Term Project

빅데이터 01분반

Team name:	넵네	
Leader:	20150711	송희성
Team Members:	20165968	맹채정
	20164980	유경민
	20173954	유 진
	20176815	이현지

Topic

영화별 review를 보고 해당 영화의 장르를 예측해보기

- review속의 keyword를 추출 -

Target Data: **imdb** 영화 리뷰 데이터

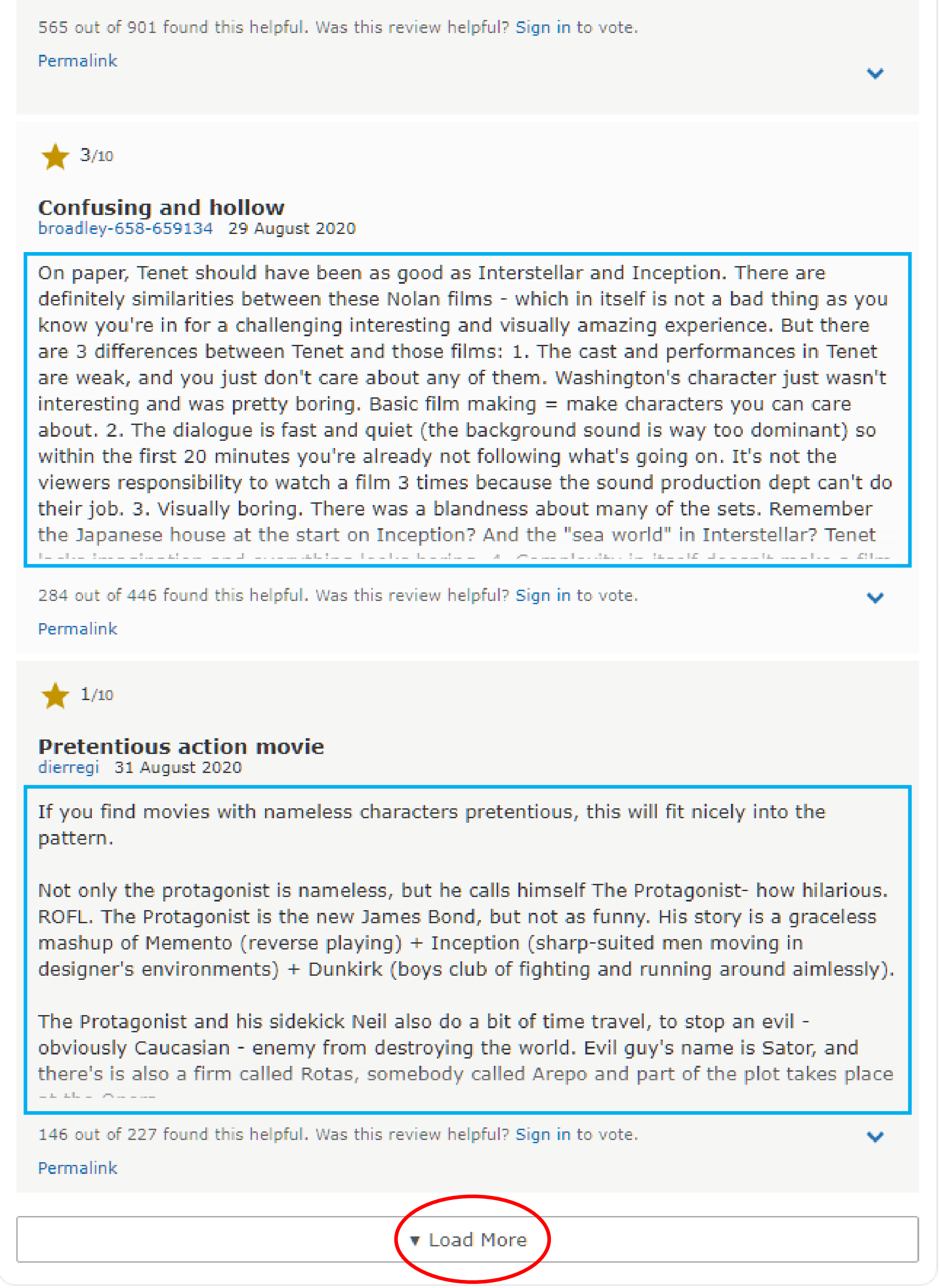
Github: https://github.com/yukyeongmin/NenepBigData

1
2
3
4
5
6

Data Scraping

imdb의 영화 리뷰 데이터:
Load More 버튼을 누르면 숨겨진 리뷰가 보이는 구조

->python selenium & bs4 이용
->스크롤을 끝까지 내리고 Load More버튼 누르기를 반복
->드러난 모든 리뷰 데이터를 scraping
->csv파일로 만들기



565 out of 901 found this helpful. Was this review helpful? Sign in to vote.
Permalink

★ 3/10

Confusing and hollow
broadley-658-659134 29 August 2020

On paper, Tenet should have been as good as Interstellar and Inception. There are definitely similarities between these Nolan films - which in itself is not a bad thing as you know you're in for a challenging interesting and visually amazing experience. But there are 3 differences between Tenet and those films: 1. The cast and performances in Tenet are weak, and you just don't care about any of them. Washington's character just wasn't interesting and was pretty boring. Basic film making = make characters you can care about. 2. The dialogue is fast and quiet (the background sound is way too dominant) so within the first 20 minutes you're already not following what's going on. It's not the viewers responsibility to watch a film 3 times because the sound production dept can't do their job. 3. Visually boring. There was a blandness about many of the sets. Remember the Japanese house at the start on Inception? And the "sea world" in Interstellar? Tenet lacks imagination and everything looks boring. 4. Complexity in itself doesn't make a film

284 out of 446 found this helpful. Was this review helpful? Sign in to vote.
Permalink

★ 1/10

Pretentious action movie
dierregi 31 August 2020

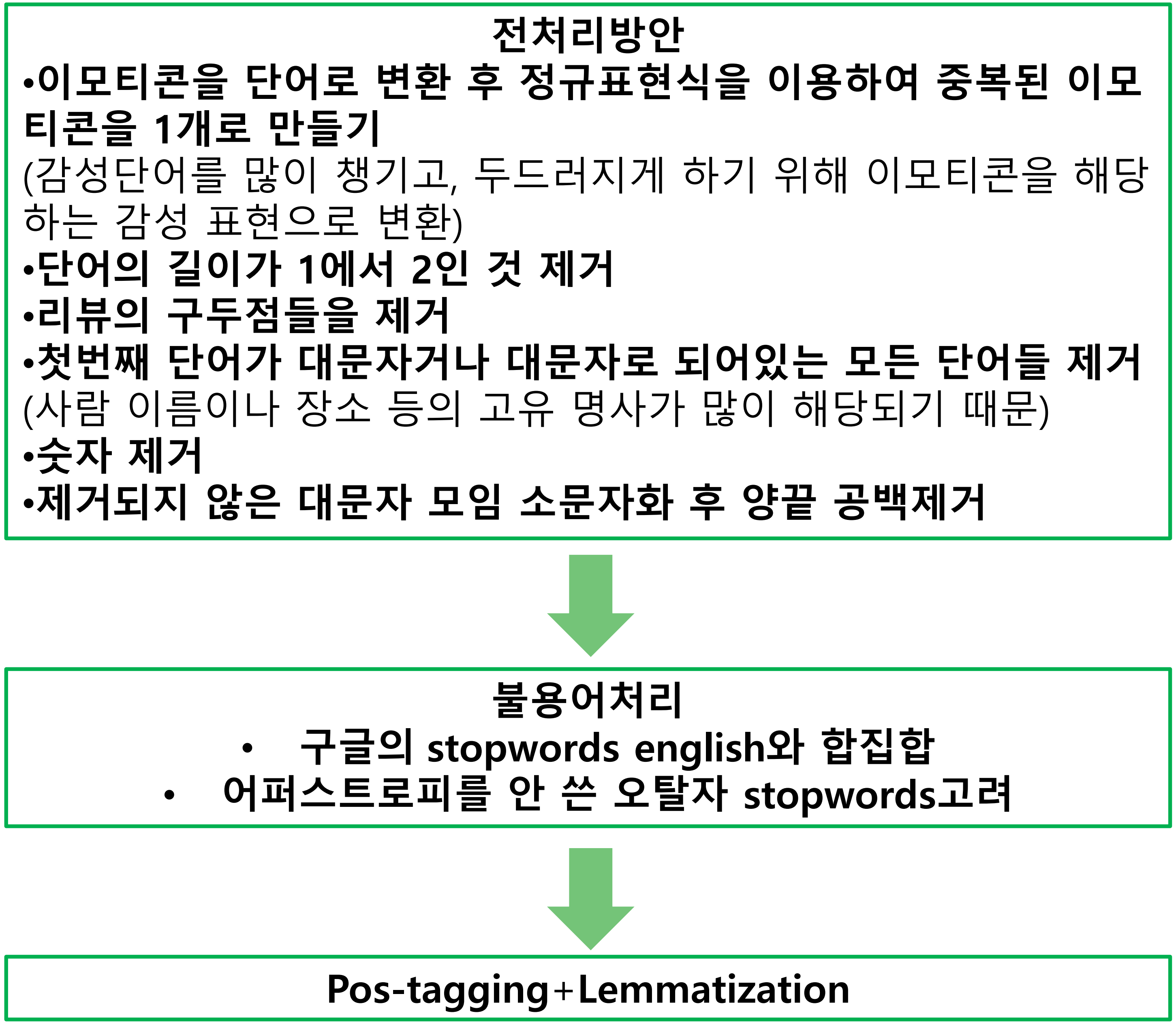
If you find movies with nameless characters pretentious, this will fit nicely into the pattern.

Not only the protagonist is nameless, but he calls himself The Protagonist- how hilarious. ROFL. The Protagonist is the new James Bond, but not as funny. His story is a graceless mashup of Memento (reverse playing) + Inception (sharp-suited men moving in designer's environments) + Dunkirk (boys club of fighting and running around aimlessly).

The Protagonist and his sidekick Neil also do a bit of time travel, to stop an evil - obviously Caucasian - enemy from destroying the world. Evil guy's name is Sator, and there's is also a firm called Rotas, somebody called Arepo and part of the plot takes place at the Opera

146 out of 227 found this helpful. Was this review helpful? Sign in to vote.
Permalink

1
2
3 Data Preprocessing
4
5
6



전처리방안

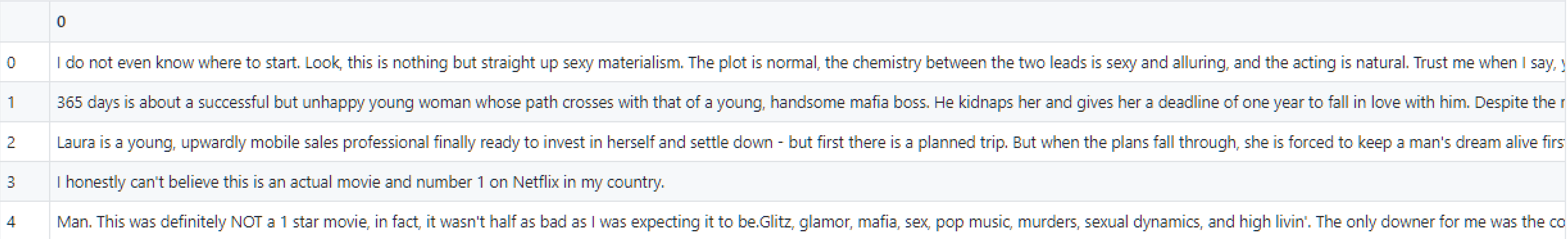
- **이모티콘을 단어로 변환 후 정규표현식을 이용하여 중복된 이모티콘을 1개로 만들기**
(감성단어를 많이 챙기고, 두드러지게 하기 위해 이모티콘을 해당하는 감성 표현으로 변환)
- **단어의 길이가 1에서 2인 것 제거**
- **리뷰의 구두점들을 제거**
- **첫번째 단어가 대문자거나 대문자로 되어있는 모든 단어들 제거**
(사람 이름이나 장소 등의 고유 명사가 많이 해당되기 때문)
- **숫자 제거**
- **제거되지 않은 대문자 모임 소문자화 후 양끝 공백제거**

불용어처리

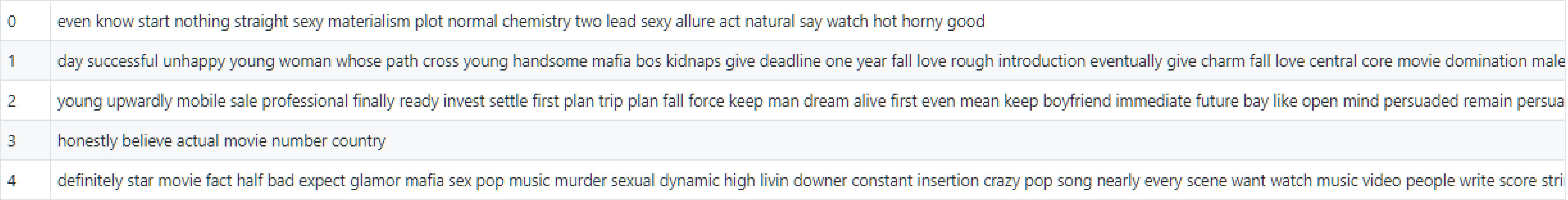
- **구글의 stopwords english와 합집합**
- **어퍼스트로피를 안 쓴 오탈자 stopwords고려**

Pos-tagging+Lemmatization

Token화 된 Data



	0
0	I do not even know where to start. Look, this is nothing but straight up sexy materialism. The plot is normal, the chemistry between the two leads is sexy and alluring, and the acting is natural. Trust me when I say, y
1	365 days is about a successful but unhappy young woman whose path crosses with that of a young, handsome mafia boss. He kidnaps her and gives her a deadline of one year to fall in love with him. Despite the r
2	Laura is a young, upwardly mobile sales professional finally ready to invest in herself and settle down - but first there is a planned trip. But when the plans fall through, she is forced to keep a man's dream alive firs
3	I honestly can't believe this is an actual movie and number 1 on Netflix in my country.
4	Man. This was definitely NOT a 1 star movie, in fact, it wasn't half as bad as I was expecting it to be.Glitz, glamor, mafia, sex, pop music, murders, sexual dynamics, and high livin'. The only downer for me was the co

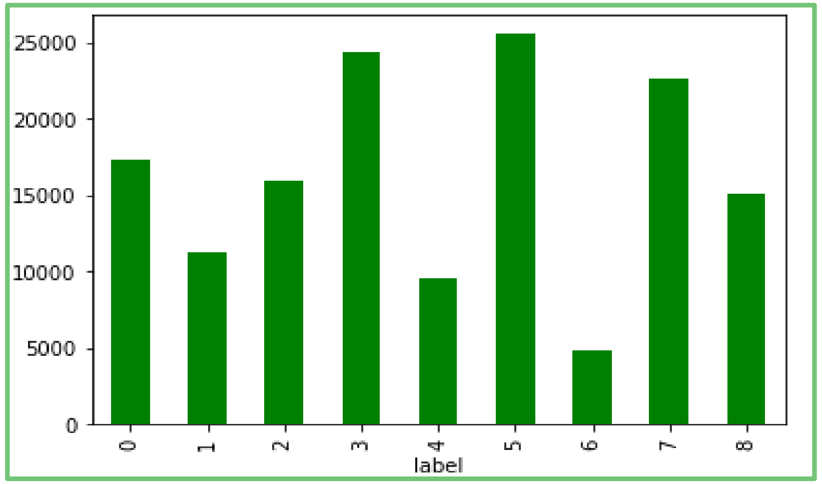


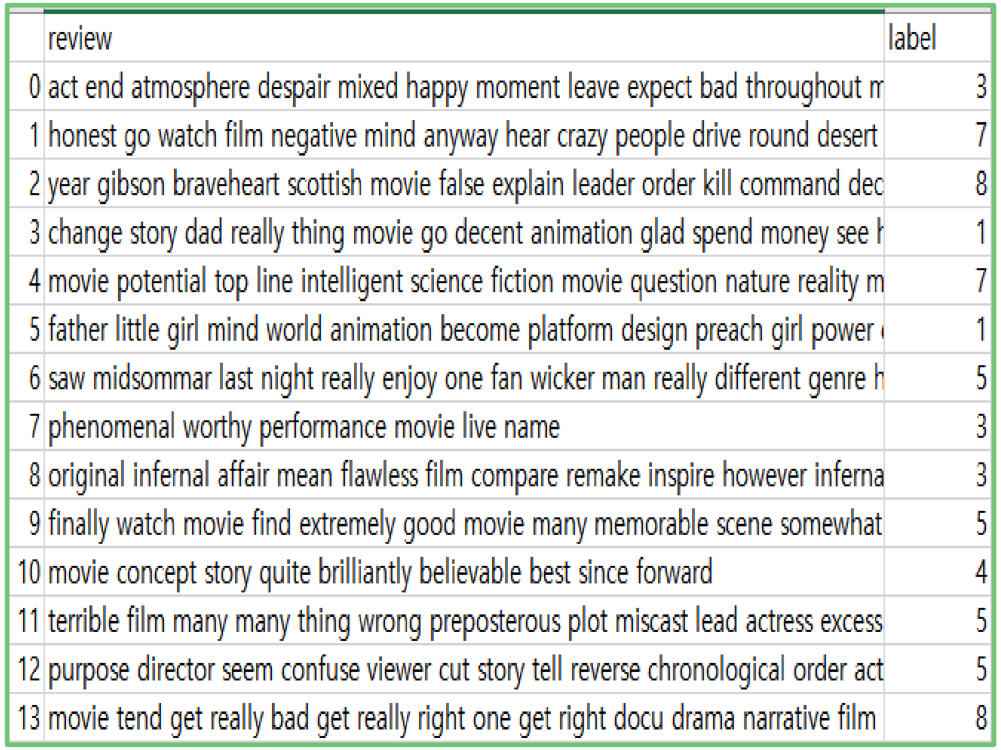
0	even know start nothing straight sexy materialism plot normal chemistry two lead sexy allure act natural say watch hot horny good
1	day successful unhappy young woman whose path cross young handsome mafia bos kidnaps give deadline one year fall love rough introduction eventually give charm fall love central core movie domination male
2	young upwardly mobile sale professional finally ready invest settle first plan trip plan fall force keep man dream alive first even mean keep boyfriend immediate future bay like open mind persuaded remain persua
3	honestly believe actual movie number country
4	definitely star movie fact half bad expect glamor mafia sex pop music murder sexual dynamic high livin downer constant insertion crazy pop song nearly every scene want watch music video people write score stri

1
2
3 Data Preprocessing
4
5
6

(Data,label)형태로 변환

장르Label
0:actionadventure (17369)
1:animation(11303)
2:biography(15984)
3:crimeaction(24327)
4:horror(9552)
5:mysterythriller(25553)
6:romanticcomedy(4821)
7:scifi(22674)
8:war(15117)





	review	label
0	act end atmosphere despair mixed happy moment leave expect bad throughout m	3
1	honest go watch film negative mind anyway hear crazy people drive round desert	7
2	year gibson braveheart scottish movie false explain leader order kill command dec	8
3	change story dad really thing movie go decent animation glad spend money see h	1
4	movie potential top line intelligent science fiction movie question nature reality m	7
5	father little girl mind world animation become platform design preach girl power	1
6	saw midsommar last night really enjoy one fan wicker man really different genre h	5
7	phenomenal worthy performance movie live name	3
8	original infernal affair mean flawless film compare remake inspire however inferna	3
9	finally watch movie find extremely good movie many memorable scene somewhat	5
10	movie concept story quite brilliantly believable best since forward	4
11	terrible film many many thing wrong preposterous plot miscast lead actress excess	5
12	purpose director seem confuse viewer cut story tell reverse chronological order act	5
13	movie tend get really bad get really right one get right docu drama narrative film	8

1

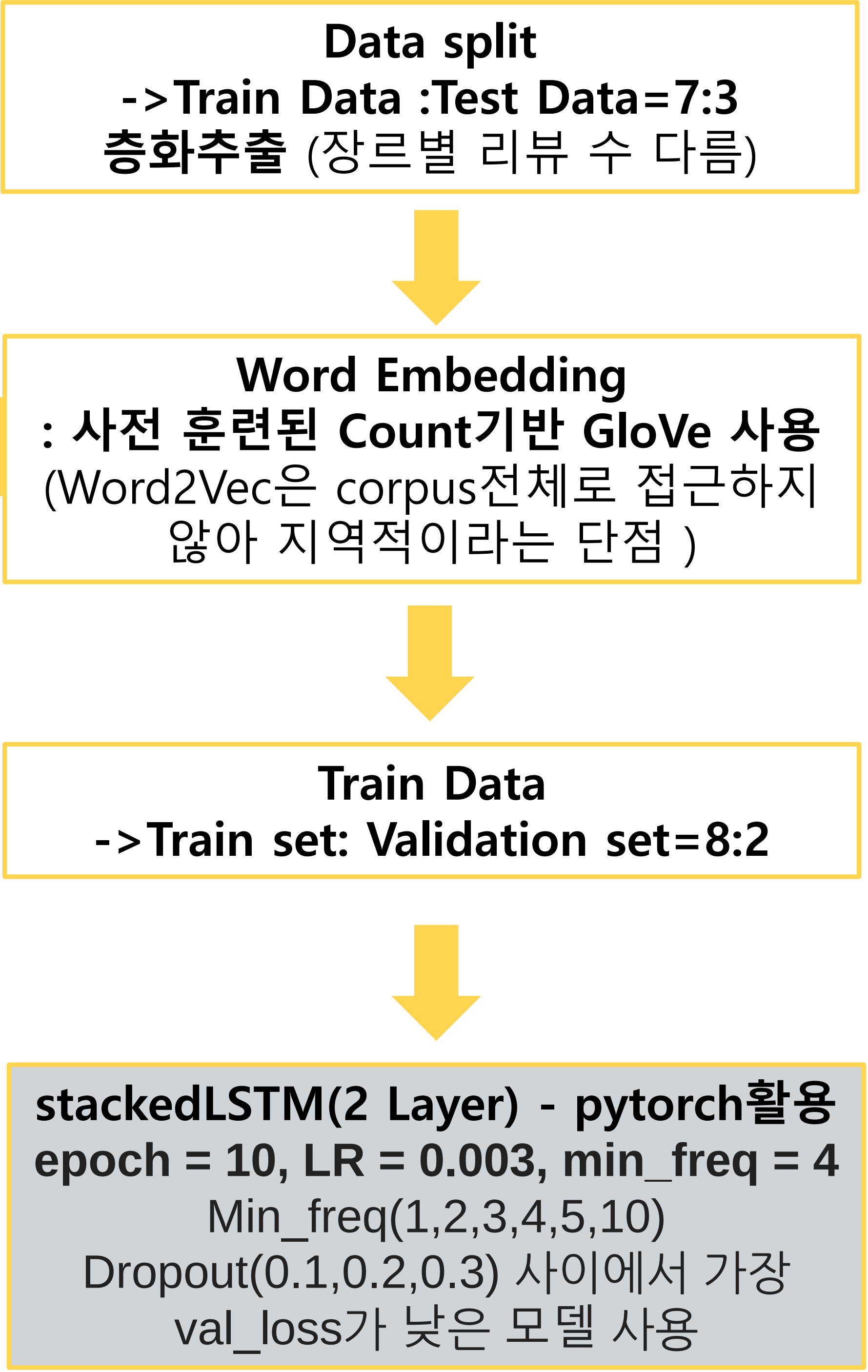
2

3

4 Model

5

6



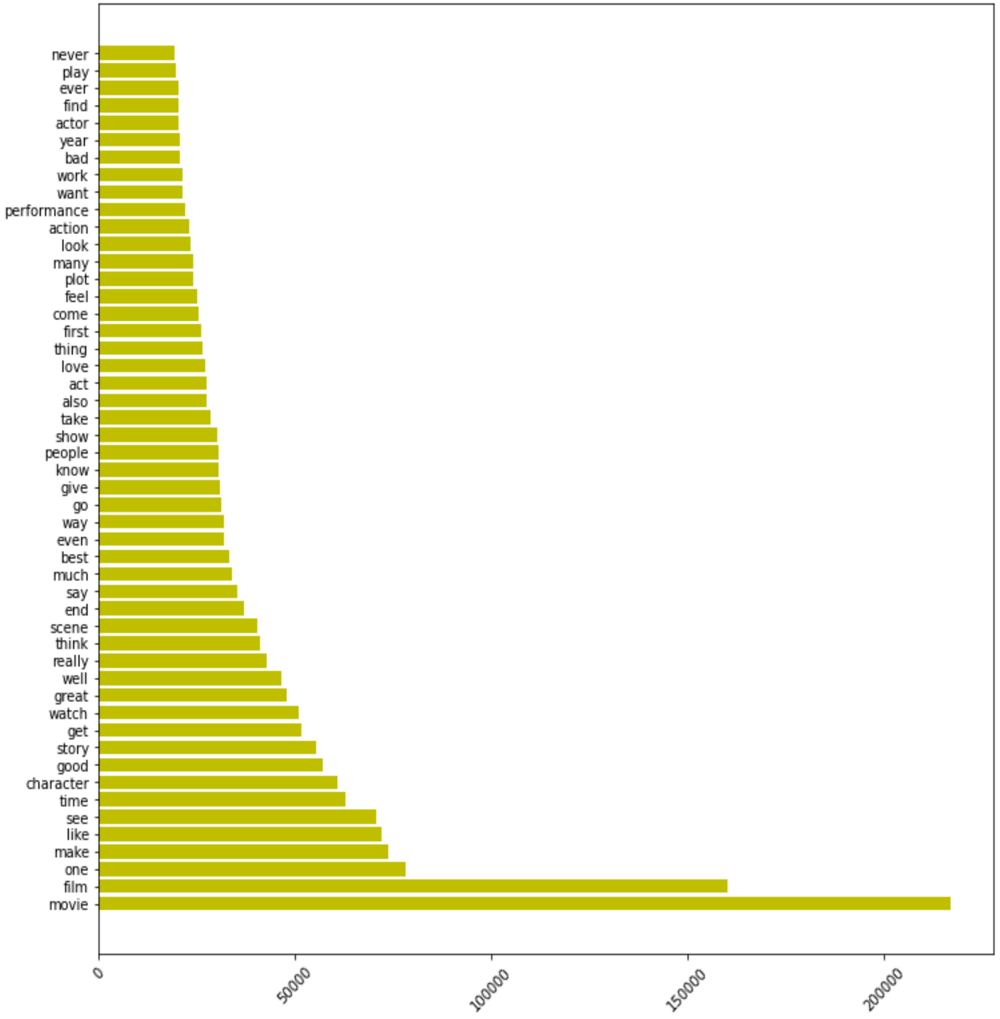
Data split
->Train Data :Test Data=7:3
층화추출 (장르별 리뷰 수 다름)

Word Embedding
: 사전 훈련된 Count기반 GloVe 사용
(Word2Vec은 corpus전체로 접근하지 않아 지역적이라는 단점)

Train Data
->Train set: Validation set=8:2

stackedLSTM(2 Layer) - pytorch활용
epoch = 10, LR = 0.003, min_freq = 4
Min_freq(1,2,3,4,5,10)
Dropout(0.1,0.2,0.3) 사이에서 가장 val_loss가 낮은 모델 사용

자주 나온 상위 50개의 단어들



1

2

3

4

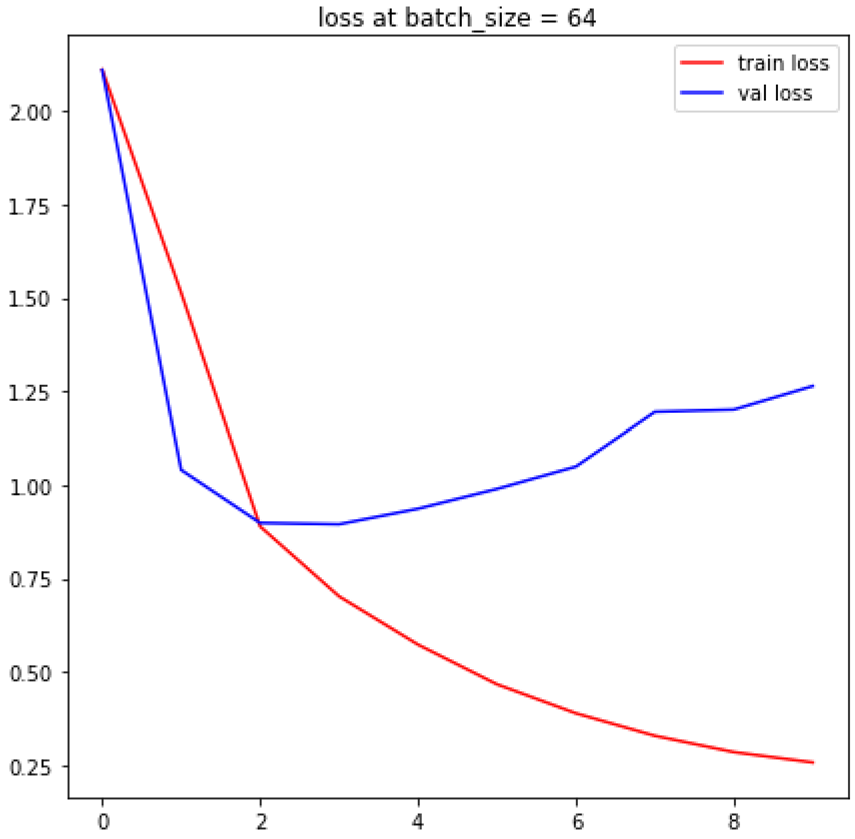
5 Measurement

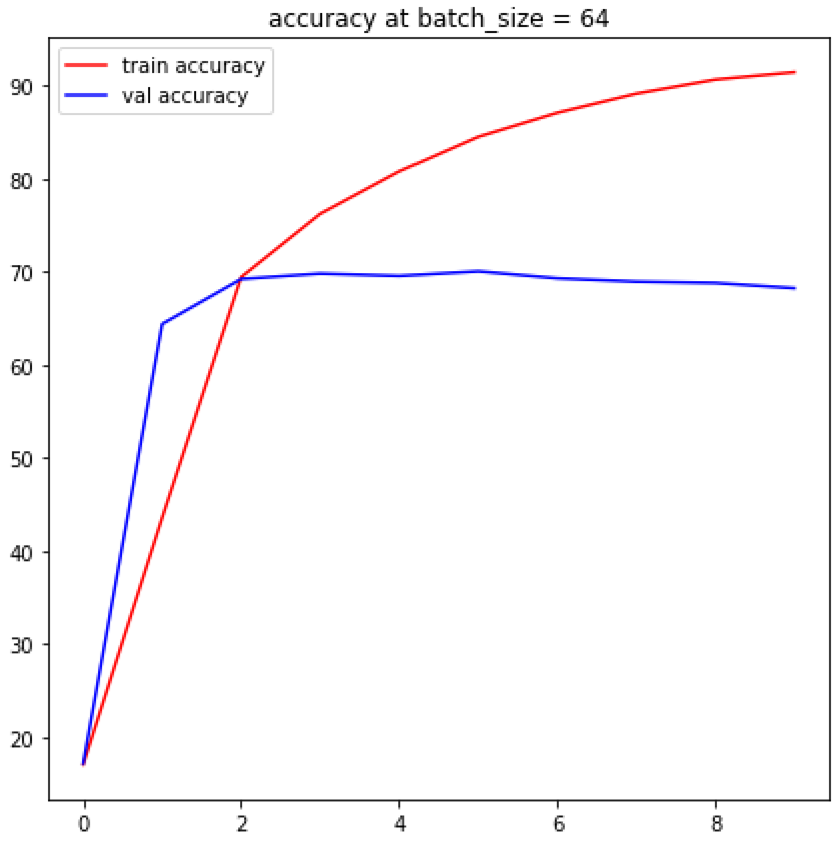
6

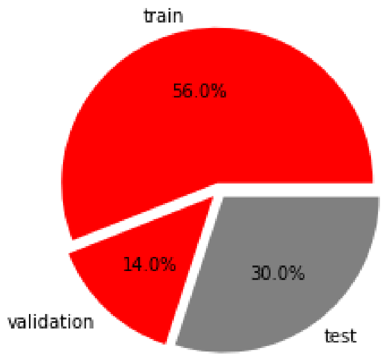
학습결과

Loss curve

Accuracy curve







1

2

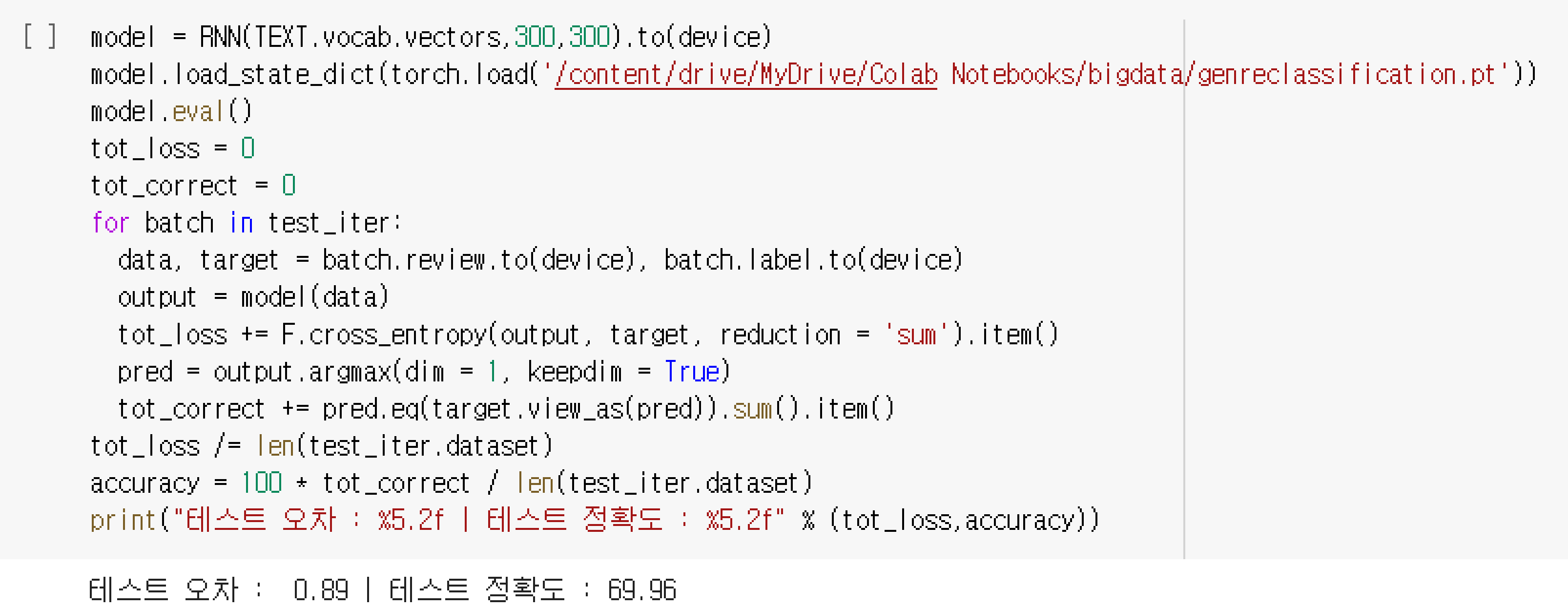
3

4

5 Measurement

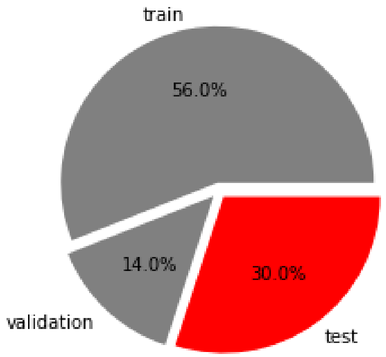
6

모델 테스트 결과 약 70% 정확도 확보



```
[ ]  model = RNN(TEXT.vocab.vectors,300,300).to(device)
     model.load_state_dict(torch.load('/content/drive/MyDrive/Colab Notebooks/bigdata/genreclassification.pt'))
     model.eval()
     tot_loss = 0
     tot_correct = 0
     for batch in test_iter:
       data, target = batch.review.to(device), batch.label.to(device)
       output = model(data)
       tot_loss += F.cross_entropy(output, target, reduction = 'sum').item()
       pred = output.argmax(dim = 1, keepdim = True)
       tot_correct += pred.eq(target.view_as(pred)).sum().item()
     tot_loss /= len(test_iter.dataset)
     accuracy = 100 * tot_correct / len(test_iter.dataset)
     print("테스트 오차 : %5.2f | 테스트 정확도 : %5.2f" % (tot_loss,accuracy))

     테스트 오차 :  0.89 | 테스트 정확도 : 69.96
```



1

2

3

4

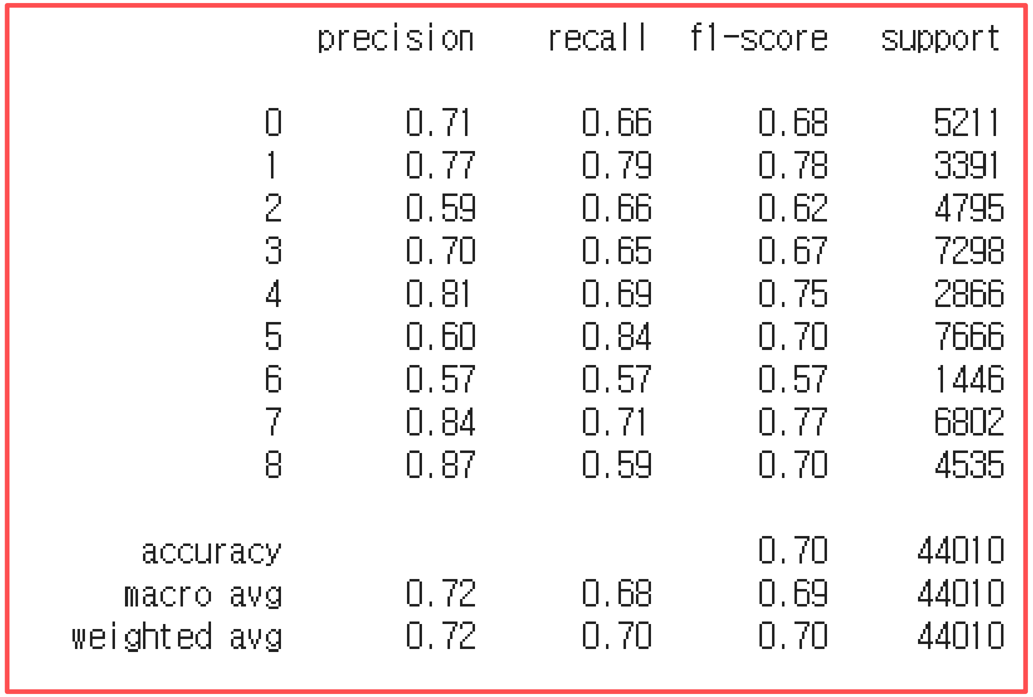
5 Measurement

6

MultiLabel Confusion Matrix

- 전반적으로 precision과 recall이 어느 한 쪽에 치우치지 않고 대략 70%의 확률로 잘 분포 됨.
- label 수 대비 성능이 괜찮음
- 2번, 5번, 6번 장르의 정확도가 상대적으로 낮음.

Word-embedding 알고리즘 BERT사용시 성능 향상 기대



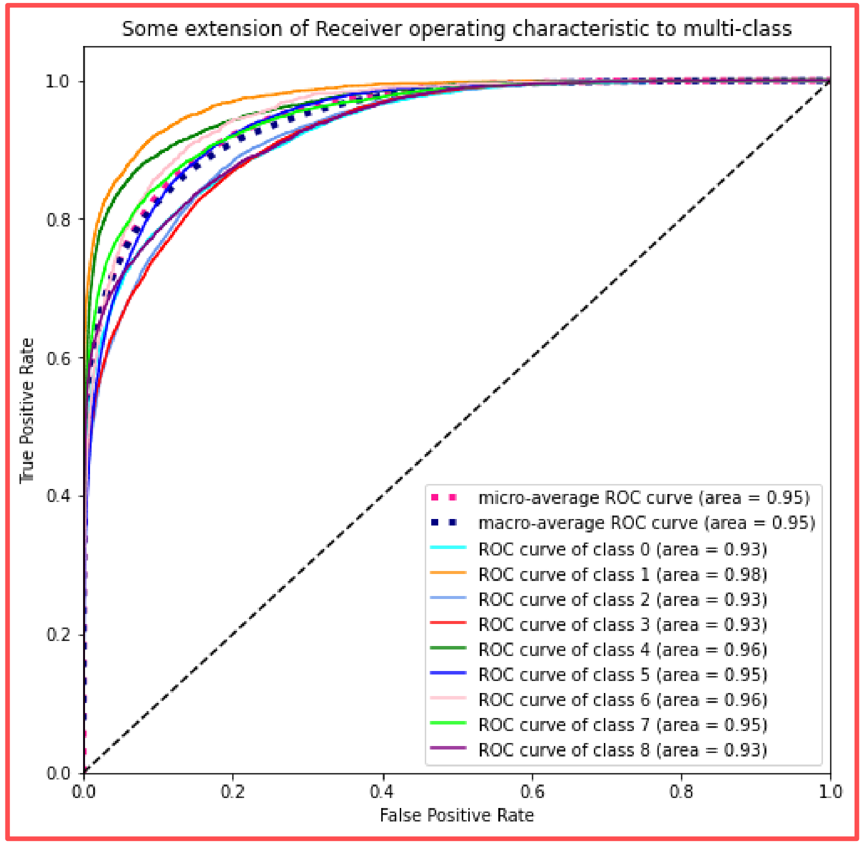
```
              precision    recall  f1-score   support

           0       0.71      0.66      0.68      5211
           1       0.77      0.79      0.78      3391
           2       0.59      0.66      0.62      4795
           3       0.70      0.65      0.67      7298
           4       0.81      0.69      0.75      2866
           5       0.60      0.84      0.70      7666
           6       0.57      0.57      0.57      1446
           7       0.84      0.71      0.77      6802
           8       0.87      0.59      0.70      4535

    accuracy                           0.70     44010
   macro avg       0.72      0.68      0.69     44010
weighted avg       0.72      0.70      0.70     44010
```

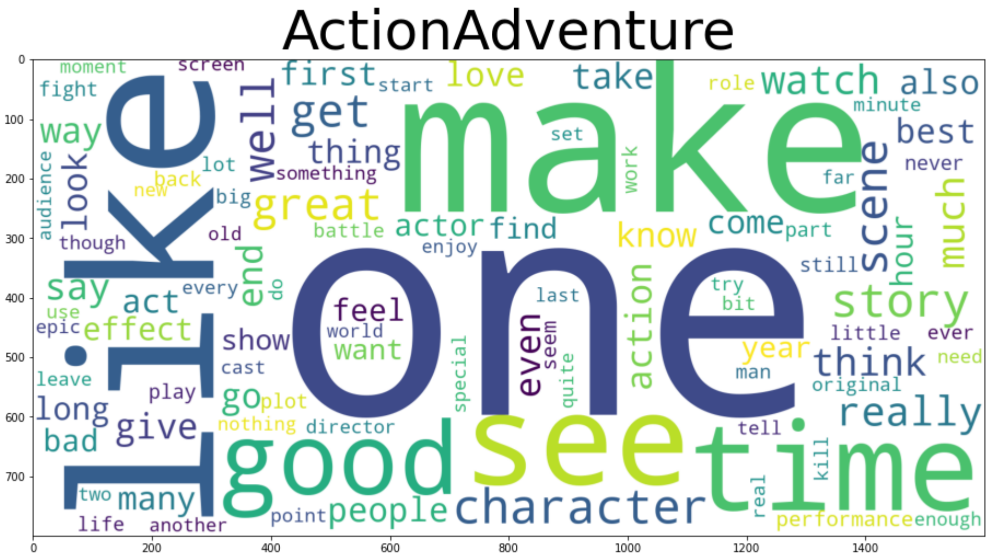
MultiClass ROC Curve

모델에서 나온 Outcome을 Softmax 하여 합을 1로 만든 뒤 ROC curve를 만듦
각 영화 장르별로 해당 장르, 나머지 장르의 이진 관점에서 보았을 때,
각 장르별 AUC가 약 95% 정도로 예측력이 정말 좋은 모델이라는 것을 알 수 있다.

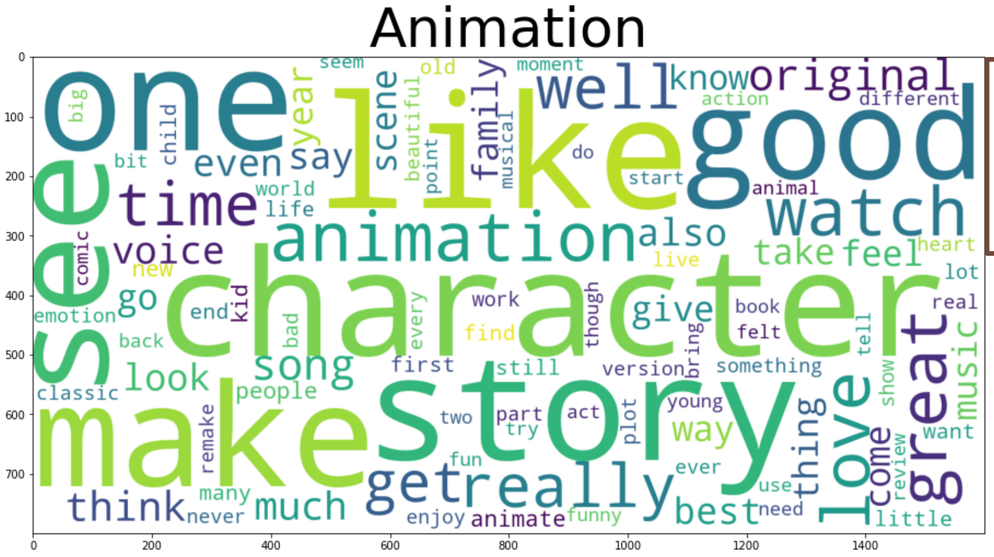


1
2
3
4
5
6 Visualization

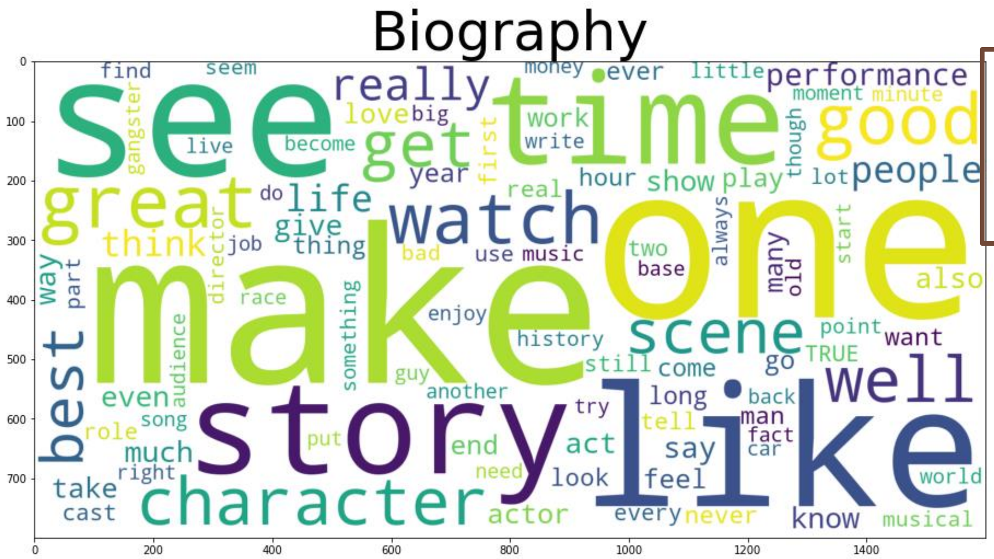
각 장르별 많이 나온 단어들


ActionAdventure

act
action
fight


Animation

animate
music
child


Biography

character
history
fact

1

2

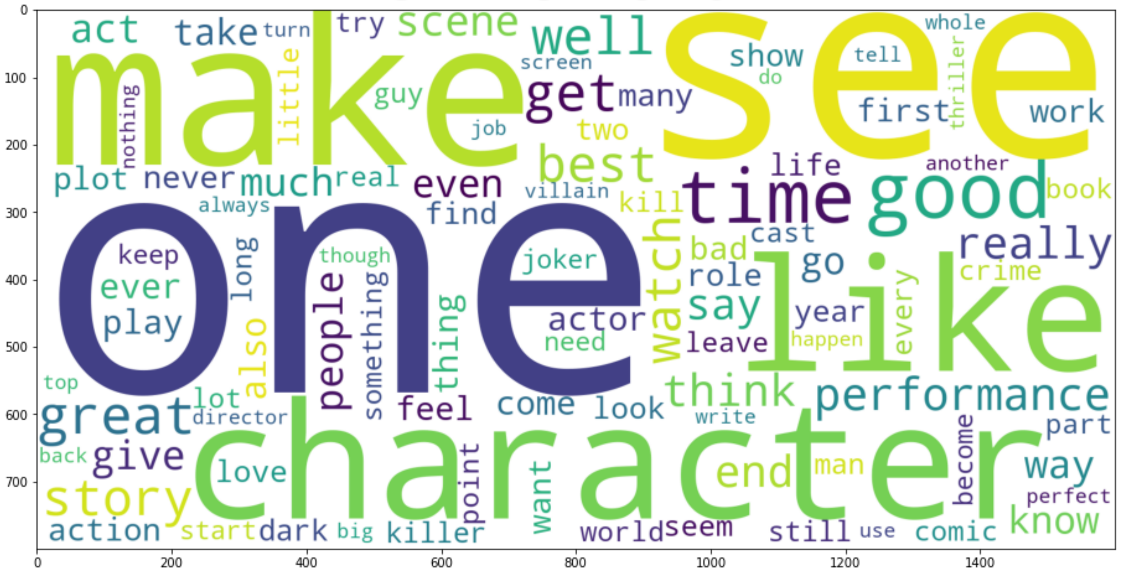
3

4

5

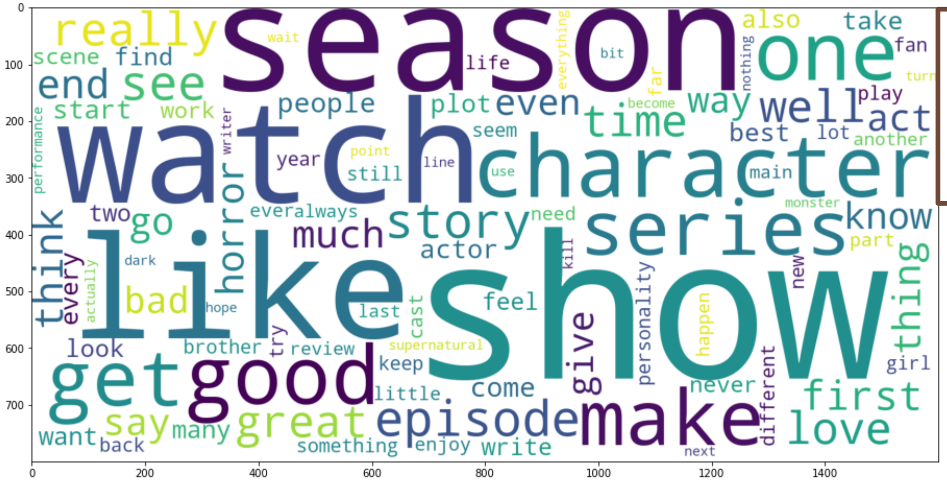
6 Visualization

Crime Action



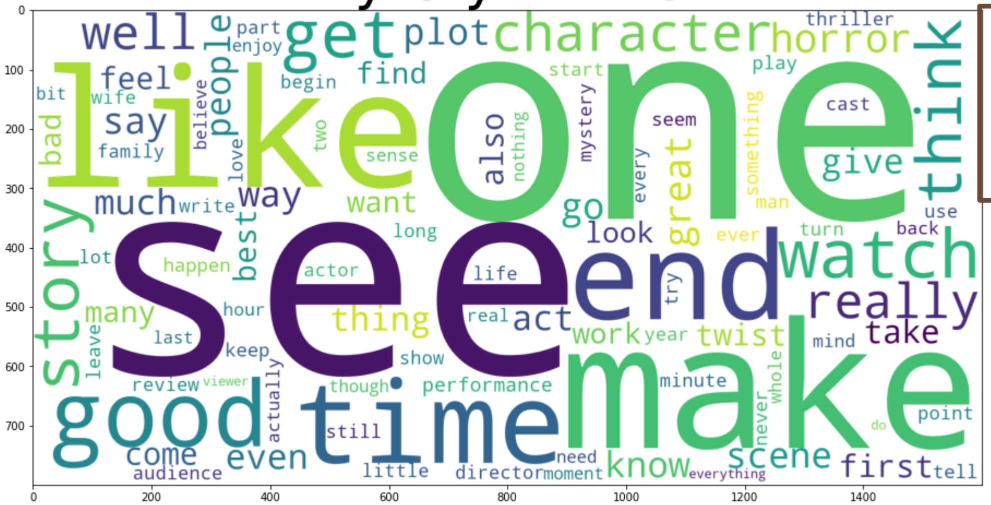
bad
killer
action

Horror



kill
horror
monster

Mytery Thriller



thriller
mystery
bad

1

2

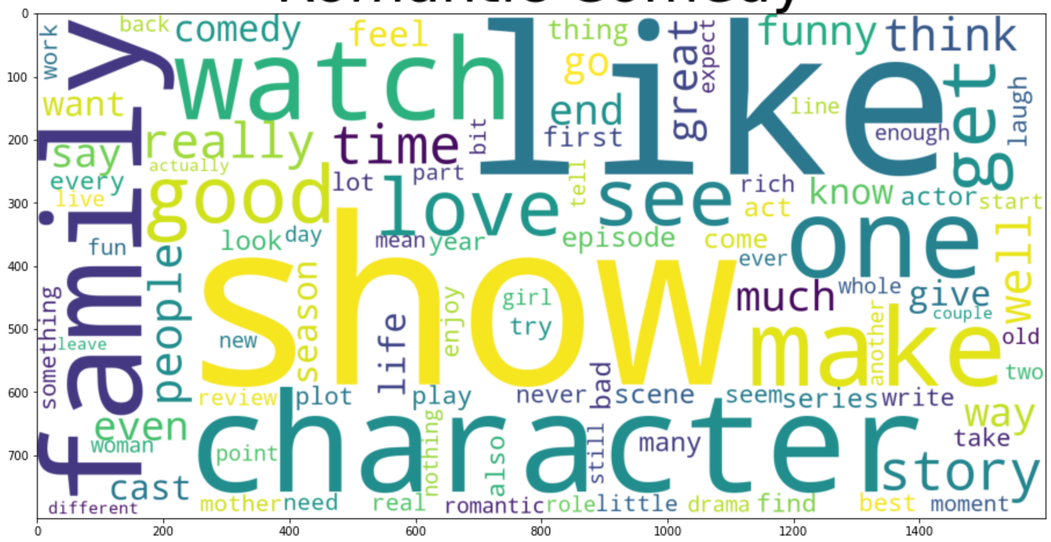
3

4

5

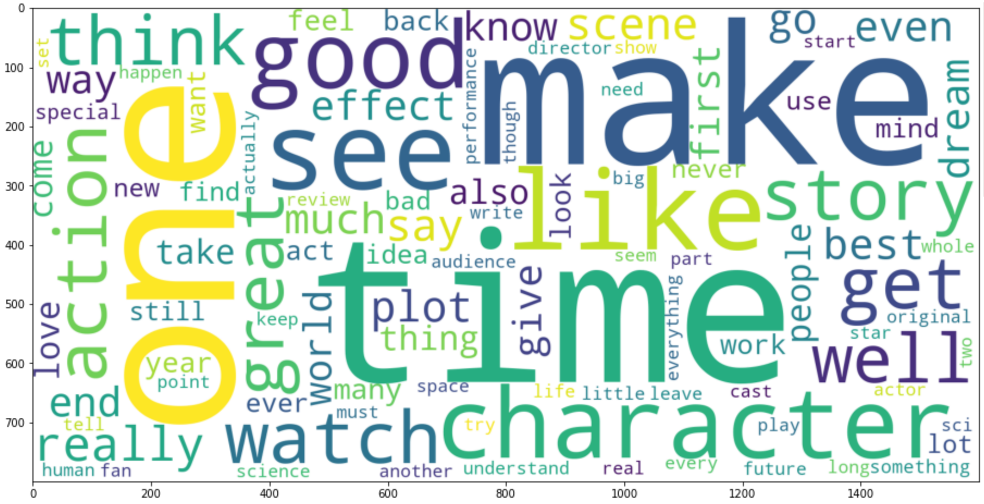
6 Visualization

Romantic Comedy



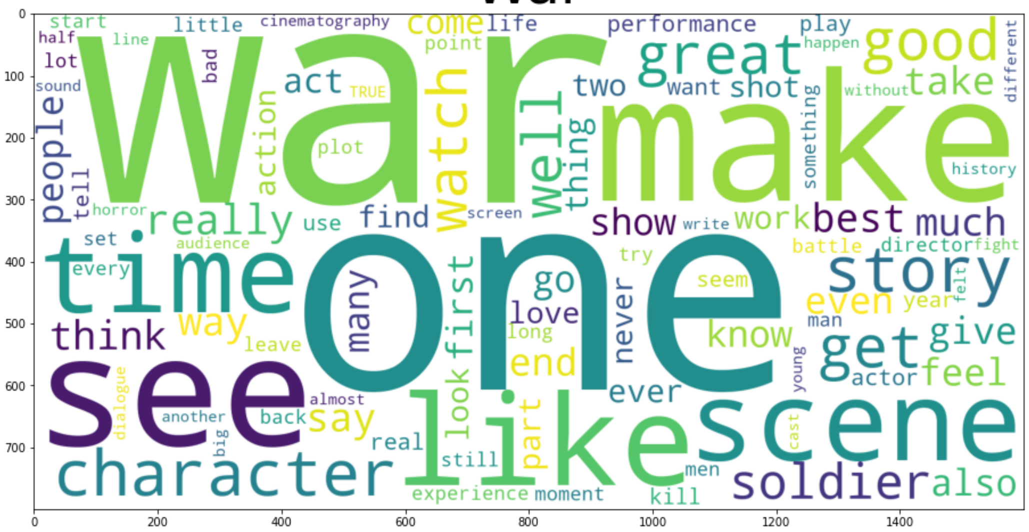
laugh
romantic
comedy

Sci - Fi



science
idea
sci

War



war
soldier
history

감사합니다