

# 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



미세먼지가 무서운 이유  
소개해주는 남자 조회수 1만회 · 5일 전  
영화 [인더덕스트] 스토리텔링 리뷰영상입니다. 감사합니다.  
새 동영상



미국 해병대가 뽀센 이유  
소개해주는 남자 조회수 10만회 · 1주 전  
영화 [어퓨굿맨] 스토리텔링 리뷰영상입니다. 감사합니다.



남의 행운을 훔치는 초능력자의 삶  
소개해주는 남자 조회수 96만회 · 1년 전  
영화 [인택토] 줄거리 요약 및 리뷰 영상입니다. 영상에 스포일러 및 결말이 포함되어있으니 시청하실 때 주의해주시길 바랍니다.



낮에는 왕따, 밤에는 천재해커인 소년의 삶  
소개해주는 남자 조회수 67만회 · 5개월 전  
영화 [후엘아이]의 스토리텔링 및 리뷰영상입니다. 결말과 결정적인 스포일러는 포함되어있지 않습니다. 이 점에 있어서 양해부탁 ...

- 국내 유튜브 콘텐츠에 적합한 새로운 카테고리

# 분석 방향 및 목표

## 2. 알고리즘 성능 확인

- 웹 크롤링을 통한 분류 알고리즘의 성능 확인

chrome 웹 스토어 로그인

tags for youtube x

« 홈

☐ 확장 프로그램

☐ 테마

특성

☐ 오프라인 실행 가능

☐ Google 제공

☐ 무료

☐ Android에서 사용 가능

☐ Google 드라이브에서 작동

평점

☐ ★★★★★

☐ ★★★★★ 이상

확장 프로그램

Tags for YouTube™

제공자: M. M. Bos

Years ago, YouTube™ hid video tags from view. This extension puts them right

★★★★★ 599 검색 도구

평가하기

vidIQ Vision for YouTube

제공업체: [vidiq.com](https://vidiq.com)

Uncover the secrets to success behind your favorite YouTube videos.

★★★★★ 7,541 소셜 및 커뮤니케이션

Chrome에 추가

# 분석 방향 및 목표

## 2. 알고리즘 성능 확인

- 웹 크롤링을 통한 분류 알고리즘의 성능 확인

[어벤져스: 엔드게임] 메인 예고편

조회수 3,318,229회

👍 3.7만 💬 648 ➦ 공유 📌 저장 ...



MarvelKorea ✓  
게시일: 2019. 3. 14.

구독 12만

'어벤져스: 엔드게임'  
2019년 4월 국내 개봉

카테고리 [엔터테인먼트](#)

[어벤져스](#) [어벤져스: 엔드게임](#) [엔드게임](#) [영화](#) [영화 예고편](#) [마블](#) [예고편](#)

간략히

댓글 8,976개 ≡ 정렬 기준

# 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, removePunctuation, 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 콘텐츠 특징 파악
- 추천 동영상 제시

# 해결해야 할 문제

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

**감사합니다.**