



저평점작 요인 분석

네이버 영화리뷰를 바탕으로
저평점, 흥행 실패작 분석

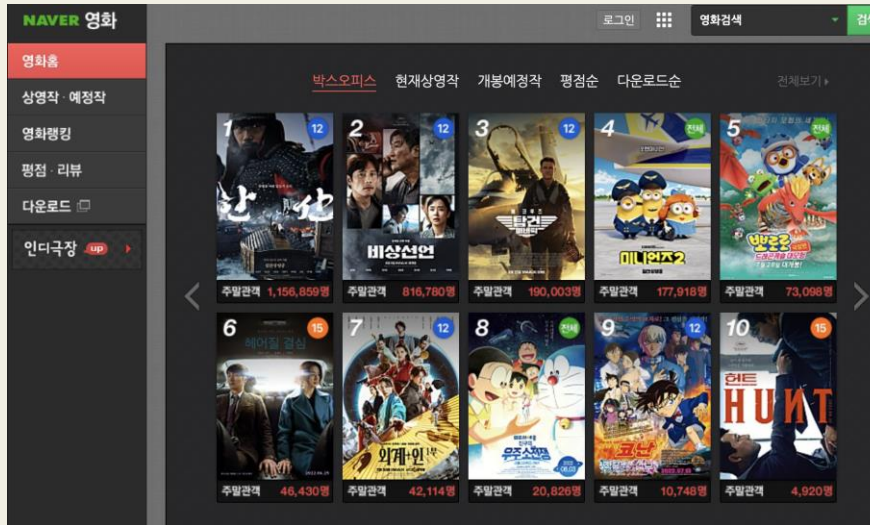


주의. 스포가 포함되어 있을 수 있습니다.

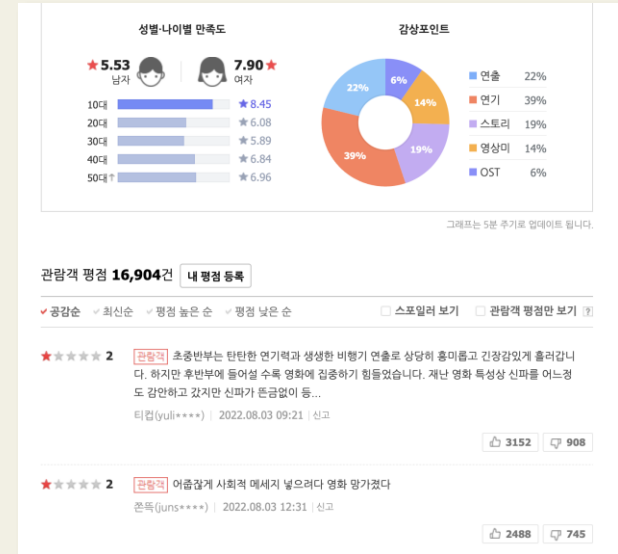


15기 고태영, 조우영, 정영희
16기 박민규

네이버 영화리뷰를 보자!



Data Source: 네이버 영화 리뷰 크롤링,
관람객 평점, 리뷰, 공감 & 비공감 수 사용



```
# 관람객 평점
review_df = pd.DataFrame(columns=['평점', '리뷰', '공감', '비공감'])

for j in range(10):
    for i in range(10):
        grade = driver.find_element("xpath", f'/html/body/div/div/div[5]/ul/li[{i+1}]/div[1]/em').text
        review = driver.find_element("xpath", f'//*[@id="_filtered_ment_{i}"]').text
        good = driver.find_element("xpath", f'/html/body/div/div/div[5]/ul/li[{i+1}]/div[3]/a[1]/strong').text
        bad = driver.find_element("xpath", f'/html/body/div/div/div[5]/ul/li[{i+1}]/div[3]/a[2]/strong').text
        review_df.loc[10*j+i] = [grade, review, good, bad]
        driver.find_element("xpath", f'//*[@id="pagerTagAnchor{j+1}"]').click()

review_df
```

Selenium 활용!



누적관객 2,041,377(08.27 기준)

- ★★★★★ 2 **[관람객]** 초중반부는 탄탄한 연기력과 생생한 비행기 연출로 상당히 흥미롭고 긴장감있게 흘러갑니다. 하지만 후반부에 들어설 수록 영화에 집중하기 힘들었습니다. 재난 영화 특성상 신파를 어느정도 감안하고 갔지만 신파가 뜬금없이 등...
- 티컬(yuli****) | 2022.08.03 09:21 | 신고
- 👍 3151 🗨 908
-
- ★★★★★ 2 **[관람객]** 어름잡게 사회적 메시지 넣으려다 영화 망가졌다
- 존득(juns****) | 2022.08.03 12:31 | 신고
- 👍 2487 🗨 745
-
- ★★★★★ 2 **[관람객]** 잘가다가 무리한 반미, 반일과 억지스러운 신파, 무리한 설정이 영화를 산으로 보냅니다. 보면서 계속 어이없는 웃음이 나와요. 진짜 억지 눈물 짜낼때는 하.... 배우들이 너무 아깝네요
- 궁(hyun****) | 2022.08.03 15:17 | 신고
- 👍 2175 🗨 651
-
- ★★★★★ 1 **[관람객]** 이 영화가 칸으로 갔다는게 부끄럽다
- 준혁(wnsg****) | 2022.08.03 17:12 | 신고
- 👍 1910 🗨 583

- ★★★★★ 2 **[관람객]** 본인 돈은 잘 챙깁시다. 다시 생각해 볼 시간은 많습니다.
- 동원참치(tlaw****) | 2022.07.20 12:19 | 신고
- 👍 4301 🗨 2362
-
- ★★★★★ 2 **[관람객]** K울트론 + 진격의 거인 + 속수무 + 원피스 고무고무 환장의 콜라보
- 김달밥(dong****) | 2022.07.20 16:01 | 신고
- 👍 2710 🗨 1451
-
- ★★★★★ 1 **[관람객]** 니들이 어떤걸 좋아할지 몰라서 다 쓰까봤어.. 이건가요??
- 산뜻한저녁(psgp****) | 2022.07.21 00:11 | 신고
- 👍 1938 🗨 969
-
- ★★★★★ 2 **[관람객]** 내가 본 관에선 다 안웃던데... 이게 진짜 재밌음..??
- zzz(qluv****) | 2022.07.20 17:02 | 신고
- 👍 2158 🗨 1258



누적관객 1,533,269(08.27 기준)

? 리뷰가 좋지 않거나,
평점이 낮은 이유는 무엇일까? ?

전처리 과정

전처리

1. 중복값 제거, 결측치 제거
2. 불용어 제거
3. 비공감 > 공감이면 제거



불용어:

조사, 동사 맺음말 + **영화 제목**

```
stopwords = [x.strip() for x in stopwords]  
stopwords.extend(['하다', '이다', '이에요', '은', '한', '거', '던', '라', '고', '비상', '선언', '영화'])  
stopwords.extend("아 휴 아이구 아이쿠 아이고 어 나 우리 저희 따라 의해 을 를 에 의 가 으로 로 에게")
```

- » 영화 제목을 리뷰에서 언급해, 큰 의미를 주지 않으므로 불용어로 설정
- » 비공감이 공감보다 높을 경우, 대중적인 의견이라고 판단하기 어렵기에 결측치로 처리해 제거

전처리 과정

전처리

4. 정규 표현식으로 **한글**만 남기기
5. 형태소 태깅 & 원형으로 보정
6. 띄어쓰기 교정
with PyKoSpacing



```
def preprocessing(review, okt, remove_stopwords=False, stop_words=[]):
    review = spacing(review)
    review_text = re.sub('[^ㄱ-ㅣ가-힣]+', '', review)
    # word_tokenize
    word_review = okt.morphs(review_text, stem=True)
    word_review_noun = ' '.join([word[0] for word in okt.pos(review_text) if word[1] == 'Noun'])
    # stop_words
    if remove_stopwords:
        word_review = [token for token in word_review if not token in stop_words]
    return word_review, word_review_noun
```

» Okt 토큰화 결과

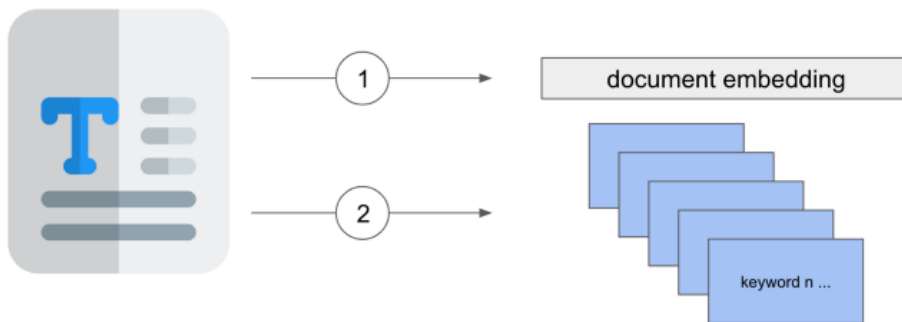
리뷰	공 감	비 공 감	전처리1
고종 없는 막장 영화. 이걸 정말 아 닌듯	0	0	[고종, 없다, 막장, 이걸, 정말, 아니다]
아직도 이런 억지스러운 반일 반 미 감정만 부추기는 영화들이 개 봉하는게... 답없다	0	0	[아직도, 이렇다, 억지스럽다, 반일, 반미, 감정, 만, 부추기다, 개봉, 답, 없다]
감독님이 공부안하고 찍음ㅋㅋㅋ	0	0	[감독, 님, 공, 부안, 하고, 찍다, ㅋㅋㅋ]

전처리 과정

토큰화

1. 품사 태깅
2. 명사 추출
3. **Keybert** 활용해 키워드 추출 >> 코사인 유사도 활용

Q. KeyBERT는 어떻게 키워드를 추출할까?



메인 분석

Kmeans Clustering

```
for i in tqdm(range(3,10)):
    km_cluster = KMeans(n_clusters=i, max_iter=10000, random_state=0)
    km_cluster.fit(feature_vect)
    cluster_label = km_cluster.labels_
    cluster_centers = km_cluster.cluster_centers_
    text1['label'] = cluster_label
    var_score1.append(sorted(text1['label'].value_counts().values)[-1]/text1['label'].value_counts().sum()*100)
    var_score2.append(sorted(text1['label'].value_counts().values)[-2]/text1['label'].value_counts().sum()*100)
    var_score3.append(sorted(text1['label'].value_counts().values)[-3]/text1['label'].value_counts().sum()*100)
```

» 추출한 키워드의 거리를 기반으로 클러스터링

Q. Kmeans란? 주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 작동



WHY?

Q. 왜 group이 5개 인가요?

〈외계+인〉

안공금하시다크요?...궁금해해주세요...

	Ratio 4 1st	Ratio 4 2nd	Ratio 4 3rd
iter			
3	68.778510	20.527116	10.694374
4	61.023822	20.121642	10.846427
5	50.126711	20.881906	18.955905
6	59.503294	19.310694	8.819057
7	40.293969	20.983274	15.154587
8	52.965028	9.224531	7.754688
9	36.594019	13.076533	12.873796

〈비상선언〉

	Ratio 4 1st	Ratio 4 2nd	Ratio 4 3rd
iter			
3	75.093953	16.433208	8.472839
4	66.757772	15.203280	11.923471
5	60.744790	12.845917	11.342672
6	55.722583	11.240178	11.069354
7	55.278442	12.060130	10.044414
8	48.616331	11.991800	9.941920
9	42.944995	10.864366	9.873591

[illegible]

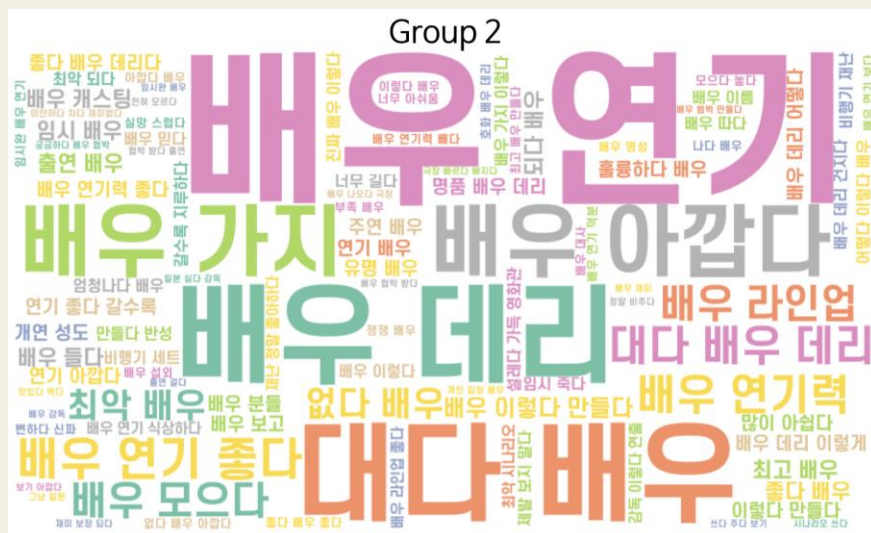
» 클러스터링한 워드 클라우드를 어떻게?

기름 없다더니 여기저기 많이 도네요ㅋㅋ김남길 혼자 생명력 끈질기고ㅋㅋ현실감 너무 떨어지고 쥐어짜는 느낌

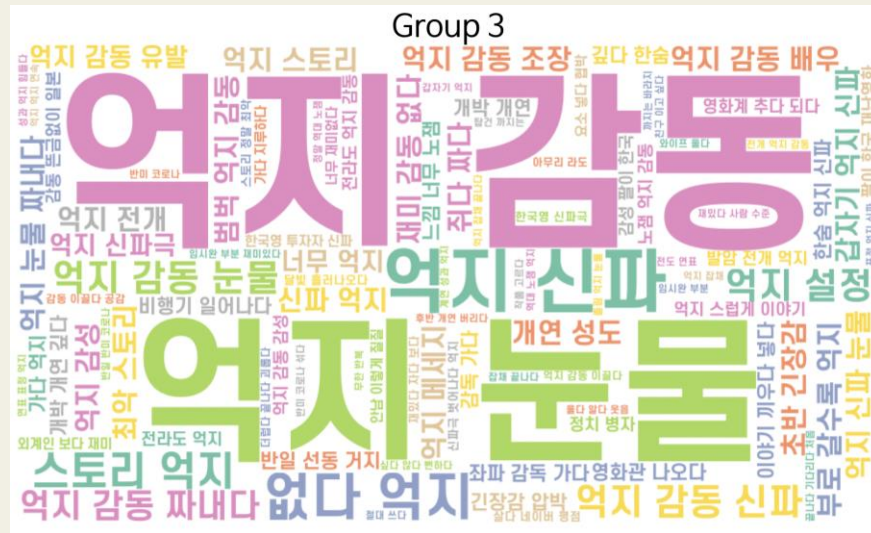
10% 차지
원색적인 비난



12% 차지

배우에 못 미치는
영화에 대한 아쉬움

워드 클라우드 시각화(비상선언)



6% 차지

“억지 신파와 눈물을 빼려는 설정이 지루하다”

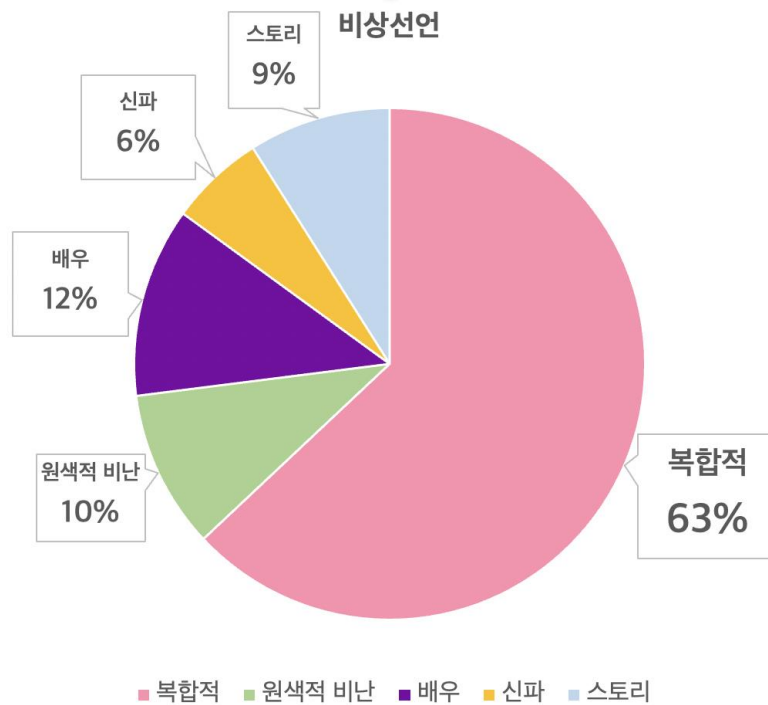
신파에 대한 비판



9% 차지

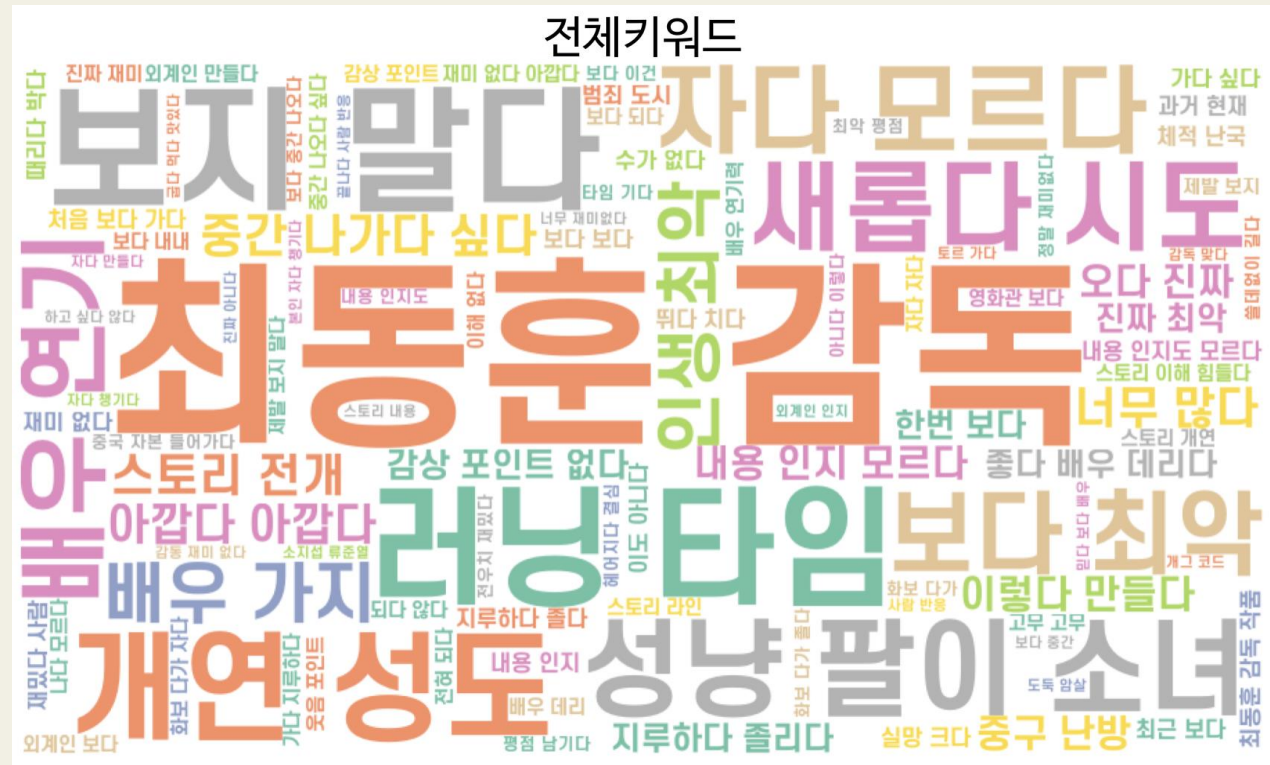
“스토리 개연성이 없고 억지스럽다”

스토리에 대한 평가



	Category	Grade	Ratio	Cat_Mean
0	0	4827.00	63.306994	복합적
1	1	767.25	10.062625	원색적 비난
2	2	921.50	12.085642	배우에 못 미치는 영화에 대한 아쉬움
3	3	455.00	5.967409	신파
4	4	654.00	8.577330	스토리

워드 클라우드 시각화



» 클러스터링한 워드 클라우드는 어떨까?

워드 클라우드 시각화(외계+인)



55% 차지

감독, 성냥팔이, 러닝타임 등
‘복합적’인 요인 존재» 군집화의 문제로 명확히
분류되지 않는 한계점 존재

실제 리뷰 내용

‘성냥팔이소녀의 재림’이라는 희대의 망작이 있었다. 그 쓰레기와 굉장히 유사한 졸작 중의 졸작임. 감독이 한심하다.

성냥팔이 소녀의 재림 2022년판인듯

성냥팔이 소녀의 재림을 능가하는 초특급SF액션대작.

최동훈 감독은 관객을 원하는 영화보다 본인이 평소 하고 싶은 영화 마음껏 찍은 느낌이다. 이전에 망작 성냥팔이 소녀의 재림 향수가 느껴진다...

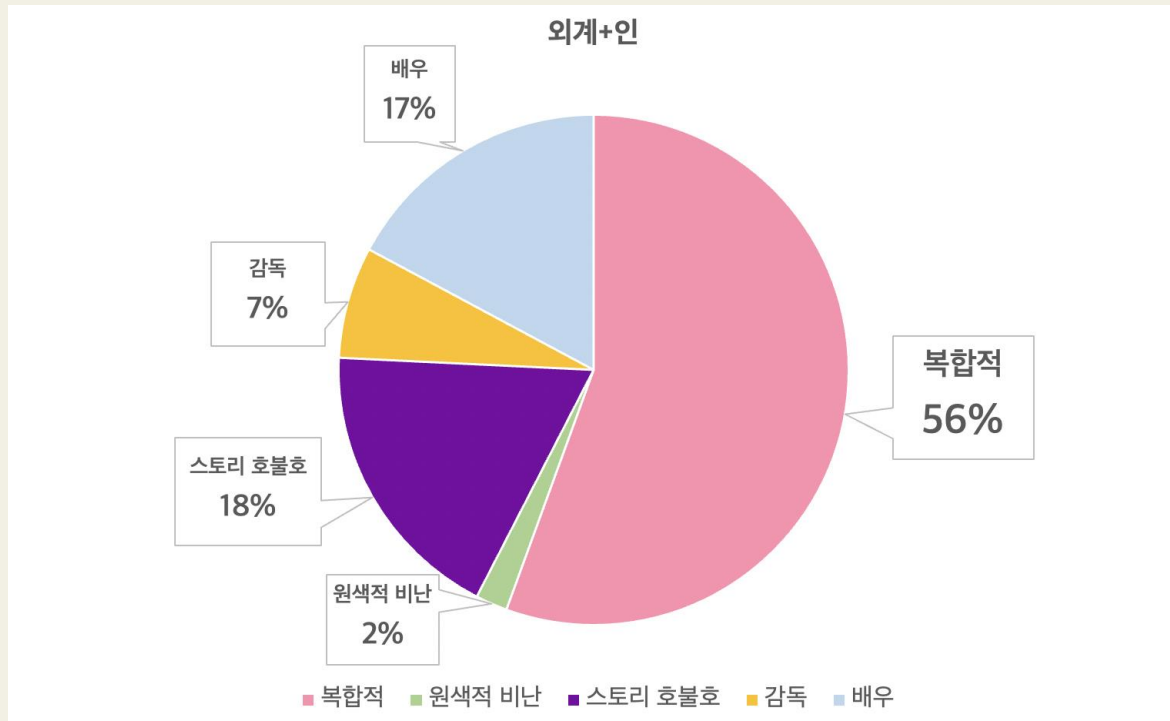
성냥팔이소녀의 재림의 재림

정말 잘만든 초딩용 시간죽이는 SF. 종합선물세트를 꿈꾸는 종합쓰레기 영화. 배우를 한국이 스토리는 중국이 감독,연출은 일본이 한것같은 잡탕영화. 옛날 SF명작 성냥팔이소녀의 아성에 도전하려는 영화

성냥팔이소녀의 재림, 7광구, 리얼을 이을 K- SF 괴작의 탄생!

스토리 전개에 대한 호불호

배우에 미치지 못한
영화에 대한 아쉬움



	Category	Grade	Ratio	Cat_Mean
0	0	3267.50	55.704727	복합적
1	1	132.50	2.258876	원색적 비난
2	2	1085.00	18.497208	스토리에 대한 호불호
3	3	387.50	6.606146	최동훈 감독
4	4	993.25	16.933044	배우에 못 미치는 영화에 대한 아쉬움

감사합니다