2018 – 가을학기 - 온라인데이터수집과 분석

**Final Project Proposal**

2018321084 안주영

2017321696 박가람

**1. 목적**

트위터, 페이스북 같은 소셜미디어가 최근 급격히 발달하였고 해당 플랫폼에 거의 모든 영화에 대한 다수의 평이 존재한다.

영화평점 사이트에 댓글을 남기는 사람은 제한적이며 따라서 영화평점사이트에 올라와 있는 평점과 한 줄 평을 통해서 영화에 대해 평가하는 것은 sampling관점에서 biased되었을 확률이 높다. 반면 많은 SNS 혹은 인터넷 게시판에는 대부분의 사람들이 자유롭게 영화에 대해서 이야기를 나누고 있다. 따라서 SNS와 영화관련 커뮤니티의 글들을 통해서 영화에 대한 review를 평가하는 것이 좀더 unbiased한 정보를 얻을 수 있다고 생각된다. 또한 예술영화 같은 비주류 영화의 경우 평점의 수가 매우 적은 데 소셜미디어를 통한 분석을 통해 이러한 취약점을 극복할 수 있으리라 기대한다.

영화평점 사이트에서 한 줄 평과 동반하는 평점데이터를 동시에 얻을 수 있다. 이는 각 텍스트에 대한 labeling이 되어 있음을 의미한다. 따라서 영화평점사이트의 데이터를 training set으로 활용하여 감정분석 모형을 모델링한다, 이후, 만들어진 감정분석 모형을 SNS와 인터넷 게시판에 올라온 텍스트에 적용하여 그 적합성을 확인하고자 한다.

**2. 수집방법**

영화평점 사이트로는 가장 접근성이 좋은 네이버영화를 선택하였고, 해당 사이트의 한 줄 평, 평점 데이터를 web crawling 방법을 통해 수집할 계획이다.

SNS의 경우 twitter에서 영화제목을 검색하여서 나오는 텍스트 일부를 분석에 사용하고자 한다. 영화에 대해서 검색을 했을 때 나오는 데이터들의 양이 매우 많고, 따라서 상단에 검색되는 게시물 중 일부를 모델링이 잘 적합 되었는지 확인하는 test set으로 사용할 계획이다.

인터넷 게시판의 경우 한국에서 가장 대중적인 영화 사이트로 알려진 ‘익스트림 무비’의 ‘영화수다’ 게시판에서 영화 제목을 검색하여 댓글수가 많은 글의 댓글을 분석하거나, 게시글의 제목들을 활용하여서 모형 적합성 테스트를 진행하고자 한다.

**3. 분석 방법**

영화평점 사이트의 댓글을 활용해 빈도분석을 진행하고 각각 긍정, 중립, 부정적인 글에서 많이 표현되는 단어를 선별해 모델링을 진행할 계획이다. 단어별로 각각 긍정, 부정, 중립의 평가로 이어질 확률을 계산하고 새로운 데이터를 입력 받았을 때 특정 감정으로 판별할 확률 값을 계산하는 naïve bayes 방법으로 모델링을 진행하겠다. 또한 이외의 추가적으로 적용할 수 있는 알고리즘을 찾아보고 accuracy rate를 개선하고자 한다.