1.주제 의식 및 개요

주제: 네이버 웹툰 평점 예측 프로그램

개요: 네이버 웹툰 서비스에 제공하는 웹툰들의 다양한 특성을 활용하여 해당 웹툰의 평점을 예측하는 데이터 기반 분석입니다. 평점 예측을 통해 웹툰의 인기도와 평가에 영향을 미치는 요인들을 파악하고, 웹툰 제작자와 독자들에게 유용한 정보를 제공하는 것을 목표로 합니다.

2.데이터 수집 방법(데이터 출처 포함):

a)네이버 웹툰 API를 활용하여 웹툰 정도(제목,작가, 장르, 업데이트 날짜, 조회수, 그림 작가, 스토리 작가 등)및 평점 데이터를 수집합니다.

b)웹 크롤링을 통해 웹툰 관련 정보 및 추가적인 특성 데이터를 수집합니다.

c)활용 가능한 외부 데이터 소스가 있다면 해당 데이터를 수집하여 활용합니다.(웹툰 평론 블로그)

#데이터 수집 시 데이터 라이선스 및 저작권을 준수하며, 개인정보 보호에 특별히 신경쓸 것

3.예상 되는 데이터 전처리 방법:

수집한 데이터들은 다양한 형식과 품질을 가지고 있을 것으로 예상 된다.

도전 만화와 같이 작가의 이름이 불문명한 경우 noname값으로 대체하여 이름 없는 값의 작가로 통일 시킬 예정입니다. 사람들이 아직 리뷰를 남기지 않아 별점이 없는 웹툰은 많지 않고 있더라도 유의미한 데이터가 아닐것으로 판단하여 그 행을 삭제할 예정입니다.

데이터를 학습용과 테스트용으로 나누어 모델 성능 평가를 위한 데이터를 준비하려합니다.

4.데이터 분석 방법

다양한 데이터 분석 방법을 활용하여 웹툰 평점 예측 모델을 개발할 예정입니다.

회귀 분석: 특성들과 평점 간의 선형 관계를 분석하여 평점 예측에 활용하여 XGBoost 모델, LightGBM(범주혁 특성을 원-핫 인코딩 필요 x)을 사용하여 예측할 예정입니다. 모델의 성능을 평가하고 최적화를 위해 교차 검증과 하이퍼파라미터 튜닝 등을 수행할 예정입니다.

5.기대 효과 및 예상되는 결과

웹툰 평점 예측 모델 개발로 웹툰 제작자에게 인기 웹툰을 개발하는데 도움을 제공합니다.

독자들에게 개인 취향에 맞는 웹툰을 추천하는데 활용될 수 있습니다.

웹툰 평가에 영향을 미치는 요인들을 파악하여 웹툰 콘텐츠의 질을 향상시키는데 기여합니다.