



유튜브 타이틀 크롤링 카테고리 선별 모델

엄재식 한희준 이가원 허진호

유튜브중독 / 2024.01.31

Contents

01. 데이터 크롤링

- 사이트 조사
- 학습 데이터 수집
- 검증 데이터 수집

02. 모델 생성

- 데이터 전처리
- 모델 학습

03. 모델 검증

- 유튜브 알고리즘 테스트 결과
- 검색한 데이터 결과

04. 개선안

- 개선방안
- 결론



Contents

01. 데이터 크롤링

- 사이트 조사
- 학습 데이터 수집
- 검증 데이터 수집


02. 모델 생성

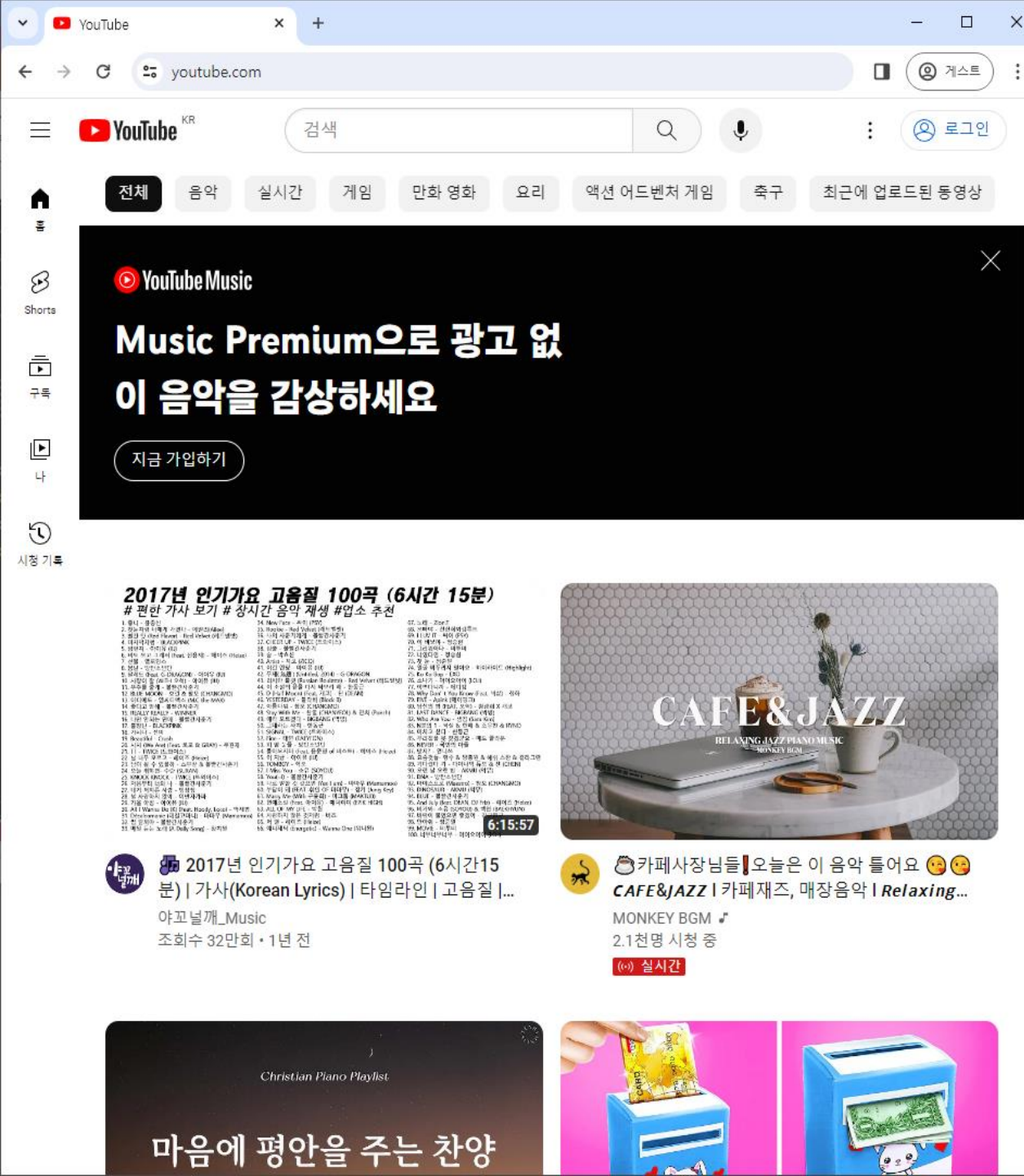
- 데이터 전처리
- 모델 학습

03. 모델 검증

- 유튜브 알고리즘 테스트 결과
- 검색한 데이터 결과

04. 개선안

- 개선방안
 - 결론
- 



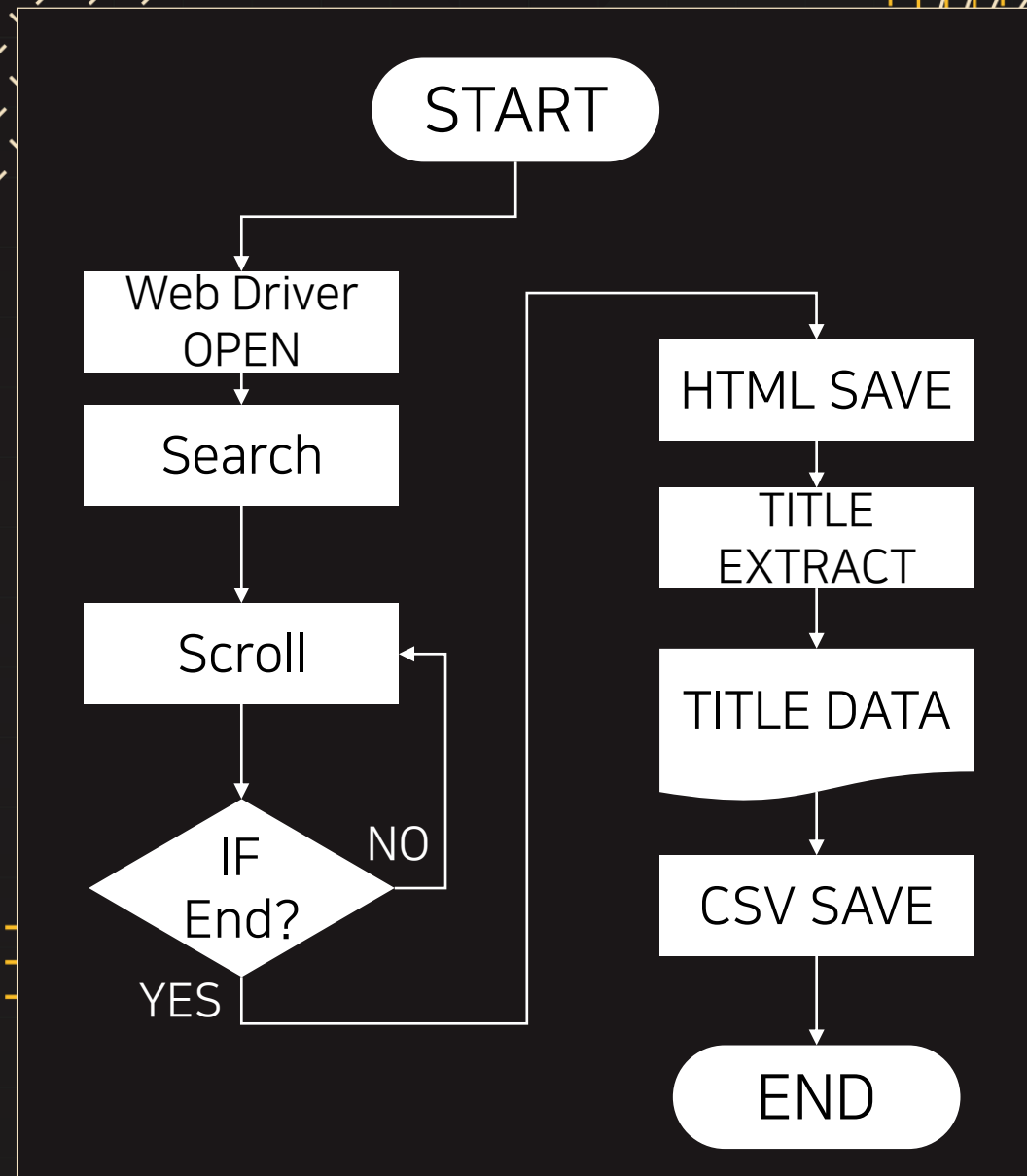
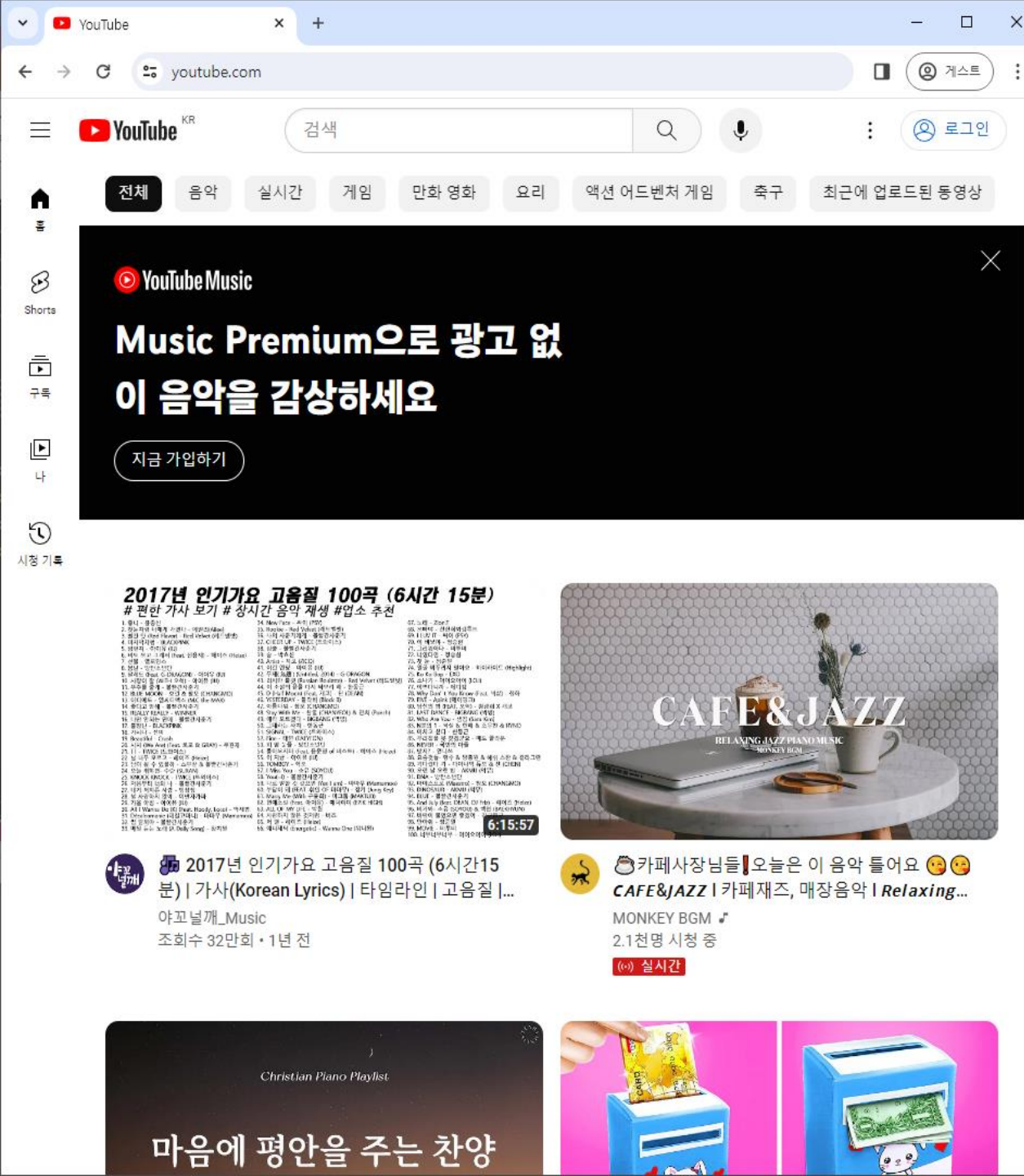
1. 검색 URL은 고정

https://www.youtube.com/results?search_query=%EA%B2%80%EC%83%89
https://www.youtube.com/results?search_query=%EC%9D%8C%EC%95%85

2. 한번에 나오는 데이터는 한정 되어 있음

많은 데이터를 보려면 스크롤 하여 서버로부터 자료를 요청 해야 함

3. Selenium으로 HTML 저장하고 BeautifulSoup으로 타이틀 추출




```
def scroll():
    last_page_height = driver.execute_script("return
                                              document.documentElement.scrollHeight")
    while True:
        driver.execute_script("window.scrollTo(0,
                                              document.documentElement.scrollHeight);")
        time.sleep(2)
        new_page_height = driver.execute_script("return
                                              document.documentElement.scrollHeight")
        if new_page_height == last_page_height:
            print('END SCROLL!!')
            break
        else:
            last_page_height = new_page_height
```

Scroll Function

Crawling Function

```
driver.get(url)

time.sleep(2)

scroll()

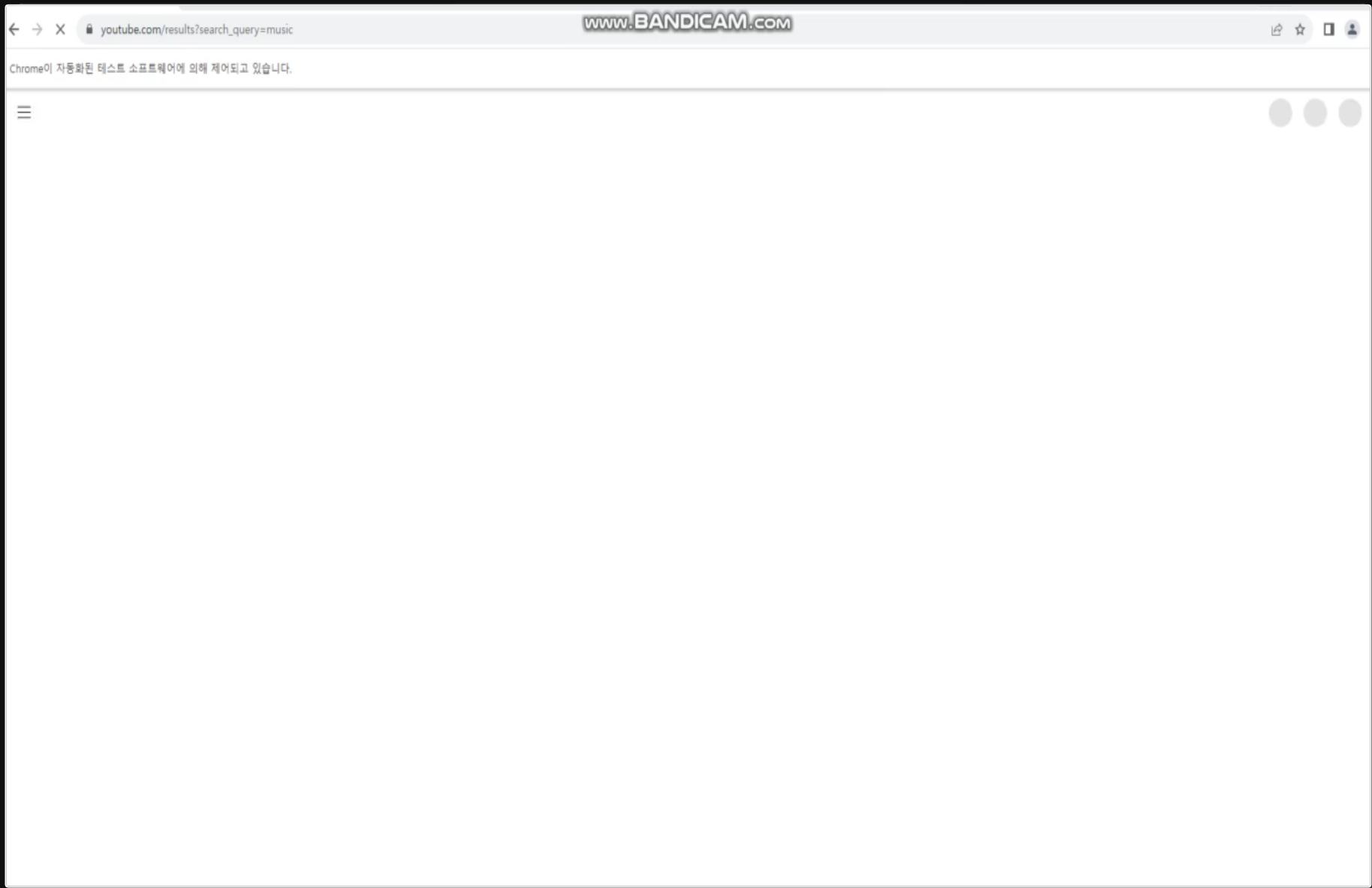
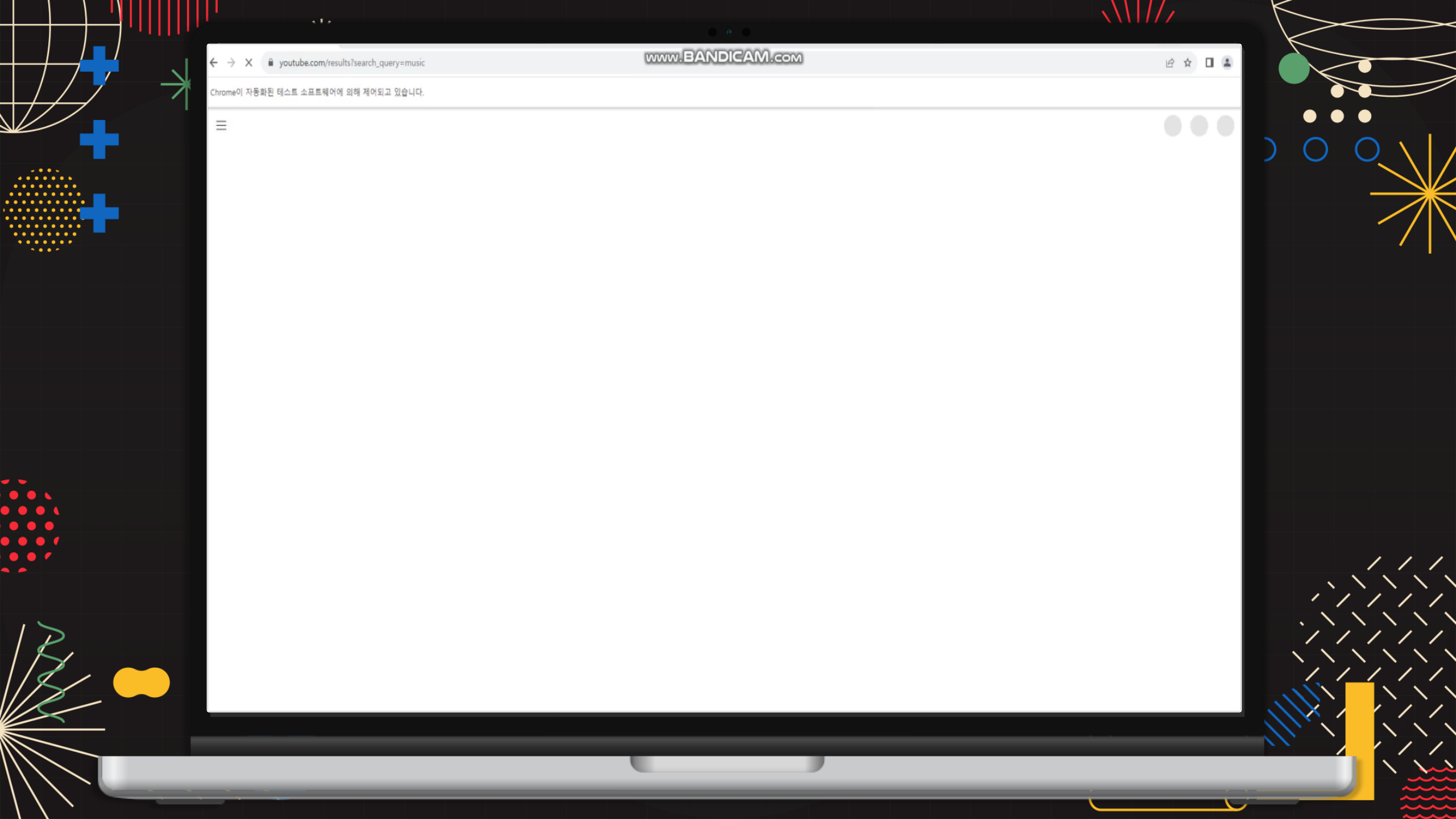
html = bs(driver.page_source, 'html.parser')
driver.close()

for content in html.select('a#video-title'):
    title = content.get('title')
    title = re.compile('[^가-힣|a-z|A-Z]').sub(' ', title)

    titleList.append(title)
```

Crawling_Data.csv

titles	category
내 남친은 고양이 라라 라이프의 필수 애완동물 가젯과 꿀팁	pets
애교만점 늑대 강아지랑 놀다보니 자기가 강아지인 줄 아는 늑대 KBS 주주클럽 방송	pets
광족들 새 멤버 공개 학교에서 애완동물 키우기	pets
동물을 사랑하는 할머니 Thug Life Grandma Is A Legit Animal Lover	pets
반려동물을 돌보는 방법 가난한 애완동물이 부자가 됐어요 라라 라이프의 반려동물 주인들을 위한 가젯	pets
반려동물 관련 직업을 갖고 싶다면 주목 펫 고등학교가 따로 있다고 Shorts	pets
물에 젖은 종이 빨대처럼 흐물흐물 거리는 애완동물	pets
asmr shorts pet dog 개 애완동물	pets
심장을 강타하는 귀여운 동물 마리 애완동물 반려동물	pets
애완견 비글 아시나요 애완동물 반려동물 강아지 shorts	pets
요즘 키우는 사람들이 늘고 있는 애완동물 feat 클로버더빙	pets
못 생겨진 주인 얼굴을 본 강아지의 귀여운 반응 반전주의	pets
ai로 만든 오늘 봐야 할 가지 귀여운 애완동물 슬라이드쇼	pets
웃긴 동물들 모음 탄	pets





Contents

01. 데이터 크롤링

- 사이트 조사
- 학습 데이터 수집
- 검증 데이터 수집


02. 모델 생성

- 데이터 전처리
- 모델 학습

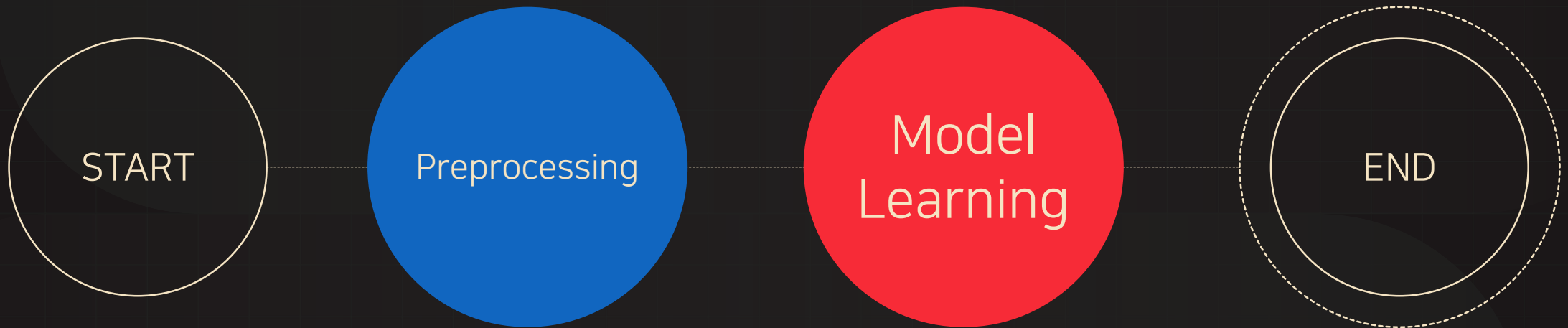
03. 모델 검증

- 유튜브 알고리즘 테스트 결과
- 검색한 데이터 결과

04. 개선안

- 개선방안
 - 결론
- 

Process



Data Concat
Labeling
Konlpy OKT
Training Data Create

models
label_encoder.pickle
news_category_classification_model_0.909443736076355.h5
news_category_classification_model_0.9256144762039185.h5
news_token.pickle

label_encoder & token
이후 Predict Data 생성 시 사용 예정

모델 생성 완료

```
Epoch 1/15  
49/49 [=====] - 14s 171ms/step - loss: 1.6064 - accuracy: 0.2723 - val_loss: 1.1447 - val_accuracy: 0.5166  
Epoch 2/15  
49/49 [=====] - 7s 148ms/step - loss: 0.7885 - accuracy: 0.6533 - val_loss: 0.6005 - val_accuracy: 0.8042  
Epoch 3/15  
49/49 [=====] - 7s 148ms/step - loss: 0.2816 - accuracy: 0.9131 - val_loss: 0.3090 - val_accuracy: 0.9083  
49/49 [=====] - 8s 154ms/step - loss: 0.1110 - accuracy: 0.9699 - val_loss: 0.2895 - val_accuracy: 0.9219  
49/49 [=====] - 7s 147ms/step - loss: 0.0438 - accuracy: 0.9899 - val_loss: 0.3142 - val_accuracy: 0.9258  
49/49 [=====] - 7s 147ms/step - loss: 0.0309 - accuracy: 0.9927 - val_loss: 0.3334 - val_accuracy: 0.9226  
49/49 [=====] - 7s 147ms/step - loss: 0.0274 - accuracy: 0.9933 - val_loss: 0.3293 - val_accuracy: 0.9226  
49/49 [=====] - 7s 152ms/step - loss: 0.0262 - accuracy: 0.9950 - val_loss: 0.3544 - val_accuracy: 0.9232  
Epoch 9/15  
49/49 [=====] - 8s 160ms/step - loss: 0.0217 - accuracy: 0.9954 - val_loss: 0.3492 - val_accuracy: 0.9258  
Epoch 10/15  
49/49 [=====] - 7s 151ms/step - loss: 0.0188 - accuracy: 0.9956 - val_loss: 0.3904 - val_accuracy: 0.9232  
Epoch 11/15  
49/49 [=====] - 7s 152ms/step - loss: 0.0205 - accuracy: 0.9938 - val_loss: 0.3357 - val_accuracy: 0.9291  
Epoch 12/15  
49/49 [=====] - 7s 147ms/step - loss: 0.0179 - accuracy: 0.9948 - val_loss: 0.3315 - val_accuracy: 0.9291  
Epoch 13/15  
49/49 [=====] - 7s 146ms/step - loss: 0.0151 - accuracy: 0.9958 - val_loss: 0.3401 - val_accuracy: 0.9304
```



Contents

01. 데이터 크롤링

- 사이트 조사
- 학습 데이터 수집
- 검증 데이터 수집


02. 모델 생성

- 데이터 전처리
- 모델 학습

03. 모델 검증

- 유튜브 알고리즘 테스트 결과
- 검색한 데이터 결과

04. 개선안

- 개선방안
 - 결론
- 

검증 데이터 수집 방법

검색한 데이터 & 유튜브 알고리즘

01.

상단 메뉴바에 있는 검색창

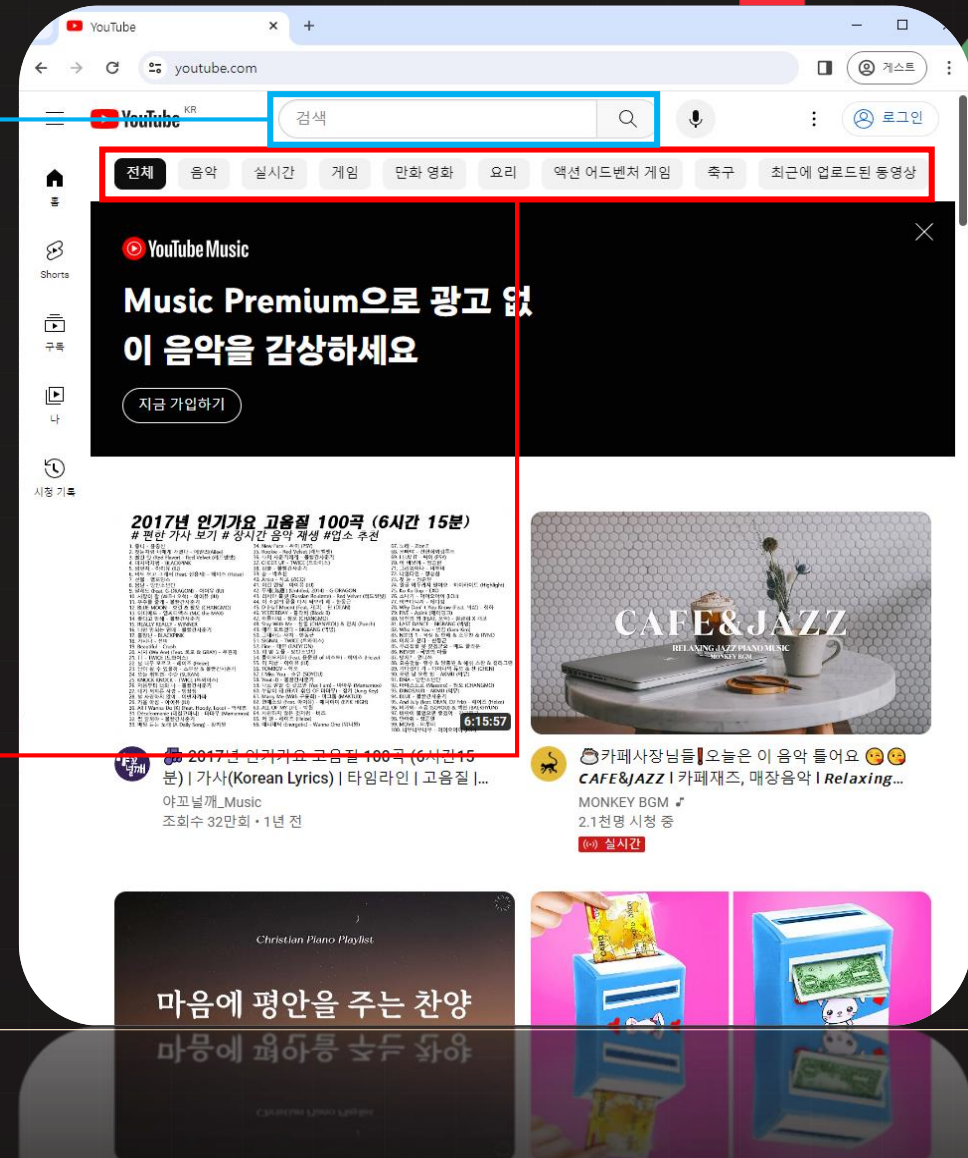
이용하여 원하는 데이터 카테고리

검색하여 수집한 검증 데이터

02.

유튜브가 만들어 놓은 알고리즘

데이터 카테고리 검증 데이터



검증 결과

```
titles ... predict
0 Playlist 첫사랑 생기다 설레다 땃국 노래 모음 플레이 리스트 ... [music, nature]
1 Beautiful Ghibli Collection 아름답다 피아노 지브리 멜로디 궁... ... [music, nature]
2 AFC 카타르 아시안컵 AFC 카타르 아시안컵 이라크 VS 요르단 하이라이트 ... [sports, music]
3 하루 종일 듣기 좋다 노래 듣기 좋다 발라드 추천 광고 노래 모음 김범수 임창정 박... ... [music, nature]
4 건물 페이커 Faker 강림 알다 형님 JTBC 방송 ... [game, sports]
...
1456 옛날 메이플 유저 접다 이유 해결 해보다 메이플스토리 테스트 토크 ... [game, sports]
1457 언제나 되어다 ... [game, cook]
1458 수준 높다 교육 기회 누리 생후 개월 시작 부의 불평등 제일 고다 다큐프라임 고르다 다큐 ... [nature, game]
1459 친구 동생 항상 오빠 꼬리표 가족 집중 오빠 vs 동생 성적 비교 티쳐스 ... [sports, pets]
1460 연인 냄새 허용 범위 ... [game, music]

[1461 rows x 3 columns]
OX
0 836
X 625
Name: count, dtype: int64
OX
0 0.572211
X 0.427789
Name: count, dtype: float64

Process finished with exit code 0
```

유튜브 알고리즘에 의한 모델 검증
57.2%

```
titles ... predict
0 편안하다 음악 Soothing Relaxation 수면 음악 부드럽다 피아노 음악 ... [music, nature]
1 Chill aesthetic music playlist hours lofi NO ADS ... [music, nature]
2 Best Christian Songs Worship Instrumental Musi... ... [music, nature]
3 편안하다 음악 수면 음악 마음 진정 피아노 음악 오다 음악 재생 ... [music, nature]
4 유령 힐링 음악 ... [music, cook]
...
265 자연과의 만남 몽클 숨터 생명 우포늪 KBS 방송 ... [nature, music]
266 멀리 아름답다 가까이 가면 위험하다 자연 ... [nature, music]
267 카메라 포착 강력 자연재해 Top ... [nature, music]
268 소리 비다 소리 힐링 소리 ASMR ... [nature, pets]
269 자연 마음 우주 통족 불통 vs 고집멸도 사성제 고미숙 ... [nature, music]

[270 rows x 3 columns]
OX
0 249
X 21
Name: count, dtype: int64
OX
0 0.922222
X 0.077778
Name: count, dtype: float64
```

검색 데이터에 의한 모델 검증
92.2%



Contents

01. 데이터 크롤링

- 사이트 조사
- 학습 데이터 수집
- 검증 데이터 수집


02. 모델 생성

- 데이터 전처리
- 모델 학습

03. 모델 검증

- 유튜브 알고리즘 테스트 결과
- 검색한 데이터 결과

04. 개선안

- 개선방안
 - 결론
- 

개선안



데이터 전처리

유튜브 알고리즘에서 분류한
데이터 카테고리의 부정확함



다언어 처리

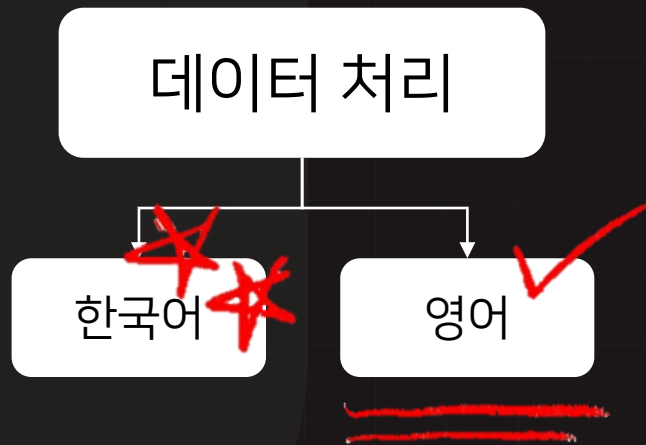
한국어 & 영어 다중 처리



썸네일 연동

자연어 처리 모델과 함께
썸네일을 통한 카테고리
분류 모델과의 연동 분류

결과



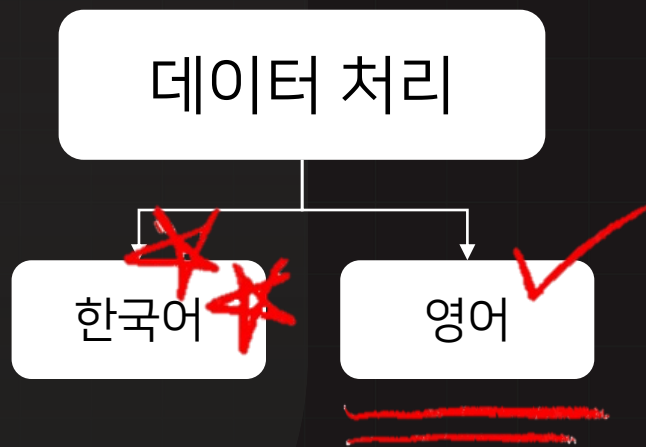
1. 부정확한 검증 데이터

추출한 검증 데이터셋의 부정확한 데이터 제거하는 알고리즘 필요

2. 영어 데이터 전처리 과정

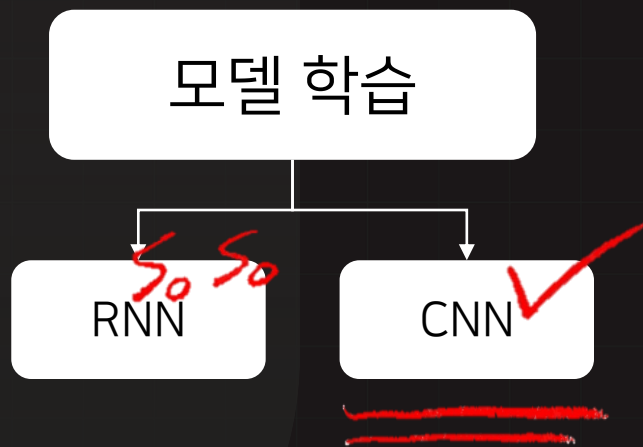
한국어 처리에만 집중하여 영어의 stopwords, 전처리 하지 않음

결과



- 시즌 북불북 여행 에서는 라면 차다 많이 먹다 에도 장면 재미 예쁘다
music, "['pets', 'game']", X
- 티전드 수근 김장 개그 장인 이수근 인간문화재 시급하다 서유기
music, "['game', 'sports']", X
- 물가 극악 샌프란시스코 충격 한국인
music, "['sports', 'cook']", X

결과



1. RNN 모델 학습 과정

모델의 다양한 Layer Argument 설정

2. Thumbnail CNN 모델 생성

모델의 정확도를 높이기 위해 RNN, CNN 모델의 연동 가능성 판단



Thanks !

유튜브중독
2024. 01. 31

