수요자 맞춤형 한국소비자원 웹페이지 구현

('23. 12. 8. 중앙대학교 권동구, 김기남, 유태권)

1. 분석 개요

□ **(분석목적)** 한국소비자원이 생산하는 소비자 정보를 수요자 맞춤으로 제공하는 웹페이지 구현

□ **(분석방법)** 고객 결제데이터를 분석하여 수요자 별 맞춤형 소비자 정보를 제공하는 서비스 구현

○ **분석 요약**

- BeautifulSoup라이브러리를 활용한 데이터 수집
- KoBERT를 활용한 Fine-Tunning 및 보도자료 분류
- TextRank를 활용한 보도자료 요약
- keyBERT를 활용한 보도자료 핵심 키워드 추출
- Cosine Similarity를 활용한 추천시스템
- Streamlit을 활용한 Web Page 구축

□ **(분석데이터)** 대구은행 BC카드 결제정보, 한국소비자원 제공 소비자 데이터

○ 대구은행 BC카드 결제정보

- 고객 ID, 성별, 연령, 가맹점업종코드, 가맹점업종명, 승인금액, 승인건수를 변수로 가지는 데이터

○ 한국소비자원 보도자료, 리콜정보, 소비자상담 통계 등

2. 데이터 수집 및 전처리

□ 데이터 수집

○ 학습 데이터

- 5년간의 뉴스들을 (2018.06 ~ 2023.05) 계절성을 고려하여 관련도순으로 약 10,000개를 수집했다. BC카드 가맹점 업종 코드표를 참고하여 결제 데이터에서 도출한 8개의 업종을 기준으로 뉴스를 수집했다.

○ 추천 데이터

- 한국소비자원 보도자료를 크롤링하여 데이터 수집하였고 데이터베이스를 구축하여 추천에 활용했다.

○ 서비스 데이터

- 한국소비자원 소비자 상담데이터, 국내 리콜 데이터를 크롤링하여 데이터베이스를 구축하여 서비스에 활용했다.

□ 텍스트 전처리

- 모든 알고리즘에 획일화된 텍스트 전처리 과정을 진행하지 않고 알고리즘별로 그 쓰임에 맞게 텍스트 전처리를 차별화하여 진행했다.

	KoBERT	TextRank	keyBERT	Word Cloud
형태소 분석	조사, 어미 제거	X	명사 추출	명사 추출
불용어 처리	O	X	O	O
짧은 단어 제거	X	X	O	O

3. 분석 결과

□ 고객 결제 데이터 분석

○ 고객 Feature Vector 도출

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

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

- 고객ID 별 업종에 대한 승인금액 및 승인건수의 값을 업종별 총 승인금액 및 승인건수로 나눠서 '업종별 총 승인금액 대비 각 고객의 승인금액 ', '업종별 총 승인건수 대비 각 고객의 승인건수 '를 비율로 생성하고 더했다. 이로써 자연스럽게 많이 소비할 수밖에 없는 업종의 영향을 줄인 파생변수를 도출했다. 즉, 절대적인 금액이 아닌 고객 간 상대적인 구매력을 고객의 관심으로 정의했다.

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

- 업종 별 고객 Feature Vector 값의 총합이 1이 되도록 스케일을 변환해주었다. 이로써 고객의 소비 패턴을 반영한 고객 Feature Vector 데이터 프레임을 도출했다.

분류 고객ID	IT_전자_자동차	교육	생활	여행	외식	의료	취미	패션_뷰티
200001351	0.402504	0.0	0.010102	0.219834	0.140320	0.000000	0.227240	0.000000
200004132	0.000000	0.0	0.162976	0.000000	0.096085	0.032402	0.628817	0.079721
200004601	0.000000	0.0	0.100042	0.000000	0.351386	0.000000	0.000000	0.548571
200005514	0.000000	0.0	0.073351	0.000000	0.302408	0.000000	0.624241	0.000000
200006021	0.553392	0.0	0.141997	0.048074	0.027693	0.000000	0.076986	0.151858

□ KoBERT

○ 보도자료 Feature Vector 도출

- KoBERT는 위키피디아나 뉴스 등에서 수집한 수백만 개의 한국어 문장으로 이루어진 대규모 말뭉치를 학습한 모델이다. 한국어 문장으로만 사전학습이 이루어졌기에 기존 BERT보다 우수한 성능을 보일 것

이라 판단했다. 비슷한 도메인의 데이터를 학습한 모델이기에 보도자료 분류에 적합한 모델이라 판단했다.

- 수집한 학습 데이터를 전처리한 후, 사전학습된 KoBERT를 전이학습을 통해 활용했다. 사전 학습 모델의 layer들은 모두 동결시킨 뒤, 크롤링한 학습 데이터를 추가로 학습시키는 Fine-Tunning 과정을 진행했다. KoBERT layer들 아래에 뉴스 분류를 위한 Linear layer를 한 층 쌓았고, 출력 크기를 8로 설정하여 8개의 클래스로 보도자료를 분류하는 모델을 구축했다. 해당 모델은 고객 Feature Vector와 같은 형식을 가지는 보도자료 Feature Vector를 생성하고 추천서비스에 사용된다.

□ 추천 알고리즘

- 고객 Feature Vector는 결제 패턴의 특성상 0으로 채워진 값이 많은 희소 벡터(Sparse Vector)이기 때문에, 코사인 유사도를 기반으로 고객에게 적합한 보도자료를 추천하는 알고리즘을 구축했다. numpy.linalg의 norm 라이브러리와 numpy의 dot 라이브러리로 유사도 산식을 함수로써 구현한 뒤, 가장 높은 유사도를 보이는 보도자료를 추천하도록 했다. 단순히 소비금액이 많은 업종에 대한 보도자료를 추천하기보다 고객의 결제패턴과 가장 유사한 보도자료를 추천함으로써 세밀한 개인화 추천서비스가 가능하다.

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

□ TextRank

- 추출적 요약은 별도의 학습 데이터가 필요하지 않고, 모델 특성상 학습도 매우 빠르다는 장점이 있다. 뉴스의 의도를 바꾸지 않고, 빠르게 뉴스를 요약하기 위해 추출적 요약 알고리즘인 Text Rank를 사용했다.

☐ keyBERT

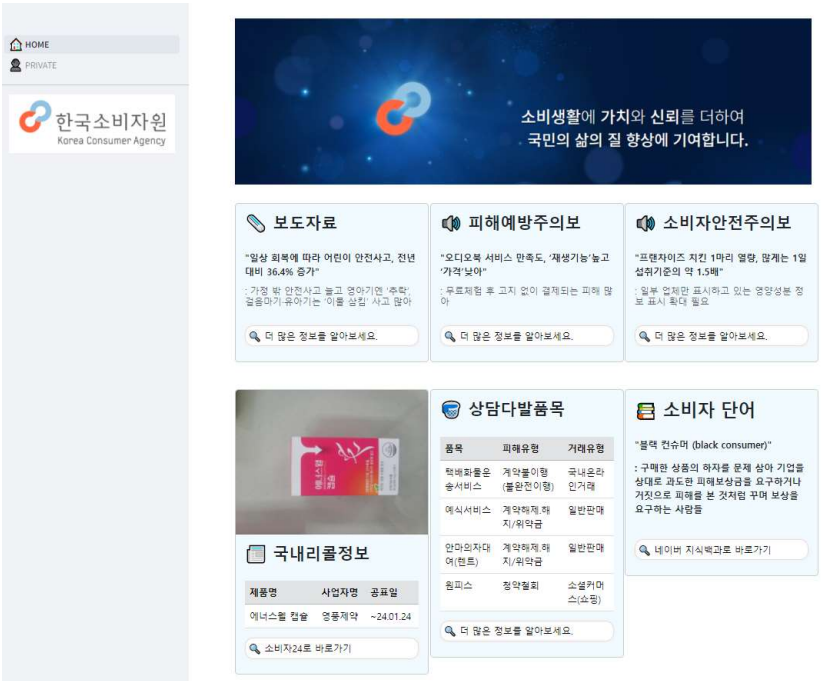
- 기존의 키워드 추출 방법론들 (TF-IDF 등)은 단어의 등장 빈도에 의존하는 경우가 많았다. 하지만 KeyBERT는 단어의 의미를 잘 표현할 수 있는 임베딩을 BERT로부터 도출해 사용하기에 기존의 방법론들에 비해 문서 전체의 문맥을 고려한 핵심적인 키워드가 추출된다.

☐ Streamlit

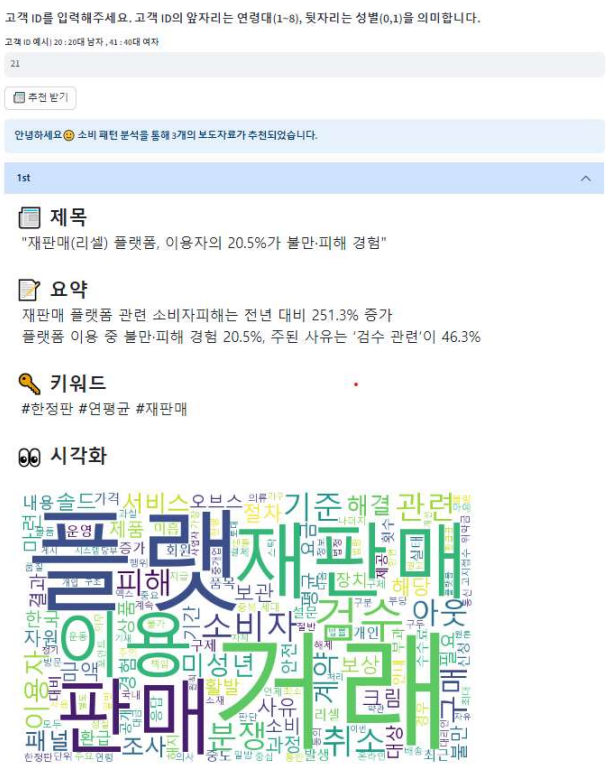
- Streamlit은 데이터 기반 웹 애플리케이션을 만드는 라이브러리이다. 개인화된 수요자 맞춤형 Web Page를 구축 및 배포하여 분석에서 그치는 것이 아닌 서비스 상용화를 이루었다.

- Streamlit의 특성상 화면이 업데이트 될때마다 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.

〈Home〉



〈Private〉



4. 활용 및 개선 방향

□ 활용 방안

○ 시계열을 고려한 서비스

- 월별 고객 결제데이터를 활용하여 분석을 진행했으므로 소비패턴이 바뀌는 월별, 계절별 맞춤으로 서비스를 제공할 수 있다.

○ 광고 및 소비 생활 레포트

- 고객의 결제패턴을 바탕으로 광고 추천이 가능하고 결제 패턴 분석을 통해 소비자별 소비생활 레포트를 제공할 수 있다.

□ 개선 방향

○ 다국어 BERT 활용

- 다국어 BERT모델을 활용한다면 해외 뉴스에 대해서도 똑같은 서비스를 적용할 수 있게 되어 언어에 구애받지 않고 폭넓은 소비자에게

정보를 제공해 줄 수 있을 것이다.

○ Appication 제작

- 현재는 Web Page를 제작하여 배포하였지만 APP을 제작한다면 접근성 측면에 있어서 더 발전된 서비스를 구현할 수 있을 것이다.

○ 한국소비자원 사용자 로그데이터

- 현재는 한국소비자원의 사용자 로그데이터를 활용할 수 없었지만 동기화된 고객 결제데이터, 소비 상담 데이터, 웹 로그데이터를 지급받을 수 있다면, 초개인화된 소비자 서비스를 구현할 수 있을 것이다.

【참고자료】

□ https://korea-consumer-agency.streamlit.app/ : Web Page Link

□ 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