**진로탐색학점제 진행 전 세부계획**

**1. 추진일정**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 팀명 : Youtube Follow Support(변경가능)  팀원 : 김정훈, 이연준, 노태원  진행 시간 : 10시 ~15시(점심시간 1시간 포함, 변동가능)  진행 방식 : 대면(기본 비대면) 및 비대면(불가피한 경우) 섞어서 진행 예정  진행 장소 : 한성대학교 내 카페 혹은 상상파크   | **주차** | **활동 목표 및 내용** | **투입시간**  **(1주일기준)** | | --- | --- | --- | | 1주차 | 파이썬 언어로 웹 크롤링을 통한 영상 댓글 수집 모듈 개발 및 문장 벡터화( 벡터화 방법론 탐색 ) | 20시간 | | 2주차 | 벡터화 된 문장을 TF-IDF방식으로 변환하고, SQL DB에 연결하여 URL별 테이블 생성 및 유튜버의 영상 테이블들의 집합을 유튜버로 설정하여 데이터베이스 구조 작성 | 20시간 | | 3주차 | 교수님에게 1차 지도점검 후 방법론 변화 및 중간보고서 작성(7월 12일 중간 보고서 제출) | 20시간 | | 4주차 | 사이킷 런을 사용한 지도학습-회귀 방식 및 텐서플로우를 사용한 인공신경망 방식 중 더 잘 맞는다고 생각하는 방식을 택해 댓글 감정 분석 진행. | 20시간 | | 5주차 | 교수님에게 2차 지도점검 후 댓글 평가 AI 모델을 개선하여 정확도 증진 후 구글 vertex ai를 사용한 멀티 모달리티 영상분석 도전 | 20시간 | | 6주차 | 지지도가 낮고 조회수가 높은 나락 베스트 및 모든 영상의 평균지지도가 높고, 조회수가 높은 홍보효과 베스트 그리고 사용자에게 직접 댓글을 입력받아 영상에 대한 감정을 분석해보는 참여형 댓글분석 기능을 웹페이지에 구현한다. | 20시간 | | 7주차 | 최종 보고서 및 ppt작성 | 20시간 | |

**2. 활동 수행 결과물**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **가. 주요 결과물**  **1. 유튜브 댓글 웹크롤링 모듈(python)**   |  | | --- | | 파이썬으로 모듈 개발 예정이며, youtube 동영상의 url링크를 기반으로 댓글을 리스트로 가져오는 웹 크롤링 모듈을 개발할 것이다.  (※만약, 웹 크롤링이 유튜브 정책상 불가능 하다면, google video intelligence의 OCR기능을 사용하여,  youtube댓글들을 pdf로 캡쳐해서 이 pdf를 pdf-poppler라는 라이브러리와 연계하여, 댓글을 추출할 수 있다.(해당 방법은 사전에 개발완료되어 있음)) |   **2. 문장 벡터화(단어 빈도 기반 벡터화 예상하나, 변동 가능) 모듈(python)**   |  | | --- | | 이렇게 얻은 댓글 리스트를, 단어 단위로 나눈 후, 벡터화 할 것이다. 현재로서는 단어 빈도를 기반으로 한 BOW방식으로 벡터화를 진행할 예정이나, 여러 방법론들을 찾는 과정을 거치고(보고서로 작성예정) 이 중 더 잘 맞는 방식이 있다면 그 방식을 채택할 예정이다. |   **3. TF-IDF 전처리 모듈(다른 방법론 변경가능, python)**   |  | | --- | | 위 BOW방식과 연계되어 있기 때문에 다른 방식을 선택한다면 변경될 수 있다.  인공지능이 댓글을 평가하기 위한 data로 단어빈도를 기반으로 벡터화 한다면, 이는 TF-IDF방식으로 중요도를 파악할 수 있다. 물론 문장을 평가하기 위해서는 단어의 순서 등 고려해야 할 요소가 많기에 방법론이 변화할 수 있다. |   **4. SQL DB**   |  | | --- | | 전처리가 완료된(벡터화 후 TF-IDF까지 완료) 데이터들은 DB에 들어갈 것이며, url을 이름으로 한, table을 생성할 것이다. 이러한 table들이 속하는 집합이 Youtuber가 될 것이며 Mysql을 사용할 예정이다. |   **5. 문장 벡터를 기반으로 댓글 평가하는 AI model**   |  | | --- | | 전처리 된 데이터를 기반으로 댓글을 평가하는 AI모델을 파이썬으로 구성할 것인데,  댓글을 분석해서 얻는 것을 0~1사이의 연속된 실수값으로 생각 중이며, 지도학습-회귀 방식으로 하려고한다.  데이터 마이닝 강의에서 배운 사이킷런을 사용하여 하고싶지만, 다른 방법론도 찾아보고 결정하려고 한다. |   **5. 웹 페이지**   |  | | --- | | 1. 지지도가 낮지만, 조회수가 높다면 어떠한 유튜버가 잘못을 한 영상일 가능성이 높다.  이러한 영상을 사람들은 좋아하며 지지도가 낮고 조회수가 높은 경우를 나락수치라고 정리하여,  나락수치가 높은 list를 보여줄 수 있다.  2. 모든영상의 지지도가 평균적으로 높고 조회수 또한 높다면 해당 유튜버는 맹목적으로 유튜버를 좋아하는 팬층이 형성되어있을 가능성이 높다. 이러한 유튜버들은 불가침영역이고, 홍보효과가 뒤어날 가능성이 높기에 홍보효과 best list를 만들어 보여줄 수 있다.  3. 한 영상을 보여주고, 그 영상을 시청한 사람( test사용자 ) 에게 댓글을 입력받아, 방금쓴 댓글로 사용자가 영상에 대해 어떤 감정을(호,불호 및 얼마나 좋아하는지 혹은 얼마나 싫었는지) 파악해서 답해주는 기능을 본인이 직접 참여해볼 수 있도록 한다. 이 프로젝트의 완성도를 보여줄 수 있다. |   **6(Challenge). 구글 vertex ai api를 사용한 멀티 모달리티 영상분석**   |  | | --- | | 구글 vertex api(변경가능)를 통해 멀티 모달리티 영상분석을 진행하고, 영상의 조회수가 높은 이유와 낮은 이유등을 분석해, 어떠한 영상을 올리면 사람들이 좋아할지에 대해서 list를 구성한다. | |

**3. 지도 점검 계획**

|  |
| --- |
| 7월 12일에 중간보고서 제출이 있기 때문에,  1주차 및 2주차에 웹크롤링을 통한 데이터 마이닝 및 단어 벡터화까지 진행한 후,  2주차의 마지막인 7월 5일(교수님 일정에 맞춰 무조건 변경가능합니다.) 교수님에게  단어 벡터화 방법론 선택에 대해서 여러 벡터화 방법을 제시한 후 어떠한 방법론이 좋을지 상담 받고싶습니다.  이후, 4주차의 마지막인 7월 19일(교수님 일정에 맞춰 무조건 변경가능합니다.)에 교수님에게  문장벡터를 기반으로 댓글을 평가하는 ai 모델을 구성할 때, 인공신경망을 통한 텐서플로우를 사용할지,  사이킷 런을 사용한 회귀분석을 사용할지 혹은 다른 방법론을 찾아본 후 ai 모델 구성 방법론에 대한 상담을 받고싶습니다.  지금은 이렇게 2회 지도 점검 계획을 가지고 있습니다. |