[양식 3-0] 중간보고서

[캡스톤디자인 중간보고서]

■ 연구과제

과 제 명 (작품명)	커뮤니티 기반 게시판 기반 상품 추천 시스템	참여학기	2021 년 2학기

■ 강좌정보

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

■ 팀구성

팀 명	핫모아		팀구성 총인원	2명
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 중간보고서

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

일자 : 2021 년 11 월 4 일

신청자(또는 팀 대표) 옥현종 (인)

[양식 4-0] 과제 중간보고서

[캡스톤디자인 과제 중간보고서]

과 제 명	커뮤니티 핫딜 게시판 기반 상품 추천 시스템

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. 핫딜 제품 가격 분석 및 추천도 제공

-> 회원의 초기 입력 데이터 (현재는 보유 카드 이용사, 음료와 같은 회원의 관심 물품 분류)를 통해 회원별 추천도 가중치를 주어 반영한다. 예로 물품이 신한카드에서 추가 할인이 있으면 이를 보유한 자의 추천도를 높여준다

-> 현재 여러 쇼핑몰에 동일 제품의 가격을 실시간 크롤링하여 현재 가격이 쇼핑몰에 올라온 가격보다 얼마나 더 싼지 보여준다

3. 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.

4. 관심 있는 제품의 시세 분석 및 비교한 내용 시각화

-> 유저가 편하게 볼 수 있게 1번에 내용인 여러 쇼핑몰 가격 비교 표, 추천도, 과거 가격 데이터 추이 그래프를 시각화 해준다.

5. 지속적으로 크롤링 할 수 있게 클라우드 서비스를 이용하여 크롤링을 한 후 데이터베이스에 저장해준다.

6. 어플리케이션 혹은 웹으로 개발한다.

7. 크롤링을 한 후 과거 데이터를 EDA하여 다른 유용한 데이터 혹은 방법이 있는지 탐색하고 추가적으로 진행한다. (예로 카테고리 세부분류를 현재 생각중인데 이것이 유의미한 데이터인지 잘 따져보고 진행한다. 만약 유의미하다면 세부분류를 라벨링하여 test set, train set, validation set을 잘 나누어 f1-score을 기반으

로 여러 가지 모델을 실험하여 최적의 모델을 탐색한다.

다. 최종결과물의 목표

1. 동일 제품 중복 게시글의 분류는 중복이라고 인지를 잘못한 경우가 더 치명적이기 때문에 f1-score기반으로 실제로 상용가능한 점수를 목표로 한다. 현재는 0.9 이상을 목표로 하고 있다.
2. 한 눈에 볼 수 있는 UI/UX와 데이터 시각화를 구현한다.
3. 모델API 와 여러 API를 잘 만들어 안정화된 서버를 만든다.
4. 어플리케이션이 완성되면 최대한 많은 홍보를 하여 추천도에 대한 피드백을 받을 수 있게 하고 그 유저들로 인해 발생한 데이터를 분석하여 의미 있게 추천시스템을 강화한다.

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. 추천도 로직 만들기 + EDA를 통한 추가적인 기능 혹은 모델 구현 (추가 기능)
4. 웹 혹은 앱 서비스 구현 및 시각화

3. 진행내용

가. 과제진행 내용

1. 뽐뿌, 에펨코리아에서 자동으로 게시글을 크롤링하는 시스템 구현
2. 게시글 별 제품 세부분류 기준 결정
3. 제품 세부분류를 위한 라벨링
4. 제품 세부분류 모델 공부 및 구현

나. 진행내용의 주요특징 및 설명

1. 구글 클라우드 플랫폼의 cloud function을 이용하여 크롤링하고 big query에 저장하는 시스템 구현.
2. 게시글 제목을 기반으로, 각 카테고리 별 최빈단어를 10~15개 정도 추출.
3. 추출한 최빈단어 중 의미 없는 것들은 제거. 모델의 성능을 높이기 위해, 하나의 카테고리에 라벨링 작업을 집중하기로 결정. 이에 '디지털' 카테고리의 약 2200개 가량의 데이터에 라벨링 작업 완료.
4. 간단한 전처리를 적용하고 fasttext, gpt2, albert, bert, funnel 등 몇 개의 모델에 대하여 조사 및 구현하였으며, 각 모델에 대해 성능을 비교.

4. 향후계획

1. 마저 못한 각 카테고리 별 데이터 라벨링.
2. 뽐뿌, 에펨코리아 핫딜 게시판 간에 겹치는 게시글이 있을 것이므로, 이를 선별해내기 위한 유사도 검사 시스템 모델링.
3. 웹앱 개발을 위한 풀스택 공부 및 구현.

팀 학생대표 성명 : 옥현종 (인)