자연어 처리 및 정보검색

(프로젝트 보고서)

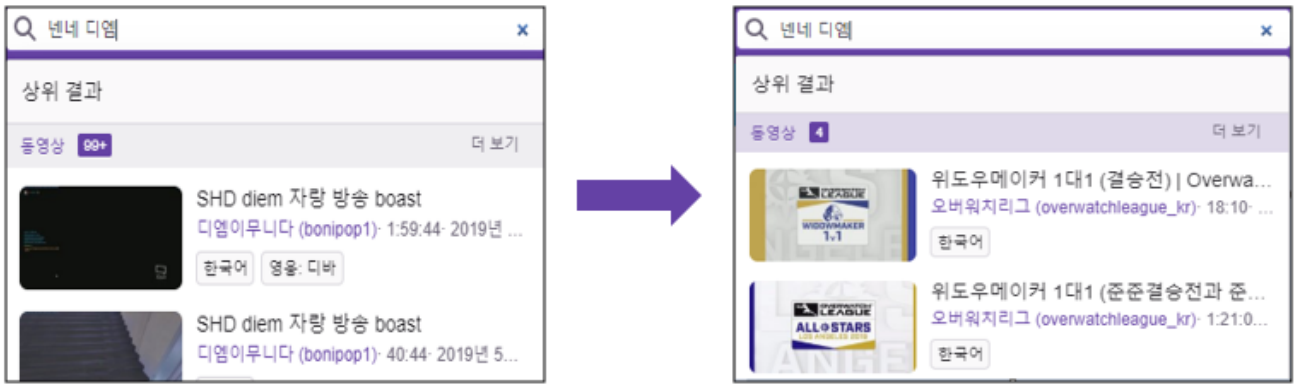
담당 교수님	정재은 교수님
팀 이름	AND GAME
팀 원	20145034 홍성현 20146363 강민승 20133950 공찬형 20142921 이승현 20143567 이희상

목 차

1. 프로젝트 주제

Video retrieval system using twitch chat log (트위치 채팅로그를 이용한 영상목록 검색 시스템)

구체적인 내용을 넣어도 관련된 영상이 나오는 검색 시스템 (ex) 넨네, 디엠(오버워치 프로게이머)

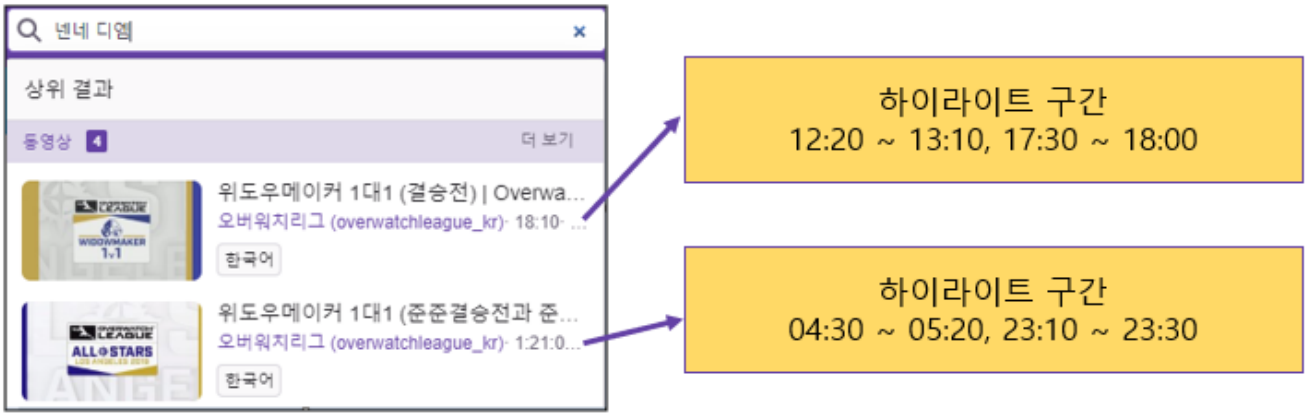


관계없는 내용이 나온다 (기존 시스템)

찾고 싶은 영상 검색됨 (예상 결과)

추가로, 검색 결과 영상에 대해 하이라이트 구간을 알려 줌

(검색 키워드 정보 이용 X, 시간에 대해 단어 빈도수만 이용)



2. 팀이름 및 팀원 및 깃허브 주소

- And Game

학번	학부	학년	성명	역할
20145034	컴퓨터공학부	4	홍성현	팀장, IR 시스템
20146363	컴퓨터공학부	4	강민승	데이터 수집
20133950	컴퓨터공학부	4	공찬형	데이터 전처리
20142921	컴퓨터공학부	4	이승현	IR 시스템
20143567	컴퓨터공학부	4	이희상	평가를 위한 TC 생성

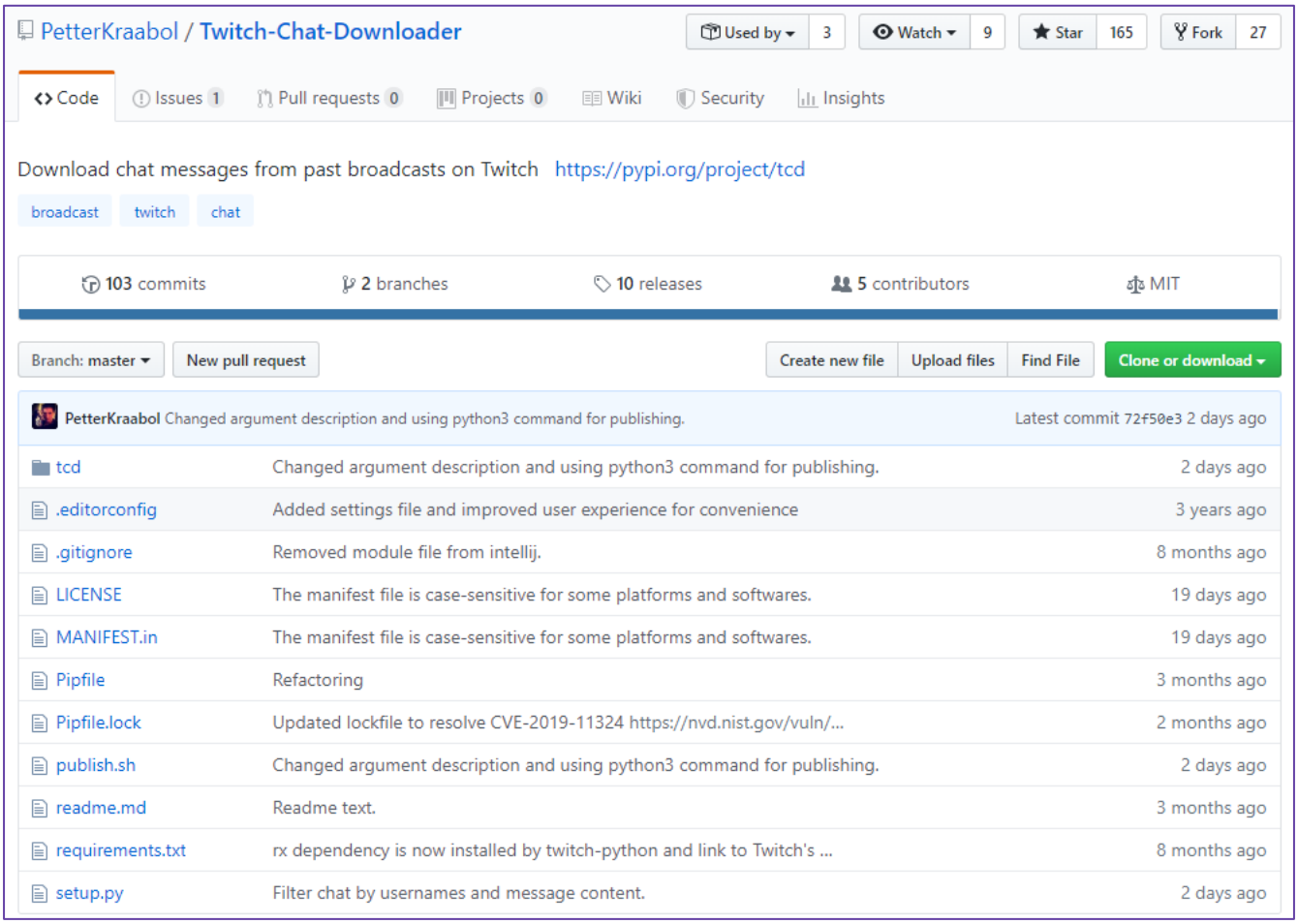
- https://github.com/cau-andgame

3. 프로젝트 일정

	5월			6월		
	17	20	27	3	8	14
프로젝트 주제 발표						
프로젝트 세부 사항 결정 및 설계						
데이터 수집 및 전처리						
영상 검색 시스템 구현						
하이라이트 시간 추출						
IR 시스템 평가(evaluation)						

4. 데이터 수집

- 'TwitchChatDownloader'를 사용




```
tcd
```

```
# Download chat from VODs by video id
tcd --video 789654123,987456321 --format irc --output ~/Downloads
```

```
# Download chat from the first 10 VODs from multiple streamers
tcd --channel sodapoppin,nymn,lirik --first=10
```

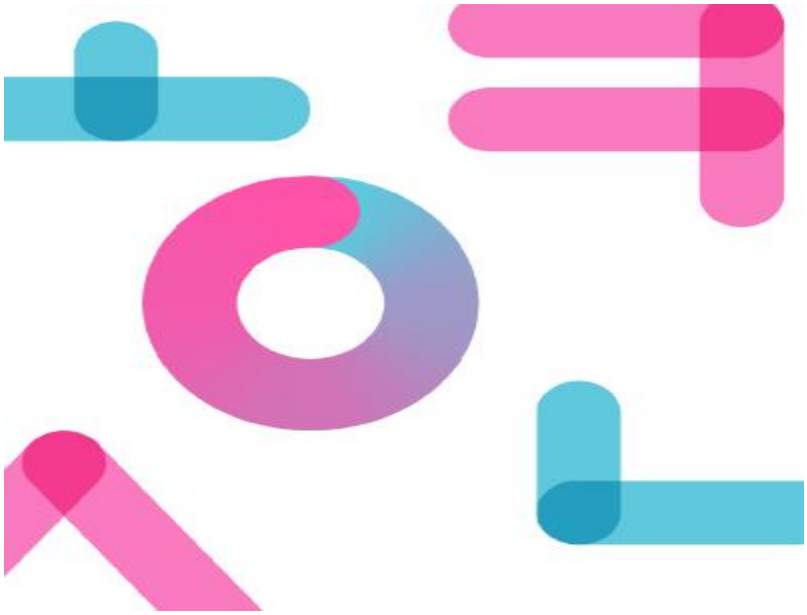
```
~/workspace/nlp-project/data tcd -c yapyap30 --first=10
```

채팅 로그 추출 대상 : 한국의 게임 스트리머, 공식 대회 중계 채널 등

전체 373 개의 동영상에 대해 8,960,777 개의 채팅을 수집했다. 영상 별 평균 채팅은 88,720 개다.

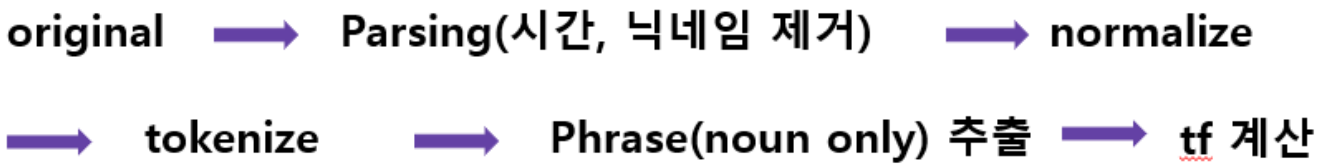
5. 데이터 전처리

'Open Korean Text' 사용



데이터 전처리는 여러 가지 한글 전처리 프로그램(Nori, kkma, KoNLPy)들을 이용해보고, 그 중 가장 우수한 'open Korean text'를 선택했다.

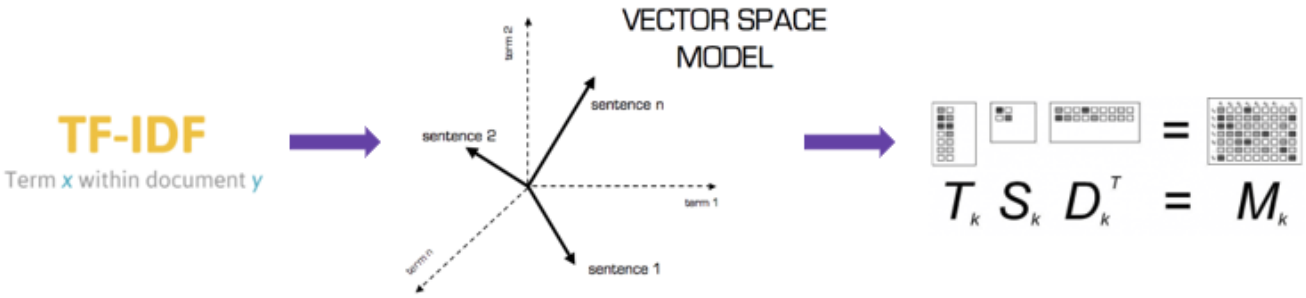
- 전처리의 과정은 다음과 같다.



우선, 원래의 data 에서 시간과 닉네임을 제거한 후, normalize, tokenize, phrase 추출 과정을 거친 후, tf 를 계산했다. 그 후 결과를 관찰해보니 예상했던 용어들이 높은 빈도로 언급되는 것을 확인할 수 있었다. 그 후, 커뮤니티나 공식 사이트에서 사용되는 명칭들을 모아 사용자 corpus 를 추가해봤으나 성능이 하락했다(mAP 74.71% → 65.87%).

6. IR 시스템

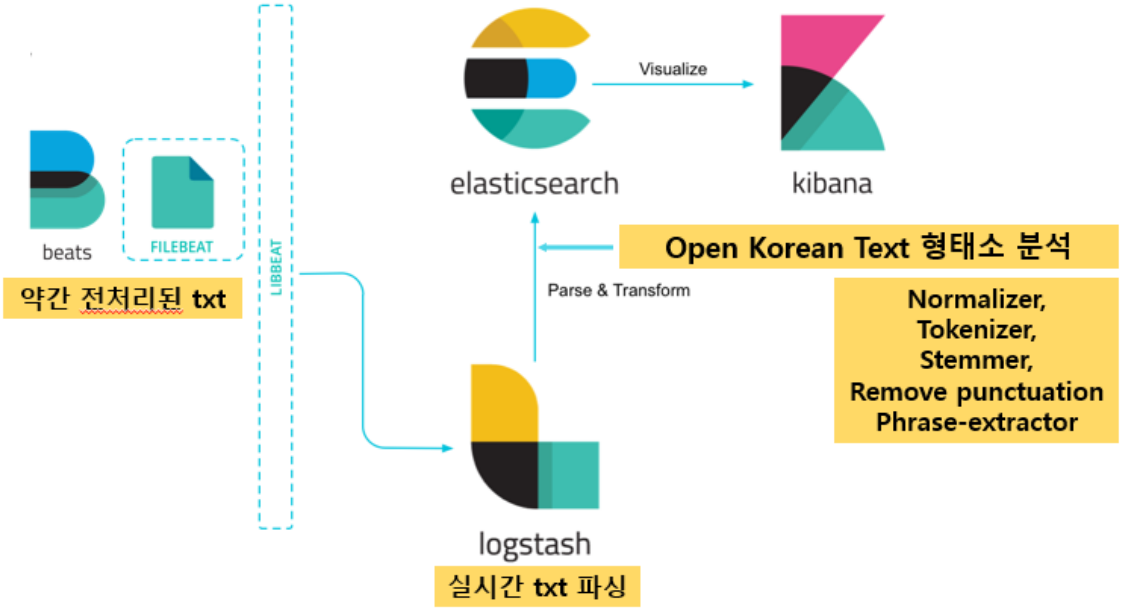
기존에는 수업 시간에 배운 개념들을 이용해 직접 코딩하여 구현하려고 했다.



먼저 각각의 동영상을 document 로 하고 채팅에서 추출한 token 들을 term 으로 하여 tf 와 idf 를 계산하고, 이를 바탕으로 Vector Space Model 을 만든 뒤, 잠재 의미 분석(LSA)을 수행하여 IR 시스템을 구현하려고 했다.

하지만, 실제로 직접 구현을 해본 결과, 메모리가 초과하는 문제가 발생했고, 적은 데이터로 테스트를 돌려본 결과, 시간이 지나치게 오래 걸려 사용하기 어렵다는 결론을 내렸다.

그래서 beat 와 엘라스틱 서치를 이용해서 IR 시스템을 구현하고, 키바나를 이용해 시각화했다. 이를 도식화한 것은 아래와 같다.



엘라스틱 서치를 사용하면서 tf-idf 대신 BM25 를 사용했는데, 그 이유는 여러 논문들과 TREC 등의 챌린지에서도, 여러 커뮤니티의 사용자들도, 루씬의 개발자들도 잘 동작한다고 했기 때문에 채택했다.

처음에는 소수의 영상들만 계속 상위로 랭크되는 문제점이 있었다. 다양한 가설을 내리고 실험을 해 본 결과 채팅 로그 사이즈가 큰 영상들이라는 공통점을 찾을 수 있었고, 기존에는 score 의 값들을 단순히 더해주었는데 이를 평균으로 바꿈으로써 normalize 를 수행했다. 그 결과, 성능이 개선되었다.

(normalization 수행 전 mAP: 69.85% → normalization 수행 후 mAP: 73.87%)

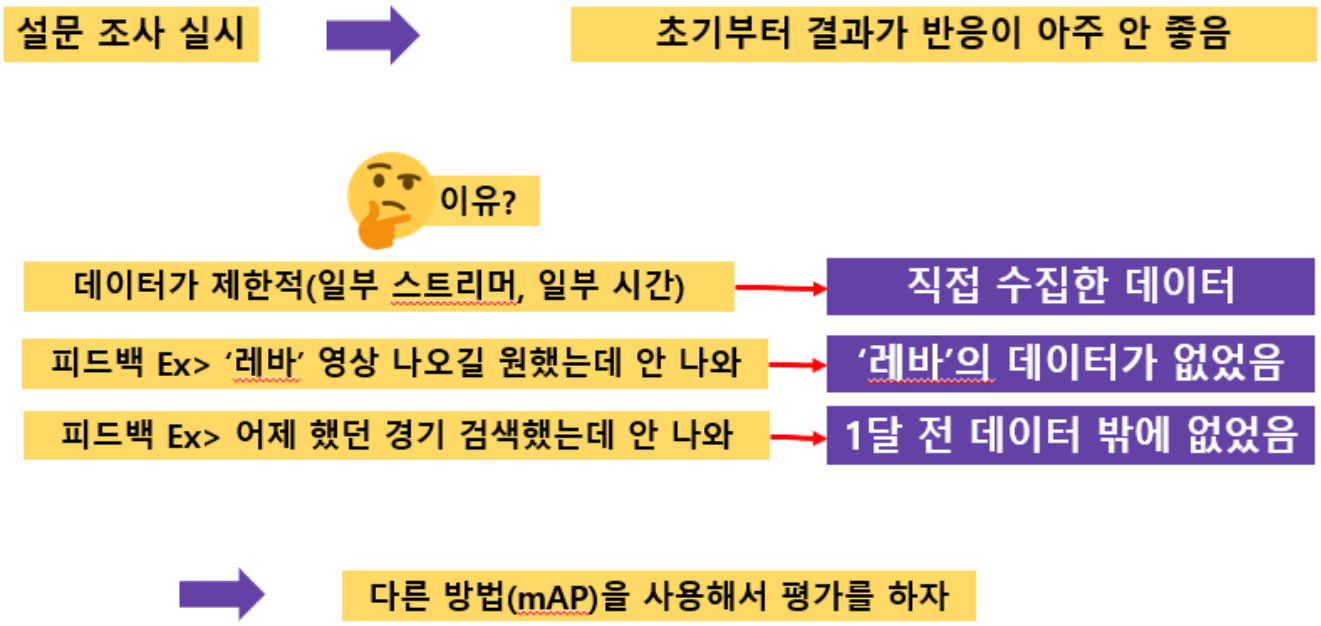
7. 하이라이트 시간 추출

최초 계획에는 없었던 부분이지만, 2 주차 브리핑 이후 교수님께서 추가적으로 생각해보라고 하셔서 반영했다. 하이라이트에는 많은 시청자들이 채팅을 칠 것이라고 예상했기 때문에 특정 시간 동안 채팅 로그의 빈도 수가 급증하는 구간을 추출하기로 했다. 그 결과는 다음과 같다.

주로 '와', '오' 등의 감탄사나 실수를 책망하는 부분, 혹은 'ㅋㅋㅋㅋㅋㅋㅋ' 등의 웃음이 자주 등장하는 부분들이 하이라이트로 추출되었다. 실제 하이라이트 영상을 보면 놀랄 만한 멋진 장면이거나, 아주 웃긴 장면이거나, 혹은 결정적인 실책을 한 장면이었고, 이들을 하이라이트라고 부르기에 무리가 없을 것이라고 판단했다.

8. 평가

설문 조사를 실시해서 평가를 하고자 했다. 하지만, 초기부터 반응이 매우 안 좋았다. 그래서 바로 원인 조사에 착수했는데, 받았던 피드백을 예로 들어 "레바 영상이 나오길 기대했는데 나오질 않았다.", "어제 했던 경기 검색했는데 안 나온다." 등이 있었다. 이에 대한 이유는 실시간으로 모든 채널의 데이터를 수집하는 것이 아니기 때문이었다. 우리가 수집한 채널들이 한정적이고, 그 시간 역시 과거의 데이터이므로 사용자 만족도로 평가하면 안 좋게 나올 수 밖에 없었다고 생각했다. 사용자를 만족시키기 위해서는 실시간으로 모든 채널의 데이터를 수집해와야 하지만, 프로젝트 기한 내로는 불가능하다고 판단했다.



그래서 현재 프로젝트를 평가할 방법으로 mAP 를 채택했다. 프로젝트 초기부터 팀원 한 명이 꾸준히 Test Case 를 생성했고, 이를 기반으로 mAP 를 계산했다. 검색어에는 스트리머의 별명, 스트리머의 애완동물, 프로게이머, 게임 캐릭터 이름, 게임 아이템 이름, 특정 상황 묘사 등 채팅에 나올만한 내용들을 대입했고 그 결과로 나올 것이라고 예상되는 동영상의 목록들을 만들었다.

검색어	예상결과	ap
감자	419690936, 420168205, 420687628...	1
루밍	413283283, 414286782, 414735658...	1
ymq	412334252, 412813845, 413216289...	0.7
새봄추	427826743, 428215290, 428256596...	0.8
차보해	428601939, 429051750, 429564491...	0.1
김진효	413283283, 414286782, 414735658...	1
이희주	412334252, 412813845, 413216289...	0.9
김성태	419690936, 420168205, 420687628...	1
이상혁	398393566, 400494624, 400994662...	0.875
윤현우	412831781, 414318985, 414755911...	1
조현수	424993287, 425425439, 427256306...	0.8
은돌이	427826743, 428215290, 428256596...	0.6
첵스	419690936, 420168205, 420687628...	1
오즈	419690936, 420168205, 420687628...	1
스웩이	413283283, 414286782, 414735658...	0.8
페이커 나르 5인궁	426391802 ,426377000	1
lck 결승	410315530	0.5
msi 결승	426877410, 426869245	0.29
페이커 눈물	410315530	0.5
페이커 라이즈 2대1	410315530	1
샌프란시스코 쇼크 우승	424066354	1
누누 챌린저 과외	431440690	0.5
페이커 솔랭	398393566, 400494624, 400994662...	0
2633	419690936, 420168205, 420687628...	0.9
lck 준결승 코르키 죽음의 무도	407457306	1
lck 쇼메이커	405423139, 436555262	0.33
소드 바이퍼 타잔 쵸비 리헨즈	410315530, 403960817	0.4
페이커 마타 테디 클리드 칸	426377076, 424670121, 424250883...	1
넨넨 위도우메이커	425409699	1
몰왕 초시계 평타	436555262	0.166
	평균 (mAP)	0.7389

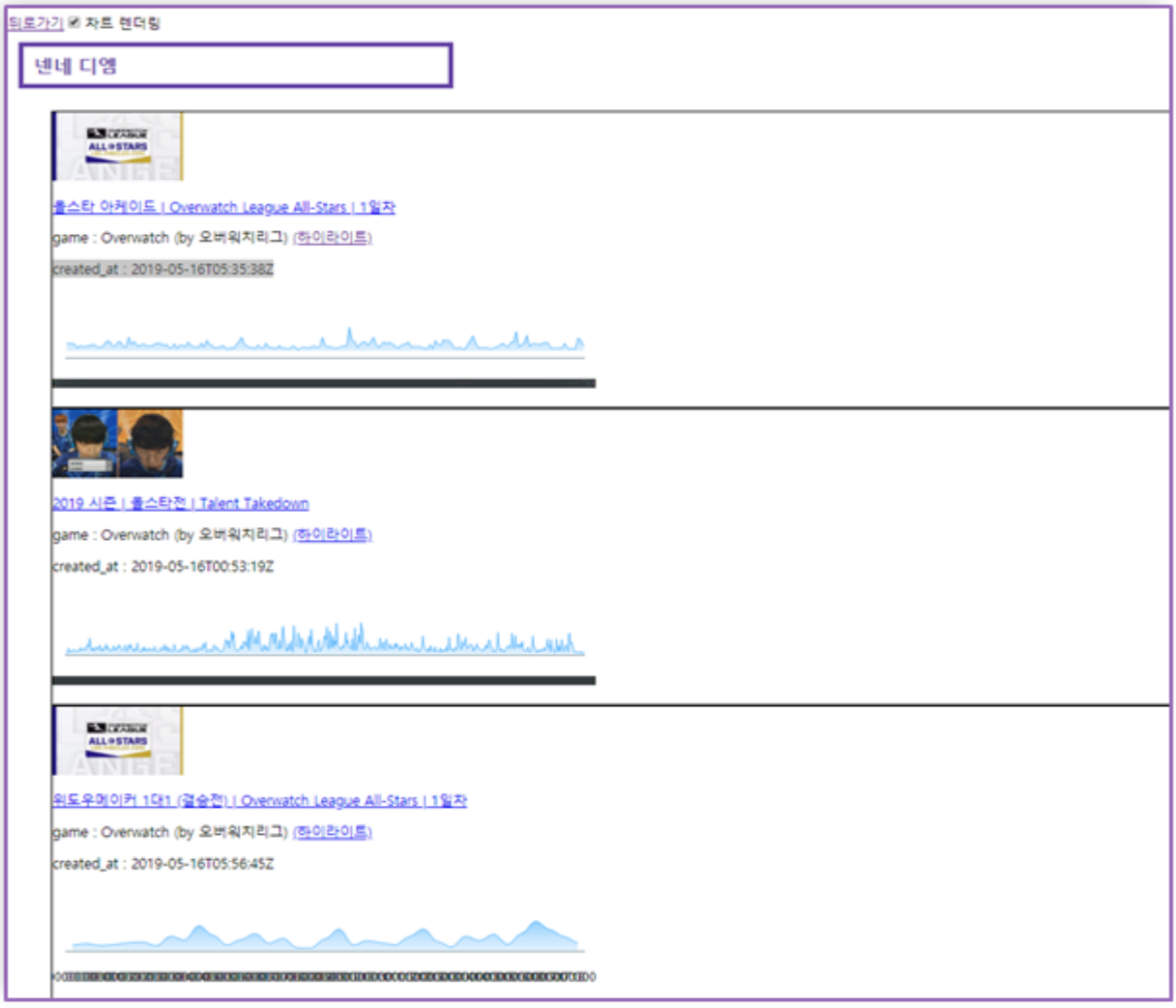
- 검색어와 예상 결과 목록(31 개), 그에 따른 ap 와 최종적인 mAP 를 계산한 테이블

9. 결과

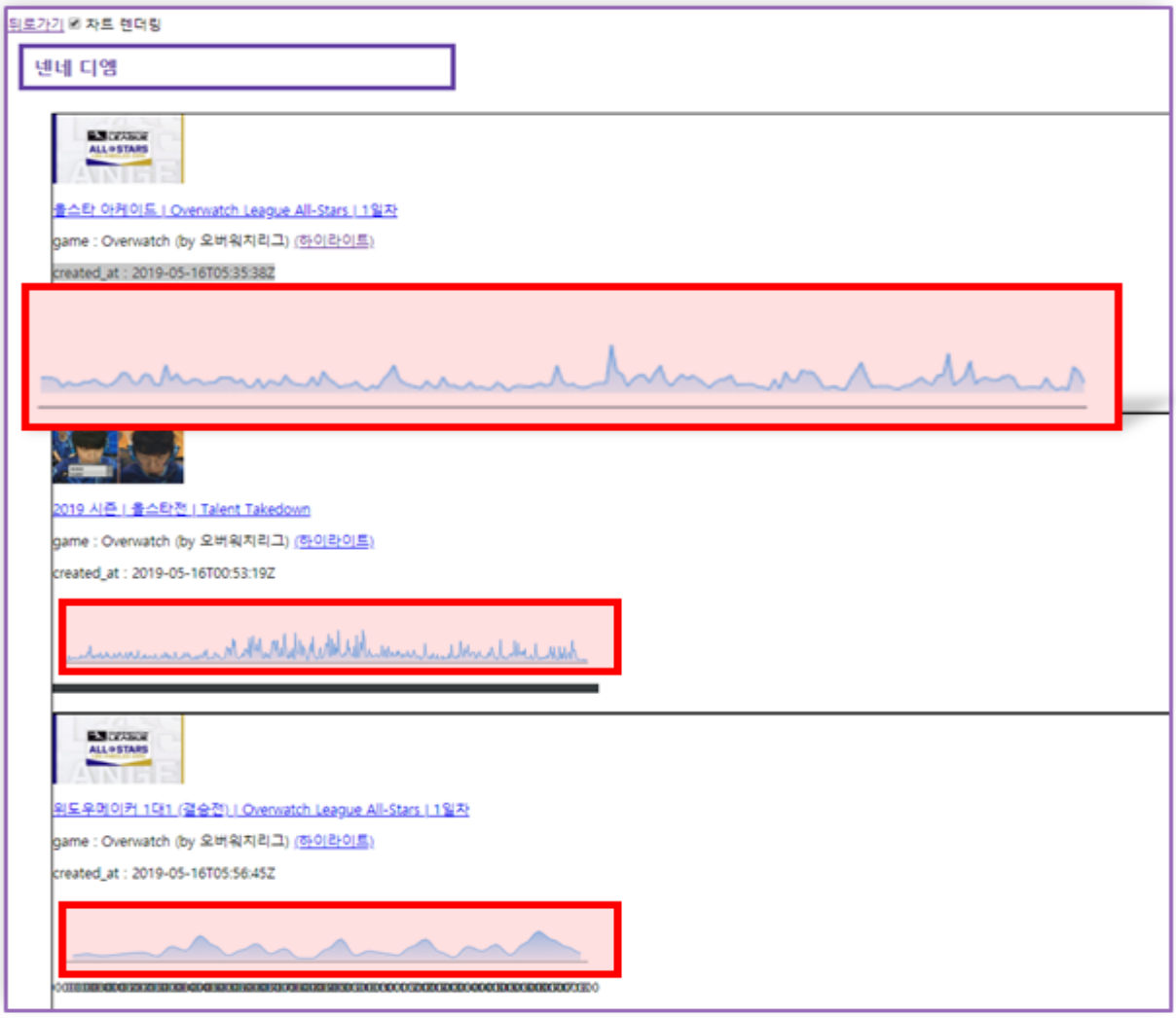
https://cau-andgame.github.io/twitch-demo-web/

위의 주소로 들어가면 실제로 검색을 하고 그 결과를 살펴볼 수 있다.

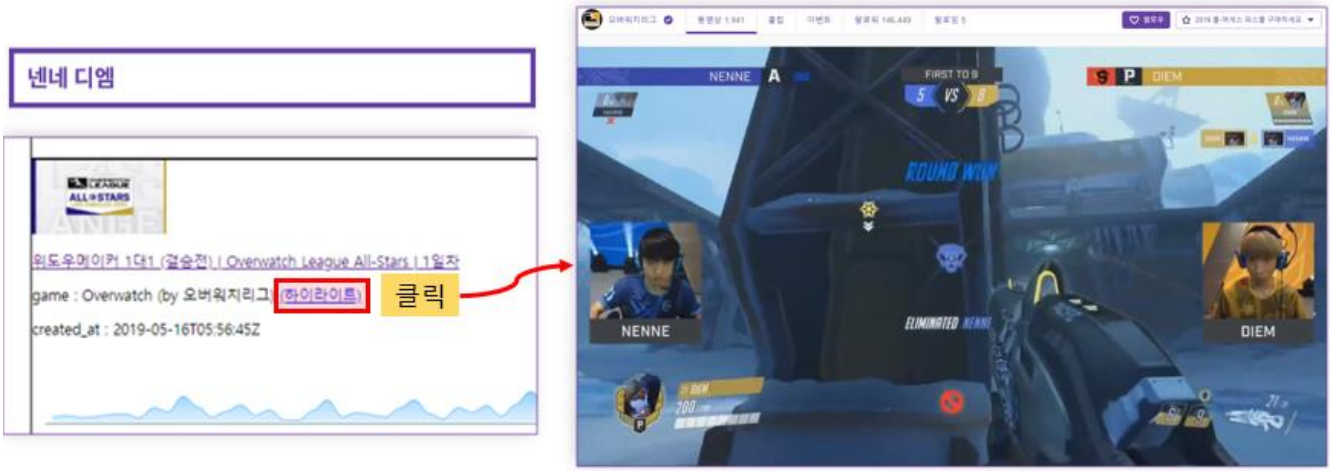
- 초기 화면



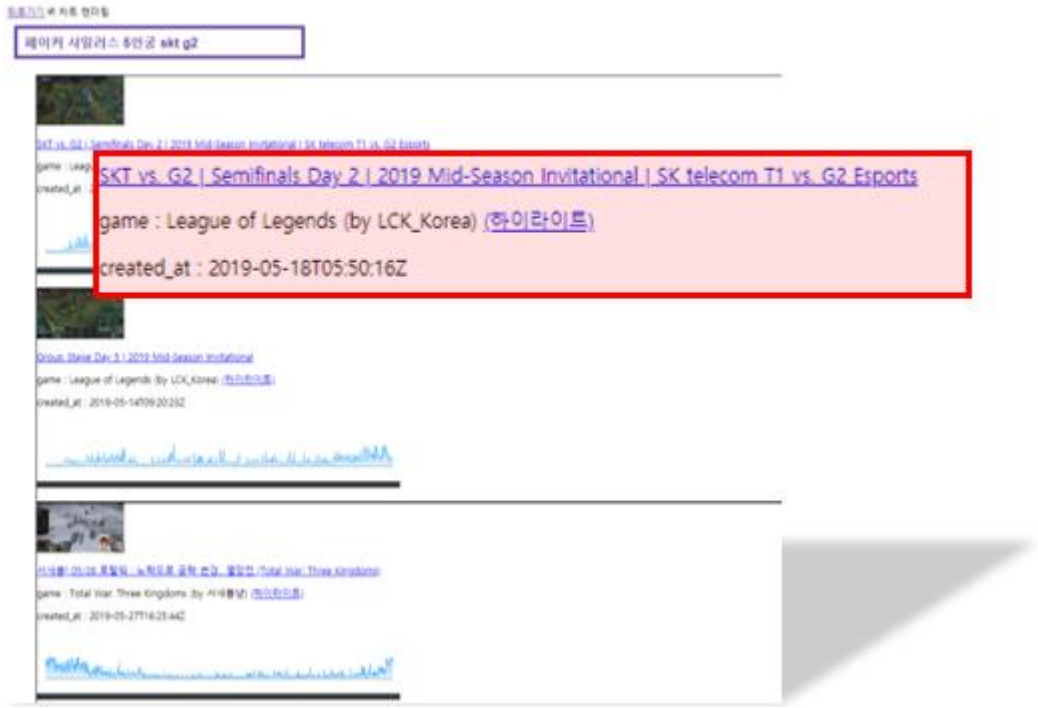
- '넨네 디엠' 검색 결과(* 넨네와 디엠은 오버워치 프로게이머로, 위도우 메이커라는 캐릭터를 이용한 1 대 1 결승 경기에 진출했다[3 번째 검색 결과])



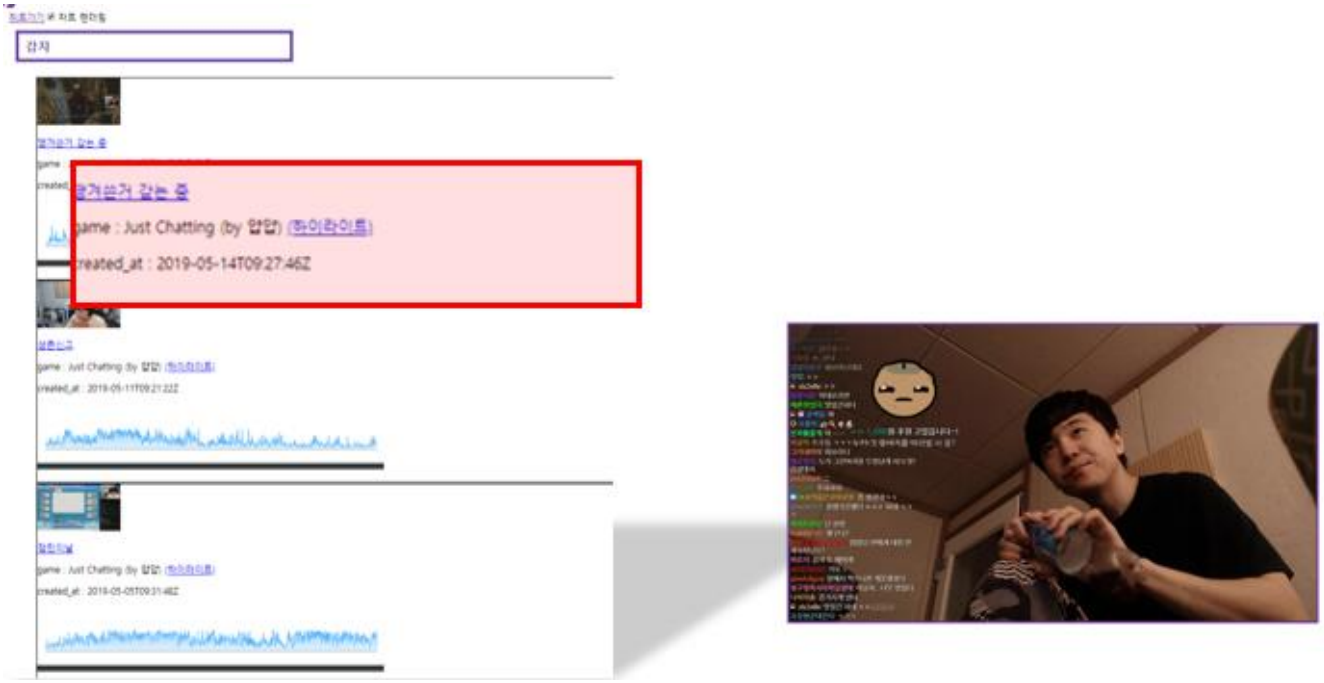
- 시간별 채팅 빈도 수의 변화를 그래프로 보여줌으로써, 하이라이트 구간이 어디인지 알 수 있도록 했다.



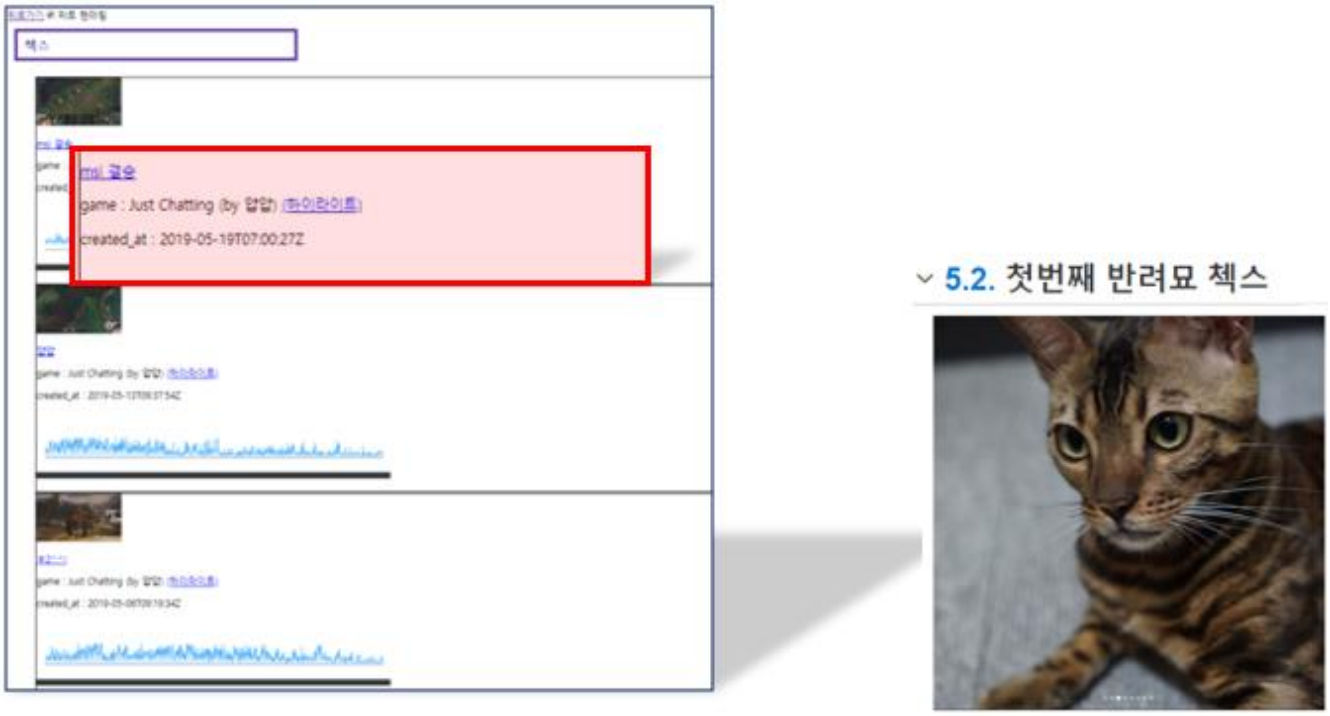
- 영상 결과에 있는 하이라이트를 클릭하면 순간적으로 채팅이 가장 높았던 부분부터 보여준다. (실제 트위치 사이트로 연결된다.)



- MSI 결승전 명장면에 대한 내용(페이커 사일러스 5 인궁 skt g2)으로 검색할 경우, 공식 중계 영상과 스트리머가 중계한 영상 모두 검색되었다.



- 스트리머의 별명을 입력해도 결과가 잘 나왔다.



- 스트리머가 키우는 애완동물 이름을 입력해도 결과가 잘 나왔다.

- Corpus 추가 전: 73.87%, Corpus 추가 후: 64.73% (성능 개선 X)
- Normalization 수행 전: 69.85% Normalization 수행 후: 73.87% (성능 개선 O)
- **최종 mAP: 73.87 %**