 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

캡스톤 디자인 I

종합설계 프로젝트

프로젝트 명	요바(요기 바바)
팀 명	안진마
문서 제목	중간보고서

Version	1.3
Date	2020-04-23


팀원	신 상훈 (조장)
	김 연수
	송 성재
	박 형준
	허 진선
	윤 정연
지도교수	강 승식 교수

CONFIDENTIALITY/SECURITY WARNING


이 문서에 포함되어 있는 정보는 국민대학교 전자정보통신대학 컴퓨터공학부 및 컴퓨터공학부 개설 교과목 캡스톤 디자인 I 수강 학생 중 프로젝트 “요바(요기바바)”를 수행하는 팀 “안진마”의 팀원들의 자산입니다. 국민대학교 컴퓨터공학부 및 팀 “안진마”의 팀원들의 서면 허락없이 사용되거나, 재가공 될 수 없습니다.

문서 정보 / 수정 내역

Filename	중간보고서-요바(요기바바).doc
원안작성자	신상훈
수정작업자	신상훈, 송성재


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

수정날짜	대표수정자	Revision	추가/수정 항목	내 용
2020-04-17	신상훈	1.0	최초 작성	프로젝트 목표 작성
2020-04-19	송성재	1.1	내용 수정	수행 내용 및 중간결과 작성
2020-04-21	신상훈	1.2	내용 수정	수정된 연구내용 및 추진 방향 작성
2020-04-23	송성재	1.3	내용 수정	향후 추진계획, 고충 및 건의사항 작성
2020-04-23	신상훈	1.3	최종 검토	

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

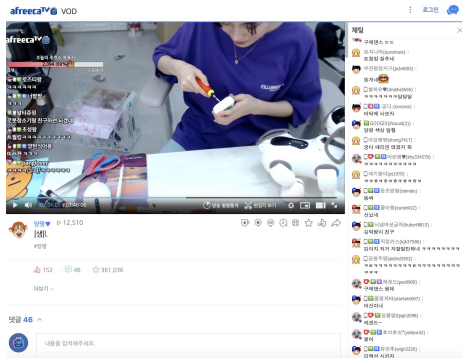
목 차

프로젝트 목표	4
수행 내용 및 중간결과	6
계획서 상의 연구내용	6
수행내용	9
수정된 연구내용 및 추진 방향	20
수정사항	20
향후 추진계획	21
향후 계획의 세부 내용	21
고충 및 건의사항	22

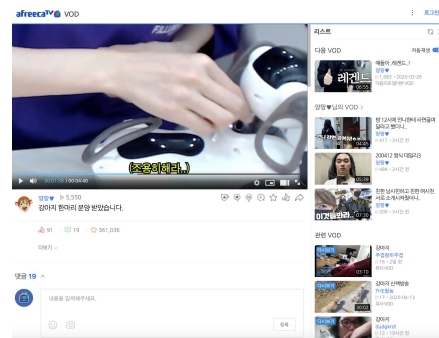
 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

1 프로젝트 목표

본 프로젝트는 인터넷 방송 크리에이터가 장시간의 방송을 마친 후 업로드 된 다시보기 영상을 분석하여 하이라이트 추천지점을 제공한다. 본 프로젝트에서 추출된 하이라이트 지점을 통해 편집자는 하이라이트 위치를 쉽게 찾을 수 있고 이를 토대로 편집시간을 줄일 수 있다.



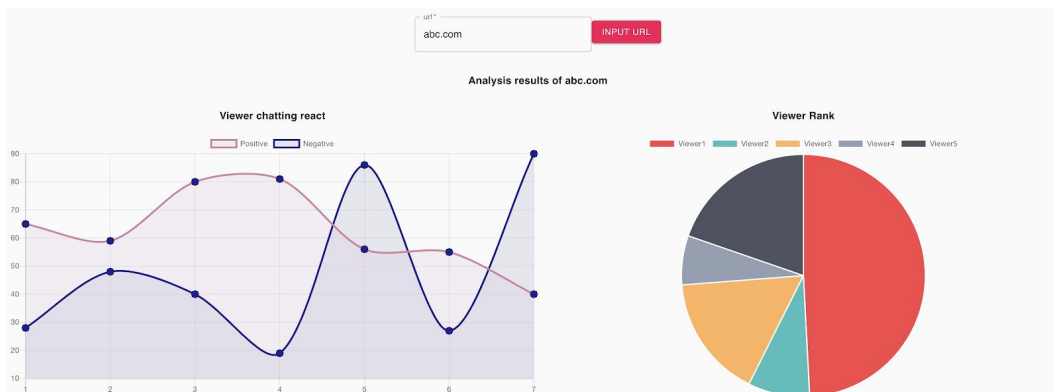
[그림 1-1] 다시보기 영상 (3시간 동영상, 채팅 로그)



[그림 1-2] 하이라이트 편집 영상 (4분 동영상)


Highlight	Start Point	End Point	chat or audio
Highlight 1	07:00	08:00	chat
Highlight 2	19:00	20:00	audio
Highlight 3	32:00	33:00	chat
Highlight 4	52:00	53:00	audio

[표 1-1] 하이라이트 추천 지점 예시



[그림 1-3] 다시보기 영상 데이터 지표화 예시

즉, 1인 크리에이터 인터넷 방송 편집자를 돕는 툴을 만드는 것이다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

주요 기능은 다음과 같다.


- 하이라이트 추출
- 시청자의 반응 긍정/부정 분류 및 가시화
- 시청자별 채팅 참여도 가시화
- 채팅에서 등장 빈도가 높은 단어 가시화
- 선택한 단어가 나오는 채팅 등장 시기 및 빈도 가시화
- 동영상 음성 평준화

이를 통해서 편집자가 영상 편집에 있어서 하이라이트를 선별 하는데 도움을 주는 것이다. 또한 오디오 파일의 볼륨 크기를 평준화하여 이를 다운로드 할 수 있게 제공한다.

시장조사 결과 현재 인터넷 실시간 방송 국내 시청자 선호 플랫폼 중 상위 3개에 랭크된 플랫폼은 유튜브, 아프리카TV, 트위치인 것으로 확인된다.[그림 1-4] 이에 따라 본 프로젝트는 해당 3사를 타겟으로 한 서비스로 기획했다.



[그림 1-4] 국내 시청자 선호 인터넷 방송 플랫폼, 나스미디어

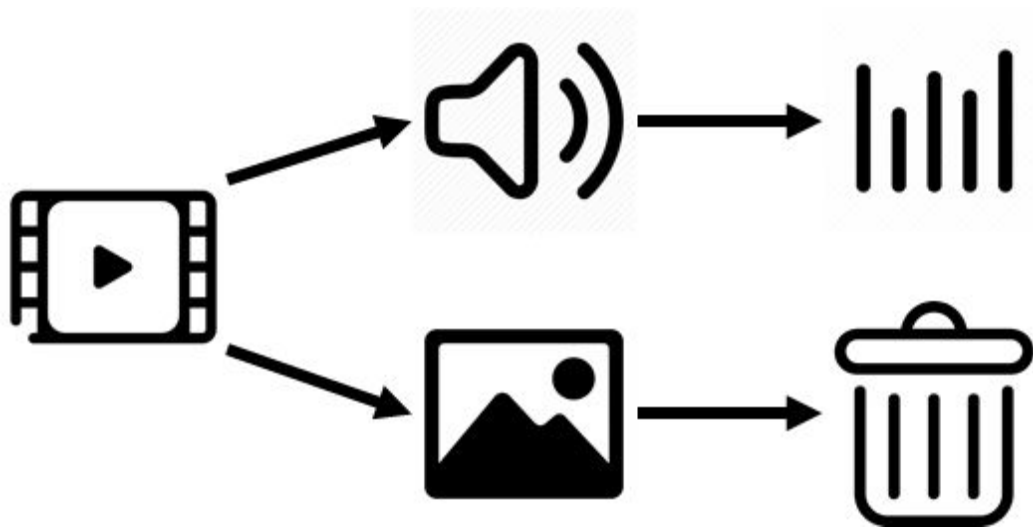
 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2 수행 내용 및 중간결과


2.1 계획서 상의 연구내용

1. 음성파일 분류 모듈

1인방송의 특성상 자신을 촬영해주는 사람이 따로 없기 때문에 카메라가 고정된 위치에 있을 수 밖에 없다. 그렇기 때문에 크리에이터는 한정된 공간밖에 사용하지 못하기 때문에 신체적인 움직임 보다는 목소리를 이용해 자신의 감정을 나타내고 시청자들과 공감하며 활동성을 나타내기 위해 집중한다. 따라서 크리에이터의 활동성을 분석하는데 오디오 파일을 사용하기로 하였으며 동영상 파일은 크기가 매우 크고 불필요 하기 때문에 걸러내는 작업이 필요하다. 추출된 오디오 파일은 스펙트럼을 분석하여 하이라이트 추천 알고리즘에 사용할 수 있도록 수치화한다.

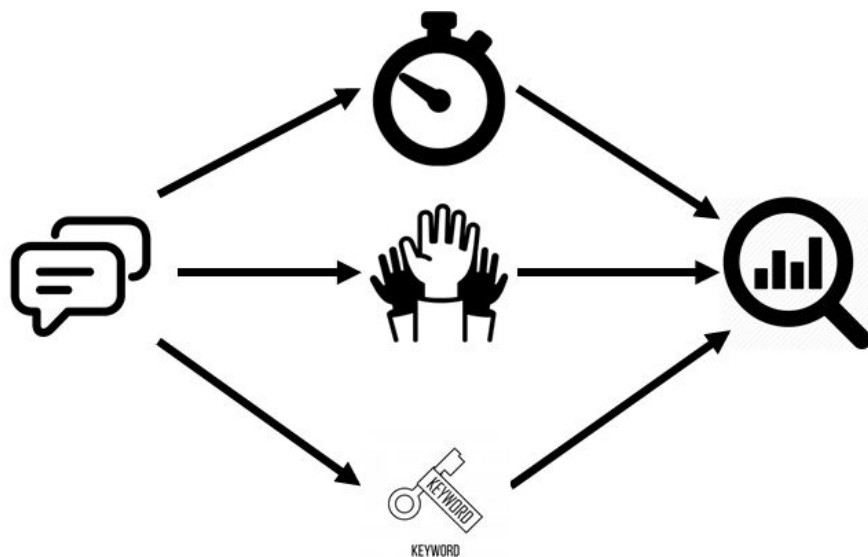


[그림 2-1] 음성파일 분류 모듈

 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2. 채팅 로그 분석 모듈

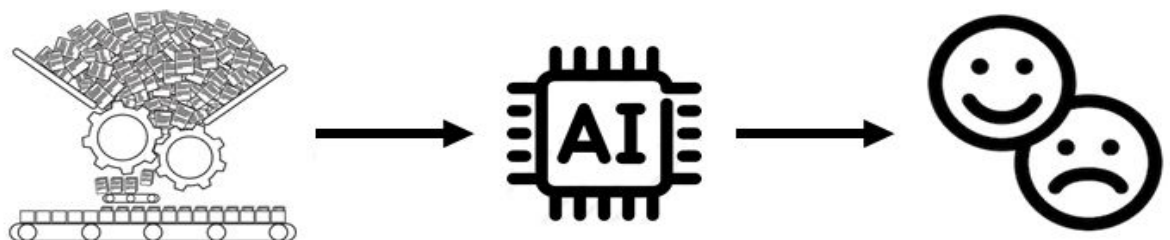
크리에이터는 실시간 스트리밍을 통해 시청자와 즉각 소통하며 영상을 제작할 수 있다. 시청자의 채팅을 읽고 수행하는 콘텐츠도 있을 정도로 크리에이터들은 시청자들의 참여를 꾸준히 유도한다. 즉 시청자는 이제 제작된 영상을 소비하기만 하는것이 아니라 영상의 제작에도 참여하는 소비자이자 제작참여자 가 된 것이다. 때문에 크리에이터에게 시청자의 반응을 체크하는 것은 선택이 아닌 필수가 됐다. 그래서 채팅로그를 추출하여 채팅 시간을 이용하여 시간대별 채팅 수를 분석하고, 채팅 아이디를 이용하여 채팅에 참여한 시청자 수를 분석한다. 또한 형태소 분석을 통해 자주 출현하는 명사를 키워드로 추출 하고 이 분석 내용들을 수치화한다.




[그림 2-2] 채팅 로그 분석 모듈

3. 긍부정 감정 분석 모듈

시청자들의 반응에는 절대적인 채팅량많이 중요한 것이 아니다. 만약 특정 구간에서 채팅량이 폭발적으로 증가했지만 대다수가 부정적인 댓글이라면 이 구간은 하이라이트 지점으로 선정해서는 안 될 것이다. 따라서 인공지능을 통해 채팅 내용을 감정 분석을 해 줄 필요가 있다. 이를 위해 인공지능 모델을 구축 하고, 채팅 로그를 전처리하여 라벨링 하여 모델을 학습 시켜 채팅을 감정 분석 한 뒤 수치화 한다. 긍부정 모델이 성공적으로 아웃풋을 뽑아 낸다면 차후 긍부정 뿐 아니라 더 다양한 감정을 분석 할 수 있도록 기능을 더할 예정이다.



[그림 2-3] 긍부정 감정 분석 모듈

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

4. 사용자가 서비스를 쉽게 이용할 수 있도록 웹 페이지를 구성

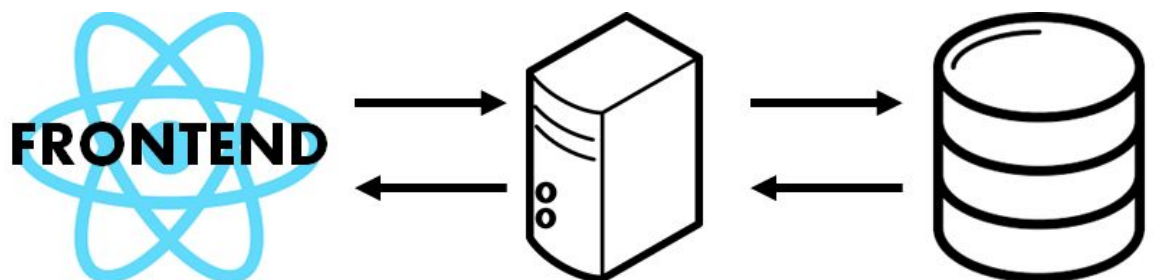
우리의 서비스는 편리함을 추구한다. 따라서 프로그램 설치 없이 간편하게 사용할 수 있도록 웹으로 제작한다. 또한 메뉴얼을 읽지 않고도 자연스럽게 사용할 수 있도록 UI / UX 디자인을 적용해 사용성을 높이려고 한다. 분석이 종료된 이후에는 수학적 지식이 많지 않더라도 쉽게 읽을 수 있는 그래프 종류를 이용하여 가시화하여 보여줄 것이다.




[그림 2-4] 웹 페이지

5. 서비스를 제공할 수 있는 서버를 구성

웹 페이지에서 정적인 화면을 보여줄 것이 아니기 때문에 서버는 필수적으로 구성하게 된다. 프레임워크는 플라스크를 사용할 것이며 프론트 서버와 연동을 위해 CORS를 적용할 것이다. 또한 웹 페이지가 구성이 되어 있지 않아도 서버를 개발 할 수 있도록 테스트 코드를 작성하며 API를 개발하려고 한다. 관계형 데이터베이스인 Postgresql을 사용하고 ORM은 SQLAlchemy를 사용할 것이며 DB 스키마는 alembic을 이용하여 버전관리를 할 것이다.



[그림 2-5] API 연결 구성

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23


2.2 수행내용

2.2.1 다시보기 URL을 통해 오디오와 채팅로그 다운

- Youtube, Twitch, AfreecaTV 3사에서 제공하는 스트리밍 영상 다시보기 서비스의 URL을 입력받고 정규식을 이용하여 입력된 URL이 어떤 플랫폼에 업로드된 영상인지 구분하여 해당 플랫폼에서 데이터 다운
- 아프리카TV 채팅이 전부 다 읽히지 않는 문제 분석과 해결
 - 아프리카TV에서는 각 동영상마다 rowKey라는 고유의 값이 존재하는데 이 값을 아프리카의 채팅로그가 저장된 사이트(<http://videoimg.afreecatv.com/php/ChatLoad.php>) 의 URL과 함께 requests get 메소드의 params 값으로 넣어주면 채팅로그를 추출
 - 일부 아프리카TV의 동영상의 채팅 로그를 추출할때 채팅이 일부분만 받아지는 문제가 발생했는데 이는 동영상에 존재하는 rowKey가 1개가 아닌 1개 이상이 존재하기 때문임을 발견
 - 이를 해결하기위해 아프리카의 동영상 정보를 제공하는 API에서 동영상에 존재하는 rowKey를 모두 찾아 이를 params 값으로 넣어주어 일부 동영상에서 채팅이 일부분만 받아지는 문제를 해결

2.2.2 다시보기 영상과 하이라이트 영상을 매칭시키는 데이터셋 구축 모색

- 하이라이트를 분류하기 위해 스트리머의 행동 변화량(영상), 영상의 볼륨 크기(사운드), 채팅 로그 데이터를 피쳐로 하여 인공지능을 설계하려고 했었음. 인공지능에 영상 데이터를 사용하기 위해서는 풀 영상에서 어떤 지점이 하이라이트인지 라벨링 된 데이터가 필요함. 조사 결과, 제공되는 데이터셋이 존재하지 않아 풀 영상과 편집 영상의 프레임을 매칭하여 직접 데이터셋을 구축하고자 함
- 편집된 영상에 특수 효과나 자막등이 삽입되어 프레임 비교의 정확도가 떨어질 뿐더러 프레임 간의 매칭이 연산량이 매우 많아 시간 내 데이터셋 구축이 불가능하다고 판단함
- 휴리스틱한 방법(채팅의 수, 영상의 볼륨 크기, 스트리머의 행동 변화량)에 의존하고 있는데 이에 대한 타당성을 검증하는 과정에서 채팅 수와 볼륨만으로도 충분히 하이라이트 탐지가 가능하다고 판단하여 영상데이터를 활용하지 않기로 함


 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.3 하이라이트를 추출하는 휴리스틱한 방법에 대한 검증

- 다시보기 영상(2시간 이상)과 하이라이트 영상(10분 가량) 데이터를 60쌍 탐색 및 수집
- 영상의 소리가 커지는 지점, 채팅량이 많아지는 지점을 분석하여 하이라이트 영상에 포함 여부 탐색
- 휴리스틱의 타당성 검증 과정
 - a. 데이터(풀영상의 오디오, 채팅로그) 다운로드
 - b. 오디오 파일에서 분당 평균 볼륨크기를 계산해 평균 볼륨이 가장 큰 3, 5, 10개 지점을 추출
 - c. 채팅 로그 파일에서 초당 채팅 수를 계산해 채팅 수가 가장 많은 3, 5, 10개 지점을 추출
 - d. 풀 영상에서 추출된 6, 10, 20개의 지점과 하이라이트 영상을 비교하여 추출된 지점이 하이라이트 영상에 존재하는지 확인
- 휴리스틱 검증에 사용한 데이터는 아래와 같다. 전체데이터: [데이터소스](#)

1	번호	URL	영상 제목	되는 스트림 또는 게시 일	성별	게임, 토크, 먹:분:초	플랫폼	형태(풀, 편집)	
2	1	https://www.youtube.com/watch?v=...	송대익 생방송 풀버전 [20.03 송대익(송	2020.03.24	남	토크	1:09:23	유튜브	풀
3		http://vod.afreecatv.com/PL...	장모님 안산에 오셨습니.. 송대익	2020.03.24	남	토크	1:09:31	아프리카	풀
4		https://www.youtube.com/watch?v=...	장모님 사할합니다. 송대익(sc	2020.03.27	남	토크	0:07:13	유튜브	편집
5	2	http://vod.afreecatv.com/PL...	[생] 댓글로 웃겨라! 웃기면 1BJ★일다	2020.03.25	남	토크	0:41:46	아프리카	풀
6		https://www.youtube.com/watch?v=...	[하이라이트] 최초의 컨텐츠! 일다TV	2020.03.26	남	토크	0:11:13	유튜브	편집
7	3	http://vod.afreecatv.com/PL...	화이트 데이 기념 철구오빠 : 외질혜	2020.03.14	여	토크? 요리?	1:09:37	아프리카	풀
8		https://www.youtube.com/watch?v=...	철구오빠한테 크림파스타 하 외질혜 (C	2020.03.16	여	토크? 요리?	0:07:00	유튜브	편집
9	4	http://vod.afreecatv.com/PL...	육박 조강력 자석 자동차 끝! 최고다육	2019.12.18	남	토크, 야방	1:26:11	아프리카	풀
10		youtube.com/watch?v=...	[절대] 안 떨어지는 조강력 자 최고다육	2019.12.25	남	토크, 야방	0:06:43	유튜브	편집
11	5	https://www.twitch.tv/videos/...	엠비션) 5번째승급전 엠비션_	2020.02.24	남	게임	7:26:49	트위치	풀
12		https://www.youtube.com/watch?v=...	마스터에 너무너무너무 가고 엠비션 유	2020.02.29	남	게임	0:12:10	유튜브	편집
13	6	https://youtu.be/VUzT3OEP...	대도서관 생방송! 동물의 숲! 대도서관	2020.03.20	남	게임	5:22:31	유튜브	풀
14		https://youtu.be/GDRPKMm...	나 대사로이... 무인도에서 저 대도서관	2020.03.24	남	게임	0:10:55	유튜브	편집
15		https://youtu.be/R_QUmptq...	그동안 시청해주셔서 감사합 대서관	2020.03.27	남	게임	0:09:06	유튜브	편집
16	7	http://vod.afreecatv.com/PL...	[생]킹기훈 가현이랑 400번 킹기훈	2020.03.06	남	토크, 요리	1:26:02	아프리카	풀
17		https://youtu.be/sn8GL73R...	400번 저어만드는 분노의 달 사나이 킹	2020.03.10	남	토크, 요리	0:05:57	유튜브	편집
18	8	http://vod.afreecatv.com/PL...	[생]. 양팡♥	2020.03.21	여	토크	3:46:06	아프리카	풀
19		http://vod.afreecatv.com/PL...	강아지 한마리 분양 받았습니 양팡♥	2020.03.26	여	토크	0:04:40	아프리카	편집
20	9	http://vod.afreecatv.com/PL...	[생]개박이 엄기적인 할랄님 개박이개	2020.03.08	남	토크	2:16:31	아프리카	풀
21		https://youtu.be/3c2sBU07g...	서로 얼굴 칠해주고 확인하기 개박이	2020.03.10	남	토크	0:10:00	유튜브	편집
22		https://youtu.be/H3lBfrllj0...	방송 시작하자마자 서러워서 개박이	2020.03.13	남	토크	0:10:09	유튜브	편집
23	10	http://vod.afreecatv.com/PL...	[생]집순이수인이♥갓심들이 수인이♥	2020.03.19	여	토크	4:00:39	아프리카	풀
24		https://youtu.be/Fwx3vl-Ww7...	누구나 너 멋여아쓰기 챌린수이이네!	2020.03.29	여	토크	0:05:07	유튜브	편집

[그림 2-6] 데이터 소스

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

- 결과: 알고리즘이 하이라이트라고 판단한 지점이 실제로 하이라이트일 확률(True Positive)은 87.2%이며, 하이라이트라고 판단했지만 실제로 하이라이트가 아닐 확률(False Positive)은 12.8%로 볼륨크기와 초당 채팅 수를 3개씩 추출하는 것이 준수한 분류 성능을 보인다.

	3개 추출	5개 추출	10개 추출
Actual/Predict	True	True	True
True	273(75.8%)	367(61.2%)	456(38.0%)
False	87(24.2%)	233(38.8%)	744(62.0%)

[표 2-1] 휴리스틱 검증 결과

2.2.4 시청자 채팅 긍정 부정 분류


- keras의 MLP모델을 사용했으며 네이버의 영화리뷰 데이터셋을 사용하여 모델을 학습
- 데이터셋을 Okt 토큰라이저를 이용하여 토큰화 시킨 후 BOW 기법을 이용하여 벡터화
- 벡터화된 데이터를 이용해 모델 학습
- 토큰라이저와 분석 모델을 바꿔 가며 성능 측정 실험 연구 진행

```

Train on 150000 samples
Epoch 1/10
2020-04-20 22:44:01.192650: I tensorflow/stream_executor/platform/default/dso_loader.cc:44] Successfully opened dynamic library cublas64_100.dll
150000/150000 [=====] - 6s 40us/sample - loss: 0.3937 - binary_accuracy: 0.8289
Epoch 2/10
150000/150000 [=====] - 2s 16us/sample - loss: 0.3277 - binary_accuracy: 0.8585
Epoch 3/10
150000/150000 [=====] - 2s 13us/sample - loss: 0.3068 - binary_accuracy: 0.8694
Epoch 4/10
150000/150000 [=====] - 2s 14us/sample - loss: 0.2903 - binary_accuracy: 0.8794
Epoch 5/10
150000/150000 [=====] - 2s 14us/sample - loss: 0.2746 - binary_accuracy: 0.8873
Epoch 6/10
150000/150000 [=====] - 2s 14us/sample - loss: 0.2577 - binary_accuracy: 0.8953
Epoch 7/10
150000/150000 [=====] - 2s 15us/sample - loss: 0.2407 - binary_accuracy: 0.9038
Epoch 8/10
150000/150000 [=====] - 2s 14us/sample - loss: 0.2241 - binary_accuracy: 0.9107
Epoch 9/10
150000/150000 [=====] - 2s 13us/sample - loss: 0.2084 - binary_accuracy: 0.9183
Epoch 10/10
150000/150000 [=====] - 2s 13us/sample - loss: 0.1938 - binary_accuracy: 0.9247

```

[그림 2-7] 모델 학습

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.4.1 네이버 영화 리뷰 데이터셋을 이용해 모델을 학습

- 시청자 채팅 긍정 부정 분류를 함에 있어서 토큰라이저를 Okt와 SPM 을 이용 했을 때 성능을 비교 분석, 분석 결과 accuracy는 미세하게나마 Okt 가 성능이 높음을 확인

<pre> Preprocessing Model define and train 2020-04-20 23:01:42.792775: I tensorflow/stream_executor 2020-04-20 23:01:43.050767: I tensorflow/core/common name: GeForce GTX 960 major: 5 minor: 2 memoryClockR pciBusID: 0000:01:00.0 2020-04-20 23:01:43.208535: I tensorflow/stream_executor 2020-04-20 23:01:43.227322: I tensorflow/core/common 2020-04-20 23:01:43.241622: I tensorflow/core/platform 2020-04-20 23:01:43.311320: I tensorflow/core/common name: GeForce GTX 960 major: 5 minor: 2 memoryClockR pciBusID: 0000:01:00.0 2020-04-20 23:01:43.319165: I tensorflow/stream_executor 2020-04-20 23:01:43.324533: I tensorflow/core/common 2020-04-20 23:01:51.448072: I tensorflow/core/common 2020-04-20 23:01:51.454392: I tensorflow/core/common 2020-04-20 23:01:51.457189: I tensorflow/core/common 2020-04-20 23:01:51.558203: I tensorflow/core/common ice: 0, name: GeForce GTX 960, pci bus id: 0000:01:0 2020-04-20 23:01:53.866980: I tensorflow/stream_executor accuracy: 0.84936 </pre>	<pre> Preprocessing Model define and train 2020-04-20 23:24:49.044689: I tensorflow/stream_executor 2020-04-20 23:24:49.178280: I tensorflow/core/common name: GeForce GTX 960 major: 5 minor: 2 memoryClockR pciBusID: 0000:01:00.0 2020-04-20 23:24:49.192208: I tensorflow/stream_executor 2020-04-20 23:24:49.203156: I tensorflow/core/common 2020-04-20 23:24:49.212021: I tensorflow/core/platform 2020-04-20 23:24:49.221338: I tensorflow/core/common name: GeForce GTX 960 major: 5 minor: 2 memoryClockR pciBusID: 0000:01:00.0 2020-04-20 23:24:49.238090: I tensorflow/stream_executor 2020-04-20 23:24:49.253282: I tensorflow/core/common 2020-04-20 23:24:56.890323: I tensorflow/core/common 2020-04-20 23:24:56.896674: I tensorflow/core/common 2020-04-20 23:24:56.899407: I tensorflow/core/common 2020-04-20 23:24:56.946989: I tensorflow/core/common ice: 0, name: GeForce GTX 960, pci bus id: 0000:01:0 2020-04-20 23:24:58.843789: I tensorflow/stream_executor accuracy: 0.82922 </pre>
---	---

[그림 2-8] Okt 토큰라이저를 사용한 모델

[그림 2-9] SPM 토큰라이저를 사용한 모델


- 긍정 부정의 분석 결과를 확인한 바 영화 리뷰에 대해서 학습 했기 때문에 분류가 제대로 되지 못했다. 그래서 채팅에 대한 데이터셋을 구성해야 할 필요성 확인

```

[엠브로 반가워요 🍌🍌🍌❤️]는 부정
[슬기님 안녕하세요]는 부정
[브로~~~~~]는 부정
[오~~]는 부정
[슬기님은 엠브로님이랑같이알하시나요]는 부정
[유튜브 잘보고있는데 ]는 부정
[오~~~]는 부정
[오 슬기님 안녕하세요!!!!!!]는 부정
[슬기님 하이 좋음]는 부정
[둘이케미 좋음]는 긍정
[공방인가]는 부정
[오! 자꾸옆쪽을보시지]는 부정
[옆쪽에 사람들이많대요]는 부정
[생방인가요]는 부정
[브로 ~ 슬기 ~ ]는 긍정
[안녕하신교]는 부정

```

[그림 2-10] 채팅 데이터에 대한 분류 결과

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.4.2 새로운 채팅 데이터셋 구축


- 네이버 영화 리뷰 데이터셋의 텍스트와 실제 채팅 텍스트의 양상이 상이하여 자체 데이터셋을 구축하여 추가 학습하고 모델을 검증하기 위해 라벨링 진행 (2만개 라벨링)

```

긍정 1 부정 0 중지 q:0
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
쓰는 돈이 뭘 문제야 활동필이지 상 ㅋㅋㅋ
긍정 1 부정 0 중지 q:0
ㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
ㅋㅋㅋㅋㅋㅋ 배달의민족이지
긍정 1 부정 0 중지 q:1
전투민족...ㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
전투의민족 ㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
오늘 올라온 영상 썸네일 실황가 아직 안봤는데 ㅋㅋㅋ
긍정 1 부정 0 중지 q:1
와 아프리카수준 진짜 ㅋㅋ 티베랑 글자날이랑 친해서언급해도되는데
긍정 1 부정 0 중지 q:0
ㅋㅋㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
아프리카나 유튜브나 트위치나 카카오톡 다 똑같은데 ㅋㅋ 소속감가지는건가?
긍정 1 부정 0 중지 q:0
근데 뽀빠는 하는 거 보고 있으면 데카가 모잘라 보이는데 나만 그러나 ㅋㅋ
긍정 1 부정 0 중지 q:1
진짜 재밌는걸하는데 역경다 ㅋㅋ 말이라도안꺼내면 얼마나 클린해
긍정 1 부정 0 중지 q:0
ㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
아 ㅋㅋㅋ 엘름 스나이핑 ㅋㅋㅋㅋㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:1
ㅋㅋㅋㅋ
긍정 1 부정 0 중지 q:

```

[그림 2-11] 데이터셋 구축

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.4.3 직접 구축한 데이터셋을 통해 긍정 부정 분류

- 2만개의 채팅 라벨링 데이터를 15000개는 train-set, 5000개는 test-set으로 사용함
- 이를 토대로 토큰라이저를 Okt와 SPM로 했을 때 정확도와 소요 시간을 비교 분석함
- SPM 토큰라이저 모델은 네이버 영화평 데이터셋을 사용해 학습하였고, vocabulary 사이즈는 10000, 모델 타입은 word로 사용.
- 이 실험의 결과로 한국어 형태소 분석 기능이 뛰어난 Okt보다 SPM 토큰라이저를 사용하는 것이 채팅 데이터 분류에 유리하다는 것을 검증
- 정확도는 직접 구축한 데이터셋을 이용하여 평가하였고, 소요시간은 BJ 엠브로의 다시보기 영상의 채팅로그 (<http://vod.afreecatv.com/PLAYER/STATION/46443514>) 1401라인을 사용하여 계산하였다.


	Okt	SPM
Accuracy	71.38%	86.03%
Time	47.7s	47.0s

[표 2-2] 토큰라이저 성능 비교

- 긍정 부정 분류가 채팅에 맞춰서 제대로 나눠짐을 파악함

[답변 부탁요]는 긍정
 [엠브로형 잘생겼어요]는 긍정
 [2+1도아닌데]는 긍정
 [사진도찌금]는 긍정
 [ㅇㅇㅇㅇ ?]는 긍정
 [스레기님 못생긴 이적달았어요]는 부정
 [빠연이]는 긍정
 [저는요 ?]는 긍정
 [1+2 인 듯]는 긍정
 [이적보단 잘생겼죠]는 긍정
 [아]는 긍정
 [저 살짝 이현우닮음]는 긍정
 [그냥 우유를 더넣으면되지 —]는 부정
 [아닌데]는 긍정
 [슬기님이 훨잘생김]는 긍정
 [은쥬님 까불지 마세요 디집니다]는 부정

[그림 2-12] 구축한 데이터셋을 통한 분류 결과


 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.5 채팅에서 가장 많이 나온 키워드 추출

- 방송중에 채팅에서 가장 많이 언급된 단어를 추출
- 형태소 분류기 4개를 토대로 가장 적합한 단어를 추출하는 분류기 선택 실험 진행
 - Hannanum: 한나눔. KAIST Semantic Web Research Center 개발
 - Kkma: 꼬꼬마. 서울대학교 IDS(Intelligent Data Systems) 연구실 개발
 - Komoran: 코모란. Shineware에서 개발
 - Okt(Open Korean Text): 오픈 소스 한국어 분석기. 과거 트위터 형태소 분석기
- AfreecaTV 킹기훈의 채팅로그(<http://vod.afreecatv.com/PLAYER/STATION/53773494>)를 사용하여 테스트하였으며 ‘다’, ‘노’, ‘마’와 같은 실질적 의미가 적은 길이가 1인 명사를 제외한 길이가 2 이상인 명사만을 사용
 - 채팅에서 많이 출연한 상위 10개의 명사를 추출한 결과
 - 같은 채팅로그에서 명사를 추출했음에도 같은 명사의 빈도수가 다르게 측정되는 것은 형태소 분석기 간 띄어쓰기가 없는 채팅에서 명사를 추출하는 성능의 차이가 존재하기 때문

◦ Hannanum			◦ Kkma			◦ Komoran			◦ Okt		
	명사	출연 빈도수		명사	출연 빈도수		명사	출연 빈도수		명사	출연 빈도수
1	ㅋㅋㅋㅋ	510	1	얼음	264	1	기후	226	1	기훈	276
2	ㅋㅋ	279	2	기훈	220	2	얼음	252	2	얼음	169
3	기훈	240	3	기훈이	167	3	가현	238	3	가현	173
4	ㅋㅋㅋ	192	4	가현	164	4	돼지	135	4	돼지	145
5	ㅋㅋㅋㅋ	178	5	돼지	142	5	우유	124	5	우유	118
6	얼음	137	6	우유	129	6	커피	91	6	기후	102
7	돼지	191	7	기후	112	7	오늘	84	7	그냥	98
8	우유	80	8	커피	93	8	유하	72	8	오늘	90
9	기후	79	9	오늘	85	9	김정은	62	9	커피	82
10	가현	73	10	술가락	47	10	달고나	56	10	유하	74

[그림 2-13] 형태소 분류기 키워드 추출 성능 비교

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

- 채팅에서 많이 출연한 상위 10개의 명사 추출에 소요된 시간


형태소 분석기	소요시간(초)
Hannanum	11
Kkma	132
Komoran	8
Okt	15

[그림 2-14] 형태소 분류기 명사 추출 소요 시간

- 띄어쓰기가 올바르게 올바르지 않은 채팅에서 명사를 추출한 결과
 - 테스트 문장 1: 오늘레전드컨텐츠 들고왔땅, 테스트 문장 2: 질혜님생일파티안가세요
 - 두번의 테스트 모두 Okt가 가장 좋은 성능을 보여줬다.

형태소 분석기	결과	형태소 분석기	결과
Hannanum	오늘레전드컨텐츠, 들고왔땅	Hannanum	질혜님생일파티안가
Kkma	오늘, 오늘레전, 컨텐츠, 왔땅	Kkma	파티, 질혜님생일, 생일, 안가, 안가세요, 세요
Komoran	추출된 명사 없음	Komoran	생일, 파티, 가세
Okt	오늘, 레전드, 컨텐츠	Okt	질혜, 생일, 파티


[그림 2-15] 띄어쓰기가 올바르게 올바르지 않은 테스트 문장으로부터의 명사 추출 성능 비교

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

- 매우 긴 실행시간이 소요되지 않고 띄어쓰기가 정확하지 않은 문장에서 명사를 추출하는 성능이 좋은 Okt 형태소 분석기가 채팅에서 많이 출연한 명사를 추출하는 기능을 구현하는데 사용하기 가장 적합했다.

형태소 분석기	장점	단점
Hannanum	- 소요시간이 적음	- "ㅋㅋㅋ", "ㄱㅅ"과 같은 완전하지 않은 문자를 명사로 인식함 - 띄어쓰기가 정확하지 않은 문장에서 명사를 추출하지 못함
Kkma	- 감전수용소'에서 명사를 추출하면 '감점', '수용소', '감전수용소'와 같이 자세히 추출됨	- 소요시간이 큼 - "ㅋㅋㅋ", "ㄱㅅ"과 같은 완전하지 않은 문자를 명사로 인식함 - 긴 문자열을 분석하면 메모리 오류가 발생함으로 길이가 20 이상인 채팅 문자열은 쪼개어 분석해야 함 - 띄어쓰기가 정확하지 않은 문장에서 명사를 추출하지 못함
Komorani	- 소요시간이 적음	- 띄어쓰기가 정확하지 않은 문장에서 명사를 추출하지 못함 - 😊와 같은 특수한 이모티콘을 인식하지 못해 유니코드 에러가 발생함으로 이모티콘은 별도의 예외 처리가 필요함
Okt	- 소요시간이 적음 - 띄어쓰기가 정확하지 않은 문장에서 명사를 잘 추출함	


[그림 2-16] 형태소 분석기 성능 비교 결과 정리

 <div> 국민대학교 컴퓨터공학부 캡스톤 디자인 I </div>	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.6 프론트엔드 구성

- 리엑트를 사용해서 Model 부분 구현.
- 로그인 후 사용자가 URL을 입력하면 해당 다시보기 링크를 분석해 결과 가시화
- Mocking을 통해 데이터를 지표화
- 중간평가 이후 API 서버와 연결하여 실제 데이터를 나타낼 예정




 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

2.2.7 백엔드 구성 및 구현 내용

- 프레임 워크로 플라스크를 사용
- 프론트 서버에서 API를 호출해서 사용할 수 있도록 CORS를 적용
- 에러 내역을 남겨 디버깅에 용이하도록 파일로깅
- gevent의 WSGI를 사용하여 구동
- alembic을 이용하여 스키마를 버전 관리
- ORM으로 SQLAlchemy 사용
- pytest로 테스트코드 작성해 검증된 API 제공
- 테스트용 database 구축
- 환경변수를 이용해 개발 / 테스트 / 배포 세가지 환경으로 구성

2.3 계획서상의 진도와 비교 분석

- 음성파일 분류 모듈, 채팅 로그 분석, 긍부정 감정 분석, 웹 페이지, 서버 각각의 코드는 모두 구현이 완료되었으나 모듈로 분리되어있지 않다. 따라서 각 소스를 모듈화 하고 연동 시키는 작업이 추가적으로 필요하다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

3 수정된 연구내용 및 추진 방향

3.1 수정사항

[삭제] 기존에는 하이라이트를 인공지능을 이용하여 추출 할 계획이었으나 인공지능을 배제

- 다시보기 영상을 초당 10 프레임 별로 하이라이트 영상과 매칭시켜 라벨링 시도
- 프레임별로 매칭시간이 매우 오래 걸림
- 움직임이 많이 없는 BJ의 경우 매칭정도가 유효하지 못해 라벨링의 문제가 발생
- 편집영상의 경우 효과로 인해 매칭이 제대로 이뤄지지 않음
- 프레임 연산을 하는 시간이 너무 오래 걸려 데이터셋을 구축하는 시간이 일정 수준을 넘어섬
- 프레임을 다루는 순간 부터 연산이 오래 걸리기 때문에 이를 배제하기로 결정
- 채팅량과 음성 변화도 만으로 하이라이트 지점 추출이 가능하다고 판단

[추가] 영상 편집에 있어서 음성 평준화를 해주면 편집자에게 도움을 줄 수 있다고 판단


- 다시보기의 경우 음성이 어떻게 녹음 되었냐에 따라 음량의 크기가 너무 크거나 작게 녹음되는 현상 발견
- 이는 시청자가 영상을 시청할 때, 음량의 크기에 따라 수동으로 볼륨을 조절해야 한다는 불편함이 있음
- 음성파일을 분석해 moviepy 라이브러리를 사용하여 볼륨 수준을 평준화하여 다운로드 받을 수 있도록 구현

[추가] 채팅의 다양한 감정 분류

- 채팅을 긍정/부정으로 나누는것을 확장하여 다양한 감정으로 분류
- 이를 통해 편집자 및 스트리머는 시청자가 영상의 특정 지점에서 어떤 감정을 느꼈는지 파악할 수 있고, 영상의 썸네일 또는 편집의 컨셉을 정하는데 도움을 줄 수 있음

[추가] 채팅 감정 분류를 통한 BGM 추천

- 채팅을 다양한 감정으로 분류한 결과를 통해 영상을 보는 시청자의 특정 감정이 도드라질때 감정에 맞는 음악을 추천해줌으로써 편집에 용이하게 이용가능

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

4 향후 추진계획

4.1 향후 계획의 세부 내용

4.1.1 개발한 각 모듈들을 백엔드와 연결 시키기

- 오디오와 채팅 로그를 다운 받는 기능 모듈화 한다.
- 채팅 로그를 통해 분석하는 기능들 모듈화 한다.

4.1.2 채팅 긍정/부정 분류

- 토큰나이징하기 전에 맞춤법을 검사를 통해 띄어쓰기, 오타자 등을 수정하고 토큰화 하여 정확도를 향상시킨다.

4.1.3 채팅의 다양한 감정 분류


- 채팅 긍정/부정에서 더 나아가 시청자들의 반응을 다양한 감정으로 분류하여 스트리머들이 시청자의 반응을 구체적으로 살필수 있게 한다.

4.1.4 분석한 데이터 시각화

- 웹페이지에 실제 오디오 및 채팅 로그들을 분석한 결과를 시각화하기
- 영상의 볼륨 크기를 그래프로 시각화
- 가장 등장 빈도가 많은 키워드들 시각화
- 시청자의 반응 긍정/부정 분류 및 시각화
- 시청자별 채팅 참여도 시각화
- 선택한 단어가 나오는 채팅 등장 시기 및 빈도 가시화

4.1.5 채팅 감정 분류를 통한 BGM 추천

- 하이라이트로 추출된 지점의 채팅 감정 분류를 통해 해당 지점에 삽입할만한 BGM을 추천하여 편집이 용이하게 한다.

 국민대학교 컴퓨터공학부 캡스톤 디자인 I	중간보고서		
	프로젝트 명	요바(요기바바)	
	팀 명	안진마	
	Confidential Restricted	Version 1.3	2020-APR-23

5 고충 및 건의사항

- 컴퓨팅 자원이 부족해 인공지능을 활용한 기능을 더 추가하지 못한 것이 아쉽습니다.