# Youtube

진행상황

문헌정보학과 전은지 2019년 4월 25일

# 진행상황

- ㆍ 주제 선정
- 데이터 수집
- ・ 앞으로의 분석방향

### 1. 분류 알고리즘 생성

- · 제시된 데이터 분석을 통한 카테고리 분류 알고리즘 **개선**
- · 텍스트 데이터에 **가중치**를 부여하여 카테고리 분류 기준 생성



미세먼지가 무서운 이유

소개해주는 남자 ♥ 조회수 1만회 • 5일 전 영화 [인더더스트] 스토리텔링 리뷰영상입니다. 감사합니다. 세 등영상



미국 해병대가 빡센 이유

소개해주는 남자 ② 조회수 10만회 • 1주 전 영화 [어퓨굿맨] 스토리텔링 리뷰영상입니다. 감사합니다.



남의 행운을 훔치는 초능력자의 삶

소개해주는 남자 ② 조회수 96만회 • 1년 전 영화[인택토] 즐거리 요약 및 리뷰 영상입니다. 영상에 스포일러 및 결말이 포함되어있으니 시청하실 때 주의해주시길 바랍니다.



낮에는 왕따, 밤에는 천재해커인 소년의 삶

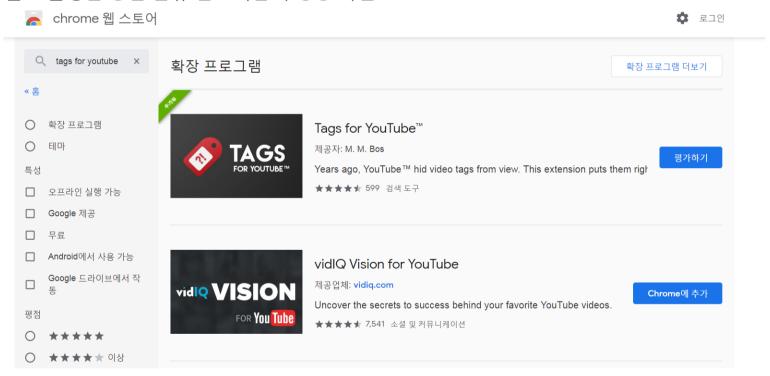
소개해주는 남자 🔗 조회수 67만회 • 5개월 7

영화[후엠아이]의 스토리텔링 및 리뷰영상입니다. 결말과 결정적인 스포일러는 포함되어있지않습니다 이 점에 있어서 양해부탁 ...

• 국내 유튜브 컨텐츠에 적합한 새로운 카테고리

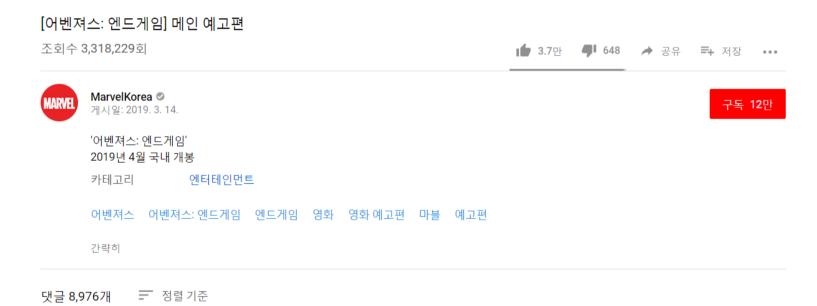
### 2. 알고리즘 성능 확인

· 웹 크롤링을 통한 분류 알고리즘의 성능 확인



### 2. 알고리즘 성능 확인

· 웹 크롤링을 통한 분류 알고리즘의 성능 확인



### 패키지 소개

#### 1. Stringr

```
• str_remove, str_replace
• str_split
• str_detect

extract_tag <- function(category_num){
    df <- subset.data.frame(KRvideos, (category_id==category_num)&(tags!='[none]'))
    keywords_vector <- vector()
    for (tag in df$tags){
        keywords <- tag %>%
            tolower() %>%
            str_remove_all('"') %>%
            str_replace_all("", "\\|\|\|\|\|\") %>%
            unlist() %>%
            unique()
        keywords_vector <- append(keywords_vector, keywords)
    }
}</pre>
```

### 패키지 소개

#### 1. tm

- TermDocumentMatrix
- · stopwords, removePuntuation, stripWhitespace

```
text <- readLines(paste(getwd(), "/wordcount_tb1.txt", sep=""), encoding="UTF-8") %>%
   VectorSource() %>%
   Corpus() %>%
   TermDocumentMatrix()
Encoding(tdm1$dimnames$Terms) = 'UTF-8'
tdm <- tdm %>% as.matrix()
```

### 패키지 소개

#### 1. KoNLP

- extractNoun
- buildDictionary

```
## Warning: package 'KoNLP' was built under R version 3.5.3
## Checking user defined dictionary!

sentence <- "안녕하세요 데이터 사이언스 입문 수강생 여러분"
extractNoun(sentence)
## [1] "안녕" "데이터" "사이언스" "입문" "수강생" "분"
```

### 3. 추가적인 기대 효과

- ・ 트렌드 분석
- ・ 특정 채널 및 크리에이터의 컨텐츠 특징 파악
- ・ 추천 동영상 제시

## 해결해야 할 문제

- 시점에 따른 카테고리 변동성, 연속성
- 비어있거나 키워드와 연관성이 없는 데이터

감사합니다.