2018 – 가을학기 - 온라인데이터수집과 분석

**Final Project Proposal**

2018321084 안주영

**1. 목적**

트위터, 페이스북 같은 소셜미디어의 발달

영화평점 사이트에 댓글을 남기는 사람은 제한적이다. 따라서 영화평점사이트에 올라와 있는 평점과 한줄평을 통해서 영화에 대해 평가하는 것은 sampling관점에서 biased된 sampling이 될 수 있다. 반면 많은 SNS 혹은 인터넷 게시판에는 대부분의 사람들이 자유롭게 영화에 대해서 이야기를 나누는 것을 볼 수 있다. 따라서 SNS와 영화관련 커뮤니티의 글들을 통해서 영화에 대한 review를 평가하는 것이 좀더 unbiased한 정보를 얻을 수 있다고 생각된다.

영화 평점사이트의 리뷰는 단순히 영화 리뷰에 대한 텍스트 데이터뿐만이 아니라 평점을 함께 작성하기 때문에 각 텍스트에 대한 labeling이 되어 있다는 장점이 있다. 따라서 영화평점사이트의 데이터를 training set으로 활용하여서, SNS와 인터넷 게시판에 올려진 텍스트들의 감정분석을 시행하고 이를 통해 다른 매체에서 만들어진 감정분석 모델이 SNS에서도 적용될 수 있음을 확인하고자 한다.

**2. 수집방법**

영화 평점사이트는 가장 접근성이 좋은 네이버영화의 한줄평 데이터를 web crawling 방법을 사용해서 수집하고자 한다.

SNS의 경우 twitter에서 영화제목으로 검색하여서 나오는 텍스트들을 분석에 사용하겠다. 실제로 영화에 대해서 검색을 했을 때 나오는 데이터들의 양이 매우 많기 때문에 상단에 검색되는 게시물 중 일부를 모델링이 잘 적합 되었는지 확인하는 테스트셋으로 사용하고자 한다.

인터넷 게시판의 경우는 가장 대중적으로 이용하는 사이트로 알려진 ‘익스트림무비’의 ‘영화수다’게시판에서 영화의 제목으로 검색해서 나오는 글 중 댓글이 많은 글의 댓글을 분석, 또는 게시글들의 제목을 활용하여서 테스트를 진행하고자 한다.

**3. 분석 방법**

영화평점사이트의 댓글을 활용해 빈도분석을 통해 각각 긍정, 중립, 부정적인 글에서 많이 표현되는 단어를 활용해서 분석을 진행하도록 하겠다.

단어별로 각각 긍정, 부정, 중립의 확률을 계산하여서 새로운 input을 받았을 때 특정 감정으로 판별할 확률 값을 계산하여서 naïve bayes방법으로 모델링을 진행하겠다.

추가적으로 적용할 수 있는 알고리즘을 활용할 수 있는지 찾아보고 accuracy rate를 높일 수 있는 알고리즘을 구축해 보도록 하겠다.