

캡스톤 프로젝트 최종 보고서



커뮤니티 거래를 위한
제목 변경 웹 확장 프로그램

팀 명: 제몯
20176058 박시현
20176758 박진영
20172634 이민희

1. 프로젝트 개요 및 동기

최근 쉽게 접근할 수 있고, 이용할 수 있는 커뮤니티 상에서의 거래가 활발하게 이루어지고 있다. [그림 1]에서 볼 수 있듯이 중고 거래에서 가장 유명한 커뮤니티인 중고나라(네이버 카페)는 연간 방문자 수가 1억 9000 명에 다름 만큼 활성화되어 있다. 커뮤니티 상에서의 거래는 판매자의 거래 글 등록, 구매자의 거래 요청과 같은 순서로 이루어진다. 구매자는 게시글에서 자신이 원하는 상품을 찾기 위해 많은 탐색 과정을 거친다. 하지만, 거래 글 작성자가 다양하다 보니 글 제목이 매우 다양하다. 또, 글 제목은 일반 판매 글과 같은데 내용은 광고성 글이거나 매입 글인 경우, 글 제목과 본문 내용이 다른 경우도 상당하다. 정상적인 판매 글이더라도 글 제목에 중요한 정보, 심지어는 상품명까지 명시되어 있지 않거나 글 제목에 너무 많은 내용이 있어 가독성이 떨어지는 불편함이 있다. 이를 해결하기 위해 검색어의 단어 수를 늘리거나 검색 조건을 더 좁게 설정하곤 하는데, 이러한 방법은 검색 결과의 수를 줄여 선택의 폭을 감소시킨다.

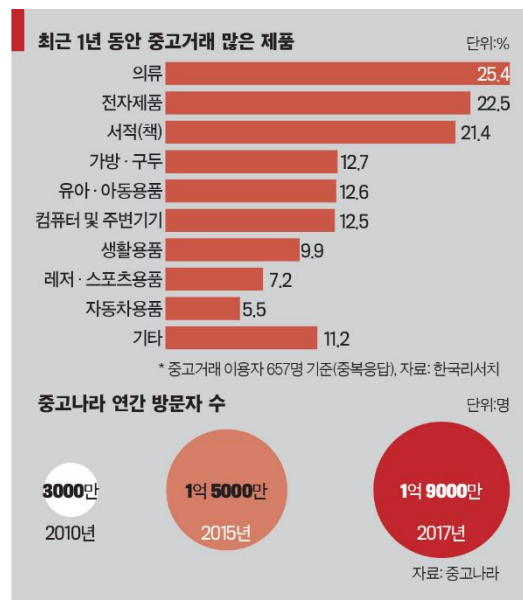


그림 1 중고나라 연간 방문자 수

564542664	아이폰X 실버 64,256기가 두대있습니다 개인서둘직거래만합니다	dlrbdnjs56	18:27	2
564542243	아이폰X 실버 64,256기가 두대있습니다 개인서둘직거래만합니다	dlrbdnjs56	18:26	3
564542185	아이폰x 64기가 팝니다	폴드스펠	18:25	4
564542092	[역삼] 아이폰 X 64기가 그레이 리퍼 10월 26일까지	DereK	18:25	2
564541919	삼니다))) 가게통 박스폰데일 아이폰x 아이폰xs 아이폰xs max 노트9 노트8 갤럭시9+ 갤럭시9 수도권출장매입합니다	Dheh2	18:25	6
564541808	아이폰X 실버 64,256기가 두대있습니다 개인서둘직거래만합니다	dlrbdnjs56	18:24	4
564541790	아이폰X 64G 스페이스그레이 외관완전A급 배터리성능 88프로 아주 저렴하게 판매합니다(강북직거래/전국택배)	중나 최고회원	18:24	5
564541332	아이폰X 실버 64,256기가 두대있습니다 개인서둘직거래만합니다	dlrbdnjs56	18:23	2
564541185	아이폰x 64G8 스페이스그레이	pinkmobile	18:22	3
564541143	9H 아이폰 고강도글라스 아이폰젤백케이스 아이폰7 아이폰8 아이폰7플러스 아이폰8플러스 아이폰x 아이폰xs 아이폰xs max 아이폰xr 강화유리필름	굿글라스	18:22	0
564541124	[역삼] 아이폰 X 256기가 그레이 리퍼 만료	DereK	18:22	3
564540812	가게통 박스폰 새폰삼니다))) 아이폰x 아이폰xs 아이폰xs max 갤럭시s9 노트9 서울경기천지역 출장 매입합니다	포르리	18:21	0

1 2 3 4 5 6 7 8 9 10 다음 >

전체기간 ▼ 제목+내용 ▼ 검색어를 입력해주세요 검색

그림 2 중고나라 게시글 제목

2. 프로젝트 목표

[그림 2]와 같이 중고거래 게시판의 게시글 제목은 한눈에 알아보기 힘들다. 위와 같이 많은 사용자의 불편함을 줄이기 위해 커뮤니티의 거래 게시판 목록의 글 제목을 사용자가 원하는 형태로 한 번에 변환해주는 프로그램을 개발한다.

구매자 측면에서는 게시판 목록의 제목만으로 원하는 글의 본문 내용을 파악 가능하고, 정해진 형태로 제목이 변경되기 때문에 쉽게 비교를 할 수 있다.

판매자 측면에서는 다른 사용자가 올린 양식을 보고 어떻게 제목을 지을지, 게시물에 어떤 내용을 담을지 고민하는 과정을 줄일 수 있다.

거래 커뮤니티 구매자와 판매자 모두의 편의성을 크게 증가시키는 것이 목표이다.

3. 페르소나

다음은 프로젝트의 대상이 되는 가상의 인물들과 프로젝트가 제시하는 각 상황 별 해결방안을 나열한 것이다.

- 윤연경 (나이: 22, 여, 직업: 대학생, 웹이용능력: 상)

대학교 수업을 들을 때 필기용으로 사용할 아이패드 구매를 예정에 두고 있다. 용돈을 받는 대학생이라 새 상품보다 저렴한 중고 제품으로 구매하고 싶어 중고거래를 하는 커뮤니티를 탐색 중이다. 그런데, 게시판에 “아이패드”를 검색했을 때 게시글 제목만 보고 아이패드와 관련 없는 글, 광고성 글, 원하는 조건이 아닌 글이 선택되는 경험을 많이 했다. 또, 게시글마다 원하는 조건의 정보들을 다 따로 기록해야 해 시간이 오래 걸려 구매가 하루하루 늦춰지고 있는 상황이다.

- A. 원하는 키워드로 보이게 변경할 시 게시글 제목만 보고 본문을 읽을 가치가 있는지, 광고이거나 상품에 대한 정보가 있는지를 더 정확하게 판단할 수 있다.
- B. 같은 상품에 대한 조건을 제목만 보고도 알 수 있어 게시글 목록을 보고 다양한 상품을 비교할 수 있다.

- 김민철(나이: 56, 남, 직업: 은행원, 웹이용능력: 중)

매일 앉아서 근무하는 김민철 씨는 나이가 들수록 운동의 필요성을 느낀다. 퇴근 후에도 집에서 쉽게 운동하며 습관을 들이기 위해 집에 둘 만한 운동기구를 찾고 있다. 운동기구가 생각보다 비싸 중고로 알아보기로 결심을 하고 중고 거래에서 제일 유명한 커뮤니티라 할 수 있는 중고나라에서 알아보기 시작했다. 운동기구에 대해 잘 알지 못해 각 운동기구 종류와 가격대가 어느 정도인지 알지 못하는데 게시글에 들어갔다가 가격을 보고 놀라 나오기를 반복한다. 운동기구명이나 종류와 함께 가격도 제목에 같이 써준다면 이런 수고를 덜어줄 것 같다고 생각해 ‘가격’이라는 단어로 검색했지만 ‘가격’이란 단어가 제목이나 본문에 있는 게시글 수가 적어 선택의 폭이 좁아지는 점도 마음에 들지 않는다.

- A. 검색어 추가, 상세조건 검색으로 선택의 폭이 좁아지는 불상사를 막을 수 있다.
- B. 제목만 보고도 원하는 조건에 맞는지 판단할 수 있다.

- 이준호(나이: 30, 남, 직업: 공무원, 웹이용능력: 상)

최근 공무원 시험을 통과한 이준호 씨는 발령지로의 이사를 앞두고 불필요한 물건 정리에 나섰다. 시간이 촉박해 여러 커뮤니티에 중고 거래 글을 게시하고 있는데도 관심을 보이는 사람이 없다. 사람들이 관심을 가질 만하게 제목을 써보려 하지만 과장된 제목은 제재를 받을 수 있고, 하루에 같은 글을 몇 번 이상 올려도 제재를 받는 규칙이 있는 커뮤니티도 있어 혼란스럽다. 사람들이 원하는 조건을 제목에 넣고 싶지만, 사람들마다 원하는 조건이 다 다르고 그것을 모두 넣기엔 제목이 길어져 가독성이 떨어

어지는 것이 난관이다. 어떻게 제목을 결정해야 하는지 중고 판매의 고수의 특별 과외라도 받고 싶은 심정이다.

- A. 판매자가 제목을 고려하지 않아도 구매자는 원하는 조건이 담긴 제목을 볼 수 있다.
- B. 판매자는 본문에 정보를 나열하기만 하면 제목을 고민할 필요가 없다.

4. 개발 배경과 구현 내용

4.1 개발 배경

- 검색, 정보 처리, 거래 등 여러 목적으로 많은 사람들이 사용하는 웹 브라우저

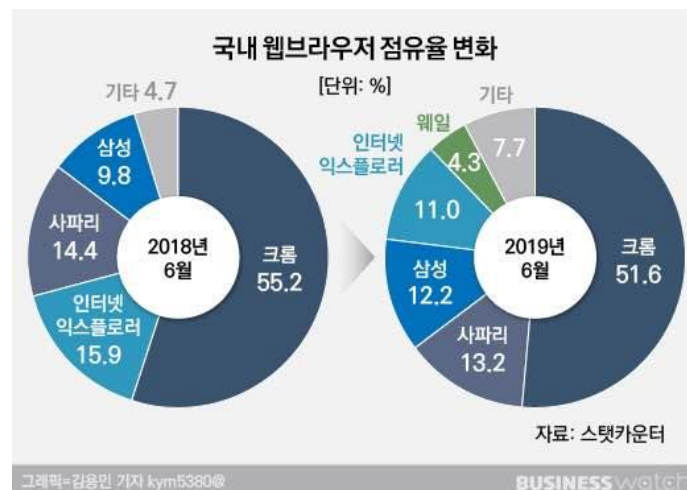


그림 3 웹 브라우저 점유율

많은 사람들에게 웹을 사용할 때 웹 브라우저의 사용은 뭘 해야 뭘 수 없다. [그림 3]처럼 여러 브라우저들이 있고, 브라우저들은 단순히 html 문서나 파일을 출력해주는 것 이외에도 북마크, 즐겨찾기 등을 제공하고, 그 외에도 다양하고 편리한 기능을 사용자에게 제공하기 위해 여러 확장 앱을 제공한다. IE Tab, Google 번역 등이 그 예이다. 웹 기능에 불편함이 있어 원하는 기능을 추가하려면 이러한 확장 앱으로 구현을 하는 것이 적합하다고 판단하였다.

4.2 개발 환경

- Python 3.7
- Html / Css / Javascript

4.3 구현 사항

- 원하는 제목 형태 키워드 선택 기능

가능한 구분 키워드를 카테고리에 따라 나누어 보여주어 간편하게 제목 형식을 만들 수 있게 한다.

1. 키워드 버튼 선택 기능
2. 선택된 키워드를 보여주는 창 내에서 드래그 앤 드롭으로 순서 변경 기능
3. X 표시 클릭을 통한 삭제 기능
4. 카테고리에 따른 버튼 활성화 / 비활성화 기능

- 선택한 형태로 제목 변경 기능

키워드를 다 선택하고 변경 버튼을 누르면 다음의 일련의 과정을 한다.

1. 게시글 목록에 있는 각 게시글 페이지 html 을 크롤링
2. 크롤링한 정보들을 자연어 처리해 키워드에 따른 값들을 추출
3. 키워드와 추출된 값들을 조합해 새로운 형식의 제목으로 변경
4. 변경 후 제목에 커서 올릴 시 이전 제목 나타내기

5. 상세 개발 내용

5.1 서버

- 데이터 크롤링

대상 카테고리인 옷과 IT 에 대한 자연어 처리를 진행하기에 앞서 기존 데이터를 수집하고 학습하는 과정이 필요해 데이터 크롤링을 진행하게 되었다. 크롤링 대상은 중고나라 게시판으로써 옷과 IT 에 관련한 카테고리를 선별했다. 선별한 옷 카테고리는 [그림 4]와 같이 여성상의, 여성하의, 여성신발, 남성상의, 남성하의, 남성신발이고 IT 카테고리는 [그림 5]와 같이 노트북, 핸드폰(SKT, KT, LG U+), 태블릿 PC 이다.

여성의류잡화	남성의류잡화
 여성상의 	 남성상의 
 여성하의 	 남성하의 
 여성신발 	 남성신발 

그림 4 옷 카테고리

모바일	컴퓨터
☑ 스마트폰 고가 매입	☑ 컴퓨터 자유게시판 N
☑ 중고나라 중고폰	☑ 노트북/맥북/넷북 N
☑ SKT N	☑ 노트북-미개봉상품
☑ SKT-미개봉상품	☑ 태블릿PC N
☑ KT N	☑ 태블릿PC-미개봉
☑ KT-미개봉상품	
☑ LGU+ N	
☑ LGU+-미개봉상품	

그림 5 IT 카테고리

언어는 python 을 사용했고 크롤링 시 필요한 beautifulsoup, 네이버 로그인에 필요한 selenium 을 사용했다. 네이버 카페의 경우 중고거래 게시판을 사용하기 위해서는 네이버에 로그인이 되어 있어야 하고, 카페에 가입이 되어 있어야 한다. 최근 네이버 보안 정책인 네이버캡차를 우회하기 위해 클립보드에 아이디와 비밀번호를 임시저장한 후 키 조합을 통해 불러오는 방식을 사용했다. 한 카테고리당 게시글이 1000 페이지까지 있는 것을 감안해 한 번에 최대 만개의 게시글을 크롤링했다. 카테고리별로 업데이트되는 주기가 천차만별이라 크롤링 주기를 각각 다르게 해주었다. 빠르면 하루, 늦으면 한 달에 걸쳐서 1000 페이지가 리셋됨을 확인할 수 있었다. 빠르게 업데이트되는 카테고리의 경우에는 크롤링을 진행하면서도 게시글이 실시간으로 올라올 것이라고 판단되어 겹치는 데이터가 없도록 1000 페이지를 시작으로 뒤에서부터 크롤링을 진행해왔다.

☑ 여성상의.txt	☑ 남성상의.txt
☑ 여성상의2.txt	☑ 남성상의2.txt
☑ 여성상의3.txt	☑ 남성상의3.txt
☑ 여성상의4.txt	☑ 남성상의4.txt
☑ 여성신발.txt	☑ 남성신발.txt
☑ 여성신발2.txt	☑ 남성신발2.txt
☑ 여성하의.txt	☑ 남성신발3.txt
☑ 여성하의2.txt	☑ 남성하의.txt
	☑ 남성하의2.txt
	☑ 남성하의3.txt

그림 6 옷 카테고리 데이터셋

KT.txt	태블릿PC.txt
LGU.txt	태블릿PC2.txt
SKT.txt	태블릿PC3.txt
노트북.txt	태블릿PC4.txt
노트북2.txt	태블릿PC5.txt
노트북3.txt	

그림 7 IT 카테고리 데이터셋

IT 카테고리의 경우 카테고리 특성상 광고글이나 매입 글이 많아 본문에 광고성 글이 지나치게 많은 것을 볼 수 있었는데 이러한 데이터는 학습에 좋지 않은 영향을 끼칠 것이라고 판단하여 본문의 글자 수를 세서 지나치게 긴 글은 광고성 글이라고 판단하고 크롤링을 진행하지 않았다. 또한, 한 작성자가 같은 글을 여러 번 올리는 등과 같은 학습에 악영향을 줄 수 있는 요소들을 피하기 위해 한 페이지의 15 개의 게시물 중 랜덤으로 10 개의 게시글을 크롤링했다. 여러 예외 상황들의 경우에 크롤링을 진행하지 않다 보니 1000 페이지의 게시물들을 크롤링해도 만개를 채우지 못하는 경우도 있었다. 이러한 경우에는 일정 시간이 지난 후 새로 업데이트되는 게시물에 대해 추가적으로 크롤링을 진행했다. [그림 6]과 [그림 7]에서 볼 수 있듯이 데이터 셋을 총 29 만 개 확보했다.

- 웹서버

python 기반의 Django 웹 프레임워크를 사용하여 모듈을 연결하는 웹서버를 구축했다. 프론트엔드단에서 dictionary 형태의 json 값을 넘겨주면 그 json 값을 dictionary 로 받는다. 그다음 dictionary 를 파싱 하여 필요한 값인 제목, 가격, 본문을 list 로 저장한 후 카테고리나 선택된 키워드들과 함께 자연어 처리 함수로 넘겨준다. 자연어 처리 함수에서 list 형식으로 받은 결과값과 키워드를 조합해 새로운 제목을 생성한다. 프론트엔드로 다시 값을 넘겨줄 때는 게시글 번호와 새로운 제목을 dictionary 형태의 json 값으로 전달한다.

django 의 views.py 에 request 가 들어오면 위의 과정이 실행되도록 구현하였고 값을 넘겨주는 방식은 POST 로 결정하였다. 통신 테스트는 Postman 을 사용하여 테스트했다. 예시 데이터 여러 개를 사용해 로컬 서버와 aws 서버 통신도 이루어지는지 확인했다. [그림 8]은 로컬 서버인 <http://127.0.0.1:8000/server/>로 POST 요청을 보냈을 때 예시 데이터를 받아 새 제목을 생성해서 성공적으로 반환을 해줌을 확인할 수 있다. [그림 9]은 같은 방식으로 aws 서버인 <http://15.164.195.30/server>로 POST 요청이 들어왔을 때 성공적으로 값을 반환해줌을 확인할 수 있다.

POST http://127.0.0.1:8000/server/

Params Authorization Headers (9) **Body** Pre-request Script Tests Settings

none form-data x-www-form-urlencoded raw binary GraphQL BETA JSON ▼

```

1 {
2   "category": "남성상의",
3   "keyword": "[\"이름\", \"브랜드\", \"가격\", \"색상\", \"거래방식\", \"착용횟수\", \"사이즈\"]",
4   "crawlingData": [{"number": "674586330", "title": "아르캐니트 일괄급쳐", "price": "70,000원", "content": "\n\n두트 합쳐서 택배7에 팝니다. 그래요, 네이 안있었습니다. 얼마이고 둘다 택배중입니다 구매원하시면 연락처 남겨주세요\n"}, {"number": "674586077", "title": "파타고니아 신질라", "price": "\n\n\n사이즈 s이나 m사이즈 정도 됩니다. 거의 새상품입니다. 010 35팔구 43팔구"}, {"number": "674586280", "title": "파타고니아 레트로x 팝니다.", ": \n\n파타고니아 레트로x 팝니다.19년 국내 매장판 제품입니다.\n네추럴 S // 카키 M // 엘리컨 M // 블랙 S, L 있습니다.\n운포 31입니다.\n010-233 : 674586362", "title": "파타고니아 레트로x 팝니다.", "price": "310,000원", "content": "\n\n파타고니아 레트로x 팝니다.19년 국내 매장판 제품입니다.\nM // 블랙 S, L 있습니다.\n운포 31입니다.\n010-2335-5988연락주세요\n"}, {"number": "674586342", "title": "[9천원]강골(KANGOL) 후드티 블랙 M(95~ ,000원", "content": "\n\n\n강골 후드티입니다\n블랙 색상 M사이즈입니다\n평소 95~100입으시면 잘 맞습니다\n마지막으로 세탁 한 번 하고 작은 사진 업로드할게요 하시는 분은 편하게 연락을 부탁드립니다 감사합니다. 꼭 확인하시고 구매하시길 바랍니다"}]
```

Body Cookies Headers (7) Test Results Status: 201 Created Time: 7.42s

Pretty Raw Preview Visualize BETA JSON ≡

```

11      "number": "674586280",
12      "newTitle": " | 이름 : 카키 엘리컨 | 브랜드 : 파타고니아 | 가격 : 310000 | 색상 : ? | 거래방식 : 택배 | 착용횟수 : ? | 사이즈
13    },
14    {
15      "number": "674586362",
16      "newTitle": " | 이름 : 카키 엘리컨 | 브랜드 : 파타고니아 | 가격 : 310000 | 색상 : ? | 거래방식 : 택배 | 착용횟수 : ? | 사이즈
17    },
18    {
19      "number": "674586342",
20      "newTitle": " | 이름 : 후드 티 | 브랜드 : 강골 | 가격 : 9000 | 색상 : 후드티입니다|블랙색 | 거래방식 : 택배+직거래 | 착용횟수 : 한번
21    },
22    {
23      "number": "674586337",
24      "newTitle": " | 이름 : 파이버덕 카모 저지 | 브랜드 : 카모 | 가격 : 54000 | 색상 : ? | 거래방식 : 택배+직거래 | 착용횟수 : 아래
```

그림 8 로컬 서버 통신 확인

POST

http://15.164.195.30/server/

Params

Authorization

Headers (9)

Body

Pre-request Script

Tests

Settings

none

form-data

x-www-form-urlencoded

raw

binary

GraphQL

BETA

JSON

```

1 {
2   "category": "남성상의",
3   "keyword": "[\"이름\", \"브랜드\", \"가격\", \"색상\", \"거래방식\", \"착용횟수\", \"사이즈\"]",
4   "crawlingData": [{"number": "674586330", "title": "아르캐니트 일괄급쳐", "price": "70,000원", "content": "\n두니트 합쳐서 택포7에 합니다. 그레이,네이
안였습니다. 염사이고 출다 택보유중입니다 구매원하시면 연락처 남겨주세요\n"}, {"number": "674586077", "title": "파타고니아 신질라", "price":
: "\n\n사이즈 S이나 M사이즈 정도 됩니다. 거의 새상품입니다. 010 35말구 43말구"}, {"number": "674586280", "title": "파타고니아 레트로x 합니다.",
: "\n\n파타고니아 레트로x 합니다.19년 국내 매장판 저품입니다.\n네추럴 S // 카키 M // 펠리컨 M // 블랙 S, L 있습니다.\n운포 31입니다.\n010-2338
: "674586362", "title": "파타고니아 레트로x 합니다.", "price": "310,000원", "content": "\n\n파타고니아 레트로x 합니다.19년 국내 매장판 저품입니다.\n
M // 블랙 S, L 있습니다.\n운포 31입니다.\n010-2335-5988연락주세요\n"}, {"number": "674586342", "title": "[9천원]강골(KANGOL) 후드티 블랙 M(95~
00원", "content": "\n\n강골 후드티입니다.\n블랙 색상 M사이즈입니다.\n핑크 95~100입으시면 잘 맞습니다.\n마지막으로 세탁 한 번 하고 찍은 사진입
짜고 후기 남겨주세요. 반품요구 내함으로 인한 환불자의 정해정도를 주고 이어서 받되자한 수 없습니다. 이 점 꼭 이해하시고 구매하시면 됩니다.

```

Body

Cookies

Headers (7)

Test Results

Status: 201 Created

Time: 1107ms

Pretty

Raw

Preview

Visualize

BETA

JSON

```

30 {
31   "number": 674586300,
32   "newTitle": " | 이름 : 자켓 오트밀 | 브랜드 : 파타고니아 | 가격 : 99000 | 색상 : 후리스입니다오트밀색 | 거래방식 : 택배+직거래 | 착용횟수
33 },
34 {
35   "number": 674586164,
36   "newTitle": " | 이름 : 데님 | 브랜드 : 아크 | 가격 : 130000 | 색상 : 네이비색 | 거래방식 : 택배 | 착용횟수 : 사용안함 | 사이
37 },
38 {
39   "number": 674586362,
40   "newTitle": " | 이름 : 카키 펠리컨 | 브랜드 : 파타고니아 | 가격 : 310000 | 색상 : ? | 거래방식 : 택배 | 착용횟수 : ? | 사이즈
41 },

```

그림 9 aws 서버 통신 확인

5.2 프론트엔드

- 확장앱

사용자가 확장 앱을 사용할 부분을 html, css, javascript 및 jquery 를 이용해 구현하였다. 웨일 브라우저에서 지원하는 사이드바를 이용해 확장 앱을 사용할 수 있으며, 우측 배너에 위치한 ‘제몫’의 아이콘을 클릭하면 [그림 10]과 같은 ‘제몫’의 사이드바를 볼 수 있다.



그림 10 확장앱 사이드바

키워드는 접속한 카테고리에 따라 비활성화된다. 예를 들어, 현재 접속한 카테고리가 노트북 카테고리라면 의류 키워드인 ‘착용횟수’, ‘사이즈’ 키워드가 비활성화된다. 의류나 IT/전자기기가 아닌 다른 카테고리에 접속한다면 공통 키워드를 제외한 키워드가 모두 비활성화된다.

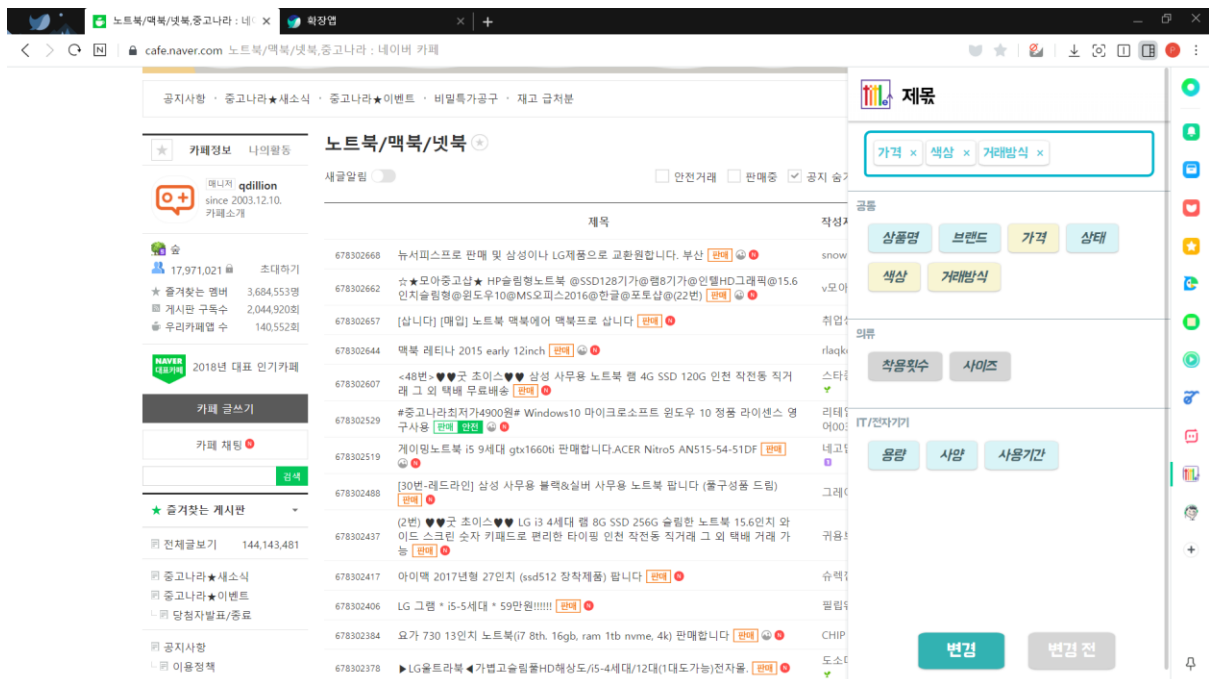


그림 11 의류 키워드 비활성화

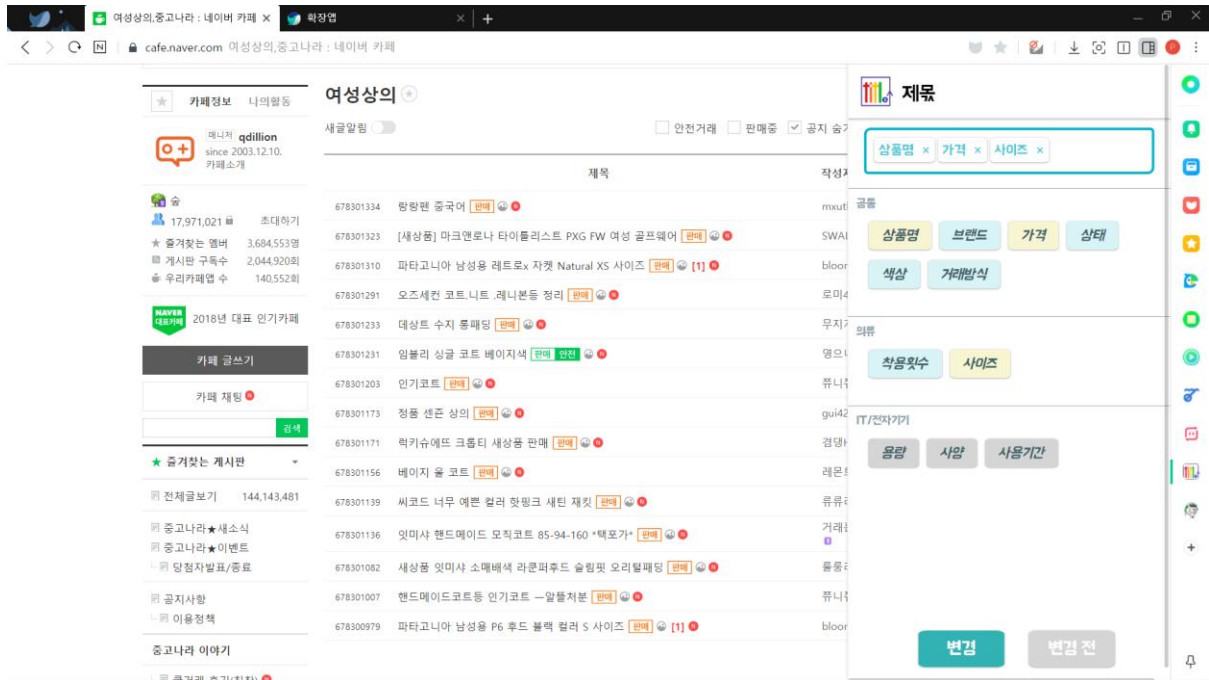


그림 12 IT/전자기기 키워드 비활성화

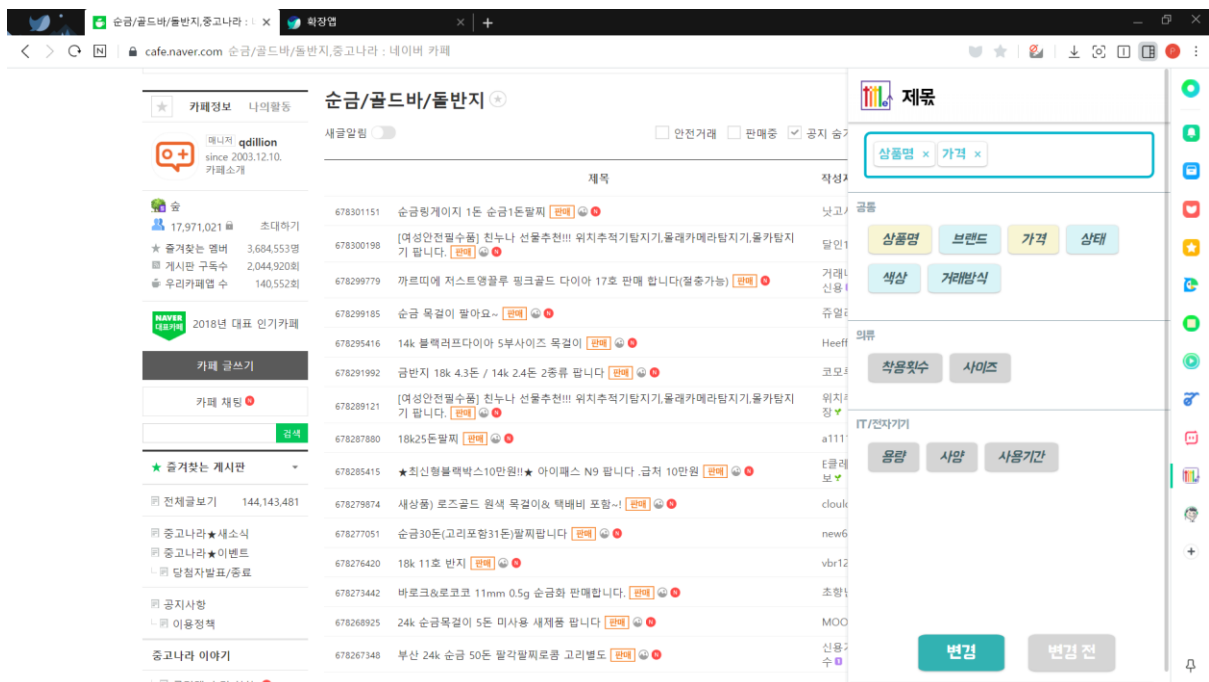


그림 13 공통 키워드 제외 비활성화

또한, [그림 14]처럼 키워드를 선택 및 삭제할 수 있으며 드래그 앤 드롭을 통해 키워드의 순서를 바꿀 수 있도록 구현하였다.



그림 14. 키워드 선택(왼) 키워드 순서 변경(오)

변경 버튼을 클릭하면, 알림 창이 뜬 후 로딩이 시작되며 서버와의 통신을 통해 변경 제목 데이터를 전달받으면 로딩이 멈추고 사이드바가 자동으로 닫히게 된다.

변경 전 버튼을 클릭하면 변경되었던 제목이 원래 상태로 되돌아간다. 변경 전 버튼은 현재 접속 중인 페이지에 변경된 제목이 있을 때에만 활성화된다.

- 콘텐츠 스크립트

제목을 변경하는 것과 같은 작업은 확장 앱 상이 아닌 제목이 변경될 사이트 상에서 하여야 한다. 따라서 확장 앱에선 콘텐츠 스크립트를 통해 이러한 작업을 가능하게 한다. 콘텐츠 스크립트¹란 웹 페이지에 파일을 삽입하는 것을 말하며, CSS 파일 또는 JS 파일을 추가할 수 있다.

위치한 페이지의 게시글 제목과 게시글 각각의 본문 내용을 크롤링하는 작업, 서버와의 통신, 통신 결과(변경된 제목)를 페이지 상의 원래 제목과 대치하는 작업, 변경 요청과 변경 완료 상태를 확장 앱과 메시지로 주고받는 작업 등을 삽입된 JS 파일 상에서 구현하였다. 또한, jquery 의 메소드와 툴팁 기능을 사용하기 위해 jquery 파일과 툴팁의 스타일 적용을 위해 CSS 파일도 삽입해주었다. 파일을 삽입할 사이트가 iframe 으로 내부 프레임이 존재하여 원하는 부분이 제대로 인식되지 않는 문제가 발생하였다. 이는 manifest 에 삽입될 프레임을 전체로, 삽입될 시점을 페이지가 다 로딩되었을 때로 설정해주어 문제를 해결하였다.

크롤링은 jquery 의 \$.post()를 사용하여 페이지 접속 시 바로 진행된다. 중고나라의 경우 제목과 함께 가격이 적혀 있어 가격도 함께 크롤링하여 전달한다. 결과적으로

¹ 콘텐츠 스크립트, Whale developers,

https://developers.whale.naver.com/getting_started/anatomy_1/#%EC%BD%98%ED%85%90%EC%B8%A0-%EC%8A%A4%ED%81%AC%EB%A6%BD%ED%8A%B8

게시글 번호, 제목, 가격, 본문을 크롤링하여 서버에 전달한다. 개발자 도구(F12)를 사용해 콘솔에 크롤링한 데이터를 출력한 결과는 다음과 같다.

```
▼ (15) [{"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}]
▶ 0: {number: "678401561", title: "우락군 토틀링 95사이즈(M) 팝니다. 호치민주면 중정 실직3회", price: "140,000원", content: "우락군 토틀링 팬이어서 샀었는데 추위를 너무 안타서 작년엔 3번 입었고 말..."}
▶ 1: {number: "678401373", title: "[ 새상품 ] 디스커버리 찰튼 패딩 (옐로우)", price: "80,000원", content: "디스커버리 찰튼 덕다운 패딩 옐로우 색상입니다. 완전 새상품이며 색도 그대로 달려있습니다..."}
▶ 2: {number: "678401521", title: "칼하트 후리스 M 새상품 싸게 팝니다.", price: "55,000원", content: "칼하트 후리스 M 새상품 싸게 팝니다. 정품문인x e o 발 제품입니다2주걸려 받았으나 사이...건비용..."}
▶ 3: {number: "678401637", title: "[ 새상품 ] FILA 휠라/필라 데이피 패딩", price: "100,000원", content: "휠라/필라 데이피 다운 패딩입니다. 색상은 화이트 / 완전 새상품 / 색도 그대로 있음/사이즈 100..."}
▶ 4: {number: "678401199", title: "타임즈 패딩 코트 105", price: "79,900원", content: "브렌드 : 타임즈/종류 : 패딩 코트/색상 : 블랙/표기사이즈/실사이즈 : 2/105/...는 만원이상 물품 2개이상 구..."}
▶ 5: {number: "678401498", title: "모노소잉(헨스) / 윈터 셋업자켓 블랙 / 50", price: "60,000원", content: "작년에 완판된 핸드메이드 더플입니다. 윈터 셋업자켓 블랙 / 50 / 윈터 셋업자켓 블랙 / 50 / 윈터 셋업자켓 블랙 / 50..."}
▶ 6: {number: "678401104", title: "준지 18fw 핸드메이드 더플코트 46사이즈", price: "900,000원", content: "작년에 완판된 핸드메이드 더플입니다. 준지는 오버핏이라 취향에 따라 사이즈 선택하시면 됩..."}
▶ 7: {number: "678401493", title: "*세일* 노스페이스 슈퍼에어다운 정품 판매.smx1859095100105", price: "265,000원", content: "미작품 "새상품 정품" 노스페이스 플 패딩. * 2019년 FW 국내 매장 정..."}
▶ 8: {number: "678401457", title: "통글래어 자수패치 트레이닝복 상하의세트 새제품 전사이즈 판매합니다.", price: "82,000원", content: "통글래어 자수패치 트레이닝복/색상: 차콜/사이즈: 95 . 100 . ..."}
▶ 9: {number: "678401459", title: "마마가리 플 로딩 브라운 48", price: "200,000원", content: "마마가리 플 로딩 브라운 48 / 플 로딩 브라운 48 / 플 로딩 브라운 48 / 플 로딩 브라운 48..."}
▶ 10: {number: "678401108", title: "[L] 정품 스루시 랑오버 아노락 블랙", price: "140,000원", content: "정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙..."}
▶ 11: {number: "678401657", title: "디키즈점퍼(백포2만원)", price: "20,000원", content: "디키즈점퍼(백포2만원) / 디키즈점퍼(백포2만원) / 디키즈점퍼(백포2만원) / 디키즈점퍼(백포2만원)..."}
▶ 12: {number: "678401363", title: "[95] 나이키 베이직 라운드 반팔티셔츠 네이비 (백포1.5)", price: "15,000원", content: "정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙..."}
▶ 13: {number: "678401606", title: "블루, 빈폴, 아디다스, 나이키, 라코스테, 헤지스 남성외투 100-110 팔아요", price: "10,000원", content: "정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙..."}
▶ 14: {number: "678401247", title: "경제사정 너무 어려워져 먼저 연락주시는 분과 거래합니다.. 사진 有", price: "8,000원", content: "정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙 / 정품 스루시 랑오버 아노락 블랙..."}
length: 15
__proto__: Array(0)
```

그림 15. 크롤링 결과

서버와의 통신 또한 \$.ajax()를 사용하며, 데이터를 성공적으로 받았을 때, 제목을 서버에서 전달받은 데이터로 변경하게 된다.

```
▼ (15) [{"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}, {"-"}]
▶ 0: {number: 678401561, newTitle: "상품명 : 토틀링 | 가격 : 140000 | 색상 : ? | 착용횟수 : 3회 | 사이즈 : 95"}
▶ 1: {number: 678401373, newTitle: "상품명 : 패딩 | 가격 : 80000 | 색상 : 옐로우 | 착용횟수 : 사용안함 | 사이즈 : 100"}
▶ 2: {number: 678401521, newTitle: "상품명 : 후 리스 | 가격 : 55000 | 색상 : ? | 착용횟수 : 사용안함 | 사이즈 : ?"}
▶ 3: {number: 678401637, newTitle: "상품명 : 패딩 | 가격 : 100000 | 색상 : 화이트 | 착용횟수 : 사용안함 | 사이즈 : 100"}
▶ 4: {number: 678401199, newTitle: "상품명 : 패딩 코트 | 가격 : 35000원 | 색상 : 코트색 | 착용횟수 : ? | 사이즈 : 105"}
▶ 5: {number: 678401498, newTitle: "상품명 : ? | 가격 : 60000 | 색상 : 블랙 | 착용횟수 : 한두번 | 사이즈 : ?"}
▶ 6: {number: 678401104, newTitle: "상품명 : 코트 | 가격 : 900000 | 색상 : ? | 착용횟수 : 2회 | 사이즈 : 46"}
▶ 7: {number: 678401493, newTitle: "상품명 : 다운 | 가격 : 265000 | 색상 : 사이즈 | 착용횟수 : 사용안함 | 사이즈 : 85"}
▶ 8: {number: 678401457, newTitle: "상품명 : 자수 패치 트레이닝복 | 가격 : 82000 | 색상 : 차콜 | 착용횟수 : 사용안함 | 사이즈 : ?"}
▶ 9: {number: 678401459, newTitle: "상품명 : 플 | 가격 : 200000 | 색상 : 브라운 | 착용횟수 : ? | 사이즈 : ?"}
▶ 10: {number: 678401108, newTitle: "상품명 : 아노락 | 가격 : 140000 | 색상 : 블랙 | 착용횟수 : ? | 사이즈 : L"}
▶ 11: {number: 678401657, newTitle: "상품명 : 점퍼 | 가격 : 20000 | 색상 : ? | 착용횟수 : ? | 사이즈 : 105"}
▶ 12: {number: 678401363, newTitle: "상품명 : 라운드 티셔츠 | 가격 : 15000 | 색상 : 네이비 | 착용횟수 : 6번 | 사이즈 : M"}
▶ 13: {number: 678401606, newTitle: "상품명 : 후 | 가격 : 10000 | 색상 : 베이지 | 착용횟수 : ? | 사이즈 : ?"}
▶ 14: {number: 678401247, newTitle: "상품명 : 트랜치코트 | 가격 : 8000 | 색상 : 루톤 | 착용횟수 : 사용안함 | 사이즈 : 105"}
length: 15
__proto__: Array(0)
```

그림 16 서버 통신 결과

	제목	작성자	작성일	조회
678401657	상품명 : 점퍼 가격 : 20000 색상 : ? 착용횟수 : ? 사이즈 : 105 판매  	asumir 	01:20	0
678401637	상품명 : 패딩 가격 : 100000 색상 : 화이트 착용횟수 : 사용안함 사이즈 : 100 판매  	gfsc10 	01:20	0
678401606	상품명 : 후 가격 : 10000 색상 : 베이지 착용횟수 : ? 사이즈 : ? 판매  	huh8080 	01:20	0
678401561	상품명 : 롱패딩 가격 : 140000 색상 : ? 착용횟수 : 3회 사이즈 : 95 판매  	생각없이쓰느네임 	01:20	5
678401521	상품명 : 후 리스 가격 : 55000 색상 : ? 착용횟수 : 사용안함 사이즈 : ? 판매  	후로로 	01:19	1
678401498	상품명 : ? 가격 : 60000 색상 : 블랙 착용횟수 : 한두번 사이즈 : ? 판매  	개미 	01:19	1
678401493	상품명 : 다운 가격 : 265000 색상 : 사이즈 착용횟수 : 사용안함 사이즈 : 85 판매  [1] 	y17 	01:19	2
678401459	상품명 : 롱 가격 : 200000 색상 : 브라운 착용횟수 : ? 사이즈 : ? 판매  	개미 	01:19	3
678401457	상품명 : 자수 패치 트레이닝복 가격 : 82000 색상 : 차콜 착용횟수 : 사용안함 사이즈 : ? 판매  	산업공구 	01:19	2
678401373	상품명 : 패딩 가격 : 80000 색상 : 옐로우 착용횟수 : 사용안함 사이즈 : 100 판매  	gfsc10 	01:18	2
678401363	상품명 : 라운드 티셔츠 가격 : 15000 색상 : 네이비 착용횟수 : 6번 사이즈 : M 판매  	안심상점 	01:18	0
678401247	상품명 : 트랜치코트 가격 : 8000 색상 : 투톤 착용횟수 : 사용안함 사이즈 : 105 판매  	ioytuytgrt 	01:18	5
678401199	상품명 : 패딩 코트 가격 : 3500원 색상 : 코트색 착용횟수 : ? 사이즈 : 105 판매  	솔타시질알띠 	01:17	3
678401108	상품명 : 아노락 가격 : 140000 색상 : 블랙 착용횟수 : ? 사이즈 : L 판매  	안심상점 	01:17	2
678401104	상품명 : 코트 가격 : 900000 색상 : ? 착용횟수 : 2회 사이즈 : 46 판매  	씩씩이3 	01:17	0

그림 17. 서버 통신 결과로 제목 변경

변경 완료된 제목에 커서를 위치하면 툴팁으로 변경 전 제목을 확인할 수 있다. 툴팁은 jquery ui 에서 지원하며 툴팁 생성을 원하는 태그에 'title' 속성을 추가하고 CSS 파일을 추가해 원하는 모양대로 툴팁이 보이게 할 수 있다.










	제목	작성자
678401657	상품명 : 점퍼 가격 : 20000 색상 : ? 착용횟수 : ? 사이즈 : 105 판매  	asumir 
678401637	상품명 : 패딩 가격 : 100000 색상 : 화이트 착용횟수 : 사용안함 사이즈 : 100 판매   [새상품] FILA 휠라/필라 테이퍼 패딩	gfsc10 
678401606	상품명 : 후 가격 : 10000 색상 : 베이지 착용횟수 : ? 사이즈 : ? 판매  	huh8080 

그림 18 변경 전 제목 확인

확장 앱과 콘텐츠 스크립트는 변경 버튼 클릭 시, 변경 전 버튼 클릭 시, 변경 완료 시 로딩을 끝내기 위해 서로 통신해야 한다. 확장 앱과 콘텐츠 스크립트는 다른 문맥에서 작동하므로 메시지를 이용해 통신하며 확장 앱 API 인 콘텐츠 스크립트는 whale.runtime.sendMessage, 확장 앱은 whale.tabs.sendMessage 를 이용해 메시지를 보내며 whale.runtime.onMessage.addListener 를 이용해 메시지를 받을 수 있다.

- 웹 스토리지 사용

제목이 변경된 후 새로 고침을 하거나 다른 페이지에 접속 후 되돌아왔을 때 변경 내용이 유지되게 하기 위해 웹 스토리지를 이용해 변경 내용을 저장하였다. 스토리지는 key 와 value 쌍으로 유지되며, 게시물 번호를 key 로, 변경 전 제목과 변경 후 제목을 value 로 저장하였다.

Key	Value
674924182_changed	상품명 : 아이폰 브랜드 : LG 가격 : 260000 상태 : 기존 용량 : 64GB
674699583_changed	상품명 : 가디건 상태 : 옷장 가격 : 20000 색상 : ? 착용횟수 : n한번 사이즈 : 호 브랜드 : ?
674846857_changed	상품명 : 상태 : 대장급 가격 : 200000 색상 : ? 착용횟수 : ? 사이즈 : ? 브랜드 : 노스페이스
674927678_changed	상품명 : 우븐 조거 팬츠 브랜드 : 대상트 가격 : 69000 착용횟수 : ? 사이즈 : 85
674931279_changed	용량 : ? 사양 : 사용기간 : ?
674829091_changed	상품명 : 후드 코트 브랜드 : 빈폴 가격 : 180000 사이즈 : ? 상태 : 있다
"674584373"_changed	"상품명" : "브랜드" : "가격" :
674829251_changed	상품명 : 바람막이 브랜드 : 나이키 가격 : 30000 사이즈 : L 상태 : 새상태
675041805_changed	상품명 : 최초 브랜드 : ? 가격 : 75000
675041777_changed	상품명 : 브랜드 : ? 가격 : 5000
674829120_changed	상품명 : 후드 패딩 브랜드 : 행방 가격 : 335000 사이즈 : ? 상태 : 고급
675041565_changed	상품명 : 패딩 베스트 브랜드 : JEEP 사이즈 : ?
674859097_changed	브랜드 : 라코스테 가격 : 25000 상품명 : 김영 색상 : ? 사이즈 : 105
674936867_changed	상품명 : 코트 외투 브랜드 : 닉스
678384473_changed	상품명 : 블라우스 가격 : 38,000원 색상 : 브라운색 착용횟수 : ? 사이즈 : ?
674847613_changed	상품명 : 슬리브 상태 : 새상품 가격 : 220000 색상 : ? 착용횟수 : 사용안함 사이즈 : XS 브랜드 : ?
674916202_changed	상품명 : 니트 브랜드 : 랑방 가격 : 220000
674916101_changed	착용횟수 : 3회 사이즈 : 프리 상품명 : 청바지 브랜드 : ? 색상 : ?

그림 19 스토리지에 변경 후 제목 저장

Key	Value
674860059	망고 MANGO 남자 가죽 자켓
674988653	소녀레시피 플라운스프리트 그레이
674563524	수퍼드라이(super dry) 자켓 M사이즈
674908727	◆소형 위치추적기◆ 즉시사용가능!! 무소음 전용배터리 사용시 최대 90일 맵시 최대 5년 / 전국 무료배송가능!!...
674939251	@@ 남색 패딩 조끼 운포 13000원 @@
674824111	나이키 윈드러너 바람막이 M 사이즈
675034138	햇한후드집업
675012139	새제품) 모노키니 레오파드 비키니 수영복 + 호피 여행가방 빅사이즈 비치백
678384646	헤지스글루프다운패딩90
674523035	리버풀레시 셔츠니트 95사이즈 4만
678300864	지오다노 니트일괄
674936817	미친가격ㅠㅠ 급해서요... 풀로 체크셔츠
674823996	클림모나코 롱코트
674539677	양가죽 패딩(구스다운), 라이더 가죽바지
674936877	** ◆◆ 미사 사슬리 제이크루 SJ50 트윗 디젤 Enc 아메리칸어퍼럴 ◆◆5
671572453	(무료배송)메종키즈네 예코백

그림 20 스토리지에 변경 전 제목 저장

5.3 자연어처리

- 전반적인 코드 진행 순서

웹 확장 프로그램에서 현재 카테고리 이름, 선택된 키워드 리스트, 게시물 제목, 가격, 본문으로 구성된 리스트를 받으면 서버에서 자연어 처리 후 추출한 키워드 리스트를 반환하는 코드를 작성하였다.

카테고리 이름과 선택된 키워드로 현재 카테고리를 인식하고 그에 해당하는 자연어 처리 모델 연결, 전역 변수 cate 에 현재 카테고리 정보를 저장한다. 이 전역변수는 후에 카테고리마다 이름을 추출하는 방법이 다른데 이름을 추출하는 함수 호출할 때 어느 함수를 호출할지 정하는 등에 사용된다. 그리고 카테고리에 해당하는 이름, 브랜드, 색상 사전을 읽어와 전역 변수 name, brand, color 에 리스트로 저장을 한다.

그 후 반복문을 사용하여 게시물 정보 리스트와, 선택된 키워드 리스트를 돌며 선택된 키워드에 해당하는 함수들을 호출하게 되고 그 함수의 반환 값을 리스트에

붙인다. 모든 리스트들을 다 돌면 각 게시글 별 키워드에 해당하는 값이 담긴 리스트가 생기는데 이를 최종 리스트에 순서대로 붙여주고 반환하며 코드가 종료된다.

이때 자유로운 길이의 리스트가 들어올 수 있도록 구현하여, 게시판 목록에 15 개의 게시글이 보이는 중고나라에만 한정하지 않고 확장성을 추구하였다.

- 게시글 정보 전처리, 함수 호출

게시글 정보를 전처리하고, 각 키워드에 해당하는 함수를 적합한 인자를 넘겨주며 호출한다.

전처리 과정은 다음과 같다.

KoNLPy API 중 가장 속도가 빠른 Okt 를 사용해 토큰나이징하였다. 특정 상황에서는 게시글 정보 전체를 토큰화하지 않는다. 게시글 본문 밑에 판매하는 상품 이외의 모든 상품명을 적어 놓는 등 게시글 정보가 터무니없이 긴 경우가 빈번했기 때문이다. 이는 총 자연어 처리 중 토큰화 과정에서 시간이 많이 소요된다는 것으로부터, 불필요한 문장은 토큰화하지 않아야 빠른 프로그램 실행으로 이어질 것이라고 생각해, 카테고리 별로 일정 길이 이상이 넘어가는 게시글 본문은 잘라서 사용했다. 정상적인 긴 글이라고 해도, 앞부분에 중요한 정보가 있을 것이라 생각한다.

가격 키워드 추출 시에는 미리 제공된 가격이 존재할 경우 따로 처리를 해주지 않고 그 값을 반환한다. 중고나라는 가격 기입이 필수이기 때문에 값이 항상 존재하였지만, 가격 기입이 필수가 아닌 카페에서 가격이 나와있지 않은 상태이거나, 가격이 5000 원 이상이 아니거나, 1000 원 단위가 아니라 의미 있는 가격이 아니라고 판단할 때 구현한 자연어 처리 가격 추출 함수를 호출한다.

인자로 넘겨주는 값은 다음과 같다.

- A. ~번, ~색, ~급 ~GB 등 특정 형태 있는 단어를 찾을 때 사용할 원본 스트링 (string_list)
- B. [], (), ‘, ’, ‘아’ 같이 키워드와는 관련 없는 의미 없는 값을 무시하고 토큰 간의 관계를 유용하게 보기 위해, 키워드가 될 수 있는 가능성이 있는 품사라고 판단한 명사, 형용사, 숫자, 알파벳, 동사 품사인 것으로만 이루어져 있는 토큰화된 리스트 (docs_list)
- C. 연달아 나오는 품사 종류에 따라 키워드로 인식되는 단어를 찾아내는 상황에서 사용하기 위해 문장 전체가 토큰화된 리스트 (all_docs_list)

- 이름 / 브랜드 키워드 추출 함수 설명

이름 키워드를 추출하는 함수는 카테고리를 저장한 전역 변수 값에 따라 다른 함수를 호출한다.

옷 카테고리 일 때는 docs_list 를 반복문으로 돌면서 이름으로 추정되는 단어가 있으면 반환 리스트에 담는 방법으로 구현하였다. 앞에서 설명한 사전에서 얻은 값이 담겨있는 name 리스트를 사용한다. docs_list 의 단어들을 앞에서부터 읽어가며 name 에 일치하는 단어가 등장하면 리스트에 담고 flag 를 1 로 변경한다. 그 후 단어 뒤에 단어들을 읽어가며 여전히 name 과 일치하는지 비교한다. 일치하면 리스트에

담고, 일치하지 않으면 반복문을 탈출한다. 반복문 탈출 후 리스트의 길이를 통해 이름이라고 예상되는 게 아무것도 없으면, 찾은 리스트의 길이가 0 이면 ‘?’ 를 1 이상이면 리스트를 join 해서 string 으로 반환한다.

flag 를 사용한 이유는 ‘기모가 두툽한 맨투맨’ 같은 예시 문장에서 볼 수 있다. 이 문장을 필요한 품사만 있게 토큰화할 경우 ‘기모/Noun’, ‘두툽하다/Adjective’, ‘맨투맨/Noun’ 같이 토큰화가 될 것이다. 이러한 상황이 빈번하여 flag 를 사용해 이름이 아닌 단어가 한 번 나와도 반복문을 탈출하지 않도록 구현하였다. 이름이 아닌 단어가 나오면 flag 가 1 일때 2 로, 2 일 때 3 으로 만들고, flag 가 3 일 때 반복문을 탈출하게 만들었다. 또한 같은 이름이 반복적으로 나오는 경우에 반복문을 탈출하지 않는 경우도 발생하여, 반환하는 리스트에 존재하는 단어인 경우에도 반복문을 탈출하도록 구현했다.

신발, IT 기기는 ‘구두’, ‘노트북’ 같은 특징 카테고리가 기입되어 있지 않고, ‘에어맥스’, ‘갤럭시’ 등 대표 명과, 영어, 숫자로 이뤄진 상품코드가 나오는 경우가 많았다. 영어, 숫자로 이뤄진 상품코드를 다 사전에 넣기에는 너무 많은 경우의 수가 있어서 실행 속도와 효율성에 문제가 있을 것이라고 판단하였고, 연속적으로 나오는 유효한 정보가 키워드일 가능성이 높지 않을까라는 생각을 하여, all_docs_list 를 사용했다.

신발 카테고리 일 때는 all_docs_list 를 돌면서 명사, 알파벳, 숫자로 이뤄진 연속되는 조합을 3 개 찾고, 가장 긴 것을 반환했다. 이때 다양한 예외 처리를 해주었다. 숫자 중엔 사이즈를 나타내는 경우가 많아 5 단위인 200 에서 300 사이인 숫자가 나오면 제외를 하고, ‘사이즈’를 뜻하는 단어가 앞뒤에 나오면 근처 숫자를 제외하는 등의 방법을 사용했다. 단순히 저 사이의 값을 배제해 버리는 것은 너무 큰 리스크가 아닌지 생각될 수 있지만, 실제 많은 게시글을 보았을 때 사이즈라는 표현을 하지 않고, 숫자만 기입하는 경우가 많았으며, 상품명들 중 저 범위의 5 단위인 경우는 매우 드물어서 이러한 방법을 사용하였다. 그리고 브랜드에 해당되는 정보는 브랜드 키워드에서 나타낼 것이기 때문에 중복을 피하기 위해 명사인 경우 브랜드인지 아닌지를 판단하여 중복된 값을 제외하고, 전역 변수 brand_name 에 저장해 브랜드를 검색하는 함수를 호출하였을 시 작업을 두 번 하지 않도록 하였다. 이 외에도 명사지만 ‘제품’, ‘정품’, ‘판매’ 등 다른 키워드를 찾을 시 사용하는 단어 즉 이름과는 관련 없는 단어가 나올 때에도 처리를 하도록 구현하였다.

IT 카테고리인 경우 신발 이름과 비슷한 방법을 사용하였지만, 다르게 처리한 부분이 있다. 첫째로 IT 는 신발보단 대표 명으로 자주 나오는 이름이 한정되어 있어, 사전으로 구축하였다. 이를 찾고 연속해서 나오는 뒤 단어를 보아 영어나, 숫자가 나오면 상품명으로 인식해 추출한다. 둘째로, 브랜드명을 이름으로 인식하지 않도록 처리하는 과정에 있어서, IT 기기의 경우 영어로 된 브랜드명이 많아 별도로 처리하였다. 영어는 토큰화 시 모두 Alpha 품사가 되는데 직접 구축한 사전은 명사로만 이루어져 있어 따로 영어 처리 코드를 추가해주었다. 이때, 대소문자 구별을 하지 않아 여러 경우 모두 인식해주도록 하였다. 마지막으로 -, + 와 같은 punctuation 일부를 허용하고, i 후에 나오는 숫자는 상품명이라고 인식, GB 등의 단어 인식 후 앞뒤 숫자 파악, 지나치게 앞쪽에 위치한 유효하지 않은 숫자 제외, 상품명이 아닌 제작 연도, 가격 등의 숫자 인식 등의 예외 처리를 해주었다.

브랜드 키워드를 추출하는 함수는 직접 구축한 브랜드 사전에 일치하는 단어만 반환한다. ‘바나나리퍼블릭’ 같은 브랜드는 토큰화 시 ‘바나나’ ‘리퍼블릭’ 으로 되는 등 하나의 브랜드 이름이 여러 개로 쪼개져서 들어가 있는 경우가 많았다. 이를 위해 연속된 브랜드 이름은 하나의 브랜드로 처리를 해주었다.

- 이름 / 브랜드 이외의 키워드 추출 함수 설명

가격, 상태, 색상, 착용횟수, 사용기간, 상세 스펙, 용량 키워드를 호출하는 함수는 ~원, ~급, ~색, ~번 등의 특징이 있는 단어를 검색하고, 취급, 발급, 지급, 배색, 변색 등 원치 않는 단어를 후처리하고, 검색된 개수에 따라 다른 활동을 한다.

가격 함수에서 보면 ~원으로 추출된 유효한 값이 여러 개인 경우 ‘정가’, ‘구입’이라는 단어로부터 먼 단어를 반환하고 하나도 없는 경우 ‘가격’, ‘판매가’ 등의 단어 근처에서 숫자인 단어를 찾는다. 예를 들어 ‘15000 원’ 같이 숫자와 단위가 합쳐진 경우도 숫자 품사로 분리가 된다. 찾아진 단어들의 개수에 따라 또 같은 알고리즘을 반복한다.

상태, 착용 횟수, 사용기간 함수도 마찬가지로 ‘상태’, ‘사용감’ 등과 같은 단어 근처의 명사와 형용사를 추출하고, ‘미착용’, ‘미개봉’ 같이 한 번도 사용되지 않은 경우도 인식한다.

색상 함수도 ‘색상’, ‘컬러’ 등의 단어 근처에서 단어를 추출하고, 모델을 사용하여 예시 토큰 ‘베이지’와 비슷한 정도가 일정 값 이상이고, 가장 비슷한 단어를 추출한다. 딱히 적합한 단어가 없다고 판단되면 docs_list 를 탐색하면서 color 사전과 일치하는 값을 반환한다. color 사전은 각 카테고리별로 색상을 지칭하는 법이 다르다고 생각해 카테고리에 옷, 신발, IT 로 각각 구축했다. 옷은 벽돌, 머스타드, 소라 등 다양한 표기법이 있었고, 신발은 흰과, 검노 등 여러 섞인 색을 줄여 말하는 경우가 많았고, IT 는 스그, 로콜 같이 상세한 색상 단어를 줄여 말하는 경우가 많았기 때문이다.

거래 방식은 직접 설정한 단어와 일치하는 단어의 유무로 정리를 하여 출력하였다.

사이즈 함수는 ‘사이즈’, ‘size’ 등의 단어 근처의 토큰을 가져와 Free 같은 단어인지, 알파벳인지, 숫자인지 확인하고, 전부 숫자라면 가장 작은 숫자를 출력한다. 만약 아무것도 추출하지 못했다면 ~인치, ~mm, ~호 등의 특징이 있는 단어를 추출한다.

용량 함수의 경우 IT 기기는 용량을 뜻하는 단어가 아주 많아 여러 경우에서 추출할 수 있는 값에 따라 다르게 출력을 한다. 우선 SSD, HDD 같은 단어의 존재 유무를 판단하고, 추출한 유효한 크기를 맞춰서 반환을 한다.

상세 스펙 함수는 ~인치, ~kg, i 뒤에 나오는 숫자 영어 등으로 추출한다.

위와 같은 처리를 해주기 위해, 유사도 비교, 근처 토큰 추출, 특정 토큰 간 거리 획득 등의 함수를 구현하였다.

- NER 사전 구축

```
≡ ccolor.txt
≡ fbrand.txt
≡ fname.txt
≡ ibrand.txt
≡ icolor.txt
≡ iname.txt
≡ mbrand.txt
≡ mname.txt
≡ sbrand.txt
≡ scolor.txt
```

그림 21 구축한 사전들

[그림 21]에서 구축한 사전들을 나타낸다. m 은 남자, f 는 여자, s 는 신발, i 는 IT 기기를 뜻한다. 각 카테고리별로 따로 사전을 구축하였고, NER 을 사용해서 name, brand, color 를 구분해 추출하였다. Named Entity Recognition (NER) 은 문장에서 특정한 종류의 단어를 찾아내는 information extraction 문제 중 하나로, 우리의 키워드 추출 프로젝트에 꼭 필요한 기술이다.

wordnet 을 사용해서 해당하는 단어를 찾고, 위로 올라가며 어느 분야의 단어인지 파악하거나, 부모가 같은 것으로 묶는 방법이 있지만, 다양한 문제가 있다. 첫째로 한국어가 wordnet 은 잘 제공되지 않으며, 둘째로 brand 와 같은 도메인은 매우 넓고, 지속적인 업데이트가 필수인데 wordnet 은 여러 사람이 직접 구축하는 것으로 이를 감당하기에 어려움이 있을 것이라고 판단했다. 마지막으로 원치 않는 부분도 포함하고 있어 너무 많은 도메인이라 검색에 오랜 시간이 걸릴 것으로 예상되어 빠른 시간에 돌아가야 하는 우리의 프로그램에는 적합하지 않다고 생각하였다.

또한 nltk 에 이미 구현된 NER 을 사용하는 등의 방법도 있긴 하지만, 나라 이름인지 사람 이름인지 등 우리가 원치 않는 도메인을 판단하는 NER 이라 직접 데이터를 모으고 NER 후 사전을 구축하였다.

NER 을 구현하기 위해 text CNN 같이 window 를 이동해가며 각 단어의 위치적 특성으로 도메인을 구분해내는 방법을 사용했다. 이때, word2vec 모델로 seed_word 를 설정해 추출할 도메인을 지정하고, Logistic Regression 을 사용해 seed_word 이외에도 seed_word 와 비슷한 성향을 보이는 토큰들을 추출했다.

NER 은 같은 단어이지만 품사가 다른 경우를 다 다르게 봐주고 위치적 특성을 뽑아낸다. 이렇듯 품사 형식에 큰 영향을 받아 토큰화된 문장에서 진행을 하였다. 또한, 마치 한 문장씩 가로로 윈도우가 지나가듯이 1*N 크기의 윈도우로 지나가는 것이다 보니, 문장 구성의 영향을 많이 받는다. 판매자가 문장 분리를 잘 하지 않고 게시글을 올려도 NER 의 성능을 올리기 위해 sentence slice 의 필요하다고 생각했고, kss 라이브러리를 사용하여 문장 분리를 하여 진행하였다.

각 사전 구현에서 다양한 하이퍼 파라미터들이 있어서 추출하고자 하는 카테고리, 도메인에 따라 다 다르게 설정해주는 작업에 오랜 시간이 걸렸다. 처리해야 하는 범위, 포함해야 하는 단어 수에 따라 min_count 를 조정해보고, 구문적 의미적 추출에 의해 window_size 를 조정하는 작업을 했다. 또한 참고 자료에서는 seed_word 선택 과정만 나와있었지만, 생각보다 값이 잘 나오지 않아 seed_word 에서 제외할 단어들을 설정해주는 부분과 맨 마지막 추출 값 후처리해주는 부분을 추가 구현하였다.

6. 프로젝트 제약사항

- 한 게시글 내에서 두 개이상의 물건 판매 상황 제외

위의 상황을 처리하기 위해선 여러 물건을 판매하는 상황을 인식하는 모델이 필요하고, 게시글 내에 기입되어있는 각 정보들이 어떤 물건에 대한 정보인지 매칭해주어야 한다. 하지만 각 물건에 해당하는 정보의 양이 다른 경우가 많아 위의 기능까지 처리하기에는 시간이 부족할 것이라 판단하였다. 이름을 제외한 키워드들은 가장 가능성이 높은 정보 한 개만 띄워주고 이름은 연속적으로 나오는 경우에만 띄워주는 것으로 진행하였다.

- 여러 사이트 상에서의 크롤링 및 제목 변경 어려움

크롤링과 제목 변경 작업을 할 때 해당 사이트의 html 을 가져와 사용한다. 이때 원하는 내용을 html 태그를 사용해 가져오게 되는데, 사이트마다 태그의 종류와 태그를 선택할 때 사용하는 Class 와 Id 가 다르다는 문제가 있다. 따라서 우선 대상으로 '중고나라' 사이트를 선택해 진행하였으며 중고나라와 같은 도메인인 네이버 카페는 조금의 수정을 거쳐 사용할 수 있을 것이라 생각한다. 같은 네이버 카페이더라도 게시글 목록에서 사진이 보이는 경우, 본문에 가격 부분이 따로 없는 경우 다르게 처리를 해주어야 할 것으로 보인다. 하지만, 다음 카페와 같은 경우 html 이 완전히 다르게 구성되어 있어 현재 구현한 것과 별도로 구현해야 하며 접속한 도메인이 무엇인지 판단하는 작업도 해줘야 한다는 어려움이 있다.

7. 발전 가능성

- 접목 가능한 추가 기능

게시글 내에 많은 사람들의 유입을 위한 다양한 물건 이름이 태그처럼 기입되어 있거나, 너무 많은 특수문자, 사족 그리고 제목과 관련 없는 게시물 내용 등의 상황을 인지하는 모델을 만든다면 광고인지 아닌지 판단 후 키워드로 제공이 가능할 것이다.

현재는 텍스트 형태의 데이터만 처리하지만 이미지나 동영상 형태의 정보도 처리를 해준다면 사용자에게 더 다양한 정보들을 제공할 수 있을 것이다. 게시글 내에 색상이 기입되어 있지 않아도 사진을 인식해 자동적으로 색상 정보를 추출하는 방법 등으로 현재 제공하는 키워드들의 성능을 높일 수 있다. 추가적으로는 사진을 인식해 단순히 판매 홈페이지에 나와있는 사진인지, 실제 판매자가 찍은 사진인지

구별을 하거나, 이미지 내 글자 인식을 통해 판매자 정보가 있는지 없는지 판단해 인증이 되어있는 사진인지 등의 키워드도 제공할 수 있을 것이다.

- 넓은 도메인 활용 가능성

현재는 네이버 카페 중고나라의 일부 카테고리에서만 기능이 제공되지만, 더 넓은 도메인에서도 활용이 가능할 것이라고 생각한다. 거래 도메인이 아닌 다른 분야에서 키워드를 직접 추출해보았을 때 [그림 22]와 [그림 23]과 같은 결과를 예상해 볼 수 있다.

상항으로 합격한 늦은 19학번 사회복지학과 6종합 후기 [39]	서울시립대 사회복지학과 종합 2.3 상항
가톨릭대 프랑스어문학과 내신 4.2 합격 [12]	가톨릭대 프랑스어문학과 교과 4.2 극상항
<<<<< 천안권 간호학과+응급구조학과 >>>>> [2]	천안대 간호학과 학종 3.8 안정

그림 22 네이버 카페 수만휘 변경 예시

500/45 흑석동 풀오피스 원룸 입주가능 보증금조정가능 흑석역 중앙대 동작역 [12]	500/45 원룸 3층 <u>흑석역</u> 6분
서울대입구역 7분, 낙성대역 8분 신축원룸 (500/50), 복비지원 [12]	500/50 <u>신축원룸</u> 4층 서울대입구역 7분
9호선 흑석역 1분거리 풀오피스 옥탑방 월세임대 500/55 [12]	500/55 옥탑 6층 <u>흑석역</u> 1분

그림 23 네이버 카페 피터팬 집 구하기 변경 예시

또한 더 나아가면 [그림 24]와 같이 카페 게시글이 아닌 일반적인 검색창에서도 활용이 가능할 것이라고 생각한다. 화장품, 맛집 등 많은 사용자의 관심 분야에 따라 손쉽게 검색하고 정보를 얻는 데 도움이 클 것이라 전망한다.



그림 24 네이버 검색창 활용 예시

현재는 옷, IT 에 대한 모델을 따로 만들어 주었지만, 위와 같이 다양한 도메인에서도 적용되기 위해선 직접 하나하나 정의해주고 분류하는 것은 무리가 있을 것이라 판단했다. 추후에는 모든 게시글을 대상으로 학습을 시키고, 도메인 구별, 키워드 추출 등을 인공지능을 통해 자동적으로 판단, 추출해내는 형태로 진행을 하여야 할 것이다.

8. 업무 분담과 프로젝트 스케줄

8.1 업무 분담

- 박시현

데이터 크롤링 및 데이터셋 구축

서버 구축

웹서버 구축 및 통신 구현

- **박진영**

프론트엔드 화면 및 동작 구현

텍스트 대치 구현

웹페이지 쿠키 구현

- **이민희**

데이터 전처리

자연어 처리 알고리즘 구현

키워드 사전 및 모델 구축

8.2 개발 일정

(공통: 하늘색 / 박시현: 분홍색 / 박진영: 연두색/ 이민희: 보라색)

		9 월				10 월				11 월				12 월			
	주차	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	주제 확정 및 제안서 작성																
	구현 방법 학습 및 자료조사																
박 시 현	웹 데이터 크롤링																
	데이터셋 구축																
	서버 구축																
	서버 단 통신 모듈 작업																
	웹서버 구축 및 서버 통신 구현																
	중간 데모 준비																
이 민 희	라이브러리 사용 전처리																
	알고리즘 구체화																
	문장분리																
	주변 토큰 처리																
	키워드 사전 변경																
	모델 구축 및 파라미터 변경																
	성능 및 정확도 향상																
박 진 영	프론트엔드 화면 구현																
	프론트엔드 동작 구현																
	텍스트 대치 구현																
	확장 앱 통신 구현																

최
종
리
포
트
제
출
기
말
고
사

웹페이지 쿠키 구현															
데모 준비 및 서비스 테스트															
최종 보고서 준비															

9. GitHub 주소

- https://github.com/celi1004/Team4_Capstone/settings