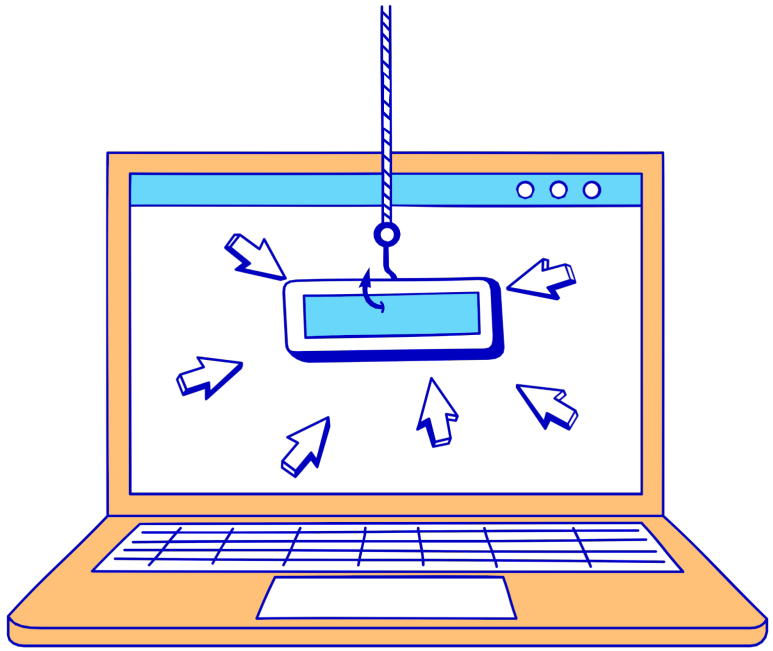
딥러닝 모델을 통한 클릭베이트 기사 감지



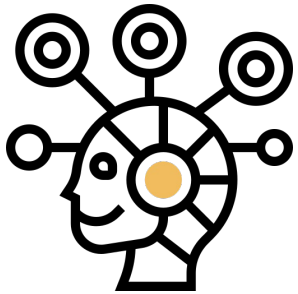
1조- 데이터구조

김예담 김지예 김현민 우형석 이한

multicampus

목차

1. 팀구성



김현민

딥러닝 모델 개발

김예담, 김지예

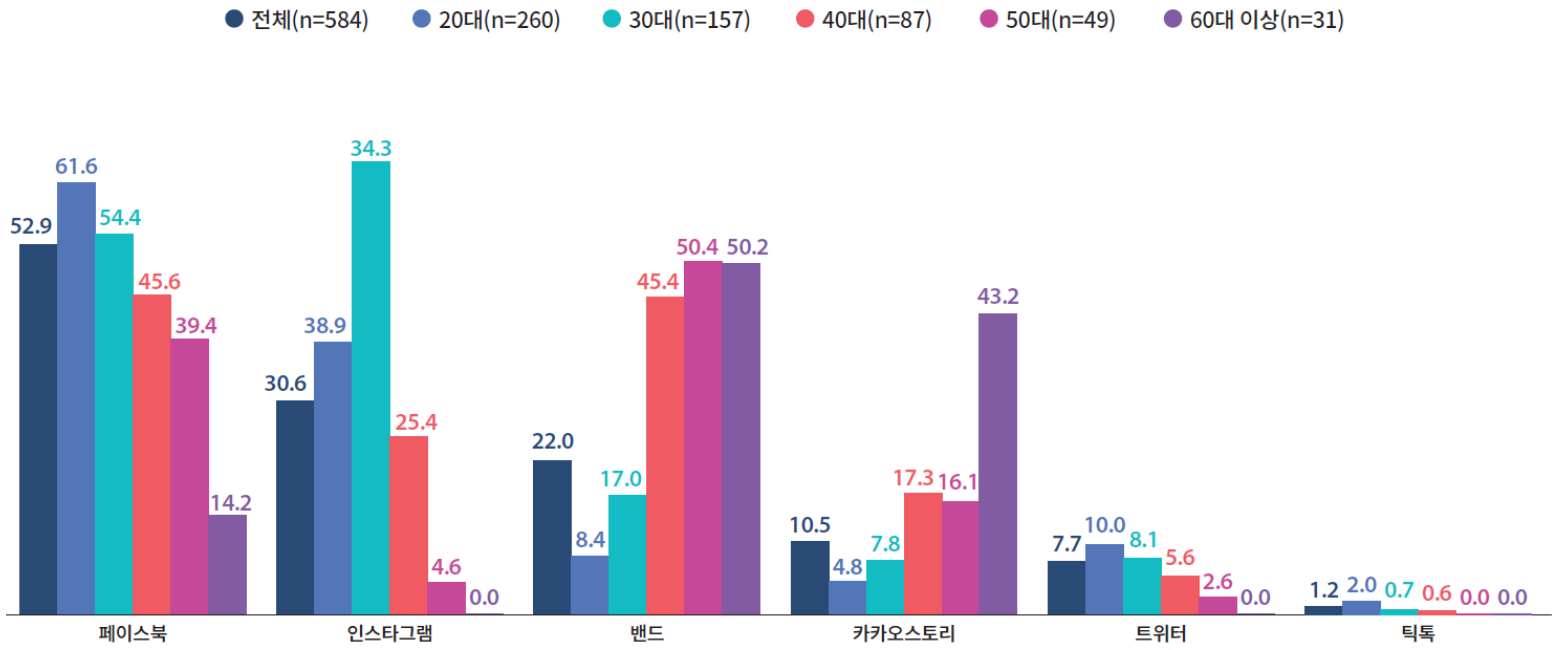
환경구축
워드클라우드
Crontab
DB기획

우형석, 이한

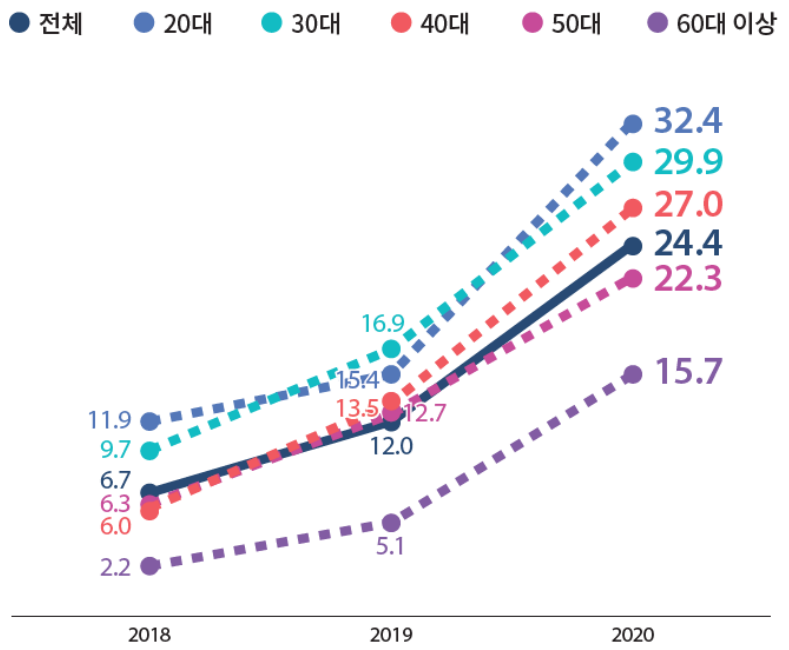
크롤링

프론트

SNS 뉴스 이용자의 SNS별 뉴스 이용률 (단위: %)



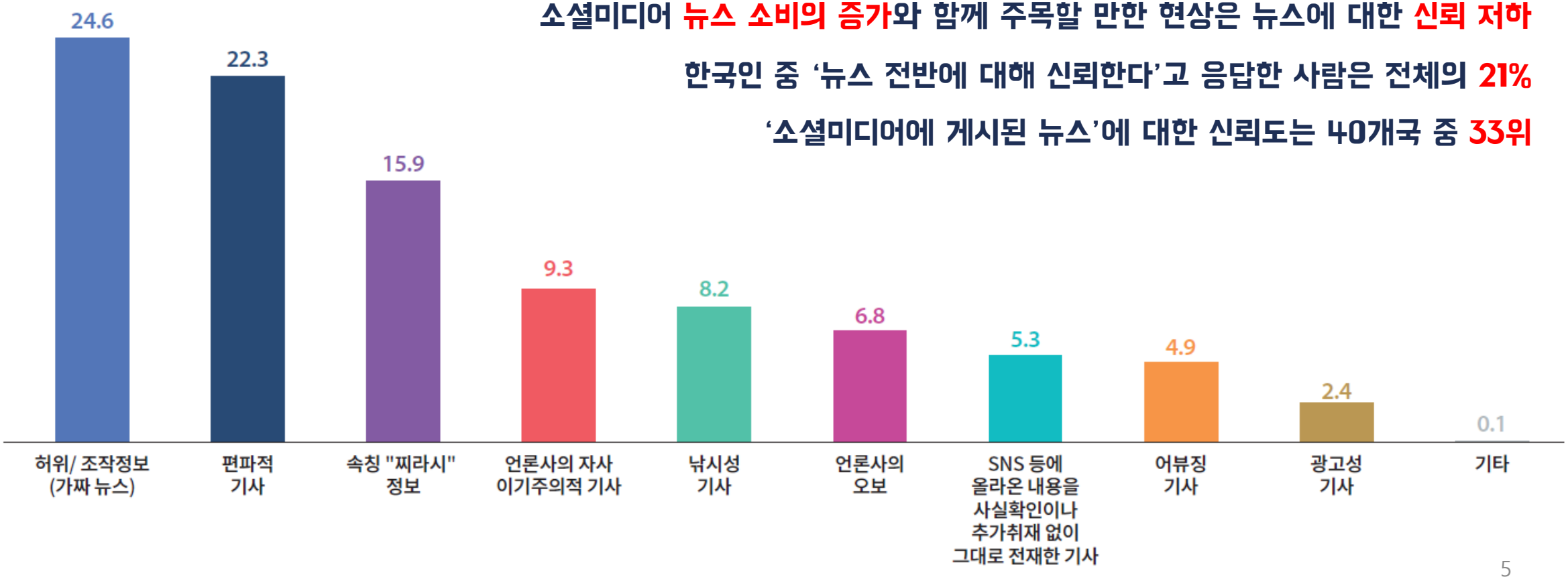
연령대별 온라인 동영상 플랫폼을 통한 뉴스 이용률 추이(2018~ 2020년) (단위: %)



한국언론진흥재단의 <2020년 언론수용자 조사>

- 2020년 유튜브 등 온라인 동영상 플랫폼을 통한 뉴스 이용률은 **24.4%**로 최고치를 기록
- 페이스북을 비롯한 SNS, 메신저, 유튜브를 통한 뉴스 소비가 보편적으로 자리 잡아가는 상태.

한국 언론의 가장 큰 문제점 (단위: %)



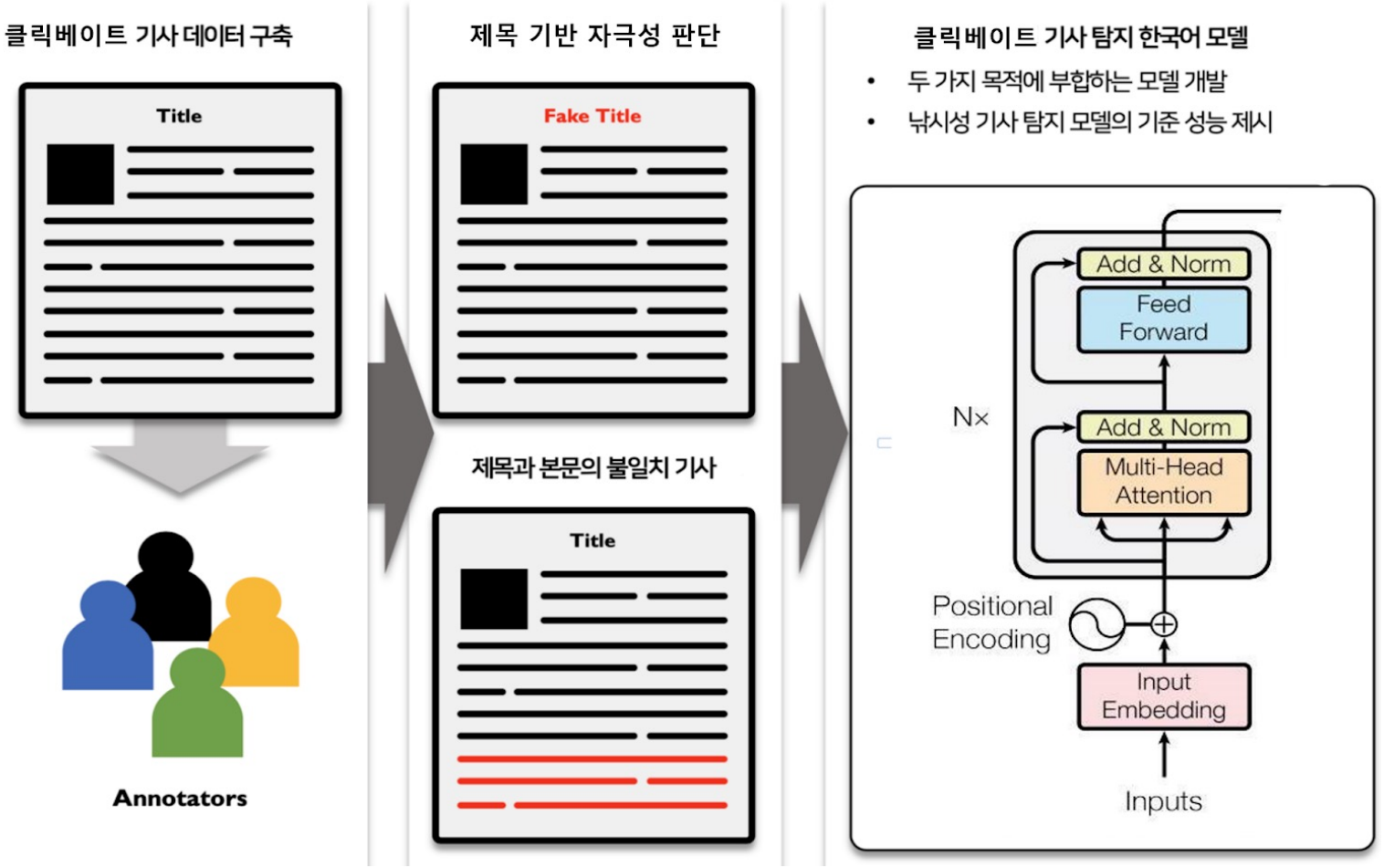
소셜미디어 뉴스 소비의 증가와 함께 주목할 만한 현상은 뉴스에 대한 신뢰 저하

한국인 중 '뉴스 전반에 대해 신뢰한다'고 응답한 사람은 전체의 21%

'소셜미디어에 게시된 뉴스'에 대한 신뢰도는 40개국 중 33위

<디지털 뉴스 리포트 2020>

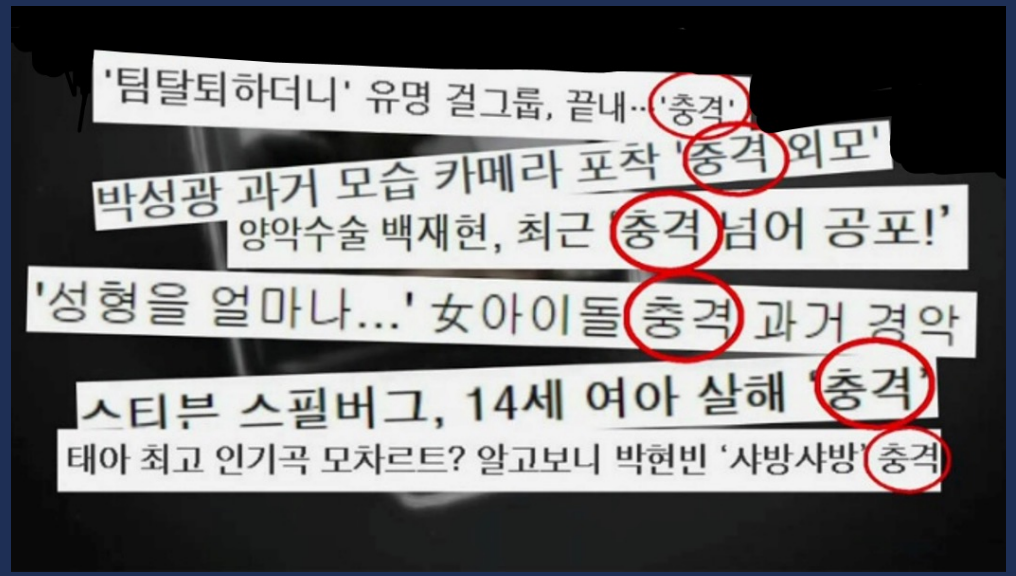
클릭베이트 기사 탐지용 AI 모델 개발



- 독자들에게 뉴스 제목과 본문의 일관성, 제목의 낚시성 등을 예측하고 점수를 제시하는 **클릭베이트 기사 탐지 시스템** 및 웹사이트 구현

- 이러한 시스템은 올바른 뉴스 기사를 선택할 수 있도록 하여 바람직한 뉴스 생태계를 만드는데 기여

클릭베이트(ClickBait)란?



- **클릭베이트**는 '**클릭(Click)**'과 '**미끼(Bait)**'의 합성어

- 국내 클릭베이트는 소위 말하는 '**낚시 기사**'나 '**쓰레기 기사**'의 형태로 널리 알려져 있다.

- '**헉**', '**충격**', '**경악**' , '**역대급**' 등의 지나치게 **감정적**인 표현이나 '숨 막히는 뒤태' 같은 **선정적**인 표현을 동반

-

DATA

모델 학습용 기사 데이터

	title	content	label
0	전주시, 부동산 불법행위 신고 1000만원 포상	전북 전주시는 단속 사각지대에서 음성적으로 이뤄지는 부동산 거래 불법행위 근절을 위...	1
1	한기영 서울시의원, '서울청년 선거대책본부' 구성	한기영 서울시의원이 오는 4월 7일 치러지는 보궐선거에 청년들의 투표 독려를 위해 ...	1
2	대구시·시장 평가 급상승	지난달 대구시의 주민생활만족도와 시장에 대한 평가가 급상승 한 것으로 나타났다 대구...	1
3	北 영변 핵시설서 의문의 연기... "재가동 징후"	북한이 영변 핵시설 단지의 일부 시설을 재가동했다는 주장이 나왔다 단지 내 방사화학...	1
4	"몸 상태 최악..." 19살에 암 투병 고백한 '스걸파' 클루씨 리더 이채린 근황	지난 1월 종영한 스트릿댄스 걸스 파이터에 댄스 크루 클루씨 리더로 출연한 이채...	0

데이터 출처

a. AI Hub 뉴스 기사 기계독해 데이터 (약 4만개)

b. 인터넷 언론사 크롤링 데이터 (약 2만개)

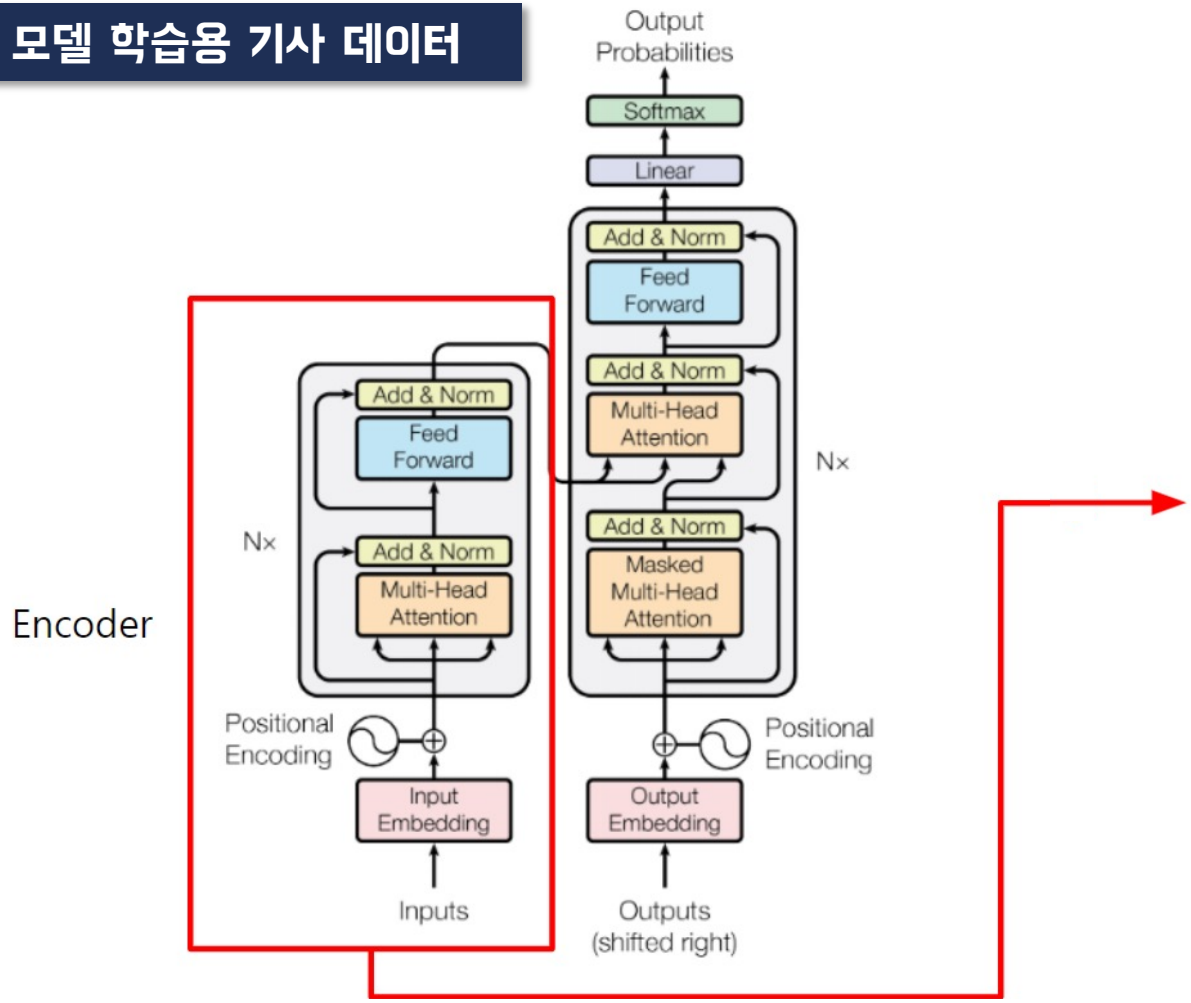
데이터 형태: 제목(title), 내용(content), 라벨

전처리: 중복제거, 특수문자 제거 등

최종 학습 데이터 약 **5만개**

Transformer and BERT

모델 학습용 기사 데이터



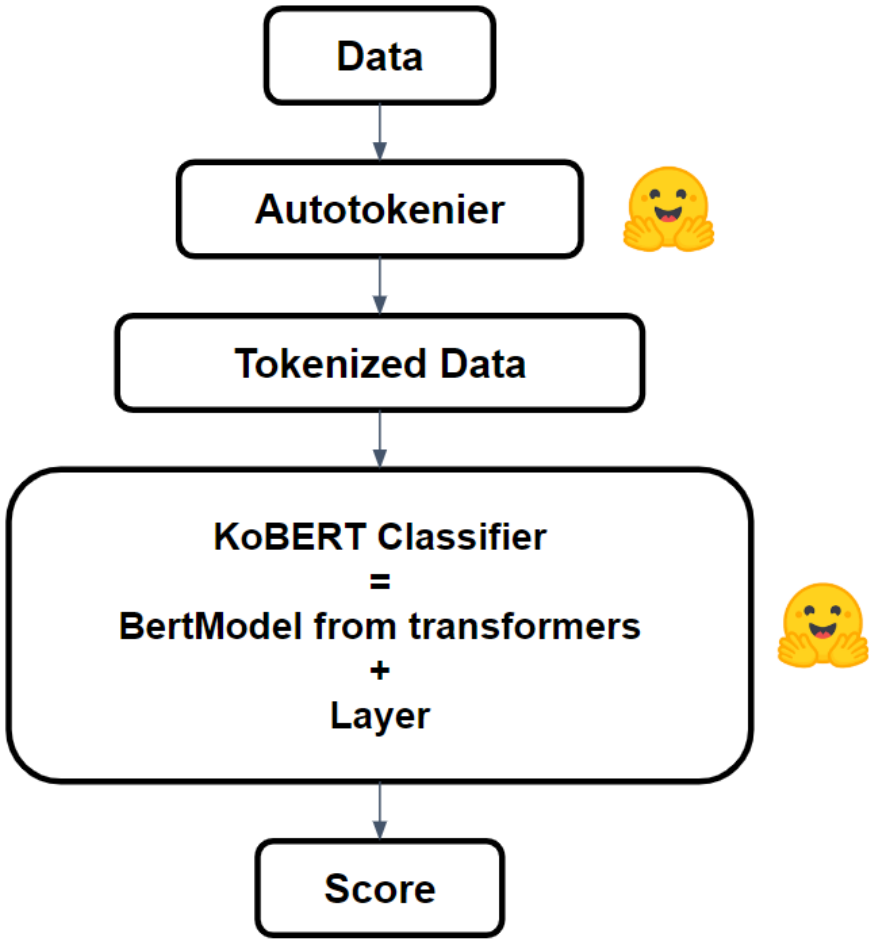
Transformer

- Long-term Dependency 문제 해결

BERT

- Transformer의 Encoder 구조 사용
- 사전 학습 모델
- 양방향 모델을 적용하여
- 문장의 앞과 뒤의 문맥을 고려

KoBERT Classifier



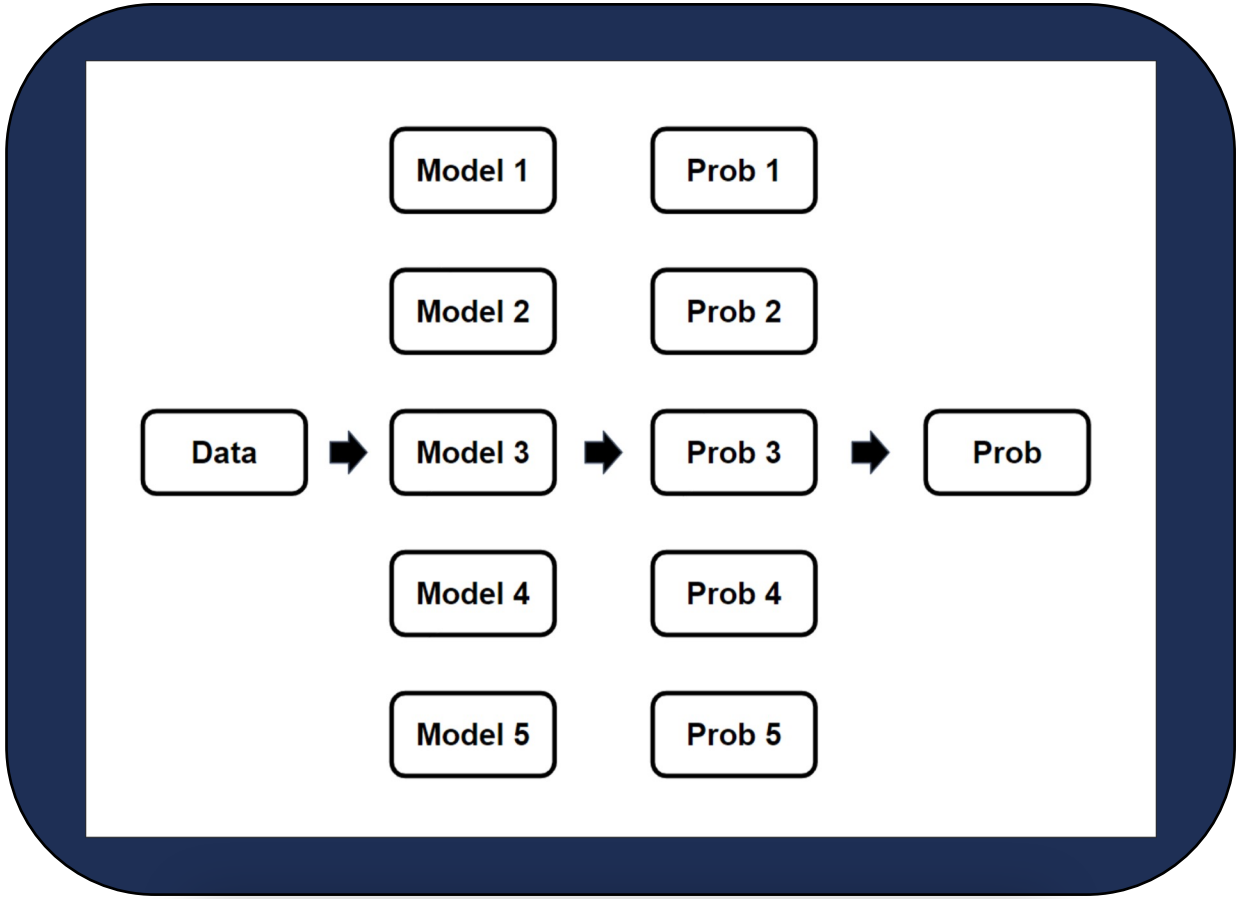
- 데이터 -> 사전학습 Tokenizer -> Bert Input (Tensor 형태)
- 사전학습 KoBERT Model
- BertModel의 Output이 786이기 때문에 추가 Layer 사용, Binary Classifier 정의

Training

- Stratified K-Fold : 5 Fold
- 10 Epoch
- Early stopping
- Optimizer : AdamW
- Scheduler : Linear
- Loss : Label Smoothing CrossEntrophy Loss
- Score : F1 Score

Predict Score

모델 학습용 기사 데이터

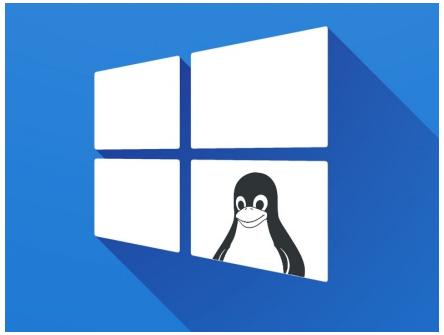


점수 산출 방식

- 정상 기사를 1, 비정상 기사를 0
- 학습된 모델을 사용한 예측 결과가 0일 확률 = 해당 기사가 클릭베이트 기사일 확률

점수는 5개 모델의 **예측값 평균** 사용 (Soft Voting)

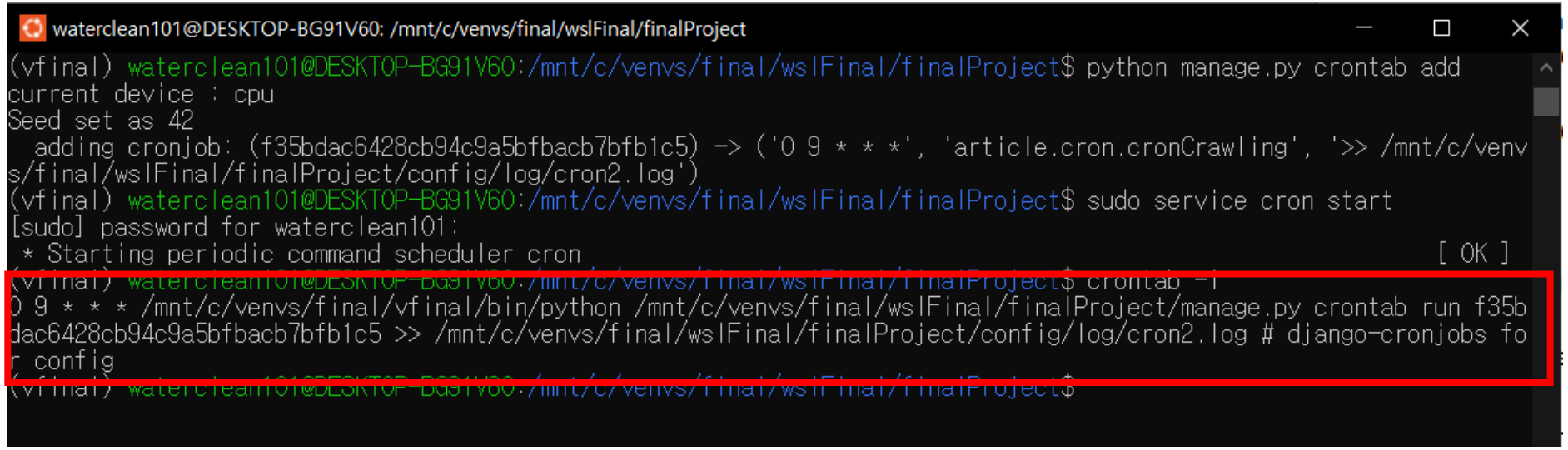
환경구축




Client
Web Server
WSGI
Application Server
Request
Request
Request
NGINX
gunicorn
django
SQLite
Static(CSS, JS, Image ETC) Return

작업 스케줄링(Crontab)

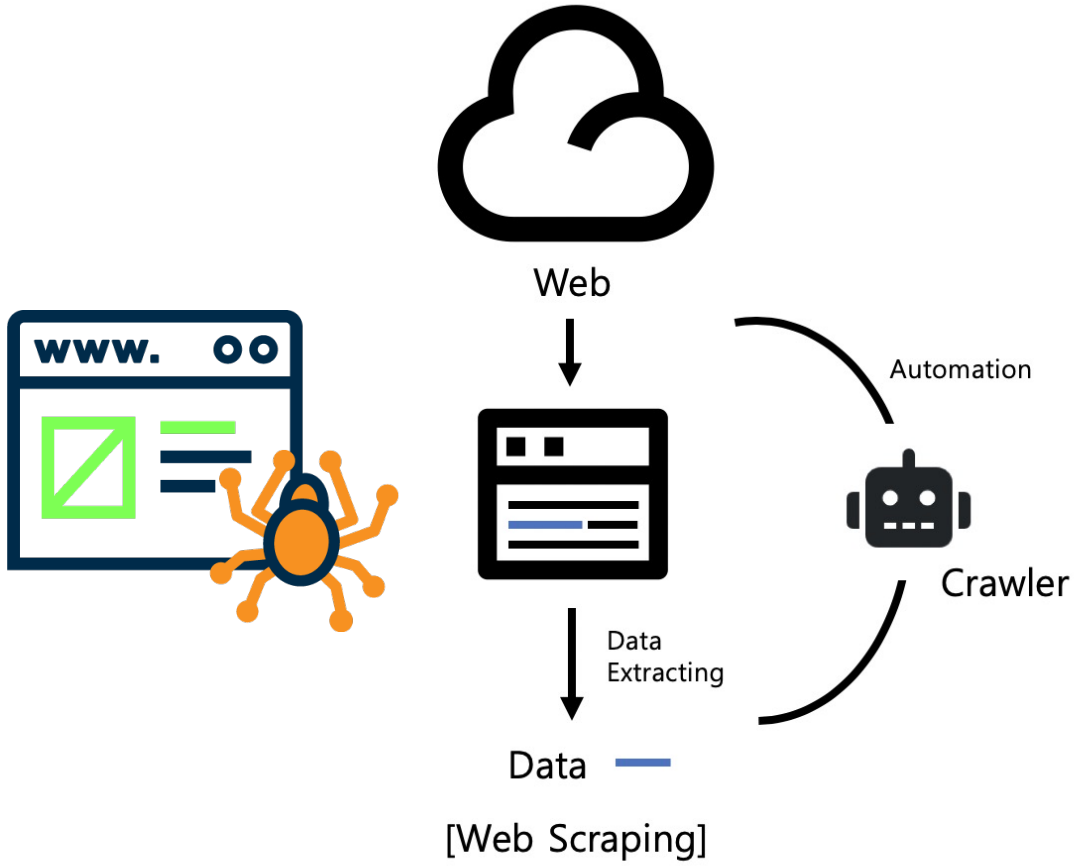
```
# CRONTAB_DJANGO_SETTINGS_MODULE= 'finalProject.settings'

CRONJOBS = [
    ('0 9 * * *', 'article.cron.cronCrawling', '>> '+os.path.join(BASE_DIR, 'config/log/cron2.log')),
]
```



워드클라우드란?

- 자료의 빈도 시각화
- 빅데이터(big data)를 분석할 때 데이터의 특징을 도출하기 위해 활용
- KoNLPy(한국어 형태소 파서 라이브러리)사용

Web Crawling



목표

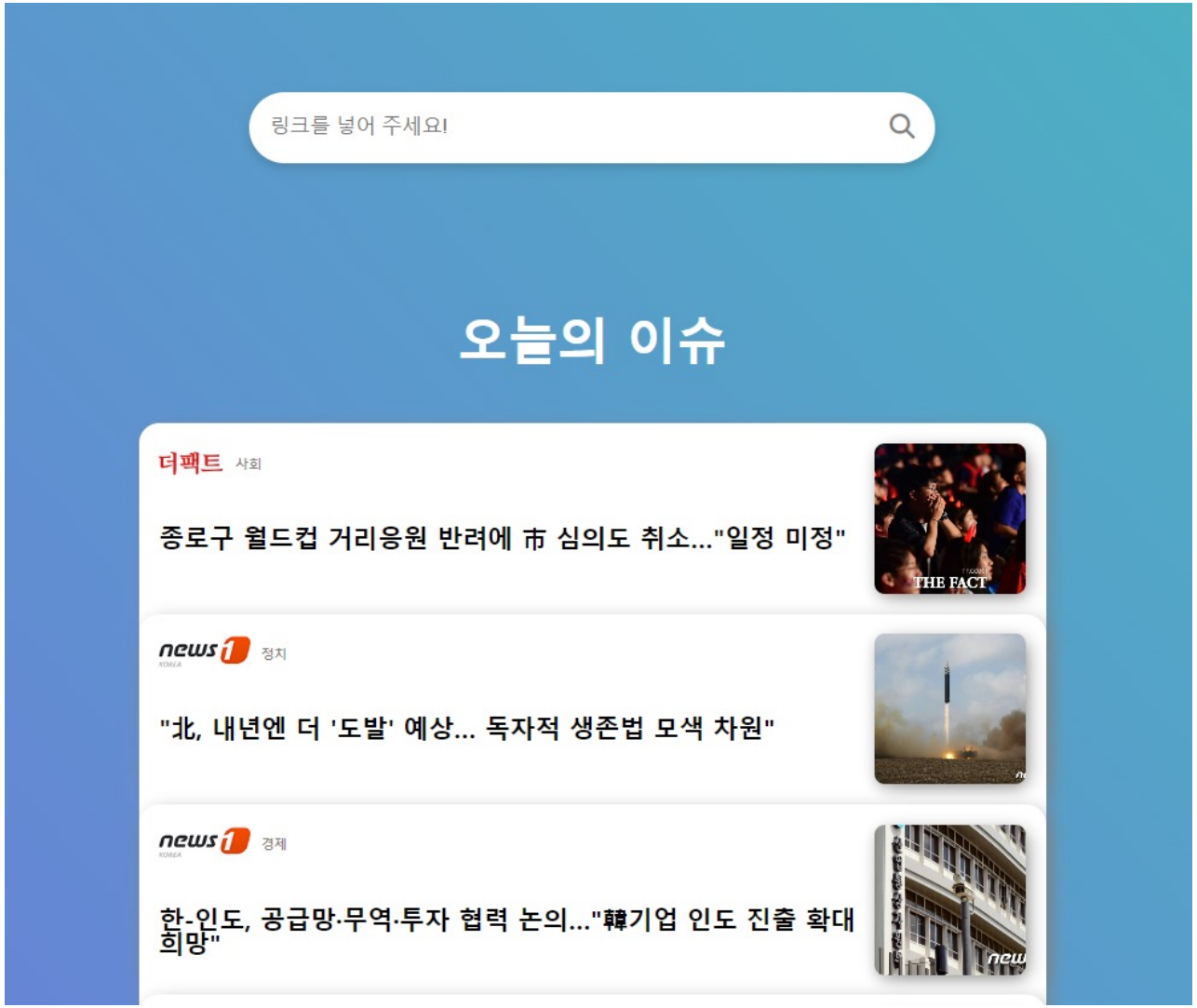
- 딥러닝 학습과 웹페이지의 결과물 출력을 위한 뉴스 기사 내용과 제목 추출
- 백엔드, 프론트엔드, 딥러닝에 필요한 모든 요소 수집

Selenium & BeautifulSoup

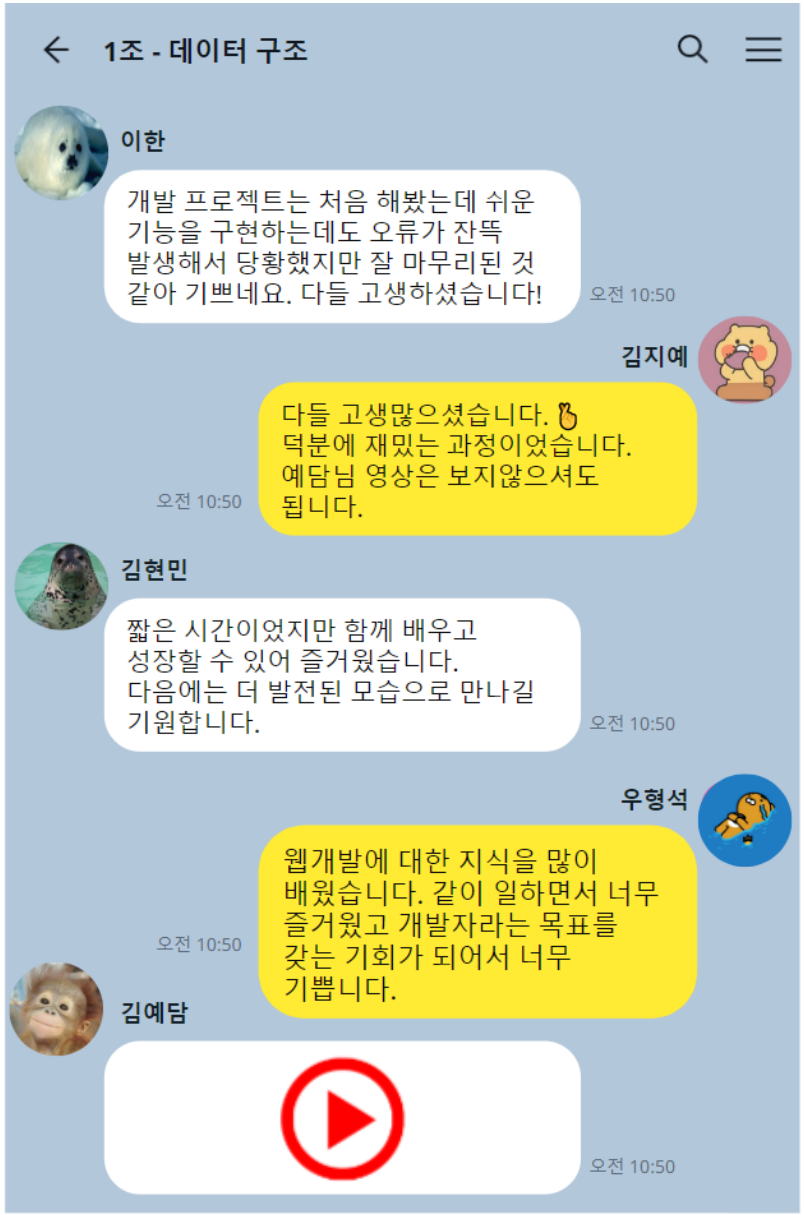
- 자동화된 웹 테스트 프레임워크
- 웹과 동작 기능 자동 실행
- 작동시간 김

- HTML parser
- HTML 요소 추출
- 작동시간 빠름

3. 결과물 시현


링크를 넣어 주세요!
오늘의 이슈
더팩트 사회
종로구 월드컵 거리응원 반려에 市 심의도 취소..."일정 미정"
THE FACT
news1 정치
"北, 내년엔 더 '도발' 예상... 독자적 생존법 모색 차원"
news1 경제
한-인도, 공급망·무역·투자 협력 논의..."韓기업 인도 진출 확대 희망"

3. 소감


1조 - 데이터 구조
이한
개발 프로젝트는 처음 해봤는데 쉬운 기능을 구현하는데도 오류가 잔뜩 발생해서 당황했지만 잘 마무리된 것 같아 기쁘네요. 다들 고생하셨습니다!
오전 10:50
김지예
다들 고생많으셨습니다.
덕분에 재밌는 과정이었습니다.
예담님 영상은 보지않으셔도 됩니다.
오전 10:50
김현민
짧은 시간이었지만 함께 배우고 성장할 수 있어 즐거웠습니다.
다음에는 더 발전된 모습으로 만나길 기원합니다.
오전 10:50
우형석
웹개발에 대한 지식을 많이 배웠습니다. 같이 일하면서 너무 즐거웠고 개발자라는 목표를 갖는 기회가 되어서 너무 기쁩니다.
오전 10:50
김예담
오전 10:50

공식문서

- Transformers Documentation (https://huggingface.co/docs/transformers/main/en/index)
- Google BERT - Pre Training and Fine Tuning for NLP Tasks (https://ranko-mosic.medium.com/googles-bert-nlp-5b2bb1236d78)
- Django Documentation

논문

- 2020언론수용자조사, 한국언론진흥재단, 2020.12.15
- 박아란, 이소은, “디지털 뉴스 리포트 : 한국 2020”, 한국언론진흥재단, 2020
- 김태균, 박건우, 차미영, “뉴스 기사의 일관성 탐지를 위한 딥러닝 시스템”, 한국컴퓨터종합학술대회 논문집, 2018, 2201-2203
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention Is All You Need”, 2017

Q & A