임베딩 모델을 통한 알고리즘 편향도 평가 서비스 개발

30118 이수민, 30219 윤영서

I. 개발 동기

현재 SNS, 쇼핑몰 등 다양한 서비스에서 사용자의 취향에 맞춰 콘텐츠를 추천해주는 알고리즘이 사용되고 있다. 이는 사용자가 자신에게 필요한 정보를 검색하거나 하는 등의 추가적인 노력 없이도 알맞은 정보를 얻을 수 있도록 도와준다. 하지만 사용자가 관심을 가지고 있는 주제의 콘텐츠를 더욱 자주 노출시키게 되면서 사용자는 좁은 범위의 정보만을 접하게 되는 '필터 버블' 현상이 사회적 문제로 떠오르게 되었다. 이 현상은 사용자의 확증 편향을 강화하여 건전한 사고를 해친다. 이를 해결하기 위해 추천 알고리즘 콘텐츠를 분석하여 편향도를 측정하는 서비스를 개발하고자 한다.

II. 개요

교내 학생 18명에게 설문 조사를 실시한 결과 3명이 필터 버블에 대해 알고있다고 답하였고, 전원 모두 필터 버블을 경험한 적이 있다고 답하였다. 이들 중 대부분이 유튜브에서 필터 버블을 경험하였고 SNS, 쇼핑몰 순으로 경험한 플랫폼이 많았다. 응답 중 필터 버블이 가장 많이 경험된 유튜브의 알고리즘 편향도 평가 서비스를 개발한다.

III. 서비스 작동 원리

본 서비스에서는 임베딩이라는 기술을 활용한다. 임베딩은 인간이 사용하는 자연어를 고차원 공간 (본 서비스에서는 3072차원)에 벡터로 표현하는 것을 말한다.

유튜브 추천 알고리즘을 통해서 콘텐츠의 제목을 불러온다. 제목에서 명사만을 추출한 뒤에 임베딩한다. 추출된 명사를 미리 임베딩 시켜둔 정치, 동물등의 주제와 비교하여 유사도를 측정한다. 비교는 벡터의 각도를 이용한 코사인 유사도를 통해 진행한다. 각 명사와 주제들의 유사도의 평균을 내어서 해당 제목을 가진 영상이 어떤 주제를 다루고 있는지 판단한다. 본 과정을 수차례 반복하여 사용자의 추천 알고리즘이 어떤 주제의 영상을 어떤 비중으로 추천하는지 측정한다. 이때 일정 수준 이상으로 한 주제가 추천된다면 이는 알고리즘이 편향되었다고 간주한다.

IV. 향후 계획

서비스를 사용자들이 사용할 수 있도록 웹페이지로 공개할 계획이다.

콘텐츠의 주제를 잘 나타내는 것은 대체로 제목이겠지만 예외는 존재한다. 유튜브의 경우 제목 뿐 아니라 썸네일, 설명, 태그 등 사용자가 콘텐츠의 주제를 짐작할 수 있는 요소가 존재하므로 이 또한 고려하면 정확도가 증가할 것이다. 따라서 설명과 태그, 제목과 같은 텍스트 형식 정보를 취합하여 판단할 수 있도록 개선하고 이미지 모델을 활용하여 썸네일 또한 다뤄 서비스 개선을 목표로 하고 있다.