InfoSearcher

2021 Alpha Project 최종발표

소프트웨어학부 20191632 윤상건
소프트웨어학부 20181656 이민종
소프트웨어학부 20191608 서형빈

Content

1. 프로젝트 소개

기존의 정보 탐색은 “수동적” 이다.

- 평소 인터넷에서 글, 동영상 등의 다양한 매체를 접하는 과정에서 보고 있는 내용에 대해 자세히, 혹은 추가적인 정보를 알고 싶은 경우가 자주 발생
- 새로 탐색하는 과정이 번거롭기도 하며 흐름을 끊기도 함

정보 탐색 과정을 “자동적” 으로 만들 순 없을까?

InfoSearcher

크롬 웹 브라우저에서 사용자가 보고 있는 화면의 글, 이미지, 소리를
인식하여 관련된 정보나 자료를 보기좋게 제공하는 크롬 확장 프로그램

텍스트 분석

현재 사용자가 보고 있는 글과 관련된 네이버 검색 결과 추천

이미지 분석

현재 사용자가 보고 있는 페이지에 존재하는 이미지로 부터 쇼핑 검색어 추천

오디오 분석

현재 사용자가 듣고있는 소리를 처리하여 음악 인식 결과 제공 및 STT를 통한 텍스트 분석 결과 제공

2. Text

목적

현재 사용자가 보고 있는 글과 관련된 정보를 전달

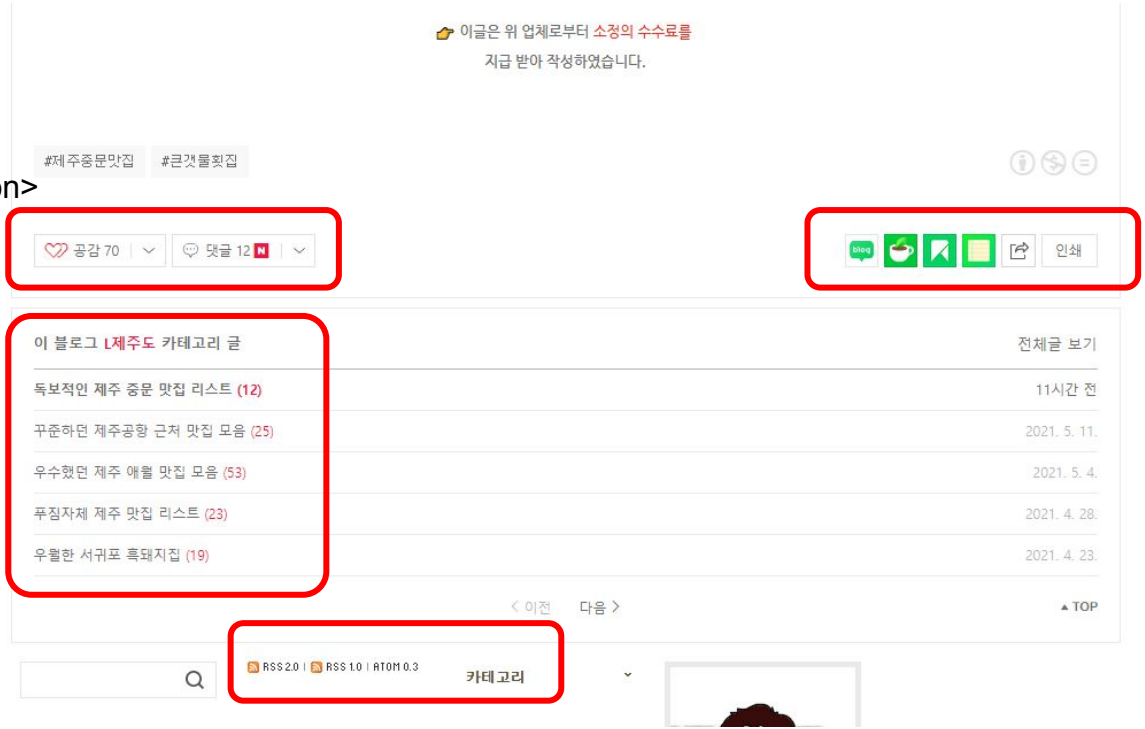
뉴스, 블로그 등의 장문의 글에서 **핵심 텍스트** 추출

↓

해당 텍스트의 중심이 되는 **키워드** 추출

↓

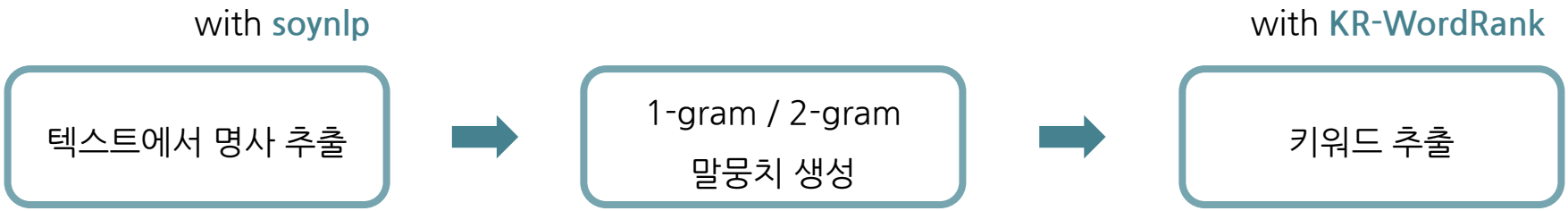
키워드와 관련된 글 **링크** 제공



<button>

<a>

a,
button,
script,
style,
nav,
aside,
footer,
figure,
noscript,
li, ul, ol,
dl, dd, dt,
tr, td, th,



soynlp

- 한국어 자연어 처리 라이브러리

KR-WordRank

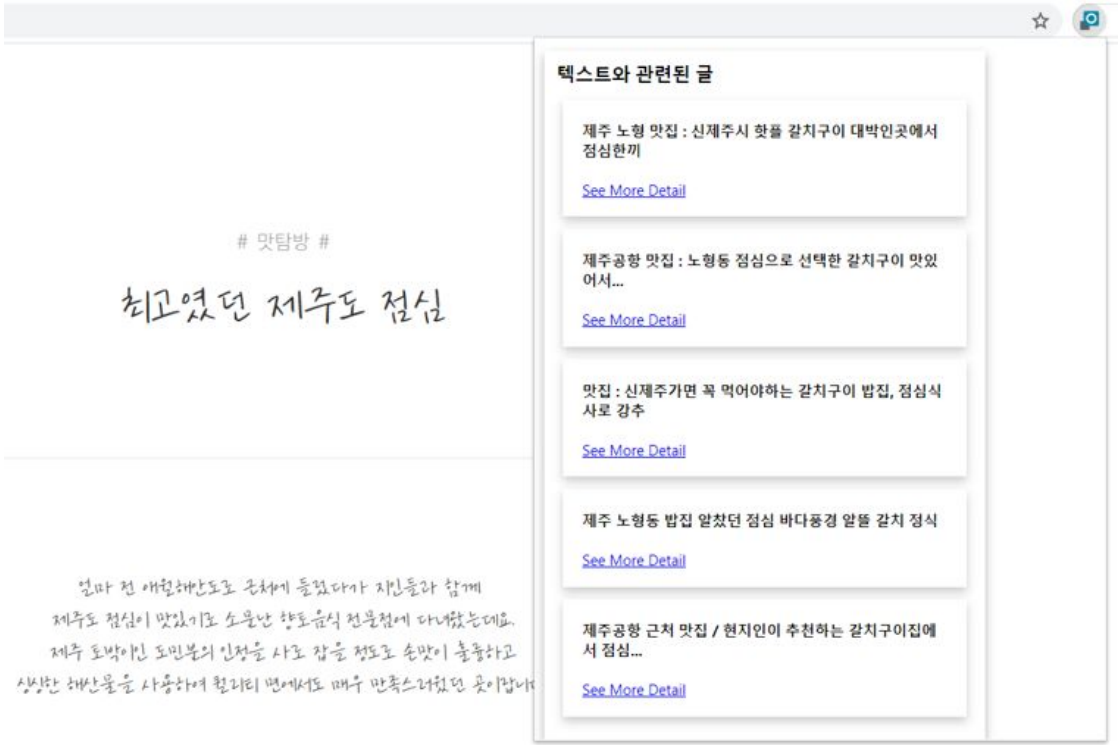
- Ranking 알고리즘 기반 비지도학습 키워드 추출 알고리즘
- 한국어 텍스트에 특화되어 있음.

네이버 API '검색'

- 블로그, 뉴스, 책, 영화, 쇼핑, 백과사전 등 일부 분야에 대해 검색 가능
- 검색 쿼리, 결과 개수, 정렬 가능

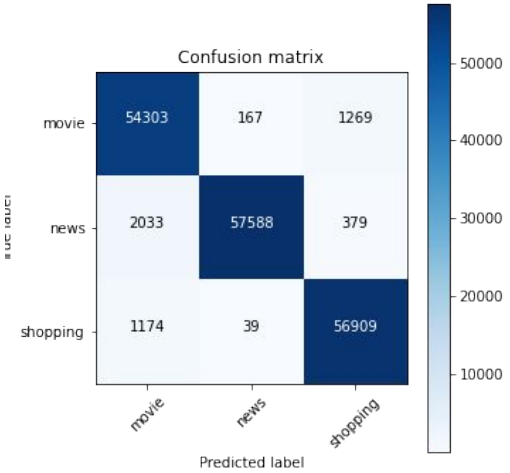
개발

- 추출된 **키워드** 이용
- **블로그**에 대해 검색하여 결과 URL을 **UI**를 통해 사용자에게 전송
- 블로그 뿐만 아니라 뉴스, 책, 쇼핑 등 확장 가능


맛탐방
최고였던 제주도 점심
얼마 전 애월해안도로 근처에 들렀다가 지인들과 함께
제주도 점심이 맛있기로 소문난 향토음식 전문점에 다녀왔는데요.
제주 토박이인 도민분의 인정을 사고 갑을 정도로 손맛이 훌륭하고
싱싱한 해산물을 사용하여 퀄리티 면에서도 매우 만족스러웠던 곳이랍니
텍스트와 관련된 글
제주 노형 맛집 : 신제주시 핫플 갈치구이 대박인곳에서 점심한끼
See More Detail
제주공항 맛집 : 노형동 점심으로 선택한 갈치구이 맛있어서...
See More Detail
맛집 : 신제주가면 꼭 먹어야하는 갈치구이 밥집, 점심식사로 강추
See More Detail
제주 노형동 밥집 알찼던 점심 바다풍경 알뜰 갈치 정식
See More Detail
제주공항 근처 맛집 / 현지인이 추천하는 갈치구이집에서 점심...
See More Detail

텍스트 분야 분류 시도

- 중심 텍스트의 분야(**뉴스, 쇼핑, 영화**)에 대해 **딥러닝**적 접근으로 분류 시도
- 네이버 쇼핑몰 리뷰, 네이버 영화평, 뉴스 본문 데이터 각 **20만개 수집**. (총 60만개)
- **BERT** 이용하여 분류 시도 -> 의미있는 결과 X
- **LSTM** 이용하여 분류 시도
    - 테스트셋에 대해 약 **97%** 의 정확도



테스트 데이터셋 confusion matrix

3. Image

목적

현재 사용자가 보고 있는 페이지에 존재하는
이미지로 부터 쇼핑 검색어 추천

HTML DOM에 접근하여 Image 태그 추출 및 **이미지 추출**

추출한 이미지로 부터 **사물 인식** 진행

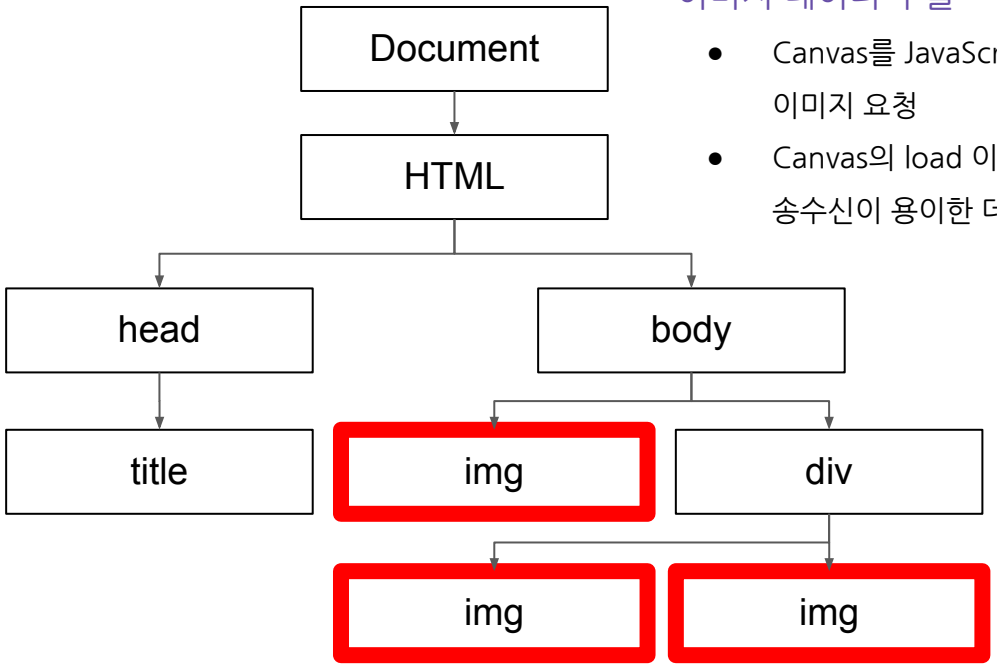
사물 인식 결과로 부터 **검색 키워드 추출**

DOM에서 부터 이미지 태그 추출

- chrome-extension의 in-content는 사용자가 탐색중인 페이지의 DOM 객체에 접근 가능
- document.images 를 활용하여 페이지 내의 모든 이미지 태그 추출

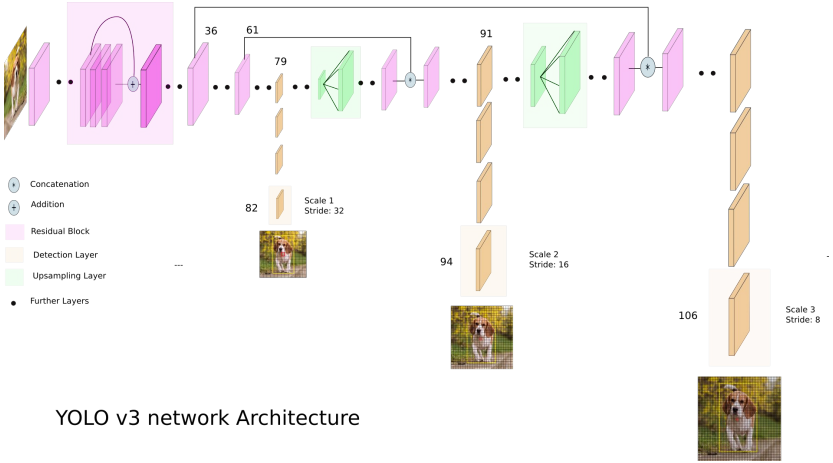
이미지 데이터 추출

- Canvas를 JavaScript 상에서 동적으로 생성하여 이미지 태그로 부터 얻은 이미지 src로 이미지 요청
- Canvas의 load 이벤트가 발생하면 가져온 이미지 데이터를 base64로 인코딩하여 송수신이 용이한 데이터 형태로 변형



현재 이미지 추출 방식 한계 및 개선 방법

- 현재 이미지 추출 방식은 이미지 요청을 웹 페이지 내부에서 처리하기 때문에 Cross-Origin Resource Sharing (CORS) 보안 정책에 의해 가져올 수 없는 이미지가 존재함
- proxy서버를 구성하여 CORS를 회피하여 이미지를 가져올 수 있음



YOLO v3 network Architecture

YOLOv3를 이용한 object detection

- Object detection을 수행하기 위한 모델로 YOLOv3를 사용함
- YOLOv3는 실시간 object detection 기술로, 빠른 속도와 높은 정확도를 보이기 때문에 사용자가 탐색중인 웹 페이지가 변경될 때마다 실행되어야 하는 환경에 적합

인식 데이터 셋

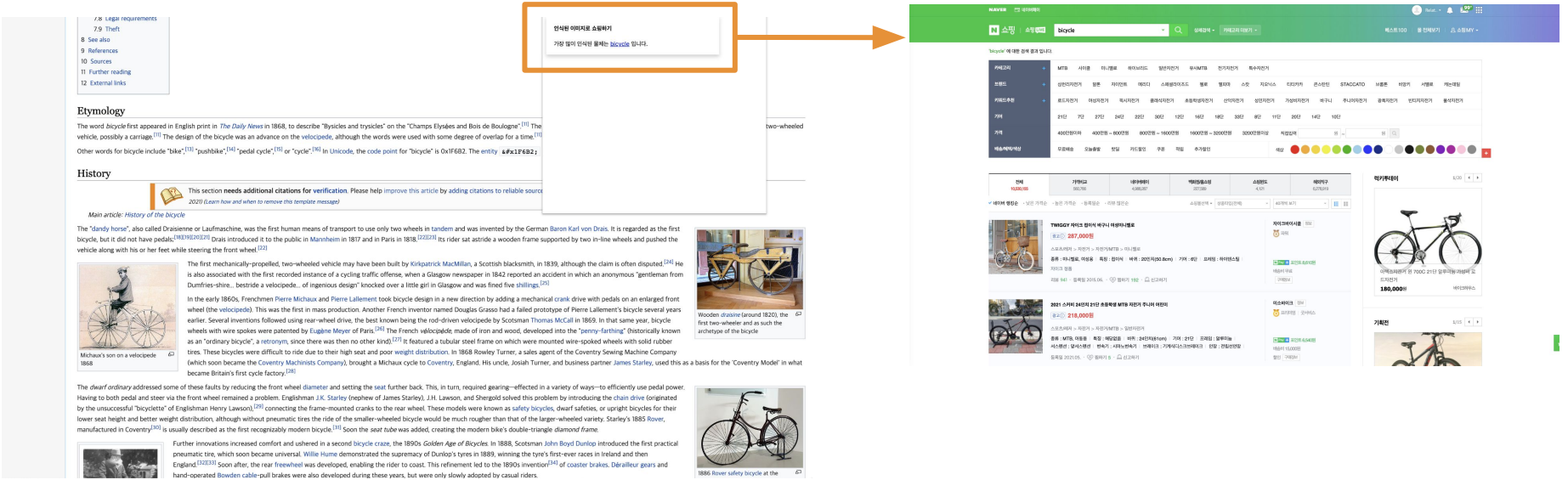
- 80가지의 사물에 대한 라벨링이 되어있는 COCO dataset을 활용했음

사물 인식 결과에서 부터 검색 키워드 추출

- 인식된 결과 중 가장 많이 인식된 사물을 찾아 해당 사물을 검색 키워드로 활용

검색 키워드 활용 방법

- 인식된 사물에 대한 쇼핑 검색을 도와줄 수 있는 결과를 사용자에게 출력
- 만약 웹 페이지에서 자전거에 대한 인식 결과가 높으면, 사용자에게 자전거에 대한 네이버 쇼핑 링크를 안내



4. Audio

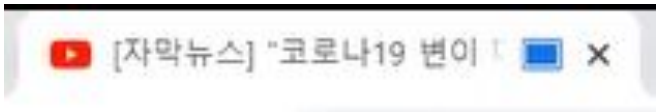
목적

현재 사용자가 듣고있는 소리와 관련된 정보를 전달

뉴스, 유튜브 등의 매체에서 **소리 데이터**를 추출

추출한 소리로 **텍스트, 음악 검색** 진행

검색한 **정보** 제공



TapCapture 기능으로 Audio Capture

- chrome-extension의 background에서 현재 탭의 audio stream 접근

문제점 해결

- 무음 데이터가 너무 많음 - mute 여부 확인
- capture를 시작하면 소리가 안나옴 - audiostream 다시 재생
- 사용자가 원하지 않는 정보, 너무 잦은 api 호출 - 단축키로 녹음 시작 및 종료

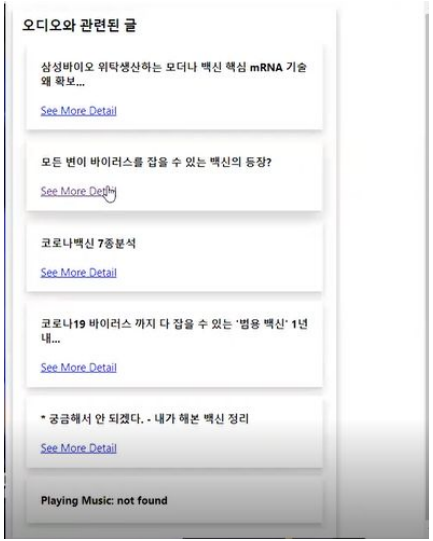
데이터 encoding → with google-cloud-speech: STT 결과 생성 → with naver search api: 관련 정보 검색

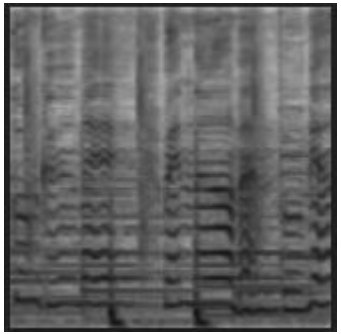
google-cloud-speech

- 구글에서 제공하는 STT api

naver search api

- 네이버 검색엔진을 이용하여 정보 검색
- text 파트에서 자세히 설명





Efficientnet을 이용한 image classifier

- melspectrogram 변환 후 이미지 분류 수행

인식 데이터 셋

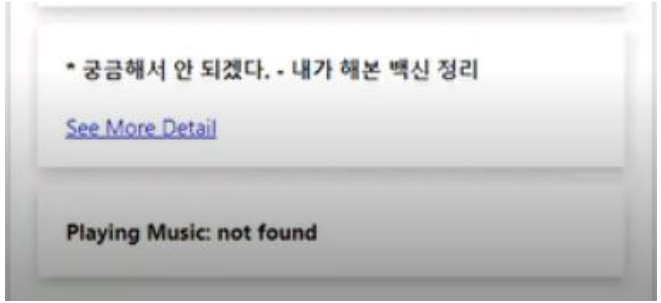
- 최근 3년간 bugs 주간 차트에서 100위 안에 수록된 곡들을 크롤링, mp3파일 수집

Out-of-Distribution 문제

- DB에 없는 음악이거나, 음악이 재생되지 않는 상황도 DB에 있는 특정 음악이라고 분류

문제점 해결

- ODIN 기법을 사용하여 문제점 완화
- hard voting 으로 가능성이 낮은 분류는 제거



5. 예상 효과

5. 예상 효과

- 사용자는 콘텐츠를 소비하여 발생한 **새로운 정보에 대한 탐색 비용을 줄일 수 있다**.
- 자동으로 표시되는 추가 정보를 통해 사용자가 **가짜 정보에 대응할 수 있는 기회**를 만들어 준다.
- 직접 검색하여 탐색하는 수동적인 정보 전달 방식에서 **자동화된 정보 전달**이 수행되면서 사용자에게 **새로운 정보 탐색 기회를 제공**할 수 있다.

Thanks