

## 프로젝트명: <유튜브 댓글 긍정/부정 분석과 워드 클라우드 표현>

### 1. 프로젝트 배경 (주제선정에 대한 배경)

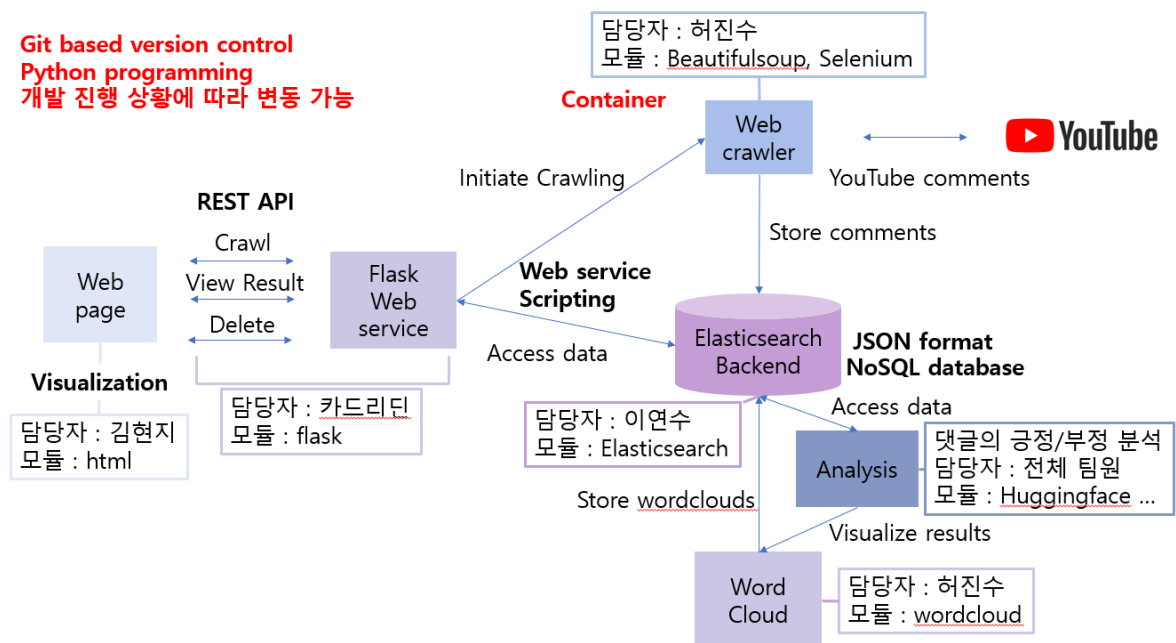
User가 유튜브에서 볼 영상에 대해 미리 다른 시청자들의 의견이 그 영상에 대해 평균적으로 부정적인지 긍정적인지 알 수 있도록 도와주는 웹 페이지를 만들고자 한다. 이를 통해 해당 영상이 불만한 영상인지 아닌지를 알 수 있다. 또한 유튜브에서 내용이 심한 영상들은 18+ 허용이 있어야 볼 수 있지만 내용이 심한 댓글들은 그런 제한이 없다. 따라서 유튜브 영상의 댓글을 분석해서 댓글의 심각성을 나타낼 수 있는 웹 페이지가 필요하다고 판단하였다.

뿐만 아니라 요즘은 유튜브에서 영상을 업로드해서 수익을 창출하는 유튜버들도 많이 증가하고 있는 추세이다. 유튜버들에게 수익을 얻기 위해 영상 조회 수를 높이는 것이 매우 중요하다. 그리고 영상 조회 수를 높이기 위해 유튜버가 시청자들의 요구 사항 및 관심도를 파악하는 것은 필수적이다. 보통 유튜버들은 댓글들을 통해서 영상에 대한 반응을 확인하여 유튜브 이용자들의 관심도를 파악한다. 하지만 댓글 수가 200~300개 이상이 되면 댓글들을 하나 하나씩 읽어서 분석하는 것이 매우 어렵다. 이러한 문제를 풀기 위해 유튜브 영상의 댓글들을 분석해 영상에 대한 반응, 관심도 등을 워드 클라우드를 통해서 한 눈에 볼 수 있도록 하는 서버를 만들고자 한다.

유튜브 댓글을 크롤링하고 워드 frequency을 통해서 워드 클라우드로 표현하고 댓글들의 각 문장을 분석해서 해당 영상에 대한 부정적 인식이 많은 지, 긍정적 인식이 많은 지 결과적으로 보여주는 것을 목표로 한다.

### 2. 시스템 구조

- Git based version control
- Python programming
- 개발 진행 상황에 따라 변동 가능



김현지

- HTML Page 디자인 총괄/제작: 회의를 통해 page 디자인을 구체화하고, html과 css를 이용해 사용자에게 보여지는 web page를 제작하는 역할을 맡았다.

이연수

- 크롤링한 데이터 관리: 댓글 데이터를 크롤링한 후에 Elasticsearch를 이용하여 데이터를 저장하고 관리하는 역할을 맡았다.

카드리딘

- Flask를 이용하여 Rest API 생성: HTML Page 제작이 완료되면 Flask에 연결하여 REST API를 생성하는 역할을 맡았다

허진수

- 유튜브 영상 댓글 크롤링: 댓글 분석에 필요한 데이터를 유튜브를 통해 BeautifulSoup과 Selenium을 이용하여 지속적으로 모으는 역할을 맡았다
- 워드 클라우드 제작: 댓글 데이터와 분석 방법을 이용하여 추출된 결과를 워드 클라우드로 나타내는 역할을 맡았다.

(팀원 전체) 김현지, 이연수, 카드리딘, 허진수

- 댓글의 긍정/부정 분석: Huggingface와 같은 자연어 처리 알고리즘이 있는 모듈을 이용하여 댓글이 긍정적인 반응인지, 부정적인 반응인지 판단하는 역할을 맡았다. 모듈은 좀 더 정밀한 조사 후에 추가될 예정이다.

### 3. 시스템 동작

① 사용자에게 보여지는 **Web page**에서 사용자로부터 youtube 영상에 대한 url을 입력받으면, **Web crawler**로 해당 youtube url 내의 영상, 제목, 좋아요/싫어요 수, 댓글 정보 등을 crawling해 온다.

② 1번에서 crawling한 정보를 url ID, 영상, 제목, 좋아요/싫어요 수, 댓글 정보 등의 데이터 묶음으로 **Elasticsearch**를 통해 저장한다.

③ Elasticsearch의 설정 파일에서 IP와 포트설정 부분을 재설정하여 외부에서도 접속이 가능하게 하여, 함께 협업하는 작업자들이 모두 데이터에 접근하고 사용하는 것이 가능하도록 한다.

④ **Elasticsearch**를 통해 저장된 데이터에서 영상, 제목, 좋아요/싫어요 수는 **Flask Web service**를 기반으로 사용자에게 보여지며, 댓글 정보는 **sentence analysis**를 통해 영상에 대한 댓글의 긍정/부정 반응을 판단한다.

⑤ 또한, youtube 영상 댓글에 대한 정보를 **Word Cloud**로 표현하여, 사용자가 해당 영상에 대한 일반 사용자들의 반응을 시각적으로 한 눈에 확인할 수 있도록 돕는다.

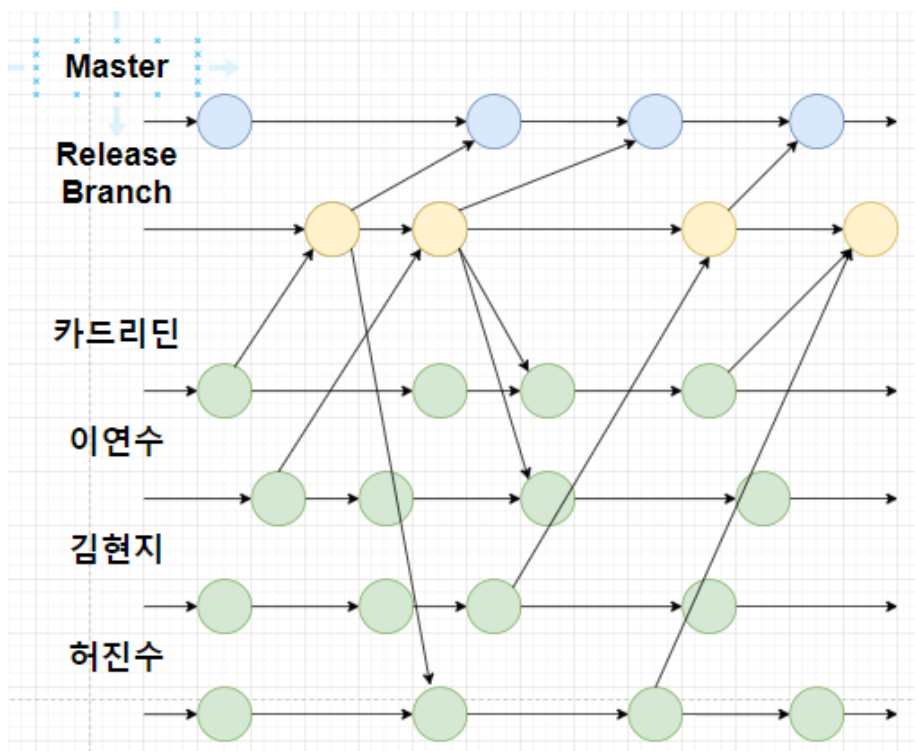
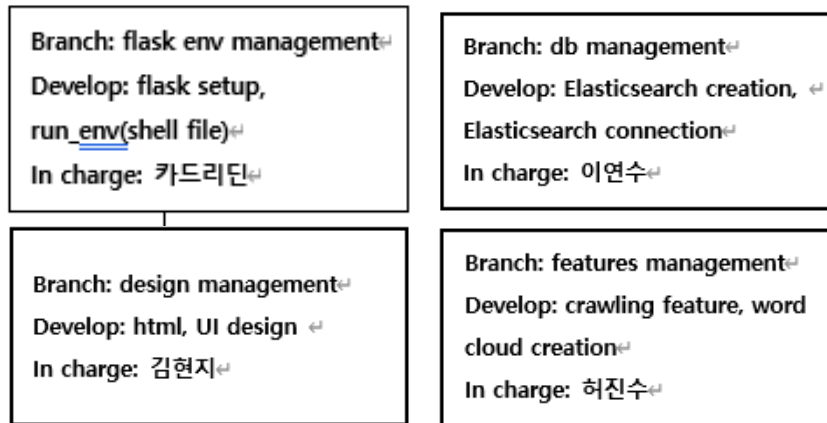
⑥ 3번과 4번에서 분석한 댓글의 긍정/부정 판단 결과와 Word Cloud는 **Elasticsearch**를 통해 저장되고, 이후 **Flask Web service**를 기반으로 사용자가 확인할 수 있는 **Web page**에 보여준다.

#### 4. 협업 환경

이 프로젝트는 Github를 통해서 팀원들과 같이 협업한다. 각 멤버는 자기의 branch에서 개발한다. 구현한 다음에 unit testing을 해보고 문제없는 경우에 pull request로 합친다.

- GitHub URL: [https://github.com/Qad00/OSP\\_Final\\_Pro](https://github.com/Qad00/OSP_Final_Pro)

- Branch 작업 환경 계획



## 5. 전체 일정

	1주차 (05.23 ~ 29)							2주차 (05.30 ~ 06.05)							3주차 (06.06 ~ 12)							4주차 (06.13 ~ 19)						
	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일	월	화	수	목	금	토	일
카드리딘																												
개발 환경 Setting																												
Rest API 생성																												
알고리즘, 모델 자료 수집																												
김현지																												
Page 디자인																												
Page 제작																												
허진수																												
영상 댓글 크롤링																												
워드 클라우드 제작																												
이연수																												
크롤링한 데이터 관리																												
카드리딘, 허진수, 김현지, 이연수																												
댓글의 긍정/부정 분석																												

### 구체적인 일정

※ 모든 Task에 대한 일정은 개발 기간 중에 변동이 있을 수 있음.

① 개발 환경 설정: 개발 환경을 먼저 설정해야 개발을 진행할 수 있기에 기간을 5월 23일부터 29일까지로 정하고 나중에 필요한 library가 생길 시 일정을 조율할 예정. 담당자는 카드리딘이 맡음.

② Flask를 이용하여 Rest API 생성: HTML Page 제작이 완료되면 Flask에 연결을 바로바로 해야 하기 때문에 기간을 5월 25일부터 6월 5일까지로 정함. 담당자는 카드리딘이 맡음.

③ 댓글의 긍정/부정 분석을 위한 알고리즘 및 모델 자료 수집: 개발 환경과 Rest API를 어느 정도 설정했을 때 수집할 여유가 있을 것 같아서 기간을 5월 28일부터 6월 12일까지로 정함. 담당자는 카드리딘이 맡음.

④ HTML Page 디자인: 팀원과의 회의를 통해 디자인이 계속 변경될 것이기 때문에 기간을 5월 23일부터 6월 5일까지로 정함. 담당자는 김현지가 맡음.

⑤ HTML Page 제작: Page 디자인이 어느 정도 되었을 때 만들 수 있을 것 같아서 기간을 5월 25일부터 6월 5일까지로 정함. 담당자는 김현지가 맡음.

⑥ 유튜브 영상 댓글 크롤링: 댓글 분석에 필요한 데이터를 지속적으로 모아야 하기에 기간을 5월 23일부터 6월 5일까지로 정함. 담당자는 허진수가 맡음.

⑦ 워드 클라우드 제작: 댓글 데이터가 어느 정도 모이고 분석 방법이 정해졌을 때 제작을 할 수 있기에 기간을 마지막 주(6월 13일부터 6월 19일까지)로 정함. 담당자는 허진수가 맡음.

⑧ 크롤링한 데이터 관리: 댓글 데이터를 크롤링한 후에 데이터를 저장하고 관리할 수 있기 때문에 기간을 5월 26일부터 6월 5일까지로 정함. 담당자는 이연수가 맡음.

⑨ 댓글의 긍정/부정 분석: 댓글 분석 방법이 어느 정도 모였을 때 시작하는게 좋고 분석이 오래 걸릴 것 같아서 기간을 5월 30일부터 6월 12일까지로 정함. 담당자는 카드리딘, 김현지, 허진수, 이연수가 맡음.