

캡스톤 디자인 I 최종결과 보고서

프로젝트 제목(국문): 자연어 처리 기반 영화 리뷰 필터링 및 시각화

프로젝트 제목(영문): Filter and visualize movie reviews based on NLP

프로젝트 팀(원): 학번: 20171602 이름: 한재혁
프로젝트 팀(원): 학번: 20171615 이름: 서민석
프로젝트 팀(원): 학번: 20171598 이름: 정인상

1. 중간보고서의 검토결과 심사위원의 '수정 및 개선 의견'과 그러한 검토의견을 반영하여 개선한 부분을 명시하시오.

- 기존에 영화를 선택하면 실시간으로 해당 영화에 대한 리뷰 데이터를 수집해 모델을 통과시켜 반환하는 방식에서 리뷰 데이터와 모델 통과 결과를 DB에 저장하여 요청시 바로 반환하는 방식으로 수정.
- DB에 저장되는 리뷰 데이터와 모델 통과 결과는 주기적으로 업데이트시켜 최신화.

2. 기능, 성능 및 품질 요구사항을 충족하기 위해 본 개발 프로젝트에서 적용한 주요 알고리즘, 설계방법 등을 기술하시오.

- BERT를 활용한 감성 분석

각 영화별 감상포인트를 분석하기 위해 크롤링한 리뷰들을 한국어 오픈소스 라이브러리인 Kiwi를 이용해 리뷰를 문장 단위로 분리한다. 이후 명사를 추출해 미리 정의한 감상포인트별 키워드에 따라 분류해 감성 분석을 진행한다. 감성 분석은 Hugging Face로 구현한 BERT를 사용하였고, 모델 훈련을 위해 Naver Sentiment Movie Corpus 150,000개를 사용하였다.

```
Batch 100 of 1,563. Elapsed: 0:00:09.
Batch 200 of 1,563. Elapsed: 0:00:17.
Batch 300 of 1,563. Elapsed: 0:00:26.
Batch 400 of 1,563. Elapsed: 0:00:35.
Batch 500 of 1,563. Elapsed: 0:00:43.
Batch 600 of 1,563. Elapsed: 0:00:52.
Batch 700 of 1,563. Elapsed: 0:01:01.
Batch 800 of 1,563. Elapsed: 0:01:10.
Batch 900 of 1,563. Elapsed: 0:01:18.
Batch 1,000 of 1,563. Elapsed: 0:01:27.
Batch 1,100 of 1,563. Elapsed: 0:01:36.
Batch 1,200 of 1,563. Elapsed: 0:01:44.
Batch 1,300 of 1,563. Elapsed: 0:01:53.
Batch 1,400 of 1,563. Elapsed: 0:02:02.
Batch 1,500 of 1,563. Elapsed: 0:02:10.
```

```
Accuracy: 0.87
Test took: 0:02:16
```

그림 1 테스트셋으로 성능을 측정한 결과
87%의 정확도를 달성함

3. 요구사항 정의서에 명시된 기능 및 품질 요구사항에 대하여 최종 완료된 결과를 기술하시오.

- 감상포인트별 긍·부정 비율 계산

크롤링한 리뷰를 json 파일 형태로 읽어들인 뒤, 각 리뷰에서 명사를 추출해 미리 정의해 놓은 감상 포인트별 키워드에 따라 이를 분류한 뒤, 감성 분석을 진행하는 방식으로 구현하였다.

```
review len: 32074
[8.493150684931507, 37.36986301369863, 5.26027397260274, 48.273972602739725, 0.6027397260273972]
[65.80645161290323, 34.193548387096776, 81.23167155425219, 18.7683284457478, 89.58333333333334, 10.416666666666668, 7
0.03405221339388, 29.96594778660613, 81.81818181818183, 18.181818181818183]
point extraction time : 13.715265035629272
```

그림 2 크롤링한 32,074개의 리뷰를 처리하는데 13.72초가 소요됨

4. 구현하지 못한 기능 요구사항이 있다면 그 이유와 해결방안을 기술하시오,

최초 요구사항	구현 여부 (미구현, 수정, 삭제 등)	이유 (일정부족, 프로젝트 관리미비, 팀원변동, 기술적 문제 등)
리뷰 필터링 모델 개발	미구현	ABSA에 대한 지식 부족 및 일정 부족으로 인해 추후 구현 예정
감상 포인트 추출 모델 개발	수정	ABSA에 대한 지식 부족으로 룰 베이스로 대체하여 최소 구현. 향후 ABSA로 대체 예정

5. 요구사항을 충족시키지 못한 성능, 품질 요구사항이 있다면 그 이유와 해결방안을 기술하시오.

분류(성능, 속도 등) 및 최초요구사항	충족 여부 (현재 측정결과 제시)	이유 (일정부족, 프로젝트 관리미비, 팀원변동, 기술적 문제 등)
감성 분석 모델 성능	정확도: 87%	리뷰 데이터의 특성으로 인한 문제로 클렌징 및 맞춤법·띄어쓰기 교정을 통해 데이터를 정제할 예정

6. 최종 완성된 프로젝트 결과물(소프트웨어, 하드웨어 등)을 설치하여 사용하기 위한 사용자 매뉴얼을 작성하시오.

- (1) 사용자는 데스크탑과 모바일을 통해 해당 서비스에 접근할 수 있습니다. (그림 3-1)
- (2) 서비스에 접근한 사용자는 상단의 검색바를 통해 원하는 영화를 검색하게 됩니다. (그림 3-2)
- (3) 검색된 결과 중, 원하는 영화를 클릭하면 해당 영화의 상세 정보를 확인할 수 있습니다. (그림 3-3 상단)
- (4) 상세 정보 아래의 ‘리뷰보기’버튼을 클릭하면 필터링된 해당 영화의 리뷰들을 볼 수 있습니다. (그림 3-3 하단)
- (5) 상세 정보 아래의 ‘분석보기’버튼을 클릭하면 관람객들의 해당 영화의 관람 포인트 별 긍정 및 부정 퍼센테이지(%)를 볼 수 있습니다. (그림 3-4)

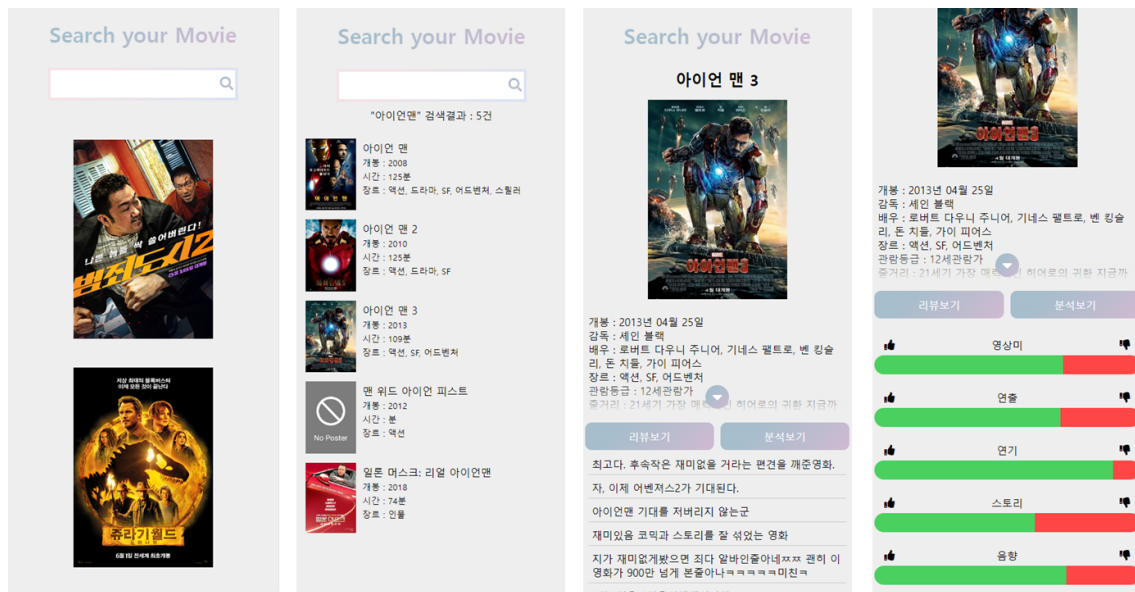


그림 3 왼쪽부터 그림 3-1, 3-2, 3-3, 3-4

7. 캡스톤디자인 결과의 활용방안

- 사회적 기대효과

대다수의 소비자들이 소비 후 불만족을 줄이기 위해 온라인 리뷰를 확인한다. 하지만 이를 악용하여 일부로 부정적인 리뷰를 작성하거나 실제로 소비를 하지 않았음에도 한 것처럼 작성한 리뷰들이 많아 실제 소비자들이 리뷰를 확인하는 것에 많은 피로도를 느끼고 있다. 이를 필터링 해 소비자들에게 도움이 될만한 리뷰를 보여줌으로써 사용자들에게 좀 더 객관적인 평가와 정확한 사실을 제공받을 권리를 보장하고, 또한 만족 포인트를 텍스트 내에서 추출해 요약하여 보여 주기 때문에 소비자들의 질 높은 소비 활동이 가능해질 것으로 기대된다.

- 기술적 기대효과

다양한 상품에 대한 리뷰가 넘쳐나는 현재, 리뷰 데이터를 분석해 소비자의 만족도를 끌어올리는 것은 많은 이커머스 기업들의 목표이다. 속성 기반 감성분석은 기존 감성 분석 모델보다 발전된 형태의 감성 분석 기술로 국내에서도 관련 연구가 이루어지고 있지만 현재까지는 적극적으로 활용되고 있지 않다. 본 캡스톤 디자인은 영화 리뷰에만 이를 적용했지만 이에 국한되지 않고, 다양한 소비 영역으로 확장할 수 있을 것이라 생각된다.