[양식 3] 결과보고서

# [캡스톤디자인 결과보고서]

## ■ 연구과제

| 과 제 명 (작품명) | 커뮤니티 핫딜 게시판 기반 상품 모음 웹(Hotmoa) | 참여학기 | 2021 년 2 학기 |
|---|---|---|---|

## ■ 강좌정보

| 과 목 명 | 데이터분석캡스톤디자인 | 학수번호 | SWCON32101 |
|---|---|---|---|
| 과제기간 | 2021년 9월 1일 ~ 2021년 12월 26일 | 학 점 | 3 |

## ■ 팀구성

| 팀 명 | 핫모아 | | 팀구성 총인원 | 2 명 |
|---|---|---|---|---|
| 구 분 | 성명 | 학번 | 소속학과 | 학년 |
| 대표학생 | 옥현종 | 2019102109 | 소프트웨어융합학과 | 3 |
| 참여학생 | 유종석 | 2019102111 | 소프트웨어융합학과 | 3 |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

## ■ 지도교수 확인

| 지도교수 | 성 명 | 이대호 | 직 급 | 전임교수 |
|---|---|---|---|---|
| | 소속학과 | 소프트웨어융합대학 | 지도교수 확인 | 성명 : (인) |

## ■ 붙임

[첨부1] 과제 요약보고서

[결과물] 최종결과물 (최종작품 사진/도면/발표자료 등)

**본 팀은 과제를 성실히 수행하고 제반 의무를 이해하여 이에 따른 결과보고서를 제출합니다.**

**일자 : 2021 년 12 월 19 일**

**신청자(또는 팀 대표)** 옥현종 (인)

**[양식 4] 과제 요약보고서**

# [캡스톤디자인 과제 요약보고서]

| 과 제 명 | 커뮤니티 핫딜 게시판 기반 상품 모음 웹(Hotmoa) |
|---|---|

1. 과제 개요

가. 과제 선정 배경 및 필요성

최근 물가 상승으로 인해, 여러 게시판에 사람들끼리 "핫딜"이라는 것 공유하기 시작하였음. 그러나 사람들의 관심사가 제각각이기에, 다양한 종류의 제품들이 게시글로 올라오기 시작하였고, 개개인이 원하는 카테고리의 물품을 찾기가 어려워짐. 또한, 플랫폼 간 핫딜 게시판에 중복된 아이템들이 올라와 검색하는 입장에서 혼동이 있을 수 있음.

나. 과제 주요내용

1, 제품 카테고리 세부분류 -> 커뮤니티에서 제공하는 카테고리는 분류가 크게 되어있어 세부적으로 더 분류하여 유저가 원하는 결과를 세부적으로 볼 수 있게 해준다. 크롤링한 데이터를 분석하여 세부분류 기준을 어떻게 정하면 좋을지 생각하고 그 분류기준을 바탕으로 라벨링을 진행한 후 딥러닝 모델을 이용하여 성능을 낸다.

2. 뽐뿌, 에펨코리아 등 다수의 플랫폼에 올라오는 같은 내용의 게시글 분류

| index | fmkorea | fmkorea_category | ppomppu | ppomppu_category |
|---|---|---|---|---|
| 0 | 비비고 왕교자 385g 10개 | 먹거리 | CJ 비비고 왕교자 385g*10개 | 식품/건강 |
| 1 | 펩시 제로 355ml 48개, 210ml 60개(현대/씨티/롯데/우리) | 먹거리 | 펩시 제로 355ml x 48개 | 식품/건강 |
| 2 | 비비고 사골곰탕 500g 18봉 (국민X스마일페이) | 먹거리 | 비비고 사골곰탕 500G X 18개 | 식품/건강 |
| 3 | 아이폰 SE2 레드 256GB | 가전제품 | 아이폰SE2 256GB 자급제 레드 | 디지털 |
| 4 | 로지텍 G304 LightSpeed 화이트/블랙 2년보증 | PC제품 | 로지텍 G304 LIGHTSPEED WIRELESS 무선 게이밍 마우스 | 디지털 |
| 5 | 진라면(매운맛or순한맛) 1박스(20봉) KT사용자 역대가 | 먹거리 | 진라면20개 1박스,적립금 5000증정 | 식품/건강 |
| 6 | 투썸 스트로베리 초콜릿 케이크 (아메리카노 증정) | 먹거리 | 투썸플레이스 X mas 케이크 15%할인 + 아메리카노 | 기타 |

위는 동일 제품에 대한 서로 다른 커뮤니티에 유저가 올린 글의 제목이다. 거의 제목이 똑같은 글도 있지만 유사한 글도 있다. 이를 올린 시간 범위, 쇼핑몰, 가격, 제목 등 다양한 요소를 이용하여 동일제품에 대한 글인지 유사도를 검사한다. (완전 정해진 것은 아니지만) 예로 쇼핑몰, 가격 등은 완전 일치하는지 비교를 하고 제목은 벡터화하여 유사도를 높은 threshold를 적용하여 100퍼센트 정확한 동일 제품 다른 글 데이터를 일정량 수집한 후 임의로 몇 개의 데이터는 글을 스위칭하여 동일 제품일 경우 1, 동일 제품이 아닐 경우 0으로 라벨링 한 후 test set, train set, validation set을 잘 나누어 f1-score을 기반으로 여러 가지 모델을 실험하여 최적의 모델을 탐색한다. 생각중인 모델 중 하나는 Bert모델에 제목을 입력하고 나머지 데이터를 정규화 한 후 DNN Classfication을 해주는 모델을 Baseline으로 생각중이다.

3. 지속적으로 크롤링 할 수 있게 클라우드 서비스를 이용하여 크롤링을 한 후 데이터베이스에 저장해준다.
4. 어플리케이션 혹은 웹으로 개발한다.
5. 크롤링을 한 후 과거 데이터를 EDA하여 다른 유용한 데이터 혹은 방법이 있는지 탐색하고 추가적으로 진행한다. (예로 카테고리 세부분류를 현재 생각 중인데 이것이 유의미한 데이터인지 잘 따져보고 진행한다. 만약 유의미하다면 세부분류를 라벨링하여 test set, train set, validation set을 잘 나누어 f1-score을 기반으로 여러 가지 모델을 실험하여 최적의 모델을 탐색한다.

다. 최종결과물의 목표

1. 한 눈에 볼 수 있는 UI/UX와 데이터 시각화를 구현한다.
2. 모델API 와 여러 API를 잘 만들어 안정화된 서버를 만든다.

2. 과제 수행방법

가. 과제수행을 위한 도구적 방법 (활용 장비, 조사 방법론 등)

Google Colab을 이용하여 최적 모델 구현, AWS, GCP을 이용하여 클라우드 서버 구현.

뽐뿌, 에펨코리아의 핫딜 게시판 게시글 크롤링

나. 과제수행 계획

1. 커뮤니티 핫딜 게시판의 게시글 크롤링 서버 구현
2. 커뮤니티 간 동일 내용 중복성 검사 모델 구현
3. 웹 혹은 앱 서비스 구현
4. 커뮤니티 카테고리 세부 분류 모델 구현

3. 수행결과

가. 과제수행 결과

1. 모델 구현
2. 백엔드 구현
3. 프론트엔드 구현

나. 최종결과물 주요특징 및 설명

1. 모델 구현

라벨링 작업에 상당한 시간이 소요됨.
라벨링 작업에 한계로 부족한 데이터를 BackTranslation, Pseudo Labeling을 통해 해결함.
위에 2가지 기법을 포함하여 그 밖에 여러가지 실험을 통해 첫 Fasttext Baseline 0.804 -> Final Kor-Bert-Base 0.934 까지 성능 향상.
다소 큰 모델 사이즈 (450MB)를 Teacher - Student(DistilBert) 기법을 이용해 250MB로 감소시킴.
유사도 모델을 구현하였으나 검증과 실제 서비스에 적용은 시간관계상 하지 못함. ( 일부 Test 샘플로 Test 해본결과 동일 글 잘 감지 )

2. 백엔드 구현

GoogleCloudFunction을 통해 게시글을 스크래핑하고 모델에 적용하여 SQL로 보내는 전체 과정을 구현
SQL을 프론트엔드에게 URL로 전달하기 위해 Flask를 활용한 API Server 구현하고 EC2에 Docker와 Nginx로 배포

3. 프론트엔드 구현

Firebase를 이용한 로그인 기능 및 유저 관심 카테고리 데이터베이스 구축
React를 이용해 Backend 데이터 시각화 및 주요 기능 구현

4. 기대효과 및 활용방안

가. 기대효과

뽐뿌, 에펨코리아의 게시글들을 세부적으로 더욱 자세하게 카테고리를 분류하였기 때문에, 이제 유저들은 이 애플리케이션 사용을 통해 자신이 원하는 것들을 더욱더 쉽고 빠르게 접근할 수 있게 됨.
예를 들어 디지털 카테고리에서 핸드폰 관련 게시글들을 찾고 싶은 경우 기존에는 유저가 직접 검색하고 자신이 원하는 핸드폰들을 일일이 찾아봐야 하는 고충이 있었는데, 이제 게시글이 올라오면 자동적으로 세부 카테고리로 분류하여 정리해놓기 때문에 유저는 자신이 원하는 핸드폰 카테고리만 눌러 검색하면 핸드폰 관련 게시글만 볼 수 있음.

나. 활용방안

본 프로젝트는 에펨코리아, 뽐뿌 핫딜 게시글들을 세부적으로 나누는 것을 구현하였음. 제목을 기반으로 카테고리를 분류한 것이기 때문에, 핫딜 게시글들에 대해 한정적으로 적용되는 것이 아닌, 타 게시판 글들에 대해서도 정리할 수 있고, 이로 인해 더욱 다양한 주제로 확장하여 세부 분류를 할 수 있을 것으로 보임.

5. 결론 및 제언

이번 커뮤니티 핫딜 게시판 기반 상품 모음 웹 구현은 기존에 있던 핫딜 게시글 크롤링과는 차별점이 있음. 게시글들을 제목 기반 세부 분류할 수 있도록 모델링한 이후, 이를 세부 카테고리별 출력하여 기존 핫딜 게시글 검색 시스템에서 어려움을 겪던 유저들에게 유용한 편의를 제공할 수 있고, 나아가 다른 게시판 게시글 세부 분류에 대한 확장 가능성도 포함하고 있음.

본 프로젝트를 진행하는데 있어 아쉬웠던 점은 시간임. 세부 카테고리에 대한 각 게시글 라벨링에 시간이 많이 소요된 탓에 다른 구현하는데 있어 시간이 촉박한 점이 있었음.

프로젝트에 참여하면서 어려웠던 점도 많았지만, 이번 프로젝트를 통해 데이터 가공을 위한 라벨링, 모델링, 백엔드 및 프론트엔드 구현 등 정말 많은 것을 새로 접하고 또 활용해보는 기회가 되었음.

※ 본 양식은 요약보고서이며, 최종결과물을 필히 추가 제출하여야 함.

팀 학생대표 성명 : 옥현종 (인)