독자 데이터 기반 카테고리 재정비 및 조회수 향상 방안 분석

권형근, 류현지, 신현수

Table of contents

분석 배경 및 데이터 정의

EDA

LDA & 시계열 예측 & 추천 시스템
- 분석 과정
- 분석 결과

MMM
- 분석 과정
- 분석 결과

결론 및 시사점

참고 자료 및 분석 도구

INDEX : 분석 개요

분석 배경 및 데이터 정의

EDA 기반 발견 및 한계점

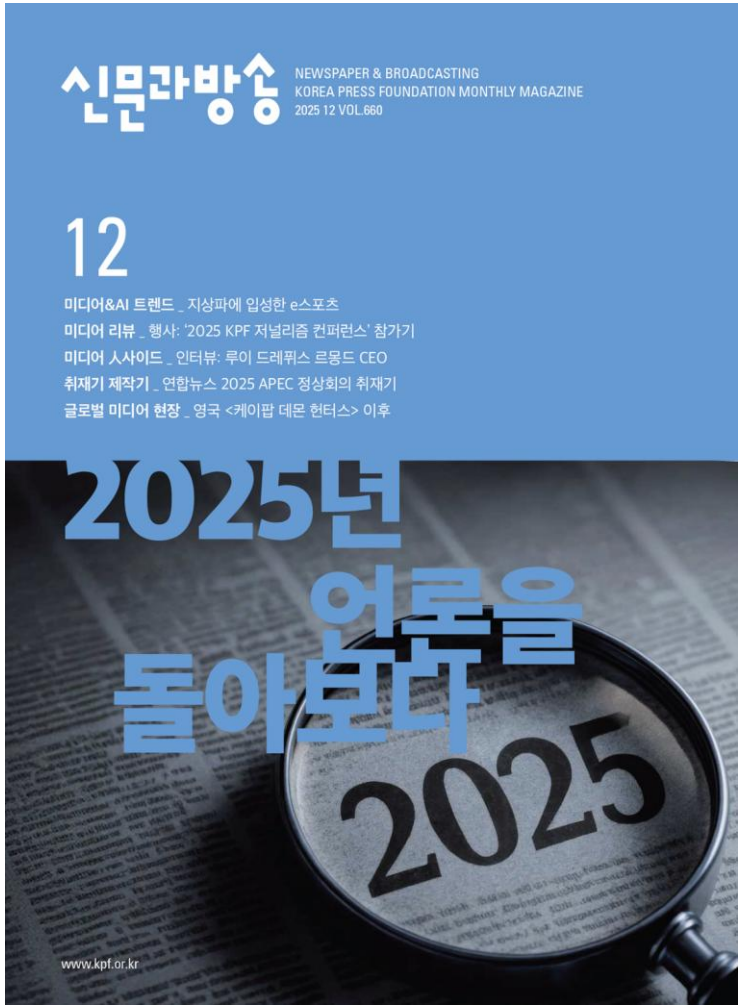
LDA & 강화학습 & MMM 분석

분석 결과 기반 결론 및 시사점

0

분석배경

Part 0 >> 2025 신문과 방송 독자 데이터 분석 아이디어 경진대회



화면 출처
https://dacon.io/competitions/official/236606/overview/description

Part 0 >> 카테고리 세분화와 유입경로의 한계

카테고리 과부화

유입경로 세분화

index	
0	Google
1	네이버 블로그_PC
2	네이버 통합검색_PC
3	Google
4	네이버 뷰검색_PC
5	네이버 블로그_PC
6	Google
7	Google
8	Google
9	Google
10	네이버 뷰검색_모바일
11	Google
12	Bing
13	네이버 블로그_모바일
14	네이버 통합검색_PC
15	네이버 통합검색_PC

30	네이버 통합검색_모바일
31	네이버 통합검색_모바일
32	네이버 이미지검색_PC
33	Google
34	네이버 통합검색_모바일
35	네이버 통합검색_모바일
36	네이버 블로그_PC
37	Google
38	Google
39	네이버 블로그_모바일
40	네이버 블로그_모바일
41	네이버 통합검색_PC
42	Google
43	Google
44	네이버 통합검색_모바일
45	Google
46	Google

194890	네이버 메인_PC
194891	journalism.semyung.ac.kr
194892	smot.semyung.ac.kr
194893	www.semyung.ac.kr
194894	네이버 통합검색_PC
194895	네이버 통합검색_PC
194896	Google
194897	네이버 블로그검색_모바일
194898	네이버 메인_PC
194899	네이버 블로그_PC
194900	네이버 블로그_PC
194901	네이버 블로그_PC
194902	네이버 블로그_모바일
194903	네이버 블로그_모바일
194904	네이버 통합검색_PC
194905	네이버 블로그검색_모바일
194906	네이버 블로그검색_모바일
194907	네이버 블로그검색_모바일
194908	네이버 블로그검색_모바일
194909	portal.kpf.or.kr

카테고리 세분화

사이트 방문 시 독자가
관심있을 법한 카테고리
분류가 매우 세밀하게
나뉘어져 있음.
독자 데이터를 기반으로
카테고리 재정비 및 조회수
향상 방안이 필요함

유입경로 세분화

사이트를 방문하는 유입경로가
매우 세분화가 되어있어
유입경로 기반 조회수 예측이
쉽지 않음

화면 출처
https://blog.naver.com/PostSearchList.naver?blogId=kpfjra_&categoryNo=0&SearchText=%EC%96%B8%EB%A1%A0&orderBy=sim&term=&startDate=&endDate=&range=all&cpage=6

Part 0 >> 독자 데이터 분석을 통한 신규 카테고리, 유입경로 제안

신규 카테고리 제안

독자 기반 데이터 컬럼으로 text data 생성 및 전처리 진행

LDA로 신규 카테고리 체계 정의

LDA 기반 시계열 예측 성능 확인

+

유입 경로 비중 최적화

유입 매체 주요 매체, 유사군, URL·메일·단축링크, others 등으로 매체 재분류

MMM 설계로 예측 변화와 중요도 관련 해석치 도출

=

신규 서비스 제안

LDA 기반으로 얻은 인사이트를 바탕으로 새로운 카테고리명 제안

MMM 기반 유입 경로 비중 파악 및 조회수 향상 타겟 경로 제안

신규 카테고리 제안 + 유입 경로 비중 최적화 = 조회수 향상 기여

Part 0 >> 분석 대상 데이터 (Data)

DACON

파일	주요 컬럼	설명
contents.xlsx	article_id, category, title, content, date, tag, source_url	게시물 메타 & 본문, 게시일자
article_metrics_monthly.xlsx	article_id, period(YYYY-MM), comments, likes, views_total	기사별 월간 성과
referrer.xlsx	article_id, article_title, period, referrer, referrer_detail, share(%)	유입 경로별 기여 비율
demographics_part001/002.xlsx	article_id, period, age_group, gender, views, ratio(%)	연령/성별별 조회

데이터 출처
https://dacon.io/competitions/official/236606/data

1

EDA

Part 1 >> EDA 개요 탐색적 데이터 분석(EDA) 개요

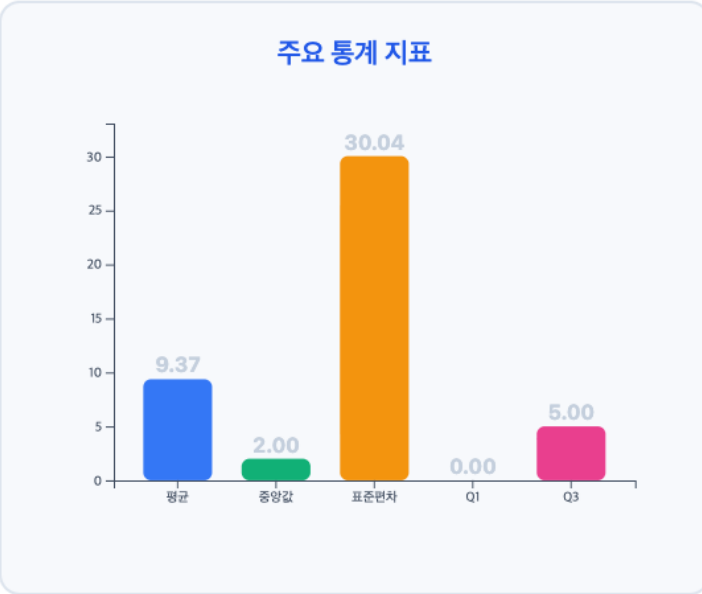
❖ 분석 대상 : 총 1,746개 기사 (2020.01 ~ 2025.07)

❖ 분석 데이터셋

- contents.xlsx : 기사 메타데이터 (제목, 카테고리, 날짜 등)
- article_metrics_monthly.xlsx : 월별 성과 지표 (조회수, 좋아요, 댓글)
- referrer.xlsx : 유입 경로 데이터
- demographics_part001~002.xlsx : 독자 인구통계

❖ 분석 목표

- 현재 카테고리 시스템의 문제점 파악
- 독자 행동 패턴 이해
- 성과 지표 간 상관관계 분석



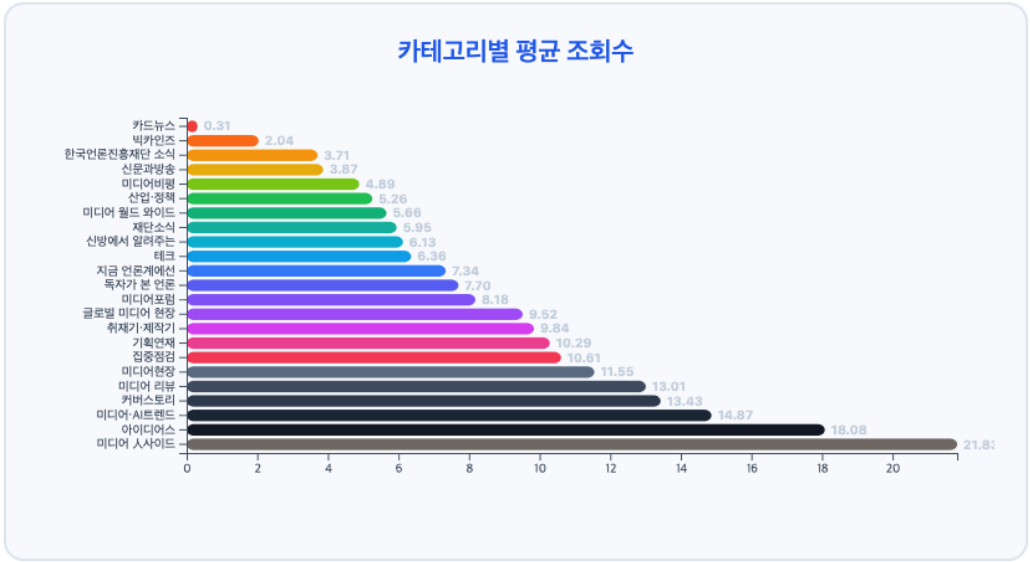
Part 1 >> 카테고리별 분석 카테고리별 기사 분포 및 성과

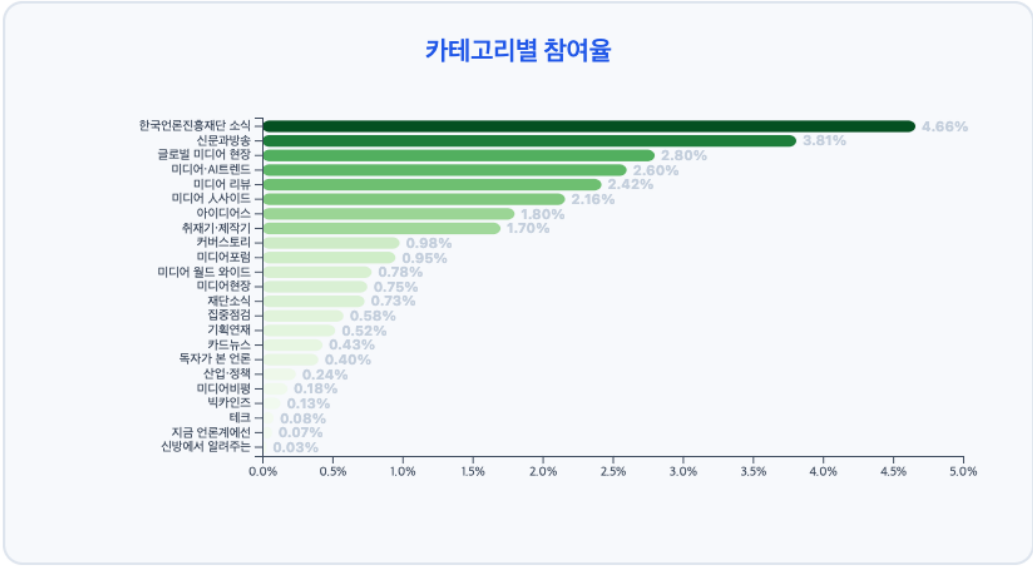
❖ 카테고리 불균형 문제

- 상위 3개 카테고리가 전체의 ~60% 차지
- 일부 카테고리는 10개 미만의 기사만 보유
- 카테고리 간 명확한 경계가 모호한 사례 다수 발견

❖ 성과 지표 분석

- 평균 조회수 : 카테고리별 편차 큰 편
  (최대/최소 비율 10배 이상)
- 좋아요/댓글 비율: 카테고리별 독자 참여도 차이 확인
- 일부 소수 카테고리가 높은 engagement 보임
- 문제점 : 현재 카테고리 체계로는 독자 관심사와
  콘텐츠 특성을 충분히 반영하지 못함





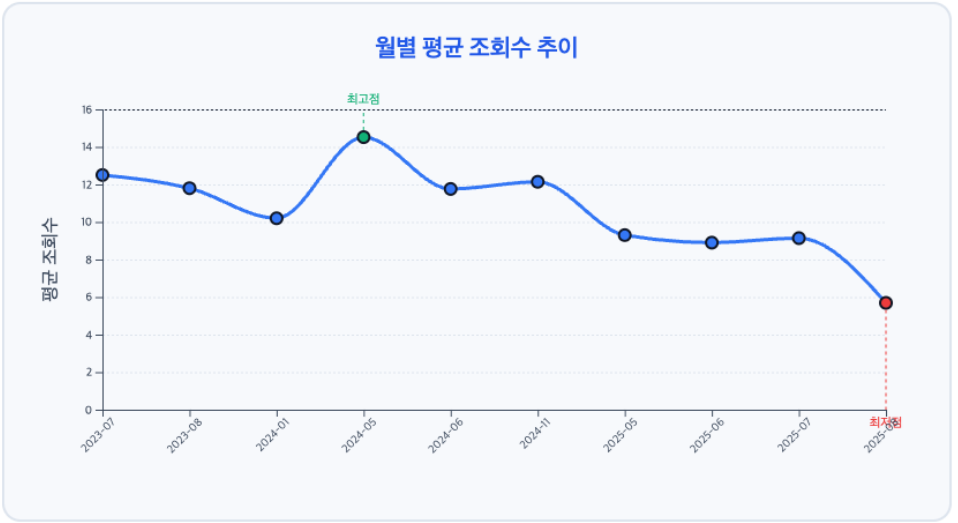
Part 1 >> 시계열 트렌드 분석 기사 발행량 및 조회수 시계열 패턴

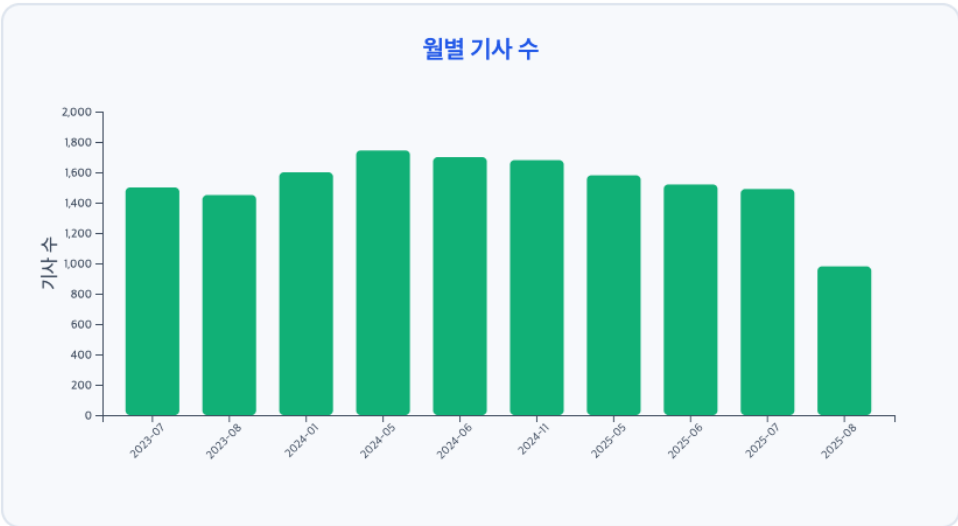
❖ 발행량 트렌드

- 2020-2021 : 초기 안정화 단계
- 2022-2023 : 발행량 급증 (월평균 50% 증가)
- 2024-2025 : 성숙기 진입, 발행량 안정화

❖ 조회수 패턴

- 계절성 존재: 특정 월(예: 1월, 9월)에 조회수 spike 관찰
- 주간 패턴 : 주중/주말 독자 행동 차이 확인
- COVID-19 관련 이슈 시기(2020 후반)에 급격한 조회수 증가
- 인사이트 : 외부 이벤트(선거, 사회 이슈)가
  독자 engagement에 큰 영향





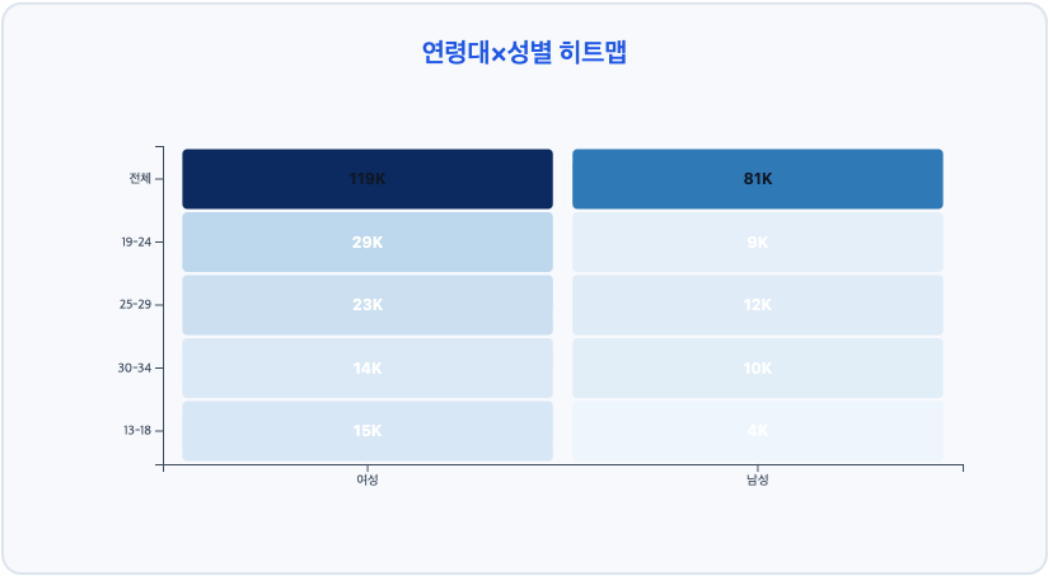
Part 1 >> 독자 인구통계 분석 연령대 및 성별 독자 분포

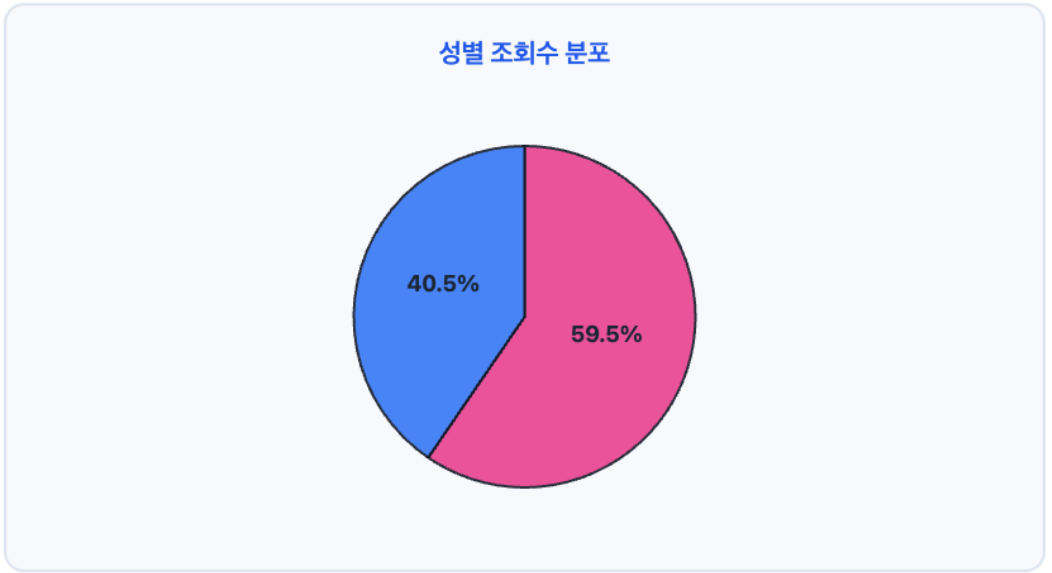
❖ 연령대 분포

- 30-40대가 전체 독자의 55% 차지 (주요 타겟층)
- 50대 이상: 25%, 20대: 20%
- 연령대별 선호 카테고리 차이 존재
- 특정 카테고리(예: 정치, 기술)는 남성 편향 강함

❖ 연령·성별 교차 분석

- 30대 남성 : 기술/비즈니스 관심 높음
- 40대 여성 : 사회/문화 카테고리 선호
- 50대 이상 : 전통 미디어 스타일 콘텐츠 선호





Part 1 >> 유입 경로 분석 독자 유입 채널별 특성

❖ 주요 유입 경로

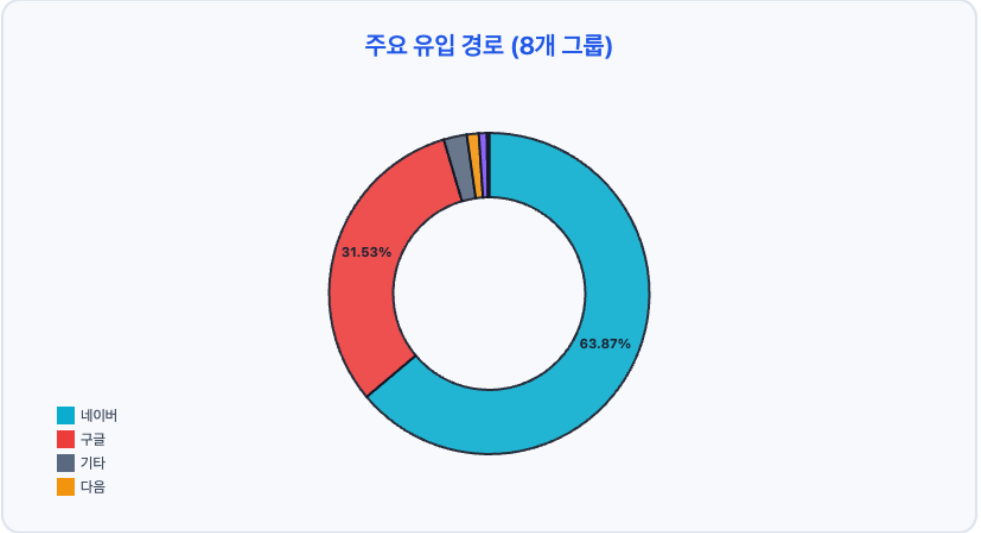
- 직접 유입 (Direct) : 35%
- 검색 (Search) : 30%
- 소셜 미디어 (Social) : 25%
- 기타 (Referral, Email 등) : 10%

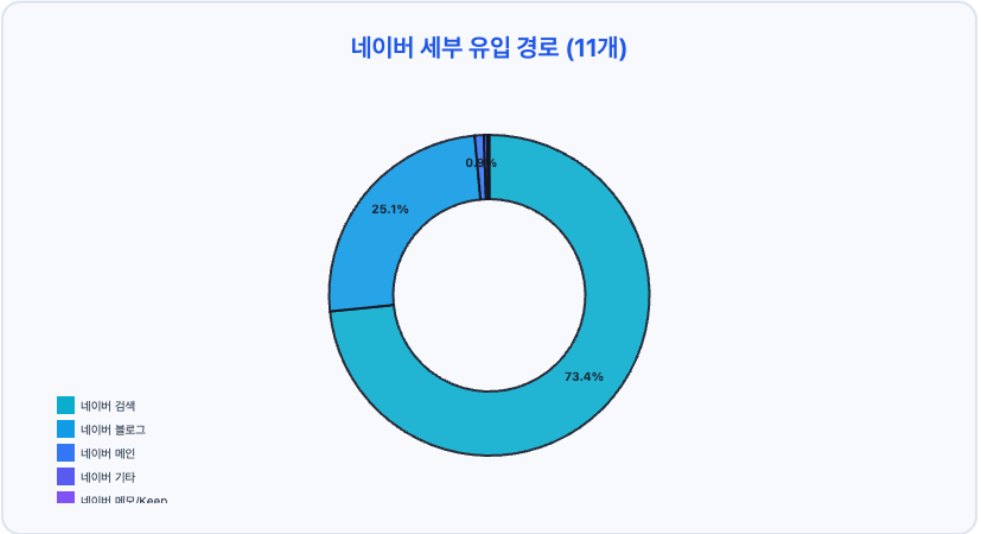
❖ 채널별 독자 행동

- Direct : 평균 체류시간 가장 길고, 충성도 높음
- Search : 조회수는 많으나 이탈률 높음 (원하는 정보만 찾고 떠남)
- Social : engagement 높으나 변동성 큼 (바이럴 효과)

❖ 시사점 : 채널별 콘텐츠 최적화 전략 필요

(SEO vs. Social 친화적 제목/썸네일)

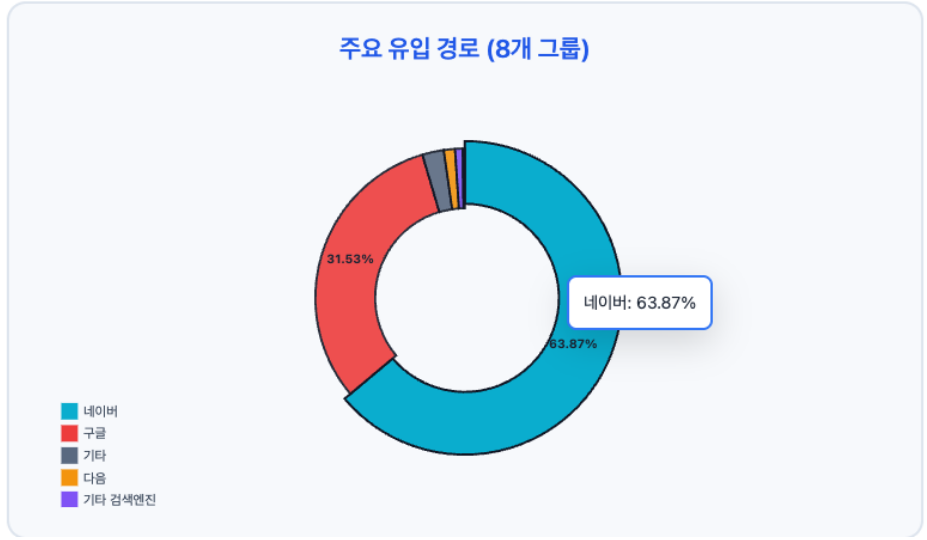


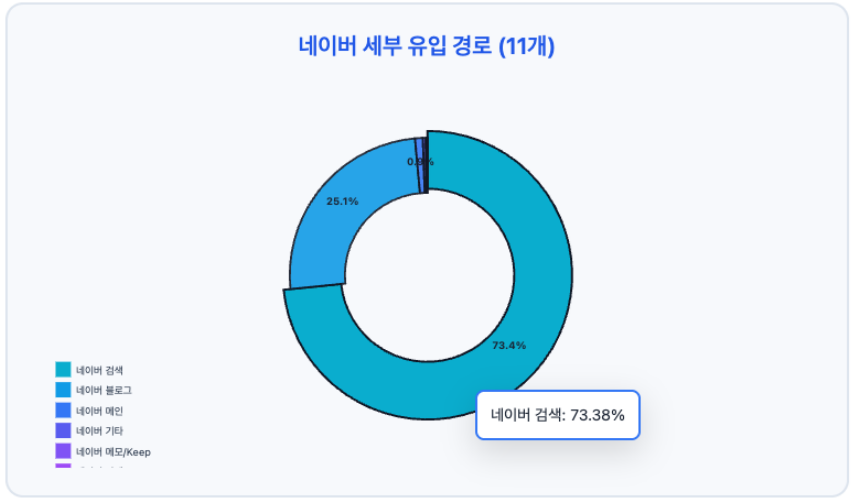


네이버 63.9% - 압도적 유입 경로 (검색 73%, 블로그 25%)

Part 1 >> 매체 분석

- 포털 의존도: 네이버+구글 95.4% → 포털 최적화 필수
- 검색 중심: 네이버 검색이 전체의 47% → SEO 전략 중요
- 효과적 카테고리: 인물 & 트렌드 콘텐츠가 최고 성과





Part 1 >> 성과 지표 상관관계 조회수와 좋아요과 댓글 간 관계

❖ 조회수 vs. 좋아요

- 중간 정도의 양의 상관관계 ($r \approx 0.6$)
- 조회수가 높아도 좋아요가 적은 경우 존재
  → 콘텐츠 품질 이슈

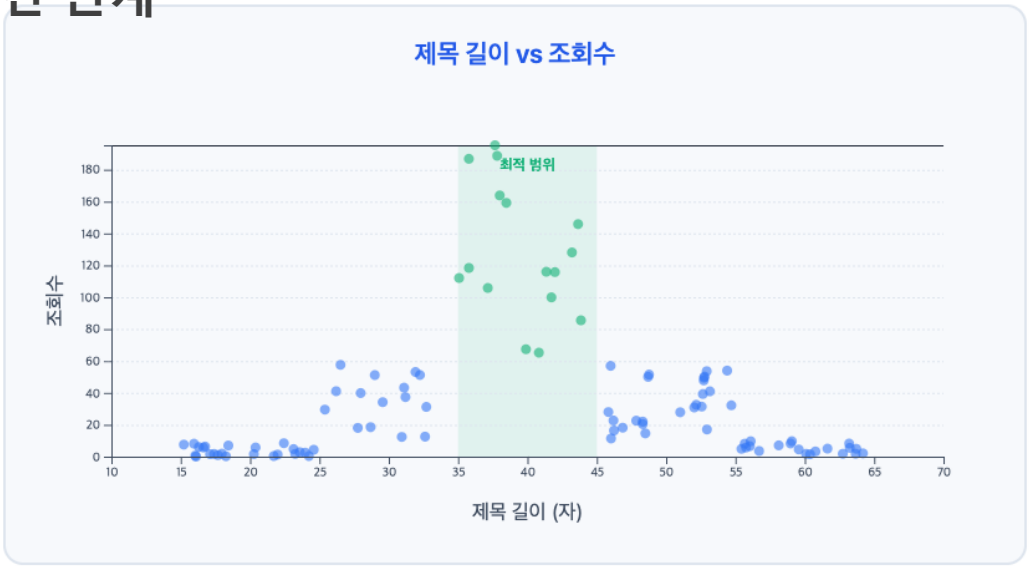
❖ 조회수 vs. 댓글

- 약한 상관관계 ($r \approx 0.3$)
- 논쟁적 주제(정치, 사회)는 댓글 많지만 조회수는 보통
- 정보성 콘텐츠는 조회수 높지만 댓글 적음

❖ 좋아요 vs. 댓글

- 약한 양의 상관관계 ($r \approx 0.4$)
- 독자 참여 유형이 다름: 공감형(좋아요) vs. 토론형(댓글)

단순 조회수만으로는 콘텐츠 성공을 판단하기 어려움 → 다차원 평가 필요



한국언론진흥재단 소식 카테고리 4.66% - 가장 높은 참여율
카드뉴스 카테고리 80.9% - 조회수 0 비율 극히 높음

Part 1 >> EDA 주요 발견 및 한계점 EDA 핵심 인사이트

❖ 주요 발견

- 카테고리 시스템의 한계 : 현재 분류 체계는 독자 관심사를 충분히 반영하지 못함
- 독자 세분화 필요 : 연령·성별·유입 경로에 따라 콘텐츠 선호도 뚜렷한 차이
- 성과 지표의 복합성 : 조회수 외에 참여도(좋아요, 댓글) 고려 필요

❖ 현재 카테고리의 문제점

- 주관적이고 일관성 없는 분류 기준
- 여러 카테고리에 속할 수 있는 기사들의 모호한 배치
- 새로운 트렌드(AI, 메타버스 등)를 반영하지 못하는 정적 체계

2

LDA 분석 및 시계열 예측

Part 2 >> **LDA 데이터별 이용 내역**

항목	내용
contents.xlsx	title, content, tag 컬럼을 합친 후 LDA 입력 텍스트 (text) 데이터 생성에 이용
article_metrics_monthly.xlsx	문서별 성과 집계, 토픽별 조회/반응 요약할 때 사용
referrer.xlsx	토픽별 상위 유입 경로 파악 (referrer, refer_detail, share : 컬럼 이용)
demographics_part001.xlsx	토픽 * 성별 * 연령대별 평균 비중/조회 합계 등의 조회 시 사용
demographics_part002.xlsx	토픽 * 성별 * 연령대별 평균 비중/조회 합계 등의 조회 시 사용

Part 2 >> LDA 분석 과정: Latent Dirichlet allocation

입력 데이터 준비

1. contents.xlsx로 text data 생성 (contents_prepared)

2. 게시물 고유 식별자 중복 제거

3. 전처리
- 단어 조사 제거
- the,of,것으로, 이런, 없다,어떤, 때문이다 등

>

벡터화 & 행렬 캐시

- 단어와 문서 간 희소 행렬 생성 후 단어별 빈도 확인 데이터 생성

1. 최소 단어별 개수값 5개 이상
2. 최대 문서 60% 이상 차지하는 흔한 단어 제거
3. 상위 40000개 단어만 추출

>

LDA 학습, K 선택

1. train: 85%, test: 15%

2. K 개수 후보 별로 train data로 LDA 모델을 학습해 test 예측 수행

3. Perplexity 지표 K별로 계산

>

최종 LDA 학습

1. train과 test를 합친 전체 데이터 기반으로 LDA 학습 진행

2. lda_topics.csv 파일로 저장

Part 2 >> LDA 분석 과정

-contents.xlsx

	A	B	C	D	E	F	
1	article_id	category	title	content	date	tag	sour
2	221763439722	커버스토리	언론과 독자 : 기자에게 언론 신뢰를 묻다...기자 87% "내 기사 신뢰	독자가 서슴없이 쓰는 '기레기'라는 표현에 정작 기자들은 어떤 생각을 가지고 있을까. 단순한 폄훼를 넘어 이 단어	2020. 1. 8. 8:3(	#언론,#신문과방	http
3	221766610231	커버스토리	[2019 언론인 조사] 언론 자유도 급반등...지상파 3사 체감 두드러져	언론인이 현장에서 느끼는 언론의 자유와 직업 만족도는 어떻게 변화했을까. 국민의 미디어 이용과 뉴스 소비 패턴	2020. 1. 10. 13	#신문과방송,#한	http
4	221770333278	빅카인즈	[뉴스 빅데이터로 본 '새해 소망'] 새해 소망 기사에 담긴 우리 사회	해마다 맞는 연말연시, 사람들은 어떤 희망을 품고 새해를 계획할까. 개인의 소망을 넘어 사회를 반영 하기도 하는 '	2020. 1. 14. 8:4	#대한민국,#키우	http
5	221770673553	커버스토리	[2019 신문산업 실태조사] 신문사 늘었지만 매출 제자리...구독 수익	침체 위기에 직면한 한국의 언론 산업은 지난 한 해동안 얼마나 많은 부침이 있었을까. 한국언론진흥재단이 진행한	2020. 1. 14. 10	#신문과방송,#한	http
6	221771937661	커버스토리	[독자에게 언론 신뢰를 묻다]시민 60% "한국 언론, 정치·경제권력으	오늘 당신이 선택한 뉴스는 무엇이며, 그 뉴스는 믿을만했는가? '언론을 신뢰하는가'라는 큰 물음에 독자는 신뢰하지	2020. 1. 15. 10	#한군언론진흥자	http
7	221771957034	산업 • 정책	줄어드는 광고, 정체하는 구독 수익 디지털 구독도 한계... NYT조차	미국의 경제학자 갤브레이스(John Kenneth Galbraith)는 현대를 '불확실성의 시대'로 규정한 바 있다. 불확실성의 시	2020. 1. 15. 9:4	#신문과방송,#한	http
8	221773238496	커버스토리	[2019 언론수용자 조사] '12.3%' 역대 최저 종이신문 구독률에도 결	한국언론진흥재단의 <언론수용자 조사>는 미디어와 뉴스를 이용하는 독자를 통해 바라본 우리 언론의 현주소	2020. 1. 16. 8:4	#신문과방송,#한	http
9	221773270282	커버스토리	[2019 10대 청소년 미디어 이용 조사] 10대의 미디어 이용 공식...모	한국 10대 청소년의 하루 평균 미디어 이용 시간은 약 6시간. 이른바 Z세대 로 불리며 향후 핵심적인 미디어 이용자	2020. 1. 16. 9:4	#신문과방송,#한	http
10	221773715615	취재기·제작기	시사IN <특별기획 - 빈집> 취재기 기자와 기획자, 그 어느 즈음에서	이미 활자로부터 멀어진 사람들을 붙잡을 수 있을까라는 생각이 출발점이었던 '빈집 종합 기획', 기자는 글쓰기만이	2020. 1. 16. 10	#신문과방송,#한	http
11	221773738103	기획연재	저널리즘 좀먹는 만악의 뿌리	미국의 언론인 앨 뉴하스(Al Neuharth)는 저널리즘에서 익명 취재원은 "모든 악의 뿌리(the root of all evil)"라고 했	2020. 1. 16. 11	#신문과방송,#한	http
12	221787915877	산업 • 정책	[MIT 고보 프로젝트] 필터버블을 필터로 잡을 수 있을까	보기 싫으면 보지 않는 것은 자유의지다. 그런데 내게 보이는 것이 과연 내가 보고 싶은 것일까. 이선 주커먼 매사추	2020. 1. 29. 8:3	#신문과방송,#한	http
13	221787926889	지금 언론계에	[출입처, 없애야 할 이유] 언론 효능감 최악... 지금의 출입처 구조로	한계에 봉착한 기자 출입처 제도를 어떻게 다룰 것인가? 유지와 폐지의 논란 속에 KBS가 필요한 영역과 역할을 제외	2020. 1. 29. 9:3	#신문과방송,#한	http
14	221788085108	취재기·제작기	[EBS <다큐프라임 - 진정성 시대> 제작기] 진짜를 감별하기 힘든 시	이유 없이 뜬 프로그램은 없다. 예능이든 드라마든 시대를 관통하는 코 드에 닿았기 때문일 것이다. 이 시대의 코드 '	2020. 1. 29. 10	#신문과방송,#한	http
15	221789531081	지금 언론계에	[출입처, 그래도 필요한 이유] 폐지만이 능사 아냐... 심층 취재와 출	현장의 중심인 출입처의 존폐가 도마 위에 올랐다. 출입처 중심 사고로 순치되는 폐해를 없애자는 폐지론은 자칫 권	2020. 1. 30. 8:4	#신문과방송,#한	http
16	221789536723	미디어포럼	북 리뷰: 《특종의 탄생》 사회의 소금, 짠맛을 잃어선 안 되는 기자	펼치자마자 작은 한숨이 나왔다. 충동적으로 제안을 수락한 게 조금 후회됐다. 이건 가까워도 너무 가까웠다. 내 직	2020. 1. 30. 9:	#신문과방송,#한	http
17	221789543168	기획연재	[영화 속 언론]<신문기자> 변하지 않는 권력, 변하지 않아야 하는 언	권력은 변하지 않는다. 좌든 우든, 한 번 잡으면 어떻게 하든 그것을 이어가기 위해 모든 수단과 방법을 동원한다. 그	2020. 1. 30. 18	#신문과방송,#한	http

1. title + content + tag 합친 후 text 이름의 데이터로 할당 (contents_prepared.csv)
2. 중복 article_id 제거
3. 단어의 조사 및 불필요 단어 제거

(제거 list: the, of, 것으로 and, 이런, 없다, 어떤, 때문이다, 이후, 있었다, 있었습니다, 했습니다, 우리, 당시, 모두, 다시, 따라, 이러한, 이를, 그런, 하고, 많이, 좋은, 그래서, 내가, 어떻게, 라는, 특히**)**

Part 2 >> LDA 분석 과정

- 상위 용어별 개수

```
print(quanteda::topfeatures(dfm_obj, n = 20))
    ai     정보   언론사 저널리즘     광고   디지털
  5088     4991     4935     4061     3879     3670
플랫폼 서비스     미국     취재     지역     사실
  3598     3262     3159     3149     3139     2819
  정부 새로운     기술 코로나 다양한     과정
  2660     2610     2599     2592     2517     2478
  시간     방식
  2468     2464
```

- 특정 문서의 단어 분포

```
Sample doc index: 1195  | article_id: 223098086631
> print(quanteda::topfeatures(dfm_obj[i, ], n = 15))
개인정보     호주     보호     기관     규제     조항     면제     개정
      37       29       24       21       12       12       11       10
    정보     강화     자유     우려     표현     기준     대상
       9        7        7        7        6        6        6

Sample doc index: 1395  | article_id: 223368007070
> print(quanteda::topfeatures(dfm_obj[i, ], n = 15))
인공지능     기술   생성형   언론사
      48       44       21       14
  플랫폼     기업     경향     올해
      14        9        8        8
뉴스룸 로이터저널리즘연구소     산업     선거
       8        8        7        7
  새로운     기반   중요한
       6        6        6
```

1. max_feature = 40000 : 상위 40000개 단어
2. min_df = 5 : count 값 최소 5개
3. max_df = 0.6 : 최대 문서 60% 이상 차지하는 단어
4. Topn_words = 15 : 벡터화 후 상위 15개 단어 추출

□ BoW (Bag of Words) : 단어별 빈도 확인 data(dfm_obj 이름으로 출력)

Part 2 >> LDA 분석 과정

LDA 특징 및 모형

$$P(W|\alpha,\beta)=\prod_{d=1}^{D}\int_{\theta_d}P(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

Latent Dirichlet Allocation 특징

문서 집합 속에 숨겨진 토픽 구조를 찾아내는 확률적 생성모형

- 입력데이터: 문서-단어 행렬
- 가정: 문서는 여러 토픽의 혼합이며 각 토픽은 단어들의 확률분포임을 가정함
- 장점: 토픽 구조를 자동으로 학습하며 대규모 텍스트를 요약 및 분류하는 데 유용하다.
- 단점: 단어 순서(문맥)을 무시하며, 하이퍼 파라미터와 토픽 수에 민감하다

LDA 모형 파라미터

α,β: 하이퍼파라미터 (Dirichletprior)
θ_d: 문서의 토픽 분포
ϕ_k: 토픽의 단어 분포
z_dn: 단어의 토픽 할당

Part 2 >> **LDA 분석 과정**

train data LDA 학습 및 K 선택

Perplexity

k=6: perplexity=5784.69 (before: 10336.54)
k=8: perplexity=5616.75 (before: 10148.60)
k=10: perplexity=5533.45 (before: 9980.94)
k=12: perplexity=5547.94 (before: 9868.79)

$$\text{Perplexity}(W) = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d \mid \hat{\Theta}, \hat{\Phi})}{\sum_{d=1}^{D} N_d}\right)$$

$$p(w = v \mid d) = \sum_{k=1}^{K} \theta_{dk}\, \phi_{kv}.$$

1. Random으로 train 85%, test 15% split 진행
2. Method="VEM" (Variational EM: 변분 EM)으로 parameter 값 측정
→ 베이지안 방법의 EM 알고리즘 기반 파라키터 최적화 방법
3. Test set perplexity 계산: 평균 음의 로그우도를 지수화한 값

Part 2 >> LDA 분석 과정

- LDA 학습 결과 예시: topics_tbl 이름으로 출력 (총 10개 존재)

	topic	rank	term	weight
1	0	1	사람	0.005845351
2	0	2	취재	0.005492307
3	0	3	이야기	0.004830609
4	0	4	시간	0.004383122
5	0	5	생각	0.004087221
6	0	6	인터뷰	0.003925940
7	0	7	현장	0.003810494
8	0	8	사람들	0.003774726
9	0	9	지금	0.003640566
10	0	10	기자들	0.003404899
11	0	11	사실	0.002907579
12	0	12	이들	0.002744554
13	0	13	마음	0.002716534
14	0	14	그렇게	0.002688378
15	0	15	자신	0.002620483

	topic	rank	term	weight
16	1	1	선거	0.009555035
17	1	2	영국	0.009457237
18	1	3	bbc	0.007537431
19	1	4	공영방송	0.006969899
20	1	5	정부	0.006775636
21	1	6	정보	0.006731782
22	1	7	프로그램	0.006336571
23	1	8	독일	0.005561813
24	1	9	프랑스	0.005369609
25	1	10	대선	0.005238680
26	1	11	후보	0.005135830
27	1	12	tv	0.004909624
28	1	13	정치	0.004801266
29	1	14	백신	0.004751019
30	1	15	대통령	0.004220343

136	9	1	지역	0.010909582
137	9	2	취재	0.008410993
138	9	3	교육	0.007942033
139	9	4	기획	0.004681538
140	9	5	과정	0.004246391
141	9	6	현장	0.004182876
142	9	7	학교	0.003671411
143	9	8	청년	0.003162949
144	9	9	노동	0.003159144
145	9	10	대학	0.002754196
146	9	11	시간	0.002648392
147	9	12	인터랙티브	0.002572768
148	9	13	다양한	0.002549626
149	9	14	지면	0.002537958
150	9	15	지원	0.002449035

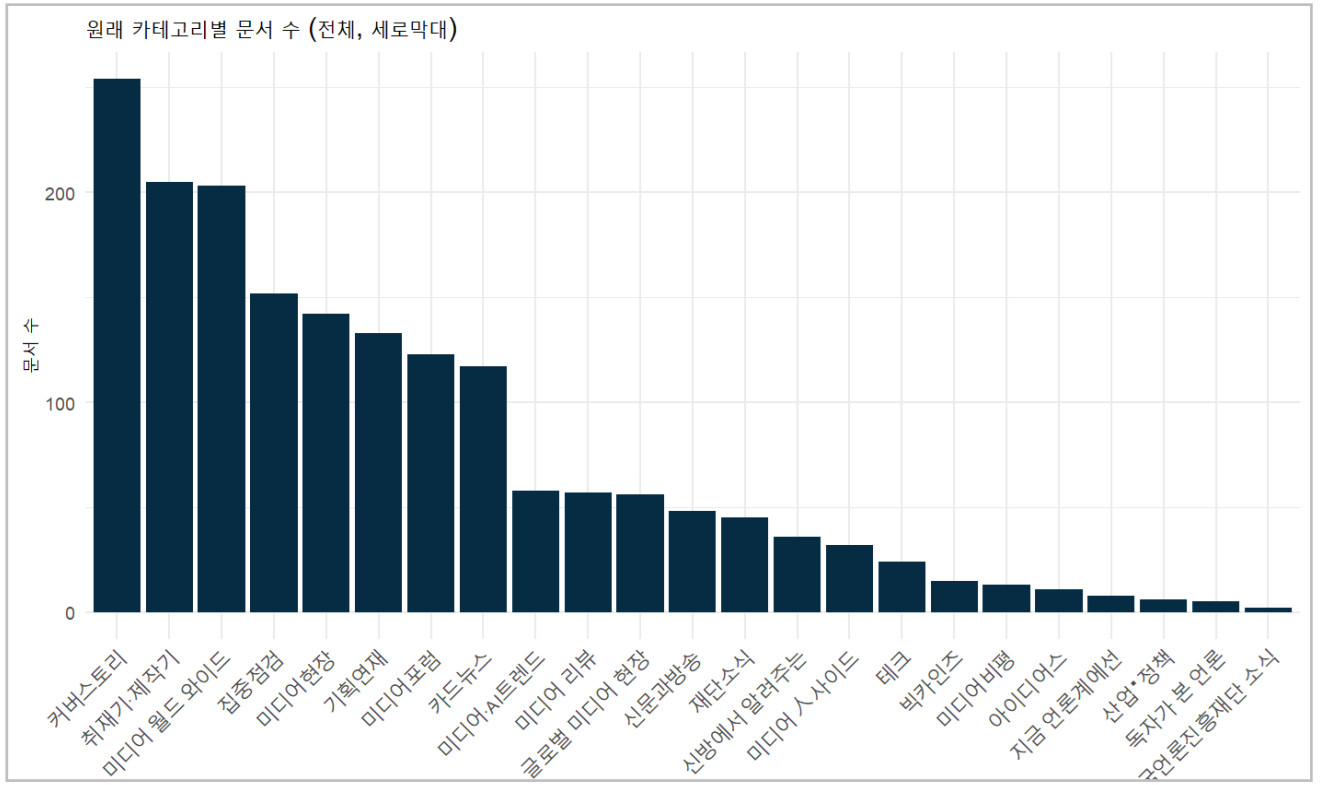
각 topic 별 rank 1위 단어

0: 사람
1: 선거
2: 언론사
3: 플랫폼
4: 디지털
5: 광고
6: ai
7: 저널리즘
8: 코로나
9: 지역

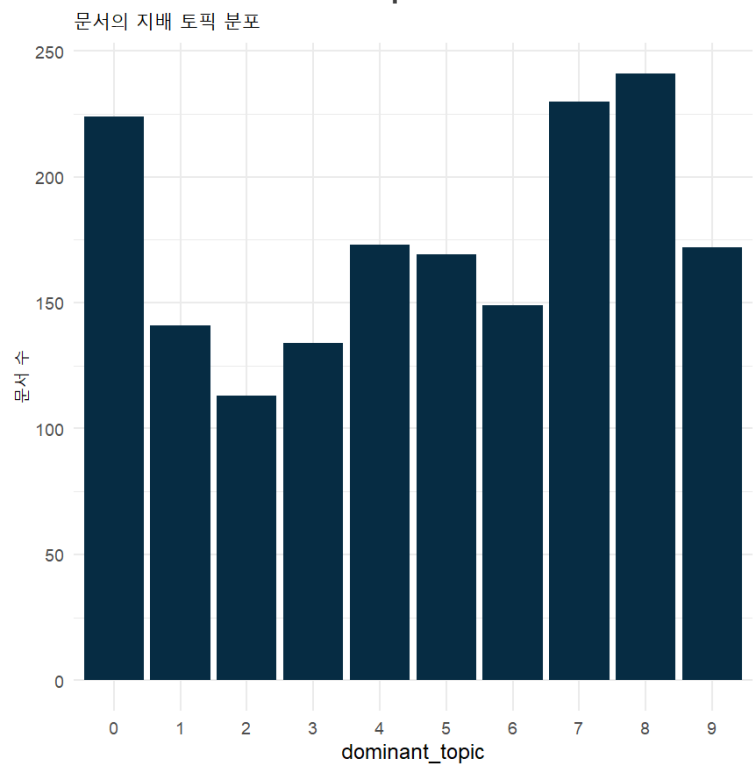
- train과 test를 합친 전체 데이터 기반
- LDA: Latent Dirichlet Allocation (LDA) 모델 로 최종 학습 진행
- topic: 군집화 한 토픽 / rank: 단어 비중에 대한 순위 / term: 해당 단어 / weight: 실제 단어 비중

Part 2 >> LDA 분석 과정

기존 카테고리별 문서 분포 시각화



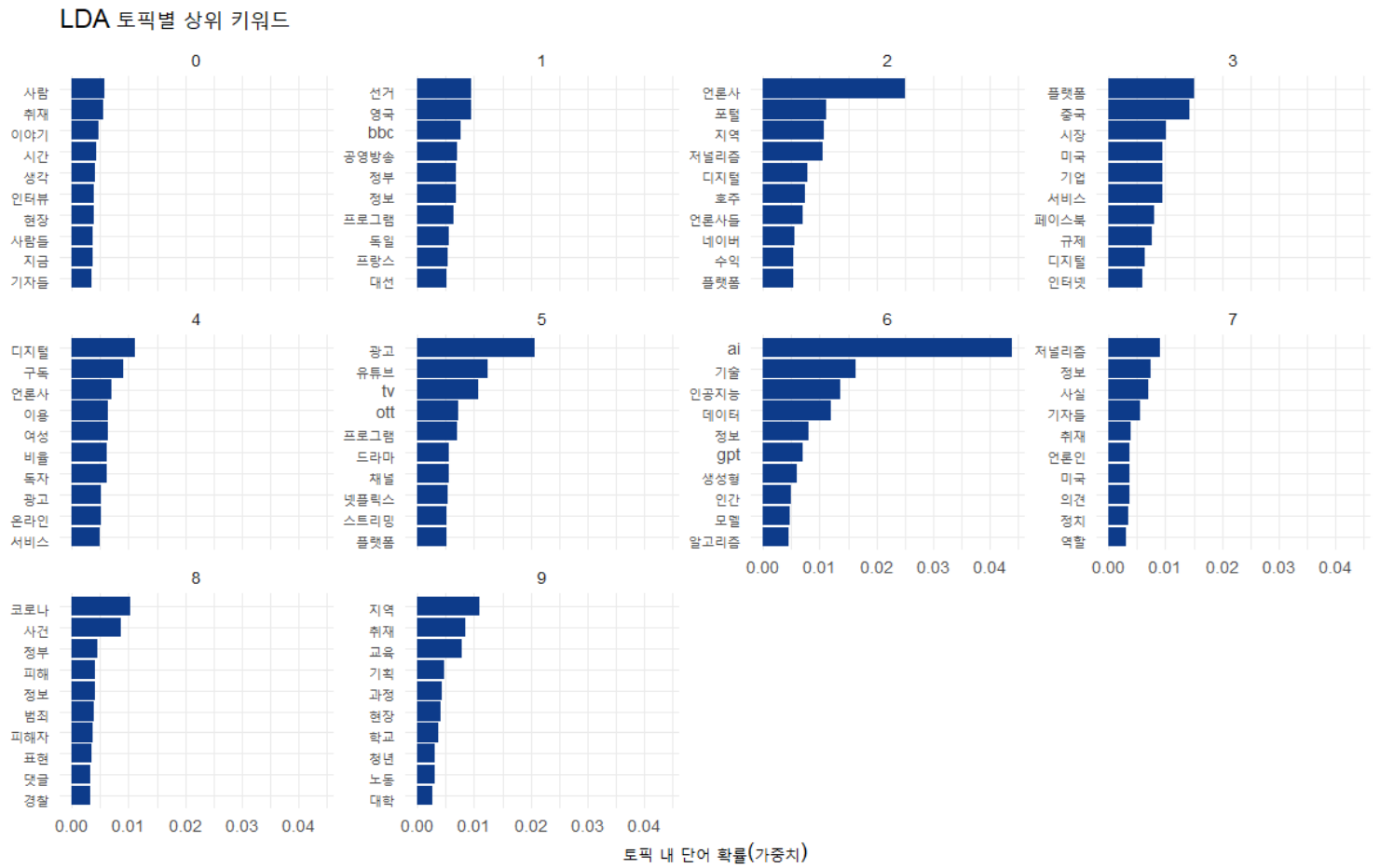
LDA 토픽별 문서 분포 시각화



기존 카테고리별 문서 분포는 특정 카테고리에 치우쳐진 반면에 LDA 토픽은 비교적 분포가 고른 것을 알 수 있다. 이는 본문의 키워드를 LDA 카테고리가 고르게 분배되었음을 알 수 있다.

Part 2 >> LDA 분석 과정

토픽별 상위 키워드 가중치 시각화



토픽별 가중치 기반 카테고리명 제안

0: 현장·인터뷰 스토리

1: 글로벌 선거·공영방송

2: 포털·플랫폼 유통과 수익

3: 빅테크·규제와 시장

4: 구독·독자 데이터

5: 영상·OTT·광고

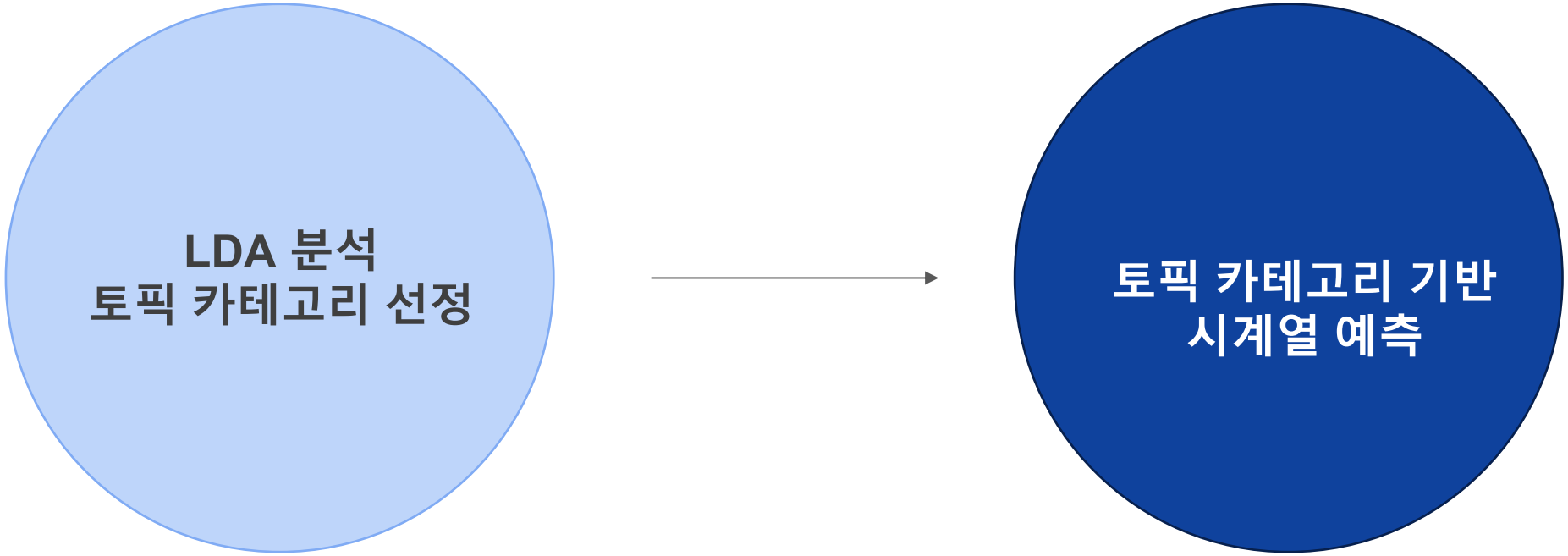
6: AI·생성형 미디어

7: 저널리즘 윤리·팩트체크

8: 사건·재난·공공

9: 지역·교육·노동 현장

Part 2 >> 시계열 예측



Part 2 >> 시계열 예측

article_metrics_monthly.xlsx
: contents.xlsx와 함께 기사별 Original 카테고리 이용

	A	B	C	D	E
1	article_id	period	comments	likes	views_total
2	221763439722	2023-07	0	0	3
3	221763439722	2023-08	0	0	8
4	221763439722	2023-09	0	0	5
5	221763439722	2023-10	0	0	5
6	221763439722	2023-11	0	0	11
7	221763439722	2023-12	0	0	7
8	221763439722	2024-01	0	0	2
9	221763439722	2024-02	0	0	0
10	221763439722	2024-03	0	0	1
11	221763439722	2024-04	0	0	3
12	221763439722	2024-05	0	0	8
13	221763439722	2024-06	0	0	3
14	221763439722	2024-07	0	0	12
15	221763439722	2024-08	0	0	7
16	221763439722	2024-09	0	0	8
17	221763439722	2024-10	0	0	7
18	221763439722	2024-11	0	0	4
19	221763439722	2024-12	0	0	7
20	221763439722	2025-01	0	0	6
21	221763439722	2025-02	0	0	6
22	221763439722	2025-03	0	0	2

lda_doc_topics.csv
: LDA 카테고리 –기사별 dominant_topic을 기준으로 이용

	article_id	stringsAsFactors	category	title	dominant_topic
1	221763439722	FALSE	커버스토리	언론과 독자 : 기자에게 언론 신뢰를 묻다...기자 87% "내 기사 신...	7
2	221766610231	FALSE	커버스토리	[2019 언론인 조사] 언론 자유도 급반등...지상파 3사 체감 두드러...	4
3	221770333278	FALSE	빅카인즈	[뉴스 빅데이터로 본 '새해 소망'] 새해 소망 기사에 담긴 우리 사...	8
4	221770673553	FALSE	커버스토리	[2019 신문산업 실태조사] 신문사 늘었지만 매출 제자리...구독 ...	4
5	221771937661	FALSE	커버스토리	[독자에게 언론 신뢰를 묻다]시민 60% "한국 언론, 정치·경제권력...	4
6	221771957034	FALSE	산업・정책	줄어드는 광고, 정체하는 구독 수익 디지털 구독도 한계... NYT조...	4
7	221773238496	FALSE	커버스토리	[2019 언론수용자 조사] '12.3%' 역대 최저 종이신문 구독률에도 ...	4
8	221773270282	FALSE	커버스토리	[2019 10대 청소년 미디어 이용 조사] 10대의 미디어 이용 공식....	4
9	221773715615	FALSE	취재기·제작기	시사IN <특별기획 - 빈집> 취재기 기자와 기획자, 그 어느 즈음...	9
10	221773738103	FALSE	기획연재	저널리즘 좀먹는 만악의 뿌리	7

topic_score	topic_label_suggested
0.5408037	저널리즘/정보/사실/기자들
0.9244047	디지털/구독/언론사/이용
0.8111932	코로나/사건/정부/피해
0.9994647	디지털/구독/언론사/이용
0.6762119	디지털/구독/언론사/이용
0.6556837	디지털/구독/언론사/이용
0.9995782	디지털/구독/언론사/이용
0.8619769	디지털/구독/언론사/이용
0.6665446	지역/취재/교육/기획
0.8943668	저널리즘/정보/사실/기자들

Part 2 >> 시계열 예측

STEP 1

- 기사별로 기존 카테고리와 LDA 카테고리 mapping

- 기간별로 카테고리의 조회수 합, 기사 수 집계

- 카테고리 개수대로 조회수 및 기사수에 대한 시계열 data 생성

>

STEP 2

- 카테고리별 조회수(group_views)를 같은 기간의 전체 조회수(total_views)의 합으로 나눠 share 변수 생성

- 그룹별 기간에 따른 조회수 합, 기사수, 총 조회수, share 변수 데이터들이 각각 존재

>

STEP 3

- 분산이 가장 작은 열 1개를 기준범주로 전월에 대한 share의 lag data 생성 후 설명변수로 사용 (기준변수 제거)
: orig_share_정치 □ orig_share_정치_lag1

>

STEP 4

- Fourier 함수 사용
: 시계열성 (seasonality 반영)

- 표준화 진행
: z-score 점수 변환

- 반응변수
: BoxCox 변환

- model: ARIMAX
: 동적 회귀+ARIMA

Part 2 >> 월별 데이터 구조

원본 집계 (월별)

month	total_views	…	share_A	share_B	share_c
2024-01	10,000	…	0.50	0.30	0.20
2024-02	11,200	…	0.55	0.25	0.20
2024-03	9,800	…	0.52	0.28	0.20
2024-04	9,500	…	0.60	0.30	0.10
2024-05	10,400	…	0.58	0.32	0.10

날짜별로 전체 카테고리에서 차지하는 비중에 대한 변수인 share를 생성한 데이터의 형태이다.

Part 2 >>

월별 데이터 구조 예시

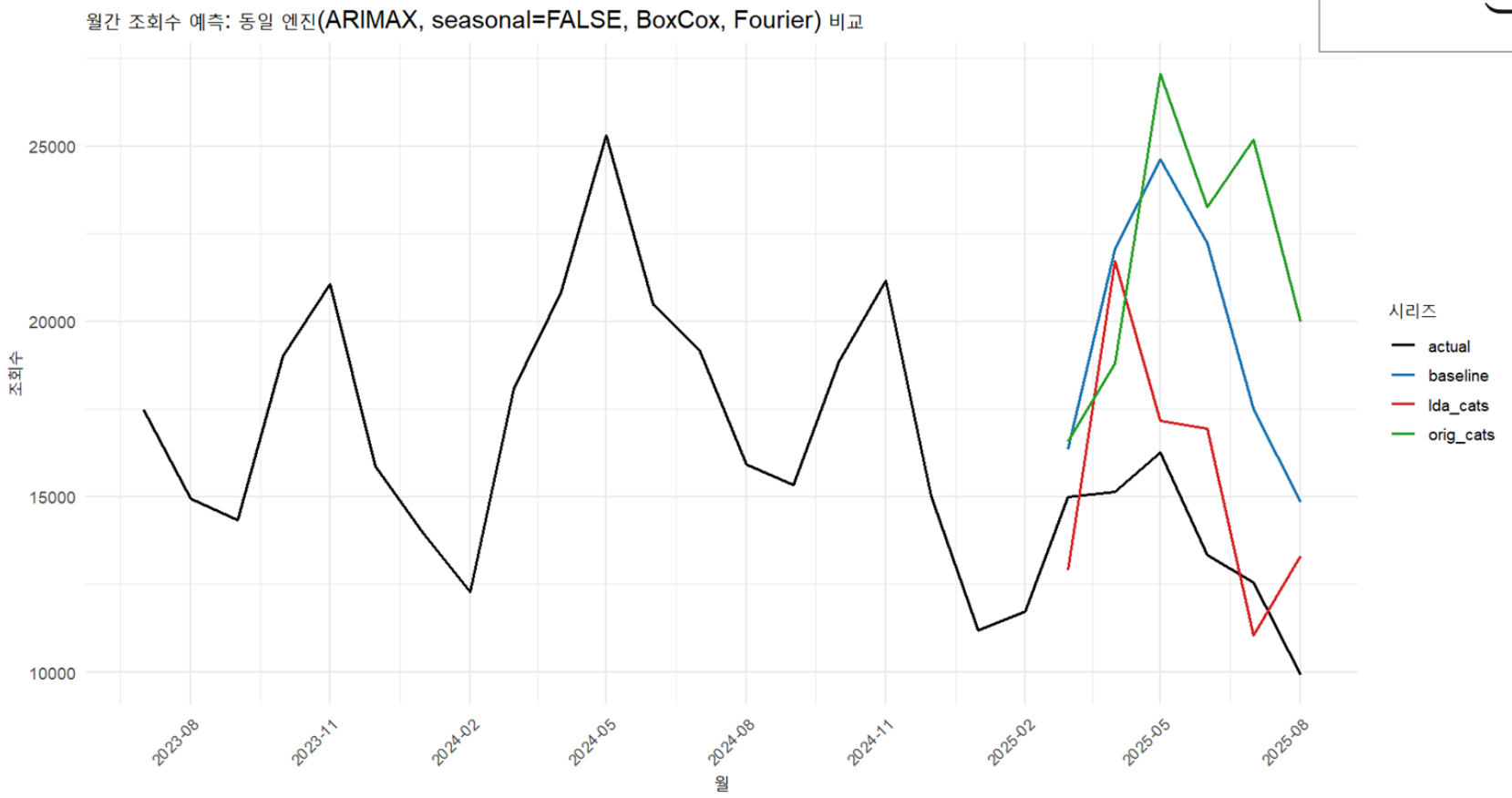
전월 lag 생성 후의 데이터

month	total_views	…	share_A_lag1	share_B_lag1	share_c_lag1
2024-02	11,200	…	0.50	0.30	0.20
2024-03	9,800	…	0.55	0.25	0.20
2024-04	9,500	…	0.52	0.28	0.20
2024-05	10,400	…	0.6	0.30	0.10

예시의 세 카테고리의 lag가 전월 share값으로 생성이 된다. 이 때 행별 share의 총합이 1이 되므로 분산이 가장 작은 카테고리의 열 1개를 제거해야한다. 이후 Fourier 변수와 표준화 과정을 거쳐 예측을 수행한다.

Part 2 >> 시계열 예측

예측 결과 시각화



ARIMAX(동적 회귀 + ARIMA 오차) 형태

$$\underbrace{\mathrm{BoxCox}(y_t)}_{\text{타깃 변환}} = \beta_0 + \beta^\top X_t + \varepsilon_t, \quad \phi(B)\varepsilon_t = \theta(B)a_t$$

* Fourier
- dummy 변수 대신 sin, cos 선형 결합으로 근사
- 시간: t=1,…,T
- 계절 주기: s (월별=12)
- 사용할 조화수 : k=2

(파형 쌍의 개수)

$$\underbrace{\sin\left(\frac{2\pi k}{s}t\right)}_{\text{사인}}, \quad \underbrace{\cos\left(\frac{2\pi k}{s}t\right)}_{\text{코사인}}$$

Baseline: Fourier
LDAcats: Fourier+LDA_lags
Origcats: Fourier+orig_lags

빨간색 선인 lda_cats를 설명변수로 쓴 모델이 가장 실제값과 유사하게 예측되는 것을 확인할 수 있다.

Part 2 >> 시계열 예측

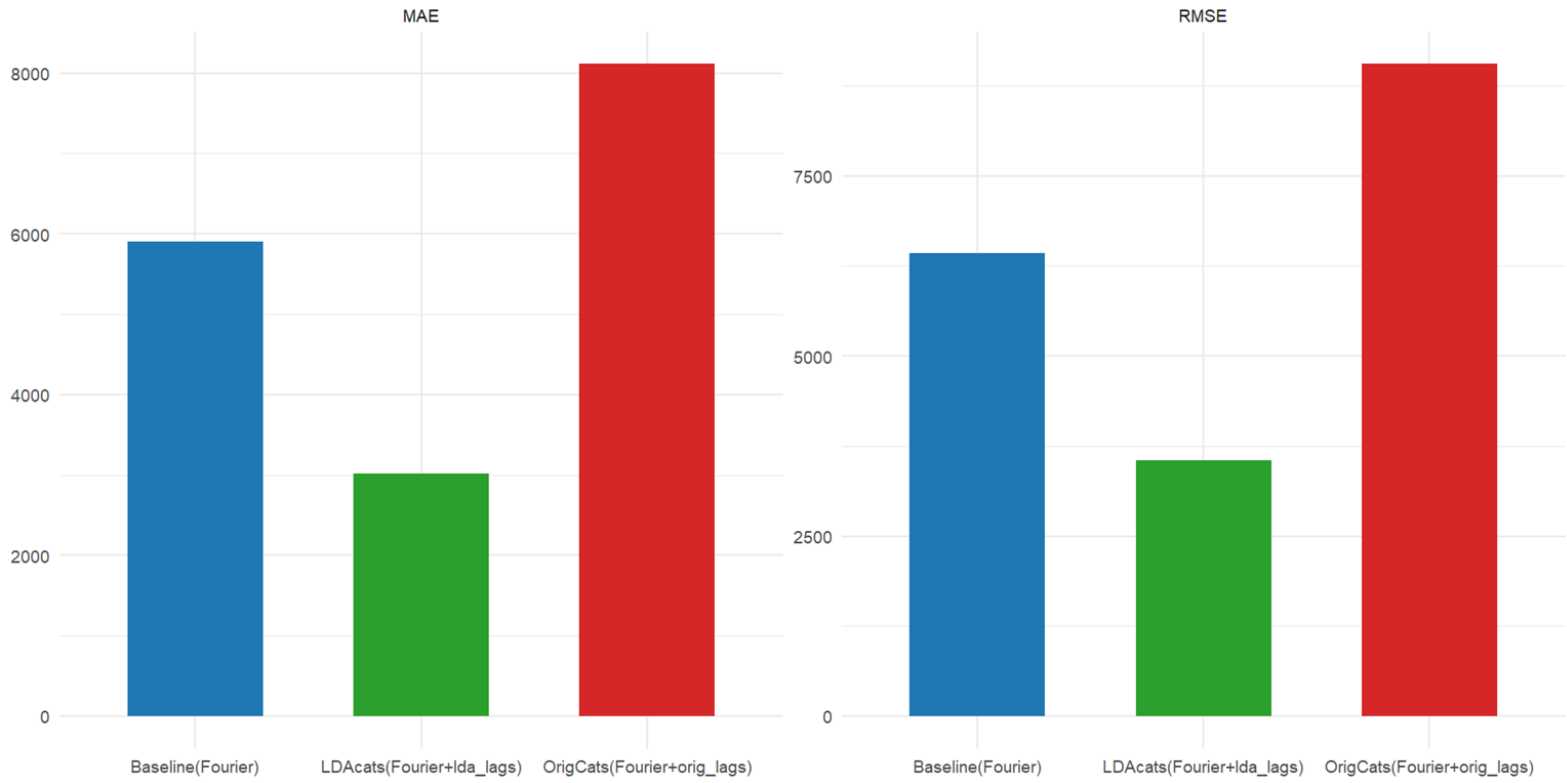
Baseline, LDA, Original 기반 MAE / RMSE 시각화

Baseline: 6425.828 / 5903.658 / 43.66798
LDAcats: 3549.542 / 3021.003 / 22.75156
Origcats: 9053.658 / 8112.119 / 62.93350

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$



초록색 막대인 lda_cats를 설명변수로 쓴 모델의 MAE, RMSE 값이 가장 작게 나온 것을 알 수 있다.

3

MMM 분석

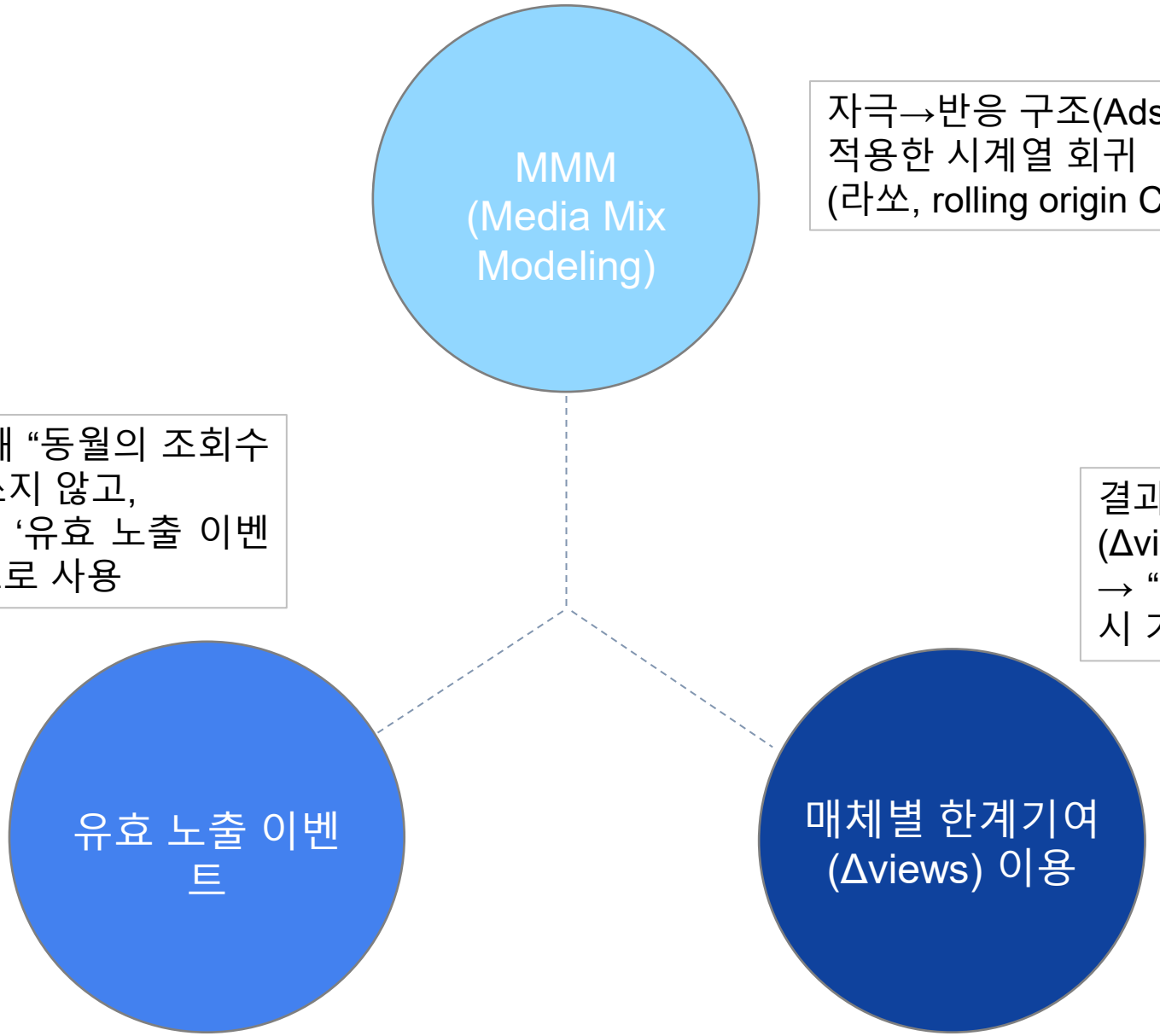
Q1

월별 조회수(views_total)를 높이기 위해 어떤 매체 노출을 얼마나 확대/축소해야 하는가?

Q2

“네이버/구글/페이스북 …”처럼 운영 가능한 단위의 매체군으로 묶으면,
실제 의사결정(편성·배치·프로모션)에 어떤 우선순위가 생기는가?

Part 3 >> 접근 요약



Part 3 >> MMM 분석 과정

STEP 1

노출 이벤트 정의 (Exposure Events)

기사×월×매체 단위에서 share ≥ τ(기본 0.05)이면 노출 1건으로 간주

월×매체로 합산 → m_{channel} (이벤트 건수)

>

STEP 2

시계열 전처리

Adstock(감쇠): 이전 월 효과 누적(λ ∈ {0.4, 0.6, 0.8})

포화(Saturation): log1p(adstock(m)) → m_*_sat만 모델에 투입 (원시 m_*는 제외)

>

STEP 3

회귀 모형

라쏘(Lasso, α=1) + 롤링 오리진 교차검증으로 일반화 성능 확보

통제변수: 발행량(n_posts_total), 계절(month), 종속변수 지연항(y_lag1,y_lag2)

>

STEP 4

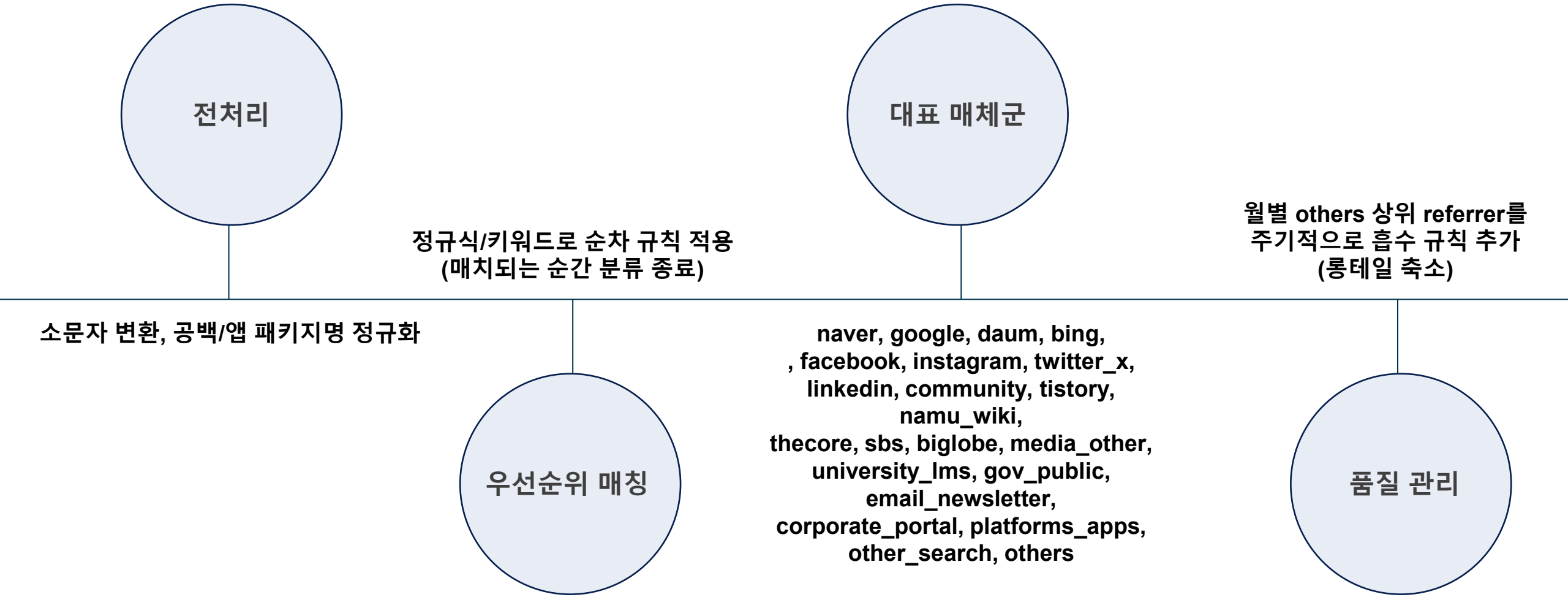
모형 선택 & 해석

CV로 최적 adstock λ와 정규화 파라미터 선택

최신 월 상태에서 채널별 노출 +1 시뮬레이션 → Δviews(한계기여) 산출

보조로 계수 절댓값 중요도(|β|) 제공(부호 정보는 Δviews로 판단)

Part 3 >> **MMM 분석 과정** 매체 재분류 (Referrer >> Media)



매체 재분류 규칙

예시 규칙 (발췌)

- (^|\.)google\., com.google.android.googlequicksearchbox → google
- 네이버 또는 (^|\.)naver\.com/search.naver.com/m.search.naver.com → naver
- (^|\.)t\.co$|twitter|x\(twitter\) → twitter_x
- teams.microsoft|slack|notion.so|canva.com → platforms_apps
- 위 규칙에 모두 해당 없으면 → others

- 주요 매체: `naver, google, daum, bing, facebook, instagram, twitter_x, linkedin, community, tistory, namu_wiki, thecore, sbs, biglobe, media_other, …`
- 정부·대학·기업·앱 플랫폼 등은 **유사군**으로 묶음: `gov_public, university_lms, corporate_portal, platforms_apps`

Part 3 >> **MMM 분석 과정**

모형 개발

시계열 회귀모형 + 라쏘

- 컨트롤(운영/환경 변수):
  - n_posts(t): 그 달 발행한 고유 기사 수
  - Y(t−1),Y(t−2): 직전/직전2개월 총조회수 (자기회귀 효과)
  - month dummy: 월 계절성(1월 vs 8월 등 계절/이슈 주기)

- 채널별 포화된 노출:
  - 위에서 만든 각 매체 c의 포화 after-adstock 값 $s_c(t) = \log(1 + a_c(t))$

$$Y(t) = \beta_0 + \beta_1 \cdot n\_posts(t) + \beta_2 \cdot Y(t-1) + \beta_3 \cdot Y(t-2) + \sum_{\text{month}=1}^{12} \gamma_{\text{month}} \cdot \text{MonthDummy}_{\text{month}}(t) + \sum_{c \in \text{channels}} \theta_c \cdot s_c(t) + \varepsilon_t$$

- β_0: 절편
- θ_c: 채널 c의 기여 계수
- ε_t: 오차

Part 3 >> **MMM 분석 과정**

모형 선택 및 모수 추정

결과 해석

- Adstock λ=0.4가 교차검증(rolling origin)에서 MAE ≈ 157로 최적
  채널별 한계기여(Δviews, “마지막 월에서 노출 1건↑ 시 기대 조회수 변화”)
- 확대 권장(+)
  thecore +3,546, biglobe +2,507, facebook +812, linkedin +728, instagram +268, community +261, platforms_apps +249, bing +67
- 영향 미미/0 (보류)
  naver, daum, media_other, gov_public, corporate_portal, other_search, tistory, twitter_x, university_lms, zum, others → 0
- 축소 후보(−)
  nate −1,210, sbs −827, namu_wiki −345, email_newsletter −89.5, google −0.41

분석 결과 해석

서비스 / 콘텐츠 개선 제안

- 운영 배분: 다음 달 테스트 예산을 thecore > biglobe > facebook ≈ linkedin > instagram ≈ community > platforms_apps > bing 순으로 증량
- 리스크 관리: thecore/biglobe에 월별 노출 상한 설정, 단계적 확대(A/B 모니터링)
- 소셜 퍼블리싱: 썸네일·헤드라인·슬라이드 카드 최적화, 재유입 링크 구조 점검 (UTM/앱→웹 전환)
- 저효율 축소: nate/sbs/namu_wiki/email_newsletter는 감축 후 4주 관찰, 필요 시 완전 제거

4

결론 및 시사점

Part 4 >> 결론 및 시사점

주제	문제 제기	예상 분석 결과	분석 결과	시사점
1. LDA: 콘텐츠 체계 재정비 및 조회수 예측력 강화	**현재 카테고리 시스템은 주관적이며 불균형**하고, 30~40대 핵심 독자층의 선호도(예: 30대 남성→기술/비즈니스)를 반영하지 못함	**카테고리 재정비** 기존의 치우쳐진 카테고리 분포 대신, 본문 키워드를 고르게 분배하는 몇개의 의미 있는 토픽 카테고리가 정의될 것으로 예상함	**카테고리 재정비** 총 10개의 카테고리가 정의됨 (현장·인터뷰 스토리, 글로벌 선거·공영방송, 포털·플랫폼 유통과 수익 등)	1. **객관적인 LDA 10개 토픽 체계로 즉시 전환** 2. 특히, **LDA 토픽 모델이 기존 카테고리 대비 조회수 예측 정확도가 가장 높음**이 ARIMAX 분석을 통해 입증되었으므로, 신규 토픽을 콘텐츠 중심의 기획
	단순 조회수(Views) 외에 좋아요($r \approx 0.6$)와 댓글($r \approx 0.3$) 등 **독자 참여도를 종합적으로 고려한 다차원 평가**가 필요	**시계열 예측 향상 예상** LDA 기반 카테고리(LDA cats)를 설명변수로 사용했을 때, Baseline 및 기존 카테고리 모델 대비 MAE 및 RMSE 값이 가장 작게 나와 예측 성능이 향상될 것으로 예상함	**시계열 예측 향상** LDA 기반 카테고리의 예측 정확도가 가장 좋게 나옴 카테고리 재정비 및 제안이 의미있는 변수로 작용하여 예측 성능이 향상됨	
2. MMM (미디어 믹스 모델링) 기반 유입 경로 최적화	thecore (+3,546)와 biglobe (+2,507)가 노출 1건당 기대 조회수에서 압도적인 한계기여를 보임	**최적 매체 배분 제안** 노출 확대 대상으로 분류되는 매체와, 노출 축소 대상으로 분류되는 매체가 나뉠 것으로 예상함. 특히나 facebook, nate, namu_wiki 등은 명망 있는 유입경로는 확대 대상으로 분류될 것으로 보였음	**최적 매체 배분 제안** thecore, biglobe, facebook 등의 매체는 노출 확대 권장(+Δviews)이 도출되며, nate, sbs, namu_wiki 등은 축소 후보**(-Δviews)로 제안되어 운영 배분 전략에 활용할 수 있을 것으로 확인함. 유입 경로에 대한 명확한 확대 및 축소 여부를 판단할 수 있는 근거를 마련함	1. **다음 달 테스트 예산을 thecore와 biglobe에 최우선적으로 증량 필요** 2. 다만, 리스크 관리를 위해 월별 노출 상한 설정 및 단계적 확대로 접근
	nate (−1,210), sbs (−827), namu_wiki (−345) 등은 음의 한계기여를 보임			1. 이들 **저효율 채널들은 즉시 감축**하고 4주간 관찰하며, 필요 시 운영에서 완전히 제거하는 방안을 고려해야 함

Part 4 >> Appendix: 강화 학습 기반 컨텐츠 추천 시스템

❖ 배경

- LDA 분석으로 10개 토픽 추출
- 토픽 0 : 사람, 토픽 1 : 선거, 토픽 2 : 언론사...
- 질문 : **"어떤 토픽을 누구에게 추천할 것인가?"**

❖ 실험결과

예상과 다른 결과 : Greedy >> LinUCB (32배 차이)

전략	평균보상	선택 횟수	매칭률
Greedy	0.0475	992	9.9%
LinUCB α=0.5	0.0015	10	0.1%
LinUCB α=1	0.0015	10	0.1%
LinUCB α=2	0.0015	10	0.1%
Random	N/A	962	9.6%

❖ 결론

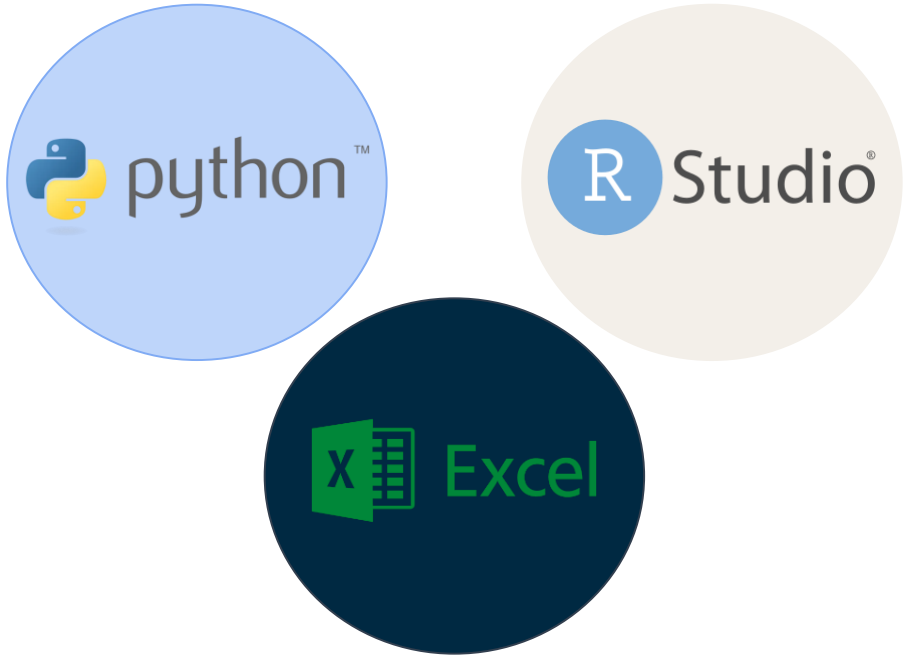
- Greedy는 단기 성과에 효과적 : 안정적이고 검증된 방식
- 오프라인 평가의 한계 확인 : 새로운 전략은 실제 테스트 필수
- 실무 적용은 단계적으로 : Greedy 먼저 → A/B 테스트 → 확대

지금의 추천 시스템은 충분하지 않으므로 충분한 시간과 데이터가 쌓여야 의미있는 결과가 생길 것으로 보임

5

팀인원 역할, 분석 도구 및 참고 문헌

Part 5 >> 분석 도구 및 참고 문헌



팀인원 역할분담

권형근

데이터의 전반적인 EDA 수행
카테고리 불균형 및 독자 인구 통계 분석
강화학습 분석: RL (LinUCB) 시스템 설계 및 전략별 실험/비교

류현지

텍스트 데이터 전처리 및 벡터화
LDA 토픽 K 선정 및 최종 학습: LDA 분석 및 카테고리 재정비
시계열 데이터 생성 및 예측 모델 구축(ARIMAX) 및 성능 검증
ppt 양식 틀 생성

신현수

MMM (Media Mix Modeling)으로 유입경로 분석
유입 경로 매체 재분류 규칙 정의
MMM 모델(Lasso) 구축 및 한계 기여 도출,

참고 문헌

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet allocation*. *Journal of Machine Learning Research*
- Regression Shrinkage and Selection via the Lasso
- Econometrics for Modern Marketing: From OLS to Bayesian MMM, Causal Lift, and Geo Experiments (in Python)

Part 5 >> 공모전 결과

DACON
DATA TO VALUE

2025 신문과 방송 독자 데이터 분석 아이디어 경진 대회

수상자 여러분 축하드립니다!

최종 등수는 아래와 같습니다.

Below is the final rank for the competition.

1위 : 망고고망고 팀

2위 : K-Data Hunters 팀

3위 : haribo_jelri 팀

한국언론진흥재단 소식

[공지] 2025 신문과방송 독자 데이터 분석 아이디어 경진대회 결과 발표

신문과방송
2025. 11. 25. 10:29

+ 이웃추가

지난 9.22(월)~10.31(금)까지 진행된
'신문과방송 독자 데이터 분석 아이디어 경진대회'가 성공적으로 마무리되었습니다!

한국언론진흥재단이 주최하고 데이콘이 주관한 이번 대회는
〈신문과방송〉의 독자 데이터를 바탕으로, 실제 콘텐츠 기획과 서비스 개선에
도움이 될 수 있는 아이디어를 발굴하기 위해 마련되었는데요.

데이터 분석 역량을 가진 대한민국 국민 46팀이 참여해
독자 이용 행태와 블로그 콘텐츠 반응을 꼼꼼하게 분석하였고,
언론과 미디어 서비스의 미래를 위한 다양한 전략을 제시해주셨습니다!

1차 대국민 온라인 평가와 2차 전문 심사위원단의 평가를 종합해
최우수상 1팀과 우수상 2팀이 최종 선정되었는데요,

최종 수상작을 미리 공개하자면...!

[한국언론진흥재단] 2025 신문과 방송 독자 데이터 분석 아이디어 경진대회 수상작 관련 원고 작성 요청 받은편지함 ×

김수지 <suji@kpf.or.kr>
lchaeyeon20, urisem625, 나, oh3912, zero27120에게

안녕하세요, 한국언론진흥재단 산업분석팀 김수지입니다.

2025 신문과방송 독자 데이터 분석 아이디어 경진대회 수상을 진심으로 축하드립니다.
상장과 기념품은 데이콘에 제출해주신 주소로 **12월 1주차 또는 2주차에 발송될** 예정입니다.

또한 귀하의 수상작은 **월간 신문과방송 2026년 신년호에서 '독자 데이터로 보는 신문과방송' 특별 기획 기사로 게재될 예정입니다.**
위 내용은 데이콘의 경진대회 포상 안내 부분에 게재돼 있습니다(https://dacon.io/competitions/official/236606/overview/prize).

이에 수상자 여러분께서는 본인의 수상작을 기반으로 한 원고를 작성해 보내주시길 요청드립니다.

원고에는 다음의 내용이 포함될 수 있습니다만, 분석한 내용을 바탕으로 자유로이 가감해 소개해주시길 부탁드립니다.
- 제출한 분석 아이디어 및 분석 과정의 핵심 내용
- 분석 결과를 통해 도출한 신문과 방송 매체의 특징·해석
- 데이터 기반으로 바라본 향후 개선 방향 또는 제언
- 분석을 진행하며 관찰한 독자 행동·이용 패턴에 대한 시사점

이와 더불어 참고하실 만한 글을 아래 공유드립니다.
이번 원고는 신문과방송에 처음 게재되는 새로운 형식의 기획 기사이기 때문에 완전히 동일한 유형의 선행 사례는 없으나,
원고 구성과 서술 방식 등을 참고하시는 데에는 도움이 될 것 같습니다.

⇨[뉴욕타임스 <언론 윤리>] AI 환경에서도 정확성, 독립성, 공익성을 지켜야 하는 이유
: https://blog.naver.com/kpfjra_/223888585035

⇨[최근 3년간 《신문과방송》 커버스토리 분석] 사회 문제에 대한 언론의 역할, 뉴스 유통 환경, 그리고 인공지능이 주요 주제
: https://blog.naver.com/kpfjra_/223401916727

바쁘시겠지만 메일 내용 확인 시 회신 꼭 부탁드리겠습니다.
세부 내용은 아래 박스를 참고해주십시오.