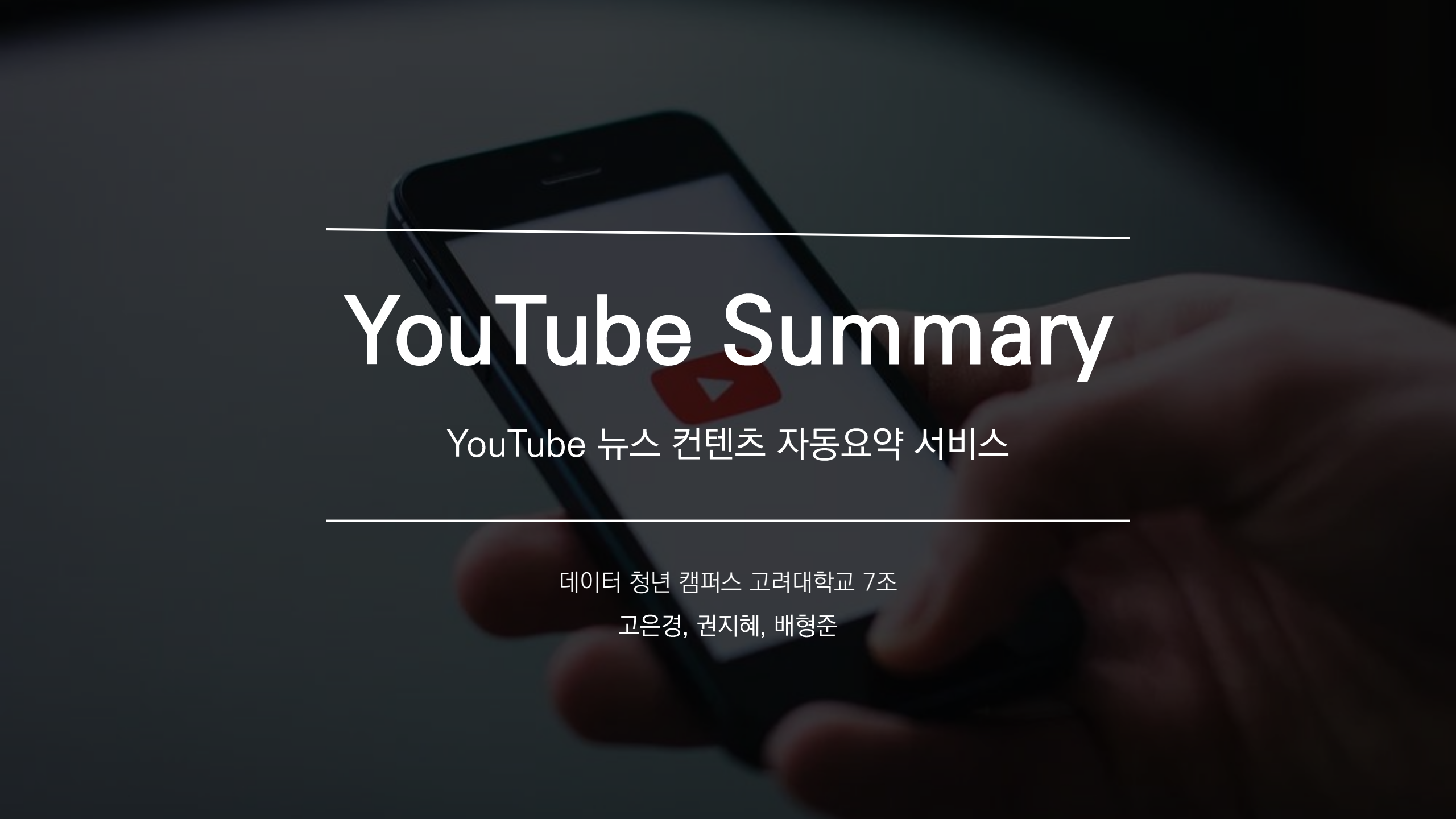


YouTube Summary

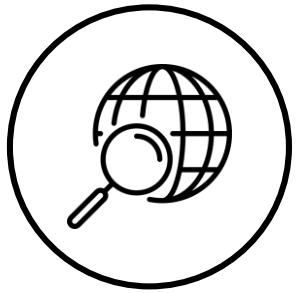
YouTube 뉴스 컨텐츠 자동요약 서비스

데이터 청년 캠퍼스 고려대학교 7조

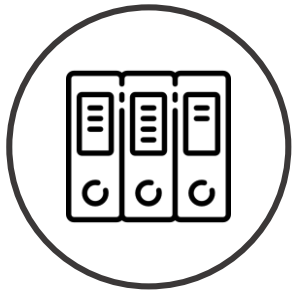
고은경, 권지혜, 배형준

Contents

배경



활용데이터 정의



데이터 처리방안
& 분석기법

분석결과

활용방안 및 기대효과

Contents

- YouTube의 성장
- 서비스의 필요성

배경

YouTube의 성장

SNS를 통해 뉴스를 접하는 사람들이 점점 많아지고 있다.

[표 3] 언론사 페북·트위터·인스타·유튜브 메인 뉴스계정 구독자 수 총합

순위	언론사별 메인뉴스 계정	구독자 수 합계	순위	언론사별 메인뉴스 계정	구독자 수 합계
1	JTBC뉴스	255만	2	YTN	198만 7,000
3	SBS뉴스	191만 7,000	4	KBS뉴스	127만 8,000
5	경향신문	113만 4,000	6	중앙일보	95만
7	한겨레신문	92만	8	조선일보	87만 8,000
9	MBC News	78만 6,000	10	연합뉴스	67만 5,000
11	한국일보	51만 7,000	12	MBN	45만 6,000
13	TV조선	32만 2,000	14	채널A	31만 5,000
15	세계일보	30만 2,000	16	매일경제	29만 1,000
17	연합뉴스TV	28만 9,000	18	동아일보	25만 1,000
19	서울신문	23만 5,000	20	한국경제신문	17만 9,000
21	한국경제TV	15만 1,000	22	국민일보	13만 4,000

출처: 한국언론진흥재단 신문과방송 2019년 4월호

2019년 1월 기준

주요 22개 언론사

메인 뉴스 계정 SNS (Facebook, Twitter, Instagram, YouTube)

구독자 수 총합 **약 1600만 명**

배경

YouTube의 성장

특히 유튜브는 빠르게 성장하고 있는 플랫폼으로, 타 플랫폼에 비해 사용 시간도 많다.

출처: WISEAPP 안드로이드 앱 사용시간 통계

배경

YouTube의 성장

홈
인기
구독
보관함

From 영국 로이터저널리즘연구소
‘디지털 뉴스 리포트 2020’

국내 응답자 중

SNS로 뉴스 접하는 비율

44%

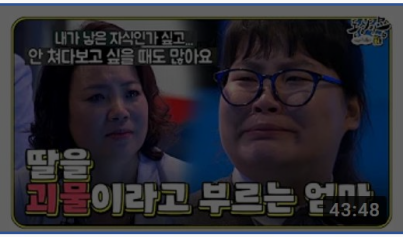
(복수응답)

SNS를 통한 뉴스 이용 매체 중
YouTube 사용

45%

"소설 쓰시네"로 뒤집어진 후...법사위 2차전은 어땠을까?
News1 눈TV • 조회수 23만회 • 1일 전

(서울=뉴스1) 송영성 기자,문동주 기자 = 미래통합당이 21대 국회 개원 이후 처음으로 국회 법제사법위원회에 참석했지만 파행으로 얼룩졌습니다....



[#동상이몽★레전드] 딸을 '괴물'이라고 부르는 엄마😟 "엄마가 그리워요..." | 동상이몽, 괜찮아 괜찮아! | SBS ENTER
SBS Entertainment • 조회수 81만회 • 2일 전

00:00 엄마와 딸 등장 03:48 맘say 18:49 패널들의 의견 20:15 딸say 33:45 패널들의 의견 #동상이몽 #모녀갈등 #김구라 동상이몽, 괜찮아 괜찮아! 4회 20150331...

하태경 '적과 내통' VS 박지원 '면책특권 숨지마'
노컷브이 • 조회수 76만회 • 2일 전

#박지원 #하태경 #적과내통 [하태경 '적과 내통' VS 박지원 '면책특권 숨지마'] 미래통합당 하태경 의원은 27일 오후 국회에서 열린 박지원 국가정보...

[용터뷰] "진짜 죽을 수 있었구나" 이재명 경기도지사
김용민TV • 조회수 15만회 • 1일 전

김용민TV 7월 LIVE 편성표 월~목 17시 : 김용민브리핑 LIVE 월요일 20시 30분 : 관훈라이트클럽 (with 민동기, 정상근, 이언경) 화요일 20시 30분 : 정치부심...

배경

서비스의 필요성

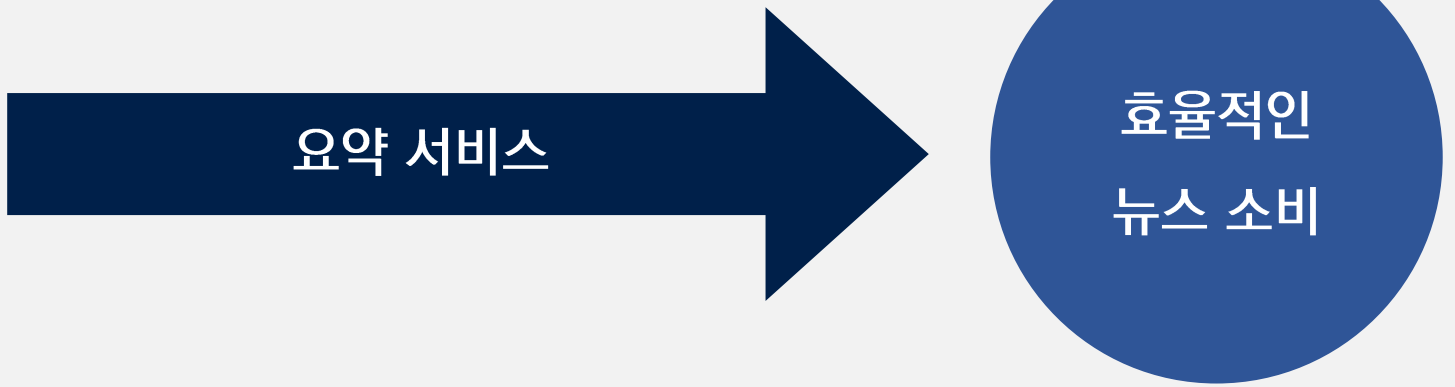
바쁜 현대인들에게 빠르게 뉴스 내용을 습득할 수 있게 하는 YouTube 뉴스 컨텐츠 요약 서비스 필요!!

YouTube 뉴스 소비 증가

바쁜 현대인

끊임 없이 생산되는 뉴스

서머리 산업 부상



Contents

- 활용데이터 개요
- 분석 · 개발환경

활용데이터 & 분석 · 개발환경

활용 데이터 개요

	Train / Validation		Test
	Naver	Daum	YouTube
수집정보	• 게시일자 • 제목 • 기사본문 • 요약문	• 게시일자 • 제목 • 기사본문 • 요약문	• 영상제목 • 영상 스크립트 • 댓글
수집방안	Beautifulsoup		pytube & YouTube API
수집건수	112,466건	109,419건	실시간
기타	날짜별, 카테고리별 상위 랭킹 뉴스		YTN, SBS, KBS, MBC, JTBC, MBN, 채널A, 연합뉴스

활용데이터 & 분석 · 개발환경

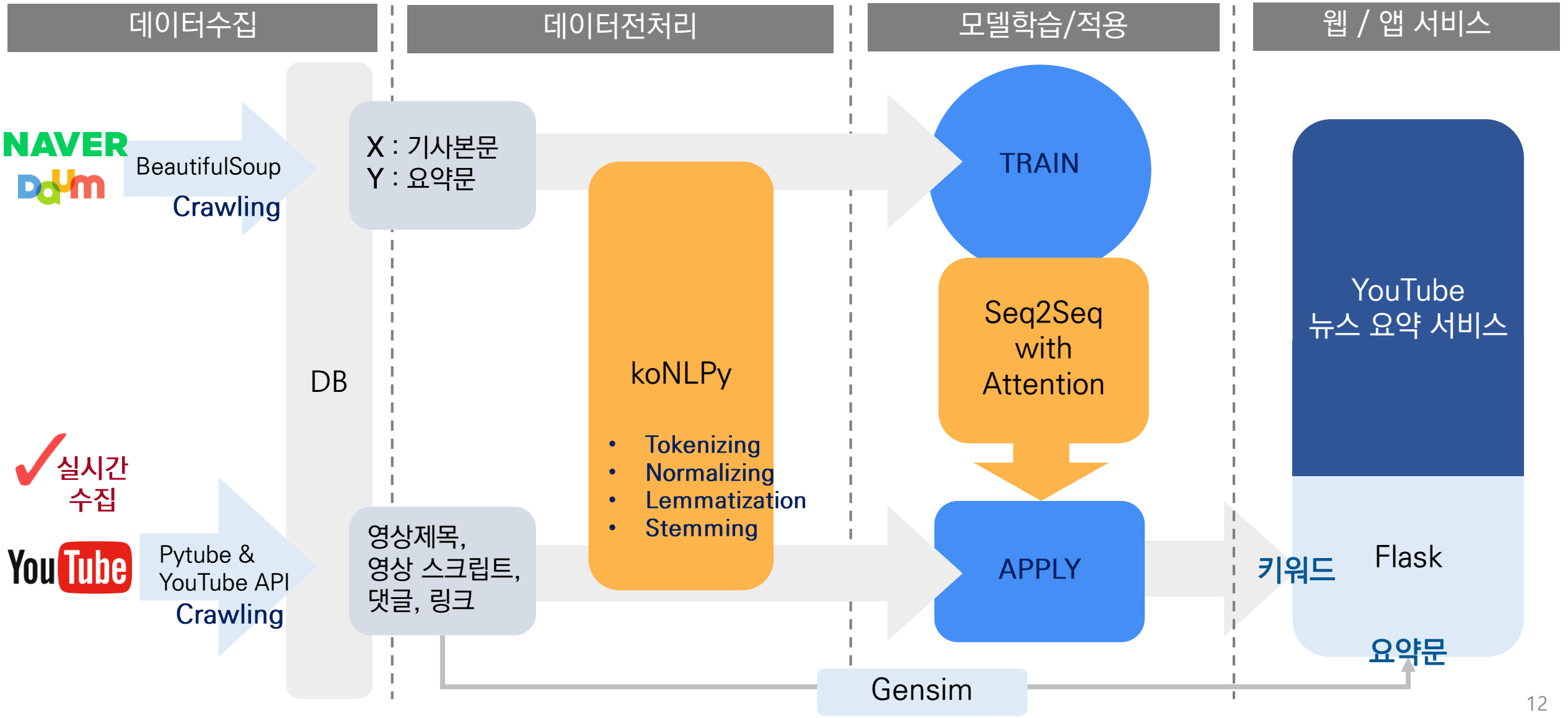
Contents

배경 — 활용데이터 정의 — **데이터 처리방안 & 분석기법** — 분석결과 — 활용방안 및 기대효과

- 시스템 흐름도
- 데이터 전처리
- 활용 모델 및 학습과정

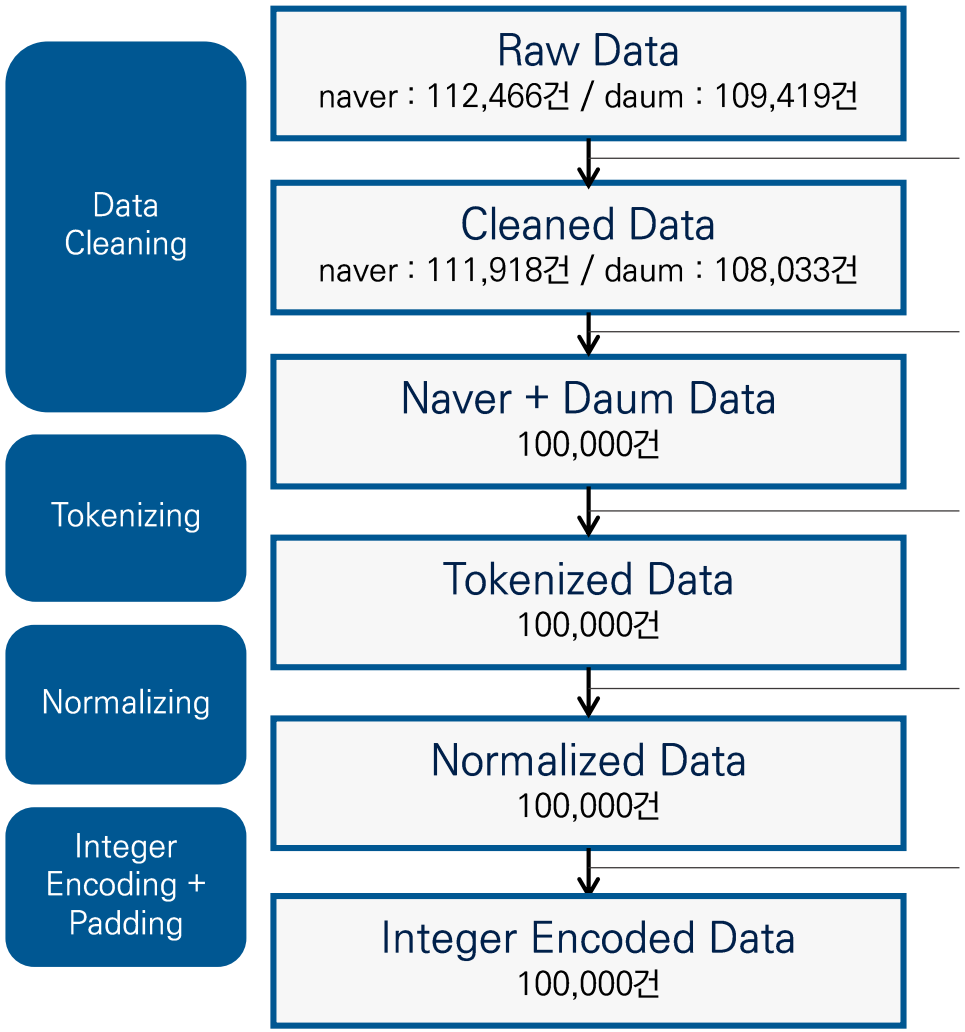
데이터 처리방안 & 분석기법

시스템흐름도



데이터 처리방안 & 분석기법

데이터 전처리 - 네이버 / 다음 데이터



기사 본문 길이가 250보다 큰 데이터만 남기기
Ex. naver.loc[naver['body'].str.len() > 250]

Data Cleaning을 위한 함수 생성
def cleaning(text)

이후 학습에서의 처리속도 향상 위해 각각 5만 건으로 한정
Ex. naver_copy.iloc[61918:].copy()

Naver + Daum concat
pd.concat()

Body, Summary에서 명사추출
def get_nouns(row, col_name='body')

소괄호, 숫자 제거
def remove_parentheses(nouns_list)
def remove_nums(nouns_list)

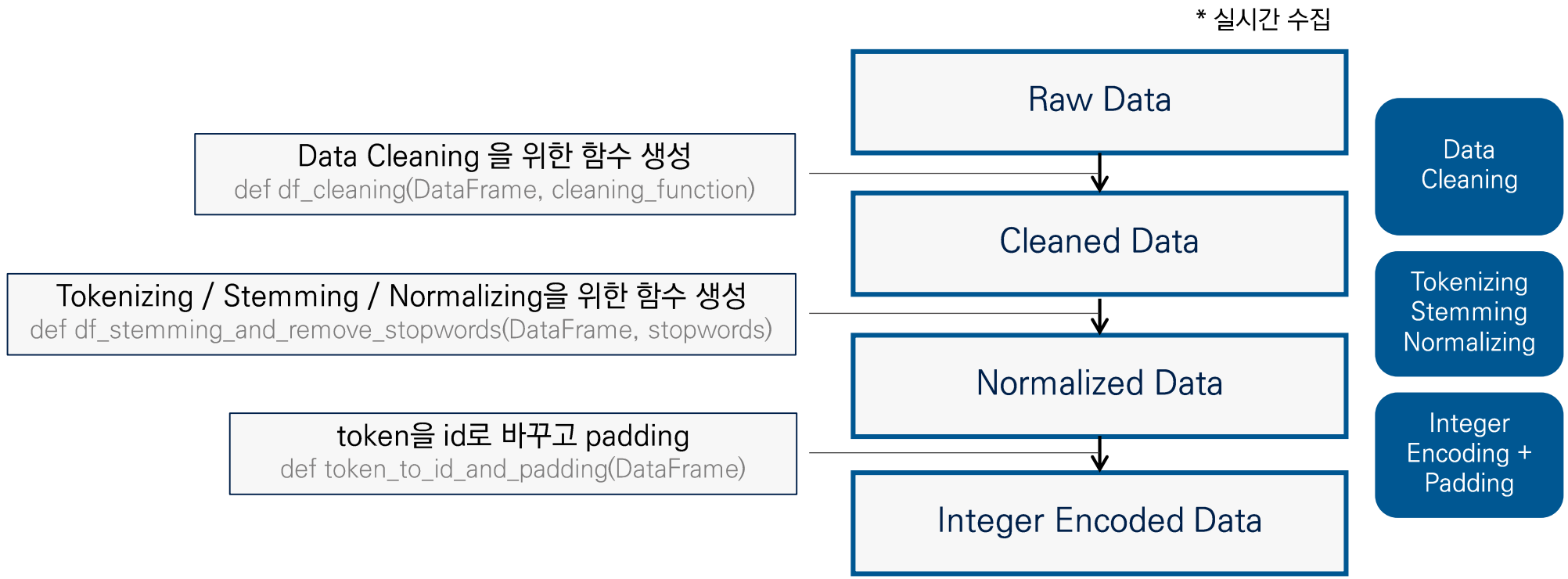
불용어 처리
def remove_stopwords(nouns_list, stopwords_list)

정수인코딩, 패딩

참고: 2. preprocessing / 01_naverdaum / 01_train_cleaning.ipynb
2. preprocessing / 01_naverdaum / 02_train_tokenizing_normalizing.ipynb

데이터 처리방안 & 분석기법

데이터 전처리 - 유튜브 데이터

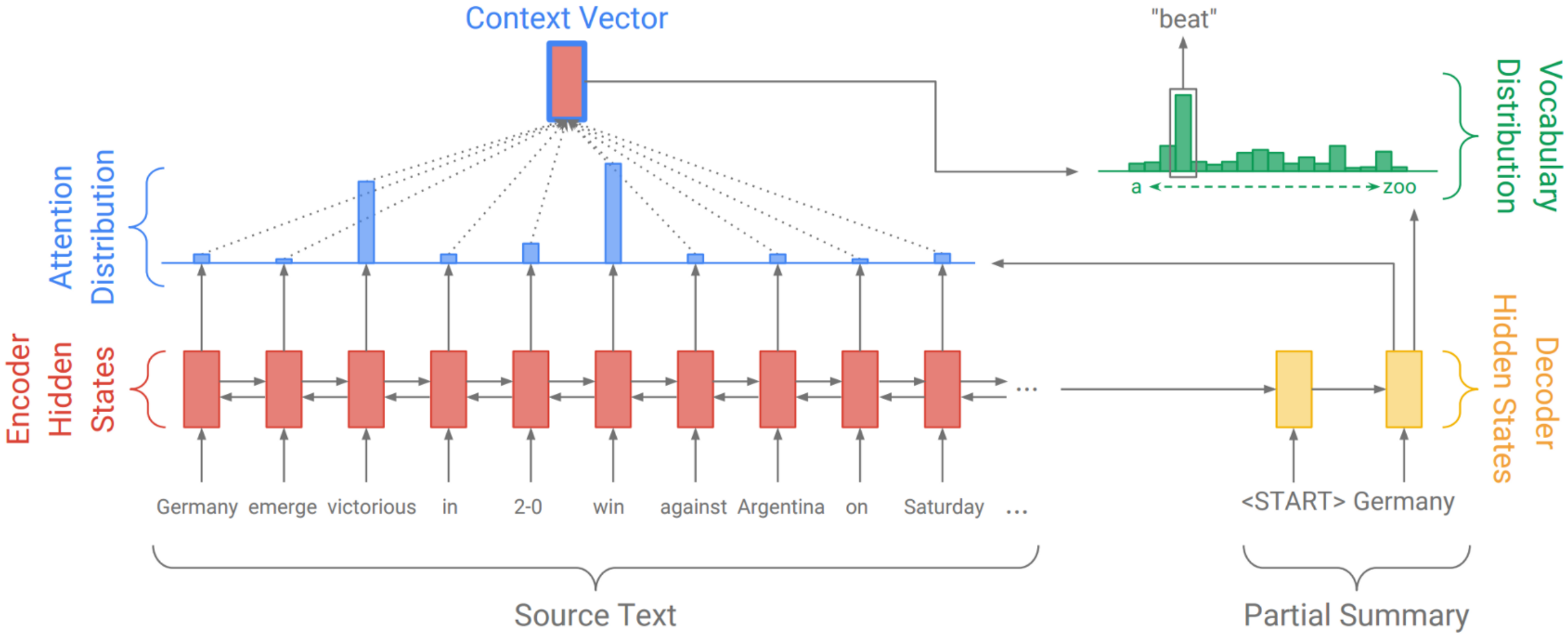


참고: 2. preprocessing / 02_youtube_pipeline / 01_main.ipynb
2. preprocessing / 02_youtube_pipeline / 02_main.ipynb

데이터 처리방안 & 분석기법

활용 모델 및 학습과정

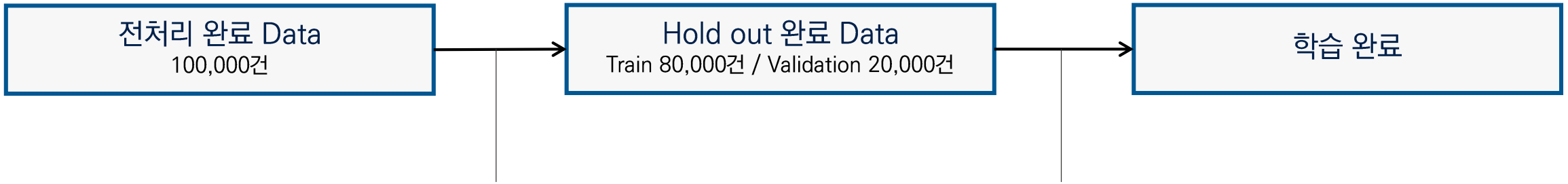
Seq2Seq with Attention



데이터 처리방안 & 분석기법

활용 모델 및 학습과정

전처리 완료한 데이터에 Seq2Seq with Attention을 적용하였다.



Train / Validation Split

- enc_train : 80,000건
- enc_val : 20,000건
- dec_in_train : 80,000건
- dec_in_val : 20,000건
- dec_out_train : 80,000건
- dec_out_val : 20,000건

Model Train

- 정수 인코딩된 자료 ➔ Embedding
- 활용 모델

  class Seq2Seq_Attention()

Contents

배경 — 활용데이터 정의 — 데이터 처리방안 & 분석기법 — **분석결과** — 활용방안 및 기대효과

- 학습 결과
- 요약문 생성

분석결과

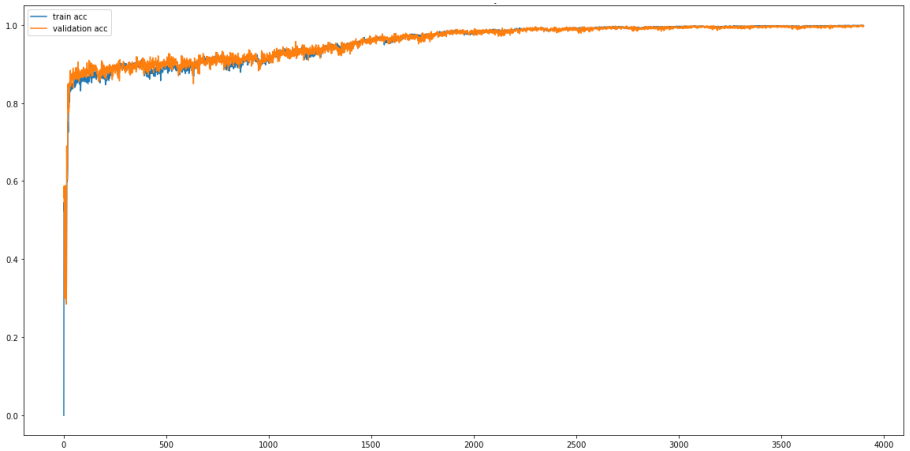
학습 결과

Attention 모델의 텍스트 요약 학습 결과 (koBERT embedding 사용)

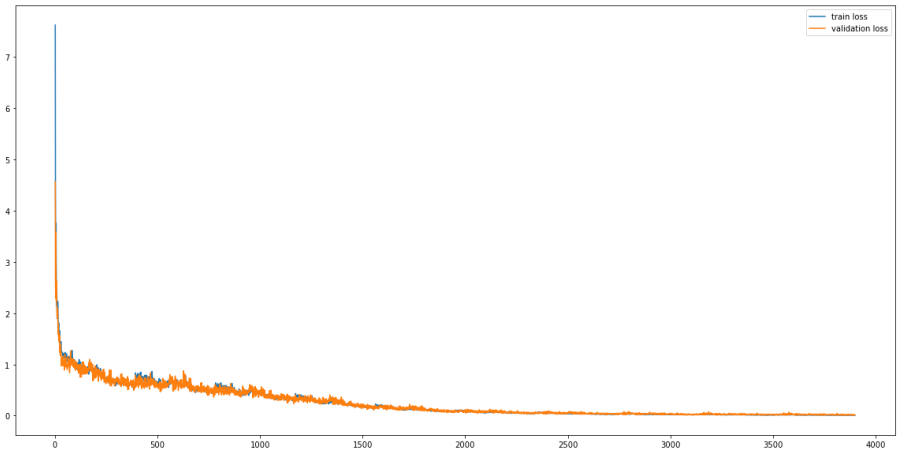
- Train / Validation 성능 정확도는 각각 99.85%, 99.60%로 높았음
- **But** 자연스러운 문장 형태로 요약되지 않음 (unknown token 지나치게 많이 발생)

```
(128, 300, 768) (128, 100, 768) (128, 100, 2017) (25, 300, 768) (25, 100, 768) (25, 100, 2017)
Time Elapsed 1 day, 2:53:06.290519 [Epoch 9/10] [Batch 0/390] Train Loss: 0.016088 Train Acc:0.996875 Valid Loss: 0.018769 Valid Acc:0.996400
Time Elapsed 1 day, 3:33:16.005473 [Epoch 9/10] [Batch 100/390] Train Loss: 0.013623 Train Acc:0.996406 Valid Loss: 0.030443 Valid Acc:0.994000
Time Elapsed 1 day, 4:13:17.650991 [Epoch 9/10] [Batch 200/390] Train Loss: 0.012345 Train Acc:0.997344 Valid Loss: 0.020870 Valid Acc:0.994400
Time Elapsed 1 day, 5:09:56.257216 [Epoch 9/10] [Batch 300/390] Train Loss: 0.009991 Train Acc:0.998516 Valid Loss: 0.022921 Valid Acc:0.996000
```

train and validation accuracy at each batch



train and validation loss at each batch



분석결과

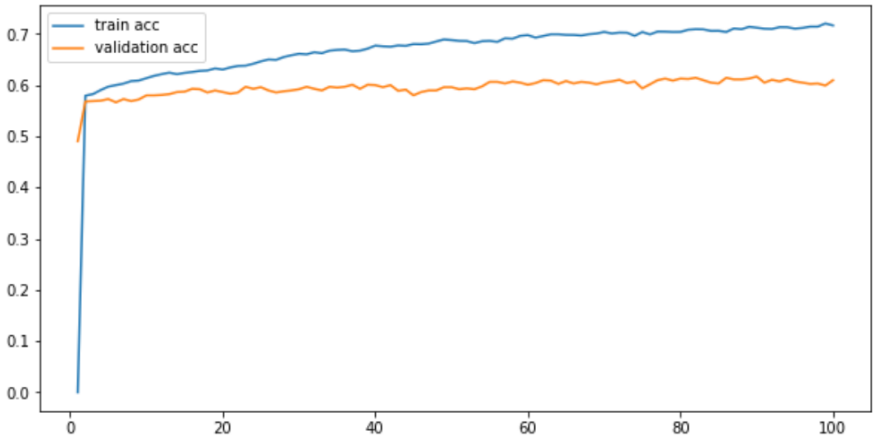
학습 결과

Attention 모델의 키워드 요약 학습 결과 (정수 encoding & embedding layer 사용)

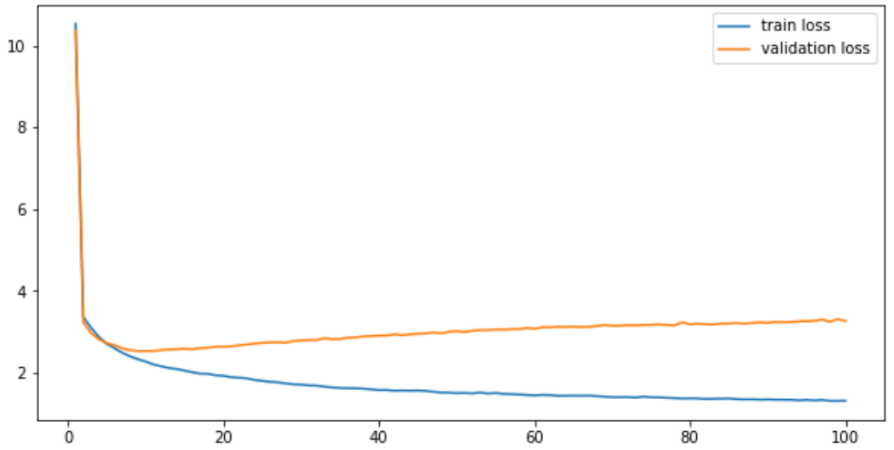
- Train / Validation 성능 정확도는 각각 71.65%, 60.96%
- 만족스러운 키워드 결과물 생성

```
Time Elapsed 4:46:24.363722 [Epoch 97/100] [Batch 0/312] Train Loss: 1.311911 Train Acc:0.714219 Valid Loss: 3.245837 Valid Acc:0.603200
(128, 300) (128, 50) (128, 50, 4879) (25, 300) (25, 50) (25, 50, 4879)
Time Elapsed 4:49:18.289958 [Epoch 98/100] [Batch 0/312] Train Loss: 1.313512 Train Acc:0.720469 Valid Loss: 3.311233 Valid Acc:0.599200
(128, 300) (128, 50) (128, 50, 4879) (25, 300) (25, 50) (25, 50, 4879)
Time Elapsed 4:52:12.845327 [Epoch 99/100] [Batch 0/312] Train Loss: 1.314143 Train Acc:0.716563 Valid Loss: 3.266664 Valid Acc:0.609600
```

train and validation accuracy at each epoch



train and validation loss at each epoch



분석결과

요약문 생성

Gensim Summarize

(extractive summarization)

기사본문 요약 결과

0
수도권 코로나19 비상 서울, 경기 거리두기 2단계 코로나19 확산세가 크게 늘어 그제 신규 확진자 수가 다섯달 만에 가장 많은 166명에 달했는데요. 방역당국과 지방자치단체의 자체 집계에 따르면 어제 0시 이후 서울과 경기에서 새로 확진 판정을 받은 감염자만, 150명을 넘었습니다. 여기에 해외유입 환자도 더해지면 확진자 수는 더 늘어날 것으로 보입니다. 정부가 확산세가 심각한 서울시와 경기도의 사회적 거리두기를 1단계에서 2단계로 높이기로 했죠? 오늘부터 2주간 서울과 경기도에서는 사회적 거리두기가 2단계로 격상됩니다.

1
엄중 경고에도 자가격리 위반 천태만상 방역당국은 코로나19 확산세가 심상치 않다며 연일 방역수칙을 지켜달라고 당부하고 있죠. 코로나19 확산세가 다시 가팔라지면서 방역당국은 지금을 위기 라고 진단했습니다. 이런 때일수록 기본 방역수칙을 지키는 게 중요한데, 자가격리 위반 사례가 계속 나오고 있습니다. 미국에 머물던 20대 정 모 씨는 자가격리 기간 다시 미국으로 출국했다가 덜미를 잡혔습니다. 경기도 의정부에서는 자가격리 무단이탈자에게 처음으로 실형이 확정됐습니다. 정부와 수사기관은 국가적 재난 상황을 감안해 앞으로도 자가격리 위반자는 엄정조치할 계획이라고 밝혔습니다.

댓글 요약 결과

0
진보는 코로나 막으려고 애쓰고... 보수는 코로나 퍼트릴려고 애쓰고... 왜?갑자기 늘어났죠? 이시국에.집회 강행 빤스 코로나 민폐충들 때문에 한국도 2차 웨이브 발생하는구나 코로나 수칙 안지킨 인간들은 죽게 생겨도 살려달란 소리 하지말길 갑자기 왜 늘어나냐, 지금까지는 숨겼냐,, 또 교회만 방송하냐,, 지금까지코로나조작했다는증거!

1
자가격리 위반 와 미친 엄중경고에도 안지끼는놈들은 해왜놈은 벌금5억때리고 추방하고 우리나라사람은 감옥 독방에 쳐넉코 벌금5억때리면 대지 참나 그냥 자가 격리자 얼굴 이름 동선 싹 공개해라 저놈들 자업자득이다 자가 격리 어기는 사람들 모두 얼굴 공개하면 된다 그러면 안지킬수가 없다 법이 약해 그러니 않 지키지 강력하게 처벌해야 자가격리 잘 지키지 벌금 대폭 올리고!

Contents

- 활용방안
- 기대효과

활용방안 & 기대효과

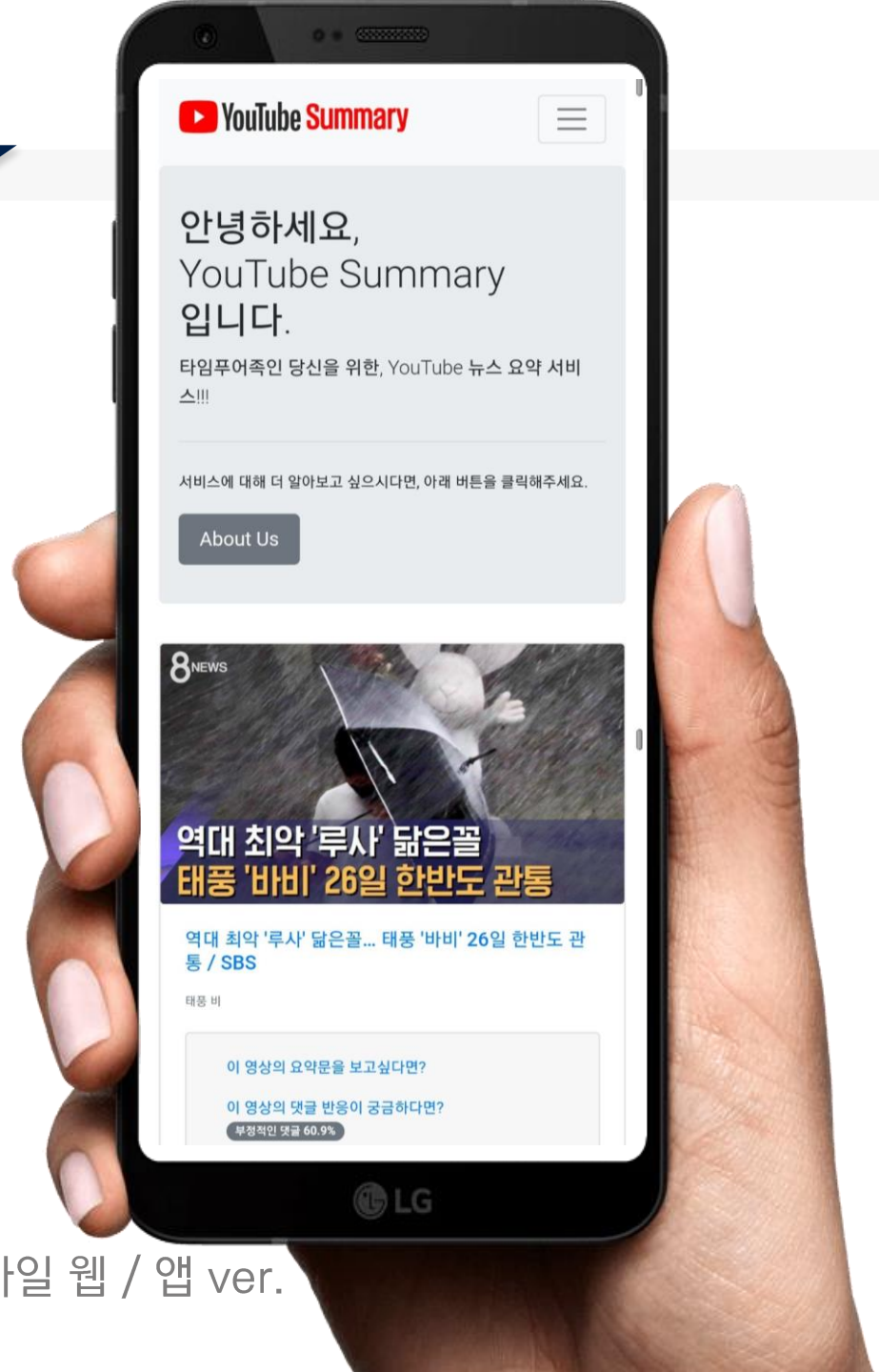
웹 / 어플리케이션 구현

내용요약

- 네이버와 다음 뉴스의 요약 결과를 학습한 모델에 YouTube 스크립트를 적용해 생성한 키워드 제공
- YouTube 스크립트에서 생성한 요약문 제공
- 영상을 클릭하면 원본 영상이 있는 유튜브 페이지로 이동

댓글요약

- 댓글 내용을 요약한 요약문 제공
- 감성분석으로 생성한 댓글의 긍정/부정 정도 제공



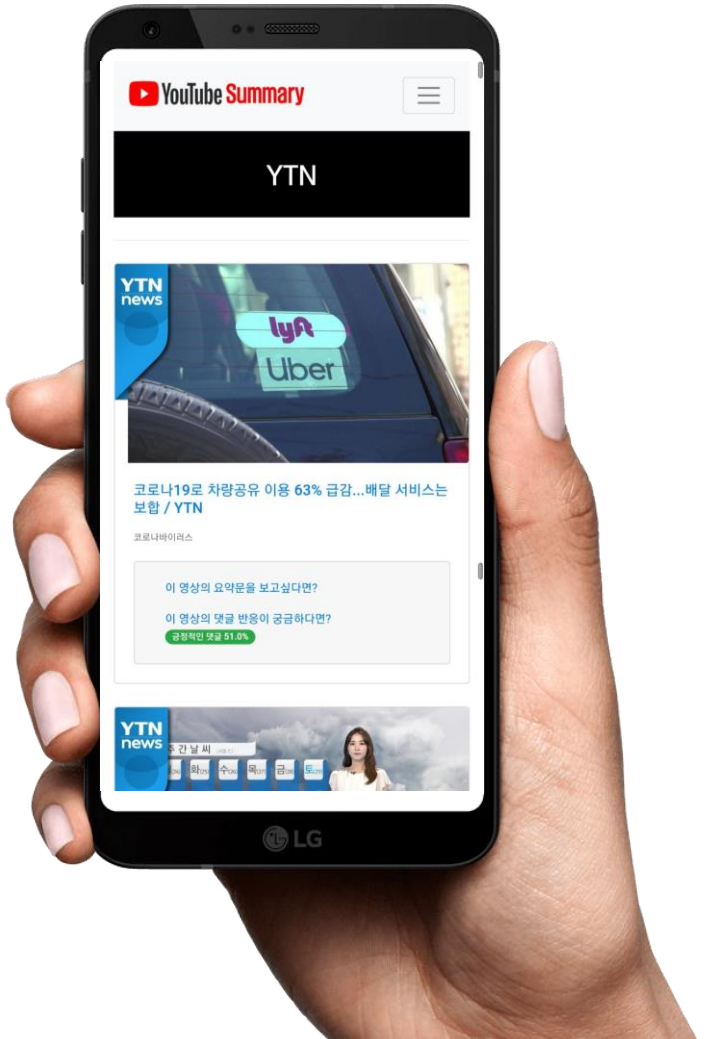
모바일 웹 / 앱 ver.

활용방안 & 기대효과

웹 / 어플리케이션 구현

일간리포트

- 생성된 키워드의 랭킹 정보 제공
- 일주일간의 Word Cloud 제공



언론사별 보기

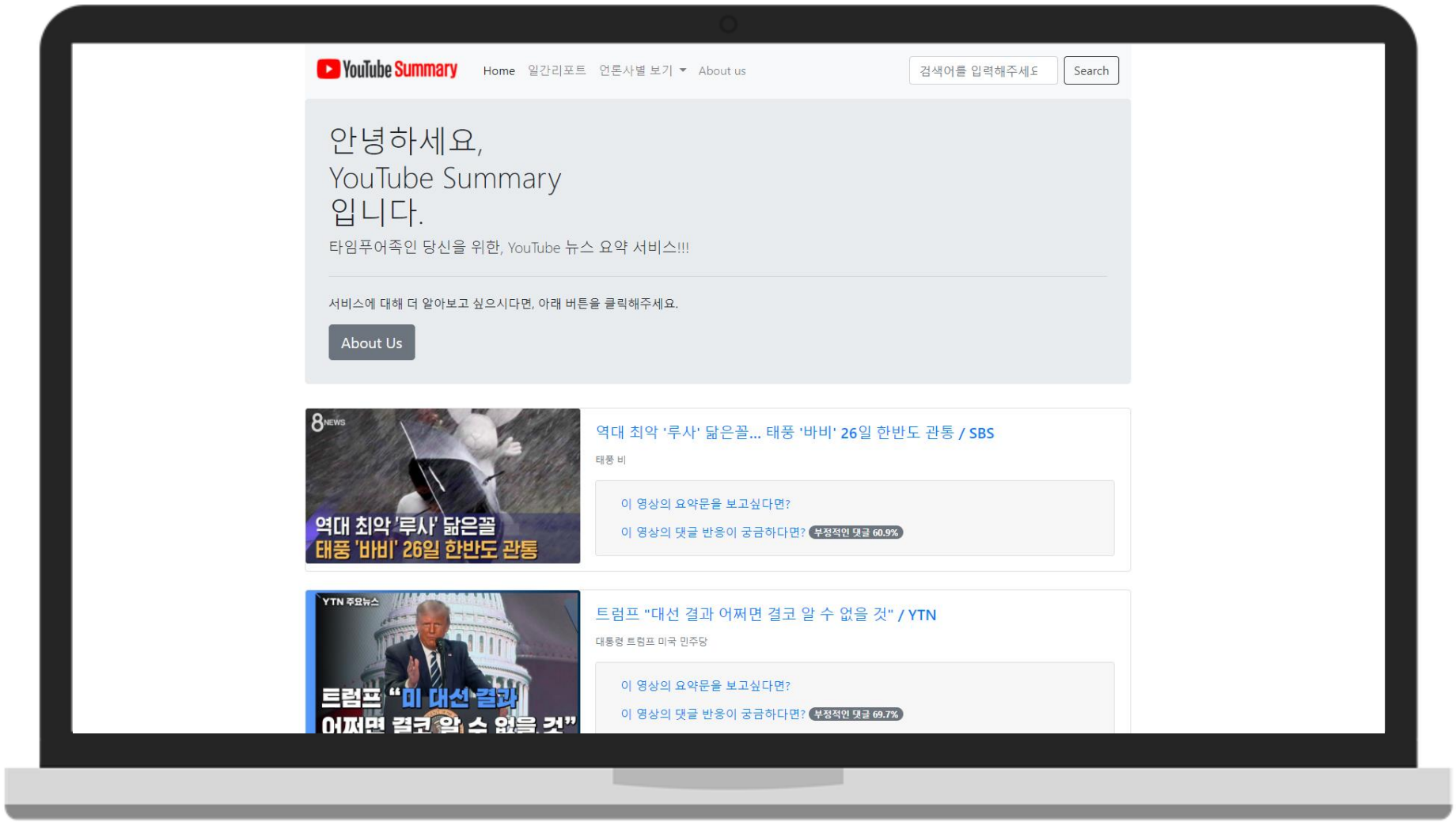
- 원하는 언론사별로 모아 볼 수 있는 페이지 제공

모바일 웹 / 앱 ver.

활용방안 & 기대효과

활용방안

웹 / 어플리케이션 구현



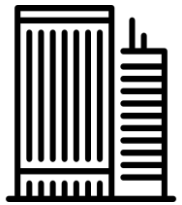
PC 웹 ver.

활용방안 & 기대효과

기대효과

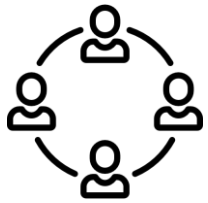
소비자

- 정보의 홍수 속에서 양질의 정보 선택 용이
- 시간 단축을 통한 효율적 컨텐츠 소비 가능



언론채널

- 시청자들의 선택의 폭이 증가하여 경쟁력 확보를 위한 정확한 정보전달로 언론 신뢰도 향상
- 기존에 긴 기사 / 영상은 보지 않던 신규 시청자 유입 창출



정부(사회)

- 유튜브 알고리즘으로 인한 확증편향을 방지하여 국민 지식수준 향상
- 타임푸어(Time Poor)족의 뉴스 소비를 촉진시킴으로써 정보격차 해소에 일조

References

- dongjun-Lee / text-summarization-tensorflow Github

  https://github.com/dongjun-Lee/text-summarization-tensorflow

- SKTBrain / KoBERT Github

  https://github.com/SKTBrain/KoBERT

- RaRe-Technologies / genism Github

  https://github.com/RaRe-Technologies/gensim

- 한국언론진흥재단 신문과방송 2019년 4월호
- WISEAPP 안드로이드 앱 사용시간 통계
- 영국 로이터저널리즘연구소 '디지털 뉴스 리포트 2020'