Youtube

진행상황

문헌정보학과 전은지
2019년 4월 25일

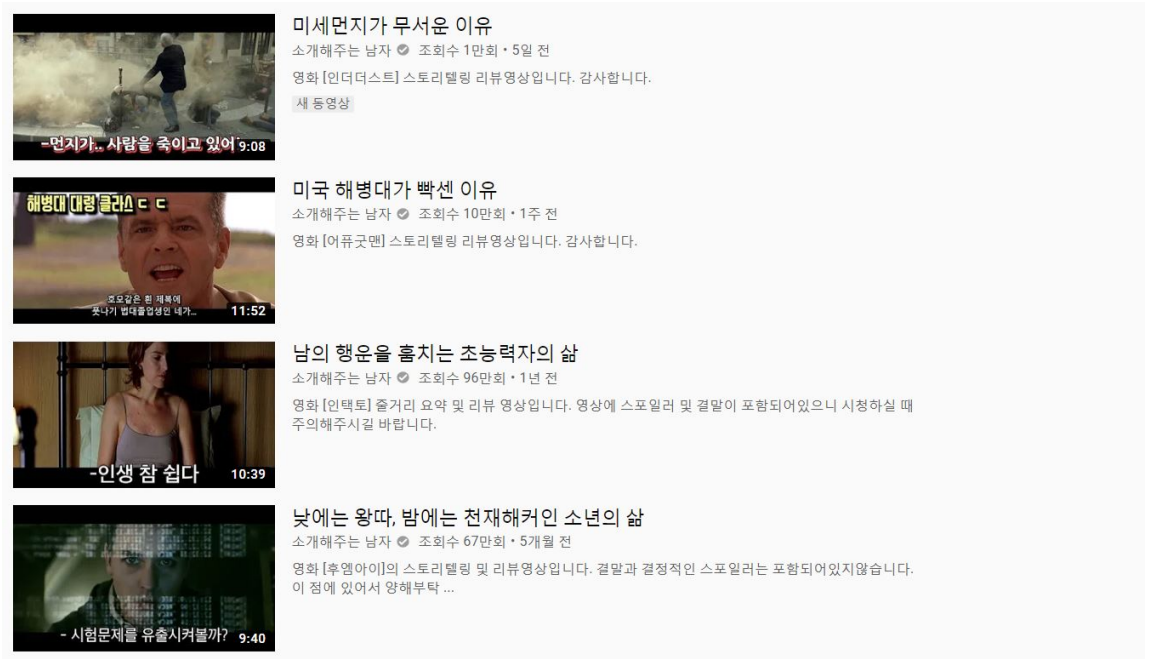
진행상황

- 주제 선정
- 데이터 수집
- **앞으로의 분석방향**

분석 방향 및 목표

1. 분류 알고리즘 생성

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

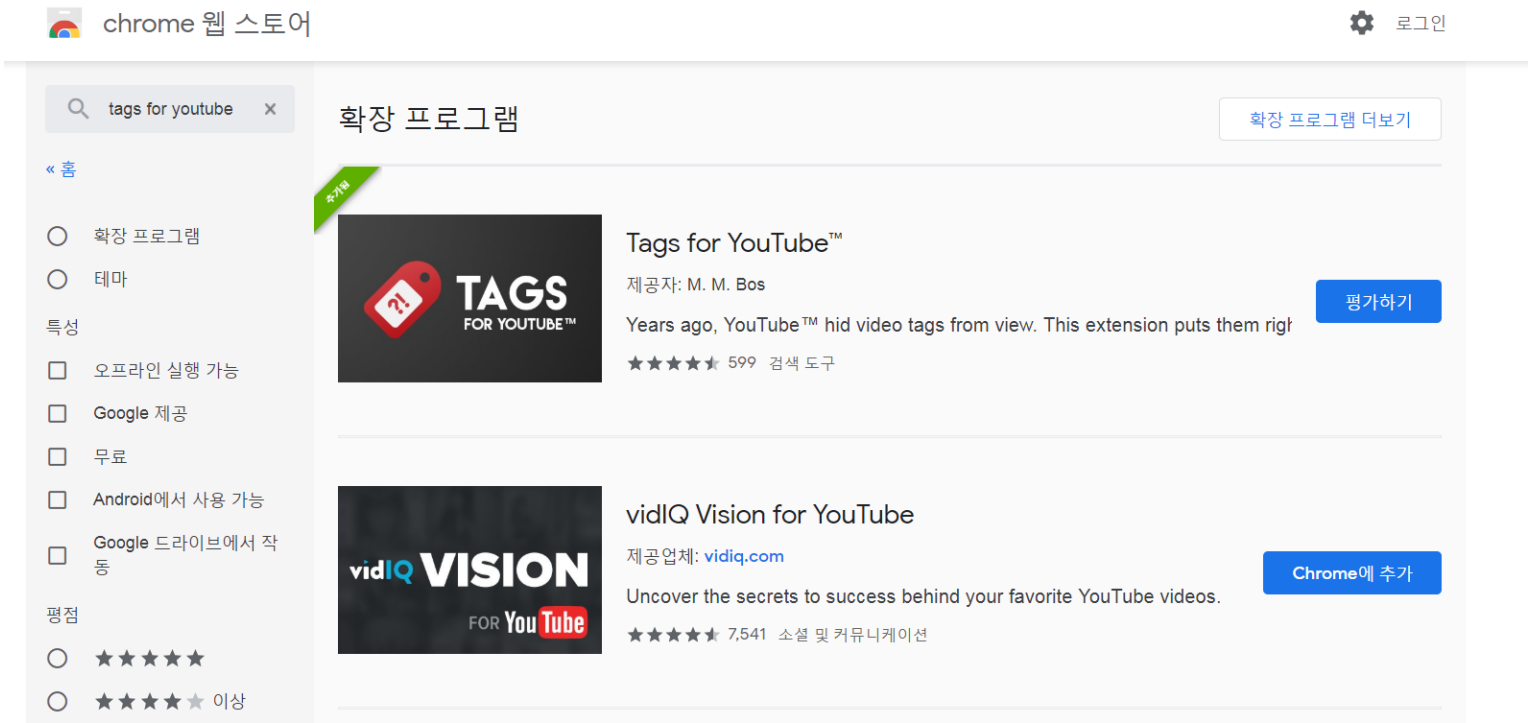


- **국내** 유튜브 컨텐츠에 적합한 새로운 카테고리

분석 방향 및 목표

2. 알고리즘 성능 확인

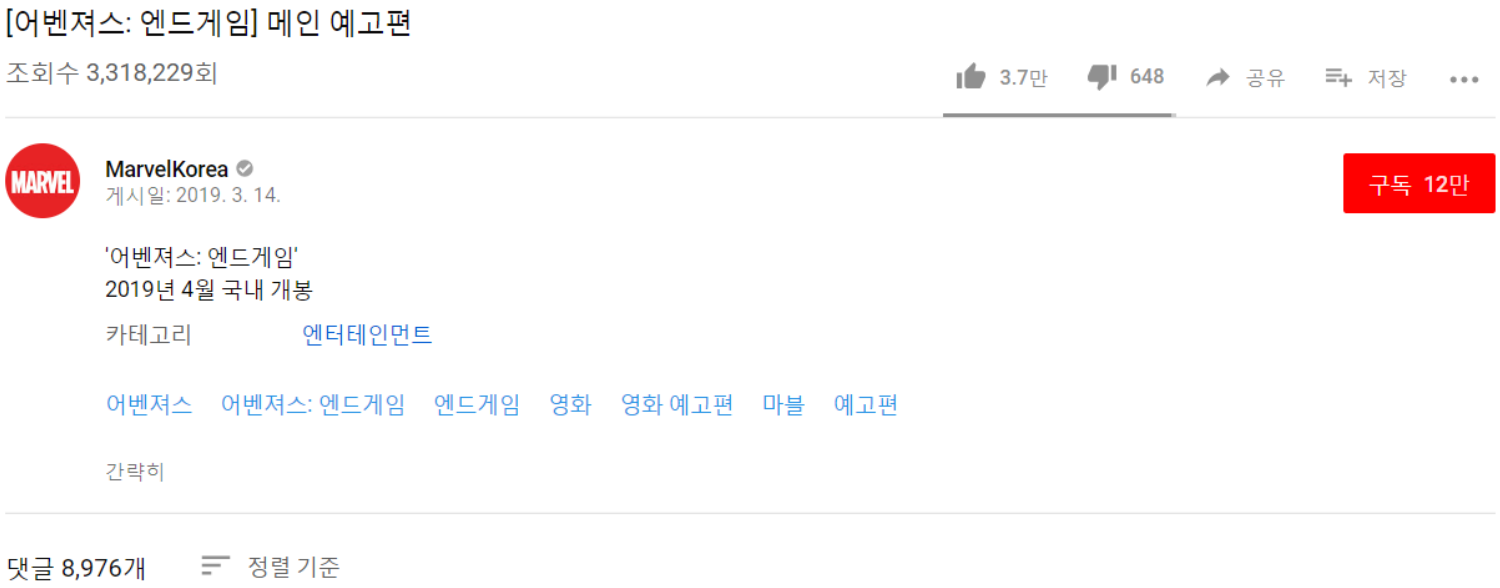
- **웹 크롤링**을 통한 분류 알고리즘의 성능 확인



분석 방향 및 목표

2. 알고리즘 성능 확인

- **웹 크롤링**을 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(" ", "WW|") %>%
      str_split("WW|") %>%
      unlist() %>%
      unique()
    keywords_vector <- append(keywords_vector, keywords)
  }
}
```

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

```
library(KoNLP)
```

```
## 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"
```

분석 방향 및 목표

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.