「제5회 대구 빅데이터 분석 경진대회」
분석 결과 보고서

	※ 작성하지 않음

성명(팀명)	CAUsumer
분석과제명	소비자 맞춤형 뉴스 추천서비스

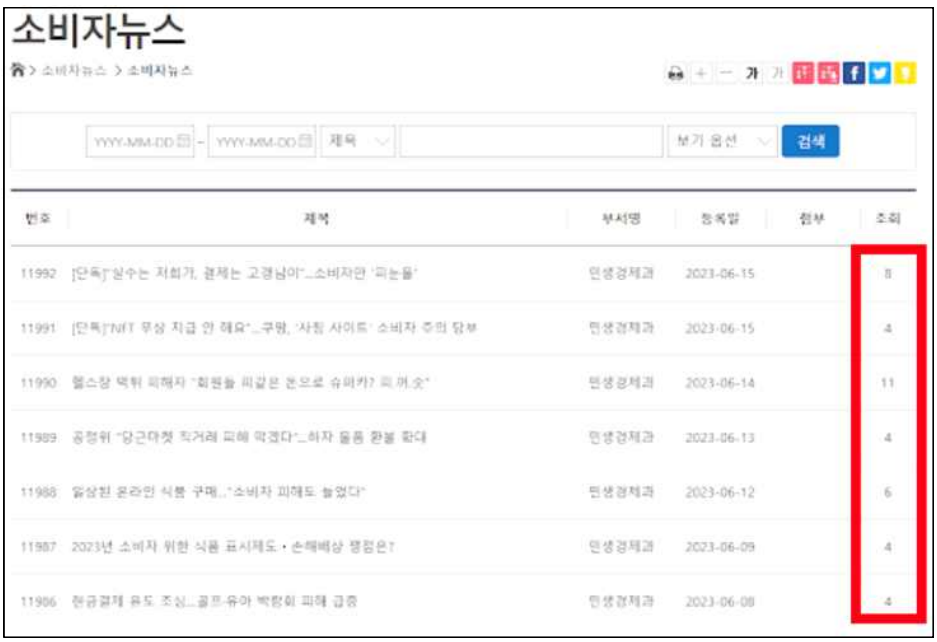
Ⅰ. 분석개요

▢ 배경 및 필요성

○ 1. 뉴스 접근성의 중요성

< 소비자뉴스에 대한 낮은 관심도와 접근성 >

대구광역시 소비생활 센터에서는 소비자들에게 필요한 뉴스를 선별하여 제공하고 있다. 대구통계(stat.daegu.go.kr)에 따르면 2022년 12월 기준 대구광역시의 총인구는 2,363,691명, 그중 활발한 소비를 하는 경제 활동 인구는 약 1,263,000명이다. 하지만 최근 6개월을 보았을 때, 선별하여 제공하는 뉴스의 조회수는 100회 이상을 기록하지 못하였다. <그림-1>



<그림-1> 소비자뉴스 조회수. 출처 : 대구광역시 소비생활센터

또한 '유사 투자자문'이 2021년 10월부터 2022년 6월까지 소비자 월별 상담다발품목 1위를 유지하고 있다. 유사투자자문서비스(주식리딩방)에 대한 경각심을 일깨워 주는 소비자 뉴스가 많았음에도 장기간 1위를 유지하는 것을 보았을 때 <그림-2>, 양질의 뉴스라 하더라도 접근성이 좋지 않고 쉽게 읽을 수 없다면 효과가 없음을 알 수 있다. 대구광역시 소비생활 센터는 양질 정보를 제공하는 데 반하여 접근성이 떨어지고 사람들의 관심을 끌기엔 부족했다. 이에 따라 접근성 높은 전달 방법의 필요성을 느꼈다.

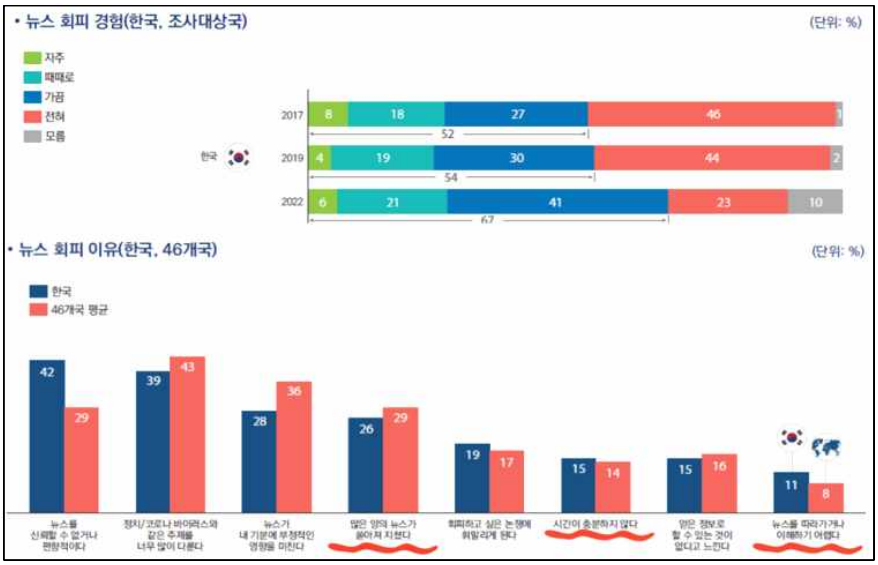
순위	2021년 10월		2021년 11월		2021년 12월	
	소분류 품목	건수	소분류 품목	건수	소분류 품목	건수
1위	유사투자자문	176	유사투자자문	235	유사투자자문	163
2위	이동전화서비스	68	헬스장	90	헬스장	74
3위	기타의류·섬유	58	이동전화서비스	72	이동전화서비스	68
순위	2022년 1월		2022년 2월		2022년 3월	
	소분류 품목	건수	소분류 품목	건수	소분류 품목	건수
1위	유사투자자문	119	유사투자자문	138	유사투자자문	114
2위	이동전화서비스	69	이동전화서비스	59	이동전화서비스	61
3위	기타의류·섬유	57	헬스장	39	헬스장	53
순위	2022년 4월		2022년 5월		2022년 6월	
	소분류 품목	건수	소분류 품목	건수	소분류 품목	건수
1위	유사투자자문	97	유사투자자문	77	유사투자자문	76
2위	이동전화서비스	49	이동전화서비스	64	이동전화서비스	61
3위	헬스장	44	헬스장	51	기타의류·섬유	44

<그림-2> 소비자 월별 상담다발품목 (2021.10~2022.06) 출처 : 대구 통계

○ 2. 맞춤형 뉴스 요약의 필요성

< 소비자들의 뉴스 회피 경험 >

뉴스를 회피하는 국민의 비율은 연을 거듭할수록 증가하고 있고 2022년 통계자료를 볼 때, 한국인의 67%. 즉, 한국인의 3명 중 2명은 뉴스를 회피한 경험이 있다고 응답했다. 뉴스 회피에는 다양한 이유가 있지만 뉴스의 내용이 어렵고 양이 많아 읽을 시간이 충분하지 않다는 것이 주를 이루었다. 미디어의 발달로 어렵고 긴 글을 읽고 싶어 하지 않으며 관심 없는 뉴스를 보지 않는 소비자들에게 관심을 가질만한 뉴스의 핵심만을 파악하여 짧고 간결한 뉴스를 제공할 필요가 있다.



<그림-3> 소비자 뉴스 회피 경험률, 뉴스 회피 이유. 출처 : 한국언론진흥재단

▢ 분석목적

배경 및 필요성에 본 것처럼 소비자들의 소비생활에 도움 되는 뉴스 정보를 제공할 때, 3가지가 중요하다고 생각한다. 첫 번째, **접근성**이 좋아야 한다. 넷플릭스, 유튜브와 같은 미디어 플랫폼이 상용화된 오늘날 접근성이 좋지 않은 뉴스를 직접 찾아보는 사람은 드물다. 두 번째, 소비자가 **관심 있어 하는 분야**의 뉴스를 추천해야 한다. 좋아하는 것만 봐도 시간이 부족한 정보화 사회 속에서 관심 없는 주제를 찾아보려 하는 소비자는 없다. 마지막으로 **핵심만을 전달**해야 한다. 미디어의 발달로 Short-form 형식의 영상에 익숙해져 핵심만을 전달받고자 하는 수요가 늘어나고 있다.

따라서 CAUsumer는 소비자들의 결재 패턴을 분석하여 접근성 좋은 소비자 맞춤형 뉴스 추천 서비스를 제공하고자 한다.

▢ 분석요약

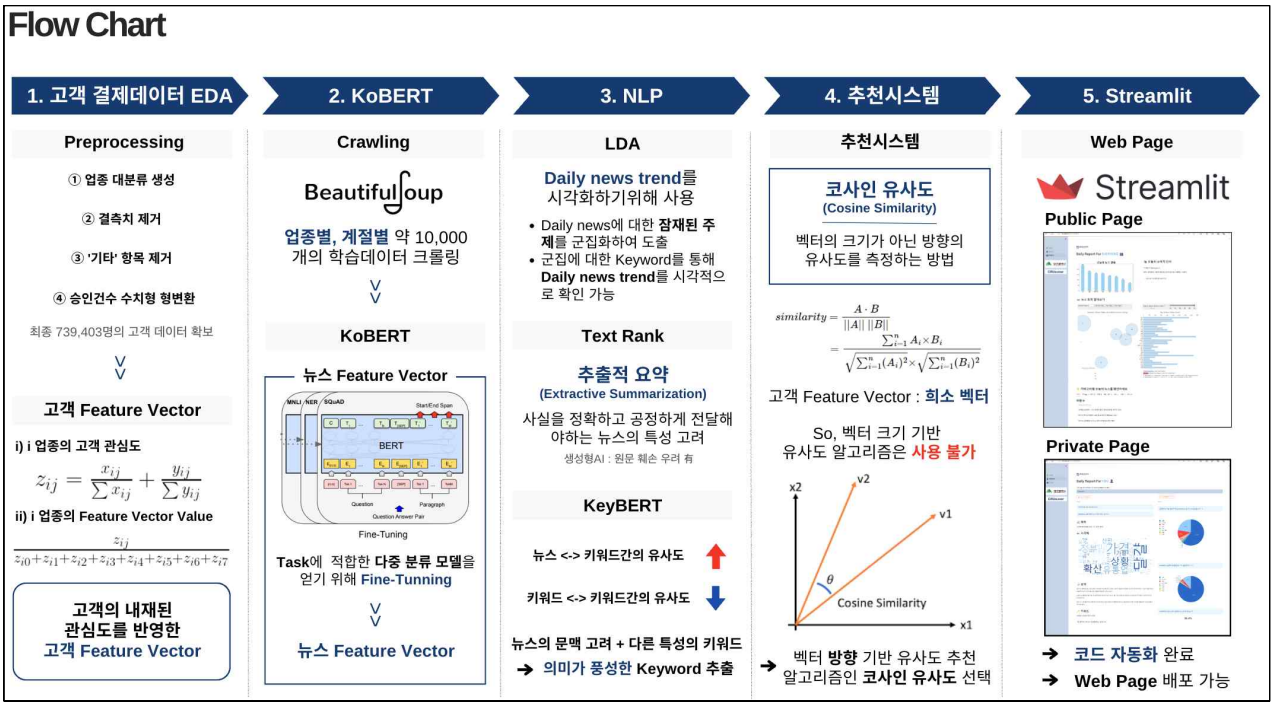
- **활용데이터**
  - 대구은행 BC카드 결제 데이터
  - BC카드 가맹점 업종코드
  - 네이버 뉴스 크롤링 데이터

- **분석도구**
  - 환경 : Jupyter Notebook, Google Colab, VScode
  - 언어 : Python

- **분석결과**
  - Streamlit 라이브러리를 활용하여 웹 페이지 구축

- **분석기법**
  - Crawling : 네이버 뉴스 수집
  - KoBERT : 뉴스 분류
  - LDA : 토픽 모델링, 뉴스 토픽 시각화
  - Text Rank : 뉴스 요약
  - KeyBERT : 뉴스 핵심 키워드 추출

○ **Flow Chart**



<그림-4> 서비스 Flow Chart

Ⅱ. 분석결과

※ 분석 결과물을 먼저 보여드린 뒤, 분석 기법을 설명하는 순서로 조정했습니다.

▢ 활용데이터

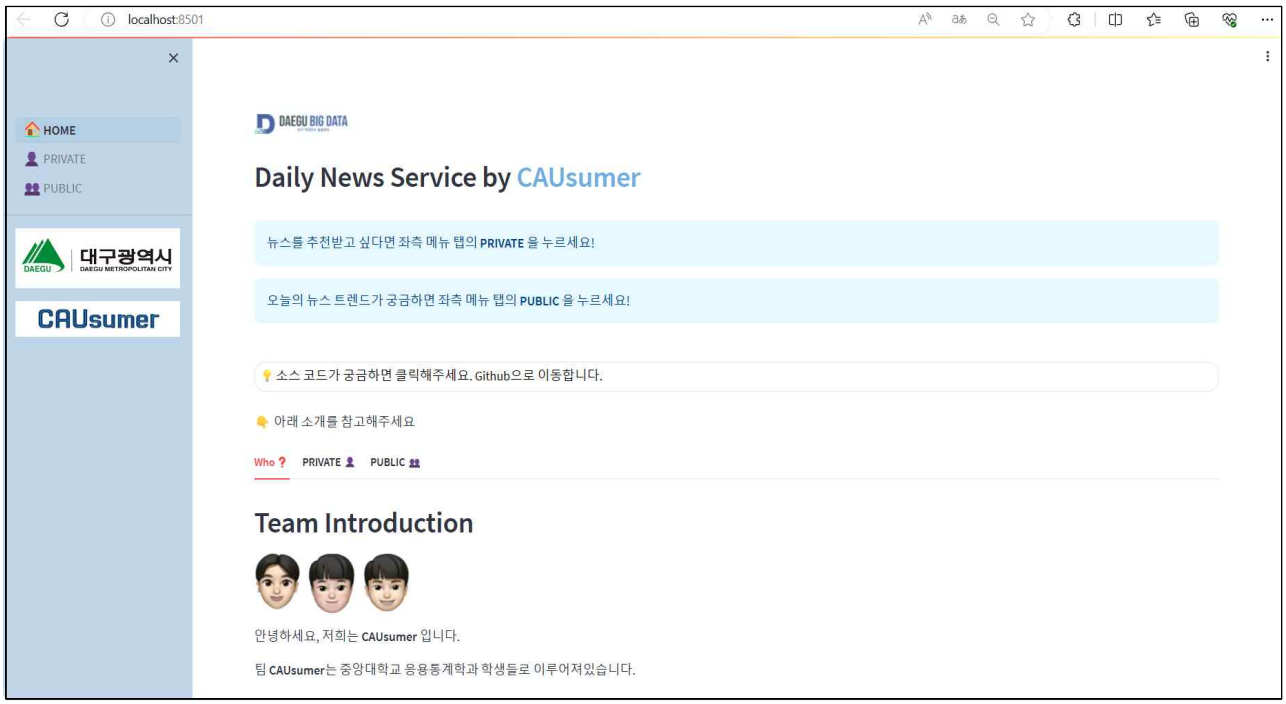
데이터명	형식	대상기간	사용변수	출처	비고
대구은행 BC카드 결제 데이터	csv	2022년.12월1일~ 2022년.12월31일	고객ID, 가맹점업종코드, 가맹점업종명, 승인금액, 승인건수	대구은행	-
BC카드가맹점 업종코드	csv	-	가맹점업종, 가맹점분류1, 가맹점분류2, 가맹점업종명	주최 측	-
네이버 뉴스 크롤링 학습데이터	csv	2018년. 6월 ~ 2023년 .5월	url, title, content, label	네이버 뉴스	https://news.naver.com
네이버 뉴스 크롤링 daily 데이터	csv	활용시점 기준 만 1일 (ex. 2022.12.12)	url, title, content, label	네이버 뉴스	https://news.naver.com

▢ 구현된 모습

streamlit 라이브러리를 활용하여 분석 결과를 담은 WepPage 구축

○ 홈 화면

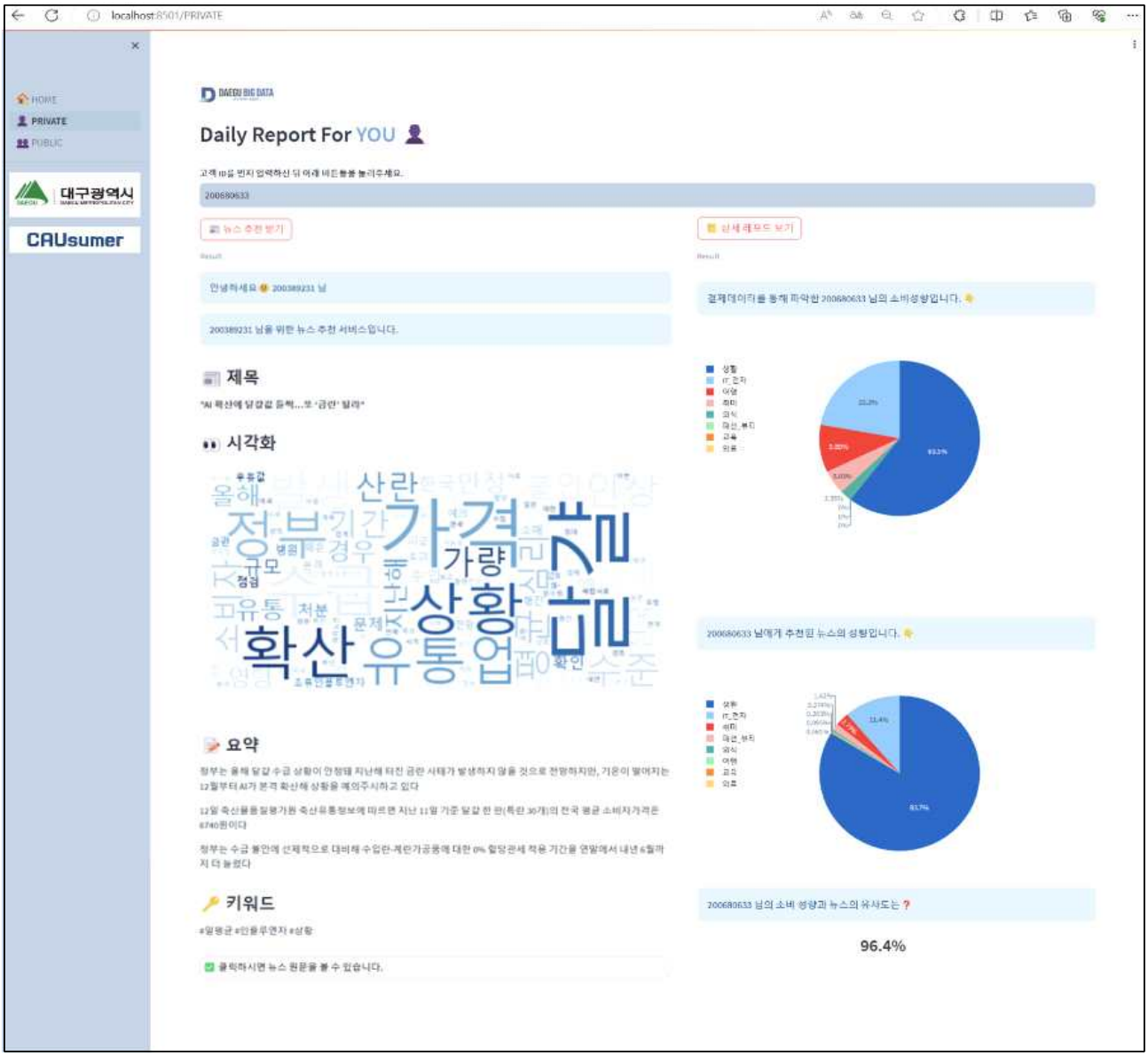
- CAUsumer 팀 소개, 서비스 소개
- Github(소스코드) 하이퍼링크 버튼



<그림-5> Home 화면

○ **Private 화면 (개인화 서비스)**

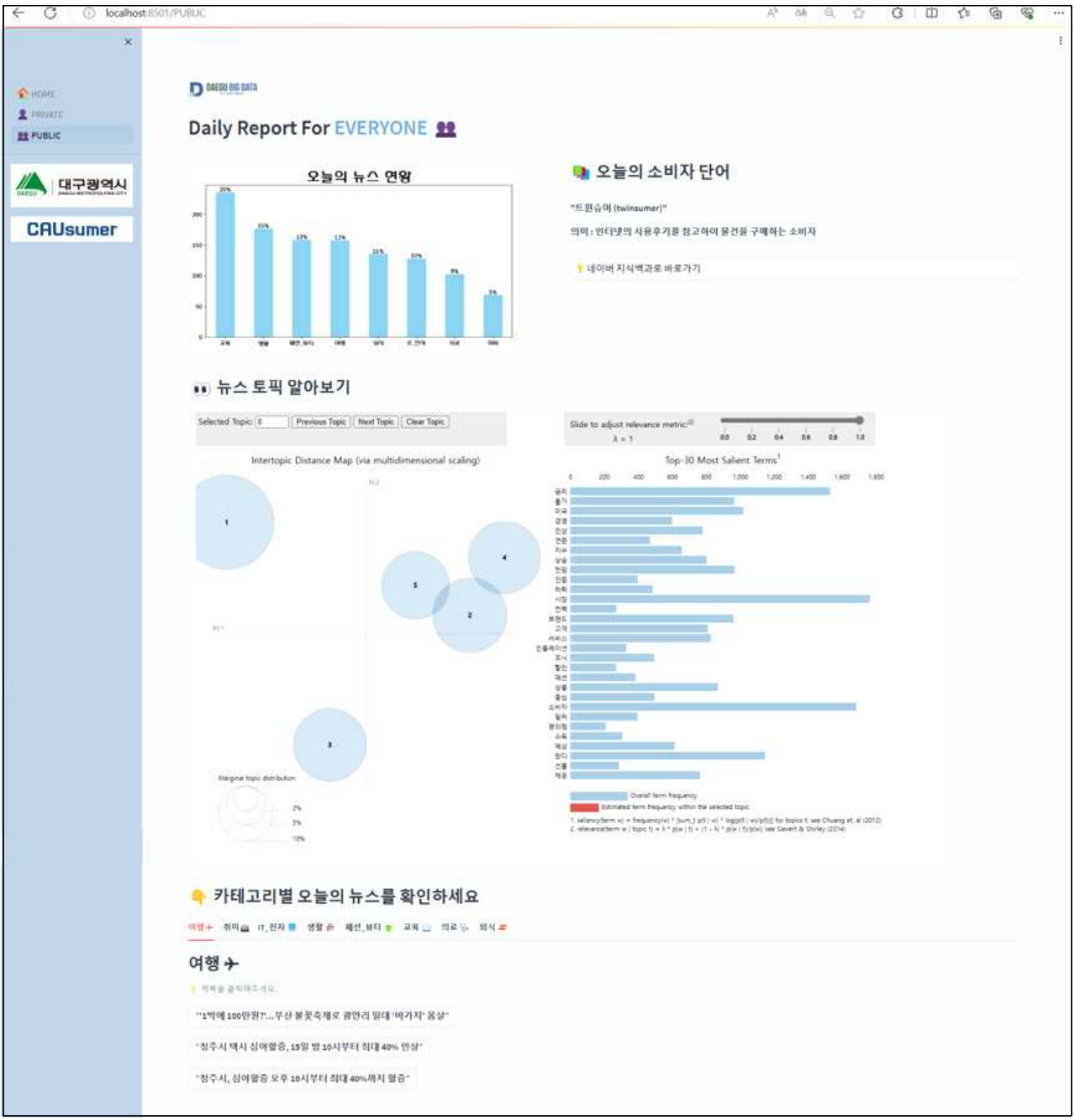
- 고객 ID 입력 받기
- "뉴스 추천 받기" 클릭
  - 고객에 적합한 뉴스 추천
  - 뉴스 제목 출력
  - 시각화(Word Cloud)
  - 뉴스 요약(Text Rank)
  - 뉴스 핵심 키워드(KeyBERT)
- "상세 레포트 보기" 클릭
  - 고객 Feature Vector 시각화
  - 뉴스 Feature Vector 시각화
  - 유사도 출력



<그림-6> Private 화면

○ Public 화면 (공공 서비스)

- 오늘의 뉴스 현황
  - 데일리 뉴스를 KoBERT 모델을 이용하여 8개 업종으로 분류
- 오늘의 소비자 단어
  - 소비자들이 알면 좋을 소비자 단어 출력
  - 네이버 지식백과 하이퍼링크 버튼
- 뉴스 토픽 알아보기
  - 토픽 모델링(LDA)을 통해 뉴스 트렌드 파악
- 업종별 뉴스 확인
  - 데일리 뉴스들을 업종별로 3개 출력
  - 제목 클릭 시, 우측에 시각화, 요약, 키워드, 원문 링크 제시



<그림-7> Public 화면1

<그림-8> Public 화면2

Ⅲ. 분석방법

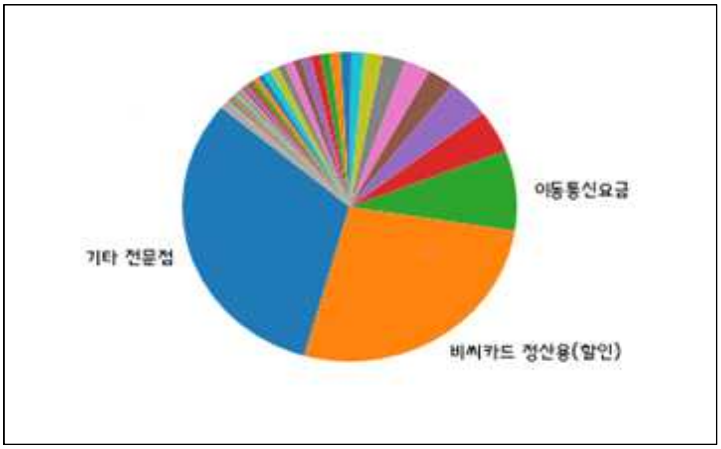
○ 고객 결제 데이터 분석 (고객 Feature Vector 도출)

▪ 사용 패키지 : pandas, numpy, matplotlib, re

< 8개의 업종대분류 도출 >

BC카드 업종 분류표, 대한민국 표준 산업 분류표 그리고 소비자들의 주된 소비 유형을 고려하여 9개('여행', '취미', 'IT/전자/자동차/', '생활', '패션/뷰티', '교육', '의료', '외식', '기타')의 업종대분류를 설정하였다. 각 업종대분류에 가맹점업종명을 대응시키고 결측치를 처리하였다.

업종대분류가 '기타'인 가맹점의 업종명을 확인해 본 결과, <그림-9>처럼 '기타 전문점'이 31.8%, '비씨카드 정산용(할인)'이 26.9%임을 파악했고 과반수의 업종이 어떤 사업을 영위하는지 알 수 없다고 판단하였다. 이에 따라 개인의 결제 패턴을 고려하여 맞춤 뉴스를 추천할 때, '기타' 업종이 불필요한 항목이라 판단하여 삭제하고 총 8개의 업종대분류 항목을 설정하였다.



<그림-9> 업종대분류가 '기타'인 가맹점의 업종명 pie 차트

< 변수 변환 >

'대구은행 BC카드 결제 데이터'의 '승인건수' 변수는 5개의 고유한 값('5번이하', '5번초과 10번이하', '10번초과 15번이하', '15번초과 20번이하', '20번 초과')을 갖는 범주형 변수이다. 고객 Feature Vector 도출 과정에서 수치형 연산이 필요하여 '승인건수'변수를 '승인건수_수치' 변수로 변환하였다.

'5번이하'의 경우, 대응하는 승인금액이 0인 경우에는 0으로 설정하고 그 외에는 0과 5의 평균인 2.5로 변환하였다.

'5번초과 10번이하', '10번초과 15번이하'와 '15번초과 20번이하'는 두 숫자의 평균인 7.5, 12.5와 17.5로 변환하였다.

'20번 초과'는 평균처리를 해줄 수 없어서 다른 방식으로 변환하였다.

승인건수가 '20번 초과'가 아닌 데이터에서 업종대분류별 '승인금액', '승인건수_수치'의 합을 groupby 메소드를 통해 집계하였다. 집계된 데이터 프레임에서 '승인금액'의 합을 '승인건수_수치'의 합으로 나눈 '건당_승인금액' 변수를 생성하였다.

승인건수가 '20번 초과'인 고객의 '승인건수_수치'를 '승인금액 / 건당_승인금액'으로 대체하였고 그 값이 21보다 작은 경우, 해당 범주의 최솟값인 21로 대체해 주었다.

< 고객 Feature Vector 도출 >

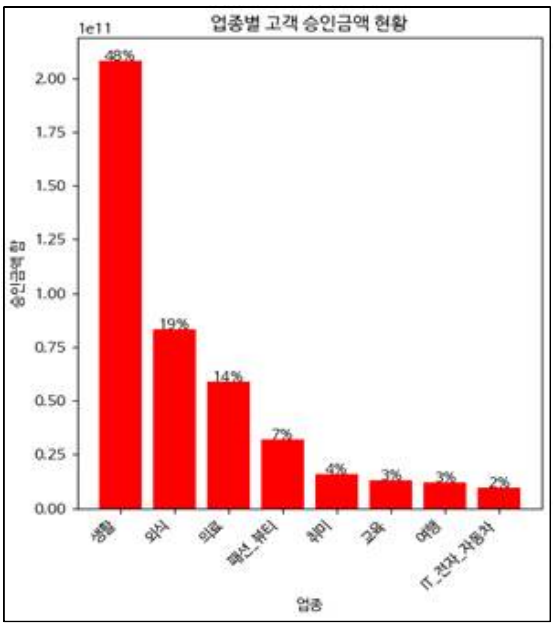
전처리된 데이터프레임에서 '고객ID', '승인금액', '승인건수_수치', '업종대분류' 변수만을 사용하여 '고객ID 별 업종대분류에 대한 승인금액의 합' 피벗테이블과 '고객ID 별 업종대분류에 대한 승인건수의 합' 피벗테이블을 생성한 후 열 방향으로 병합하였다.

즉, <그림-10>처럼 각 고객ID의 8가지 업종대분류 항목에 대한 승인금액 및 승인건수를 한 행에 나타내도록 하였다.

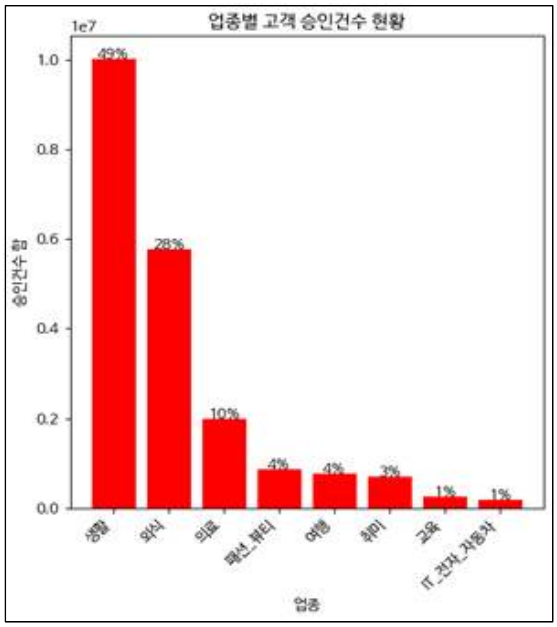
분류	IT_전자_자동차_승인금액	교육_승인금액	생활_승인금액	여행_승인금액	외식_승인금액	의료_승인금액	취미_승인금액	패션_뷰티_승인금액	IT_전자_자동차_승인건수	교육_승인건수	생활_승인건수	여행_승인건수	외식_승인건수	의료_승인건수	취미_승인건수	패션_뷰티_승인건수
고객ID																
200001351	390000.0	0.0	210000.0	440000.0	1820000.0	0.0	500000.0	0.0	7.5	0.0	10.000000	5.0	35.0	0.0	10.0	0.0
200004132	0.0	0.0	410000.0	0.0	210000.0	10000.0	220000.0	20000.0	0.0	0.0	52.500000	0.0	10.0	2.5	10.0	2.5
200004601	0.0	0.0	30000.0	0.0	80000.0	0.0	0.0	20000.0	0.0	0.0	5.000000	0.0	7.5	0.0	0.0	2.5
200005514	0.0	0.0	50000.0	0.0	60000.0	0.0	10000.0	0.0	0.0	0.0	2.500000	0.0	7.5	0.0	2.5	0.0
200006021	1700000.0	0.0	4710000.0	40000.0	390000.0	0.0	80000.0	690000.0	2.5	0.0	258.294584	10.0	27.5	0.0	15.0	26.0

<그림-10> 고객ID 별 업종대분류에 대한 승인금액 및 승인건수 데이터프레임 예시

업종별 총 고객 승인금액은 1위 생활, 2위 외식, 3위 의료 순이며 각각의 승인금액 비중은 전체 대비 각각 48%, 19%, 14%이다. <그림-11>
업종별 총 고객 승인건수은 1위 생활, 2위 외식, 3위 의료 순이며 각각의 승인건수 비중은 전체 대비 각각 49%, 28%, 10%이다. <그림-12>



<그림-11> 업종별 고객 승인금액 현황



<그림-12> 업종별 고객 승인건수 현황

이처럼 상위 항목들의 순위는 동일하지만, 하위항목의 순위는 약간의 차이가 있음을 알 수 있다. 또한, 업종별 고객 승인금액 현황과 업종별 고객 승인건수 현황에서 생활, 외식과 같은 항목에 지나치게 편중된 것을 볼 수 있다. 사람들의 일상적인 소비 비중을 생각해 보았을 때, 생활 및 외식 등 일상적인 항목에 많이 소비하는 것은 합리적인 결과이다. 자연스럽게 많이 소비할 수밖에 없는 항목이 고객의 관심도를 대변하지는 않는다고 판단했다.

예를 들어 A고객이 외식에 50만 원, 패션/뷰티에 30만 원을 소비했을 때, 전체 고객의 외식에 대한 총 승인금액이 500만 원, 패션/뷰티에 대한 총 승인금액이 200만 원이라고 한다면 절대적인 금액은 50만 원을 사용한 외식이 더 크다. 그러나 외식에 대한 A고객의 관심도는 50/500 = 0.1, 패션/뷰티에 대한 관심도는 30/200 = 0.15로 패션/뷰티에 더 큰 관심이 있음을 알 수 있다. 이처럼 단순히 승인금액과 승인건수가 높은 업종을 추천하는 것이 아닌 상대적인 비율로써 업종에 대한 고객의 구매력을 관심도로써 정의했다.
따라서 '[그림-2] 고객ID 별 업종에 대한 승인금액 및 승인건수'의 값을 업종별 총 승인금액 및 승인건수로 나눠서 '업종별 총 승인금액 대비 각 고객의 승인금액', '업종별 총 승인건수 대비 각 고객의 승인건수'를 비율로 생성하고 더하였다. 이는 <수식-1>과 같다.

$$z_{ij} = \frac{x_{ij}}{\sum_i x_{ij}} + \frac{y_{ij}}{\sum_i y_{ij}}$$

<수식-1>
x: 승인금액, y: 승인건수, i: 고객 index, j: 업종대분류 index(0~7)

이로써 자연스럽게 많이 소비할 수밖에 없는 업종의 영향을 줄인 파생변수를 도출하였다. 업종별 고객 Feature Vector 값의 총합이 1이 되도록 스케일을 변환해 주었다. 이는 <수식-2>와 같다.

$$\frac{z_{ij}}{z_{i0}+z_{i1}+z_{i2}+z_{i3}+z_{i4}+z_{i5}+z_{i6}+z_{i7}}$$

<수식-2>

이로써 고객의 소비 패턴을 반영한 고객 Feature Vector 데이터 프레임(<그림-13>)을 도출했다.

분류 고객ID	여행	취미	IT_전자_자동차	생활	패션_뷰티	교육	의료	외식
200001351	0.219834	0.227240	0.402504	0.010102	0.000000	0.0	0.000000	0.140320
200004132	0.000000	0.628817	0.000000	0.162976	0.079721	0.0	0.032402	0.096085
200004601	0.000000	0.000000	0.000000	0.100042	0.548571	0.0	0.000000	0.351386
200005514	0.000000	0.624241	0.000000	0.073351	0.000000	0.0	0.000000	0.302408
200006021	0.048074	0.076986	0.553392	0.141997	0.151858	0.0	0.000000	0.027693
200007203	0.000000	0.000000	0.000000	1.000000	0.000000	0.0	0.000000	0.000000
200007404	0.000000	0.000000	0.000000	0.172281	0.000000	0.0	0.827719	0.000000
200009319	0.093632	0.380982	0.000000	0.066941	0.234598	0.0	0.032754	0.191094
200011167	0.000000	0.000000	0.000000	1.000000	0.000000	0.0	0.000000	0.000000
200011228	0.000000	0.355534	0.000000	0.105617	0.274093	0.0	0.129485	0.135271

<그림-13> 고객 Feature Vector 데이터프레임 예시

○ 텍스트 전처리

■ 사용 패키지 : re, konlpy(Okt, Mecab)

모든 알고리즘에 획일화된 텍스트 전처리 과정을 진행하여 입력값으로 활용하는 것이 아닌 알고리즘별로 그 쓰임에 맞게 텍스트 전처리를 차별화하여 진행하였다. 이를 요약하면 아래 <표-1>과 같다.

알고리즘 / 전처리	KoBERT	LDA	Text Rank	KeyBERT	Word Cloud
형태소 분석	조사, 어미 제거	조사, 어미 제거	X	명사 추출	명사 추출
불용어 처리	O	O	X	O	O
짧은 단어 제거	X	O	X	O	O

<표-1>

KoBERT의 입력값에 대한 전처리로 불용어 처리를 해주었다. 불용어는 8개의 모든 뉴스 업종에 빈번하게 많이 등장하는 단어들이기에 분류 모델에 불용어가 노이즈로 작용할 것이라고 판단하여 불용어 처리를 해주었다. 또한, 조사와 어미를 제거해 줌으로써 단어의 의미를 담고 있는 부분만을 남기고자 했다. 길이가 1인 짧은 단어 중에도 의미를 담고 있는 단어들(ex 책, 차, 몸, 병)이 일부 존재했기 때문에 분류 모델이 이를 학습하도록 짧은 단어는 제거하지 않았다.

LDA는 토픽 모델링 알고리즘으로 그날 올라온 뉴스 기사들의 토픽을 파악할 수 있는 알고리즘이다. 토픽 파악에 방해가 되는 불용어와 짧은 단어를 제거해 주었고, 단어에서 의미를 담고 있는 부분만을 남기기 위해 조사와 어미를 제거해 주었다.

Text Rank는 뉴스를 요약해 주는 알고리즘이다. 원문을 최대한 훼손하지 않고 요약해 주기 위해 별다른 전처리를 하지 않고 특수문자, 구두점 제거 정도만 진행해 주었다.

WordCloud와 KeyBERT는 동일한 전처리를 진행해 주었다. 단어들을 시각적으로 제시함으로써 뉴스의 핵심을 파악하는 알고리즘이기 때문에, 핵심 파악에 방해가 되는 불용어와 짧은 단어들을 제거해 주었다.

○ 크롤링 (학습 및 데일리 데이터 수집)

■ 사용 패키지 : BeautifulSoup, schedule

< 학습 데이터 >

뉴스 분류를 위한 네이버 뉴스 즉, 학습 데이터를 수집하였다. 5년간의 뉴스들을 (2018.06~2023.05) 관련도순으로 약 10,000개를 수집하였다. 네이버 뉴스의 구조상 1년의 기간을 두고 관련도순으로 데이터를 수집하면 최신 뉴스에 높은 가중치가 부여됨을 확인했다. 또한, 특정 계절에만 자주 등장하는 뉴스들이 있는 것처럼 뉴스는 계절성이 존재한다. 이 점을 고려하여, 수집되는 뉴스가 특정 계절에 편향되지 않도록 기간을 설정하여 사계절(봄(3~5월), 여름(6~8월), 가을(9~11월), 겨울(12~2월))의 뉴스를 고루 수집하였다.

고객 결제 데이터에서 도출한 8개의 업종으로 뉴스를 분류하기 위해, 8개의 업종을 기준으로 뉴스 데이터의 검색 키워드를 결정하였다. 그 결과는 아래 <표-2>와 같다.

업종명(label)	검색 항목
여행(0)	소비,여행 | 소비,숙소 | 소비,펜션 | 소비,호텔 | 소비,모텔 | 소비,항공 | 소비,택시 | 소비,버스 | 소비,기차 | 소비,교통 | 소비,렌트카
취미(1)	소비,레져 | 소비,공연 | 소비,골프 | 소비,악기 | 소비,스포츠 | 소비,영화 | 소비,티켓팅
IT_전자_자동차(2)	소비,스마트폰 | 소비,아이폰 | 소비,갤럭시 | 소비,가전제품 | 소비,국산차 | 소비,외제차
생활(3)	소비,할인 | 소비,마트 | 소비,편의점 | 소비,주유 | 소비,생활용품 | 소비,생필품 | 소비,가구 | 소비,주방용품
패션_뷰티(4)	소비,의류 | 소비,패션 | 소비,뷰티 | 소비,패션잡화 | 소비,화장품 | 소비,미용 | 소비,백화점
교육(5)	소비,교육 | 소비,문제집 | 소비,학원 | 소비,도서 | 소비,교재 | 소비,입시 | 소비,수능 | 소비,내신 | 소비,학종
의료(6)	소비,병원 | 소비,약국 | 소비,건강식품 | 소비,보험 | 소비,한약 | 소비,진료
외식(7)	소비,외식 | 소비,간편식 | 소비,한식 | 소비,중식 | 소비,양식 | 소비,일식 | 소비,주점 | 소비,제과점

<표-2>

검색 항목의 형식 ‘,’는 and를 의미하고, ‘|’는 or를 의미한다. 소비자들을 위한 뉴스를 수집하기 위해 “소비”라는 단어를 포함했다.

예를 들어, “소비,여행 | 소비,숙소”가 검색항목이라면 소비와 여행 또는 소비와 숙소가 들어간 뉴스만을 수집하도록 하는 의미를 가진다. 아래<그림-14>는 의료 업종에 대해 수집된 학습데이터의 일부이다. 의료 업종의 뉴스가 잘 수집된 것을 확인할 수 있다.

	company	url	title	content	label
0	조세일보	https://n.news.naver.com/mnews/article/123/000...	동아제약, 챔프시럽 '빨간색'만 회수…"책임 다할 것"	- 챔프 이부펜 시럽 등 문제없어…소비자 환불 등 게시- "챔프는 빨간색(아세트아...	6
1	이데일리	https://n.news.naver.com/mnews/article/018/000...	'손해사정사 선임권' 알고 계셨나요?…보험사 안내문 들어보니	'보험사 셀프 논란 방지' 손해사정사 선임권, 법제화 3년"달라진게 없다" 국내 ...	6
2	한국경제	https://n.news.naver.com/mnews/article/015/000...	시능만 내는 '비대면 초진'	복지부, 6월 1일부터 시범사업 시행소아과, 야간·휴일 상담만 가능…처방은 못 받...	6

<그림-14> 의료 업종 학습데이터 일부

< 데일리 데이터 >

데일리 데이터란 뉴스 추천에 활용할 데이터를 의미한다. CAUsumer의 서비스를 활용한 시점으로부터 만 하루의 뉴스 데이터를 추천받도록 하기 위해 schedule 라이브러리를 활용하여 1시간 간격으로 최신 뉴스를 업데이트하고 서비스 활용 시점으로부터 24시간 전의 뉴스를 갱신하도록 코드를 자동화하였다.

예를 들어, 25일 저녁 6시에 5~6시에 올라온 기사를 수집하여 업데이트한다면 24일 5시~6시의 뉴스 데이터는 데이터베이스에서 사라지는 구조로 추천 시점으로부터 24시간의 데이터베이스가 항상 유지되도록 구축하였다.

소비자 뉴스만을 추천하기 위해 "소비" 키워드의 데일리 뉴스를 수집하였고 추천에 이용되는 뉴스 Feature Vector를 만드는 데에 사용하였다.

○ KoBERT (뉴스 분류 -> 뉴스 Feature Vector 도출)

▪ 사용 패키지 : torch, transformer, gluonnlp, sklearn

고객 결제 데이터를 기반으로 도출한 8개의 업종으로 뉴스를 분류하기 위해 한국어 뉴스 분류 모델을 구축하였다.

< 모델 선정 >

BERT(Bidirectional Encoder Representations from Transformers)는 문장 분류에 뛰어난 성능을 보이는 사전학습(pre-trained) 모델이다. 그 성능이 입증된 만큼 우리는 BERT 계열의 모델을 사용하였다. KoBERT(Korean BERT)는 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치를 학습한 모델이다. 한국어 문장으로만 사전학습이 이루어졌기 때문에 기존 BERT(base multilingual cased)보다 우수한 성능을 보일 것이라고 판단했다. 또한, KoBERT의 사전학습 과정에서 뉴스에서 수집된 한국어 문장이 사용되었기 때문에, 즉 **비슷한 도메인의 데이터를 학습한 모델이기에 한국어 뉴스 분류에 더욱 적합한 모델**이라고 판단하였다.

< 모델 입력 >

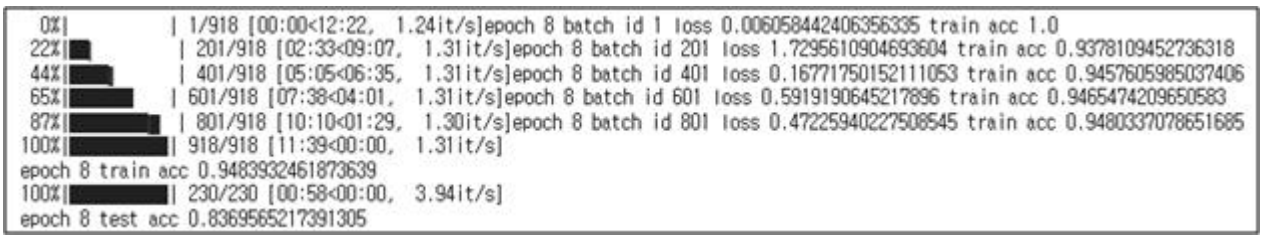
우선, 수집한 학습 데이터를 전처리한 뒤 KoBERTTokenizer로 모델 입력 값으로 사용할 수 있게 처리해주었다. 입력 데이터에 대한 임베딩 벡터, Segment 임베딩, Attention Mask를 만들어 주었고, 모델의 입력 값으로 사용하였다.

< 전이학습 (Transfer Learning) >

전이학습이란 사전 학습된 모델의 지식을 다른 Task에서 활용하는 기법을 말한다. 우리는 사전 학습된 KoBERT를 전이학습을 통해 활용했다. 사전 학습 모델의 layer들은 모두 동결시킨 뒤, 크롤링한 학습 데이터를 추가로 학습시키는 미세 조정(Fine-Tuning) 과정을 진행했다. <그림-15>은 설계한 모델의 구조이다. KoBERT layer들 아래에 뉴스 분류를 위한 Linear layer를 한 층 쌓았고, 출력 크기를 8로 설정하여 8개의 클래스로 뉴스를 분류하는 모델을 구축하였다.

<그림-15> 모델 구조

학습 데이터를 train_test_split 함수를 통해 8:2로 분할하였고, stratify 파라미터를 통해 검증용 데이터 셋에 8개의 클래스가 학습용 데이터 셋과 유사한 분포로 분할되게 하였다. 옵티마이저는 다양한 옵티마이저(Nadam, Adam, AdamW, RMSProp)를 시도한 뒤 가장 좋은 성능을 보인 AdamW를 사용하였고, 손실함수로는 CrossEntropy를 사용하였다. 하이퍼 파라미터 중 배치 크기, 학습률, 에포크를 조절하면서 가장 높은 성능을 보이는 모델을 선정하였다. 그 결과 최종적으로 <그림-16>처럼 validation accuracy가 약 84%인 분류 모델을 구축할 수 있었다.



<그림-16> 모델 성능 평가

< 모델 출력 >

모델의 출력층에서는 Softmax 함수를 사용했다. Softmax 함수는 다중 클래스 분류에서 사용되는 함수로, 각 클래스에 속할 확률을 추정할 수 있다. 출력값을 아래의 예시처럼 뉴스의 Feature Vector로써 사용할 수 있다.

예시)

모델을 통해 A뉴스를 분류한 결과 "여행"일 확률이 0.1, "패션_뷰티"일 확률이 0.6, "교육"일 확률이 0.3 이라고 한다면, A뉴스는 [0.1, 0, 0, 0, 0.6, 0.3, 0, 0] 라는 뉴스 Feature Vector를 갖게 된다.

○ 추천 알고리즘

■ 사용 패키지 : numpy

고객 결제 데이터 분석을 통해 도출한 고객 Feature Vector와 KoBERT 모델을 통해 도출한 뉴스 Feature Vector 간의 유사도를 파악한 뒤, 유사도가 가장 높은 뉴스를 고객에게 추천하였다.

유클리디안 거리(Euclidean distance)는 아래 <수식-3>처럼 벡터 간의 직선거리를 계산함으로써 벡터 간의 유사도를 파악할 수 있는 척도이다. 고객 Feature Vector는 결제 패턴의 특성상 0으로 채워진 값이 많은 희소 벡터(Sparse Vector)이기 때문에, 유클리디안 거리를 통해 유사도를 측정한다면 벡터 공간상에서 데이터 포인트 간의 거리 계산 시 0 값에 큰 영향을 받게 되어 잘못된 거리 측정이 발생할 수 있다. 따라서 다른 유사도 척도를 고려하였다.

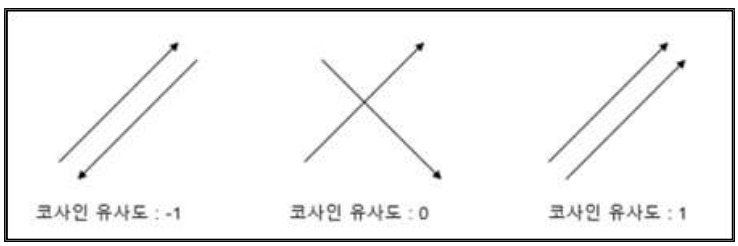
$$\sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\ \cdots\ +(q_n-p_n)^2}=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$

<수식-3>

코사인 유사도(Cosine Similarity)란 아래 <수식-4>, <그림-17>처럼 벡터의 크기보단 방향성에 초점을 맞추어 두 벡터의 방향이 얼마나 유사한지 나타내는 척도이다. 분모에서 두 벡터 간의 성분 곱을 통해 연산이 이루어지기 때문에 0 값에 강건하고, 희소 벡터 연산에 적합하다고 판단하였다. 따라서 코사인 유사도를 기반으로 고객에게 적합한 뉴스를 추천하는 알고리즘을 구축하고자 하였고, numpy.linalg의 norm 라이브러리와 numpy의 dot 라이브러리로 유사도 산식을 함수로써 구현한 뒤, 가장 높은 유사도를 보이는 뉴스를 추천하도록 하였다.

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

<수식-4>



<그림-17>

○ Text Rank (뉴스 요약)

■ 사용 패키지 : numpy, math, re

< 왜 추출적 요약인가? >

텍스트 요약은 크게 추출적 요약(Extractive Summarization)과 추상적 요약(Abstractive Summarization)으로 나눌 수 있다. 추출적 요약이란 원문에서 중요한 핵심 문장 혹은 단어구를 그대로 추출하여 요약하는 방식이고, 추상적 요약이란 원문에 없던 문장이라도 핵심 문맥을 반영한 새로운 문장을 생성해서 원문을 요약하는 방식이다.

추상적 요약은 원문의 의도를 바꿀 우려가 있어 객관적 사실을 전달하는 뉴스의 특성을 고려하여 추상적 요약이 아닌 추출적 요약을 사용하기로 했다. 또한, 추상적 요약에 비해 추출적 요약은 별도의 학습 데이터가 필요하지 않고, 모델 특성상 학습도 매우 빠르다는 장점이 있다. 뉴스의 의도를 바꾸지 않고, 빠르게 뉴스를 요약하기 위해 대표적인 추출적 요약 알고리즘인 Text Rank를 사용하였다.

< 구현 >

Text Rank 알고리즘은 페이지랭크(Page Rank)를 기반으로 한 요약 알고리즘이다. 특이한 점은 문장 간 유사도 측정을 위해 코사인 유사도가 아니라 아래<수식-5>와 같은 다른 척도를 사용했다는 점이다. 문장 간 유사도를 구할 때, 두 문장에 공통으로 등장한 단어의 개수를 각 문장의 단어 개수의 log 값의 합으로 나누어 준다.

$$\mathrm{Sim}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \wedge w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)}$$

<수식-5>

$$\mathrm{TR}(V_i) = (1-d) + d \sum_{j=0}^{N-1} \frac{\mathrm{Sim}(S_i, S_j)}{\sum_{k=0}^{N-1} \mathrm{Sim}(S_j, S_k)} \times \mathrm{TR}(V_j)$$

<수식-6>

이를 통해 구한 유사도를 바탕으로 <수식-6>와 같이 문장 별 RANK 값을 계산한 뒤, 높은 RANK 값이 부여된 문장이 문서의 핵심 문장이 된다. 이 과정을 Python math 라이브러리를 활용하여 함수로써 구현하였다.

뉴스의 문장에 적용해 본 결과, 뉴스에서 여러 번 등장한 단어가 핵심 문장에 포함 되어있었다. <그림-18>은 **"AI 확산에 달걀값 들썩.... 또 '금란' 될라"** 라는 제목의 기사를 요약한 것이다. 직관적으로 뉴스가 잘 요약되었음을 알 수 있다.

정부는 올해 달걀 수급 상황이 안정돼 지난해 터진 금란 사태가 발생하지 않을 것으로 전망하지만, 기온이 떨어지는 12월부터 AI가 본격 확산해 상황을 예의주시하고 있다
12일 축산물품질평가원 축산유통정보에 따르면 지난 11일 기준 달걀 한 판(특란 30개)의 전국 평균 소비자가격은 6740원이다
정부는 수급 불안에 선제적으로 대비해 수입란 · 계란가공품에 대한 0% 할당관세 적용 기간을 연말에서 내년 6월까지 더 늘렸다

<그림-18> Text Rank 출력 결과

○ KeyBERT (뉴스 핵심 키워드 추출)

▪ 사용 패키지 : keybert

KeyBERT는 BERT의 임베딩을 활용한 키워드 추출 방법론이다. 기존의 키워드 추출 방법론들(Text Rank, TF-IDF 등)은 단어의 등장 빈도에 의존하는 경우가 많았다. 하지만 KeyBERT는 단어의 의미를 잘 표현할 수 있는 임베딩을 BERT로부터 도출해 사용하기 때문에 기존의 방법론들에 비해 더 핵심적인 키워드가 추출되겠다고 판단하였다.

또한, KeyBERT는 키워드의 의미적 다양성을 조절할 수 있는 파라미터가 존재한다. diversity 파라미터와, use_mmr 파라미터를 다양한 조합으로 시도해 보면서 핵심적인 키워드가 잘 추출되는 경우의 수를 탐색하였고, 핵심 키워드 추출에 적합한 파라미터를 아래<표-4>와 같이 설정하였다.

< KeyBERT 파라미터 설정 >

파라미터	값	비고
keyphrase_ngram_range	(1,1)	단어 단위의 출력을 위해 ngram의 윈도우 사이즈를 1로 설정
diversity	0.2	다음과 같이 설정하여 다양성 조절
use_mmr	False	
top_n	3	상위 3개의 키워드 추출

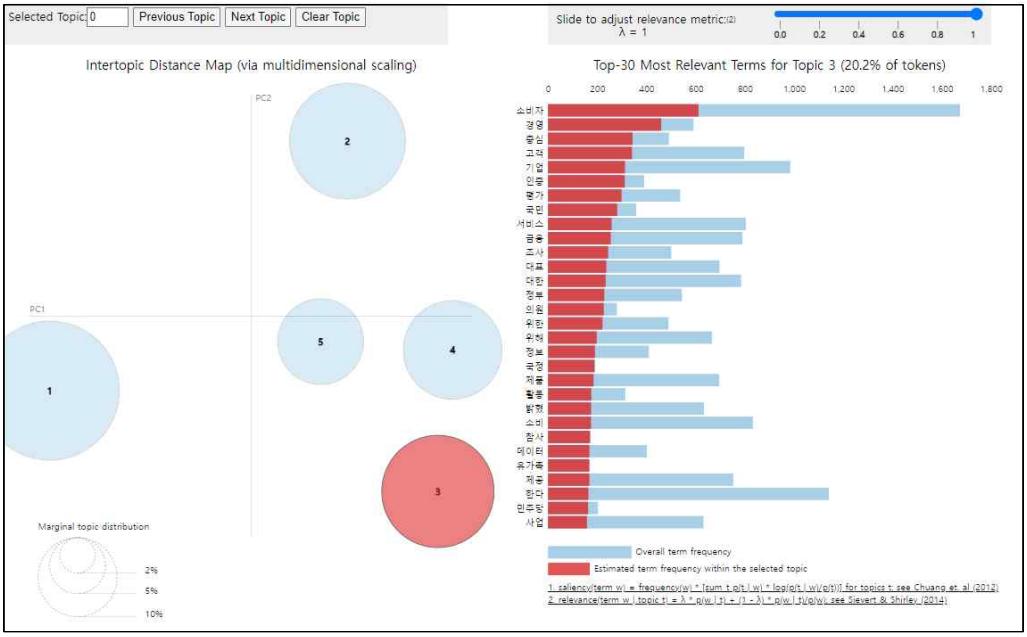
<표-4>

○ LDA (토픽 모델링)

▪ 사용 패키지 : gensim, pyLDAvis

LDA(Latent Dirichlet Allocation, 잠재 디리클레 할당)는 문서 데이터에서 핵심 주제(Topic)를 찾는 알고리즘이다. LDA는 확률 기반의 모델링 기법을 통해 방대한 양의 문서 데이터를 분석함으로써 문서 내에 어떤 Topic이, 어떤 비율로 구성되어 있는지 분석한다. 또한, Topic 별로 어떤 키워드가 구성되어있는지 정보를 제공하여 이에 따른 인사이트를 도출할 수 있다.

LDA의 결과물은 pyLDAvis 라이브러리를 활용하여 시각화하기 용이하기 때문에 <그림-19>, 데일리 뉴스들의 트렌드를 파악하기에 적합한 알고리즘이라고 판단하였다.



<그림-19> LDA 출력 결과

○ Streamlit (웹 페이지 구축 및 다양한 서비스 구현)

▪ 사용 패키지 : streamlit, plotly, matplotlib, WordCloud, Counter, random

Streamlit은 데이터 기반 웹 애플리케이션을 만드는 라이브러리이다. 쉽게 배포까지 가능하다는 장점이 있기 때문에 분석 결과를 Streamlit을 통해 웹 페이지로 구현함으로써 서비스의 실질적인 활용이 가능하게 하였다. Streamlit의 set_page_config 함수를 사용하여 Home 화면, Private 화면, Public 화면 총 3가지 화면으로 UI를 구성하였다.

※ <Ⅱ. 분석결과> 참고

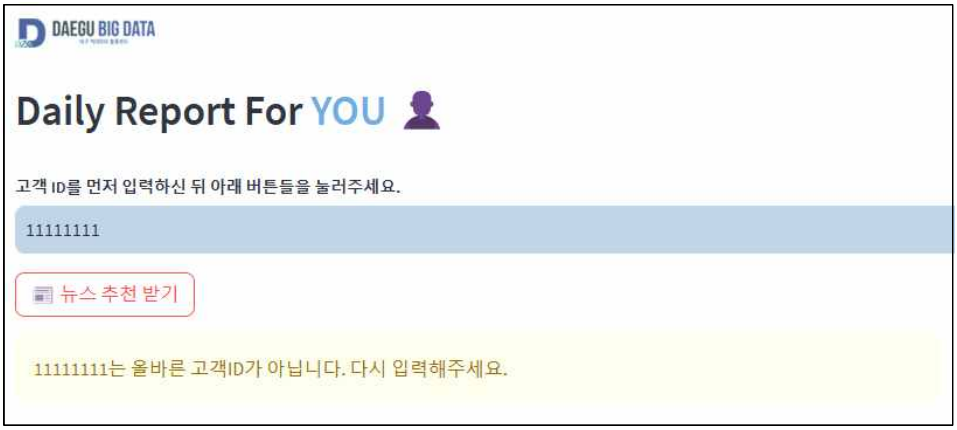
< Home 화면 >

CAUsumer의 팀 소개와 우리의 서비스를 접하는 고객에게 서비스를 소개해 주는 내용을 구현하였다. 또한, Markdown 문법을 활용하여 Github(소스코드)으로 바로 이동이 가능한 하이퍼링크 버튼을 추가하였다.

< Private 화면 >

개인화 서비스를 구현하였다. Streamlit의 특성상 화면이 업데이트될 때마다 변수가 새로 할당된다. 그 과정에서 많은 시간이 소요되는 문제를 해결하기 위해 @st.cache_data 데코레이터를 활용하였다. 데코레이터로 시간이 오래 걸리는 출력값들을 캐싱함으로써 반복적인 로딩을 방지하였다.

text_input함수로 고객 ID를 입력받는 입력창을 구현하였다. 입력받은 값을 변수로 할당한 뒤, 고객 결제 데이터의 고객 ID와 비교하여 고객을 특정한 후 개인화 추천을 진행하였다. 데이터에 존재하지 않는 고객 ID가 입력되었을 경우, 아래 <그림-20>처럼 예외 처리를 해주었다.



<그림-20> 예외 처리 예시

고객 ID 입력 후, **"뉴스 추천 받기"** 버튼을 누르면 화면 좌측에 간단한 인사 문구와 추천받은 뉴스의 제목 그리고 뉴스 텍스트의 Word Cloud 시각화가 출력된다. 이어서 Text Rank 알고리즘을 활용한 뉴스 요약 3문장과 KeyBERT를 활용한 핵심 키워드 3개, 마지막으로 뉴스 원문을 읽을 수 있도록 하이퍼링크 기능을 담은 버튼을 구현하였다.

고객 ID 입력 후, **"상세 레포트 보기"** 버튼을 누르면 화면 우측에 plotly 라이브러리를 활용한 고객 Feature Vector pie차트와 추천된 뉴스의 Feature Vector의 pie차트가 출력된다. 마지막에는 코사인 유사도를 통해 도출된 유사도가 출력된다.

< Public 화면 >

데일리 뉴스를 통해 트렌드를 파악할 수 있게 하였다. 좌측 상단에는 matplotlib 라이브러리를 통해 KoBERT 모델이 데일리 뉴스를 분류한 결과를 시각화하였다. 우측 상단에는 random 라이브러리를 활용하여 소비자들이 알아야 하는 소비자 단어를 데이터프레임에서 무작위로 추출하여 소개해 주는 기능을 구현하였다.

화면 중간에는 LDA 결과를 pyLDAvis 라이브러리를 활용하여 시각화함으로써 데일리 뉴스의 트렌드를 파악할 수 있게 하였다.

마지막으로 화면 하단에는 KoBERT 모델이 분류한 데일리 뉴스의 8개의 클래스 중 하나에 속할 확률이 가장 높은 뉴스를 업종별로 3개씩 제시하였다. 각 뉴스의 제목을 클릭하면 우측 하단에 해당 뉴스의 Word Cloud 시각화, 뉴스 요약, 키워드, 원문 링크가 제시되도록 구현하였다.

Ⅳ. 독창성 및 차별성

○ **고객의 관심도를 반영한 고객 Feature Vector**

해당 업종에 대한 전체 고객의 총 승인금액 합, 총 승인건수 합과 비교하여 해당 업종에 얼마나 큰 비용을 지불하는지 수치화했다. 절대적인 승인금액이 크고 승인건수가 많은 것을 관심도로 보지 않고 전체 고객 대비 각 고객이 해당업종에 얼마나 결제했는지로 관심도를 파악했다. 즉, 업종 내 해당 고객의 구매력을 관심도로 평가하였다.

○ **계절성을 반영한 크롤링**

네이버 뉴스를 관련도순으로 정렬할 때, 최신뉴스에 가중치를 두고 정렬하는 구조임을 파악했다. 특정 계절에만 기재되는 소비자 뉴스가 존재하기에 계절성을 고려하여 사계절로 나누어 소비자 뉴스와 관련도가 높은 학습데이터를 수집하였다.

○ **코드 자동화 및 WebPage 배포**

코드를 자동화하여 서비스 활용 시점으로부터 만 하루의 뉴스 데이터를 1시간 간격으로 갱신하여 최신 뉴스를 바로바로 추천 받을 수 있도록 자동화하였다. 또한, WebPage로 배포가 가능하도록 구현하여 고객들의 접근성을 고려하였다. 즉, 분석에만 그치지 않고 분석 결과를 실제로 활용할 수 있게 한 것이다.

Ⅴ. 활용방안

○ 고객에 대한 승인금액과 승인건수 변수만을 제공받아 추천서비스를 제공할 때 제약이 많았는데 대구은행과의 연계를 통해 지속적으로 데이터를 업데이트 받고 고객의 성별, 연령, 메뉴 클릭 로그 데이터 등의 데이터를 추가로 얻을 수 있다면 더욱더 세분화된 맞춤형 추천서비스가 될 것이다.

○ 현재는 WebPage를 구축하여 배포가 가능하도록 한 것에 그쳤지만 CAUsumer APP을 만든다면 접근성 측면에 있어서 더 발전된 서비스를 구현할 수 있을 것이다.

○ 다국어 BERT모델을 활용한다면 해외 뉴스에 대해서도 똑같은 서비스를 적용할 수 있게 되어 언어에 구애받지 않고 폭넓은 소비자에게 정보를 제공해 줄 수 있을 것이다.

○ 대구광역시의 2023년[제5차 소비자 정책 기본계획]을 보면 총 13개의 과제 중 '디지털 전환기 소비자역량 강화' 및 '국민 체감 소비생활 안전 확대'에 대한 주제가 5개에 해당하며 이는 우리의 서비스와 밀접한 주제이다. 그 중 구체적인 과제내용으로 '찾아가는 맞춤형 소비자교육 추진'이 있다. 이는 교육 신청을 접수받으면 직접 찾아가 방문교육을 하는 오프라인 교육이기에 교육받는 인원 및 횟수가 한정되어 있다. 이와 더불어 WebPage 및 APP으로 발전시킨 우리의 서비스를 활용한다면 소비자 능력 향상 및 소비자 피해 사전 예방 등 목표하는 성과를 달성할 수 있을 것이다.

VI. 기타

▢ 참고 자료

https://github.com/SKTBrain/KoBERT : KoBERT
https://github.com/MaartenGr/KeyBERT : KeyBERT
https://www.ranks.nl/stopwords/korean : 한국어 불용어 리스트
https://streamlit.io/ : streamlit
홍진표, 차정원(2009), TextRank 알고리즘을 이용한 한국어 중요 문장 추출