토픽 모델링 기반 마케팅 솔루션:
“워드온(WordON)”

5조 (이보성 박혜연 손소연 백종휘)

빅데이터 예측분석 조별프로젝트 (정화민 교수님)

2021-2 서강대학교 정보통신대학원

역할 구성.

이보성(조장)

- 조장 및 디렉팅
- Data 전처리 및 분석
- 프로젝트 발표 (목차 5-6)

박혜연

- Data 전처리 및 분석
- 솔루션 메인 코딩

손소연

- Data 분석
- 사업화 전략 구성
- PPT 자료 제작

백종휘

- 자료 수집 및 구성
- BM 구성 및 프로젝트 발표

CONTENTS

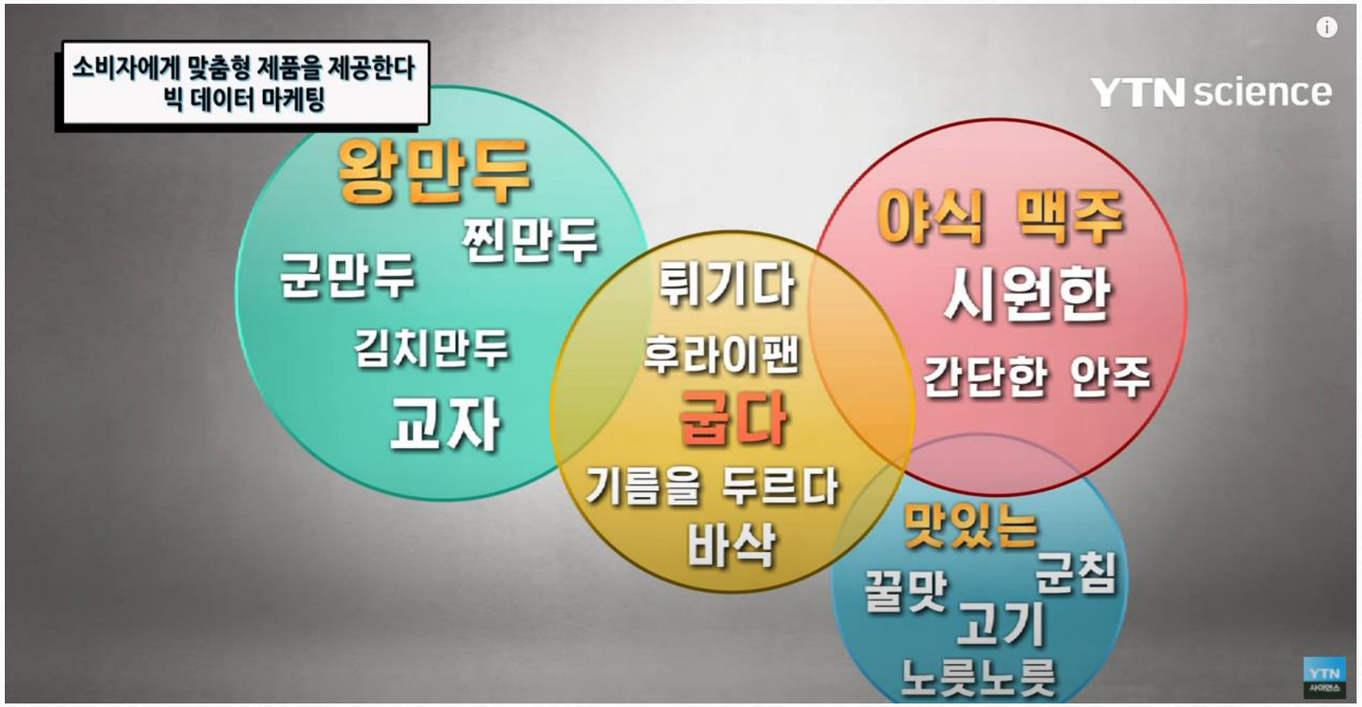
01

제안배경

제안 배경 및 관련 사례

01.
제안 배경 - 관련사례1

"소비자에게 맞춤형 제품을 제공한다, 빅 데이터 마케팅 / YTN 사이언스"
2016. 11. 29.



소비자에게 맞춤형 제품을 제공한다, 빅 데이터 마케팅 / YTN 사이언스 - YouTube

1. 식품, 유통업계의 빅데이터 수집, 분석을 통한 사례
1) '김치찌개 맛 감자칩'이 편의점에서 히트
→ 장바구니 분석을 통해 편의점에서 김치찌개 맛 컵라면을 살 때 감자 스낵도 함께 결제된 경우가 많다는 것을 알아내고, 인사이트를 얻어 상품 출시

2) 대용량 요구르트
→ 편의점 계산대에서 축적되는 소비자들의 구매 패턴에서 한 번에 3개 이상을 마시는 고객이 많은 것에서 착안해 용량을 3배로 늘린 제품 출시

3) 맥주와 만두를 묶어 판매해 판매량 증대
→ 일반인들의 SNS 글 41억 건을 분석하여 집에서 맥주를 마실 때 만두를 안주로 자주 먹는다는 패턴을 알아냄

01.
제안 배경 - 관련사례2

불스원의 “불스원샷”
이미 국내 시장 점유율 **90%**를 차지하는 **1**위 기업이지만, 빅데이터 분석으로 기존 고객의 이탈 방지에 성공

불스원샷 연관 검색어 Top 20

1	불스원샷 넣는법	980	11	LPG연료첨가제	438
2	불스원샷 경유	908	12	연료첨가제	432
3	불스원샷 70000	905	13	불스원샷 휘발유	420
4	불스원샷 효과	824	14	불스원샷 프로	409
5	불스원샷 디젤	717	15	엔진세정제	308
6	불스원샷 프리미엄	707	16	불스원샷 주기	284
7	불스원샷 뉴카	663	17	불스파워	200
8	불스원샷 사용법	536	18	불스원	174
9	디젤연료첨가제	514	19	불스원샷 신제품	158
10	검아웃	447	20	카텍몰	75

[불스원샷 연관 검색어]

1) 불스원샷 제품에 대한 뉴스, 커뮤니티, 블로그, 카페 상의 빅데이터를 수집해 키워드 분석을 진행

2) 불스원샷 연관 검색어와 화제어 분석 결과, 토픽 상위 20개 중 ‘넣는 법'과 ‘사용법'과 같은 의미의 키워드가 상위에 나타났고,
이를 심층 분석하여 많은 고객들이 불스원샷의 용도와 사용법을 모른다는 시사점을 도출

3) 제품을 알면서도 사용 방법과 효과를 제대로 인지하지 못한 고객군을 타겟으로 교육적 차원의 마케팅을 진행

4) 해외 경쟁사의 국내 진출로 기존 고객마저도 빼앗길 수 있다는 위기에서 탈출!

02. 시장규모

2021년 국내 광고시장 규모 12조원
PC/모바일 광고시장 6조 2천억원. 매년 증가세.

□ 2019-2021 년 매체 별 총 광고비

(단위: 억 원, %)

구분	매체	광고비(억 원)			성장률(%)		구성비(%)	
		'19 년	'20 년	'21 년(F)	'20 년	'21 년(F)	'20 년	'21 년(F)
디지털	PC	17,708	18,548	19,410	4.7	4.6	15.5	15.5
	모바일	32,824	38,558	42,570	17.5	10.4	32.1	33.9
	디지털 계	**50,532**	**57,106**	**61,980**	**13.0**	**8.5**	**47.6**	**49.4**
제작		5,101	4,384	4,585	-14.1	4.6	3.7	3.7
총 계		**120,926**	**119,951**	**125,500**	**-0.8**	**4.6**	**100.0**	**100.0**

[출처 : 제일기획, 대한민국 총 광고비 결산 및 전망 발표 2021]

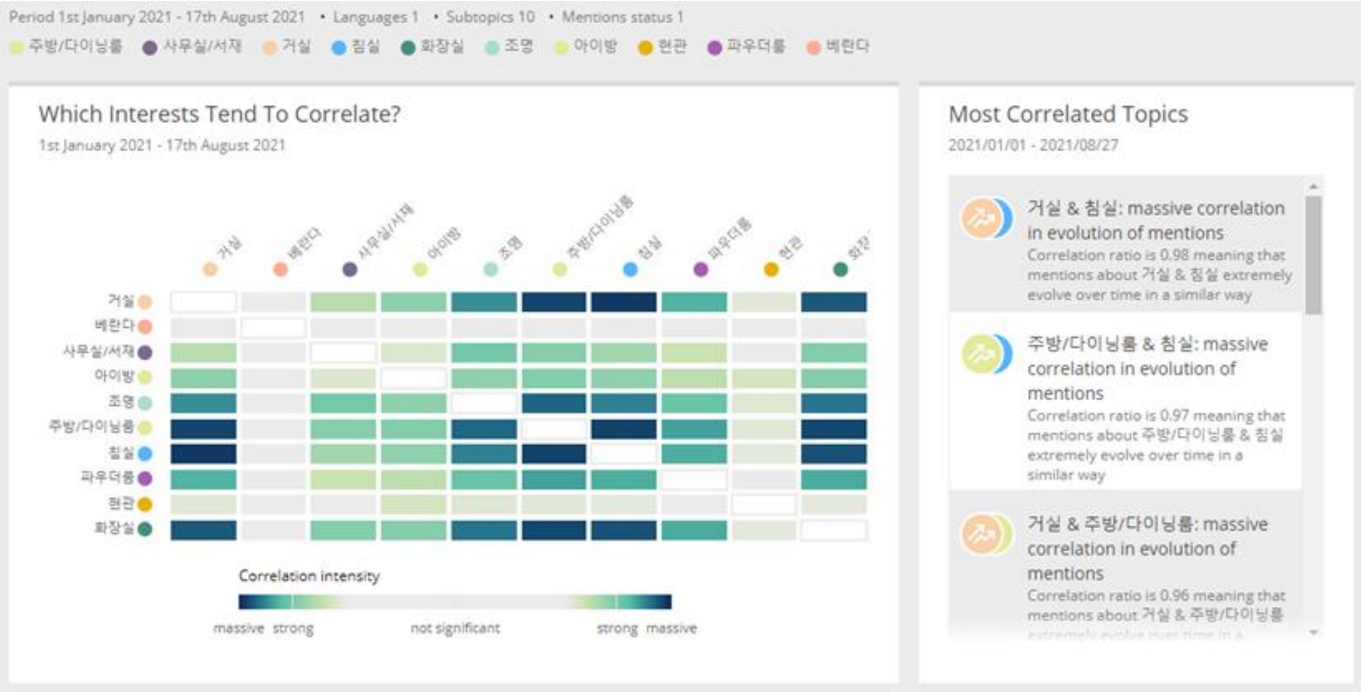
02

시장 조사

Market Research

01. 시장조사

신디지오(SYNTHESIO)
https://www.sm2marketing.co.kr/synthesio



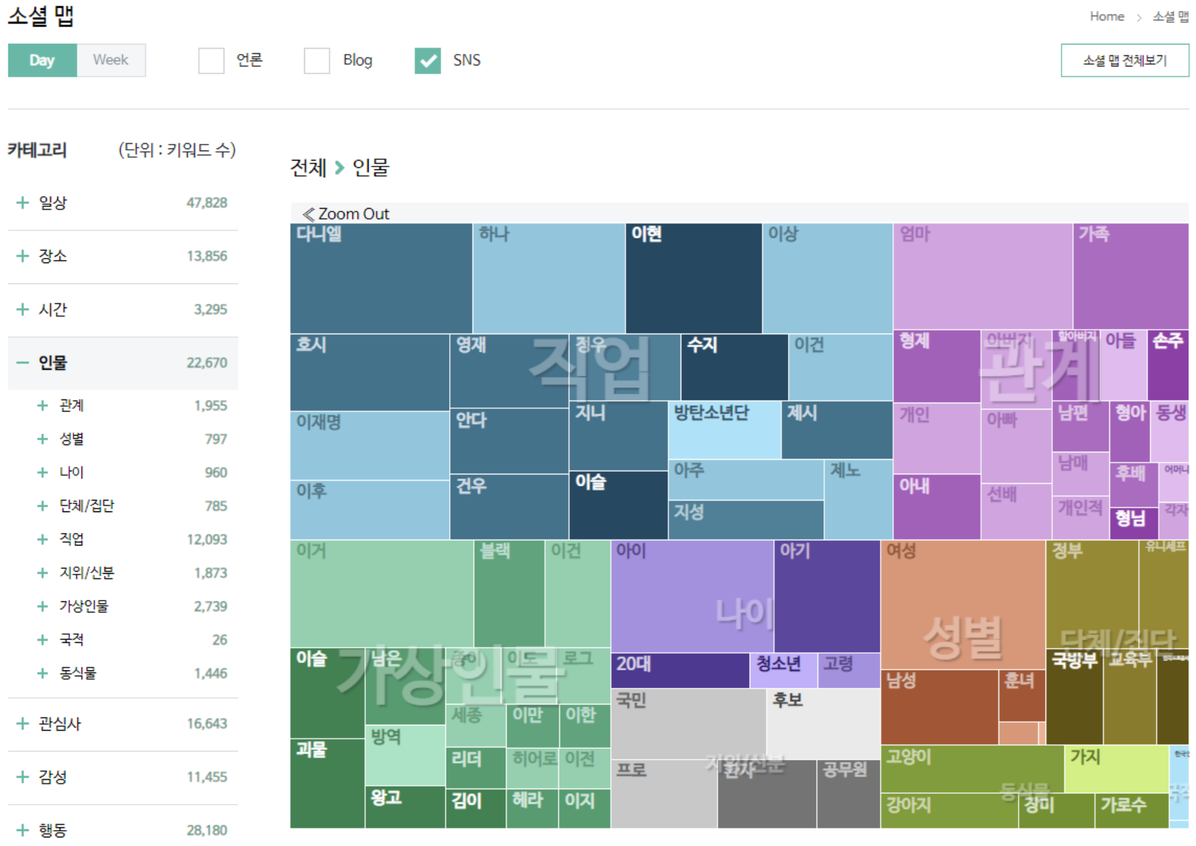
- 언론, Blog와 SNS 같은 소셜 미디어에서 언급되는 주제별 단어들을 라이프 스타일에 맞춰 측정한 데이터를 제공.

- 실시간으로 소셜 미디어 데이터를 수집 및 분석함으로써 마케팅 캠페인과 소비자, 경쟁 업체에 대한 인사이트를 제공하여 스마트한 마케팅 전략을 수립하는데 효과적인 플랫폼

\+ 다양한 시각화 결과 제공하여 활용도가 높다는 점
\- **유료 솔루션으로, 소상공인의 활용이 어렵다는 점**

02 시장조사

오디피아(ODPIA)
https://www.odpia.org



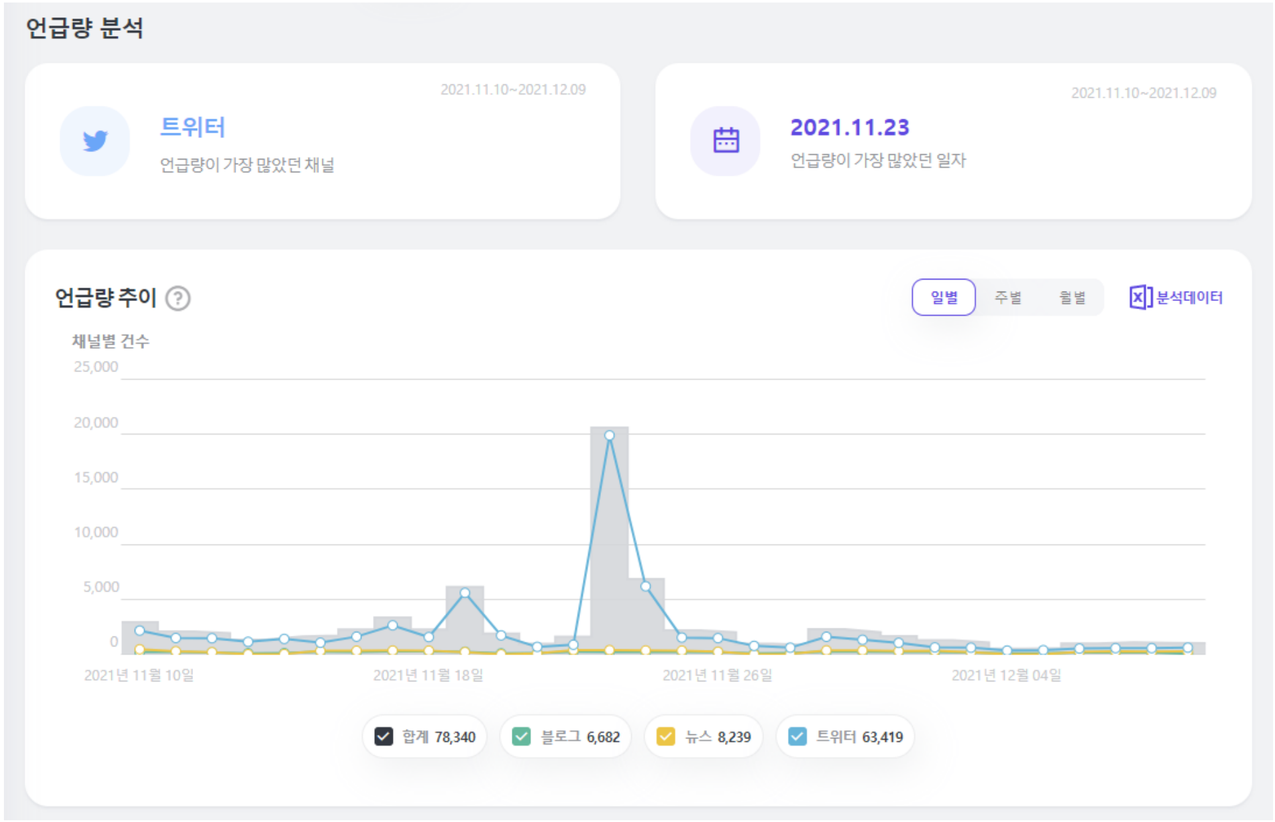
- 국내 주요 포털 사이트, SNS, 커뮤니티, 블로그 등 소셜 채널에서 언급되는 데이터를 실시간으로 분석해 라이프스타일 트렌드와 소셜미디어 상에서의 기업 평판 등 흐름을 한 눈에 볼 수 있다.

- 소셜 인덱스를 분석해 보면 해당 기업 및 산업 분야에 취업 준비를 하거나 기업의 마케팅, 홍보를 하는 경우에 유용하게 활용할 수 있다.

\+ 한 눈에 쉽게 들어오는 화면구조와 쉬운 사용
\- **SNS 키워드를 소셜맵 형태로만 제공하여 다른 도식으로는 볼 수 없음**

03. 시장조사

썸트렌드(Sometrend)
https://some.co.kr/



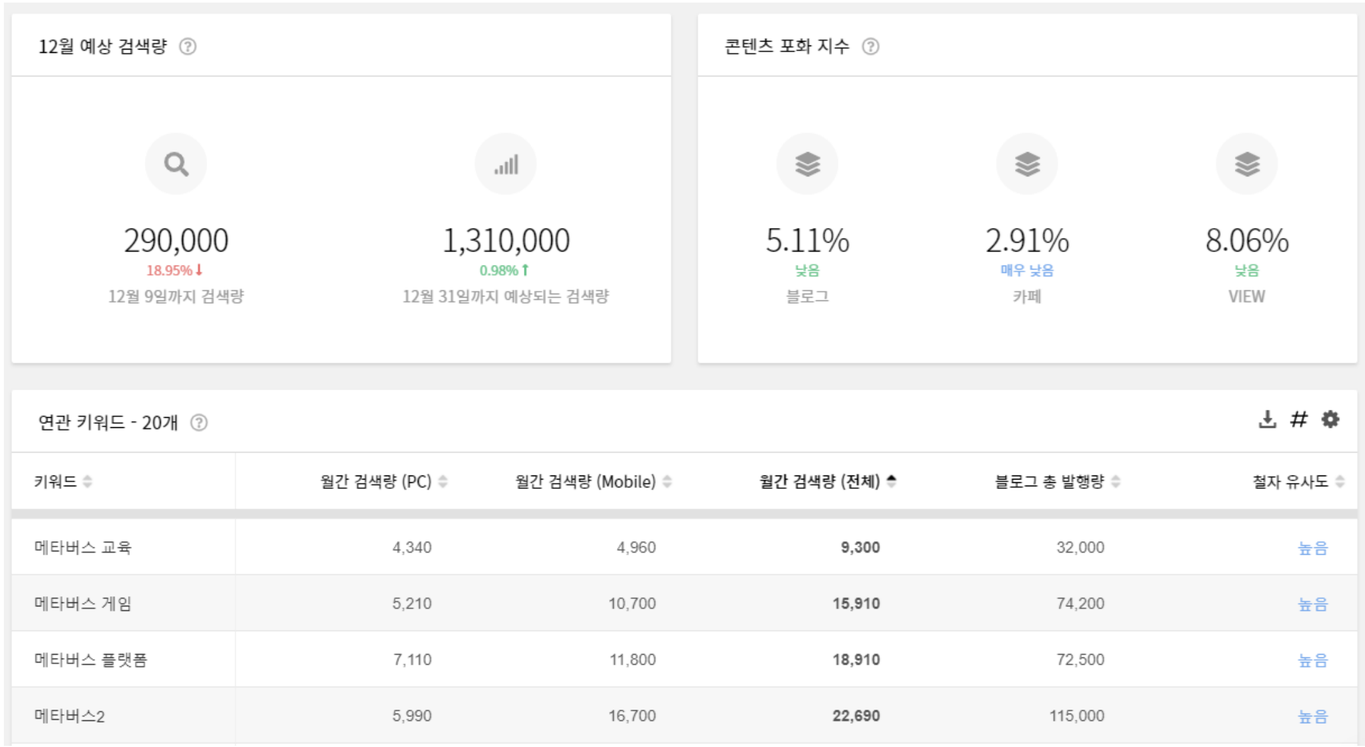
- 키워드별로 소셜에 관련된 정보를 보여주는 서비스. 검색한 키워드에 관련된 트렌드를 한눈에 파악 가능
- 키워드의 다양한 연관어를 알려주며, 해당 데이터로 소비자의 추가 욕구를 확인 가능
- 긍정/부정, 감성, SNS Feed까지 소개

\+ 많이 언급된 토픽 키워드의 언급 원문을 볼 수 있다는 장점

\- **부분 유료서비스이며, 향후 트렌드에 대한 예측보다는 현황파악 위주**

04 시장조사

블랙키위(Blackkiwi)
https://blackkiwi.net/



12월 예상 검색량

290,000
18.95% ↓
12월 9일까지 검색량

1,310,000
0.98% ↑
12월 31일까지 예상되는 검색량

콘텐츠 포화 지수

5.11%
낮음
블로그

2.91%
매우 낮음
카페

8.06%
낮음
VIEW

연관 키워드 - 20개

키워드	월간 검색량 (PC)	월간 검색량 (Mobile)	월간 검색량 (전체)	블로그 총 발행량	철자 유사도
메타버스 교육	4,340	4,960	**9,300**	32,000	높음
메타버스 게임	5,210	10,700	**15,910**	74,200	높음
메타버스 플랫폼	7,110	11,800	**18,910**	72,500	높음
메타버스2	5,990	16,700	**22,690**	115,000	높음

- 빅데이터 기반 키워드 분석 플랫폼으로, 직접 키워드를 입력해 원하는 키워드의 분석결과 열람 가능. 해당 키워드의 정보성과 상업성에 대한 지표 제공

- 검색 동향, 연령별/성별에 따른 검색 비율 통계 확인 가능하며, 연관키워드를 검색량에 따라 제시

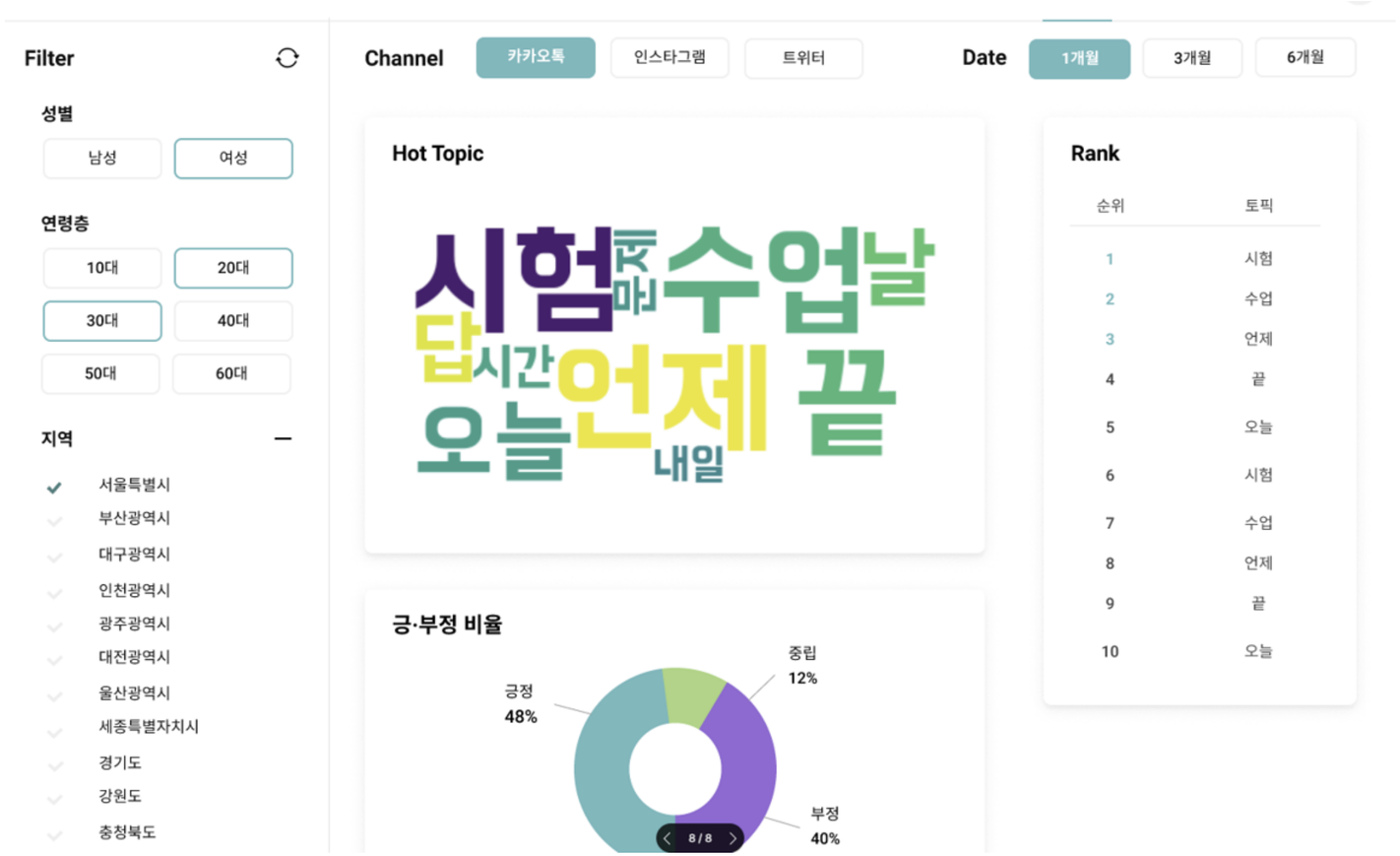
\+ 해당 키워드의 “정보성”과 “상업성”에 대한 지표 제공

\- **개인이 운영하는 곳으로 타사에서 제공하는 서비스보다 데이터의 신뢰성이 다소 떨어짐**

03

비즈니스 모델 “워드온”

01. 서비스 개요



- **Hot Topic**을 추출하여 마케팅 인사이트 제공
- 긍·부정 비율 및 키워드 랭킹 제공
- 성별, 연령층, 지역에 따른 강력한 타겟팅 필터
- 카카오톡, 인스타그램, 트위터 등의 다채로운 **SNS** 채널 필터
- 기간별 랭킹을 통한 다음분기 트렌드 예측

예) 서울에 사는 20대 여성은 이런 "키워드"에 관심이 있다

토픽모델링을 기반으로 한 키워드 분석 마케팅 솔루션 “가제”

02. 비즈니스 모델 캔버스

SNS(카카오톡) 기반 토픽 모델링 Business Canvas

Key Partners

1. 마케팅/광고대행사
2. 쇼핑몰 운영자
3. 브랜드 사업자
4. 투자자 VC
5. 기관(정부, 연구소, 단체)

Key Activities

1. 광고 효율 증대
2. 신규 아이템 발굴
3. 시장 트랜드 분석

Key Resources

1. 데이터 사이언티스트
2. 개발자
3. 마케터

Value Proposition

1. 마케팅 인사이트 제공 (키워드, 시장 동향, 아이템)

2. 광고 최적화 제안 (채널 선정, 노출 방안, ROAS/ROI 등 KPI 개선)

3. 연구과제 수주

Customer Relationships

1. 마케팅 커뮤니티
2. SNS 채널로 홍보
3. 연구과제 수주

Channels

1. 웹(PC)
2. 모바일 앱(마켓)
3. 박람회

Customer Segments

개인 광고주(쇼핑몰)

법인 광고주(중견규모 브랜드 사업자, 온라인 플랫폼)

기관 (정부, 연구소, 단체)

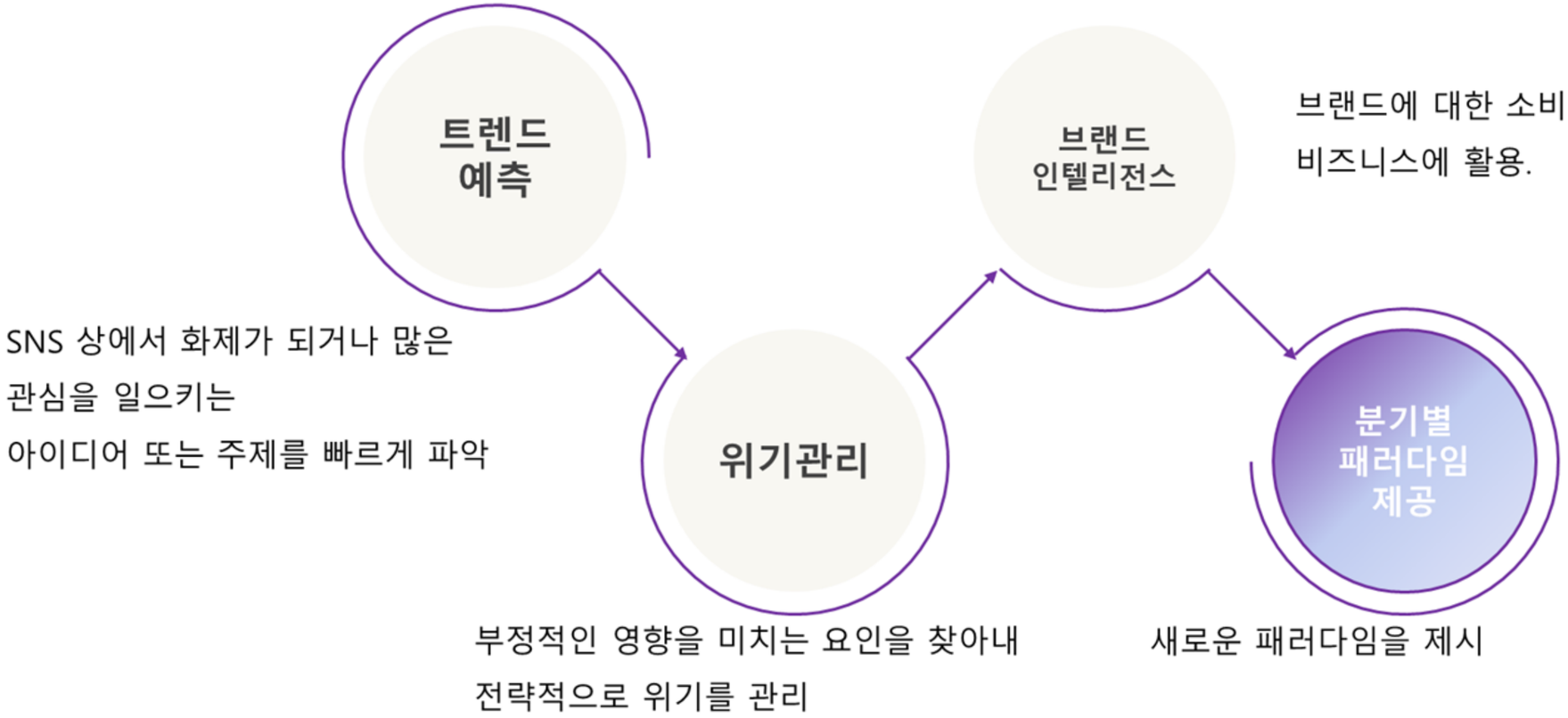
Cost Structure

1. 유지관리비(서버/임대료/부대비용)
2. 인건비
3. 광고비

Revenue Streams

1. 가입비 (데이터 수집/분석)
2. 솔루션 이용료(월간/연간 단위)
3. 연구용역비 (데이터 모델링/분석)
4. 자문료 (마케팅 인사이트)

03
서비스 목적


트렌드
예측
SNS 상에서 화제가 되거나 많은
관심을 일으키는
아이디어 또는 주제를 빠르게 파악
위기관리
부정적인 영향을 미치는 요인을 찾아내
전략적으로 위기를 관리
브랜드
인텔리전스
브랜드에 대한 소비자 인식과 니즈를
비즈니스에 활용.
분기별
패러다임
제공
새로운 패러다임을 제시

04
목표 고객

1) 마케팅을 위한 인사이트를 얻고자 하는 모든 분야의 기업, 개인 고객

2) 복잡하고 고가인 마케팅 솔루션을 이용하기 부담스러운 중소기업이나 소상공인 고객

3) SNS 화제 키워드를 통해 사업아이템을 선별하고자 하는
예비 창업자

04

사업화 전략

BM Strategy

01. 사업화 과정

사업화 일정

02
차별화 전략

1) **한국형 솔루션** : SNS 데이터로 많이 사용되는 전세계 사용자 대상의 인스타그램, 페이스북, 트위터 뿐 만 아니라
한국사람들이 가장 많이 사용하는 APP 카카오톡 데이터를 기반으로 한 한국형 트렌드 예측

2) **월간 구독형 요금제 도입** : 고가의 마케팅 솔루션이 아니더라도,
필요한 시점에만 손쉽게 구독/해지 가능하여 저렴하고 간편하게 사용자 needs 파악 가능

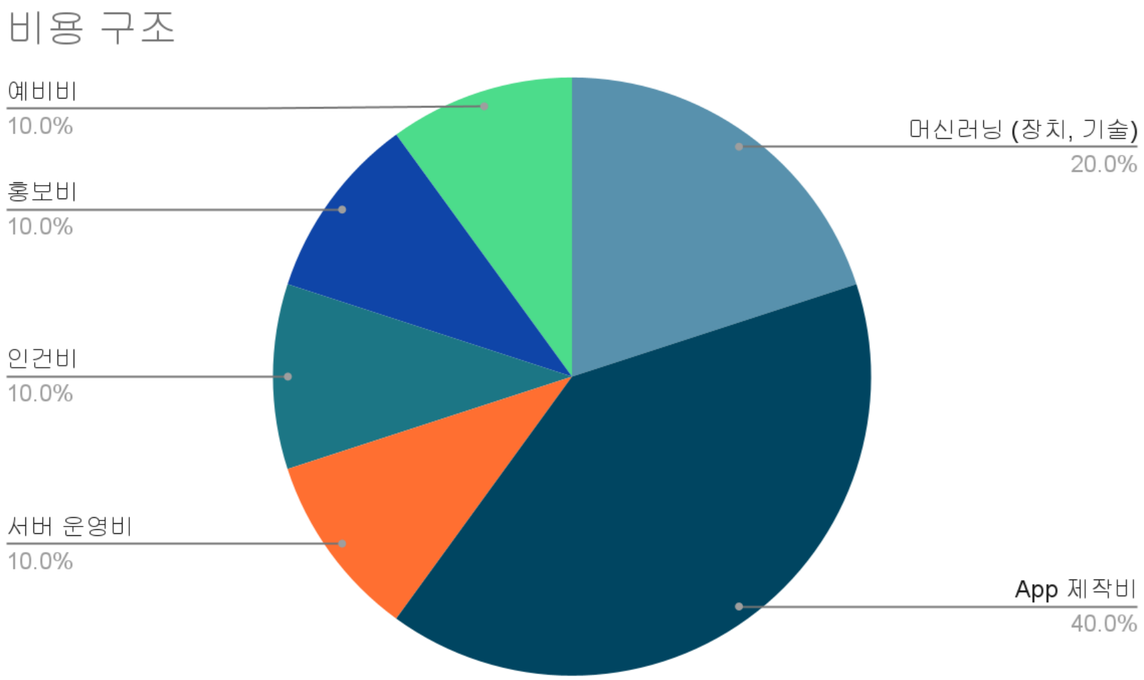
3) **뉴스레터형 트렌드 리포트 제공** : 자주찾는 필터의 트렌드를 설정해 둔 회원은 1주, 한달, 분기별로
이메일로 자동 트렌드 리포트를 받아볼 수 있게 함
예) “서울특별시 지역의 20대 여성” 필터를 저장한 회원에게 해당 필터의 주기적인 리포트를 이메일로 발송. 트렌드 리포트에 대한 접근/열람의 용이성 확보

4) **분석결과 제공** : 신뢰성있는 공공데이터 및 SNS 크롤링을 통해 해당 토픽이 언급된 원본에 사용자가 직접 접근할 수 있도록 제공
예) “개강” 토픽 언급된 트위터 원문 URL 로 이동

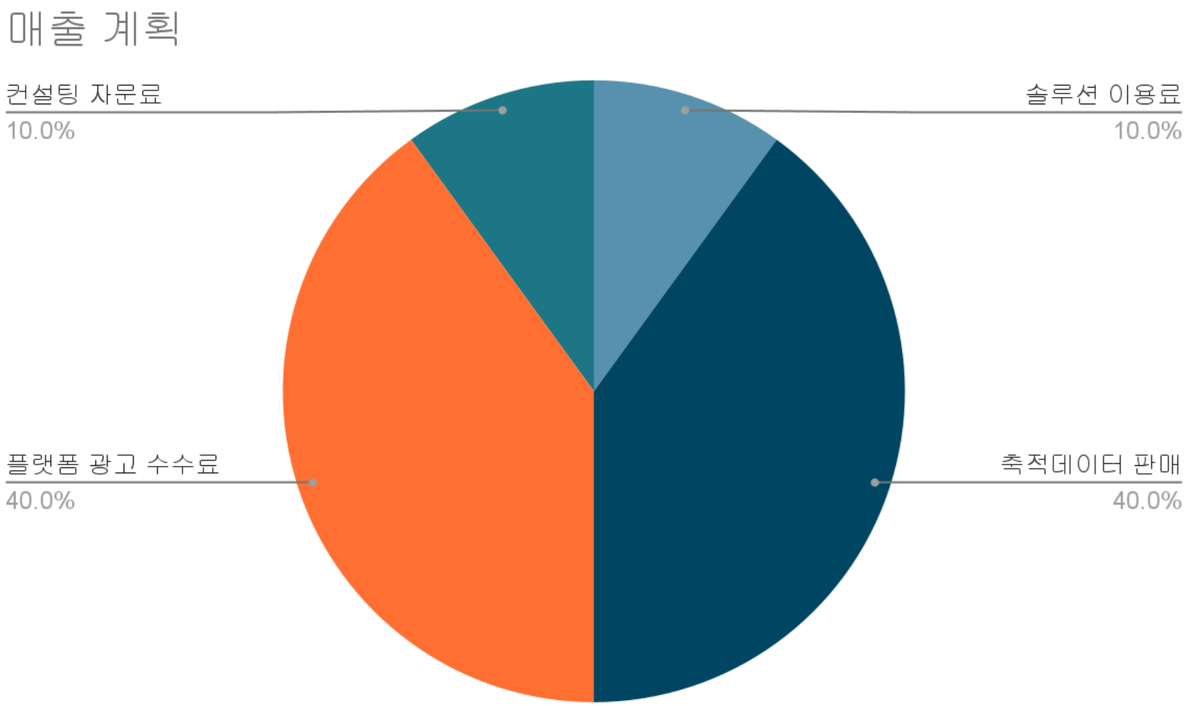
5) **필터 기반의 쉬운 사용성** : 사용하기 복잡하고 무거운 프로그램이 아닌 중소기업, 소상공인도
사용법을 쉽게 익히고 이용할 수 있는 마케팅 솔루션

03
비용 구조 및 매출 계획

비용구조



매출계획

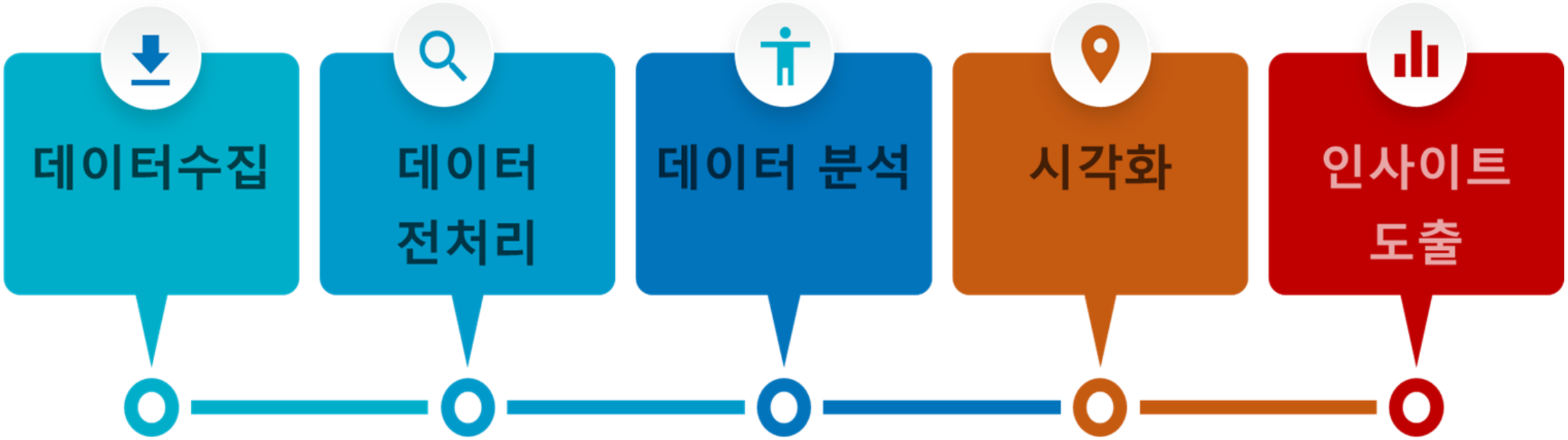


05

데이터 분석

Research Result

01. 분석 과정.



- AI HUB 공개 데이터 중 카카오톡 SNS 데이터셋 활용

- 파라미터 추출
- KoNLPy 활용한 한국어 자연어 처리
- 명사 단위 추출

- LDA 모델 생성 및 학습
- 토픽 일관성 점수 계산
- 적합한 모델 선정
- 토픽 별 출연 확률 확인

- DashBoard 형태의 시각화
- 사용자의 편의성 고려
- 채널,기간,연령대,거주지 별 필터 설정

- B2B 형태의 마케팅 제안

02
사용 데이터

데이터셋: 한국어 **SNS**(카카오톡 대화 내용) **https://aihub.or.kr/aidata/30718**
한국어 구어체 텍스트 기반의 정보검색, 대화분석, 질의응답, 명령어 이해, 언어모델 학습 등의 자연어처리 **AI** 기술 개발을 위한 한국인의 일상대화 **SNS** 데이터 구축

- 4차산업혁명의 핵심 요소인 인공지능의 대표적인 응용 분야 가운데 하나인 대화 처리 기술의 연구 개발에 활용할 수 있는 학습 데이터셋으로 총 200만건의 SNS 대화로 구성.
- 일상적인 대화에 흔히 포함될 수 있는 개인정보에 대하여 비식별화.
- 대화의 유형 : SNS로 이루어지는 가장 일반적인 대화 형태인 일상 대화와 토론 대화, 그리고 질의 응답대화
- 대화의 내용 : 일상 속에서 이루어지는 대화는 다양한 주제를 다루는데, 본 데이터셋은 개인 및 관계, 주거와 생활, 상거래(쇼핑), 식음료, 공공 서비스 등의 주제분류를 적용하였다.


```
{
    “numberOfItems”: 100000,
    “data” : [
        {
            “header”: {
                “dialogueInfo”: {
                “dialogueID”: “D000001”,
                “type”: “일상 대화”,
                “topic”: “행사”,
                “numberOfParticipants”: 3,
                “numberOfTurns”: 2,
                “numberOfUtterances”: 5
                },
                “participantsInfo”: [
                {
                        “participantID”: “P1”,
                        “gender”: “남”,
                        “age”: “50대”,
                        “residentialProvince”: “서울”
                },
                {
                        “participantID”: “P2”,
                        “gender”: “남”,
                        “age”: “20대”,
                        “residentialProvince”: “서울”
                },
                {
                        “participantID”: “P3”,
                        “gender”: “남”,
                        “age”: “20대”,
                        “residentialProvince”: “서울”
                }
                ]
    },
```

03. 파라미터 정보.

변수명	변수유형	설명	사용여부
numberOfParticipants	범주형	화자 인원	X
numberOfUtterances	범주형	대화 횟수	X
numberOfTurns	범주형	화자의 전환 횟수	X
type	범주형	대화내용의 타입	X
topic	범주형	대화내용의 주제	X
dialogueID	문자형	대화 구분 값	X
age	범주형	나이	O
residentialProvince	범주형	거주지 - 서울특별시 ~ 부산광역시 총 17개	O
gender	범주형	성별 - 남성 or 여성	O
participantID	범주형	화자의 구분 값 - P01,P02~	O
utterance	문자형	대화내용	O
utteranceID	범주형	화자 ID	X
date	일시형	발언날짜 ex. 2021-01-01	O
turnID	범주형	화자 변경 횟수	X
time	일시형	발언시간	X

04 데이터 전처리

Raw Data


```
"data":[
 {
  "header":{
   "dialogueInfo":{
    "numberOfParticipants":2,
    "numberOfUtterances":18,
    "numberOfTurns":5,
    "type":"일상 대화",
    "topic":"시사₩/교육",
    "dialogueID":"30fab051-9111-5972-add3-21a9f9bb90c8"
   },
   "participantsInfo":[
    {
     "age":"20대",
     "residentialProvince":"부산광역시",
     "gender":"여성",
     "participantID":"P01"
    },
    {
     "age":"20대",
     "residentialProvince":"인천광역시",
     "gender":"여성",
     "participantID":"P02"
    }
   ]
  },
  "body":[
   {
    "utterance":"한국에사 드꺼운 마스크 두개 사왔는데",
    "utteranceID":"U1",
    "participantID":"P01",
    "date":"2020-01-24",
    "turnID":"T1",
    "time":"21:24:00"
   },
   {
    "utterance":"기침이랑 그런걸로",
    "utteranceID":"U2",
    "participantID":"P02",
    "date":"2020-01-24",
    "turnID":"T2",
    "time":"21:24:00"
   },
```


Preprocessing

	Unnamed: 0	participantID	age	gender	residentialProvince	utterance	date
0	0	P01	20대	여성	부산광역시	한국에사 드꺼운 마스크 두개 사왔는데	2020-01-24
1	1	P02	20대	여성	인천광역시	기침이랑 그런걸로	2020-01-24
2	2	P02	20대	여성	인천광역시	옮는데	2020-01-24
3	3	P02	20대	여성	인천광역시	공기중 침	2020-01-24
4	4	P01	20대	여성	부산광역시	들어왔을땜 일케 심각 안해서	2020-01-24
...	...	...	...	...	...	...	...
1116382	1116382	P01	20대	여성	서울특별시	ㅇㅇ 뭐 영어도아니고,, 폴란드,,?	2020-03-18
1116383	1116383	P01	20대	여성	서울특별시	아니 기억안나 아무튼 외국어래	2020-03-18
1116384	1116384	P02	20대	여성	서울특별시	당연히 한자나 한글인줄 알았는데	2020-03-18
1116385	1116385	P02	20대	여성	서울특별시	외래어 였군요^^	2020-03-18
1116386	1116386	P01	20대	여성	서울특별시	네 상식 하나 알려드렸습니다 ㅎㅎ	2020-03-18

1116387 rows × 7 columns

05.
분석 알고리즘 .

토픽 모델링 - LDA (Latent Dirichelt Allocation, 잠재 디리클레 할당)
- 토픽 모델링은 문서의 집합에서 토픽을 찾아내는 프로세스.
- 토픽들은 확률 분포에 기반하여 단어들을 생성한다고 가정하고, 데이터가 주어지면, LDA는 문서가 생성되던 과정을 역추적한다.

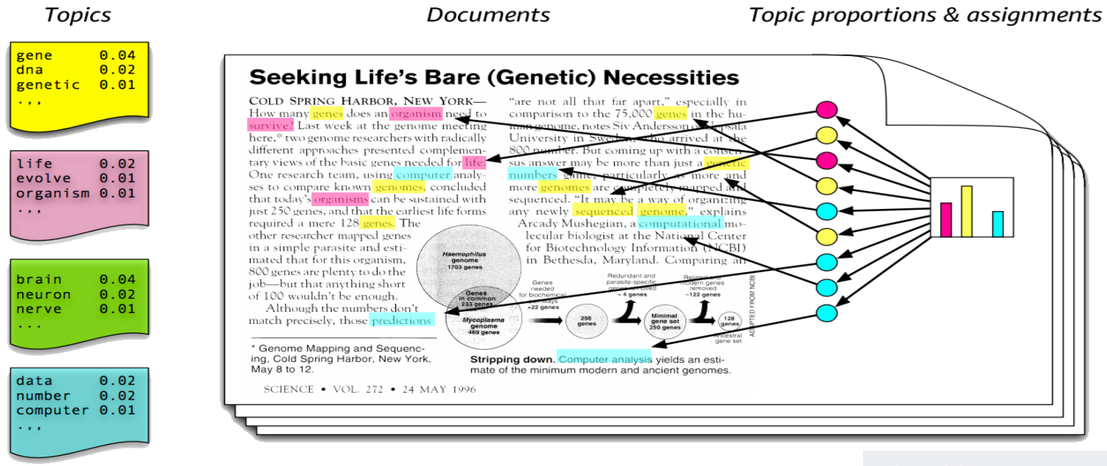
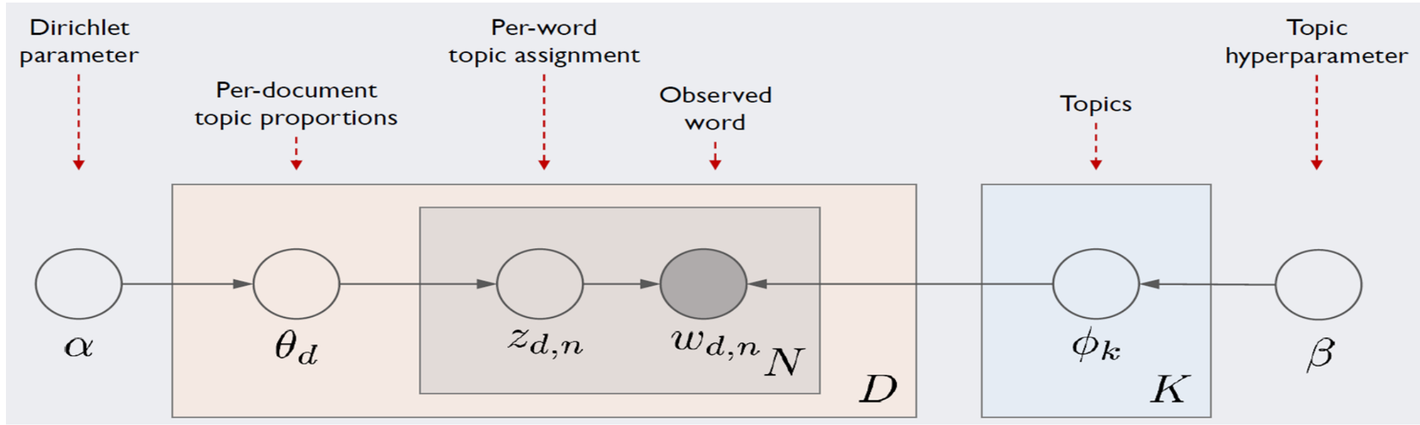


그림 1. LDA 도식

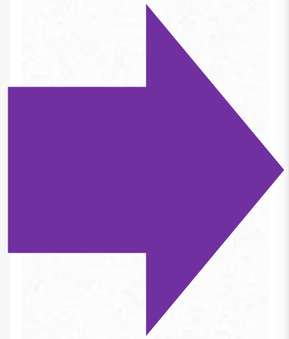
그림 2. LDA Model Architecture



출처 : '단문의 주제 분석을 위한 LDA 와 BTM 토픽모형 평가 - 문길성'

06
자연어 처리

```
[6]  def getList(key, value) :
       return result.loc[result[key] == value]

     # 연령대 별 결과물 추출
     # list = getList('age', '20대')

     # 성별 별 결과물 추출
     # list = getList('gender', '남성')

     # 지역 별 결과물 추출
     # list = getList('residentialProvince', '부산광역시')

     # 복합 조건
     list = result.loc[(result['age'] == '20대') & (result['gender'] == '여성')]

     from konlpy.tag import Okt
     okt = Okt()
     words = []
     for i in range(len(list)) :
       ut = list['utterance'].iloc[i]
       n = okt.nouns(ut) # 명사 단위 추출

       # print(ut, '======>',n)
       words.append(n)

       # 빈 아이템 삭제 필요
```



```
words
```

```
[['한국', '사', '드꺼운', '마스크', '개'],
 ['땜', '일케', '심각', '안해'],
 ['당일', '당일', '폐쇠'],
 ['어젠', '엇그제', '우한', '도시', '폐쇠'],
 ['어제', '개', '도시', '확대', '폐쇠'],
 ['오늘', '상해', '관광지', '폐쇠'],
 ['긴급'],
 ['상해', '관광', '어늘', '긴급'],
 ['발표', '된거'],
 [],
 ['기차', '지하철'],
 ['뱅기'],
 ['기침', '걸'],
 ['옮는데'],
 ['공기', '침'],
 ['대박'],
 ['아예', '안밖못'],
 ['미가', '능', '사람', '어케'],
 ['또', '한국', '여행', '사람'],
 ['한국', '사람', '곳', '또', '또'],
 ['만', '한국', '보고'],
 ['여'],
 [],
 ['어제', '서비스', '차원'],
```

07. LDA Topic 모델링

```
limit=21; start=4; step=2;
x = range(start, limit, step)
topic_num = 0
count = 0
max_coherence = 0
for m, cv in zip(x, coherence_values):
    print("Num Topics =", m, " has Coherence Value of", cv)
    coherence = cv
    if coherence >= max_coherence:
        max_coherence = coherence
        topic_num = m
        model_list_num = count
    count = count+1


# Select the model and print the topics
optimal_model = model_list[model_list_num]
model_topics = optimal_model.show_topics(formatted=False)
#print(optimal_model.print_topics(num_words=10))
```

```
Num Topics = 4  has Coherence Value of 0.7923684456397492
Num Topics = 6  has Coherence Value of 0.7544151999737972
Num Topics = 8  has Coherence Value of 0.744853961049716
Num Topics = 10  has Coherence Value of 0.7344126922919546
Num Topics = 12  has Coherence Value of 0.7204341666654631
Num Topics = 14  has Coherence Value of 0.7107238907065149
Num Topics = 16  has Coherence Value of 0.7021090388966431
Num Topics = 18  has Coherence Value of 0.6883052273973249
Num Topics = 20  has Coherence Value of 0.6941387030873936
```

그림 3. 각 모델의 빈도수 파악

```
def format_topics_sentences(ldamodel=optimal_model, corpus=corpus, texts=texts):
    # Init output
    sent_topics_df = pd.DataFrame()

    # Get main topic in each document
    #ldamodel[corpus]: lda_model에 corpus를 넣어 각 토픽 당 확률을 알 수 있음
    for i, row in enumerate(ldamodel[corpus]):
        row = sorted(row, key=lambda x: (x[1]), reverse=True)
        # Get the Dominant topic, Perc Contribution and Keywords for each document
        for j, (topic_num, prop_topic) in enumerate(row):
            if j == 0:  # => dominant topic
                wp = ldamodel.show_topic(topic_num,topn=10)
                topic_keywords = ", ".join([word for word, prop in wp])
                sent_topics_df = sent_topics_df.append(pd.Series([int(topic_num),
            else:
                break
    sent_topics_df.columns = ['Dominant_Topic', 'Perc_Contribution', 'Topic_Keywor
    sent_topics_df = pd.concat([sent_topics_df], axis=1)
    return(sent_topics_df)

df_topic_sents_keywords = format_topics_sentences(ldamodel=optimal_model, corpus=c

# Format
df_topic_tweet = df_topic_sents_keywords.reset_index()
# df_topic_tweet.columns = ["participantID", "age", "gender", "residentialProvince
df_topic_tweet
```

```
[17] lda_inform.to_csv("lda_inform.csv", index = None)
lda_inform
```

	Dominant_Topic	Topic_Perc_Contrib	Keywords	Topic_Counts	Topic_Contribution
0.0	0.0	0.3444	사람, 명, 코로나, 거기, 그, 확, 진자, 때문, 난리, 번	163174	0.2354
1.0	1.0	0.2173	시험, 이번, 데, 학기, 언제, 주, 전, 날, 저번, 합격	36849	0.0532
2.0	2.0	0.2986	거, 문제, 개, 책, 저, 과목, 그냥, 인강, 보고, 번	30908	0.0446
3.0	3.0	0.1938	말, 거, 뭐, 알, 자기, 그게, 무슨, 줄, 걸, 보	30714	0.0443
4.0	4.0	0.1949	시스템, 사진, 응, 기타, 그거, 웅웅, 마자, 강, 오, 검색	35815	0.0517
5.0	5.0	0.1855	나, 오늘, 내일, 다시, 낼, 구, 바로, 차, 기사, 걱정	30640	0.0442
6.0	6.0	0.2693	햇, 내, 니, 친구, 도, 나, 자, 얘기,	29108	0.0420

그림 4. 특정 토픽 당 확률

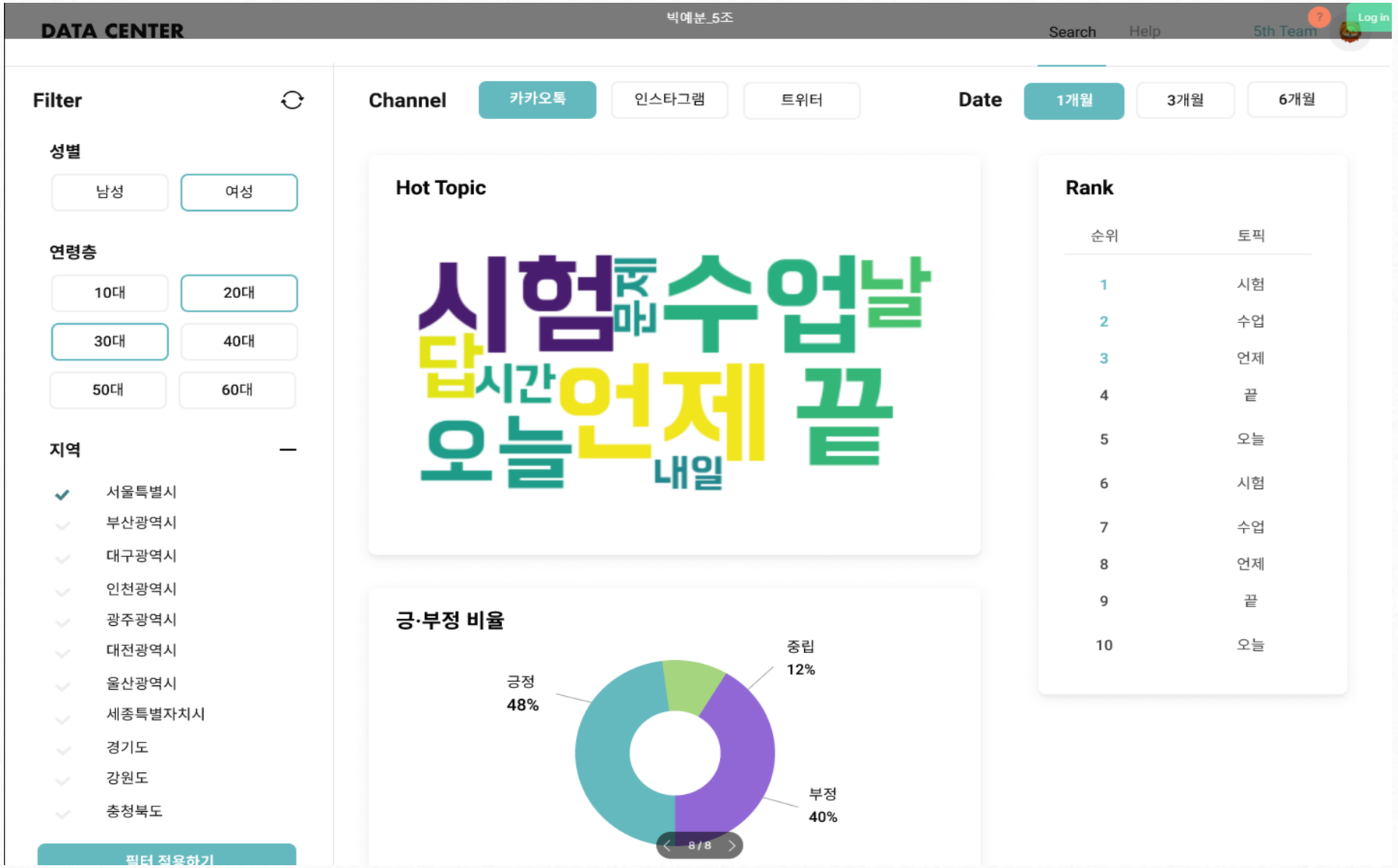
06

솔루션 소개 & Demo

Research Result & Demo

01. Visualization.

그림 5. ProtoType – Dash Board



QnA

질의 응답

- 자유롭게 질문 부탁드립니다.

THANK YOU.

감사합니다.