딥러닝 기반 핵심 산업별 빅데이터 분석

<키워드 분석 프로젝트 – 1팀>

일 자	2018년 11월 19일	주 제	영화 키워드 분석
팀 장	김하준<khajuni@gmail.com>	팀 원	문정연, 이상훈, 주상훈

주제 : 2018 청룡영화제 남녀주연상 수상자 예측

2013-2017 5년간의 청룡영화제 남녀주연상 수상작 댓글 분석을 통해 2018 수상자 예측하기

가설1) 이전 수상작 댓글에서 연기력,배우와 관련된 단어들이 많이 언급될 것이다.

가설2) 가설1에서 도출한 단어들이 많이 언급된 영화의 배우가 2018 청룡영화제 남녀주연상을 수상할것이다.

1단계. 영화 데이터 크롤링 (사이트URL, 수집데이터 항목)

1.데이터 수집

- 2018 영화흥행순 70개 영화 api 다운 (영화진흥위원회)
- 2013-2017 수상작 데이터 (csv로)
- 리뷰 크롤링

1) 2018 영화흥행순 70개 영화 api 다운

- http://www.kobis.or.kr/kobis/business/stat/offc/searchOfficHitTotList.do?searchMode=year
  - 청룡영화제 수상 대상에 해당하는 2018년 한국 국적의 상업영화
  - 총 434건 중 70위권 이하의 경우 상영관이 10개 미만이며 관객수 1000명 이하인 경우가 대다수

  ➔ 흥행 순위 70위권으로 분석 데이터 선정
- 영화데이터분석/2018 전체작/movie2018_69.csv

2) 2013-2017 수상작 데이터 (csv로)

- 영화데이터분석 / 이전 수상작 / awarded.csv

3) 리뷰크롤링

- 네이버 영화의 평점,리뷰 항목에서 selenium 을 통해 미리 만들어놓은 데이터 중 영화명을 순차적으로 입력
- 입력된 영화의 코드명과 총 리뷰수, 마지막 페이지 수를 구함
- 마지막 페이지수까지 for구문을 통해 반복적으로 리뷰를 크롤링해서 리스트에 저장
- 영화 코드명,리뷰수,마지막 페이지 등의 정보가 담긴 csv를 데이터 프레임화 하여 csv로 저장
- 영화 전체 리뷰들 저장
- 위의 과정을 2018 전체작, 2018이전 수상작들 모두 시행하여 데이터 추출

2단계. 분석 검증 + 워드 클라우드

크롤링한 리뷰 데이터 형태분석 및 단어 빈도수 추출

가설 1 검정결과

가설1) 이전 수상작 댓글에서 연기력,배우와 관련된 단어들이 많이 언급될 것이다.

- 실패작과 수상작 형태소 분석
- 워드클라우드에서 주요 키워드 비교
- 수상작에서 나타나는 키워드 도출

  ➔ 워드 클라우드 그려본 결과 도출된 키워드 : 연기 / 연기력 / 배우이름 / 감동 ….

- 실패작과 수상작을 비교하여 확인하려 했으나 청룡영화제에서 수상하지 못한 영화이더라도 대종상 등 다른 영화제에서 수상한 경우를 고려하지 못했음
- 이미 타 영화제에서 수상한 경우 상을 몰아주지 않는 영화제들간의 관계를 충분히 고려하지 못한점이 한계

3단계. 분석 예측

크롤링한 리뷰 데이터 형태분석 및 단어 빈도수 추출

- 실패작과 수상작 형태소 분석
- 워드클라우드에서 주요 키워드 비교
- 수상작에서 나타나는 키워드 도출

결론

- 연기, 감동이라는 단어가 많은 작품이라고 해서 무조건적으로 수상을 하는 것은 아니다.
- 즉, 수상작에서 많이 나타나는 단어의 비율이 높다고 해서 수상가능성이 높을 것이라는 가설 2 기각
- 연기라는 단어가 리뷰에서 사용되는 경우 실제 배우의 연기력과 연관성이 크지 않은 것으로 보임

  ➔ 오히려 특정 배우와 함께 사용되는 경우가 더욱 많았기에 특정 배우들의 티켓파워를 검증하는 척도로 사용될 수 있다는 새로운 가능성 제시
- 우리의 분석대상은 네티즌 리뷰였으나 영화제 수상의 최종결정권자는 평론가들이기에 이러한 측면을 더욱 고려하는 것이 필요해보임
- 단순 네티즌 리뷰로는 영화제 수상 예측이 어렵다.

감사합니다