이 파일은 이슈사항을 정리하는 파일입니다

이슈내용제목

- 이슈 내용: 이것은 이러이러한 내용으로 이슈가 있어서 이러이러한 방법으로 개선할 예정.

예시)
이미지 크롤링 자동화 부분

- 이미지 크롤링 자동화 부분이 매끄럽게 되지 못해서, 이후에 다른 api 등을 찾아서 추가적으로 시도해볼 예정.

이슈 1. Fine tuning의 어려움

- data set의 문제일 가능성 낮음
- 모델 간 성능 편차 확인
- 과대적합 문제 등

조건	learning_r	batch_size	weight_de	epoch		Train_loss	Val loss	Val accur
1	4.50E-03	8	0.1	1		0.6958	0.6937	0.4882
2	4.50E-04	8	0.1	2		0.6932	0.693	0.5118
3	4.50E-05	8	0.1	2		0.349	0.3853	0.8253
4	4.50E-06	8	0.1	1		0.3431	0.3766	0.8264
5	4.50E-03	16	0.1	3		0.693	0.6929	0.5118
6	4.50E-04	16	0.1	1		0.6954	0.6929	0.5118
7	4.50E-05	16	0.1	2		0.3219	0.3724	0.8241
8	4.50E-06	16	0.1	1		0.17	0.4604	0.8373
9	4.50E-03	32	0.1	4		0.6933	0.6929	0.811
10	4.50E-04	32	0.1	3		0.6941	0.693	0.811
11	4.50E-05	32	0.1	1		0.2105	0.4236	0.8294
12	4.50E-06	32	0.1	1		0.1791	0.4483	0.8339
13	4.50E-03	8	0.01	3		0.6933	0.6929	0.5118
14	4.50E-04	8	0.01	1		0.6951	0.6929	0.5118
15	4.50E-05	8	0.01	1		0.4923	0.401	0.8155
16	4.50E-06	8	0.01	1		0.6933	0.6929	0.5118
17	4.50E-03	16	0.01	4		0.6935	0.6929	0.5118
18	4.50E-04	16	0.01	2		0.6932	0.6929	0.5118
19	4.50E-05	16	0.01	1		0.2342	0.3899	0.8313
20	4.50E-06	16	0.01	1		0.1782	0.4581	0.8331
21	4.50E-03	32	0.01	2		0.6933	0.6929	0.5118
22	4.50E-04	32	0.01	3		0.6934	0.6959	0.5118
23	4.50E-05	32	0.01	2		0.3485	0.3544	0.8429
24	4.50E-06	32	0.01	4		0.376	0.3966	0.8245

1. 학습 초기(1~2 epoch)에 가장 낮은 val loss 값이 나오고 그 값 자체가 충분히 낮지 않음. 조건을 수정하여 재 학습이 필요

2. 3차 fine tuning 과정
   (train set : 36000/test set : 2667) 91 min. L40s 사용
   ASHAScheduler, AdamW, Token Max : 256
   Learning rate r : 5e-6 ~ 1e-3
   batch_size : 8, 16, 32
   epoch(max) : 5 + early stop; patient 2
   grad accumulate steps : 4
   Weight decay : 0.01 (constant)
   Result : lr : 2.37e-5, batch_size : 8, epoch : 5
   val loss = 0.386

이슈 사항 2. 장르 별로 감정 분류(긍정, 부정) 어떻게 할 것인지??

- 신파가 있는 드라마 장르의 경우 슬프다, (주인공의 이야기가)안타깝다 와 같은 리뷰를 부정적으로 해석할 가능성.

- 평점을 네이버평점(GT)과 함께 plot 후 특이점에 대해 경향 파악(장르…) 하여 미세조정 예정

이슈 사항 3. 영화 포스터 크롤링에서 부정확한 이미지를 가져오는 경우 발생

- 게시물로 사용하기에 저해상도 이미지 -> 해상도를 확인하여 저장여부를 결정하도록 수정
- 키노라이츠에서 이미지 크롤링시, 영화와 드라마의 구분이 없어 영화가 아닌 드라마 포스터를 가져오는 경우 발생.
- 네이버에서 이미지 크롤링시, 포스터와 스틸컷 구분이 필요하여, easyocr을 통해 문자가 있는 이미지를 포스터로 구분하였으나, 일부 배경 또는 의상에 문자를 인식하여 스틸컷이 포스터로 저장되는 경우 발생.

이슈 사항 4. 넷플릭스 게시물 한줄 리뷰 생성

- 전반적으로 넷플릭스의 리뷰가 적어, 다른 영화지만 비슷한 한줄 리뷰가 생성되는 경우가 발생 -> 리뷰에 소개글을 추가하여, 영화 내용이 일부 들어가는 리뷰로 개선

이슈 사항 5. 양산 모델 문제

- 대문 게시물 글을 3줄이 아닌 경우로 반환 / 프롬프트 혹은 모델 변경
- 추천 영화 제목을 제목이 아닌 dict형식으로 출력하는 경우 발생
- 게시글을 json 형식이 아닌 경우로 출력하는 경우 발생.

이슈 사항 6. 크롤링 데이터 저작권

- 현재 네이버, 키노라이츠에서 실시
- 향후 KMDB로 변경(opensource data)

시장 분석

- 인플루언서 :
  인스타그램을 통해 영화, 드라마 시리즈를 소개하는 인플루언서. OTT 서비스에 접속한 뒤 목록을 뒤적이다 입맛에 맞는 영화를 찾지 못해 결국 한편도 시작하지 못한 현대인들의 속을 시원하게 해줄 알잘딱깔쎈 영화 떠먹이기 “니, 무비 무봤나!?”

한계 극복

- 업데이트 되는 데이터에 대해 독립성이 필요한 경우:
  영화 순위, 평점 환산방식의 가중치, ...
- 기존 인플루언서와의 차별점(시장 가치)
- 개봉 예정영화의 경우 서비스 수요 및 가치가 높은데 비해 데이터확보가 비교적 어려움