네이버 웹툰 텍스트 분석

빅데이터 경영통계
20172848 이지평

Index

01 서론
네이버 웹툰 텍스트 분석

고수 류기운 / 문정후

<용비불패> 최강의 콤비가 무협의 전설을 다시 쓰다!
천하제일의 고수 강룡. 그리고 수많은 다른 고수들의 이야기.

스토리, 판타지 | 15세 이용가

소녀의 세계 모랑지

완벽해 보이지만 사실 외로웠던 백조들과 맘씨 착한 오리가 만나
여러 갈등을 함께 겪으며 진짜 친구가 되어가는 소녀들의 찐 우정을

스토리, 드라마 | 전체연령가

웹툰 소개글 너무 추상적으로 묘사

↓

입문자들 소개글 통해 새로운 웹툰을 시작
하기 어려움 있음

↓

댓글에서 해당 웹툰의 특징을 뽑아내서 해시태그
형식으로 키워드를 제공한다면,
좀 더 쉽게 새로운 웹툰에 접근 할 수 있지 않을까?

02 데이터 수집
네이버 웹툰 텍스트 분석

사이트	https://comic.naver.com/webtoon/weekday.nhn
크롤링 방법	Selenium + lxml
대상	현재 연재 중 & 완결되었지만 모든 화에 접근가능한 일부 웹툰의 베스트 댓글들
장르	판타지, 드라마(장르별로 분류했을 때 가장 많은 수를 차지함)
데이터 수	판타지(4470화) , 드라마(2885화) 각 행(화)마다 15개의 베스트 댓글들이 들어있음.

	index	comment	score	episode	date	sentiment	year
0	1	정주행하는 사람 조용히 추천 전세계로 퍼졌는데 우리나라만 탐에 열쇠에 손오...	9.87	1화	2011.04.08	0	2011
1	2	가 100억을 겨우 그거라고 말할처지가 아닐텐데 빚 75억있는게 누구...	9.93	2화	2011.04.15	1	2011
2	3	좋은 씨앗의 강한 씨앗 찾을꺼면 진모리랑 결혼해 ㅋㅋㅋ 둘이 잘어울려 ...	9.91	3화	2011.04.22	1	2011
3	4	돈뿌릴때 11개던데 십만원이라네 ㄷㄷ 11만원 찾을려다가 5622만원 날리...	9.89	4화	2011.04.29	1	2011
4	5	이때까진 가 개그캐가 될줄은 몰랐지 1ㅅ 봉사 아저씨 ㅠㅠㅠ 지금보니까...	9.95	5화	2011.05.06	1	2011
...	...	...	...	...	...	...	...
1412	1413	불평불만 하면서 매번 바로 들어오는 덴경대가 레전드 뭔 우주 최강의 퀑이 ...	5.06	3-410화 2.에필로그(4-7)	2019.12.22	0	2019
1413	1414	자꾸 보러오는 우리가 제일 문제 아닐까 낭심ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 저번에 누가...	6.24	3-411화 2.에필로그(4-8)	2019.12.24	0	2019
1414	1415	그러든가 말든가 ㅋㅋㅋㅋㅋㅋㅋ작가의 말 저기다 써놨네 다음편 에드레이의 박...	6.32	3-412화 2.에필로그(4-9)	2019.12.25	0	2019
1415	1416	덴마가 컸으니 덴마크 앜ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 아니 씨 귀여운 셀 어디가고 ...	5.12	3-413화 2.에필로그(5-1)	2019.12.27	0	2019
1416	1417	이걸 인생웹툰이라도 소개하고 다녔던 내 10년이 아깝다 용두사미급도 아니고...	2.13	3-414화 2.에필로그(5-2)	2019.12.29	0	2019

03 전처리

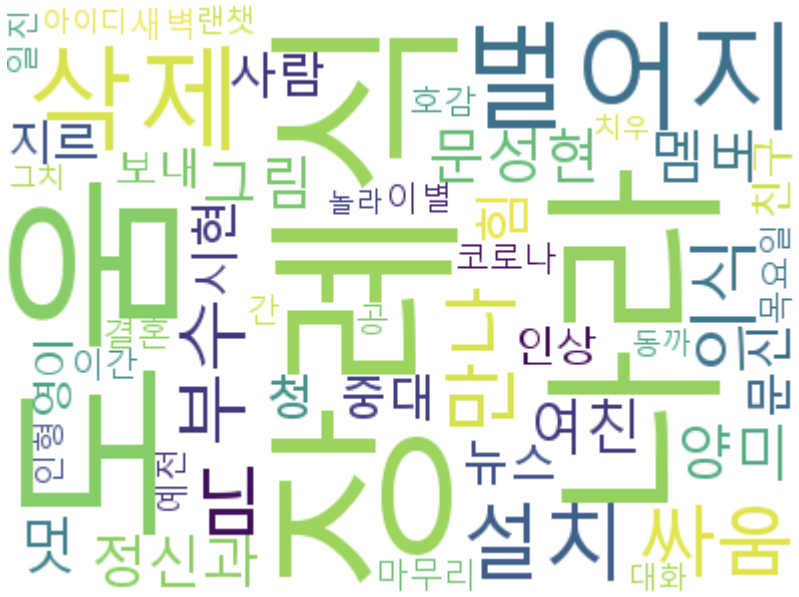
네이버 웹툰 텍스트 분석

감정(sentiment) 처리	웹툰 당 평점의 분포가 다르므로 각 웹툰의 median 기준, 크거나 같으면 긍정, 작으면 부정으로 처리
정규 표현식	한글과 숫자를 제외한 불필요한 영어, 이모티콘 삭제
불용어 사전	https://www.ranks.nl/stopwords/korean + 전처리간 발생한 불용어
형태소 분석	Kiwi 라이브러리 사용, load_user_dictionary 통해서 주요한 고유명사 가중치 강화
토큰화	명사와 동사 추출 - 명사 : 불용어 사전에 포함되지 않고, 의존명사가 아닌 것 추출 - 동사 : 한 글자는 해석이 힘드므로, 두 글자 이상 동사 추출

04 단어 빈도

네이버 웹툰 텍스트 분석

	단어	빈도
1433	장례식	154.999886
426	도움	141.636224
275	나라	131.688294
715	벌어지	127.904107
861	삭제	118.007253
918	설치	107.852260
1049	싸움	98.817624
535	만나	92.934402
349	님	92.789338
776	부수	88.711477
1322	의식	81.048240
608	문성현	78.980507
1110	양미	78.458861
567	멋	76.637301
1158	여친	76.423403
1494	정신과	76.235167
202	그림	75.959251
571	멤버	75.796405
1997	힘	75.198124
1560	중대	72.329137



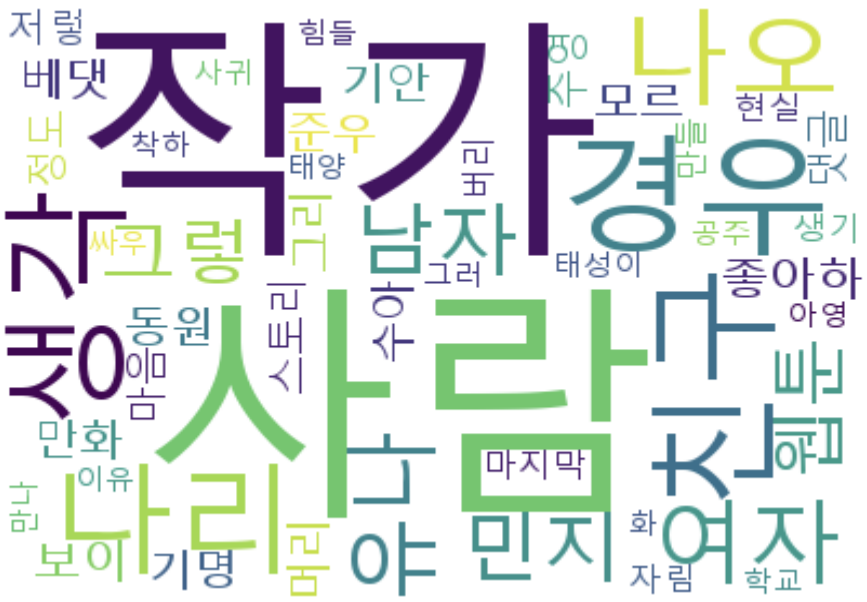
판타지 장르

- 명사보다는 동사 위주
- 장례식, 삭제, 싸움, 부수-, 정신과, 힘 등 다소 과격한 단어 등장

04 단어 빈도

네이버 웹툰 텍스트 분석

	단어	빈도
1418	작가	116.172479
843	사람	111.843773
95	경우	91.911152
1750	친구	86.243420
891	생각	84.279518
278	나리	78.102515
282	나오	76.935379
1152	여자	73.422213
1290	유나	69.966196
296	남자	63.452367
633	민지	60.494329
200	그렇	57.947205
1282	웹툰	57.548417
1533	좋아하	56.535151
742	보이	55.131151
228	기안	52.281773
576	모르	52.159966
1557	준우	50.322392
442	동원	49.514102
561	머리	48.560952



드라마 장르

- 동사보다는 명사 위주
- 등장인물의 이름 비중 높음

05 감성 분석

네이버 웹툰 텍스트 분석


```
comments_fan['sentiment'].value_counts()
```

```
1    2526
0    1944
Name: sentiment, dtype: int64
```

긍정(1) : 부정(0) = 0.56 : 0.44

```
comments_dra['sentiment'].value_counts()
```

```
1    1589
0    1296
Name: sentiment, dtype: int64
```

긍정(1) : 부정(0) = 0.55 : 0.45

웹툰 별로 평점의 분포가 다르므로 각 웹툰의 중앙값을 기준으로 sentiment 열 만듦.

05 감성 분석

네이버 웹툰 텍스트 분석

판타지 부정감성

	토큰	가중치
1935	현봄이	-0.529252
1462	전개	-0.441121
19	가지	-0.434661
1111	양민	-0.417701
291	남녀	-0.406110
617	뭣	-0.383105
1934	혁진이	-0.377843
954	소림이	-0.372258
1260	우기	-0.371046
1461	적응	-0.365096
149	관련	-0.360103
1688	천재	-0.357350
1298	유이건	-0.354524
1921	행복	-0.339866
829	뻐	-0.325702
1722	최한	-0.318824
1284	위기	-0.316709
549	맛있	-0.314739
766	부끄럽	-0.311499
1297	유명	-0.311148

- 웹툰 내용에 대한 부정적 감성
  - 특정 인물이 잘못을 했거나 보는 이들로 하여금 답답함을 유발할 때
- 웹툰 자체에 대한 부정적 감성
  - 웹툰의 전개가 개연성 없이 진행될 때 혹은 무리한 진행으로 가지가지 한다는 느낌을 받을 때

05 감성 분석

네이버 웹툰 텍스트 분석

판타지 긍정감성

	토큰	가중치
1789	탓	0.320078
123	고요	0.320154
1541	주위	0.333787
1177	연예인	0.334491
129	유일	0.334495
341	느낌	0.338589
695	방향	0.338993
572	면	0.341457
1290	유감	0.342475
520	마감	0.347562
787	분홍	0.353964
361	다희	0.357313
1153	여자	0.361151
1960	확률	0.368005
553	망치	0.370801
176	커버	0.387515
1815	특정	0.403292
1053	쌍이	0.404948
1874	피곤	0.437441
597	못생기	0.536564

- 웹툰 내용에 대한 긍정적 감성
  - 특정인물이 주인공을 커버쳐준다거나 도움이 되는 행동을 할 때
- 웹툰 자체에 대한 긍정적 감성
  - 마감을 빨리하고, 웹툰의 방향성이나 느낌이 유일하다고 느껴질 때

05 감성 분석

네이버 웹툰 텍스트 분석

드라마 부정감성

	토큰	가중치
997	스토리	2.574345
751	복수	-2.240108
1462	전개	2.173629
961	소재	2.039986
862	산	2.022372
162	진행	1.986383
1390	임대	-1.960252
1538	주식	-1.936832
712	번화	-1.935711
116	역겹	1.924543
732	별점	-1.911017
142	작품	1.821203
1849	평점	-1.794600
121	고생	-1.791385
1838	페미	-1.759145
123	왕오춘	-1.735338
137	공주	-1.731037
153	광어	-1.708277
1624	질	-1.694447
1821	팀	-1.666801

- 웹툰 내용에 대한 부정적 감성
  - 특정인물의 행동에 대한 부정적 감성
- 웹툰 자체에 대한 부정적 감성
  - 스토리가 산으로 간다는 느낌, 작품 소재도 식상, 역겨움을 느낌

05 감성 분석

네이버 웹툰 텍스트 분석

드라마 긍정감성

	토큰	가중치
238	길채현	.650417
1560	중대	1.680470
129	유나	.685073
1282	월요일	1.693061
1137	엄청나	1.725447
1099	앞뒤	1.732145
993	숨기	1.735984
235	기회	1.764320
1639	짠하	1.782037
417	데려오	1.782514
1078	아침	1.846411
1362	인기	1.872829
413	덕화	.890535
176	칼업뎃	.905096
1196	예쁘	1.913493
656	박태준	.923244
725	베뎃	1.997508
1907	할머니	2.053169
578	모바일	2.301643
176	칼업	.548125

- 웹툰 내용에 대한 긍정적 감성
  - 사이다 캐릭터이거나, 노력하는 모습을 보이는 캐릭터
- 웹툰 자체에 대한 긍정적 감성
  - 작가의 웹툰 업로드 속도가 칼같이 빠를 때

과정

1. 토큰화 때 사용했던 함수를 이용해 docs_fan(판타지)과 docs_dra(드라마) 생성
2. docs_fan과 docs_dra를 dictionary 형태로 바꿔 줌
3. Filter_extremes()을 통해 최소 몇 문서에서 나와야 하는지, 몇퍼센트 이하로 나와야 하는지 설정
4. Doc2bow를 통해 bag of words 형태로 바꿔 줌
5. LDA 모델에 적용
6. LDAvis를 통해 시각화

Num_topics

1. 판타지 장르와 드라마 장르 웹툰을 각 10개씩 크롤링했으므로, 최적의 num_topics가 10 주변에 있을 것이라 가정
2. 10 주변으로 log_perplexity(혼란도)와 coherence(응집도)를 비교해가며 최적의 num_topics 선택

Corpus_fan – num_topics = 14
Corpus_dra – num_topics = 14

06 주제 분석

네이버 웹툰 텍스트 분석

판타지

토픽	토픽 1	토픽 2	토픽 3	토픽 4	토픽 5	토픽 6	토픽 7
단어	추이, 무커, 귀신, 소환, 동물	열렙전사, 다크, 특성, 악몽, 게임	성재, 군단장, 군대, 식당, 맛있-	게임, 패치, 컨티뉴, 수호대, 전자오락	밤, 탑, 자하드, 신수, 가문, 슬레이어	다이크, 덴마, 엘, 가이린, 제트	고산, 덴마, 퀑, 우주, 경대
주제	호랑이형님	열렙전사	취사병 전설이 되다	전자오락 수호대	신의 탑	덴마	덴마
토픽	토픽 8	토픽 9	토픽 10	토픽 11	토픽 12	토픽 13	토픽 14
단어	용비, 구휘, 강룡, 지존, 싸움	이령, 산, 호랑이/ 파천신군, 사패천	자하드, 우렉, 공주, 탑, 밤	진모리, 제천대성, 힘, 차력, 신	리즈, 신, 마법, 쿠베라, 인간	발락, 공자, 아비가일, 퀑,	게이머, 한지한, 스토리, 반장,
주제	고수	호랑이형님 +고수	신의 탑	갓 오브 하이스쿨	쿠베라	덴마	더 게이머

06 주제 분석

네이버 웹툰 텍스트 분석

드라마

토픽	토픽 1	토픽 2	토픽 3	토픽 4	토픽 5	토픽 6	토픽 7
단어	소연, 시아, 사귀-, 헤어지-, 연애	태성, 펀치, 작가,	박형석, 크루, 싸움, 바스코	수아, 선임, 부대, 계급, 군대	주영, 자림, 연애, 사귀-, 싸우-	소대, 중수, 전역, 의경,	아영, 동원, 대두, 좋아하-, 빨개지-
주제	연놈	프리드로우	외모지상주의	뷰티풀 군바리	연애혁명	뷰티풀 군바리	평범한 8반
토픽	**토픽 8**	**토픽 9**	**토픽 10**	**토픽 11**	**토픽 12**	**토픽 13**	**토픽 14**
단어	정수아, 고효원, 전설, 후임	찬양, 나리, 승하, 호감, 커플	박태준, 형석, 일진, 정체	준우, 태양, 학교, 성아, 반장	하린, 동까, 박세준, 칼업뎃,	기안, 두치, 초심, 달라지-, 기명	태성이, 민지, 도봉산, 백도화, 하린
주제	뷰티풀 군바리	소녀의 세계	외모지상주의	랜덤채팅의 그녀	프리드로우	복학왕	프리드로우

07 결론

네이버 웹툰 텍스트 분석

- 서론에서 제시했던 문제인 웹툰의 소개 글이 너무 추상적으로 작성되어 있다는 점을 보완하기 위해 웹툰 댓글을 텍스트 분석함
- 단어 빈도 분석을 통해 장르별로 어떤 단어가 많이 있는지 확인
- 감성 분석 통해 웹툰 평점에 긍정적인 요소와 부정적인 요소 확인
- 추상적인 소개 글을 보완할 키워드를 뽑기 위해 주제 분석을 실시
- 판타지와 드라마 장르로 나눠서 LDA 모델 적용
- 각 모델에 맞는 num_topics 찾고 모델 학습 실시
- 각 웹툰 별로 주제분석을 통해 키워드 추출
- 네이버 웹툰 소개 글에 해시태그를 통해 추가한다면, 웹툰 입문자들에게 해당 웹툰의 정보를 더 잘 전달할 수 있을 것임.

소녀의 세계 모랑지

완벽해 보이지만 사실 외로웠던 백조들과 맘씨 착한 오리가 만나
여러 갈등을 함께 겪으며 진짜 친구가 되어가는 소녀들의 찐 우정물

스토리, 드라마 | 전체연령가

#나리
#찬양
#승하
#호감
#커플

이러한 해시태그를 추가함으로써 소녀의 세계 웹툰의 현재 흐름은 주인공인 나리와 옆의 남자애들(좌 승하, 우 찬양)과 호감을 가지고 커플이 되는 스토리임을 예측할 수 있다.

항목	점수	평가 근거
서론	2 / 2	서론을 통해 다루고자 하는 주제, 현황, 문제점 등 파악가능함.
데이터 수집	3 / 3	수업 내용에 더하여 효과적인 방법으로 Selenium+lxml 을 사용함. 현재 연재하고 있는 웹툰 외에 완결 혹은 휴재 중인 웹툰도 크롤링함.
전처리	3 / 3	수업 내용에 더하여 한국어 불용어 사전을 추가했고, 토큰화 함수를 응용하여 개발함.
단어 빈도	3 / 3	주제를 반영하여 효과적인 방식을 추가로 조사함
감성 분석	3 / 3	장르별로 나눠서(판타지, 드라마) 효과적인 분석 및 해석
주제 분석	3 / 3	처음에 세웠던 서론과 관련하여 효과적인 분석 및 해석
결론	2 / 2	서론 및 본론과 부합하는 결론을 제시
합계	19 / 19	