**카페추천시스템**

**[목표]**

카페를 방문하는 목적이 다양해졌으므로 목적에 맞는 카페를 추천해주는 시스템이 필요

**[역할]**

데이터 수집 자동화, 전처리, 평가, 시각화

**[데이터]**

텍스트, 폰트, 번호, 스크린샷이(가) 표시된 사진

자동 생성된 설명

원하는 데이터가 없어 카페 데이터를 직접 크롤링을 통해 수집하기로 했다. 여러 웹 지도들을 살펴본 결과 ‘네이버 지도’의 리뷰 탭에서 이미 labeling된 지표들이 있어 활용하기로 했고, 영수증 인증을 통해 실제로 구매한 사람들만 리뷰를 남길 수 있다는 점에서 신뢰성도 확보했다.

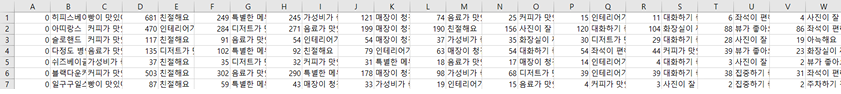
텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

위 스크린샷은 크롤링 코드의 일부분으로 실제 크롤링은 서울시에 위치한 역사명을 받아, 자동화된 크롤링을 수행하게 된다

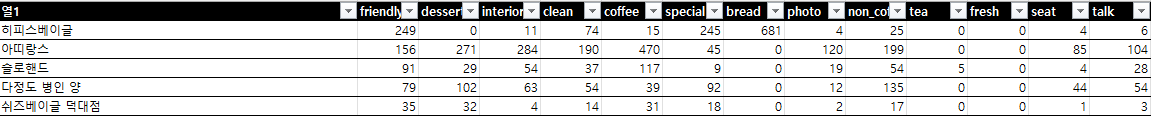
텍스트, 스크린샷, 소프트웨어, 멀티미디어 소프트웨어이(가) 표시된 사진

자동 생성된 설명



‘서울시 지하철역 + 카페’ 라는 쿼리로 크롤링을 진행하여 총 2만여개의 카페 데이터와 103개의 feature를 수집 완료했다. 그 후 데이터들을 하나의 파일로 합치고 전처리를 진행했다.

1. 중복제거 : 역간 거리가 짧은 경우 중복되는 카페들이 있다.
2. 이상치 제거 : 방탈출카페, 팝업스토어 등 목적 데이터와 다른 데이터들을 제거했다.
3. Relabeling : 성능을 높이기 위해 지표를 통합/폐기하여 relabeling을 진행했다.



1. 스케일링 : 추후 모델링을 위해서 모든 데이터를 최대값으로 나눠서 0~1 사이 값을 갖도록 했다. (정규화)

텍스트, 스크린샷, 문서, 영수증이(가) 표시된 사진

자동 생성된 설명

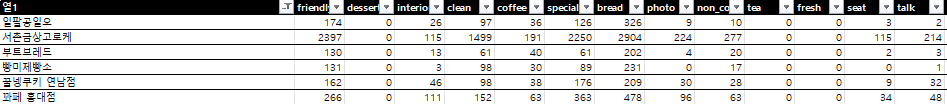
추천을 위해 알고리즘을 선택해야 했다. 유저 피드백이 없는 Cold start 상태이기에 content based filtering을 바탕으로 하기로 했고 LightFM이라는 모델을 사용하여 cold start 문제를 그나마 해결하려고 했다.

**[평가]텍스트, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명**

유저가 긍정적으로 평가한 카페 명을 입력값으로 받으면, 해당 카페와 비슷한 5개의 카페들을 추천해주도록 설계했다.

왼쪽은 예시로 ‘서촌금상고로케’와 비슷한 feature들을 가진 카페들을 추천해주었고 아래는 해당 카페들의 feature로 비슷한 값을 갖고 있음을 알 수 있다.

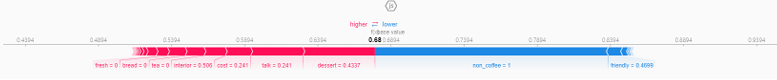


텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

LightFM을 통해 정확도 평가를 진행했을 때 RMSE값이 0.1278로 높은 정확도를 갖고 있음을 알 수 있다.

LightFM 모델을 사용해 예측한 결과에 대해서 해석/평가하기 위해 SHAP를 사용했다. SHAP는 우리가 relabeling한 특성들이 모델의 예측에 얼마나 기여하는지를 나타내고 특성들간의 상관관계를 나타낼 수 있었다.



텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

해당 그래프들은 ‘커피가 맛있다’라는 지표를 기준으로 양/음의 상관관계를 가진 칼럼들을 보여준 것이다. 예를 들면 ‘커피가 맛있다’고 평가한 고객들은 ‘디저트가 맛있다’라는 평가도 같이 남기지만 ‘논커피가 맛있다’라는 평가는 남기지 않는 경향이 있다는 것을 알 수 있었다.