<Youtube 콘텐츠 썸네일에 따른 조회수 예측: 영화 및 드라마 리뷰를 중심으로>

**2024.01.22**

1. 분석 대상

지무비, 김시선, 고몽, 달빛뮤즈, 디토이야기, 무드킹, 영화한입만, 24프레임, 이쁜영화, 영화보는하루, 삐맨, Popcorn & Coke Review, 엘플릭스, 미쓰김씨네, 어퍼컷Tube, U무비, 1인실 영화관, 라떼한편, 9불(본인 이야기 나오는 공지 영상 있음), 영화쉼, 하씨네, 카랑, 어바웃타임, 무비체인, 화양영화, 빵또와영화채널, 동네비디오방, 뭅픽, 안보면쏜다, 무비톡, 바나나리뷰, 준또리뷰TV, 오티티: OTT, 하리봉무비, 추적드라마60분 (35개)

제외: 뭅뭅 ( 예고편 및 배우 관련 영상 다수)

2. 영상 제외 기준(전처리)

* 제작된 지 6개월 이상이 된 영상(선행논문에 따르면, 유튜브 콘텐츠는 제작된지 6개월 안에 최종 영상 조회 수가 결정되며, 이후에는 유의한 변화를 가지지 않는다.)
* 해당 유튜버들의 평균 조회수와 신뢰구간을 확인하여 이에 포함되지 않는 극단치들은 제외
* 쇼츠와 같은 일정길이 넘지 않는 영상 제외

3. 관련 데이터

* 구독자 수 추이: [https://socialblade.com/youtube/user/rladndgussla (3](https://socialblade.com/youtube/user/rladndgussla%20(3)년 간 구독자 수 추이 확인 가능 사이트)
* 유튜브 썸네일, 영상 제목, 영상 게시자, 조회수, 평균평점, 영상길이(초), 연령제한여부, 영상 설명 : 파이튜브를 활용하여 수집

**2024.01.23**

**영상 관련 feature**

**MAIN ANALYTICS**

1. 채널이름+ 영상~번째수(추후 dropped) O

2.길이 (s) O

3.게시날짜(데이터 불러온 날 기준 며칠전 영상인지?)

4.조회수

**4-1. 구독자 수 대비 조회 수 (y값)**

4-2. 평균 조회 수 대비 해당 영상의 조회 수

5. 수집날짜: 2024.01.24

6. 썸네일 주소(나중에 모델 돌릴때 지울 것임)

**이미지 feature**

**MAIN ANALYTICS**

1.유해성 스코어(5개 항목 sum()) O

2. 명도

3. 썸네일에 등장하는 인물: 인물 없음, 인물 있음 (개인: 인물 1-2명 / 집단: 인물 3명 이상)

4. 등장하는 인물의 성별 (구글은 PERSON만 나오고 아마존은 성별 구분 가능)

**SUB**

1. 썸네일에 사용된 색상 [top?] > colorful??? : 스펙트럼으로 뽑기 (+ 진행한다면 글자색과 같이)

2. 등장하는 인물의 표정

**썸네일 / 제목 텍스트 feature**

**MAIN ANALYTICS**

1. 감정분석 (전처리 필요): 네이버 클로바

2. 글자 배치 (상/중/하/없음)

3. 텍스트의 내용: 숫자 유무 => 이건 몇 개국 1위, 수익, 제작비 등 성과 강조로 시청 유도

* 시즌 및 회차별 정보가 걸릴 수 있으니
* 필터: 박스오피스, 수상, 넷플릭스, 평점, 제작비, r{숫자}위, 몇개국 등의 단어로 걸어두기

**SUB**

1. 텍스트의 내용: 시간 되면 자연어 처리

* 주인공의 먼치킨 특성 강조(신, 천재, 영재, 재벌 등) (전처리 하면서 딕셔너리 카운트해보고 할지 말지 판단)
* [결말 포함]의 유무 (2222)
* [몰아보기]의 유무(있다면 전체 회차/ 회차별 등)
* 주인공의 변신/사이다 강조(찐따가 일진 참교육, 왕따가 커서…, 통쾌, 복수)

2. 글자색: (어떻게 가져올지는 더 고민해볼 것): 색조합? 제일 많이 나온 글자 색깔?

3. 글자 스타일

**y값 (예측)**

조회수

- 구독자 수 대비 조회 수

- 평균 조회 수 대비 해당 영상의 조회 수

- 영상별 조회수 추이

수집한 데이터:

1. Youtube 채널 ID -> <https://ytlarge.com/youtube/channel-id-finder/ko> 사용하여 획득
2. Youtube 채널의 모든 영상 url 주소 -> Youtube Data API
3. Youtube 영상의 채널명, 제목, 길이, 게시날짜, 조회수, 썸네일 주소..etc -> Pytube(영상 url로 접근)

구조 : 이상한 사이트에서 획득한 채널 ID -> Youtube Data API에 사용하여 영상 url 획득 -> 획득한 Url 영상 주소를 통해서 pytube 라이브러리로 영상 정보 조회

**분석대상 추가** ( +기존 = 총 84개 채널, 데이터 10,000개 이상)

1. 덕봉이의 영사실 (51)
2. 고체극장 (35)
3. 찐뷰 영화리뷰 (59)
4. 기무리뷰 (207)
5. 영화천국 (27)
6. 무비야한 (52)
7. 하울무비 (67)
8. 콜라냥 (318)
9. 명작소녀 (356)
10. 프리리뷰 (14)
11. 풍산비디오 (116)
12. 시네마홀릭 (50)
13. 무비콕콕 (106)
14. 무비디바비디View (98)
15. 로투스 무비 (92)
16. 알락투무비 (71)
17. B급 리뷰 (262)
18. 어쩌다영화한편 (136)
19. 시네마톡 (100)
20. 씨네마유 (61)
21. 시네마헌터 (172)
22. 정씨네 (91)
23. 메트로 무비: [Metro Movie] (71)
24. 주말의 띵화 (203)
25. 리뷰 MASTER (181)
26. 영화중독 (109)
27. 영화맛집 (225)
28. 무비띵크\_Movie Think (792)
29. 무비튜브 (201)
30. 무비방 (112)
31. 퐉스오피스 (106)
32. 드림텔러 (227)
33. 나홀로팝콘 (15)
34. 엔스Ens (413)
35. 진솔한리뷰 (371)
36. 리씨네 (154)
37. 다크무비 (35)
38. 씨네 토피아 (75)
39. 곰줄 (57)
40. 제이퀀도 (261)
41. 청불곧추(50)
42. 나인무비(47)
43. 띵잘(175)
44. 채널휸\_리뷰하는 남자(66)
45. 소개해주는 남자(339)
46. 쩔무비(322)
47. 아재영화 Commani Movie(123)
48. 스토리머(65)
49. 잼플릭스(92)

—-------------------------------------------

여기서는 24.01.24 에서 추가

* 만약 더 추가한다면 ?!?!?!

1. 대나무숲 극장 (31)
2. 영화와 드라마 : Movie & Drama (15)
3. 순삭영화 (134)
4. 핫무비(95)
5. 리뷰봐야지(307)
6. 무비월드(90)
7. 시네마틱(19)
8. 리뷰아재-[드라마리뷰] (463)

**2024.01.24**

1. 썸네일 이미지: 아마존 ( 이미지 속성만: 이미지당 0.00075 USD -> 11,000개의 이미지를 대상으로 분석한다고 가정했을 때 8.25 USD) = 11,032원

2. 썸네일 글씨: 구글 (90일 무료)

3. 구독자 수 추이: 플레이보드 (29,000원): 결제 - 은선

4. 제목 글씨 감정분석 (썸네일은 전처리 후에 보는 걸로): 네이버 클로바 -> 1,000회 초과시 호출당 1원 -> 여유롭게 30,000개로 가정했을 때 30,000원

**썸네일의 크기 조정 및 과도한 글씨 인식 문제**

- 이미지 크기 모두 일치 시킨 후 > 이미지에서 좌표 넓이 기준으로 문자 인식 시키기

- 이미지에서 텍스트 크기 filtering 기준치 확인하기

**수집 데이터 수정**

- 크리에이터 ‘무비톡’에서 총 8개의 영상에서 ‘수정된 영상으로 보실 수 없습니다’ 오류 발견

- 비공개 및 수정된 영상 등 제대로 읽히지 않는 데이터는 전처리 과정에서 코드를 사용하여 제거 필요

**최종 데이터 관련 정보 (2023.01.23 feature 참조)**

**1.** 분석대상: 총 85개 채널, 데이터 10,000개 이상

* (숏폼 포함, 코드를 활용하여 1분 이내 영상은 제외하고 수집)

2. playboard 사이트에서 구독자 추이를 확인한 후, 구독자 추이 별 조회 수에 가중치를 부여하여 조회수 조정

**역할 분담**

1. 코드 및 데이터 수집: 오현진, 임형섭, 이승규
2. 데이터 분석 및 시각화(EDA): 이은선, 정은혜
3. 데이터 검수: 이은선, 정은혜
4. PPT 제작(데이터 수집 어떻게 했는지, EDA 및 데이터 분포, 전처리: feature 정한 이유, 목표, 무슨 모델을 사용했는지, 해당 모델의 hyperparameter 등, 앙상블 및 모델 결과, 이걸 어떻게 사용할 수 있을까 -> 결과): 전체
5. 선행 논문 및 관련 자료(기사 등) 수집: 전체
6. 발표: 정은혜

**일정**

1. 데이터 수집: 1/24 (수)
2. 데이터 전처리: 1/27 (토)
3. 데이터 검수 (썸네일의 text 가 잘 들어갔는지, 오류난 데이터가 있는지): 1/28 (일)
4. 모델링: 1/29 (월)
5. 모델링 향상 1/28 (화) ~
6. 데이터 시각화 및 PPT 제작: 1/29 (월) ~

배치(왼오위아래), 글씨의 글씨, 이미지 스타일

썸네일 잘 만드는 방법 등을 보고 이를 활용하여 기존의 정보를 바탕으로 체크해보기

이미지 스타일은 모델 만들어야 함

**2024.01.25**

**결제**

1. 썸네일 이미지: 아마존 ( 이미지 속성만: 이미지당 0.00075 USD -> 11,000개의 이미지를 대상으로 분석한다고 가정했을 때 8.25 USD) = 11,032원

2. 썸네일 글씨: 구글 (90일 무료)

3. 구독자 수 추이: 플레이보드 **(29,000원: 은선)**

4. 제목 글씨 감정분석 (썸네일은 전처리 후에 보는 걸로): 네이버 클로바 -> 1,000회 초과시 호출당 1원 -> 여유롭게 30,000개로 가정했을 때 30,000원

**구독자 수 추이**

* 플레이보드의 해당 요금제로는 1년치만 표로 불러올 수 있음.
* 전체 년도의 구독자 수 추이를 확인해야 할 필요가 있기에 social blade 사이트에서 최대 2021.02 ~ 2023.12 까지의 구독자 수 추이를 직접 입력하기로 결정 (~ 42번까지 완료: 은혜, ~ 84번까지 완료: 은선)

**썸네일 글자 위치**

* 상, 중, 하로 구분하여 썸네일에서 글자가 어디에 위치해 있는지 확인
* 좌, 우까지 고려하게 된다면 코딩하기가 어려워 이를 구분하지 않기로 결정

channel data: 채널명, 채널 고유 ID, 유튜브 링크

**2024.01.26**

**최종 데이터**

* **데이터: 13,097개**
* 제외
  + 6개월 이내에 업로드 된 영상( 수집일 2024.01.25 기준 180일 전: 2023.07.24 ~ ) (총 2,269개)
  + 조회수 추이를 확인할 수 없는 채널 ( 프리리뷰(2), 빵또와 영화채널(82)) (총 84개)
  + 영상 썸네일 이미지 404 Error (2개)

**총 데이터 개수: 10,742개**

**썸네일 및 제목 feature에 대한 논의**

1. 기계적 방법: 자언어처리 과정을 통해 임베딩(글자를 벡터로 변환) -> 한국어 자연어 처리 library 활용 (글자를 형태소로 나누어 벡터를 만들 수 있다)
2. 우리만의 기준을 정해 feature 값을 만들어서 원하는 코딩 (0, 1) : 근거 필요