset과 하위 N% Dataset은 유의미한 차이가 관측되지 않는다.
- ** 평일과 휴일의 뚜렷한 매출 차이가 관측된 관계로, 훨씬 많은 Data가 확보된 평일을 이용해 분석



날짜 / 일자 별 패턴으로 인한 요인 분석 불가



(우측) Topic 별로 Split 된 날짜의 표준편차 비율. Topic Vector는 미세먼지와 같은 계절적인 이슈를 미약하게 따르기 때문에 표준편차 비가 너무 작아서도, 너무 커서 날짜 / 일자 별 패턴 요인을 강화 시켜서도 안 된다.

Topic	Ratio
1	0.88
2	1.35
3	1.33
4	1.87
5	1.14
6	1.47
7	1.74
8	1.24
9	1.42
10	1.90
11	1.34
12	0.90
13	0.89
14	1.15
15	1.30

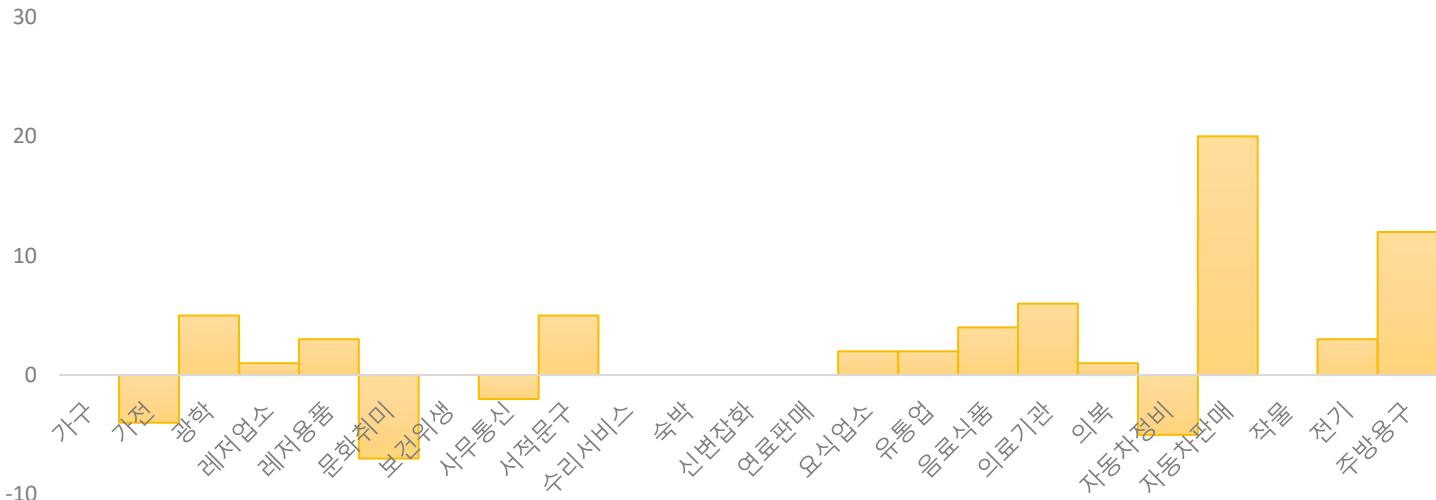
Data Analysis

분석결과 0. Baseline – 미세먼지 농도 별 분석

미세먼지 농도를 기준으로 Data 구분.

- 자동차 판매 업종, 주방용구에서 10% 이상의 매출 증가
- 미세먼지 농도와의 상관관계 입증 가능.
- 단 두 업종만 20%, 11%의 매출 증대 확인.
- 그러나 이외의 강력한 상관관계 단서 관측 불가.

미세먼지 농도에 따른 업종별 매출 변화

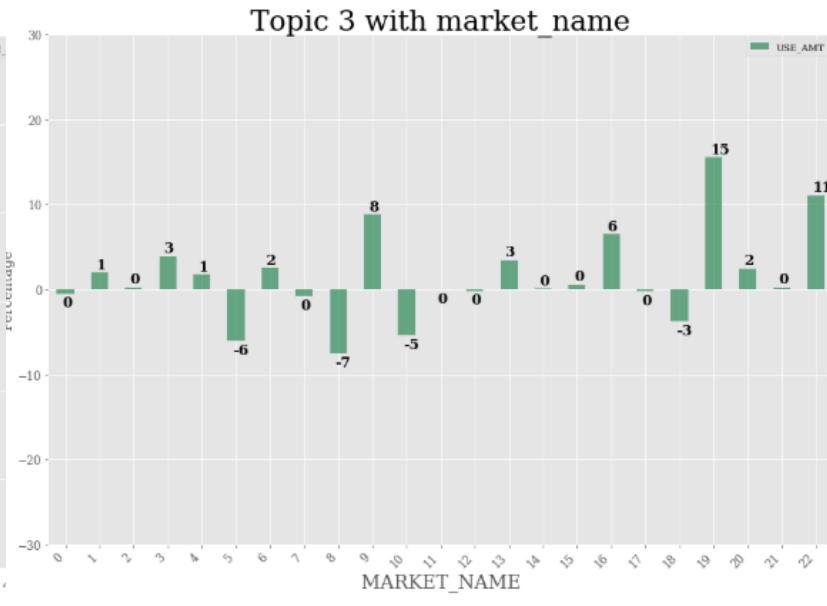
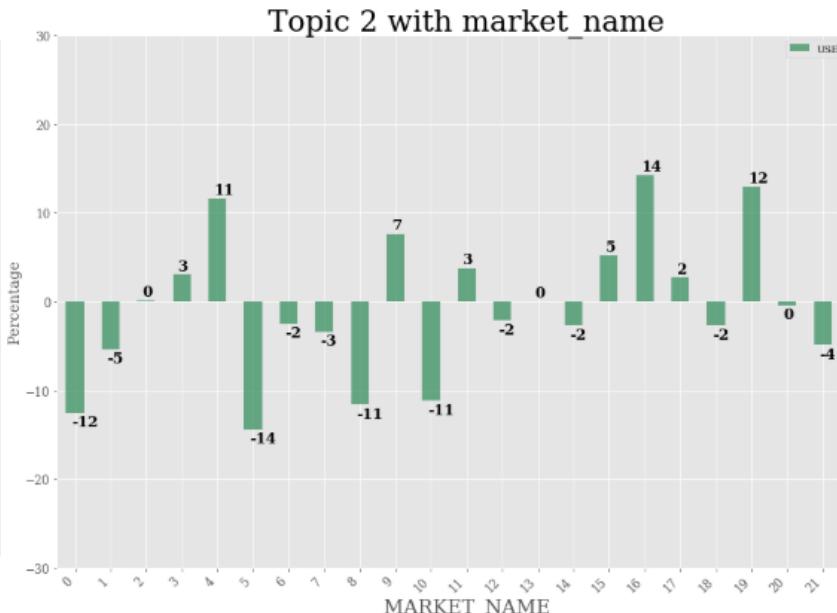
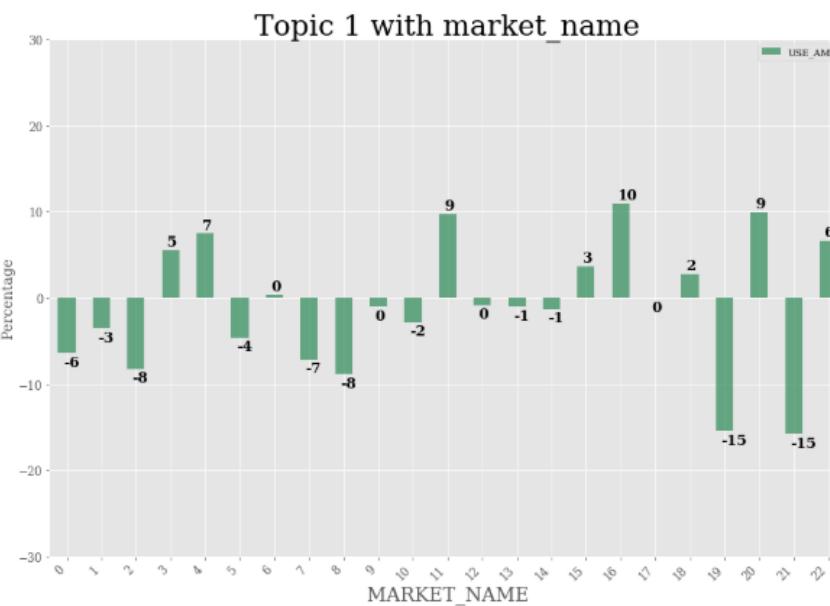


미세먼지 농도에 따른 연령별 매출 변화



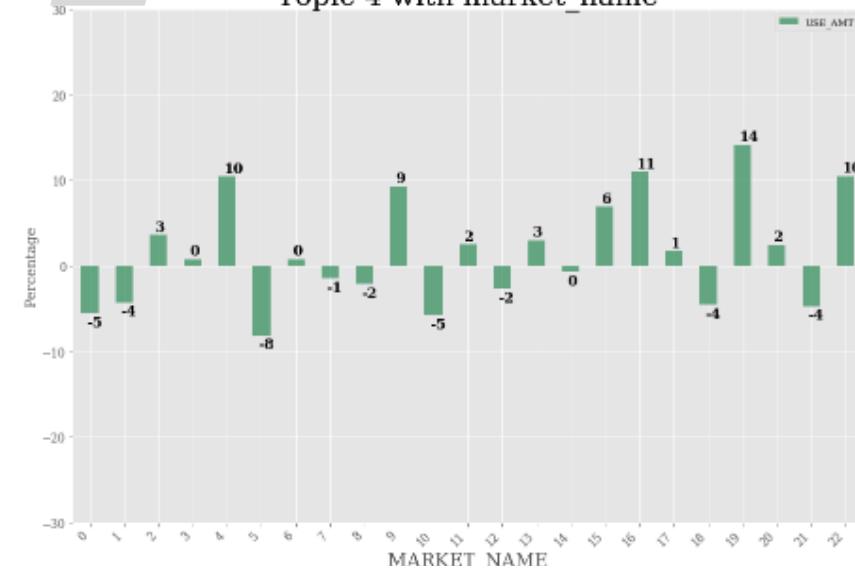
Data Analysis

분석결과 1. Topic 별, 업종별 분석결과 *주요 11개 (세부 사항은 첨부 코드)

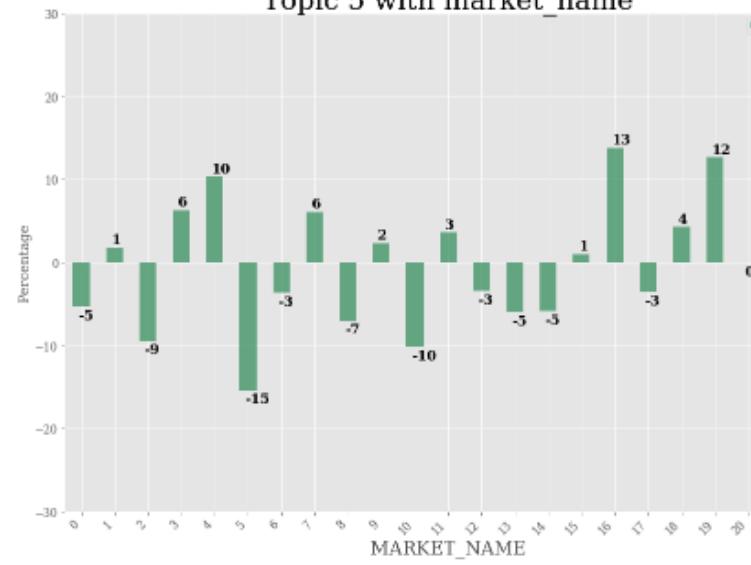


0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
가구	가전	광학제품	레저업소	레저용품	문화취미	보건위생	사무통신	서적문구	수리서비스	숙박	신변잡화	연료판매	요식업소	유통업	음료식품	의료기관	의복	자동차정비	자동차판매	작물	전기	주방용구

Topic 4 with market_name



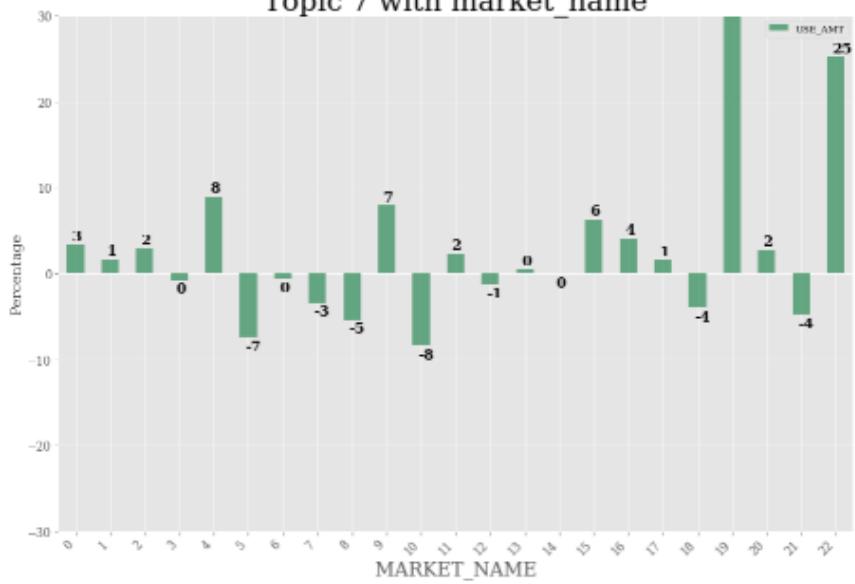
Topic 5 with market_name



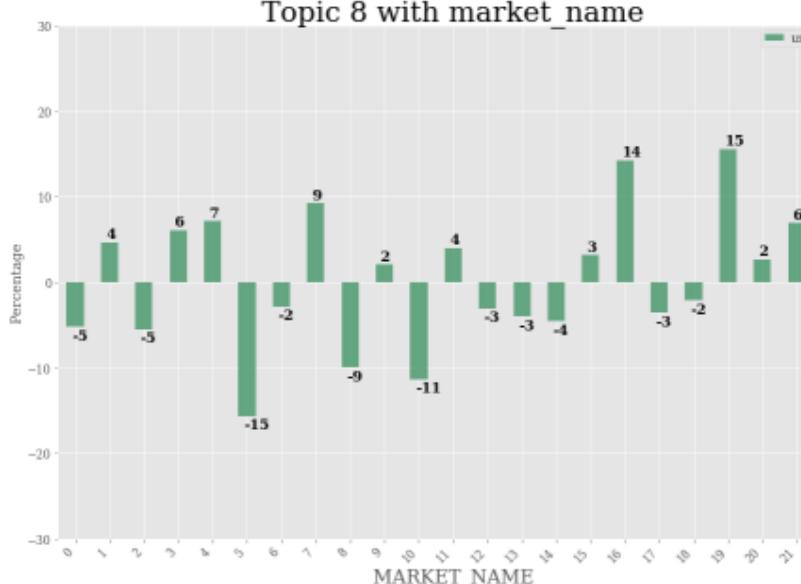
Topic 6 with market_name



Topic 7 with market_name



Topic 8 with market_name



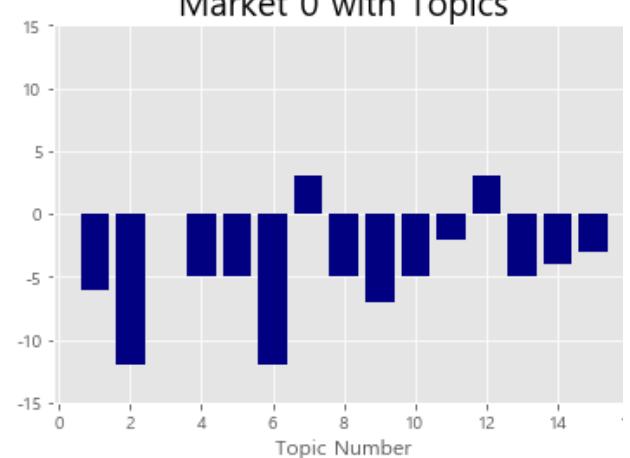
Topic 9 with market_name



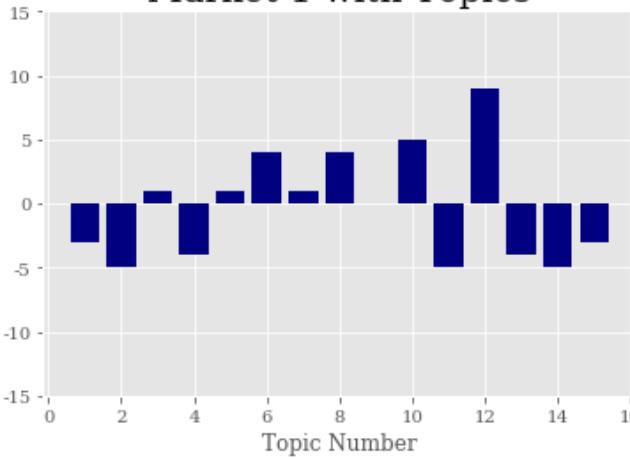
Data Analysis

분석결과 2. 업종별, Topic 별 분석결과 *주요 4개 (세부 사항은 첨부 코드)

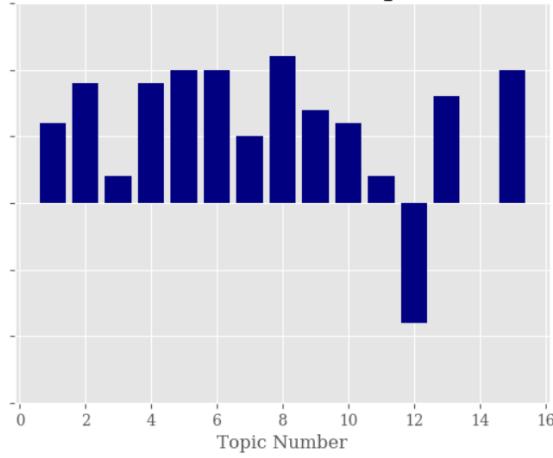
Market 0 with Topics



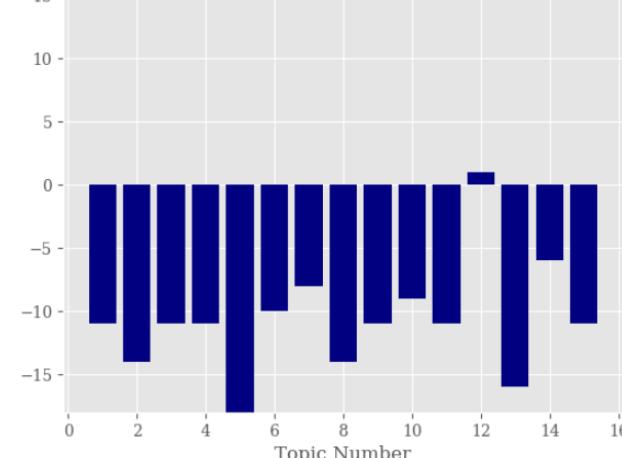
Market 1 with Topics



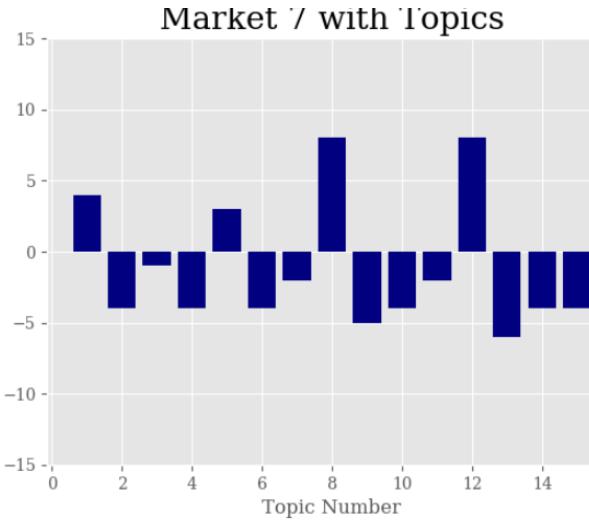
Market 4 with Topics



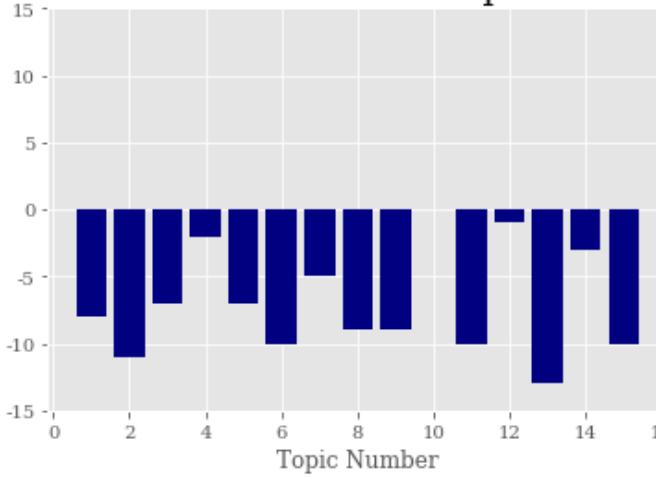
Market 5 with Topics



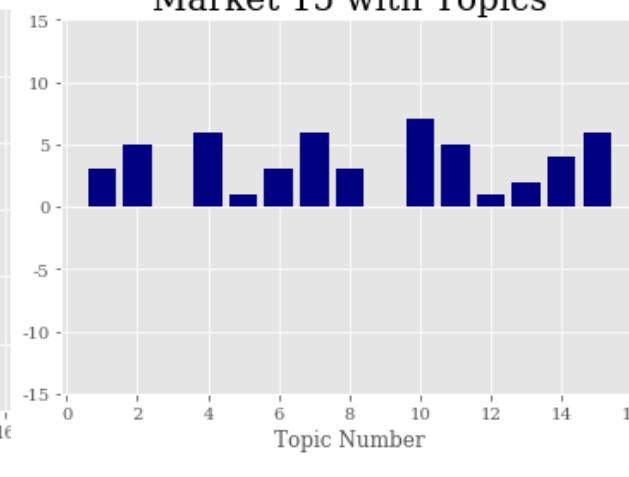
Market 7 with Topics



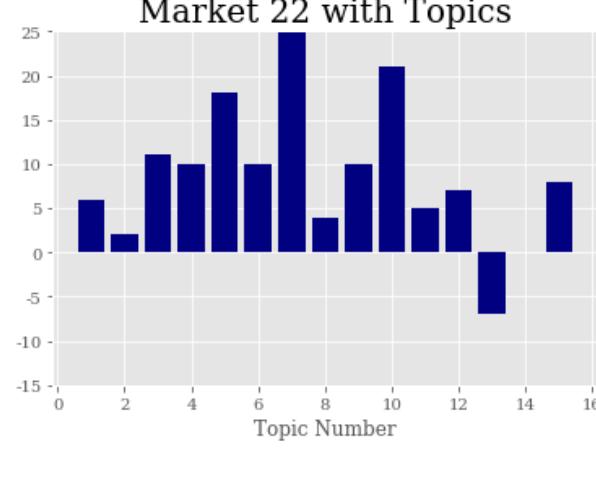
Market 8 with Topics



Market 15 with Topics



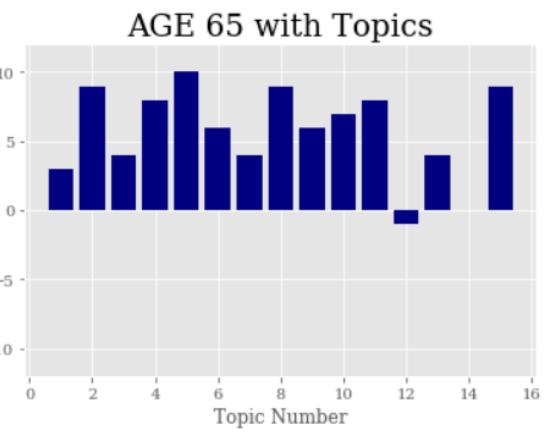
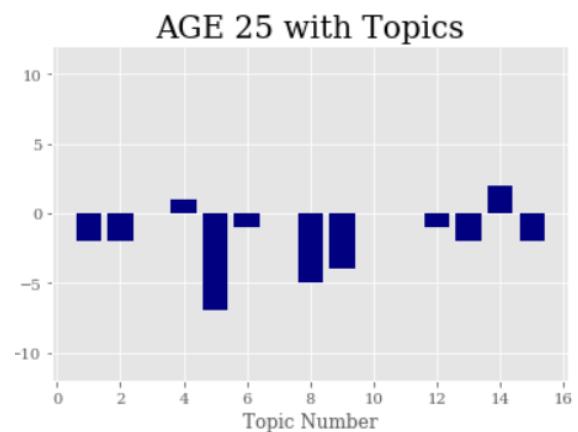
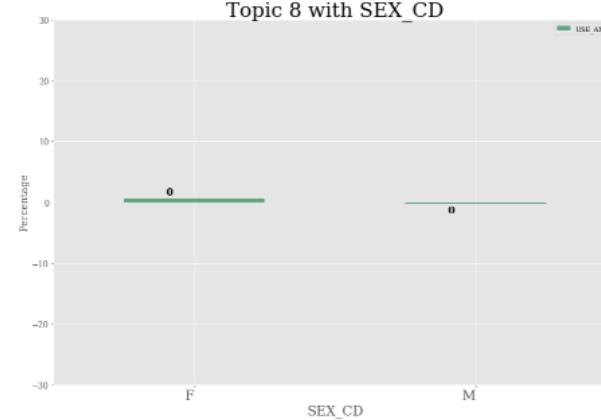
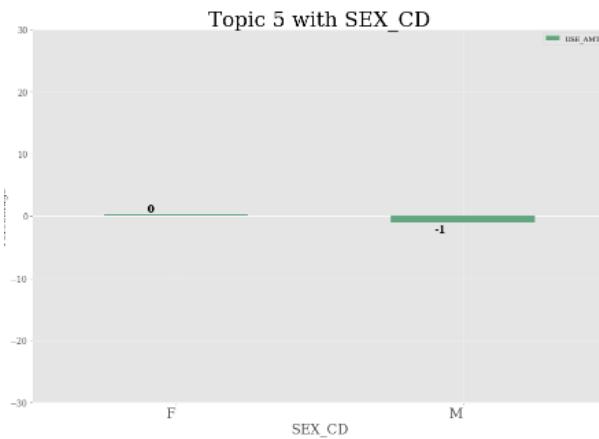
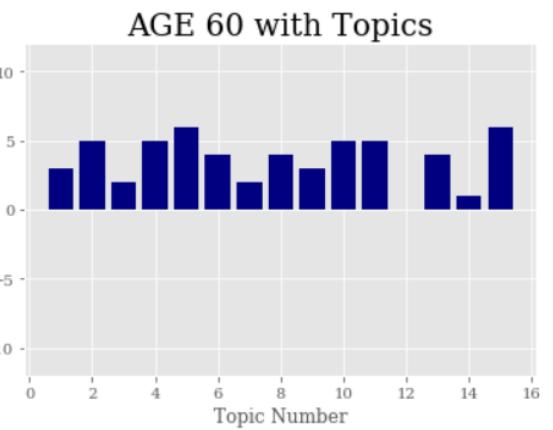
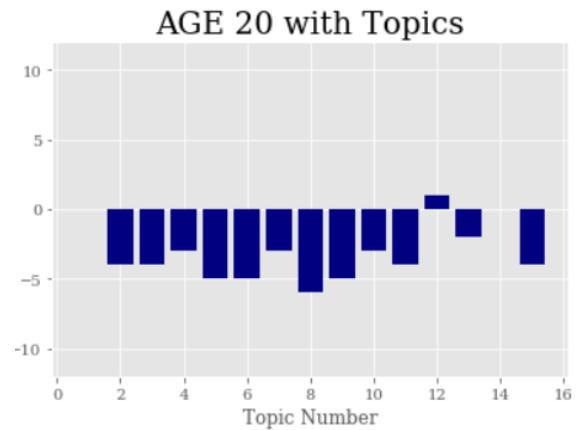
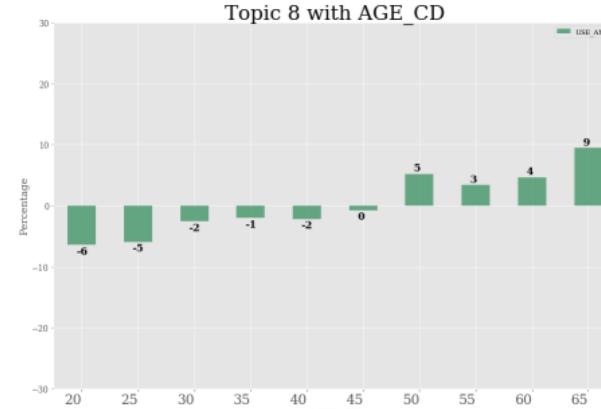
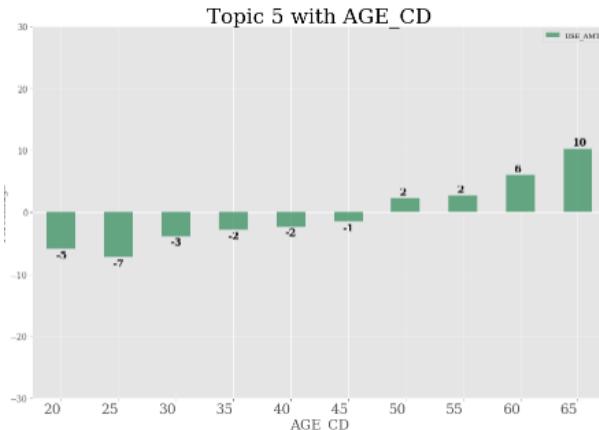
Market 22 with Topics



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
가구	가전	광학제품	레저업소	레저용품	문화취미	보건위생	사무통신	서적문구	수리서비스	숙박	신변잡화	연료판매	요식업소	유통업	음료식품	의료기관	의복	자동차정비	자동차판매	작물	전기	주방용구

Data Analysis

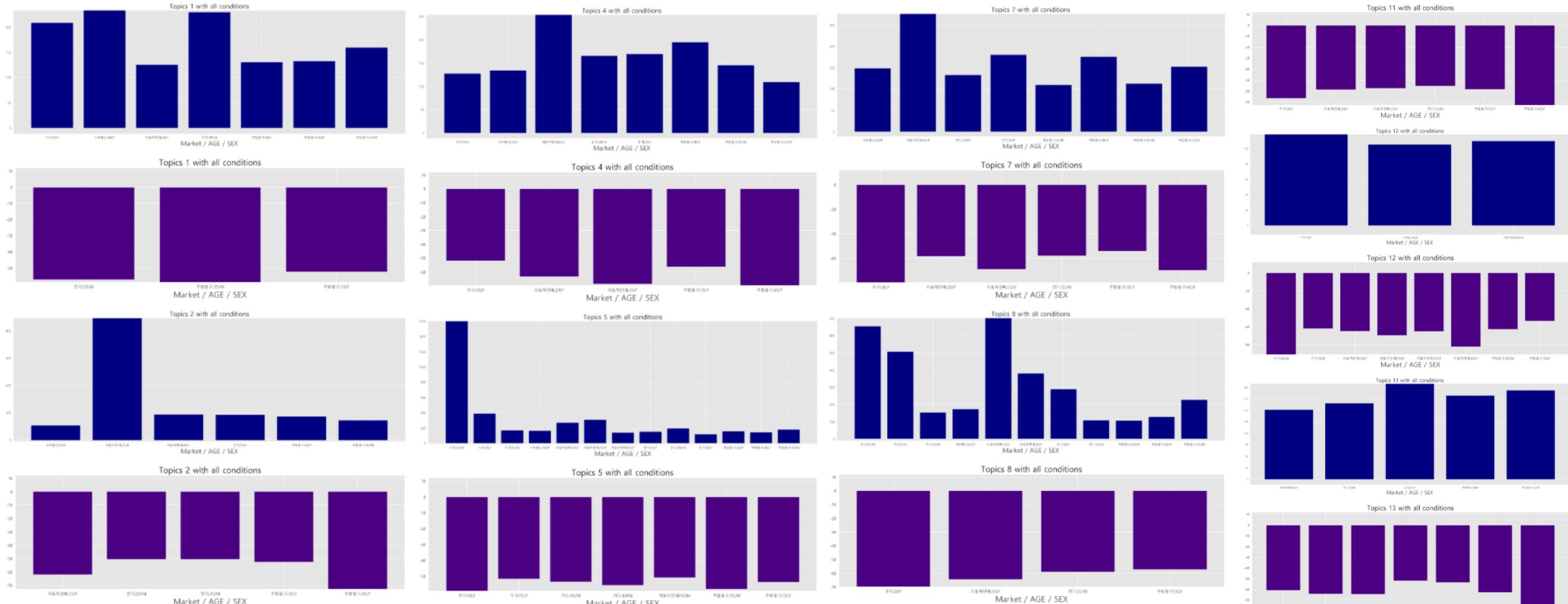
분석결과 3. 성/연령별, Topic 별 분석결과 *주요 4/4 개 (세부 사항은 첨부 코드)



Data Analysis

분석결과 4. (업종, 연령, 성)별, Topic 별 분석결과 *주요 8개 (세부 사항은 첨부 코드)

해당 분석은 매출 증가 100% 이상, 혹은 매출 감소 50% 이하인 것만 의미 있는 것으로 보고 해당 사항만 기록

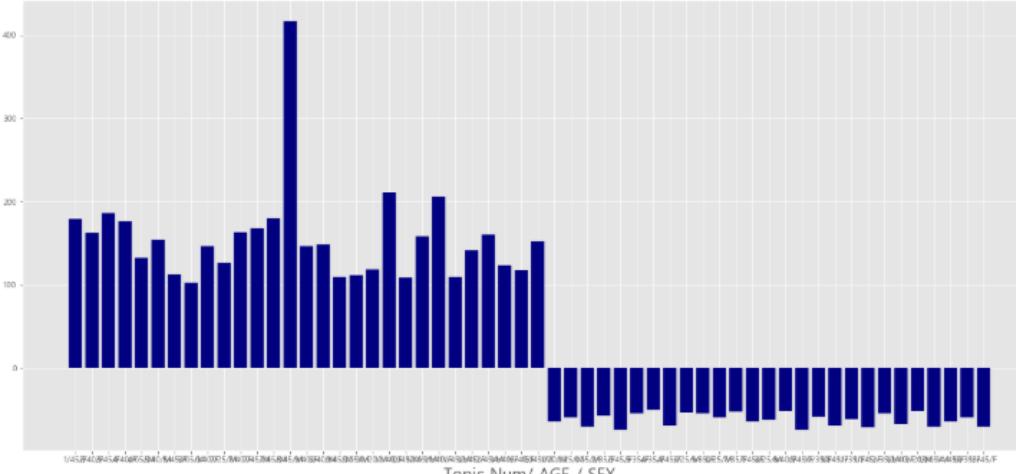


Data Analysis

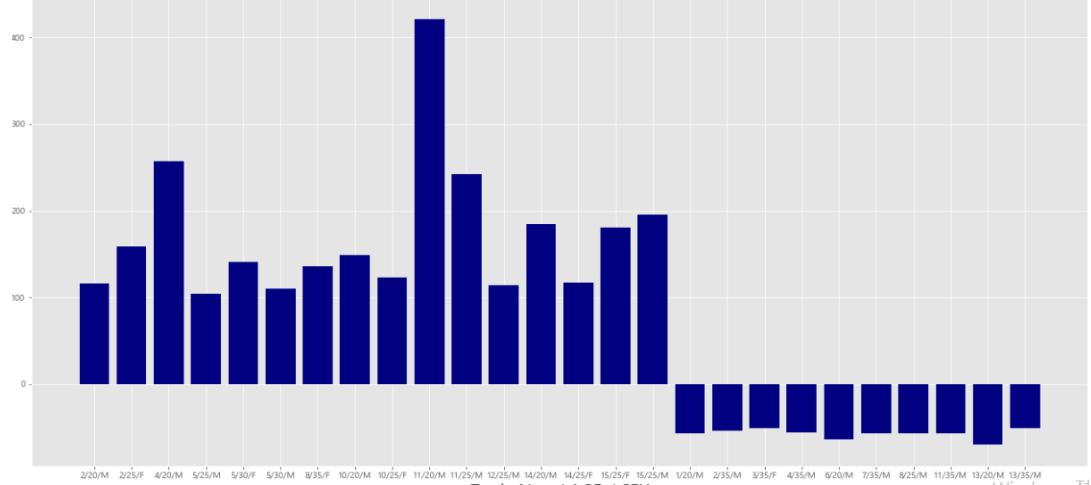
*해당 분석은 매출 증가 100% 이상, 혹은 매출 감소 50% 이하인 것만 의미 있는 것으로 보고 해당 사항만 기록

분석결과 5. (Topic, 연령, 성)별, 업종별 분석결과 *주요 4개 (세부 사항은 첨부 코드)

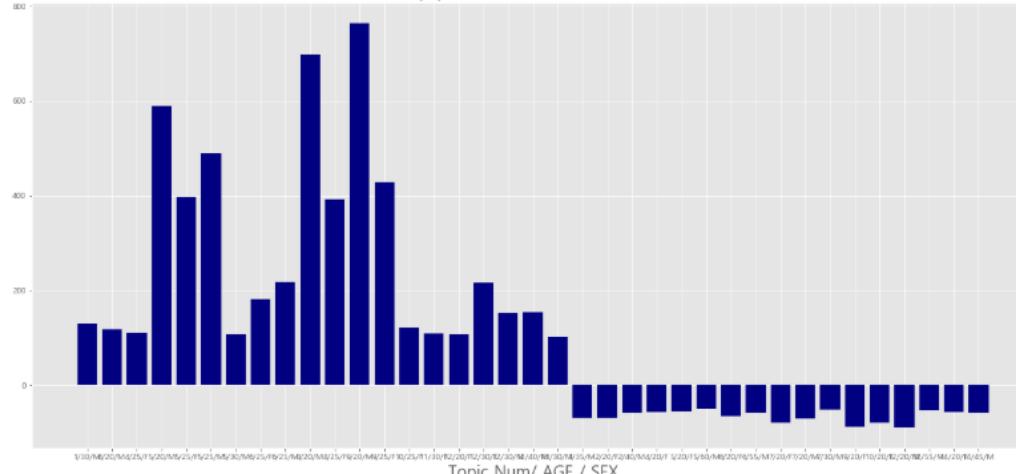
주방용구 with all conditions



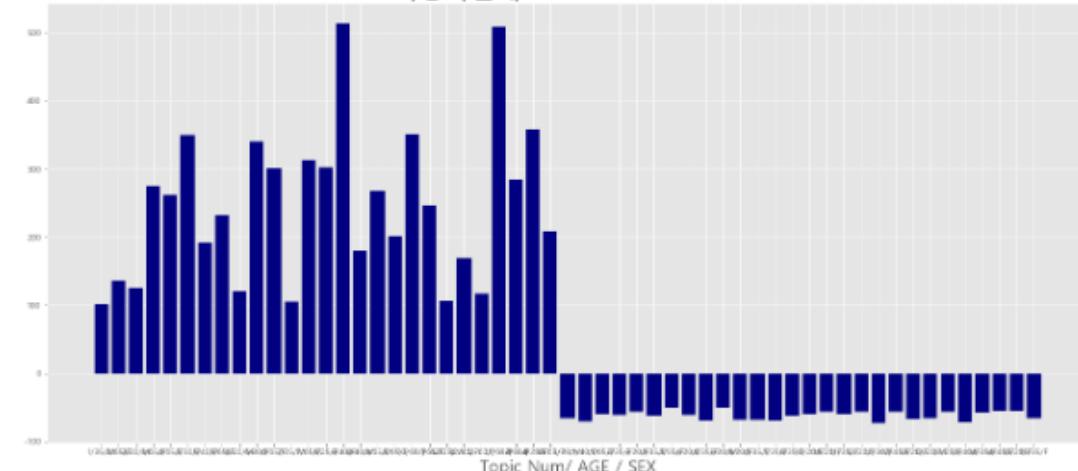
전기 with all conditions



가구 with all conditions



자동차판매 with all conditions



Data Analysis

주방용품

- Topic 7의 분포는 주방용품 매출 증감과 강한 상관관계가 있다.
- 주방용품 중 매출 증가에 기여한 제품은 인덕션*으로 추정.

* 물품 별 매출 데이터는 없으므로 증명 불가



[중앙일보] 미세먼지 피해 줄이는 생활가전 3종은?

[SBS] 주방공기 이럴 수가... 환기 안 시켰다가 '아찔'

[영향력 Top 5]

Topic 7 집, 살림

25% 매출 증가

Topic 10
기업 및 주식
21% 증가

Topic 5
정치
17 % 증가

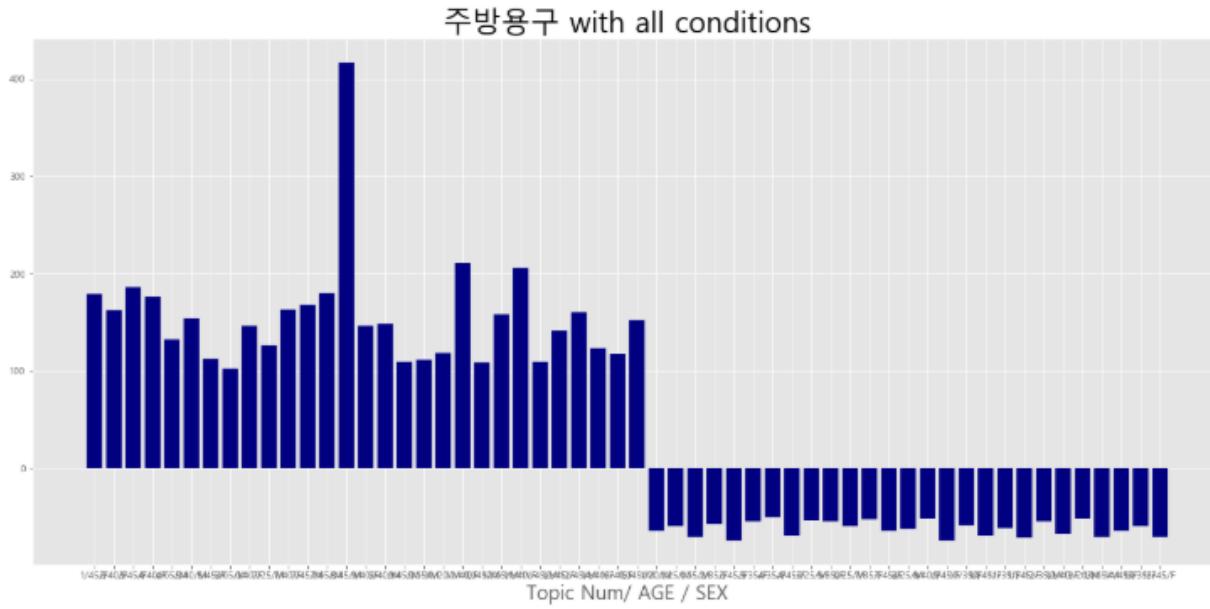
Topic 3
11 %

Topic 4
10 %

- 가스레인지에 대한 경각심이 올라간 상태에서 외부 미세먼지로 환기 불가

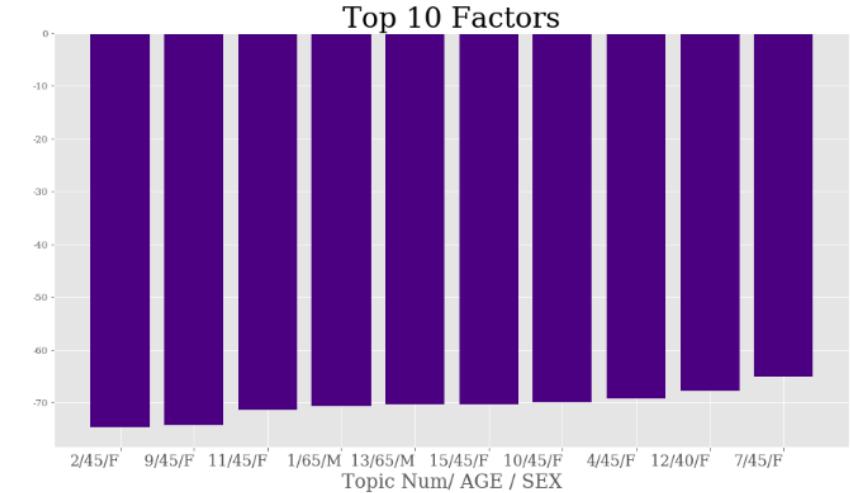
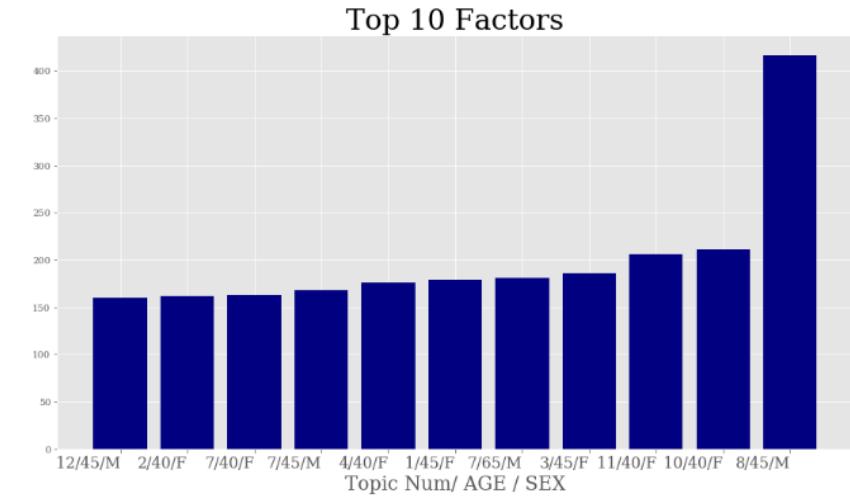
- 이 때문에 지속적인 미세먼지 노출에 의한 불안 고조로 매출 증가로 이어졌다고 추정.

Data Analysis



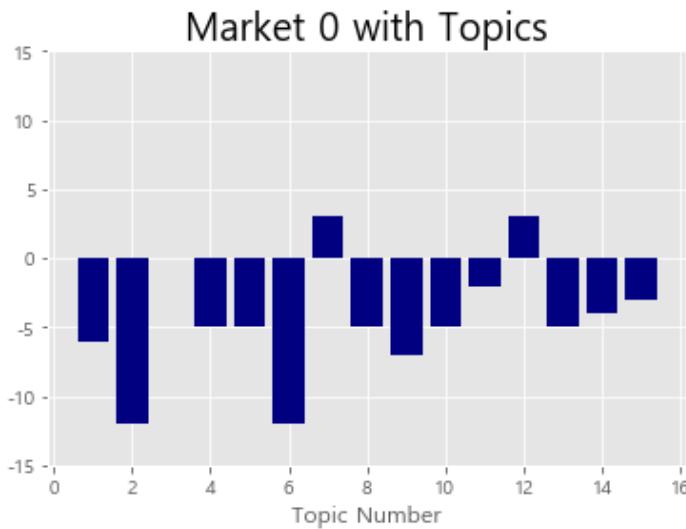
- Topic 8 상위 50 %, 45세 남성의 주방용품은 400% 이상의 매출 증가를, 거의 대부분의 Topic 상위 50%, 45세 여성의 주방용품은 70% 이하의 매출 감소를 가져왔다.
- 40~45 대 인구의 매출 증감은 대부분 100% 이상, 50% 이하로 매우 극단적으로 나타났으며 Topic 분포에 가장 강력한 상관관계를 가진 것으로 분석.

매출 증감이 가장 큰 10개



Data Analysis

가구

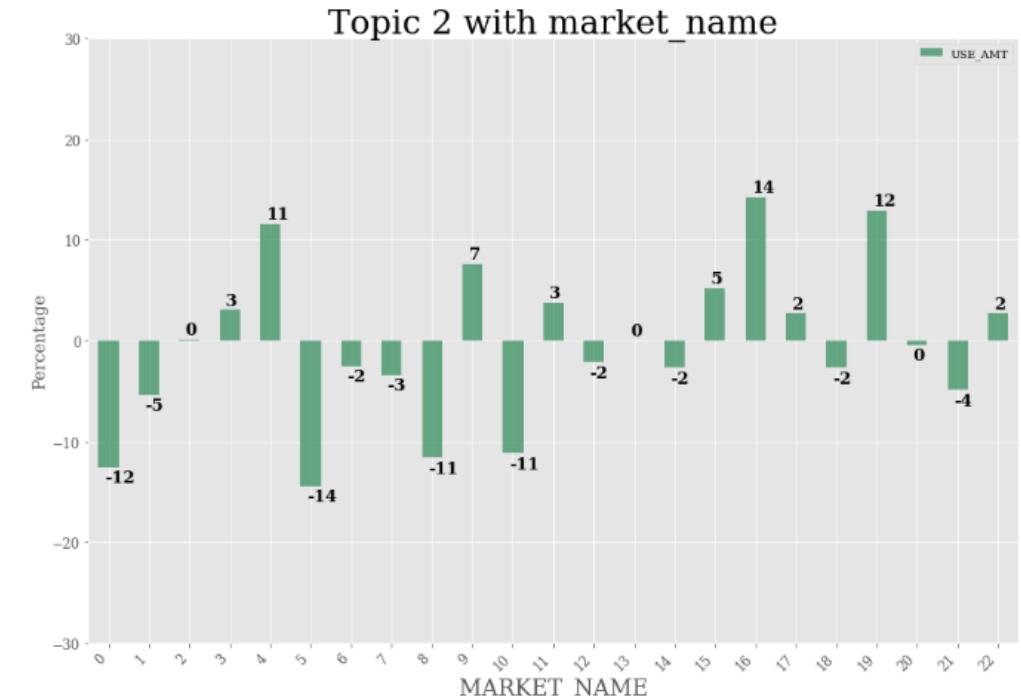


- 일부 Topic은 가구 매출 증감과 강한 상관관계가 있다.
- Topic 유형과 오프라인 가구 매출의 증감이 보다 상관관계가 존재.

[영향력 Top 2]

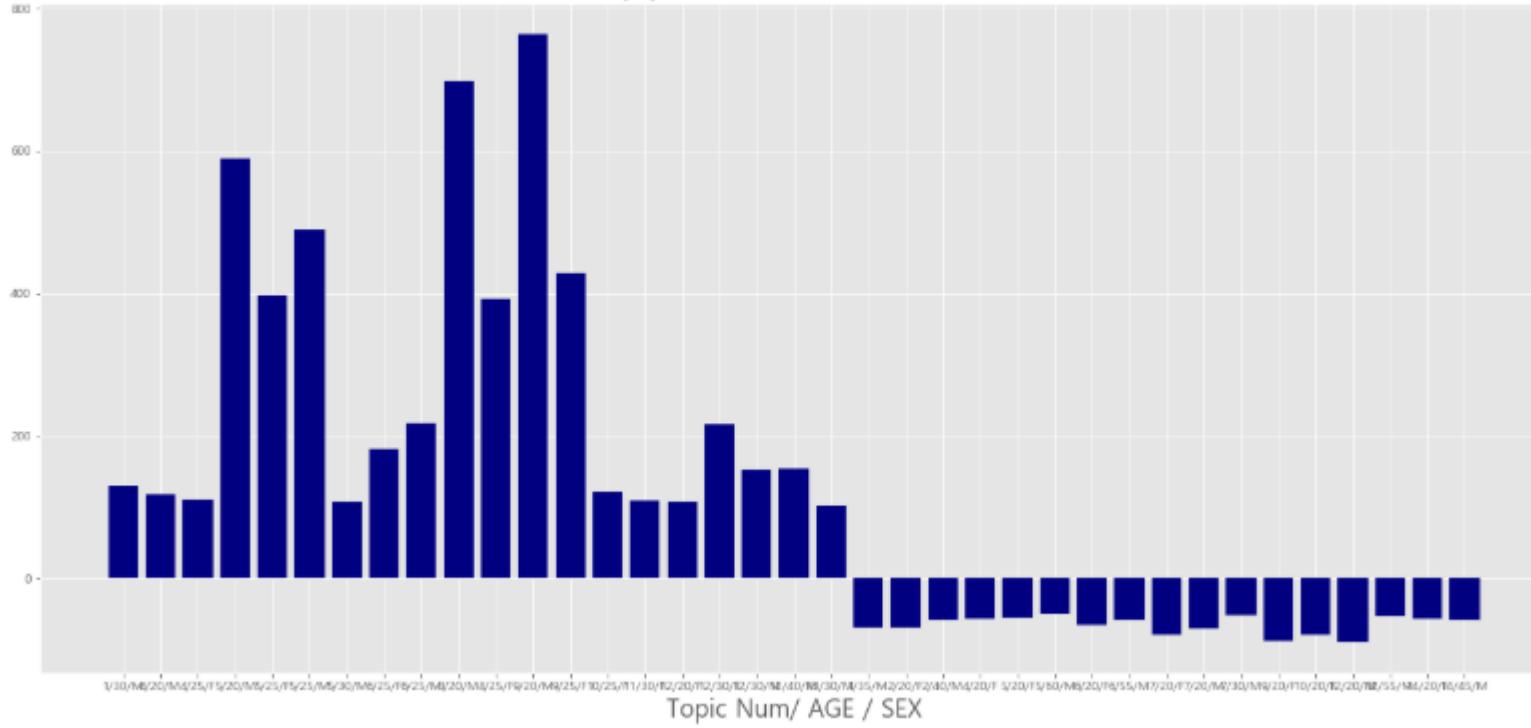
Topic 2
일상, 블로그
12 % 감소

Topic 6
피부미용
12% 감소



Data Analysis

가구 with all conditions



- 20대 남성은 Topic 8, 9 하위 50% 대비 상위 50%인 날 600% 이상의 매출 증가.
- 특정 Topic은 20대, 25대 남성과 25세 여성의 매출 증감에 강력한 영향력을 미침.



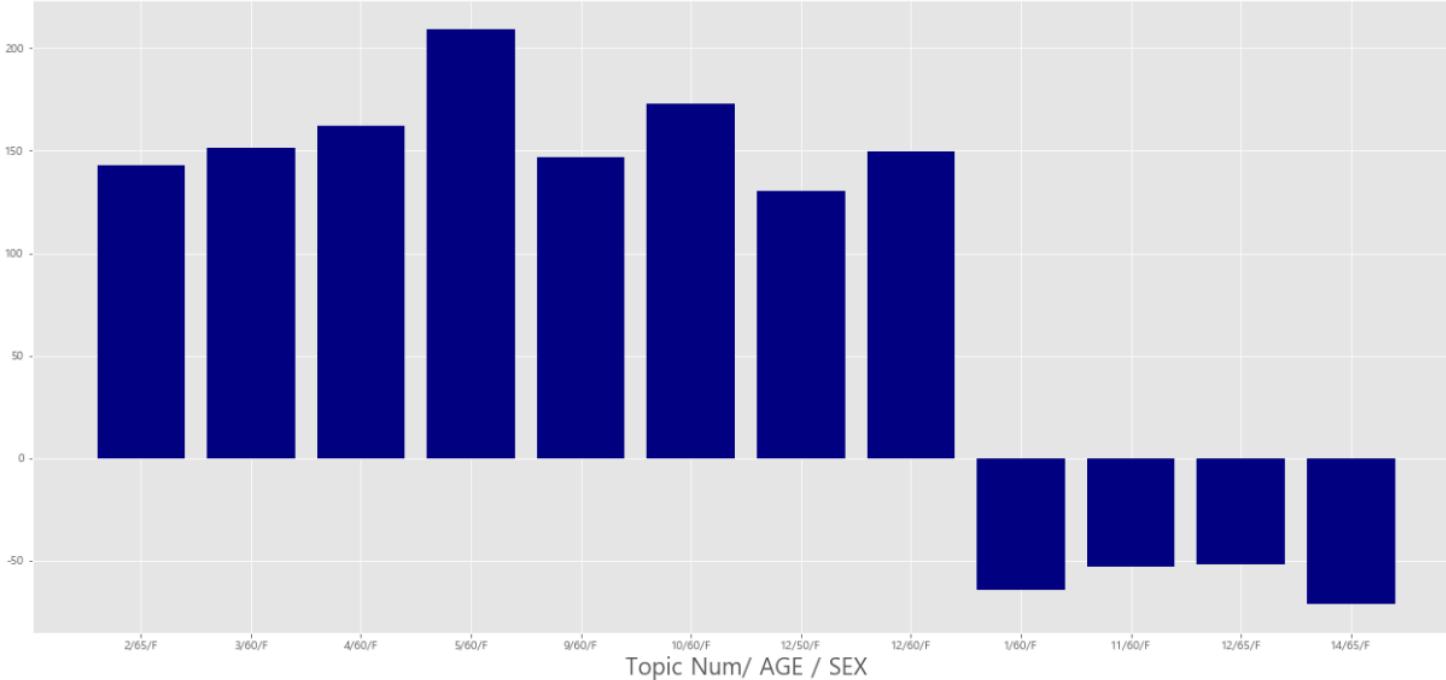
[현대엠앤티소프트]

- 이는 기존 연구에서 20대, 25대 남성/여성이 오프라인 가구 매장 IKEA를 미세먼지 나들이 장소로 인식했다는 연구와 부합
- 또한 20대, 25대가 주 고객층인 만큼 SNS에 보다 민감하게 반응한다는 상식에도 부합

Data Analysis

사무통신

사무통신 with all conditions



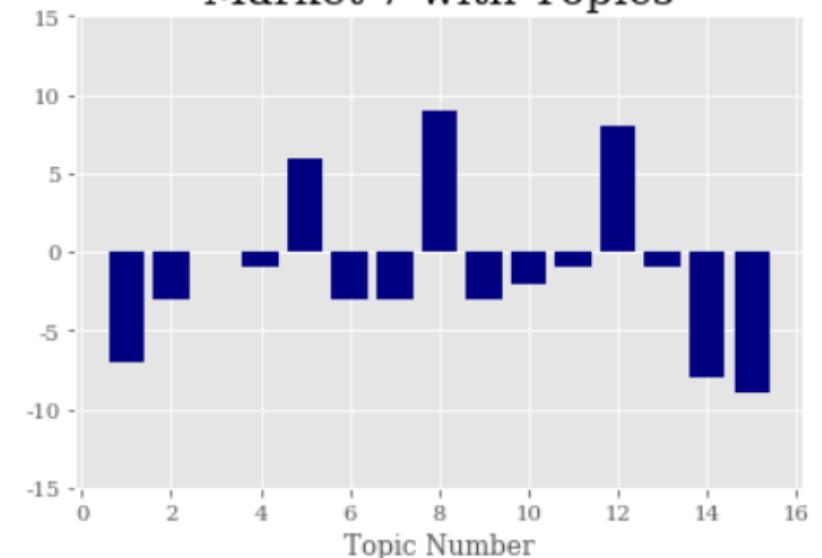
- 오프라인 사무통신은 전반적으로 Topic 8과 Topic 12 와 상관관계가 존재한다.
- 날짜 Random 이 잘 보장된 Topic 8 과 Topic 12 임으로 보다 강하게 두 토픽과의 영향 생각 가능.
- 두드러지는 특징은 Topic 분포와 60대 여성의 강한 상관관계 추정 가능.

[영향력 Top 2]

Topic 8
미세먼지 질환
9 % 증가

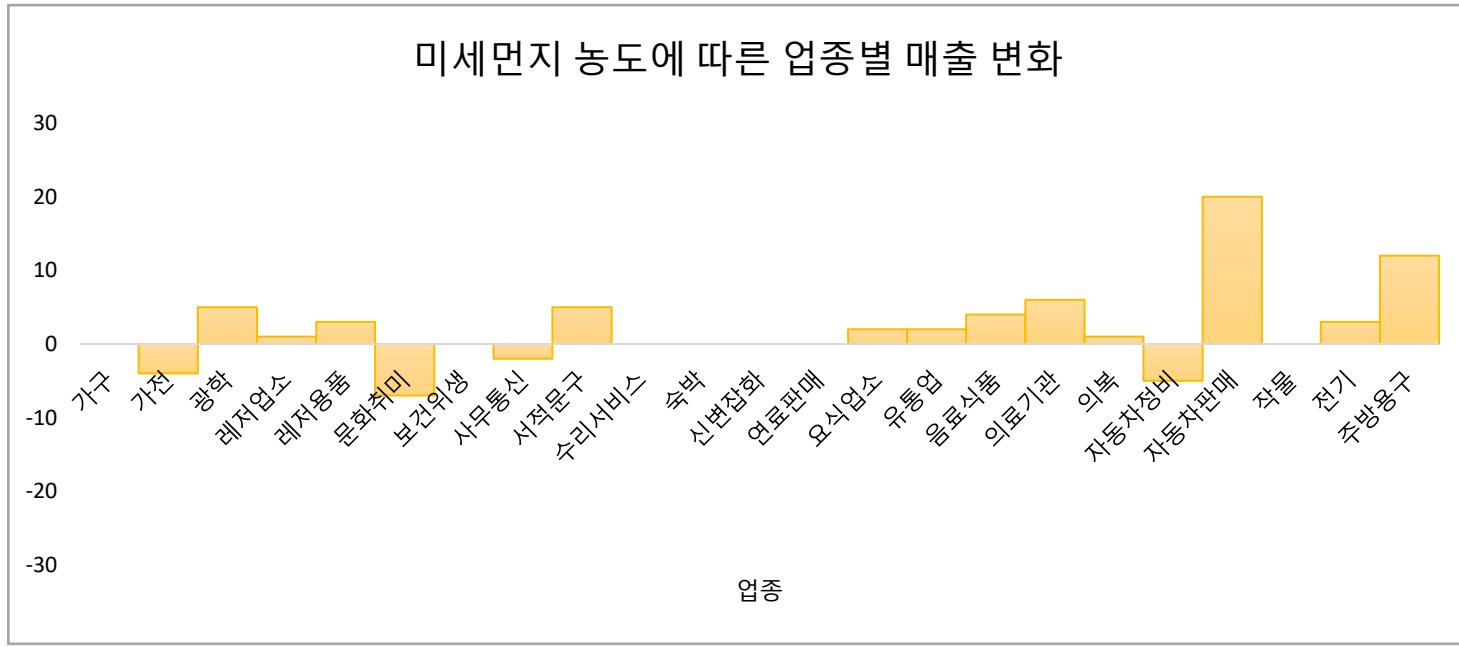
Topic 12
성형 시술
7% 증가

Market 7 with Topics

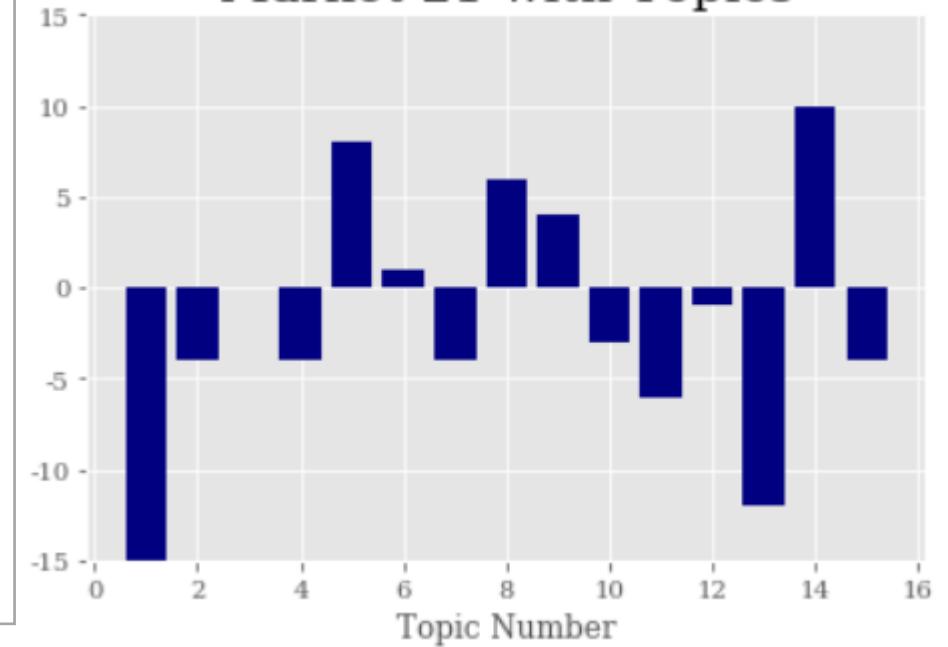


Data Analysis

전기



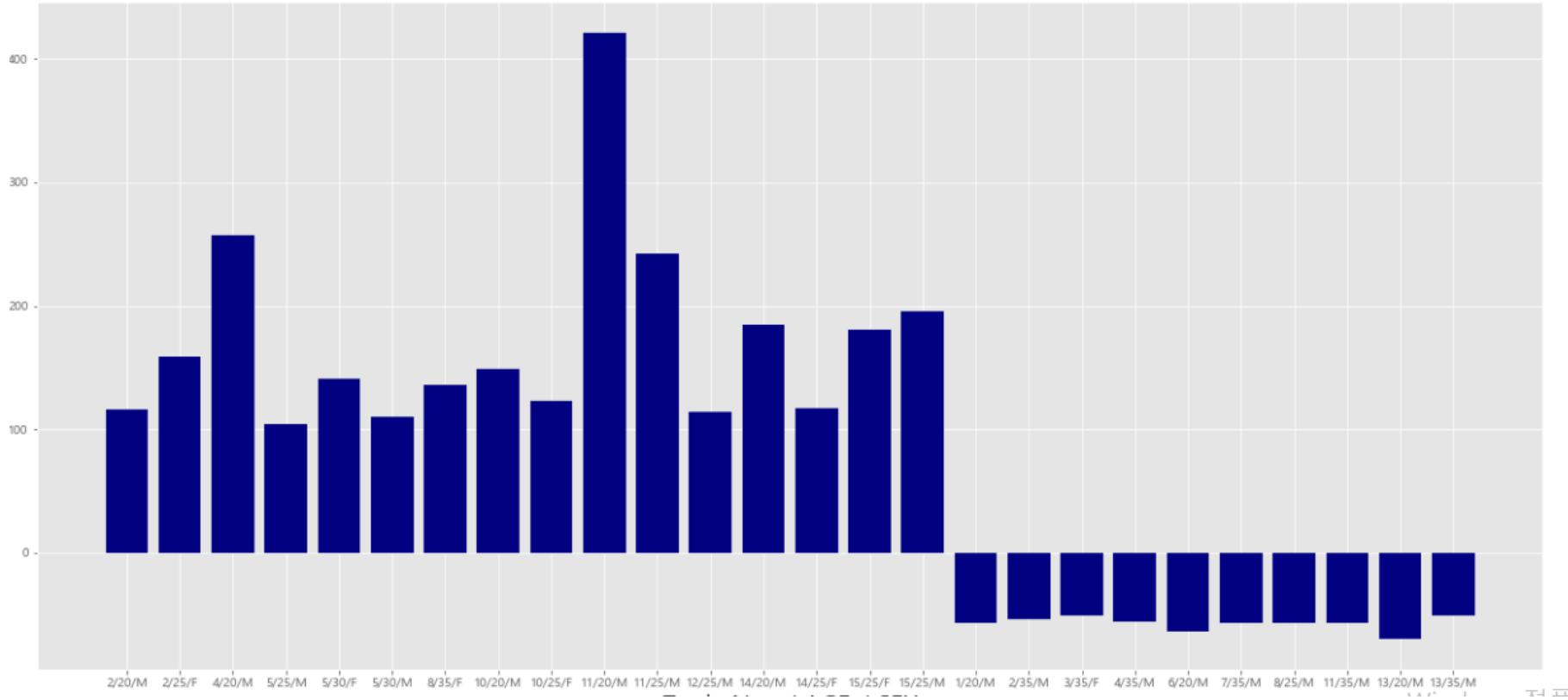
Market 21 with Topics



- 상식적으로 냉난방 기구는 미세먼지 농도나 민감도와 시기적으로 비슷함으로 매출 증대를 예상해볼 수 있으나 상관관계 매우 약함.
- 오히려 날짜 Random Split 이 잘 되어있는 Topic 13, Topic 14과 가장 강한 상관관계가 있다.

Data Analysis

전기 with all conditions



- 20대, 30대는 두드러지게 Topic 분포에 강한 상관관계를 가짐.
- 다양한 Topic에 매출 증감이 뚜렷하게 다른 것으로 나타남.

Data Analysis

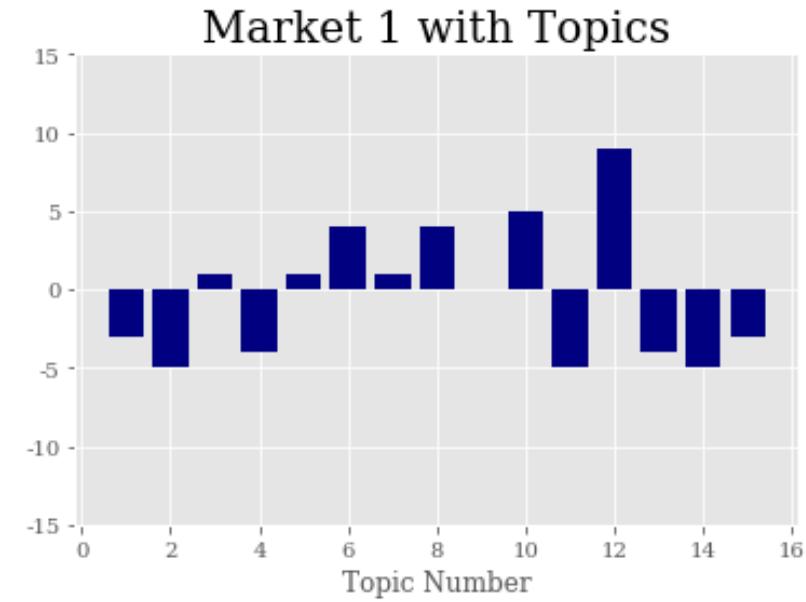
가전제품

- 일부 Topic은 가전 매출 증감과 상관관계가 있다.
- 일반적인 통념, 미세먼지 농도나 이슈에의 노출은 가전 제품 매출 증가와 상관관계가 없다.
- Topic 유형과 오프라인 가전 매출의 증감이 보다 상관관계가 존재.

[영향력 Top 2]

Topic 12
성형 및 시술
9% 증가

Topic 10
주가 및 주식
5 % 증가



Data Analysis

광학 제품

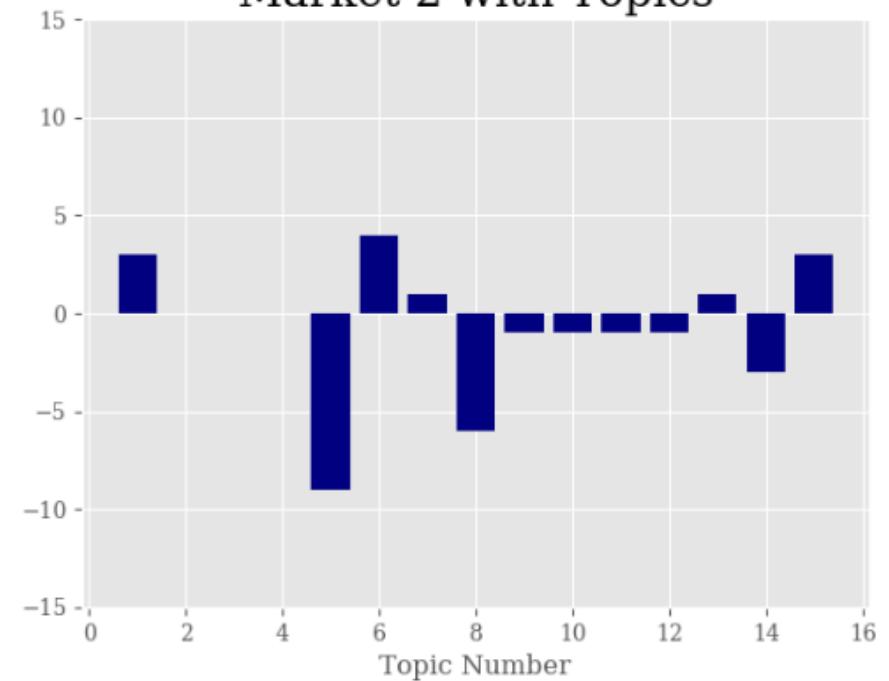
- 일부 Topic은 광학 제품 매출 증감과 상관관계가 있다.
- Topic 유형과 오프라인 광학 제품 매출의 증감이 보다 상관관계가 존재.

[영향력 Top 2]

Topic 5
정치
9 % 감소

Topic 8
미세먼지 질환
6% 감소

Market 2 with Topics



Data Analysis

레저 용품

레저 용품은 가장 독특하고 상식과 위배되는 결과

1. 미세먼지 농도와 매출은 상관관계가 없다.
2. 대부분의 Topic 유형은 뚜렷한 매출 증가 상관관계가 있다.
3. Topic2, 12는 뚜렷한 매출 증감 상관관계가 있다.

언론 및 연구도 각기 다른 결과를 보이고 있어서, 선별은 추측도 힘들다.

다른 야외 스포츠 용품은 자전거와 보드(7%)·야구(-3%)·배드민턴(8%) 등은 이 기간 감소

골프용품, 미세먼지 속에도 지난해 대비 판매량 25% 상승

미세먼지 넘어선 봄 맞이 캠핑 열풍

텐트, 캠핑의자, 코펠 등 '캠핑용품' 매출 46% 증가

기간: 2019년 4월 1일 ~ 4월 12일 매출 기준 (2019년 3월 20일 ~ 3월 31일 대비)

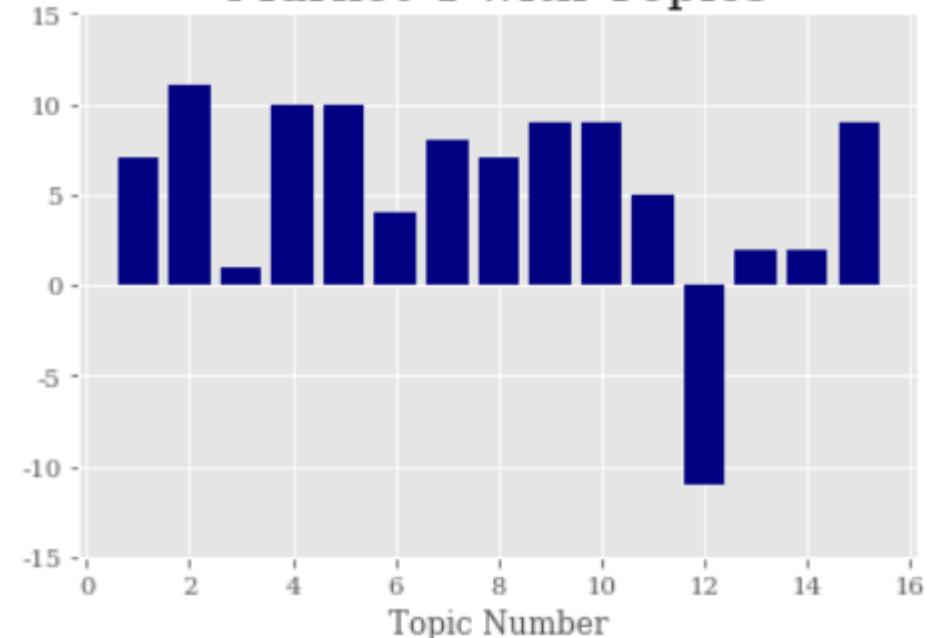


[영향력 Top 2]

Topic 2
정부 정책
11% 매출 증가

Topic 12
성형 및 시술
11 % 감소

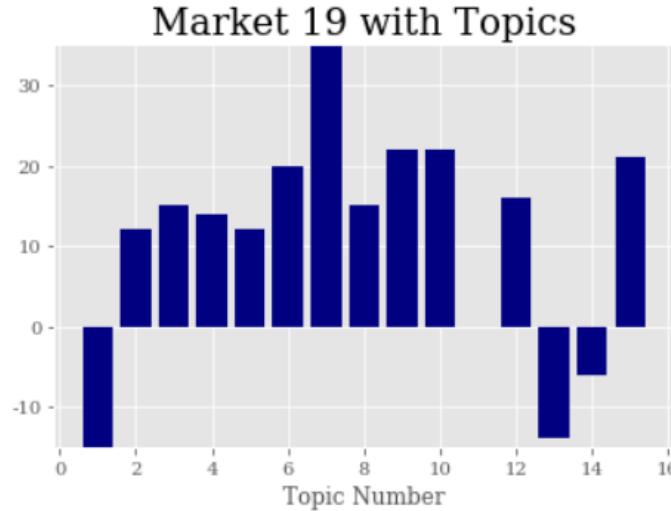
Market 4 with Topics



Data Analysis

자동차 판매

- 대부분의 Topic 및 미세먼지 농도는 자동차 판매 매출 증가와 강한 상관관계가 있다. 그러나 Topic 1 , Topic 13과는 음의 상관관계를 가짐.
- 이는 “The Psychological Effect of Weather on Car Purchases” 를 비롯한 다양한 연구에서 이미 확인 된 바로, 소비자들은 날씨가 좋지 않은 날 신차 구매를 희망한다. 이는 본 연구와도 부분적으로 상충하는 결과.



[영향력 Top 5]

Topic 7
집, 살림
35% 증가

Topic 9
기상예보
21 % 증가

Topic 1
일상, 블로그
15 % 감소

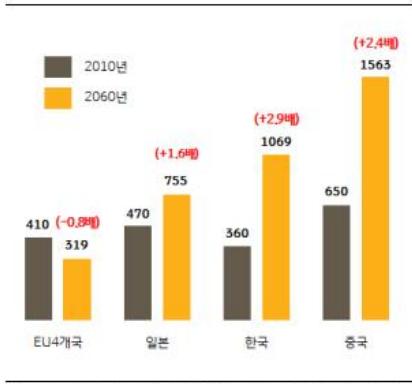
Topic 13
광고 및 홍보
12 % 감소

Data Analysis

의료기관

- 대부분의 Topic은 의료기관 매출 증가와 강한 상관관계가 있다.
- 단, Topic의 주제에 크게 다르지 않고 미세먼지 농도와 강한 상관관계는 측정되지 않으므로 대중의 미세먼지 이슈 노출 여부에 더 큰 영향을 받는다고 분석.

[그림 6] 대기오염 영향 연간 조기사망자수 전망



[표 5] 대기오염에 따른 건강영향(글로벌 기준)

구분	2010	2060
호흡기질환 (백만 건)		
기관지염(6-12세)	12	36
만성기관지염(성인)	3.5	10
천식 증상 (백만 일)		
천식증상일수(5-19세)	118	360
건강보험비용 (백만 일)		
입원일수	3.6	11
활동제약일수 (백만 일)		
근무일수상실	1,240	3,750
활동제약	4,930	14,900

주: 분석대상 조기오염 물질은 초미세먼지와 오зон으로

100만명당 조기 사망자수임

자료: 보험연구원, OECD

[KB금융지주 연구소] 일상화된 미세먼지 이용 행태와 의료 행위 변화

[서울신문] 미세먼지로 추가 의료비 연 450억

[영향력 Top 5]

Topic 5 정치

13% 매출 증가

Topic 13
광고 및 홍보
13% 증가

Topic 11
자동차 정보 및 거래
11 % 증가

Topic 15
다이어트 및 운동
11 % 증가

Topic 2
정부 미세먼지 정책
11 % 증가

Data Analysis

문화 취미, 숙박, 서적문구

- 대부분의 Topic 은 문화취미, 숙박, 서적 문구 등 여가활동 매출 감소로 이어졌다.
- 단, Topic 의 주제에 크게 다르지 않고 미세먼지 농도와 강한 상관관계는 측정 되지 않으므로 대중의 미세먼지 이슈 노출 여부에 더 큰 영향을 받는다고 분석.

[포토]고농도 미세먼지에 서점으로 피신

f t d 최종수정 2019.01.13 14:35 기사입력 2019.01.13 14:35 댓글 쓰기

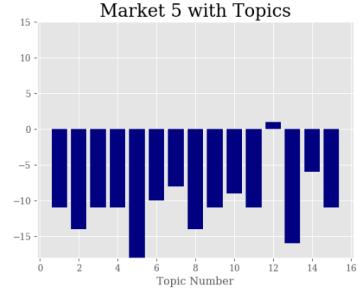


[아시아 경제]

*특히 일부 언론은 미세먼지가 높으면 실내 서점, 영화관 등으로 발길을 옮기는 서술을 하나 명확히 확인된 바는 없다.

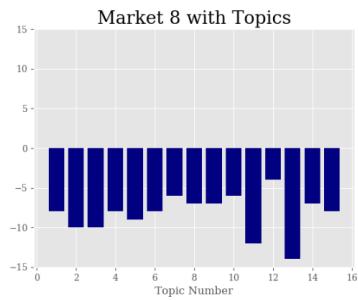
[문화 취미 영향력 Top 1]

Topic 4 / Topic 8
-14 % 매출 감소



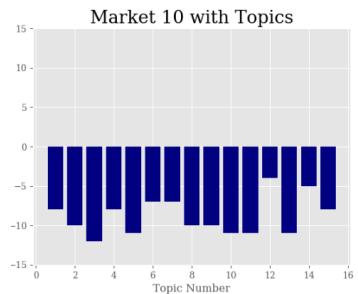
[서적 문구 영향력 Top 1]

Topic 13
-14 % 매출 감소



[숙박 영향력 Top 1]

Topic 11 / Topic 13
-11 % 매출 감소



Data Analysis

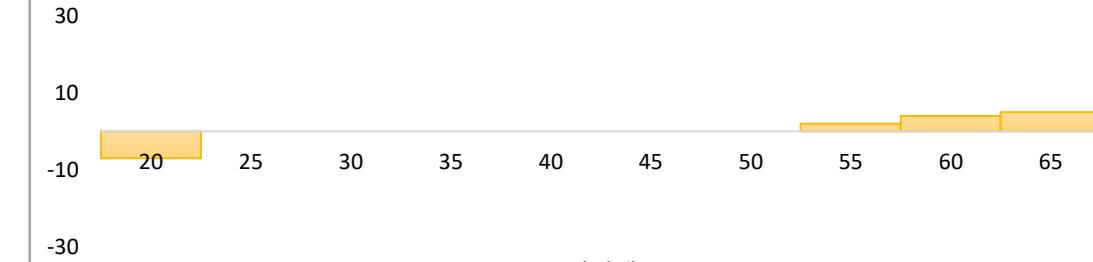
연령대별

- 20대는 미세먼지 농도, 미세먼지 이슈와 음의 상관관계를, 60~70세는 양의 상관관계를 보여준다. 외의 연령대는 차이가 존재하지 않는다.
- 사회적 통념인 나이가 많아질수록 미세먼지에 둔감에 어느정도 부합
- 다만 25~55 세가 직업 인구임을 감안하면 설불리 인과관계를 추론할 수 없다.

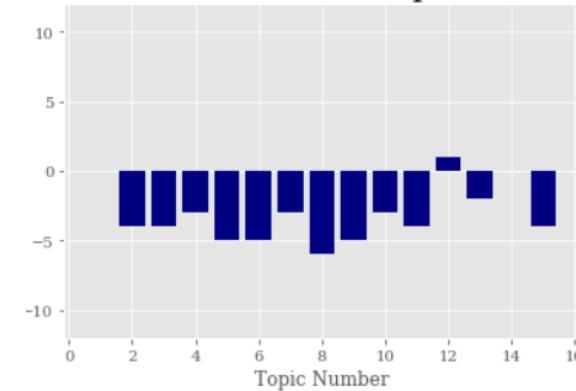


[연합뉴스]

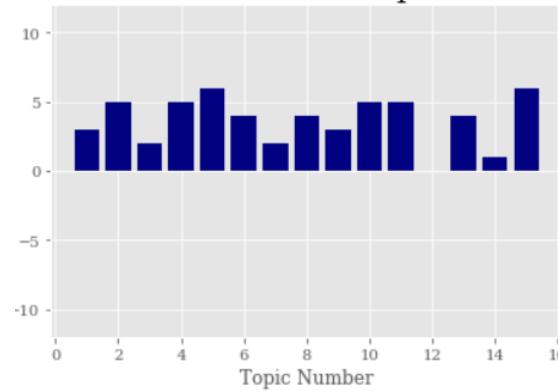
미세먼지 농도에 따른 연령별 매출 변화



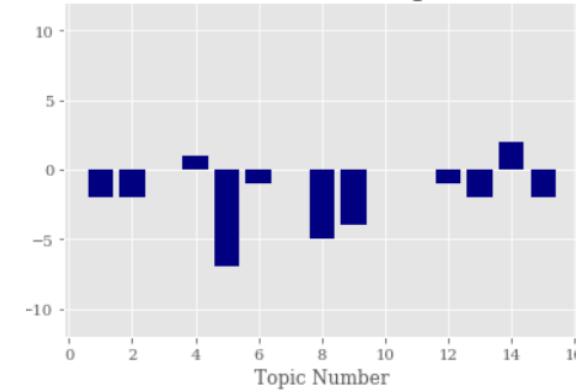
AGE 20 with Topics



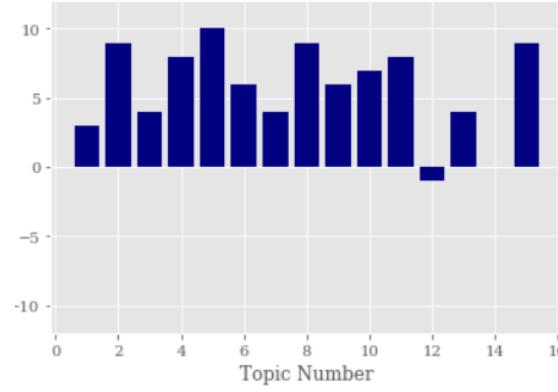
AGE 60 with Topics



AGE 25 with Topics



AGE 65 with Topics



Result

결과 정리

- 업종 별 매출은 미세먼지 농도보다 Topic 분포와 강한 상관관계가 존재
- 상식과 위배되는 매출 추이가 많고 Topic 분포와의 강한 상관관계가 존재
- 매출과 Topic 분포와의 상관관계는 연령 / 성별로 인구분포를 특정 지었을 시 보다 강한 상관관계가 존재

Result

제안 - 1 Topic Tracking

- Topic 과 특정 인구 분포, 업종과의 강한 상관관계를 통해 SNS 및 자연어 데이터 분석 방향 제시 = 본 연구로 Topic Modeling 가치 증명
- 기업은 실시간으로 변화하는 SNS/온라인 Topic을 매우 적은 비용으로 Tracking 가능.

An Online Topic Modeling Framework with Topics Automatically Labeled

Fenglei Jin¹ Cuiyun Gao¹ Michael R. Lyu¹

Abstract

In this paper, we propose a novel online topic tracking framework, named IEDL, for tracking the topic changes related to deep learning techniques on Stack Exchange and automatically interpreting each identified topic. The proposed framework combines the prior topic distributions in a time window during inferring the topics in current time slice, and introduces a new ranking scheme to select most representative phrases and sentences for the inferred topics in each time slice. Experiments on 7,076 Stack Exchange posts show the effectiveness of IEDL in tracking topic changes and labeling topics.

1. Introduction

Recent advances in deep learning promote the innovation of many intelligent systems and applications such as autonomous driving and image recognition. Tracking the changes of focus for deep learning engineers and researchers is helpful to identify current emerging deep learning-related topics. In this work, we choose Stack Exchange to col-

are involved during inferring the current topic distribution. In (Espinoza et al., 2018), the proposed approach also selects sentiment words for each topic based on Dynamic Topic Model (DTM) (Blei & Lafferty, 2006).

Inspired by the recent work (Gao et al., 2018), where an adaptively online latent Dirichlet allocation approach, named IDEA, is introduced to track user opinions in user feedback, and outperforms the OLDA approach (AlSumait et al., 2008), we propose a new framework **IEDL** for Identifying Emerging Deep Learning-related topics. The difference between IDEA and our approach lies in the combination styles of the prior topic distributions. In IDEA, the similarities between the topics in previous time slices and those in the previous one time slice are taken into account for inferring the topics in current time slice, while we introduce an exponential decay function in a time window. Besides, we propose a novel topic labeling approach based on the unique characteristics of Stack Exchange posts.

The experimental results on 7,076 Stack Exchange posts verify the effectiveness of IEDL in detecting topic changes and topic labeling.

The contributions of our paper are elaborated as below.

- 실제로 실시간으로 Topic을 Tracking 할 수 있는 효율적인 방법인 Online Topic Modeling이 개발되고 많은 IT 기업들이 응용 중.
- 실시간 Topic 추이를 관측함으로써 보다 강력한 시계열 분석 수행 가능 / 또한 밝혀진 상관관계를 통해 특정 인구 분포에 대한 업계 매출 증감 대비 및 전략 수립 가능.
- Topic Modeling 이 Semantic 분석, Document Filtering에도 응용되는 등 빠르게 그 유용성 입증 중. 기업은 현재의 가치와 앞으로의 폭발적인 성장에 대비하여 양질의 Dataset을 지금부터 확보해야함. = Topic Tracking 유용성 인식

Result

제안 - 2 추가적인 Data와 결합

- 해당 분석을 통해 경험적 / 상식적으로 유추하기 힘든 다양한 상관관계 확보
- 그러나 이는 업종 내 물품 별 Data / 소비자의 소득 분포 배제, / 1년 치 Dataset, 매우 제한적인 Dataset만으로 이끌어낸 결과
- 기업이 가지고 있는 Domain Specific Dataset 과 Topic Modeling을 결합하면 데이터 분석 및 통찰 능력 향상.

제안 - 3 상관관계를 넘은 인과관계 증명

- Data 분석을 통해 인과관계를 증명하는 것은 어려움
- 그러나 Topic Model과 인과관계를 증명하면 그 활용성은 무한.
- ex) 기업이 인위적으로 Topic Distribution을 변형해서 원하는 매출 증 / 감 유도
- 타겟 시장과 타겟 물품의 온라인 자연어 영향력 특정 가능, 명확한 마케팅 전략 수립
- 확보된 Data를 바탕으로 예측 Modeling, Topic을 이용한 업종 매출 예측.

Result

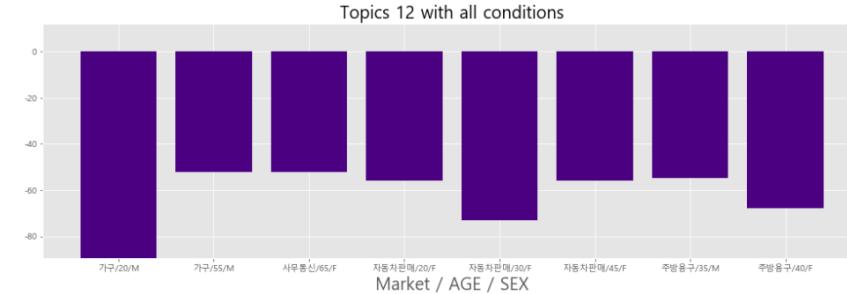
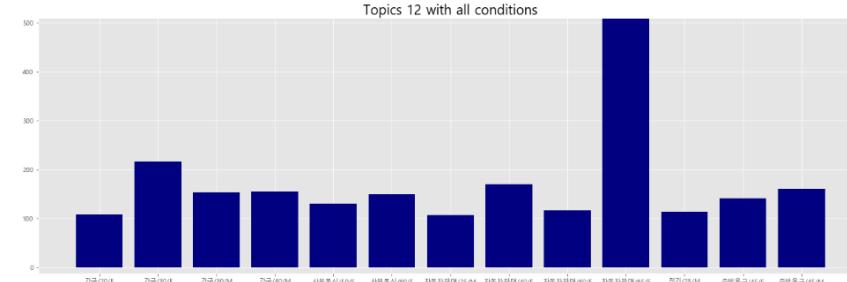
제안

- 특히 특정 인구 분포에 대한 상관관계 및 후속 연구는 카드사 수익악화 실적 완화에 큰 도움이 될 것으로 예상됨.

카드사 수익악화 비상...전문가들이 제시한

박호식 기자 hspark@bizwatch.co.kr
2018.12.04(화) 17:50

여신협회 '위기와 지속성장 모색' 포럼
수수료 개편, 3년간 순익 1.5조·고객혜택 8천억 감소 전망
전문가, NFC결제 확대·빅데이터 공동사업·송금업 허용 등 제시



- 특히 Topic 5는 20~40대, Topic 12 40대 ~60대와는 몇몇 토픽은 강력한 상관관계를 가짐에 따라 Topic과 인구분포의 상관관계를 시사함.
- 또한 Topic 5는 정치적인 뉴스가 대부분이며 대중의 매출이 정치적인 뉴스로 이어지는 인과가 있다고 판단이 힘듦.
- Topic 5 와 Topic 12는 계절적 패턴 등의 영향력도 최소화 함.
- 조심스럽게 Topic의 분포가 기업의 매출에 영향을 준다고 짐작해볼 수 있음.
- 또한 Topic Modeling은 앞으로 더욱 많은 활용가치를 지니고 있음으로 기업은 빠르게 데 이터 확보 필요.

Result

제안 - 예시

- Topic 12 와 40대 남성 자동차 업종 매출과의 인과관계 관측

1. 판촉

Topic 12 급증이 관측되는 날 40대 남성 고객층에게 고객유치전략(할인, 제휴 등) 제시, 소비 유도. 카드 가맹점과 유도전략을 통해 보다 최소의 판촉으로 최대의 효과



2. 수익 증대

Topic 12 를 인위적으로 증가시켜 40대 남성 고객층의 자동차 업종 판매량 증대

3. 업종 가치 판단

Topic 분포 추이를 통해 향후 자동차 업종의 매출 증 / 감 예측





감사합니다