회의내용

저번주 회의에서 유튜브 api를 이용하여 인기 동영상 내의 video_id, title, publishedAt, channelId, channelTitle, categoryId, trending_date, tags, view_count, likes, dislikes, comment_count, thumbnail_link, comments_disabled, ratings_disabled, description 정보를 csv파일로 추출하였다. Title과 channelTitle, description 태그에서 한국어 형태소 분석을 하여 많이 세어진 키워드 순서별로 순위를 나열해봤다. 첫번째 문제점으로 하나의 csv 파일로는 충분한 count 수를 나타내지 못해서 유의미한 데이터가 아니라는 생각이 들었다. 그래서 하루에 4번정도 데이터를 추출하여 형태소 분석을 해본 결과 각 키워드 별로 count 수가 차이를 알 수 있는 유의미한 데이터가 되었다. 서버를 할당 받아 3시간별로 데이터를 추출해 주간 키워드를 추출하면 더욱 정확한 데이터가 될 것 같다는 생각이 들었다. 상위 키워드에 대한 가중치를 부여하자는 얘기가 나왔는데 이 부분에 대해서는 우리가 상위 키워드에 임의의 가중치를 부여하기에는 객관성이 부족하다는 얘기가 나와서 이 내용에 대해서는 멘토님의 의견을 여쭤 보기로 하였다. 그리고 키워드를 추출해보니, '몰카', '드라마', '운동' 등 특정 인기 동영상을 유추할 수 없는 '카테고리'성 키워드에 대한 필터링이 필요할 것 같다는 생각이 들었다. 이 부분에 대해서는 다른 형태소 분석 모델을 쓰던가 아니면 필터링 기준에 대해서 멘토님께 여쭤 봐야할 것 같다. 그리고 데이터에서 영어 키워드에 대해 추출하자는 얘기가 있었는데, 한글과 영어 혼합하여 형태소 분석하는 모델을 찾기 힘들고 추출한 데이터 자체가 영어와 한글이 혼합되어 있어서 영어 키워드를 추출하려면 영어 형태소 분석기 모델을 따로 구해서 데이터를 추출해야 한다. 한글 데이터와 영어 데이터를 추출하여 합쳤을 때도 문제가 생기는데, 예를 들어 한글에서 '롤린'과 영어로 'Rollin' 키워드가 중복 된다는 것이다. 이 중복 처리를 하지 못할 시 서비스 측면에서 퀄리티가 떨어져 보인다는 문제점이 있다.

역할분담 & To-Do

중간발표가 얼마 남지 않아서 각자 역할을 분담을 하였다. 일단 '프론트엔드'를 맡은 정용훈은 발표 자료와 프론트엔드 쪽 발표 대본을 만들기로 하고 간단한 웹서비스 화면 구현을 맡기로 하였다. '백엔드'를 맡은 유지석은 발표를 위해 발표 대본을 만들고 이번 데이터 추출을 위한 모델 말고 다른 모델로도 데이터를 추출 하기로 하였다. 그리고 발표 자료를 위해 백엔드 프로세스 흐름도를 만들기로 하였다.

평가 교수님께 보여줄 깃허브 소개 페이지도 구성하고 각자 개발한 작업들을 깃에 공유

하기로 하였다. 그리고 중간 보고서와 발표 동영상 중에 무엇을 할지는 아직 논의 중이다.