

나 만 의 뉴 스 콘 텐 츠 큐 레 이 셴 서 비 스 - 뉴 스 캣 ( N E W S C A T )

오늘의 이슈, 이견 어때?

# 뉴스캣

## (NewsCat)

# CONTENTS

---



- 01** 서비스 소개
- 02** 분석 과정
- 03** 서비스 구현
- 04** 서비스 시연
- 05** 향후 계획 및 발전 방향



# 01

## 서비스 소개

- 01. 아이디어 제안 배경
- 02. 서비스 설명
- 03. 사용 기술 스택

# 01

## 배경

# 편향적 추천 기사, 조작되는 선호



- ▶ 네이버 실시간 검색어 폐지, 대중의 **공통적 관심사 및 이슈 확인** 어려움
- ▶ 보수 성향의 기사 혹은 언론사를 더 많이 노출하는 포털들의 **편중된 편집 알고리즘**
- ▶ 자극적인 뉴스 기사 전면 노출, 대형 언론사 위주의 뉴스 구독 시스템으로 인지도 낮은 지역 매체 및 전문지 외면

**채영길** / 교수 / 한국외대 미디어커뮤니케이션학과

지금 포털은 우리에게 '선호하는 것들이 이런 것이 되어야 돼'라고 하는 것을 보여준다는 것이죠.

그래서 우리의 선호가 우리의 것이 아니라, 포털이 줘준 선호 속에서 우리의 선호가 결정되는 경향이 크고요.

그렇기 때문에 뉴스의 소비가 굉장히 편향적이 되는 것이고, 확증 편향도 더 강화되는 이유가 바로 거기에 있습니다.

서비스 기업의 의도/편향 개입 없이, 순수하게 사용자 선호만 고려하는 뉴스 추천 서비스 必

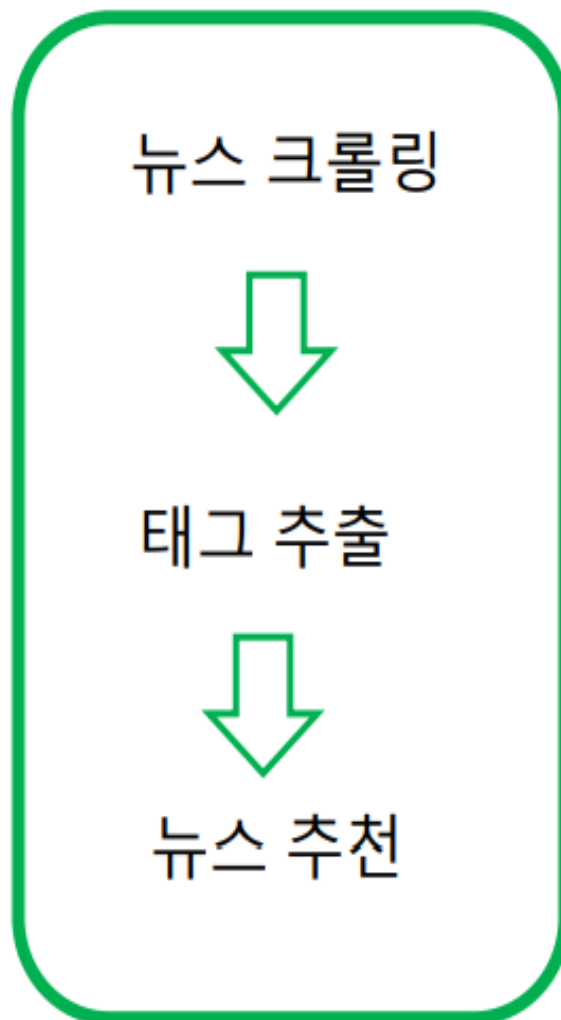
# 02

## 서비스 설명



### 추천 알고리즘을 활용한 뉴스 콘텐츠 큐레이션 서비스

뉴스 추천 서비스



뉴스 추천 내역 전송 ( 매일 update )



User의 Daily뉴스 선호도 조사



챗봇 서비스



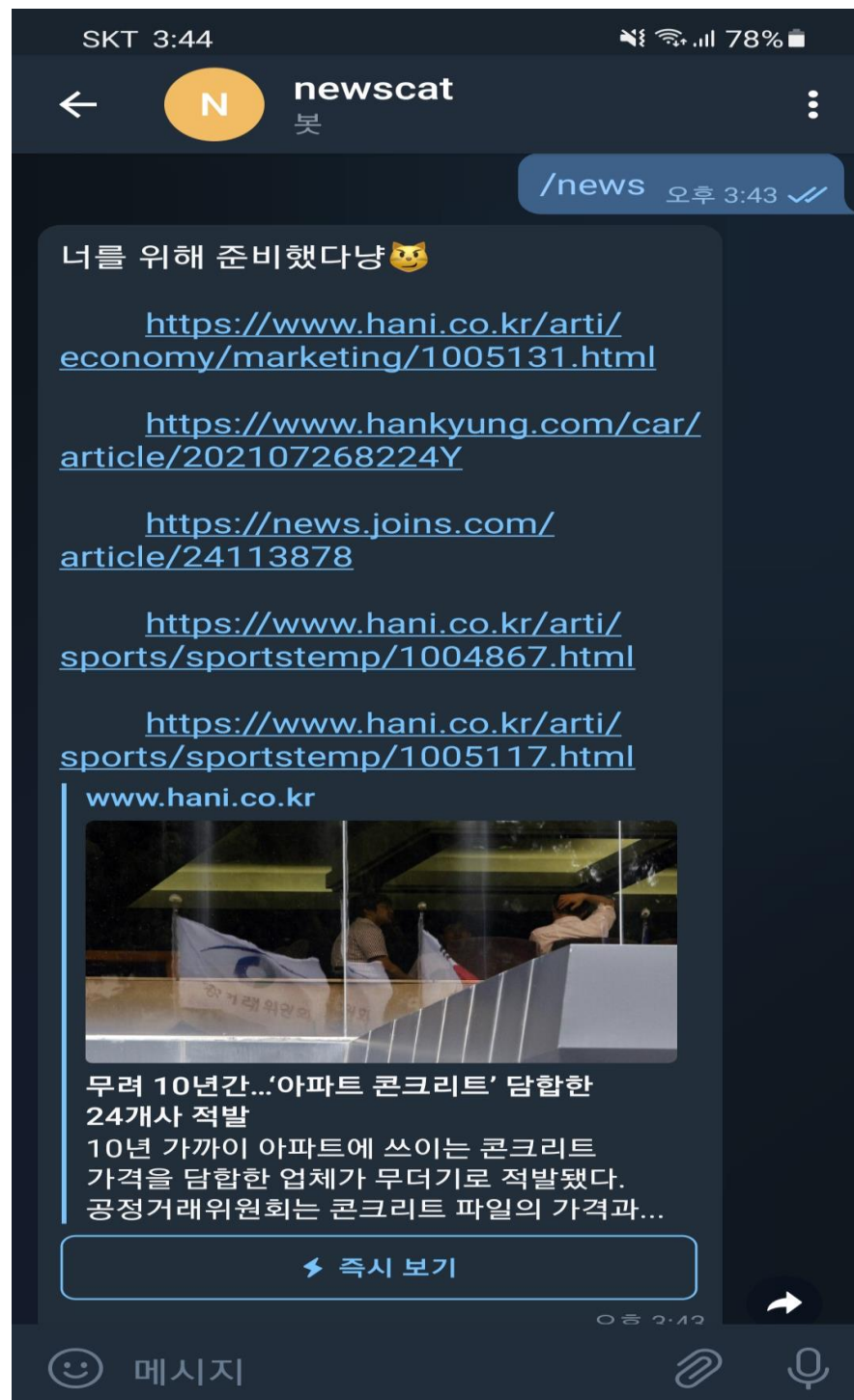
나 만 의 뉴 스 콘 텐 츠 큐 레 이 셴 서 비 스 - 뉴 스 캣 ( N E W S C A T )

# 02

## 서비스 설명



### 챗봇 형식의 구독형 뉴스 콘텐츠 큐레이션 서비스



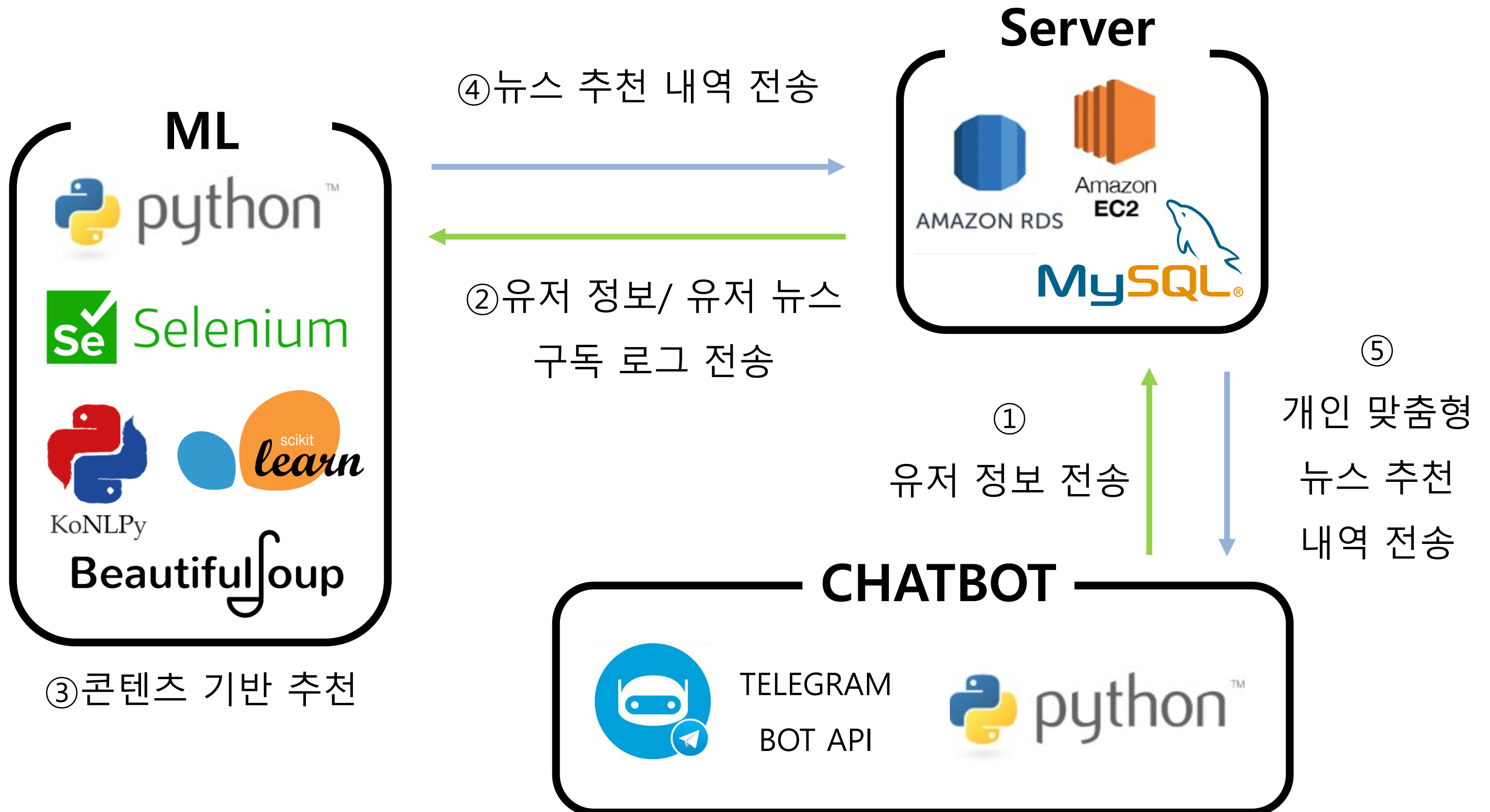
기존	VS	뉴스캣
단순 뉴스 클리핑		추천 알고리즘을 바탕으로 한 선호 기사 학습 고도화
개인을 구분하지 않는 일괄적 발행		해시태그 설정 등 <b>유저 선호 세부적으로 조절</b> 가능한 커스터마이징 서비스

- 일일 정기 뉴스 발행. **Daily Recommendation** (매일 9시)
- 요청으로 추가적인 뉴스 set 더 추천 받을 수 있음
- 추천 받은 뉴스 set에 대해 **피드백 시스템** 제공(Like, Dislike)
- 현재 5가지 카테고리 내 제공, 확장 예정  
: 정치, 경제, 사회, 국제, 스포츠
- **해시태그** 설정하여 세부 관심 항목 입력 받을 예정  
ex) #CBDC #비트코인 #부동산



# 03

## 사용 기술 스택





# 02

## 분석 과정

- 01. 계획 및 일정
- 02. 데이터 수집
- 03. 데이터 전처리
- 04. 모델링



# 01 계획 및 일정

## ▶ 분석 일정

데이터 수집 및 전처리 → 데이터 분석 및 추천 알고리즘 적용(모델링) → 서비스  
구현(챗봇/DB/서버)

스프린트 및 전체 과업 기록							
Aa 과업명	항목구분	담당자	과업 상태	오전/오후	진행기간	작업 기간	
보고 및 조사 작성	상위항목	한유진 <span>예진</span> 김예진 <span>현근</span> 송현근	완료	오후	07/22/2021 →	3일 소요 예정	07,
주제 선정 및 계획서 작성	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근 <span>최</span> 최윤석	완료	오후	07/22/2021 →	3일 소요 예정	07,
WBS 작성	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근	완료	오후		1일 소요 예정	까,
유저 스토리 작성	하위항목	<span>예진</span> 김예진 <span>현근</span> 송현근 <span>한유진</span> <span>최</span> 최윤석	완료	오후		1일 소요 예정	까,
시스템 아키텍처 구성	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근 <span>최</span> 최윤석	완료	오후		1일 소요 예정	까,
비즈니스 배경조사	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근 <span>최</span> 최윤석	완료	오전		1일 소요 예정	까,
포털 신문사 현황 조사	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근 <span>최</span> 최윤석	완료	오후		1일 소요 예정	까,
챗봇 사용 사례 분석	하위항목	<span>예진</span> 김예진 <span>한유진</span> <span>현근</span> 송현근 <span>최</span> 최윤석	완료	오전		1일 소요 예정	까,
						1일 소요 예정	까,
챗봇 공부	상위항목	<span>예진</span> 김예진 <span>최</span> 최윤석	완료	오전	07/23/2021	1일 소요 예정	07,
챗봇 api 조사	하위항목	<span>예진</span> 김예진 <span>최</span> 최윤석	완료	오전	07/23/2021	1일 소요 예정	07,
카톡 API 공부 & 계정 발급	하위항목	<span>예진</span> 김예진 <span>최</span> 최윤석	완료	오전	07/23/2021	1일 소요 예정	07,
						1일 소요 예정	까,
크롤링	상위항목	<span>한유진</span> <span>현근</span> 송현근	완료	오전	07/23/2021 →	2일 소요 예정	07,
뉴스 카테고리 나누기	하위항목	<span>한유진</span> <span>현근</span> 송현근	완료	오전	07/23/2021 →	2일 소요 예정	07,

김예진 6 days

유튜브 콘텐츠 시청

-4일 남음

시스템 아키텍처 구성

완료

1일 남음

추천 알고리즘

진행 예정

0일 남음

보고 및 조사 작성

완료

-3일 남음

챗봇 공부

완료

-4일 남음

서버 및 DB 생성

진행 예정

-2일 남음

한유진 6 day

EDA

완료

-1일 남음

시스템 아키텍처

완료

1일 남음

보고 및 조사 작성

완료

-3일 남음

크롤링

완료

-3일 남음

데이터전처리

진행 중

-3일 남음

주제 선정 및 계획

완료

-3일 남음

협업 툴 : notion을 활용한 프로젝트 상황 공유 및 진행

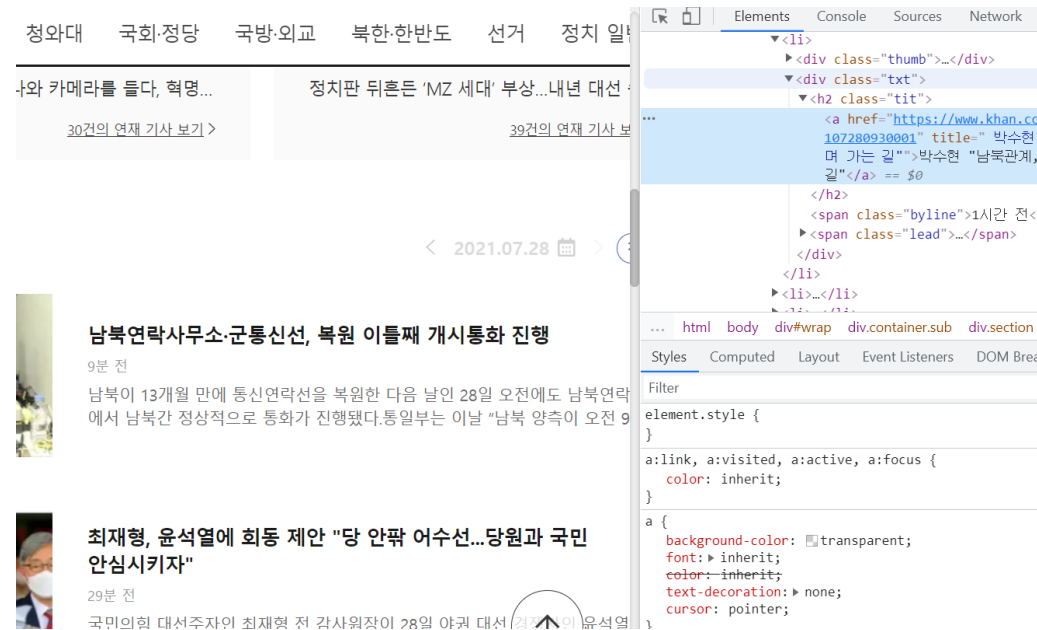


# 02

## 데이터 수집



### 언론사 사이트 일간 뉴스 크롤링 진행



- 6개 신문사 : 조선, 중앙, 동아, 경향, 한겨레, 한국경제
- 5개 카테고리 : 정치, 경제, 사회, 국제, 스포츠
- 5개 Column : headline, contents, url, category, name
- Request, Urllib, Selenium 활용
- 약 **3,000개** row의 DataFrame 생성

	headline	contents	url	category	name
2929	'이것이 패자의 품격' 은메달 따낸 이다빈, 금메달 축하하는 엄지척!	[도쿄올림픽]졌지만 '패자의 품격'이 빛났다. 태권도 '스마일 퀸' 이다빈이 생애 ...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선
2930	중주국 체면 구겼다.. 태권도, 첫 '노골드' 올림픽 '은1등2' 마감	이다빈(25, 서울시청)이 태권도 금메달 획득에 실패하면 서 올림픽 사상 처음으로 '...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선
2931	배드민턴 안세영, 조 1위로 16강 진출	[도쿄2020]배드민턴 유망주 안세영(19·삼성생명)이 자신 의 첫 올림픽에서 2연승...	https://www.chosun.com/sports/tokyo-2020/2021/...	스포츠	조선
2932	포수→외야수→투타겸업... '탁타니' 첫 선, 이유는 "1군 경기 절실해서"	[OSEN=부산, 조형래 기자] 현재 2군에서 타자로 성과를 내고 있었다. 포지션 ...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선
2933	한화 새 외인 타자 페레즈 28일 가족과 함께 입국, 충북 옥천서 2주 격리	[스포츠조선 박상경 기자] 한화 이글스 새 외국인 타자 에 르난 페레즈가 입국한다.페...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선

2934 rows × 5 columns

# 03

## 데이터 전처리



### 크롤링 RAW 데이터 전처리

#### ▶ DATA 전처리

- 결측치 : 본문 중 문자열 형식이 아니어서 추출되지 않거나(NaN) 본문 내용에 특정한 의미가 없는 기사 제거.

- 중복 제거 : 보도자료 형식으로 배포된 중복 기사

#### ▶ 형태소 분석 및 키워드 추출

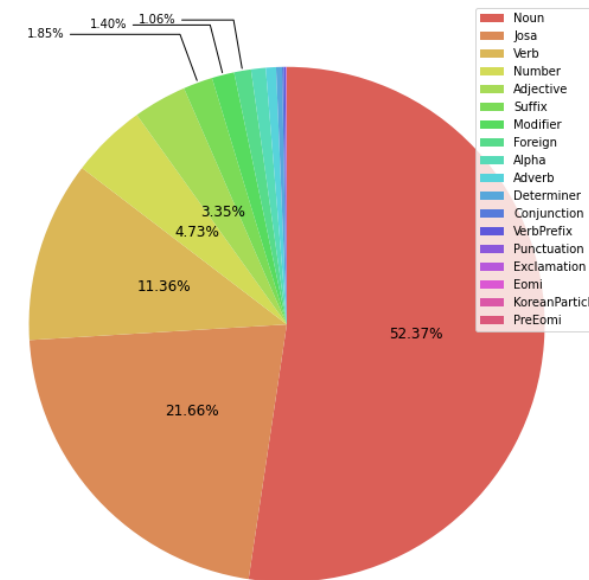
- Konlpy 활용, 형태소 구분에 따른 데이터 분할
- 의미적으로 활용할 수 있는 'Noun' 형태소 선택
- 핵심 단어 추출 : TF-IDF 활용하여 키워드

```
0 조해진 “김어준 ‘대법관 좌표찍기’ 전체주의적
120 문준용 “SNS 안할 순 없어...불편한 분들께는 죄
3 문준용 “SNS 안할 순 없어... 불편한 분들께는 죄
108 진 중권 “수 120시간’ 윤석열, 왜 성지식 오해 사는지...미
59 류호정 “이준석, 경쟁에 미쳐있는 것 같다
```

중복 기사 제거



본문 내용 없는 기사 제거



형태소 분포 파이차트

TF(d, t)  
특정 문서 d에서의 특정 단어 t의 등장 횟수

DF(t)  
특정 단어 t가 등장한 문서의 수

IDF(d, t)  
DF(t)에 반비례하는 수

$idf(d, t) = \log\left(\frac{n}{1 + df(t)}\right)$

$TF(d, t) * IDF(d, t) = TF-IDF(d, t)$

TF-IDF (단어-역문서 빈도) 자연어 처리

# 03

## 데이터 전처리



### 크롤링 RAW 데이터 전처리

#### DATA 전처리

결측치 : 본문 중 문자열 형식이 아니어서 추출되지

#### 최종 DATAFRAME 형태

	headline	contents	url	category	name	hot_word
2683	'이것이 패자의 품격' 은메달 따낸 이다빈, 금메달 축하하는 엄지척!	도쿄올림픽했지만 패자의 품격이 빛났다 태권도 스마일 퀸 이다빈이 생애 첫 올림픽에서...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선	[이다빈, 태권도, 은메달, 도쿄올림픽, 획득]
2684	중주국 체면 구겼다.. 태권도, 첫 '노골드' 올림픽 '은1등2' 마감	이다빈25 서울시청이 태권도 금메달 획득에 실패하면서 올림픽 사상 처음으로 노골드에...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선	[금메달, 이다빈, 태권도, 은메달, 동메달]
2685	배드민턴 안세영, 조 1위로 16강 진출	도쿄2020배드민턴 유망주 안세영19삼성생명이 자신의 첫 올림픽에서 2연승을 달리며...	https://www.chosun.com/sports/tokyo-2020/2021/...	스포츠	조선	[안세영, 복식, 리그, 김소영, 승찬]
2686	포수→외야수→투타겸업... '탁타니' 첫 선, 이유는 "1군 경기 절실해서"	OSEN부산 조형래 기자 현재 2군에서 타자로 성과를 내고 있었다 포지션 전향도 순...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선	[투수, 원탁, 포지션, 야수, 타석]
2687	한화 새 외인 타자 페레즈 28일 가축과 함께 입국, 충북 옥천서 2주 격리	스포츠조선 박상경 기자 한화 이글스 새 외국인 타자 에르난 페레즈가 입국한다페레즈는...	https://www.chosun.com/sports/sports_photo/202...	스포츠	조선	[페레즈, 한화, 포지션, 타자, 소화]

2688 rows × 6 columns

핵심 단어 추출 : TF-IDF 활용하여 키워드

형태소 분포 파이차트

$$idf(d, t) = \log\left(\frac{n}{1 + df(t)}\right)$$

TF-IDF (단어-역문서 빈도) 자연어 처리

서태



# 04

## 모델링

### 컨텐츠 기반 필터링, 추천 알고리즘 유사도 계산

1 컨텐츠 기반 모델

Represented Items  
Items을 벡터 형태로 표현. 도메인에 따라 다른 방법이 적용

item1  
·  
·  
·  
itemN

벡터1  
·  
·  
·  
벡터N

벡터들간의 유사도를 계산

1 유사도 함수

코사인 유사도  
- 문서간의 유사도를 계산

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

과일이 있고 노란 먹고 바나나 사과 싶은 저는 좋아요

	과일이	있고	노란	먹고	바나나	사과	싶은	저는	좋아요
문서1	0	0	0	1	0	1	1	0	0
문서2	0	0	0	1	1	0	1	0	0
문서3	0	1	1	0	2	0	0	0	0
문서4	1	0	0	0	0	0	0	1	1

[장점]  
- 벡터의 크기가 중요하지 않은 경우에 거리를 측정하기 위한 메트릭으로 사용. (예 : 문서내에서 단어의 빈도수 - 문서들의 길이가 고르지 않더라도 문서내에서 얼마나 나왔는지를 비율을 확인하기 때문에 상관없음.)

[단점]  
- 벡터의 크기가 중요한 경우에 대해서 잘 작동하지 않음

```
1 from sklearn.metrics.pairwise import cosine_similarity
2 cosine_similarity(countvect_df, countvect_df)
```

array([[1., 0.66666667, 0., 0.],  
 [0.66666667, 1., 0.47140452, 0.],  
 [0., 0.47140452, 1., 0.],  
 [0., 0., 0., 1.]])

기준행

0	1	7	2
1	1	2	4
2	0	8	3
3	2	0	3

cosine\_similarities(적용)

비교 대상 행

	0	1	2	3
0	1 (0행 자신의 유사도)	0.68 (0행과 1행의 유사도)	0.99 (0행과 2행의 유사도)	0.3 (0행과 3행의 유사도)
1	0.68 (1행과 0행의 유사도)	1 (1행 자신의 유사도)	0.72 (1행과 2행의 유사도)	0.85 (1행과 3행의 유사도)
2	0.99 (2행과 0행의 유사도)	0.72 (2행과 1행의 유사도)	1 (2행 자신의 유사도)	0.29 (2행과 3행의 유사도)
3	0.3 (3행과 0행의 유사도)	0.85 (3행과 1행의 유사도)	0.29 (3행과 2행의 유사도)	1 (3행 자신의 유사도)

- **컨텐츠 기반 필터링** : 아이템 고유의 정보 바탕으로 **아이템 간 유사성**을 파악해서 추천
- = **이후 word 토큰화, 임베딩, 유사도 계산**을 통해 **맞춤기사 3개**, 다양한 뉴스채널을 위한 **한 토픽 기사 2개** 제공
- **유저의 취향을 반영하면서 편협한 시야를 방지**하기 위해 랜덤 기사 제공 & 처음에는 유저 정보가 없는 "콜드 스타트" 해결 위해, 처음에는 유저가 선택한 카테고리 내에서 랜덤하게 5개 뉴스 제공



# 컨텐츠 기반 필터링, 추천 알고리즘 유사도 계산

Tacademy

## 2 Neighborhood based method

✓ 정의  
Neighborhood based Collaborative Filtering은 메모리 기반 알고리즘으로 협업 필터링을 위해 개발된 초기 알고리즘입니다.

✓ 알고리즘

1. User-based collaborative filtering  
사용자의 구매 패턴(평점)과 유사한 사용자를 찾아서 추천 리스트 생성
2. Item-based collaborative filtering  
특정 사용자가 준 점수간의 유사한 상품을 찾아서 추천 리스트 생성

- **협업 필터링** : 다수의 **사용자의 간 유사성(로그)**을 파악하여 좋아할만한 항목을 추천
- 유저 log 누적에 따른 **하이브리드 필터링** 방법 확장 예정



# 03

## 서비스 구현

01. 서버 및 DB

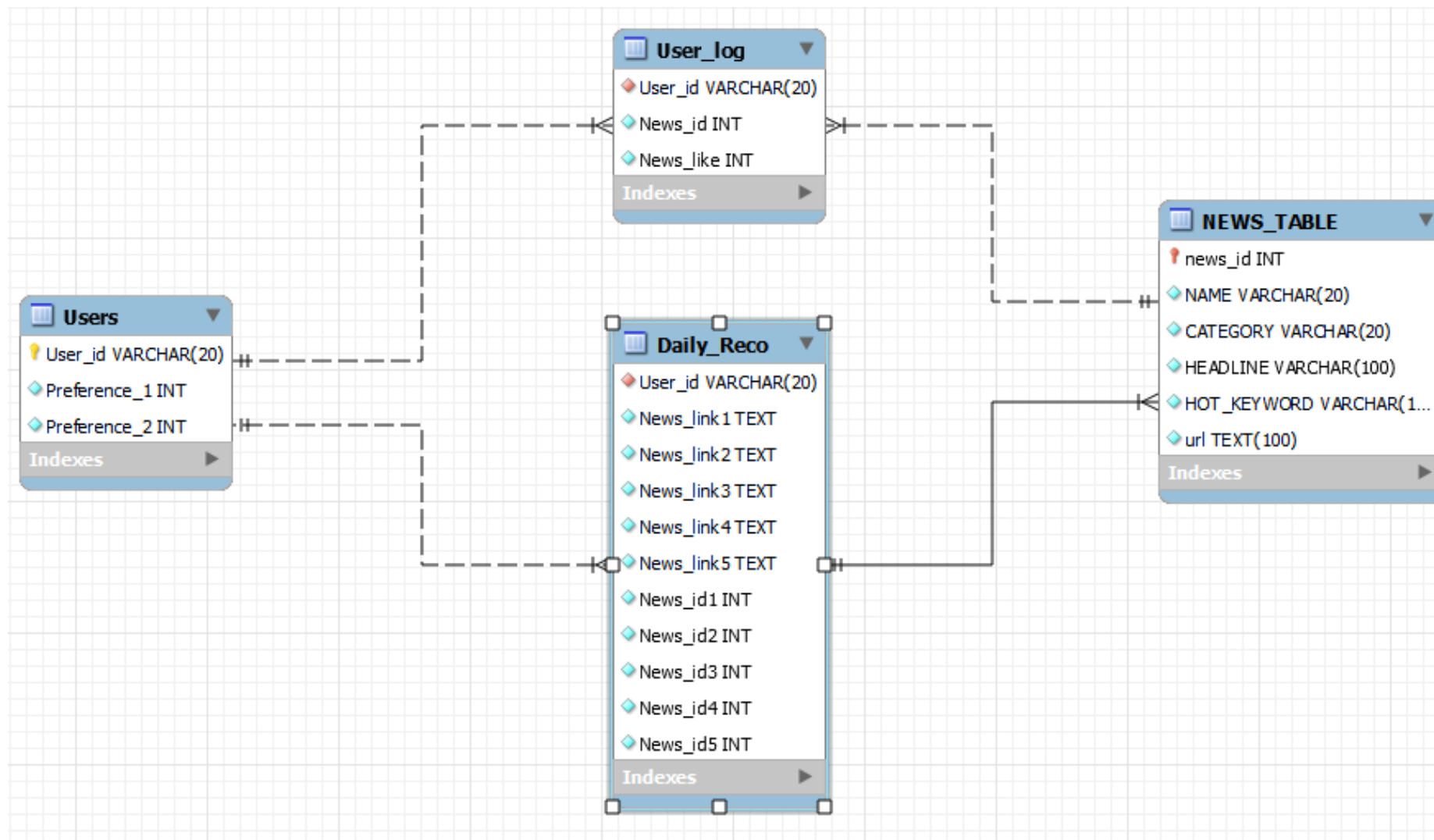
02. 챗봇



# 01

# 서버/ DB

## DB TABLE 설명 및 Key 관계



### [USER LOG 無]

랜덤한 5개 뉴스 제공,  
이후 해당 유저 log에 관심사  
및 피드백 업데이트

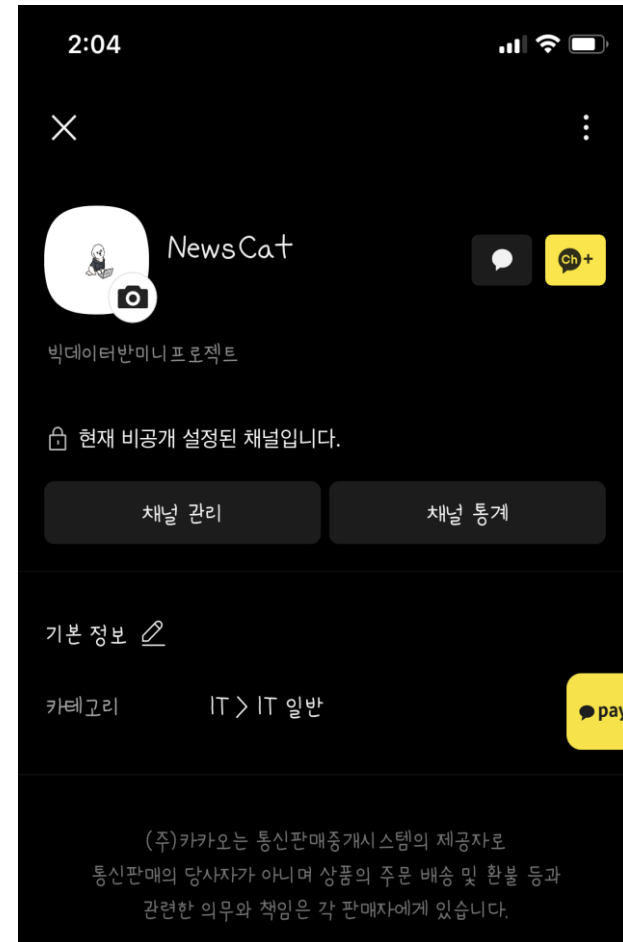
### [USER LOG 有]

유저 취향 학습하여 맞춤형 고도화

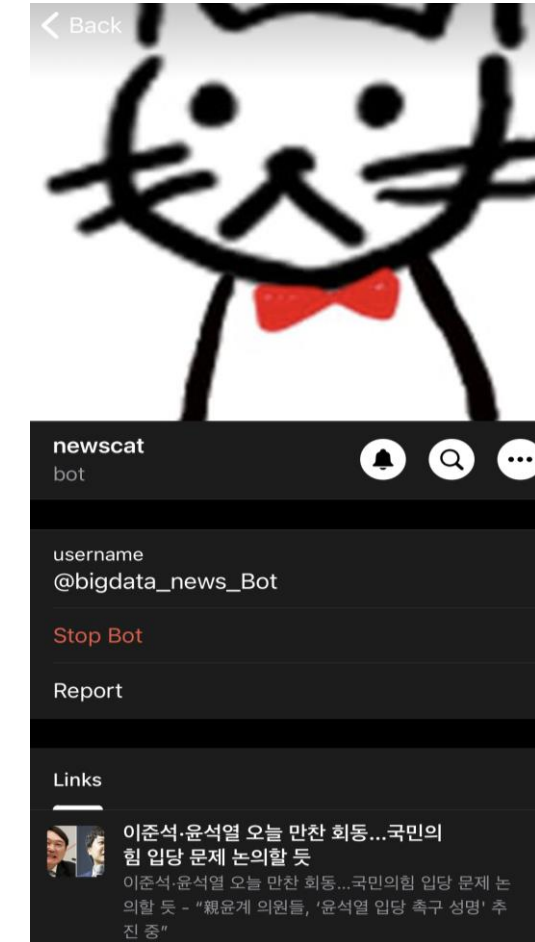
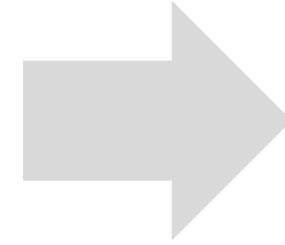
# 02

## 챗봇

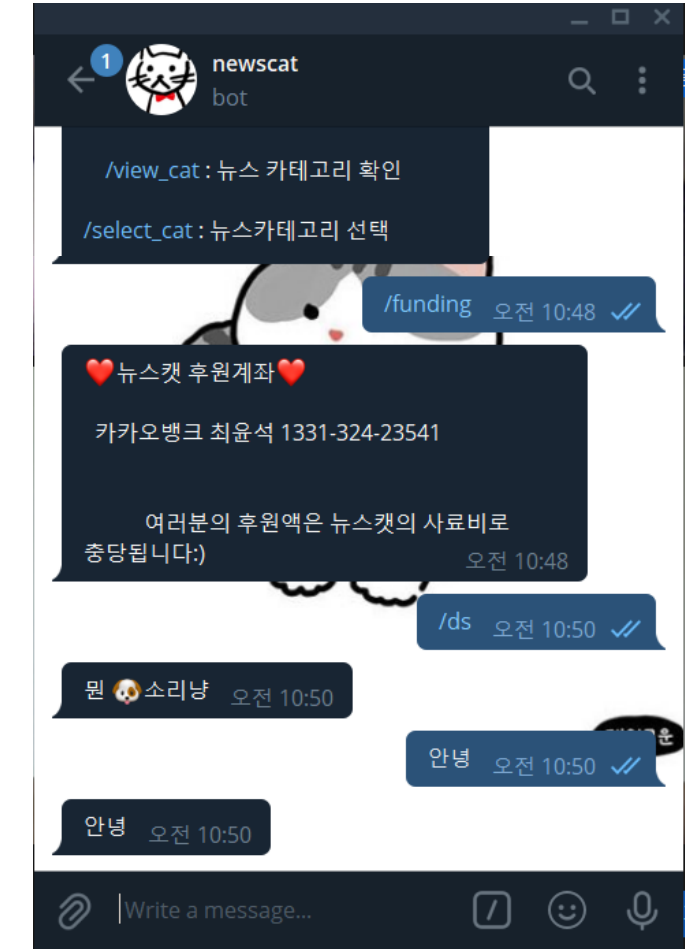
### 챗봇 소개 및 기능



카카오톡 챗봇



텔레그램 챗봇



- 시간 내 카카오톡 챗봇 API 권한 획득 불가 → 텔레그램 챗봇 API 활용
- **CUI 커맨드** 명령어 구현 → 추후 GUI 커맨드 추가 예정
- MZ세대들의 성향을 반영한 **B급의 키치한 뉴스 캣(category-cat) 컨셉**.  
매번 다른 인사말 제공 등으로 캐릭터와 대화하는 느낌 전달하며 소비자 경험 증대



# 02

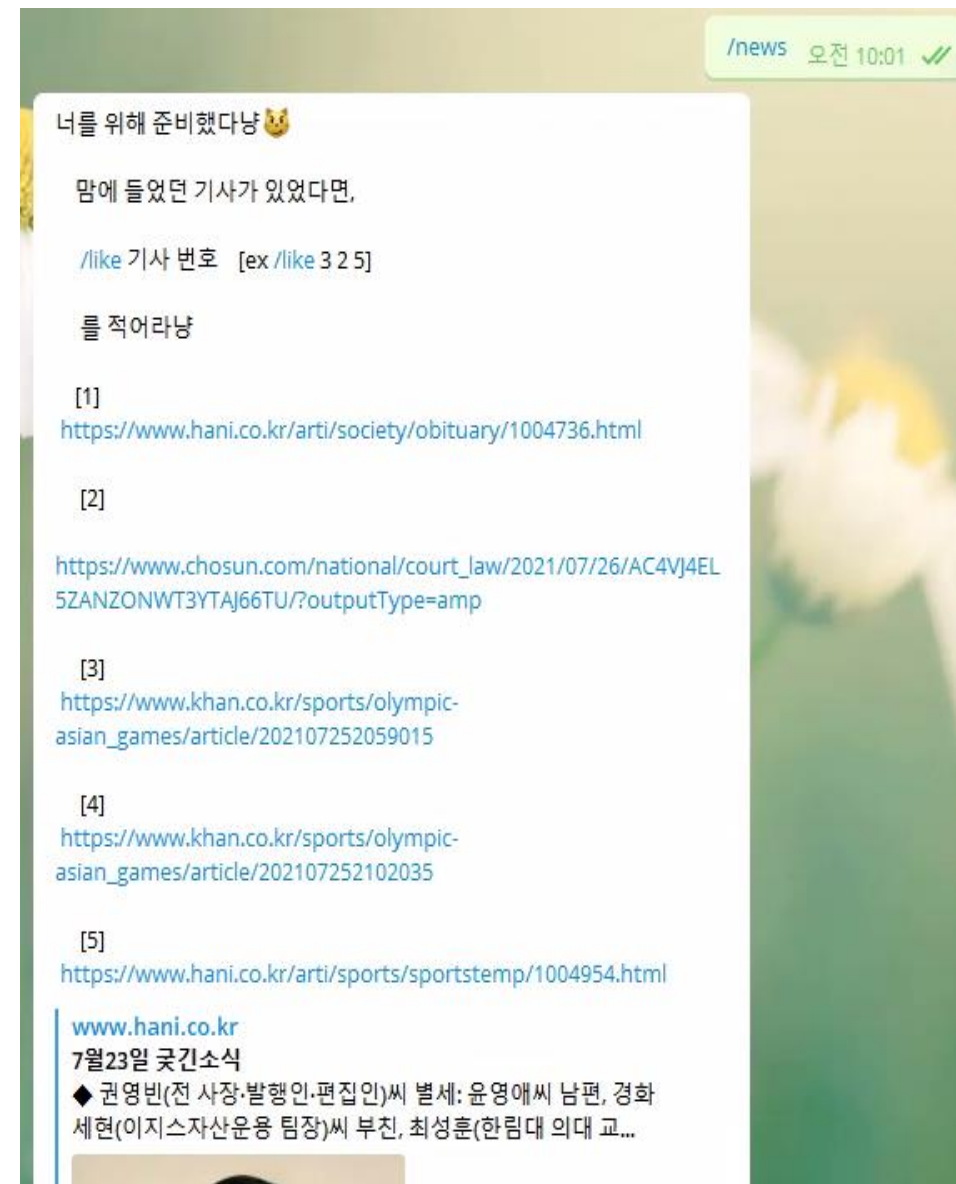
## 챗봇



## 이용 방법



카테고리 선택



추천 뉴스 set 요청



뉴스 피드백(만족도 반영)



04



서비스 시연



05





향 후  
방향

계 획    및    발 전

# 01

## 발전 방향

 : 유저 수 증가 & 유저 피드백 log 누적에 따른 콘텐츠 및 협업 필터링 **학습 고도화** → 추천 정확도 향상

 : 대용량의 뉴스 Data 누적 및 활용하도록 서버/DB 고도화 → 정확도 향상, 실시간 크롤링 확장

 : **실시간 크롤링** → 기사가 업로드 되는 즉시의 가장 최신 뉴스 제공

 : 추천 시 상위 유사도 n개의 순위가 아닌, 특정 점수 이상으로 범위 지정 →

요청 시 마다 매번 다른, 다양한 set의 뉴스 콘텐츠 제공 가능

 : 성능 지표를 활용한 모델링 검정 적용 → 추천 정확도 향상

 : 파일럿 형식의 **실제 서비스**를 단기간 운영하며 추천 시스템 고도화 예정

카테고리 및 해시태그 활용, **특정 분야 전문 뉴스 큐레이션** 서비스 발전 가능

ex) 해외 축구 이슈 큐레이션 서비스

나 만 의 뉴 스 콘 텐 츠 큐 레 이 셴 서 비 스 - 뉴 스 캣 ( N E W S C A T )

---

# 감사합니다

---

