Visualization in Sentiment-based Color and Comparative Analysis of Sales Figures in Harry Potter Series

Sungbin Kim^{a,1}, Sungho Gong^{a,1}, Seungpyo Hong^{a,1}

^aDepartment of History, Ajou University

¹ These authors contributed equally to this work.

Introduction to Data for Humanities (Capstone Design) Team Project

서 론

영국 작가 J. K. Rowling이 쓴 『해리포터(Harry Potter)』시리즈는 현대 문학의 상징과도 같다. 1997년 『해리포터와 마법사의 돌 (Harry Potter and the Philosopher's Stone)』로 출판을 시작한 해 리포터의 이야기는 79개 언어로 번역되어 전 세계적으로 5억 부 이상이 팔렸다. 『해리포터』시리즈는 또한 영화로 제작되어 성공 적인 영화 프랜차이즈의 본보기가 되었고, 대중문화 발달에 큰 발 자취를 남겼다. 다양한 연령층으로 이루어진 수많은 독자의 상상 력을 자극한 『해리포터』시리즈는 아직도 사람들에게 큰 사랑을 받고 있다.

본 프로젝트는 자연어 처리를 이용하여 『해리포터』시리즈의 감 정 관련 단어들을 분류·분석하고 시리즈 전반의 감정의 궤적을 파 악하고자 한다. 나아가 그 궤적이 작품의 분위기와 매력에 어떤 영 향을 미치고 궁극적으로는 시리즈별 판매량에 어떤 영향을 미쳤는 지 분석하고자 한다. 이는 지금까지 사회적·문화적·문학적인 관점 에서 주로 분석되었던 문학 작품에 새로운 해석의 층을 제시할 수 있다.

요컨대, 본 프로젝트는 1차적으로『해리포터』전권에 등장하는 감정 관련 단어의 정량을 측정하고자 한다. 그다음 시리즈별 지배 적인 감정 흐름과 기복을 조명하고 그것을 심리학자 Robert Plutchik가 제시한 감정 원형모형의 색상에 대입하여 시각화하고 자 한다. 마지막으로 출력된 시각 자료 도서 판매량을 바탕으로 분 석함으로써 독자들로부터 많은 인기를 끌어올 수 있었던 비결을 찾고자 한다.

연구 방법론

① 텍스트 전처리 및 형태소 분석

『해리포터』전권 텍스트를 구하고 마침표마다 줄을 나누었음.

한글을 제외한 다른 모든 문자를 제거하고 **토큰화**한 후 의미가 있는 **명사, 동사, 형용사** 등을 추출함.

result = eachOut[0]+"/"+each sentencePOS = sentencePOS +

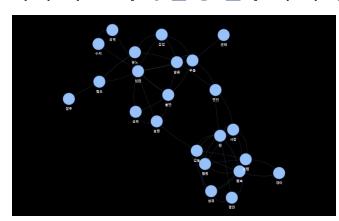
② 감정 점수 부여 및 분류

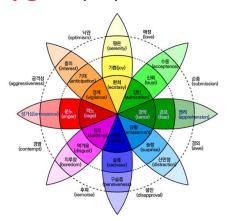
감정 사전에 배정된 점수에 따라 단어에 **감정 점수**를 부여함. TF-IDF **알고리즘**에 따라 점수를 조정함.

0 후		공포	권 태	놀라	분노	불 안	사 랑	설 렘	섭 섭	성 취	수치	슬픔	심란	연민	열정	아애	재 미	정	죄 책	질 투	평 안	행복	혐 오	중립
비찬성	- o	1.2	0	0.6	0	0	0	0	0	0	0.4	4.4	1.8	0	0	1.8	0	0	0.4	0	0	0	1.2	0
7 7 7	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0.6	4.8	0	0

감정들을 감정 유사도 네트워크와 Robert Plutchik(1980)의 감정 원형 모형에 따라 분류하고 대표 색상을 정함.

단어가 가진 최대 감정 점수에 따라서 색상을 부여함.

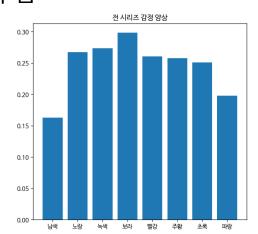




문장 내에 우세한 색상이 그 문장의 대표 색상이 됨. (90,825개 문장이 각각 색을 가짐)

③ 시각화 (전체)

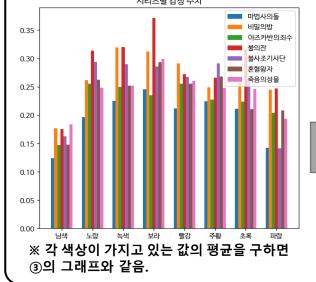
Python의 Matplotlib 패키지로『해리포터』전권의 감정 양상을 시각화 함

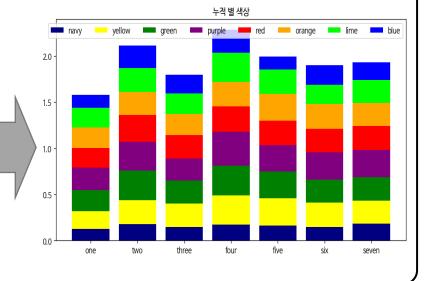


남색(navy)	슬픔, 우울, 섭섭, 연민, 권태
노랑(yellow)	행복, 재미, 설렘, 평안
녹색(green)	공포, 불안, 수치, 죄책
보라(purple)	혐오, 질투
빨강(red)	분노
주황(orange)	열정, 성취
초록(lime)	정, 사랑, 감동
파랑(blue)	놀람, 심란

⑤ 시리즈별 시각화

시리즈마다 색상을 다르게 하여 다중 그래프를 출력함. 시리즈별 감정의 양상을 파악하기 용이하나 감정의 궤적을 추적하기 어려우므로 x축을 감정이 아닌 시리즈로 설정

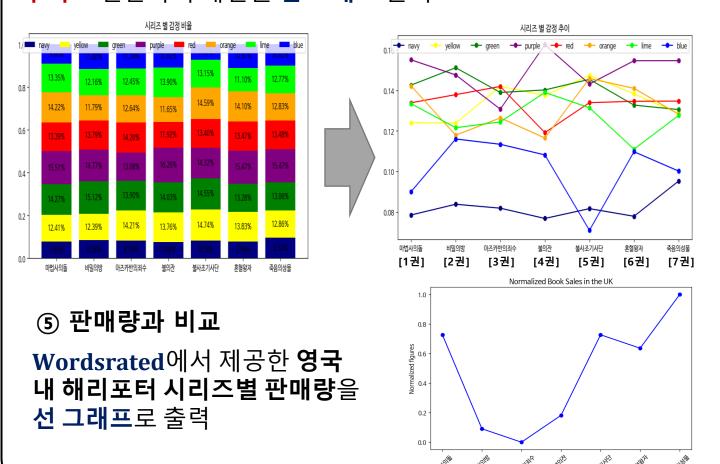




연구 방법론 (계속)

각 감정이 시리즈에서 차지하는 비율을 알기 위해 막대의 높이를 모두 1(100%)로 설정하고 백분율을 표시함.

감정의 **차이**가 도드라지지 않아 분석이 어려우므로 y축을 정규화한 수치로 변환하여 개선된 선 그래프 출력



TF- IDF 알고리즘

TF-IDF Term Frequency – Inverse Document Frequency

TF: 한 문서 내에서 특정 단어의 빈도 수

IDF: 전체 문서에서 특정 단어가 출현한 빈도 수가 클수록 작아지는 값 def tf(t, d):

 $w_{i,j} = t f_{i,j} \times log(\frac{N}{df_i})$

• N은 문서 전체 개수

• $tf_{i,j}$ 는 단어 i가 문서 j에 출현하는 빈도수

for doc in docs: return log(N/(df+1)) def tfidf(t, d): • df_i 는 전체 문서에서 단어 i가 출현하는 빈도수를 나타낸다.

return d.count(t)

return tf(t,d)* idf(t)

def idf(t):

문장은 출현하는 단어의 빈도 수에 영향을 받음.

중요한 단어는 일상적으로 쓰이는 단어들보다 빈도 수가 낮을 것이라고 예상하여 너무 많이 나오지도, 너무 적게 나오지도 않는 단어가 문장에서 중요한 역할을 한다고 가정

해석과 의의

① 일반 소설의 문법

해리포터 전체의 서사가 '발단-전개-위기-절정-결말' 순으로 이루어진 일반 소설의 문법을 따른다는 것을 감정 분석을 통해 확인하였다.

『해리포터와 불의 잔』: 보라(혐오, 질투)의 빈도수가 높고, 주황(열정, 성취)의 빈도가 낮 음. -> '**위기**' (주인공이 새로운 사태를 만나 목표의 달성에 실패하거나 절망에 빠짐)

『해리포터와 불사조 기사단』, 『해리포터와 혼혈 왕자』

: 빨강(분노)과 주황의 빈도가 높고, 파랑(놀람·심란)의 빈도가 낮음 -> '**절정**' (주인공이 위기에서 얻은 수단과 해결책을 통해 최후의 행동으로 나아감)

② 판매량 증가 원인 1: 카타르시스

※『해리포터와 마법사의 돌』과『해리포터와 죽음의 성물』은 각각 발단과 결말이라 는 점에서 판매량에 영향을 주었을 것으로 예상되므로 해석에서 제외

빨강, 주황, 초록, 파랑의 변화에서 **판매량과 인기**의 **증가 원인**을 예측

『해리포터와 불사조 기사단』 이전 작품

- <mark>분노</mark>, 공포, 불안, 놀람, 심란이 높음.

『해리포터와 불사조 기사단』

- 열정, 성취가 높아짐 - 놀람, 심란이 낮음. - 열정, 성취가 낮음

(주인공이 침착하게 사건 해결)

... 독자들의 피로와 답답함

... 카타르시스와 주인공의 활약

③ 판매량 증가 원인 2: 절정과 해피엔딩

『해리포터와 불사조 기사단』

『해리포터와 혼혈 왕자』

『해리포터와 죽음의 성물』

- 볼드모트를 쓰러뜨릴 방법을 알게 됨 - 조력자(시리우스, 덤블도어)의 사망

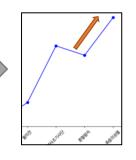
- 보라(혐오, 질투)가 높음 - 남색(슬픔, 우울, 섭섭, 연민)이 가장 높음.

··· 마지막 편까지 '절정'이 계속됨

슬프고 우울한 사건을

모두 **해결**하고 **행복한 결말**을 맞음 ... 더욱 큰 카타르시스

... 숙원 해소와 목표 달성의 만족감



※『해리포터와 죽음의 성물』이 시리즈의 마지막이기 때문에 판매량이 높을 것이 라는 일반적인 분석에 또 다른 원인을 제시하는 새로운 분석

한계점과 개선 방안

① 세부 분석의 필요성

시리즈별로 분석하는데 성공하였으나 이를 연결하여 『해리포터』의 서사 전반의 **감정 궤도**를 분석하기에는 한계가 있었음.

→ 장(Chapter)과 같은 소단위를 기준으로 분석한다면 연속적인 감정의 궤도를 분석 할 수 있고 시리즈마다 가진 패턴 또한 분석할 수 있을 것으로 기대

② 자료의 한계

본 프로젝트는 'https://wordsrated.com/'에서 제공한 영국 내『해리포터』판매량 통 계를 사용함. 그러나 『해리포터』의 영국 내 판매량은 전 세계 판매량의 3~5%밖에 미 치지 못함.

→ 영화의 각본과 대중에게 정확하게 공개된 '영화 <해리포터> 시리즈의 전 세계 매 출액' 통계를 활용한다면 더욱 정확하고 의미 있는 조사가 이루어질 것으로 기대